many systematists who consider compositional stability to be important.

ACKNOWLEDGMENTS

We thank J. Anderson, M. J. Donoghue, T. Eriksson, M. Härlin, M. Lee, and M. Thollesson for their helpful comments on the manuscript.

REFERENCES

Bertrand, Y., and M. Härlin. 2006. Stability and universality in the application of taxon names in phylogenetic nomenclature. Syst. Biol. 55:848–858.

Bininda-Emonds, O. R. P. 2004. The evolution of supertrees. Trends Ecol. Evol. 19:315–322.

Bryant, H. N., and P. D. Cantino. 2002. A review of criticisms of phylogenetic nomenclature: Is taxonomic freedom the fundamental issue? Biol. Rev. 77:39–55.

Cantino, P. D., and K. de Queiroz. 2007. International Code of Phylogenetic Nomenclature, Version 4b. Available at http://www.ohiou.edu/phylocode.

Cantino, P. D., and R. G. Olmstead. 2004. Phylogenetic nomenclature of Lamiaceae. Abstracts of the First International Phylogenetic Nomenclature Meeting (Paris): 13. Available at http://www.ohiou.edu/phylocode/events.html.

Cantino, P. D., R. G. Olmstead, and S. J. Wagstaff. 1997. A comparison of phylogenetic nomenclature with the current system: A botanical case study. Syst. Biol. 46:313–331.

Cantino, P. D., and R. W. Sanders. 1986. Subfamilial classification of Labiatae. Syst. Bot. 11:163–185.

de Queiroz, K., and J. Gauthier. 1990. Phylogeny as a central principle in taxonomy: Phylogenetic definitions of taxon names. Syst. Zool. 39:307–322.

de Queiroz, K., and J. Gauthier. 1992. Phylogenetic taxonomy. Annu. Rev. Ecol. Syst. 23:449–480.

Driskell, A. C., A. Cécile, J. G. Burleigh, M. M. McMahon, B. D. O'Meara, and M. J. Sanderson. 2004. Prospects for building the tree of life from large sequence databases. Science 306:1172–1174.

Hibbett, D. S., R. H. Nilsson, M. Snyder, M. Fonseca, J. Costanzo, and M. Shonfeld. 2005. Automated phylogenetic taxonomy: An example in the Homobasidiomcetes (mushroom-forming fungi). Syst. Biol. 54:660–668.

Kaufmann, M., and M. Wink. 1994. Molecular systematics of the Nepetoideae (Family Labiatae): Phylogenetic implications from *rbcL* gene sequences. Z. Naturforsch. 49c:635–645.

Sanderson, M. J., A. Purvis, and C. Henze. 1998. Phylogenetic supertrees: Assembling the trees of life. Trends Ecol. Evol. 13:105–109.

Wagstaff, S. J., and R. G. Olmstead. 1997. Phylogeny of Labiatae and Verbenaceae inferred from *rbcL* sequences. Syst. Bot. 22:165–179.

First submitted 15 January 2007; reviews returned 1 May 2007; final acceptance 24 October 2007
Associate Editor: Michael Lee

Syst. Biol. 57(1):160–166, 2008

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701884640

Taxon Sampling Affects Inferences of Macroevolutionary Processes from Phylogenetic Trees

TRACY A. HEATH,[1] DERRICK J. ZWICKL,[1,3] JUNHYONG KIM,[2] AND DAVID M. HILLIS[1]

[1]*Section of Integrative Biology and Center for Computational Biology and Bioinformatics, University of Texas at Austin, Austin, Texas 78712, USA; E-mail: tracyh@mail.utexas.edu (T.A.H.)*
[2]*Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA*
[3]*Current Address: National Evolutionary Synthesis Center, Durham, North Carolina 27705, USA*

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 we can detect the patterns resulting from the evolutionary processes that shape trees. These patterns can be obscured by nonbiological 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 nonbiological 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., 1994; 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 nonrandom 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, we investigated the influence of varying levels of random taxon sampling on phylogenetic tree imbalance. We 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. We 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.

METHODS

Simulations

We simulated non-ERM trees 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. We then enforced a gamma-distributed prior on speciation and extinction rates 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. We show by simulation that increasing the shape parameters results in a decrease in the diversification rate variation and produces more balanced topologies (Fig. 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 the ERM or PDA models, because 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). Our 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).

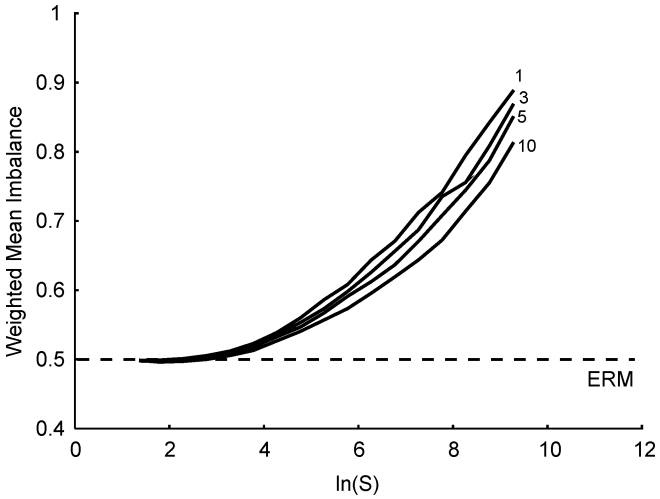

FIGURE 1. The functional relationship between weighted mean imbalance and ln(node size) for four sets of trees simulated under a range of variance parameters. The parameter, alpha, of the gamma-distributed rate prior was changed for each set of simulations to 1, 3, 5, and 10 (for both the speciation 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 (equal rates Markov model; dashed line).

The above method for simulating tree topologies was implemented by modifying code from the program *Phylo-gen* (Rambaut, 2002; the modified code is available from the authors). We 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 (Fig. 1). We also generated trees under constant speciation and extinction rates (ERM model) and the proportional-to-distinguishable arrangements (PDA) model.

Empirical Phylogenies

We collected trees from recently published studies of empirical data (online Appendix 1; http://systematicbiology.org). When surveying the literature, we 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 we selected the tree based on the combined analysis. When we encountered more than one study on a particular taxonomic group, we selected the most recently published tree. The trees in our 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), we only used trees with species as terminal taxa so that we could directly calculate the amount of species-level sampling and avoid subjective aspects of higher level taxonomic grouping. We determined the proportion of taxon sampling based on the number of described species in the group. Our 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. We then sorted the empirical phylogenies based on the proportion of taxon sampling and the method used to reconstruct the tree. In this study, we only present the imbalance of phylogenies with sampling densities lower than 10% because our collection of published studies contained relatively few trees with more complete species sampling.

Measure of Imbalance

We 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

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

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:

$$\begin{aligned} &\text{if } S \text{ is odd, } w = 1,\\ &\text{if } S \text{ is even, and } I > 0,\ w = (S-1)/S,\\ &\text{if } S \text{ is even, and } I = 0,\ w = 2(S-1)/S. \end{aligned}$$

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; 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; 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 multifurcating 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, we use I_w to determine the impact of reduced species sampling by comparing the imbalance of complete trees with that of incomplete trees.

Results and Discussion

The Effect of Node Size on Tree Imbalance

The nodal weighted mean imbalance for the empirical trees is summarized in Fig. 2. We 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, we did not detect a significant difference in the pattern of imbalance between trees reconstructed under maximum parsimony versus those reconstructed using parametric methods (Fig. 2). Although there appear 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, we combined the trees into a single set of empirical phylogenies for our subsequent analyses. When combining the trees, if a single paper presented both a parsimony tree and a maximum likelihood or Bayesian tree, we selected the tree at random. This combined collection of trees consisted of 77 parsimony trees and 78 maximum likelihood/Bayesian trees.

Figure 3 shows the weighted mean imbalance of our combined collection of empirical trees and a set of trees simulated under our model of varying speciation and extinction rates (where $\alpha = 2$ for the gamma-distributed rate priors for both speciation and extinction rates). We also show the imbalance expected under the ERM and PDA models. Although we used a different collection of empirical trees than used in previous studies (Purvis and Agapow, 2002; Holman, 2005; Blum and François, 2006), our 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 (Fig. 3). The trees simulated under our model of 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 our model show a positive association between

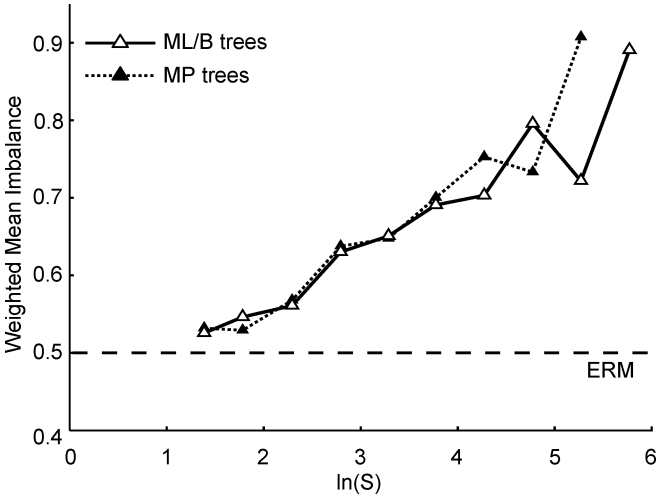

FIGURE 2. The weighted mean imbalance of empirical trees plotted as a function of the natural log of the node size (S). The dashed line at 0.5 indicates the imbalance expected under the ERM model. One hundred and twenty-four trees reconstructed using maximum parsimony (MP) are indicated by the dotted 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.

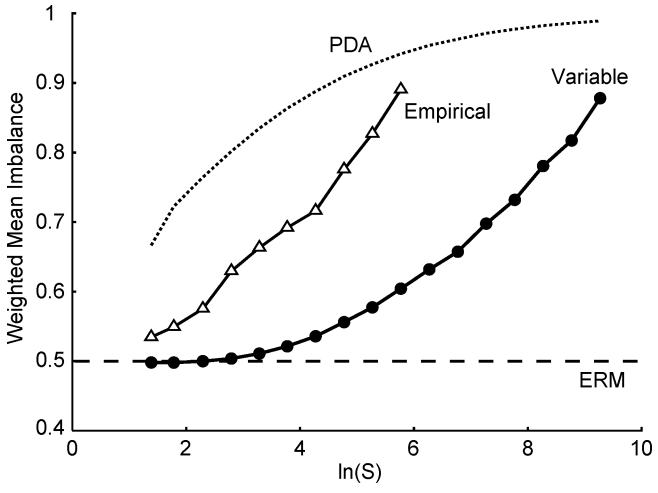

FIGURE 3. The nodal imbalance for the combined collection of empirical trees (triangles; 157 total trees) and the collection of trees simulated under varying rates of speciation and extinction (circles). The upper dotted line represents the imbalance expected for trees generated under the PDA model (proportional-to-distinguishable arrangements model) and the dashed line at 0.5 indicates the imbalance expected under the ERM model.

node size and node age, as well as a positive correlation between node age and imbalance.

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), our collection of empirical trees all had low percentages of sampled taxa because we 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 we randomly pruned taxa from the trees simulated with variable and autocorrelated speciation/extinction rates, we observed an increase in nodal imbalance and a very good approximation of the imbalance found in the empirical trees (Fig. 4). In contrast, we 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 (Fig. 5). 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.

We randomly pruned 50% of the taxa from trees in our combined set of empirical phylogenies to determine whether or not an additional reduction in taxon sampling would increase the imbalance in empirical phylogenies (Fig. 6). The results shown in Fig. 6 are from 100 replicates of randomized pruning and suggest that, on average, random reduced taxon sampling does indeed increase the imbalance in these trees.

Our results indicate that incomplete taxon sampling in the presence of diversification rate variation may be sufficient to explain much of the imbalance observed in our collection of empirical trees, because as species are removed from a phylogeny, the apparent variation in the rates of diversification is increased. Our 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

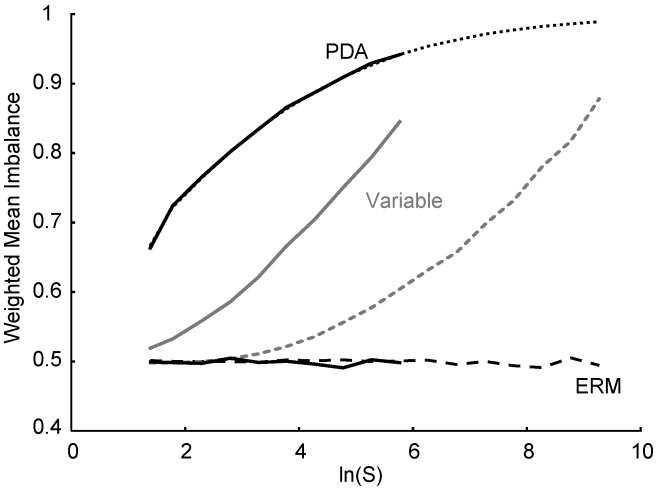

FIGURE 5. Weighted mean imbalance as a function of the natural log of the node size for trees simulated under the PDA model (black), the ERM model (black), and variable rates model (gray). The sets of trees with 100% taxon sampling are indicated by dashed lines. Sets of trees with 3% taxon sampling are 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.

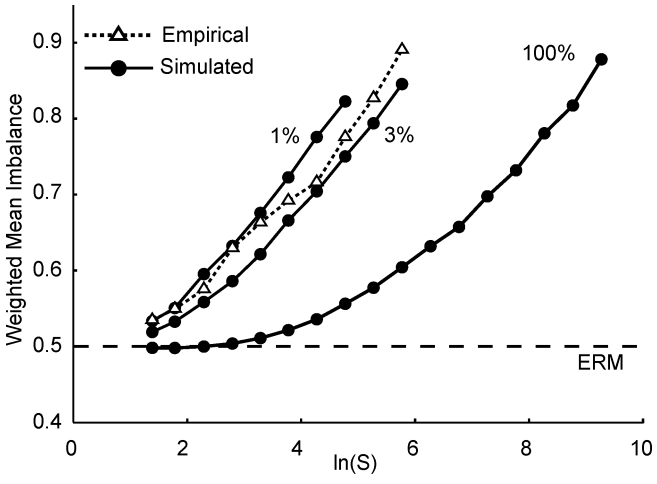

FIGURE 4. Weighted mean imbalance for empirical trees (dotted line/triangles) and trees simulated under varying rates with different levels of taxon sampling (solid line/circles). The simulated trees were reduced to 3% and 1% taxon sampling. The dashed line at 0.5 indicates the imbalance expected for trees generated under the ERM model.

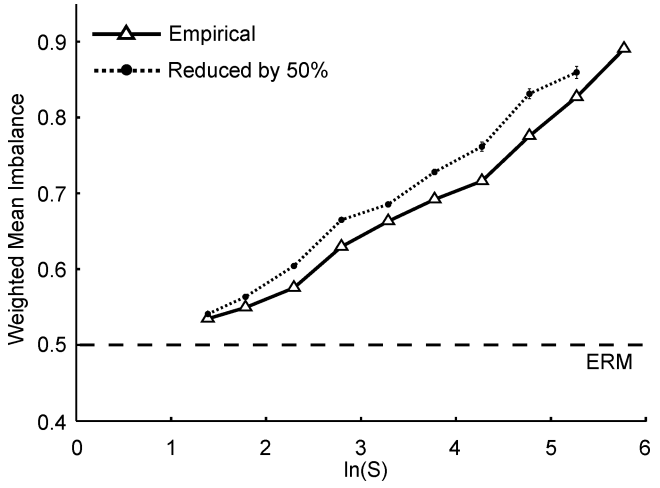

FIGURE 6. The weighted mean imbalance of empirical trees with reduced taxon sampling. The imbalance of the published phylogenies without a reduction in taxon sampling is represented by the solid line. The dotted line indicates the same set of trees with a 50% reduction in taxa averaged over 100 replicates with standard error bars. The dashed line at 0.5 indicates the imbalance expected under the ERM model.

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. Our 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 our model can be adjusted to produce varying levels of tree imbalance (Fig. 1), the general conclusions of our simulations remain consistent across a wide range of parameter values. Our 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 (Fig. 4). 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.

CONCLUSIONS

Variation in the relative rates of speciation and extinction produces tree topologies with greater imbalance than trees generated under the equal rates model (Fig. 3). 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 (Fig. 4), 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 nonbiological 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 nonbiological 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.

ACKNOWLEDGMENTS

We thank Vincent Savolainen, Rod Page, 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). 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. Computational resources were provided by the Texas Advanced Computing Center (TACC) at the University of Texas at Austin (http://www.tacc.utexas.edu).

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. Syst. Biol. 51:866–872.

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

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

DeBry, R. W. 2005. The systematic component of phylogenetic error as a function of taxonomic sampling under parsimony. Syst. Biol. 54:432–440.

Dial, K. P., and J. M. Marzluff. 1989. Nonrandom diversification within taxonomic assemblages. Syst. Zool. 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.

Fusco, G., and Q. C. B. Cronk. 1995. A new method for evaluating the shape of large phylogenies. J. Theor. Biol. 175:235–243.

Good-Avila, S. V., V. Souza, B. S. Gaut, and L. E. Eguiarte. 2006. Timing and rate of speciation in Agave (Agavaceae). Proc. Natl. Acad. Sci. USA 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.

Graybeal, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. 47:9–17.

Guyer, C., and J. B. Slowinski. 1991. Comparisons of observed phylogenetic topologies with null expectations among 3 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.

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. Syst. Biol. 45:115–118.
Heard, S. B., and A. O. Mooers. 2002. Signatures of random and selective mass extinctions in phylogenetic tree balance. Syst. Biol. 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. Syst. Biol. 55:522–529.
Hendy, M. D., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297–309.
Hillis, D. M. 1996. Inferring complex phylogenies. Nature 383:130–131.
Hillis, D. M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. 47:3–8.
Hillis, D. M., D. D. Pollock, J. A. McGuire, and D. J. Zwickl. 2003. Is sparse taxon sampling a problem for phylogenetic inference? Syst. Biol. 52:124–126.
Holman, E. W. 2005. Nodes in phylogenetic trees: The relation between imbalance and number of descendent species. Syst. Biol. 54:895–899.
Huelsenbeck, J. P., and M. Kirkpatrick. 1996. Do phylogenetic methods produce trees with biased shapes? Evolution 50:1418–1424.
Huelsenbeck, J. P., and K. M. Lander. 2003. Frequent inconsistency of parsimony under a simple model of cladogenesis. Syst. Biol. 52:641–648.
Huelsenbeck, J. P., B. Larget, and D. Swofford. 2000. A compound Poisson process for relaxing the molecular clock. Genetics 154:1879–1892.
Kim, J. 1998. Large-scale phylogenies and measuring the performance of phylogenetic estimators. Syst. Biol. 47:43–60.
Kirkpatrick, M., and M. Slatkin. 1993. Searching for evolutionary patterns in the shape of a phylogenetic tree. Evolution 47:1171–1181.
Kishino, H., J. L. Thorne, and W. J. Bruno. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. Mol. Biol. Evol. 18:352–361.
McPeek, M. A., and J. M. Brown. 2007. Clade age and not diversification rate explains species richness among animal taxa. Am. Nat. 169:E97–E106.
Mitter, C., B. Farrell, and B. Wiegmann. 1988. The phylogenetic study of adaptive zones—Has phytophagy promoted insect diversification. Am. Nat. 132:107–128.
Mooers, A. O. 1995. Tree balance and tree completeness. Evolution 49:379–384.
Mooers, A. O., and S. B. Heard. 1997. Evolutionary process from phylogenetic tree shape. Q. Rev. Biol. 72:31–54.
Mooers, A. O., R. D. M. Page, A. Purvis, and P. H. Harvey. 1995. Phylogenetic noise leads to unbalanced cladistic tree reconstructions. Syst. Biol. 44:332–342.
Nee, S., E. C. Holmes, R. M. May, and P. H. Harvey. 1994. Extinction rates can be estimated from molecular phylogenies. Philos. Trans. R Soc. Lond. B Biol. Sci. 344:77–82.
Poe, S. 2003. Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. Syst. Biol. 52:423–428.
Poe, S., and D. L. Swofford. 1999. Taxon sampling revisited. Nature 398:299–300.
Pollock, D. D., D. J. Zwickl, J. A. McGuire, and D. M. Hillis. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51:664–671.
Purvis, A., and P. M. Agapow. 2002. Phylogeny imbalance: Taxonomic level matters. Syst. Biol. 51:844–854.
Purvis, A., A. Katzourakis, and P. M. Agapow. 2002. Evaluating phylogenetic tree shape: Two modifications to Fusco & Cronk's method. J. Theor. Biol. 214:99–103.
Pybus, O. G., and P. H. Harvey. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proc. Roy. Soc. B 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. Syst. Biol. 51:881–888.
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. Syst. Biol. 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. J. Geol. 81:525–542.
Ricklefs, R. E. 2006. Global variation in the diversification rate of passerine birds. Ecology 87:2468–2478.
Robinson, M., M. Gouy, C. Gautier, and D. Mouchiroud. 1998. Sensitivity of the relative-rate test to taxonomic sampling. Mol. Biol. Evol. 15:1091–1098.
Rosen, D. E. 1978. Vicariant patterns and historical explanation in biogeography. Syst. Zool. 27:159–188.
Salisbury, B. A., and J. Kim. 2001. Ancestral state estimation and taxon sampling density. Syst. Biol. 50:557–564.
Sanderson, M. J., and M. J. Donoghue. 1994. Shifts in diversification rate with the origin of angiosperms. Science 264:1590–1593.
Savage, H. M. 1983. The shape of evolution—Systematic tree topology. Biol. J. Linn. Soc. 20:225–244.
Savolainen, V., S. B. Heard, M. P. Powell, T. J. Davies, and A. O. Mooers. 2002. Is cladogenesis heritable? Syst. Biol. 51:835–843.
Shao, K. T., and R. R. Sokal. 1990. Tree balance. Syst. Zool. 39:266–276.
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.
Thorne, J. L., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15:1647–1657.
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. Available at www.bio.utexas.edu/faculty/antisense/garli/Garli.html.
Zwickl, D. J., and D. M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588–598.

First submitted 7 May 2007; reviews returned 9 July 2007;
final acceptance 11 September 2007
Associate Editor: Vincent Savolainen
Editors in Chief: Rod Page and Jack Sullivan