The Dissertation Committee for Tracy Anne Heath Certifies that this is the approved version of the following dissertation:

Understanding the Importance of Taxonomic Sampling for Large-scale Phylogenetic Analyses by Simulating Evolutionary Processes under Complex Models

Committee:

David M. Hillis, Supervisor

David C. Cannatella

Robert K. Jansen

Junhyong Kim

Lauren Ancel Meyers

Understanding the Importance of Taxonomic Sampling for Large-scale Phylogenetic Analyses by Simulating Evolutionary Processes under Complex Models

by

Tracy Anne Heath, B. A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August, 2008

Dedication

To the Lonestar HPC cluster at the Texas Advanced Computing Center. Without you, I would be here for three more years.

Understanding the Importance of Taxonomic Sampling for Large-scale Phylogenetic Analyses by Simulating Evolutionary Processes under Complex Models

Publication No._____

Tracy Anne Heath, Ph.D.

The University of Texas at Austin, 2008

Supervisor: David M. Hillis

Appropriate and extensive taxon sampling is one of the most important determinants of accurate phylogenetic estimation. In addition, accuracy of inferences about evolutionary processes obtained from phylogenetic analyses is improved significantly by thorough taxon sampling efforts. Much of the previous work examining the impact of taxon sampling on phylogenetic accuracy has focused on the effects of random taxon sampling or directed taxon addition/removal. Therefore, the effect of realistic, nonrandom taxon sampling strategies on the accuracy of large-scale phylogenetic reconstruction is not well understood. Typically, broad systematic studies of diverse clades select species according to current classification to span the diversity within the group of interest. I simulated phylogenies under a realistic model of cladogenesis and used these trees to generate sequence data. Using these simulations, I

explored the effect of taxonomy-based taxon sampling on the accuracy of maximum likelihood reconstruction. The results demonstrate that taxonomy-based sampling has a stronger, negative, effect on phylogenetic accuracy than random taxon sampling. Therefore, it is recommended that systematists conducting phylogenetic analyses of diverse clades concentrate on improving sampling density within their group of interest by selecting multiple representatives from each taxonomic level.

Phylogenetic tree imbalance is often used to make inferences about macroevolutionary processes that generate patterns of tree shape. However these patterns may be obscured by non-biological factors that can bias tree shape. Using published trees inferred from biological data and trees simulated under a realistic branching model; I investigated the affect of random taxon omission on phylogenetic tree imbalance. My results indicate that incomplete taxon sampling in the presence of variable rates of speciation and extinction may be sufficient to explain much of the imbalance observed in empirical phylogenies.

Previous research has indicated that some methods of phylogenetic inference can produce biased tree topologies and shapes. Using simulated model tree topologies and sequence data, I investigated the non-biological factors that lead to biases in phylogenetic tree imbalance. Based on my results, I concluded that phylogenetic noise is the primary cause of tree shape bias. Methods that account for unobserved substitutions, such as maximum likelihood, can overcome the systematic bias toward imbalanced topologies.

Table of Contents

List of Tables

List of Figures

Chapter 1: Taxon Sampling and the Accuracy of Phylogenetic Analyses

1.1 Dense taxon sampling improves phylogenetic accuracy

Phylogeneticists have long acknowledged that data sets containing a large number of taxa create a more complex computational problem for phylogenetic analysis. As more taxa are added to a phylogenetic data set, the number of possible tree topologies increases very rapidly. In addition, the degree of homoplasy (convergent changes or reversals) increases with the number of taxa (Sanderson and Donoghue, 1989). Regardless, numerous studies on the importance of dense taxon sampling have indicated that introducing additional taxa into a phylogenetic analysis results (on average) in more accurate estimates of evolutionary relationships (Lecointre et al., 1993; Philippe and Douzery, 1994; Hillis, 1996, 1998; Graybeal, 1998; Rannala et al., 1998; Zwickl and Hillis, 2002; Pollock et al., 2002; Poe, 1998a, 1998b, 2003; DeBry, 2005; Hedtke et al., 2006). These studies represent a broad range of approaches including simulations, examinations of well-studied biological groups, and comparisons to known phylogenies. Each of these approaches has distinct advantages and disadvantages (Hillis, 1995) and together they provide a strong and consistent message about the importance of dense taxon sampling. The benefits of denser taxon sampling are especially evident in conjunction with more thorough searches of solution space (Figure 1.1). Additionally, evaluations of phylogenetic analyses often attribute problematic reconstruction and low resolution to inadequate taxon sampling (e.g. Bremer et al., 1999; Johnson, 2001; Lin et

al., 2002; Braun and Kimball, 2002; Chen et al., 2003; Freudenstein et al., 2003; Sorenson et al., 2003; Albrecht et al., 2007).

Although the importance of taxonomic sampling has been intensely investigated, many studies have focused primarily on parsimony and distance methods. Felsenstein (1978) demonstrated that, under certain circumstances, parsimony methods are inconsistent, meaning they converge on an incorrect topology as more and more characters are added for a limited sample of taxa. When two non-adjacent taxa share many homoplastic character states along long branches, parsimony methods often interpret such similarity as homology. The resulting tree depicts the two taxa as sister to one another, attributing the shared changes to a branch joining them; this effect is termed long-branch attraction (LBA). Inconsistency is not restricted to parsimony, however, as all phylogenetic reconstruction methods can exhibit this behavior if their assumptions are seriously violated or if there are not enough taxa in the analysis to accurately estimate the parameters of the evolutionary model (Felsenstein, 1978; Hendy and Penny, 1989; DeBry, 1992; Huelsenbeck and Hillis, 1993; Yang, 1994; Huelsenbeck, 1995; Lockhart et al., 1996; Gascuel et al., 2001; Huelsenbeck and Lander, 2003; Susko et al., 2004; Philippe et al., 2005). For example, maximum likelihood estimation has been shown to be inconsistent in the presence of severe branch-length heterogeneity (heterotachy, a form of non-stationarity) if the substitution process is assumed to be homogeneous across all lineages (Kolaczkowski and Thornton, 2004; Spencer, et al., 2005; Philippe et al., 2005). This example emphasizes the need for probabilistic models that incorporate complex evolutionary processes, which may improve topology estimates by reducing method

inconsistency (Yang and Roberts, 1995; Galtier and Gouy; 1998; Foster, 2004; Blanquart and Lartillot, 2006; Gowri-Shankar and Rattray, 2007; Blanquart and Lartillot, 2008; Kolaczkowski and Thornton, 2008).

Including additional taxa in a phylogenetic analysis will increase the accuracy of the inferred topology by dispersing homoplasy across the tree and reducing the effect of long-branch attraction. Hillis (1996) analyzed data simulated on a 228-taxon tree and showed that simple parsimony and distance methods accurately reconstruct the true topology when provided with sequences 5,000 nucleotides in length. At the time, this result was surprising because it seemingly contradicted the common belief that accurate phylogenetic reconstruction from very large data sets was infeasible. Moreover, Hillis et al. (1994b) had previously shown that analyses of much smaller data sets, containing only 4 taxa, required considerably longer sequences to attain the same level of accuracy. The results of Hillis's (1996) large-scale simulation indicated that for phylogenies containing many taxa, convergent substitutions or reversals (homoplasy) are distributed among the many lineages in the tree and therefore such misleading information is less likely to overwhelm the true phylogenetic signal.

Because inadequate species sampling can result in trees containing relatively long terminal branches, sparsely sampled data sets are more likely to be affected by LBA. Rannala et al. (1998) simulated ultrametric trees under a simple model of cladogenesis to investigate the impact of removing ingroup taxa on the distribution of branch lengths. They demonstrated that decreasing the proportion of sampled taxa leads to an increase in

the average length of terminal branches and generates tree shapes that may be susceptible to long-branch attraction (Figure 1.2). Huelsenbeck and Lander (2003) simulated sequences using simple substitution models on trees generated under a simple branching process and determined that the frequency that parsimony is inconsistent becomes greater as the proportion of taxa sampled decreases and substitution rates increase. Even under very simple models of evolution, unweighted parsimony underestimated the number of changes along branches and converged on an incorrect topology (Huelsenbeck and Lander, 2003).

In general, many studies have shown that adding taxa to bisect long branches can mitigate the effect of LBA (Hendy and Penny, 1989; Graybeal, 1998; Poe and Swofford, 1999; Poe, 2003). However, taxon addition should be practiced judiciously to ensure that enough taxa are added to sufficiently partition multiple long branches (Graybeal, 1998; Poe, 2003) and that the new taxa do not result in a tree model that is difficult to estimate with long terminal branches and short internal branches (Kim, 1998). Prudent taxon addition is particularly important when conducting parsimony analyses since this method is especially liable to inconsistency due to long-branch attraction. Because parametric methods, such as maximum likelihood, incorporate models that account for unobserved substitutions, these methods are less prone to the effects of long-branch attraction, as long as the models of evolution are adequate. However, enough taxa must be sampled to parameterize these models effectively (Pollock et al., 2002). In addition, longer branches require more accurate models of evolution (because more unobserved changes must be

inferred), so increased taxon sampling (which breaks up long branches) greatly benefits parametric methods as well as nonparametric methods.

Apart from its effect on topological accuracy, the density of taxon sampling also has an impact on branch-length estimation. Branch lengths provide important information about the amount of change that has occurred over the tree and are critical for applications using phylogenies to make inferences about evolution. Under the parsimony criterion, branch lengths are often underestimated in sparsely sampled regions of the tree because less information is available to infer the history of unobserved substitutions (Fitch and Bruschi, 1987; Fitch and Beintema, 1990). This artifact has been termed the node-density effect (NDE) and may mislead studies that investigate correlations between rates of molecular evolution and biodiversity (Webster et al., 2003; Venditti et al., 2006; Hugall and Lee, 2007). Maximum likelihood, Bayesian, and distance methods are also susceptible to node-density effects, particularly when the assumed model of sequence evolution is overly simple and substitution rates are high (Gojobori et al., 1982; Bruno and Halpern, 1999; Hugall and Lee, 2007). If the density of taxon sampling is increased, additional internal nodes can reveal undetected substitutions and improve estimates of branch lengths.

It has been shown that missestimation of branch lengths can, in turn, lead to biased tree topologies (Xia, 2006). Errors in estimates of genetic distance become greater as the amount of divergence between two sequences increases. Pairwise distance methods for phylogenetic reconstruction typically use log-transformed formulae to account for

unobserved substitutions (Swofford et al., 1996, Hoyle and Higgs, 2003). When using logarithmic formulae to calculate genetic distances, particularly at high levels of sequence divergence, there is a significant probability that the distance estimates will be undefined even if the "true" model of sequence evolution is assumed (Hoyle and Higgs, 2003). Therefore, when conducting distance-based analyses, it is very important to consider how taxa are sampled and avoid inclusion of highly divergent sequences.

Many advances in phylogenetic analysis over the past two decades have involved model-based approaches, such as maximum likelihood and Bayesian analyses (Swofford et al., 1996; Ronquist and Huelsenbeck, 2003; Felsenstein, 2004). In general, these parametric methods outperform nonparametric methods in both simulations and experimental studies (Hillis et al., 1994a; Huelsenbeck, 1995; Cunningham et al., 1997). However, accurate phylogenetic results from model-based studies depend, at least in part, on reasonably accurate parameter estimates for the models of evolution (Goldman, 1993; Hillis et al., 1994b; Cunningham et al., 1998; Lemmon and Moriarty, 2004; Brown and Lemmon, 2007). One of the reasons that increased taxon sampling results in more accurate phylogenetic estimation for these model-based methods is that sampling additional taxa also improves parameter estimation (Pollock et al., 1999; Sullivan et al., 1999; Pollock and Bruno, 2000; Pollock et al., 2002). In addition, as branch lengths are shortened, there are fewer unobserved changes that need to be inferred, so the accuracy of the inference becomes less dependent on the model of evolution.

In addition to their effect on phylogenetic analyses, the parameters of evolutionary models are themselves of interest to biologists. These parameters are often gene-specific, so collecting genomic-scale data from many genes across only a few taxa does little to improve our estimates of the details of evolutionary models. Instead, a thorough taxon-sampling approach is needed for each gene. Of course, the evolutionary processes may not be static across the Tree of Life for any given gene, so models that account for non-stationarity in these processes can provide better descriptions of evolutionary history (Yang and Roberts, 1995; Galtier and Gouy; 1998; Foster, 2004; Blanquart and Lartillot, 2006; Boussau and Gouy, 2006; Gowri-Shankar and Rattray, 2007; Blanquart and Lartillot, 2008; Kolaczkowski and Thornton, 2008). These models relax the assumption of time-homogeneity and can be used to detect signatures of complex evolutionary processes, such as base composition heterogeneity or heterotachy (branch-length heterogeneity), that exist in biological data (Lockhart et al. 1992; Foster et al., 1997; Mooers and Holmes, 2000; Lopez et al., 2002; Jermiin et al., 2004; Ane et al. 2005). Non-stationary, parameter-rich models can greatly increase the need for even more thorough taxon sampling. It is important to note, however, that under non-stationary models, the number of parameters can increase as more sequences are added, thus increasing the computational difficulty of phylogenetic reconstruction from large data sets. Nonetheless, this obstacle may be mitigated by the use of carefully constructed priors in a Bayesian MCMC framework (Yang, 2006) and with the development of computational methods for calculating likelihoods from non-reversible models (Boussau and Gouy, 2006).

Parameters that have been shown to be important for phylogenetic estimation include site-specific rates of evolutionary change; rates of change across first, second, and third positions of codons; rates of change relative to changes in functional groups of amino-acid residues; relative rates of the various classes of transitions and transversions between nucleotide states; branch-specific rates of evolutionary change; and taxon-specific differences in base composition (Olsen, 1987; Steel et al., 1993; Hasegawa and Hashimoto, 1993; Hillis et al., 1993; Leipe et al., 1993; Goldman and Yang, 1994; Steel, 1994; Swofford et al., 1996). The number of taxa that are needed to effectively estimate these parameters differ greatly across the parameters, but all of the estimates are improved by more thorough taxon sampling. For instance, Pollock and Bruno (2000) noted significant improvement in parameter estimation (and in turn, phylogenetic estimation) as their taxon samples increased from 4 to 8 to 16 to 24 taxa. They concluded that both phylogenetic reconstruction and estimation of unknown evolutionary processes show greater improvement through increasing taxon sampling than by increasing sequence length. In some cases, reasonable parameter estimates may be obtained from external data sources, such as the HIV database, and then applied to a more limited set of taxa in the phylogenetic analysis (Hillis, 1999). However, for most taxa, the appropriate comparative data must be obtained by the investigator for a specific group of species under study.

1.2 Dense taxon sampling improves inferences of evolutionary processes from phylogenetic trees

Beyond simply broadening our understanding of species relationships, phylogenetic trees are essential tools used in many areas of biology. Phylogenies are often used to explain broad evolutionary patterns and processes such as the evolution of adaptive traits, ancestral character states, the timing of species divergences, and variation in evolutionary rates. Many of the applications developed for these types of analyses require robust and accurate estimates of phylogeny (topology, branch lengths, and root position). This is an important consideration in and of itself; however, post-tree reconstruction applications are sensitive to reduced levels of data sampling, even when provided with an accurate phylogenetic tree.

Comparative methods.—Comparative analyses are a fundamental component in the fields of evolutionary biology, behavior, and ecology. The development of statistical methods that incorporate phylogenetic trees (Felsenstein, 1985) have allowed for robust and reliable tests of the evolution of adaptive traits and the processes that might drive diversification. For example, these methods have been used to reveal patterns in the biodiversity of marine teleost fishes (Alfaro et al., 2007) and to show that independent origins of dietary specialization have been a major factor in the evolution of defensive mechanisms in neotropical poison frogs (Darst et al., 2005). Comparative analyses of character evolution using phylogenetic comparative methods require attention to adequate sampling at many levels. At the intraspecific level, poor sampling of organismal attributes can lead to measurement error, which may result in an underestimation of the variance of contrasts between sister taxa (Ricklefs and Starck, 1996). Generation of a robust phylogeny is extremely important since different comparative methods have

different ways of dealing with topological uncertainty (Purvis et al., 1994). In addition, fewer taxa (and thus fewer internal nodes for calculating contrasts) can lead to increased variance and uncertainty in the results. Ackerly (2000) used simulated data to show that the statistical power of several comparative tests decreased as the sample size of taxa decreased, and that careful attention should be paid to how species are sampled for these analyses. Biased taxon sampling, particularly with respect to the characters of interest, can lead to systematic biases in the calculation of statistical correlations between characters. The results presented by Ackerly (2000) indicate that uniform, random sampling of taxa does not introduce error in phylogenetic comparative methods.

Ancestral character states.—An integral component of phylogenetic comparative analyses and other evolutionary applications is the reconstruction of ancestral character states. These methods use phylogenetic trees and branch lengths to infer the states of discrete or continuous characters at ancestral nodes, and have been used to reconstruct such diverse ancestral characters as the advertisement calls of frogs in the genus *Physalaemus* (Ryan and Rand, 1995, 1998), the fruiting-body forms of homobasidiomycetes (Hibbett, 2004), and ancient bacterial protein sequences (Gaucher et al., 2003). Dense taxon sampling is also an important consideration for ancestral-state reconstruction methods. Salisbury and Kim's (2001) analyses of simulated data and trees indicated that the accuracy of parsimony ancestral-state estimation decreases with reduced taxon sampling and increased rates of character evolution (Figure 1.3). Because parsimony methods do not account for unobserved changes, they usually underestimate the number of changes along a branch (Fitch and Bruschi, 1987; Fitch and Beintema,

1990; Huelsenbeck and Lander, 2003). Dense taxon sampling can reduce this effect and improve the accuracy of parsimony ancestral-state estimates. Maximum likelihood and Bayesian methods for reconstructing ancestral states have also been developed (Pagel, 1994; Schluter et al., 1997; Pagel, 1999; Huelsenbeck and Bollback, 2001; Pagel et al., 2004). These parametric ancestral-state reconstruction methods are also sensitive to high rates of character evolution. However, Schluter et al. (1997) showed that parsimony ancestral-state reconstruction methods often fail to identify ambiguous-node state estimates. Conversely, maximum likelihood and Bayesian methods are less likely to provide misleading results because these methods incorporate branch-length information and explicit models of character evolution and quantify uncertainty in ancestral-state estimates (provided that the model assumptions are adequate). Bayesian approaches, in particular, use Markov chain Monte Carlo sampling to accommodate and quantify uncertainty in the tree topology, branch lengths, ancestral states, and model parameters (Huelsenbeck and Bollback, 2001; Pagel et al., 2004). Denser taxon sampling reduces the number of unobserved evolutionary events, and so is also expected to simplify and improve the reconstruction of ancestral states in model-based analyses.

Divergence time estimation.—A primary field of research in evolutionary biology involves estimation of the timing and rate of evolutionary processes. In these applications, phylogenetic trees are used to date speciation events and infer lineage-specific substitution rates. Reliable estimates of species divergence times are fundamental components for understanding historical biogeography, testing hypotheses of adaptive character evolution, and estimating speciation and extinction rates. However, divergence

time estimation is hindered by the fact that the rate of evolution and time are intrinsically linked when inferring genetic distances between lineages. Several methods test for variation in the rates of molecular evolution or tease apart the rate of substitution and time by applying models for estimating lineage-specific substitution rates. These methods include strict molecular clock models (Zuckerkandl and Pauling, 1962; Langley and Fitch, 1974), local molecular clocks (Kishino and Hasegawa, 1990; Rambaut and Bromham, 1998; Yoder and Yang, 2000; Yang and Yoder, 2003), non-parametric and semi-parametric methods for estimating autocorrelated substitution rates (Sanderson, 1997, 2002), and Bayesian methods for estimating autocorrelated and uncorrelated rates (Thorne et al, 1998; Huelsenbeck et al., 2000; Kishino et al., 2001; Thorne and Kishino, 2002; Drummond et al., 2006; Lepage et al., 2006). These various approaches have been applied to a number of biological data sets (e.g., Yang and Yoder, 2003; Smith et al., 2006; Bell, 2007; Hugall et al., 2007; Roelants et al., 2007; Zhou and Holmes, 2007). Current implementations of most of these methods require a fixed tree topology and sometimes fixed branch lengths (Thorne and Kishino, 2002; Sanderson, 2003; Lepage et al., 2007; for exceptions see Drummond et al., 2006). Because of their reliance on phylogenetic data, these methods can be sensitive to taxon sampling density. Robinson et al. (1998) evaluated the effect of reduced taxon sampling on the performance of the relative-rates test. The relative-rates test (Sarich and Wilson, 1973; Wu and Li, 1985) is used to compare the substitution rates between two species and has been extended for analyzing larger phylogenetic trees to detect rate variation (Li and Bousquet, 1992; Takezaki et al., 1995). The simulation study of Robinson et al. (1998) showed that increased proportions of taxon sampling improved the accuracy of the relative-rates test.

Most of the work exploring the accuracy of molecular dating methods has revealed that these methods are very sensitive to the fossil calibrations used and little is known about the impact of taxon sampling on divergence time estimates (Yang and Rannala, 2006; Rutschmann et al., 2007; Hugall et al., 2007). A recent study by Hug and Roger (2007) used two biological data sets with low levels of taxon sampling (30 metazoan taxa with two outgroup species, and a 36 taxon data set that spanned all eukaryotes) and concluded that, for these data sets, reduced taxon sampling was not an important factor in the estimation of node times. However, their analyses showed that the choice and application of fossil calibration points resulted in a significant impact on the estimates of node ages. From their results, Hug and Roger (2007) recommended that biologists should focus on improving the number and quality of their fossil calibrations and not on increasing taxon sampling, provided there are enough taxa to obtain a reliable estimate of phylogeny. However, because of the sparsely sampled data sets used in this study and the demonstrated extreme sensitivity of these data to fossil constraints, Hug and Roger's (2007) results may not apply to a more general set of conditions and the importance of dense taxon sampling for estimating species divergence times is still an open question.

Node-density effects, as a result of uneven taxon sampling, may adversely affect molecular dating analyses (Hugall and Lee, 2007). Based on the studies demonstrating the sensitivity of divergence time estimation methods to fossil calibration choice (Near and Sanderson, 2004; Near et al., 2005; Roger and Hug, 2006; Yang and Rannala, 2006;

Ho, 2007; Hugall et al., 2007; Rutschmann et al., 2007), together with studies emphasizing the importance of increased taxon sampling on phylogenetic reconstruction methods and the estimation of evolutionary parameters (Lecointre et al., 1993; Hillis, 1996, 1998; Graybeal, 1998; Rannala et al., 1998; Pollock and Bruno, 2000; Zwickl and Hillis, 2002; Pollock et al., 2002; Poe, 2003; DeBry, 2005; Hedtke et al., 2006), it is recommended that biologists focus on increased collection of fossils and improved taxon sampling density for these types of analyses, whenever possible. Maximizing the number of fossil calibration points goes hand-in-hand with increasing taxon sampling because densely sampled trees provide a greater number of internal nodes on which an investigator can place a fossil calibration. Moreover, investigators are far more restricted by the availability of fossils and other types of information for calibrating divergences than by the availability of extant taxa. Further investigation using simulations and well-sampled data sets of living and fossil taxa should help shed light on this issue. Because extensive taxon sampling (especially of fossil taxa) is sometimes impractical, Bayesian methods for divergence time estimation present promising opportunities to account for uncertainty in phylogenies by simultaneously estimating the tree topology and branching times (Drummond et al., 2006). These methods can also incorporate information on taxon sampling density in the form of priors on the distribution of divergence times (Yang and Rannala, 1997, 2006).

Evaluating diversification rates.—Phylogenetic trees are fundamental for understanding variation in species diversity. Methods for elucidating patterns of speciation and extinction measure the shape of phylogenies to detect shifts in diversification rates or to

estimate global net diversification rates. Phylogenetic tree shape can be measured by quantifying how node ages are distributed over time or by calculating the degree of asymmetry among lineages in the tree. Measures of tree shape can be compared to a null model that assumes all lineages have experienced the same rate of diversification (Shao and Sokal, 1990; Kirkpatrick and Slatkin, 1993; Nee et al., 1994b; Pybus and Harvey, 2000; Agapow and Purvis, 2002). Analyses of the temporal distribution of diversification events use branch lengths obtained from time-adjusted phylogenies to estimate and detect large shifts in speciation and extinction rates (Nee et al., 1994b; Pybus and Harvey, 2000). For example, Becerra (2005) applied these methods to investigate temporal and biogeographic processes that may have shaped the diversity of the plant genus *Bursera*. The results of this study indicate that the radiation of this group is associated with the establishment of tropical dry forest habitat in Mexico. However, inadequate taxon sampling has a significant impact on these methods. Nee et al. (1994a) used lineages-by-time plots to show that incomplete taxon sampling can result in an apparent reduction in the rate of diversification over time, even when the tree evolved under constant rates of speciation and extinction.

Analyses based on topology measure asymmetry in the distribution of lineages over a tree to test for changes in diversification rates. These methods evaluate the balance either at a single node or over the entire tree (Shao and Sokal, 1990; Kirkpatrick and Slatkin, 1993; Agapow and Purvis, 2002) and are often used to detect patterns characteristic of rapid radiations in phylogenetic trees (Guyer and Slowinski, 1993; Chan and Moore, 1999, 2002). The degree of taxon sampling is an important consideration

when conducting these analyses. Several studies have shown that published phylogenies are (on average) much more imbalanced than expected under a model assuming constant diversification rates (Guyer and Slowinski, 1991; Heard, 1992; Mooers, 1995; Purvis and Agapow, 2002; Holman, 2005; Blum and François, 2006; and see chapter 3). Mooers (1995) compared the level of tree imbalance in a collection of published phylogenies and found that incomplete trees are more imbalanced than completely sampled phylogenies. The bias caused by incomplete species sampling must be considered when using phylogenies to test hypotheses about species diversity.

Notwithstanding the results of numerous studies demonstrating the importance of dense taxon sampling, some researchers have argued that increasing the number of taxa does not have a large impact on the accuracy of phylogenetic analyses. For example, some contend that large character data sets are sufficient to get an accurate phylogeny. Rosenberg and Kumar (2001) conducted a simulation study indicating that adding taxa to a problematic phylogeny is less effective than adding additional characters. This paper led to a debate in the literature and a reanalysis of the Rosenberg and Kumar (2001) data (Zwickl and Hillis, 2002; Pollock et al., 2002, Rosenberg and Kumar, 2003; Hillis et al., 2003). Pollock et al. (2002) reanalyzed the Rosenberg and Kumar (2001) data using a different approach to summarizing results (measurement of error), and Zwickl and Hillis (2002) re-conducted the Rosenberg and Kumar (2001) study with a different approach to study design that examined a fuller spectrum of taxon sampling strategies. Both studies concluded that taxon sampling has a very strong and positive effect on the accuracy of phylogenetic reconstruction. However, because of the increasing availability and

accumulation of genomic data and the difficulty of obtaining sequence data for many taxa, the debate about the relative importance of taxon sampling versus character sampling continues in the literature (Hillis et al., 2003; Rosenberg and Kumar, 2003; Rokas et al., 2003; Cummings and Meyer, 2005; Rokas and Carroll, 2005; Hedtke et al., 2006; Gatesy et al., 2007) and the importance of dense taxon sampling for large-scale tree inference using parametric methods requires further investigation.

ACKNOWLEDGEMENTS

My co-authors, Shannon M. Hedtke and David M. Hillis, and I would like to thank professors Deyuan Hong, Zhiduan Chen, Chengxin Fu, Michael J. Donoghue, and Yin-Long Qiu for hosting the symposium on "Evolutionary Biology in the 21st Century—Tracing Patterns of Evolution through the Tree of Life," and for the invitation to D.M.H. to present this paper. The symposium was supported by a grant from National Natural Science Foundation of China. A.J. Abrams, J. Brown, M. Morgan, G. Pauly, R. Springman, M. Swenson, R. Symula, D. Zwickl, Hervé Philippe, and an anonymous reviewer provided helpful comments on this manuscript. D.M.H. gratefully acknowledges grant support from the United States National Science Foundation (NSF). T.A.H. was funded by a graduate research traineeship provided by an NSF IGERT grant in Computational Phylogenetics and Applications to Biology awarded to the University of Texas, Austin. S.M.H. was supported by an NSF pre-doctoral fellowship.

FIGURE 1.1. Error in phylogenetic reconstruction typically decreases with increased taxon sampling of a given taxonomic group. The benefits of increased taxon sampling are particularly evident when searches of the solution space are more thorough. In this graph (adapted from Zwickl and Hillis, 2002: fig. 6), phylogenetic error decreases with increased taxon sampling across all analyses. However, the benefits of adding additional taxa are smaller if only the stepwise-addition algorithm (*SA*) is used to find an approximate solution, compared to the more thorough searches provided by stepwise-addition plus nearest-neighbor-interchanges branch-swapping (*SA + NNI*) or tree-bisection-reconnection branch-swapping (*SA + TBR*). Analyses of larger data sets generally require more thorough search algorithms (and thus more computational effort), but result in greatly decreased phylogenetic error.

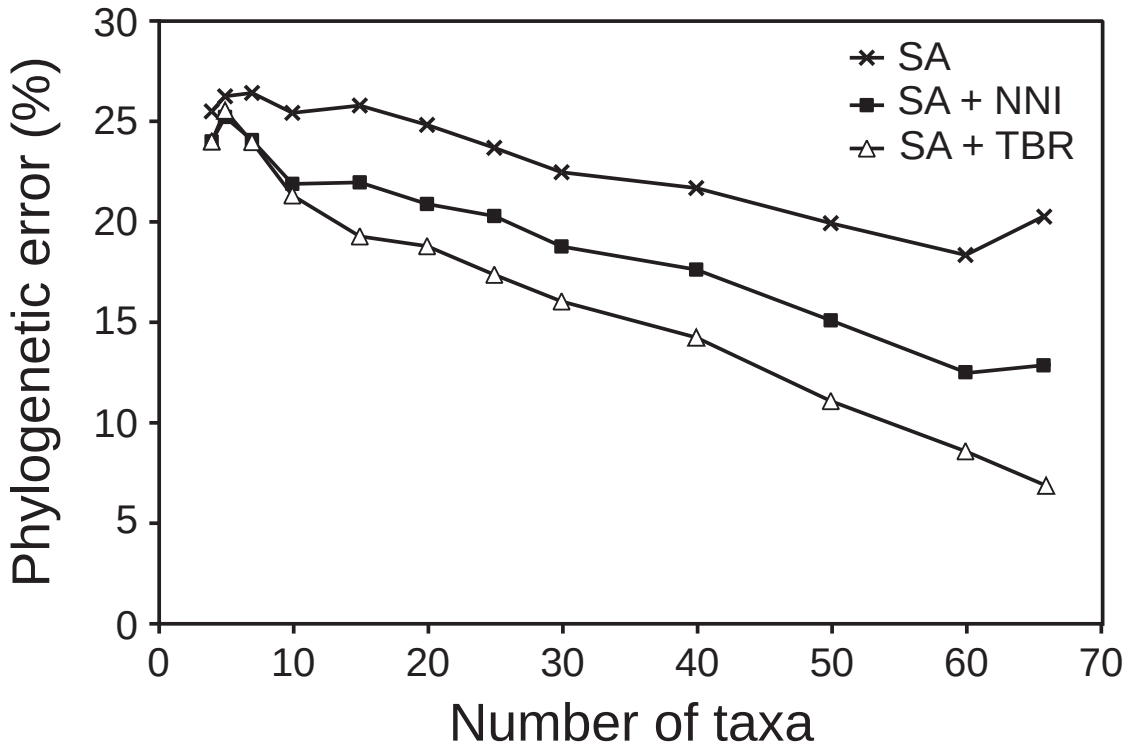


FIGURE 1.2. Two simulations of a birth-death process to model cladogenesis. The speciation rate (λ) and extinction rate (μ) were fixed throughout the simulation and arbitrarily set to $\lambda / \mu = 2$. (**A**) Phylogenetic tree with complete (100%) taxon sampling (20 taxa total). (**B**) Phylogenetic tree with 10% taxon sampling (20 taxa sampled from 200 taxa total). Adapted from Rannala et al. (1998: figs. 1 and 2).

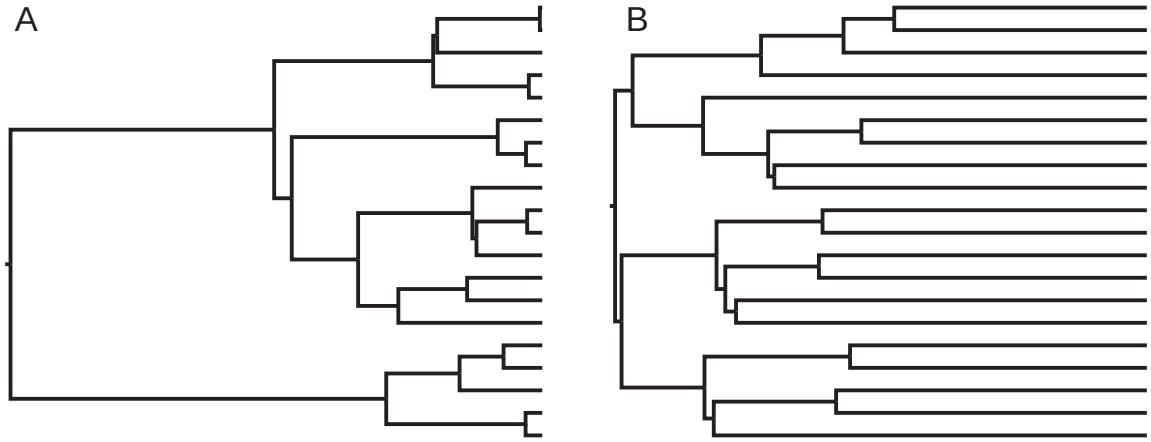
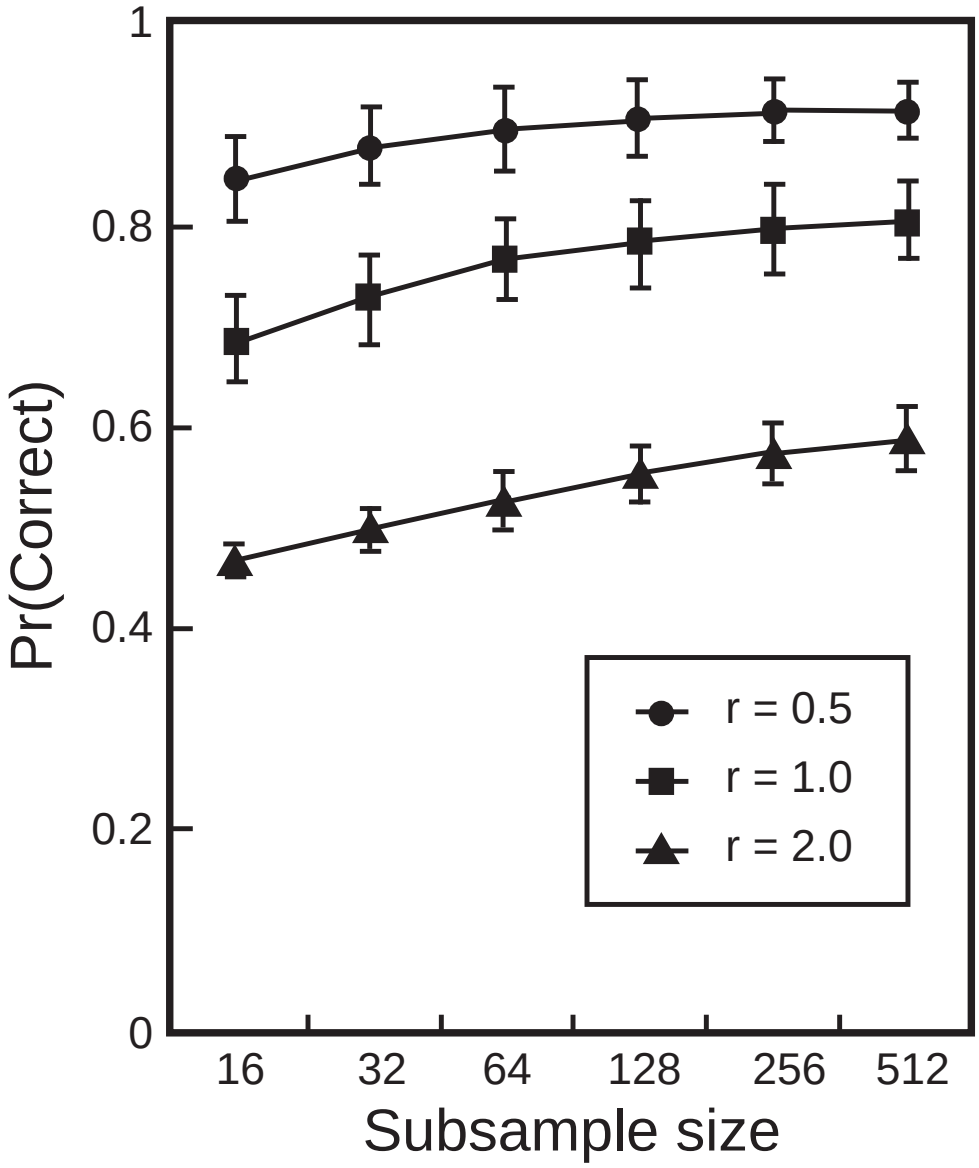


Figure 1.3: The mean probabilities, *Pr(Correct)*, of correctly estimating the root state of a binary character evolving at 3 different rates (*r*) on subsamples of 512-taxon, pure-birth model tree topologies. Each point is the mean for a sample of 100 trees and the error bars represent the ± 1 standard deviation. Adapted from Salisbury and Kim (2001: fig. 1).



Chapter 2: Phylogenetic Reconstruction of Diverse Clades and the Importance of Dense Taxonomic Sampling

2.1 INTRODUCTION

Recent advances in computational resources and innovations in phylogenetic algorithms have made it feasible to reconstruct robust phylogenies from large-scale, multi-locus data sets. Furthermore, because of the rapid accumulation of genomic data, and funding initiatives for elucidating the phylogenies of large taxonomic groups – and ultimately the Tree of Life – it is critical to understand how phylogenetic methods perform on large data sets and how phylogenetic inference may be affected by non-biological factors.

Appropriate and thorough taxon sampling is an important consideration for any broad investigation of phylogenetic relationships. Numerous studies have investigated the impact of reduced taxon sampling on the accuracy of phylogenetic reconstruction. The results of these studies, which used simulated data, biological data, and data from known phylogenies, have demonstrated that, on average, increased taxon sampling from within the clade of interest can improve the accuracy of estimates of tree topology (Lecointre et al., 1993; Philippe and Douzery, 1994; Hillis, 1996, 1998; Graybeal, 1998; Rannala et al., 1998; Zwickl and Hillis, 2002; Pollock et al., 2002; Poe, 1998a, 1998b, 2003; DeBry, 2005; Hedtke et al., 2006), branch lengths (Gojobori et al., 1982; Fitch and Bruschi,

1987; Fitch and Beintema, 1990; Bruno and Halpern, 1999; Hugall and Lee, 2007), and model parameter values (Pollock et al., 1999; Sullivan et al., 1999; Pollock and Bruno, 2000; Pollock et al., 2002); as well as inferences of evolutionary processes based on phylogenetic relationships (Nee et al., 1994a; Mooers, 1995; Robinson et al., 1998; Ackerly, 2000; Salisbury and Kim, 2001; and see chapter 3).

The effect of realistic, non-random taxon sampling strategies on the accuracy of large-scale phylogenetic reconstruction is still unclear, however. Previous simulation studies have investigated the importance of taxon sampling density by assessing random taxon sampling (Kim, 1998; Rannala et al., 1998; Zwickl and Hillis, 2002) or directed taxon addition/removal (Graybeal, 1998; Poe and Swofford, 1999; Poe, 2003). These studies have also primarily evaluated smaller data sets compared to recent large-scale analyses of diverse taxonomic groups. Consequently, the results of these studies do not necessarily extrapolate to broad systematic studies of diverse clades. It is unlikely that analyses of large taxonomic groups employ random sampling as a strategy for assembling a phylogenetic data set (Hillis, 1998). Many factors play a role in determining if a species is sampled, such as difficulty in amplifying genetic material, inaccessibility due to geographic distribution, or limited resource availability. Typically, species are sampled according to their taxonomic ranking so as to cover much of the perceived diversity within the group of interest (Hillis, 1998). For example, Freitas and Brown (2004) conducted a phylogenetic analysis of the higher-level relationships within the butterfly family Nymphalidae. They included representatives from all 13 subfamilies by selecting 95 species based on taxonomy, which represented 94 of the 542 different genera

identified in this species-rich group. This non-random sampling is likely to have a different effect on the shape of the tree (in terms of the distribution of branch lengths as well as topological asymmetry) than sampling uniformly from all species in the group. However, there are no published studies exploring this type of non-random sampling and its effect on phylogenetic accuracy. Additionally, very few studies have evaluated the effect of reduced taxon sampling on the performance of model-based methods, such as maximum likelihood or Bayesian inference. In this study, I use simulated phylogenies (generated under a realistic model of cladogenesis) and sequence data to explore the relative effects of taxonomy-based and random species sampling on the accuracy of large-scale phylogenetic reconstruction under the maximum likelihood optimality criterion.

2.2 METHODS

2.2.1 *Tree Simulation*

Model tree topologies and branching times were simulated under a simple model of exponential waiting time for speciation/extinction events with variable lineage-specific speciation and extinction rates. Each tree started with a single root lineage and initial values for speciation and extinction rates. The time to the next event (lineage splitting or extinction) was drawn from an exponential distribution based on the sum of the rates for all extant lineages. The type and location of each event was chosen in proportion to the speciation and extinction rates for each of the extant lineages. When the next event

resulted in extinction, the lineage was removed and a new waiting time was drawn. At a speciation event, the parent lineage bifurcated into two daughter lineages. The speciation/extinction rates of each daughter lineage were obtained by multiplying the parent rate by a random number (m). The value of m was drawn from a gamma distribution with a shape parameter (α) and scale parameter (β), where $\beta = \alpha$ so that $E(m) = 1$ and the rates were autocorrelated. A gamma-distributed prior on speciation and extinction rates was enforced to discourage the rates from going to infinity or zero. Therefore when the rate of a new daughter lineage was drawn, that rate was accepted in proportion to the gamma-distributed prior. The prior distributions on the rates were also assigned shape and scale parameters. These parameters were responsible for regulating much of the rate variation. I show by simulation that increasing the shape parameters results in a decrease in the diversification-rate variation and produces more balanced topologies (Figure 2.1). This model is a biologically motivated method for generating variable and autocorrelated speciation/extinction rates. Trees generated under this model should produce more biologically realistic tree topologies than models assuming constant rates of diversification (the equal rates Markov model; ERM model), since it is an empirical observation that speciation and extinction rates do vary across groups, and these rates are correlated among related species (Dial and Marzluff, 1989; Guyer and Slowinski, 1991; Heard, 1992; Sanderson and Donoghue, 1994; Savolainen et al., 2002; Holman, 2005). This model for generating variable speciation/extinction rates is analogous to probabilistic models of the rate of molecular evolution implemented in methods used to estimate divergence times (e.g. Thorne et al., 1998; Huelsenbeck et al.,

2000; Kishino et al., 2001). This method for simulating tree topologies was implemented by modifying code from the program *Phyl-o-gen* (Rambaut, 2002).

Three sets of model tree topologies were generated under the variable-rate model of cladogenesis with the gamma-shape parameter set to 3.0. Figure 2.2 shows an example from each model tree set: (A) 50, 500-taxon trees with a tree depth of 1.0 substitutions/site, (B) 50, 1,000-taxon trees with a tree depth of 1.0 substitutions/site, and (C) 50, 500-taxon trees with a tree depth of 0.5 substitutions/site.

2.2.2 *Data Simulation*

The trees simulated with 500 taxa and scaled to 1.0 substitutions/site (tree set A: Figure 2.2A) were used to generate sequence data sets under a number of substitution models (Table 2.1). For each tree in this collection, three data sets were simulated under the HKY model varying in the total number of characters: 500, 1000, and 2000 nucleotides. Additionally, sequence data were generated on this set of trees under the JC model and the GTR+I model, each 1000 nucleotides in length. The remaining sets of model trees (tree sets B and C; Figure 2.2B and 2.2C) were used as model topologies to generate sequence data sets under the HKY model (each 1000 nucleotides in length).

2.2.3 *Taxon Sampling*

In this study, the effect of clade-based sampling (also referred to as taxonomy-based sampling) on phylogenetic accuracy was compared to the accuracy of estimates from randomly sampled data sets. Taxonomic classification was simulated by identifying the sub-trees in each of the simulated phylogenies. Figure 2.3 illustrates how taxa were sampled according to their taxonomic rank. If the target data set size was 10 taxa, then sub-clades were defined by first identifying the 10 basal nodes in the complete phylogeny that formed distinct sub-trees. The descendants of each node made up a single taxonomic group. Each sub-clade contained a fraction of the total taxa and no terminal taxon was excluded from classification. As a result, some taxonomic groups were very diverse, whereas others only contained one or two species. A single representative of each sub-clade was selected randomly from each group. Therefore, this sampling strategy mimics a situation where the species are classified according to their phylogenetic position and a single exemplar species is chosen to represent each taxonomic group. Because the taxonomy of a particular biological group is rarely in perfect accord with the true species phylogeny, this sampling scheme is an ideal case. However, this strategy more closely mimics the sampling procedure practiced in many systematic studies, which attempt to cover the taxonomic diversity of a group of interest. Clade-based taxon sampling is similar in theory to phylogenetic diversity measures for prioritizing conservation efforts. Measures of phylogenetic diversity (PD) quantify the overall biodiversity of a sub-set of taxa based on a larger phylogenetic tree and can be used to determine the set of taxa that maximizes PD for biodiversity management (Faith, 1992; Barker, 2002; Mooeres et al., 2005; Crozier et al., 2005).

The randomly sampled data sets were assembled by selecting taxa from a uniform distribution spanning all of the species in the complete tree until the target data set size was reached. When a data set is comprised of taxa selected based on taxonomy, the total tree depth is always the same as that of the full tree. For example if the complete phylogeny had a root-to-tip length of 1.0 substitutions/site, then a clade-based sample of that tree would also have a tree depth of 1.0. However, when taxa are randomly selected, the sub-sample does not necessarily span the total diversity of the complete tree and may have a shorter tree depth than a clade-based sample. If the two sampling strategies are to be compared, it is important to ensure that the sub-sampled data sets are within the same scope. Therefore, when taxa were sampled at random, reduced data sets that did not span the entire depth of the full phylogeny were discarded and the tree was resampled. A range of target data set sizes were sampled from each complete, simulated data set ranging from 2 – 100% sampling density. For every target sample size, 100 iterations of each sampling strategy were performed on each simulated tree.

2.2.4 *Phylogenetic Reconstruction*

The sub-sampled and complete data sets were analyzed under the maximum likelihood criterion using GARLI version 0.951 (Zwickl, 2006). The serial GARLI algorithm uses a type of evolutionary algorithm to heuristically explore the possible set of tree topologies, branch lengths, and model parameters and find the solution that maximizes the likelihood. The phylogenetic reconstruction algorithm implemented in GARLI allows for rapid estimation of phylogenetic relationships from very large data

sets. All analyses were run on the Lonestar Dell Dual-Core Linux Cluster (5,200 compute-node processors) at the Texas Advanced Computing Center (TACC: http://www.tacc.utexas.edu/). Each data set was analyzed under the true simulation model or an intentionally misspecified model (see Table 2.1) and with all other program specific settings set to their default values. This procedure was followed to produce base-line results. It is important to note, however, that when conducting analyses using GARLI, multiple replicate runs should be conducted for each data set, and alternate program settings should be explored.

2.2.5 *Measuring Phylogenetic Error*

Each estimated tree topology was compared to the true topology used to generate the sequences after pruning unsampled taxa. The accuracy of a phylogenetic estimate was quantified using the absolute error. This measure of error uses the Robinson-Foulds (RF) distance (also called the symmetric difference; Robinson and Foulds, 1981) which is the total number of branches in the estimated topology that must be collapsed and/or expanded to reproduce the topology of the true tree. For trees with n taxa, the minimum RF distance is 0 (identical topologies) and the maximum RF distance is equal to twice the number of internal edges in the true tree (for an unrooted tree there are $n - 3$ internal edges):

$$RF_{MAX} = 2(n - 3)$$

An RF value equal to the maximum distance indicates that no correct bipartitions were reconstructed in the estimated topology. Because RF distances are dependent on the

number of taxa in the tree, they cannot be used to compare the accuracy of trees estimated from data sets of varying sizes. Therefore the error must be normalized by calculating the proportion of error (or absolute error; Zwickl and Hillis, 2002). The proportion error (*E*) is calculated by dividing the observed RF distance by the maximum RF distance. The value of this measure ranges from 0 for identical topologies, to 1 for topologies that do not share any bipartitions.

An alternative measurement of error (adjusted error) has been used in previous studies (Zwickl and Hillis, 2002; Pollack et al., 2002) and avoids a major disadvantage of the absolute error. When comparing the accuracy of reconstructions from very small data sets (fewer than 10 taxa) with the accuracy of trees estimated from larger data sets, the absolute error does not provide a consistent measure. This is because the expectation of the absolute error (*E*) asymptotically approaches *1* as the number of taxa increases (Zwickl and Hillis, 2002). To correct this problem, the adjusted error uses the expected RF distance rather than the maximum RF distance to normalize the observed value (Zwickl and Hillis, 2002). In their investigation of the different measures of error, Zwickl and Hillis (2002) found that for sufficiently large data sets (10 taxa or greater), the absolute error and adjusted error are equivalent. Because the sample sizes considered in the present study range from 10 taxa to 1000 taxa, I chose the absolute error to determine accuracy.

The absolute error was calculated for each replicate data set. Then the proportion of error was averaged across every estimated topology of a given sampling density and taxon-sampling strategy.

2.3 Results and Discussion

Increasing the density of sampled species greatly improves the accuracy of tree topology estimates (Figure 2.4): as the proportion of species sampling increases, the proportion of error decreases. This holds for data sets with both 500 and 1000 total taxa. In general, these results are consistent with previous studies using trees simulated under birth-death branching processes; demonstrating that the proportion of taxa sampled from a monophyletic group – rather than the total number of taxa – strongly influences the accuracy of the estimated phylogeny (Rannala et al., 1998).

Clade-based sub-sampling has a stronger, negative, effect on phylogenetic accuracy than random taxon sampling for all of the simulation conditions considered in this study (Figures 2.4, 2.7, 2.8, 2.9, 2.10). For data sets sampled based on taxonomy, the estimated topologies are significantly less accurate than analyses of random data sets (Figure 2.4). Reduced taxon sampling generally results in trees that are more star-like, with longer terminal branch lengths relative to internal branch lengths (Kim, 1998; Rannala et al., 1998). Kim (1998) showed that such trees are more difficult to reconstruct under the parsimony criterion. My results demonstrate that the type of sampling strategy has a strong effect on the distribution of branch lengths in the tree. When a clade-based

sampling strategy is used to compile a phylogenetic data set, the average terminal branch length of the true (sampled) tree is greater than when the taxa are sampled at random (Figure 2.5). This is illustrated by the example in figure 2.6. Uniform, random sampling of terminal taxa (Figure 2.6B) produces a tree that contains a greater number of recent bifurcations than when taxa are sampled based on their taxonomic rank (Figure 2.6C). Clade-based sampling introduces many long terminal branches, which can produce a data set with more homoplasy, which can mislead phylogenetic reconstruction methods. This increase in phylogenetic noise can lead to missestimation of model parameter values and branch lengths which can, in turn, lead to decreased topological accuracy. As the density of sampling within the group of interest increases, the distribution of branch lengths becomes less skewed and accuracy is increased (Kim, 1998). The high error rates resulting from the taxonomy-based sampling strategy are independent of the depth of the initial, complete phylogeny (Figure 2.7). When the simulated topologies have a total depth of 0.5 substitutions/site, the general pattern of improved accuracy with increasing sampling density remains for these trees.

Because of the increasing availability of genomic data for a limited number of organisms, some researchers continue to reconstruct phylogenies with very few taxa (e.g. Cannarozzi et al., 2007), despite the conclusions reached by many studies indicating the importance of dense species sampling for accurate phylogenetic analysis. Moreover, biologists are often less restricted by their ability to collect more genes for a given set of species than by their ability to sample additional taxa. This conflict has led to continued debate focused on the relative importance of increased taxonomic sampling versus

increased character sampling (Rosenberg and Kumar, 2001; Zwickl and Hillis, 2002; Pollack et al., 2002; Hillis et al., 2003; Rosenberg and Kumar, 2003; Rokas et al., 2003; Cummings and Meyer, 2005; Rokas and Carroll, 2005; Hedtke et al., 2006; Gatesy et al., 2007). In this study, I show that both increased taxon sampling and increased character sampling have a strong effect on phylogenetic accuracy (Figure 2.8). In contrast, some studies have found that increased taxon sampling (compared to genomic sampling) has little impact on the accuracy of phylogenetic reconstruction (Rosenberg and Kumar, 2001; Rokas et al., 2003; Rokas and Carroll, 2005). My results emphasize that increasing the amount of data in a phylogenetic analysis, both in terms of the number of characters and the density of taxa sampled, is important for accurate topological reconstruction. However, unlike the simple simulations in this study, it is not always safe to assume that, for biological data, additional genes or characters were generated under the same model of evolution. Increasing the number of characters in phylogenetic data sets can also introduce greater heterogeneity. Thus, with multi-locus data, it may be necessary to apply complex, parameter-rich models of sequence evolution (Bull et al., 1993; Nylander et al., 2004; Brandley et al., 2005; Brown and Lemmon, 2007). These models may require denser taxon sampling to improve the estimates of the many parameters. Moreover, different genes do not always share the same evolutionary history due to incomplete lineage sorting or hybridization (Maddison, 1997; Degnan and Salter, 2005; Degnan and Rosenberg, 2006; Maddison and Knowles, 2006; Edwards et al., 2007) and, in some cases in which the gene tree does not mach the species tree, it may be best to analyze these loci separately (Ane and Sanderson, 2005). Therefore, data sampling for phylogenetic analysis is not a matter of sequencing many genes or many taxa. Instead, it

is important to adequately sample in both dimensions and to consider complex gene-histories to get a reliable estimate of phylogenetic relationships.

As the generating model becomes more complex, the difficulty of phylogenetic reconstruction increases (Figure 2.9). Estimates from data sets simulated and analyzed under the GTR+I model are less accurate than estimates of data generated under simpler models. However, for very small data sets with fewer than 20% of taxa sampled, analyses of data generated under the simplest model, JC, are less accurate than analyses of HKY data sets. In this case, misinformation in the JC data sets overwhelms the phylogenetic signal of the true tree. Furthermore, my results also underscore the importance of proper model specification (Figure 2.10). For data generated under the HKY model and analyzed under an under-parameterized model (JC), the true model (HKY), and an over-parameterized model (GTR), my results are consistent with other studies demonstrating that assuming an overly simple substitution model leads to a reduction in topological accuracy (Kuhner and Felsenstein, 1994; Lockhart et al., 1996; Sullivan and Swofford, 2001; Lemmon and Moriarty, 2003; Huelsenbeck and Rannala, 2004; Brown and Lemmon, 2007). When the data are analyzed under an over-parameterized model, however, the error rates are not significantly different from analyses under the true model because, in this case, HKY is a sub-model of the more general GTR model.

2.4 CONCLUSIONS

The results of this study stress the importance of dense taxon sampling for accurate phylogenetic inference. This is particularly significant for studies seeking to understand the relationships of higher-level taxonomic groups. Clade-based sampling introduces more long terminal branches than random sampling and increases the difficulty of phylogenetic analysis. However, these results should not be taken as an endorsement for random taxon sampling, because such a strategy is unrealistic. Additionally, assembling a phylogenetic data set by randomly sampling from all of the species within a clade will not necessarily address questions that may be of interest to the investigator. Sampling taxa using a taxonomy-guided strategy, however, is a practical approach for systematists seeking to determine the resolution of particular nodes, such as the relationships among genera within a family. Since it is typically impossible to obtain samples for every member of a given taxonomic group, it is recommended that systematists conducting large-scale phylogenetic analyses should focus on improving the density of sampling within their group of interest by selecting multiple representatives from each taxonomic level. Furthermore, within some clades there may be species rich groups that will require denser sampling than groups with fewer taxa. Such sampling may, in turn, help to reveal polyphyletic or paraphyletic taxonomic groups. For example, Wiens et al. (2005) conducted a phylogenetic analysis of hylid frogs and sampled species based on the current classification of the family. Although many genera in their study were only represented by a single species, they included multiple samples from several of the larger genera. They also heavily sampled the largest genus, *Hyla*, which contained 337 described species (at that time). Their coverage of this species-rich genus included 86 samples, spanning the geographic range of the group. Based on their results, which

indicated that the diverse *Hyla* genus was polyphyletic, Wiens et al. (2005) proposed a new taxonomic classification of hylid frog species.

Although the results of this simulation study can be generalized to some extent to empirical phylogenetic analyses, these simulations are simple and may not emulate all of the complex processes responsible for generating biological sequence data. Assembling a phylogenetic data set from biological data requires an understanding of the group of interest and a clear definition of the scope of the relationships that the systematist wishes to infer. Appropriate taxon selection of a group of organisms is typically data-set dependent and, therefore, simulation studies cannot produce a general "rule of thumb" for the number of taxa or genes that must be sequenced to obtain a robust phylogenetic estimate. Nevertheless, this study and many others have indicated that improving sampling density within the taxonomic group of interest will, on average, result in improved inferences of phylogenetic relationships. Moreover, my results indicate that the way in which taxa are sampled has a significant impact on the accuracy of tree reconstruction and that systematic studies of large groups should consider appropriate strategies for representing species diversity.

ACKNOWLEDGEMENTS

Derrick Zwickl generously provided helpful advice during the development of the tree simulation model. Computational resources were provided by the Center for Computational Biology and Bioinformatics and the Texas Advanced Computing Center. This work was supported by an NSF IGERT grant in Computational Phylogenetics and Applications to Biology awarded to the University of Texas, Austin and the UT Ecology, Evolution, and Behavior Hartman Merit Award.

TABLE 2.1: Conditions for simulation and analysis of sequence data sets. Three different sets of model trees were used that differed in their total tree depth or in the number of total taxa. Sequence data sets were generated under three different substitution models. Data sets generated on tree set A under the HKY model varied in sequence length. Most of the simulated data sets were analyzed under their true (T) simulation model. Model misspecification was assessed by analyzing data sets generated under HKY (with 1000 nucleotide length sequences on model tree set A) under an over-parameterized model (GTR) and an under-parameterized model (JC).

Model Trees			**Simulation Model**	**Sequence Length**	**Analysis Model** T = true
Name	**Total Taxa**	**Tree Depth**			
A	500	1	HKY	500	T
				1000	T
					JC
					GTR
				2000	T
			JC	1000	T
			GTR+I	1000	T
B	1000	1	HKY	1000	T
C	500	0.5	HKY	1000	T

FIGURE 2.1. The functional relationship between the nodal weighted mean imbalance and *ln*(node size) for four sets of trees simulated under a range of variance parameters. Imbalance was calculated using weighted mean imbalance (Fusco and Cronk, 1995; Purvis et al., 2002) which has an expected value of 0.5 under constant speciation/extinction rates (the equal rates Markov model – ERM; dashed line). Higher values of weighted mean imbalance indicate nodes with greater asymmetry. The parameter, alpha, of the gamma-distributed rate prior was changed for each set of simulations to 1, 3, 5, and 10 (for both the speciation rate and extinction rate). Increasing alpha decreases the amount of rate variation and, as a result, it also decreases the amount of nodal imbalance. In the case where alpha = infinity, the tree shapes should be identical to what is expected under the ERM model (dashed line).

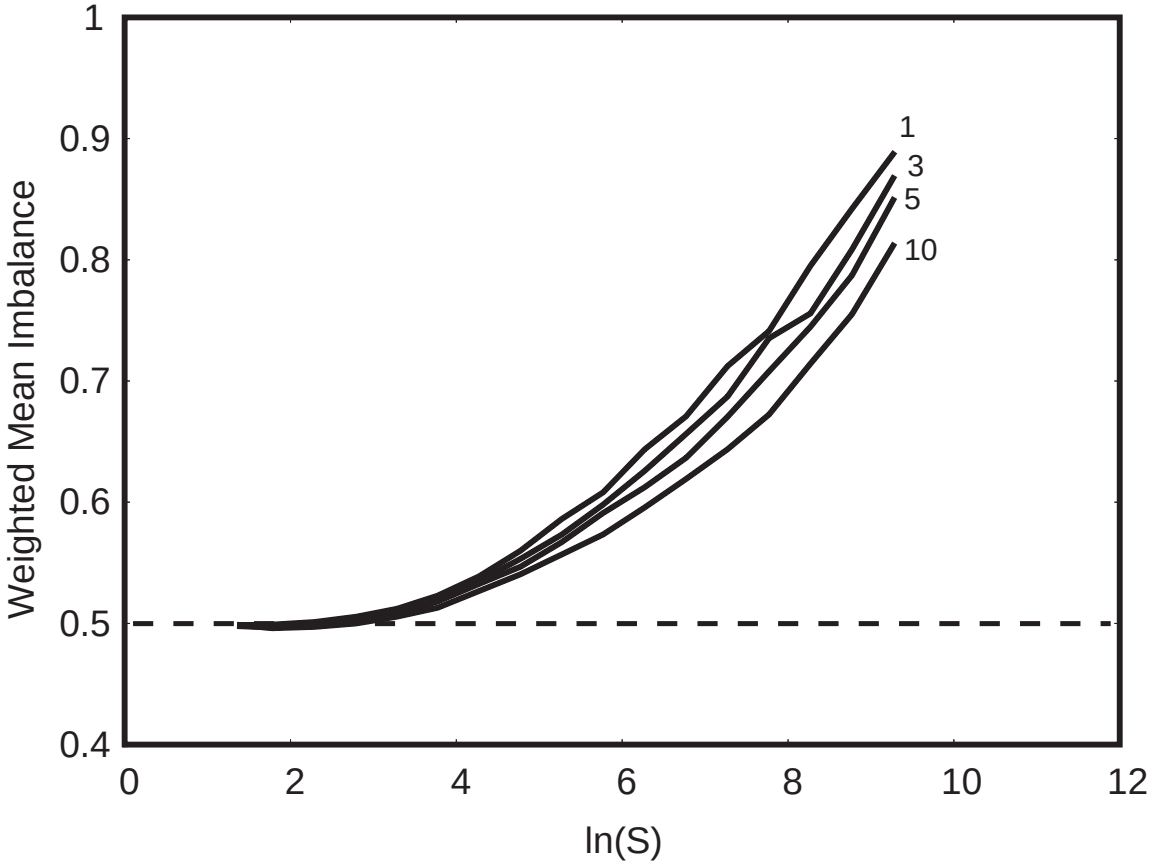


FIGURE 2.2: Examples from each of the sets of model trees. All model tree sets were simulated under variable speciation and extinction rates and consisted of 50 trees. The trees in set A contained 500 taxa and had a total tree depth of 1.0 substitutions/site. Set B trees all had 1000 taxa and were scaled to 1.0 substitutions/site. Set C contained 500-taxon trees scaled to a depth of 0.5 substitutions/site.

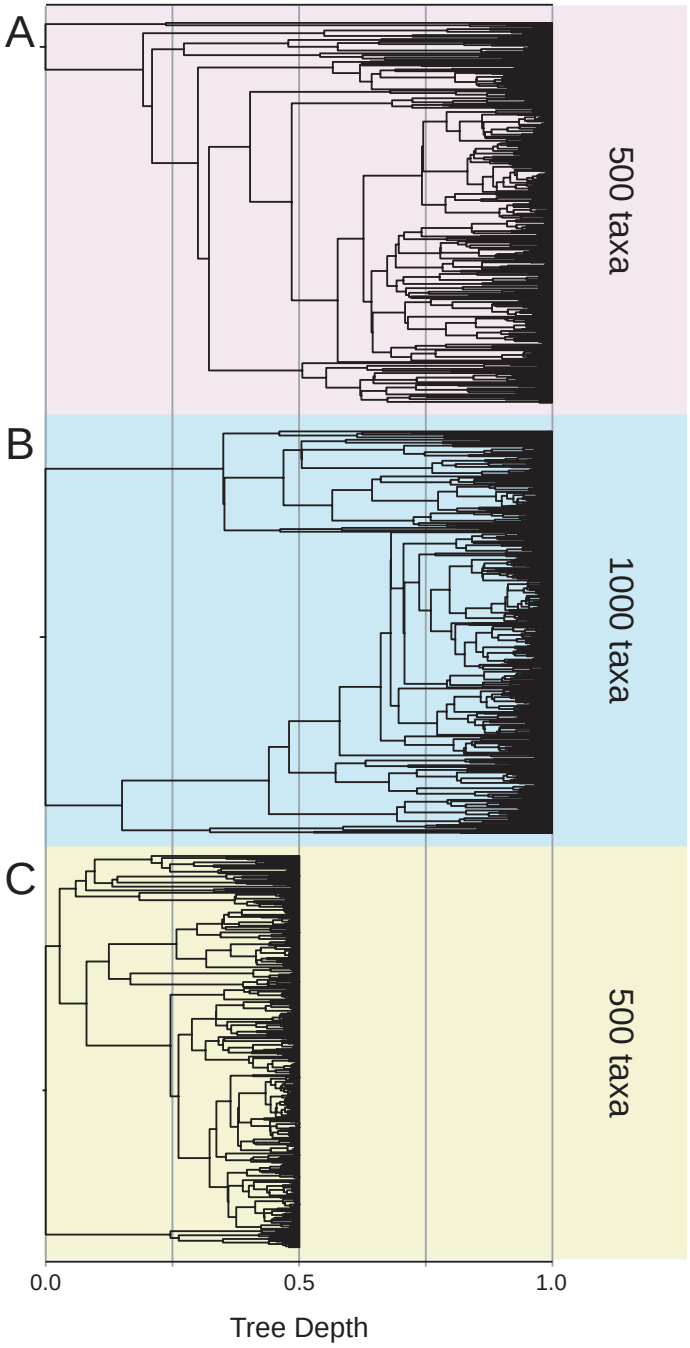


FIGURE 2.3: Clade-based sampling of terminal taxa. Taxa are assigned to a particular taxonomic group based on which of the *n* major sub-clades they belong. A single lineage is chosen from a uniform distribution on all taxa within a clade so that a single sequence represents each major group. In this example, there are a total of 100 taxa in the full tree (A) and 10% are sampled in the reduced tree (B). Thus, the target data set size is $n = 10$.

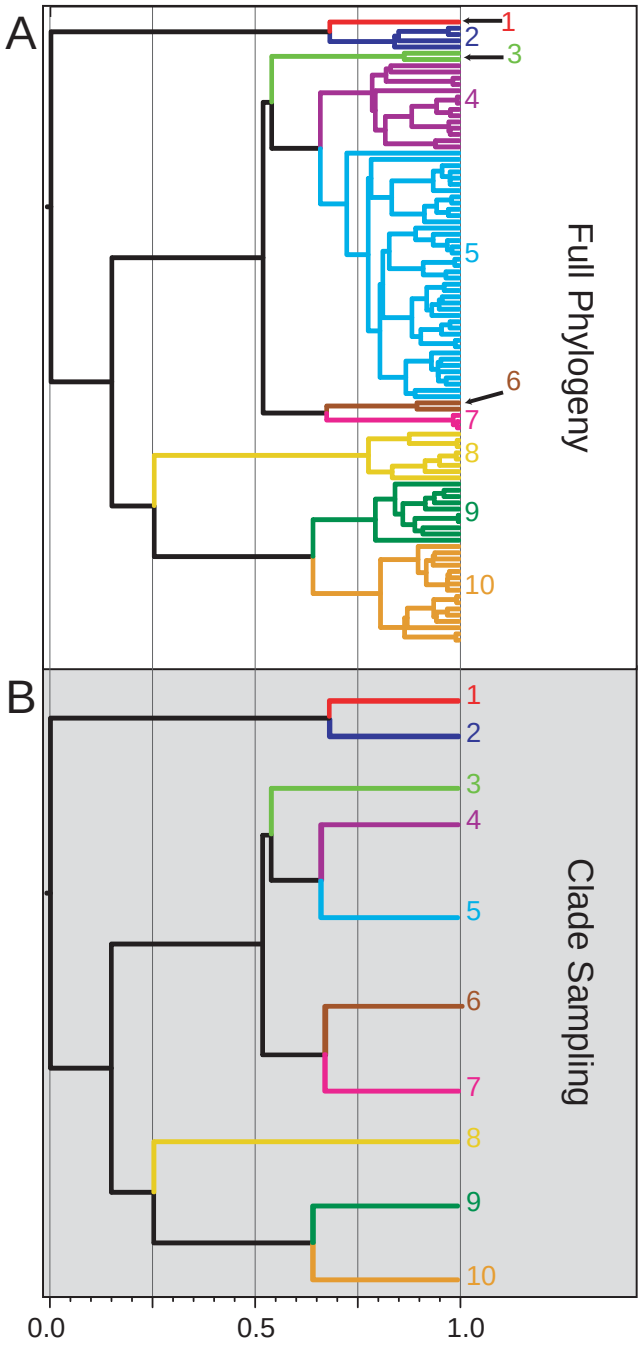


FIGURE 2.4: Absolute error of trees reconstructed from data sets with different taxon sampling densities. Results for data sets simulated on 500 taxon trees (black) and 1000 taxon trees (red). Randomly sampled data sets are represented with filled circles. Taxonomic-based samples are represented with open triangles. Increasing the proportion of taxa sampled improves the accuracy of the reconstructed topology. When taxa are sampled based on taxonomic classification, the overall accuracy of the phylogenetic estimates is significantly lower than when species are randomly sub-sampled.

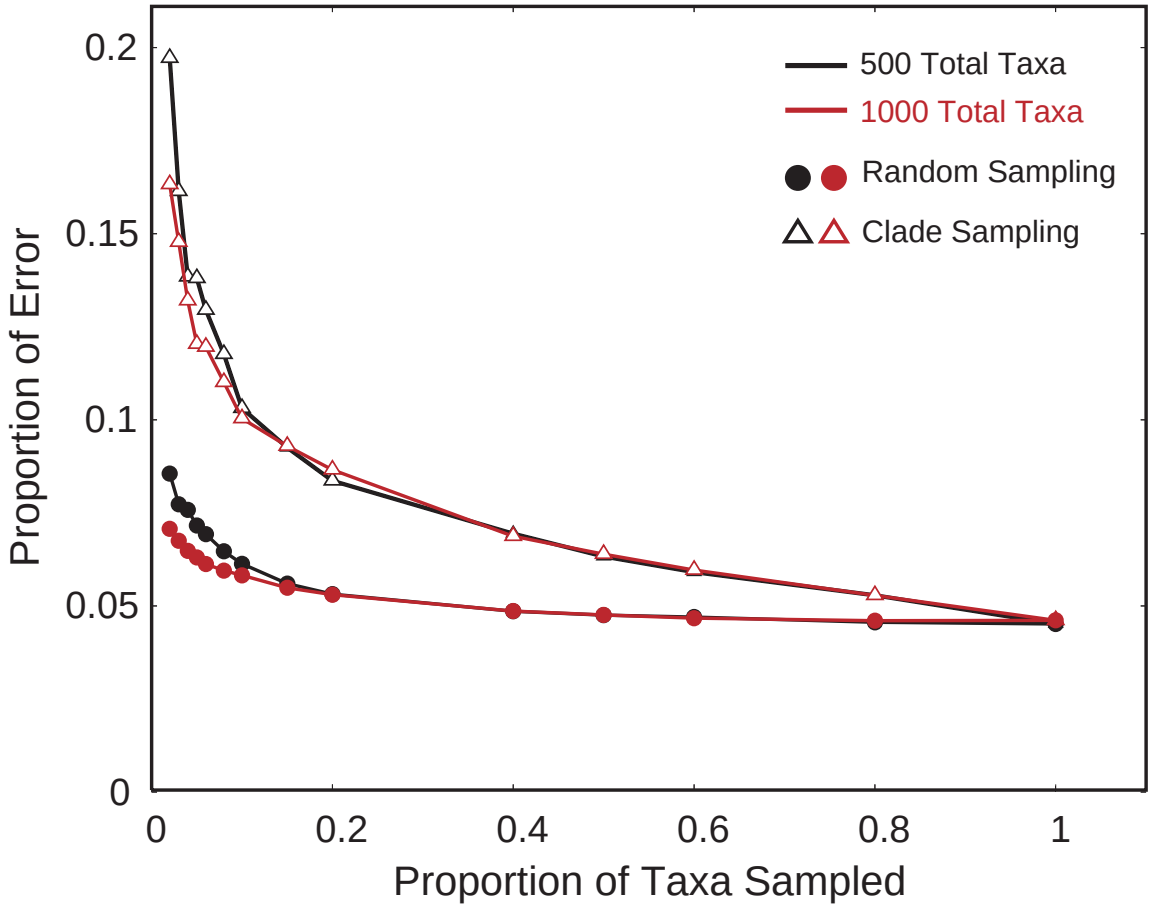


FIGURE 2.5: The average length of terminal branches when extant taxa are sampled randomly or selected based on taxonomic rank from the true trees. The complete trees have a tree depth of 1.0 substitutions/site. Random sampling is represented by solid lines and clade-based sampling is shown with dotted lines for trees with 500 (tree set A; black) and 1000 (tree set B; red) total taxa. Reduced taxon sampling results in an increase in the average terminal branch length, which contributes to the overall decrease in phylogenetic accuracy.

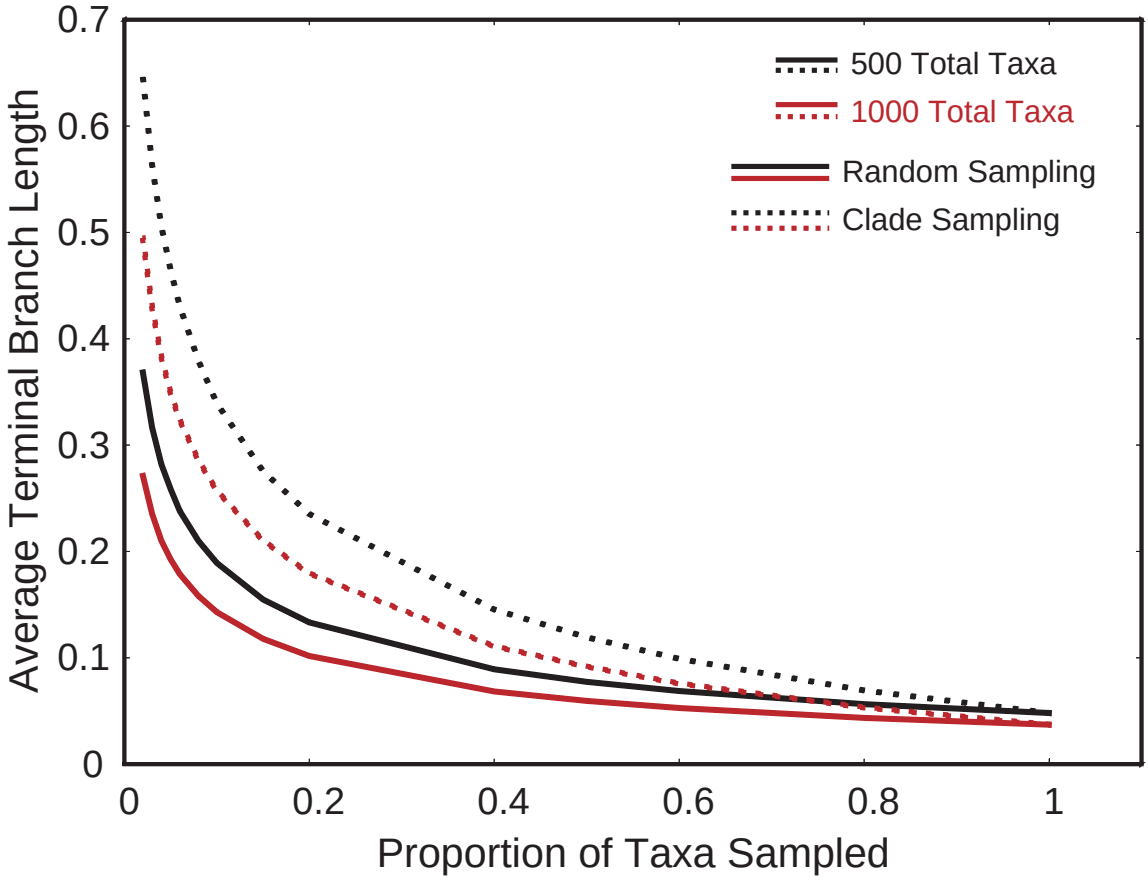


FIGURE 2.6: An example of the effect of different taxon-sampling strategies on terminal branch length. The full phylogeny (A) contains 500 taxa. 50 taxa are sampled randomly (B) and 50 taxa are sampled based on their taxonomic group (C). On average, randomly sampled trees contain shorter terminal branches relative to trees from clade-based samples.

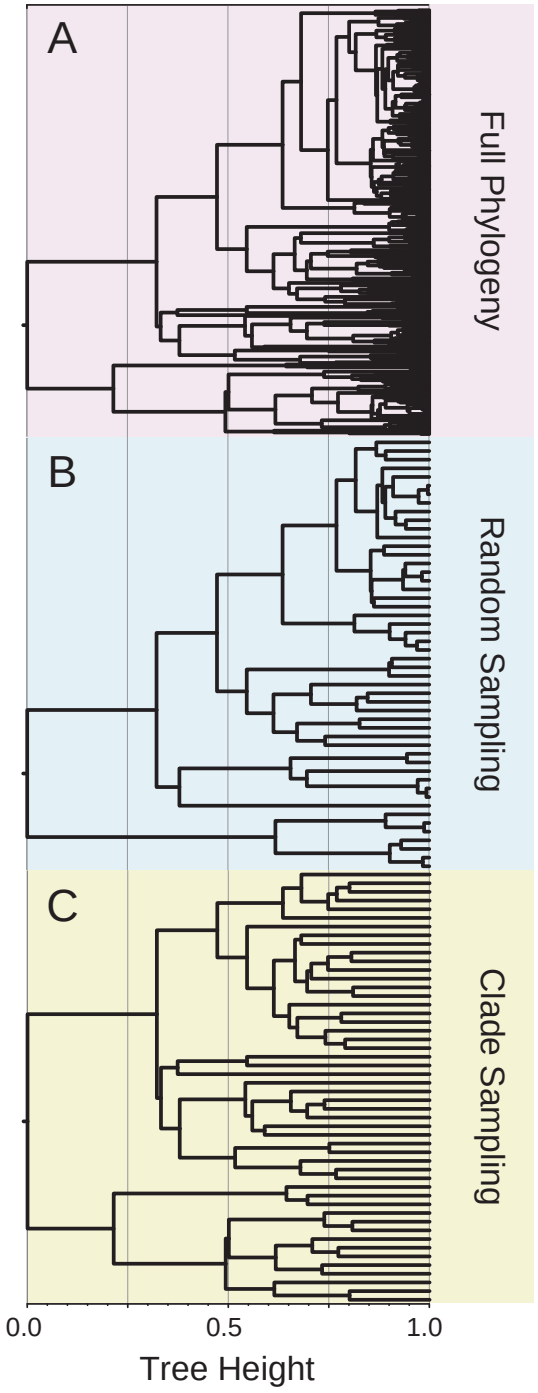


FIGURE 2.7: The proportion of error found in tree topologies reconstructed from subsampled data sets with different tree depths. Two tree depths were considered, the black lines represent TD = 1.0 substitutions/site and the red lines represent TD = 0.5 substitutions/site. The accuracy of trees reconstructed from data with a higher rate of substitution (TD = 1.0) are not significantly less accurate than the trees estimated from the slower evolving sequences. The observed patterns of error for both tree depths indicate that the importance of dense taxon sampling holds for a range of substitution rates.

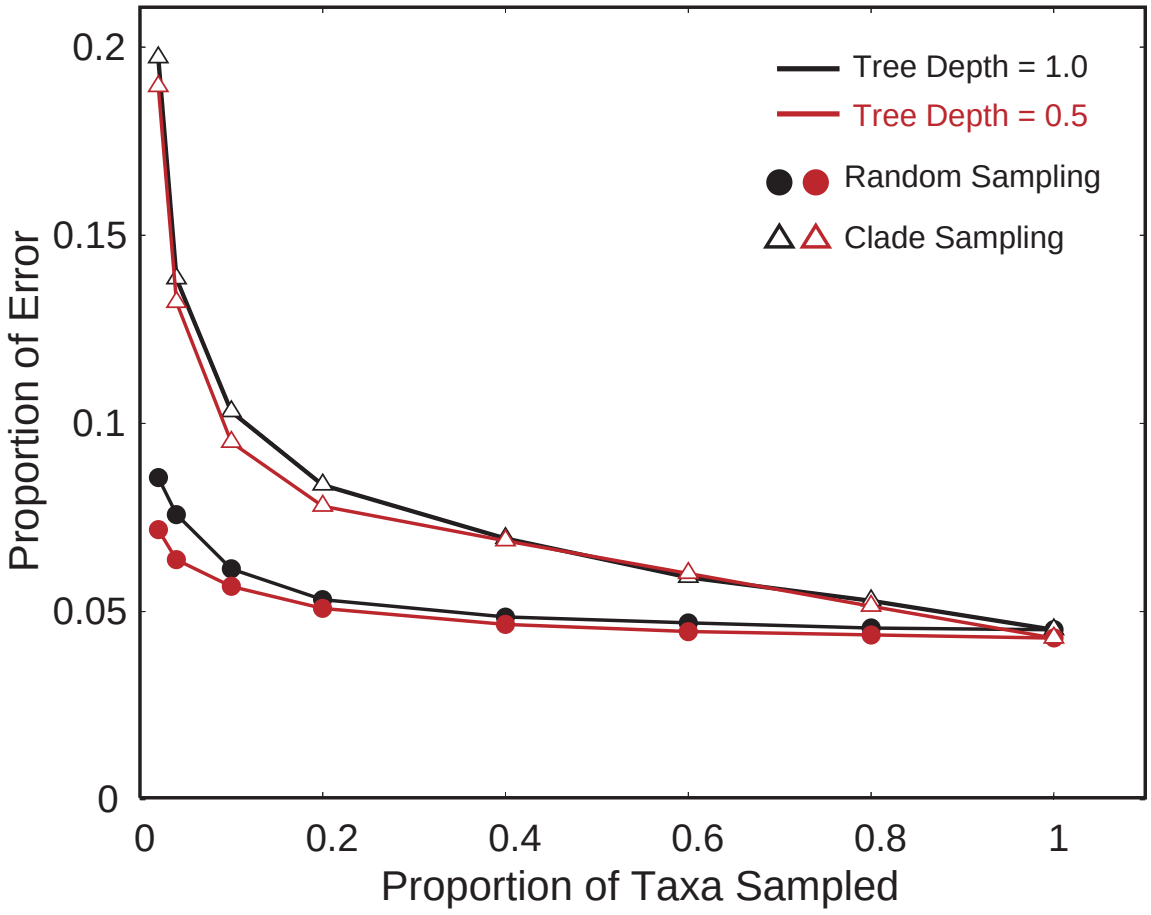


FIGURE 2.8: The effect of increased sequence length on the accuracy of trees reconstructed from data sets sampled using different strategies. Data sets generated on tree set A (500 total taxa, TD = 1.0) with 500, 1000, and 2000 nucleotides were examined. Increasing the number of characters greatly improves the accuracy for both taxon-sampling strategies when the sequences are very short (500 bases). This illustrates the importance of increasing both the proportion of taxon sampling as well as the amount of data per taxon.

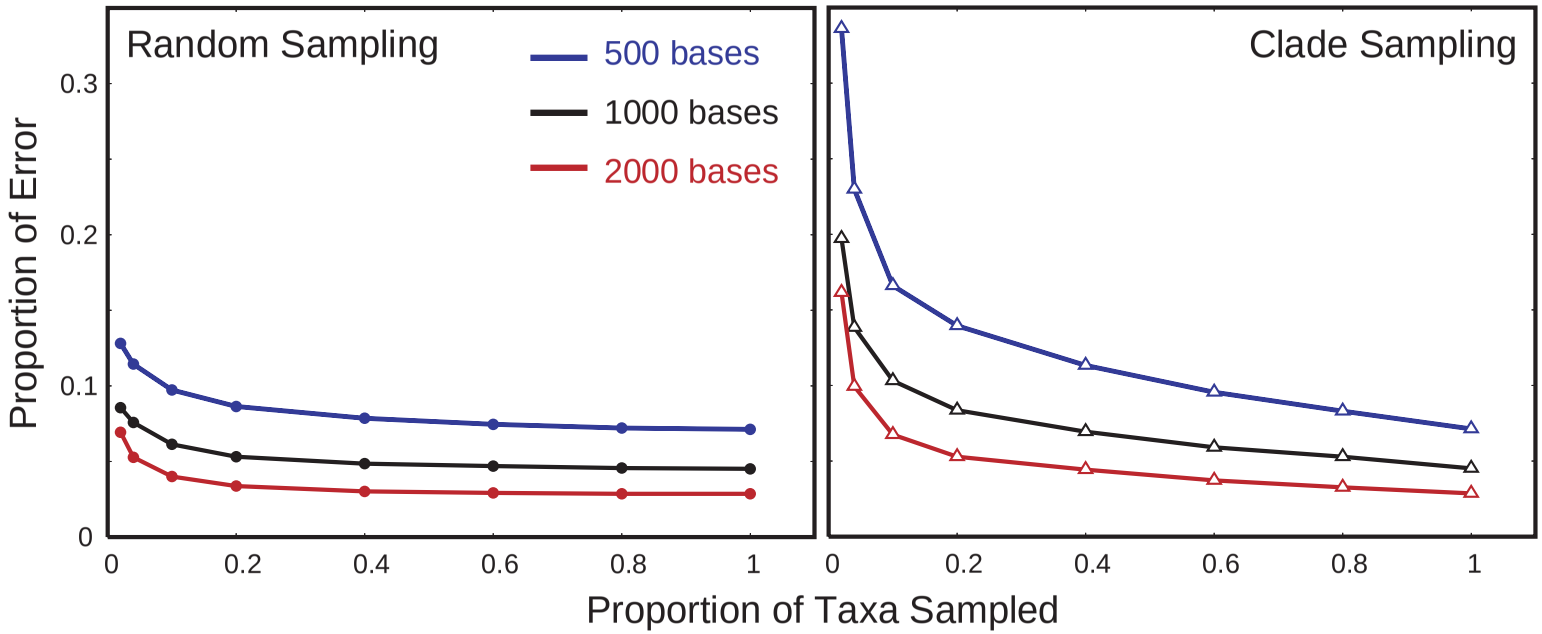


FIGURE 2.9: The effect of substitution model complexity on the accuracy of phylogenetic reconstruction. Data sets were generated using tree set A (500 total taxa, TD = 1.0) under three different substitution models (see Table 2.1). Trees estimated from data generated and analyzed under the HKY model are represented with black lines, GTR+I data sets are indicated with blue lines, and JC data sets are shown with red lines. As the model used to generate the sequences becomes more complex, accuracy is reduced when the sampling density is greater than 40%. However, if the data set contains very few taxa (fewer than 20%), reconstruction from data generated under a very simple model (JC) can be less accurate than estimates using data generated under a more heterogeneous model (HKY).

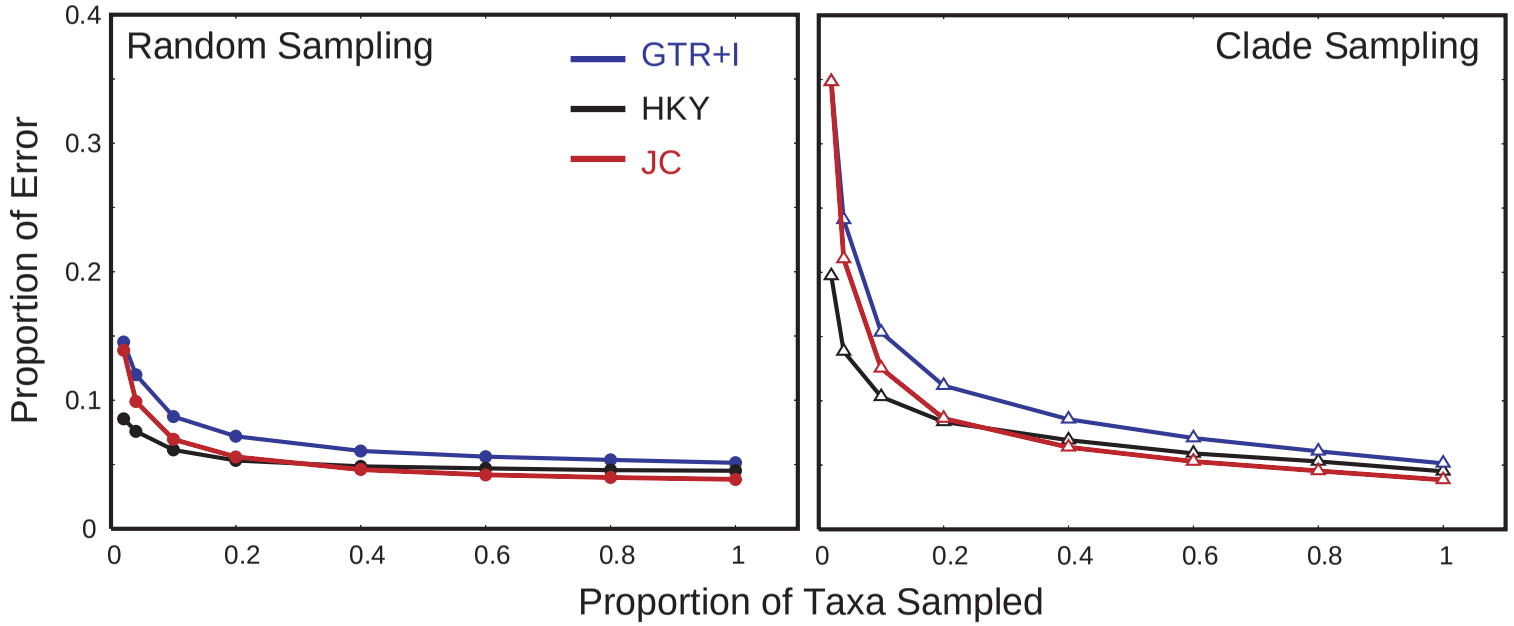
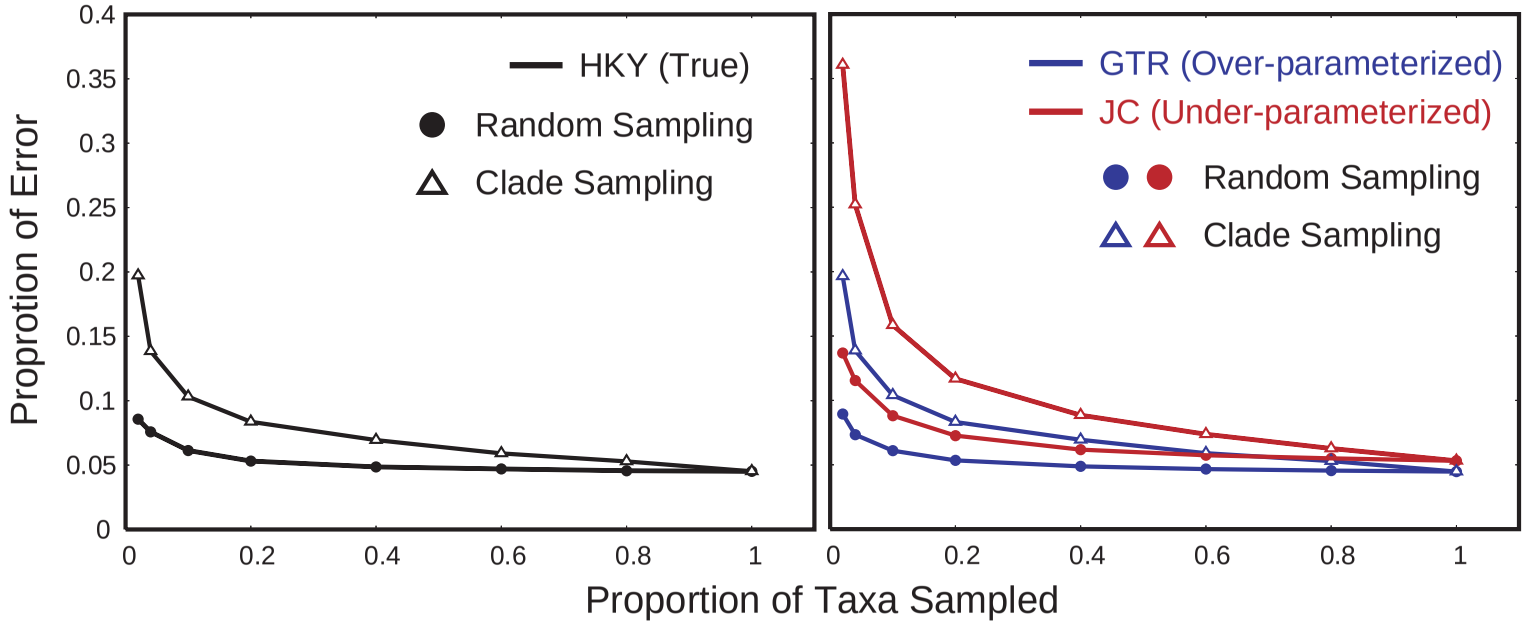


FIGURE 2.10: The effect of model misspecification on the accuracy of trees reconstructed from sub-sampled data sets. All of the data sets were simulated under the HKY model then analyzed under the true model (HKY; black), an over-parameterized model (GTR; blue), and an under-parameterized model (JC; red). The error of trees estimated using an over-parameterized model is indistinguishable from the pattern observed when the true model is used. This is due to the fact that HKY is a sub-model of the GTR model. When the data are analyzed under an overly simple model, however, the reconstructed trees are less accurate.



Chapter 3: Taxon Sampling Affects Inferences of Macroevolutionary Processes from Phylogenetic Trees

3.1 INTRODUCTION

Phylogenetic relationships across the Tree of Life form the basis for comparing and organizing the Earth's biodiversity. In addition to providing information about the evolution of individual genes, populations, or species, phylogenetic trees are often used to study broader evolutionary patterns. In particular, the shape of phylogenetic trees (e.g., the distribution of cladogenic events across the tree) has been used to understand broad speciation and extinction patterns (Raup et al., 1973; Gould et al., 1977; Rosen, 1978; Savage, 1983; Mitter et al., 1988; Heard, 1992; Guyer and Slowinski, 1993; Mooers and Heard, 1997; Dodd et al., 1999; Good-Avila et al., 2006; Ricklefs, 2006). The results of many studies on phylogenetic tree shape suggest that variation in the rates of speciation and extinction has played an important role in shaping the Tree of Life. However, it remains to be determined to what extent biologists can detect the patterns resulting from the evolutionary processes that shape trees. These patterns can be obscured by non-biological factors that can bias tree shape, such as incomplete taxon sampling (Mooers, 1995; Rannala et al., 1998; Pybus and Harvey, 2000; Purvis and Agapow, 2002; Huelsenbeck and Lander, 2003), phylogenetic reconstruction methods (Heard and Mooers, 1996; Huelsenbeck and Kirkpatrick, 1996), or phylogenetic noise (Mooers et al.,

1995; Heard and Mooers, 1996; Stam, 2002). Therefore it is important to understand how estimates of tree shapes might be biased as a result of non-biological factors.

Tree shape often refers to either the distribution of branching times over the tree (using measures such as the γ-statistic, Pybus and Harvey, 2000), or tree imbalance (Shao and Sokal, 1990; Kirkpatrick and Slatkin, 1993; Agapow and Purvis, 2002). Measures of tree imbalance (the focus of this study) assess the distribution of lineages over a tree topology and quantify the degree of asymmetry among the branches. These measures are often compared to the values expected under a null model of equal speciation/extinction rates over all lineages (the equal-rates Markov model or ERM model). Using a wide range of tree imbalance measures, many studies have found that published phylogenies reconstructed from empirical data are more imbalanced than predicted under the ERM model (Guyer and Slowinski, 1991; Heard, 1992; Mooers, 1995; Purvis and Agapow, 2002; Holman, 2005; Blum and François, 2006). An alternative to the ERM null model is the proportional-to-distinguishable arrangements (PDA) model (or uniform model). Under this model, every labeled tree topology is equally likely (Rosen, 1978). Trees generated under this model are on average more imbalanced than those generated under the ERM model, and studies have shown that the PDA model predicts more tree imbalance than what is observed in empirical phylogenies (Cunningham, 1995; Holman, 2005; Blum and François, 2006).

Numerous researchers have found that taxon sampling has a strong influence on the accuracy of phylogenetic reconstruction methods (Hendy and Penny, 1989; Hillis,

1996, 1998; Graybeal, 1998; Kim, 1998; Rannala, et al. 1998; Poe and Swofford, 1999; Pollack et al., 2002; Zwickl and Hillis, 2002; Hillis et al., 2003; Poe, 2003; DeBry, 2005; Hedtke et al., 2006). Taxon sampling also has an impact on the distribution of branching times and phylogenetic tree imbalance. Removing ingroup taxa creates longer terminal and/or internal branches compared to a phylogeny containing all extant lineages (Rannala et al., 1998; Huelsenbeck and Lander, 2003). In addition to the problems this effect produces for phylogenetic inference, it also can confound estimates of diversification rates, divergence times, rates of molecular evolution, and ancestral state reconstruction (Nee et al., 1994a; Robinson et al., 1998; Ackerly, 2000; Pybus and Harvey, 2000; Salisbury and Kim, 2001; Pybus et al., 2002).

Studies investigating the influence of taxon sampling on tree imbalance have primarily surveyed published phylogenies. Mooers (1995) compiled 39 "full" phylogenies (e.g. trees missing no more than one taxon, where the taxa could be species or higher taxonomic groups) each consisting of 8 to 14 terminal taxa. He compared the imbalance of the full trees to the imbalance in a collection of 82 incomplete phylogenies obtained from a study by Heard (1992). This comparison showed that incomplete trees are more imbalanced than trees comprised of almost all of the members of the group in question. In another study, Purvis and Agapow (2002) collected 61 phylogenies of superspecific taxa and showed that tree imbalance is, on average, greater when the terminal taxa are higher-level taxonomic units than when they are species. It has been suggested that the change in tree imbalance that results from sparse taxon sampling might be due in part to the non-random way in which systematists sample taxa, and that a truly

random selection of taxa may not bias tree imbalance (Guyer and Slowinski, 1991; Kirkpatrick and Slatkin, 1993; Mooers, 1995; Purvis and Agapow, 2002). Heard and Mooers (2002), however, used simulated tree topologies to show that random mass extinctions caused an increase in tree imbalance after a period of recovery if the speciation and extinction rates were allowed to vary.

In this study I investigated the influence of varying levels of random taxon sampling on phylogenetic tree imbalance. I compared the patterns of imbalance found in recently published phylogenies with very low taxon sampling to the expectations of tree imbalance under different branching models and sampling levels. I show that the observed levels of tree imbalance in empirical studies are consistent with the expectations from simulations that include variable and autocorrelated rates of speciation and extinction combined with low levels of taxon sampling.

3.2 METHODS

3.2.1 *Simulations*

Tree topologies were generated under the model incorporating variable and autocorrelated speciation and extinction rates described in chapter 2.2.1. I simulated sets of 500 trees each consisting of 10,000 terminal taxa under a range of parameters for the amount of rate variation. Sets of trees simulated across the range of parameter values showed very similar patterns of imbalance (see chapter 2, Figure 2.1). I also generated

trees under constant speciation and extinction rates (ERM model) and the proportional-to-distinguishable arrangements (PDA) model.

3.2.2 *Empirical Phylogenies*

The set of biological trees was assembled from recently published studies of empirical data (Appendix A). When surveying the literature, I selected trees from studies if their analyses included molecular data and used maximum likelihood, Bayesian, and/or maximum parsimony methods to infer the tree. When a study presented trees estimated using more than one data partition, I selected the tree based on the combined analysis. When I encountered more than one study on a particular taxonomic group, I selected the most recently published tree. The trees in the collection of published phylogenies were then pruned of redundant species, and outgroups were removed so as not to increase the tree imbalance, but retain the root position. Unlike previous studies using published phylogenies (Mooers, 1995; Purvis and Agapow, 2002; Holman, 2005), I only used trees with species as terminal taxa so that I could directly calculate the amount of species-level sampling and avoid subjective aspects of higher-level taxonomic grouping. I determined the proportion of taxon sampling based on the number of described species in the group. These estimates of the proportions of taxon sampling are necessarily dependent on the monophyly of the sampled groups and undiscovered biodiversity, but the overall results do not depend on the exact value of the sampled proportions. Empirical phylogenies were then sorted based on the proportion of taxon sampling and the method used to reconstruct the tree. In this study, I only present the imbalance of phylogenies with sampling

densities lower than 10% because the collection of published studies contained relatively few trees with more complete species sampling.

3.2.3 *Measure of Imbalance*

I calculated the imbalance of simulated and empirical topologies using the imbalance measure first introduced by Fusco and Cronk (1995) and later modified by Purvis et al. (2002). Fusco and Cronk (1995) imbalance is calculated for an individual node such that:

$$I = \frac{B - m}{S - m - 1}$$

where for a given node with S extant descendants, B is the number of terminal taxa descended from the larger daughter lineage and $m = S/2$ (rounded up to the next integer if S is odd). For any node with more than three descendants, I has a maximum value of 1 for a node that is completely imbalanced ($B = S - 1$), and a minimum value of 0 for a node where each daughter lineage has the same number of descendants (or differing by 1 if S is odd). One property of this imbalance measure is that the expected value of I under the ERM model depends on whether S is even or odd (Purvis et al., 2002). Therefore, Purvis et al. (2002) introduced a set of weights (w) to calculate an expected weighted mean of I (I_w) so that the measure has an expected value of 0.5 for all node sizes under equal rates:

if S is odd, w = 1,

if S is even, and I > 0, w = (S – 1) / S,

if S is even, and I = 0, w = 2(S – 1) / S.

For a single node, I_w is the product of I and w divided by the mean of the node weights across the entire tree (Purvis et al., 2002 and Purvis and Agapow, 2002). Using these weights, the imbalance for a collection of nodes can also be measured by calculating the weighted mean of I (Purvis et al., 2002 and Holman, 2005).

Unlike many other measures of tree imbalance (for examples see Agapow and Purvis, 2002), I_w does not require fully resolved topologies (because the imbalance at multi-furcating nodes is not measured), nor is it dependent on the size of the tree. Additionally, I_w can be used to evaluate the imbalance of a collection of trees to assess the relationship between imbalance and node size (Holman, 2005), and compare unique sets of trees to detect differences in macroevolutionary patterns (assuming that there is homogeneity across a set of trees). For each set of trees, the bifurcating nodes with more than three descendants were binned according to the natural log of node size, *ln(S)* in intervals of 0.5, and the weighted mean imbalance for the nodes in each bin was calculated (see Holman, 2005). Although this measure of imbalance was developed for complete trees, or phylogenies of higher level taxonomic groups incorporating species richness data, in this study, I use I_w to determine the impact of reduced species sampling by comparing the imbalance of complete trees with that of incomplete trees.

3.3 Results and Discussion

3.3.1 *The Effect of Node Size on Tree Imbalance*

The nodal weighted mean imbalance for the empirical trees is summarized in Figure 3.1. I observed a pattern of imbalance in empirical trees similar to that reported by Holman (2005), with imbalance increasing as node size increases. A recent study by McPeek and Brown (2007) offers a plausible biological explanation for this positive correlation between node size and imbalance. They observed that clade size increases with clade age, therefore larger nodes are typically older nodes and their descendant lineages have had more time to experience the pressures that may cause shifts in diversification rates. This implies that there is also a positive association between node age and imbalance.

For nodes with fewer than 140 descendants, I did not detect a significant difference in the pattern of imbalance between trees reconstructed under maximum parsimony versus those reconstructed using parametric methods (Figure 3.1). Although there appears to be somewhat greater differences in the imbalance at larger nodes, these differences are largely attributable to the smaller number of observations in those categories. Therefore, I combined the trees into a single set of empirical phylogenies for subsequent analyses. When combining the trees, if a single paper presented both a parsimony tree and a maximum likelihood or Bayesian tree, I selected the tree at random. This combined collection of trees consisted of 77 parsimony trees and 78 maximum likelihood/Bayesian trees.

Figure 3.2 shows the weighted mean imbalance of the combined collection of empirical trees and a set of trees simulated under the model for varying speciation and extinction rates (where $\alpha = 2$ for the gamma-distributed rate priors for both speciation and extinction rates). I also show the imbalance expected under the ERM and PDA models. Although I used a different collection of empirical trees than used in previous studies (Purvis and Agapow, 2002; Holman, 2005; Blum and François, 2006), my results are similar to those found by Holman (2005) and Blum and François (2006). Specifically, the PDA and ERM models do not adequately represent the imbalance found in empirical phylogenies (Figure 3.2). The trees simulated under speciation and extinction rate variation, however, have nodal imbalance that is more representative of empirical phylogenies than the ERM model and are much less imbalanced than trees generated under the PDA model. As with the empirical observations of McPeek and Brown (2007), trees generated under the model developed for this study show a positive association between node size and node age, as well as a positive correlation between node age and imbalance.

3.3.2 *The Effect of Reduced Taxon Sampling on Tree Imbalance*

Unlike some of the previous surveys of tree imbalance (Mooers, 1995; Purvis and Agapow, 2002; Holman, 2005), my collection of empirical trees all had low percentages of sampled taxa because I treated the tips as individual species instead of considering higher taxonomic rank with species richness information. The empirical trees presented in this study all had less than 10% of the described species represented in the phylogeny

(with a median of ~2%). When I randomly pruned taxa from the trees simulated with variable and autocorrelated speciation/extinction rates, I observe an increase in nodal imbalance and a very good approximation of the imbalance found in the empirical trees (Figure 3.3). In contrast, I show that for trees simulated under the ERM and PDA models, random taxon sampling does not alter the functional relationship between imbalance and node size (Figure 3.4). This result was also demonstrated by Heard and Mooers (2002) who showed that random mass extinctions of ERM topologies did not affect tree imbalance after a period of recovery under constant diversification rates.

I randomly pruned 50% of the taxa from trees in the combined set of empirical phylogenies to determine whether or not an additional reduction in taxon sampling would increase the imbalance in empirical phylogenies (Figure 3.5). The results shown in Figure 3.5 are from 100 replicates of randomized pruning and suggest that, on average, random reduced taxon sampling does indeed increase the imbalance in these trees.

These results indicate that incomplete taxon sampling in the presence of diversification-rate variation may be sufficient to explain much of the imbalance observed in the collection of empirical trees, because as species are removed from a phylogeny, the apparent variation in the rates of diversification is increased. My simulations show that older nodes are, on average, more imbalanced than younger nodes. Therefore, pruning taxa from these trees results in an increase in the average age of the internal nodes, and additionally, removal of terminal branches increases the average imbalance for nodes of a given size. However, it remains unclear exactly how much

reduced taxon sampling biases tree imbalance. The published phylogenies used in this study most likely do not contain random samples of taxa, so it is difficult to determine the relative influence of biased taxon sampling versus random sampling on tree imbalance. Because so many factors influence whether or not a species is included, it is difficult to emulate the way in which systematists sample taxa. Using a simple model of biased taxon sampling, however, Mooers (1995) was able to show that nonrandom exclusion of terminal lineages can increase the imbalance of ERM trees. More investigation into the impact of biased taxon omission on phylogenetic tree shape and tree reconstruction is required.

When incomplete species sampling is taken into account, the model for varying speciation and extinction rates presented in this paper is a better representation of the tree shapes observed in published phylogenies than the ERM model or the PDA model. However, it is a parametric, stochastic model and not based on detailed biological processes. The described model does not attempt to capture all of the biological and environmental factors by which diversification rates vary over the course of evolution. Although the specific values of parameters in the heterogeneous-rate model can be adjusted to produce varying levels of tree imbalance (see chapter 2, Figure 2.1), the general conclusions of the simulations remain consistent across a wide range of parameter values. These simulations demonstrate that it is important to consider the interaction between diversification-rate variation and reduced taxon sampling when assessing the shapes of empirical phylogenies (Figure 3.3). Inferences of macroevolutionary processes based on incomplete phylogenies should be interpreted with

caution, and, when available, information on species diversity should be included in the calculation of I_w (Fusco and Cronk, 1995). This may result in a less biased estimate of tree imbalance even without relatively complete taxon sampling.

3.4 CONCLUSIONS

Variation in the relative rates of speciation and extinction produces tree topologies with greater imbalance than trees generated under the equal rates model (Figure 3.2). Removal of taxa from trees generated under variable and autocorrelated rates results in a disproportionate representation of older divergences and increases the apparent variation in diversification rates among the lineages on the tree. Consequently, reduced taxon sampling causes an increase in tree imbalance (Figure 3.3), which, in turn, may mislead analyses using tree shape to detect shifts in diversification rates.

It is also important to note that there are other non-biological factors that can contribute to imbalance in empirical phylogenies. Methods of phylogenetic reconstruction have been shown to be biased toward imbalanced trees (Huelsenbeck and Kirkpatrick, 1996), at least for trees of few taxa. Additionally, incorrect rooting of the tree can result in a more imbalanced topology. These factors may make it very difficult to tease apart the biological processes that contribute to tree imbalance.

It will be important to understand and account for these non-biological contributors to tree imbalance if tree shape is to be used to study large-scale patterns of

diversification. However, it is clear that in addition to producing more accurate estimates of phylogenetic relationships, increased taxon sampling also improves inferences about macroevolutionary events based on phylogenetic tree shape. As more complex and realistic models of diversification-rate variation are developed, we will improve our understanding of the macroevolutionary forces that shape the Tree of Life. In addition, as phylogenetic reconstruction programs become capable of handling larger data sets (e.g., Stamatakis, 2006; Zwickl, 2006), models of complex branching processes can be used to generate model tree topologies for large-scale simulation studies on these new algorithms.

ACKNOWLEDGEMENTS

My co-authors (Derrick J. Zwickl, Junhyong Kim, and David M. Hillis) and I gratefully acknowledge Vincent Savolainen, Rod Page, Jack Sullivan, Arne Mooers, and an anonymous reviewer as well as Mike Steel, members of the CIPRES project, the UT-IGERT discussion group, and members of the Hillis/Bull/Cannatella lab groups for helpful comments and advice. Financial support for this study was provided by the National Science Foundation (NSF EF 0331453 to the University of Texas and NSF EF 0331654 to the University of New Mexico). TAH was funded by a graduate research traineeship provided by an NSF IGERT grant in Computational Phylogenetics and Applications to Biology awarded to the University of Texas, Austin. Computational resources were provided by the Texas Advanced Computing Center (TACC) at The University of Texas at Austin (http://www.tacc.utexas.edu).

FIGURE 3.1. The weighted mean imbalance of empirical trees plotted as a function of the natural log of the node size (S). The dashed line at 0.5 indicates the imbalance expected under the ERM model. 124 trees reconstructed using maximum parsimony (MP) are indicated by the dotted line with black triangles and 107 trees reconstructed by maximum likelihood or Bayesian methods (ML/B) are represented by the solid line and white triangles.

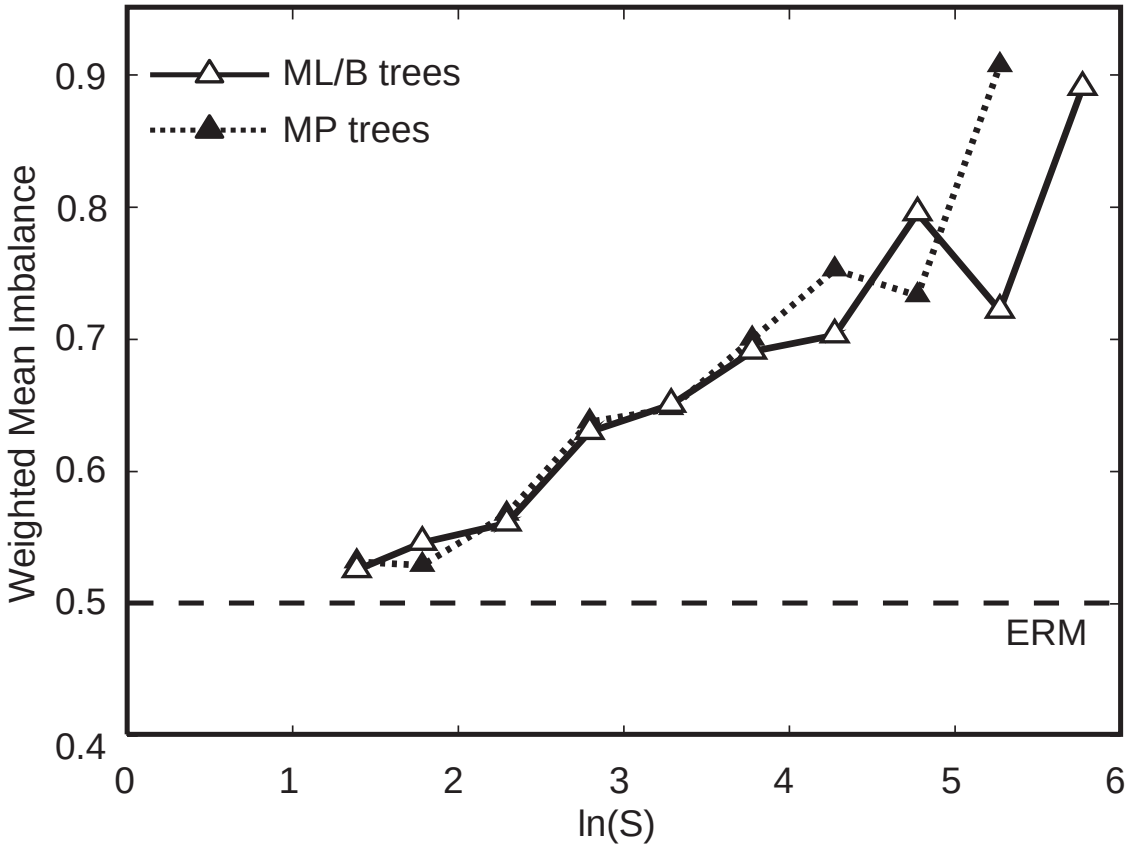


FIGURE 3.2. The nodal imbalance for the combined collection of empirical trees (triangles; 157 total trees) and the collection of trees simulated under varying rates of speciation and extinction (circles). The upper dotted line represents the imbalance expected for trees generated under the PDA model and the dashed line at 0.5 indicates the imbalance expected under the ERM model.

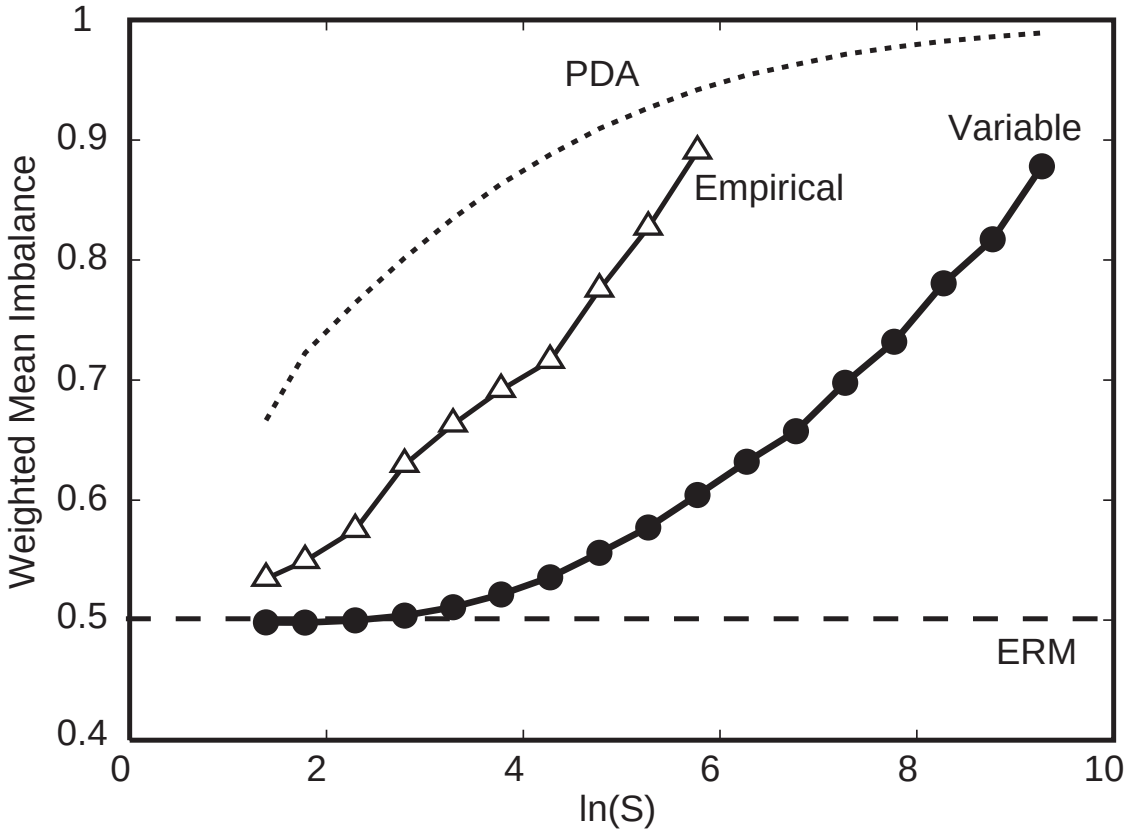


FIGURE 3.3. Weighted mean imbalance for empirical trees (dotted line/triangles) and trees simulated under varying rates with different levels of taxon sampling (solid line/circles). The simulated trees were reduced to 3% and 1% taxon sampling. The dashed line at 0.5 indicates the imbalance expected for trees generated under the ERM model.

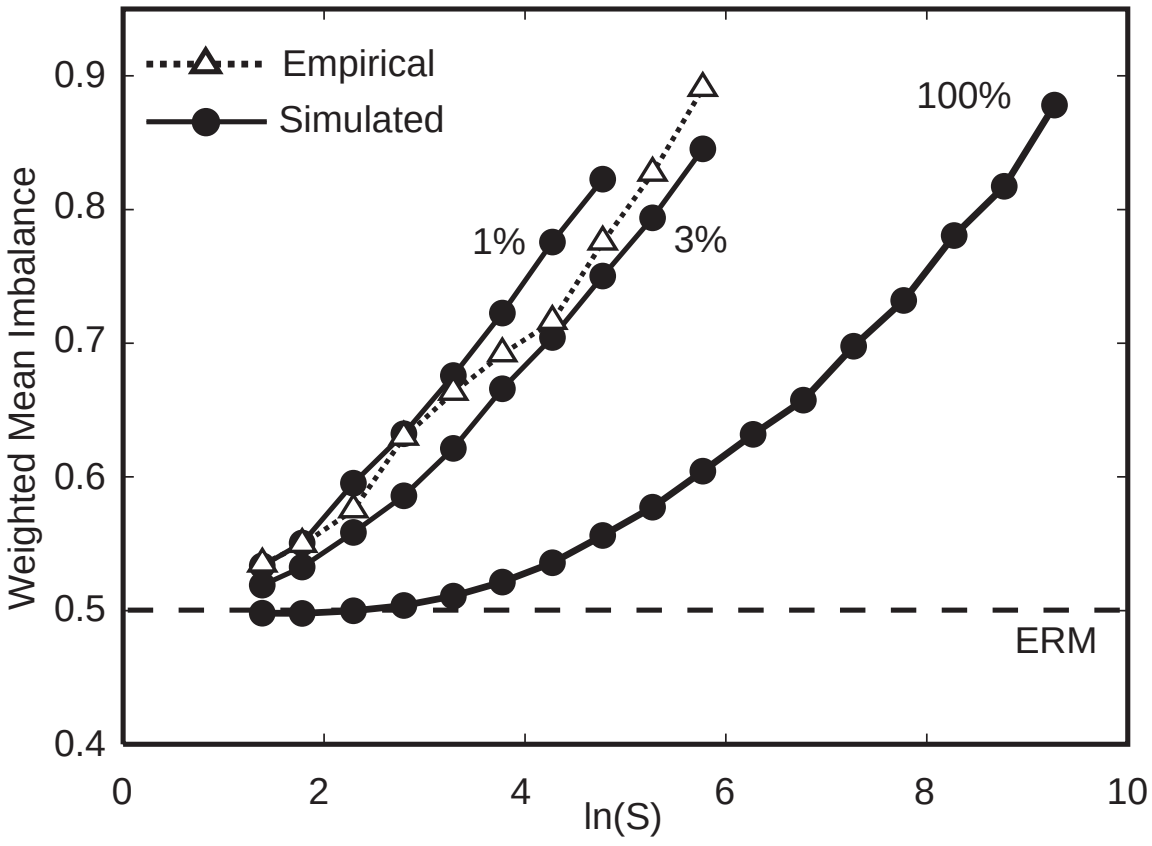


FIGURE 3.4. Weighted mean imbalance as a function of the natural log of the node size for trees simulated under the PDA model (black), the ERM model (black), and variable rates model (gray). The sets of trees with 100% taxon sampling are indicated by dashed lines. Sets of trees with 3% taxon sampling are represented by the solid lines. These simulations indicate that random taxon sampling of trees generated either by the PDA model or the ERM model does not result in a change in the relationship between imbalance and node size, whereas there is a strong taxon-sampling effect for the variable rates model.

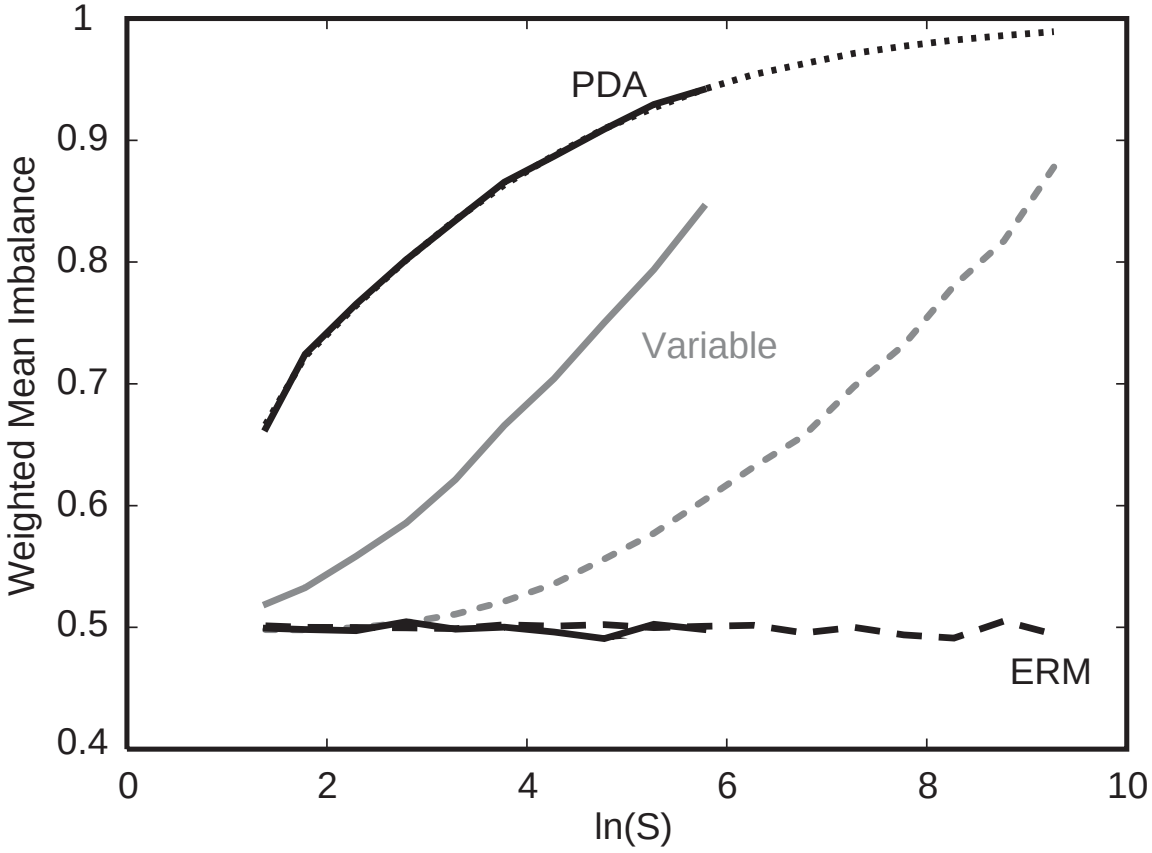
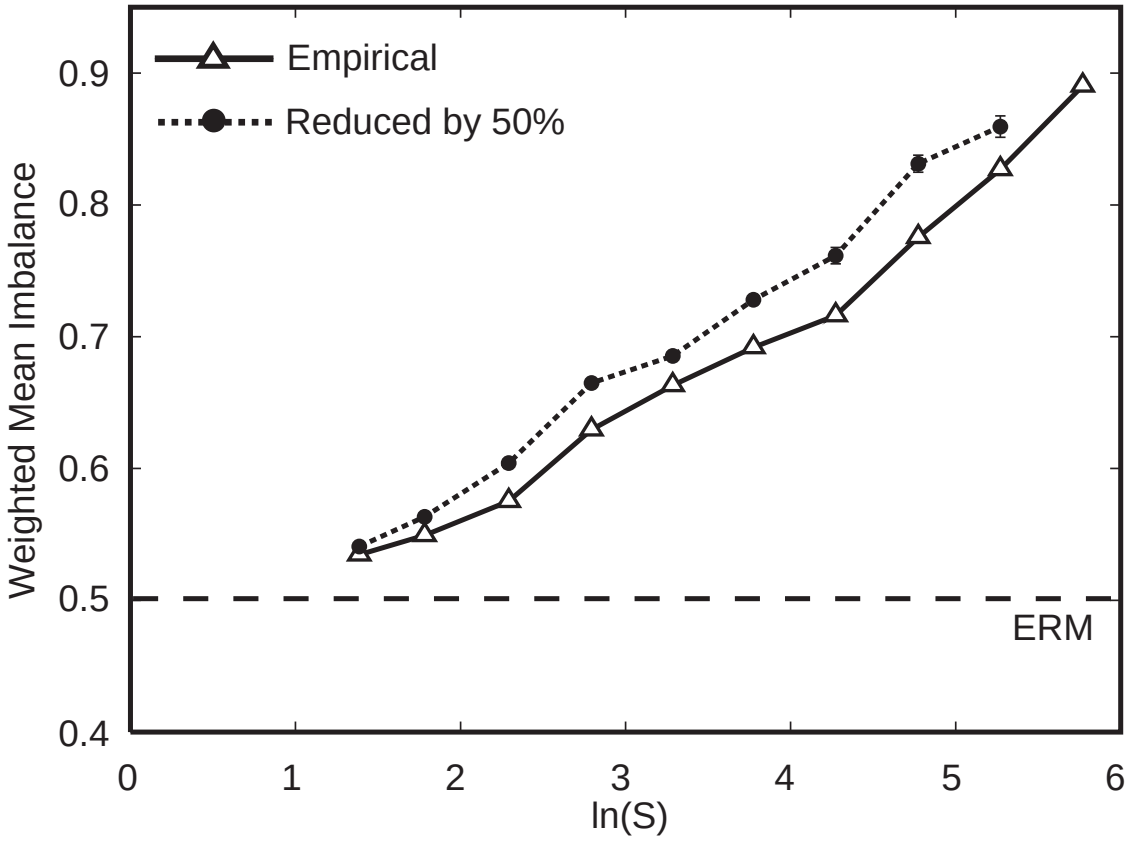


FIGURE 3.5. The weighted mean imbalance of empirical trees with reduced taxon sampling. The imbalance of the published phylogenies without a reduction in taxon sampling is represented by the solid line. The dotted line indicates the same set of trees with a 50% reduction in taxa averaged over 100 replicates with standard error bars. The dashed line at 0.5 indicates the imbalance expected under the ERM model.



Chapter 4: Factors Contributing to Systematic Biases in Phylogenetic Tree Imbalance

4.1 Introduction

Phylogenetic trees are fundamental components of studies that seek to broaden our understanding of evolutionary processes and are widely used throughout biology. Because of their important role, it is critical that the methods for estimating phylogenies are thoroughly understood and developed. Extensive investigation has revealed conditions under which the available methods produce accurate or inaccurate estimates of topology, branch lengths, or parameter values (Felsenstein, 1978; Hendy and Penny, 1989; Huelsenbeck and Hillis, 1993; Zharkikh and Li, 1993; Yang, 1994; Huelsenbeck, 1995; Hoyle and Higgs, 2003; Huelsenbeck and Lander, 2003). Notwithstanding all of these valuable studies, it is still not well understood whether phylogenetic reconstruction methods are biased toward particular trees when estimating phylogenies from large data sets. The goal of this study is to explore the range of conditions under which methods for reconstructing phylogenetic trees can be biased toward certain tree shapes.

Estimates of phylogenetic relationships reveal patterns of diversity that can be used to elucidate the evolutionary processes acting on lineages in the Tree of Life. Differential rates of speciation and extinction or extrinsic environmental factors, such as those that cause mass extinction, can leave signatures on species phylogenies. Methods

used to detect these signature patterns often measure asymmetry in the branching pattern of a rooted phylogeny (tree imbalance) and compare the observed values of imbalance to the values expected under a null model that assumes constant rates of speciation and extinction (equal rates Markov model, ERM). As I discussed in chapter 3, variation in diversification rates results in tree topologies that are less balanced than expected under the ERM model. Additionally, many studies have observed greater imbalance (on average) in empirical phylogenies (see chapter 3; Guyer and Slowinski, 1991; Heard, 1992; Mooers, 1995; Aldous, 2001; Purvis and Agapow, 2002; Holman, 2005; Blum and François, 2006). Therefore, it is important to determine whether the degree of tree imbalance detected in published, empirical phylogenies results from biases in the methods used to reconstruct the trees.

A study by Heard (1992) examined empirical phylogenies to test the prediction that trees reconstructed by parsimony methods would be more imbalanced than those estimated using distance methods (Colless, 1982; Shao and Sokal, 1990). However, Heard (1992) was unable to detect a difference in the imbalance of trees produced by either method. Mooers et al. (1995) collected empirical phylogenies estimated using parsimony and found that trees with low support (as measured by nonparametric bootstrapping or jackknifing) were less balanced than those that were well supported. A second component of the Mooers et al. (1995) study evaluated the imbalance of trees estimated from simulated data and their results indicated that incorrectly inferred trees were generally more imbalanced than true trees (Mooers et al., 1995). Based on their results, Mooers et al. (1995) concluded that inaccurate reconstructions of phylogenetic

trees, due to conflicting information in the data, may contribute to the levels of imbalance observed in published phylogenies.

Huelsenbeck and Kirkpatrick (1996) simulated phylogenetic trees and data sets under simple models to investigate whether certain reconstruction methods were biased toward particular tree shapes. They simulated 8-taxon data sets under the Jukes-Cantor model, and demonstrated that maximum parsimony, maximum likelihood, and distance methods were all biased toward imbalanced topologies. This bias became more exaggerated as substitution rates increased, and at high rates of evolution, maximum likelihood was the most biased method. Although they simulated small data sets under a simple substitution model, Huelsenbeck and Kirkpatrick (1996) predicted that their results would hold for more complex simulations.

Advances in computational resources and recent developments in phylogenetic reconstruction algorithms have allowed us to analyze larger data sets under more complex models. Therefore, it is important to determine if phylogenetic methods are biased under a broader range of conditions. In this study, I used simulated 100-taxon phylogenies and data sets to examine the non-biological factors contributing to tree shape bias. Specifically, I examined the effects of substitution rate, reconstruction method, model misspecification, sequence length, and outgroup branch length. All of these factors are important considerations for phylogenetic analysis and can lead to inaccurate topological reconstruction.

4.2 METHODS

4.2.1 *Simulations*

Model tree topologies and branch lengths were simulated under a constant-rate birth/death process (see Figure 4.1 for examples). Phylogenies generated under this model of cladogenesis are consistent with the equal rates Markov (ERM) model. This null model of diversification assumes that speciation and extinction rates remained unchanged over the course of evolution. Under this model, tree topologies are much more balanced than trees generated in the presence of diversification rate variation (see chapter 3). For this study, 1000 tree topologies, each with 100 terminal taxa, were generated using the program Phylogen (Rambaut, 2002). Birth and death rates were arbitrarily set to 0.4 and 0.3, respectively. Simulating tree topologies under this random branching process produces ultrametric trees, where branch lengths are proportional to time. Because the trees were simulated until a specified number of terminal taxa (100) had been generated, the model trees varied in tree depth (*TD* = distance from the root to tips). After the trees were simulated, every tree was rescaled to twelve different tree depths, producing 12 sets of 1000 trees. The trees were rescaled to 0.05, 0.1, 0.15, 0.2, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, and 2.0 (in units of expected substitutions/site). Because all of the simulated phylogenies were ultrametric, the depth of each tree corresponds to the substitution rate.

Although identifying the root of a tree is an important consideration when using phylogenies as tools for understanding evolutionary processes, rooting can be a very

difficult problem for phylogenetic inference. Measures of tree imbalance and other applications that use tree shape to detect shifts in diversification rates require rooted trees. To examine the effect of outgroup selection on tree shape bias, I added three different outgroup taxa to each of the simulated phylogenies (Figure 4.2). The outgroup lineages were attached to each tree in a basal polytomy, each differing in their distance to the root (*d*). The outgroup taxon used in the simulation study by Huelsenbeck and Kirkpatrick (1996) was contemporary with the ingroup taxa. This is outgroup taxon A in this study (Figure 4.2), where the length of the outgroup branch is equal to the total tree depth ($d = TD$). In contrast, outgroup taxon B was the ancestral sequence of the ingroup ($d = 0$). The length of the third outgroup lineage (outgroup C) was half the total tree depth ($d = 0.5 * TD$).

Each of the 12,000 trees was used to generate sequence data under a range of substitution models. The simplest model considered was the Jukes-Cantor (JC) model (Jukes and Cantor, 1969). Data were also simulated under JC with gamma-distributed rate-heterogeneity (JC+G), the Kimura 2-parameter model (K2P; Kimura, 1980) which includes differential rates of substitution for transitions and transversions, the Hasegawa-Kishino-Yano model (HKY; Hasegawa, Kishino, and Yano, 1985), and the general-time-reversible model (GTR; Tavare, 1986). Model parameters (Appendix A) were chosen from values estimated using empirical sequence data (Murphy et al., 2001). For each simulated data set, the total number of nucleotides was set to 1000. Additional data sets of 500 and 2000 nucleotides were simulated under the JC and HKY models. Generating

sequence data that varied in the number of characters provided a way to investigate the impact of sequence length on tree shape bias.

An additional set of 1000, 100-taxon trees was simulated under a model for varying rates of speciation and extinction. These non-ERM trees were generated under the model described in chapter 2 (see Figure 4.3 for examples of the non-ERM trees). Trees generated under a model of cladogenesis that incorporates variation in diversification rates are more imbalanced than expected under the ERM model (see chapter 3). The variable rates trees were scaled to different substitution rates ranging from 0.5 to 2.0 substitutions/site. Sequence data sets were simulated on the imbalanced trees under the JC model, with outgroup A.

The goal of this study was to determine the non-biological factors that contribute to tree shape bias and inaccurate phylogenetic reconstruction. In total, 276,000 data sets were constructed so as to best test a range of properties: substitution rate, complexity of the simulation model, model misspecification, sequence length, outgroup branch length, and model of cladogenesis (Table 4.1).

4.2.2 *Phylogenetic Reconstruction Methods*

Phylogenetic trees for each simulated data set were reconstructed using three commonly applied methods: neighbor joining (NJ), maximum parsimony (MP), and maximum likelihood (ML). Although Bayesian inference methods are frequently used by

systematists and known to produce robust estimates of evolutionary relationships, the trees produced by these methods are summaries of the posterior distribution of topologies, rather than point estimates. As a result, Bayesian methods were not considered in this study.

Neighbor joining (NJ; Saitou and Nei, 1987) analyses were conducted using maximum likelihood corrected distances under the true simulation model in PAUP* version 4.0b10 (Swofford, 1998). Distance-based phylogenetic reconstruction is often conducted using logarithmic formulae to correct for unobserved substitutions according to a model of sequence evolution. However, Hoyle and Higgs (2003) showed that distance methods using logarithmic formulae are error-prone at high rates of substitution. This problem can be alleviated if maximum likelihood distances are used. For each data set, a NJ tree was constructed using uncorrected p-distances. This initial tree was then used to estimate model parameters under maximum likelihood. The parameter values were set to equal the maximum likelihood estimates and the final tree was generated using NJ with distances calculated based on the fixed ML parameter estimates.

Maximum parsimony (MP) reconstruction from each simulated data set was carried out using a heuristic search in PAUP* version 4.0b10 (Swofford, 1998) with 10 replicates, each using a random step-wise addition sequence starting tree and tree-bisection and reconnection (TBR) branch swapping with no collapsing of zero-length branches. For each analysis, tree statistics were averaged across all of the most-parsimonious trees retained by the search. Because they considered data sets with only

eight taxa, Huelsenbeck and Kirkpatrick (1996) used the branch-and-bound search strategy to find the optimal topology under parsimony, an algorithm that is guaranteed to locate the best tree. With much larger data sets (100 sequences), I was restricted to using a heuristic search, which may not have found the optimal tree for every data set.

Maximum likelihood (ML) estimates of phylogeny were reconstructed using the program GARLI v0.951 (serial version; Zwickl, 2006). Each data set was analyzed under the true simulation model (or in some cases, an intentionally misspecified model) with all other program-specific settings set to default values. GARLI uses an evolutionary algorithm to search through tree space and find the optimal tree topology, branch lengths, and model parameters under maximum likelihood. The GARLI algorithm is very efficient and capable of analyzing large data sets in a reasonable amount of time, making it feasible to conduct robust, parametric reconstruction of thousands of data sets. All GARLI analyses were run on the Lonestar Dell Dual-Core Linux Cluster (configured with 5,200 compute-node processors) at the Texas Advanced Computing Center (TACC: http://www.tacc.utexas.edu/).

Following reconstruction, each estimated tree was rooted using the outgroup taxon. Then, the outgroup was removed so that the shape of the ingroup topology could be measured.

4.2.3 *Measures of Tree Imbalance*

Tree shape was calculated using six different tree imbalance/balance measures: I_C, *mean* I_w, *N-bar*, σ^2, *B1*, and *B2* (Agapow and Purvis, 2002). Because the results are very similar across all tree statistics, I will focus the discussion on the analyses from Colless's imbalance (I_C; Colless, 1982; Heard, 1992) and *mean* I_w imbalance (*mean* I_w; Purvis et al., 2002).

Colless's imbalance statistic (I_C) is commonly used to evaluate tree asymmetry. This measure was first introduced by Colless (1982) and later corrected by Heard (1992). For every internal node (for rooted trees, the number of internal nodes is equal to the number of taxa, *n*, minus *1*), the number of descendants from each of the daughter lineages is compared, where *r* is the number of taxa in the larger daughter clade and *s* is the number of taxa in the smaller daughter clade ($r \geq s$).

$$I_C = \frac{2}{(n-1)(n-2)} \sum_{i=1}^{n-1} (r_i - s_i)$$

A perfectly symmetrical tree (with an even number of taxa) will have a value of 0 for this measure, and a completely pectinate tree will have a value of 1. I_C requires a rooted and bifurcating tree topology. Heard (1992) and Rogers (1994) gave formulas for calculating the expected value of I_C under the assumption of constant speciation and extinction rates (for 100 taxa: *ERM* $I_C = 0.072$). This tree shape measure (like many others) is dependent on the size of the tree and gives high weight to the imbalance at the root of the tree (Kirkpatrick and Slatkin, 1993; Mooers et al., 1995).

Imbalance at a single node can be measured using a statistic introduced by Fusco and Cronk (1995) and later modified by Purvis et al (2002). This measure of imbalance (described in chapter 3.2.3) can be calculated for the whole tree by averaging the weighted imbalance (I_w) across all nodes in the tree (*mean* I_w). *Mean* I_w does not require strictly bifurcating topologies and the expectation under the ERM model is size independent (*ERM mean* $I_w = 0.5$).

The values for all of the tree statistics were compared to the values of tree shape for the true (ERM) topologies. In addition, because simulated data were used, I also calculated the accuracy of the reconstructed (unrooted) ingroup topologies. Accuracy was measured using the absolute error (*E*), which uses the Robinson-Foulds (RF) distance (also called the symmetric distance; Robinson and Foulds, 1981; Zwickl and Hillis, 2002). The absolute error is calculated by dividing the RF distance, a measure of the number of incorrect bipartitions in the estimated tree compared to the true tree, by the maximum RF distance. The maximum RF distance is the number of internal branches in the true tree multiplied by two. The absolute error gives a value ranging from zero (identical topologies) to 1 (no shared branches).

4.3 Results and Discussion

Inaccurate estimates of tree topology are (on average) more imbalanced than the true topology. The simulation results show that all three of the phylogenetic reconstruction methods considered are biased toward imbalanced tree shapes. In general,

different reconstruction methods are affected by various properties of the data and have different levels of bias toward imbalanced trees. Figure 4.4 depicts the relationship between tree imbalance (using four different measures of tree shape) and the substitution rate of the trees reconstructed from data simulated and analyzed (for NJ and ML) under the Jukes-Cantor (JC) model. At very high substitution rates, all three methods become increasingly biased toward imbalanced topologies. In contrast to the findings of Huelsenbeck and Kirkpatrick (1996), whose results indicated that ML was the most biased method, I found that NJ, using the minimum evolution method, produced trees that were much more imbalanced than those reconstructed using ML or MP when data were simulated under simple models of sequence evolution (Figure 4.4). Neighbor joining analyses from JC-corrected distances produce extremely biased topologies as substitution rates increase above 0.75 substitutions per site. This is consistent with our understanding of the problems with distance estimation because at such high rates of evolution, pairwise distance estimates can be undefined and lead to biased topological estimates (Hoyle and Higgs, 2003; Xia, 2006).

The results indicate that at high rates of substitution, phylogenetic methods produce tree topologies with greater imbalance. A similar pattern is observed when comparing the accuracy of the estimated trees with the rate of change. Figure 4.5 shows the proportion of error (RF/RF_{max}) of the estimated trees increasing as substitution rates increase. Perhaps a surprising result is that maximum parsimony (MP) produces more accurate estimates of topology than likelihood (ML) at high substitution rates under this simple model of sequence evolution. Similar results have been found from data simulated

on trees that satisfy a strict molecular clock and are analyzed by methods assuming an over-simplified substitution model (Rzhetsky and Sitnikova, 1996; Yang, 1997; Shavit et al., 2007). It is likely that many of the ultrametric model trees simulated in this study contain long branches that are joined as sister taxa. Because parsimony reconstruction tends to group long terminal branches together, this method is favored and appears to outperform ML and NJ (Bruno and Halpern, 1999). This property of parsimony reconstruction was described in detail by Swofford et al. (2001). A simulation study by Hulesenbeck and Lander (2003) showed that the probability that parsimony reconstruction is inconsistent (converges on the wrong topology as more data are added) becomes greater as rates of evolution increase and if the number of sequences examined is low. Their results were based on simulations from ERM trees that satisfied a strict molecular clock. Based on the conclusions of Huelsenbeck and Lander (2003), there is a non-zero probability that parsimony is inconsistent for the data sets simulated in my study. Theory suggests that if ML reconstruction was conducted under the assumption of a strict molecular clock, or much longer sequences were used, ML would be more accurate than MP. Moreover, in their examination of estimates of tree imbalance, Huelsenbeck and Kirkpatrick (1996) showed that UPGMA was the least biased method. The UPGMA method assumes a strict molecular clock and the results of Huelsenbeck and Kirkpatrick (1996) indicate that UPGMA outperformed other methods because the data met the assumptions of the model. Therefore, the observed performance of parsimony and under-parameterized model-based methods is not an indication these methods are superior, but instead, the data are simulated under conditions (ultrametric trees and small data set sizes) such that the true tree is favored by biased methods. The

simulations presented in this study are oversimplified and it is unlikely that such highly divergent biological sequences will satisfy a strict molecular clock and support a simple model of sequence evolution (JC). In fact, my results for data simulated under heterogeneous substitution models show a very different pattern.

When sequence data are generated under models that incorporate greater complexity in the substitution process, the degree of bias toward imbalanced topologies is reduced for ML and NJ (Figures 4.6 and 4.7). Figure 4.6 shows the imbalance, calculated using I_C, for trees simulated and analyzed under JC+G, K2P, HKY, and GTR (with outgroup A). In contrast to the pattern of imbalance in figure 4.4A, likelihood and distance methods are much less biased toward imbalanced trees. When comparing methods using *mean* I_w imbalance, NJ and MP are both much more biased than ML. Figure 4.8 shows the proportion of error for trees estimated from datasets generated under the K2P model. This can be compared with the patterns of imbalance observed in figure 4.7B. Imbalance measured by *mean* I_w shows a pattern that is similar to the error of the methods. Although, NJ appears to produce more accurate estimates of tree shape than MP, the topologies are actually less accurate. This discrepancy is due, in part, to how the trees are rooted and how polytomies are resolved. For the MP analyses, the method was forced to resolve zero-length branches and these resolutions translate to strong imbalance in the inferred topologies, particularly when using I_C to measure imbalance (Figure 4.6). At high substitution rates, methods that account for unobserved substitutions, are less likely to underestimate branch lengths and may also be less likely to include polytomies.

Parametric methods for phylogenetic reconstruction are known to be consistent when the model assumptions are satisfied (Chang, 1996). However, it is very important to consider the effect of model misspecification on the accuracy of these methods. Under-parameterization of a model can lead to decreased accuracy in estimates of topology, branch length, and parameter values. Though, in some cases, over-simplified models have been shown to produce more accurate estimates of tree topology than the true model (Yang, 1997), because such model violation can lead to biases in the method that favor the true tree (Bruno and Halpern, 1999; Swofford et al., 2001; Sullivan and Swofford, 2001). Several studies have shown that, under certain conditions, maximum likelihood estimation under an under-parameterized model can be an inconsistent method and prone to the same long-branch attraction effects as parsimony (Kuhner and Felsenstein, 1994; Yang et al., 1994; Yang, 1996; Bruno and Halpern, 1999; Philippe et al., 2005; Lartillot et al., 2007). For data generated under the HKY model, under-parameterization leads to an increase in estimates of tree imbalance (Figure 4.9) and reduced topological accuracy (Figure 4.10). When the data are analyzed under a very simple model (JC) the estimated tree topologies approach the levels of imbalance observed in parsimony reconstructions. However, when an over-parameterized model (GTR) is used, the imbalance in the estimated topologies is not significantly different from the estimates obtained from analyzing under the true model (HKY). In this case, the true model is a special case of the over-parameterized model and analyses under the true model and the overly complex model produce similar estimates of tree topology and imbalance (Figures 4.9 and 4.10). These results are consistent with our understanding of the effects of model misspecification on phylogenetic inference (Kuhner and Felsenstein, 1994; Lockhart et

al., 1996; Sullivan and Swofford, 2001; Lemmon and Moriarty, 2003; Brown and Lemmon, 2007). Moreover, with the increasing availability of genomic data, these results emphasize the importance of developing new models and phylogenetic methods that can better encompass the complex evolutionary processes responsible for generating multi-locus, biological data sets.

Another factor observed to influence estimates of phylogenetic tree imbalance is the number of characters included in an analysis. In their study, Huelsenbeck and Kirkpatrick (1996) simulated data sets comprised of sequences of 100 and 500 nucleotides in length. They found that all methods produced biased tree shapes regardless of the number of characters analyzed. Their comparison indicated that although increasing sequence length did not cure the bias toward imbalanced topologies, including more data resulted in an overall increase in topological accuracy. In the present study, I simulated additional data sets of 500 and 2000 bases under the JC (results not shown) and HKY models. Analyses of these data sets show that increasing sequence length improves tree imbalance estimates and leads to greater topological accuracy for all methods of reconstruction (Figure 4.11). However, it is important to state that, when dealing with biological data, increasing the number of characters is not straightforward. The simulations generated in this study satisfy the assumption that the data evolved under a single, stationary model and share a single evolutionary history and these results may not be general for large, multi-gene data sets.

If phylogenies are used to address questions about the timing, rate, or directionality of evolutionary processes, then rooting is a necessary step in analysis. For example, methods for assessing diversification rate variation and evaluating tree shape typically require rooted trees and many tree imbalance measures place a considerable amount of weight on asymmetry at the root of the tree. Including taxa from outside the group of interest is the most common method for rooting phylogenetic trees, and selection of these outgroup taxa is an important consideration for any phylogenetic analysis. Holland et al. (2003) used simulated data to demonstrate that highly divergent outgroup taxa can disrupt the topology of the ingroup or misplace the root of the tree because the long outgroup branch is joined to a long terminal branch in the ingroup. I assessed three different types of outgroup taxa: outgroup A was the same distance to the root of the tree as any given ingroup taxon, outgroup B was the ancestral sequence of the ingroup, and outgroup C was half the distance to the root of the tree as any given ingroup (Figure 4.2). Figure 4.12 depicts the relationship between the *mean* I_w imbalance and the substitution rate for trees generated from data sets simulated under K2P with each of the different outgroup taxa. These analyses indicate that methods that correct for unobserved substitutions (ML and NJ) are unaffected by the distance of the outgroup taxon to the root. Maximum parsimony (MP), however, becomes less biased toward imbalanced topologies as the outgroup branch is shortened. This result can also be seen when calculating imbalance using I_C (Figure 4.13). Colless's imbalance weights imbalance at the root of the tree much more than *mean* I_w imbalance. When measuring imbalance with I_C, MP appears to be biased toward more balanced topologies when using the ancestral sequence as an outgroup. In this case, parsimony favors placing the outgroup taxon closer

to the midpoint of the tree. When more distant outgroup taxa are used, however, long-branch attraction causes the outgroup to be placed with other long-branched terminal taxa. This is illustrated in Figure 4.14 for two trees estimated using parsimony. When outgroup taxon A is included in the analysis, long-branch attraction causes the outgroup and another long-branched terminal taxon (labeled IN and in blue) to be grouped together as sister taxa. This leads to an overall increase in tree imbalance, particularly when measured with I_C, because the tree is rooted on a single terminal branch (Figures 4.12 and 4.11). Conversely, when the ancestral sequence (outgroup B) is included, the root is placed more centrally on the tree, resulting in greater symmetry at the root (Figure 4.13). Additionally, for parsimony reconstruction, when a highly divergent outgroup sequence is selected, the resolution of the ingroup can be affected. Shorter outgroup branches have less impact on the reconstruction of the ingroup and, on average, inclusion of a shorter outgroup sequence resulted in greater accuracy. Typically, one should consider using more than one outgroup taxon to reduce the effects of long-branch attraction (Graybeal, 1998; Holland et al., 2003; Shavit et al., 2007). Maddison et al. (1984) found that single outgroup taxa were more likely to incorrectly root the tree. Inclusion of a monophyletic outgroup results in an overall reduction of long branches and can lead to more accurate reconstruction of the root of the tree (Smith, 1994).

The results of this simulation study demonstrate that the extreme bias toward imbalanced topologies, when trees are reconstructed from data generated under simple models, does not hold under a more general set of conditions (provided the assumptions of the model are satisfied). On average, inaccurate estimates of topology are imbalanced

and therefore, it is important to understand the sets of conditions that lead to inaccuracy for all methods. It is imperative to state, however, that the simulations in this study (and previous studies on tree shape bias) are very simple and do not adequately explore the parameter space occupied by biological data. For example, phylogenetic trees estimated from biological data are much more imbalanced than expected under the ERM model, and it is not clear from results based on ERM simulations whether methods are biased toward imbalanced topologies under a more realistic model of cladogenesis. When data are simulated on trees generated under variable speciation and extinction rates, inaccurate reconstructions are more imbalanced than the true trees (Figure 4.15). I observed the same pattern of increasing bias toward imbalanced trees with increasing substitution rates in non-ERM data sets as was observed when the ERM data sets were analyzed (Figure 4.15). These results indicate that even when the true tree is imbalanced, methodological bias toward greater asymmetry remains.

4.4 CONCLUSIONS

When using tree shape to address questions about macroevolutionary processes, it is important to consider the effects of non-biological factors which can lead to inaccurate estimates of tree topology and, in turn, inflate levels of tree imbalance. The results of this study have shown that different properties of the data and different inference methods can lead to elevated levels of phylogenetic tree imbalance. Based on these results, it is apparent that bias toward asymmetric topologies has contributed to the degree of imbalance observed in published phylogenies. However, previous surveys of published

phylogenies have failed to detect a difference in the imbalance of trees reconstructed using different phylogenetic methods (Heard, 1992; and see chapter 3). Therefore, an important next step in understanding tree shape bias will be to compare methods using a number of biological data sets. Although one would be unable to assess the accuracy of the methods using these data, tree shape measures would be useful for comparing topological estimates and detecting directional biases.

My results, and those of previous studies (Mooers et al., 1995; Heard and Mooers, 1996; Huelsenbeck and Kirkpatrick, 1996; Salisbury, 1999), have shown that reconstruction methods are inaccurate and biased toward imbalanced trees when model assumptions are violated or when the data are inundated with large amounts of homoplasy. Perhaps model violation is prevalent in empirical phylogenetic analyses and this model inadequacy is a possible explanation for the apparent lack of difference in imbalance estimates of biological trees produced by different phylogenetic methods. It is acknowledged that biological data are more complex than commonly used substitution models. Additionally, with the rapid accumulation of genomic data, phylogenetic analyses are being conducted with large, multi-locus data sets. The complexity of molecular data and our inability to model complex evolutionary histories underlines the necessity for new methods and biologically-realistic models that can better accommodate the heterogeneity we know is present in empirical data. Improvement in our ability to realistically model evolutionary processes will lead to fewer inaccurate estimates of phylogenetic relationships which can exacerbate phylogenetic tree imbalance.

ACKNOWLEDGEMENTS

Helpful comments and advice were given by the UT-IGERT discussion group, members of the Hillis/Bull/Cannatella lab groups, John Huelsenbeck, Derrick Zwickl, and Erick Matsen. This work was funded by a graduate research traineeship provided by an NSF IGERT grant in Computational Phylogenetics and Applications to Biology awarded to the University of Texas, Austin. Computational resources were provided by the Center for Computational Biology and Bioinformatics and the Texas Advanced Computing Center (TACC) at The University of Texas at Austin (http://www.tacc.utexas.edu).

TABLE 4.1: The range of conditions addressed by simulating and analyzing different data sets. The variables considered include the model tree shape (simulation tree type), the substitution model for simulation, the model for analysis, the outgroup, and the length of the sequences. Every set of conditions was simulated on every tree for each of the 12 sets of trees that ranged in tree depth (12,000). Only two cases involved model misspecification, where the simulation model was HKY and was analyzed under an under-parameterized model (JC) and an over-parameterized model (GTR).

Simulation Tree Type	**Simulation Model**	**Analysis Model** T = true	**Outgroup**	**Sequence length (# nucleotides)**
ERM	JC	T	A	500
				1000
				2000
		T	B	500
				1000
				2000
		T	C	1000
ERM	JC+G	T	A	1000
		T	B	1000
ERM	K2P	T	A	1000
		T	B	1000
		T	C	1000
ERM	HKY	T	A	500
				1000
				2000
		JC	A	1000
		GTR	A	1000
		T	B	500
				1000
				2000
ERM	GTR	T	A	1000
			B	1000
Variable Rates	JC	T	A	1000

FIGURE 4.1: Two examples of constant-rate birth/death tree topologies. The branch lengths correspond to the divergence times. The simulation parameters were set so that $\lambda = 0.4$ (birth rate) and $\mu = 0.3$ (death rate).

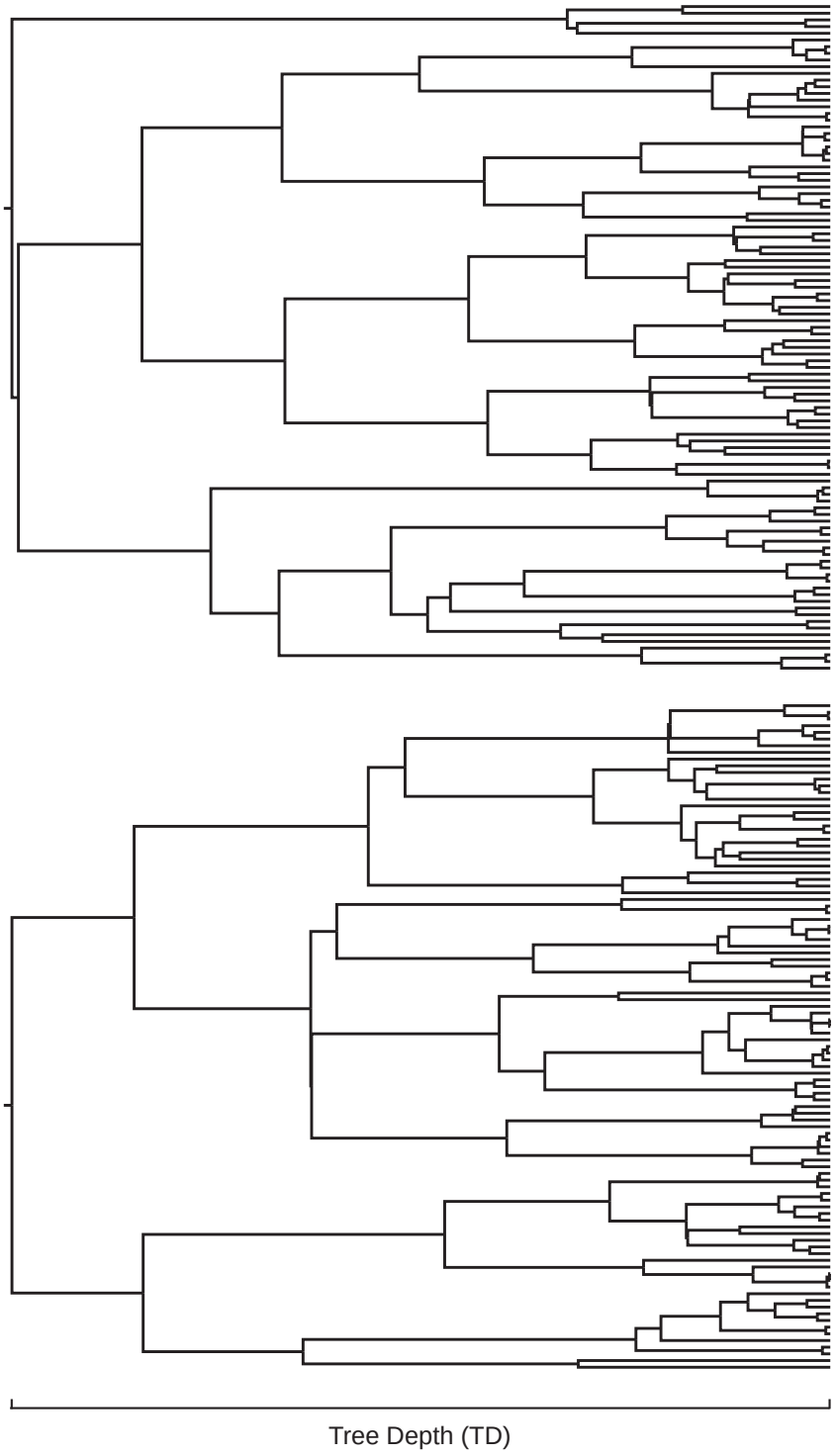


FIGURE 4.2: The three different outgroup branch lengths considered for this simulation study. The distance (*d*) from the root of the tree to the tip of the outgroup lineage was varied from so that one outgroup (outgroup A) was the same distance to the root as any given ingroup taxon (*d = TD*), the second outgroup (outgroup B) was the ancestral sequence of the ingroup (*d = 0*), and the length of the third outgroup branch (outgroup C) was half of the total tree depth (*d = 0.5 * TD*).

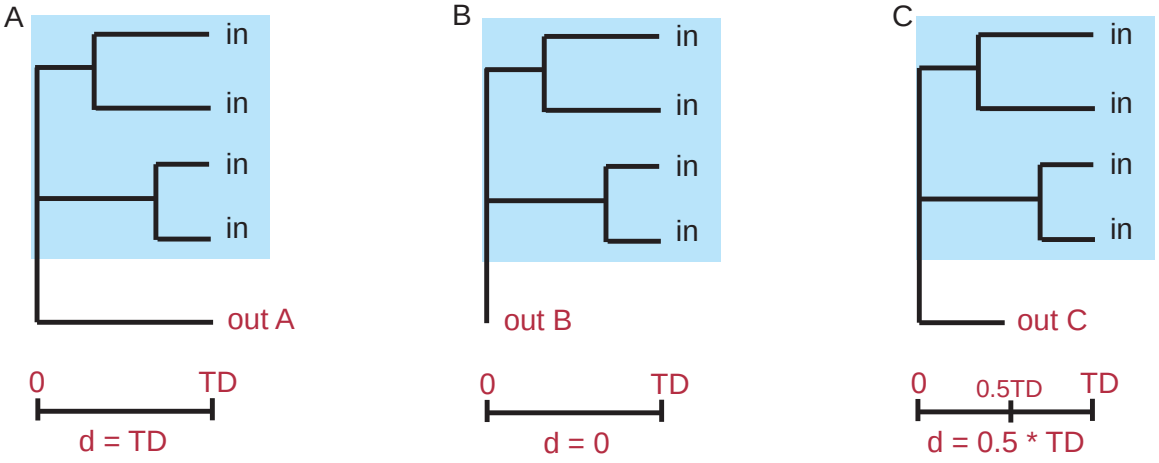


FIGURE 4.3: Two examples of trees simulated under variable rates of speciation and extinction. The gamma shape parameter that controls the level of rate variation was set to 3 (see chapter 2). The branch lengths correspond to the divergence times.

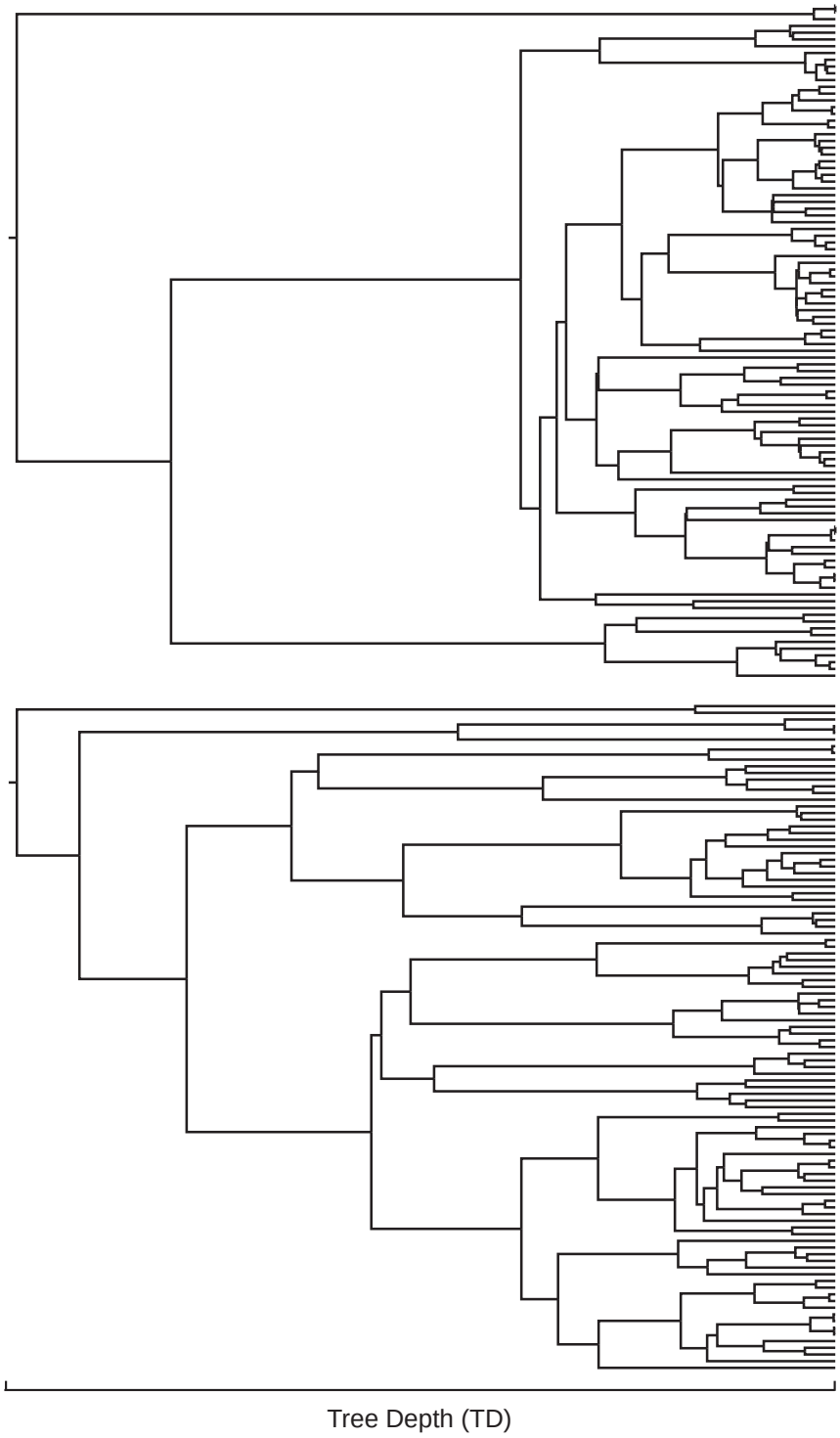


FIGURE 4.4: Tree imbalance/balance as a function of substitution rate for four different tree imbalance measures. Tree shape was measured for each estimated tree using (A) I_C, (B) *N-bar*, (C) *B2*, and (D) *mean* I_w and averaged across all topologies reconstructed from data sets simulated under a given substitution rate (0.05, 0.1, 0.15, 0.2, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, and 2.0 substitutions/site) for each of the three methods with outgroup taxon A. Neighbor joining (NJ) trees are indicated by blue lines, maximum parsimony (MP) trees are depicted using black lines, and maximum likelihood (ML) trees are shown with red lines. The average values for the true (ERM) trees are indicated by the grey dotted lines. For substitution rates above 1.25, *mean* I_w could not be calculated for NJ trees because of high numbers of undefined distances or zero-length branches.

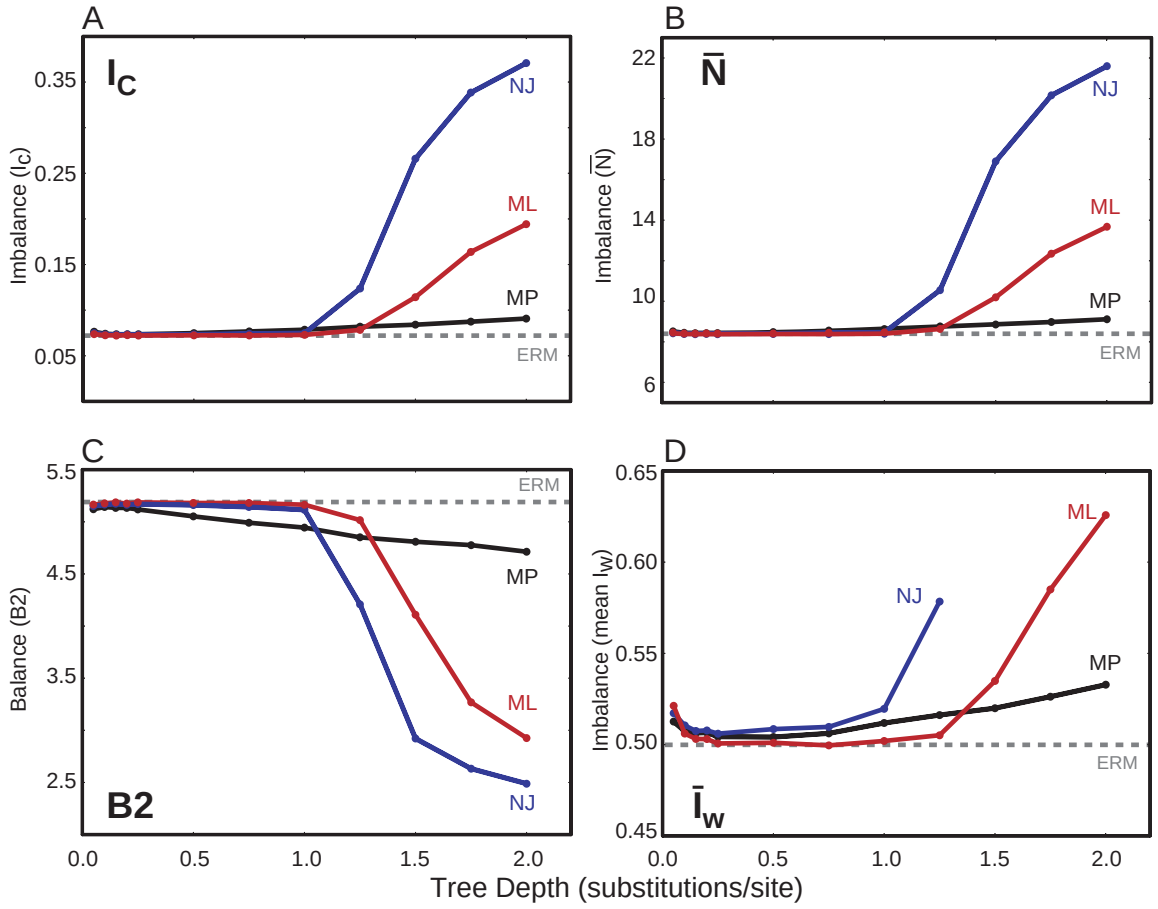


FIGURE 4.5: The proportion error (or absolute error) of reconstructed topologies as a function of substitution rate calculated using Robinson-Foulds distances and normalized by the maximum RF distance. The error is shown for trees reconstructed using NJ (blue), MP (black), and ML (red), with outgroup taxon A.

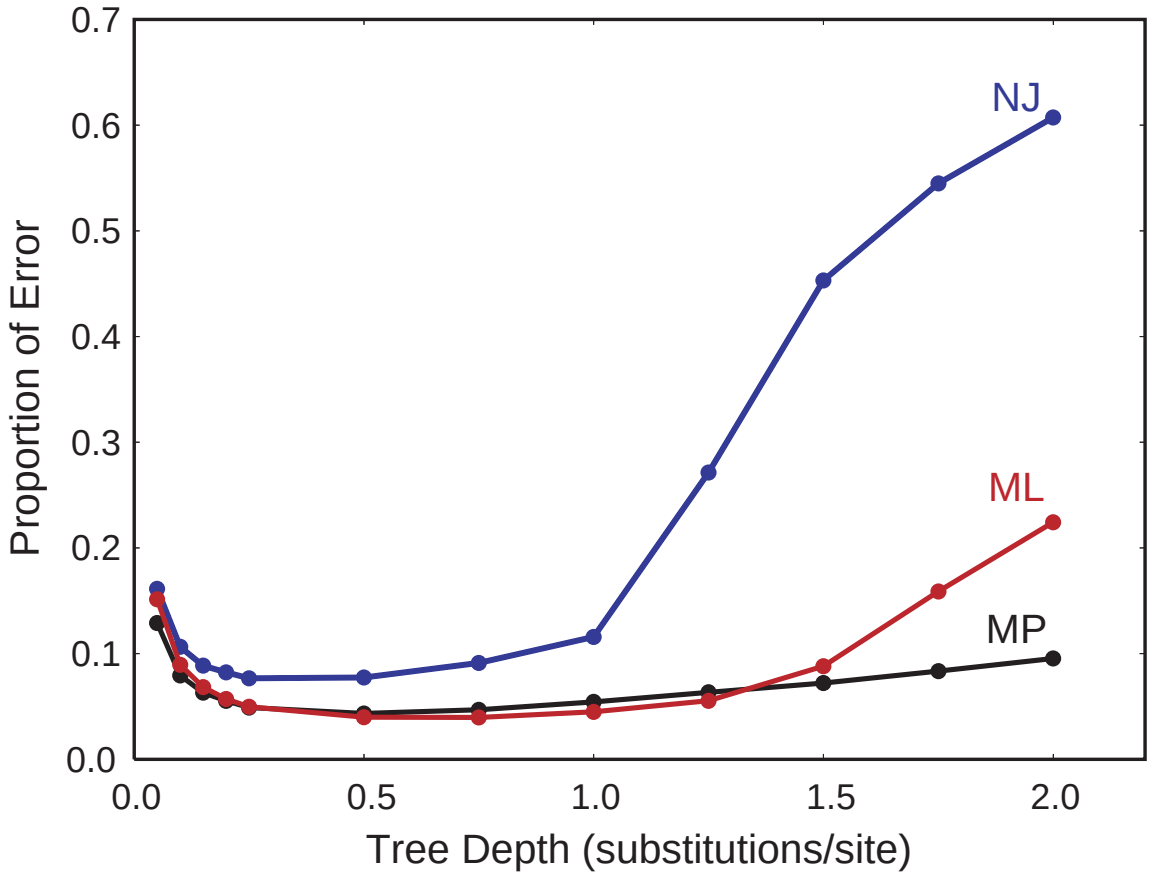


FIGURE 4.6: Colless's imbalance (I_C) for data simulated and analyzed under heterogeneous models. The imbalance is shown for data sets simulated under (A) the JC model with gamma-distributed rate heterogeneity (JC+G), (B) K2P, (C) HKY, and (D) GTR, with outgroup taxon A.

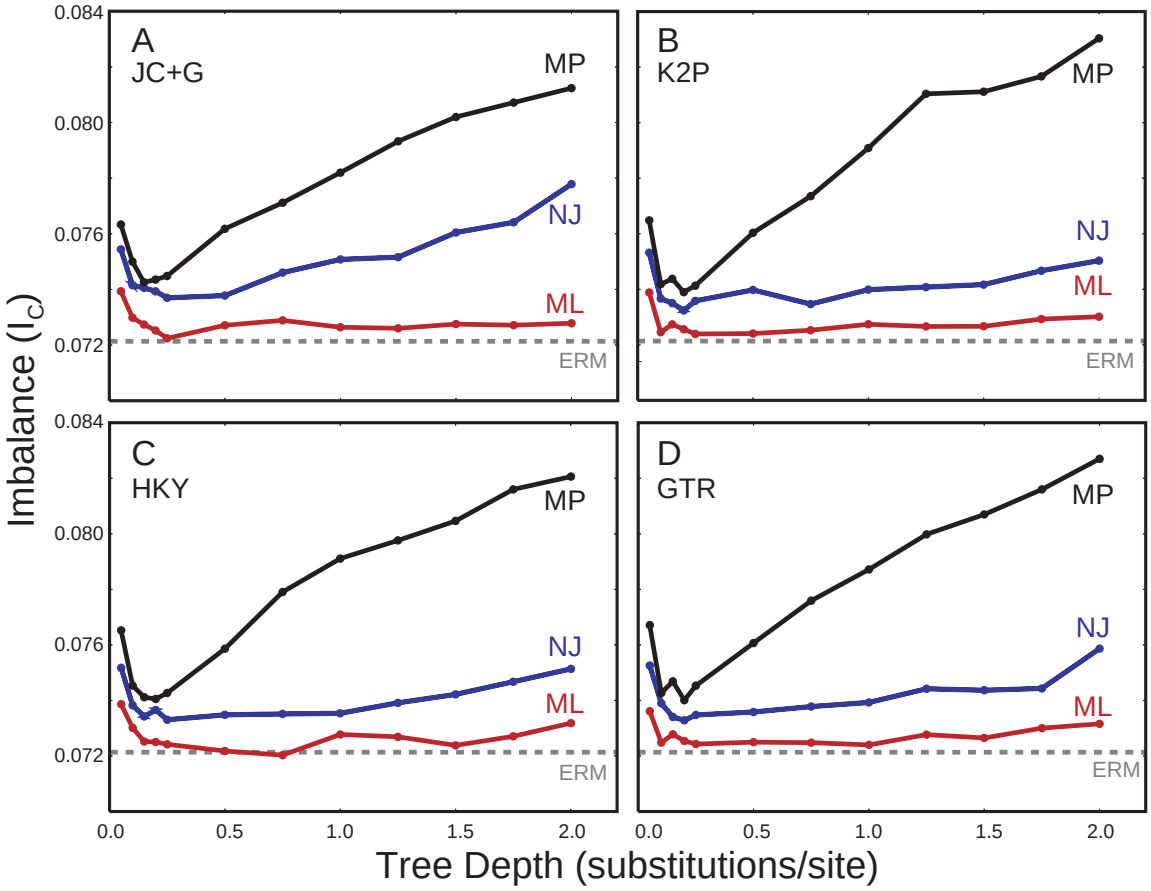


FIGURE 4.7: *Mean* I_w imbalance for data simulated and analyzed under heterogeneous models. The imbalance is shown for data sets simulated under (A) the JC model with gamma-distributed rate heterogeneity (JC+G), (B) K2P, (C) HKY, and (D) GTR, with outgroup taxon A.

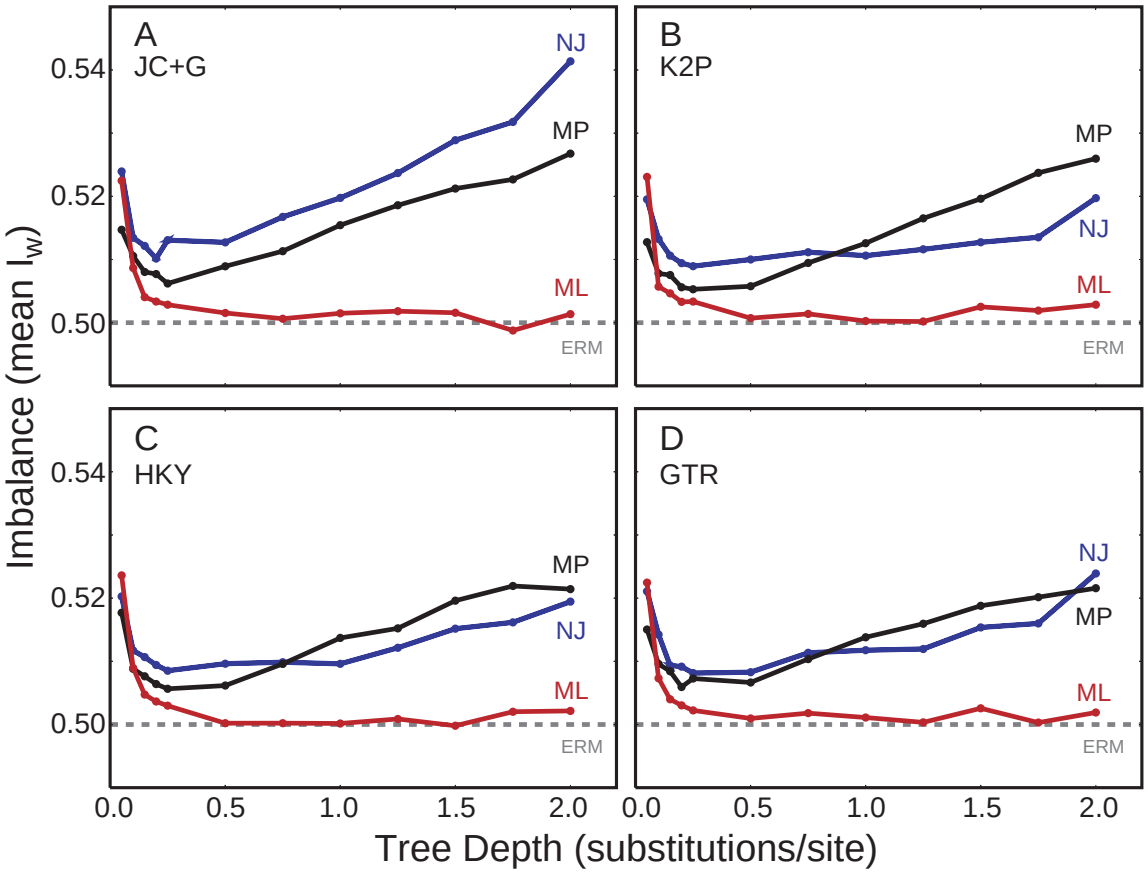


FIGURE 4.8: The proportion of error (absolute error) for trees reconstructed from data sets simulated under the K2P model. NJ and ML analyses were conducted using the same model for analysis that was used for simulation (K2P) and outgroup taxon A was included in the analysis to root the tree.

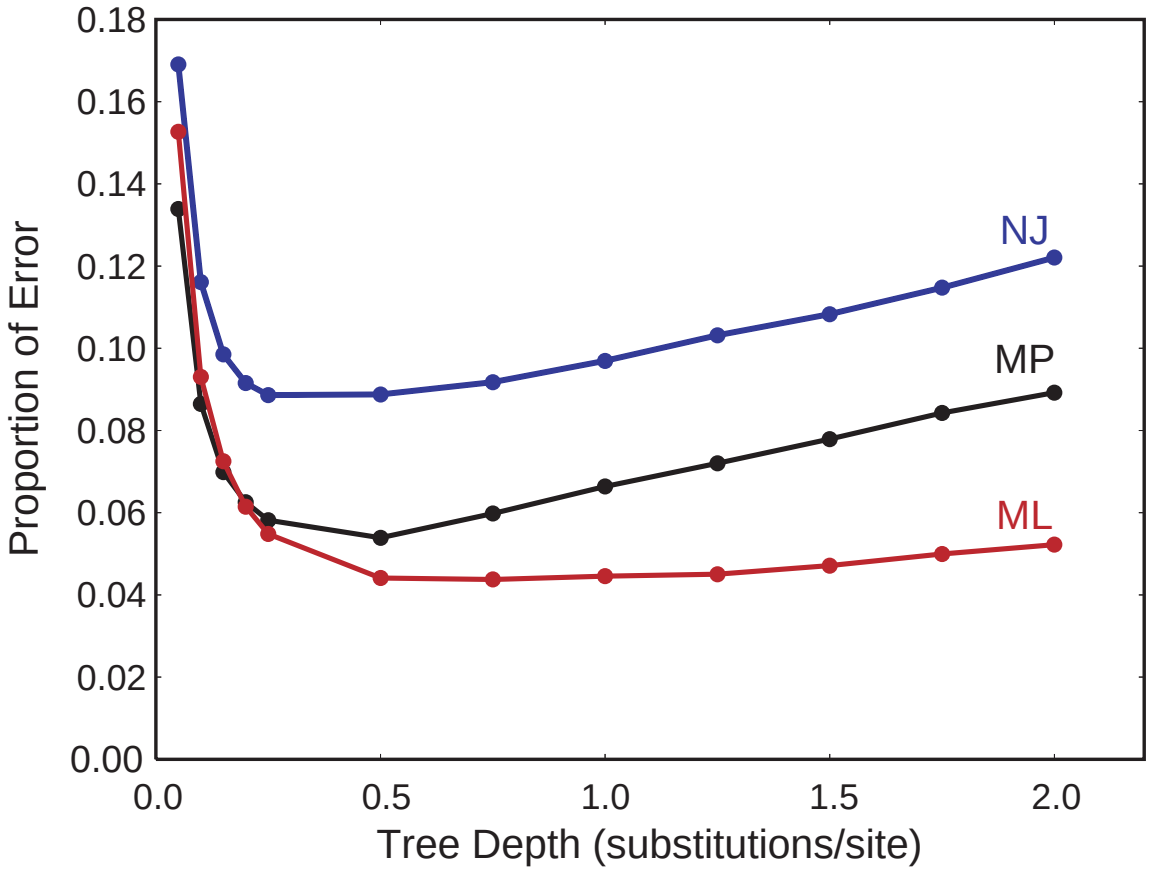


FIGURE 4.9: The *mean* I_w imbalance for trees reconstructed from data sets simulated under HKY and analyzed using parsimony (MP) and misspecified models using maximum likelihood (ML). Outgroup taxon A was included in the analyses for rooting. When the true model (ML/HKY solid-red line) or an over-parameterized model (ML/GTR dotted-red line) is assumed, no difference in the estimate of imbalance was observed. When the analysis assumes an under-parameterized model (ML/JC dashed red line), the estimated trees are more imbalanced.

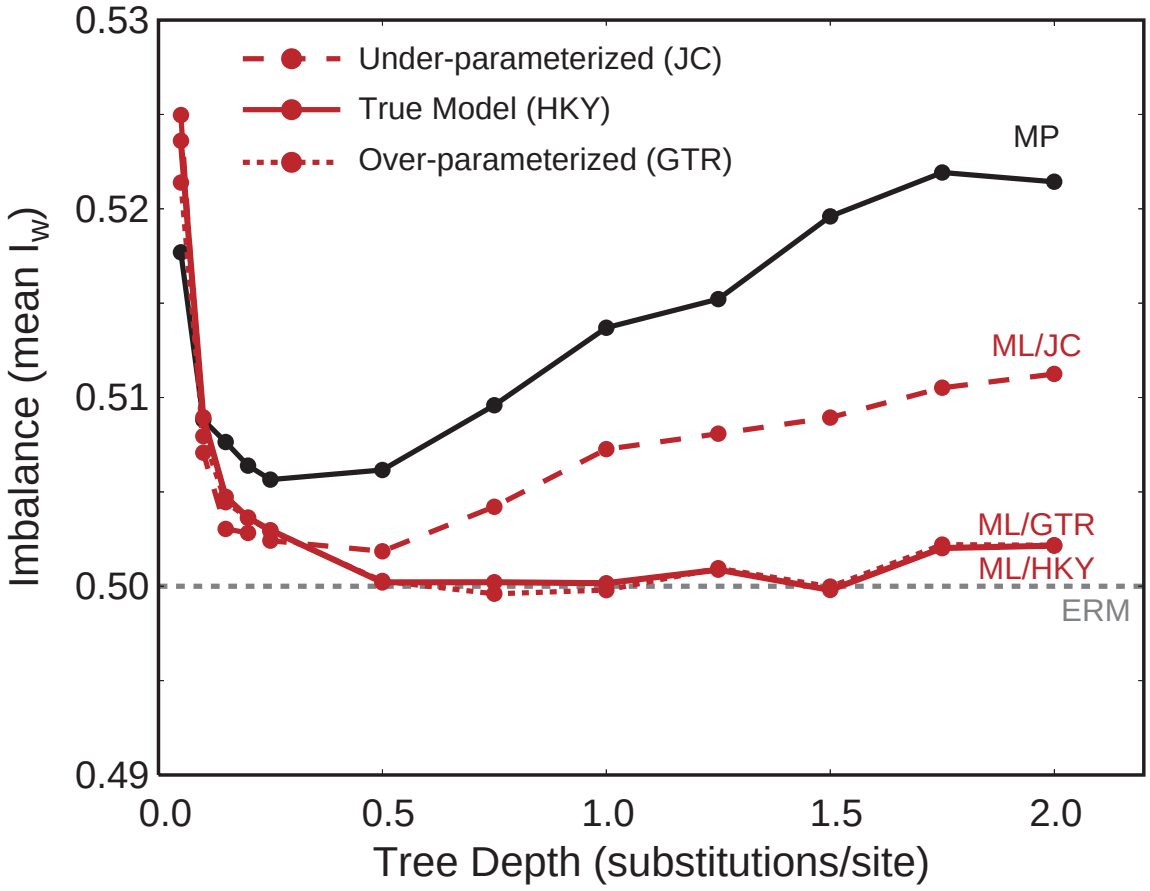


FIGURE 4.10: The proportion of error for trees estimated under MP (black line) and ML (red lines) with misspecified models. The data sets analyzed were simulated under the HKY model with outgroup taxon A. ML reconstruction under the under-parameterized model (JC) is represented by the dashed red line and is less accurate than trees estimated under the true model (HKY, solid red line) or the over-parameterized model (GTR, dotted red line). The true/simulation model is a special case of the excessively complex model, and as a result, there is no difference in the average error of estimated topologies (in this case).

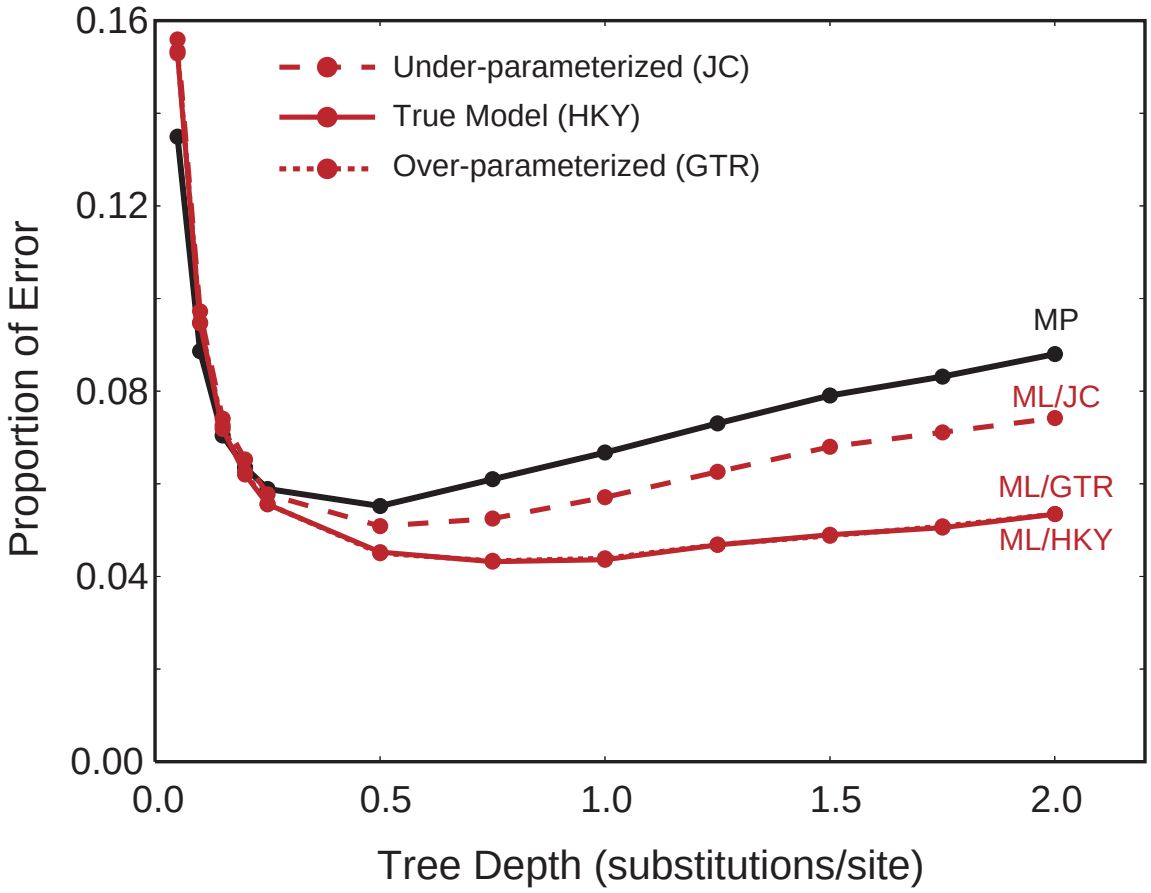


FIGURE 4.11: The imbalance (measured using I_C) of trees estimated from data sets simulated under HKY with a range of sequence lengths using each of the three methods. Data sets with 500 bases are shown using dashed lines, the solid lines represent the data sets with 1000 bases, and the dotted lines indicate the trees estimated from data sets contain 2000 nucleotides. Outgroup taxon A was used in the analyses to root the trees.

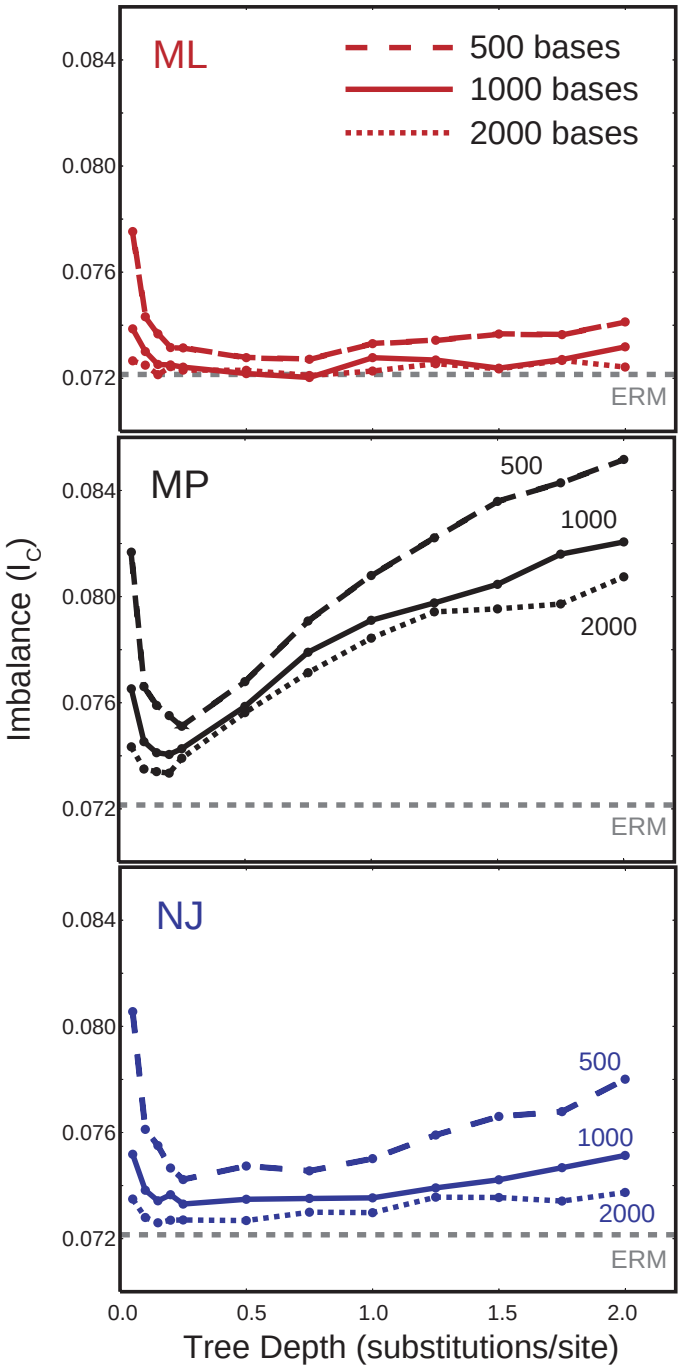


FIGURE 4.12: The effect of outgroup branch length on tree imbalance. *Mean* I_w imbalance of trees reconstructed from data sets with one of three outgroup taxa. Outgroup A (solid line) had a branch length equal to the total tree depth, outgroup B (dashed line) was the ancestral sequence, and the length of outgroup C (dotted line) was half of the total tree depth. These results indicated that for maximum parsimony (MP) reconstruction, the topology of the ingroup is greatly affected by the length of the outgroup branch. NJ and ML analyses appear to be unaffected by the selection of the outgroup.

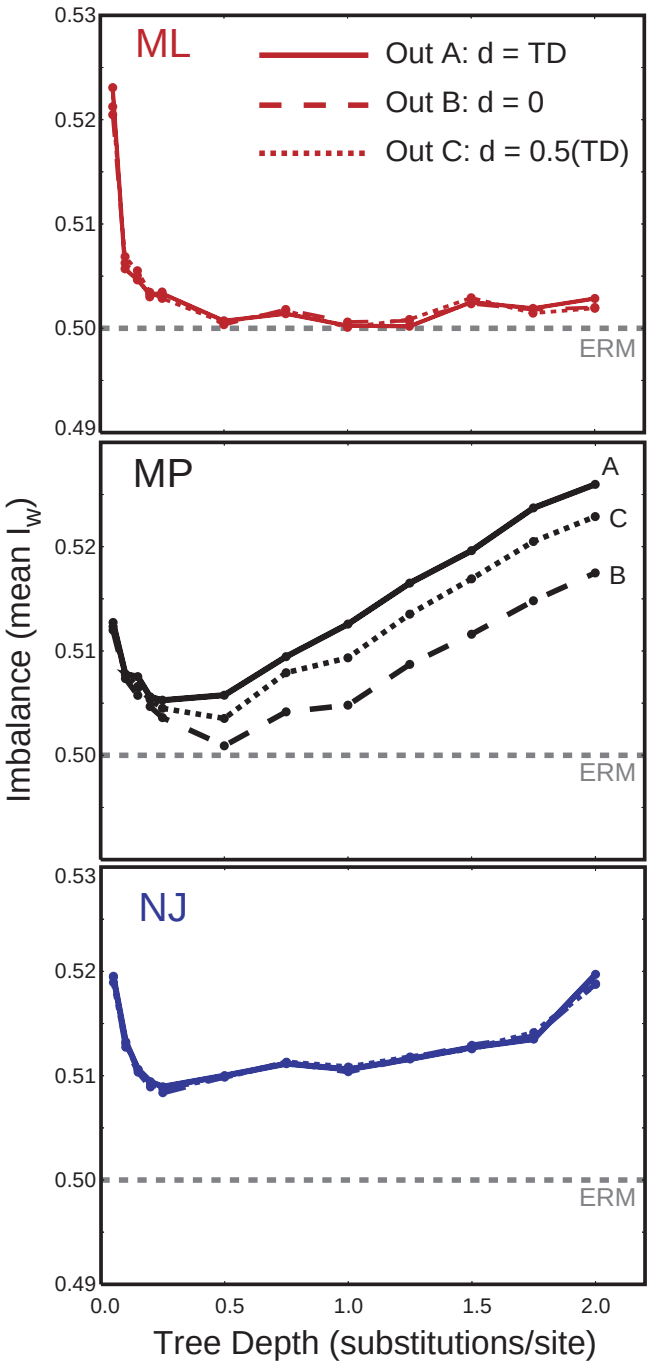


FIGURE 4.13: The effect of outgroup branch length on imbalance (measured using I_C). Colless's imbalance places greater weight on the imbalance at the root of the tree. When the ancestral sequence is used as an outgroup, there is a tendency for that taxon to be placed in a more central position on the tree when using parsimony.

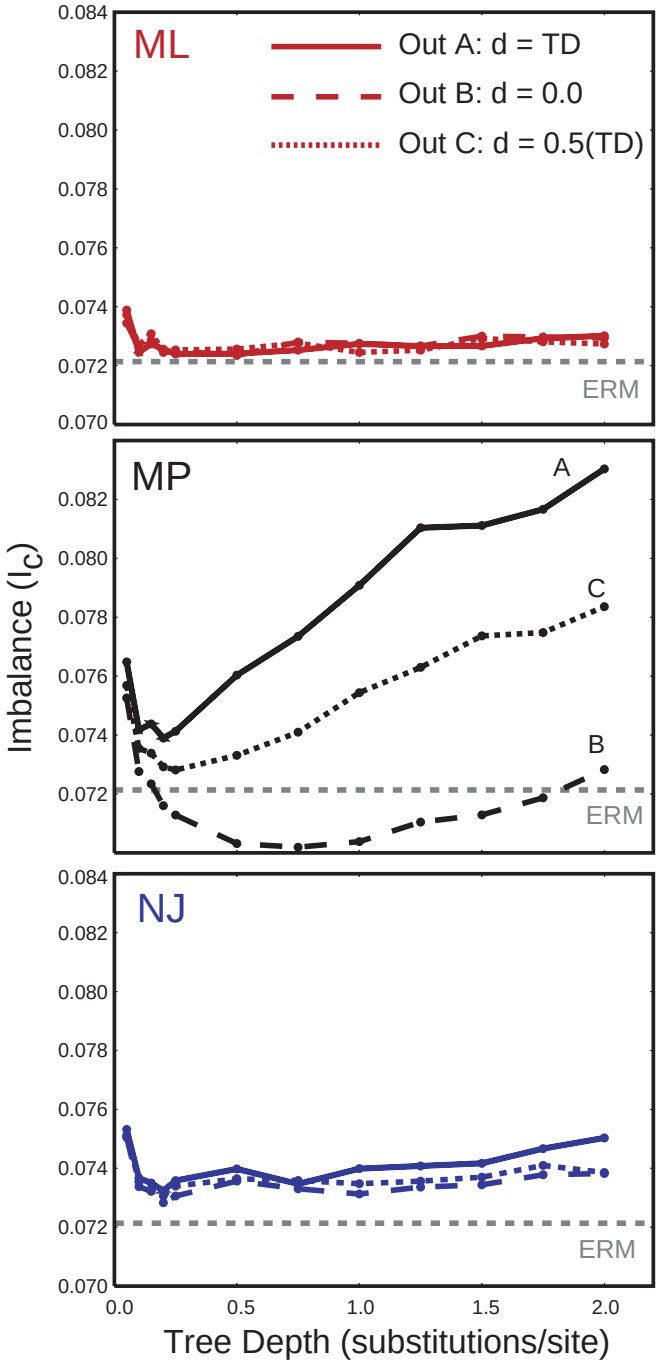


FIGURE 4.14: An example of the effect of outgroup branch length on the ingroup topology under the maximum parsimony (MP) optimality criterion. The unrooted trees were reconstructed from the same simulated data sets with different outgroups (OUT; indicated in red). The length of the branch leading to outgroup A was equal to the total tree depth, which in this case was 0.75 substitutions/site ($d_A = 0.75$). Outgroup B was the ancestral sequence of the ingroup ($d_B = 0.0$). When an outgroup with a long branch is used, there is a greater chance the topology of the ingroup will be affected if there are ingroup taxa on long branches. In this example, homoplasy in the outgroup sequence and an ingroup sequence on a long branch (IN; colored in blue) causes the two taxa to be drawn together because of long-branch attraction. When the ancestral sequence (outgroup B) is used, parsimony is more likely to accurately reconstruct the root of the tree and the outgroup is less likely to affect the topology of the ingroup.

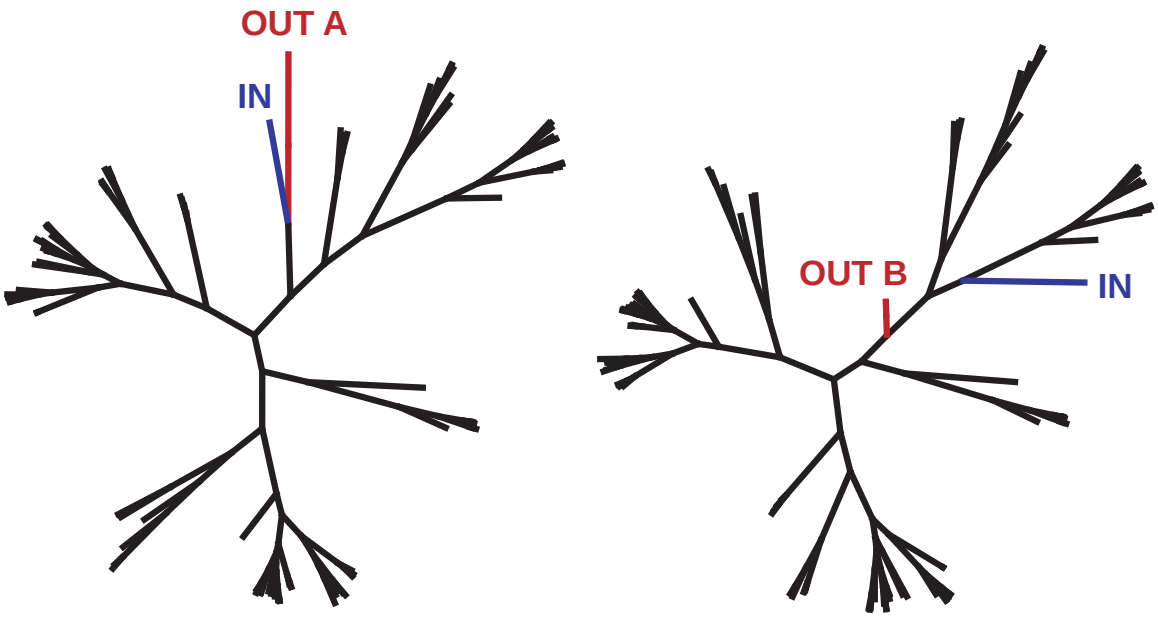
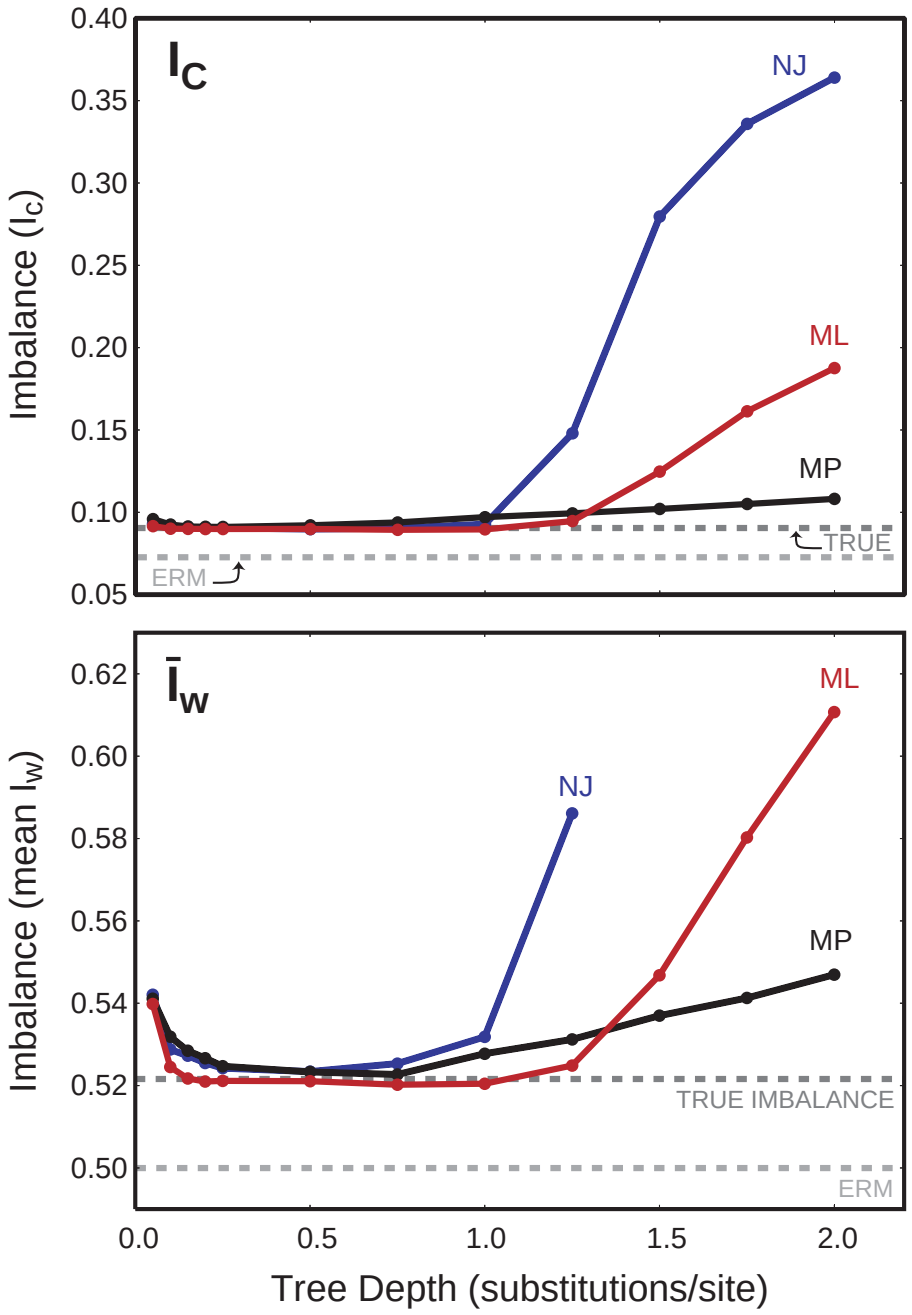


FIGURE 4.15: Imbalance of trees reconstructed from data sets simulated on non-ERM trees, calculated using I_C and *mean* I_w imbalance. The model trees generated under variable rates of diversification had an average I_C imbalance of 0.09 and an average *mean* I_w imbalance of 0.52 and are indicated by the dark grey, dashed lines. The average imbalance expected under constant rates of speciation and extinction for each measure (ERM I_C = *0.072*; ERM *mean* I_w = *0.5*) are represented by the light grey, dashed lines. A similar pattern of increasing imbalance with increasing substitution rates was observed for these analyses.



Appendix A: Simulation Model Parameters

TABLE A.1: Model parameters estimated from mammalian gene sequences (Murphy et al., 2001). The various model parameters were used to simulated data sets used in chapters 2 and 3. Parameter values for the simulation models (JC+G, K2P, HKY, GTR, GTR+I) were selected from these estimates.

		Base Frequencies				Relative substitution rates						Proportion invariant	
Gene	Preferred Model	A	C	G	T	AC	AG	AT	CG	CT	GT	sites	Alpha
ADORA3	*K2P*	0.25	0.25	0.25	0.25	1	3	1	1	3	1	0	
ZFX	*HKY+I+G*	0.35	0.23	0.18	0.23	1	7.94	1	1	7.94	1	0.49	1.24
mtRNA	*GTR+I+G*	0.34	0.2	0.21	0.25	5.86	14	3.85	0.58	29.3	1	0.41	0.53

Appendix B: Collection of Published Phylogenies and References

TABLE B.1: Taxonomic groups and references for empirical phylogenies. ML = maximum likelihood, B = Bayesian, P = maximum parsimony

Group	Method	Citation
"PROTISTS"		
Dinoflagellates	ML	Murray et al. (2005)
Euglenozoa	ML	Simpson and Roger (2004)
Phaeophyceae	ML/P	Yoon et al. (2001)
FUNGI		
Ascomycota	ML/P	Kauff and Lutzoni (2002)
Cortinarius	B	Froslev et al. (2005)
Homobasidiomycetes	P, ML	Binder and Hibbett (2002), Bodensteiner et al. (2004)
Lichenicolous fungi	B	Lawrey et al. (2007)
Pertusariaceae	B	Schmitt and Lumbsch (2004)
Sordariales	P	Miller and Andrew (2005)
PLANTS		
Anthemideae	P	Watson et al. (2000)
Araceae	P	Lewis and Doyle (2002)
Aristolochiaceae	ML/P	Neinhuis et al. (2005)
Begonia	P	Forrest et al. (2005)
Bryophyta	B	Shaw and Renzaglia (2004)
Burseraceae	P	Weeks et al. (2005)
Calamoideae	P	Baker et al. (2000)
Cardueae	ML/P	Garcia-Jacas et al. (2002)
Cariceae	P	Starr et al. (2004
Caryophyllaceae	B/P	Fior et al. (2006)
Cornales	ML/P	Xiang et al. (2002
Croton	ML/P	Berry et al. (2005)
Cucurbitales	P	Zhang et al. (2006)
Epidendroideae	B	van den Berg et al. (2005)
Eucalyptus	P	Steane et al. (1999)
Gentianaceae	P	Yuan et al. (2003)
Houstonia	ML	Church (2003)
Lamiales	P	Wortley et al. (2005)
Leguminosae	B	Lavin et al. (2005)
Lythraceae	B/P	Graham et al. (2005)
Malaxideae	P	Cameron (2005)
Malvatheca	B/P	Baum et al. (2002)
Marchantiidae	P	Boisselier-Dubayle et al. (2002)

Marchantiophyta	B	Shaw and Renzaglia (2004)
Monilophytes	B	Pryer et al. (2004)
Myrtaceae	P	Wilson et al. (2005)
Ocimeae	B/P	Paton et al. (2004)
Orchidaceae	P	Cameron (2004)
Phyllanthaceae	P	Kathriarachchi et al. (2005)
Physematieae	P	Sano et al. (2000)
Poaceae	MRP	Salamin et al. (2002)
Rhododendron	ML/P	Goetsch et al. (2005)
Sapindaceae	P	Harrington et al. (2005)
Saxifragales	ML/P	Fishbein et al. (2001)
Solanaceae	B	Martins and Barkman (2005)
Vernonieae	B	Keeley et al. (2007)
Vitaceae	P	Soejima and Wen (2006)

ANIMALS

ARTHROPODS

Agromyzidae	P	Scheffer et al. (2007)
Anostraca	ML	Weekers et al. (2002)
Aphididae	P, ML	Ortiz-Rivas et al. (2004), von Dohlen et al. (2006)
Aphodiini	B	Cabrero-Sanudo and Zardoya (2004)
Arthropoda	ML	Pisani (2004)
Asilidae	ML/P	Bybee et al. (2004)
Avenzoariinae	P	Dabert et al. (2002)
Bactrocera	P	Smith et al. (2003)
Braconidae	B/P	Shi et al. (2005)
Branchiopoda	B/P	deWaard et al. (2006)
Carabidae	ML/P	Ober (2002)
Ceratopogonidae	P	Bekenbach and Borkent (2003)
Cheilosia	P	Stahls and Nyblom (2000)
Cicadomorpha	P	Cryan (2005)
Cicindela	P	Pons et al. (2004)
Coccoidea	ML/P	Cook et al. (2002)
Cotesia	ML/P	Kankare and Shaw (2004)
Curculionoidea	P	Marvaldi et al. (2002)
Dermaptera	P	Jarvis et al. (2005)
Drosophilidae	P	Remsen and O'Grady (2002)
Ephemeroptera	P	Ogden and Whiting (2005)
Euglossini	P	Michel-Salzat et al. (2005)
Eumolpinae	P	Gomez-Zurita et al. (2005)
Eurytomidae	B	Chen et al. (2004)
Formicidae	P	Astruc et al. (2004)
Formicinae	ML/P	Johnson et al. (2003)
Gelechioidea	P	Bucheli and Wenzel (2005)
Geometridae	P	Abraham et al. (2005)

Gerromorpha	P	Damgaard et al. (2005)
Harpalini	ML/P	Martinez-Navarro et al. (2005)
Insecta	B	Kjer (2004)
Membracoidea	P	Dietrich et al. (2001)
Microgastrinae	ML/P	Mardulyn and Whitfield (1999)
Mysida	ML/P	Remerie et al. (2004)
Noctuoidea	ML/P	Fang et al. (2000)
Nymphalidae	P	Freitas and Brown (2004)
Orthoptera	B/P	Jost and Shaw (2006)
Papilionini	ML/P	Aubert et al. (1999)
Pentatomomorpha	ML/P	Li et al. (2005)
Pholcidae	ML/P	Bruvo-Madaric et al. (2005)
Phyllopoda	ML/P	Braband et al. (2002)
Pipunculidae	P	Skevington and Yeats (2000)
Pycnogonida	ML/P	Arango (2003)
Salticidae	ML/P	Maddison and Hedin (2003)
Simuliidae	ML/P	Pruess et al. (2000)
Staphyliniformia	ML/P	Caterino et al. (2005)
Syrphidae	P	Skevington and Yeats (2000)
Tephritoidea	ML/P	Han and Ro (2005)
Therevidae	P	Yang et al. (2000)
Theridiidae	B/P	Arnedo et al. (2004)
Trichoptera	ML/P	Kjer et al. (2001)
VERTEBRATES		
Amniota	P	Hill (2005)
Amphibia	ML	Zhang et al. (2005)
Anura	ML	Roelants and Bossuyt (2005)
Carcharhiniformes	P	Iglesias et al. (2005)
Caudata	B/P	Weisrock et al. (2005)
Chiroptera	ML/MRP	Teeling et al. (2005), Jones et al. (2002)
Chondrichthyes	ML	Douady et al. (2003)
Clupeiforms	B	Li and Orti (2007)
Colubroidea	ML/P	Lawson et al. (2005)
Cyprinidae	ML/P	Wang and He (2007)
Elapidae	ML/P	Slowinski and Keogh (2000)
Elopomorpha	P, B	Obermiller and Pfeiler (2003), Inoue et al. (2004)
Emberizidae	ML/P	Carson and Spicer (2003)
Euteleostei	ML/P	Ishiguro et al. (2003)
Eutheria	B/MRP	Murphy et al. (2001), Beck et al. (2006)
Gobioidei	P	Thacker (2003)
Gymnophiona	ML	San Mauro et al. (2004)
Hyloidea	ML/P	Darst and Cannatella (2004)
Lygosominae	P, ML	Honda et al. (2000), Reeder (2003)
Mammalia	ML	Phillips and Penny (2003)

Marsupialia	P, ML	Palma and Spotorno (1999), Amrine-Madsen et al. (2003)
Muridae	ML/P	Jansa and Weksler (2004)
Myliobatiformes	ML/P	Dunn et al. (2003)
Osteoglossomorpha	ML/P	Lavoue and Sullivan (2004)
Passeriformes	B/P	Spicer and Dunipace (2004)
Reptilia (w/ birds)	ML	Rest et al. (2003)
Rodentia	ML/P	Adkins et al. (2003)
Scincidae	B	Brandly et al. (2005)
Serpentes	B/P	Slowinski and Lawson (2002)
Sigmodontinae	ML	Weksler (2003)
Siluriformes	B/P	Hardman (2005)
Squamata	ML/P	Townsend et al. (2004)
Sylvioidea	B	Alstrom et al. (2006)
Testudines	ML/P	Krenz et al. (2005)
Tyranni	ML/P	Chesser (2004)
OTHER ANIMAL GROUPS		
Acanthocephala	P, ML	Near et al. (1998), Garcia-Varela et al. (2002)
Acoela	P	Hooge et al. (2002)
Anthozoa	ML	Berntson et al. (1999)
Brachiopods	ML/P	Saito et al. (2000)
Calcarea	ML	Manuel et al. (2003)
Cestoda	P	Olson et al. (2001)
Conus	ML	Cunha et al. (2005)
Demospongiae	B	Nichols (2005)
Gastrotricha	P	Wirz et al. (1999)
Mollusca	ML/P	Passamaneck et al. (2004)
Mytilidae	ML/P	Distel (2000)
Nematoda	B/P	Meldal et al. (2007), Smythe et al. (2006)
Nemertea	ML/P	Sundberg and Saur (1998)
Octocorallia	B/P	McFadden et al. (2006)
Opisthobranchia	ML/P	Grande et al. (2004)
Pectinidae	ML/P	Barucca et al. (2004)
Platyhelminthes	ML	Campos et al. (1998)
Polychaeta	ML/P	Bleidorn et al. (2003)
Proseriata	P	Littlewood et al. (2000)
Rotifera	ML/P	Sorenson and Giribet (2006)
Scleractinia	B	Le Goff-Vitry et al. (2004)
Spatangoida	B/P	Stockley et al. (2005)
Tubificidae	B/P	Sjolin et al. (2005)
Tunicata	ML/P	Stach and Turbeville (2002)
Venerinae	B/P	Kappner and Bieler (2006)

REFERENCES FOR COLLECTED PUBLISHED PHYLOGENIES

Abraham, D., N. Ryrholm, H. Wittzell, J. D. Holloway, M. J. Scoble, and C. Lofstedt. 2001. Molecular phylogeny of the subfamilies in geometridae (Geometroidea : Lepidoptera). *Molecular Phylogenetics and Evolution* 20:65-77.

Adkins, R. M., A. H. Walton, and R. L. Honeycutt. 2003. Higher-level systematics of rodents and divergence time estimates based on two congruent nuclear genes. *Molecular Phylogenetics and Evolution* 26:409-420.

Alstrom, P., P. G. Ericson, U. Olsson, and P. Sundberg. 2006. Phylogeny and classification of the avian superfamily Sylvioidea. *Molecular Phylogenetics and Evolution* 38:381-97.

Amrine-Madsen, H., M. Scally, M. Westerman, M. J. Stanhope, C. Krajewski, and M. S. Springer. 2003. Nuclear gene sequences provide evidence for the monophyly of australidelphian marsupials. *Molecular Phylogenetics and Evolution* 28:186-196.

Arango, C. P. 2003. Molecular approach to the phylogenetics of sea spiders (Arthropoda : Pycnogonida) using partial sequences of nuclear ribosomal DNA. *Molecular Phylogenetics and Evolution* 28:588-600.

Arnedo, M. A., J. Coddington, I. Agnarsson, and R. G. Gillespie. 2004. From a comb to a tree: Phylogenetic relationships of the comb-footed spiders (Araneae, Theridiidae) inferred from nuclear and mitochondrial genes. *Molecular Phylogenetics and Evolution* 31:225-245.

Astruc, C., I. Julien, C. Errard, and A. Lenoir. 2004. Phylogeny of ants (Formicidae) based on morphology and DNA sequence data. *Molecular Phylogenetics and Evolution* 31:880-893.

Aubert, J., L. Legal, H. Descimon, and F. Michel. 1999. Molecular phylogeny of swallowtail butterflies of the tribe Papilionini (Papilionidae, Lepidoptera). *Molecular Phylogenetics and Evolution* 12:156-167.

Baker, W. J., T. A. Hedderson, and J. Dransfield. 2000. Molecular phylogenetics of subfamily Calamoideae (Palmae) based on nrDNA ITS and cpDNA rps16 intron sequence data. *Molecular Phylogenetics and Evolution* 14:195-217.

Barucca, M., E. Olmo, S. Schiaparelli, and A. Canapa. 2004. Molecular phylogeny of the family Pectinidae (Mollusca : Bivalvia) based on mitochondrial 16S and 12S rRNA genes. *Molecular Phylogenetics and Evolution* 31:89-95.

Baum, D. A., S. D. Smith, A. Yen, W. S. Alverson, R. Nyffeler, B. A. Whitlock, and R. L. Oldham. 2004. Phylogenetic relationships of Malvatheca (Bombacoideae and Malvoideae; Malvaceae Sensu Lato) as inferred from plastid DNA sequences. *American Journal of Botany* 91:1863-1871.

Beck, R. M., O. R. Bininda-Emonds, M. Cardillo, F. G. Liu, and A. Purvis. 2006. A higher-level MRP supertree of placental mammals. *BMC Evolutionary Biology* 6:93.

Beckenbach, A. T., and A. Borkent. 2003. Molecular analysis of the biting midges (Diptera : Ceratopogonidae), based on mitochondrial cytochrome oxidase subunit 2. *Molecular Phylogenetics and Evolution* 27:21-35.

Berntson, E. A., S. C. France, and L. S. Mullineaux. 1999. Phylogenetic relationships within the Class Anthozoa (Phylum Cnidaria) based on nuclear 18S rDNA sequences. *Molecular Phylogenetics and Evolution* 13:417-433.

Berry, P. E., A. L. Hipp, K. J. Wurdack, B. Van Ee, and R. Riina. 2005. Molecular phylogenetics of the giant genus *Croton* and tribe Crotoneae (Euphorbiaceae sensu stricto) using its and trnL-trnF DNA sequence data. *American Journal of Botany* 92:1520-1534.

Binder, M., and D. S. Hibbett. 2002. Higher-level phylogenetic relationships of homobasidiomycetes (mushroom-forming fungi) inferred from four rDNA regions. *Molecular Phylogenetics and Evolution* 22:76-90.

Bleidorn, C., L. Vogt, and T. Bartolomaeus. 2003. New insights into polychaete phylogeny (Annelida) inferred from 18S rDNA sequences. *Molecular Phylogenetics and Evolution* 29:279-288.

Bodensteiner, P., M. Binder, J. M. Moncalvo, R. Agerer, and D. S. Hibbett. 2004. Phylogenetic relationships of cyphelloid homobasidiomycetes. *Molecular Phylogenetics and Evolution* 33:501-515.

Boisselier-Dubayle, M. C., J. Lambourdiere, and H. Bischler. 2002. Molecular phylogenies support multiple morphological reductions in the liverwort subclass Marchantiidae (Bryophyta). *Molecular Phylogenetics and Evolution* 24:66-77.

Braband, A., S. Richter, R. Hiesel, and G. Scholtz. 2002. Phylogenetic relationships within the Phyllopoda (Crustacea, Branchiopoda) based on mitochondrial and nuclear markers. *Molecular Phylogenetics and Evolution* 25:229-244.

Brandley, M. C., A. Schmitz, and T. W. Reeder. 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. *Systematic Biology* 54:373-390.

Bruvo-Madaric, B., B. A. Huber, A. Steinacher, and G. Pass. 2005. Phylogeny of pholcid spiders (Araneae : Pholcidae): Combined analysis using morphology and molecules. *Molecular Phylogenetics and Evolution* 37:661-673.

Bucheli, S. R., and J. Wenzel. 2005. Gelechloidea (Insecta : Lepidoptera) systematics: A reexamination using combined morphology and mitochondrial DNA data. *Molecular Phylogenetics and Evolution* 35:380-394.

Bybee, S. M., S. D. Taylor, C. R. Nelson, and M. F. Whiting. 2004. A phylogeny of robber flies (Diptera : Asilidae) at the subfamilial level: molecular evidence. *Molecular Phylogenetics and Evolution* 30:789-797.

Cabrero-Sanudo, F. J., and R. Zardoya. 2004. Phylogenetic relationships of Iberian Aphodiini (Coleoptera : Scarabaeldae) based on morphological and molecular data. *Molecular Phylogenetics and Evolution* 31:1084-1100.

Cameron, K. M. 2004. Utility of plastid psaB gene sequences for investigating intrafamilial relationships within Orchidaceae. *Molecular Phylogenetics and Evolution* 31:1157-1180.

Cameron, K. M. 2005. Leave it to the leaves: A molecular phylogenetic study of malaxideae (Epidendroideae, Orchidaceae). *American Journal of Botany* 92:1025-1032.

Campos, A., M. P. Cummings, J. L. Reyes, and J. P. Laclette. 1998. Phylogenetic relationships of platyhelminthes based on 18S ribosomal gene sequences. *Molecular Phylogenetics and Evolution* 10:1-10.

Carson, R. J., and G. S. Spicer. 2003. A phylogenetic analysis of the emberizid sparrows based on three mitochondrial genes. *Molecular Phylogenetics and Evolution* 29:43-57.

Caterino, M. S., T. Hunt, and A. P. Vogler. 2005. On the constitution and phylogeny of staphyliniformia (Insecta : Coleoptera). *Molecular Phylogenetics and Evolution* 34:655-672.

Chen, Y., X. A. Hui, J. Z. Fu, and D. W. Huang. 2004. A molecular phylogeny of eurytomid wasps inferred from DNA sequence data of 28S, 18S, 16S, and COI genes. *Molecular Phylogenetics and Evolution* 31:300-307.

Chesser, R. T. 2004. Molecular systematics of New World suboscine birds. *Molecular Phylogenetics and Evolution* 32:11-24.

Church, S. A. 2003. Molecular phylogenetics of *Houstonia* (Rubiaceae): descending aneuploidy and breeding system evolution in the radiation of the lineage across North America. *Molecular Phylogenetics and Evolution* 27:223-238.

Cook, L. G., P. J. Gullan, and H. E. Trueman. 2002. A preliminary phylogeny of the scale insects (Hemiptera : Sternorrhyncha : Coccoidea) based on nuclear small-subunit ribosomal DNA. *Molecular Phylogenetics and Evolution* 25:43-52.

Cryan, J. R. 2005. Molecular phylogeny of Cicadomorpha (Insecta : Hemiptera : Cicadoidea, Cercopoidea and Membracoidea): adding evidence to the controversy. *Systematic Entomology* 30:563-574.

Cunha, R. L., R. Castilho, L. Ruber, and R. Zardoya. 2005. Patterns of cladogenesis in the venomous marine gastropod genus *Conus* from the Cape Verde islands. *Systematic Biology* 54:634-650.

Dabert, J., M. Dabert, and S. V. Mironov. 2001. Phylogeny of feather mite subfamily Avenzoariinae (Acari : Analgoidea : Avenzoariidae) inferred from combined analyses of molecular and morphological data. *Molecular Phylogenetics and Evolution* 20:124-135.

Damgaard, J., N. M. Andersen, and R. Meier. 2005. Combining molecular and morphological analyses of water strider phylogeny (Hemiptera-Heteroptera, Gerromorpha): effects of alignment and taxon sampling. *Systematic Entomology* 30:289-309.

Darst, C. R., and D. C. Cannatella. 2004. Novel relation-ships among hylold frogs inferred from 12S and 16S mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution* 31:462-475.

deWaard, J. R., V. Sacherova, M. E. Cristescu, E. A. Remigio, T. J. Crease, and P. D. Hebert. 2006. Probing the relationships of the branchiopod crustaceans. *Molecular Phylogenetics and Evolution* 39:491-502.

Dietrich, C. H., R. A. Rakitov, J. L. Holmes, and W. C. Black. 2001. Phylogeny of the major lineages of Membracoidea (Insecta : Hemiptera : Cicadomorpha) based on 28S rDNA sequences. *Molecular Phylogenetics and Evolution* 18:293-305.

Distel, D. L. 2000. Phylogenetic relationships among Mytilidae (Bivalvia): 18S rRNA data suggest convergence in mytilid body plans. *Molecular Phylogenetics and Evolution* 15:25-33.

Douady, C. J., M. Dosay, M. S. Shivji, and M. J. Stanhope. 2003. Molecular phylogenetic evidence refuting the hypothesis of Batoidea (rays and skates) as derived sharks. *Molecular Phylogenetics and Evolution* 26:215-221.

Dunn, K. A., J. D. McEachran, and R. L. Honeycutt. 2003. Molecular phylogenetics of myliobatiform fishes (Chondrichthyes : Myliobatiformes), with comments on the effects of missing data on parsimony and likelihood. *Molecular Phylogenetics and Evolution* 27:259-270.

Fang, Q. Q., A. Mitchell, J. C. Regier, C. Mitter, T. P. Friedlander, and R. W. Poole. 2000. Phylogenetic utility of the nuclear gene dopa decarboxylase in noctuoid moths (Insecta : Lepidoptera : Noctuoidea). *Molecular Phylogenetics and Evolution* 15:473-486.

Fior, S., P. O. Karis, G. Casazza, L. Minuto, and F. Sala. 2006. Molecular phylogeny of the Caryophyllaceae (Caryophyllales) inferred from chloroplast MATK and nuclear rDNA its sequences. *American Journal of Botany* 93:399-411.

Fishbein, M., C. Hibsch-Jetter, D. E. Soltis, and L. Hufford. 2001. Phylogeny of saxifragales (angiosperms, eudicots): Analysis of a rapid, ancient radiation. *Systematic Biology* 50:817-847.

Forrest, L. L., M. Hughes, and P. M. Hollingsworth. 2005. A phylogeny of *Begonia* using nuclear ribosomal sequence data and morphological characters. *Systematic Botany* 30:671-682.

Freitas, A. V. L., and K. S. Brown. 2004. Phylogeny of the Nymphalidae (Lepidoptera). *Systematic Biology* 53:363-383.

Froslev, T. G., P. B. Matheny, and D. S. Hibbett. 2005. Lower level relationships in the mushroom genus *Cortinarius* (Basidiomycota, Agaricales): A comparison of RPB1, RPB2, and ITS phylogenies. *Molecular Phylogenetics and Evolution* 37:602-618.

Garcia-Jacas, N., T. Garnatje, A. Susanna, and R. Vilatersana. 2002. Tribal and subtribal delimitation and phylogeny of the Cardueae (Asteraceae): A combined nuclear and chloroplast DNA analysis. *Molecular Phylogenetics and Evolution* 22:51-64.

Garcia-Varela, M., M. P. Cummings, G. P. P. de Leon, S. L. Gardner, and J. P. Laclette. 2002. Phylogenetic analysis based on 18S ribosomal RNA gene sequences supports the existence of class polyacanthocephala (acanthocephala). *Molecular Phylogenetics and Evolution* 23:288-292.

Goetsch, L., A. J. Eckert, and B. D. Hall. 2005. The molecular systematics of *Rhododendron* (Ericaceae): A phylogeny based upon RPB2 gene sequences. *Systematic Botany* 30:616-626.

Gomez-Zurita, J., P. Jolivet, and A. P. Vogler. 2005. Molecular systematics of Eumolpinae and the relationships with Spilopyrinae (Coleoptera, Chrysomelidae). *Molecular Phylogenetics and Evolution* 34:584-600.

Graham, S. A., J. Hall, K. Sytsma, and S. H. Shi. 2005. Phylogenetic analysis of the Lythraceae based on four gene regions and morphology. *International Journal of Plant Sciences* 166:995-1017.

Grande, C., J. Templado, J. L. Cervera, and R. Zardoya. 2004. Phylogenetic relationships among Opisthobranchia (Mollusca : Gastropoda) based on mitochondrial cox 1, trnV, and rrnL genes. *Molecular Phylogenetics and Evolution* 33:378-388.

Han, H. Y., and K. E. Ro. 2005. Molecular phylogeny of the superfamily Tephritoidea (Insecta : Diptera): new evidence from the mitochondrial 12S 16S, and COII genes. *Molecular Phylogenetics and Evolution* 34:416-430.

Hardman, M. 2005. The phylogenetic relationships among non-diplomystid catfishes as inferred from mitochondrial cytochrome b sequences; the search for the ictalurid sister taxon (Otophysi : Siluriformes). *Molecular Phylogenetics and Evolution* 37:700-720.

Harrington, M. G., K. J. Edwards, S. A. Johnson, M. W. Chase, and P. A. Gadek. 2005. Phylogenetic inference in Sapindaceae sensu lato using plastid matK and rbcL DNA sequences. *Systematic Botany* 30:366-382.

Hill, R. V. 2005. Integration of morphological data sets for phylogenetic analysis of amniota: The importance of integumentary characters and increased taxonomic sampling. *Systematic Biology* 54:530-547.

Honda, M., H. Ota, M. Kobayashi, J. Nabhitabhata, H. S. Yong, and T. Hikida. 2000. Phylogenetic relationships, character evolution, and biogeography of the subfamily Lygosominae (Reptilia : Scincidae) inferred from mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution* 15:452-461.

Hooge, M. D., P. A. Haye, S. Tyler, M. K. Litvaitis, and I. Kornfield. 2002. Molecular systematics of the Acoela (Acoelomorpha, Platyhelminthes) and its concordance with morphology. *Molecular Phylogenetics and Evolution* 24:333-342.

Iglesias, S. P., G. Lecointre, and D. Y. Sellos. 2005. Extensive paraphylies within sharks of the order Carcharhiniformes inferred from nuclear and mitochondrial. *Molecular Phylogenetics and Evolution* 34:569-583.

Inoue, J. G., M. Miya, K. Tsukamoto, and M. Nishida. 2004. Mitogenomic evidence for the monophyly of elopomorph fishes (Teleostei) and the evolutionary origin of the leptocephalus larva. *Molecular Phylogenetics and Evolution* 32:274-286.

Ishiguro, N. B., M. Miya, and M. Nishida. 2003. Basal euteleostean relationships: a mitogenomic perspective on the phylogenetic reality of the "Protacanthopterygii". *Molecular Phylogenetics and Evolution* 27:476-488.

Jansa, S. A., and M. Weksler. 2004. Phylogeny of muroid rodents: relationships within and among major lineages as determined by IRBP gene sequences. *Molecular Phylogenetics and Evolution* 31:256-276.

Jarvis, K. J., F. Haas, and M. F. Whiting. 2005. Phylogeny of earwigs (Insecta : Dermaptera) based on molecular and morphological evidence: reconsidering the classification of Dermaptera. *Systematic Entomology* 30:442-453.

Johnson, R. N., P. M. Agapow, and R. H. Crozier. 2003. A tree island approach to inferring phylogeny in the ant subfamily Formicinae, with especial reference to the evolution of weaving. *Molecular Phylogenetics and Evolution* 29:317-330.

Jones, K. E., A. Purvis, A. MacLarnon, O. R. Bininda-Emonds, and N. B. Simmons. 2002. A phylogenetic supertree of the bats (Mammalia: Chiroptera). *Biological Reviews of the Cambridge Philosophical Society* 77:223-59.

Jost, M. C., and K. L. Shaw. 2006. Phylogeny of Ensifera (Hexapoda: Orthoptera) using three ribosomal loci, with implications for the evolution of acoustic communication. *Molecular Phylogenetics and Evolution* 38:510-30.

Kankare, M., and M. R. Shaw. 2004. Molecular phylogeny of *Cotesia* Cameron, 1891 (Insecta : Hymenoptera : Braconidae : Microgastrinae) parasitoids associated with

Melitaeini butterflies (Insecta : Lepidoptera : Nymphalidae : Melitaeini). *Molecular Phylogenetics and Evolution* 32:207-220.

Kappner, I., and R. Bieler. 2006. Phylogeny of venus clams (Bivalvia: Venerinae) as inferred from nuclear and mitochondrial gene sequences. *Molecular Phylogenetics and Evolution* 40:317-331.

Kathriarachchi, H., P. Hoffmann, R. Samuel, K. J. Wurdack, and M. W. Chase. 2005. Molecular phylogenetics of Phyllanthaceae inferred from five genes (plastid atpB, matK, 3 ' ndhF, rbcL, and nuclear PHYC). *Molecular Phylogenetics and Evolution* 36:112-134.

Kauff, F., and F. Lutzoni. 2002. Phylogeny of the Gyalectales and Ostropales (Ascomycota, Fungi): among and within order relationships based on nuclear ribosomal RNA small and large subunits. *Molecular Phylogenetics and Evolution* 25:138-156.

Keeley, S. C., Z. H. Forsman, and R. Chan. 2007. A phylogeny of the "evil tribe" (Vernonieae: Compositae) reveals Old/New World long distance dispersal: Support from separate and combined congruent datasets (trnL-F, ndhF, ITS). *Molecular Phylogenetics and Evolution* 44:89-103.

Kjer, K. M. 2004. Aligned 18S and insect phylogeny. *Systematic Biology* 53:506-514.

Kjer, K. M., R. J. Blahnik, and R. W. Holzenthal. 2001. Phylogeny of Trichoptera (caddisflies): Characterization of signal and noise within multiple datasets. *Systematic Biology* 50:781-816.

Krenz, J. G., G. J. P. Naylor, H. B. Shaffer, and F. J. Janzen. 2005. Molecular phylogenetics and evolution of turtles. *Molecular Phylogenetics and Evolution* 37:178-191.

Lavin, M., P. S. Herendeen, and M. F. Wojciechowski. 2005. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. *Systematic Biology* 54:575-94.

Lavoue, S., and J. P. Sullivan. 2004. Simultaneous analysis of five molecular markers provides a well-supported phylogenetic hypothesis for the living bony-tongue fishes (Osteoglossomorpha : Teleostei). *Molecular Phylogenetics and Evolution* 33:171-185.

Lawrey, J. D., M. Binder, P. Diederich, M. C. Molina, M. Sikaroodi, and D. Ertz. 2007. Phylogenetic diversity of lichen-associated homobasidiomycetes. *Molecular Phylogenetics and Evolution* 44:778-789.

Lawson, R., J. B. Slowinski, B. I. Crother, and F. T. Burbrink. 2005. Phylogeny of the Colubroidea (Serpentes): New evidence from mitochondrial and nuclear genes. *Molecular Phylogenetics and Evolution* 37:581-601.

Le Goff-Vitry, M. C., A. D. Rogers, and D. Baglow. 2004. A deep-sea slant on the molecular phylogeny of the Scleractinia. *Molecular Phylogenetics and Evolution* 30:167-177.

Lewis, C. E., and J. J. Doyle. 2001. Phylogenetic utility of the nuclear gene malate synthase in the palm family (Arecaceae). *Molecular Phylogenetics and Evolution* 19:409-420.

Li, C., and G. Orti. 2007. Molecular phylogeny of Clupeiformes (Actinopterygii) inferred from nuclear and mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution* 44:386-398.

Li, H. M., R. Q. Deng, J. W. Wang, Z. Y. Chen, F. L. Jia, and X. Z. Wang. 2005. A preliminary phylogeny of the Pentatomomorpha (Hemiptera : Heteroptera) based on nuclear 18S rDNA and mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution* 37:313-326.

Maddison, W. P., and M. C. Hedin. 2003. Jumping spider phylogeny (Araneae : Salticidae). *Invertebrate Systematics* 17:529-549.

Manuel, M., C. Borchiellini, E. Alivon, Y. Le Parco, J. Vacelet, and N. Boury-Esnault. 2003. Phylogeny and evolution of calcareous sponges: Monophyly of Calcinea and Calcaronea, high level of morphological homoplasy, and the primitive nature of axial symmetry. *Systematic Biology* 52:311-333.

Mardulyn, P., and J. B. Whitfield. 1999. Phylogenetic signal in the COI, 16S, and 28S genes for inferring relationships among genera of Microgastrinae (Hymenoptera; Braconidae): Evidence of a high diversification rate in this group of parasitoids. *Molecular Phylogenetics and Evolution* 12:282-294.

Martinez-Navarro, E. M., J. Galian, and J. Serrano. 2005. Phylogeny and molecular evolution of the tribe Harpalini (Coleoptera, Carabidae) inferred from mitochondrial cytochrome-oxidase I. *Molecular Phylogenetics and Evolution* 35:127-146.

Martins, T. R., and T. J. Barkman. 2005. Reconstruction of Solanaceae phylogeny using the nuclear gene SAMT. *Systematic Botany* 30:435-447.

Marvaldi, A. E., A. S. Sequeira, C. W. O'Brien, and B. D. Farrell. 2002. Molecular and morphological phylogenetics of weevils (Coleoptera, Curculionoidea): Do niche shifts accompany diversification? *Systematic Biology* 51:761-785.

McFadden, C. S., S. C. France, J. A. Sanchez, and P. Alderslade. 2006. A molecular phylogenetic analysis of the Octocorallia (Cnidaria: Anthozoa) based on mitochondrial protein-coding sequences. *Molecular Phylogenetics and Evolution* 41:513-27.

Meldal, B. H., N. J. Debenham, P. De Ley, I. T. De Ley, J. R. Vanfleteren, A. R. Vierstraete, W. Bert, G. Borgonie, T. Moens, P. A. Tyler, M. C. Austen, M. L. Blaxter, A. D. Rogers, and P. J. Lambshead. 2007. An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa. *Molecular Phylogenetics and Evolution* 42:622-636.

Michel-Salzat, A., S. A. Cameron, and M. L. Oliveira. 2004. Phylogeny of the orchid bees (Hymenoptera : Apinae : Euglossini): DNA and morphology yield equivalent patterns. *Molecular Phylogenetics and Evolution* 32:309-323.

Miller, A. N., and N. M. A. Andrew. 2005. Multi-gene phylogenies indicate ascomal wall morphology is a better predictor of phylogenetic relationships than ascospore morphology in the Sordariales (Ascomycota, Fungi). *Molecular Phylogenetics and Evolution* 35:60-75.

Murphy, W. J., E. Eizirik, S. J. O'Brien, O. Madsen, M. Scally, C. J. Douady, E. Teeling, O. A. Ryder, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* 294:2348-2351.

Murray, S., M. F. Jorgensen, S. Y. W. Ho, D. J. Patterson, and L. S. Jermiin. 2005. Improving the analysis of dinoflagellate phylogeny based on rDNA. *Protist* 156:269-286.

Near, T. J., J. R. Garey, and S. A. Nadler. 1998. Phylogenetic relationships of the Acanthocephala inferred from 18S ribosomal DNA sequences. *Molecular Phylogenetics and Evolution* 10:287-298.

Neinhuis, C., S. Wanke, K. W. Hilu, K. Muller, and T. Borsch. 2005. Phylogeny of Aristolochiaceae based on parsimony, likelihood, and Bayesian analyses of trnL-trnF sequences. *Plant Systematics and Evolution* 250:7-26.

Nichols, S. A. 2005. An evaluation of support for order-level monophyly and interrelationships within the class Demospongiae using partial data from the large subunit rDNA and cytochrome oxidase subunit I. *Molecular Phylogenetics and Evolution* 34:81-96.

Ober, K. A. 2002. Phylogenetic relationships of the carabid subfamily Harpalinae (Coleoptera) based on molecular sequence data. *Molecular Phylogenetics and Evolution* 24:228-248.

Obermiller, L. E., and E. Pfeiler. 2003. Phylogenetic relationships of elopomorph fishes inferred from mitochondrial ribosomal DNA sequences. *Molecular Phylogenetics and Evolution* 26:202-214.

Ogden, T. H., and M. F. Whiting. 2005. Phylogeny of Ephemeroptera (mayflies) based on molecular evidence. *Molecular Phylogenetics and Evolution* 37:625-643.

Olson, P. D., D. T. J. Littlewood, R. A. Bray, and J. Mariaux. 2001. Interrelationships and evolution of the tapeworms (Platyhelminthes : Cestoda). *Molecular Phylogenetics and Evolution* 19:443-467.

Ortiz-Rivas, B., A. Moya, and D. Martinez-Torres. 2004. Molecular systematics of aphids (Homoptera : Aphididae): new insights from the long-wavelength opsin gene. *Molecular Phylogenetics and Evolution* 30:24-37.

Palma, R. E., and A. E. Spotorno. 1999. Molecular systematics of marsupials based on the rRNA 12S mitochondrial gene: The phylogeny of didelphimorphia and of the living fossil microbiotheriid *Dromiciops gliroides* Thomas. *Molecular Phylogenetics and Evolution* 13:525-535.

Passamaneck, Y. J., C. Schander, and K. M. Halanych. 2004. Investigation of molluscan phylogeny using large-subunit and small-subunit nuclear rRNA sequences. *Molecular Phylogenetics and Evolution* 32:25-38.

Paton, A. J., D. Springate, S. Suddee, D. Otieno, R. J. Grayer, M. M. Harley, F. Willis, M. S. J. Simmonds, M. P. Powell, and V. Savolainen. 2004. Phylogeny and evolution of basils and allies (Ocimeae, Labiatae) based on three plastid DNA regions. *Molecular Phylogenetics and Evolution* 31:277-299.

Phillips, M. J., and D. Penny. 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. *Molecular Phylogenetics and Evolution* 28:171-185.

Pisani, D. 2004. Identifying and removing fast-evolving sites using compatibility analysis: An example from the arthropoda. *Systematic Biology* 53:978-989.

Pons, J., T. G. Barraclough, K. Theodorides, A. Cardoso, and A. P. Vogler. 2004. Using exon and intron sequences of the gene Mp20 to resolve basal relationships in *Cicindela* (Coleoptera : Cicindelidae). *Systematic Biology* 53:554-570.

Pruess, K. P., B. J. Adams, T. J. Parsons, X. Zhu, and T. O. Powers. 2000. Utility of the mitochondrial cytochrome oxidase II gene for resolving relationships among black flies (Diptera : Simuliidae). *Molecular Phylogenetics and Evolution* 16:286-295.

Pryer, K. M., E. Schuettpelz, P. G. Wolf, H. Schneider, A. R. Smith, and R. Cranfill. 2004. Phylogeny and evolution of ferns (monilophytes) with a focus on the early leptosporangiate divergences. *American Journal of Botany* 91:1582-1598.

Reeder, T. W. 2003. A phylogeny of the Australian Sphenomorphus group (Scincidae: Squamata) and the phylogenetic placement of the crocodile skinks (*Tribolonotus*): Bayesian approaches to assessing congruence and obtaining confidence in maximum likelihood inferred relationships. *Molecular Phylogenetics and Evolution* 27:384-97.

Remerie, T., B. Bulckaen, J. Calderon, T. Deprez, J. Mees, J. Vanfleteren, A. Vanreusel, A. Vierstraete, M. Vincx, K. J. Wittmann, and T. Wooldridge. 2004. Phylogenetic relationships within the Mysidae (Crustacea, Peracarida, Mysida) based on nuclear 18S ribosomal RNA sequences. *Molecular Phylogenetics and Evolution* 32:770-777.

Remsen, J., and P. O'Grady. 2002. Phylogeny of Drosophilinae (Diptera : Drosophilidae), with comments on combined analysis and character support. *Molecular Phylogenetics and Evolution* 24:249-264.

Rest, J. S., J. C. Ast, C. C. Austin, P. J. Waddell, E. A. Tibbetts, J. M. Hay, and D. P. Mindell. 2003. Molecular systematics of primary reptilian lineages and the tuatara mitochondrial genome. *Molecular Phylogenetics and Evolution* 29:289-297.

Roelants, K., and F. Bossuyt. 2005. Archaeobatrachian paraphyly and pangaean diversification of crown-group frogs. *Systematic Biology* 54:111-126.

Saito, M., S. Kojima, and K. Endo. 2000. Mitochondrial COI sequences of brachiopods: Genetic code shared with protostomes and limits of utility for phylogenetic reconstruction. *Molecular Phylogenetics and Evolution* 15:331-344.

Salamin, N., T. R. Hodkinson, and V. Savolainen. 2002. Building supertrees: an empirical assessment using the grass family (Poaceae). *Systematic Biology* 51:136-50.

San Mauro, D., D. J. Gower, O. V. Oommen, M. Wilkinson, and R. Zardoya. 2004. Phylogeny of caecilian amphibians (Gymnophiona) based on complete mitochondrial genomes and nuclear RAG1. *Molecular Phylogenetics and Evolution* 33:413-427.

Sano, R., M. Takamiya, M. Ito, S. Kurita, and M. Hasebe. 2000. Phylogeny of the lady fern group, tribe Physematieae (Dryopteridaceae), based on chloroplast rbcL gene sequences. *Molecular Phylogenetics and Evolution* 15:403-413.

Scheffer, S. J., I. S. Winkler, and B. M. Wiegmann. 2007. Phylogenetic relationships within the leaf-mining flies (Diptera: Agromyzidae) inferred from sequence data from multiple genes. *Molecular Phylogenetics and Evolution* 42:756-75.

Schmitt, I., and H. T. Lumbsch. 2004. Molecular phylogeny of the Pertusariaceae supports secondary chemistry as an important systematic character set in lichen-forming ascomycetes. *Molecular Phylogenetics and Evolution* 33:43-55.

Shaw, J., and K. Renzaglia. 2004. Phylogeny and diversification of bryophytes. *American Journal of Botany* 91:1557-1581.

Shi, M., X. X. Chen, and C. van Achterberg. 2005. Phylogenetic relationships among the Braconidae (Hymenoptera : Ichneumonoidea) inferred from partial 16S rDNA, 28S rDNA D2, 18S rDNA gene sequences and morphological characters. *Molecular Phylogenetics and Evolution* 37:104-116.

Simpson, A. G. B., and A. J. Roger. 2004. Protein phylogenies robustly resolve the deep-level relationships within Euglenozoa. *Molecular Phylogenetics and Evolution* 30:201-212.

Sjolin, E., C. Erseus, and M. Kallersjo. 2005. Phylogeny of Tubificidae (Annelida, Clitellata) based on mitochondrial and nuclear sequence data. *Molecular Phylogenetics and Evolution* 35:431-441.

Skevington, J. H., and D. K. Yeates. 2000. Phylogeny of the Syrphoidea (Diptera) inferred from mtDNA sequences and morphology with particular reference to classification of the Pipunculidae (Diptera). *Molecular Phylogenetics and Evolution* 16:212-224.

Slowinski, J. B., and J. S. Keogh. 2000. Phylogenetic relationships of elapid snakes based on cytochrome b mtDNA sequences. *Molecular Phylogenetics and Evolution* 15:157-164.

Slowinski, J. B., and R. Lawson. 2002. Snake phylogeny: evidence from nuclear and mitochondrial genes. *Molecular Phylogenetics and Evolution* 24:194-202.

Smith, P. T., S. Kambhampati, and K. A. Armstrong. 2003. Phylogenetic relationships among *Bactrocera* species (Diptera : Tephritidae) inferred from mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution* 26:8-17.

Smythe, A. B., M. J. Sanderson, and S. A. Nadler. 2006. Nematode small subunit phylogeny correlates with alignment parameters. *Systematic Biology* 55:972-92.

Soejima, A., and J. Wen. 2006. Phylogenetic analysis of the grape family (Vitaceae) based on three chloroplast markers. *American Journal of Botany* 93:278-287.

Sorensen, M. V., and G. Giribet. 2006. A modern approach to rotiferan phylogeny: Combining morphological and molecular data. *Molecular Phylogenetics and Evolution* 40:585-608.

Spicer, G. S., and L. Dunipace. 2004. Molecular phylogeny of songbirds (Passeriformes) inferred from mitochondrial 16S ribosomal RNA gene sequences. *Molecular Phylogenetics and Evolution* 30:325-335.

Stach, T., and J. M. Turbeville. 2002. Phylogeny of Tunicata inferred from molecular and morphological characters. *Molecular Phylogenetics and Evolution* 25:408-428.

Stahls, G., and K. Nyblom. 2000. Phylogenetic analysis of the genus *Cheilosia* (Diptera, Syrphidae) using mitochondrial COI sequence data. *Molecular Phylogenetics and Evolution* 15:235-241.

Starr, J. R., S. A. Harris, and D. A. Simpson. 2004. Phylogeny of the unispicate taxa in Cyperaceae tribe Cariceae I: Generic relationships and evolutionary scenarios. *Systematic Botany* 29:528-544.

Steane, D. A., G. E. McKinnon, R. E. Vaillancourt, and B. M. Potts. 1999. ITS sequence data resolve higher level relationships among the eucalypts. *Molecular Phylogenetics and Evolution* 12:215-223.

Stockley, B., A. B. Smith, T. Littlewood, H. A. Lessios, and J. A. Mackenzie-Dodds. 2005. Phylogenetic relationships of spatangoid sea urchins (Echinoidea): taxon sampling density and congruence between morphological and molecular estimates. *Zoologica Scripta* 34:447-468.

Sundberg, P., and M. Saur. 1998. Molecular phylogeny of some European heteronemertean (Nemertea) species and the monophyletic status of Riseriellus, Lineus, and Micrura. *Molecular Phylogenetics and Evolution* 10:271-280.

Teeling, E. C., M. S. Springer, O. Madsen, P. Bates, S. J. O'Brien, and W. J. Murphy. 2005. A molecular phylogeny for bats illuminates biogeography and the fossil record. *Science* 307:580-584.

Thacker, C. E. 2003. Molecular phylogeny of the gobioid fishes (Teleostei : Perciformes : Gobioidei). *Molecular Phylogenetics and Evolution* 26:354-368.

Townsend, T. M., A. Larson, E. Louis, and J. R. Macey. 2004. Molecular phylogenetics of Squamata: The position of snakes, Amphisbaenians, and Dibamids, and the root of the Squamate tree. *Systematic Biology* 53:735-757.

van den Berg, C., D. H. Goldman, J. V. Freudenstein, A. M. Pridgeon, K. M. Cameron, and M. W. Chase. 2005. An overview of the phylogenetic relationships within Epidendroideae inferred from multiple DNA regions and recircumscription of Epidendreae and Arethuseae (Orchidaceae). *American Journal of Botany* 92:613-624.

von Dohlen, C. D., C. A. Rowe, and O. E. Heie. 2006. A test of morphological hypotheses for tribal and subtribal relationships of Aphidinae (Insecta: Hemiptera: Aphididae) using DNA sequences. *Molecular Phylogenetics and Evolution* 38:316-29.

Wang, X., J. Li, and S. He. 2007. Molecular evidence for the monophyly of East Asian groups of Cyprinidae (Teleostei: Cypriniformes) derived from the nuclear recombination activating gene 2 sequences. *Molecular Phylogenetics and Evolution* 42:157-70.

Watson, L. E., T. M. Evans, and T. Boluarte. 2000. Molecular phylogeny and biogeography of tribe Anthemideae (Asteraceae), based on chloroplast gene ndhF. *Molecular Phylogenetics and Evolution* 15:59-69.

Weekers, P. H. H., G. Murugan, J. R. Vanfleteren, D. Belk, and H. J. Dumont. 2002. Phylogenetic analysis of anostracans (Branchiopoda : Anostraca) inferred from nuclear 18S ribosomal DNA (18S rDNA) sequences. *Molecular Phylogenetics and Evolution* 25:535-544.

Weeks, A., D. C. Daly, and B. B. Simpson. 2005. The phylogenetic history and biogeography of the frankincense and myrrh family (Burseraceae) based on nuclear and chloroplast sequence data. *Molecular Phylogenetics and Evolution* 35:85-101.

Weisrock, D. W., L. J. Harmon, and A. Larson. 2005. Resolving deep phylogenetic relationships in salamanders: Analyses of mitochondrial and nuclear genomic data. *Systematic Biology* 54:758-777.

Weksler, M. 2003. Phylogeny of Neotropical oryzomyine rodents (Muridae : Sigmodontinae) based on the nuclear IRBP exon. *Molecular Phylogenetics and Evolution* 29:331-349.

Wilson, C. A. 2004. Phylogeny of Iris based on chloroplast matK gene and trnK intron sequence data. *Molecular Phylogenetics and Evolution* 33:402-412.

Wirz, A., S. Pucciarelli, C. Miceli, P. Tongiorgi, and M. Balsamo. 1999. Novelty in phylogeny of Gastrotricha: Evidence from 18S rRNA gene. *Molecular Phylogenetics and Evolution* 13:314-318.

Wortley, A. H., P. J. Rudall, D. J. Harris, and R. W. Scotland. 2005. How much data are needed to resolve a difficult phylogeny? Case study in Lamiales. *Systematic Biology* 54:697-709.

Xiang, Q. Y., M. L. Moody, D. E. Soltis, C. Z. Fan, and P. S. Soltis. 2002. Relationships within Cornales and circumscription of Cornaceae - matK and rbcL sequence data and effects of outgroups and long branches. *Molecular Phylogenetics and Evolution* 24:35-57.

Yang, L. L., B. M. Wiegmann, D. K. Yeates, and M. E. Irwin. 2000. Higher-level phylogeny of the Therevidae (Diptera : Insecta) based on 28S ribosomal and elongation factor-1 alpha gene sequences. *Molecular Phylogenetics and Evolution* 15:440-451.

Yoon, H. S., J. Y. Lee, S. M. Boo, and D. Bhattacharya. 2001. Phylogeny of Alariaceae, Laminariaceae, and Lessoniaceae (Phaeophyceae) based on plastid-encoded RuBisCo spacer and nuclear-encoded ITS sequence comparisons. *Molecular Phylogenetics and Evolution* 21:231-243.

Yuan, Y. M., S. Wohlhauser, M. Moller, P. Chassot, G. Mansion, J. Grant, P. Kupfer, and J. Klackenberg. 2003. Monophyly and relationships of the tribe Exaceae (Gentianaceae) inferred from nuclear ribosomal and chloroplast DNA sequences. *Molecular Phylogenetics and Evolution* 28:500-517.

Zhang, L. B., M. P. Simmons, A. Kocyan, and S. S. Renner. 2006. Phylogeny of the Cucurbitales based on DNA sequences of nine loci from three genomes: implications for morphological and sexual system evolution. *Molecular Phylogenetics and Evolution* 39:305-22.

Zhang, P., H. Zhou, Y. Q. Chen, Y. F. Liu, and L. H. Qu. 2005. Mitogenomic perspectives on the origin and phylogeny of living amphibians. *Systematic Biology* 54:391-400.

References

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

Agapow, P. M., and A. Purvis. 2002. Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. *Systematic Biology* 51:866-872.

Albrecht, C., K. Kuhn, and B. Streit. 2007. A molecular phylogeny of Planorboidea (Gastropoda, Pulmonata): insights from enhanced taxon sampling. *Zoologica Scripta* 36:27-39.

Aldous, D. J. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statistical Science* 16:23-34.

Alfaro, M. E., F. Santini, and C. D. Brock. 2007. Do reefs drive diversification in marine teleosts? Evidence from the pufferfish and their allies (order tetraodontiformes). *Evolution* 61:2104-2126.

Ane, C., J. G. Burleigh, M. M. McMahon, and M. J. Sanderson. 2005. Covarion structure in plastid genome evolution: A new statistical test. *Molecular Biology and Evolution* 22:914-924.

Ane, C., and M. J. Sanderson. 2005. Missing the forest for the trees: Phylogenetic compression and its implications for inferring complex evolutionary histories. *Systematic Biology* 54:146-157.

Barker, G. M. 2002. Phylogenetic diversity: a quantitative framework for measurement of priority and achievement in biodiversity conservation. *Biological Journal of the Linnean Society* 76:165-194.

Becerra, J. X. 2005. Timing the origin and expansion of the Mexican tropical dry forest. *Proceedings of the National Academy of Sciences of the United States of America* 102:10919-10923.

Bell, C. D. 2007. Phylogenetic placement and biogeography of the North American species of Valerianella (Valerianaceae: Dipsacales) based on chloroplast and nuclear DNA. *Molecular Phylogenetics and Evolution* 44:929-41.

Blanquart, S., and N. Lartillot. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Molecular Biology and Evolution* 23:2058-2071.

Blanquart, S., and N. Lartillot. 2008. A site- and time-heterogeneous model of amino acid replacement. *Molecular Biology and Evolution* 25:842-58.

Blum, M. G. B., and O. Francios. 2006. Which random processes describe the Tree of Life? A large-scale study of phylogenetic tree imbalance. *Systematic Biology* 55:685-691.

Boussau, B., and M. Gouy. 2006. Efficient likelihood computations with nonreversible models of evolution. *Systematic Biology* 55:756-768.

Brandley, M. C., A. Schmitz, and T. W. Reeder. 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. *Systematic Biology* 54:373-390.

Braun, E. L., and R. T. Kimball. 2002. Examining basal avian divergences with mitochondrial sequences: Model complexity, taxon sampling, and sequence length. *Systematic Biology* 51:614-625.

Bremer, B., R. K. Jansen, B. Oxelman, M. Backlund, H. Lantz, and K. J. Kim. 1999. More characters or more taxa for a robust phylogeny--case study from the coffee family (Rubiaceae). *Systematic Biology* 48:413-435.

Brown, J. M., and A. R. Lemmon. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. *Systematic Biology* 56:643-655.

Bruno, W. J., and A. L. Halpern. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. *Molecular Biology and Evolution* 16:564-566.

Bull, J. J., J. P. Huelsenbeck, C. W. Cunningham, D. L. Swofford, and P. J. Waddell. 1993. Partitioning and combining data in phylogenetic analysis. *Systematic Biology* 42:384-397.

Cannarozzi, G., A. Schneider, and G. Gonnet. 2007. A phylogenomic study of human, dog, and mouse. *PLoS Computational Biology* 3:9-14.

Chan, K. M. A., and B. R. Moore. 1999. Accounting for mode of speciation increases power and realism of tests of phylogenetic asymmetry. *American Naturalist* 153:332-346.

Chan, K. M. A., and B. R. Moore. 2002. Whole-tree methods for detecting differential diversification rates. *Systematic Biology* 51:855-865.

Chang, J. T. 1996. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Mathematical Biosciences* 134:189-215.

Chen, W. J., C. Bonillo, and G. Lecointre. 2003. Repeatability of clades as a criterion of reliability: A case study for molecular phylogeny of Acanthomorpha (Teleostei) with larger number of taxa. *Molecular Phylogenetics and Evolution* 26:262-288.

Colless, D. H. 1982. Phylogenetics: the theory and practice of phylogenetic systematics II (book review). *Systematic Zoology* 31:100-104.

Crozier, R. H., L. J. Dunnett, and P. M. Agapow. 2005. Phylogenetic biodiversity assessment based on systematic nomenclature. *Evolutionary Bioinformatics* 1:11-36.

Cummings, M. P., and A. Meyer. 2005. Magic bullets and golden rules: Data sampling in molecular phylogenetics. *Zoology* 108:329-336.

Cunningham, S. A. 1995. Problems with null models in the study of phylogenetic radiation. *Evolution* 49:1292-1294.

Cunningham, C. W., K. Jeng, J. Husti, M. Badgett, I. J. Molineux, D. M. Hillis, and J. J. Bull. 1997. Parallel molecular evolution of deletions and nonsense mutations in bacteriophage T7. *Molecular Biology and Evolution* 14:113-116.

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

Darst, C. R., P. A. Menendez-Guerrero, L. A. Coloma, and D. C. Cannatella. 2005. Evolution of dietary specialization and chemical defense in poison frogs (Dendrobatidae): A comparative analysis. *American Naturalist* 165:56-69.

DeBry, R. W. 1992. The consistency of several phylogeny-inference methods under varying evolutionary rates. *Molecular Biology and Evolution* 9:537-551.

DeBry, R. W. 2005. The systematic component of phylogenetic error as a function of taxonomic sampling under parsimony. *Systematic Biology* 54:432-440.

Degnan, J. H., and L. A. Salter. 2005. Gene tree distributions under the coalescent process. *Evolution* 59:24-37.

Degnan, J. H., and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genetics* 2:762-768.

Dial, K. P., and J. M. Marzluff. 1989. Nonrandom diversification within taxonomic assemblages. *Systematic Zoology* 38:26-37.

Dodd, M. E., J. Silvertown, and M. W. Chase. 1999. Phylogenetic analysis of trait evolution and species diversity variation among angiosperm families. *Evolution* 53:732-744.

Drummond, A. J., S. Y. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology* 4:e88.

Edwards, S. V., L. Liu, and D. K. Pearl. 2007. High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences of the United States of America* 104:5936-5941.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* 27:401-410.

Felsenstein, J. 1985. Phylogenies and the comparative method. *American Naturalist* 125:1-15.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer, Sunderland, MA.

Fitch, W. M., and M. Bruschi. 1987. The evolution of prokaryotic ferredoxins - with a general-method correcting for unobserved substitutions in less branched lineages. *Molecular Biology and Evolution* 4:381-394.

Fitch, W. M., and J. J. Beintema. 1990. Correcting parsimonious trees for unseen nucleotide substitutions - the effect of dense branching as exemplified by Ribonuclease. *Molecular Biology and Evolution* 7:438-443.

Foster, P. G., L. S. Jermiin, and D. A. Hickey. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *Journal of Molecular Evolution* 44:282-288.

Foster, P. G. 2004. Modeling compositional heterogeneity. *Systematic Biology* 53:485-95.

Freitas, A. V. L., and K. S. Brown. 2004. Phylogeny of the Nymphalidae (Lepidoptera). *Systematic Biology* 53:363-383.

Freudenstein, J. V., K. M. Pickett, M. P. Simmons, and J. W. Wenzel. 2003. From basepairs to birdsongs: phylogenetic data in the age of genomics. *Cladistics* 19:333-347.

Fusco, G., and Q. C. B. Cronk. 1995. A new method for evaluating the shape of large phylogenies. *Journal of Theoretical Biology* 175:235-243.

Galtier, N., and M. Gouy. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Molecular Biology and Evolution* 15:871-9.

Gascuel, O., D. Bryant, and F. Denis. 2001. Strengths and limitations of the minimum evolution principle. *Systematic Biology* 50:621-627.

Gatesy, J., R. DeSalle, and N. Wahlberg. 2007. How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence. *Systematic Biology* 56:355-363.

Gaucher, E. A., J. M. Thomson, M. F. Burgan, and S. A. Benner. 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* 425:285-288.

Gojobori, T., K. Ishii, and M. Nei. 1982. Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. *Journal of Molecular Evolution* 18:414-423.

Goldman, N. 1993. Simple diagnostic statistical tests of models for DNA substitution. *Journal of Molecular Evolution* 37:650-661.

Goldman, N., and Z. Yang. 1994. Codon-based model of nucleotide substitution for protein-coding DNA-sequences. *Molecular Biology and Evolution* 11:725-736.

Good-Avila, S. V., V. Souza, B. S. Gaut, and L. E. Eguiarte. 2006. Timing and rate of speciation in Agave (Agavaceae). *Proceedings of the National Academy of Sciences of the United States of America* 103:9124-9129.

Gould, S. J., D. M. Raup, J. J. Sepowski, and T. J. M. Schopf. 1977. The shape of evolution: A comparison of real and random clades. *Paleobiology* 3:23-40.

Gowri-Shankar, V., and M. Rattray. 2007. A reversible jump method for Bayesian phylogenetic inference with a nonhomogeneous substitution model. *Molecular Biology and Evolution* 24:1286-1299.

Graybeal, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? *Systematic Biology* 47:9-17.

Guyer, C., and J. B. Slowinski. 1991. Comparisons of observed phylogenetic topologies with null expectations among three monophyletic lineages. *Evolution* 45:340-350.

Guyer, C., and J. B. Slowinski. 1993. Adaptive radiation and the topology of large phylogenies. *Evolution* 47:253-263.

Hasegawa, M., H. Kishino, and T. A. Yano. 1985. Dating of the human ape splitting by a molecular clock of mitochondrial-DNA. *Journal of Molecular Evolution* 22:160-174.

Hasegawa, M., and T. Hashimoto. 1993. Ribosomal-RNA trees misleading. *Nature* 361:23-23.

Heard, S. B. 1992. Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. *Evolution* 46:1818-1826.

Heard, S. B., and A. O. Mooers. 1996. Imperfect information and the balance of cladograms and phenograms. *Systematic Biology* 45:115-118.

Heard, S. B., and A. O. Mooers. 2002. Signatures of random and selective mass extinctions in phylogenetic tree balance. *Systematic Biology* 51:889-897.

Hedtke, S. M., T. M. Townsend, and D. M. Hillis. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Systematic Biology* 55:522-529.

Hendy, M. D., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. *Systematic Zoology* 38:297-309.

Hibbett, D. 2004. Trends in morphological evolution in homobasidiomycetes inferred using maximum likelihood: a comparison of binary and multistate approaches. *Systematic Biology* 53:889-903.

Hillis, D. M., M. W. Allard, and M. M. Miyamoto. 1993. Analysis of DNA sequence data: phylogenetic inference. *Methods in Enzymology* 224:456-487.

Hillis, D. M., J. P. Huelsenbeck, and C. W. Cunningham. 1994a. Application and accuracy of molecular phylogenies. *Science* 264:671-677.

Hillis, D. M., J. P. Huelsenbeck, and D. L. Swofford. 1994b. Hobgoblin of phylogenetics. *Nature* 369:363-364.

Hillis, D. M. 1995. Approaches for assessing phylogenetic accuracy. *Systematic Biology* 44:3-16.

Hillis, D. M. 1996. Inferring complex phylogenies. *Nature* 383:130-131.

Hillis, D. M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Systematic Biology* 47:3-8.

Hillis, D. M. 1999. Phylogenetics and the study of HIV. Pages 105-121 in *Molecular Evolution of HIV* (K. A. Crandall, ed.) Johns Hopkins University Press, Baltimore.

Hillis, D. M., D. D. Pollock, J. A. McGuire, and D. J. Zwickl. 2003. Is sparse taxon sampling a problem for phylogenetic inference? *Systematic Biology* 52:124-126.

Ho, S. Y. W. 2007. Calibrating molecular estimates of substitution rates and divergence times in birds. *Journal of Avian Biology* 38:409-414.

Holland, B. R., D. Penny, and M. D. Hendy. 2003. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock - A simulation study. *Systematic Biology* 52:229-238.

Holman, E. W. 2005. Nodes in phylogenetic trees: The relation between imbalance and number of descendent species. *Systematic Biology* 54:895-899.

Hoyle, D. C., and P. G. Higgs. 2003. Factors affecting the errors in the estimation of evolutionary distances between sequences. *Molecular Biology and Evolution* 20:1-9.

Huelsenbeck, J. P., and D. M. Hillis. 1993. Success of phylogenetic methods in the 4-taxon case. *Systematic Biology* 42:247-264.

Huelsenbeck, J. P. 1995. Performance of phylogenetic methods in simulation. *Systematic Biology* 44:17-48.

Huelsenbeck, J. P., and M. Kirkpatrick. 1996. Do phylogenetic methods produce trees with biased shapes? *Evolution* 50:1418-1424.

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

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

Huelsenbeck, J. P., and K. M. Lander. 2003. Frequent inconsistency of parsimony under a simple model of cladogenesis. *Systematic Biology* 52:641-648.

Huelsenbeck, J., and B. Rannala. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Systematic Biology* 53:904-913.

Hug, L. A., and A. J. Roger. 2007. The impact of fossils and taxon sampling on ancient molecular dating analyses. *Molecular Biology and Evolution* 24:1889-1897.

Hugall, A. F., and M. S. Y. Lee. 2007. The likelihood node density effect and consequences for evolutionary studies of molecular rates. *Evolution* 61:2293-2307.

Hugall, A. F., R. Foster, and M. S. Lee. 2007. Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene RAG-1. *Systematic Biology* 56:543-63.

Jermiin, L. S., S. Y. W. Ho, F. Ababneh, J. Robinson, and A. W. D. Larkum. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Systematic Biology* 53:638-643.

Johnson, K. P. 2001. Taxon sampling and the phylogenetic position of passeriformes: Evidence from 916 avian cytochrome b sequences. *Systematic Biology* 50:128-136.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. In *Mammalian Protein Metabolism* (H. N. Munro, ed.) Academic Press, New York.

Kim, J. 1998. Large-scale phylogenies and measuring the performance of phylogenetic estimators. *Systematic Biology* 47:43-60.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16:111-20.

Kirkpatrick, M., and M. Slatkin. 1993. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution* 47:1171-1181.

Kishino, H., and M. Hasegawa. 1990. Converting distance to time: application to human evolution. *Methods in Enzymology* 183:550-70.

Kishino, H., J. L. Thorne, and W. J. Bruno. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular Biology and Evolution* 18:352-361.

Kolaczkowski, B., and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431:980-984.

Kolaczkowski, B., and J. W. Thornton. 2008. A mixed branch length model of heterotachy improves phylogenetic accuracy. *Molecular Biology and Evolution* In Press.

Kuhner, M. K., and J. Felsenstein. 1994. Simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution* 11:459-468.

Langley, C. H., and W. M. Fitch. 1974. An examination of the constancy of the rate of molecular evolution. *Journal of Molecular Evolution* 3:161-77.

Lartillot, N., H. Brinkmann, and H. Philippe. 2007. Suppression of long-branch attraction artifacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology* 7 Suppl 1:S4.

Lecointre, G., H. Philippe, H. L. Van Le, and H. Le Guyader. 1993. Species sampling has a major impact on phylogenetic inference. *Molecular Phylogenetics and Evolution* 2:205-224.

Leipe, D. D., J. H. Gunderson, T. A. Nerad, and M. L. Sogin. 1993. Small subunit ribosomal RNA+ of *Hexamita inflata* and the quest for the first branch in the eukaryotic tree. *Molecular and Biochemical Parasitology* 59:41-48.

Lemmon, A. R., and E. C. Moriarty. 2004. The importance of proper model assumption in Bayesian phylogenetics. *Systematic Biology* 53:265-277.

Lepage, T., S. Lawi, P. Tupper, and D. Bryant. 2006. Continuous and tractable models for the variation of evolutionary rates. *Mathematical Biosciences* 199:216-233.

Lepage, T., D. Bryant, H. Philippe, and N. Lartillot. 2007. A general comparison of relaxed molecular clock models. *Molecular Biology and Evolution* 24:2669-80.

Li, P., and J. Bousquet. 1992. Relative-rate test for nucleotide substitutions between two lineages. *Molecular Biology and Evolution* 9:1185-1189.

Lin, Y. H., P. A. McLenachan, A. R. Gore, M. J. Phillips, R. Ota, M. D. Hendy, and D. Penny. 2002. Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. *Molecular Biology and Evolution* 19:2060-2070.

Lockhart, P. J., C. J. Howe, D. A. Bryant, T. J. Beanland, and A. W. D. Larkum. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. *Journal of Molecular Evolution* 34:153-162.

Lockhart, P. J., A. W. D. Larkum, M. A. Steel, P. J. Waddell, and D. Penny. 1996. Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. *Proceedings of the National Academy of Sciences of the United States of America* 93:1930-1934.

Lopez, P., D. Casane, and H. Philippe. 2002. Heterotachy, an important process of protein evolution. *Molecular Biology and Evolution* 19:1-7.

Maddison, W. P., M. J. Donoghue, and D. R. Maddison. 1984. Outgroup analysis and parsimony. *Systematic Zoology* 33:83-103.

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

Maddison, W. P., and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology* 55:21-30.

McPeek, M. A., and J. M. Brown. 2007. Clade age and not diversification rate explains species richness among animal taxa. *American Naturalist* 169:E97-E106.

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

Mooers, A. O. 1995. Tree balance and tree completeness. *Evolution* 49:379-384.

Mooers, A. O., R. D. M. Page, A. Purvis, and P. H. Harvey. 1995. Phylogenetic noise leads to unbalanced cladistic tree reconstructions. *Systematic Biology* 44:332-342.

Mooers, A. O., and S. B. Heard. 1997. Evolutionary process from phylogenetic tree shape. *Quarterly Review of Biology* 72:31-54.

Mooers, A. O., and E. C. Holmes. 2000. The evolution of base composition and phylogenetic inference. *Trends in Ecology and Evolution* 15:365-369.

Mooers, A. O., S. B. Heard, and E. Chrostowski. 2005. Evolutionary heritage as a metric for conservation. Pages 120-138 in *Phylogeny and Conservation* (A. Purvis, T. Brooks, and J. Gittleman, eds.). Cambridge University Press, New York.

Murphy, W. J., E. Eizirik, S. J. O'Brien, O. Madsen, M. Scally, C. J. Douady, E. Teeling, O. A. Ryder, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* 294:2348-2351.

Near, T. J., and M. J. Sanderson. 2004. Assessing the quality of molecular divergence time estimates by fossil calibrations and fossil-based model selection. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* 359:1477-1483.

Near, T. J., P. A. Meylan, and H. B. Shaffer. 2005. Assessing concordance of fossil calibration points in molecular clock studies: An example using turtles. *American Naturalist* 165:137-146.

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

Nee, S., R. M. May, and P. H. Harvey. 1994b. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* 344:305-411.

Nylander, J. A. A., F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. *Systematic Biology* 53:47-67.

Olsen, G. J. 1987. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. *Cold Spring Harbor Symposia on Quantitative Biology* 52:825-37.

Pagel, M. 1994. Detecting correlated evolution on phylogenies - a general-method for the comparative-analysis of discrete characters. *Proceedings of the Royal Society of London Series B-Biological Sciences* 255:37-45.

Pagel, M. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology* 48:612-622.

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

Philippe, H., and E. Douzery. 1994. The pitfalls of molecular phylogeny based on four species, as illustrated by the cetacea/artiodactyla relationships. *Journal of Mammalian Evolution* 2:133-152.

Philippe, H., Y. Zhou, H. Brinkmann, N. Rodrigue, and F. Delsuc. 2005. Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology* 5:50.

Poe, S. 1998a. The effect of taxonomic sampling on accuracy of phylogeny estimation: Test case of a known phylogeny. *Molecular Biology and Evolution* 15:1086-1090.

Poe, S. 1998b. Sensitivity of phylogeny estimation to taxonomic sampling. *Systematic Biology* 47:18-31.

Poe, S., and D. L. Swofford. 1999. Taxon sampling revisited. *Nature* 398:299-300.

Poe, S. 2003. Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. *Systematic Biology* 52:423-428.

Pollock, D. D., W. R. Taylor, and N. Goldman. 1999. Coevolving protein residues: Maximum likelihood identification and relationship to structure. *Journal of Molecular Biology* 287:187-198.

Pollock, D. D., and W. J. Bruno. 2000. Assessing an unknown evolutionary process: Effect of increasing site-specific knowledge through taxon addition. *Molecular Biology and Evolution* 17:1854-1858.

Pollock, D. D., D. J. Zwickl, J. A. McGuire, and D. M. Hillis. 2002. Increased taxon sampling is advantageous for phylogenetic inference. *Systematic Biology* 51:664-671.

Purvis, A., J. L. Gittleman, and H. K. Luh. 1994. Truth or consequences - effects of phylogenetic accuracy on 2 comparative methods. *Journal of Theoretical Biology* 167:293-300.

Purvis, A., and P. M. Agapow. 2002. Phylogeny imbalance: taxonomic level matters. *Systematic Biology* 51:844-854.

Purvis, A., A. Katzourakis, and P. M. Agapow. 2002. Evaluating phylogenetic tree shape: Two modifications to Fusco & Cronk's method. *Journal of Theoretical Biology* 214:99-103.

Pybus, O. G., and P. H. Harvey. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society of London Series B-Biological Sciences* 267:2267-2272.

Pybus, O. G., A. Rambaut, E. C. Holmes, and P. H. Harvey. 2002. New inferences from tree shape: Numbers of missing taxa and population growth rates. *Systematic Biology* 51:881-888.

Rambaut, A., and L. Bromham. 1998. Estimating divergence dates from molecular sequences. *Molecular Biology and Evolution* 15:442-448.

Rambaut, A. 2002. Phyl-o-gen: Phylogenetic tree simulator package v1.1. http://evolve.zoo.ox.ac.uk/software.html?id=phylogen.

Rannala, B., J. P. Huelsenbeck, Z. Yang, and R. Nielsen. 1998. Taxon sampling and the accuracy of large phylogenies. *Systematic Biology* 47:702-710.

Raup, D. M., S. J. Gould, T. J. M. Schopf, and D. S. Simberloff. 1973. Stochastic-models of phylogeny and the evolution of diversity. *Journal of Geology* 81:525-542.

Ricklefs, R. E., and J. M. Starck. 1996. Applications of phylogenetically independent contrasts: A mixed progress report. *Oikos* 77:167-172.

Ricklefs, R. E. 2006. Global variation in the diversification rate of passerine birds. *Ecology* 87:2468-2478.

Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53:131-147.

Robinson, M., M. Gouy, C. Gautier, and D. Mouchiroud. 1998. Sensitivity of the relative-rate test to taxonomic sampling. *Molecular Biology and Evolution* 15:1091-1098.

Roelants, K., D. J. Gower, M. Wilkinson, S. P. Loader, S. D. Biju, K. Guillaume, L. Moriau, and F. Bossuyt. 2007. Global patterns of diversification in the history of modern amphibians. *Proceedings of the National Academy of Sciences of the United States of America* 104:887-92.

Roger, A. J., and L. A. Hug. 2006. The origin and diversification of eukaryotes: problems with molecular phylogenetics and molecular clock estimation. *Philosophical Transactions of the Royal Society B-Biological Sciences* 361:1039-1054.

Rogers, J. S. 1994. Central moments and probability-distribution of Colless's coefficient of tree Imbalance. *Evolution* 48:2026-2036.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798-804.

Rokas, A., and S. B. Carroll. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Molecular Biology and Evolution* 22:1337-1344.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572-1574.

Rosen, D. E. 1978. Vicariant patterns and historical explanation in biogeography. *Systematic Zoology* 27:159-188.

Rosenberg, M. S., and S. Kumar. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. *Proceedings of the National Academy of Sciences of the United States of America* 98:10751-10756.

Rosenberg, M. S., and S. Kumar. 2003. Taxon sampling, bioinformatics, and phylogenomics. *Systematic Biology* 52:119-124.

Rutschmann, F., T. Eriksson, K. Abu Salim, and E. Conti. 2007. Assessing calibration uncertainty in molecular dating: The assignment of fossils to alternative calibration points. *Systematic Biology* 56:591-608.

Ryan, M. J., and A. S. Rand. 1995. Female responses to ancestral advertisement calls in tungara frogs. *Science* 269:390-392.

Ryan, M. J., and A. S. Rand. 1998. Evoked vocal response in male tungara frogs: pre-existing biases in male responses? *Animal Behavior* 56:1509-1516.

Rzhetsky, A., and T. Sitnikova. 1996. When is it safe to use an oversimplified substitution model in tree-making? *Molecular Biology and Evolution* 13:1255-1265.

Saitou, N., and M. Nei. 1987. The neighbor-joining method - a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406-425.

Salisbury, B. A. 1999. Misinformative characters and phylogeny shape. *Systematic Biology* 48:153-169.

Salisbury, B. A., and J. Kim. 2001. Ancestral state estimation and taxon sampling density. *Systematic Biology* 50:557-564.

Sanderson, M. J., and M. J. Donoghue. 1989. Patterns of variation in levels of homoplasy. *Evolution* 43:1781-1795.

Sanderson, M. J., and M. J. Donoghue. 1994. Shifts in diversification rate with the origin of angiosperms. *Science* 264:1590-1593.

Sanderson, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution* 14:1218-1231.

Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Molecular Biology and Evolution* 19:101-109.

Sanderson, M. J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19:301-302.

Sarich, V. M., and A. C. Wilson. 1973. Generation time and genomic evolution in primates. *Science* 179:1144-7.

Savage, H. M. 1983. The shape of evolution - Systematic tree topology. *Biological Journal of the Linnean Society* 20:225-244.

Savolainen, V., S. B. Heard, M. P. Powell, T. J. Davies, and A. O. Mooers. 2002. Is cladogenesis heritable? *Systematic Biology* 51:835-843.

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

Shao, K. T., and R. R. Sokal. 1990. Tree balance. *Systematic Zoology* 39:266-276.

Shavit, L., D. Penny, M. D. Hendy, and B. R. Holland. 2007. The problem of rooting rapid radiations. *Molecular Biology and Evolution* 24:2400-2411.

Smith, A. B. 1994. Rooting molecular trees - problems and strategies. *Biological Journal of the Linnean Society* 51:279-292.

Smith, A. B., D. Pisani, J. A. Mackenzie-Dodds, B. Stockley, B. L. Webster, and D. T. Littlewood. 2006. Testing the molecular clock: molecular and paleontological estimates of divergence times in the Echinoidea (Echinodermata). *Molecular Biology and Evolution* 23:1832-51.

Sorenson, M. D., E. Oneal, J. Garcia-Moreno, and D. P. Mindell. 2003. More taxa, more characters: the hoatzin problem is still unresolved. *Molecular Biology and Evolution* 20:1484-1498.

Spencer, M., E. Susko, and A. J. Roger. 2005. Likelihood, parsimony, and heterogeneous evolution. *Molecular Biology and Evolution* 22:1161-1164.

Stam, E. 2002. Does imbalance in phylogenies reflect only bias? *Evolution* 56:1292-1295.

Stamatakis, A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688-2690.

Steel, M., L. Szekely, P. L. Erdos, and P. Waddell. 1993. A complete family of phylogenetic invariants for any number of taxa under Kimura 3ST model. *New Zealand Journal of Botany* 31:289-296.

Steel, M. 1994. Recovering a tree from the leaf colourations it generates under a Markov model. *Applied Mathematics Letters* 7:19-23.

Sullivan, J., D. L. Swofford, and G. J. P. Naylor. 1999. The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. *Molecular Biology and Evolution* 16:1347-1356.

Sullivan, J., and D. L. Swofford. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Systematic Biology 50:723-729.

Susko, E., Y. Inagaki, and A. J. Roger. 2004. On inconsistency of the neighbor-joining, least squares, and minimum evolution estimation when substitution processes are incorrectly modeled. *Molecular Biology and Evolution* 21:1629-1642.

Swofford, D. L., J. L. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic Inference. Pages 407-514 in *Molecular Systematics* (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer Associates, Sunderland, Massachusetts.

Swofford, D. L. 1998. PAUP*. Phylogenetic analysis using parsimony (*and other methods), version 4. Sinauer Associates, Sunderland, Massachusetts.

Swofford, D. L., P. J. Waddell, J. P. Huelsenbeck, P. G. Foster, P. O. Lewis, and J. S. Rogers. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Systematic Biology* 50:525-539.

Takezaki, N., A. Rzhetsky, and M. Nei. 1995. Phylogenetic test of the molecular clock and linearized trees. *Molecular Biology and Evolution* 12:823-833.

Tavare, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* 17:57-86.

Thorne, J. L., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* 15:1647-1657.

Thorne, J. L., and H. Kishino. 2002. Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology* 51:689-702.

Venditti, C., A. Meade, and M. Pagel. 2006. Detecting the node-density artifact in phylogeny reconstruction. *Systematic Biology* 55:637-643.

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

Wiens, J. J., J. W. Fetzner, C. L. Parkinson, and T. W. Reeder. 2005. Hylid frog phylogeny and sampling strategies for speciose clades. *Systematic Biology* 54:719-748.

Wu, C. I., and W. H. Li. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proceedings of the National Academy of Sciences of the United States of America* 82:1741-1745.

Xia, X. 2006. Topological bias in distance-based phylogenetic methods: problems with over- and under-estimated genetic distances. *Evolutionary Bioinformatics* 2:375-387.

Yang, Z. 1994. Statistical properties of the maximum-likelihood method of phylogenetic estimation and comparison with distance matrix-methods. *Systematic Biology* 43:329-342.

Yang, Z., N. Goldman, and A. Friday. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Molecular Biology and Evolution* 11:316-324.

Yang, Z., and D. Roberts. 1995. On the use of nucleic-acid sequences to infer early branchings in the Tree of Life. *Molecular Biology and Evolution* 12:451-458.

Yang, Z. 1996. Phylogenetic analysis using parsimony and likelihood methods. *Journal of Molecular Evolution* 42:294-307.

Yang, Z., and N. Goldman. 1997. Are big trees indeed easy? *Trends in Ecology and Evolution* 12:357-357.

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

Yang, Z. 1997. How often do wrong models produce better phylogenies? *Molecular Biology and Evolution* 14:105-108.

Yang, Z., and A. D. Yoder. 2003. Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Systematic Biology* 52:705-716.

Yang, Z. 2006. *Computational Molecular Evolution*. Oxford University Press, New York.

Yang, Z., and B. Rannala. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Molecular Biology and Evolution* 23:212-226.

Yoder, A. D., and Z. Yang. 2000. Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution* 17:1081-90.

Zharkikh, A., and W. H. Li. 1993. Inconsistency of the maximum-parsimony method - the case of 5 taxa with a molecular clock. *Systematic Biology* 42:113-125.

Zhou, Y., and E. C. Holmes. 2007. Bayesian estimates of the evolutionary rate and age of hepatitis B virus. *Journal of Molecular Evolution* 65:197-205.

Zuckerkandl, E., and L. Pauling. 1962. Molecular disease, evolution, and genetic heterogeneity. Pages 189-225 in *Horizons in Biochemistry* (M. Kasha, and B. Pullman, eds.). Academic Press, New York.

Zwickl, D. J., and D. M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology* 51:588-598.

Zwickl, D. J. 2006. Genetic algorithm approaches for the phylogenetic analyses of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas at Austin. www.bio.utexas.edu/faculty/antisense/garli/Garli.html.

Vita

Tracy Anne Heath and her family lived in San Carlos, AZ, Albuquerque, NM, Ada, OK, and Santa Fe, NM before moving to Gaithersburg, MD in 1989. After graduating from Gaithersburg High School in 1996, Tracy attended Boston University where she majored in Biology. Tracy earned a Bachelor of the Arts degree in May of 2000. Upon graduation from B.U., Tracy was hired as a research assistant in the laboratory of Christopher Schneider in the Boston University biology department where she worked on a phylogeographic study of Cameroonian skinks. In the following fall, Tracy was hired to manage the core DNA sequencing facility in the Boston University Biology department. In this position, she conducted research under the supervision of Michael Sorenson on the systematics of waterfowl. In the fall of 2001, Tracy enrolled in the PhD program in Evolution, Ecology and Behavior at the University of Texas at Austin under the supervision of David Hillis. Shortly after arriving at UT, Tracy's interests shifted from lizard phylogeography to computational phylogenetics and she began pursuing research in this area.

Permanent address: 3928 Garcia Street NE, Albuquerque, NM 87111

This dissertation was typed by the author.