Modeling the dynamics of phenotypic diversity across deep time

Presented in Partial Fulfillment of
the Requirements for the Degree of

Doctor of Philosophy

with a Major in

Bioinformatics and Computational Biology

in the

College of Graduate Studies

University of Idaho

by

Matthew Wesley Pennell

April 2015

Major Professor
Luke J. Harmon, Ph.D.

Committee
Jack Sullivan, Ph.D.
Scott L. Nuismer, Ph.D.
Paul Joyce, Ph.D.
Arne Ø. Mooers, Ph.D.

Department Administrator
Eva M. Top, Ph.D.

Authorization to Submit Dissertation

This dissertation of Matthew Wesley Pennell, submitted for the degree of Doctor of Philosophy with a Major in Bioinformatics and Computational Biology and titled "Modeling the dynamics of phenotypic diversity across deep time," has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: ______________________ ____________
Luke J. Harmon, Ph.D. Date

Committee Members: ______________________ ____________
Jack Sullivan, Ph.D. Date

______________________ ____________
Scott L. Nuismer, Ph.D. Date

______________________ ____________
Paul Joyce, Ph.D. Date

______________________ ____________
Arne Ø. Mooers, Ph.D. Date

Department Administrator: ______________________ ____________
Eva M. Top, Ph.D. Date

Abstract

One of the most enduring puzzles in evolutionary biology is how processes operating at the level of populations (microevolution) scale up to large-scale patterns of diversity (macroevolution). Recent advances in our ability to infer the historical pattern of evolutionary branching—the phylogeny—for many groups of organisms have provided opportunities to gain new perspectives on this question. In this dissertation I develop statistical methods, computational machinery, and theoretical frameworks that will enable researchers to make more meaningful inferences about the processes that have driven diversity through deep time using phylogenetic data.

In my opening chapter, I develop a theoretical foundation for how researchers can use models of trait evolution to test hypotheses related to the long-controversial theory of punctuated equilibrium, which asserts that speciation causes rapid evolution against a backdrop of stasis. I break the hypothesis down into four key elements and argue that combining these conceptually distinct ideas under the single framework of punctuated equilibrium is distracting and confusing, and more likely to hinder progress than to spur it.

Next, I present a suite of statistical software, written in the R programming language, for fitting evolutionary models to phylogenetic data. This is a complete overhaul of the popular GEIGER package designed to facilitate analyses of large and complex comparative datasets.

As an example of how phylogenetic models of trait evolution can provide complimentary insights to population-level models, I investigate the evolution of sex chromosome-autosome fusions. Using discrete character models and a recently compiled database of sexual systems, I find that Y-autosome fusions occur at a much higher rate than X-, Z-, or W-autosome fusions in fish and squamate reptiles. This result grounded a theoretical investigation into the evolutionary forces driving sex chromosome fusions—the phylogenetic results allowed my collaborators and I to exclude from consideration several existing theories for why fusions become fixed in populations. Specifically, we found that the phylogenetic results cannot be accounted for by either direct or sexually antagonistic selection on their own. We argue that the observed patterns can be best explained when chromosomal fusions occur more frequently in males, are slightly deleterious, and are primarily fix by drift.

In the final two chapters, I address two outstanding statistical problems that hinder the use and interpretation of phylogenetic models of trait evolution. First, I develop a novel statistical framework for assessing the absolute fit, or adequacy, of phylogenetic models of trait evolution. To date, researchers have focused almost exclusively on the relative explanatory power of alternative models, rather than the ability of a model to provide a good explanation for data on its own terms. I use my approach to evaluate the statistical performance of commonly used trait models on 337 comparative datasets covering three key functional traits of angiosperms ("flowering plants"). In general, the models I considered often provide poor statistical explanations for the evolution of these traits. This was true for many different groups and at many different scales. Whether such statistical inadequacy will qualitatively alter inferences drawn from comparative datasets will depend on the context. Regardless, assessing model adequacy can provide interesting biological insights—how and why a model fails to describe variation in a dataset gives us clues about what evolutionary processes may have driven trait evolution across time.

Second, I develop a new technique that leverages taxonomic information to assess and overcome sampling biases in trait datasets; such sampling biases are likely prevalent and have the potential to confound both tests of macroevolutionary and macroecological hypothesis. As an example of the utility of this method, I use it to provide the first estimate of the global distribution of woody and herbaceous plants from a database of 39,313 records and find that the world is likely much woodier than researchers thought.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Luke Harmon has been a tremendous mentor, teacher, collaborator, and friend. He came to know me better than I knew myself. He found a way to bring out the best in me and restrain the worst. And he made science fun—there is no one else I would rather make mistakes with.

I thank my committee: Jack Sullivan, for supporting and challenging me in science and in life; Scott Nuismer, for making sure I knew where I stood; Paul Joyce, for always making time when I got stuck on a problem; and Arne Mooers, for introducing me to evolutionary biology and for keeping me honest over the years.

David Tank was practically a second advisor to me. He took me under his wing and taught me how to think and how to be a scientist. And somehow he was able to convince me that plants are actually pretty damn awesome (I will never forget digging up fossils in Clarkia). I am also grateful to Larry Forney, for taking the time to talk to me, and for calling me on my bullshit when I needed to be called on it.

Simon Uribe-Convers, Brice Sarver, Travis Hagey, and Tyler Hether were my brothers in the trenches—we worked late together, argued endlessly about biology and everything else, and kept each other sane during the ups and downs of grad school. I am extremely grateful for their friendship.

There is an old and persistent (not to mention, pernicious) myth that working towards a doctorate is, or at least ought to be, a lonely journey. My own experience could not be farther from this; collaboration and camaraderie have been central to all aspects of my intellectual life. A great deal of this work, and so much of the thought behind it, has sprung from my interactions with other (and, in many cases, young) scientists. In particular, I would like to thank Richard FitzJohn and Josef Uyeda. They challenged me to think harder and to never take the easy way out. I have learned so much from both of them and enjoyed ever minute (or at least, most) of the many that we spent together brainstorming ideas, writing papers, and hacking code. I also thank Jon Eastman for getting off his rocket ship to help me whenever I fell off my tricycle. And David Bapst has been my constant sounding board and sparring partner throughout my dissertation. It is only because of him that I can converse with paleontologists without complete embarrassment. I would also specifically like to thank some of my collaborators and mentors—Michael Alfaro, Steve Arnold, Frank Burbrink, Will Cornwell, Bernie Crespi, Joe Felsenstein, David Green, Paul

Harnik, Mark Kirkpatrick, Craig Miller, Brian O'Meara, Sally Otto, Erica Bree Rosenblum, Carl Simpson, Graham Slater, David Swofford, and Amy Zanne—whose intellectual and personality generosity is deeply humbling.

It has been an incredible privilege, and a whole lot of fun, to do science in the Harmon/RoHa lab. You guys are the best: I thank Joseph Brown, Daniel Caetano, Simone Des Roches, Kayla Hardwick, Denim Jochimsen, Suzanne Joneson, Rafael Maia, Eliot Miller, Tom Poorten, James Rosindell, and Jamie Voyles.

I loved being an IBESTian at the University of Idaho. I am thankful to all the people who made it so fun, both in and out of the office. In particular, I would like to thank Daniel Beck, Lucius Caldwell, Matthieu Delcourt, Anahi Espindola, Michael France, Sam Hunter, Sarah Jacobs, CJ Jenkins, Maribeth Latvis, Wesley Loftie-Eaton, Hannah Marx, Genevieve Metzger, Diego Morales-Briones, Matt Settles, Matt Singer, Chloe Stenkamp-Strahm, and ET Thornquist. I also want to give an extra shoutout to all the members, past and present, of the Phylogenetics Reading Group (PuRGe); I loved arguing with you all for the past five years. I thank all the IBEST staff and faculty, especially Lisha Abendroth and Eva Top, for without their boundless patience, I would literally not have finished my Ph.D.

I thank Roxana Hickey for all the things: for celebrating with me when things were going well; for commiserating when things were not; for listening to me rant even when she didn't care; for being my friend; for suffering me when I was insufferable; and for always pushing me to be better. Last, I would like to thank my family: my parents, Richard and Brenna Pennell, and my siblings, Stephen, Alec, and Laura. I cannot even begin to articulate the support they have given me over the years nor my gratitude for it. For once, I have no words.

Dedication

To my grandparents

Betty and Joseph Pennell

Frieda and Earl Harder

CHAPTER 1

Introduction: Models, meanings, and macroevolution [1]

1.1 Objectives and structure of this dissertation

The primary goal of my dissertation is to improve the statistical and conceptual foundations that underlie phylogenetic tests of macroevolutionary hypotheses. To address the statistical component, I have: led the development of statistical software for fitting models and analyzing data (Chapter 3), which I used to study the macroevolution of sex chromosomes across vertebrates (Chapter 4); created a novel framework for assessing the absolute fit, or adequacy, of models of trait evolution (Chapter 5); and developed a new approach for evaluating and dealing with sampling biases in comparative datasets (Chapter 6), a problem that is becoming increasingly pertinent, if understudied, as researchers rely more on curated collections of data. Additionally, I have been involved in a number of other projects during my Ph.D. (not included in this dissertation) in which I have examined the statistical properties of existing phylogenetic comparative approaches (Pennell *et al.*, 2012; Uyeda *et al.*, 2015), developed new ones (Slater and Pennell, 2014; Cornwell *et al.*, 2014) and applied these to test empirical questions in angiosperms (Cornwell *et al.*, 2014; Tank *et al.*, 2015), ascidians (Maliska *et al.*, 2013), and whales (Slater and Pennell, 2014).

As mentioned above, I have coupled my work in statistical methods with more theoretical work, in which I aimed to provide roadmaps for better interpreting the results of tests using these methods (Rosenblum *et al.*, 2012; Pennell and Harmon, 2013; Pennell *et al.*, 2014b,c; Pennell, 2015). One of these projects, an investigation into tests of the theory of punctuated equilibrium (Eldredge, 1971; Eldredge and Gould, 1972) is included as a chapter in this dissertation (Chapter 2).

As an introduction to my dissertation, I briefly overview the field of comparative biology (for a more comprehensive discussion, see Pennell and Harmon, 2013) and point out what I believe is the biggest challenge facing the field—that we often have a poor understanding of precisely what

[1] This chapter was previously published in a modified form as: Pennell M.W. 2015. Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice—Book Review. Systematic Biology 64:161–163

we are measuring and explaining when we use phylogenetic comparative methods (PCMs)—and highlight possible ways forward.

1.2 OVERVIEW OF PHYLOGENETIC COMPARATIVE METHODS

Investigating similarities and differences across species (the "comparative method") has been an esssential conceptual tool in the study of adaptation since Darwin (1859). Interspecific comparisons are especially valuable when there is little or no variation in the trait of interest within species; in these cases, complimentary approaches such as field experiments are of limited utility. Even when there is variation within a species, the comparative approach allows biologists to assess the generalities of patterns. PCMs for the study of adaptation arose out of the recognition that shared evolutionary history can confound statistical comparisons (Harvey and Pagel, 1991). As a result of the process of descent with modifications, closely related species share many traits and trait combinations and therefore individual species cannot be considered independent observations. In the 1980s and early 1990s, a number of highly influential statistical approaches were developed to incorporate phylogeny into interspecific comparisons (Ridley, 1983; Felsenstein, 1985; Grafen, 1989; Maddison, 1990; Harvey and Pagel, 1991; Lynch, 1991; Pagel, 1994).

While initially controversial (e.g., Westoby *et al.*, 1995), PCMs have gained near universal acceptance in the ensuing decades, such that today, it is near impossible to publish an interspecific study without considering phylogeny. This victory for phylogenetics is so decisive that some researchers have expressed concern that the pendulum has swung too far toward phylogenetic approaches in the study of evolutionary ecology (Losos, 2011). While PCMs are still routinely used to test for adaptation, the field has evolved in subtle yet substantial ways: researchers recognized that the same models developed for comparative questions could also be used to test macroevolutionary questions—for example, what is the pattern of trait change through deep time and what processes drove these trends?—that were long the exclusive domain of paleobiology (Hansen and Martins, 1996; Hansen, 1997; Schluter *et al.*, 1997; Pagel, 1997; Mooers and Schluter, 1998; Pagel, 1999; Mooers *et al.*, 1999). The rate of development of novel PCMs has been incredible and this pace has been matched by the ever-increasing availability of more reliable phylogenetic trees along with large-scale efforts to aggregate phenotypic data from across the Tree of Life.

1.3 CURRENT CHALLENGES IN COMPARATIVE BIOLOGY

Despite the incredible progress of phylogenetic comparative methods over the last few decades, there remain some fundamental issues that are deeply unsettling: while we have sophisticated machinery for fitting many different types of models to comparative data, we often lack a clear interpretation of what exactly they mean. Reading many papers in the field (including my own!), I cannot help but recall a sentiment expressed by Houle *et al.* (2011) in their lucid review of measurement theory and its applications in biology. They criticize statisticians who advocate that data transformations are justifiable whenever they result in distributions that meet the assumptions of a particular analysis: "If that is statistics, we want no part of it, as science is about nature, not numbers" [p. 18]. I argue that our ability to analyze phylogenetic comparative data has outpaced our ability to understand it.

Consider for example, regression models of the form

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

In phylogenetic regression (Grafen, 1989; Lynch, 1991), it is usually assumed that the tree only enters into the model in the error term ϵ such that $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$ where $\mathbf{V}$ is the expected variance-covariance matrix for the traits given an evolutionary model. In other words, the evolutionary model is used to model the structure of the residuals and not the actual traits. Formulating the model in such a way allows us to make use of well-established statistical theory from generalized least squares (GLS) and generalized linear mixed-effects (GLM) models (Lynch, 1991; Rohlf, 2001, 2006; Housworth *et al.*, 2004; Hadfield and Nakagawa, 2010). Including the phylogenetic structure in the error variance is no different from including any other type of covariance. By recognizing this equivalency, we can now fit phylogenetic regression models with a variety of distributions for the response variable Y (Ives and Garland Jr., 2010; Hadfield and Nakagawa, 2010), incorporate measurement error (Ives *et al.*, 2007; Hansen and Bartoszek, 2012), and take advantage of a large number of other standard statistical tricks (see Garamszegi, 2014, for a recent review)

There are a variety of different models one can use to create $\mathbf{V}$. The most popular is to assume that the residuals are distributed according to the expectations of a Brownian motion (BM) model. Indeed, the original independent contrasts method (Felsenstein, 1985) produces identical results to a phylogenetic regression model when this assumption is made (Blomberg *et al.*, 2012). A number of the researchers have advocated that a λ tree transformation (Pagel, 1999; Freckleton *et al.*, 2002, 2011) is often more appropriate than simply assuming BM for constructing the error variance term $\mathbf{V}$. (The λ transformation involves multiplying the off-diagonals of $\mathbf{V}$ by a estimated parameter between 0 and 1.) This is a purely phenomenological construct—by shrinking every branch except those leading to the tips, it implies that there is something special about extant taxa, which is clearly not the case. Nonetheless, researchers often use such models to claim that one trait is adapted to the value of another. In a series of papers, Hansen and colleagues have clearly articulated the problem with such inferences (Hansen and Orzack, 2005; Hansen *et al.*, 2008; Labra *et al.*, 2009; Hansen and Bartoszek, 2012). Effectively, standard regression models assume adaptation to a new environment is instantaneous and that maladaption is phylogenetically structured; closely related species will have similar deviations from the optimal trait value even if the optimum differs between them. From a biological perspective, this seems very odd.

Perhaps even more confusing is the use of Ornstein-Uhlenbeck (OU) models to construct the error variance term. OU is attractive for modeling the residual variance because, unlike the λ transformation, it is a coherent stochastic process and is directly analagous to a population level model from quantitative genetics—quadratic stabilizing selection on a fixed adaptive landscape (Lande, 1976; Hansen and Martins, 1996). While the λ transformation is obviously just a statistical construct, OU *seems* biologically motivated. Indeed, researchers commonly interpret the OU-structured variance term as representing stabilizing selection or constraints. But these does not get around Hansen's criticisms. These models still assume phylogenetically structured maladapation and they do not allow researchers to make specific inference about stabilizing selection or evolutionary constraints—it is completely unclear precisely what is being constrained or how the residuals are under stabilizing selection. OU error structures may often fit data better than BM error structures but it is likely that this is simply because OU can accomodate more variance towards the tips of the phylogeny than a BM model can (including λ has a similar effect). The evolutionary argument here seems merely window dressing for a purely statistical argument.

The arguments I have made here apply equally well to models without predictor variables—where what we want to explain with comparative methods are the distribution of traits through time without considering predictor variables. It is now a common exercise in both phylogenetic comparative biology and paleobiology to compare alternative models of trait evolution and then to interpret the best-fitting model in terms of evolutionary processes (e.g., Mooers *et al.*, 1999; Hunt, 2007; Harmon *et al.*, 2010; Hopkins and Lidgard, 2012; Burbrink *et al.*, 2012; Hunt, 2012)

1.4 PATHS FORWARD

How then are we to make sense of comparative analyses? In my view, there are three possible frameworks with which to think about comparative biology. First, we can take the view that what we are measuring are strictly patterns and that we are not necessarily making inferences about specific evolutionary processes. This is certainly a defensible position: the patterns may be interesting in and of themselves and documenting commonalities and differences among clades and through time may provide a broader picture of the history of life on earth. In practice, this is what researchers are often actually doing, even if they are hesitant to admit this. And since the models we used in comparative biology predict trait distributions that conform to common probability distributions, there are undoubtedly a huge number of processes that could generate the patterns we observe (Jaynes, 2003; Frank, 2009, 2014). A benefit of openly adopting this perspecitve is that we can consider a much broader suite of models that may provide a much better fit to our data and predictive power than current models—if we are not interested in making specific evolutionary inferences, then we need not be beholden to specific evolutionary models. Such alternatives may include macroevolutionary diffusion processes (e.g., Clauset and Erwin, 2008), models derived from macroecological theories (Brown *et al.*, 2004; Harte, 2011) or making use of statistical learning approaches divorced from any process whatsoever.

The second framework is the quantitative genetics view: the models we fit in comparative biology should be taken as literally representing microevolutionary hypotheses. Many of the commonly used models can be directly interpreted in terms of population-level parameters (Hansen and Martins, 1996; Pennell and Harmon, 2013). We can compare the estimated model parameters to within-population measures to test if macroevolutionary divergences are consistent with evolution by drift, stabilizng selection, etc. This project is certainly interesting and worth pursuing. But given the results of studies that have explicitly examined this connection (Lynch, 1990; Estes and Arnold, 2007; Hohenlohe and Arnold, 2008; Harmon *et al.*, 2010; Bolstad *et al.*, 2014) using rather simple models, it appears that translating the parameters estimated from comparative data to the terms of quantitative genetics (e.g., if we assume that BM is strictly a model of drift with fixed additive genetic variance **G**, the estimated rate parameter σ^2 is equal to **G** divided by the effective population size N_e ; Lande 1976) will often result in nonsensical numbers.

The third perspective is to take seriously the idea that macroevolutionary models reflect the dynamics of adaptive landscapes through deep time (Arnold *et al.*, 2001; Hansen, 2012; Pennell *et al.*, 2014b). Comparative biologists have a tendency to discuss many of these ideas in scare quotes. The optimum of OU models is referred to as "clade level optimum". A model with decelerating rates of change depicts an "early burst". I argue that a much richer and more meaningful connection can potentially be made. Theoretical work over the last century has produced a beautiful and fairly comprehesive understanding of how populations move across adaptive landscapes and empricists have tested the theoretical predictions in a wide variety of systems and contexts. In contrast, we have only a preliminary understanding of how the landscapes themselves evolve on longer time scales. This is a fundamentally important question in evolutionary biology and one which I believe, phylogenetic comparative biology and paleobiology can help address.

There is a lot of work to be done before we will really able to get at these types of questions. Once we recognize that some of the classic concepts in evolutionary biology—such as adaptive zones, adaptive radiations and key innovations—are actually hypotheses about the structure and dynamics of adaptive landscapes (Hansen, 2012), we can start developing statistical models that actually capture their essential properties. Current models are, at best, loosely tied to these ideas (hence the scare quotes). Additionally, there are a number of exisiting mathematical frameworks that make predictions about these higher order processes and trait evolution over longer time periods (see for example, Holt *et al.*, 2003; Gavrilets, 2004; Doebeli, 2011). But there is currently no way to estimate the relevant parameters of these models from comparative data.

1.5 CONCLUDING REMARKS

Both the development of new PCMs and the interest in using them has grown tremendously over the last decade. Nevertheless, I feel that we, as a field, are somewhat stuck. First, the same handful of statistical models are employed over and over again with most of the progress representing relatively minor variations on similar themes. (That is not the say that such improvements are not challenging or worthwhile; indeed a portion of my dissertation is aimed in precisely this direction.) Second, we are often much too vague on what exactly we want to explain with PCMs. I argue that these two problems are deeply intertwined. The standard collection of models available today, namely those based on BM and OU, have had such staying power in part because they can be useful for detecting patterns, can be interpreted in light of evolutionary genetics and can loosely be tied to questions about adaptive landscapes. Requiring this sort of conceptual flexibility is also a limitation. More focused, question-specific approaches to modeling that are directly tied to the inferences we actually want to make will likely get us much further than sticking to models that are more general but address no questions particularly well.

CHAPTER 2
Revisiting the punctuated equilibrium debate in light of emerging phylogenetic data and methods[2]

2.1 SUMMARY

The long-controversial theory of punctuated equlibrium (PE) asserts that speciation causes rapid evolution against a backdrop of stasis. PE is currently undergoing a resurgence driven by new developments in statistical methods. However, we argue that PE is actually a tangle of four unnecessarily conflated questions: i) is evolution gradualistic or pulsed?; ii) does trait evolution occur mainly at speciation or within a lineage?; iii) are changes at speciation adaptive or neutral?; and iv) how important is species selection in shaping patterns of diversity? We discuss progress towards answering these four questions but argue that combining these conceptually distinct ideas under the single framework of PE is distracting and confusing, and more likely to hinder progress than to spur it.

2.2 INTRODUCTION: THE RESURGENCE OF PUNCTUATED EQUILIBRIUM

The following three quotations were all drawn from abstracts of recent papers purporting to use statistical models to empirically evaluate punctuated equilibrium (PE):

> A long-standing debate in evolutionary biology concerns whether species diverge gradually through time or by punctuational episodes at the time of speciation. We found that approximately 22% of substitutional changes at the DNA level can be attributed to punctuational evolution, and the remainder accumulates from background gradual divergence. (Pagel *et al.*, 2006, p. 119)

> This controversy, widely known as the 'punctuated equilibrium' debate, remained unresolved, largely owing to the difficulty of distinguishing biological species from fossil remains. We analyzed body masses of 2143 existing mammal species on a

[2] Previously published as: Pennell M.W., Harmon L.J., and Uyeda J.C. 2014. Is there room for punctuated equilibrium in macroevolution? Trends in Ecology & Evolution 29:23–32

> phylogeny comprising 4510 (i.e., nearly all) extant species to estimate rates of gradual (anagenetic) and speciational (cladogenetic) evolution. (Mattila and Bokma, 2008, p. 2195)

> Under such processes, observations at the tips of a phylogenetic tree have a multivariate Gaussian distribution, which may lead to suboptimal model specification under certain evolutionary conditions, as supposed in models of punctuated equilibrium or adaptive radiation. (Landis *et al.*, 2013, p. 193)

These three papers are representative of a substantial number of other recent high-profile studies that have discussed their research in the context of PE (e.g., Bokma, 2002, 2008; Webster *et al.*, 2003; Hunt, 2007, 2008, 2012; Hunt *et al.*, 2008; Atkinson *et al.*, 2008; Ingram, 2011; Uyeda *et al.*, 2011; Rabosky, 2012; Rabosky *et al.*, 2013; Simpson, 2013; Baca *et al.*, 2013; Bartoszek, 2014). This is somewhat remarkable given that arguably no idea has had such a turbulent history in modern evolutionary thought as punctuated equilibrium. In the early 1970s, Eldredge and Gould (Eldredge, 1971; Eldredge and Gould, 1972; Gould and Eldredge, 1977) proposed that the predominant pattern of evolution throughout deep time is that of stasis "punctuated" by brief intervals of rapid evolution, which often occurred during speciation events. This was originally conceived as a way of bridging the gap between prevailing ideas about speciation, i.e., Mayr's allopatric model (1942), and observations from the fossil record (Sepkoski, 2012). However, PE has expanded and shifted in definition to become a much more far-reaching hypothesis to many researchers. Consequently, it has been viewed as both a a rather innocuous statement about the general patterns found in the fossil record and as an affront to the central tenets of evolutionary theory (Stanley, 1975, 1979; Gould, 1980; Charlesworth *et al.*, 1982; Levinton, 2001). For some researchers, the stakes of the debate over the prevalence of PE could not have been higher:

> If most evolutionary changes occurs during speciation events, and if speciation events are largely random, natural selection, long viewed as the process guiding evolutionary change, cannot play a significant role in determining the overall course of evolution. Macroevolution is decoupled from microevolution... (Stanley, 1975, p. 648)

In the wake of such claims, much of the intellectual history of PE was been characterized by fierce, and often vitriolic, theoretical debates—exhaustively catalogued in (Levinton, 2001; Gould, 2002)—and the theory remains divisive (for more on the history of the idea, see Sepkoski, 2012).

The field of macroevolution has recently witnessed a resurgence of interest in PE as paleobiologists and, increasingly, comparative biologists armed with molecular phylogenies, have applied sophisticated statistical models to quantitatively test the major hypotheses of PE. In this review, we ask whether these new statistical advances have "rescued" PE from intellectual extinction. We answer this question in the negative. The challenges inherent in elucidating macroevolutionary processes and patterns from paleontological and comparative data are only exacerbated by the muddled historical legacy of PE. Although a number of studies have indeed discussed their findings in light of PE, they have actually addressed a wide variety of conceptual issues; the studies from which we have quoted above exemplify this—each one asks a fundamentally distinct question.

What then, exactly, defines PE? The central definitions and concepts of PE have shifted substantially over time, including the views of the theory's chief advocates (for analysis, see Mayr, 1982; Ruse, 1989; Sepkoski, 2012). We argue that the key to disentangling this Gouldian knot, lies not in attempting to parse the literature in search of the true "essence" of PE, but rather in recognizing that the myriad concepts often associated with the theory can be conceptually dissociated and evaluated independently. We believe that dissociating the different components of PE will lead to a more productive discussion of these ideas and facilitate progress in some of the most fundamental questions in macroevolution. In this essay, we identify four key questions that have been lumped under the topic of PE, discuss how their association with each other has led to confusion, and comment on recent methodological developments, using a variety of types of data, that may provide novel insights into large-scale patterns of diversity.

2.3 PUNCTUATED EQUILIBRIUM AS A CONGLOMERATE OF CONCEPTS

In our view, the theory of PE, and the extensive discussion surrounding it, conflates four separate primary research questions: i) what is the relative importance of gradualistic versus pulsed evolution?; ii) what is the role of speciational events (cladogenesis) vesus within lineage evolution (anagensis) in generating trait divergence?; iii) when change is cladogenetic, are these changes adaptive or driven by neutral processes?; and iv) how important is higher level selection (species selection) in shaping patterns of diversity?

2.3.1 *Gradualistic versus pulsed evolution*

In principle, it is quite feasible to distinguish gradualistic versus pulsed evolution using either phylogenetic comparative or paleontological data. Constant-rate gradualism is typically modeled as a random walk or Brownian motion process (BM; see BOX 1) in both phylogenetic and paleobiological studies. Several recent studies have examined whether fossil time-series conform better to predictions from constant-rate BM, phenotypic stasis, or directional evolution, which each predict different distributions of trait values through time and can be distinguished using model selection techniques (Hunt, 2012). These studies have found mixed support for each mode of evolution in different lineages and traits (Hunt, 2007; Hunt *et al.*, 2008; Hunt, 2008; Grey *et al.*, 2008; Hopkins and Lidgard, 2012; Hunt, 2012). An exceptional demonstration of pulsed evolution in the fossil record was examined by Hunt *et al.* (2008), who found support for a rapid pulse of evolution in sticklebacks as they colonized a novel adaptive peak. Similar model-fitting approaches have been used to demonstrate that in particular fossil time-series shifts in the mode of evolution (i.e., directional evolution, stasis, or BM) are separated by phenotypic bursts (Hunt, 2008). This pulsed pattern of evolution is supported by large collections of micro- and macroevolutionary data (Estes and Arnold, 2007; Uyeda *et al.*, 2011). Studies of fossil time-series often include a number of caveats that may complicate inference of evolutionary modes. These include unequal sampling probabilities and uncertain stratigraphic position, as well as issues relating to range shifts, time-averaging, and phenotypic plasticity (Patzkowsky and Holland, 2012). A particularly promising, albeit data-intensive, method developed by Hannisdal (2007) incorporates some of these additional sources of uncertainty in a Bayesian framework.

While fossil time-series provide direct observations of phenotypes over time, they are limited by the difficulty in confidently assembling sequences of ancestor-descendant relationships. Phylogenetic comparative methods provide a complementary means to study departures from constant-rate gradualism; these can be applied to both extant and extinct data, if the fossil data can be placed in a phylogenetic context (Pennell and Harmon, 2013). Several methods allow the detection of rate shifts across clades by allowing the BM rate parameter σ^2 to differ across branches of the phylogeny (e.g., O'Meara *et al.*, 2006; Eastman *et al.*, 2011; Slater *et al.*, 2012b). However, these methods model sustained shifts in evolutionary rates, rather than pulsed patterns suggested by PE. Pure-burst models, in which all change accumulates in pulses, can also be fit to phylogenies and more closely align with PE (Hansen and Martins, 1996; Khaitovich *et al.*, 2005; Uyeda *et al.*, 2011). Landis *et al.* (2013) modeled both gradual and punctuational patterns of evolution using jump-diffusion models, in which both jumps and gradual evolution come from a single, long-tailed distribution (see also, Eastman *et al.*, 2013b). In addition to these methods, discrete shifts in adaptive optima separated by long periods of stabilizing selection have been extensively implemented in in phylogenetic comparative methods by using Ornstein-Uhlenbeck (OU) models (Felsenstein, 1988; Hansen, 1997; Butler and King, 2004). OU models are attractive alternatives to BM that can incorporate stasis, stabilizing selection and adaptive hypotheses (Pennell and Harmon, 2013, see BOX 1 for further details).

Advances in quantitative model-fitting of evolutionary processes have allowed us to explore much wider range of evolutionary hypotheses and processes than simple BM (Pennell and Harmon, 2013), including pulsed evolutionary patterns. However, to what extent does the framework of punctuated equilibrium contribute to interpretation of the results of these models? Note that none of the models described in this section can distinguish between cladogenetic or anagenetic change (see next section). Furthermore, PE is tied to a very specific pattern of evolution and a specific temporal frame: stasis over the lifespan of a species—typically millions of years—followed by geologically brief bursts of phenotypic evolution occurring at speciation (Eldredge and Gould, 1972; Gould and Eldredge, 1977; Gould, 2002). Therefore, even robust support for a pattern of pulsed evolution, represented by shifts in trait values along branches that are not accounted for by gradual evolution, may be incompatible with PE if the pulses occur too infrequently for conventional PE theory, which predicts pulses at all, or nearly all, speciation events. In addition, exactly as paleontologists have long recognized that repeated burst-stasis episodes can appear gradualistic if viewed at too coarse a scale, gradualistic evolution with variable rates can appear pulse-like at the same coarse scale. A pulsed pattern detected from phylogenies, which typically have much longer timescales and coarser sampling than a fossil time-series, may not reflect phenotypic bursts between species. Instead, model-fits may reflect "jumps" between higher-level niche space or adaptive zones, within which whole clades or groups of species may cluster (Simpson, 1944; Hansen, 1997, 2012; Eastman *et al.*, 2013b). The observation that groups of species cluster around different phenotypic optima says nothing about whether individual lineages exhibit a pattern of stasis and phenotypic bursts of evolution over the lifespan of individual species. Tying patterns measured at phylogenetic scales to species-level and not clade-level change is fraught with difficulty. However, we can still address other interesting macroevolutionary questions such as whether evolution is characterized by pulses, how often they occur and what ecological factors may be associated with them (Eastman *et al.*, 2013b).

2.3.2 *Anagenetic versus cladogenetic change*

Although speciation is undoubtedly associated with genetic and trait divergence (Nosil, 2012, and references therein), its relative importance compared to evolutionary change within a lineage is currently poorly understood. Several studies have attempted to evaluate the contribution of cladogenetic change to trait evolution (Wagner and Erwin, 1995; Jackson and Cheetham, 1999; Aze *et al.*, 2011; Strotz and Allen, 2013) using paleontological data. This is evaluated by determining whether the stratigraphic ranges of descendant species overlap with their progenitor species, indicative of coexistence and cladogenesis. However, robustly distinguishing between cladogenetic and anagenetic changes using fossil data crucially depends on several assumptions, such as the accurate reconstruction of ancestor-descendant relationships, the equivalency of species concepts applied to fossil and extant taxa, the robust estimation of species' temporal ranges and enough sampling to eliminate the possibility of gradual evolution (Jackson and Cheetham, 1999). Disputes over the validity of these assumptions have been well played out in the punctuated equilibrium literature (e.g., the Turkana Basin molluscs, Williamson, 1981; Fryer *et al.*, 1983; Van Bocxlaer *et al.*, 2008). Approaches have been developed to account for potential biases, such as estimating ancestor-descendant relationships (Marshall, 1995; Foote, 1996) and stratigraphic ranges (Marshall, 1990, 1994, 1997; Wagner, 2000), accounting for sampling, but difficulties remain. As a recent example, Strotz and Allen (2013) found a predominance of cladogenetic change among fossil Foraminifera, using assumed ancestor-descendant relationships assembled from stratigraphic and phenotypic data (Aze *et al.*, 2011). We view such claims with considerable skepticism because it is impossible to detect cryptic speciation—which is increasingly being inferred in extant groups (Fujita *et al.*, 2012)—in the fossil record, and therefore distinguish decisively between anagenetic and cladogenetic change.

Another tactic to assess the relative contributions of anagenetic and cladogenetic change has been to look for correlations between speciation rates (or species richness, as a proxy for speciation rates) and rates of evolution using phylogenetic comparative data. This has been done by fitting regression models between inferred lineage-specific rates of evolution and diversification.

Several studies have demonstrated such a correlation using a variety of characters, including genetic changes (Webster *et al.*, 2003; Pagel *et al.*, 2006; Venditti and Pagel, 2010), morphological traits (Ricklefs, 2004; Adams *et al.*, 2009; Rabosky and Adams, 2012; Rabosky *et al.*, 2013) and linguistic characters (Atkinson *et al.*, 2008). However, demonstrating a correlation between speciation rates and trait evolution does not demonstrate that the actual speciation events themselves are associated with evolutionary change. For example, higher rates of speciation and trait evolution might both be driven by a common cause (Rabosky, 2012, see below, BOX 2).

A more promising avenue for partitioning out the influence of cladogenetic versus anagenetic change is to use statistical models, which explicitly parameterize both of these components and simultaneously estimate them using maximum likelihood or Bayesian inference. Early forms of such models assumed that all speciation events were captured by the reconstructed phylogeny. These methods partition the variance in trait values between the speciation events and the background evolution occurring within a lineage (Pagel, 1997; Mooers *et al.*, 1999; Bokma, 2002; Wagner, 2000; Wagner and Marcot, 2010). More sophisticated approaches attempt to simultaneously model the diversification process together with trait evolution (Bokma, 2002; Mattila and Bokma, 2008; Bokma, 2008, 2010; Goldberg and Igić, 2012; Magnuson-Ford and Otto, 2012; Simpson, 2013, see BOX 3 for details) to account for the fact that extinction has erased many of the speciation events in the inferred phylogeny (Nee *et al.*, 1992, 1994; Nee, 2006; Ricklefs, 2007). Such model-based approaches are not without caveats. Importantly, violations of simplifying assumptions may strongly affect inferences, and methods to evaluate model adequacy are sorely needed. Furthermore, unaccounted measurement error may be erroneously folded into estimates of cladogenetic change. In fact, we should expect samples of recently diverged species to differ substantially regardless of whether evolution is punctuational or gradual—even after accounting for simple forms of sampling error—due to within-lineage processes such as local adaptation (Uyeda *et al.*, 2011; Hansen, 2012). These processes may or may not be important for macroevolutionary patterns (Futuyma, 1987, 2010), and are difficult to model using current comparative methods (Stone *et al.*, 2011).

Even when speciation is inferred to be associated with divergence, a broader conceptual issue remains: what are the causal mechanisms that could generate such an association against a general backdrop of apparent stasis (Benton and Pearson, 2001; Eldredge *et al.*, 2005)? Speciation has long been thought of as a major driver of phenotypic change, both in the context of PE and in evolutionary biology more broadly (Sætre, 2013). In their original conception of PE, Eldredge and Gould (Eldredge, 1971; Eldredge and Gould, 1972) viewed the pattern posited by PE as a consequence of Mayr's (1942) model of allopatric speciation; as such, speciation is considered a mechanism that interrupts stasis (Futuyma, 1987, 2010). However, the causes of stasis in macroevolutionary data are still unclear (Hansen and Houle, 2004; Estes and Arnold, 2007; Walsh and Blows, 2009; Futuyma, 2010). In particular, the direction of causality cannot be elucidated from the statistical methods we have described. Alternative explanations remain such that trait evolution often generates reproductively isolated lineages. Regardless, it is important to recognize that a central tenet of PE theory—that speciation causally leads to phenotypic evolution—remains impossible to evaluate from either phylogenetic comparative or paleontological data.

2.3.3 *Adaptive versus neutral evolution at speciation*

One of the most contentious ideas surrounding PE is that changes associated with speciation are random or neutral; this is what led Gould, Stanley and others to claim that macro- and microevolution were effectively "decoupled". There are actually two specific versions of this question and these, similarly to many of the ideas we discuss throughout the paper, have often been conflated. The first version is that the changes that occur are random with respect to the direction of a macroevolutionary trend. This is referred to as "Wright's rule" in the paleobiology literature (Gould and Eldredge, 1977) and has been evaluated by testing whether trait differences between ancestors and descendants are directionally biased. More precisely, researchers have tested whether the mean of the distribution of changes in significantly different than zero, the null expectation under most models of trait evolution (Wagner, 1996, 2001). For example, if there is a trend of increasing body size throughout the history of a clade (Cope's rule), then Wright's rule requires that daughter species, at speciation, are on average no bigger (or smaller) than their ancestor. To Gould and Eldredge (1977), as well as to other researchers (for example, Stanley, 1975, 1979), Wright's rule was a key justification for including species selection in a PE framework; if change only occurs at speciation and that change is random with respect to macroevolutionary trends, those trends can only be explained by species selection. However, if we recognize that the nature of change at speciation is independent of species selection (see below), then establishing Wright's rule has no bearing on the strength of higher level selection (Simpson, 2013). At the same time, biased transmission may still be involved in macroevolutionary trends (McShea, 1994, 1998; Wagner, 1996).

The second, broader version of the claim is that change at speciation is driven by neutral processes rather than adaptive evolution (Stanley, 1979; Gould, 1980, 2002). This is much more complex to address. There have been several attempts to investigate the hypothesis that past trait changes were adaptive using phylogenetic comparative or paleobiological data (see for example, Rose and Lauder, 1996). For example, phylogenetic methods can attempt to associate trait changes with changes in the selective regimes experienced by those lineages (Baum and Larson, 1991; Butler and King, 2004; Beaulieu *et al.*, 2012), or studies of functional morphology can provide specific hypotheses about relationships between trait states and the environment, which can then be tested statistically (Wainwright, 2007). However, these necessarily rely on either detailed information about form, function, and the environment (e.g., Vermeij, 1987, and references therein) or researcher's *a priori* hypotheses regarding what was adaptive at some period in the past. We know from studies of wild populations that the direction of selection is often temporally and spatially variable (Grant and Grant, 2002; Siepielski *et al.*, 2009, 2011) and it is therefore extremely tenuous to draw conclusions regarding the adaptive value of changes during speciation from comparative or paleontological data alone.

There has been a great deal of study investigating the patterns of evolution throughout the course of speciation using natural populations, experimental systems and mathematical models (Schluter, 2000; Coyne and Orr, 2004; Gavrilets, 2004; Rundle and Nosil, 2005; Doebeli, 2011; Nosil, 2012, and references within). In particular, many recent studies have explored the distinction between ecological speciation, where speciation is driven by divergent natural selection between lineages, and other forms of speciation (e.g., Bateson-Dobzhansky-Muller incompatibilities, speciation driven by sexual selection, etc.; reviewed in Nosil, 2012). As a result of these studies, we have learned a great deal about the mechanisms involved in speciation and are beginning to understand the relative importance of adaptive and neutral processes during speciation across a broad suite of taxa, although it is much too early to draw any sweeping conclusions. We strongly suggest that this avenue of research is far more appropriate for addressing this aspect of PE than analyzing either phylogenetic comparative or paleontological data alone.

2.3.4 *Species selection as a macroevolutionary process*

Though long controversial in its own right (FitzJohn, 2012b), the idea that natural selection can act on species-level characteristics is becoming more widely appreciated (Coyne and Orr, 2004; Jablonski, 2008; Rabosky and McCune, 2010; FitzJohn, 2012b). Here we follow the lead of other authors (Williams, 1992; Coyne and Orr, 2004; Rabosky and McCune, 2010) and define species selection as "repeatable effects of that trait on the rate of diversification of species possessing it" (Coyne and Orr, 2004, p. 444), regardless of whether or not the trait is an emergent property of the lineage or the aggregate of individual-level traits. We therefore ignore the (in our view, unnecessary) distinction between "species selection" and "species sorting" (*sensu* Vrba and Gould, 1986). For an alternative perspective on this issue, see Jablonski's excellent review (2008) of the topic.

The idea that the tempo and mode of evolutionary change is inexorably linked to selection at the lineage level is an old and persistent one and is, in the minds of at least some researchers, part and parcel of a broader macroevolutionary theory (Stanley, 1975, 1979; Gould and Eldredge, 1977; Gould, 1980; Charlesworth *et al.*, 1982; Dennett, 1995; Levinton, 2001; Gould, 2002). The reasoning behind this is that, in some researcher's conception of the process, selection can only act on "evolutionary individuals" (Hull, 1980) and species can only operate as such if they have a definite beginning and end (Gould and Eldredge, 1977; Gould, 2002)—a pattern that is produced if evolutionary change only occurs at cladogenesis. Although this may seem intuitive, such an association is logically and mathematically unnecessary. Species selection does not require any particular mode of evolutionary change and it certainly does not require the majority of change to be concentrated at speciational events (Van Valen, 1975; Bookstein *et al.*, 1978; Slatkin, 1981; Arnold and Fristrup, 1982; Rice, 1995; McShea, 2004; Rice, 2004; Okasha, 2006; Jablonski, 2008; Simpson, 2013). The conflation of species selection with punctuated change has been cited by some authors to be a cause of antagonism towards species selection (Turner, 2010; FitzJohn, 2012b).

Species selection has been recently reviewed in depth (Jablonski, 2008; Rabosky and McCune, 2010) and we will not attempt to be comprehensive here. Instead, we focus on recent methodological developments that have improved our ability to detect species selection. Conventionally, inference regarding the influence of a trait on diversification rate from molecular phylogenies has been carried out by comparing the diversities or diversification rate between independent pairs of sister taxa (Mitter *et al.*, 1988; Sargent, 2004; Vamosi and Vamosi, 2004; Rabosky and McCune, 2010). However, this is problematic for several reasons including statistical power (Slowinski and Guyer, 1989, 1993; Vamosi and Vamosi, 2005) and that asymmetries in character transition rates can confound asymmetries in diversification rate, and *vise versa* (Maddison, 2006). A major innovation to simultaneously deal with this issue and investigate the correlation between traits and speciation and extinction was made by Maddison *et al.* (2007), with their Binary State Speciation and Extinction, or BISSE, model (see BOX 3). This has been extended beyond binary traits to investigate the effect of multiple discrete traits (MUSSE; FitzJohn, 2012a), quantitative traits (QUASSE; FitzJohn, 2010) and geographic range (GEOSSE; Goldberg *et al.*, 2011) on lineage diversification.

Although these are certainly very promising statistical approaches, they rely on some large sample sizes and potentially dubious assumptions, such as that diversification can be modeled as constant-rate branching process (i.e., a "birth-death" model; Kendall, 1948), that rates of evolution are constant across the phylogeny, and that the directionality and strength of species selection is consistent. There is substantial evidence that suggests that diversification rates are not constant through time or across clades (Rabosky *et al.*, 2007; McPeek, 2008; Phillimore and Price, 2008; Alfaro *et al.*, 2009; Rabosky, 2012; Rabosky *et al.*, 2013), perhaps owing to diversity-dependent diversification (Sepkoski, 1984; Alroy, 2008; Rabosky, 2009), a focus of much modeling work in both paleobiology (Roy, 1996; Eble, 2000; Sepkoski *et al.*, 2000) and phylogenetic comparative methods (Rabosky and Lovette, 2008; Etienne *et al.*, 2012; Etienne and Haegeman, 2012). Similarly, rates of trait evolution are likely often quite heterogeneous (Eastman *et al.*, 2011; Beaulieu *et al.*, 2013) and the vector of species selection has been inferred to be variable in some groups (Jablonski, 1986; Simpson, 2010; Harnik *et al.*, 2012). Some of these assumptions can potentially be relaxed (Rabosky and Glor, 2010) but in general, the robustness of these methods to severe violations awaits further investigation.

In paleontological research, there has been increasing development of multivariate methods to partition out the effect of various correlated traits on speciation and or extinction, which is key to elucidating causal mechanisms. This has been accomplished using statistical techniques such as general linear models (i.e., predict lineages' diversification rates or durations in the fossil record from lineage-specific traits; Jablonski and Hunt, 2006; Harnik, 2011; Harnik *et al.*, 2012). Alternatively researchers have used the Price Equation (Price, 1972; Rice, 2004; Okasha, 2006; Frank, 2012) to examine the covariance between traits and diversification rates. The Price equation was first proposed for the purposes of studying macroevolution by Arnold and Fristrup (Arnold and Fristrup, 1982), and this has recently been expanded upon by Simpson and colleagues (Simpson and Harnik, 2009; Simpson, 2010). Adopting the Price equation also allows for the possibility of a unified approach to the study of species selection across data-types that could potentially be applied to both phylogenetic comparative data and fossil time-series (Jablonski, 2008; Simpson, 2013).

2.4 CONCLUDING REMARKS

We have described quantitative approaches to addressing four fundamental macroevolutionary questions that have long been conflated with each other in the literature on PE. Confusion among these disparate and independent questions has led many researchers to consider PE as being robustly verified, whereas others believe the theory bankrupt of empirical support. Either view may be justified depending on which component an individual researcher considers the essence of PE theory. If macroevolutionary researchers dissociate these concepts, the fact that some may be more difficult to evaluate or are less theoretically sound should not impede progress on other questions.

Although we argue throughout that the questions that make up PE can be addressed independently, this does not preclude synthesis. Instead, multiple processes could be important (and often, probably are) to understanding the accumulation of diversity and disparity through deep time. For example, Goldberg *et al.* (2010) used the BISSE model to demonstrate that species selection was important for maintaining self-incompatibility in the plant family Solanaceae (the "nightshade" family). In a subsequent paper, Goldberg and Igić (2012), re-analyzed the same data but used a model that allowed for trait evolution within a lineage, species selection, and (additionally) trait evolution occurring at cladogenesis (see BOX 3 for details). They found that all three processes appear to be important in this group. Nonetheless, these are independent processes that may or may not be linked mechanistically, in this group or others.

Instead of bringing new insight into PE—and thereby rescuing the term from its historical problems—novel developments have demonstrated that the terminology associated with PE can be problematic. We believe that emerging statistical models and datasets are best suited for testing independent components of PE theory. Evaluation of these methods in the context of PE will only lead to confusion. Although PE undoubtedly served as a catalyst in the development of concepts and methods discussed above, we think it is time to move on, and encourage researchers in macroevolution to look forward rather than look back.

Paleontologically oriented readers may view our discussion of PE as being too harsh on theoretical constructs from their disciplines—and, admittedly, all of us were raised in the traditions of population biology and evolutionary genetics. However, our reading of much of the literature using phylogenetic comparative methods, we have found a recurrent theme of comparative biologists adopting concepts from the paleobiological literature (including, but not limited to PE), but doing so rather blithely. Although it is widely recognized that incorporating fossil data into comparative studies will dramatically improve the inferences we can draw from them (Quental and Marshall, 2010; Slater *et al.*, 2012a; Pennell and Harmon, 2013; Fritz *et al.*, 2013), in our opinion, concordant attention has not been paid to the conceptual foundations which underlie the studies. Comparative biologists have much to gain by engaging more seriously with the arguments and ideas from the rich literature in paleontology on rates of evolution, macroevolutionary trends, species selection, adaptive radiations, and so forth. A truly synthetic macroevolutionary research programme will involve the melding of data and theory from different disciplines, and a thoughtful examination of precisely what the fundamental questions are and how we can go about answering them.

2.5 BOX 1: MODELING TRAIT EVOLUTION

The same basic set of stochastic models are often fit to both fossil time-series and phylogenetic comparative methods. Phyletic gradualism is formulated statistically as constant-rate BM. This model describes a continuous-time random-walk in which the amount of phenotypic change in the population trait mean ($\bar{z}$) over time-interval t is:

$$\Delta\bar{z} = \sigma dW \tag{2.1}$$

where dW is a non-directional diffusion process with mean 0 and variance t. Because change over each time interval is independent of previous time intervals (i.e., the process is Markovian), the amount of variance among replicate lineages increases linearly through time such that $\mathrm{Var}(\bar{z}) = \sigma^2 t$ (Figure 2.1). The covariance between observations is proportional to the shared evolutionary history of samples, which for comparative methods is provided by the phylogeny.

To model discontinuous processes, a shift location is estimated either on a fossil timeseries or a phylogeny. For pulsed models, a shift corresponds to a burst in phenotypic evolution, against a background of a single, constant-rate BM parameter (σ^2). Multiple bursts can be modeled, for example, as a compound Poisson point process, in which bursts occur stochastically at exponentially distributed time-intervals at rate λt and magnitudes drawn from a normal distribution with parameters ($\mu_{\text{burst}}, \sigma^2_{\text{burst}}$) (see Figure 2.1). However, several comparative methods do not model bursts, but instead fit different parameters or models on either side of the shift, corresponding to either an increased or decreased rate of evolution (O'Meara *et al.*, 2006; Hunt, 2008; Eastman *et al.*, 2011). Thus, for k shifts, there would be $k + 1$ BM rate parameters, ($\sigma^2_1, \ldots, \sigma^2_{k+1}$). These models can also be combined to jointly model both bursts and rate shifts (Eastman *et al.*, 2013b).

BM models predict that divergence can increase without bounds, which is unrealistic under adaptive scenarios of trait evolution or under models of stasis, where traits are expected to evolve around adaptive optima. A simple extension of a BM model is an OU model of trait evolution. The per unit time change in mean phenotype under this model is:

$$\Delta\bar{z} = -\alpha(\bar{z} - \theta) + \sigma dW \tag{2.2}$$

where σdW is identical to a BM process and contributes stochastically to divergence, θ is the optimum trait value, and α is a "pull" parameter that governs how strongly the population mean is pulled toward θ. Thus, divergence is a balance between the stochastic diffusion parameter (σ^2) and the deterministic pull parameter (α) toward the optimum value (θ) (Figure 2.1). As with BM models, discontinuous OU models can allow for shifts in rate parameters (α, σ^2), which has been implemented in a phylogenetic context (Beaulieu *et al.*, 2012). Rapid shifts in optima are more naturally included in OU models via shifts in the θ parameter, and have been used extensively in phylogenetic comparative methods (Hansen, 1997; Butler and King, 2004; Beaulieu *et al.*, 2012). Population trait means approach a new optimum at a rate proportional to the strength of α.

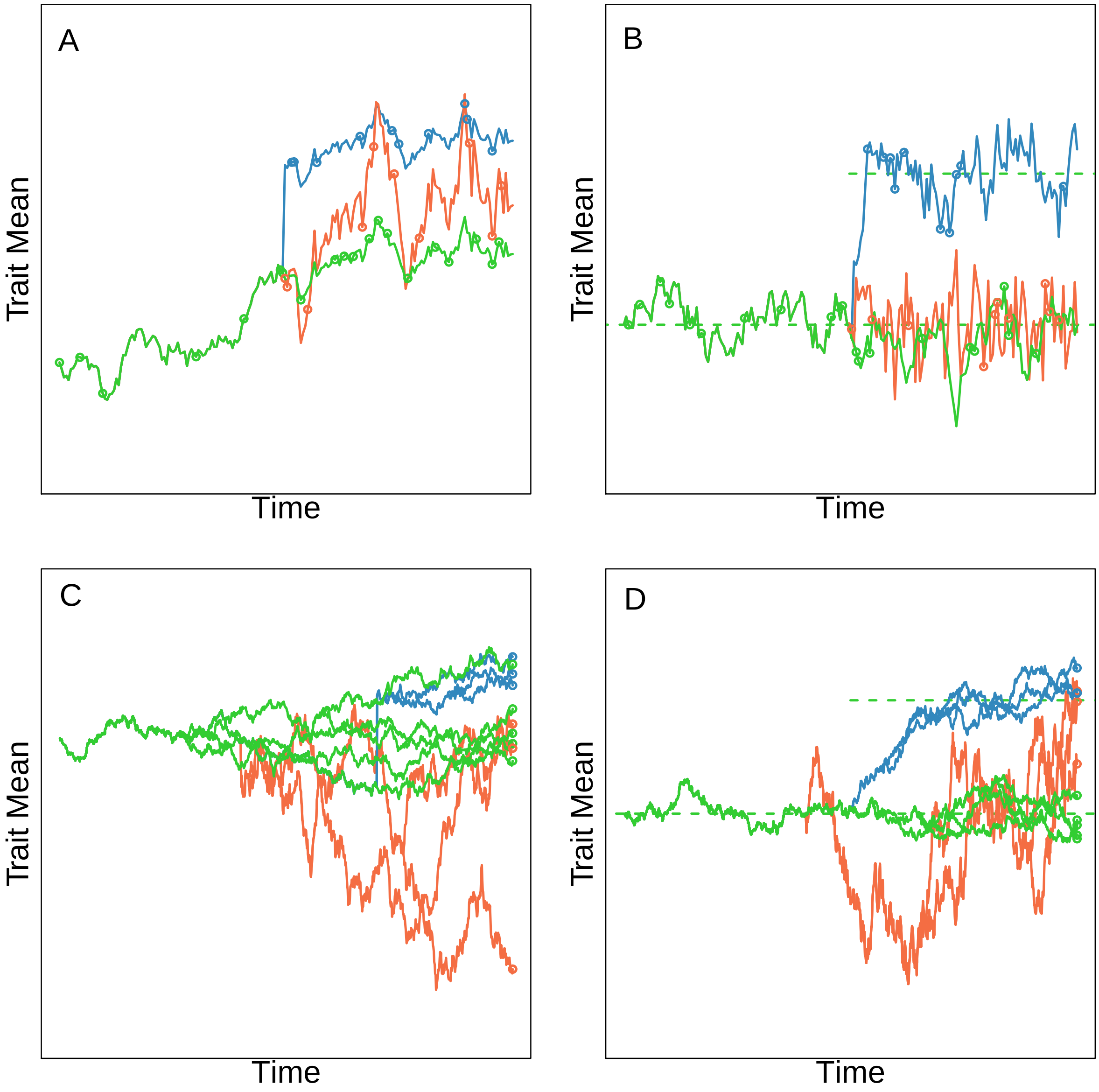

Figure 2.1: Simulated datasets for different models of trait evolution from fossil time-series (A, B) and phylogenetic comparative data (C, D) under Brownian motion models (A, C) and Ornstein-Uhlenbeck models (B, D). Green lines are simulated as constant rate BM and OU processes, with circles indicating sampled data. Blue lines are discontinuous processes in which a burst of evolution occurs in the form of a single displacement (for BM models) or a walk to a new optimum (θ_2) for OU models. However, all other parameters are kept constant. By contrast, red lines are models in which rate parameters (σ^2 and α for BM and OU models, respectively) shift to higher values and remain constant thereafter, but are not burst-like (no shift in the expected value of the process).

Models with alternative patterns of selective regimes can then be compared via model selection techniques to evaluate adaptive hypotheses (Butler and King, 2004). OU models have proven to be very useful for inferring various processes using both phylogenetic comparative (Hansen *et al.*, 2008; Mahler *et al.*, 2013) and paleontological (Hunt, 2008; Reitan *et al.*, 2012) data.

2.6 BOX 2: AN EXAMPLE OF HOW PE CAN MISLEAD INFERENCES

From a historical perspective, it is undoubtedly accurate that numerous comparative methods owe their genesis to the framework of PE. However, the temptation to frame these methods as tests of PE is, in our opinion, unwarranted. For example, Webster *et al.* (2003) developed a method to correlate the total genetic distance between the root and the tip of the tree (hereafter, the path length) with the number of nodes along that path (Figure 2.2). A significant correlation between path length and the number of nodes rejects constant-rate gradualism in molecular evolution, purportedly in favor of a PE model. This correlation has been repeatedly demonstrated in a variety of datasets in traits ranging from molecular sequences to human languages (Webster *et al.* 2003; Pagel *et al.* 2006; Atkinson *et al.* 2008; Lanfear *et al.* 2010; but see Goldie *et al.* 2011).

However, to what extent is there evidence in these cases for PE? We argue: very little. Eldredge and Gould (Eldredge and Gould, 1972) hypothesized that allopatric speciation causes pulsed phenotypic divergence. However, the direction of causality can just as easily be reversed. Genetic divergence is expected to promote speciation under many models of speciation (Nosil, 2012). Alternatively, divergence and speciation may result indirectly from causal links with a third factor, such as shorter generation times, higher fecundity, or increased genetic variation, to name a few (Goldie *et al.*, 2011). Furthermore, trait evolution need not be pulsed for a positive correlation to exist. This effect was demonstrated by Rabosky (Rabosky, 2012), who showed that correlations between path length and speciation are expected whenever trait evolutionary rates are correlated with rates of speciation, even under purely gradual models.

Finally, trait change may not be correlated with speciation at all, but instead be correlated with extinction rates. This may occur, for example, if higher evolvability decreases extinction risk (Lanfear *et al.*, 2010). It is certainly a worthwhile avenue of research to establish a correlation between diversification and trait evolutionary rates, but the available tests demonstrate nothing about whether or not trait evolution is pulsed, whether trait change accumulates anagenetically

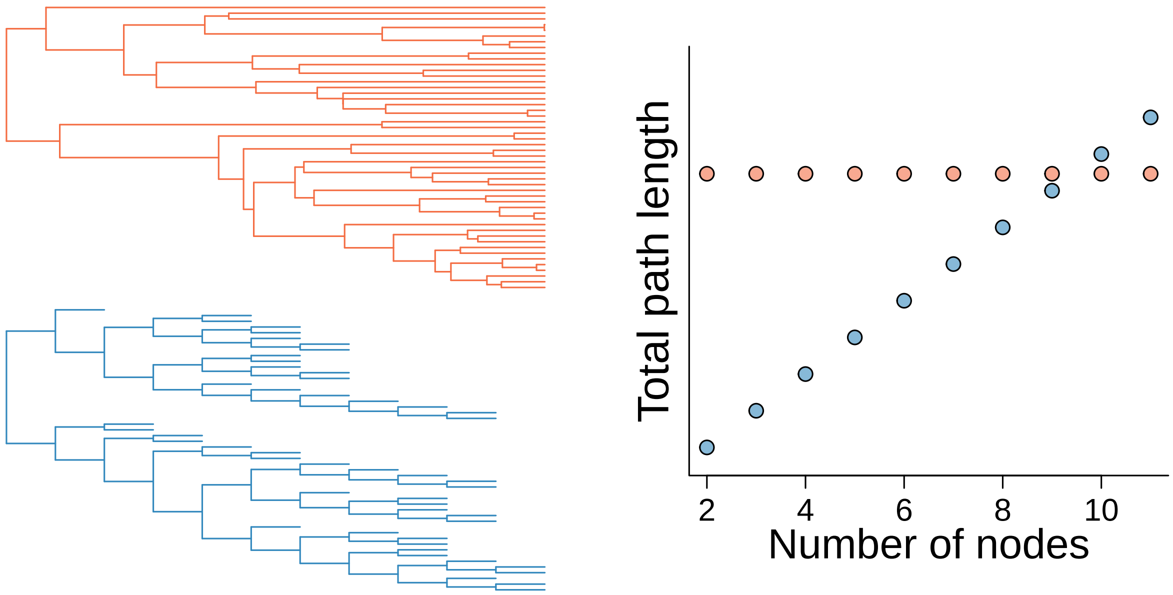

Figure 2.2: Illustration of a method for correlating evolutionary divergence with speciation. Branch lengths for both phylogenies are in units of evolutionary change. The total path length from the root of the tree to the tips is plotted against the total number of nodes along that path. A positive correlation (blue) is indicative of a relationship between the number of speciation events and evolutionary change, while under constant-rate gradualism no such relationship exists (red). Adapted from (Pagel *et al.*, 2006).

or cladogenetically or the direction of causality. Taking the "off-the-shelf" interpretation of these macroevolutionary patterns in the form of PE only obfuscates understanding, and worse, could lead to recapitulating four decades worth of often unproductive and contentious debates. Instead, we argue that we should focus on inferences that may be effectively tested using our available statistical tools. These tools should be integrated with more narrowly-defined theories that are free of the unwanted assumptions of PE.

2.7 BOX 3: MODELING SPECIES SELECTION AND CLADOGENETIC CHANGE ON PHYLOGENIES

In a ground-breaking paper, Maddison and colleagues (2007) developed a statistical framework that has opened up investigation into two major components of PE: the influence of traits on diversification ("species selection", *sensu* Coyne and Orr, 2004; Rabosky and McCune, 2010) and cladogenetic character change. The premise of the approach is that instead of specifying a full likelihood of the model, one need only to describe the probabilities of all possible events that could occur in a very short time interval, Δt, solve a differential equation and then use numerical integration to evaluate the likelihood of the model given the phylogeny and trait data at the tips (see Maddison *et al.* 2007, for full details). The initial model considered by Maddison *et al.* (2007) was the BISSE model in which different states for a single character resulted in different diversification rates.

Consider that lineage diversification can be modeled by a birth-death process (Kendall, 1948), in which there is a constant rate of speciation λ and extinction μ across the clade. Lineages with a trait in state 0 diversify at rates λ_0 and μ_0 and lineages in state 1 diversify at rates λ_1 and μ_1. Transitions (anagenetic evolution) between states $0 \rightarrow 1$ occur at rate q_{01} and transitions from $1 \rightarrow 0$ occur at rate q_{10}. The probabilities of all possible events that can occur during Δt can be described as a set of differential equations. One can then use the integration machinery, as described in Maddison *et al.* (2007), to simultaneously estimate all parameters using either maximum likelihood or Bayesian inference to test for a statistical difference between λ_0 and λ_1 (or between μ_0 and μ_1) in order to infer the strength of species selection.

The BISSE model was extended by Magnuson-Ford and Otto (2012) (BISSE-NESS) to allow for the possibility of character transitions at speciation (cladogenetic change). (An identical model was independently derived by Goldberg and Igić 2012, and related approaches were also developed by Bokma 2002, 2008, 2010.)

In addition to the 6 parameters of the BISSE model ($\lambda_0, \lambda_1, \mu_0, \mu_1, q_{01}, q_{10}$), their model includes the probabilities of a change occuring at a speciation event (p_{0c} and p_{1c}, for the two states, respectively) as well as the probabilities that the character changes are asymmetrical, where the change only occurs in one of the two daughter lineages, p_{0a} and p_{0b} (often referred to as "budding cladogenesis" in the paleobiological literature). This allows one to simultaneously evaluate the importance of species selection as well as the relative importance of cladogenetic versus anagenetic change. This model also highlights the general message of our paper; the questions can be evaluated independently of each other if parameter sets are constrained:

$\lambda_0 = \lambda_1, \mu_0 = \mu_1$	Estimate cladogenetic and anagenetic rates only
$q_{01}, q_{10} = 0$	Estimate species selection with only cladogenetic change
$p_{0c}, p_{1c}, p_{0a}, p_{1a} = 0$	Estimate species selection with only anagenetic change

thus making it an excellent statistical framework, though certainly not the only one, for evaluating the questions associated with PE.

CHAPTER 3

Software for fitting evolutionary models to phylogenetic data[3]

3.1 SUMMARY

Phylogenetic comparative methods are essential for addressing evolutionary hypotheses with interspecific data. The scale and scope of such data has increased dramatically in the last few years. Many existing approaches are either computationally infeasible or inappropriate for data of this size. To address both of these problems, we present GEIGER V2.0, a complete overhaul of the popular R package GEIGER (Harmon *et al.*, 2008). We have re-implemented existing methods with more efficient algorithms and have developed several new approaches for accomodating heterogeneous models and data types.

3.2 INTRODUCTION

In the past few decades, phylogenetic trees have become an key component of evolutionary research. This development has been fueled by the increased availability of robust time-calibrated phylogenies for many groups, in addition to an expanding number of statistical techniques for inferring patterns and processes from comparative data (reviewed in Pennell and Harmon, 2013). Among the many R packages developed for phylogenetic and comparative data, GEIGER (Harmon *et al.*, 2008) has been a primary utility for making macroevolutionary inferences from phylogenetic trees.

However, in the six years since the initial release of GEIGER, the data available for comparative biology have changed substantially. For some groups, we now have phylogenies and corresponding trait data with thousands, and even tens of thousands, of species (e.g., Jetz *et al.*, 2012; Rabosky, 2012; Pyron and Burbrink, 2014; Cornwell *et al.*, 2014; Zanne *et al.*, 2014a). GEIGER V2.0 is a complete overhaul of the previous release (Harmon *et al.*, 2008), designed to scale up compar-

[3] Previously published as: Pennell M.W., Eastman J.M., Slater G.J., Brown J.W., Uyeda J.C., FitzJohn R.G., Alfaro M.E., and Harmon L.J. 2014. GEIGER V2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Bioinformatics 15:2216–2218.

Function	Description	Citations
`fitContinuous`	Fit continuous trait models with ML	Felsenstein (1973); Hansen (1997); Pagel (1997, 1999); Blomberg *et al.* (2003); Hunt (2006); Harmon *et al.* (2010); FitzJohn (2012a)
`fitDiscrete`	Fit discrete trait models with ML	Pagel (1994); Lewis (2001); FitzJohn *et al.* (2009)
`rjmcmc.bm`	Fit multi-rate models to continuous traits	Eastman *et al.* (2011)
`rjmcmc.bm`	Fit jump diffusion models to continuous traits	Eastman *et al.* (2013b)
`mecca`	Fit continuous models to unresolved clades with ABC	Slater *et al.* (2012b)
`fitContinuousMCMC`	Fit simple models of continuous trait evolution with MCMC and incorporate fossil data	Slater *et al.* (2012a)
`pp.mcmc`	Posterior predictive simulations to assess model adequacy	Slater and Pennell (2014)
`medusa`	Estimate shifts in diversification rates	Alfaro *et al.* (2009)
`congruify.phylo`	Time-scale large phylogenies	Eastman *et al.* (2013a)

Table 3.1: Major functions of GEIGER v2.0 with description and citations

ative methods to large data sets. To do so, we have taken two complementary tacks. The first is to improve algorithms and implementations to increase computational efficiency of existing methods. The second is to expand the suite of statistical approaches to allow for heterogeneity in both models and data types across the phylogeny.

In this chapter, we briefly describe the methods now in GEIGER, with a particular focus on novel implementations and algorithms. Most of these methods have been previously published elsewhere in some form and we refer readers to the relevant publications for full explanations. For an overview of the main features of the package, see Table 3.1.

3.3 METHODS

3.3.1 *Fitting simple models of character evolution with maximum likelihood*

Fitting and comparing models of trait evolution can provide insight into many macroevolutionary questions (Pennell and Harmon, 2013). The two "workhorse" functions in GEIGER for fitting models of trait evolution using maximum likelihood, `fitContinuous` and `fitDiscrete`, have both been completely re-implemented. The previous version of `fitContinuous` calculated the likelihood of a set of continuous characters (e.g., body size) having evolved under a model using a variance-covariance (vcv) matrix approach. This involves inverting the vcv, which is extremely computationally intensive, making the method infeasible for large trees (Hadfield and Nakagawa, 2010; FitzJohn, 2012a; Freckleton, 2012; Ho and Ané, 2014). FitzJohn (2012a) demonstrated that using a "pruning"-based algorithm (Felsenstein, 1973) allows for much more efficient likelihood calculations. This algorithm is used the DIVERSITREE package (FitzJohn, 2012a). (For related algorithms, see Freckleton 2012, Ho and Ané 2014). The approach has now been extended to all the models in `fitContinuous`. In addition to improving the efficiency of the algorithm, we have improved numerical optimization procedures and implemented a novel method to simultaneuously estimate model parameters and an additional term to account for measurement error.

For discrete character data—for example, the presence or absence of fur—the most commonly used model is the Mk model (Pagel, 1994; Lewis, 2001). In this model, there are n states $1, 2, \ldots, n$ and the goal is to estimate rates of transition among these, where the rate of transition from state i to state $j \neq i$ is q_{ij}. Considering just one branch in a phylogeny, let $\vec{D}$ be the vector of probabilities of the data given we are in state i, that is, the ith element of D is the probability of the data descended from this point in the tree, given that we are in state i; see Maddison *et al.* (2007) and FitzJohn (2012a) for notation.

In most R-based implementations of Mk, such as APE (Paradis *et al.*, 2004) and previous versions of GEIGER (Harmon *et al.*, 2008), to move from the tip to the base of a branch, we multiply $\vec{D}$ by $\mathbf{P}(t)$, the transition probability matrix with off-diagonal elements that describe the probability of moving from state i to state j over time t and diagonal elements that describe the probability of not changing from state i. To compute $\mathbf{P}(t)$, we take the transition rate matrix $\mathbf{Q}$ composed of q parameters above and compute $\mathbf{P}(t) = \exp(\mathbf{Q}t)$ where exp is the matrix exponential (Sidje, 1998).

For GEIGER V2.0 we have sped up these calculations by using an alternative algorithm. As the number of states gets very large, it is simpler to compute $\exp(\mathbf{Q}t)\vec{D}$ directly, rather than in two steps (Sidje, 1998). This can be done by solving the system of differential equations

$$\frac{d\vec{D}}{dt} = \mathbf{Q}\vec{D} \tag{3.1}$$

subject to the initial condtions at the branch tips.

For small state spaces (a few states to a few tens of states) there will be no speed differences between these two approaches. However, for very large state spaces (hundreds of states) this approach will be much faster than computing $\exp(\mathbf{Q}t)$ directly. Importantly, computing $\exp(\mathbf{Q}t)$ grows faster than linearly in the number of states, while the approach here should grow approximately linearly.

3.3.2 *Bayesian methods for fitting models of character evolution*

A major addition to GEIGER is the implementation of several Bayesian methods for fitting models of trait evolution to comparative data. These include the AUTEUR approach of Eastman *et al.* (2011), which uses reversible jump Markov chain Monte Carlo machinery (Green, 1995) to move across multi-rate models of various complexity. The implementation of this method in GEIGER V2.0 improves upon the original by allowing model partitions to be constrained *a priori* and alternative models to be compared (Eastman *et al.*, 2013b). Additionally, GEIGER now includes: a method for fitting models to phylogenies including unresolved clades using Approximate Bayesian Computation (MECCA; Slater *et al.*, 2012b); a method for including fossil information as priors on node states (Slater *et al.*, 2012a); and a posterior predictive simulation approach for assessing the adequacy of common models of trait evolution (Slater and Pennell, 2014). These types of approaches, which allow for greater complexity both in models and data, will be essential to making robust evolutionary inferences from large comparative datasets.

3.3.3 *Inferring shifts in the rate of lineage diversification*

Alfaro *et al.* (2009) developed an approach, MEDUSA, to detect shifts in diversification rates from molecular phylogenies using a stepwise-AIC algorithm. A single-rate birth-death model is fit to the entire tree, then the tree is partitioned into two rate classes, breaking the tree at all possible nodes. The partition which improves the fit of the model is then fixed and the process is repeated, breaking the tree into three partitions, and so on, until a stopping criterion is reached. MEDUSA can be applied to both fully bifurcating and unresolved trees.

For this release of GEIGER, the MEDUSA algorithm has been improved in a number of ways. It has been re-coded so that it is now orders of magnitude faster and scales well to large trees; this version of MEDUSA has already been applied to a phylogeny of all 9,993 extant bird species (Jetz *et al.*, 2012). We have also developed tools for summarizing MEDUSA analyses across a distribution of trees, such as from a Bayesian posterior or from non-parametric bootstrapping, so that uncertainty in both topology and branch lengths can be accomodated.

The most significant improvement for this version of MEDUSA is in the model selection procedure. As stated above, the MEDUSA algorithm involves comparing the fit of diversification models of varying dimensions (number of parameters). To select an appropriate model, we use Akaike Information Criterion (AIC; Akaike, 1974), together with the small-sample bias-correction (AICc; Burnham and Anderson, 2004b):

$$\mathrm{AICc}_i = -2\log(\mathcal{L}_i) + \frac{2k_i n}{n - k_i - 1} \tag{3.2}$$

where $\mathcal{L}_i$ is the maximized joint likelihood of model i with k_i estimable (free) parameters, and n data points. Here, k is the number of diversification parameters (net diversification rates, $r_i = \lambda_i - \mu_i$, and extinction fractions, $\epsilon_i = \mu_i/\lambda_i$) plus the number of inferred rate shifts. It is unclear whether the shift locations are indeed 'free' parameters of the model as they are not all estimated simultaneously; each additional shift that is introduced reduces the number of possible locations in the next iteration of the algorithm. This is a complex issue and we do not have a strong statistical argument for including them as parameters or not; as such, we have elected to be conservative and penalize adding them as we would any other parameter. For MEDUSA, the sample size n is taken to be the total number of "observed" nodes in a tree (internal + pendant).

However, because MEDUSA considers all possible nested piecewise diversification models, an additional concern is that of multiple testing: as trees grow larger, it becomes increasingly likely that spurious stochastic rate shifts are inferred when no real shifts exist. Indeed, simulations show that the original version of MEDUSA has a high rate of false positives for large phylogenetic trees (more than approximately 50 unresolved tips). We therefore determined an an acceptance threshold (AICc$_t$) through simulation (Figure 3.1). 10,000 single-rate birth-death trees were generated for each of a number of resulting extant taxa N = {10, 20, 50, 100, 200, 400, 500, 750, 1000, 1500, 2000, 2500} using the R package TREESIM (Stadler, 2011). For each simulation we randomly drew diversification parameters where $\lambda \sim \mathcal{U}(0,1]$ and $\mu \sim \mathcal{U}[0,\lambda)$. Each tree was analyzed in MEDUSA, and the difference (ΔAICc) in AICc values between the (true) 0-shift model and best (incorrect) 1-shift model was logged. We then fit a x-shifted power function

$$\text{AICc}_t = a(N-b)^c + x \tag{3.3}$$

to the 95th percentile ΔAICc values for each tree size (Figure 3.1). The best-fitting function had a = -35.94, b = 6.74, c = -0.10, and x = 27.52. If `partition=NA` (the default argument in the `medusa` function), the AICc$_t$ is automatically calculated for the specified tree from this best-fitting function. For trees of 20 or fewer taxa, AICc$_t$ is set to 0.

3.4 CONCLUDING REMARKS

In this note we provide a broad overview of the methods now available in GEIGER. We have not discussed some methods implemented in GEIGER (e.g., 'Congruification' for time-scaling large trees; Eastman *et al.*, 2013a) and many of the nuances of the methods described here have been left out. We refer readers to associated publications and the package documentation for more information.

It is an exciting time for macroevolutionary research. We now have access to data sets of unparalleled size and a wide variety of new statistical approaches with which to analyze them. We hope that the software presented here will help researchers address some fundamental and long-standing questions in macroevolution.

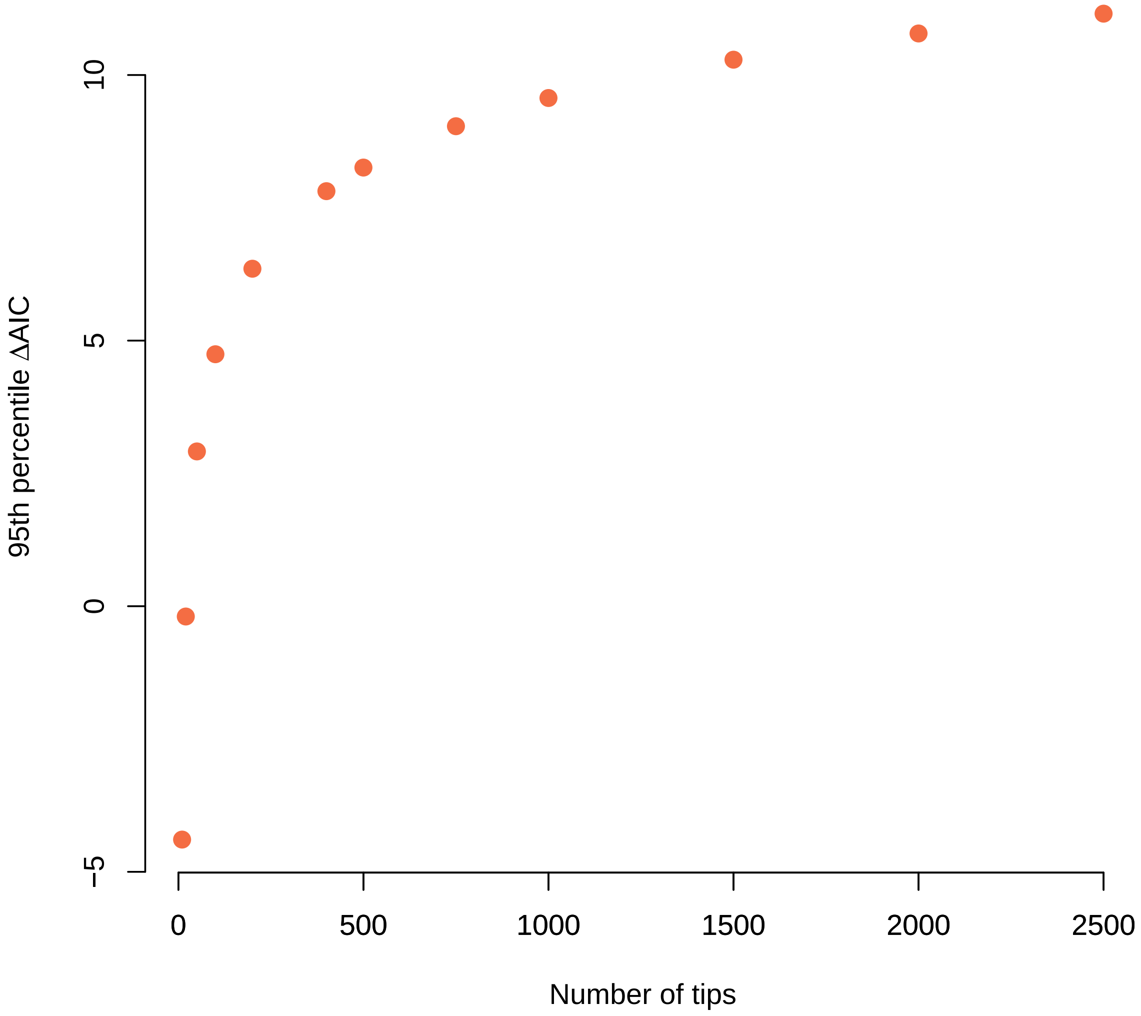

FIGURE 3.1: The 95^{th} percentile values for the difference in AICc scores (ΔAICc) between the (incorrect) 1-shift model and the (true) 0-shift model plotted against simulated tree size. We fit a x-shifted power function to these points to estimate a AICc$_t$ with a Type-1 error rate of 0.05 for a given tree with N tips.

CHAPTER 4

Y FUSE? USING PHYLOGENETIC AND POPULATION GENETIC MODELS TO UNDERSTAND SEX CHROMOSOME FUSIONS [4]

4.1 SUMMARY

The evolution of chromosome number plays an important role in divergent adaptation and speciation. Chromosomal fusion is a common mechanism of karyotypic evolution, but there is little understanding of the evolutionary forces that have driven chromosomal fusions. Because sex chromosomes (X and Y in male heterogametic systems, Z and W in female heterogametic systems) differ in their selective, mutational, and demographic environments, those differences provide a unique opportunity to dissect the evolutionary forces that drive chromosomal fusions. We estimate the rate at which fusions between sex chromosomes and autosomes establish across the phylogenies of both fishes and squamate reptiles. Both the incidence among extant species and the establishment rate of Y-autosome fusions is much higher than for X-autosome, Z-autosome, or W-autosome fusions. Using population genetic models, we show that this pattern cannot be reconciled with many standard explanations for the spread of fusions. In particular, direct selection acting on fusions or sexually antagonistic selection cannot, on their own, account for the predominance of Y fusions. We identify three plausible explanations for the excess of Y-autosome fusions: (i) fusions are deleterious, and the mutation rate is male-biased or the reproductive sex ratio is female-biased, (ii) fusions capture loci under sexually antagonistic selection, and the mutation rate is male-biased or the reproductive sex ratio is female-biased, and (iii) meiotic drive acts against fusions in females. These results may shed light on the processes that drive structural changes throughout the genome.

4.2 INTRODUCTION

The number of chromosomes is one of the most fundamental features of a eukaryotic genome. Chromosome number often varies within species or between closely related species, and such

[4] In press as: Pennell M.W., Kirkpatrick M., Otto S.P., Vamosi J.C., Piechel C.L., Valenzuela N., and Kitano J. Y fuse? Sex chromosome fusions in fishes and reptiles. PLoS Genetics.

variation can contribute to divergent adaptation and speciation (White, 1973; King, 1993; Pérez-Ortín *et al.*, 2002; Chang *et al.*, 2013; Hou *et al.*, 2014). Although genetic drift, selection for reduced recombination, and meiotic drive are hypothesized to fix chromosomal fusions (Nachman and Searle, 1995; Guerrero and Kirkpatrick, 2014), we have an incomplete understanding of the evolutionary forces that allow fusions and fissions to become established.

Sex chromosomes offer a unique opportunity to dissect these forces. The X and Y chromosomes of male-heterogametic species (as in mammals) and the Z and W chromosomes of female-heterogametic species (as in birds) differ in many aspects of their evolutionary environments, particularly with respect to hemizygosity (i.e., XX and ZZ individuals are common, but not YY and WW). While Y and W chromosomes are often thought to be evolutionarily similar, they differ in the amount of time spent in males and females: Y chromosomes spend 100% of their evolutionary history in males, while W chromosomes spend none. X and Z chromosomes also differ: X chromosomes spend 1/3 of their evolutionary history in males, while Z chromosomes spend 2/3 of their history in males. Consequently, the four types of sex chromosomes vary in how selection acts on them, in their effective population sizes, in their mutation rates, and in the relative importance of meiotic drive (Ellegren, 2011; Bachtrog *et al.*, 2011; Beukeboom and Perrin, 2014). All of these factors could play a role in the evolution of chromosomal rearrangements, and so differences in rates of rearrangement among sex chromosomes offer clues to what evolutionary conditions favor changes to genome structure.

Structurally, sex chromosomes are the most rapidly evolving parts of the genome in many groups of animals (White, 1973; Bull, 1983; Ezaz *et al.*, 2006; Beukeboom and Perrin, 2014) In some taxa, such as fishes and squamate reptiles, closely related species (and even populations within a species) differ in how sex is determined (Ezaz *et al.*, 2006; Bachtrog *et al.*, 2014). Further, a large number of fusions between sex chromosomes and autosomes have been discovered (White, 1973; The Tree of Sex Consortium, 2014). Thus there are many phylogenetically independent events, providing the opportunity to test whether fusions involving the four different types of sex chromosomes are equally likely to occur and/or establish within a species.

A fusion between a sex chromosome and an autosome can usually be detected because it creates an odd number of chromosomes in one sex (Figure 4.1; Ohno, 1967; White, 1973). With XY sex determination, a Y-autosome fusion creates an X1X2Y system, with the unfused homologue segregating as a neo-X chromosome. Likewise, X-autosome fusions generate XY1Y2 systems, Z-

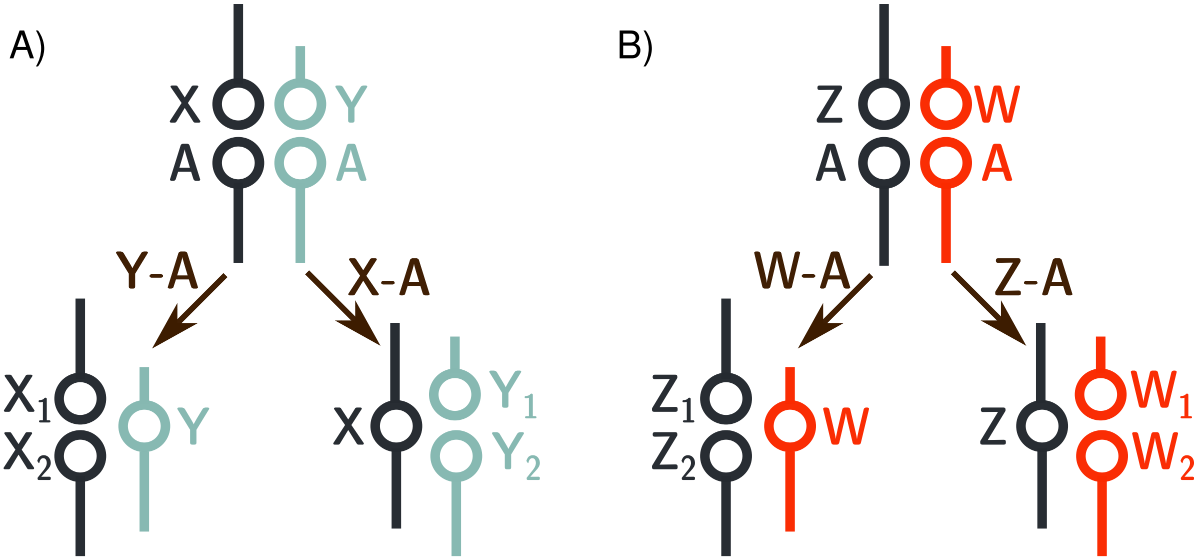

Figure 4.1: Sex chromosome-autosome fusions create multiple sex chromosome systems. (A) In XY systems, X-autosome and Y-autosome fusions make XY1Y2 and X1X2Y systems, respectively. (B) In ZW systems, Z-autosome and W-autosome fusions make ZW1W2 and Z1Z2W systems, respectively.

autosome fusions generate ZW1W2 systems, and W-autosome fusions generate Z1Z2W systems. These neo-sex chromosome systems can often be identified by light microscopy, without molecular cloning or linkage mapping. This has enabled cytogenetic studies to identify many species with sex chromosome-autosome fusions (White, 1973; Ezaz *et al.*, 2009; Kitano and Peichel, 2012; Yoshida and Kitano, 2012; Maddison and Leduc-Robert, 2013). These data have yet to be used to estimate rates of different types of sex-autosome fusions.

Three main evolutionary forces have been thought to be important to the establishment of fusions. The first is direct selection. While chromosome rearrangements are often considered deleterious (King, 1993; Gardner *et al.*, 2012), chromosomal translocations may alter the expression of genes near the breakpoint (Ohno, 1967; Dobigny *et al.*, 2004), which may sometimes be beneficial (Pérez-Ortín *et al.*, 2002; Chang *et al.*, 2013). A second mechanism that has been proposed to establish fusions is sexually antagonistic selection at an autosomal locus (Charlesworth *et al.*, 1982). A fusion with a sex chromosome can cause an allele that is beneficial in one sex to spend more than half of its evolutionary life in that sex. Meiotic drive is a third force. During female meiosis in animals, one of the products of meiosis goes into the egg, while the others are discarded in the polar bodies. In some species, female meiotic drive preferentially transmits fused chromosomes to eggs, while unfused chromosomes go into polar bodies (Pardo-Manuel de Villena and Sapienza, 2001a,b). This drive favors X-autosome fusions because they experience female meiosis in two of every three generations. In other species, female meiotic drive preferentially transmits fused chromosomes, which should select against X-autosome fusions (Yoshida and Kitano, 2012). While these evolutionary forces are known to affect the spread of sex chromosome-autosome fusions, previous work has not examined the relative rates at which fusions with different types of sex chromosomes establish within a population.

We begin this study by analyzing a large new data set that includes information on the sex determination system and karyotypes across the tree of life (The Tree of Sex Consortium, 2014). We focus on fishes and squamate reptiles because these taxa include many independent origins of XY and ZW systems (Ezaz *et al.*, 2009; Kitano and Peichel, 2012), allowing us to assess differences in the rates of fusions. We find that Y-autosome fusions fix at a much higher rate than any of the other three types of sex chromosome-autosome fusions. This then motivates us to develop an integrated body of analytic models that predict the relative fixation rates for the different types of fusions. The models incorporate a large number of potentially important factors: deleterious and beneficial effects of fusions, sexually antagonistic selection, female meiotic drive, genetic drift, sex-biased mutation rates, and biased sex ratios. We find that several of the data cannot be explained by some of the most frequently-discussed hypotheses. There are, however, several combinations of forces that are able account for the observed patterns of sex chromosomes fusions, as we highlight.

4.3 ANALYSIS OF PATTERNS OF SEX CHROMOSOME-AUTOSOME FUSIONS IN VERTEBRATES

We compiled lists of species with multiple sex chromosome systems (X1X2Y, XY1Y2, ZW1W2, and Z1Z2W systems) from the Tree of Sex database (The Tree of Sex Consortium, 2014). Although X1X2Y systems (or ZW1W2 systems) can also arise from species with XO (or ZO) systems through a reciprocal translocation between an X (or a Z) and an autosome (White, 1973; Kitano and Peichel, 2012), XO or ZO systems are rare in vertebrates (The Tree of Sex Consortium, 2014). In addition, although fission of sex chromosomes can also create multiple sex chromosome systems (White, 1973; Kitano and Peichel, 2012), such fissions are also rare in vertebrates (Ohno, 1967;

Taxa	Y-A (X1X2Y)	X-A (XY1Y2)	W-A (Z1Z2W)	Z-A (ZW1W2)	XY systems	ZW systems
Fish	42	3	0	2	109	38
Amphibians	1	0	0	0	29	16
Reptiles	40	0	2	4	120	240
Birds	-	-	2	4	0	192
Mammals	18	24	-	-	467	0

Table 4.1: Observed number of species with multiple sex chromosome systems in vertebrates. Only X1X2Y, XY1Y2, Z1Z2W, and ZW1W2 systems are counted here.

Kitano and Peichel, 2012; Yoshida and Kitano, 2012). Therefore, we considered that most multiple sex chromosome systems are derived from sex chromosome-autosome fusions in vertebrates. We address two questions with our empirical analyses. First, do Y-A (W-A) fusions occur at different rates than X-A (Z-A) fusions? Second, are there differences in rates of fusion between male and female heterogametic lineages? For both questions, we first simply tabulated the numbers in the database and computed Fisher's exact test. This ignores phylogenetic non-independence but allowed us to use all of the available data.

Examining the raw counts (Table 4.1), two interesting patterns emerge. Hereafter, we refer to the fusion between a Y chromosome and an autosome as Y-A fusion, and similarily for other sex chromosomes. First, there are more species with Y-A fusions (X1X2Y karyotype, 101 species) than with X-A fusions (XY1Y2 karyotype, 27 species). The pattern is particularly strong in both fishes and squamate reptiles, while the numbers are more nearly equal in mammals (Table 4.1). Such counts, however, do not account for the phylogenetic relatedness among many of the species. Second, sex chromosomes in XY lineages are more often fused than those in ZW lineages (Table 1). In fishes, 41.3% (45/109) of XY species have fused sex chromosomes, whereas only 5.3% (2/38) of ZW species do (Fisher's exact test $p < 0.001$). In reptiles, 33% (40/120) of XY species have fusions, whereas only 2.5% of species (6/240) of ZW species do (Fisher's exact test $p < 0.001$).

To gain a better estimate of the rates at which fusions establish with different chromosomes, we fit phylogenetic models to the fusion data. We first matched sex chromosome systems from the fish dataset to a recent time-calibrated phylogeny of teleosts (Rabosky *et al.*, 2013), containing 7811 species (we note that a small number of species were removed from the published phylogeny due to errors discovered after publication; M. Alfaro, personal communication). We matched the data of sex chromosome systems from squamates to the squamate phylogeny (Pyron *et al.*, 2013; Pyron and Burbrink, 2014) using genetic data from 4161 species. In order to maximize overlap between the trait data and the species, we used an approximate matching algorithm for unmatched species: 1) retain all species that occur in both the tree and the dataset; 2) replace an unmatched species in the tree with a randomly selected unmatched species in the dataset from the same genus as long as this did not result in more than two representatives from the genus (this assumes monophyly of genera but avoids determining node order for nodes not in the original trees). We then pruned down the phylogeny down to those tips with data assignments.This resulted in phylogenetic comparative datasets containing 163 species of fish (Figure 4.2) and 261 squamate (Figure 4.3) species.

We conducted two separate types of phylogenetic analyses on both groups. First, we examined differences between XY and ZW systems; here, we treat X-autosome and Y-autosome fusions as equivalent (and likewise, Z-autosome and W-autosome fusions). Second, we investigated autosomal fusion rates for all types of sex chromosomes individually (i.e., Y-, X-, W-, and Z-autosome fusions). While the second analysis provides more detailed resolution, some of the states are rarely observed (and in some cases, not at all). All analyses were performed using the R package DIVERSITREE (FitzJohn, 2012a), and code to reproduce all results can be found at `https://github.com/mwpennell/fuse/analysis`.

4.3.1 *Fusion rates in XY vs. ZW systems*

Using a Markov model (Pagel, 1994), we considered transitions among the following states:

1. XY: Male heterogametic unfused
2. XY_F: Male heterogametic fused (X1X2Y or XY1Y2)
3. ZW: Female heterogametic unfused
4. ZW_F: Female heterogametic fused (Z1Z2W or ZW1W2)

allowing transitions between all states with $q_{A.B}$ representing the transition rate between states A and B. We then used likelihood ratio tests to restrict the model in order to improve our ability to estimate the parameters of interest.

We first imposed the biologically reasonable constraint that prior to becoming XY_F (or ZW_F), a lineage must first be XY (or ZW); e.g., the transition rate from female heterogametic unfused to

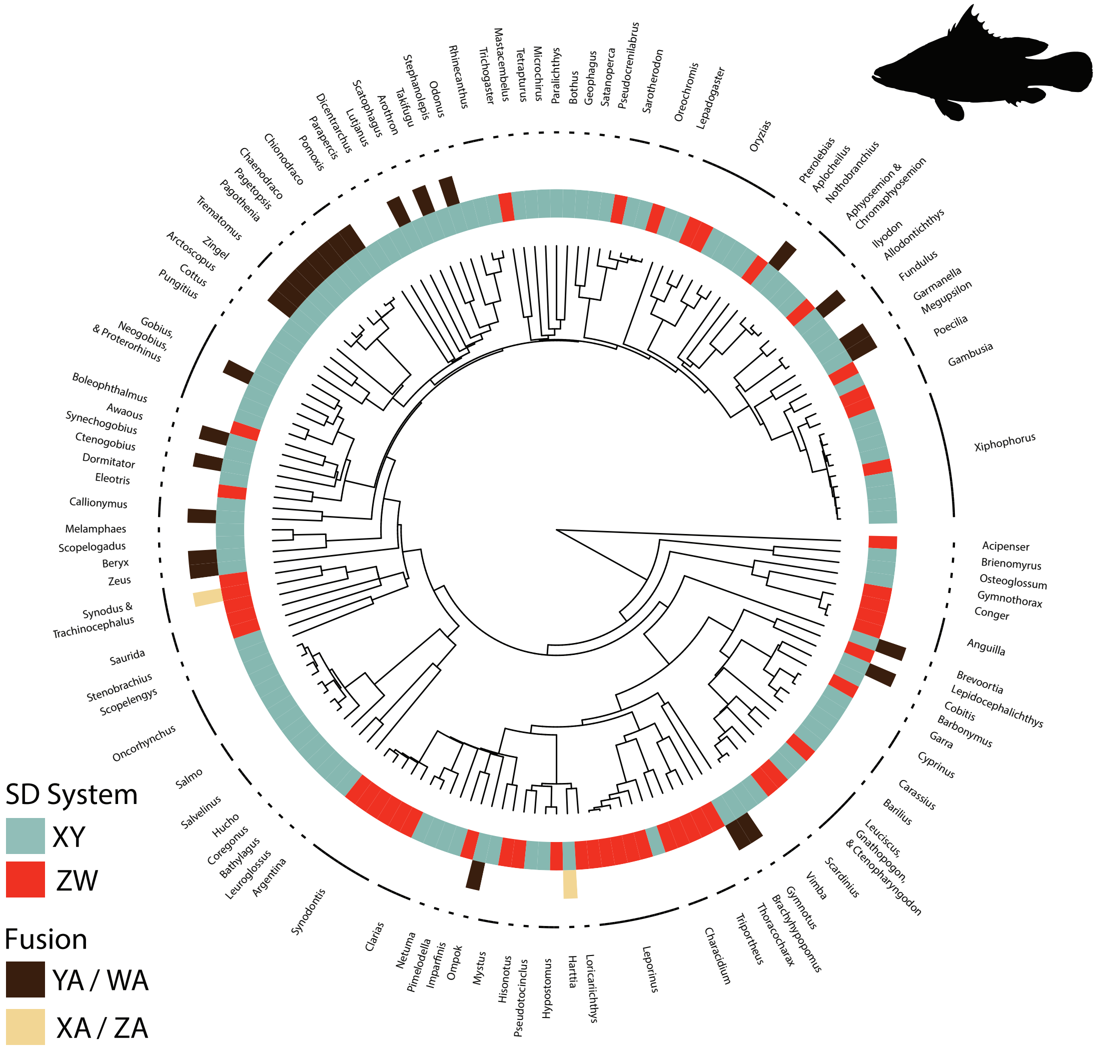

FIGURE 4.2: Sex chromosome fusions (outer circle) and sexual determination system (inner circle) mapped onto the phylogenetic trees of fish. The vast majority of fusions occur in XY systems (aqua) and involve Y-A fusions (brown).

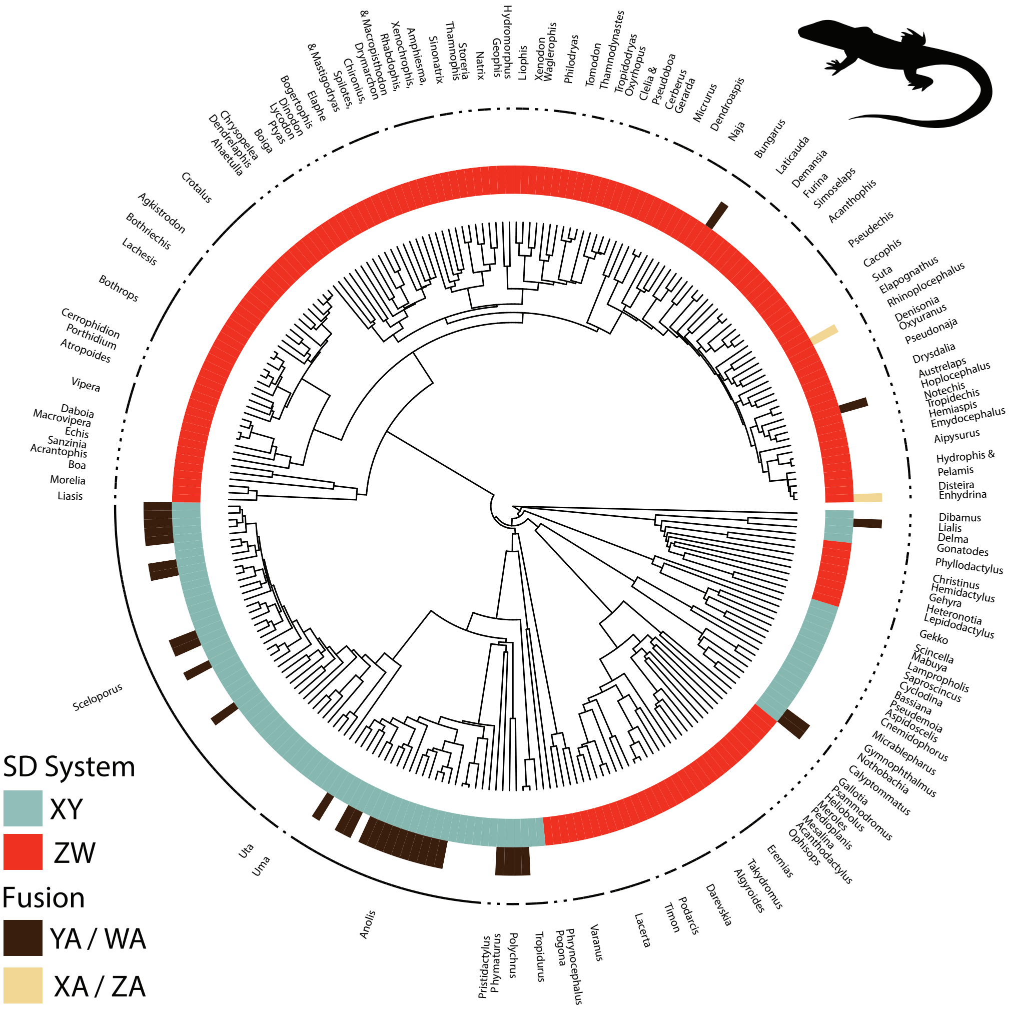

Figure 4.3: Sex chromosome fusions (outer circle) and sexual determination system (inner circle) mapped onto the phylogenetic trees of squamate reptiles. The vast majority of fusions occur in XY systems (aqua) and involve Y-A fusions (brown).

male heterogametic fused $q_{ZW.XY_F}$ would be zero. These restrictions did not lead to a significant decline in likelihood for either squamates or fish and was accepted. Next, we proposed a model in which the rate of switching the heterogametic sex, going from a XY to a ZW system and *vice versa*, did not depend on whether the lineage contained a fused sex chromosome or not (e.g., $q_{XY_F.ZW} = q_{XY.ZW}$). In both fish and squamates, this restriction was acceptable.

In the next step, we proposed a model in which the rate of chromosomal fission, going from a fused sex chromosome system to an unfused system of the same type, was the same for XY and ZW systems. In fish, a likelihood ratio test favored the more restricted model, whereas in squamates, the more general model (where $q_{XY_F.XY} \neq q_{ZW_F.ZW}$) was favored ($p = 0.012$). The support for the more general model in squamates stems from the scarcity of ZW fusions in the data; there is little information to reliably estimate the transition rate from fused female heterogametic to unfused female heterogametic ($q_{ZW_F.ZW}$) using maximum likelihood (see below). We therefore took slightly different approaches when analyzing the two clades.

For fish, we compared the resulting model ($q_{XY_F.XY} = q_{ZW_F.ZW}$, $q_{ZW.XY_F} = q_{XY.ZW_F} = 0$, $q_{XY_F.ZW} = q_{XY_Z.ZW}$, $q_{ZW_F.XY} = q_{ZW.XY}$) to an even more reduced model in which the XY and ZW fusion rates were set to be equal ($q_{XY.XY_F} = q_{ZW.ZW_F}$). We found the rate difference to be highly significant ($p = 0.014$) using a likelihood ratio test. To better accomodate uncertainty in the estimate, we ran a Bayesian analysis. We set broad exponential priors on all parameters ($\lambda = 20$) and sampled 50,000 generations of the MCMC, discarding the first 10,000 as burnin. This also supported our conclusion that XY fusions occur at a higher rate than ZW fusions (98.6% of the posterior probability supported this and the 95% credibility interval for the difference in rates did not overlap with zero; Figure 4.4).

For the squamate data, we took two approaches. First, we assumed that the 'equal fission rates model' was indeed reasonable and performed the same analysis as in fish. Using a likelihood ratio test, the difference in fusion rates for XY and ZW was found to be highly significant ($p = 0.003$). The same was true for the Bayesian analysis (99.9% of the posterior probability distribution supported this conclusion; Figure 4.4). Second, we used a Bayesian MCMC to fit a model in which the fission rate $q_{ZW_F.ZW}$ was estimated independently of $q_{XY_F.XY}$. For this model the support for the difference between XY and ZW fusion rates was not as strong (92.0% of the posterior probability supported $q_{XY.XY_F} > q_{ZW.ZW_F}$; Figure 4.5).

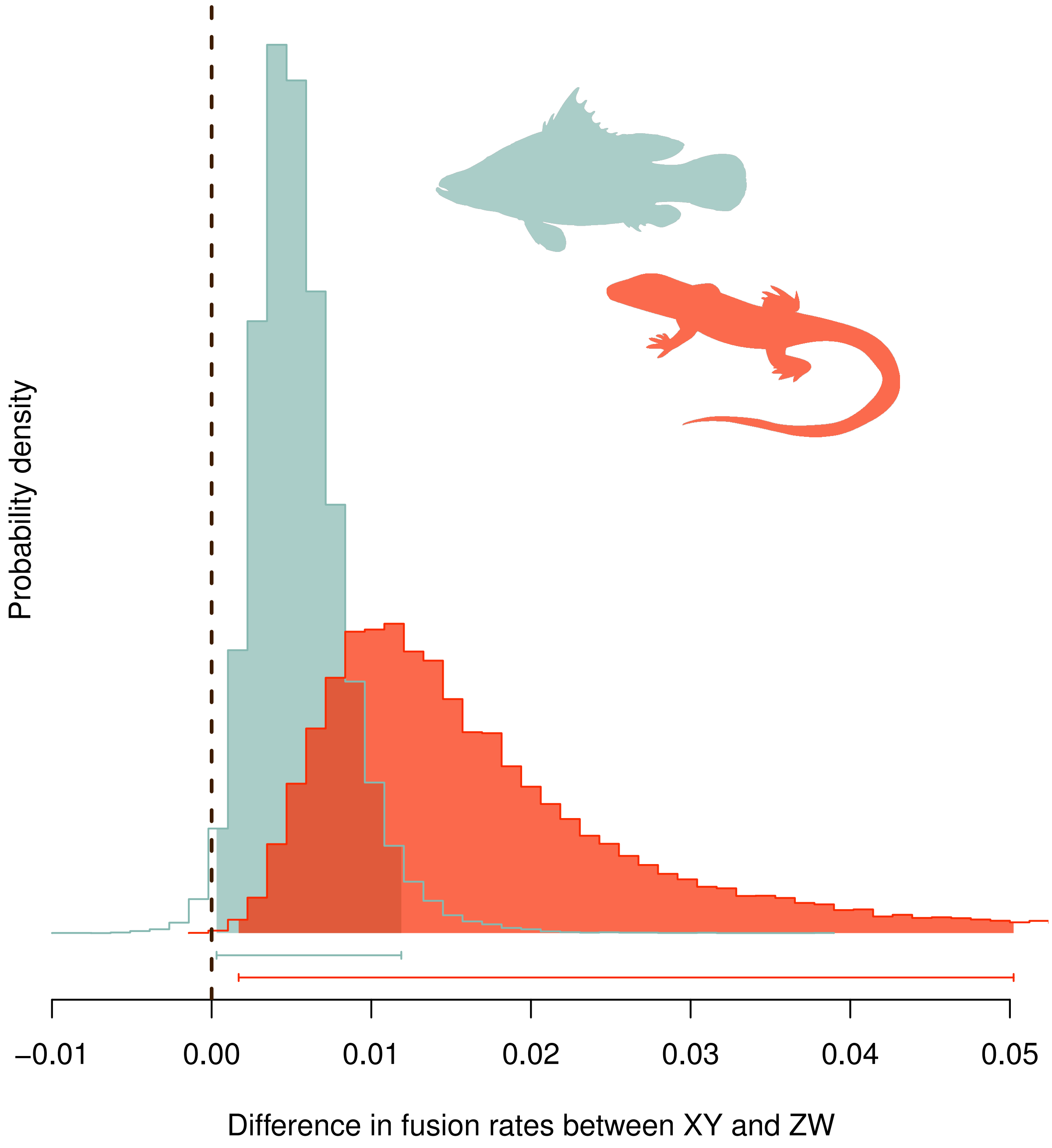

Figure 4.4: Posterior probability density of the difference in fixation rates of fusions between autosomes and sex chromosomes (rates in XY species minus in ZW species). The plot illustrates the difference in fusion rates over the last 40,000 steps of an MCMC chain, with the 95% credibility intervals shown by the horizontal bars below the figure.

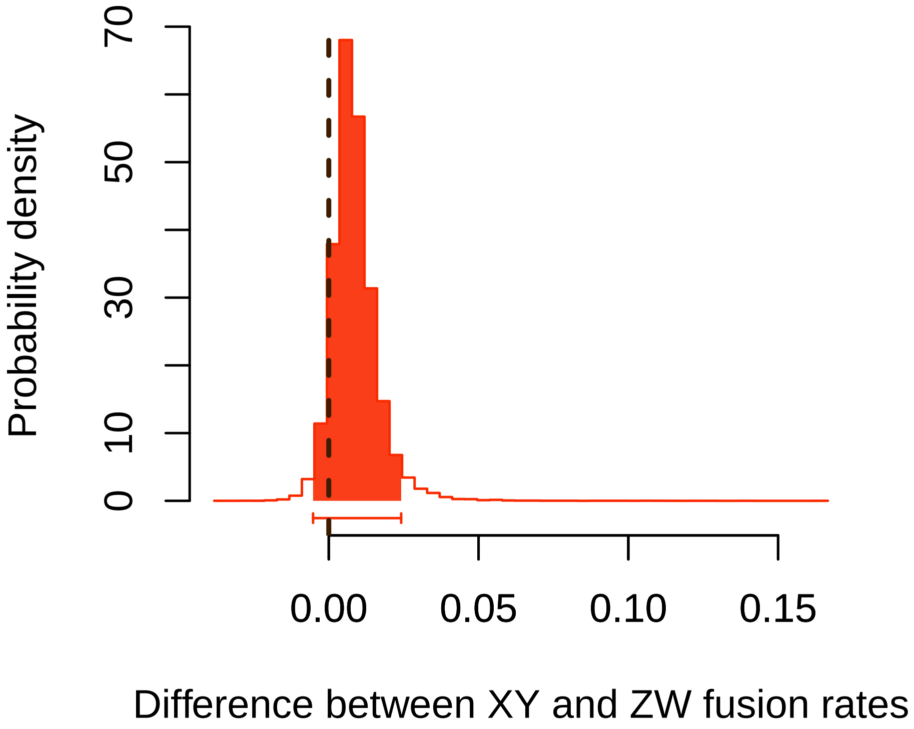

Figure 4.5: Posterior estimate of the rate difference between XY and ZW fusions ($q_{XY.XY_F} - q_{ZW.ZW_F}$) in squamate reptiles when we allow the fission rates $q_{XY_F.XY}$ and $q_{ZW_F.ZW}$ to differ.

As mentioned above, the squamate data contain very little information about fission rates, especially from ZW_F to ZW. The likelihood approach has difficulty distinguishing between two explanations for the lack of fused ZW chromosomes: rare ZW fusions or common ZW fissions. Nevertheless, there is a strong signal that ZW fusions should be less common, which we confirmed by considering residency times t_R. For XY fusions,

$$t_{R,XY_F} = \frac{q_{XY.XY_F}}{q_{XY.XY_F} + q_{XY_F.XY}} \tag{4.1}$$

and for ZW fusions

$$t_{R,ZW_F} = \frac{q_{ZW.ZW_F}}{q_{ZW.ZW_F} + q_{ZW_F.ZW}} \tag{4.2}$$

Using a Bayesian analysis, we found very strong support for the residency time being greater for XY fusions than ZW fusions (99.8% of the posterior probability supported $t_{R,XY_F} > t_{R,ZW_F}$; Figure 4.6). In the absence of direct information about fission rates for fused ZW chromosomes, we conclude that the data is more parsimoniously explained by rare ZW fusions, while acknowledging that rapid ZW fission rates may also explain the data for squamates.

4.3.2 *Comparing fusion rates between chromosomes*

Rather than classifying the states as male/female heterogametic unfused/fused, we separated out the different types of fusions (e.g., classifying X-autosome [XA] and Y-autosome [YA] fusions as different states). This allowed us to assess whether the patterns we observed were driven by an overabundance of autosomal fusions with the Y chromosome. After matching the data to the tree, we did not have any records of WA fusions in fish while in squamates, XA fusions were absent. We thus considered models with only three fused states (for fish: XA, YA, and ZA; for squamates: YA, WA, and ZA)

For both the fish and the squamates, we again restricted the model via a nested series of likelihood ratio tests. For both clades, we found it to be statistically justifiable to assume that: a) transitions from one fused state directly to another fused state were impossible; b) prior to becoming fused, a lineage had to be in the corresponding unfused state; and c) fission rates were constrained to be equal. This allowed us to reliably evaluate whether the fusion rates differed by chromosome.

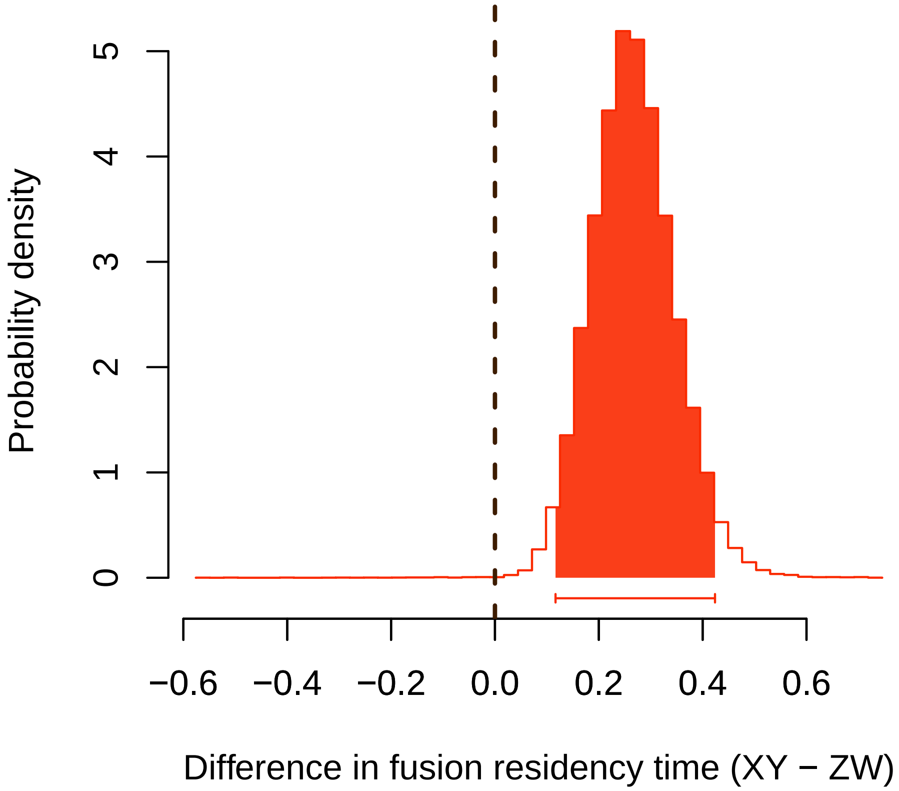

FIGURE 4.6: Posterior estimate of the difference in residency time between XY and ZW fusions (i.e., $t_{R,XY_F} - t_{R,ZW_F}$) in squamate reptiles.

For the fish, using likelihood ratio tests, we found YA fusions to be significantly higher than XA fusions (p = 0.016) and ZA fusions (p = 0.035), but that XA and ZA fusion rates were not significantly different (p = 0.658). Again, WA fusions did not exist in the fish analysis so we could not compare them to other classes. We then performed a Bayesian MCMC analysis to gain a better estimate of the relevant parameters. For the purposes of this analysis, we fixed XA and ZA fusions to occur at the same rate and then compared this rate to that for YA fusion. We found that YA fusions occur at a much higher rate than XA/ZA fusions (Figure 4.7; 99.5% of the posterior distribution supported this conclusion).

For the squamate analysis, YA fusions also occured at a higher rate than WA fusions ($p <$ 0.001) and ZA fusions ($p <$ 0.001). WA and ZA fusions rates were not significantly different from one another ($p \approx 1$). As with the fish, for the Bayesian analysis we set WA and ZA fusion rates to be equal and estimated the difference between YA fusions and other type of fusions. 99.9% of the posterior probability distribution supported YA fusions occuring at a higher rate than fusions on other chromosomes (Figure 4.8).

Taken together, these results strongly suggest that the difference between XY and ZW fusion rates is driven almost entirely by the very high rates of autosomal fusions involving the Y chromosome relative to the other sex chromosomes.

4.4 THEORETICAL ANALYSIS

To evaluate the plausibility of various mechanisms to explain the excess of fusions involving Y chromosomes, we modeled the rate of establishment of different sex chromosome-autosome fusions under various evolutionary scenarios. The core results are derived in Appendix A, where we approximate the rate, R_C, at which a given type of chromosome fusion ($C = X, Y, Z, \text{or} W$) establishes within a population, accounting for both the rates that different types of fusions arise in a population and the probabilities that they fix.

To facilitate comparison to the data, we focus on the establishment rates for Y-A, Z-A, and W-A fusions relative to the rate of X-A fusions. We begin by studying the neutral case, where selection is absent. We allow, however, for sex-biased mutation rates and sex-biased sex ratios (see Appendix A for definitions). We then ask how these neutral results are altered by the three

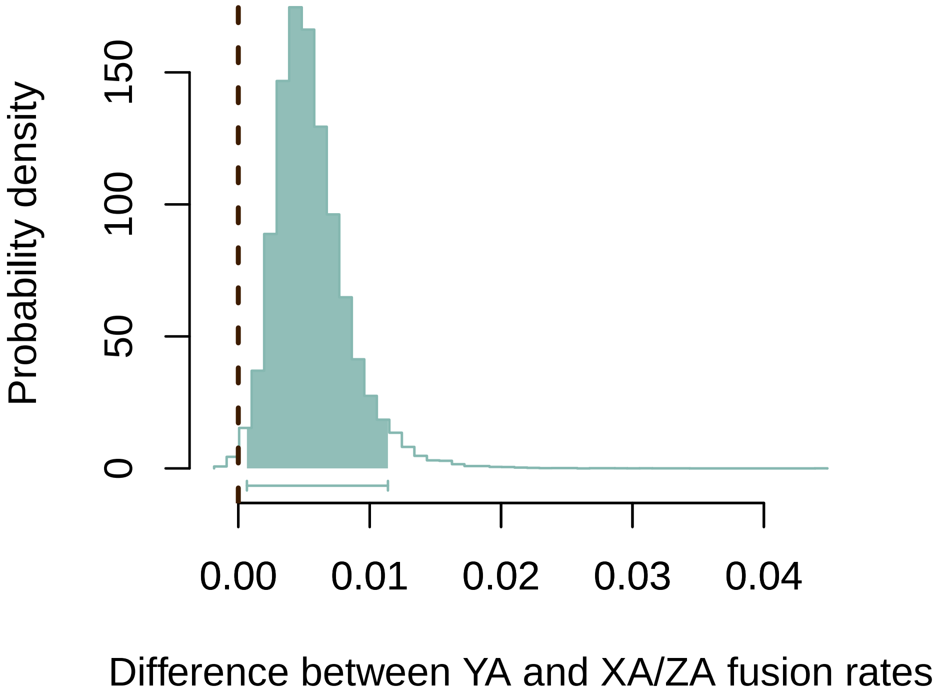

Figure 4.7: Posterior estimate of the rate difference between YA and XA/ZA fusions in fish. When the estimate is greater than zero, this means that the YA fusion rates are higher than those of the other chromosomes

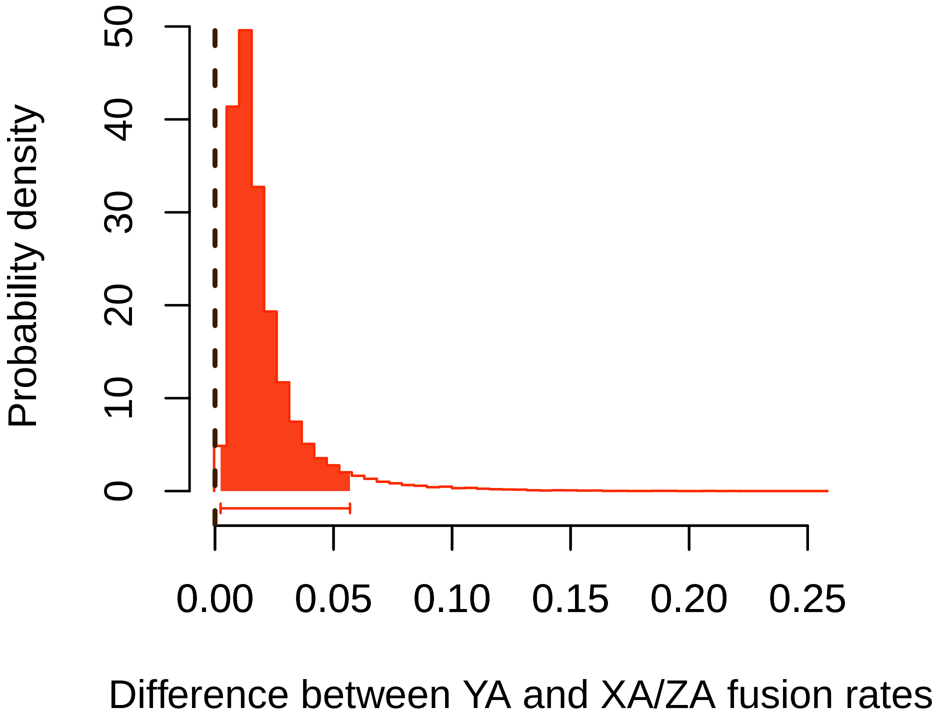

FIGURE 4.8: Posterior estimate of the rate difference between YA and WA/ZA fusions in squamate reptiles. When the estimate is greater than zero, this means that the YA fusion rates are higher than those of the other chromosomes

main evolutionary forces thought to impact the rate of fusions: direct selection acting on the fusion, meiotic drive, and sexually antagonistic selection.

Neutral case — We first consider the case without any selection or drive in the model. The overall establishment rates for fusions are given by the mutation rates generating each type of fusion (Appendix A, Equation A.6). Interestingly, the sex ratio does not enter into these results. Among newborns, each copy of a particular sex chromosome has an equal chance of being the progenitor of the entire population of that sex chromosome at some distant point in the future, regardless of subsequent changes in the survival and reproductive success of males versus females.

Sex-biased mutation would alter the relative frequencies that different types of neutral fusions become fixed. Evidence suggests that the sexes differ substantially in the rate at which fusions arise: data from humans indicates that balanced translocations are the most likely source of new fusions (Schubert and Lysak, 2011), and these seem to be predominantly paternal in origin (Batista *et al.*, 1993; Sartorelli *et al.*, 2001; Wyrobek *et al.*, 2006; Thomas *et al.*, 2010; Grossmann *et al.*, 2010; Schubert and Lysak, 2011). If mutation is male-biased but does not depend on the type of chromosome (that is, the X and Y chromosomes in a male are equally likely to fuse), then Y-A fusions will fix most frequently (see Equation A.7). In this case, however, Z-A fusions would be almost as common as Y-A fusions (at least 2/3 as common, see Equation A.7 and Figure 4.9B, black curves), which is not seen in the data (Figures 4.2 and 4.3). Thus the hypothesis that sex-autosome fusions are selectively neutral does not appear consistent with the data.

Direct fitness effects — We next ask how relative establishment rates depend on the direct fitness effects of a fusion (Appendix A). Assuming that the fusion has an additive effect on fitness and that all else is equal (unbiased reproductive sex ratios and mutation rates, and equal fitness effects for all types of fusions), the establishment rate is equal for X-A and Z-A fusions and for Y-A and W-A fusions (Figure 4.9). Fusions involving the Y or W are a factor $\frac{1}{3}(1+e^{-2Ns}+e^{-Ns})$ more common, where N is the number of reproductive adults and s is the fitness effect of the fusion. Thus, deleterious fusions ($s<0$) are much more likely to involve the Y or W chromosome, because of the smaller population size of these chromosomes (Figures 4.9A and

4.9C). Conversely, beneficial fusions are more likely to involve X or Z chromosomes because they are more numerous and so more often the targets of beneficial fusions (Figures 4.9B and 4.9D).

Direct selection alone does not, however, explain why Y-A fusions are more common than W-A fusions. Similarly, direct selection cannot, on its own, explain why fusions in XY lineages are more common than in ZW lineages. To account for the observed data, therefore, we must invoke a combination of direct selection and sex biases, either in the sex ratio or in the mutation rate of fusions.

Sexual selection is often stronger in males, which leads to a female-biased reproductive sex ratio (that is, more reproducing females than males; Bateman, 1948). This situation will favor Y-A fusions over all other types if fusions are deleterious (Figure 4.9A). In this case, fusions are established by random genetic drift. Y fusions establish most frequently because the Y has the smallest effective population size of the four types of sex chromosomes because it is both hemizygous and restricted to the sex (males) with the fewest number of breeding individuals. By contrast, if fusions are beneficial, Y-A fusions are unlikely to be the most common type of fusion (only seen when there is an extremely male-biased sex ratio, with many fewer breeding females than males; see Equation A.7 for very weak selection; Figure 4.9B). A second asymmetry that may be important to explaining the data is sex-biased mutation. As in the neutral case, we find that Y-A fusions will be most common when they are deleterious if they arise more often in males than females (blue, Figure 4.9B). These results strictly apply only when the fusion has an additive effect on fitness, but the relative frequencies of establishment for the different types of fusions are robust to changes in dominance.

Overall, selection acting against fusions combined with male-biased sex ratios and/or male-biased mutation rates can account for the observation that fusions in male heterogametic systems are substantially more common than in female heterogametic systems (Figure 4.4), and the observation that Y-A fusions are the most common (Figures. 4.7 and 4.8).

Meiotic drive — We next consider meiotic drive, which is thought to favor fused autosomes in some species of mammals and unfused chromosomes in others (Pardo-Manuel de Villena and Sapienza, 2001a,b). If meiotic drive is weak, we can treat it as a form of direct selection, and so Equations A.4 and A.5 in Appendix A continue to apply. For clarity, we focus here on meiotic drive in females. (The results apply to meiotic drive in males if we interchange the sexes and

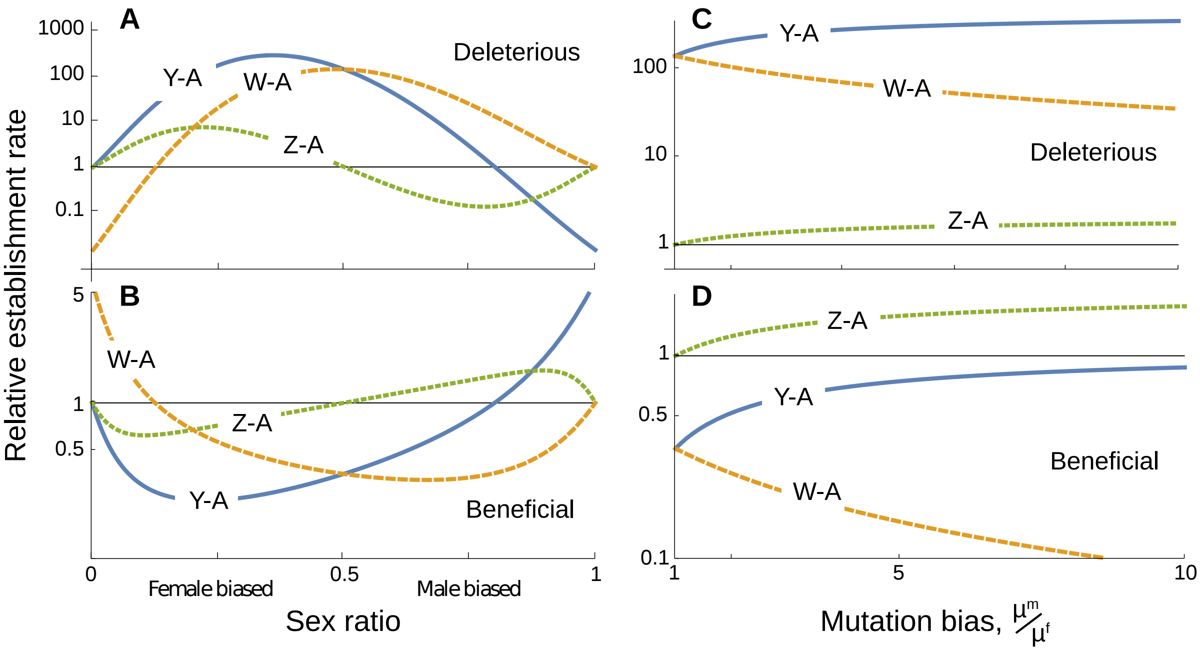

Figure 4.9: Establishment rates of sex-autosome fusions under direct selection, relative to the rate for X-A fusions. (A/B) Effect of sex ratio bias among reproductive adults, $N^m/(N^m + N^f)$, assuming $\mu^m = \mu^f$. (C/D) Effect of relative mutation rate for fusions in males versus females, μ^m/μ^f, assuming $N^f = N^m$. Mutations are deleterious ($s = -0.0003$) in panels (A) and (C), and beneficial ($s = 0.0003$) in panels (B) and (D). Total effective population size $N^f + N^m = 10000$.

the sex chromosomes, e.g., drive in ZW females becoming equivalent to drive in XY males.) Specifically, we assume that the probability that the fusion is transmitted to an egg is multiplied by a factor $1 + f$ in fusion heterozygotes. If unfused chromosomes are preferentially transmitted to the egg, f is negative. Averaging over the sexes, the effect of weak meiotic drive on an X-A fusion is equivalent to direct selection with a coefficient $s_X = 2f/3$. (The factor of 2/3 appears because drive acts only when the fusion is in a female.) Thus when female meiotic drive favors fused chromosomes, the fixation probability for a single X-A fusion is higher than that for a Y-A fusion, which never experiences female meiotic drive (that is, it has an effective selection coefficient of $s_Y = 0$). In ZW systems, a W-A fusion is always carried by females and so benefits in every generation when female meiosis is biased towards fused chromosomes ($s_W = f$), while Z-A fusions enjoy that advantage only one generation out of every three ($s_Z = f/3$).

Once we account for the numbers of each chromosome type (and assuming unbiased mutation rates and reproductive sex ratios), if female meiotic drive favors unfused chromosomes ($f < 0$), then Y-A fusions are expected to establish at the highest rate, followed by W-A fusions, Z-A fusions, and last X-A fusions. The relative rankings are reversed if female meiotic drive favors fused chromosomes ($f > 0$). Thus the observed excess of Y-A fusions can be accounted for by meiotic drive in females if unfused chromosomes benefit from drive relatively more often than fused chromosomes.

Meiotic drive in males could also account for a higher rate of Y-A fusions than X-A fusions if drive favors fusions, but under these conditions Z-A fusions would establish even more often (because there are three times as many Z chromosomes as Y chromosomes, and the Z spends 2/3 of its time in males). Thus, male meiotic drive alone cannot account for the excess of Y-A fusions over any other type of fusion, all else being equal.These effects of meiotic drive are robust to modest sex biases in mutation rates and the reproductive sex ratio. Large biases can, however, cause the relative order of establishment rates to switch in a manner that is qualitatively similar to that seen previously for fusions with direct fitness effects.

In sum, meiotic drive by itself does not seem a likely explanation for the observed excess of Y-A fusions. It could generate that pattern if drive acts in females and consistently favors unfused chromosomes. Data from mammals, however, suggest that when female meiotic drive acts on fusions, it sometimes favors fused but other times unfused chromosomes (Pardo-Manuel de Villena and Sapienza, 2001a,b).

Sexually antagonistic selection — To study fusions driven by sexually antagonistic selection, we developed a model that allows for sex-differences in selection (Appendix a). We assume that an autosomal locus segregates for alleles whose frequencies are at equilibrium before the fusion appears. (This equilibrium only occurs under some fitness values [Clark 1988], and the following results apply only when those conditions are met.)

The fixation probability of a newly arisen fusion depends on several factors: which chromosome fuses with the autosome, whether the fusion originates in a male or a female, and which of the two alleles is captured by the fusion. The outcome also depends on the recombination rate in fused chromosomes between the sexually antagonistic locus and the sex-determining region; the models developed in Appendix a assume that linkage is complete. If drift is weak relative to selection, we find that fusions establish only if they are linked to the allele favoured in the sex in which the fused chromosome spends the most time, i.e., Y-A and Z-A fusions that capture a male-beneficial allele, and X-A and W-A fusions that capture a female-beneficial allele.

Interestingly, if all else is equal (no sex biases in mutation rates or the reproductive sex ratio), the establishment rate of fusions is equal for all types of sex chromosomes (Equation a.10). Sex antagonistic selection tends to favour Y-A fusions and W-A fusions more strongly than X-A and Z-A fusions because these chromosomes are consistently found in a single sex (Charlesworth and Charlesworth, 1980). This advantage, however, is exactly balanced by the lower rate that such fusions originate in the population because there are fewer Y and W chromosomes than X and Z chromosomes. Consequently, sexually antagonistic selection alone causes no difference in establishment rates.

To explain the observed excess of Y-A fusions by sexually antagonistic selection thus requires that the sexes differ in the mutation rate of fusions and/or in reproductive sex ratio (Equation a.11). Again, Y-A fusions will be particularly common if fusions originate more frequently in males. If the mutation rates are equal in males and females, however, then Y-A fusions will only be more common than X-A fusions if the reproductive sex ratio is male-biased (that is, more males than females reproduce), which is atypical. These conditions are illustrated in Figure 4.10. In general, if there is a combination of sex-biased mutation rates and biased sex ratios, Y-A fusions establish most frequently due to sexually antagonistic selection as long as $\mu^m N^m > \mu^f N^f$, where μ^m and μ^f are the female and male mutation rates, and N^m and N^f are the effective population

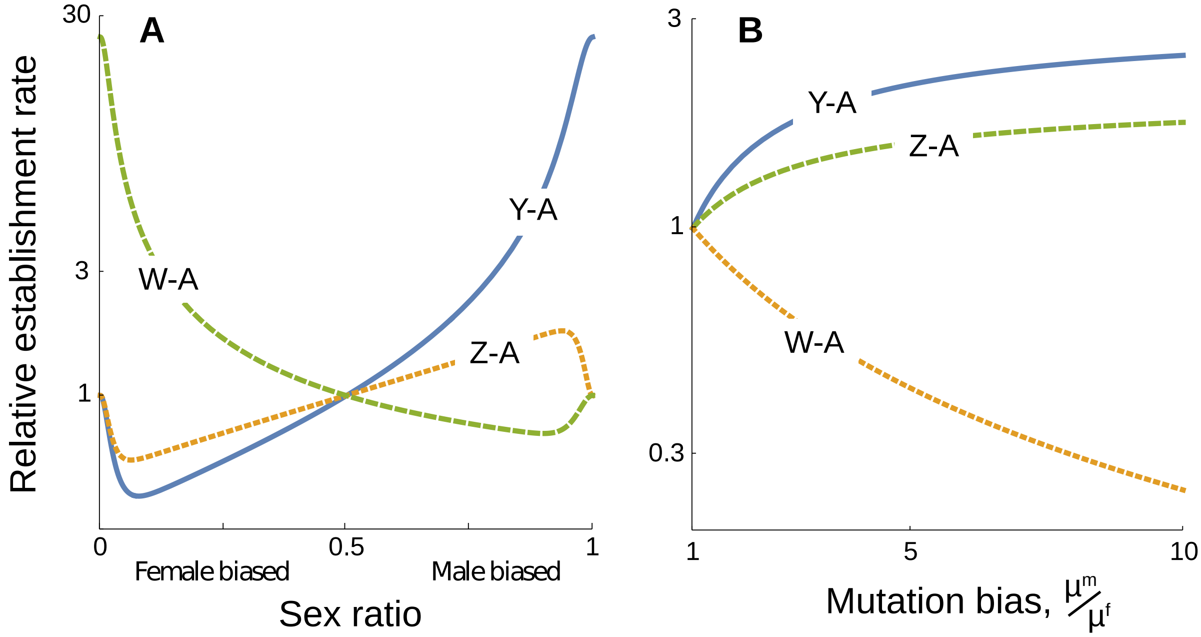

Figure 4.10: Establishment rates of sex-autosome fusions as a result of sexually-antagonistic selection, relative to the rate for X-A fusions. The fusion is assumed to be neutral except for the effects of the sexually antagonistic allele that it captures. The fittest allele in each sex has a 10% advantage when homozygous and a 9% advantage when heterozygous (results are robust to these exact numbers). (A) Effect of sex ratio bias among reproductive adults, $N^m/(N^m + N^f)$, assuming $\mu^m = \mu^f$. (B) Effect of the relative mutation rate for fusions in males versus females, μ^f/μ^m assuming $N^f = N^m$. Total effective population size $N^f + N^m = 10000$.

sizes of females and males. When this condition is met, fusions also arise more often in XY lineages than in ZW lineages.

4.5 Discussion

4.5.1 *Sex chromosome-autosome fusions are Y-biased in fishes and squamate reptiles*

A major finding in our study is that Y-autosome fusions occur more frequently than other sex chromosome fusions in vertebrates, particularly in fish and squamate reptiles. In amphibians, only one species in the database has multiple sex chromosomes, and it involves a Y-A fusion (Table 4.1). Because mammals and birds have only male heterogametic and female heterogametic systems, respectively, we could not conduct phylogenetic tests to compare the relative rates of sex chromosome fusions involving XY versus ZW chromosomes. However, there are many mammalian species with Y-A fusions, whereas there are only three avian species with fusions, supporting our conclusion that Y-A fusions tend to occur more frequently than other fusions.

Interestingly, mammals have roughly as many species with X-A fusions as with Y-A fusions, suggesting that the evolutionary forces acting on fusions may be different in mammals than in fish and reptiles. In particular, the form of female meiotic drive appears to vary among mammals, with drive favoring fused chromosomes in some species and unfused chromosomes in others, leading to a pattern in which species with X-A fusions tend to have metacentric chromosomes, while species with Y-A fusions tend to have acrocentric chromosomes (Yoshida and Kitano, 2012).

Invertebrates provide a promising system for further phylogenetic analyses, with sex chromosome variation in several groups (White, 1973; Bull, 1983; Charlesworth *et al.*, 2005; The Tree of Sex Consortium, 2014). In Diptera there are seven ZW species and 986 XY species (plus 42 XO species) in the Tree of Sex database (The Tree of Sex Consortium, 2014). Among these, there is a preponderance of fusions involving the Y: six Y-A fusions, one X-A fusion, and one species with both. Looking across all the invertebrates in the Tree of Sex database, there are many more cases of Y-A fusions (247 species) than X-A fusions (32 species), W-A fusions (8 species), and Z-A fusions (4 species); an additional 69 species have both X-A and Y-A fusions. While these data are consistent with the idea that Y-A fusions establish at a higher rate among invertebrates, a proper phylogenetic analysis is needed. A recent analysis of jumping spiders found only Y-A fusions (involving between four and seven independent events) among species that had both X and Y chromosomes (White, 1973; Maddison and Leduc-Robert, 2013). Several X-A fusions were also identified, but these occurred only in species lacking a Y. Similar analyses in other groups of invertebrates promise to shed more light on sex chromosome evolution.

4.5.2 *Accounting for the high rate of Y-A fusions*

Our theoretical analyses clarify the conditions under which fusions involving the Y chromosome are more likely to establish. Interestingly, several plausible explanations fail to account for the data. Neutral fusions could account for an excess of Y-A over X-A fusions if fusions arise more often in males, but then the theory predicts that Z-A fusions should also be common, which contradicts the data (Table 4.1, Figures 4.2 and 4.3). Beneficial fusions also cannot explain the data, as they would tend to favor the accumulation of fusions involving the X or Z, which provide more abundant targets for new fusions than the Y or W. Furthermore, hypotheses in which fusions are established because they capture sexually antagonistic alleles also fail, because the higher fixation probability of Y-A fusions capturing male-beneficial alleles or W-A fusions capturing female beneficial alleles is exactly balanced by the lower population sizes of these sex chromosomes, decreasing the rate at which Y-A and W-A fusions enter a population. To account for the preponderance of Y-A fusions thus requires more complicated explanations, involving both selection and sex biases. We consider three plausible explanations below.

Deleterious fusions with a sex biased mutation rate or reproductive sex ratio — Chromosomal fusions may often have deleterious effects because fusions can lead to the loss of genetic material, alter gene expression, or impact the rate of segregation errors (Ohno, 1967; Gardner *et al.*, 2012). Because the Y and W chromosomes have smaller effective population sizes than Z and W chromosomes, deleterious Y-A and W-A fusions are expected to fix more frequently than deleterious X-A and Z-A fusions. To account for the excess of Y-A over W-A fusions, however, requires some sort of sex bias. One promising candidate is sexual selection, which often increases the variance in reproductive success of males relative to females (Bateman's principle; Bateman, 1948). If fewer males are potentially successful as partners, the effective population size would be further reduced for the Y (but not for the W, carried by females) (Bachtrog *et al.*, 2011; Bandyopadhyay *et al.*, 2002). As a consequence, we might expect Y-A fusions to be more frequent in polygynous mating systems (Figure 4.9A).

Another promising candidate is a male-biased mutation rate. Studies in humans suggest that chromosomal translocations, a common route to fusions, are more often of paternal origin than maternal (Batista *et al.*, 1993; Thomas *et al.*, 2010; Grossmann *et al.*, 2010). By contrast, Robertsonian fusions (with two acrocentric chromosomes resulting in a fused metacentric chromosome) are more often maternal in origin (Chamberlin and Magenis, 1980; Bandyopadhyay *et al.*, 2002), but this pattern may be confounded by female meiotic drive favoring the transmission of metacentric fusions in humans (Pardo-Manuel de Villena and Sapienza, 2001a). While data from other species is needed, a preponderance of Y-A fusions can be explained if fusions are primarily deleterious and arise more often in males (Figure 4.9C). Of the three hypotheses we propose here, this may be the most compelling.

Meiotic drive — Because meiotic drive is often sex specific, it can break the symmetry between Y-A and W-A chromosomes and account for the high frequency of Y-A fusions. To do so requires female meiotic drive that selects against fused chromosomes, eliminating Z-A, W-A, and X-A fusions as they pass through female meiosis. Several cases of meiotic drive against fused chromosomes have been reported in mammals, for example in mice (Pardo-Manuel de Villena and Sapienza, 2001a,b). On the other hand, female meiotic drive favors fused chromosomes in humans (Pardo-Manuel de Villena and Sapienza, 2001a), while male meiotic favors fused chromosomes in the common shrew (Searle, 1986; Wyttenbach *et al.*, 1997). Because the nature of meiotic drive varies among taxa, it seems unlikely that one particular form—female meiotic drive against fusions—is sufficiently widespread to explain the preponderance of Y-A fusions across vertebrates, particularly among fish (Figure 4.2) and squamate reptiles (Figure 4.3). Nevertheless, meiotic drive likely plays an important role in some taxa and may underlie the variation among mammals in rates of X-A and Y-A fusions (Yoshida and Kitano, 2012).

Sexually antagonistic selection with a sex biased mutation rate — Sexually antagonistic selection is generally considered a key evolutionary factor in the turnover of sex chromosomes (Charlesworth and Charlesworth, 1980; Van Doorn and Kirkpatrick, 2007). Our models, however, indicate that fusions involving the Y will be no more common than those involving other sex chromosomes once we account for the rate that Y fusions appear in the population and the fitness they gain by capturing a male-beneficial allele. In order to break the symmetry, we must again have either a male-biased mutation rate and/or a biased reproductive sex ratio. In this case, however, the sex ratio must be male biased, with less drift among males than females so that Y-A fusions establish more frequently than W-A fusions. Assuming that sexual selection typically generates the opposite sex ratio bias, with fewer breeding males than females, sexually-antagonistic selection requires even stronger male-biased mutation to explain the preponderance of Y-A fusions, compared to an explanation based on deleterious fusions.

4.5.3 *Other considerations*

Other evolutionary forces not considered in this study may be important to the evolution of sex chromosome-autosome fusions. For example, we ignored inbreeding and spatial structure in our models. (We also did not consider fusions that capture alleles held polymorphic by heterozygote advantage, but the fate of fusions is unaffected by such loci [Charlesworth and Charlesworth 1980] unless there is inbreeding [Charlesworth and Wall 1999].) Furthermore, it is plausible that fusions may be more likely to involve some sex chromosomes for reasons that are independent of sex. For example, Y and W chromosomes often accumulate repetitive elements (Bull, 1983; Charlesworth *et al.*, 2005), which could make them more prone to fusion through nonhomologous recombination. Alternatively, the Y and W may be less likely to be captured by a fusion when they are diminutive in size relative to the X and Z. Similarly, direct selection on fusions may be chromosome specific. For example, deletions and changes to gene expression may be less problematic on degenerated Y and Z chromosomes. While our analytical results allow for mutation rates and fitness effectsto depend on the specific chromosome involved (Appendix a), our figures and conclusions were drawn assuming that there were only sex-specific and not chromosome-specific effects. As more data emerge about chromosome-specific mutation rates and selection, the analytical results can guide refinements to these conclusions.

4.6 CONCLUDING REMARKS

Using phylogenetic analyses of fish and squamate reptiles, we show that fusions between sex chromosomes and autosomes more often involve the Y than other sex chromosomes. Using population genetic models, we find that this pattern cannot be explained by models of selection unless there is also some mechanism generating a difference between the sexes, including sex-biased mutation rates, biased sex ratios, or sex-specific selection (including meiotic drive). Perhaps the most plausible hypothesis to explain the data is that fusions occur more frequently in males, are slightly deleterious, and fix by drift. Similar factors may be important to the evolution of autosome-autosome fusions. If so, we expect autosomal fusions are also typically paternal origin in origin, deleterious, and established by drift.

CHAPTER 5

Assessing the adequacy of phylogenetic models of trait evolution[5]

5.1 SUMMARY

Making meaningful inferences from phylogenetic comparative data requires a meaningful model of trait evolution. It is thus important to determine whether the model is appropriate for the data and the question being addressed. One way to assess this is to ask whether the model provides a good statistical explanation for the variation in the data. To date, researchers have focused primarily on the explanatory power of a model relative to alternative models. Methods have been developed to assess the adequacy, or absolute explanatory power, of phylogenetic trait models but these have been restricted to specific models or questions. Here we present a general statistical framework for assessing the adequacy of phylogenetic trait models. We use our approach to evaluate the statistical performance of commonly used trait models on 337 comparative datasets covering three key angiosperm functional traits. In general, the models we tested often provided poor statistical explanations for the evolution of these traits. This was true for many different groups and at many different scales. Whether such statistical inadequacy will qualitatively alter inferences drawn from comparative datasets will depend on the context. Regardless, assessing model adequacy can provide interesting biological insights—how and why a model fails to describe variation in a dataset gives us clues about what evolutionary processes may have driven trait evolution across time.

5.2 INTRODUCTION

A statistical model may provide the best explanation for a dataset compared to a few other models but still be a very poor explanation in terms of capturing the patterns of variation present in the data. For simple linear regression models, absolute model fit, or adequacy, is commonly assessed by simply plotting the data alongside the best regression line. While not quantitative, visualizing

[5] In press as: Pennell M.W., FitzJohn R.G., Cornwell W.K., and Harmon L.J. Model adequacy and the macroevolution of angiosperm functional traits. The American Naturalist.

the bivariate distribution can provide important insights regarding the fit of the model that are not captured by summaries such as the R^2 or p-value, such as whether the relationship is indeed linear (for a classic case study, see Anscombe, 1973). For these types of models, there are also a wide variety of statistical tests of model adequacy (e.g., the relationship between the residuals and the independent variable, χ^2 goodness-of-fit test, etc.) that compliment our visual intuition about model adequacy. Such formal tests used alongside informal visualizations can help researchers assess whether the inferences drawn from the fitted model are meaningful and, more interestingly, suggest how a model can be improved (Gelman and Shalizi, 2013).

Modern phylogenetic comparative methods for investigating trait evolution are almost exclusively model-based (recently reviewed in O'Meara, 2012; Pennell and Harmon, 2013), meaning that inferences are contingent on both the phylogenetic tree and the model for the traits. Selecting a good model is therefore essential for making robust inferences. Researchers typically use likelihood ratio tests or Information Theoretic measures (i.e., AIC, BIC) to select amongst models (Mooers *et al.*, 1999; Harmon *et al.*, 2010; Hunt, 2012) but these only provide a measure of relative fit. Unlike in linear regression models, for most phylogenetic models of trait evolution, it is usually very challenging to visually assess the adequacy of a model. This problem is compounded for relatively complex models such as multi-rate Brownian motion (O'Meara *et al.*, 2006; Eastman *et al.*, 2011) or multi-optima Ornstein-Uhlenbeck models (Hansen, 1997; Butler and King, 2004; Beaulieu *et al.*, 2012; Uyeda and Harmon, 2014). One can plot the trait values at the tips of the phylogeny but determining "by eye" whether this distribution is consistent with the traits having evolved under the proposed model is difficult at small scales and impossible for large phylogenies.

A number of statistical procedures have been proposed to quantitatively assess the absolute fit of a model of trait evolution (e.g., Garland *et al.*, 1992, 1993; Purvis and Rambaut, 1995; Díaz-Uriarte and Garland, 1996; Freckleton and Harvey, 2006; Boettiger *et al.*, 2012; Slater and Pennell, 2014; Beaulieu *et al.*, 2013; Blackmon and Demuth, 2014). These can be generally classified into two types of approaches. The first are tests for specific deviations from a particular model. In the early days of phylogenetic comparative biology, the focus was primarily on inferring character correlations in order to test hypotheses regarding adaptation (e.g., Felsenstein, 1985; Grafen, 1989; Harvey and Pagel, 1991; Lynch, 1991). Accordingly, a number of tests were developed to assess the reliability of assuming a Brownian motion (BM) model, which formed the basis for all phylogenetic tests of continuous character evolution at the time. Garland *et al.* (1992) proposed plotting the standardized independent contrasts (*sensu* Felsenstein, 1985) against the standard deviation of each contrast. If the contrasts and their standard deviations are correlated, this would suggest that the model (or the phylogeny) is not adequate. Purvis and Rambaut (1995) suggested using the relationship between the contrasts and the height above the root at which they were generated (see also Freckleton and Harvey, 2006, for a slight modification of this test). Similarly, Beaulieu *et al.* (2013) and Blackmon and Demuth (2014) used summary statistics to evaluate whether a set of discrete character data was consistent with some variant of a Mk model (Pagel, 1994). These are all very useful ideas, and we have adopted many of these in the method we present below, but each approach is only informative with respect to a single type of misspecification for a single type of model.

The second class of approaches is to use Monte Carlo simulations to compare an observed dataset to those expected under a model. Garland *et al.* (1993) and Díaz-Uriarte and Garland (1996) developed such an approach two decades ago. However, as this work preceded the development of analytical tools for fitting alternative (i.e., non-BM) models, the simulation parameters were not estimated directly from the data and therefore "reasonable" parameter estimates had to be chosen *a priori*. More recently, two approaches have been suggested for assessing model adequacy using parameters estimated directly from the data. Boettiger *et al.* (2012) proposed simulating data under two candidate models using the maximum likelihood parameter estimates from each model and then fitting both models to both simulated datasets. They then computed the likelihood ratio between the two candidate models for each simulating condition. After many simulations, a distribution of likelihood ratios could be obtained for each case, and these distributions compared to assess whether there was sufficient information in the data to favor one model over the other. Slater and Pennell (2014) used posterior predictive simulation (explained below) to assess the absolute fit of an "early burst" model of trait evolution, in which rates of trait evolution declined through time, compared to that of a BM model. Both Boettiger *et al.* (2012) and Slater and Pennell (2014) focused on the ability to distinguish between two models using absolute fit. Our aim here is more general: we want to compare the fit of the model to the universe of possible models.

In this paper, we propose a statistical framework for assessing the adequacy of phylogenetic models of quantitative trait evolution that generalizes previous approaches to a wide variety of alternative models. Our central thesis is that assessing model adequacy in a general way can provide valuable insights into evolutionary processes and patterns that are not evident from comparing a limited set of models. For example, one common application of phylogenetic trait models is to make inferences regarding the rate (tempo) of evolution using model selection (e.g., Mooers *et al.*, 1999; Harmon *et al.*, 2010; Hunt, 2012; Slater, 2013). Statements about rates are only informative in the context of a specific model (Hunt, 2012). It is therefore imperative to know if a model is really capturing the variation of the data in absolute terms.

In an oft-cited example of the model comparison approach, Harmon *et al.* (2010) compared three simple models of trait evolution across 49 clades and tallied the frequency with which the models were prefered in order to draw inferences about general patterns. We perform the same analysis but on a much larger scale. We analyze 337 datasets on three important angiosperm (flowering plants) functional traits using a recently published time-calibrated phylogeny (Zanne *et al.*, 2014b). We then assess the adequacy of the best-fitting model across all the datasets to determine how often one of these simple models would be adequate to make reliable inferences about rate of trait evolution.

5.3 A GENERAL FRAMEWORK FOR ASSESSING THE ADEQUACY OF PHYLOGENETIC MODELS

We focus here on models that describe the evolution of a single, continuously valued trait. More specifically, our approach works for models that predict that trait values at the tips come from a multivariate normal distribution. This applies to most models of quantitative trait evolution that have been developed to date (see below for details on the scope of the method).

If we have a phylogenetic tree consisting of n lineages and data on the trait values observed at each tip X ($X = x_1, x_2, \ldots, x_n$), we can fit a model $\mathcal{M}$ with parameters θ to describe the pattern of trait evolution along the phylogeny. There are two primary ways of fitting models to comparative data. The first is used to obtain a point estimate of θ ($\hat{\theta}$), via maximum likelihood (ML), restricted maximum likelihood (REML), least-squares, etc. The second is to estimate the posterior probability distribution $\Pr(\theta|X, \mathcal{M})$ using Bayesian approaches. For the models used in comparative biology, estimating $\Pr(\theta|X, \mathcal{M})$ requires using Markov chain Monte Carlo (MCMC) machinery to sample values of θ.

Most analyses using comparative data aim to answer one of the following questions: what values of θ best explain X given $\mathcal{M}$?; or, does $\mathcal{M}_1$ explain the data better than $\mathcal{M}_0$? Our approach is conceptually distinct in that we want to ask, how likely is it that model $\mathcal{M}$ with parameters θ would produce a dataset similar to X if we re-ran evolution?

While optimizing and Bayesian approaches to model-fitting are philosophically different from one another, our approach to assessing model adequacy is the same for both: (1) fit the model of trait evolution; (2) rescale the branch lengths of the phylogeny to place the data on a standard scale; (3) calculate a set of test statistics, $\mathcal{T}_X$, which provide statistical summmaries of the observed data; (4) simulate many new datasets $Y_1, Y_2, \ldots, Y_m$ under the model using the estimated parameters; (5) calculate test statistics on the simulated data $\mathcal{T}_{Y,1}, \mathcal{T}_{Y,2}, \ldots, \mathcal{T}_{Y,m}$; (6) compare $\mathcal{T}_X$ to the distribution of $\mathcal{T}_Y$. If $\mathcal{T}_X$ deviates significantly from the distribution of $\mathcal{T}_Y$, we can consider the model as an inadequate descriptor (see Figure 5.1)

If we have a point estimate of the model parameters, we simulate $Y_1, Y_2, \ldots, Y_m$ on the phylogeny according to $\hat{\theta}$ and $\mathcal{M}$. We then compare a single set of test statistics $\mathcal{T}_X$ calculated from our observed data to the distribution of values for $\mathcal{T}_Y$ computed across all m simulated datasets. In statistical terminology, this procedure is known as parametric bootstrapping. Parametric bootstrapping is likely familiar to phylogenetic biologists in the form of the Goldman-Cox test (Goldman, 1993) for assessing the adequacy of sequence evolution models and more recently, the phylogenetic Monte Carlo approach of Boettiger *et al.* (2012).

If we have a posterior probability distribution $\Pr(\theta|X, \mathcal{M})$, we can assess model adequacy using posterior predictive simulation (Rubin, 1984; Gelman *et al.*, 1996). We obtain new datasets by sampling from a second distribution, the posterior predictive distribution

$$\Pr(Y|X, \mathcal{M}) = \int \Pr(Y|\theta, \mathcal{M}) \Pr(\theta|X, \mathcal{M}) d\theta \tag{5.1}$$

where $\Pr(Y|X, \mathcal{M})$ is the probability of a new dataset Y given X and $\mathcal{M}$, averaged over the posterior distribution of the parameters. $\Pr(Y|X, \mathcal{M})$ can be approximated by simulating datasets using paramaters drawn from the posterior distribution. Therefore, the datasets $Y_1, Y_2, \ldots, Y_m$ are each generated from different values of θ. Posterior predictive simulation approaches have been previously developed for models in molecular phylogenetics (Bollback, 2002; Reid *et al.*,

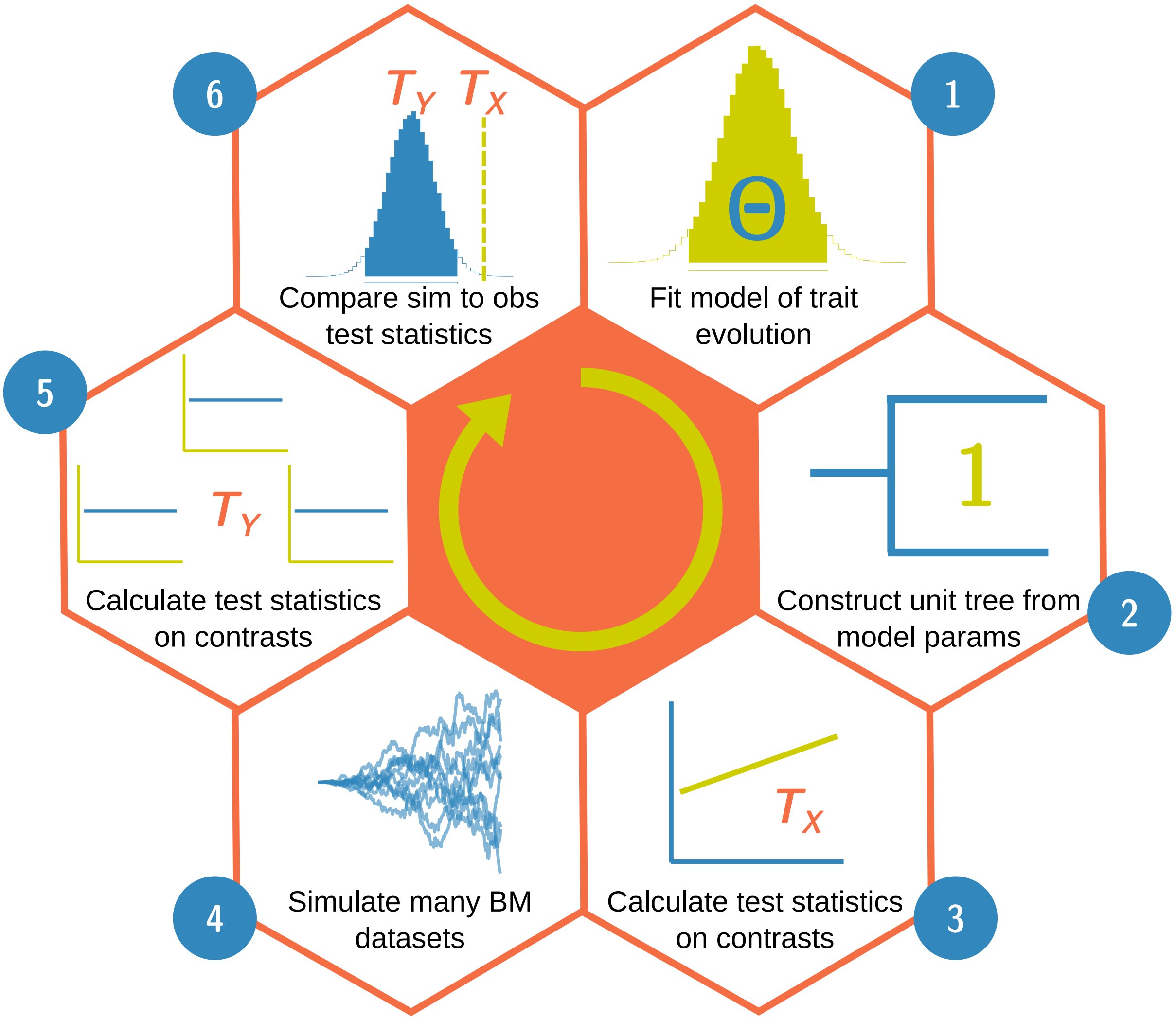
6
T_Y T_X
Compare sim to obs test statistics
1
Θ
Fit model of trait evolution
5
T_Y
Calculate test statistics on contrasts
1
Construct unit tree from model params
2
4
Simulate many BM datasets
T_X
Calculate test statistics on contrasts
3

FIGURE 5.1: Schematic diagram representing our approach for assessing model adequacy. (1) Fit a model of trait evolution to the data; (2) use the estimated model parameters to build a unit tree; (3) compute the contrasts from the data on the unit tree and calculate a set of test statistics $\mathcal{T}_X$; (4) simulate a large number of datasets on the unit tree, using a BM model with $\sigma^2 = 1$; (5) calculate the test statistics on the contrasts of each simulated dataset $\mathcal{T}_Y$; and (6) compare the observed and simulated test statistics. If the observed test statistic lies in the tails of the distribution of simulated test statistics the model can be rejected as inadequate. The rotational circle in the center of the diagram indicates that assessing model adequacy is an iterative process. If a model is rejected as inadequate, the next step is to propose a new model and repeat the procedure.

2014; Lewis *et al.*, 2014; Brown, 2014), and recently for PCMs (Slater and Pennell, 2014), but have not been widely adopted in either field.

5.3.1 *Test statistics*

No simulated dataset will ever be exactly the same as our observed dataset. We therefore need to choose informative test statistics in order to evaluate whether the model predicts datasets that are similar to our observed dataset in meaningful ways. As the states at the tips of the phylogeny are not independent—this is why we are using PCMs in the first place!—calculating test statistics on the data directly is not generally informative for models in comparative biology. We account for the non-independence of the observed data by calculating test statistics on the set of contrasts (i.e., "phylogenetically independent contrasts"; Felsenstein, 1985) computed at each node. (We refer readers to Felsenstein, 1985; Rohlf, 2001; Blomberg *et al.*, 2012, for details on how contrasts are calculated.) Under Brownian motion (BM) the contrasts will be independent and identically distributed (i.i.d.) according to a normal distribution with mean 0 and standard deviation σ, i.e., contrasts are $\sim \mathcal{N}(0, \sigma)$, where σ^2 is the BM rate parameter (Felsenstein, 1985). This i.i.d. condition allows us to perform standard statistical tests on the contrasts.

The choice of what test statistics to use for assessing model adequacy is ultimately one of balancing statistical intuition and computational effort. We have chosen the following set of six test statistics to compute on the contrasts because they capture a range of possible model violations and have well-understood statistical properties. All of these essentially evaluate whether the contrasts come from the distribution expected under BM.

M_{SIG} The mean of the squared contrasts. This is equivalent to the REML estimator of the Brownian motion rate parameter σ^2 (Garland *et al.*, 1992; Rohlf, 2001). M_{SIG} is a metric of overall rate. Violations detected by M_{SIG} indicate whether the overall rate of trait evolution is over- or underestimated.

C_{VAR} The coefficient of variation (standard deviation/mean) of the absolute value of the contrasts. If C_{VAR} calculated from the observed contrasts is greater than that calculated from the simulated contrasts, it suggests that we are not properly accounting for rate heterogeneity across the phylogeny. If C_{VAR} from the observed is smaller, it suggests that contrasts are more even than the model assumes. We use the coefficient of variation rather than the variance because the mean and variance of contrasts can be highly correlated.

S_{VAR} The slope of a linear model fit to the absolute value of the contrasts against their expected variances (following Garland *et al.*, 1992). Each (standardized) contrast has an expected variance proportional to the sum of the branch lengths connecting the node at which it is computed to its daughter lineages (Felsenstein, 1985). Under a model of BM, we expect no relationship between the contrasts and their variances. We use S_{VAR} to test if contrasts are larger or smaller than we expect based on their branch lengths. If, for example, more evolution occurred per unit time on short branches than long branches, we would observe a negative slope. If S_{VAR} calculated from the observed data deviates substantially from the expectations, a likely explanation is branch length error in the phylogenetic tree.

S_{ASR} The slope of a linear model fit to the absolute value of the contrasts against the ancestral state inferred at the corresponding node. We estimated the ancestral state using the least-squares method suggested by Felsenstein (1985) for the calculation of contrasts. (We note that this is not technically an ancestral state reconstruction [see Felsenstein, 1985]; it is more properly thought of as a weighted average value for each node.) We used this statistic to evaluate whether there is variation in rates relative to the trait value. For example, do larger organisms evolve proportionally faster than smaller ones?

S_{HGT} The slope of a linear model fit to the absolute value of the contrasts against node depth (after Purvis and Rambaut, 1995). This is used to capture variation relative to time. It is alternatively known as the "node-height test" and has been used to detect early bursts of trait evolution during adaptive radiations (see Freckleton and Harvey, 2006; Slater and Pennell, 2014, for uses and modifications of this test).

D_{CDF} The D-statistic obtained from Kolmolgorov-Smirnov test from comparing the distribution of contrasts to that of a normal distribution with mean 0 and standard deviation equal to the root of the mean of squared contrasts (the expected distribution of the contrasts under BM; see Felsenstein, 1985; Rohlf, 2001). We chose this to capture deviations from normality. For example, if traits evolved via a "jump-diffusion" type process (Landis *et al.*, 2013), in which there were occasional bursts of rapid phenotypic evolution (Pennell *et al.*, 2014b), the tip data would no longer be multivariate normal owing to a few contrasts throughout the tree being much larger than the rest (i.e., the distribution of contrasts would have heavy tails).

Alternative test statisics are certainly possible. One could, for instance, calculate the median of the squared contrasts, the skew of the distribution of contrasts, etc. If the generating model was known, we could use established procedures for selecting a set of sufficient (or, approximately sufficient; Joyce and Majoram, 2008) test statistics for that model, as is typically done when computing likelihood ratio tests. However, the aim of our approach is to assess the fit of a proposed model without reference to a true model. Our test statistics will detect many types of model misspecification but this does not mean that they will necessarily detect every type of model misspecification. We encourage researchers interested in specific questions to explore alternative test statistics that capture deviations relevant to the problem at hand.

An additional challenge is determining how to deal with the statistical problems (i.e., inflated Type-1 error rates) that may be introduced when using many test statistics. In our analyses, we chose not to correct our p-values for multiple comparisons (using Bonferroni, false discovery rates, etc.). We did this for a number of reasons. First, our tests are not truly independent and the degree of correlation between test statistics will necessarily depend on the "true" model of trait evolution. Second, as argued by Gelman (2006), we might be interested in the specific aspects of the data that differ from the expectations under the model; rather than focus on whether a model should be accepted or rejected, we "want to understand the limits of its applicability in realistic replications" (p. 175).

Beyond Brownian motion

All of our test statistics are designed to evaluate the adequacy of a BM model of trait evolution. However, if we propose a different model for the evolution of the trait, such as an Ornstein-Uhlenbeck (OU; Hansen, 1997) process, then the expected distribution of the contrasts is different. For example, under an OU model, contrasts will not be i.i.d. (Hansen, 1997). The expected distribution of contrasts under most models of trait evolution, aside from BM, is not formally characterized and even if it was, this would necessitate a specific set of test statistics for every model proposed.

Our solution to this problem is to create what we term a "unit tree", which is a phylogenetic tree transformation that captures the dynamics of trait change under a particular evolutionary model. For a particular evolutionary model $\mathcal{M}$ (with parameter values θ), we define a unit tree as a phylogenetic tree that has the following property: the length of each branch is equal to the amount of variance expected to accumulate over it under $\mathcal{M}, \theta$. The variance is standardized, such that the expected distribution of the trait data on the unit tree is equal to that of a Brownian Motion (BM) model with a rate σ^2 equal to 1.

If the fitted model is adequate, the trait data at the tips of the unit tree will have the same distribution as data generated under a BM process with a rate of 1 and the contrasts will be distributed according to a standard normal distribution (hence the name, unit tree). Creating the unit tree from the estimated model parameters prior to computing the contrasts generalizes the test statistics to most models of quantitative trait evolution (but see Landis *et al.*, 2013; Schraiber and Landis, 2014, for exceptions).

We also emphasize that because the contrasts are calculated on the unit tree, the test statistics all must depend on both the data and the model; for this reason, the Bayesian version of our approach produces a distribution of observed test statistics. Once we have created the unit tree from the estimated parameters, new datasets can be simulated under the model simply using a BM process with $\sigma^2 = 1$, which has the added benefit of being computationally efficient. The distribution of test statistics calculated on these simulated data sets can then be compared to the test statistics from the observed data.

5.3.2 *Details of unit tree construction and the scope of this approach*

Here we formalize our definition of the unit tree and delimit the scope of our approach. Readers can skip this section without missing the main point. A unit tree can be constructed from any evolutionary model where the trait has expected variance-covariance matrix $\mathbf{V}$ that satisfies the (generalized) 3-point condition proposed by Ho and Ané (2014) and the data follows a multivariate normal distribution. A matrix $\mathbf{V}$ has a strict 3-point structure if the following condition holds: for any lineages i, j, k, the two smallest of V_{ij}, V_{ik}, V_{jk} are equal. Under a simple BM model it is straightforward to show that this condition holds. If $\mathbf{C}$ is the matrix representation of the phylogeny (such that C_{ij} is the shared path length between lineages i and j), then by the nature of the tree structure, the 3-point condition will hold for $\mathbf{C}$. Since under BM $\mathbf{V} = \sigma^2\mathbf{C}$, then **V** will also be 3-point structured. The same holds true for any evolutionary model that is a branch length transformation of a BM model including the λ, δ, κ models (Pagel, 1997, 1999) and models where rates change through time (the "Early Burst" or EB model, also referred to as the Accelerating/Decelerating Change, ACDC, model; Blomberg *et al.*, 2003; Harmon *et al.*, 2010) or across the tree (O'Meara *et al.*, 2006; Thomas *et al.*, 2006; Eastman *et al.*, 2011; Revell *et al.*, 2012; Thomas and Freckleton, 2012). Standard error can be incorporated into any of these models by simply adding a species-specific scalar to each element of the diagonal. For all of the models where the 3-point condition applies, we can construct a unit tree by setting the length v of the edge $\{(i, j), k\}$ connecting the most recent common ancestor (MRCA) of lineages i and j to the MRCA of lineages i and k to be

$$v_{\{(i,j),k\}} = V_{ij} - V_{ik} \tag{5.2}$$

where V_{ij} and V_{jk} are, by the requirements of the 3-point structured condition, equal to one another. Once all branches have been transformed, the contrasts computed on the unit tree will be i.i.d. ~ $\mathcal{N}(0,1)$ under the model in question.

The OU model of trait evolution also generates 3-point structured matrices when the tree is ultrametric; this is true of both single optimum and multi-optima models (Ho and Ané, 2014). However, while the variance structure can easily be transformed to a BM-like tree, the contrasts on this tree will not necessarily be distributed according to a standard normal. For example, while it is often assumed when fitting a single regime OU model that the ancestor is at the optimum trait value (see, for example Harmon *et al.*, 2010), this need not be the case. Furthermore, if there are multiple optima on the phylogeny (Hansen, 1997; Butler and King, 2004; Ingram and Mahler, 2013; Uyeda and Harmon, 2014), lineages will necessarily be tracking optima that are different from the root state. Therefore, a transformation must also be made to the data in addition to the branch lengths of the phylogeny to produce contrasts that have are i.i.d. according to a standard normal.

To accomplish this, we again turn to the recent work of Ho and Ané (2014). In addition to 3-point structured matrices, Ho and Ané defined a broader condition: a matrix of the form

$$\mathbf{V} = \mathbf{D}_1\widetilde{\mathbf{V}}\mathbf{D}_2$$

is considered to have a generalized 3-point structure if $\widetilde{\mathbf{V}}$ is 3-point structured and $\mathbf{D}_1$ and $\mathbf{D}_2$ are diagonal matrices. Ho and Ané (2014) prove that many phylogenetic models are indeed of this class, including multi-optimum OU models (Butler and King, 2004; Ingram and Mahler, 2013; Uyeda and Harmon, 2014), those with varying rates and models across the tree (e.g., Beaulieu *et al.*, 2012) as well as to OU models fit to non-ultrametric trees. For any model that satisfies the generalized 3-point condition and where the data are assumed to come from a multivariate normal distribution, there exists some transformation to the tree (appling Equation 5.2 to $\widetilde{\mathbf{V}}$) and data (using $\mathbf{D}_1$ and $\mathbf{D}_2$) that will produce a unit tree with standard normal contrasts. We note that Slater (2014) recently pointed out that for OU models fit to non-ultrametric trees, there is no valid transformation that can make $\mathbf{V}$ BM-like. While this is indeed correct, it is however, possible to get a BM-like tree by adding a species-specific scalar to the data matrix (Ho and Ané, 2014). Therefore, once the proper tree and data transformations have been made, all the test statistics described above can apply.

The above also applies to phylogenetic regression models (Grafen, 1989; Lynch, 1991; Martins and Hansen, 1997) of the form

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

In these models, the error variance is structured by phylogeny assuming some model of trait evolution such that $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$. In these regression models $\mathbf{V}$ represents the variance-covariance matrix of the residuals rather than the traits (Rohlf, 2001). Therefore if $\mathbf{V}$ is either 3-point or generalized 3-point structured, the tree (and possibly data) can be transformed such that the contrasts on the residuals will be i.i.d. standard normal. This fact allows researchers to use our approach to assess the adequacy of a trait model for understanding correlations between traits. We note however that as $\mathbf{V}$ only affects the error structure for these models, alternative approaches (see for example Gelman *et al.*, 2003, ch. 6) will be required to assess the adequacy of the mean structure $Y = \beta_0 + \beta_1 X$ of the model.

5.3.3 *Simulations*

As a verification of our method, we conducted a brief simulation study. We focused here on assessing Type-1 error rates. As above, we emphasize that these are not necessarily the most important quantities when thinking about model adequacy, but they do provide a useful metric for demonstrating that our code is functioning correctly. The philosophy behind approaches such as ours is that the "true" model is outside of the candidate set. We want to ask whether a given model can adequately describe the variaton in the data. If it does, we can consider it statistically adequate even if it is not the true model or even the best model in our set (see Discussion for comments on the relationship between model adequacy and model selection). Furthermore, while it is certainly interesting to examine what types of deviations in model space produce what types of deviations in the various test statistics, the number of possible simulation conditions is infinitely broad.

We simulated data under BM, single-optimum OU, and EB (the same models we used in the analysis; see below). For each set of conditions, we simulated trees of 50, 100, and 200 taxa under a pure-birth process, then rescaled the tree to be unit height. For BM, we set $\sigma^2 = 1$. For OU, we used $\sigma^2 = 1$ but varied the "selection" parameter α ($\alpha = \{1,2,4\}$). For EB, we again set σ^2 to be 1 and varied a, the exponential rate of decline (see Harmon *et al.*, 2010; Slater and Pennell, 2014, for details), which was set to be $a = \{\log(0.01), \log(0.02), \log(0.04)\}$. For each parameter combination, we ran 500 simulations under two sets of conditions: (1) assuming no measurement error; and (2) assuming known error rates of 5% of the expected variance in trait values across the phylogeny. We then fit the corresponding model using maximum likelihood and evaluated the Type-1 error under each set of conditions. All simulations were conducted using DIVERSITREE (FitzJohn, 2012a).

5.4 THE ADEQUACY OF MODELS FOR THE EVOLUTION OF PLANT FUNCTIONAL TRAITS

5.4.1 *Data*

We used a phylogeny of angiosperms, containing 30,535 species, from a recent study by Zanne *et al.* (2014b). We conducted all analyses on the MLE of the phylogeny (available on DRYAD, doi:10.5061/dryad.63q27/3). We used existing large datasets on three functionally important plant traits: specific leaf area (SLA, defined as fresh area/dry mass), seed mass, and leaf nitrogen content (% mass). Seed mass is a crucial part of species' life-history strategy (Leishman *et al.*, 2000; Westoby *et al.*, 2002) and SLA and leaf nitrogen content are important and widely measured components of species' carbon capture strategies (Wright *et al.*, 2004). Understanding the macroevolutionary patterns of these three traits can provide key insights into the evolutionary processes that have shaped much of plant diversity (Cornwell *et al.*, 2014). The SLA and leaf nitrogen data comes from Wright *et al.* (2004) with additional SLA data from the LEDA project (Kleyer *et al.*, 2008). Seed mass data comes the Kew database (Royal Botanical Gardens, Kew, 2014). We used an approximate grepping approach to find and correct spelling mistakes and synonymy tools from The Plant List (2014) to match the trait databases to the Zanne et al. phylogeny. The full data set includes 3293 species for SLA, of which 2200 match species in the Zanne *et al.* (2014b) tree. For seed mass, the dataset included 22,817 species with 11,107 matched the phylogeny. For leaf nitrogen content, we have data for 1574 species with 936 included in the tree. See `https://github/richfitz/modeladequacy` for specific locations and scripts to access and process the original data.

We log-transformed all data prior to analysis. We did this for biological reasons rather than to conform the data to the assumptions of the model (Houle *et al.*, 2011). It is more meaningful to model trait evolution as a multiplicative process rather than an arithmetic one. An increase of two grams is much more significant for the seed of an orchid than the seed of a palm tree. However, we should recognize that both of these rationales are essentially statements about model adequacy and thus the validity of the log transformation can be quantitatively assessed.

Because the vast majority of the species are only represented by a single record, it was not possible to use a species-specific estimate of trait standard error (SE) to account for either measurement error or intraspecific variation. As an alternative, we estimated a single SE for each trait by calculating the mean standard deviation for all species for which we had multiple measurements. The assumption of a constant SE across all species is unlikely to be correct, but even an inaccurate estimate of error is better than assuming none at all (Hansen and Bartoszek, 2012).

5.4.2 *Analysis*

We first matched our trait data to the whole phylogeny and then extracted subclades from this dataset in a three ways: (1) by family; (2) by order; and (3) by cutting the tree at 50 My intervals and extracting the most inclusive clades (named or unnamed) for which the most recent common ancestor of a group was younger than the time-slice. (The crown age of angiosperms is estimated to be ~243 my in the MLE tree and the tree was cut at 50, 100, 150, and 200my.) We kept only subclades for which there was at least 20 species present in both the phylogeny and trait data so that we had a reasonable ability to estimate parameters and distinguish between models (Boettiger *et al.*, 2012; Slater and Pennell, 2014). For SLA, this left us with 72 clades, seed mass, 226 clades, and leaf nitrogen content, 39 clades (337 in total). We note that these datasets are not independent as many of the same taxa were included in family, order and multiple time-slice subtrees.

Following Harmon *et al.* (2010), we considered three simple models of trait evolution: (1) BM, which can be associated with genetic drift (Lande, 1976; Felsenstein, 1988; Lynch and Hill, 1986; Lynch, 1990; Hansen and Martins, 1996), randomly varying selection (Felsenstein, 1973), or the summation of many independent processes over macroevolutionary time (Hansen and Martins, 1996; Uyeda *et al.*, 2011; Pennell *et al.*, 2014b); (2) single optimum OU, which is often assumed to represent stabilizing selection (following Lande, 1976), though we think a more meaningful interpretation is that it represents an "adaptive zone" (Hansen, 2012; Pennell and Harmon, 2013); and (3) EB, which was developed as a phenomenological representation of a niche-filling process during an adaptive radiation (Blomberg *et al.*, 2003; Harmon *et al.*, 2010). We fit each of these models to all 337 subclades in our dataset. We then used the approach we developed to assess the adequacy of each fitted model.

All of the analyses conducted in this paper were conducted using both likelihood and Bayesian inference. We did so to demonstrate the scope of our approach and because both ML and Bayesian inference are commonly used in comparative biology. We emphasize that our approach is not tied to any single statistical paradigm.

For the likelihood analyses, we fit the three models (BM, OU, and EB) using ML with the DIVERSITREE package (FitzJohn, 2012a). We calculated the AIC score for each model. We then constructed a unit tree for each subtree, trait and model combination using the maximum likelihood estimates of the parameters. We calculated the six test statistics described above (M_{SIG}, C_{VAR}, S_{VAR}, S_{ASR}, S_{HGT}, D_{CDF}) on the contrasts of the data. We simulated 1000 datasets on each unit tree using a BM model with $\sigma^2 = 1$ and calculated the test statistics on the contrasts of each simulated data set.

For the Bayesian analysis, we fit the same models as above using a MCMC approach, sampling parameter values using slice sampling (Neal, 2003), as implemented in DIVERSITREE (FitzJohn, 2012a). For the BM model we set a broad uniform prior on $\sigma^2 \sim \mathcal{U}[0, 2]$, the upper bound being substantially larger than the ML estimate of σ^2 for any clade. For the OU model, we used the same prior for σ^2 and drew α values, the strength of attraction to the optimum, from a Lognormal($\mu = \log(0.5), \sigma = \log(1.5)$) distribution. A complication involved in fitting OU models is deciding what assumptions to make about the state at the root z_0. Here, we follow other authors (Butler and King, 2004; Harmon *et al.*, 2010) and assume that z_0 is at the optimum. For the EB model, we again used the same prior for σ^2 and a uniform prior on a, the exponential rate of decrease in σ^2, such that $a \sim \mathcal{U}[-1, 0]$ (the minimum value is much more negative than we would typically expect; Slater and Pennell, 2014).

Again, for each model/trait/subtree combination, we ran a Markov chain for 10,000 generations. Preliminary investigations demonstrated that this was more than sufficient to obtain convergence and proper mixing for these simple models. After removing a burn-in of 1000 generations, we calculated the Deviance Information Criterion (DIC), a Bayesian analog of AIC (Spiegelhalter *et al.*, 2002), for each model. We drew 1000 samples from the joint posterior distribution. For each of the sampled parameter sets, we used the parameter values to construct a unit tree and calculated our six test statistics on the contrasts. We then simulated a dataset on the same unit tree and calculated the test statistics on the contrasts of the simulated data.

In the likelihood analyses, for each dataset, we had one set $\mathcal{T}_X$ of observed test statistics and a 1000 sets $\mathcal{T}_{Y,1}, \mathcal{T}_{Y,2}, \ldots, \mathcal{T}_{Y,1000}$ of test statistics calculated on data simulated on the same unit tree. In the Bayesian version, we had 1000 sets of observed test statistics $\mathcal{T}_{X,1}, \mathcal{T}_{X,2}, \ldots, \mathcal{T}_{X,1000}$ using a different unit tree for each set and 1000 sets of simulated test statistics $\mathcal{T}_{Y,1}, \mathcal{T}_{Y,2}, \ldots, \mathcal{T}_{Y,1000}$, each $\mathcal{T}_{Y,i}$ corresponding to the unit tree used to compute $\mathcal{T}_{X,i}$.

For both types of analyses, we report two-tailed p-values (i.e., the probability that a simulated test statistic was more extreme than the observed). As a multivariate measure of model adequacy, we calculated the Mahalanobis distance, a scale-invariant metric, between the observed test statistics and the mean of our simulated test statistics, taking into account the covariance structure between the simulated test statistics. We took the log of the KS D-statistic, D_{CDF}, as the Mahanalobis measure assumes data is multivariate normal and the D-statistic is bounded between 0 and 1. For the Bayesian analyses, we report the mean of the distribution of Mahalanobis distances. All analyses were conducted in R v3.0.2 (R Development Core Team, 2013). Scripts to fully reproduce all analyses are available at `https://github.com/richfitz/modeladequacy`.

5.4.3 *A case study: seed mass evolution in the Meliaceae and Fagaceae*

As an illustration of our approach, we present a case study examining seed mass evolution in two tree families, the Meliaceae, the "mahogany family", and Fagaceae, which contains oaks, chestnuts and beech trees. The trait data and phylogeny for both groups are subsets of the larger dataset used in the analysis. Superficially, these datasets are quite similar. Both are of similar size (Meliaceae: 44 species in the dataset, 550 in the clade; Fagaceae: 70 species in the dataset and 600 in the clade), age (crown age of Meliaceae: ~53my; Fagaceae: ~40my) and are ecologically comparable in terms of dispersal strategy and climatic niche.

As described above, we fit three simple models of trait evolution (BM, OU, EB) to both datasets using ML and computed AIC weights (AIC_w; Akaike, 1974; Burnham and Anderson, 2004a) for the three models. For both datasets, an OU model was overwhelmingly supported ($\text{AIC}_w > 0.97$ for both groups). Therefore, looking only at relative model support, we might conclude that similar evolutionary processes are important in these two clades of trees.

Examining model adequacy provides a different perspective. We took the MLE of the parameters from the OU models for each dataset and constructed a unit tree based on those parameters. We calculated our six test statistics on the contrasts of the data, then simulated 1000 datasets on the unit tree and calculated the test statistics on the contrasts of each simulated dataset (Figure 5.2). For seed mass evolution in Meliaceae, the OU model was an adequate model; all six observed test statistics were in the middle of the distribution of simulated test statistics ($M_{\text{SIG}} : p = 0.921$, $C_{\text{VAR}} : p = 0.605$, $S_{\text{VAR}} : p = 0.979$, $S_{\text{ASR}} : p = 0.485$, $S_{\text{HGT}} : p = 0.170$, $D_{\text{CDF}} : p = 0.657$). In contrast, for Fagaceae we found that the test statistics calculated with an OU model lay outside the expected values for S_{VAR} ($p \approx 0$) and S_{HGT} ($p = 0.014$) suggesting that the process of evolution that gave rise to these data was more complex than that captured by a simple OU process.

More specifically, the slope of the contrasts against their variances S_{VAR} is negative, meaning that contrasts computed on short branches are larger than expected (or conversely, contrasts computed on long branches are too small). Such a pattern could be generated by phylogenetic error: the terminal branches in the Fagaceae tree are very short and are likely underestimated relative to the longer internal branches. This explanation is further supported by the fact that S_{HGT} is also negative—the standardized contrasts close to the tips are much larger than expected. The rest of the observed test statistics did not differ significantly from the simulated test statistics

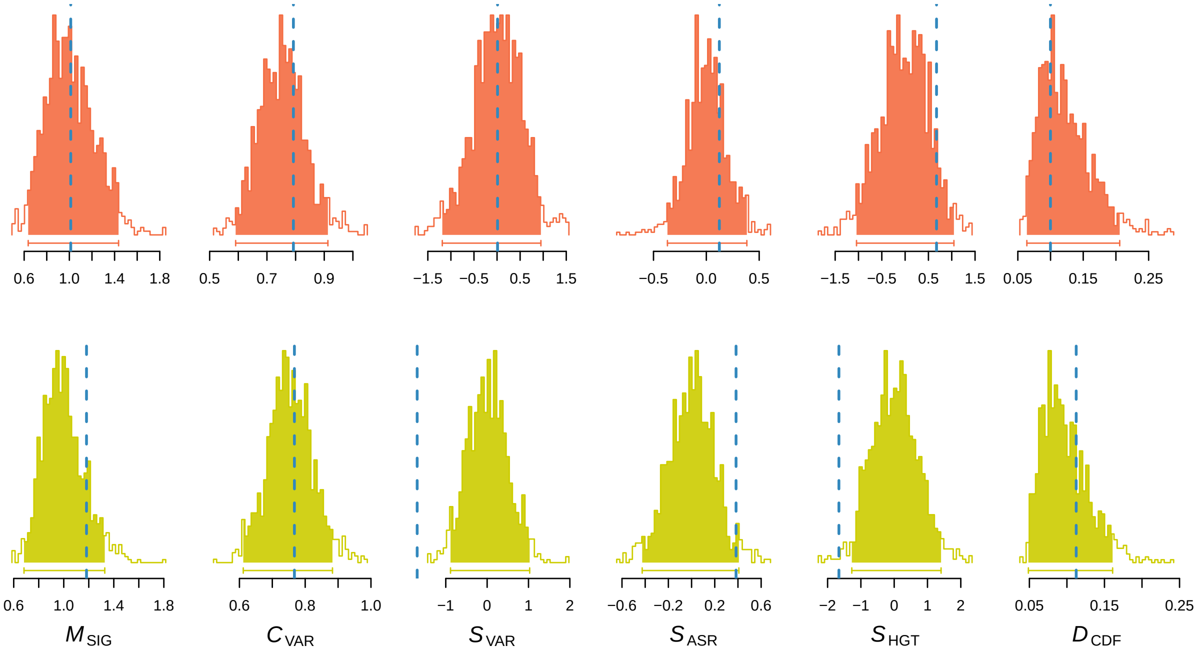

FIGURE 5.2: Illustration of our approach to model adequacy. We fit three models (BM, OU, and EB) to seed mass data from two different tree families, the Meliaceae (top panel, red) and the Fagaceae (bottom panel, yellow). In both cases, an OU model (analyzed here) was strongly supported when fit with ML. The plotted distributions are the test statistics (M_{SIG}, C_{VAR}, S_{VAR}, S_{ASR}, S_{HGT}, D_{CDF}) calculated from the contrasts of the simulated data; the bars underneath the plots represent 95% of the density. The dashed vertical lines are the values of the test statistics calculated on the contrasts of the observed data. Using our test statistics, an OU model appears to be an adequate model for the evolution of seed mass in the Meliaceae; for all of the test statistics, the observed test statistic lies in the middle of the distribution of simulated test statistics. For the Fagaceae, the slopes of the contrasts against their expected variances S_{VAR} and node height S_{HGT} are much lower than the expectations under the model.

($M_{\mathrm{SIG}} : p = 0.298$, $C_{\mathrm{VAR}} : p = 0.837$, $S_{\mathrm{ASR}} : p = 0.074$, $D_{\mathrm{CDF}} : p = 0.551$). This example illustrates the distinction between the conventional approach to model selection in PCMs and model adequacy. Selecting amongst a limited pool of models does not give a complete picture of the amount of variation that a chosen model is actually capturing.

5.5 RESULTS

5.5.1 *Simulations*

In our simulations, we found that when we assessed the adequacy of the generating model, all of the test statistics showed Type-1 errors that were consistently around or less than 0.05. This was true across models, parameters, tree sizes and did not depend on whether we included a known SE or not (Figures 5.3, 5.4, and 5.5). These results demonstrate that our unit tree construction is working properly; if the MLE is equal to the generating value, then the constrasts will be i.i.d. $\mathcal{N}(0, 1)$ and standard normal statistical properties will apply. Some of the test statistics are very conservative (have very low Type-1 error rates) under some models. We are not aware of any general statistical theory that will allow us to predict the conditions under which a test statistic will have low power to detect deviations from the expected distributions. However, there is an intiutive explanation for this pattern. Consider for example, our test statistic M_{SIG}. As mentioned above, this is equivalent to the REML estimate of σ^2. When we fit BM (or, a more general model, of which BM is a special case), and then rescale the tree with $\hat{\sigma^2}$, the observed contrasts on the unit tree will effectively be minimized with respect to this quantity and all of the contrasts on the simulated dataset will tend to be larger than our observed contrasts. So if the quantity captured by the test statistic is tightly correlated with one of the parameters being optimized in the model, this test statistic will tend to have low power to detect deviations from this model.

We also found that by using multiple test statistics and reporting a Type-1 error if *any* of the test statistics deviated significantly from expectations, the error rate increased substantially (up to around 20% under some conditions). However, as we discuss above, we do not think that this is necessarily a defect of the analysis and are not overly concerned with this error rate. Looking at what test statistics were violated and how they were violated is much more meaningful than simply rejecting or accepting a model based on the overall p-value. Furthermore, the degree to which the Type-1 error rate will rise with multiple comparisons will be a complex function of the

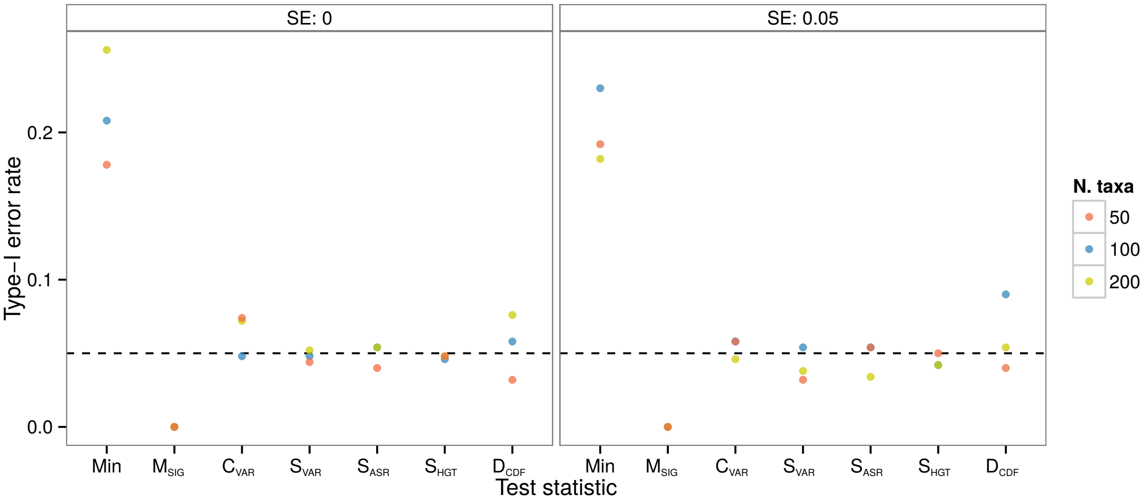

Figure 5.3: Type-1 error rates for data simulated under a Brownian motion (BM) model. We simulated 500 datasets under for 3 different tree sizes ($N = \{50, 100, 200\}$, represented by the different colors) and two known values of standard error of observed species means (0 and 0.05, left and right panel, respectively). The Type-1 error rates for each test statistic are consistently around or lower than a 0.05 threshold. However, the frequency at which *at least one* of the test statistics deviated significantly from the expectations (the variable "Min" on the left side of each plot) was substantially greater, rising to above 20% in some cases. See text for why we decided against correcting for the effect of multiple comparisons in the analysis.

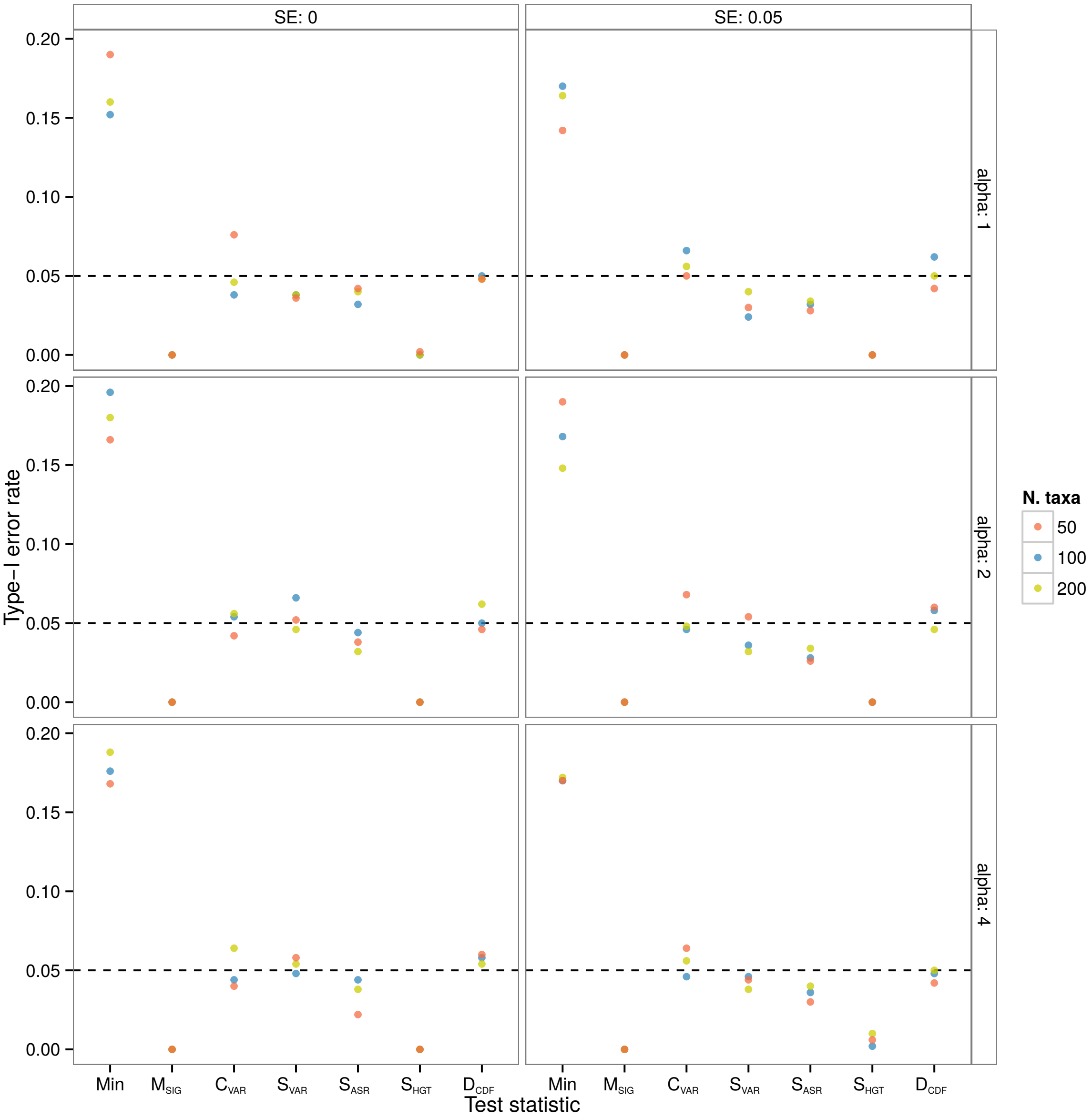

FIGURE 5.4: Type-1 error rates for data simulated under an Ornstein-Uhlenbeck (OU) model. We simulated 500 datasets under for 3 different tree sizes ($N = \{50, 100, 200\}$, represented by the different colors) and two known values of standard error of observed species means (0 and 0.05, left and right panel, respectively). We also simulated under three values for the α parameter ($\alpha = \{1,2,4\}$, top, middle and bottom panel), representing phylogenetic half-lives of 69%, 35%, 17% of total tree depth, respectively. The Type-1 error rates for each test statistic are consistently around or lower than a 0.05 threshold. However, the frequency at which *at least one* of the test statistics deviated significantly from the expectations (the variable "Min" on the left side of each plot) was substantially greater, approaching 20% in some cases. See text for why we decided against correcting for the effect of multiple comparisons in the analysis.

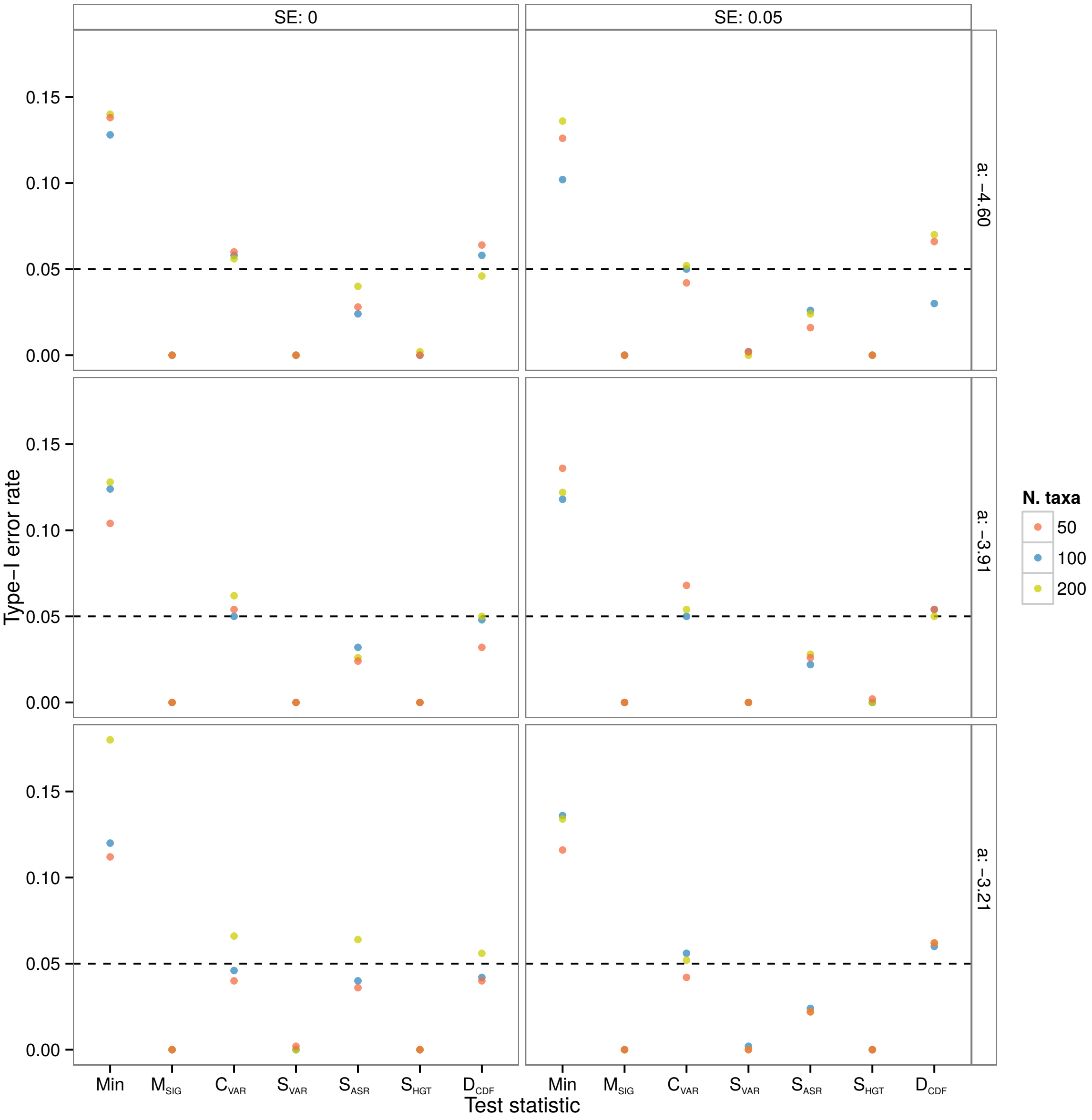

FIGURE 5.5: Type-1 error rates for data simulated under an Ornstein-Uhlenbeck (OU) model. We simulated 500 datasets under for 3 different tree sizes ($N = \{50, 100, 200\}$, represented by the different colors) and two known values of standard error of observed species means (0 and 0.05, left and right panel, respectively). We also simulated under three values for the exponential rate of slowdown, a ($a = \{\log(0.01), \log(0.02), \log(0.04)\}$, top, middle and bottom panel), which translate to the rate of trait evolution halfing every 0.15, 0.17, and 0.21 time units, respectively (note that the tree was scaled so the total depth was equal to unity). The Type-1 error rates for each test statistic are consistently around or lower than a 0.05 threshold. However, the frequency at which *at least one* of the test statistics deviated significantly from the expectations (the variable "Min" on the left side of each plot) was substantially greater, approaching 15% in some cases. See text for why we decided against correcting for the effect of multiple comparisons in the analysis.

generating model and the size of the dataset and there is no suitable general correction that we know of.

5.5.2 *Models for the evolution of angiosperm functional traits*

Our results for likelihood and Bayesian inference were broadly similar; for conciseness, we present only the results from the likelihood analyses here. Results from the Bayesian analysis are presented APPENDIX B. Full results from all analyses can be reproduced using code and workflows available at `https://github.com/richfitz/modeladequacy`.

Across the 337 subclades, we found widespread support for OU models. For 236 clades, OU had the highest AIC_w. OU had ~100% of the AIC_w in 27 clades and >75% of the weight in 189 clades (figure 5.6). Consistent with Harmon *et al.* (2010) we found that EB models rarely had high support (only 6 clades supported EB with >75% AIC_w), though we acknowledge that even if early burst dynamics were important to long-term patterns of trait evolution, these may be difficult to detect with extant species alone (Slater *et al.*, 2012a; Slater and Pennell, 2014). Larger clades commonly had very high support for a single model (of the 101 clades consisting of more than 100 taxa, 44 had >90% of the AIC weight on a single model), and that was overwhelmingly likely to be an OU model (42/44 clades).

We limit our analyses of model adequacy to only the most highly supported model in the candidate set, as supported by AIC. We did this to present a best-case scenario; if a model had very little relative support, it would be unremarkable if it also had poor adequacy (but see Ripplinger and Sullivan, 2010). Even considering only the best of the set, in general, the datasets often deviated from the expectations of the model in at least some ways (figure 5.7).

Of the 72 comparative datasets of SLA, we detected deviations from the expectations in 32 datasets (using a cut-off of $p = 0.05$), 22 by at least two, and 15 by three or more. Results were similar in the seed mass data (of the 226 seed mass datasets, we detected deviations in 153 datasets with at least one test statistic, 95 by at least two and 65 by three or more) and leaf nitrogen content (of the 39 datasets, we detected deviations in 19 by at least one, 12 by at least two, and 8 by three or more test statistic).

Some test statistics were much more likely to detect model violations than others. In 163 cases C_{VAR} revealed the data deviated significantly from the expections of the best model. In 118 cases,

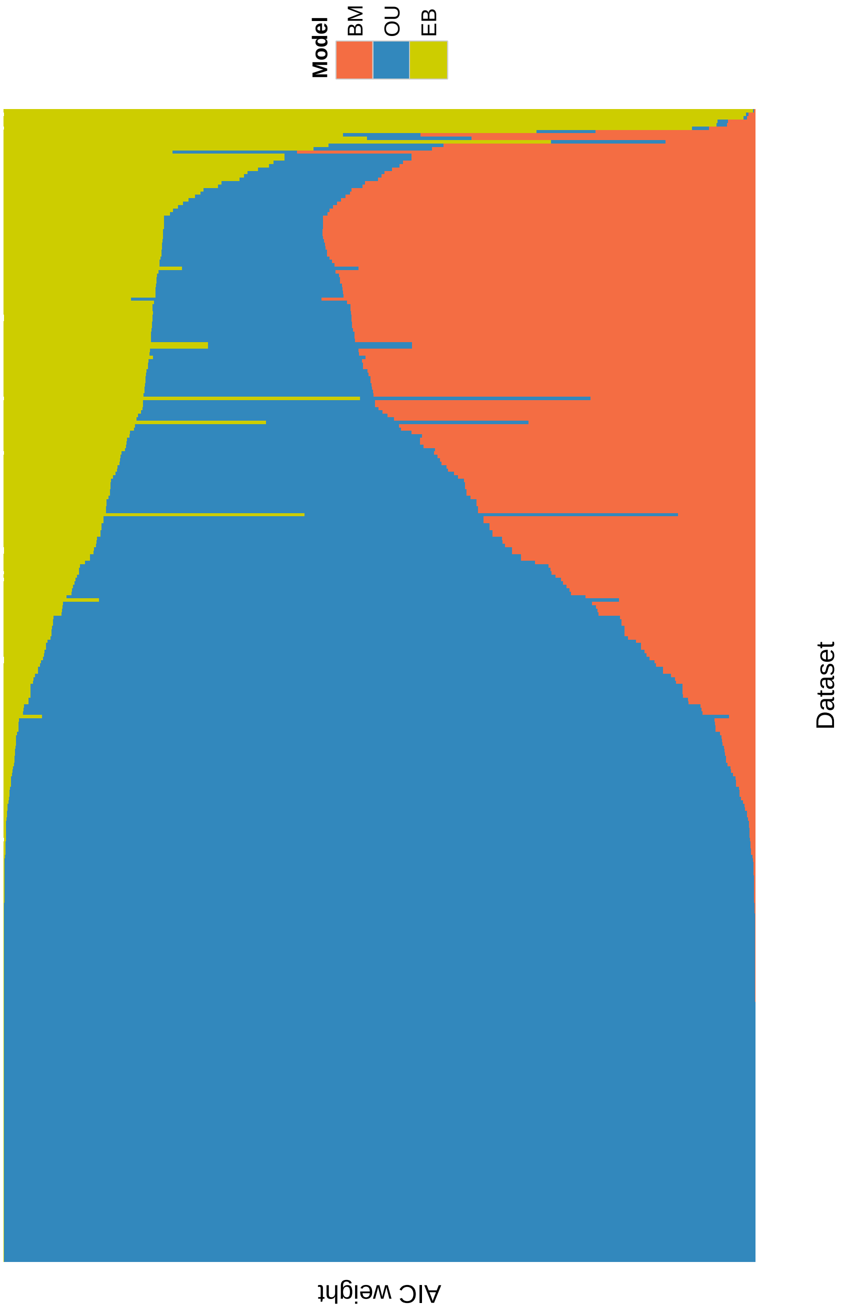

Figure 5.6: The relative support, as measured by AIC weight, for the three models used in our study (BM, OU, and EB) across all 337 datasets. An OU model is highly supported for a majority of the datasets.

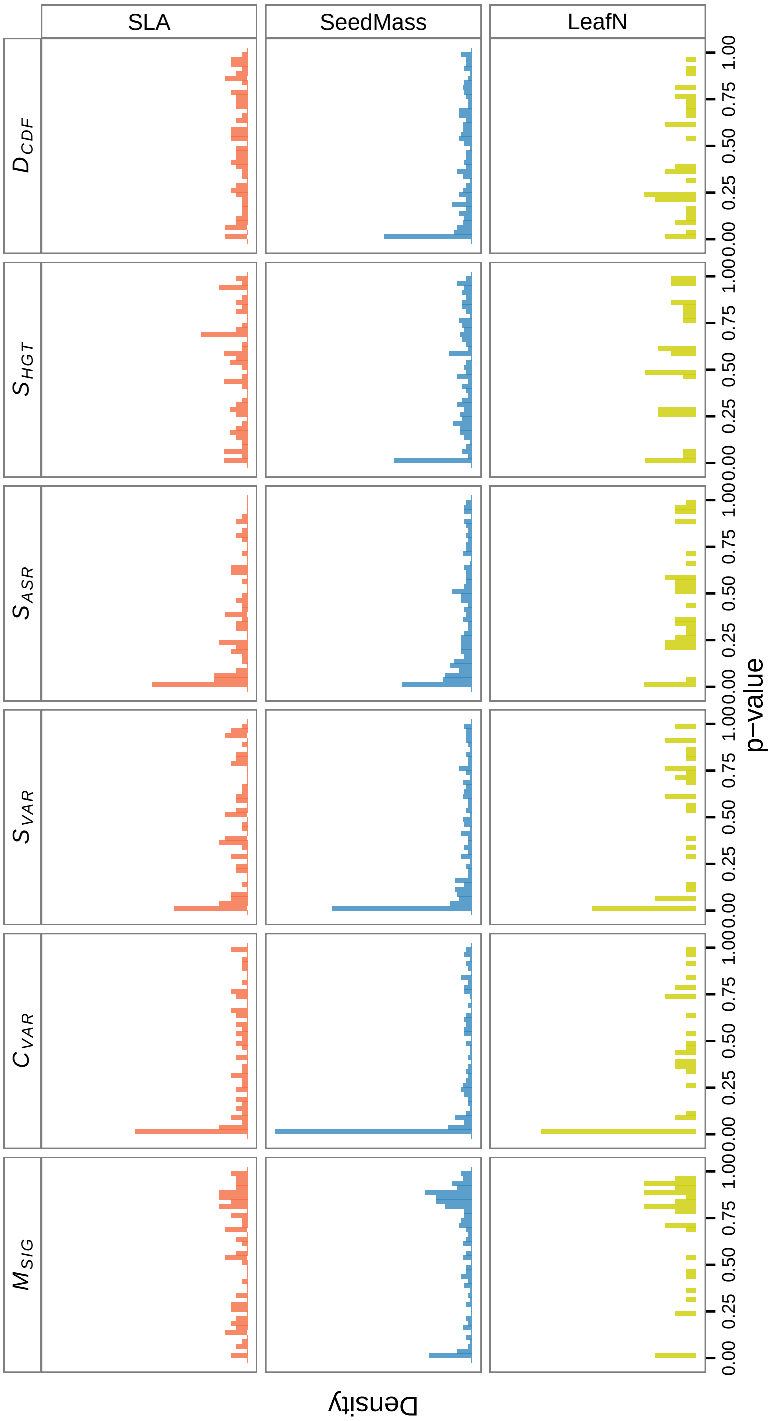

Figure 5.7: The distribution of p-values for our six test statistics over all 337 datasets in our study after fitting the models using ML. The p-values are from applying our model adequacy approach to the best supported of the three models (as evaluated with AIC). Many of the datasets deviate from the expectations under the best model along a variety of axes of variation. Deviations are particularly common for the coefficient of variation C_{VAR} and the slope of the contrasts against their expected variances S_{VAR}.

S_{VAR} did. The rate of deviation was much somewhat lower for the other test statistics (M_{SIG}: 39, S_{ASR}: 84, S_{HGT}: 54, D_{CDF}: 67).

Across all 337 datasets, 133 are adequately modeled by either BM, OU or EB. As stated above, the numbers of models that showed deviations with at least one test statistic may be somewhat overinflated. However, the proportion of clades in which p-values were less than 0.05 is much, much greater than the error rates we found in our simulations. And the proportions for each individual test statistic is much higher than would be expected by chance.

As the subclades are not independent (overlapping sets of taxa are present in family, order and time-slice phylogenies), conventional statistics, such as linear regression, are not straightforward to apply across datasets. Nonetheless, the trend is clear: the larger the phylogeny, the more likely OU is to be highly supported and the stronger the evidence the model is inadequate. There is a strong relationship between the size of a subclade and the overall distance between observed and simulated test statistics, as measured by the Mahanalobis distance (Figure 5.8). While it is likely that evolutionary rates and processes are more heterogeneous when one considers larger clades, it is also true that violations from model expectations are more easily detected when considering more data: the more contrasts that are examined, the lower the variance in the distribution of simulated test statistics, and therefore we are more likely to detect model misspecification (see Discussion). We also note here that if the model was adequate at all scales (for example, if a single OU process described the evolution of these traits across all angiosperms), then there would be no relationship between the Mahalanobis distance and the size of the phylogeny.

5.6 DISCUSSION

5.6.1 *Why does model adequacy matter?*

Whatever inferences we want to make from comparative data—e.g., characterizing broad-scale patterns of evolution through time, investigating correlations between characters or testing hypotheses about the processes that have driven trait evolution over macroevolutionary time—it is important that our chosen statistical model captures variation in the data *relevant to the question being addressed*. If, for example, the goal is to assess variation in macroevolutionary rates over time, it is essential that the model does a good job of explaning temporal heterogeneity. If we want to know about the slope of an evolutionary allometric relationship, we need a model that

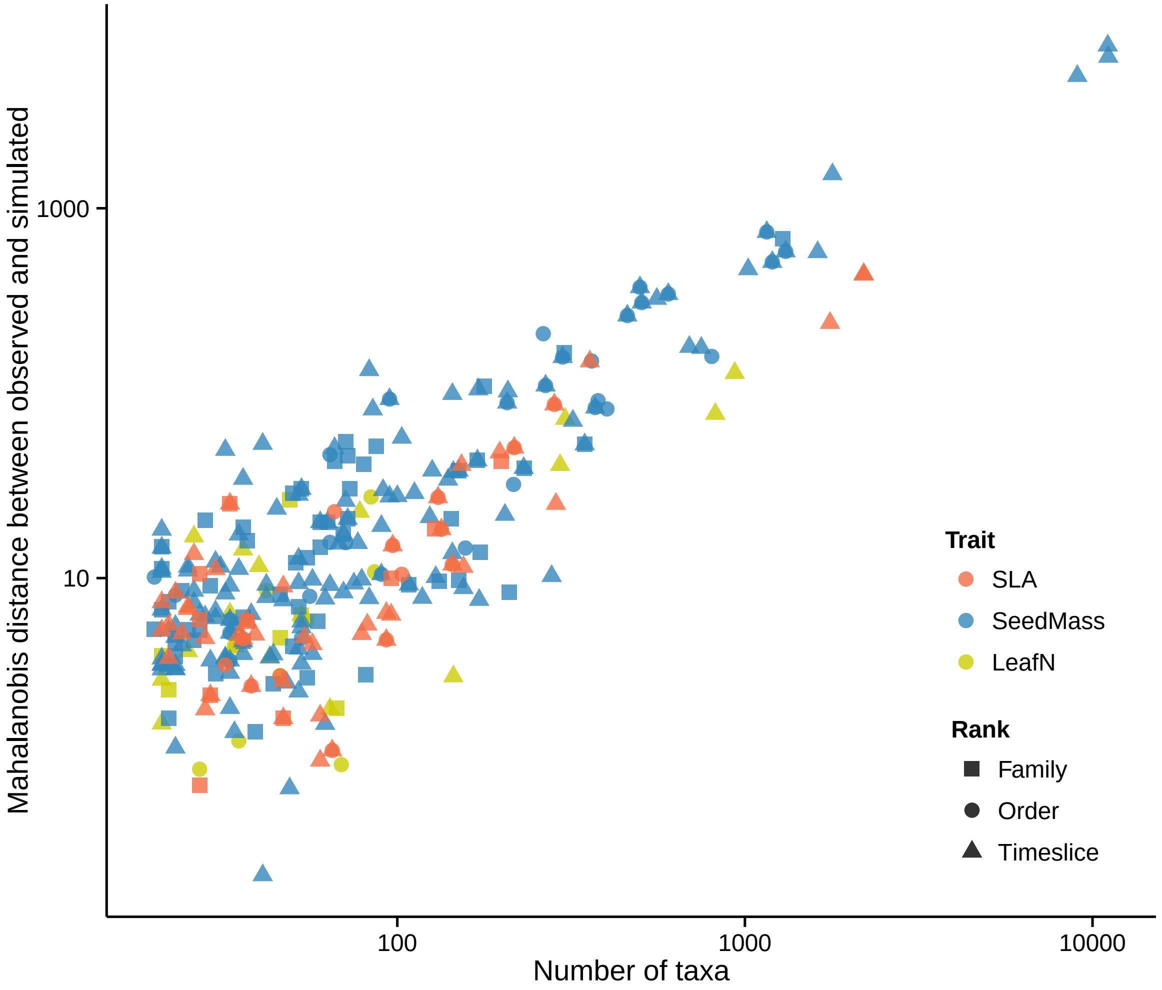

FIGURE 5.8: The relationship between clade size and a multivariate measure of model adequacy. The Mahalanobis distance is a scale-invariant metric that measures the distance between the observed and simulated test statistics, taking into account the covariance between test statistics. The greater the Mahalanobis distance, the worse the model captures variation in the data. Considering only the best supported model for each clade (as chosen by AIC), there is a striking relationship between the two—the larger the dataset, the stronger the evidence that the model does not capture variation in the data.

provides a meaningful estimate of this parameter (Hansen and Bartoszek, 2012). Comparing the fit of a model to a set of alternatives (using likelihood ratio tests, Information Theoretic metrics, Bayes Factors, etc.) can only allow for a relative assessment of the suitability of the model for the task. Such a model comparison approach does not provide any information about whether a model will allow us to actually get at the question we are interested in.

The flipside of this is that tests of model adequacy, such as ours, are designed to measure the absolute fit but not the absolute appropriateness of the model. We know that all of the models used in comparative biology are wrong. Whether they are useful or not will depend on the question being addressed. We are far from the first to suggest that model adequacy is important to consider when using comparative methods (see, for example Felsenstein, 1985, 1988; Harvey and Pagel, 1991; Garland *et al.*, 1992; Díaz-Uriarte and Garland, 1996; Hansen and Martins, 1996; Price, 1997; Garland *et al.*, 1999; Garland and Ives, 2000; Hansen and Orzack, 2005; Hansen and Bartoszek, 2012; Felsenstein, 2012; Boettiger *et al.*, 2012; Slater and Pennell, 2014; Beaulieu *et al.*, 2013; Blackmon and Demuth, 2014). The contribution of our paper is to generalize many of these previous approaches into a single, flexible statistical framework.

Again, we emphasize that simply because a dataset deviates from the expectations of the model does not imply that the model should necessarily be rejected. In our analyses of model adequacy across the 337 angiosperm clades, we were focused on whether the model was suitable for measuring rates of evolution, which is dependent on the model being a good one (Hunt, 2012). For other questions, the fact that a model fails to capture some aspects of the variation in the data may not be that important. For example, if our question was that of Harmon *et al.* (2010)—are early bursts of evolution common in macroevolution?—we could conclude with good certainty that they are not. Our datasets may not be well described by an OU model, but they are certainly nothing like what we would expect under an early burst scenario. Likewise, if we are interested primarily in whether there is a pattern of correlation between two traits, the fact that the model we used is not adequately describing much of the variation will in many cases, not greatly impact the qualitative conclusions. A nuanced view of model adequacy is particularly important when analyzing large phylogenies: the more data we consider, the greater our ability to detect subtle deviations from model expectations (Figures 5.8 and B.3). Focusing only on the test statistic p-values may lead us to reject models that are actually reasonably suitable for addressing our question of interest.

However, we view the most interesting cases to be where the best model does not adequately describe the variation of interest. The way in which a model fails can provide a richer understanding of our data and the processes that have driven the patterns we observe (Gelman and Shalizi, 2013). First, model inadequacy can point to problems in the data. We suspect that this is likely a common cause of poor model fit. For the empirical analyses, we used a very large phylogeny of angiosperms that was constructed to test specific global-scale biodiversity questions (Zanne *et al.*, 2014b). We recognize that the tree is poorly resolved in many places (particularly, near the tips) and is likely ill-suited for addressing more detailed, clade-specific questions (see the recent critique by Donoghue and Edwards, 2014). Specifically, the inaccurate placement of species will, on average, cause evolutionary rates to be inflated, which is precisely what we find (see below). However, we emphasize that phylogenetic error is likely ubiquitous and this problem is certainly not limited to the tree we used. Likewise, the dataset we assembled is rather heterogeneous in terms of quality; the data were originally collected for a diverse set of reasons and some groups have been measured much more carefully than others. And while we have done our best to clean the data, errors undoubtedly remain.

Second, and most excitingly, the failure of a model to adequately describe relevant aspects of the data can provide insight into the processes we have failed to consider in our model (Gelman and Shalizi, 2013). For example, if a model fails to capture variation relative to time (evaluated by the test statistic S_{HGT}), this suggests that temporal heterogeneity has been greater than we allowed for. The causes of such heterogeneity have long been a topic of interest in macroevolutionary studies (e.g., Simpson, 1944; Foote, 1997) and there has been a great deal of recent development towards more complex rate-varying models (e.g., O'Meara *et al.*, 2006; Thomas *et al.*, 2006; Eastman *et al.*, 2011; Weir and Mursleen, 2013; Rabosky *et al.*, 2014). Likewise, failure to adequately describe variation across the clade may indicate that the existence of multiple macroevolutionary optima (sensu Hansen, 2012) are driving the dynamics of traits over time (see Hansen, 1997; Butler and King, 2004; Beaulieu *et al.*, 2012; Ingram and Mahler, 2013; Uyeda and Harmon, 2014, for models that have been used to capture these dynamics).

Model inadequacy may also suggest types of models that have not previously been considered. For example, if recently diverged species tend to more dissimilar than can be accounted for under a simple diffusion model such as BM or OU, this may be the result of character displacement. However, almost no phylogenetic models have been put forth that explicitly model interactions between lineages (but see Nuismer and Harmon, 2015). Or if traits have lower variance than expected under an OU process, this may be the result of hard bounds. Boucher *et al.* (2014) recently argued that this is the case for climatic niches and that alternative models need to be developed for this case. Of course, a researcher may discover that her dataset is poorly described by all of the currently available models. Aside from deriving new models specific to her question and dataset, she should at least carefully examine the extent to which model misspecification is likely to affect the major conclusions and proceed forward with due caution.

5.6.2 *Implications for empirical studies*

In our analysis of angiosperm functional traits, we found common macroevolutionary models to often be poor descriptors for the patterns of variation and likely inadequate for estimating evolutionary rates. While there are certainly a number of important caveats to our analysis (discussed above), the overall trends are clear. This should certainly give researchers some pause about the models routinely used in our field—especially as they are often used in a model comparison framework to evaluate the "tempo and mode" of macroevolution. We argue that our results strongly suggest that we may often be missing a large part of the story.

The 337 comparative datasets we analyzed varied in terms of traits, size, age and placement in the angiosperm phylogeny. Nonetheless, several general patterns emerge. An OU model, was by and large, the most supported of the three we examined. In an analysis of 67 comparative datasets consisting of size and shape data from a variety of animal taxa, Harmon et al. (Harmon *et al.*, 2010) also found substantial support for OU models, though for their datasets, BM was more commonly chosen by AIC. (We note, however, that many of their datasets were quite small; see Slater and Pennell, 2014). Since their paper, a substantial number of studies conducted in a diverse array of groups have also found OU models to be preferred over BM models (e.g., Burbrink *et al.*, 2012; Quintero and Wiens, 2013; López-Fernández *et al.*, 2013; Thomas *et al.*, 2014).

The tendency of OU to explain data better than BM has inspired diverse process-based explanations, including stabilizing selection, evolutionary constraints and the presence of "adaptive zones" (Hansen and Martins, 1996; Butler and King, 2004; Hansen, 2012; Pennell and Harmon, 2013). If the widespread support for OU models was indeed caused by the biological processes that have been proposed, we would expect that an OU model would also be widely adequate.

However, this is not what we found. The datasets deviated significantly from the distributions expected under OU models, most often detected with C_{VAR} and S_{VAR} but frequently with others as well. OU models often failed to capture other important types of heterogeneity—variation with respect to rate variation (M_{SIG}), trait values (S_{ASR}) and time (S_{HGT}). Additionally, a substantial number of datasets were not well-modeled by a multivariate normal distribution (D_{CDF}). These results suggest a statistical explanation for the high support for OU models. OU predicts higher variance near the tips of the phylogeny than do BM or EB models (see Figure 1 in Harmon *et al.*, 2010).

Heterogeneous evolutionary processes, phylogenetic misestimation and measurement error could also produce such a pattern. In light of our results from model adequacy, it seems likely that OU is often supported because it is able to accommodate more "slop" (phylogenetic and trait error in addition to model misspecification) than the other models. This is not to say that the processes captured by OU models are unimportant in macroevolution, but rather that OU models may be favored for reasons that are more statistical than biological. Future, and hopefully more widely adequate, models of trait evolution could be developed that both include aspects of the OU model, especially the bounds on trait values, while incorporating additional biological realism (for a recent example of such a model, see Nuismer and Harmon, 2015).

The way in which the observed test statistics deviate from the simulated values also supports the claim that the widespread support for OU is largely a statistical artifact. Model violations were most frequently detected by the variance estimate, C_{VAR}. If the evolutionary process (or, alternatively, phylogenetic/measurement error) is heterogeneous across the tree, the lineages in some parts of the clade will be much more divergent than in others. The only way for the model to account for the highly divergent groups is to estimate a large σ^2 (and/or a small α parameter for the OU model). The unit tree formed by these parameter estimates will have long branches across the entire tree. In the less divergent parts of the tree, the contrasts calculated on this unit tree will be small, relative to what we expect under BM. So perhaps counter-intuitively, when heterogeneity in processes across taxa cause the estimated global rates of divergence to be inflated, this results in a higher value for C_{VAR}.

The second major take-home from the empirical analyses is that error, both in trait values and phylogenies, can have serious consequences for model adequacy. We frequently detected deviations from model expectations with S_{VAR}, the slope between the contrasts and their expected variances. This is indicative the rate of evolution appears to be varying with regards to branch length over which it is measured. This seems unlikely to be attributable to any biological process; it is far more probable that this reflects phylogenetic error (particularly, branch length error). Above, we outlined some of the deficiencies of the datasets we used in this paper but argue that these are likely to be widespread in comparative data. The test statistics outlined above can serve as useful diagnostics to aid researchers in identifying outliers that may be driving the pattern. We recommend that researchers faced with an inadequate model plot the magnitude of the contrasts on to the unit tree; this will usually be much more informative with regards to the model fit than plotting the magnitude of the contrasts on the original phylogeny. Exceptionally large or small contrasts on the unit tree can provide clues as to where the data may be erroneous. If phylogenetic error were causing poor model fits, we would predict that many of the anomalous contrasts would occur in parts of the tree that are poorly supported.

5.6.3 *Extensions of our approach*

There are a number of additional ways our approach could be extended. First, we have only considered a limited set of test statistics. We chose them because each of these has a clear statistical expectation and observed deviations from them have intuitive biological explanations. However, they are certainly a subset of all possible test statistics that could be applied. For example, because contrasts are i.i.d., there should be no autocorrelation between neighboring contrasts; the test statistics could be expanded to detect non-zero autocorrelation. Second, as stated above, our approach can be applied equally well to phylogenetic regression models, such as phylogenetic generalized least squares (Grafen, 1989; Martins and Hansen, 1997) or phylogenetic mixed models (Lynch, 1991; Housworth *et al.*, 2004; Hadfield and Nakagawa, 2010), where concerns regarding model adequacy are just as pertinent (Hansen and Bartoszek, 2012). While our approach can be used to assess the adequacy of the phylogenetic component of regression models “out of the box”, additional steps are required to assess the adequacy of the linear component. Third, our method was designed for quantitative trait models that assume data can be modeled with a multivariate normal distribution. We need general model adequacy approaches for other types of traits, such as: discrete traits (i.e., binary, multistate, ordinal; see Beaulieu *et al.*, 2013; Blackmon and Demuth, 2014; Maddison and FitzJohn, 2015, for recent discussions of this); traits that influence speciation rates (e.g., Maddison *et al.*, 2007; FitzJohn, 2010) and quantitative trait models that do not predict a multivariate normal distribution of traits (Landis *et al.*, 2013; Schraiber and Landis, 2014).

It may also be possible to extend our approach with an eye towards model selection. Slater and Pennell (2014) developed their posterior predictive simulation approach (which is related to our method) to distinguish between a BM model and one where rates of evolution decreased through time. They chose test statistics specifically to address this question. Slater and Pennell found using posterior predictive fit as a model selection criterion to be much more powerful than comparing models using AIC or likelihood ratio tests, particularly when "outlier taxa" (lineages where the pattern of evolution deviates from the overall model) were included in the analysis. The logic of Slater and Pennell could be extended to other scenarios; to test some evolutionary hypotheses, we may care a lot about whether a model explains varation along some axes but be less concerned about others. This is a question-specific approach to model selection and has been developed in the context of molecular phylogenetics (Bollback, 2002; Lewis *et al.*, 2014). This is also the essence of the Decision-Theoretic approach to model selection (Robert, 2007), which has also been well-used in phylogenetics (Minin *et al.*, 2003), but has not previously been considered in PCMs.

5.7 ARBUTUS

We have implemented our approach in a new R package ARBUTUS. It is available on github `https://github.com/mwpennell/arbutus`. For this project, we have also adopted code from the APE (Paradis *et al.*, 2004), GEIGER (Pennell *et al.*, 2014a) and DIVERSITREE (FitzJohn, 2012a) libraries. We have written functions to parse the output of a number of different programs for fitting trait evolution models (see the ARBUTUS website for an up-to-date list of supported models and packages). As this approach was developed to be general, we have written the code in such a way that users can include their own test statistics and trait models in the analyses.

5.8 CONCLUDING REMARKS

Attempts to assess the adequacy of phylogenetic models are almost as old as modern comparative phylogenetic biology. In the 1980s and 1990s much discussion surrounded the appropriateness of various methods and models (Felsenstein, 1985, 1988; Harvey and Pagel, 1991; Garland *et al.*, 1992; Díaz-Uriarte and Garland, 1996; Price, 1997; Garland *et al.*, 1999; Garland and Ives, 2000). We argue that this discussion is key to progressing in our field. This is not simply because we are concerned that many inferences may not be robust to model violations. Rather, we believe that considering model adequacy can help suggest new ways of thinking about how to combine data and models to test macroevolutionary hypotheses.

CHAPTER 6

How much of the world is woody? Dealing with sampling error in comparative data[6]

6.1 SUMMARY

The question posed by the title of this chapter is a basic one, and it is surprising that the answer is not known. Recently assembled trait datasets provide an opportunity to address this, but scaling these datasets to the global scale is challenging because of sampling bias. Although we currently know the growth form of tens of thousands of species, these data are not a random sample of global diversity; some clades are exhaustively characterised, while others we know little-to-nothing about. Starting with a database of woodiness for 39,313 species of vascular plants (12% of taxonomically resolved species, 59% of which were woody), we estimated the status of the remaining taxonomically resolved species by randomisation. To compare the results of our method to conventional wisdom, we informally surveyed a broad community of biologists. No consensus answer to the question existed, with estimates ranging from 1% to 90% (mean: 31.7%). After accounting for sampling bias, we estimated the proportion of woodiness among the world's vascular plants to be between 45% and 48%. This was much lower than a simple mean of our dataset and much higher than the conventional wisdom. Alongside an understanding of global taxonomic diversity (i.e., number of species globally), building a functional understanding of global diversity is an important emerging research direction. This approach represents a novel way to account for sampling bias in functional trait datasets and to answer basic questions about functional diversity at a global scale.

6.2 INTRODUCTION

The distinction between a woody and non-woody growth-form is probably the most profound contrast among terrestrial plants and ecosystems—the difference between a forest and a grassland is the presence of wood. The recognition of the fundamental importance of this divide dates back

[6] Previously published as: FitzJohn R.G., Pennell M.W., Zanne A.E., Stevens P.F., Tank D.C., and Cornwell W.K. 2014. How much of the world is woody? Journal of Ecology 102:1266–1272.

at least to *Enquiry into Plants* by Theophrastus of Eresus (371–287 BCE), a student of Plato and Aristotle, who began his investigation into plant form and function by classifying the hundreds of plants in his garden into woody and herbaceous categories (Theophrastus, 1916).

The last two thousand years of research into wood since Theophrastus classified his garden have uncovered its origin in the early Devonian (~400 Mya; Gerrienne *et al.* 2011); that prevalence of woodiness varies with climate (Moles *et al.*, 2009); that wood has been lost many times in diverse groups, both extant and extinct (Judd *et al.*, 1994), often as an adaptation to freezing temperatures (Zanne *et al.*, 2014a); that it has also been gained many times, particularly on island systems (Carlquist, 1974; Givnish, 1998); and that many different forms of pseudo-woody growth habit have appeared across groups that have lost true woodiness or diverged before true woodiness evolved (Cornwell *et al.*, 2009). We know about its mechanical properties and developmental pathways, its patterns of decomposition and their effects on ecosystem function (Cornwell *et al.*, 2009), and that woody and herbaceous species have markedly different rates of molecular evolution (Smith and Donoghue, 2008). However, we have no idea about what proportion of species in the world are actually woody.

Recently assembled functional trait datasets provide an opportunity to address this question. However, such datasets are, almost without exception, biased samples of global diversity. Researchers collect data for specific questions on a local scale, and assembling these local datasets creates a useful resource (Kattge *et al.*, 2011). But as with GenBank's assembly of genetic data (Smith *et al.*, 2011), the simple compilation of data is not an unbiased sample, and these initial sampling biases will, in turn, bias downstream analyses. Understanding and accounting for the biases in these datasets is an important and necessary next step.

We sought to develop an approach that accounts for this bias. In doing so, we were able to re-ask Theophrastus' 2000-year old question at a global scale: how many of the world's plant species are woody? We also sought to understand how well scientists were able to overcome this bias and make a reasonable estimate. To do this, we took the unconventional approach of coupling our analysis with an informal survey in which we asked our question to the broader community of botanists and other biologists.

6.3 MATERIALS AND METHODS

6.3.1 *Dataset*

We used a recently assembled database with growth-form data for 49,061 vascular plant species (i.e., lycopods, ferns, gymnosperms and angiosperms), which is the largest such database assembled to date (Zanne *et al.*, 2013, 2014a, available on the Dryad data repository;). This database uses a functional definition of woodiness: woody species have a prominent above-ground stem that persists through time and changing environmental conditions and herbaceous species lack such a stem—a definition originally suggested by Asa Gray (1887). Zanne *et al.* (2014a) chose this simple definition because it best characterised the functional aspect of growth form that they investigated, allowing them to compare species that maintain an above-ground stem through freezing conditions to ephemeral species that avoid freezing conditions. More precise definitions that rely on lignin content and/or secondary vascular tissue from a bifacial cambium are problematic because there are many exceptions depending on tissue type, times of development, or environmental conditions (Groover, 2005; Spicer and Groover, 2010; Rowe and Paul-Victor, 2012). Because our analyses and survey were based on this database, we present this functional definition of woodiness here for clarity (see Zanne *et al.* 2014a, for a discussion of the various definitions of woodiness, their merits, and pitfalls). Note that in addition to species producing secondary xylem, this definition classifies, among other groups, palms, tree ferns and bamboo as woody.

As with all large data assemblies, the underlying datasets were collected for a variety of research goals. For example, a number of the datasets come from forestry inventories, which, of course, are biased towards recording woody species. Other sources of sampling bias, including geographically restricted sampling in many sub-datasets, may be less obvious but nonetheless may have major implications for the inferences drawn from aggregate databases.

Because the effort to organise plant taxonomy, especially synonymy, is on-going, there was uncertainty regarding the status of many plant names. To bring species binomials to a common taxonomy among datasets, names were matched against accepted names in the Plant List (The Plant List, 2014). Any binomials not found in this list were matched against the International Plant Name Index (`http://www.ipni.org/`) and Tropicos (`http://www.tropicos.org/`). Potential synonymy in binomials arising from the three lists was investigated using the Plant

List tools (The Plant List, 2014). As a result of this cleaning, the number of species in the final dataset was reduced from 49,061 to 39,313.

Theophrastus recognised both the fundamental importance of the distinction between woody and herbaceous plants, and that this distinction is in some cases difficult to make. There are two ways that species were recorded as "variable" in form (Beaulieu *et al.*, 2013). First, different records of a single species may conflict in growth form (having both records of woodiness and herbaceousness); this affected 307 of the 39,313 species in the database. Second, 546 species (1.4%) were coded as variable. Following Beaulieu *et al.* (2013), we coded species in these groups as "woody" or "herbaceous" when a majority of records were either "woody" or "herbaceous", respectively, and for these species, records of "variable" do not contribute to the analysis. Our final database for the main analysis contained 38,810 records with both information on woodiness and documented taxonomy—15,957 herbs and 22,853 woody species. This included records from all flowering plant orders currently accepted by APG III (The Angiosperm Phylogeny Group, 2009) and the fern taxonomy of Stevens (2001), covering 15,232 genera and 465 families. The 503 species excluded at this step had identical numbers of records of being woody and herbaceous. We also ran analyses where we coded growth forms by treating species with *any* record of woody or variable as "woody" (and similarly for herbaceous), using all 39,313 species. Neither of these cases are likely to be biologically realistic but allowed us to evaluate the maximal possible effect of mis-coding variable species.

6.3.2 *Estimating the percentage of species that are woody*

To estimate the percentage of species that are woody, we cannot simply use the fraction of species within our trait database that are woody (22,853 of 38,810 = 59%) as these records represent a biased sample of vascular plants. For example, most Orchidaceae are probably herbaceous; we have only one record of woodiness among the 1,537 species for which we have data. However, the fraction of Orchidaceae species with known data (1,537 of 27,801 = 6%) is much lower than the overall rate of knowledge for all vascular plants (38,810 of 316,143 = 12%), which will upwardly bias the global estimate of woodiness. Conversely, systematic under-sampling of tropical species would bias the global woodiness estimate downwards, as tropical floras are thought to harbour a greater proportion of woody species than temperate ones (Moles *et al.*, 2009).

We developed a simple method to account for this sampling bias when estimating the percentage of woody species. In our approach, we treat each genus separately, and in all cases know that there are are n_w woody and n_h herbaceous species and a total of N species in the genus. For example, the genus *Microcoelia* (Orchidaceae) has 30 species in total, and we know that 12 are herbaceous and none are known to be woody ($N = 30$, $n_w = 0$, $n_h = 12$). We do not know the state of the remaining 18 species, so the true number of woody species, N_w, must lie between 0 and 18. In general, we cannot assume that these species are all herbaceous, even though both biological and mathematical intuition suggest that most of them will be.

We used two different approaches for imputing the values of these unknown species. First, we assumed that the known species were sampled without replacement from a pool of species with N_w woody and N_h herbaceous species ($N_w + N_h = N$), following a hypergeometric distribution. The probability that x of the species of unknown state are woody ($x = 0, 1, \ldots, N - n_w - n_h$) is proportional to

$$\Pr(N_w = x) \propto \binom{n_w + x}{n_w} \binom{N - n_w - x}{n_h} \tag{6.1}$$

Under this sampling model, the more species for which we do not have data, the greater the uncertainty in our estimates for the proportion of species which are woody. For *Microcoelia* this model gives a 42% probability that all species are herbaceous, and a 90% chance that at most 3 species are woody. This approach probably overestimates the number of woody species in this case, and in other cases where all known species are woody (e.g., *Actinidia* [Ericaceae]) it will probably underestimate the number of species that are woody. We see this as corresponding to a weak prior on the shape of the distribution of the fraction of woody species within a genus and will refer to this as the “weak prior” approach because it weakly constrains the state of missing species.

However, the distribution of woodiness among genera and families is strongly bimodal; most genera are either all-woody or all-herbaceous (Figure 6.1, Figure 6.2, and Sinnott and Bailey 1915). Among the 791 genera with at least 10 records, 411 are entirely woody, 271 are entirely herbaceous, and only 58 have between 10% and 90% woody species. Qualitatively similar patterns hold at both the level of family and order, though the distribution becomes progressively less bimodal as one moves up the taxonomic hierarchy (Figures 6.2 and 6.3). As a result, knowing the state of a handful of species within a genus can give a reasonable guess at the state of remaining species.

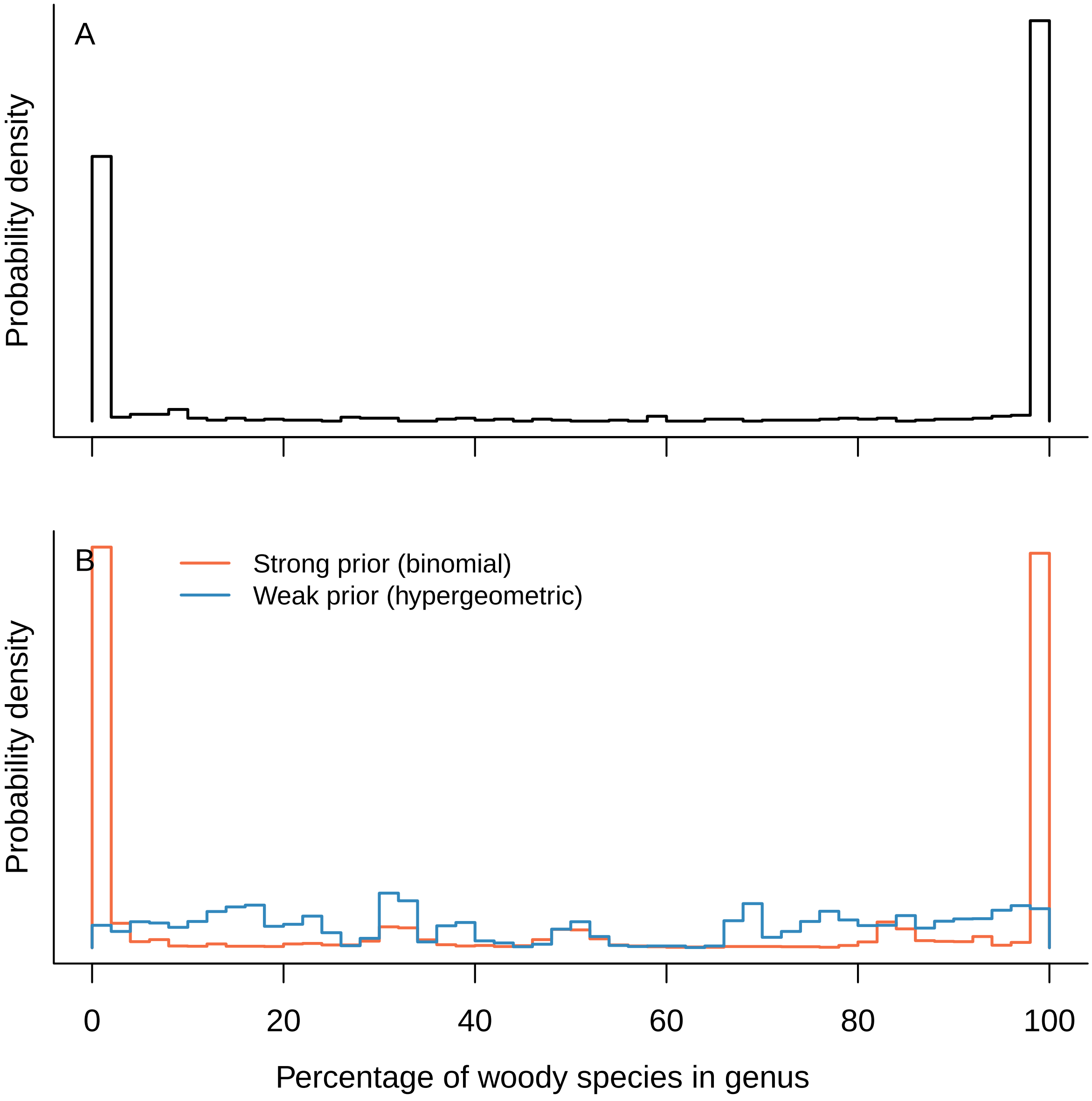

Figure 6.1: Distribution of the percentage of woodiness among genera. The distribution of the percentage of species that are woody within a genus is strongly bimodal among genera (panel A—showing genera with at least 10 species only). The two different sampling approaches generate distributions that differ in their bimodality (panel B). If we sample species with replacement from some pool, with a weak prior on the fraction of woodiness within the pool, then we generate a broad distribution with many polymorphic genera (blue line). Sampling with replacement, assuming that species are drawn from a pool of species that has a fraction of woody species equal to the observed fraction of woodiness, generates a strongly bimodal distribution (red line).

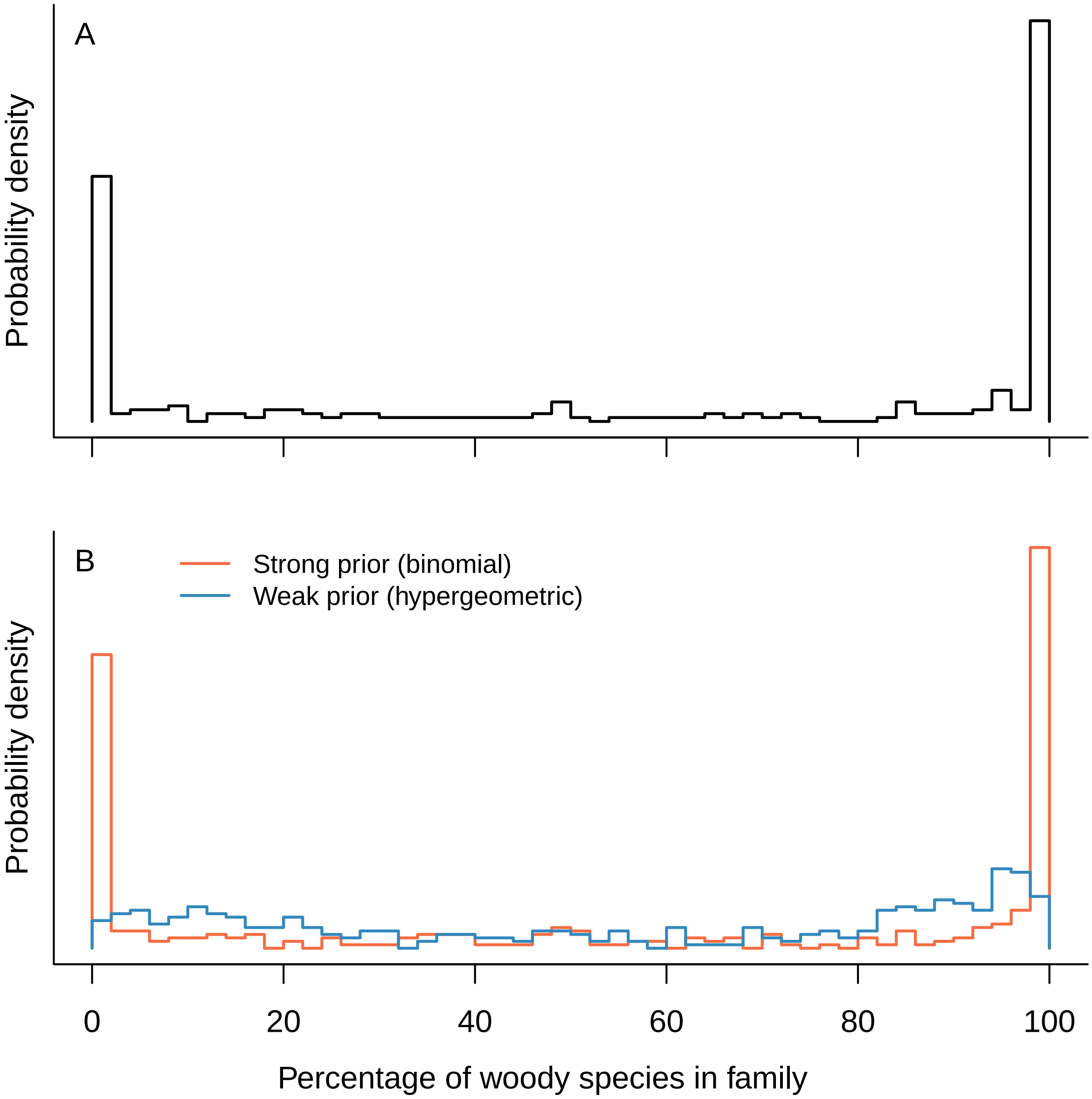

Figure 6.2: Distribution of the percentage of woodiness among families. The distribution of the percentage of species that are woody within a family is strongly bimodal among families (panel A), though less strongly bimodal than among genera. The two different sampling approaches generate distributions that differ in their bimodality (panel B). Using the weak prior approach generates a broad distribution with many polymorphic genera (blue line), while using the strong prior approach generates a strongly bimodal distribution (red line).

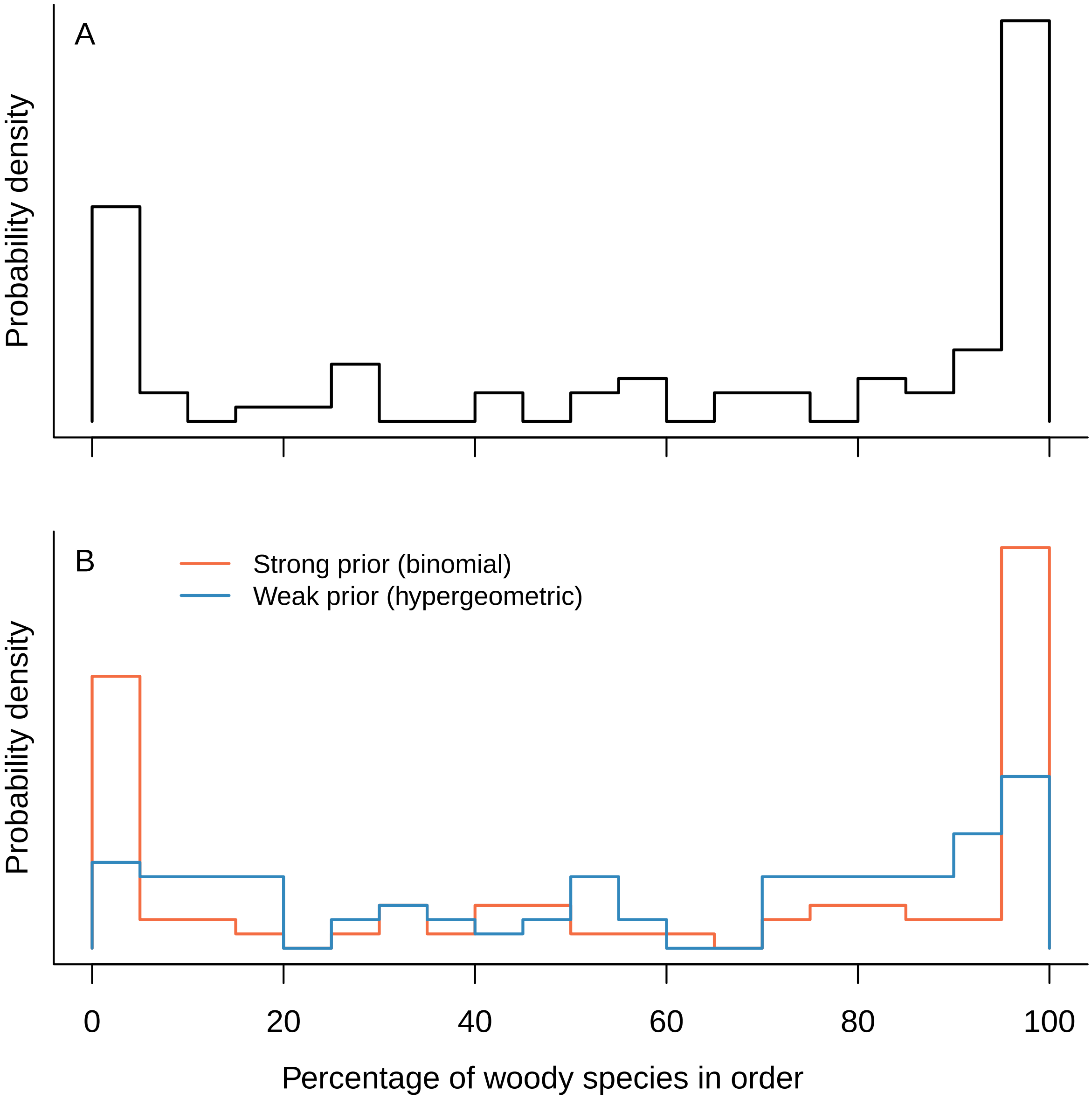

Figure 6.3: Distribution of the percentage of woodiness among orders. The distribution of the percentage of species that are woody within an order is bimodal among orders (panel A), though less strongly bimodal than among both genera and families. The two different sampling approaches generate distributions that differ in their bimodality (panel B). Using the weak prior approach generates a broad distribution with many polymorphic genera (blue line), while using the strong prior approach generates a strongly bimodal distribution (red line).

To model the other extreme of sampling, we used an approach where we computed the observed fraction of woody species

$$p_w = n_w/(n_w + n_h)$$

and sampled the state of the unobserved species using a binomial distribution, which represents the case of sampling with replacement. In this case the probability that x of the species are woody is:

$$\Pr(x = k) = \binom{N - n_w - n_h}{k - n_w} p_w^k (1 - p_w)^{N - n_h - k}. \tag{6.2}$$

In cases where all known species are woody (or herbaceous as in *Microcoelia*) this will assign all unknown species to be woody (or herbaceous). For such genera, increasing the number of unobserved species will not increase the uncertainty in the estimate, in contrast to the weak prior sampling approach. We therefore see the binomial sampling approach as corresponding to a very strong prior on the bimodal distribution of woodiness among genera, and we will refer to this as the "strong prior" approach because it more strongly constrains the state of missing species within genera with no known polymorphism. While neither of these approaches is "correct", they probably span the extremes of possible outcomes. In polymorphic genera the two approaches will give similar results, especially where the number of unknown species is relatively large.

For genera where there was no information on woodiness for any species, we sampled a fraction of species that might be woody from the empirical distribution of woodiness fractions *among genera* within the same order. We did this after imputing the missing species values within those other genera. So, if a genus is found in an order with genera that had woodiness fractions of $\{0, 0, 0.1, 1\}$ we would have approximately a 50% chance of sampling a 0% woodiness fraction for a genus, with probabilities from 0.1 to 1 being fairly evenly spread. Given this woodiness fraction, we then sampled the number of species that are woody from a binomial distribution with this fraction and the number of species in the genus as its parameters.

In addition to the number of species known to be woody and herbaceous, we also require an estimate of the number of species per genus. For this, we used the number of accepted names within each genus in the Plant List (The Plant List, 2014). The taxonomic resources were compiled by Zanne *et al.* (2014a) are on available on Dryad (Zanne *et al.*, 2013).

For each genus, we sampled the states of unobserved species, from either the hypergeometric or binomial distribution, parametrised from the observed data for that genus. For each sample we can then combine these estimates to compute the number (or fraction) of species that are woody at higher taxonomic levels (family, order or vascular plants). We repeated this sampling 1,000 times to generate distributions of the number (or fraction) of species that are woody. The R code and data to replicate this analysis are available on github (`https://github.com/richfitz/wood`).

6.3.3 *Survey*

In estimating the number of species within angiosperm families, Joppa *et al.* (2010) found that expert opinion generally agreed closely with estimates from a statistical model. We were interested in whether a consensus answer existed—even if not formalised in the literature—and if so, whether it was consistent with our estimates. We created an English-language survey (which we also translated into Portuguese) asking for an estimate of the percentage of species that are woody according to the above definition. We also asked respondents to indicate their level of familiarity with plants, level of formal training, and the country in which they received their training. We distributed the survey to the community via several electronic mailing lists with wide circulation among biologists: EVOLDIR, ECOLOG, R-SIG-PHYLO, TAXACOM, HERBARIA, as well as local lists. We also posted links on the social-networking platforms GOOGLE+, TWITTER and FACEBOOK to reach a broad audience. In order to increase representation of survey responses from Latin America, we translated the survey into Portuguese and distributed it to Brazilian biology FACEBOOK groups and university mailing lists.

To analyse the survey data, we used linear regression on logit-transformed percent woodiness as (see Warton and Hui, 2011) and treated the self-reported level of botanical familiarity and education as factors. We converted country of training to coarse latitude using shapefiles from the GBIF dataportal (GBIF, 2013), and converted these into "tropical" and "temperate" using an absolute latitude of 23° 26′.

6.4 RESULTS

Across all vascular plants, we estimated the fraction of woody species to be between 45% and 48%. Specifically, using our strong prior sampling approach (binomial distribution) we estimated 45.6% of species are woody (95% confidence interval of 45.3-45.9%) and with the weak prior (hypergeometric distribution) approach we estimated 47.6% (95% CI of 46.9-48.2%) (Figure 6.4).

The different approaches generated different distributions of the per-genus percentage of woodiness (Figure 6.1), with a less strongly bimodal distribution using the weak prior approach. (See Figures 6.2 and 6.3 for the distributions at the level of families and orders, respectively.) However, the two different approaches (strong versus weak priors) led to similar phylogenetic distributions of estimated woodiness (Figure 6.5 versus Figure 6.6), differing only in the details. We have compiled a table of the estimated number of woody species under both sampling approaches for all genera, families and orders included in our analysis. This is available on the Dryad data repository (FitzJohn *et al.*, 2014, doi:10.5061/dryad.v7m14).

As stated above, neither of these sampling approaches is "correct". However, as the observed distribution of woodiness fraction among genera is itself strongly bimodal, we believe that the true result lies closer to 45% than to 47%. A more sophisticated hierarchical modeling approach could lead to a more precise answer, but we feel that our values probably span the range of estimates that such an approach would generate. And in any case, we felt that addressing a simple question warranted a simple approach.

Different codings of variable species (see above) significantly moved our estimates, despite affecting a small minority of species. Coding all variable species as woody, our estimates increased by 1.6% to 47.1% with the strong prior approach and by 1% to 48.6% with the weak prior approach (Figure 6.7). Similarly, with coding all variable species as herbaceous, the fraction of woody species decreased by 1.9% to 43.7% under a strong prior and by 1.3% to 46.3% under a weak prior (Figure 6.7).

There was strikingly little consensus among researchers as to the percentage of species that are woody. We received 292 responses from 29 countries, with estimates that ranged from 1% to 90% with a mean of 31.7% (Figure 6.8). The lowest estimate from our analyses (45% woody) is greater than 81% of our survey estimates.

We found little effect of respondents' level of training on their estimate (Figure 6.9). There was a significant effect of the respondent's familiarity with plants on the estimates, primarily

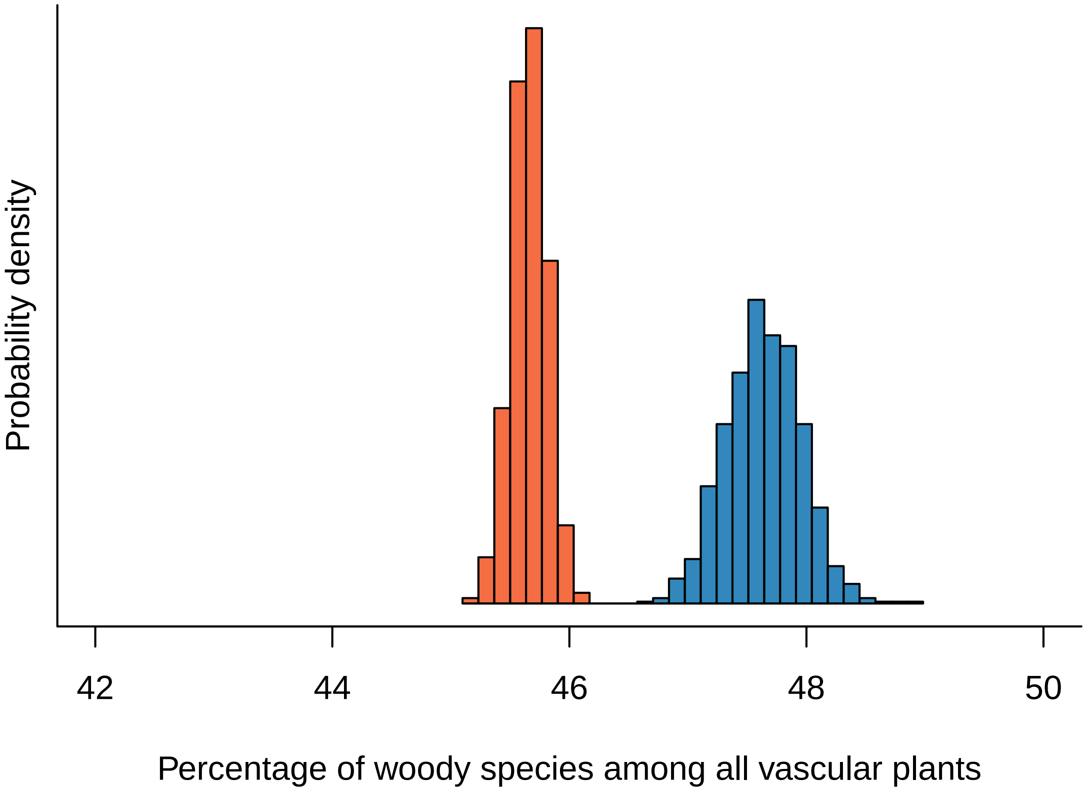

Figure 6.4: The posterior probability distribution for the proportion of the world's flora that is woody, using our two sampling approaches. The red (left) distribution samples missing species using the strong prior approach (binomial distribution), while the blue distribution (right) samples missing species using the weak prior approach (hypergeometric distribution).

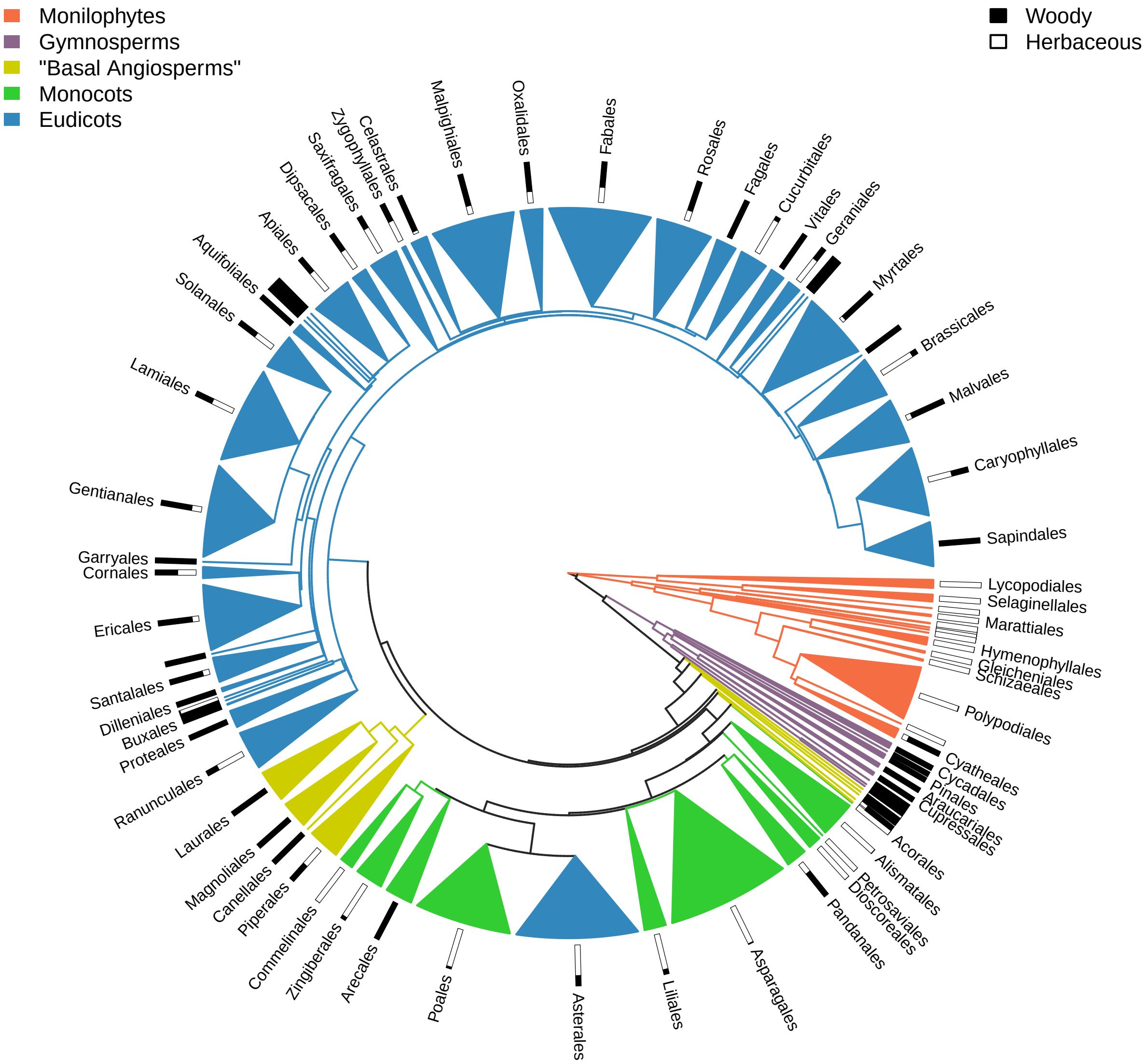

Figure 6.5: Distribution of the percentage of woodiness among orders of vascular plants. Each tip represents an order, with the width of the sector proportional to the square root of the number of recognised species in that order (data from accepted names in The Plant List 2014). The bars around the perimeter indicate the percentage of woody (black) and herbaceous (white) species, estimated using the "strong prior" (binomial) approach. Using the "weak prior" (hypergeometric) approach generally leads to an estimated percentage that is closer to 50% (see Figures 6.6 and 6.1). Phylogeny from Zanne *et al.* (2014a) (available on Dryad; doi:10.5061/dryad.63q27/3). Orders not placed by APG III (The Angiosperm Phylogeny Group, 2009) are not displayed. We note that there is some discrepancy between the Zanne et al. tree and previous well-supported phylogenetic hypotheses (e.g., Soltis *et al.*, 2011), most notably, in the position of the Magnoliids; however, the higher-level relationships do not influence any of the analyses.

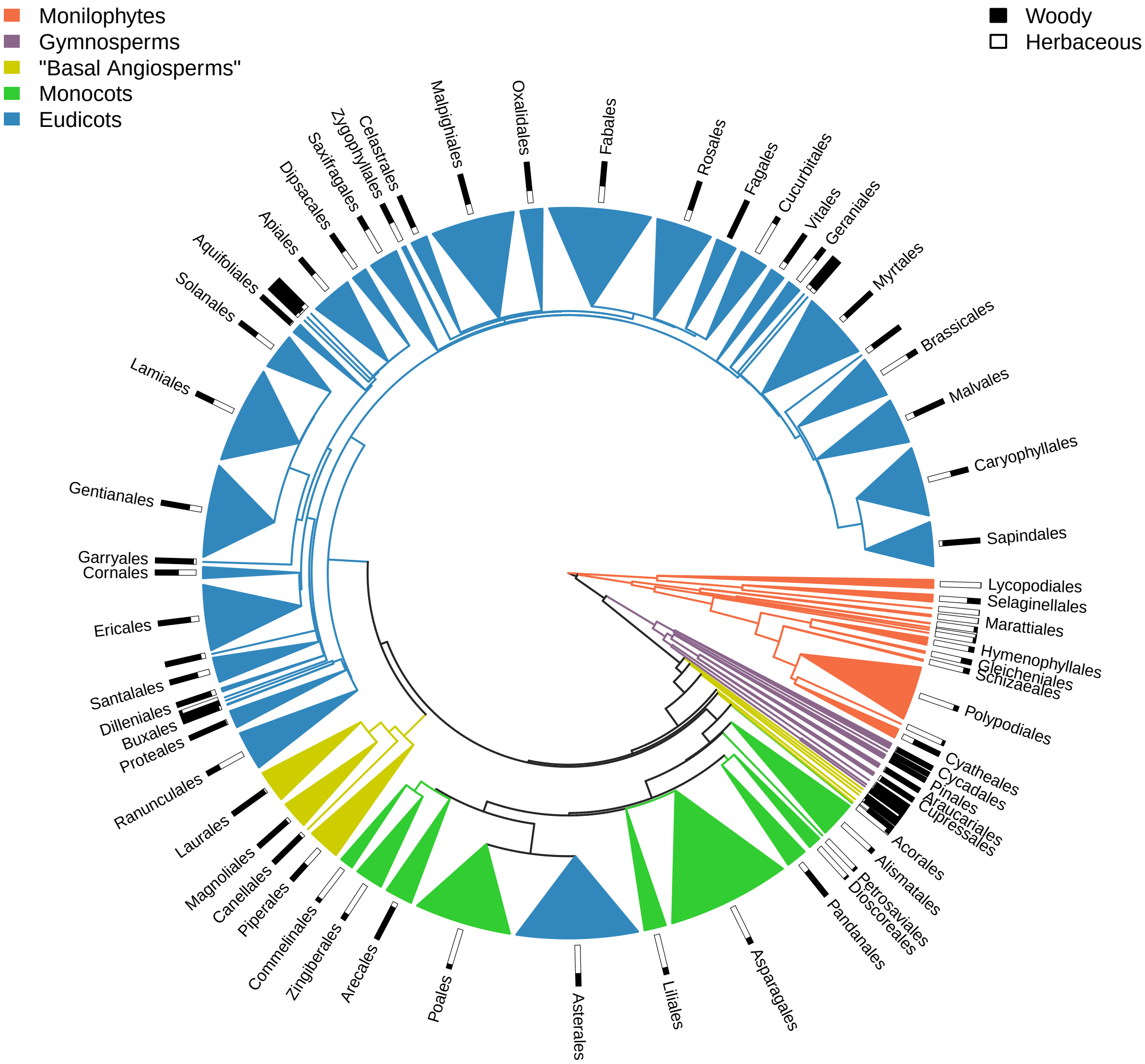

FIGURE 6.6: Distribution of the fraction of woodiness among orders of vascular plants. Each tip represents an order, with the fraction of circumference proportional to the square root of the number of recognised species in that order (data from accepted names in The Plant List 2014). The bars around the perimeter indicate the percentage of woody (black) and herbaceous (white) species, estimated using the "weak prior" (hypergeometric) approach. Using the "strong prior" (binomial) approach generally leads to an estimated percentage that is further away from 50% (see main text Figures 1 and 2). Phylogeny from Zanne *et al.* (2014a) (available on Dryad; doi:10.5061/dryad.63q27/3). Orders not placed by APG III (The Angiosperm Phylogeny Group, 2009) are not displayed.

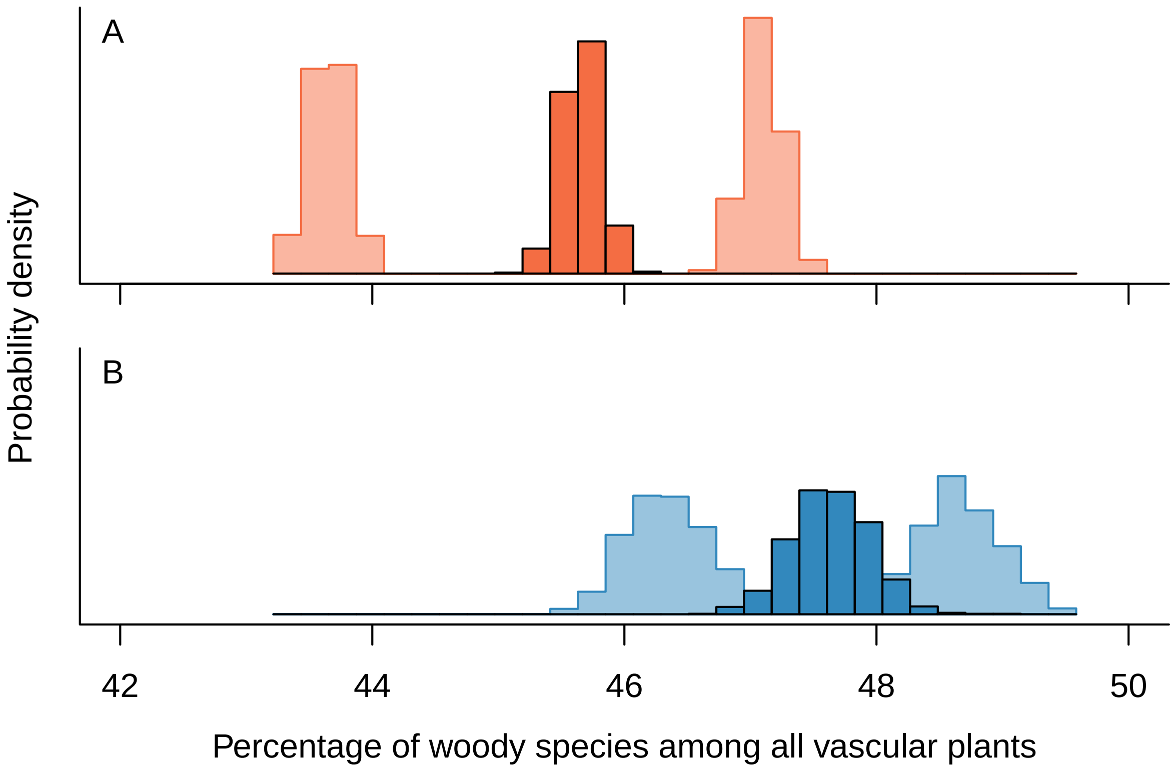

Figure 6.7: The effect of different coding on estimates of the fraction of species that are woody, under the strong prior approach (binomial; panel A) and weak prior approach (hypergeometric; panel B). The dark distributions are the results from our main analysis (Figure 6.4). Distributions to the left (with lower estimates of woodiness) code all species with any record of herbaceousness or variability as herbaceous. Similarly, distributions to the right (with higher estimates of woodiness) code all species with any record of woodiness or variability as woody.

driven by respondents with little botanical familiarity (the “What’s a Plant?” category in the survey), whose estimates tended to be lower (less woody) than the estimates of those with more familiarity. However, excluding respondents with little familiarity with plants had virtually no effect on the mean estimate of respondents (32.4% excluding this category as compared to 31.7% with them included). Restricting survey responses to only respondents at least “Familiar” with plants, and with at least an undergraduate degree in botany or a related field (143 responses), only increased the mean survey estimate to 32.9%.

Before carrying out the survey, we had hypothesised that researchers from tropical regions may perceive the world as woodier than researchers from more temperate regions due to the latitudinal gradient in woodiness (Moles *et al.*, 2009). Indeed, there was an effect of being in a tropical country, with the estimates from tropical countries being slightly higher than those from temperate countries (p=0.02), but this effect was very small (R^2=0.02, Figure 6.8).

6.5 DISCUSSION

Our estimates of woodiness differed from both the survey and the simple mean of the global database: neither simple statistics nor biologists’ intuition were accurate in this case. The difference from community knowledge is in striking contrast to Joppa *et al.* (2010), who found that that expert opinion on the number of species within different angiosperm groups agreed closely with results based on analyses of data and their bias.

The respondents to our survey perceived there to be substantially fewer woody species in the world than there probably are. This herb-centric view of the world may arise from the importance of our (mostly herbaceous) cultivated crops, or the fact that people—including most researchers—likely spend more time in the garden than in the forest, and especially not in tropical forests where diversity is high and disproportionately woody.

Our estimates of the percentage of species that are woody (45/48%) differ from the raw estimate based on species in our database (59%). This difference is caused by the interaction between biased sampling and clustered trait data at a variety of taxonomic scales. The distribution of woodiness is bimodal among genera, and the distribution of sizes of those genera differs with woodiness. Genera that are primarily herbaceous (less than 10% woody species for genera with at least 10 records) were on average larger than primarily woody genera (more than 90% woody

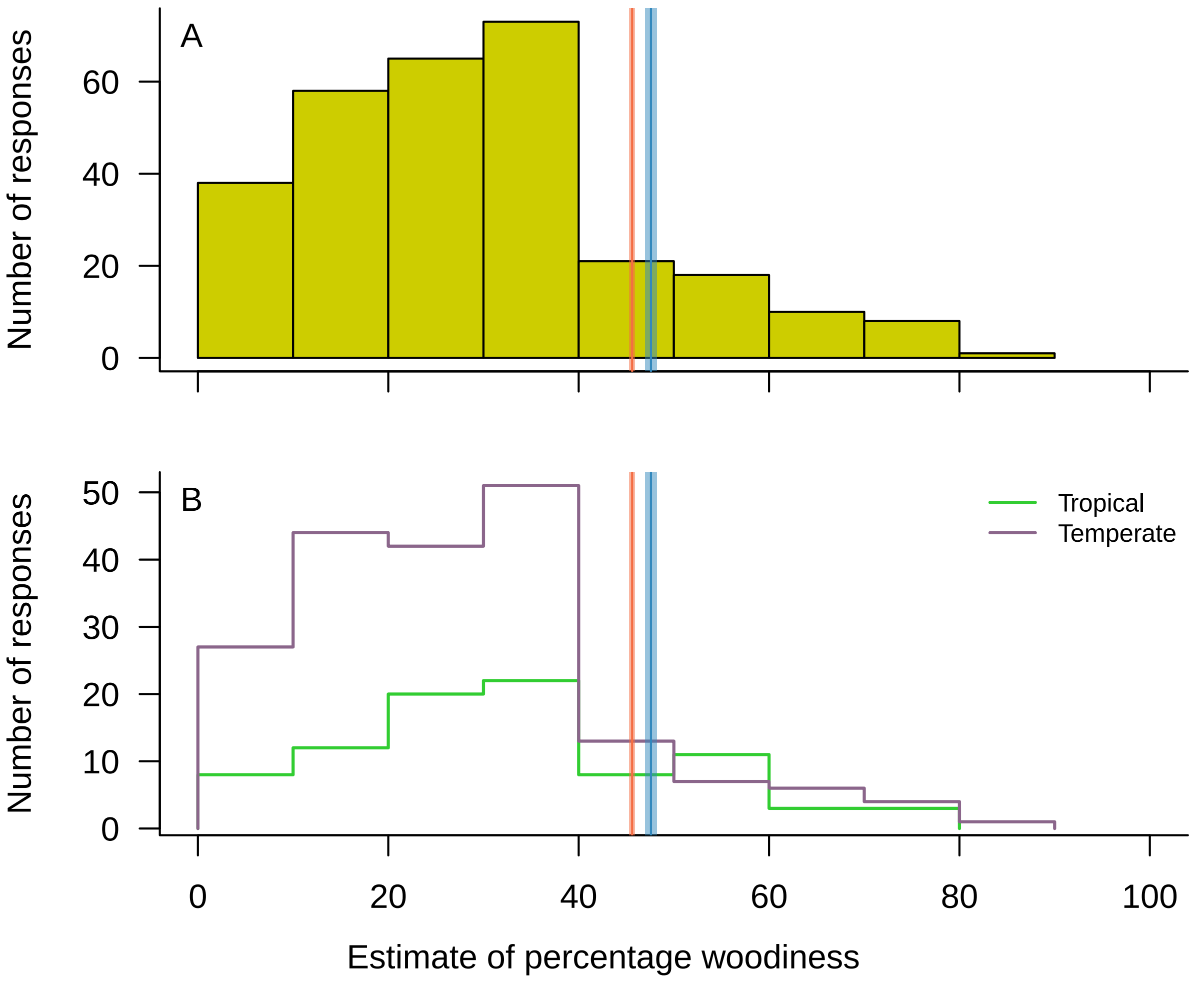

FIGURE 6.8: Distribution of all responses to the survey question "What percentage of the world's vascular plant species are woody?". The mean and 95% confidence intervals for our estimates of the proportion of woody species from the empirical data are depicted by the horizontal shaded rectangles; the blue rectangle corresponds to the "weak prior" approach and the red rectangle corresponds to the "strong prior" approach. Panel A includes all 292 responses. In panel B, the 282 responses that indicated country are shown separated into "tropical" (green distribution) and "temperate" (purple). Estimates from tropical countries were slightly, but significantly, higher than those from temperate countries (p =0.02, R^2=0.02).

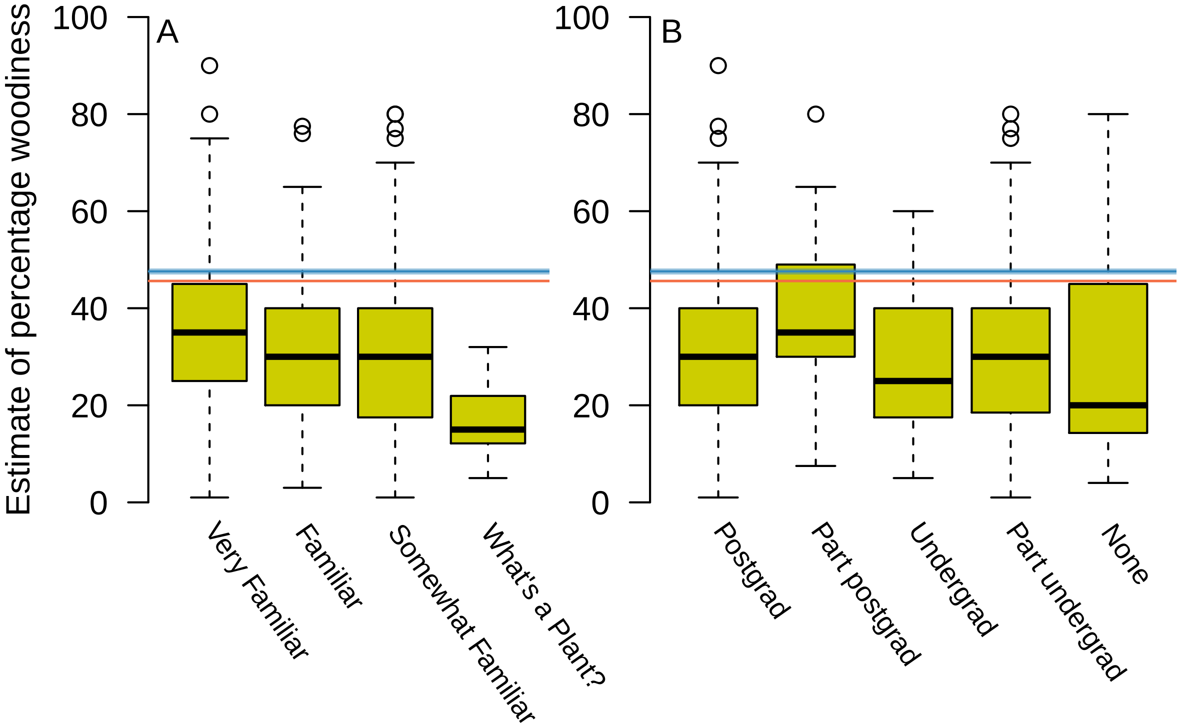

FIGURE 6.9: Distribution of responses to the survey question "What percentage of the world's vascular plant species are woody?". Responses are divided by familiarity with plants (panel A) and formal training in botany or a related discipline (panel B). The mean and 95% confidence intervals for our estimates of the proportion of woody species from the empirical data are depicted by the horizontal shaded rectangles; the blue upper rectangle corresponds to the "weak prior" approach and the red lower rectangle corresponds to the "strong prior" approach.

species), with a mean of 214 species compared to 151 (See Figure 6.10). This means that even a random sampling above the level of species will lead to a biased estimate.

The effect of sampling bias within our database on the estimate is amplified by the distribution of woodiness at higher taxonomic levels, with families or even orders often being predominantly either woody or herbaceous (Figure 6.5 and Sinnott and Bailey 1915). There are two major clades that are primarily herbaceous—the monocots (Monocotyledons) and ferns (Monilophyta). However, there are many primarily herbaceous clades nested within woody clades, and vice versa, which makes the combination of taxonomic and functional information crucial for answering this type of question.

We also found that the way in which we handled variable species significantly altered the estimates. That changing the state of such a relatively small number of species has the potential to alter inferences made at a global scale is rather surprising. Two points regarding this are worth noting here. First, we reiterate that our alternate coding schemes (all variable species coded as herbaceous and all variable species coding as woody) are rather extreme and unlikely to be biologically realistic. Second, while these alternate coding schemes certainly affected the estimates, the magnitude of their effect is much less than that of the overall sampling bias in the original database.

Higher-order classifications are at least as much a product of human pattern matching as biological processes. Genera correspond to the morphological discontinuities among species that humans deem important (Scotland and Sanderson, 2004), which likely includes woodiness (e.g., Hutchinson, 1973). The relative rarity of genera with significant numbers of both woody and herbaceous species (Figure 6.1) reinforces the importance of this trait. A significant, but unaccounted for, source of error is the likely nonrandom woodiness of undiscovered species. We would predict that there are likely more herbs to be discovered than woody plants; larger genera tend to be more herbaceous (Figure 6.10) and we think it is more likely that new species are yet to be described in these large groups. In principle, rarefaction analysis could estimate the number of species remaining to be discovered in different groups, but this is not possible for many plant clades (Costello *et al.*, 2011); for many clades the "collecting curve" shows little sign of saturation, which is required for such an analysis.

Sampling biases are pervasive in ecological datasets, and need to be addressed when using them for analyses. Global databases of functional traits (e.g., TRY; Kattge *et al.*, 2011) are central

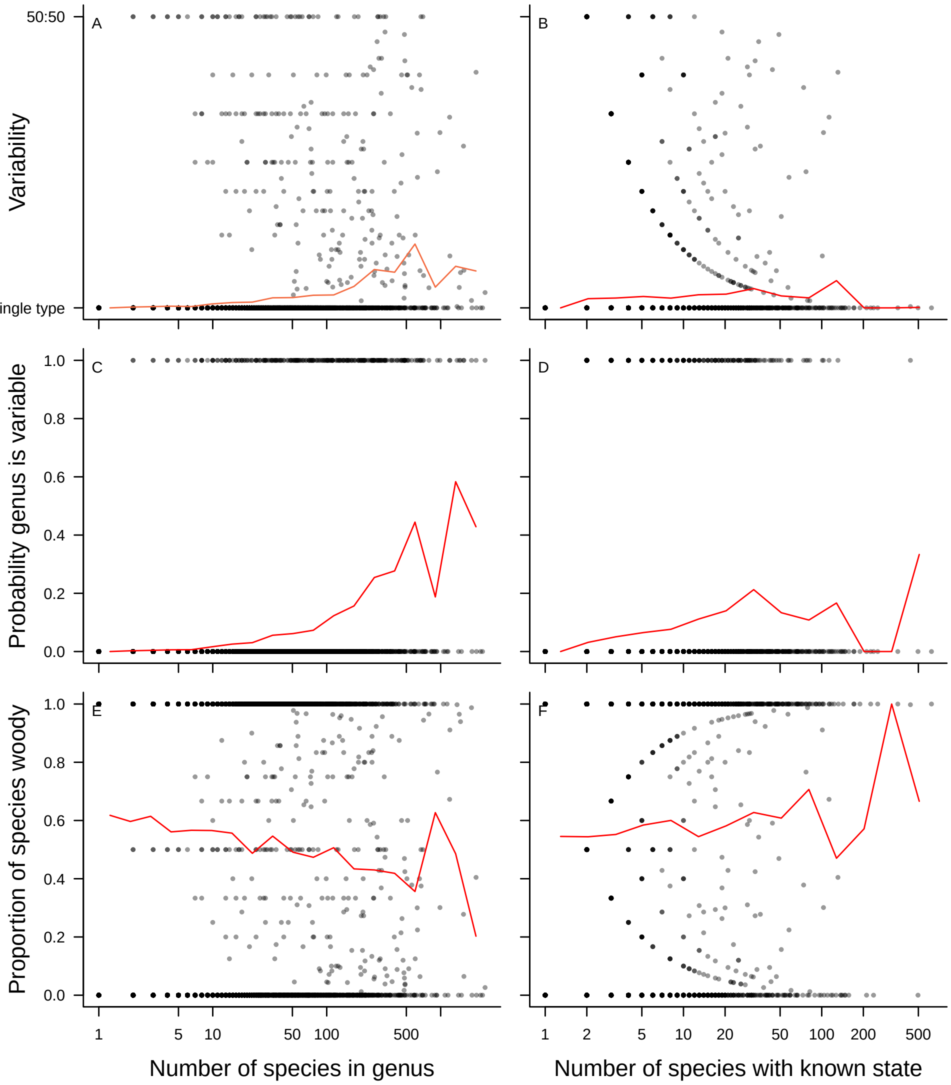

FIGURE 6.10: The relationship between the size of a genus and its chance of being "variable" for woodiness. We plotted the relationship between the level of variablitiy in the dataset (from all of a single-type to equal numbers of woody and herbaceous species) against the number of species in a genus (panel A) and the number of species with known state (panel B). Larger genera tend to be more variable although this pattern is not strong. We then coded all genera as being either variable or all of a single-type and examined the relationship between this binary characterization and the number of species per genus (panel C) and the number for which we have known states (panel D). Using the binary characterization, it is clear that large genera have a higher probability of being variable, even if few species actually vary (compare with panels A and B). Though there is a great deal of scatter, larger genera also tend to be more herbaceous than woody genera (panel E) but the genera for which we have more data tend to be more woody (panel F). This shows that the available data is generally biased towards woody species. In all panels, the red line is a moving average over 20 (left column) of 15 (right column) equally spaced bins on this log axis.

to biodiversity research, but through no fault of the database collator they are inevitably biased in terms of taxonomic breadth and this may have serious consequences for the reliability of inferences drawn from them. For example, for woodiness the economic importance of forestry species likely leads to their over-sampling in this dataset. This sampling bias also affects many commonly used methods in ecological and evolutionary research (e.g., Ackerly, 2000; Nakagawa and Freckleton, 2008; Pennell and Harmon, 2013; Pakeman, 2014) in addition to its well understood effects on conventional statistics. In our case, taking the data at face-value, we would have greatly overestimated the global percentage of woody species. Inferring the global frequency of any trait would face the same problem. For example, the ecologically important traits of nitrogen-fixing, mycorrhizal symbioses and pollinator syndrome are strongly taxonomically structured, and we would expect raw estimates to be biased in the same way that woodiness was. Our approach was developed for binary traits but similar approaches could be developed for multi-state categorical or continuous traits.

In addition to improving an estimate of the mean, the methods in this chapter can also be used to generate a probability of each unobserved species being woody. Thus, it can be used as a type of taxonomically-informed data-imputation. Recently, two related approaches have been developed to do just this, both focusing on continuous traits (Swenson, 2014; Guénard *et al.*, 2013). While their details differ, both approaches are model-based in that they impute trait values for missing species based on the fitted parameters of phylogenetic models estimated from the species already in the database. This is conceptually different from our approach; we do not assume any model for the evolution of woodiness, such as the Mk model (Pagel, 1994; Lewis, 2001), which is commonly used to model discrete characters evolving on a phylogeny. Both types of approaches—using taxonomic categories (this study) versus modeling trait evolution along a phylogeny—have advantages and disadvantages. One disadvantage of a modeling-based approach is that if the sampling is biased with respect to the character states, the parameter estimates themselves will be biased, leading to an incorrect estimation of the states for the remaining species. While our approach avoids this issue, we ignore potentially useful information on the phylogenetic relationships within genera and branch lengths separating lineages.

6.6 CONCLUDING REMARKS

As a result of centuries of effort, we now have an increasingly complete understanding of taxonomic diversity. More recent developments in assembling global trait databases offer the promise of gaining similar insights into the functional diversity of the earth's biota. While the question we ask in this chapter—what proportion of the world's flora is woody?—is simple, answering it required dealing with the pervasive biases that will be present in most large datasets. Researchers should be aware that because of these biases and the phylogenetically structured distribution of traits, the law of large numbers will not apply, and that estimates from trait databases will not converge on the true value. Our approach is just one of many potential ways to address these biases; we hope that our analysis encourages others to think critically and creatively about the problem. Just as Theophrastus' garden was a non-random sample of the Greek flora, our trait databases contain diverse biases; accounting for them will be important in making inferences about broad-scale ecological and evolutionary patterns and processes.

Bibliography

Ackerly D.D. 2000. Taxon sampling, correlated evolution, and independent contrasts. Evolution 54:1480–1492.

Adams D.C., Berns C.M., Kozak K.H., and Wiens J.J. 2009. Are rates of species diversification correlated with rates of morphological evolution? Proceedings of the Royal Society B: Biological Sciences 276:2729–2738.

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

Alfaro M.E., Santini F., Brock C., Alamillo H., Dornburg A., Rabosky D.L., Carnevale G., and Harmon L.J. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proceedings of the National Academy of Sciences, USA 106:13410–13414.

Alroy J. 2008. Dynamics of origination and extinction in the marine fossil record. Proceedings of the National Academy of Sciences, USA 105:11536–11542.

Anscombe F.J. 1973. Graphs in statistical analysis. The American Statistician 27:17–21.

Arnold A.J. and Fristrup K. 1982. The theory of evolution by natural selection: A hierarchical expansion. Paleobiology 8:113–129.

Arnold S.J., Pfrender M.E., and Jones A.G. 2001. The adaptive landscape as a conceptual bridge between micro- and macroevolution. Genetica 112:9–32.

Atkinson Q.D., Meade A., Venditti C., Greenhill S.J., and Pagel M. 2008. Languages evolve in punctuational bursts. Science 319:588.

Aze T., Ezard T.H.G., Purvis A., Coxall H.K., Stewart D.R.M., Wade B.S., and Pearson P.N. 2011. A phylogeny of cenozoic macroperforate planktonic foraminifera from fossil data. Biological Reviews 86:900–927.

Baca S.C., Prandi D., Lawrence M.S., Mosquera J.M., Romanel A., Drier Y., Park K., Kitabayashi N., MacDonald T.Y., Ghandi M., Allen E.V., Kryukov G.V., Sboner A., Theurillat J.P., Soong T.D., Nickerson E., Auclair D., Tewari A., Beltran H., Onofrio R.C., Boysen G., Guiducci C., Barbieri C.E., Cibulskis K., Sivachenko A., Carter S.L., Saksena G., Voet D., Ramos A.H., Winckler W., Cipicchio M., Ardlie K., Kantoff P.W., Berger M.F., Gabriel S.B., Golub T.R., Meyerson M., Lander E.S., Elemento O., Getz G., Demichelis F., Rubin M.A., and Garraway L.A. 2013. Punctuated evolution of prostate cancer genomes. Cell 153:666–677.

Bachtrog D., Kirkpatrick M., Mank J.E., McDaniel S.F., Pires J.C., Rice W., and Valenzuela N. 2011. Are all sex chromosomes created equal? Trends in Genetics 27:350–357.

Bachtrog D., Mank J.E., Peichel C.L., Kirkpatrick M., Otto S.P., Ashman T.L., Hahn M.W., Kitano J., Mayrose I., Ming R., Perrin N., Ross L., Valenzuela N., Vamosi J.C., and of Sex Consortium T.T. 2014. Sex determination: Why so many ways of doing it? PLoS Biology 12:e1001899.

Bandyopadhyay R., Heller A., Knox-DuBois C., McCaskill C., Berend S.A., Page S.L., and Shaffer L.G. 2002. Parental origin and timing of de novo robertsonian translocation formation. The American Journal of Human Genetics 71:1456–1462.

Bartoszek K. 2014. Quantifying the effects of anagenetic and cladogenetic evolution. Mathematical Biosciences 254:42–57.

Bateman A.J. 1948. Intra-sexual selection in drosophila. Heredity 2:349–368.

Batista D.A., Tuck-Muller C.M., Martinez J.E., Kearns W.G., Pearson P.L., and Stetten G. 1993. A complex chromosomal rearrangement detected prenatally and studied by fluorescence in situ hybridization. Human Genetics 92:117–121.

Baum D.A. and Larson A. 1991. Adaptation reviewed: A phylogenetic methodology for studying character macroevolution. Systematic Biology 40:1–18.

Beaulieu J.M., Jhwueng D.C., Boettiger C., and O'Meara B.C. 2012. Modeling stabilizing selection: Expanding the Ornstein-Uhlenbeck model of adaptive evolution. Evolution 66:2369–2383.

Beaulieu J.M., O'Meara B.C., and Donoghue M.J. 2013. Identifying hidden rate changes in the evolution of a binary morphological character: The evolution of plant habit in campanulid angiosperms. Systematic Biology 62:725–737.

Benton M.J. and Pearson P.N. 2001. Speciation in the fossil record. Trends in Ecology & Evolution 16:405–411.

Beukeboom L.W. and Perrin N. 2014. The evolution of sex determination. Oxford University Press.

Blackmon H. and Demuth J.P. 2014. Estimating tempo and mode of Y chromosome turnover: Explaining Y chromosome loss with the fragile Y hypothesis. Genetics 197:561–572.

Blomberg S.P., Garland T., and Ives A.R. 2003. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. Evolution 57:717–745.

Blomberg S.P., Lefevre J.G., Wells J.A., and Waterhouse M. 2012. Independent contrasts and PGLS regression estimators are equivalent. Systematic Biology 61:382–391.

Boettiger C., Coop G., and Ralph P. 2012. Is your phylogeny informative? measuring the power of comparative methods. Evolution 66:2240–2251.

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.

Bokma F. 2010. Time, species, and separating their effects on trait variance in clades. Systematic Biology 59:602–607.

Bollback J.P. 2002. Bayesian model adequacy and choice in phylogenetics. Molecular Biology and Evolution 19:1171–1180.

Bolstad G.H., Hansen T.F., Pélabon C., Falahati-Anbaran M., Pérez-Barrales R., and Armbruster W.S. 2014. Genetic constraints predict evolutionary divergence in dalechampia blossoms. Philosophical Transactions of the Royal Society B: Biological Sciences 369:1–15.

Bookstein F.L., Gingerich P.D., and Kluge A.G. 1978. Hierarchical linear modeling of the tempo and mode of evolution. Paleobiology 4:120–134.

Boucher F.C., Thuiller W., Davies T.J., and Lavergne S. 2014. Neutral biogeography and the evolution of climatic niches. The American Naturalist 183:573–584.

Brown J.H., Gillooly J.F., Allen A.P., Savage V.M., and West G.B. 2004. Towards a metabolic theory of ecology. Ecology 85:1771–1789.

Brown J.M. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Systematic Biology 63:334–348.

Bull J.J. 1983. Evolution of sex determining mechanisms. Menlo Park: The Benjamin/Cummings Publishing Company.

Burbrink F.T., Chen X., Myers E.A., Brandley M.C., and Pyron R.A. 2012. Evidence for determinism in species diversification and contingency in phenotypic evolution during adaptive radiation. Proceedings of the Royal Society B: Biological Sciences 279:4817–4826.

Burnham K. and Anderson D. 2004a. Model selection and multi-model inference: a practical information-theoretic approach. Springer.

Burnham K.P. and Anderson D.R. 2004b. Multimodel inference: Understanding aic and bic in model selection. Sociological Methods and Research 33:261–304.

Butler M.A. and King A.A. 2004. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. The American Naturalist 164:683–695.

Caballero A. 1995. On the effective size of populations with separate sexes, with particular refenc to sex-linked genes. Genetics 139:1007–1011.

Carlquist S. 1974. Island Biology. Columbia University Press.

Chamberlin J. and Magenis R.E. 1980. Parental origin of de novo chromosome rearrangements. Human genetics 53:343–347.

Chang S.L., Lai H.Y., Tung S.Y., and Leu J.Y. 2013. Dynamic large-scale chromosomal rearrangements fuel rapid adaptation in yeast populations. PLoS Genetics 9:e1003232.

Charlesworth B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. Genetical research 63:213–227.

Charlesworth B., Coyne J.A., and Barton N.H. 1987. The relative rates of evolution of sex chromosomes and autosomes. American Naturalist pages 113–146.

Charlesworth B., Lande R., and Slatkin M. 1982. A Neo-Darwinian commentary on macroevolution. Evolution 36:474–498.

Charlesworth B. and Wall J.D. 1999. Inbreeding, heterozygote advantage and the evolution of neo-X and neo-Y sex chromosomes. Proceedings of the Royal Society of London B: Biological Sciences 266:51–56.

Charlesworth D. and Charlesworth B. 1980. Sex differences in fitness and selection for centric fusions between sex-chromosomes and autosomes. Genetics Research 35:205–214.

Charlesworth D., Charlesworth B., and Marais G. 2005. Steps in the evolution of heteromorphic sex chromosomes. Heredity 95:118–128.

Clark A.G. 1988. The evolution of the Y chromosome with X-Y recombination. Genetics 119:711–720.

Clauset A. and Erwin D.H. 2008. The evolution and distribution of species body size. Science 321:399–401.

Cornwell W.K., Cornelissen J.H.C., Allison Steven D., Bauhus J., Eggleton P., Preston C.M., Scarff F., Weedon J.T., Wirth C., and Zanne A.E. 2009. Plant traits and wood fates across the globe: rotted, burned, or consumed? Global Change Biology 15:2431–2449.

Cornwell W.K., Westoby M., Falster D.S., FitzJohn R.G., O'Meara B.C., Pennell M.W., McGlinn D.J., Eastman J., Moles A.T., Reich P.B., Tank D.C., Wright I.J., Aarssen L., Beaulieu J.M., Kooyman R.M., Leishman M.R., Miller E.T., Niinemets U., Oleksyn J., Ordonez A., Royer D.L., Smith S.A., Stevens P.F., Warman L., Wilf P., and Zanne A.E. 2014. Functional distinctiveness of major plant lineages. Journal of Ecology 102:345–356.

Costello M.J., Wilson S., and Houlding B. 2011. Predicting total global species richness using rates of species description and estimates of taxonomic effort. Systematic Biology 61:871–883.

Coyne J.A. and Orr H.A. 2004. Speciation. Sinauer Associates.

Darwin C. 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. John Murray, London.

Dennett D.C. 1995. Darwin's Dangerous Idea: Evolution and the meanings of life. Simon & Schuster.

Díaz-Uriarte R. and Garland T. 1996. Testing hypotheses of correlated evolution using phylogenetically independent contrasts: Sensitivity to deviations from Brownian Motion. Systematic Biology 45:27–47.

Dobigny G., Ozouf-Costaz C., Bonillo C., and Volobouev V. 2004. Viability of X-autosome translocations in mammals: an epigenomic hypothesis from a rodent case-study. Chromosoma 113:34–41.

Doebeli M. 2011. Adaptive Diversification. Princeton University Press.

Donoghue M.J. and Edwards E.J. 2014. Biome shifts and niche evolution in plants. Annual Reviews of Ecology, Evolution, and Systematics 45:547–572.

Eastman J.M., Alfaro M.E., Joyce P., Hipp A.L., and Harmon L.J. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. Evolution 65:3578–3589.

Eastman J.M., Harmon L.J., and Tank D.C. 2013a. Congruification: support for time scaling large phylogenetic trees. Methods in Ecology and Evolution 4:688–691.

Eastman J.M., Wegmann D., Leuenberger C., and Harmon L.J. 2013b. Simpsonian 'evolution by jumps' in an adaptive radiation of anolis lizards. ArXiv:1305.4216.

Eble G.J. 2000. Contrasting evolutionary flexibility in sister groups: disparity and diversity in mesozoic atelostomate echinoids. Paleobiology 26:56–79.

Eldredge N. 1971. The allopatric model and phylogeny in paleozoic invertebrates. Evolution 25:156–167.

Eldredge N. and Gould S.J. 1972. Punctuated equilibria: an alternative to phyletic gradualism. *In* Models in Paleobiology (T. Schopf, ed.), pages 82–115, Freman, Cooper & Co.

Eldredge N., Thompson J.N., Brakefield P.M., Gavrilets S., Jablonski D., Jackson J.B.C., Lenski R.E., Lieberman B.S., McPeek M.A., and Miller William I. 2005. The dynamics of evolutionary stasis. Paleobiology 31:133–145.

Ellegren H. 2011. Sex-chromsome evolution: recent progress and the influence of male and female heterogamety. Nature Reviews Genetics 12:257–266.

Estes S. and Arnold S.J. 2007. Resolving the paradox of stasis: Models with stabilizing selection explain evolutionary divergence on all timescales. The American Naturalist 169:227–244.

Etienne R.S. and Haegeman B. 2012. A conceptual and statistical framework for adaptive radiations with a key role for diversity dependence. The American Naturalist 180:E75–E89.

Etienne R.S., Haegerman B., Stadler T., Aze T., Pearson P.N., Purvis A., and Phillimore A.B. 2012. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. Proceedings of the Royal Society B: Biological Sciences 279:1300–1309.

Ezaz T., Sarre S.D., O'Meally D., Marshall Graves J.A., and Georges A. 2009. Sex chromosome evolution in lizards: independent origins and rapid transitions. Cytogenetics and Genome Research 127:249–260.

Ezaz T., Stiglec R., Veyrunes F., and Graves J.A.M. 2006. Relationships between vertebrate ZW and XY sex chromosome systems. Current Biology 16:R736–R743.

Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. American Journal of Human Genetics 25:471–492.

Felsenstein J. 1985. Phylogenies and the comparative method. The American Naturalist 125:1–15.

Felsenstein J. 1988. Phylogenies and quantitative characters. Annual Review of Ecology and Systematics 19:445–471.

Felsenstein J. 2012. A comparative method for both discrete and continuous characters using the threshold model. The American Naturalist 179:145–156.

FitzJohn R.G. 2010. Quantitative traits and diversification. Systematic Biology 59:619–633.

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

FitzJohn R.G. 2012b. What Drives Biological Diversification? Detecting Traits Under Species Selection. Ph.D. thesis, University of British Columbia.

FitzJohn R.G., Maddison W.P., and Otto S.P. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Systematic Biology 58:595–611.

FitzJohn R.G., Pennell M.W., Zanne A.E., Stevens P.F., Tank D.C., and Cornwell W.K. 2014. Data from: How much of the world is woody? Journal of Ecology. Dryad Digital Repository. doi:10.5061/dryad.v7m14.

Foote M. 1996. On the probability of ancestors in the fossil record. Paleobiology 22:141–151.

Foote M. 1997. The evolution of morphological diversity. Annual Review of Ecology and Systematics 28:129–152.

Frank S.A. 2009. The common patterns of nature. Journal of Evolutionary Biology 22:1563–1585.

Frank S.A. 2012. Natural selection. IV. The Price equation. Journal of Evolutionary Biology 25:1002–1019.

Frank S.A. 2014. Generative models versus underlying symmetries to explain biological pattern. Journal of Evolutionary Biology 27:1172–1178.

Freckleton R.P. 2012. Fast likelihood calculations for comparative analyses. Methods in Ecology and Evolution 3:940–947.

Freckleton R.P., Cooper N., and Jetz W. 2011. Comparative methods as a statistical fix: The dangers of ignoring an evolutionary model. The American Naturalist 178:E10–E17.

Freckleton R.P. and Harvey P.H. 2006. Detecting non-Brownian trait evolution in adaptive radiations. PLoS Biology 4:e373.

Freckleton R.P., Harvey P.H., and Pagel M. 2002. Phylogenetic analysis and comparative data: A test and review of evidence. The American Naturalist 160:pp. 712–726.

Fritz S.A., Schnitzler J., Eronen J.T., Hof C., Böhning-Gaese K., and Graham C.H. 2013. Diversity in time and space: wanted dead and alive. Trends in Ecology & Evolution .

Fryer G., Greenwood P., and Peake J. 1983. Punctuated equilibria, morphological stasis and the palaeontological documentation of speciation: a biological appraisal of a case history in an african lake. Biological Journal of the Linnean Society 20:195–205.

Fujita M.K., Leaché A.D., Burbrink F.T., McGuire J.A., and Moritz C. 2012. Coalescent-based species delimitation in an integrative taxonomy. Trends in Ecology & Evolution 27:480–488.

Futuyma D.J. 1987. On the role of species in anagenesis. The American Naturalist 130:465–473.

Futuyma D.J. 2010. Evolutionary constraint and ecological consequences. Evolution 64:1865–1884.

Garamszegi L.Z., ed. 2014. Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology. Springer.

Gardner R.J.M., Sutherland G.R., and Shaffer L.G. 2012. Chromosome abnoralities and genetic counseling. Oxford University Press.

Garland T., Dickerman A.W., Janis C.M., and Jones J.A. 1993. Phylogenetic analysis of covariance by computer simulation. Systematic Biology 42:265–292.

Garland T., Harvey P.H., and Ives A.R. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Systematic Biology 41:18–32.

Garland T. and Ives A.R. 2000. Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. The American Naturalist 155:346–364.

Garland T., Midford P.E., and Ives A.R. 1999. An introduction to phylogenetically based statistical methods, with a new method for confidence intervals on ancestral values. American Zoologist 39:374–388.

Gavrilets S. 2004. Fitness Landscapes and the Origin of Species. Princeton University Press.

GBIF. 2013. GBIF dataportal. http://code.google.com/p/gbif-dataportal/.

Gelman A. 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1:515–534.

Gelman A., Carlin J.B., Stern H.S., and Rubin D.B. 2003. Bayesian Data Analysis. 2nd Edition. Chapman & Hall/CRC.

Gelman A., Meng X., and Stern H. 1996. Posterior predictive assessment of model fitness via realized discrepencies (with discussion). Statistica Sinica 6:733–807.

Gelman A. and Shalizi C.R. 2013. Philosophy and the practice of Bayesian statistics. British Journal of Mathematical and Statistical Psychology 66:8–38.

Gerrienne P., Gensel P.G., Strullu-Derrien C., Lardeux H., Steemans P., and Prestianni C. 2011. A simple type of wood in two early Devonian plants. Science 333:837–837.

Givnish T.J. 1998. Adaptive plant evolution on islands: classical patterns, molecular data, new insights. *In* Evolution on Islands (P. Grant, ed.), pages 281–304, Oxford University Press.

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

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

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

Goldie X., Lanfear R., and Bromham L. 2011. Diversification and the rate of molecular evolution: no evidence of a link in mammals. BMC Evolutionary Biology 11:286.

Goldman N. 1993. Statistical tests of models of DNA substitution. Journal of Molecular Evolution 36:182–198.

Gould S.J. 1980. Is a new and general theory of evolution emerging? Paleobiology 6:119–130.

Gould S.J. 2002. The Structure of Evolutionary Theory. Harvard University Press.

Gould S.J. and Eldredge N. 1977. Punctuated equilibria: The tempo and mode of evolution reconsidered. Paleobiology 3:115–151.

Grafen A. 1989. The phylogenetic regression. Philosophical Transactions of the Royal Society B: Biological Sciences 326:119–157.

Grant P.R. and Grant B.R. 2002. Unpredictable evolution in a 30-year study of Darwin's finches. Science 296:707–711.

Gray A. 1887. The elements of botany for beginners and for schools. American Book Company.

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

Grey M., Haggart J.W., and Smith P.L. 2008. Variation in evolutionary patterns across the geographic range of a fossil bivalve. Science 322:1238–1241.

Groover A.T. 2005. What genes make a tree a tree? Trends in Plant Science 10:210–214.

Grossmann V., Höckner M., Karmous-Benailly H., Liang D., Puttinger R., Quadrelli R., Röthlisberger B., Huber A., Wu L., Spreiz A., Fauth C., Erdel M., Zschocke J., Utermann G., and Kotzot D. 2010. Parental origin of apparently balanced de novo complex chromosomal rearrangements investigated by microdissection, whole genome amplification, and microsatellite-mediated haplotype analysis. Clinical Genetics 78:548–553.

Guénard G., Legendre P., and Peres-Neto P. 2013. Phylogenetic eigenvector maps: a framework to model and predict species traits. Methods in Ecology and Evolution 4:1120–1131.

Guerrero R.F. and Kirkpatrick M. 2014. Local adaptation and the evolutoin of chromosome fusions. Evolution 68:2747–2756.

Hadfield J.D. and Nakagawa S. 2010. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. Journal of Evolutionary Biology 23:494–508.

Hannisdal B. 2007. Inferring phenotypic evolution in the fossil record by Bayesian inversion. Paleobiology 33:98–115.

Hansen T.F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.

Hansen T.F. 2012. Adaptive landscapes and macroevolutionary dynamics. *In* The Adaptive Landscape in Evolutionary Biology (E. Svensson and R. Calsbeek, eds.), pages 205–221, Oxford University Press.

Hansen T.F. and Bartoszek K. 2012. Interpreting the evolutionary regression: The interplay between observational and biological errors in phylogenetic comparative studies. Systematic Biology 61:413–425.

Hansen T.F. and Houle D. 2004. Evolvability, stabilizing selection and the problem of stasis. *In* The Evolutionary Biology of Complex Phenotypes (M. Pigliucci and K. Preston, eds.), pages 130–150, Oxford University Press.

Hansen T.F. and Martins E.P. 1996. Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. Evolution 50:1404–1417.

Hansen T.F. and Orzack S.H. 2005. Assessing current adaptation and phylogenetic inertia as explanations of trait evolution: the need for controlled comparisons. Evolution 59:2063–2072.

Hansen T.F., Pienaar J., and Orzack S.H. 2008. A comparative method for stuyding adaptation to a randomly evolving environment. Evolution 62:1965–1977.

Harmon L.J., Losos J.B., Jonathan Davies T., Gillespie R.G., Gittleman J.L., Bryan Jennings W., Kozak K.H., McPeek M.A., Moreno-Roark F., Near T.J., Purvis A., Ricklefs R.E., Schluter D., Schulte II J.A., Seehausen O., Sidlauskas B.L., Torres-Carvajal O., Weir J.T., and Mooers A.Ø. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evolution 64:2385–2396.

Harmon L.J., Weir J.T., Brock C.D., Glor R.E., and Challenger W. 2008. Geiger: investigating evolutionary radiations. Bioinformatics 24:129–131.

Harnik P.G. 2011. Direct and indirect effects of biological factors on extinction risk in fossil bivalves. Proceedings of the National Academy of Sciences, USA 108:13594–13599.

Harnik P.G., Simpson C., and Payne J.L. 2012. Long-term differences in extinction risk among the seven forms of rarity. Proceedings of the Royal Society B: Biological Sciences 279:4969–4976.

Harte J. 2011. Maximum Entropy and Ecology: A Theory of Abundance, Distribution, and Energetics. Oxford.

Harvey P.H. and Pagel M.D. 1991. The comparative method in evolutionary biology. Oxford University Press.

Ho L.S.T. and Ané C. 2014. A linear-time algorithm for Gaussian and Non-Gaussian trait evolution models. Systematic Biology 63:397–408.

Hohenlohe P.A. and Arnold S.J. 2008. MIPoD: A hypothesis-testing framework for microevolutionary inference from patterns of divergence. The American Naturalist 171:366–385.

Holt R.D., Gomulkiewicz R., and Barfield M. 2003. The phenomenology of niche evolution via quantitative traits in a 'black-hole' sink. Proceedings of the Royal Society B: Biological Sciences 270:215–224.

Hopkins M.J. and Lidgard S. 2012. Evolutionary mode routinely varies among morphological traits within fossil species lineages. Proceedings of the National Academy of Sciences, USA 109:20520–20525.

Hou J., Friedrich A., de Montigny J., and Schacherer J. 2014. Chromosomal rearrangements as a major mechanism in the onset of reproductive isolation in saccharomyces cerevisiae. Current Biology 24:1153–1159.

Houle D., C. P., Wagner G., and Hansen T. 2011. Measurement and meaning in biology. Quarterly Review of Biology 86:3–34.

Housworth E.A., Martins E.P., and Lynch M. 2004. The phylogenetic mixed model. The American Naturalist 163:pp. 84–96.

Hull D.L. 1980. Individuality and selection. Annual Review of Ecology and Systematics 11:311–332.

Hunt G. 2006. Fitting and comparing models of phyletic evolution: random walks and beyond. Paleobiology 32:578–601.

Hunt G. 2007. The relative importance of directional change, random walks, and stasis in the evolution of fossil lineages. Proceedings of the National Academy of Sciences, USA 104:18404–18408.

Hunt G. 2008. Gradual or pulsed evolution: when should punctuational explanations be preferred? Paleobiology 34:360–377.

Hunt G. 2012. Measuring rates of phenotypic evolution and the inseparability of tempo and mode. Paleobiology 38:351–373.

Hunt G., Bell M.A., and Travis M.P. 2008. Evolution toward a new adaptive optimum: phenotypic evolution in a fossil stickleback lineage. Evolution 62:700–710.

Hutchinson J., ed. 1973. The Families of Flowering Plants, vol. 2,3. Clarendon Press, Oxford.

Ingram T. 2011. Speciation along a depth gradient in a marine adaptive radiation. Proceedings of the Royal Society B: Biological Sciences 278:613–618.

Ingram T. and Mahler D.L. 2013. SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise AIC. Methods in Ecology and Evolution 4:416–425.

Ives A.R. and Garland Jr. T. 2010. Phylogenetic logistic regression for binary dependent variables. Systematic Biology 59:9–26.

Ives A.R., Midford P.E., and Garland Jr. T. 2007. Within-species variation and measurement error in phylogenetic comparative methods. Systematic Biology 56:252–270.

Jablonski D. 1986. Background and mass extinctions: The alternation of macroevolutionary regimes. Science 231:129–133.

Jablonski D. 2008. Species selection: Theory and data. Annual Review of Ecology, Evolution, and Systematics 39:501–524.

Jablonski D. and Hunt G. 2006. Larval ecology, geographic range, and species survivorship in cretaceous mollusks: Organismic versus species-level explanations. The American Naturalist 168:556–564.

Jackson J.B. and Cheetham A.H. 1999. Tempo and mode of speciation in the sea. Trends in Ecology & Evolution 14:72–77.

Jaynes E.T. 2003. Probability Theory: The Logic of Science. Cambridge University Press.

Jetz W., Thomas G.H., Joy J.B., Hartmann K., and Mooers A.Ø. 2012. The global diversity of birds in space and time. Nature 491:444–448.

Joppa L.N., Roberts D.L., and Pimm S.L. 2010. How many species of flowering plants are there? Proceedings of the Royal Society B: Biological Sciences 278:554–559.

Joyce P. and Majoram P. 2008. Approximately sufficient statistics and Bayesian computation. Statistical applications in genetics and molecular biology 7:1–16.

Judd W.S., Sanders R.W., and Donoghue M.J. 1994. Angiosperm family pairs: preliminary phylogenetic analyses. Harvard Papers Botany 5:1–51.

Kattge J., Diaz S., Lavorel S., Prentice I., Leadley P., Bönisch G., Garnier E., Westoby M., Reich P.B., Wright I., *et al.* 2011. Try–a global database of plant traits. Global Change Biology 17:2905–2935.

Kendall D. 1948. On the generalized "birth-death" process. Annals of Mathematical Statistics 19:1–15.

Khaitovich P., Pääbo S., and Weiss G. 2005. Toward a neutral evolutionary model of gene expression. Genetics 170:929–939.

Kimura M. 1962. On the probability of fixation of mutant genes in a population. Genetics 47:713–719.

King M. 1993. Species evolution: the role of chromosome change. Cambridge University Press.

Kirkpatrick M. and Hall D.W. 2004. Sexual selection and sex linkage. Evolution 58:683–691.

Kitano J. and Peichel C. 2012. Turnover of sex chromosomes and speciation in fishes. Environmental Biology of Fishes 94:549–558.

Kleyer M., Bekker R., Knevel I., Bakker J., Thompson K., Sonnenschein M., Poschlod P., Van Groenendael J., Klimeš L., Klimešová J., *et al.* 2008. The LEDA traitbase: a database of life-history traits of the northwest european flora. Journal of Ecology 96:1266–1274.

Labra A., Pienaar J., and Hansen T.F. 2009. Evolution of thermal physiology in liolaemus lizards: Adaptation, phylogenetic inertia, and niche tracking. The American Naturalist 174:204–220.

Lande R. 1976. Natural selection and random genetic drift in phenotypic evolution. Evolution 30:314–334.

Landis M.J., Schraiber J.G., and Liang M. 2013. Phylogenetic analysis using Lévy processes: Finding jumps in the evolution of continuous traits. Systematic Biology 62:193–204.

Lanfear R., Ho S.Y.W., Love D., and Bromham L. 2010. Mutation rate is linked to diversification in birds. Proceedings of the National Academy of Sciences, USA 107:20423–20428.

Laporte V. and Charlesworth B. 2002. Effective population size and population subdivision in demographically structured populations. Genetics 162:501–519.

Leishman M.R., J. W.I., Moles A.T., and Westoby M. 2000. The evolutionary ecology of seed size. *In* Seeds: The Ecology of Regeneration in Plant Communities (M. Fenner, ed.), pages 31–57, CAB Int.

Levinton J.S. 2001. Genetics, Paleontology and Macroevolution. Cambridge University Press.

Lewis P.O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Systematic Biology 50:913–925.

Lewis P.O., Xie W., Chen M.H., Fan Y., and Kuo L. 2014. Posterior predictive Bayesian phylogenetic model selection. Systematic Biology 63:309–321.

López-Fernández H., Arbour J.H., Winemiller K.O., and Honeycutt R.L. 2013. Testing for ancient adaptive radiations in neotropical cichlid fishes. Evolution 67:1321–1337.

Losos J.B. 2011. Seeing the forest for the trees: The limitations of phylogenies in comparative biology. The American Naturalist 177:709–727.

Lynch M. 1990. The rate of morphological evolution in mammals from the standpoint of the neutral expectation. The American Naturalist 136:727–741.

Lynch M. 1991. Methods for the analysis of comparative data in evolutionary biology. Evolution 45:1065–1080.

Lynch M. and Hill W.G. 1986. Phenotypic evolution by neutral mutation. Evolution 40:915–935.

Maddison W.P. 1990. A method for testing the correlated evolution of two binary characters: are gains and losses concentrated on certain branches of a phylogenetic tree? Evolution 44:539–557.

Maddison W.P. 2006. Confounding asymmetries in evolution diversification and character change. Evolution 60:1743–1746.

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

Maddison W.P. and Leduc-Robert G. 2013. Multiple origins of sex chromosome fusions correlated with chiasma localization in habronattus jumping spiders (araneae: Salticidae). Evolution 67:2258–2272.

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

Magnuson-Ford K. and Otto S.P. 2012. Linking the investigations of character evolution and species diversification. The American Naturalist 180:225–245.

Mahler D.L., Ingram T., Revell L.J., and Losos J.B. 2013. Exceptional convergence on the macroevolutionary landscape in island lizard radiations. Science 341:292–295.

Maliska M.E., Pennell M.W., and Swalla B.J. 2013. Developmental mode influences diversification in ascidians. Biology Letters 9:20130068.

Marshall C.R. 1990. Confidence intervals on stratigraphic ranges. Paleobiology 16:1–10.

Marshall C.R. 1994. Confidence intervals on stratigraphic ranges: Partial relaxation of the assumption of randomly distributed fossil horizons. Paleobiology 20:459–469.

Marshall C.R. 1995. Stratigraphy, the true order of species' originations and extinctions, and testing ancestor-descendant hypotheses among Caribbean bryozans. *In* New Approaches to Speciation in the Fossil Record (D.H. Erwin and R.L. Anstey, eds.), pages 208–236, Columbia University Press.

Marshall C.R. 1997. Confidence intervals on stratigraphic ranges with non-random distributions of fossil horizons. Paleobiology 23:165–173.

Martins E.P. and Hansen T.F. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. The American Naturalist 149:646–667.

Mattila T.M. and Bokma F. 2008. Extant mammal body masses suggest punctuated equilibrium. Proceedings of the Royal Society B: Biological Sciences 275:2195–2199.

Mayr E. 1942. Systematics and the Origin of Species from the Viewpoint of a Zoologist. Harvard University Press.

Mayr E. 1982. Speciation and macroevolution. Evolution 36:1119–1132.

McPeek M.A. 2008. The ecological dynamics of clade diversification and community assembly. The American Naturalist 172:E270–E284.

McShea D. 2004. A revised Darwinism. Biology and Philosophy 19:45–53.

McShea D.W. 1994. Mechanisms of large-scale evolutionary trends. Evolution 48:1747–1763.

McShea D.W. 1998. Possible largest-scale trends in organismal evolution: Eight "live hypotheses". Annual Review of Ecology and Systematics 29:293–318.

Minin V., Abdo Z., Joyce P., and Sullivan J. 2003. Performance-based selection of likelihood models for phylogeny estimation. Systematic Biology 52:674–683.

Mitter C., Farrell B., and Wiegmann B. 1988. The phylogenetic study of adaptive zones: Has phytophagy promoted insect diversification? The American Naturalist 132:107–128.

Moles A.T., Warton D.I., Warman L., Swenson N.G., Laffan S.W., Zanne A.E., Pitman A., Hemmings F.A., and Leishman M.R. 2009. Global patterns in plant height. Journal of Ecology 97:923–932.

Mooers A.Ø. and Schluter D. 1998. Fitting macroevolutionary models to phylogenetic trees. Contributions to Zoology 68:3–18.

Mooers A.Ø., Vamosi S.M., and Schluter D. 1999. Using phylogenies to test macroevolutionary hypotheses of trait evolution in cranes (Gruinae). The American Naturalist 154:249–259.

Nachman M.W. and Searle J.B. 1995. Why is the house mouse karyotype so variable? Trends in Ecology & Evolution 10:397–402.

Nakagawa S. and Freckleton R.P. 2008. Missing inaction: the dangers of ignoring missing data. Trends in Ecology & Evolution 23:592–596.

Neal R.M. 2003. Slice sampling. The Annals of Statistics 31:705–741.

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

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

Nee S., Mooers A.Ø., and Harvey P.H. 1992. Tempo and mode of evolution revealed from molecular phylogenies. Proceedings of the National Academy of Sciences, USA 89:8322–8326.

Nosil P. 2012. Ecological Speciation. Oxford University Press.

Nuismer S.L. and Harmon L.J. 2015. Predicting rates of interspecific interaction from phylogenetic trees. Ecology Letters 18:17–27.

Ohno S. 1967. Sex chromosomes and sex-linked genes. Springer.

Okasha S. 2006. Evolution and the Levels of Selection. Oxford University Press.

O'Meara B.C. 2012. Evolutionary inferences from phylogenies: A review of methods. Annual Review of Ecology, Evolution, and Systematics 43:267–285.

O'Meara B.C., Ané C., Sanderson M.J., and Wainwright P.C. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933.

Pagel M. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. Proceedings of the Royal Society B: Biological Sciences 255:37–45.

Pagel M. 1997. Inferring evolutionary processes from phylogenies. Zoologica Scripta 26:331–348.

Pagel M. 1999. Inferring the historical patterns of biological evolution. Nature 401:877–884.

Pagel M., Venditti C., and Meade A. 2006. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. Science 314:119–121.

Pakeman R.J. 2014. Functional trait metrics are sensitive to the completeness of the species' trait data. Methods in Ecology and Evolution 5:9–15.

Paradis E., Claude J., and Strimmer K. 2004. Ape: Analyses of phylogenetics and evolution in r language. Bioinformatics 20:289–290.

Pardo-Manuel de Villena F. and Sapienza C. 2001a. Female meiosis drives karyotypic evolution in mammals. Genetics 159:1179–1189.

Pardo-Manuel de Villena F. and Sapienza C. 2001b. Non-random segregation during meiosis: the unfairness of females. Mammalian Genome 12:331–339.

Patzkowsky M.E. and Holland S.M. 2012. Stratigraphic Paleobiology: Understanding the distribution of fossil taxa in time and space. University of Chicago Press.

Pennell M.W. 2015. Modern phylogenetic comparative methods and their application in evolutionary biology: Concepts and practice—Book Review. Systematic Biology 64:161–163.

Pennell M.W., Eastman J.M., Slater G.J., Brown J.W., Uyeda J.C., FitzJohn R.G., Alfaro M.E., and Harmon L.J. 2014a. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Bioinformatics 15:2216–2218.

Pennell M.W. and Harmon L.J. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. Annals of the New York Academy of Sciences 1289:90–105.

Pennell M.W., Harmon L.J., and Uyeda J.C. 2014b. Is there room for punctuated equilibrium in macroevolution? Trends in Ecology & Evolution 29:23–32.

Pennell M.W., Harmon L.J., and Uyeda J.C. 2014c. Speciation is unlikely to drive divergence rates. Trends in Ecology & Evolution 29:72–73.

Pennell M.W., Sarver B.A.J., and Harmon L.J. 2012. Trees of unusual size: biased inference of early bursts from large molecular phylogenies. PLoS ONE 7:e43348.

Pérez-Ortín J.E., Querol A., Puig S., and Barrio E. 2002. Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. Genome Research 12:1533–1539.

Phillimore A.B. and Price T.D. 2008. Density-dependent cladogenesis in birds. PLoS Biology 6:e71.

Price G.R. 1972. Extension of covariance selection mathematics. Annals of Human Genetics 35:485–490.

Price T. 1997. Correlated evolution and independent contrasts. Philosophical Transactions of the Royal Society B: Biological Sciences 352:519–529.

Purvis A. and Rambaut A. 1995. Comparative analysis by independent contrasts (caic): an apple macintosh application for analysing comparative data. Computer applications in the biosciences 11:247–251.

Pyron R.A. and Burbrink F.T. 2014. Early origin of viviparity and multiple reversions to oviparity in squamate reptiles. Ecology Letters 17:13–21.

Pyron R.A., Burbrink F.T., and Wiens J.J. 2013. A phylogeny and revised classification of squamata, including 4161 species of lizards and snakes. BMC Evolutionary Biology 13:93.

Quental T.B. and Marshall C.R. 2010. Diversity dynamics: molecular phylogenies need the fossil record. Trends in Ecology & Evolution 25:434–441.

Quintero I. and Wiens J.J. 2013. Rates of projected climate change dramatically exceed past rates of climatic niche evolution among vertebrate species. Ecology Letters 16:1095–1103.

R Development Core Team. 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rabosky D.L. 2009. Ecological limits and diversification rate: alternative paradigms to explain the variation in species richness among clades and regions. Ecology Letters 12:735–743.

Rabosky D.L. 2012. Positive correlation between diversification rates and phenotypic evolvability can mimic punctuated equilibrium on molecular phylogenies. Evolution 66:2622–2627.

Rabosky D.L. and Adams D.C. 2012. Rates of morphological evolution are correlated with species richness in salamanders. Evolution 66:1807–1818.

Rabosky D.L., Donnellan S.C., Grundler M., and Lovette I.J. 2014. Analysis and visualization of complex macroevolutionary dynamics: An example from australian scincid lizards. Systematic Biology 63:610–627.

Rabosky D.L., Donnellan S.C., Talaba A.L., and Lovette I.J. 2007. Exceptional among-lineage variation in diversification rates during the radiation of australia's most diverse vertebrate clade. Proceedings of the Royal Society B: Biological Sciences 274:2915–2923.

Rabosky D.L. and Glor R.E. 2010. Equilibrium speciation dynamics in a model adaptive radiation of island lizards. Proceedings of the National Academy of Sciences, USA 107:22178–22183.

Rabosky D.L. and Lovette I.J. 2008. Explosive evolutionary radiations: Decreasing speciation or increasing extinction through time? Evolution 62:1866–1875.

Rabosky D.L. and McCune A.R. 2010. Reinventing species selection with molecular phylogenies. Trends in Ecology & Evolution 25:68–74.

Rabosky D.L., Santini F., Eastman J., Smith S.A., Sidlauskus B., Chang J., and Alfaro M.E. 2013. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. Nature Communications 4:1958.

Reid N.M., Hird S.M., Brown J.M., Pelletier T.A., McVay J.D., Satler J.D., and Carstens B.C. 2014. Poor fit to the multispecies coalescent is widely detectable in empirical data. Systematic Biology 63:322–333.

Reitan T., Schweder T., and Henderiks J. 2012. Phenotypic evolution studied by layered stochastic differential equations. Annals of Applied Statistics 6:1531–1551.

Revell L.J., Mahler D.L., Peres-Neto P.R., and Redelings B.D. 2012. A new phylogenetic method for identifying exceptional phenotypic diversification. Evolution 66:135–146.

Rice S.H. 1995. A genetical theory of species selection. Journal of Theoretical Biology 177:237–245.

Rice S.H. 2004. Evolutionary Theory. Sinauer and Associates.

Ricklefs R.E. 2004. Cladogenesis and morphological diversification in passerine birds. Nature 430:338–341.

Ricklefs R.E. 2007. Estimating diversification rates from phylogenetic information. Trends in Ecology & Evolution 22:601–610.

Ridley M. 1983. The explanation of organic diversity: the comparative method and adaptations for mating. Oxford University Press.

Ripplinger J. and Sullivan J. 2010. Assessment of substitution model adequacy using frequentist and Bayesian methods. Molecular Biology and Evolution 27:2790–2803.

Robert C.P. 2007. The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, 2nd Ed. Springer.

Rohlf F.J. 2001. Comparative methods for the analysis of continuous variables: geometric interpretations. Evolution 55:2143–2160.

Rohlf F.J. 2006. A comment on phylogenetic regression. Evolution 60:1509–1515.

Rose M.R. and Lauder G.V. 1996. Adaptation. Academic Press.

Rosenblum E.B., Sarver B.A.J., Brown J.W., Des Roches S., Hardwick K.M., Hether T.D., Eastman J.M., Pennell M.W., and Harmon L.J. 2012. Goldilocks meets Santa Rosalia: an emphemeral speciation model explains patterns of diversification across time scales. Evolutionary Biology 39:255–261.

Rowe N. and Paul-Victor C. 2012. Herbs and secondary woodiness–keeping up the cambial habit. New Phytologist 193:3–5.

Roy K. 1996. The roles of mass extinction and biotic interaction in large-scale replacements: A reexamination using the fossil record of stromboidean gastropods. Paleobiology 22:436–452.

Royal Botanical Gardens, Kew. 2014. Seed Information Database (SID), Version 7.1, Accessed 25 March. Http://data.kew.org/sid.

Rubin D. 1984. Bayesian justifiable and relevant frequency calculations for the applied statistician. Annals of statistics 12:1151–1172.

Rundle H.D. and Nosil P. 2005. Ecological speciation. Ecology Letters 8:336–352.

Ruse M. 1989. Is the theory of punctuated equilibria a new paradigm? *In* The Darwinian Paradigm (M. Ruse, ed.), Harvard University Press.

Sætre G.P. 2013. Hybridization is important in evolution, but is speciation? Journal of Evolutionary Biology 26:256–258.

Sargent R.D. 2004. Floral symmetry affects speciation rates in angiosperms. Proceedings of the Royal Society B: Biological Sciences 271:603–608.

Sartorelli E.M.P., Mazzucatto L.F., and de Pina-Neto J.M. 2001. Effect of paternal age on human sperm chromosomes. Fertility and Sterility 76:1119–1123.

Schluter D. 2000. The Ecology of Adaptive Radiation. Oxford University Press.

Schluter D., Price T., Mooers A.Ø., and Ludwig D. 1997. Likelihood of ancestor states in adaptive radiation. Evolution 51:1699–1711.

Schraiber J.G. and Landis M.J. 2014. Sensitivity of quantitative traits to mutational effects, number of loci, and population history. BioRxiv DOI: 10.1101/008540.

Schubert I. and Lysak M.A. 2011. Interpretation of karyotype evolution should consider chromosome structural constraints. Trends in Genetics 27:207–216.

Scotland R.W. and Sanderson M.J. 2004. The significance of few versus many in the tree of life. Science 303:643–643.

Searle J.B. 1986. Preferential transmission in wild common shrews (sorex araneus), heterozygous for robertsonian rearrangements. Genetical research 47:147–148.

Sepkoski D. 2012. Rereading the Fossil Record: The Growth of Paleobiology and an Evolutionary Discripline. University of Chicago Press.

Sepkoski J.J.J. 1984. A kinetic model of phanerozoic taxonomic diversity. iii. post-paleozoic families and mass extinctions. Paleobiology 10:246–267.

Sepkoski J.J.J., McKinney F., and Lidgard S. 2000. Competitive displacement among post-paleozoic cyclostome and cheilostome bryozoans. Paleobiology 26:7–18.

Sidje R.B. 1998. Expokit: A software package for computing matrix exponentials. ACM Transactions on Mathematical Software 24:130–156.

Siepielski A.M., DiBattista J.D., and Carlson S.M. 2009. It's about time: the temporal dynamics of phenotypic selection in the wild. Ecology Letters 12:1261–1276.

Siepielski A.M., DiBattista J.D., Evans J.A., and Carlson S.M. 2011. Differences in the temporal dynamics of phenotypic selection among fitness components in the wild. Proceedings of the Royal Society B: Biological Sciences 278:1572–1580.

Simpson C. 2010. Species selection and driven mechanisms jointly generate a large-scale morphological trend in monobathrid crnioids. Paleobiology 36:481–496.

Simpson C. 2013. Species selection and the macroevolution of coral coloniality and photosymbiosis. Evolution 67:1607–1621.

Simpson C. and Harnik P.G. 2009. Assessing the role fo abundance in marine bivalve extinction over the post-paleozoic. Paleobiology 35:631–647.

Simpson G.G. 1944. Tempo and Mode of Evolution. Columbia University Press.

Sinnott E.W. and Bailey I.W. 1915. The evolution of herbaceous plants and its bearing on certain problems of geology and climatology. The Journal of Geology 23:289–306.

Slater G.J. 2013. Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the cretaceous-palaeogene boundary. Methods in Ecology and Evolution 4:734–744.

Slater G.J. 2014. Correction to "Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the cretaceous-palaeogene boundary", and a note on fitting macroevolutionary models to comparative paleontological data sets. Methods in Ecology and Evolution 5:714–718.

Slater G.J., Harmon L.J., and Alfaro M.E. 2012a. Integrating fossils with molecular phylogenies improves inference of trait evolution. Evolution 66:3931–3944.

Slater G.J., Harmon L.J., Wegmann D., Joyce P., Revell L.J., and Alfaro M.E. 2012b. Fitting models of continous trait evolution to incompletely sampled comparative data using Approximate Bayesian Computation. Evolution 66:752–762.

Slater G.J. and Pennell M.W. 2014. Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. Systematic Biology 63:293–308.

Slatkin M. 1981. A diffusion model of species selection. Paleobiology 7:421–425.

Slowinski J.B. and Guyer C. 1989. Testing the stochasticity of patterns of organismal diversity - an improved null model. American Naturalist 134:907–921.

Slowinski J.B. and Guyer C. 1993. Testing whether certain traits have caused amplified diversification - an improved method based on a model of random speciation and extinction. American Naturalist 142:1019–1024.

Smith S.A., Beaulieu J.M., Stamatakis A., and Donoghue M.J. 2011. Understanding angiosperm diversification using small and large phylogenetic trees. American Journal of Botany 98:404–414.

Smith S.A. and Donoghue M.J. 2008. Rates of molecular evolution are linked to life history in flowering plants. Science 322:86–89.

Soltis D.E., Smith S.A., Cellinese N., Wurdack K.J., Tank D.C., Brockington S.F., Refulio-Rodriguez N.F., Walker J.B., Moore M.J., Carlsward B.S., Bell C.D., Latvis M., Crawley S., Black C., Diouf D., Xi Z., Rushworth C.A., Gitzendanner M.A., Sytsma K.J., Qiu Y.L., Hilu K.W., Davis C.C., Sanderson M.J., Beaman R.S., Olmstead R.G., Judd W.S., Donoghue M.J., and Soltis P.S. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. American Journal of Botany 98:704–730.

Spicer R. and Groover A. 2010. Evolution of development of vascular cambia and secondary growth. New Phytologist 186:577–592.

Spiegelhalter D.J., Best N.G., Carlin B.P., and Van Der Linde A. 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64:583–639.

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

Stanley S. 1979. Macroevolution. Johns Hopkins University Press.

Stanley S.M. 1975. A theory of evolution above the species level. Proceedings of the National Academy of Sciences, USA 72:646–650.

Stevens P.F. 2001. Angiosperm Phylogeny Website. Version 12, July 2012 [and more or less continuously updated since].

Stone G.N., Nee S., and Felsenstein J. 2011. Controlling for non-independence in comparative analysis of patterns across populations within species. Philosophical Transactions of the Royal Society B: Biological Sciences 366:1410–1424.

Strotz L.C. and Allen A.P. 2013. Assessing the role of cladogenesis in macroevolution by integrating fossil and molecular evidence. Proceedings of the National Academy of Sciences, USA 110:2904–2909.

Swenson N.G. 2014. Phylogenetic imputation of plant functional trait databases. Ecography 37:105–110.

Tank D., Eastman J.M., Pennell M.W., Soltis P.S., Soltis D.E., Hinchliff C.E., Brown J.W., and Harmon L.J. 2015. Nested radiations and the pulse of angiosperm diversification. New Phytologist .

The Angiosperm Phylogeny Group. 2009. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG III. Botanical Journal of the Linnean Society 161:105–121.

The Plant List. 2014. Version 1.1. Published on the Internet. http://www.theplantlist.org/ accessed 11 March.

The Tree of Sex Consortium. 2014. Tree of sex: a database of sexual systems. Scientific Data 1:140015.

Theophrastus. 1916. Enquiry Into Plants, Translated by A.F. Hort. Harvard University Press.

Thomas G.H., Cooper N., Venditti C., Meade A., and Freckleton R.P. 2014. Bias and measurement error in comparative analyses: a case study with the Ornstein-Uhlenbeck model. BioRxiv DOI: 10.1101/004036.

Thomas G.H. and Freckleton R.P. 2012. MOTMOT: models of trait macroevolution on trees. Methods in Ecology and Evolution 3:145–151.

Thomas G.H., Freckleton R.P., and Székely T. 2006. Comparative analyses of the influence of developmental mode on phenotypic diversification rates in shorebirds. Proceedings of the Royal Society B: Biological Sciences 273:1619–1624.

Thomas N.S., Morris J.K., Baptista J., Ng B.L., Crolla J.A., and Jacobs P.A. 2010. De novo apparently balanced translocations in man are predominantly paternal in origin and associated with a significant increase in paternal age. Journal of Medical Genetics 47:112–115.

Turner D. 2010. Punctuated equilibrium and species selection: what does it mean for one theory to suggest another? Theory in Biosciences 129:113–123.

Uyeda J.C., Caetano D.S., and Pennell M.W. 2015. Comparative analysis of principal components can be misleading. Systematic Biology .

Uyeda J.C., Hansen T.F., Arnold S.J., and Pienaar J. 2011. The million-year wait for macroevolutionary bursts. Proceedings of the National Academy of Sciences, USA 108:15908–15913.

Uyeda J.C. and Harmon L.J. 2014. Bayesian reversible-jump modeling of adaptive shifts for studying macroevolutionary patterns of trait evolution. Systematic Biology 63:902–918.

Vamosi J.C. and Vamosi S.M. 2004. The role of diversification in causing the correlates of dioecy. Evolution 58:723–731.

Vamosi S.M. and Vamosi J.C. 2005. Endless tests: guidelines for analysing non-nested sister-group comparisons. Evolutionary Ecology Research 7:657–579.

Van Bocxlaer B., Damme D.V., and Feibel C.S. 2008. Gradual versus punctuated evolution in the turkana basin molluscs: Evolutionary events or biological invasions? Evolution 62:511–520.

Van Doorn G.S. and Kirkpatrick M. 2007. Turnover of sex chromosomes induced by sexual conflict. Nature 449:909–912.

Van Valen L. 1975. Group selection, sex, and fossils. Evolution 29:87–94.

Venditti C. and Pagel M. 2010. Speciation as an active force in promoting genetic evolution. Trends in Ecology & Evolution 25:14–20.

Vermeij G.J. 1987. Evolution and Escalation. Princeton University Press.

Vrba E.S. and Gould S.J. 1986. The hierarchical expansion of sorting and selection: Sorting and selection cannot be equated. Paleobiology 12:217–228.

Wagner P.J. 1996. Contrasting the underlying patterns of active trends in morphologic evolution. Evolution 50:990–1007.

Wagner P.J. 2000. Likelihood tests of hypothesized durations: determining and accommodating-biasing factors. Paleobiology 26:431–449.

Wagner P.J. 2001. Rate heterogeneity in shell character evolution among lophospiroid gastropods. Paleobiology 27:290–310.

Wagner P.J. and Erwin D.H. 1995. Phylogenetic patterns as tests of speciation models. *In* New Approaches to Speciation in the Fossil Record (D.H. Erwin and R.L. Anstey, eds.), pages 87–122, Columbia University Press.

Wagner P.J. and Marcot J.D. 2010. Probabilistic phylogenetic inference in the fossil record: current and future applications. *In* Quantitative Methods in Paleobiology (J. Alroy and G. Hunt, eds.), pages 195–217, Paleontological Society.

Wainwright P.C. 2007. Functional versus morphological diversity in macroevolution. Annual Review of Ecology, Evolution, and Systematics 38:381–401.

Walsh B. and Blows M.W. 2009. Abundant genetic variation + strong selection = multivariate genetic constraints: A geometric view of adaptation. Annual Review of Ecology, Evolution, and Systematics 40:41–59.

Warton D.I. and Hui F.K.C. 2011. The arcsine is asinine: the analysis of proportions in ecology. Ecology 92:3–10.

Webster A.J., Payne R.J.H., and Pagel M. 2003. Molecular phylogenies link rates of evolution and speciation. Science 301:478.

Weir J.T. and Mursleen S. 2013. Diversity-dependent cladogenesis and trait evolution in the adaptive radiation of the auks (Aves: Alcidae). Evolution 67:403–416.

Westoby M., Falster D.S., Moles A.T., Vesk P.A., and Wright I.J. 2002. Plant ecological strategies: Some leading dimensions of variation between species. Annual Review of Ecology and Systematics 33:125–159.

Westoby M., Leishman M.R., and Lord J.M. 1995. On misinterpreting the 'phylogenetic correction'. Journal of Ecology 83:531–534.

White M.J.D. 1973. Animal cytology and evolution. Cambridge University Press.

Williams G.C. 1992. Natural Selection: Domains, Levels and Challenges. Oxford University Press.

Williamson P. 1981. Palaeontological documentation of speciation in cenozoic molluscs from turkana basin. Nature 293:437–443.

Wright I.J., Reich P.B., Westoby M., Ackerly D.D., Baruch Z., Bongers F., Cavender-Bares J., Chapin T., Cornelissen J.H., Diemer M., Flexas J., Garnier E., Groom P.K., Gulias J., Hikosaka K., Lamont B.B., Lee T., Lee W., Lusk C., Midgley J.J., Navas M., Niinemets U., Oleksyn J., Osada N., Poorter H., Poot P., Prior L., Pyankov V.I., Roumet C., Thomas S.C., Tjoelker M.G., Veneklass E.J., and Villar R. 2004. The worldwide leaf economics spectrum. Nature 428:821–827.

Wright S. 1933. Inbreeding and homozygosis. Proceedings of the National Academy of Sciences, USA 19:411–419.

Wyrobek A.J., Eskenazi B., Young S., Arnheim N., Tiemann-Boege I., Jabs E., Glaser R., Pearson F., and Evenson D. 2006. Advancing age has differential effects on dna damage, chromatin integrity, gene mutations, and aneuploidies in sperm. Proceedings of the National Academy of Sciences, USA 103:9601–9606.

Wyttenbach A., Borodin P., and Hausser J. 1997. Meiotic drive favors robertsonian metacentric chromosomes in the common shrew (sorex araneus, insectivora, mammalia). Cytogenetics and cell genetics 83:199–206.

Yoshida K. and Kitano J. 2012. The contribution of female meiotic drive to the evolution of neo-sex chromosomes. Evolution 66:3198–3208.

Zanne A.E., Tank D.C., Cornwell W.K., Eastman J.M., Smith S.A., FitzJohn R.G., McGlinn D.J., O'Meara B.C., Moles A.T., Reich P.B., Royer D.L., Soltis D.E., Stevens P.F., Westoby M., Wright I.J., Aarssen L., Bertin R.I., Calaminus A., Govaerts R., Hemmings F., Leishman M.R., Oleksyn J., Solits P.S., Swenson N.G., Warman L., and Beaulieu J.M. 2014a. Three keys to the radiation of angiosperms into freezing environments. Nature 506:89–92.

Zanne A.E., Tank D.C., Cornwell W.K., Eastman J.M., Smith S.A., FitzJohn R.G., McGlinn D.J., O'Meara B.C., Moles A.T., Reich P.B., Royer D.L., Soltis D.E., Stevens P.F., Westoby M., Wright I.J., Aarssen L., Bertin R.I., Calaminus A., Govaerts R., Hemmings F., Leishman M.R., Oleksyn J., Solits P.S., Swenson N.G., Warman L., and Beaulieu J.M. 2014b. Three keys to the radiation of angiosperms into freezing environments. Nature 506:89–92.

Zanne A.E., Tank D.C., Cornwell W.K., Eastman J.M., Smith S.A., FitzJohn R.G., McGlinn D.J., O'Meara B.C., Moles A.T., Reich P.B., Royer D.L., Soltis D.E., Stevens P.F., Westoby M., Wright I.J., Aarssen L., Bertin R.I., Calaminus A., Govaerts R., Hemmings F., Leishman M.R., Oleksyn J., Solits P.S., Swenson N.G., Warman L., and Beaulieu J.M. 2013. Data from: Three keys to the radiation of angiosperms into freezing environments. Dryad Digital Repository. doi:10.5061/dryad.63q27.2.

APPENDIX A

Supplement to Chapter 4: Theoretical results

Here we describe the theoretical models discussed in the main text. We present only the main results. We consider two main types of models: those that include direct selection on fusions and those that include sexually antagonistic selection. The direct selection models can also be used to consider the special cases of neutral evolution, weak selection, and meiotic drive. For all cases we allow for the possibility of differences between chromosomes in both effective population sizes and mutation rates. Several results have been previously derived (see Charlesworth *et al.*, 1987; Charlesworth, 1994; Kirkpatrick and Hall, 2004). We rederive them here to put them in a unified framework with a consistent notation.

A.0.1 *Direct selection*

We track the rate of appearance and establishment of a sex-autosome fusion, where the rate at which mutation generates a fusion between a sex chromosome and an autosome is μ_C^{sex} per gamete per generation for chromosome C ($C = X, Y, Z, \text{or} W$) in males ($sex = m$) and females ($sex = f$). We assume that, at birth, the population is of constant total size N, consisting of an equal number ($N/2$) of males and females. Not all individuals survive and successfully enter the reproductive pool. Specifically, we assume that the numbers of females and males that reproduce are N^f and N^m, where each of these reproductive individuals is expected to have a Poisson distributed number of offspring. The effective population sizes of Y and W chromosomes are then $N_{e,Y} = N^m$ and $N_{e,W} = N^f$, respectively, while the effective population sizes of X and Z chromosomes equal:

$$N_{e,X} = \frac{9N^f N^m}{N^f + 2N^m} \tag{A.1a}$$

$$N_{e,Z} = \frac{9N^f N^m}{2N^f + N^m} \tag{A.1b}$$

(Wright 1933; see also Caballero 1995; Laporte and Charlesworth 2002 for extensions to non-Poisson distributions). Note that the above equations define the effective number of chromosomes, not the effective number of individuals.

Once the fusion appears, we approximate its establishment rate using Kimura's (1962) diffusion approximation for the fixation probability. Dominance has little effect on which type of fusion is expected to become established most frequently. Hence, we focus here on the simpler additive case, where the fixation probability of a fusion is:

$$P_C = \frac{1 - \exp[-2 s_C N_{e,C} p]}{1 - \exp[-2 N_{e,C} s_C]} \tag{A.2}$$

where s_C is the selection coefficient acting directly upon individuals carrying the fusion when rare (as heterozygotes), p is the initial frequency of the fusion, and $N_{e,C}$ is the relevant effective population size of the chromosome C. (Recall that $N_{e,C}$ is the effective number of chromosomes, not individuals, which is why '2' rather than the standard '4' appears in Equation A.2.) We also assume that selection on the fusion is sufficiently weak that the selection coefficient can be taken as the average over many generations, accounting for the time spent in each sex:

$$s_X = \frac{2}{3} s_X^f + \frac{1}{3} s_X^m \tag{A.3a}$$

$$s_Y = s_Y^m \tag{A.3b}$$

$$s_Z = \frac{1}{3} s_Z^f + \frac{2}{3} s_Z^m \tag{A.3c}$$

$$s_W = s_W^f \tag{A.3d}$$

Below, we consider both the rate at which fusions originate and the rate at which they fix, for fusions involving different sex chromosomes.

Y-A FUSIONS — Y-A fusions appear in the population at rate $\frac{N}{2}\mu_Y^m$. The probability that the fusion fixes is the chance that the fusion is present among the adult males of the population, $N^m/(N/2)$, times the probability that the fusion will be the ultimate ancestor of the Y chromosomes among the descendants after some long period of time, given by (A.2) for the $C = Y$ chromosome with $N_{e,Y} = N^m$ and $p = 1/N^m$. Multiplying the mutation rate by the fixation probability, the overall establishment probability for a Y-A fusion is

$$\begin{aligned} R_Y &= N^m \mu_Y^m P_Y \\ &= N^m \mu_Y^m \frac{1 - \exp[-2s_Y]}{1 - \exp[-2N^m s_Y]} \end{aligned} \tag{A.4}$$

We note that (A.4) is the standard result for the establishment of a mutation in a haploid model, with the additional subscripts and superscripts added for consistency.

X-A FUSIONS — X-A fusions appear in the population at rate $2\frac{N}{2}\mu_X^f$ among females and at rate $\frac{N}{2}\mu_X^m$ among males, where the former expression accounts for the fact that females carry two X chromosomes. A fusion arising in a female has a chance $N^f/\frac{N}{2}$ of surviving to reproduce. The probability that the fusion will be the ultimate ancestor of the X chromosomes after some long period of time is then given by (A.2) for $C = X$, with $N_{e,X}$ given by (A.1a) and $p = \frac{2}{3}/(2N^f)$ accounting for the fact that $\frac{2}{3}$ of the X chromosomes in the next generation come from these mothers, among whom the fusion is at initial frequency $1/(2N^f)$. A similar calculation applies to males, so that the net establishment rate is approximately:

$$R_X = 2N^f \mu_X^f \frac{1 - \exp[-2N_{e,X}(\frac{2}{3}\frac{1}{2N^f})s_X]}{1 - \exp[-2N_{e,X}s_X]} + N^m \mu_X^m \frac{1 - \exp[-2N_{e,X}(\frac{1}{3}\frac{1}{N^m})s_X]}{1 - \exp[-2N_{e,X}s_X]} \tag{A.5}$$

W-A FUSIONS — The establishment rate of W-A fusions, R_W, is derived as for Y-A fusions, giving (A.4) but with m replaced by f and Y replaced by W.

Z-A FUSIONS — The establishment rate of Z-A fusions, R_Z, is derived as for X-A fusions, giving (A.5) but with m and f interchanged and X replaced by Z.

NEUTRAL FUSIONS — When selection is negligible, the above formulae can be simplified substantially. In the limit for neutral fusions ($s_C = 0$), the net establishment rate equals the rate at which each type of fusion arises:

$$R_Y = \mu_Y, \tag{A.6a}$$

$$R_X = \frac{2}{3}\mu_X^f + \frac{1}{3}\mu_X^m, \tag{A.6b}$$

$$R_W = \mu_W, \tag{A.6c}$$

$$R_Z = \frac{2}{3}\mu_Z^m + \frac{1}{3}\mu_Z^f. \tag{A.6d}$$

Observe that the reproductive population sizes of males (N^m) and females (N^f) are irrelevant to the relative rate of fusion establishment when there is no direct selection on the fusions (Charlesworth and Charlesworth, 1980). A neutral fusion is less likely to survive and reproduce if it first appears in the sex with the lower reproductive population size, but if it does, then it has a higher chance of being the progenitor chromosome; these effects exactly cancel out.

Weak selection — The relative establishment rates also get simplified substantially when selection is very weak: $|\theta| << 1$, where $\theta = 4Ns_C$. To leading order in θ, the establishment rate for each type of fusion, measured relative to the rate of X-A fusions, is:

$$\frac{R_Y}{R_X} = \frac{3\alpha}{2+\alpha}\left(1 + \theta\frac{1-4\gamma}{4\gamma(2+\gamma)}\right), \tag{A.7a}$$

$$\frac{R_W}{R_X} = \frac{3}{2+\alpha}\left(1 - \theta\frac{7-\gamma}{8(2+\gamma)}\right), \tag{A.7b}$$

$$\frac{R_Z}{R_X} = \frac{2\alpha+1}{2+\alpha}\left(1 + \theta\frac{9(1-\gamma)}{8(2+\gamma)(1+2\gamma)}\right), \tag{A.7c}$$

where fusions arise in males at a rate $\alpha = \mu^m/\mu^f$ times that in females and the number of reproductive females is $\gamma = N^f/N^m$ times the number of males (so that the sex ratio $N^m/(N^m + N^f) = 1/(\gamma+1)$). In the absence of a sex bias in the mutation rate ($\alpha = 1$) or number of reproductive individuals ($\gamma = 1$), we find that

$$\frac{R_Y}{R_X} = \frac{R_W}{R_X} = 1 - \frac{\theta}{4}$$

and

$$\frac{R_Z}{R_X} = 1.$$

This confirms that direct selection alone cannot explain the predominance of Y-A fusions.

Similarly, the overall rate at which fusions arise in XY systems versus ZW systems is the sum of the rates for the component chromosomes, keeping only leading order terms in θ:

$$\frac{R_X + R_Y}{R_Z + R_W} = \frac{1+2\alpha}{\alpha+2} + \frac{\theta}{2}\Bigg[\left(\frac{3\alpha}{2+\alpha}\right)\left(\frac{1-4\gamma}{4\gamma(2+\gamma)}\right) + \left(\frac{3(1+2\alpha)}{(2+\alpha)^2}\right)\left(\frac{7-\gamma}{8(2+\gamma)}\right) - \left(\frac{(1+2\alpha)^2}{(2+\alpha)^2}\right)\left(\frac{9(1-\gamma)}{8(2+\gamma)(1+2\gamma)}\right)\Bigg]. \tag{A.8}$$

A.0.2 *Sex-Antagonistic selection*

Consider an autosomal locus with selection acting in opposite directions in males and females, with allele A_0 favored in males and allele A_1 in females. If selection is weak, the allele frequency q_i of allele A_i is approximately the same in males and females. Given the sex-specific fitness of genotype ij, W_{ij}^{sex}, we can then define the selection coefficient favoring allele A_i in a particular sex as

$$s_i^{sex} = (W_{i.}^{sex} / \bar{W}^{sex}) - 1.$$

Here $W_{i.}^{sex}$ is the marginal fitness of A_i in that sex ($W_{i.}^{sex} = q_0 W_{i0}^{sex} + q_1 W_{i1}^{sex}$), and $\bar{W}^{sex}$ is the mean fitness ($\bar{W}^{sex} = q_0 W_{0.} + q_1 W_{1.}$).

Following similar logic used to derive equations (A.4) and (A.5), fusions bearing allele A_i arise with the Y chromosome and are found in a reproductive male at rate $q_i \mu_Y^m N^m$ or arise with the W and are found in a reproductive female at rate $q_i \mu_W^f N^f$. Similarly, the rate at which X-A fusions or Z-A fusions bearing allele A_i originate is $q_i(2\mu_X^f N^f + \mu_X^m N^m)$ or $q_i(\mu_Z^f N^f + 2\mu_Z^m N^m)$, respectively. If we assume selection is weak, we can average over the time the chromosome spends in a female and a male to obtain the strength of selection acting on a fusion bearing allele A_i: $s_{X,i} = \frac{2}{3}s_i^f + \frac{1}{3}s_i^m$ for an X-A fusion, $s_{Y,i} = s_i^m$ for a Y-A fusion, $s_{Z,i} = \frac{1}{3}s_i^f + \frac{2}{3}s_i^m$ for a Z-A fusion, and $s_{W,i} = s_i^f$ for a W-A fusion.

Because the X and W are more often found in females, the fixation probability of an X-A or W-A fusion is much higher if it captures the female-benefit allele A_1 than if it captures the male-benefit allele (and *vice versa* for Y-A and Z-A fusions). Using $2s_C N_{e,C} p$ to approximate the fixation probability (A.2) for a beneficial fusion initially at frequency p, the fixation probability of an X-A fusion is approximately $P_X = 2s_{X,1} N_{e,X} p$ when it captures allele A_1 and zero otherwise.

Similarly, $P_W = 2s_{w,1}N_{e,W}p$ when a W-A fusion captures A_1, $P_Y = 2s_{Y,0}N_{e,Y}p$ when a Y-A fusion captures A_0, and $P_Z = 2s_{Z,0}N_{e,Z}p$ when a Z-A fusion captures A_0.

Multiplying together the rate that fusions originate in each sex times their fixation probability (accounting for the initial frequency in that sex), we get the rate at which fusions are expected to become established for each sex chromosome:

$$R_Y = q_0 \mu_Y N^m (2s_0^m), \tag{A.9a}$$

$$R_X = 2q_1 \frac{9N^f N^m}{N^f + 2N^m} \left(\frac{2}{3}\mu_X^f + \frac{1}{3}\mu_X^m\right)\left(\frac{2}{3}s_1^f + \frac{1}{3}s_1^m\right), \tag{A.9b}$$

$$R_W = q_1 \mu_W N^f (2s_1^f), \tag{A.9c}$$

$$R_z = 2q_0 \frac{9N^f N^m}{2N^f + N^m} \left(\frac{1}{3}\mu_Z^f + \frac{2}{3}\mu_Z^m\right)\left(\frac{1}{3}s_0^f + \frac{2}{3}s_0^m\right). \tag{A.9d}$$

At an autosomal locus subject to sexually antagonistic selection, each allele has spent half of its time in males and half in females, rising in frequency in one sex and falling in the other sex. Consequently, to remain at equilibrium over the longer term, the selection coefficients for each allele must balance across the sexes, with $s_0^f = -s_0^m$ and $s_1^f = -s_1^m$. Furthermore, the fitness definitions imply that $q_0 s_0^{sex} + q_1 s_1^{sex}$ must equal zero since they sum to

$$\frac{q_0 W_{0.}^{sex} + q_1 W_{1.}^{sex}}{\bar{W}^{sex}} - 1 = \frac{\bar{W}^{sex}}{\bar{W}^{sex}} - 1 = 0.$$

Using these relationships to substitute for s_i^f and q_1, we find:

$$R_Y = 2s_0^m q_0 (\mu_Y N^m), \tag{A.10a}$$

$$R_X = 2s_0^m q_0 \left(\frac{(2\mu_X^f + \mu_X^m)N^f N^m}{N^f + 2N^m}\right), \tag{A.10b}$$

$$R_W = 2s_0^m q_0 (\mu_W N^f), \tag{A.10c}$$

$$R_Z = 2s_0^m q_0 \left(\frac{(\mu_Z^f + 2\mu_Z^f)N^f N^m}{2N^f + N^m}\right). \tag{A.10d}$$

Thus, with equal mutation rates and equal numbers of reproductive individuals of the two sexes, the establishment rates all equal one another. Otherwise, recalling that $\alpha = \mu^m/\mu^f$ and $\gamma =$

N^f/N^m, the establishment rates relative to the rate of X-A fusions become:

$$\frac{R_Y}{R_X} \approx \frac{\alpha(2+\gamma)}{\gamma(2+\alpha)}, \tag{A.11a}$$

$$\frac{R_W}{R_X} = \frac{2+\gamma}{2+\alpha}, \tag{A.11b}$$

$$\frac{R_Z}{R_X} = \frac{(1+2\alpha)(2+\gamma)}{(1+2\gamma)(2+\alpha)}, \tag{A.11c}$$

Consequently, Y-A fusions are expected to predominate (with $R_Y > \max[R_X, R_W, R_Z]$) if, and only if, $\alpha > \gamma$.

APPENDIX B

Supplement to Chapter 5: Bayesian results

As with the likelihood results (described in main text), OU models were highly supported across many datasets; 177/337 clades had the highest DIC weight (DIC_w) on an OU model; 156 of them with greater than 75% of the total DIC_w (see figure B.1). While a generally similar pattern of model support holds for both likelihood and Bayesian inference, the likelihood analyses are much cleaner (compare Figures 5.6 and B.1). This differnce can be explained by the fact that there is a tight statistical relationship between the AIC values for these three models. If two models have identical likelihoods, the AIC scores, defined as $-2\mathcal{L} + 2k$ (where $\mathcal{L}$ is the log-likelihood of the model and k is the number of parameters) will differ by 2. As BM is a special case of both OU and EB, in opposite directions in model space, the highest AIC_w possible for BM is ~0.731. The rare clades where both OU and EB have higher support than BM likely reflect problems in optimization. Calculating DIC values from posterior samples is inherently more stochastic; if there is little information in data, the best DIC model will depend on the values sampled by the chain.

For the model adequacy results, the results were also very similar to that of the likelihood analyses (compare to Results section in Chapter 5). The adequacy of these simple models was poor across the majority of the datasets (Figure B.2). Again, we limit our analyses of model adequacy to only the most highly supported model in the candidate set.

Of the 72 comparative datasets of SLA, we detected deviations from the expectations of the best supported model using at least one test statistic in 35 cases, 26 by at least two, and 19 by three or more. For the seed mass data, we detected deviations with at least one test statistic in 173 cases (by two or more in 109 datasets and by at least three in 72 cases). 24/39 leaf nitrogen datasets were found to be inadequately described by the best supported model with at least one test statistic (13 by at least two and 10 by at least three).

Also, similar to the likelihood analyses, the frequency at which deviations were found differed between the test statistics. In 171 cases, we detected model misspecification with C_{VAR} and with S_{VAR}, 141 (M_{SIG}: 24, S_{ASR}: 101, S_{HGT}: 78, D_{CDF}: 67). Again, only 105 datasets were adequately

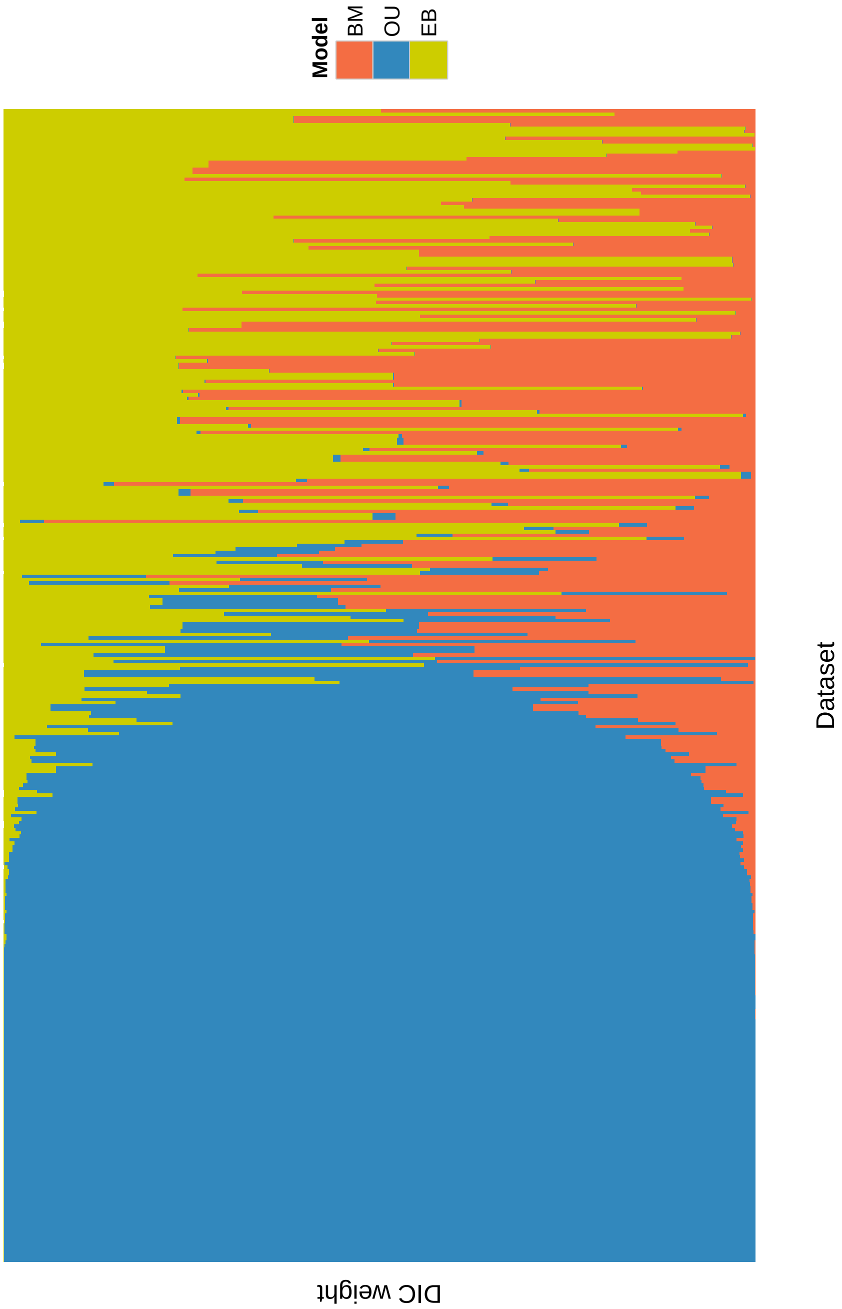

Figure B.1: The relative support, as measured by DIC weight, for the three models used in our study (BM, OU, and EB) across all 337 datasets. All models were fit with MCMC. Like the model comparisons done with AIC, an OU model is highly supported for a majority of the datasets.

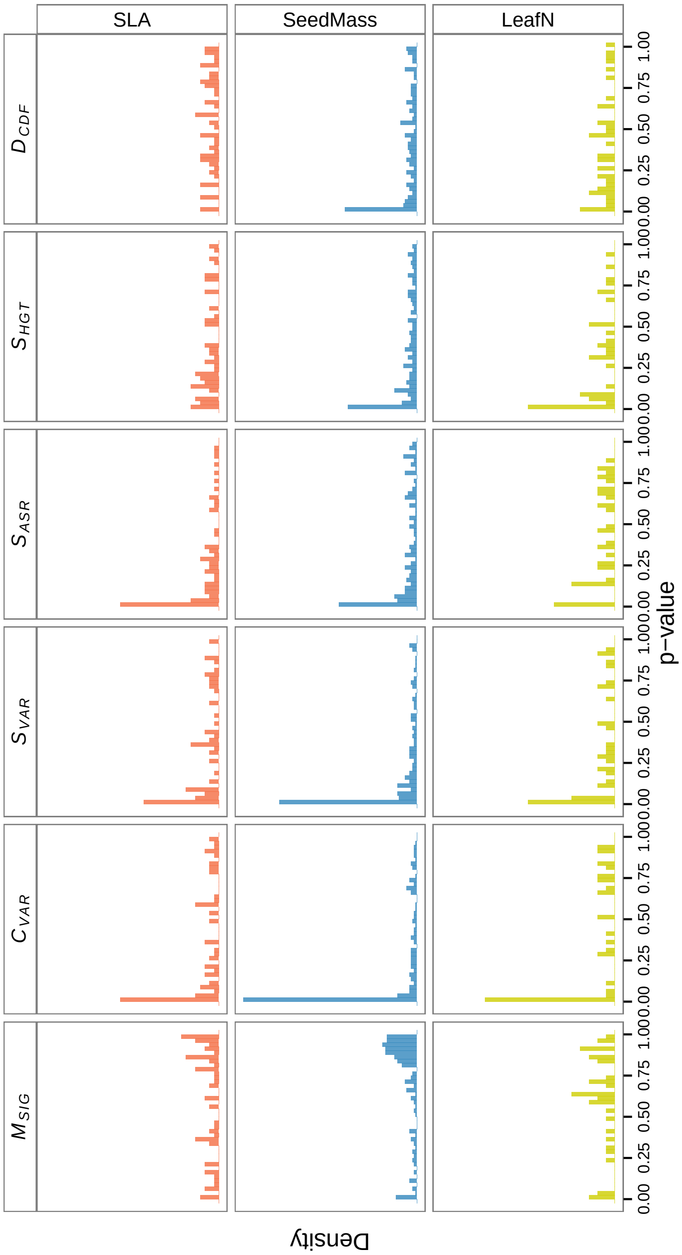

Figure B.2: The distribution of p-values for our six test statistics over all 337 datasets in our study after fitting the models using MCMC. The p-values are from applying our model adequacy approach to the best supported of the three models (as evaluated with DIC). Many of the datasets deviate from the expectations under the best model along a variety of axes of variation. Deviations are particularly common for the coefficient of variation C_{VAR} and the slope of the contrasts against their expected variances S_{VAR}.

modeled by one of the three models in our candidate set. As with the likelihood analyses, we were more likely to detect model deviations when examining larger clades (figure B.3).

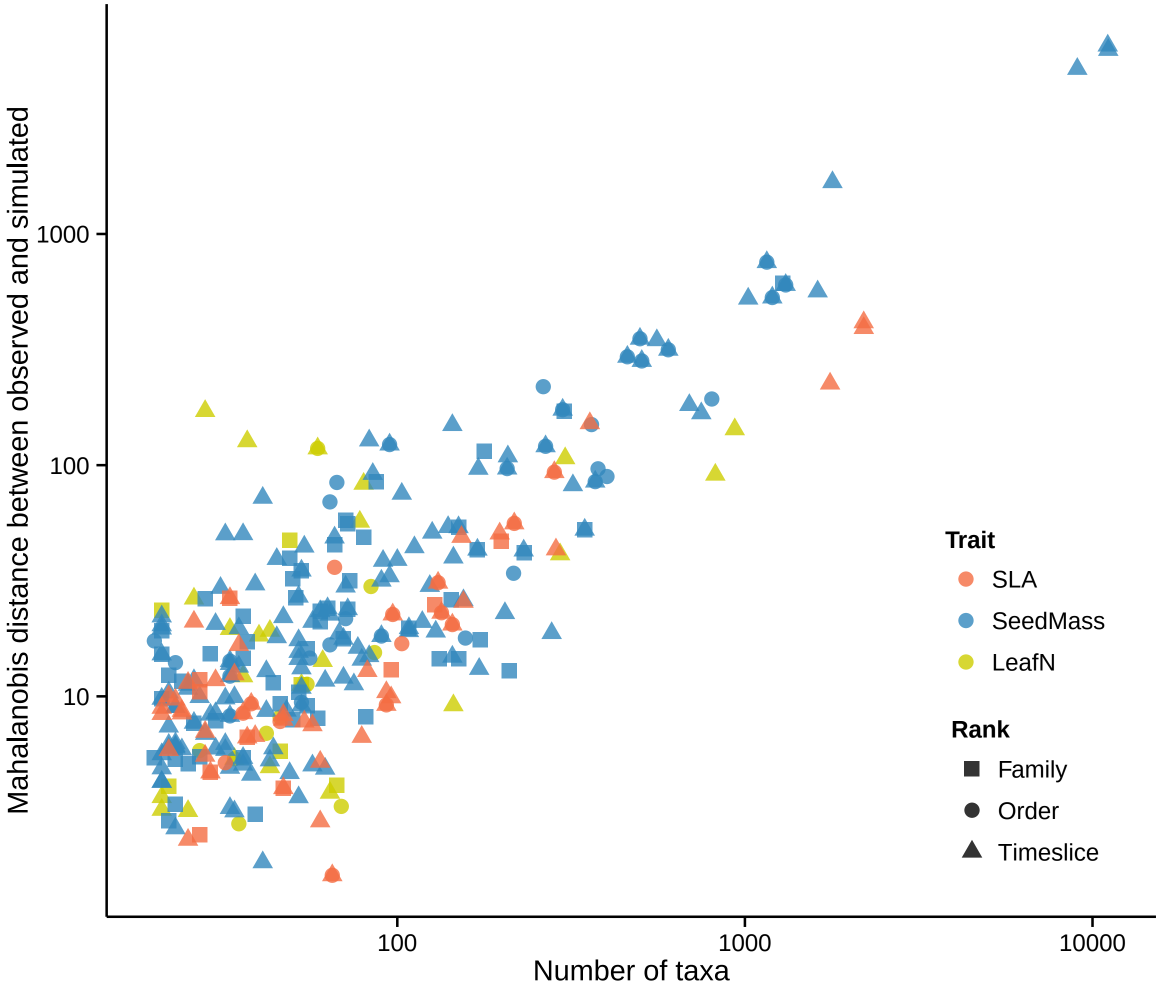

Figure B.3: The relationship between clade size and a multivariate measure of model adequacy from the Bayesian analysis. The Mahalanobis distance is a scale-invariant metric that measures the distance between the observed and simulated test statistics, taking into account the covariance between test statistics. The greater the Mahalanobis distance, the worse the model captures variation in the data. Considering only the best supported model for each clade (as chosen by DIC), there is a striking relationship between the two—the larger the dataset, the stronger the evidence that the model does not capture variation in the data.