Ecology, 93(8) Supplement, 2012, pp. S4–S13

Synthesizing phylogenetic knowledge for ecological research

Jeremy M. Beaulieu,[1,5] Richard H. Ree,[2] Jeannine Cavender-Bares,[3] George D. Weiblen,[4] and Michael J. Donoghue[1]

[1]*Department of Ecology and Evolutionary Biology, Yale University, P.O. Box 208106, New Haven, Connecticut 06520-8106 USA*
[2]*Department of Botany, Field Museum of Natural History, Chicago, Illinois 60605 USA*
[3]*Department of Ecology, Evolution, and Behavior, University of Minnesota, Saint Paul, Minnesota 55108 USA*
[4]*Department of Plant Biology, University of Minnesota, St. Paul, Minnesota 55108 USA*

Abstract. The demand for knowledge about the tree of life is steadily rising in ecology and other fields, but bioinformatic resources designed to meet these needs remain poorly developed. Ecologists pursuing phylogenetic insights into the organization of communities have come to rely on relatively conservative reference trees that, in general, are poorly resolved and documented. New methods for inferring very large trees by mining data from DNA sequence repositories will undoubtedly be useful in community phylogenetics but are not without limitations. Here we argue that the collective phylogenetic knowledge embodied in the literature of systematics is a valuable resource that can be tapped in assembling synthetic trees. Assembling a composite "literature-based" tree by the judicious grafting of clades is one way to achieve a synthesis of current knowledge, and could, under some circumstances, better represent "what we know" about phylogenetic relationships than results obtained from automated pipelines. We describe an approach and new software for storing and annotating trees from published studies and grafting clades together in a documented and repeatable manner. Using this agglomerative approach, we are in the process of assembling a literature-based tree for land plants, which presently contains 14 423 species from over 259 sources. For this strategy to be maximally effective, improvements in digital infrastructure are needed to capture and deploy advances in phylogenetic knowledge as they are published.

Key words: *community phylogenetics; Phylografter; Phylomatic; supermatrix; supertree.*

Introduction

Phylogenetic thinking has quickly become integrated into community ecology, but understanding the influence of evolutionary history on ecological patterns in space and time is hindered by the difficulty of assembling relevant phylogenetic information. Major challenges include identifying an appropriate hypothesis of relationships among clades (tree topology) and estimating temporal divergence or trait divergence (branch lengths). Ecologists working with seed plants have often relied on Phylomatic (Webb and Donoghue 2005; information *available online*),[6] which uses a reference tree that attempts to summarize current knowledge of higher level relationships (e.g., APG 2009), but provides little or no resolution of relationships among closely related taxa. Many authors have manually resolved the tips of the Phylomatic trees, and analyses using Phylomatic have yielded some important insights (e.g., Moles et al. 2005, Cavender-Bares et al. 2006, Kembel and Hubbell 2006, Strauss et al. 2006, Swenson et al. 2007, Willis et al. 2008, Kraft and Ackerly 2010). Nevertheless, for many purposes, the coarse level of phylogenetic resolution it provides and/or the lack of transparency in how ecologists manually resolve relationships is unsatisfactory. Here we explore alternative approaches to synthesizing phylogenetic knowledge for ecological uses, focusing specifically on the description of an agglomerative approach to the problem of tree topology.

Manuscript received 7 April 2011; revised 2 November 2011; accepted 15 November 2011. Corresponding Editor (ad hoc): K. Kozak. For reprints of this Special Issue, see footnote 1, p. S1.

[5] Email: jeremy.beaulieu@yale.edu

[6] http://www.phylodiversity.net/phylomatic/

Background

Several different approaches to the problem of synthesizing phylogenetic knowledge have been developed. One is the use of "supertree" methods (e.g., Sanderson et al. 1998, Bininda-Emonds 2004, Bansal et al. 2010), which assemble larger trees from smaller ones, generally by decomposing the latter into matrix representations that are then combined and analyzed using heuristic tree search methods. A supertree for mammals has been used in a number of comparative analyses (e.g., Jones et al. 2002, Ruta et al. 2003, Bininda-Emonds et al. 2007). In practice, however, supertree methods have tended to yield fairly unresolved trees owing to conflicts among the component trees (Cotton and Wilkinson 2007, McMorris and Wilkinson 2011). Some conflicts may reflect areas of genuine disagreement, but many are probably artifacts of taxon sampling given that individual studies were not designed with the goal of eventual

combination in mind. For example, small differences in the placement of poorly sampled, and often arbitrarily rooted, outgroup taxa can cause supertrees to collapse (Bininda-Emonds et al. 2005). Building a meaningful supertree requires considerable decision-making at the outset about taxon samples in the input trees, which may be challenging for nonexperts.

Another approach is to analyze a "supermatrix" of molecular sequence data assembled specifically for the species of interest (i.e., Driskell et al. 2004, McMahon and Sanderson 2006, Cadotte et al. 2008, Dunn et al. 2008, Sanderson et al. 2008, Smith et al. 2009). This is becoming increasingly popular owing to the development of automated pipelines for mining DNA sequences in GenBank (PhyLoTA, Sanderson et al. 2008; PHLAWD, Smith et al. 2009). Such pipelines make the rapid assembly of large concatenated alignments of one or more relevant genes possible. The process is further aided by improvements in tree search algorithms and access to computational power enabling analysis of increasingly large data sets (e.g., Sanderson et al. 2008, Stamatakis et al. 2008, Goloboff et al. 2009, Smith et al. 2009). For example, Smith et al. (2011) recently estimated a tree for 55 473 species of seed plants. Such large trees are now being put to use in studies of evolution, ecology, and global change (e.g., Smith and Donoghue 2008, Smith and Beaulieu 2009, Beaulieu et al. 2010, Edwards and Smith 2010, Goldberg et al. 2010, Thuiller et al. 2011). Kress et al. (2009) proposed a variant of the supermatrix approach that they argued would be especially useful for community ecologists; namely, making use of DNA barcode sequences obtained from a local species assemblage of interest (in their case, the woody plants of a forest plot on Barro Colorado Island, Panama).

Although each of these approaches has merit and might be well suited to particular circumstances, they also can be problematic. In studies that involve a large number of species, possibly spanning several different communities or regions, some (perhaps many) of the species may not have sequences in GenBank or barcodes (Ratnasingham and Hebert 2007; see Barcode of Life Data Systems [BOLD], information *available online*).[7] This will be especially true for communities and organisms where our current taxonomic knowledge is relatively limited, as in the tropics. Even for well-known systems, many species have yet to be sequenced. For example, GenBank now generally contains sequences for fewer than 20% of the species in most major angiosperm clades (Smith et al. 2011).

Furthermore, the sequence data that happen to be available may not resolve relationships adequately, accurately, or with confidence. Because DNA sequences, in general, are not generated with phylogenetic synthesis in mind, supermatrices assembled from sequence repositories generally contain very large amounts of missing data (e.g., 95% in McMahon and Sanderson [2006], 91% in Smith et al. [2009], 93% in Thomson and Shaffer [2010]). Although the absolute amount of missing data per se may not be problematical (Wiens 2005, Wiens and Moen 2008, Pyron et al. 2011, Wiens and Morrill 2011), the nonrandom distribution of missing data with respect to clades (i.e., particular genes being well sampled for some taxa, but not at all for others) can yield analytical inconsistencies, such as "rogue taxa" (Sanderson and Shaffer 2002), and render it nearly impossible to accurately resolve many nodes (Sanderson et al. 2010). Nonrandom patterns in missing data can also adversely affect the estimation of branch lengths (Lemmon et al. 2009, but see Wiens and Morrill 2011), with potentially important implications for comparative analyses. The failure of supertree and supermatrix approaches to identify particular clades that are otherwise well supported by evidence from numerous independent studies (e.g., McMahon and Sanderson 2006, Thomson and Shaffer 2010, Smith et al. 2011), frustrates the synthesis of phylogenetic knowledge in general.

Finally, it is important for the user community to appreciate that automated data-mining approaches remain imperfect, and these can yield problematic data sets and results (e.g., see Phillipe et al. [2011] for a critique of Dunn et al. [2008]). As an example, it was noted after the fact that in the matrix underlying the Smith et al. (2011) analysis, which was generated using PHLAWD (Smith et al. 2009), many species inadvertently contained extra *mat*K sequences attached to the ends of the *trn*K intron (Smith et al. 2012). Such errors could potentially compromise phylogenetic accuracy and could go undetected in the absence of expert knowledge.

A "Graft" Alternative

If the approaches outlined in the previous section do not yet guarantee a reliable topology for the species of interest, are there other options for producing comprehensive phylogenies for community ecology? An approach that has been employed (e.g., Weiblen et al. 2000, 2006, Ree and Donoghue 1999, Beaulieu et al. 2007, Novotny et al. 2010) is simply to piece together, by hand, trees from the "best" studies in the literature. This is effectively the approach that underlies the Tree of Life Web Project (Maddison and Maddison 1996, Maddison and Schulz 2007), and, for plants, Phylomatic (Webb and Donoghue 2005), as well as the APWeb (Stevens 2011). A further variation, TimeTree (Hedges et al. 2006), stitches together independently published estimates of divergence times for major clades according to the GenBank taxonomic hierarchy. The basic idea is to integrate findings from independent studies on particular clades with more broadly inclusive phylogenetic hypotheses, and to document the carefully considered choices necessary to combine expert knowledge in this way. There are, after all, thousands of published

[7] http://www.boldsystems.org

systematic studies, each one presumably representing a conscientious, targeted attempt to understand relationships in a particular part of the tree of life. If there were a good way to piece these individual trees together, and easily update them as new results become available, we could build a synthetic tree that effectively summarizes "what we think we (collectively) know." This would serve as a valuable resource for comparative biology and a reference point for downstream studies of all sorts.

There are reasons to believe that such an agglomerative tree-building approach may, in some cases if not generally, provide phylogenetic hypotheses with greater accuracy and confidence than, for example, automated mining of sequences from GenBank. For example, one might expect that a tree published in the systematics literature represents a judicious choice of taxa and characters, selected by a specialist in the organisms of interest specifically to solve a particular phylogenetic problem. Automated methods can be expected to solve such problems if the taxonomic coverage and phylogenetic signal of available data are sufficient, but this is not always the case (Sanderson et al. 2008). Trees from the systematics literature may be grounded in more phylogentically informative characters per taxon than is the case for supermatrices. These considerations are important given the potential for erroneous inferences to be drawn from incomplete samples (e.g., Zwickl and Hillis 2002). On the other hand, the grafting approach does not resolve conflict among studies and performs best when studies are structured hierarchically with nested placeholder taxa. In the end, the question is: How do we best integrate systematic knowledge in an accurate and repeatable manner for ecological research?

It is not our intention here to definitively answer this question, and certainly careful attention to the potential pitfalls can result in a reliable tree generated using automated methods. But, in the spirit of keeping our analytical options open, we describe a workflow for synthesizing phylogenetic knowledge in an annotated, transparent, and repeatable manner. We also describe a new web application, Phylografter, which aims to facilitate this process by allowing users to collaboratively upload, store, and graft together published trees. We believe that this approach could be useful to community ecologists under some circumstances. However, even if it proves to be of limited use in this context, it will be valuable as a means of summarizing current understanding and identifying gaps in phylogenetic knowledge. Of course, it will also immediately be useful as a way to update and expand the reference tree that underpins Phylomatic (Webb and Donoghue 2005; *available online*, see footnote 6) and other such community resources. We illustrate an application of this workflow by constructing a literature-based tree for 14 423 species of land plants from 259 published sources. The process of compiling a synthetic tree at this scale highlights a number of phyloinformatic challenges and opportunities.

Assembling a Literature-based Tree for Land Plants

The example described here was motivated by the need for phylogenetic information on the land plants recorded in vegetation plots across the North American Long-Term Ecological Research (LTER) network. Ecologists now routinely apply such information to investigate the phylogenetic distribution of particular species and traits among communities in relation to abiotic and biotic factors at local and regional scales (e.g., Verdú et al. 2009, Fine and Kembel 2011, Swenson et al. 2011, Cavender-Bares and Reich 2012, Helmus and Ives 2012, Knapp et al. 2012). Our example is not the broadest synthesis of land plant relationships published to date (e.g., ca. 55 000 species in Smith et al. 2011), but it follows an alternative approach that, in theory, could be applied to even larger problems. In fact, our tree is intended to be a starting point to be expanded using software.

Our procedure for assembling a large, literature-based tree for North American land plants is, at the outset, similar to that used to produce the master tree used by Phylomatic (Webb and Donoghue 2005), but differs in considering more explicit decision criteria and in documenting the provenance of clades chosen for grafting by reference to the systematic literature. We first obtained a backbone phylogeny from Qiu et al. (2006) for the major land plant groups (i.e., liverworts, mosses, hornworts, lycopods, monilophytes, acrogymnosperms, and angiosperms). We then obtained trees based on more detailed studies of each major land plant clade (i.e., lycopods, Wikstrom and Kenrick 2001; monilophytes, Schuettpelz and Pryer 2008; seed plants, Chaw et al. 2000, Davies et al. 2004) and grafted the ingroup taxa from each study onto the backbone in place of any "exemplar" taxa present in Qiu et al. (2006). This process was repeated recursively by grafting additional published phylogenies that were based on direct analysis of molecular sequence data (i.e., "source trees") to appropriate positions held by one or more placeholders (Fig. 1).

Source trees were obtained directly from TreeBASE (*available online*)[8] or were manually redrawn from published figures. Study choice was based on a set of guidelines for resolving phylogenetic conflicts among studies. First, relationships within a more detailed, more densely sampled study took priority over relationships within a broader, less densely sampled study, and source trees having more taxa and/or based on more informative characters were chosen over those with fewer. In cases where taxon sampling was approximately similar, but character sampling was not (e.g., 50 species and 1 kb of sequence vs. 40 species and 10 kb of sequence), we chose the source tree inferred from more potentially informative characters. In cases where taxon sampling differed substantially among source trees (e.g., 500

[8] http://www.treebase.org

species and 1 kb of sequence vs. 50 species and 10 kb of sequence), we chose the source tree containing more species. Such choices clearly entail judgment calls, each of which was documented. Second, we chose studies inferred from model-based approaches (i.e., maximum likelihood and Bayesian) whenever possible over trees inferred using parsimony or distance methods. Third, we selected consensus trees over any one particular hypothesis when multiple hypotheses were presented in a particular study. That is, we favored a conservative summary of the current state of phylogenetic knowledge over maximally resolved phylogenies for the clades of interest.

It is important to note that these simple rules do not specifically evaluate phylogenetic conflicts among studies or how they might be resolved. As an example, different studies differ on whether *Amborella* and Nymphaeales (water lilies) comprise a clade or a grade at the base of the angiosperms. Although our procedure simply accepted one particular study over others based on the decision criteria outlined in the previous paragraphs, in the future it might be desirable to annotate particular conflicts and incorporate uncertainty about the resolution (perhaps as polytomies). This would provide an additional mechanism to take advantage of expert knowledge and annotation in building a synthetic tree.

Following the procedures outlined in the previous two paragraphs, we produced a provisional synthetic tree consisted of 14 423 species of land plants obtained from 259 published sources (Fig. 2). We view this tree as a work in progress, which can and will change as new studies are published and are grafted into the tree. However, even at this early stage, many major clades are represented by several sources, although a large majority of these sources are studies of flowering plant clades. Angiosperms are represented by 13 595 species, the monilophytes (ferns and fern allies) by 497 species, and the acrogymnosperms (the clade containing the four major extant lineages of "gymnosperms"; Cantino et al. 2007) by 181 species. In the case of the monilophytes, almost all of the relationships are currently from a single study of 473 species of leptosporangiate ferns (Schuettpelz and Pryer 2008). Likewise, for acrogymnosperms, most of the relationships were drawn from particular studies of *Pinus* (Gernandt et al. 2005) and Cupressaceae (Gadek et al. 2000, Little 2006).

Among flowering plants, eudicots (12 457 species) are by far the best represented and a significant portion of the eudicot species belongs to the Campanulidae (5006 species), a clade of perhaps 30 000 species that includes four major lineages: Aquifoliales, Asterales, Apiales, and Dipsacales (Tank and Donoghue 2010). The production of an agglomerative tree for campanulids was an objective of one of us (J. M. Beaulieu) before the current, broader project was conceived. Hence, for campanulids, the current tree provides a fairly complete synthesis of published knowledge of phylogenetic

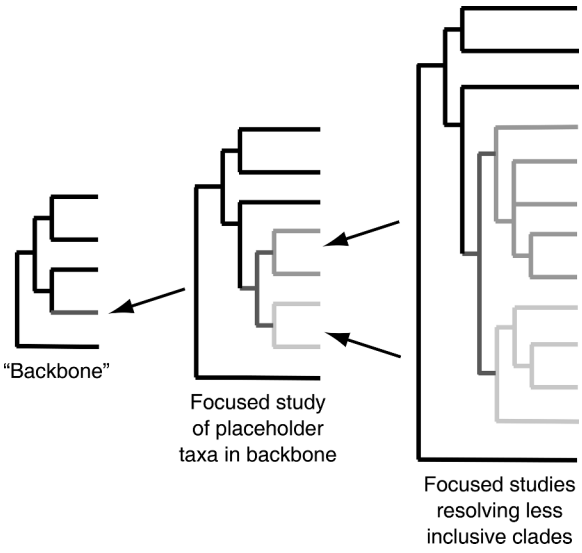

FIG. 1. The basic procedure of the agglomerative approach used to generate a synthetic tree of expert knowledge. First, a backbone tree is obtained depicting relationships among the major clades within the clade of interest (e.g., land plants). Trees based on more focused studies are then grafted onto this backbone tree, replacing placeholder or "exemplar" taxa. The agglomerative process starts with more inclusive clades and proceeds "inward," adding more and more detailed studies of included clades (gray lines).

relationships. Other major lineages will approach this level of coverage as the assembly process continues.

Several observations from campanulids are worth noting. There are some clades where targeted and coordinated efforts have already been undertaken, making it relatively easy to bring together various sources. In Dipsacales, for example, a fairly comprehensive phylogenetic study has recently been carried out for nearly every major clade. However, in other cases, we found no focused studies to replace the exemplar taxa in the "backbone" tree (e.g., Stemonuraceae and Cardiopteridaceae within the Aquifoliales). The attempt to synthesize a literature-based tree for campanulids has the benefit of clearly exposing such knowledge gaps.

In other cases, the uncoordinated nature of the sampling across studies diminishes our ability to assemble a tree. Studies within the Apiaceae provide an example. Here, several studies have resolved various portions of the tree, but have sampled variously overlapping sets of taxa for the same sets of genes (e.g., nuclear ribosomal ITS, *rps*16). By following our assembly rules we were forced, in several cases, to choose one set of taxa over another to represent certain relationships. As a result, portions of the Apiaceae are less represented here than they should be based on the sequences available in GenBank. Indeed, of the more than 1500 species of Apiaceae with relevant sequences in GenBank, fewer than 700 species are currently represented in our "literature-based" tree. This highlights a

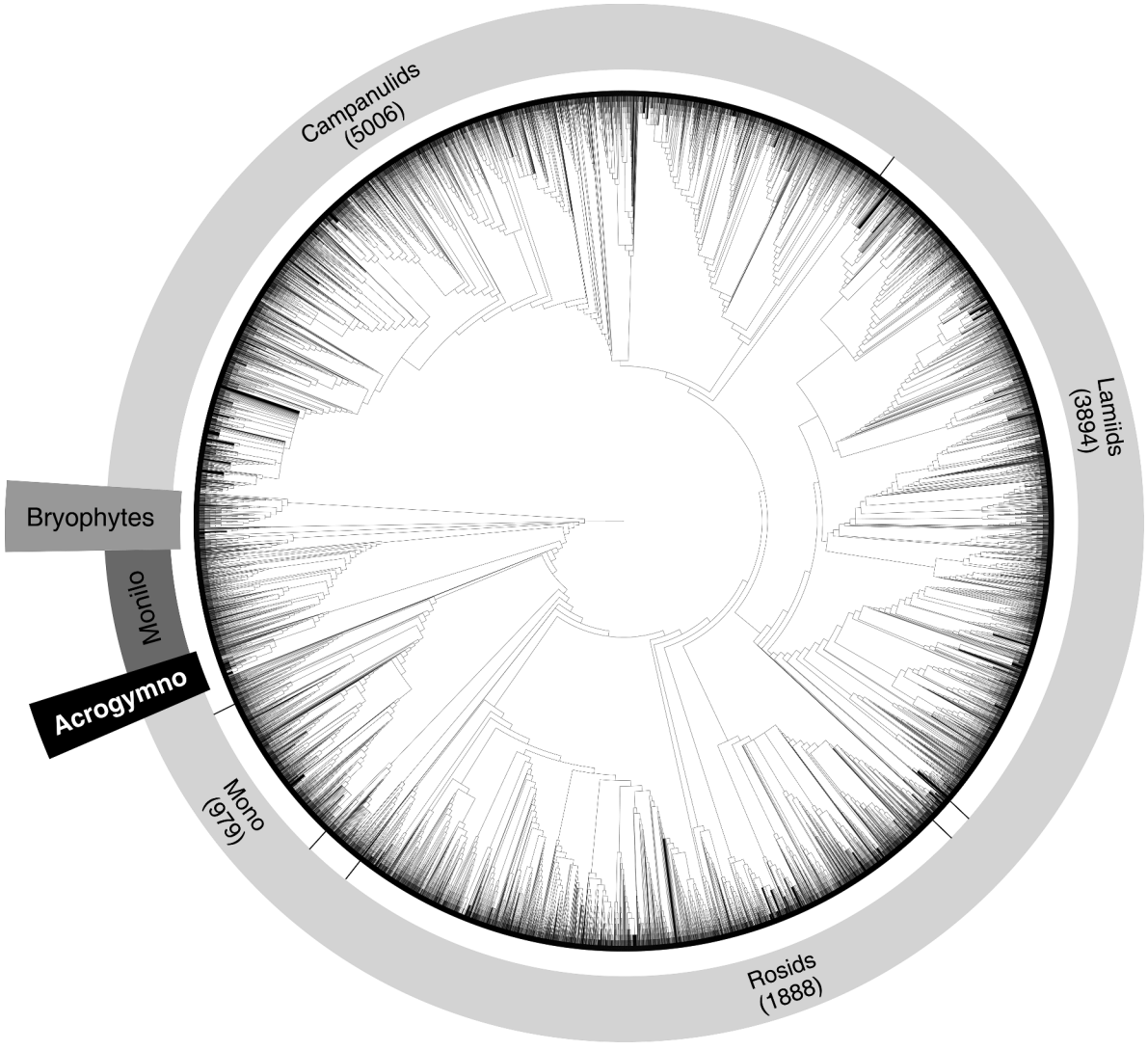

FIG. 2. A synthetic literature-based tree of 14 423 species of land plants highlighting several clades that are best sampled within angiosperms. The major clades are labeled after Cantino et al. (2007): Bryophytes include liverworts, mosses, and hornworts; Monilo represents Monilophyta; Mono represents monocots; and Acrogymno represents Acrogymnospermae. The numbers in parentheses represent the number of species contained within each of the major clades resulting from the grafting approach.

case in which a coordinated analysis of the sequence data would be highly beneficial.

PHYLOGRAFTER

To aid the task of mining the literature, storing source trees, and combining them into an increasingly comprehensive "grafted" tree, one of us has developed new software, Phylografter (Ree 2008), consisting of a database back-end and web-based front-end implemented using the web2py framework (Di Pierro 2011). Deployed on a server, Phylografter allows multiple users to collaboratively enter, edit, and assemble increasingly comprehensive phylogenetic trees from individual source trees using a graphical interface that facilitates searching for, viewing/editing, and grafting clades. The relational database stores trees as individual nodes, with each "source node" record retaining links to its source tree and publication data, including TreeBASE identification numbers when available. Each "grafted node" record is linked to its corresponding source node, ensuring the provenance and repeatability of grafted tree assembly. Phylografter thus represents a web-based content management system for phylogenetic information that provides node-by-node provenance of grafted trees. In this respect, it differs from the reference tree used by Phylomatic, as well as other software providing tree-grafting functionality, such as Phylowidget (Jordan and Piel 2008) and Mesquite (Maddison and Maddison 2010). These latter programs emphasize highly interactive graphical interfaces, while Phylograft-

er emphasizes database informatics. The objectives of its continued development are a much richer data model and tool set, both of which are needed for grafting and other strategies to be maximally effective (see next section). The source code is publicly available under a free license (Ree 2008).

Branch Lengths

The grafting procedures described in *Assembling a literature-based tree for land plants* deal only with tree topology, whereas comparative inferences also require branch lengths. This is a general problem shared by all approaches to phylogenetic synthesis, including supermatrix and supertree methods (see discussion in *Background*). One solution is the branch length adjustment (BLADJ) algorithm implemented in Phylocom (Webb et al. 2008), which fixes a subset of nodes in the tree to specified ages and evenly distributes the ages of the remaining nodes. Application of this method in Phylomatic, using plant clade age estimates from Wikström et al. (2001), has become fairly standard practice for generating branch lengths in plant ecological studies. However, such branch length estimates should be viewed with caution, in part because recent studies (e.g., Bell et al. 2010, Magallón 2010, Smith et al. 2010, Clarke et al. 2011) have revised and refined age estimates for many clades, but also because we know of no study that has critically evaluated the assumption of BLADJ approach. In the meantime, rather than relying on the Wikström et al. (2001) ages, we have applied the dates provided recently by Bell et al. (2010) in BLADJ to assign branch lengths to our assembled tree.

Supermatrix approaches, simultaneously providing topology and branch length estimates based on DNA sequences, would appear to be preferable to grafting and supertree approaches in ecological applications (e.g., Kress et al. 2009). However, a nonrandom distribution of missing sequence data can be a source of bias (Lemmon et al. 2009, but see also Wiens and Morrill 2011). There is a clear need to compare the performance of available methods for phylogenetic synthesis, as well as branch length estimation in ecological applications.

Discussion

Large, comprehensive phylogenetic trees hold the promise of providing important insights at the interface of ecology and evolution. To this end, agglomerative tree-building provides one means of delivering current knowledge of phylogenetic relationships to ecologists, and Phylografter has been developed to facilitate the process of collaborative tree-assembly while providing links to published phylogenetic hypotheses.

Considering the current state of phylogenetic knowledge, and the need for on-going synthesis, we believe that it is valuable to have at our disposal a variety of synthetic approaches. Some methods may be more appropriate or practical than others when addressing particular research questions. Although the direct estimation of community phylogenies from DNA sequences has become popular with recent advances in sequencing technology and computation (e.g., Kress et al. 2009), ecological communities rarely include all extant members of any clade, and incomplete taxon sampling is a well-known source of error in phylogenetic analysis (e.g., Graybeal 1998, Zwickl and Hillis 2002, Heath et al. 2008). Focusing only on the taxonomically heterogeneous members of particular communities may create conditions that will result in inaccurate inferences. It is possible, though untested, that the alternative approach described here, which specifically integrates expert knowledge, might perform better in terms of phylogenetic resolution and topological accuracy.

The most appropriate approach to a synthesis of phylogenetic knowledge for addressing a particular question will depend on the magnitude of the problem at hand, the quantity and quality of available data, and the expertise of the investigators. For example, when a number of genes (including barcodes) are available for many of the species of interest, it might be most appropriate to pursue a supermatrix approach. It is possible that a combination of approaches drawing on the advantages of each might simultaneously maximize accuracy and statistical power. For example, when sequence data are available for a community sample, estimation of branch lengths on an agglomerative phylogeny might simultaneously minimize topological error while providing genetic distances among community members (Whitfeld et al. 2012). Enforcing topological constraints drawn from the systematics literature in supermatrix analyses or in community-level phylogenetic analyses of sequence data provides other examples of attempts to gain similar advantage (Kress et al. 2010).

We emphasize that all approaches inherently require expert knowledge to achieve a meaningful outcome. In the case of a supermatrix approach, this arises in the context of choosing and aligning an appropriate set of genes (Smith et al. 2009, Sanderson et al. 2010). Supertree approaches require expertise from systematists to decide which input trees to include and how to curate them (e.g., removing poorly sampled outgroups from particular studies at the outset). The grafting approach entails decisions throughout the process, and, to be repeatable, also requires annotation at every step, so that decisions are properly exposed. In particular, as we move forward, it will especially valuable to capture information on uncertainty associated with conflicting phylogenetic hypotheses and on confidence in particular relationships. Synthetic trees annotated in this fashion could serve a wide variety of purposes such as tracing the development of phylogenetic hypotheses and identifying knowledge gaps in need of attention.

There is also a practical concern about how best to deploy disciplinary expertise. For example, it seems unreasonable to expect ecologists to serve as both the providers and the users of phylogenetic hypotheses. It is likely to be more efficient and productive for ecologists

to rely on syntheses generated by the systematics community, and for the foreseeable future to collaborate with systematists on such projects. This, of course, will challenge systematists to expand their thinking beyond their individual clades of interest and to seriously consider how best to estimate relationships when taxon samples are highly biased from a phylogenetic standpoint. Given the rapid integration of community ecology and phylogenetic biology (Webb 2000, Webb et al. 2002, Cavender-Bares et al. 2009, Vamosi et al. 2009), the estimation of community phylogenies must be viewed as an important objective. This requires a host of reasoned and practical choices, and must be carefully documented in order to serve as the basis for comparative studies that have meaningful implications, even for society at large (e.g., predicting the effects of global climate change; Edwards et al. 2007). We have described a grafting approach not as an ultimate solution, but as a legitimate, defensible alternative at this time, and one that has already proven to be useful under certain circumstances.

Concluding Thoughts

Several improvements are essential for agglomerative approaches to become practical and more broadly useful. These are mainly the responsibility of the systematics community. Unfortunately, to this day, many studies are published without depositing the primary products (sequence alignments and trees) in public repositories such as TreeBASE or DRYAD. Even for those that are deposited, relevant metadata and node-specific information such as support values and age estimates are not properly archived. In an era defined by advances in information synthesis (e.g., Google), it is deeply disappointing to realize that systematic knowledge continues to be disseminated almost exclusively in legacy formats that shield the most valuable information from effective capture and re-use (e.g., phylogenetic trees in the form of published figures and primary data in the form of unstructured supplementary documents). But, beyond the necessary behavioral adjustments, the lack of public tools for cross-referencing systematic studies by taxon name (species or clade), gene, sequence accession number, and so on, is crippling efforts to assemble the most up-to-date trees or, for that matter, maximally informative data sets for supermatrix approaches. In the end, the onus is on the systematics community to enable phylogenetic knowledge to be applied effectively and with confidence in ecology and other fields of biology. Some such efforts are underway in conjunction with the iPlant iPTOL project (information *available online*),[9] and this is the impetus behind the development of Phylografter: to create a community-driven resource for collectively assembling and synthesizing the results of systematic studies (not only the trees, but the underlying data) and providing links across taxa and characters.

Finally, it is important to put the idea of agglomerative trees into longer term perspective. In the very distant future, there may come a time when the phylogenetic relationships of most of the species on Earth will have been estimated with confidence. Even well before that time, for some groups of organisms, such as land plants, our understanding will have advanced to the point that the problem of community phylogeny assembly will be different. If the systematics community has fulfilled its responsibility in providing a tree of life, as chemists generated a periodic table of elements, ecologists will no longer be faced with the choices we have outlined. Nor will they need to spend their time inferring phylogenies through any of the mechanisms we have outlined. Instead, they will be able to simply access the relevant, up-to-date phylogenetic information through a service such as the current Phylomatic. In this sense, all of these approaches can be viewed as provisional measures.

In the meantime, however, it is important to appreciate that the tree of life will not miraculously assemble itself. Practicing systematists must, of course, continue to accumulate and analyze data to resolve relationships in their respective study groups. But syntheses of these efforts on a grand scale will require the conscious development of relevant infrastructure and phyloinformatics tools. This will undoubtedly involve improved methods of data mining and phylogenetic analysis, as well as some degree of coordination in character sampling. However, at a fundamental level, the process relies on the continued engagement of taxonomic specialists, and on developing the proper means to effectively capture and deploy their collective knowledge. We hope that the approach outlined here can provide one mechanism to draw together this expertise.

Acknowledgments

Thanks are due to Nicholas Deacon, Sara Branco, Ross Bernard, Rebecca Reiss, and Tim Whitfeld for their help in manually redrawing various source trees from published figures. We are also grateful to Kelly Robertson and Boris Igic for kindly providing us with their phylogeny of Solanaceae, which, at the time of the initial meetings, was unpublished. We thank Bill Piel for providing us with the entire TreeBASE database and for technical assistance during the early stages of this project. Our work was conducted as part of a working group on Ecophylogenetics, supported by the National Center for Ecological Analysis and Synthesis (NCEAS), and as part of an Encyclopedia of Life (EOL) Biodiversity Synthesis grant. Support for J. M. Beaulieu has been provided by NCEAS and the iPTOL program within the NSF-funded iPlant Collaborative (http://www.iplantcollaborative.org/).

Literature Cited

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

[9] http://www.iplantcollaborative.org

Bansal, M. S., J. G. Burleigh, O. Eulenstein, and D. Fernandez-Baca. 2010. Robinson-Foulds supertrees. Algorithms for Molecular Biology 5:18.

Beaulieu, J. M., A. T. Moles, I. J. Leitch, M. D. Bennett, J. B. Dickie, and C. A. Knight. 2007. Correlated evolution of genome size and seed mass. New Phytologist 173:422–437.

Beaulieu, J. M., S. A. Smith, and I. J. Leitch. 2010. On the tempo of genome size evolution in angiosperms. Journal of Botany 2010:989152.

Bell, C. D., D. E. Soltis, and P. S. Soltis. 2010. The age and diversification of the angiosperms revisited. American Journal of Botany 97:1296–1303.

Bininda-Emonds, O. R. P. 2004. The evolution of supertrees. Trends in Ecology and Evolution 19:315–322.

Bininda-Emonds, O. R. P., R. M. D. Beck, and A. Purvis. 2005. Getting to the roots of matrix representation. Systematic Biology 54:668–672.

Bininda-Emonds, O. R. P., M. Cardillo, K. E. Jones, R. D. E. MacPhee, R. M. D. Beck, R. Grenyer, S. A. Price, R. A. Vos, J. L. Gittleman, and A. Purvis. 2007. The delayed rise of present-day mammals. Nature 446:507–512.

Cadotte, M. W., B. J. Cardinale, and T. H. Oakley. 2008. Evolutionary history and the effect of biodiversity on plant productivity. Proceedings of the National Academy of Sciences USA 105:17012–17017.

Cantino, P. D., J. A. Doyle, S. W. Graham, W. S. Judd, R. G. Olmsead, D. E. Soltis, P. S. Soltis, and M. J. Donoghue. 2007. Towards a phylogenetic nomenclature of Tracheophyta. Taxon 56:E1–E44.

Cavender-Bares, J., A. Keen, and B. Miles. 2006. Phylogenetic structure of Floridian plant communities depends on taxonomic and spatial scale. Ecology 93(Supplement):S109–S122.

Cavender-Bares, J., K. Kozak, P. Fine, and S. Kembel. 2009. The merging of community ecology and phylogenetic biology. Ecology Letters 12:693–715.

Cavender-Barres, J., and P. B. Reich. 2012. Shocks to the system: community assembly of the oak savanna in a 40-year fire frequency experiment. Ecology 93(Supplement):S52–S69.

Chaw, S. M., C. L. Parkinson, Y. Cheng, T. M. Vincent, and J. D. Palmer. 2000. Seed plant phylogeny inferred from all three plant genomes: Monophyly of extant gymnosperms and origin of Gnetales from conifers. Proceedings of the National Academy of Science USA 97:4086–4091.

Clarke, J. T., R. C. M. Warnock, and P. C. J. Donoghue. 2011. Establishing a time-scale for plant evolution. New Phytologist 192:266–301.

Cotton, J. A., and M. Wilkinson. 2007. Majority-rule supertrees. Systematic Biology 56:445–452.

Davies, T. J., T. G. Barraclough, M. W. Chase, P. S. Soltis, D. E. Soltis, and V. Savolainen. 2004. Darwin's abominable mystery: insights from a supertree of the angiosperms. Proceedings of the National Academy of Science USA 101:1904–1909.

Di Pierro, M. 2011. Web2py for scientific applications. Computing in Science and Engineering 13:64–69.

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

Dunn, C. W., et al. 2008. Broad phylogenomic sampling improves the resolution of the animal tree of life. Nature 452:745–749.

Edwards, E. J., and S. A. Smith. 2010. Phylogenetic analyses reveal the shady history of C4 grasses. Proceedings of the National Academy of Sciences USA 6:2532–2537.

Edwards, E. J., C. J. Still, and M. J. Donoghue. 2007. The relevance of phylogeny to studies of global change. Trends in Ecology and Evolution 22:243–249.

Fine, P. V. A., and S. W. Kembel. 2011. Phylogenetic community structure and phylogenetic turnover across space and edaphic gradients in western Amazonian tree communities. Ecography 34:552–565.

Gadek, P. A., D. L. Alpers, M. M. Heslewood, and C. J. Quinn. 2000. Relationships within Cupressaceae sensu lato: a combined morphological and molecular approach. American Journal of Botany 87:1044–1057.

Gernandt, D. S., G. Geada Lopez, S. Ortiz Garcia, and A. Liston. 2005. Phylogeny and classification of *Pinus*. Taxon 54:29–42.

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

Goloboff, P. A., S. A. Catalano, J. M. Mirande, C. A. Szumik, J. S. Arias, M. Kallersjo, and J. S. Farris. 2009. Phylogenetic analysis of 73,060 taxa corroborates major eukaryotic groups. Cladistics 15:415–428.

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

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

Hedges, S. B., J. Dudley, and S. Kumar. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22:2971–2972.

Helmus, M. R., and A. R. Ives. 2012. Phylogenetic diversity–area curves. Ecology 93(Supplement):S31–S43.

Jones, K. E., A. Purvis, A. MacLarnon, O. R. P. Bininda-Emonds, and N. B. Simmons. 2002. A phylogenetic supertree of the bats (Mammalia: Chiroptera). Biological Reviews of the Cambridge Philosophical Society 77:223–259.

Jordan, G. E., and W. H. Piel. 2008. PhyloWidget: web-based visualization for the tree of life. Bioinformatics 24:1041–1042.

Kembel, S. W., and S. P. Hubbell. 2006. The phylogenetic structure of Neotropical forest tree community. Ecology 87(Supplement):S86–S99.

Knapp, S., L. Dinsmore, C. Fissore, S. E. Hobbie, I. Jakobsdottir, J. Kattge, J. Y. King, S. Klotz, J. P. McFadden, and J. Cavender-Bares. 2012. Phylogenetic and functional characteristics of household yard floras and their changes along an urbanization gradient. Ecology 93(Supplement):S83–S98.

Kraft, N. J. B., and D. D. Ackerly. 2010. Functional trait and phylogenetic tests of community assembly across spatial scales in an Amazonian forest. Ecological Monographs 80:401–422.

Kress, W. J., D. L. Erickson, F. A. Jones, N. G. Swenson, R. Perez, O. Sanjur, and E. Bermingham. 2009. Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. Proceedings of the National Academy of Sciences USA 106:18621–18626.

Kress, W. J., D. L. Erickson, N. G. Swenson, J. Thompson, M. Uriate, and J. K. Zimmerman. 2010. Advances in the use of DNA barcodes to build a community phylogeny for tropical trees in a Puerto Rican forest dynamics plot. PLoS ONE 5:e15409.

Lemmon, A. R., J. M. Brown, K. Stanger-Hall, and E. Moriarty-Lemmon. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Systematic Biology 58:130–145.

Little, D. P. 2006. Evolution and circumscription of the true cypresses (Cupressaceae: *Cupressus*). Systematic Botany 31:461–480.

Maddison, D. R., and W. P. Maddison. 1996. The Tree of Life project. University of Arizona, Tucson, Arizona, USA. http://tolweb.org/tree/

Maddison, D. R., and K. S. Schulz. 2007. The Tree of Life project. University of Arizona, Tucson, Arizona, USA. http://tolweb.org

Maddison, W. P., and D. R. Maddison. 2010. Mesquite: a modular system for evolutionary analysis. Version 2.73. Arizona Board of Regents on Behalf of the University of Arizona, Tuscon, Arizona, USA. http://mesquiteproject.org/

Magallón, S. 2010. Using fossils to break long branches in molecular dating: a comparison of relaxed clocks applied to the origin of angiosperms. Systematic Biology 59:384–399.

McMahon, M. M., and M. J. Sanderson. 2006. Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. Systematic Biology 55:818–836.

McMorris, F. R., and M. Wilkinson. 2011. Conservative supertrees. Systematic Biology 60:232–238.

Moles, A. T., D. D. Ackerly, C. O. Webb, J. C. Tweddle, J. B. Dickie, and M. Westoby. 2005. A brief history of seed size. Science 307:576–580.

Novotny, V., et al. 2010. Guild-specific patterns of species richness and host specialization in plant-herbivore food webs from a tropical forest. Journal of Animal Ecology 79:1193–1203.

Phillipe, H., H. Brinkmann, D. V. Lavrov, D. T. J. Littlewood, M. Manuel, G. Wöheide, and D. Baurain. 2011. Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS Biology 9:e1000602.

Pyron, R. A., F. T. Burbrink, G. R. Colli, A. Nieto Montes de Oca, L. J. Vitt, C. A. Kuczynski, and J. J. Wiens. 2011. The phylogeny of advanced snakes (Colubroidea), with discovery of a new subfamily and comparison of support methods for likelihood trees. Molecular Phylogenetics and Evolution 58:329–342.

Qiu, Y. L., et al. 2006. The deepest divergences in land plants inferred from phylogenomic evidence. Proceedings of the National Academy of Sciences USA 103:15511–15516.

Ratnasingham, S., and P. D. N. Hebert. 2007. The barcode of life data system. Molecular Ecology Resources 7:355–364.

Ree, R. H. 2008. Phylografter: collaborative assembly of phylogenetic trees. Field Museum of Chicago and the Biodiversity Synthesis Center, Chicago, Illinois, USA. http://phylografter.googlecode.com

Ree, R. H., and M. J. Donoghue. 1999. Inferring rates of chance in flower symmetry in asterid angiosperms. Systematic Biology 48:633–641.

Ruta, M., J. E. Jeffery, and M. I. Coates. 2003. A supertree of early tetrapods. Proceedings of the Royal Society B 270:2507–2516.

Sanderson, M. J., D. Boss, D. Chen, K. A. Cranston, and A. Wehe. 2008. The PhyLoTA browser: Processing GenBank for molecular phylogenetics research. Systematic Biology 57:335–346.

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

Sanderson, M. J., A. Purvis, and C. Henze. 1998. Phylogenetic supertrees: assembling the trees of life. Trends in Ecology and Evolution 13:105–109.

Sanderson, M. J., and H. B. Shaffer. 2002. Troubleshooting molecular phylogenetic analyses. Annual Review of Ecology, Evolution, and Systematics 33:49–72.

Schuettpelz, E., and K. M. Pryer. 2008. Fern phylogeny. Pages 395–416 *in* T. A. Ranker and C. H. Haufler, editors. Biology and evolution of ferns and lycophytes. Cambridge University Press, Cambridge, UK.

Smith, S. A., and J. M. Beaulieu. 2009. Life history influences rates of climatic niche evolution in flowering plants. Proceedings of the Royal Society B 276:4345–4352.

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

Smith, S. A., J. M. Beaulieu, and M. J. Donoghue. 2010. Uncorrelated relaxed-clock analysis suggests an early origin for flowering plants. Proceedings of the National Academy of Sciences USA 107:5897–5902.

Smith, S. A., J. M. Beaulieu, A. Stamatakis, and M. J. Donoghue. 2011. Understanding angiosperm diversification using small and large phylogenetic trees. American Journal of Botany 98:1–12.

Smith, S. A., J. M. Beaulieu, A. Stamatakis, and M. J. Donoghue. 2012. Corrigendum: Understanding angiosperm diversification using small and large phylogenetic trees. American Journal of Botany, *in press*.

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

Stamatakis, A., P. Hoover, and J. Rougemont. 2008. A rapid bootstrap algorithm for the RAxML web servers. Systematic Biology 57:758–771.

Stevens, P. F. 2011. Angiosperm phylogeny website. Version 9. University of Missouri, St. Louis, Missouri, USA. http://www.mobot.org/MOBOT/research/APweb/

Strauss, S. Y., C. O. Webb, and N. Salamin. 2006. Exotic taxa less related to native species are more invasive. Proceedings of the National Academy of Sciences USA 103:5841–5845.

Swenson, N. G., P. Anglada-Cordero, and J. A. Barone. 2011. Deterministic tropical tree community turnover: evidence from patterns of functional beta diversity along an elevational gradient. Proceedings of the Royal Society B 278:877–884.

Swenson, N. G., B. J. Enquist, J. Thompson, and J. K. Zimmerman. 2007. The influence of spatial and size scale on phylogenetic relatedness in tropical forest communities. Ecology 88:1770–1780.

Tank, D. C., and M. J. Donoghue. 2010. Phylogeny and phylogenetic nomenclature of the Campanulidae based on an expanded sample of genes and taxa. Systematic Botany 35:425–441.

Thomson, R. C., and H. B. Shaffer. 2010. Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. Systematic Biology 59:42–58.

Thuiller, W., S. Lavergne, C. Roquet, I. Boulangeat, B. Lafourcade, and M. B. Araujo. 2011. Consequences of climate change on the tree of life in Europe. Nature 470:531–534.

Townsend, J. P. 2007. Profiling phylogenetic informativeness. Systematic Biology 56:222–231.

Vamosi, S. M., S. B. Heard, J. C. Vamosi, and C. O. Webb. 2009. Emerging patterns in the comparative analysis of phylogenetic community structure. Molecular Ecology 18:572–592.

Verdú, M., P. J. Rey, J. M. Alcántara, G. Siles, and A. Valiente-Banuet. 2009. Phylogenetic signatures of facilitation and competition in successional communities. Journal of Ecology 97:1171–1180.

Webb, C. O. 2000. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. American Naturalist 156:145–155.

Webb, C. O., D. D. Ackerly, and S. W. Kembel. 2008. Phylocom: software for the analysis of phylogenetic community structure and trait evolution. Bioinformatics 24:2098–2100.

Webb, C. O., D. D. Ackerly, M. A. McPeek, and M. J. Donoghue. 2002. Phylogenies and community ecology. Annual Review of Ecology, Evolution, and Systematics 33:475–505.

Webb, C. O., and M. J. Donoghue. 2005. Phylomatic: tree assembly for applied phylogenetics. Molecular Ecology Notes 5:181–183.

Weiblen, G. D., R. K. Oyama, and M. J. Donoghue. 2000. Phylogenetic analysis of dioecy in monocotyledons. American Naturalist 155:46–58.

Weiblen, G. D., C. O. Webb, V. Novotny, Y. Basset, and S. E. Miller. 2006. Phylogenetic dispersion of host use in a tropical insect herbivore community. Ecology 87(Supplement):S62–S75.

Whitfeld, T. J. S., V. Novotny, S. E. Miller, J. Hrcek, P. Klimes, and G. D. Weiblen. 2012. Predicting tropical insect herbivore abundance from host plant traits and phylogeny. Ecology 93(Supplement):S211–S223.

Wiens, J. J. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction. Systematic Biology 54:731–742.

Wiens, J. J., and D. S. Moen. 2008. Missing data and the accuracy of Bayesian phylogenetics. Journal of Systematics and Evolution 46:307–314.

Wiens, J. J., and M. C. Morrill. 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Systematic Biology 60:719–731.

Wikström, N., and P. Kenrick. 2001. Evolution of Lycopodiaceae (Lycopsida): Estimating divergence times from rbcL gene sequences by use of nonparametric rate smoothing. Molecular Phylogenetics and Evolution 19:177–186.

Wikström, N., V. Savolainen, and M. W. Chase. 2001. Evolution of the angiosperms: calibrating the family tree. Proceedings of the Royal Society B 268:2211–2220.

Willis, C. G., B. Ruhfel, R. B. Primack, A. J. Miller-Rushing, and C. C. Davis. 2008. Phylogenetic patterns of species loss in Thoreau's woods are driven by climate change. Proceedings of the National Academy of Sciences USA 105:17029–17033.

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