American Journal of Botany

American Journal of Botany 98(3): 404–414. 2011.

Understanding angiosperm diversification using small and large phylogenetic trees[1]

Stephen A. Smith[2,3,5], Jeremy M. Beaulieu[4], Alexandros Stamatakis[3], and Michael J. Donoghue[4]

[2]Department of Ecology and Evolutionary Biology, Brown University, Providence, Rhode Island 02912 USA; [3]Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany; and [4]Department of Ecology and Evolutionary Biology, Yale University, P.O. Box 208105, New Haven, Connecticut 06520 USA

How will the emerging possibility of inferring ultra-large phylogenies influence our ability to identify shifts in diversification rate? For several large angiosperm clades (Angiospermae, Monocotyledonae, Orchidaceae, Poaceae, Eudicotyledonae, Fabaceae, and Asteraceae), we explore this issue by contrasting two approaches: (1) using small backbone trees with an inferred number of extant species assigned to each terminal clade and (2) using a mega-phylogeny of 55 473 seed plant species represented in GenBank. The mega-phylogeny approach assumes that the sample of species in GenBank is at least roughly proportional to the actual species diversity of different lineages, as appears to be the case for many major angiosperm lineages. Using both approaches, we found that diversification rate shifts are not directly associated with the major named clades examined here, with the sole exception of Fabaceae in the GenBank mega-phylogeny. These agreements are encouraging and may support a generality about angiosperm evolution: major shifts in diversification may not be directly associated with major named clades, but rather with clades that are nested not far within these groups. An alternative explanation is that there have been increased extinction rates in early-diverging lineages within these clades. Based on our mega-phylogeny, the shifts in diversification appear to be distributed quite evenly throughout the angiosperms. Mega-phylogenetic studies of diversification hold great promise for revealing new patterns, but we will need to focus more attention on properly specifying null expectation.

Key words: angiosperms; diversification rate; flowering plants; key innovation; mega-phylogeny.

Is the global success of the flowering plants a function of some feature of the group as a whole, or does it really reflect the success of one or more angiosperm subgroups (e.g., Sanderson and Donoghue, 1994)? Are such apparent successes, at whatever level they occur, best explained by key innovations or by key opportunities (e.g., Moore and Donoghue, 2007), and how has differential extinction influenced our perception of the problem? Answers to these questions depend on insights about phylogenetic relationships and about how species richness is distributed throughout the tree. Fortunately, our knowledge of angiosperm phylogeny has improved dramatically over the past decade (e.g., Cantino et al., 2007; Jansen et al., 2007; Moore et al., 2007; Angiosperm Phylogeny Group, 2009; Soltis et al., in press), as have methods for detecting the phylogenetic location of shifts in rates of diversification (e.g., Slowinski and Guyer, 1989; Guyer and Slowinski, 1993; Sanderson and Donoghue, 1994; Mooers and Heard, 1997; Moore et al., 2004; Ricklefs, 2007; Alfaro et al., 2009; Moore and Donoghue, 2009). However, it will still be a long time before we know the relationships of every angiosperm species with any level of confidence. In view of our incomplete knowledge for the foreseeable future, what are the best strategies for studying shifts in diversification rate?

One approach has been to base such analyses on a backbone tree that depicts “established” relationships among the major lineages within a clade of interest. In this case, each terminal is assigned the number of species thought to be represented by an “exemplar species,” and the analysis necessarily bypasses how diversity is distributed within these terminal lineages (e.g., Sanderson and Donoghue, 1994; Alfaro et al., 2009; Santini et al., 2009). This has the obvious drawback of not being able to identify where shifts in diversification may have occurred within a large terminal clade, and it can result in a particular kind of mistake: a shift in diversification attributed to such a composite terminal might actually be due to a shift that occurred within that clade (a form of the “trickle-down” effect; Moore et al., 2004).

Another possible approach, which has not yet been explored in detail, is to use a phylogenetic tree that includes all species for which relevant phylogenetic data are available, simply treating each terminal as a single species. In using this approach, one has to hope that there are enough representatives of the clade of interest for which data are available and that the sample of species available for phylogenetic analysis more-or-less accurately reflects the distribution of the underlying species diversity. This approach has the obvious drawback of potentially biasing the results due to the over- or underrepresentation of particular clades in the underlying data set.

One would hope that these two different approaches would largely yield similar results, but such comparisons have not yet

[1] Manuscript received 24 November 2010; revision accepted 11 February 2011.

We thank Peter Raven, J. Chris Pires, and Jon Chase for inviting us to contribute to this special issue. We are grateful to the National Science Foundation for supporting this research through the CIPRES and Angiosperm ATOL projects and through the iPToL grand challenge project supported by the iPlant Collaborative. Part of this work was also supported by the German Science Foundation (DFG) under the auspices of the Emmy-Noether program. Computational work was made possible by the facilities and the helpful staff of the Yale University High Performance Computing Center and by John Cazes at the Texas Advanced Computing Center.

[5] Author for correspondence: stephen_a_smith@brown.edu

doi:10.3732/ajb.1000481

been carried out, perhaps mainly because it has not been possible until recently to infer phylogenetic relationships at a very large scale (e.g., using all species in GenBank for major clades; see McMahon and Sanderson, 2006; Smith et al., 2009; Thomson and Shaffer, 2010). Here we begin to explore this issue by focusing on a set of major plant clades that are widely considered to be exceptionally diverse: the angiosperms as a whole, monocots, orchids, grasses, eudicots, legumes, and composites. Specifically, we first use small backbone trees for each of these groups to infer shifts in diversification, then we compare the results to those obtained when we apply diversification tests to phylogenetic trees for these clades derived from a mega-phylogenetic analysis. For this purpose, we have inferred a phylogeny for a large portion of the seed plant species represented in GenBank. We view these studies as exploratory, and the results as very preliminary. However, as we will argue, they highlight some potentially general insights into the study of diversification as well as into the patterns and processes of angiosperm diversification in particular.

IDENTIFYING MAJOR RADIATIONS IN A PHYLOGENETIC TREE

It has long been noted that the tree of life is highly imbalanced in places and that this pattern reflects differences in the net rate of diversification (speciation minus extinction) in different parts of the phylogeny. Such observations, combined especially with the rapid rise of molecular phylogenetics, have stimulated the development of methods and tools to extract information about diversification rates from phylogenies. These range from tree-balance measures that variously test for asymmetry in the partitioning of species diversity across a tree (e.g., Agapow and Purvis, 2002; Purvis et al., 2002; Chan and Moore, 2005; Holman, 2005), to methods that combine both topological and temporal information to infer speciation and extinction parameters (e.g., Nee et al., 1992; Magallón and Sanderson, 2001; Nee, 2001; Rabosky and Lovette, 2008; Alfaro et al., 2009; Moore and Donoghue, 2009).

The general question of whether a given tree has experienced significant diversification rate variation among its branches is undoubtedly important, but we often want to identify where exactly these shifts are likely to have taken place in the tree and, ultimately, the underlying causes. Early studies making use of molecular phylogenies for understanding angiosperm diversification focused on regions where shifts may have taken place (i.e., the stem subtending crown angiosperms). Tests initially evaluated highly reduced trees with just three terminals (Sanderson and Donoghue, 1994)—an outgroup and two sister clades representing a possible first branching event within crown angiosperms. Significant rate shifts were assessed by calculating a likelihood ratio for the relative fit of one or more Yule ("pure birth") rate parameters distributed over various permutations of the four branches contained within the three taxon tree. This approach did not require a comprehensive tree, only the cumulative diversity estimates for the three clades.

Chan and Moore (2002) modified the methods of Sanderson and Donoghue (1994) to allow the identification of local rate shifts in diversification across all the branches in a user-supplied tree. As with the method of Sanderson and Donoghue (1994), this approach does not require temporal information. Temporal methods for estimating diversification rates had previously been developed (reviewed in Sanderson and Donoghue, 1996; also see Ricklefs, 2007) and have since been developed to allow the integration of molecular divergence time information and also to take into account the estimated number of species in different clades (e.g., Alfaro et al., 2009; Moore and Donoghue, 2009). However, uncertainties surrounding the timing of events can confound diversification analyses (cf. Moore et al., 2004), and the significant gap that remains between molecular estimates for the origin and radiation of angiosperms (e.g., Smith et al., 2010; Magallón, 2010; Bell et al., 2010) and the stratigraphic record (e.g., Crane et al., 1995; Doyle, 2001) make this an especially contentious issue. Therefore, for the purposes of the analyses reported here, we are forced to employ methods that do not require temporal information.

Here, we use the topology-based diversification rate shift statistic, Δ_i (Chan and Moore, 2002; Moore et al., 2004), but have modified the likelihood and significance calculations to accommodate exceptionally large species numbers. For an overview of the original methods and for additional details, we refer the reader to Moore et al. (2004). In brief, local rate shifts in a tree are identified by first calculating a likelihood ratio under a one- vs. two-rate Yule branching model by comparing the extant diversity for the ingroup and its sister group. A likelihood ratio is then calculated for the focal ingroup by comparing the extant diversity of the two clades stemming from the first branching event. The Δ_i statistic is equal to the difference between these two likelihood ratios. The Δ_i attenuates the tendency for diversification rate shifts to "trickle-down" to lower nodes in the tree by conditioning the evidence for a shift along the branch leading to the ingroup by the evidence for a shift occurring within the ingroup. However, under a Yule process, and assuming a fixed branching probability, the likelihood of observing a particular number of species rapidly approaches zero as clade diversity exceeds ~1600 species. After this threshold, the estimated likelihood is automatically rounded to zero due to the fixed range in the size of the exponent used in floating-point arithmetic carried out on a standard computer. To overcome this problem, we implemented the calculation of the shift statistic (as calculated in the program SymmeTREE) in the scripting language Python (version 2.5) and incorporated a floating-point arithmetic package, mpmath (version 0.9), which uses arbitrary precision machine numbers to allow calculations involving numbers that are as small or large in magnitude (bits used for number representation) as permitted by a computer's memory.

To assess the significance of Δ_i, SymmeTREE (Chan and Moore, 2005) uses a computationally intensive Monte Carlo simulation to generate a null distribution of Δ_i assuming a single-rate Yule branching process. This involves simulating hundreds of thousands of trees with diversities equal to the study tree, which can be prohibitively time consuming for large datasets. We bypassed the need to simulate a set of trees for each test by setting the parametric shape of the null distribution, a priori. We permuted a set of 100 000 random pure birth trees for tree sizes of 100, 500, 1000, 5000, and 10 000 and calculated Δ_i for the first branch above the root. Trees of larger size were not feasible to calculate given runtime limitations, though results were consistent between different tree sizes. The results from these simulations suggested that the null distribution of Δ_i consistently approximated a Gaussian normal with $\mu = 0$ and $\sigma = 1.3$ and did not appear to be strongly related to tree size. This parametric shape of the null distribution resulted in a type I error rate close to the nominal level of 5% and was therefore used to assess the significance of a given empirical Δ_i. Significant shifts reported in the remainder of this paper assume that $\alpha = 0.05$, unless otherwise indicated. For illustrative purposes, we also examine the effect that a Bonferroni correction on α has on inferring significant shifts in diversification in our phylogenetic data set.

EXAMINING ANGIOSPERM DIVERSIFICATION USING A BACKBONE PHYLOGENY

Discussions of diversification are often oriented by taxonomy, focusing on major named groups. For example, botanists, in noting that angiosperms are the most diverse clade of land plants, have tended to attribute a shift in diversification to the origin of angiosperms, and have searched for novelties that might be responsible for this (Darwin and Seward, 1903; Davies et al., 2004; Friedman, 2009; Crepet and Niklas, 2009). Sanderson and Donoghue (1994; also see Doyle and Donoghue, 1993) argued that it may not be the angiosperms as a whole that shifted to a higher rate of diversification but, instead, that one or more major clades nested within angiosperms underwent rapid radiations. Similar arguments had been made earlier for birds by Raikow (1986); the real shifts in diversification, he suggested, took place in nested subclades.

We first explore these issues using a set of backbone trees that reflect consensus views on the relationships among the major lineages within each clade. Specifically, we focus on clades that have long been recognized and named and that are highly diverse, and where a higher diversification rate has been attributed to the group: Angiospermae, Monocotyledonae, Orchidaceae, Poaceae, Fabaceae, and Asteraceae. We also examine the eudicots, a major angiosperm clade that has only recently been identified and named (Eudicotyledonae; Cantino et al., 2007).

To test whether a significant shift in diversification rate was (or was not) associated with the origin of each major named clade, we assembled a backbone phylogeny for each one in which "terminals" generally represent many species (Fig. 1). Diversity estimates for each "terminal" in all backbone trees were obtained from Stevens (2010). All analyses used the topology-based diversification rate shift statistic, Δ_i, described above.

Our backbone trees were assembled from the primary literature, and we consider these to be the best current hypotheses of relationships within the clades examined (Fig. 1). For relationships among angiosperms and their extant relatives, we used Qiu et al. (2006), who recovered strong support for monilophytes as sister to seed plants and for acrogymnosperms as sister to angiosperms. For relationships among the major lineages within angiosperms, we relied on the recent genome-scale chloroplast phylogenies of Jansen et al. (2007) and Moore et al. (2007). These authors reported strong support for *Amborella* as sister to the rest of the angiosperms, followed by Nymphaeales, and Austrobaileyales as sister to the recently recognized Mesangiospermae (Magnoliidae, Chloranthaceae, Monocotyledonae, *Ceratophyllum*, Eudicotyledonae; Cantino et al., 2007). For relationships within monocots, we relied on Chase et al. (2006), and for eudicots, we used Worberg et al. (2007). For Orchidaceae, we relied on Cameron (2006), and we used the Grass Phylogeny Working Group (2001) tree for Poaceae. Finally, we referred to the trees of Wojciechowski et al. (2004) and Bruneau et al. (2008) for relationships within Fababceae and to the study of Panero and Funk (2008) for backbone relationships within Asteraceae.

Where considerable phylogenetic uncertainty still exists, we also examined several alternative topologies. For example, while many analyses place *Amborella* as sister to the rest of angiosperms, there is also the possibility that *Amborella* and Nymphaeales go together (Leebens-Mack et al., 2005; Soltis et al., 2007). We also examined alternative relationships among the major clades within the Mesangiospermae (i.e., the placement of Chloranthaceae; Moore et al., 2007; Jansen et al., 2007), among the early-branching lineages within Monocotyledonae (i.e., the placement of Acorales; Davis et al., 2004; Chase et al., 2006), within Orchidaceae (i.e., the arrangement of Vanilloideae and Cyprepedioideae; Cameron, 2006; Ramirez et al., 2007), and within Asteraceae (i.e., the arrangement of Stiffioideae, Mutisioideae, and Wunderlichioideae; Panero and Funk, 2002, 2008; Funk et al., 2005).

These analyses revealed a general pattern (Fig. 1). Surprisingly, significant shifts in the rate of diversification were directly associated with none of the long familiar groups. Although the exact location of significant shifts elsewhere in these trees is less certain (owing to our collapse of large clades into single terminals), it is noteworthy that in each case one or more shifts do appear to be nested not far within the focal clade. For instance, in the case of angiosperms, regardless of the exact arrangement of the early-branching lineages, we identify a major increase in diversification rate associated with Mesangiospermae. The recently recognized eudicot clade is the only one that we analyzed that is directly subtended by a significant upward shift in diversification rate (Fig. 1).

EXAMINING ANGIOSPERM DIVERSIFICATION USING A MEGA-PHYLOGENY

Molecular phylogenetic data sets have been increasing exponentially, and there has been a growing effort to use as much available sequence data as possible to construct very large phylogenetic trees. Release 178 of NCBI's GenBank contained more than 115 624 497 715 bp (excluding whole genome shotgun submissions) and 120 604 423 sequence records, an extraordinary increase of 1 275 608 944 bp and 1 492 172 records in the 58 d since the previous release. Recent progress in both multiple sequence alignment (programs MUSCLE [Edgar, 2004] and MAFFT [Katoh and Toh, 2008]) and phylogenetic analysis (programs TNT, [Goloboff, 1999], GARLI [Zwickl, 2006], and RAxML [Stamatakis, 2006]) has made it possible to assemble and analyze far larger data sets than ever before. Owing to these advances, recently published analyses have been able to include more than 1400 genes (Hejnol et al., 2009) and more than 70 000 taxa (Goloboff et al., 2009).

Large comprehensive phylogenies present new problems and amplify old ones. For example, given the nonuniform collection of data, increasing the number of taxa in a data set unavoidably increases the amount of missing data in the alignment. The potential problems of missing data, including decreased phylogenetic accuracy, have been discussed at length elsewhere (e.g., Wiens, 2003, 2005; Hartmann and Vision, 2008; Lemmon et al., 2009). It is worth noting that a recent study by Sanderson et al. (2010) explored the theoretical capabilities of phylogenetic analyses as a function of the amount and structure of missing data. They found that most data combinations contained too little information to produce a fully resolved tree, but that they can nevertheless distinguish many nodes in the tree from random.

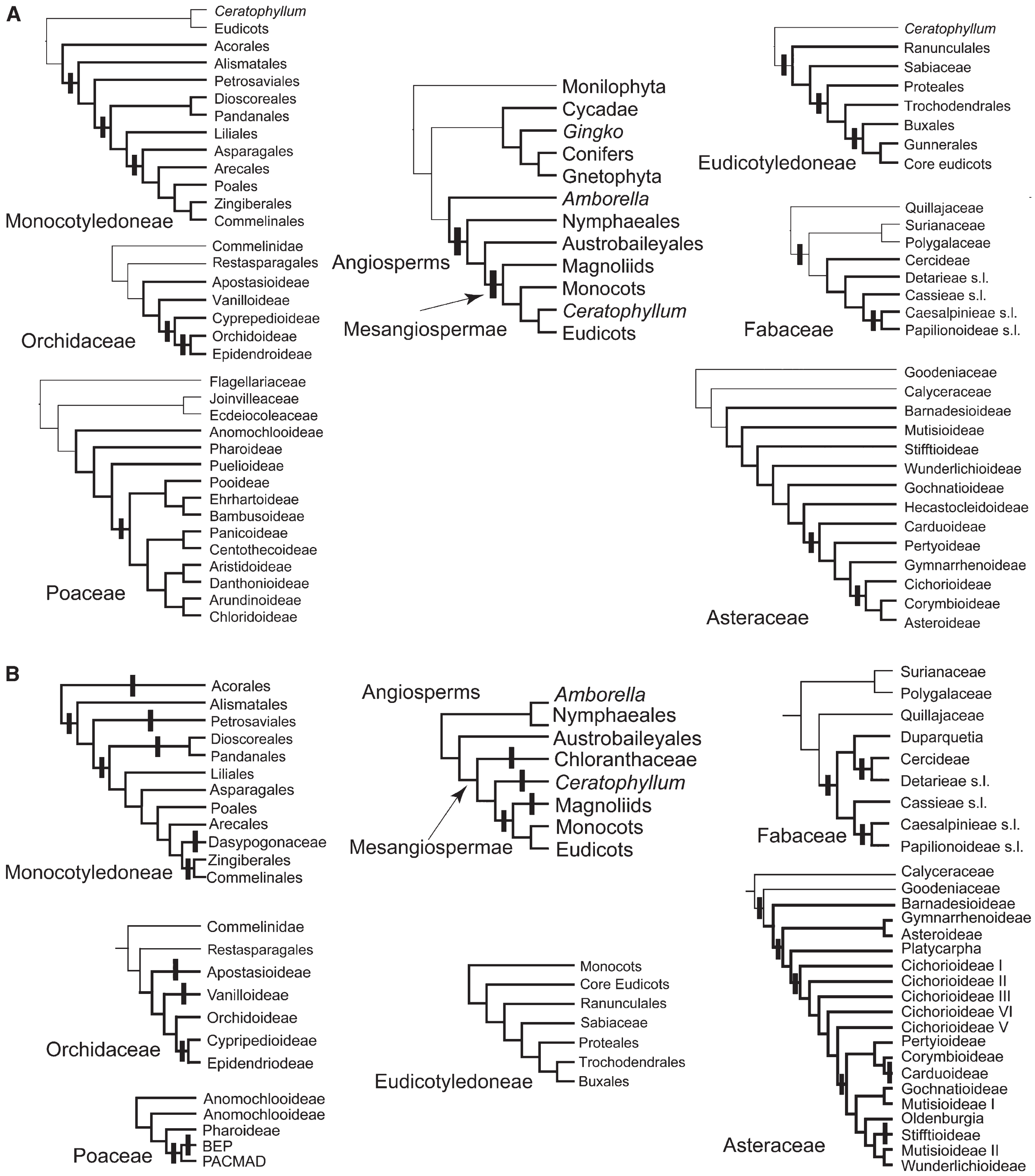

Fig. 1. The trees represent the phylogenetic relationships among the eight clades that were analyzed separately using both the (A) backbone and (B) large-scale phylogenetic approaches. The locations of rate shifts in diversification are denoted by black bars. We note that shifts are not associated with the origin of flowering plants or the origin of several highly diverse, named groups nested within. Instead, significant rate shifts are nested somewhat within each named group. The recently named Eudicotyledonae is the only clade analyzed that is directly subtended by a significant shift in diversification. (B) The locations of rate shifts when testing the large-scale phylogenetic approach are also not associated with the origin of any of the same highly diverse named groups. In general, the exact location of shifts when testing the large-scale phylogenetic approach corresponded somewhat with the locations inferred from the representative phylogenetic approach (A). The two exceptions being the Eudicotyledonae and Fabaceae, where the location of the shifts was dependent on the placement of species poor clades. We also note that the backbone approach (A) cannot detect shifts along branches leading to the "tips"; therefore, we only denote such shifts in the large-scale phylogenetic approach (B).

Despite potential pitfalls in tree inference at this scale, mega-phylogenetic analyses have proven useful in revealing broad evolutionary patterns. For example, Smith and Donoghue (2008) confirmed that rates of evolution in woody plants were slower than in herbaceous plants (Gaut et al., 1992). Smith and Beaulieu (2009) found similar results in relation to climate evolution, where woody lineages accumulated fewer changes per million years in climatic niche space than related herbaceous lineages. A large phylogenetic analysis of grasses documented that C_4 grass evolution is correlated with shifts into drier, but not necessarily warmer, climates (Edwards and Smith, 2010). Goldberg et al. (2010) found evidence for species selection using a mega-phylogeny of Solanaceae.

To explore the diversification of angiosperms using a very large phylogeny, we inferred a tree for flowering plants and acrogymnosperms (included for rooting purposes) using available data from GenBank for six gene regions: the chloroplast genes *atpB*, *matK*, *trnK*, *trnL*, *rbcL*, and nuclear ribosomal ITS. These regions were chosen because they are among the most commonly used gene regions for molecular phylogenetic studies in plants. The chloroplast regions are also relatively less complex with respect to gene duplication and loss. Although ITS is sometimes more complicated (see Alvarez and Wendel, 2003), it was included here because it is the best-sampled plant gene region in GenBank. The data set was assembled using the methods described in Smith et al. (2009), as implemented in the PHLAWD program. This method uses GenBank nucleotide data for the clade of interest and requires gene regions of interest to be identified before the analysis. It then uses a "baited" sequence comparison approach wherein a set of sequences provided by the user is used to filter GenBank sequences and to determine that sequences are homologous to the gene regions of interest. Sequences that are judged not to be homologous are removed. Once sequences are identified to belong to the gene regions of interest, saturation analyses are conducted comparing uncorrected genetic distances to corrected distances. If alignments appear to be saturated, the alignments are broken up using prior phylogenetic knowledge (classification systems) as guides, and separate alignments are carried out for the individual groups delimited in this way. These individual alignments are then aligned together using profile-to-profile alignment techniques (Edgar, 2004). Our final concatenated data set included 55 473 species and 9853 aligned sites (Appendix S1; see Supplemental Data with the online version of this article). We conducted 223 maximum likelihood analyses using the standard RAxML search algorithm with the asymptotic stopping rule and the low memory consumption flag (−F and −D options) under the GTR+CAT approximation of rate heterogeneity partitioning for each gene (Stamatakis, 2006; see online Appendix S2). We also inferred 244 bootstrap trees using the RAxML rapid bootstrap algorithm (Stamatakis et al., 2008) to provide support values for the best-scoring ML tree and to compute strict, majority-rule, and extended majority rule consensus trees. We used the majority rule consensus tree from the ML analyses in conducting tests for shifts in diversification. Some taxa, because of either misalignment, lack of convergence with the phylogenetic algorithm, or particular missing data patterns, were clear outliers. This phenomenon of rogue taxa can be common in extremely large phylogenies (e.g., McMahon and Sanderson, 2006; Smith et al., 2009), and in this case, we identified 120 such taxa and removed them from the phylogeny.

If we assume that there are around 256 000 species of flowering plants based on a rough tally of the diversity contained with the major angiosperm groups (APweb, Stevens, 2010), our data set represents ca. 21% of known angiosperm diversity. However, this sample of species is unlikely to be distributed in proportion to the actual diversity of angiosperm subclades. Some groups will likely be over- or undersampled depending on difficulties in collecting, identifying, extracting, and sequencing, and on the chance interests and inclination of individual systematists. A key difference between the "backbone" tree approach described and carrying out analyses on a mega-phylogeny is the potential for bias introduced by such nonrandom sampling.

To explore this issue, we estimated the bias in sampling for each "order" of angiosperms recognized in the APG system (see Table 1). We relied on the *Z*-test statistic to compare the observed sampling proportion for an order (i.e., total number of species contained within an order included in our phylogenetic data set/55 000) to the expected sampling proportion. In our case, if sampling were completely random, our expectation is that roughly 21% of the species contained within any given order should be contained within our phylogenetic data set. We then standardized this difference in the observed and expected proportions by assuming a standard error of 0.04 based on the fact that our phylogenetic data set of more than 55 000 species could realistically represent anywhere between 17 and 25% of flowering plants (assuming a range of 225 000 to 325 000 species surrounding our estimate of flowering plant diversity) (see Scotland and Wortley, 2003).

Based on this analysis, 25 of the 59 orders are judged to be oversampled, meaning that the phylogenetic data set includes significantly more than 21% of the inferred diversity of the group in nature. Six orders appear to be undersampled, meaning that the phylogenetic data set includes significantly less than 21% of the inferred diversity (Table 1). Phylogenetic sampling is decidedly nonrandom in being oriented by existing classification systems and specifically aiming to include a good sampling of "isolated" groups in broad phylogenetic studies. Accordingly, it is not surprising that many of the cases of oversampling involve groups that contain only one or a few species and represent relatively deep splits in the angiosperm tree, such as *Amborella*, *Acorus*, *Trochodendron*, and Petrosaviales ($z = 16$, $P < 0.001$). In contrast, several small groups remain rather poorly sampled, such as Picramniales (3 of 46 species; $z = 2.6$, $P = 0.005$). In our data set, sampling is poorest in the Pandanales, from which only 71 of the 1345 species (ca. 5%; $z = 3$, $P = 0.001$) are included in our tree. Unexpectedly, considering the uncoordinated nature of the sampling effort, 28 of the 59 orders appear to be sampled roughly in proportion to their estimated diversity (Table 1). Included among these clades are a number of exceptionally large ones, such as Asterales ($z = 0.2$, $P = 0.42$), Asparagales ($z = 0$, $P = 0.50$), and Fabales ($z = 0.40$, $P = 0.345$). In fact, most of the clades with more than 10 000 species are represented at about the 20% level; of these, Myrtales ($z = 1.0$, $P = 0.16$) are at the low end with ca. 15%, and Ericales ($z = 0.6$, $P = 0.27$) are at the high end with ca. 23%.

Although we recognize that sampling across the angiosperms is still decidedly nonrandom overall, we were satisfied by these comparisons that the existing sample could potentially yield meaningful results in analyses of diversification rate. Indeed, it is possible that the nature of the sampling to date could have some advantages. For example, a totally random selection of ca. one fifth of the species of angiosperms could well miss such small, isolated clades entirely, which might bias against the

Table 1. Proportional sampling for the 59 angiosperm "orders" recognized by the Angiosperm Phylogeny Group (2009). Sampled proportions for each order were calculated by dividing the species included in the mega-phylogeny by the estimated diversity for the group. The sampling proportions were then compared against the assumption that they should be sampled according to an expected proportion of 21% (based on the proportional sampling of our phylogenetic data set) and we assessed whether they were significantly over- or undersampled or randomly sampled with respect to the expected proportion. Differences between the two proportions exceeding critical values according to a z-distribution ($\alpha = 0.05$) were considered significantly different.

Order (APG III, 2009)	Species included in megaphylogeny	Estimated diversity	Included/Estimated
Amborellales	1	1	1.0**
Nymphaeales	73	74	0.99**
Austrobaileyales	50	100	0.50**
Chloranthales	29	75	0.39**
Magnoliales	373	2929	0.13
Laurales	730	2858	0.26
Canellales	25	105	0.24
Piperales	443	4090	0.11*
Acorales	10	10	1.0**
Alismatales	735	4490	0.16
Petrosaviales	4	4	1.0**
Dioscoreales	160	1037	0.15
Pandanales	71	1345	0.05**
Liliales	688	1558	0.44**
Asparagales	5109	26 070	0.20
Arecales	371	2361	0.18
Poales	4189	18325	0.23
Commelinales	147	812	0.18
Zingiberales	740	2111	0.35**
Ceratophyllales	3	6	0.50**
Ranunculales	933	4445	0.21
Proteales	311	1610	0.19
Trochodendrales	2	2	1.0**
Buxales	22	72	0.31*
Gunnerales	33	45	0.73**
Saxifragales	837	2470	0.34**
Vitales	118	850	0.14
Zygophyllales	140	305	0.46**
Celastrales	152	1355	0.11*
Oxalidales	395	1815	0.22
Malpighiales	2670	15935	0.17
Fabales	4419	20055	0.22
Rosales	1814	7725	0.23
Cucurbitales	507	2295	0.22
Fagales	538	1877	0.29*
Geraniales	360	836	0.43**
Myrtales	1691	11027	0.15
Crossosomatales	28	66	0.42**
Picramniales	3	46	0.07**
Sapindales	992	5670	0.17
Huerteales	4	23	0.17
Malvales	950	6005	0.16
Brassicales	1683	4450	0.38**
Santalales	281	1985	0.14
Berberidopsidales	3	4	0.75**
Caryophyllales	2354	11155	0.21
Cornales	280	590	0.48**
Ericales	2684	11515	0.23
Garryales	9	18	0.50**
Gentianales	3122	16637	0.19
Lamiales	4550	23275	0.20
Solanales	1280	4080	0.31*
Aquifoliales	78	536	0.15
Asterales	5344	25790	0.21
Escalloniales	14	130	0.11*
Bruniales	67	79	0.85**
Apiales	1631	5489	0.30*
Paracryphiales	4	36	0.11*
Dipsacales	344	1090	0.32**

Notes: *, $P < 0.05$; **, $P < 0.01$

identification of significant shifts where these actually exist. In this sense, it is a good to think that molecular systematists have specifically targeted such small clades. Given that such clades have been sampled in the first place, the fact that they are relatively oversampled should bias in favor of false negative results (failing to infer a diversification shift where one actually exists), because it would tend to diminish the real difference in species numbers between them and their larger sister groups.

In addition to flowering plants as a whole, we focused on the same six major clades considered earlier (Monocotyledonae, Orchidaceae, Poaceae, Eudicotyledonae, Fabaceae, and Asteraceae) (Fig. 1B). Importantly, the results we obtained using the mega-phylogeny is generally congruent with those obtained using the backbone tree approach described. That is, with one exception, we failed to find a significant shift in diversification associated directly with the origin of these major clades, but instead found significant shifts nested not far within the clade. As an aside, because these tests were targeted to the origin of six specific groups, we did not use the Bonferroni correction for these specific tests. In the case of the angiosperms, we identified a major increase in diversification rate associated with the branch leading to the least inclusive clade that includes the magnoliids, monocots, and eudicots within the Mesangiospermae (P = 3.5e-07). In contrast to the backbone tree approach, we did not detect a significant upward shift at the base of the eudicots in the mega-phylogeny because *Ceratophyllum* was placed as sister to the rest of the mesangiosperms (as opposed to being sister to the eudicots in the backbone tree) and the species rich monocots were placed as sister to the eudicots. Also, we note that within eudicots there are major topological differences between the mega-phylogeny and the generally accepted relationships depicted in the backbone tree (see Fig. 1). In particular, core eudicots are shown as the sister group of the remaining eudicots, as opposed to being nested within the clade. This could be a function of the inclusion of species for which sequences are missing for most loci.

In Poaceae, we recovered a shift with the origin of the BEP/PACMAD clade (P = 1.6e-09), in Orchidaceae at the origin of the clade that includes Cypripedioideae and Epidendrioideae (p = 0.018). We recovered a shift in monocots after the divergence of Acorales (P = 0.0004) and in Asteraceae just above the divergence of Barnadesioideae (P = 0.003). The one exception was Fabaceae, where we did infer a shift in diversification rate directly at the base of the group using the mega-phylogeny (P = 1.7e-07), but we note that this depends on the exact placement of the species-poor Quillajaceae clade. The location of Quillajaceae as sister to Fabaceae in the phylogeny is novel (73% bootstrap support). Focused systematic studies of Fabales have alternated between placing Quillajaceae as sister to a clade comprised of Polygalaceae, Surianaceae, and Fabaceae (Forest et al., 2002; Banks et al., 2008), and placing Quillajaceae directly with Surianaceae (Wojciechowski et al., 2004; Bello et al., 2009).

Across the entire phylogeny, we detected ca. 2700 significant shifts, on roughly 4.9% of the internal branches. Unlike the tests performed, which were targeted to particular nodes, detecting shifts across an entire tree requires correcting for multiple tests. When adjusting significance according to a Bonferroni correction, the significance threshold is $P < 9e\text{-}07$, which we believe is overly conservative as function of their being such a large number of tests on such a large tree. With this threshold, only 16 significant shifts are detected, including two discussed above—at the origin of Mesangiospermae, and at the origin of Fabaceae. Our sense is that the number of significant shifts probably lies between the 16 and 2700 significant shifts. Although we are uncertain about how to narrow in on a better number, we were fascinated by the patterns that emerged with the possibility of a very large number of shifts. Specifically, it appears that the 2700 shifts that we detected without the Bonferroni correction are relatively evenly distributed across the entire tree, as opposed to being especially concentrated in certain clades and sparse in others (Fig. 2). There is not a strong relationship between the percentage of significant shifts and the inferred diversity of the order-level clades, implying that there is no systematic bias at this level for larger clades to contain more shifts than smaller ones (Fig. 2B). In fact, considering the most diverse orders, we roughly find 4–6% of the nodes associated with shifts in diversification: e.g., Asterales = 6.4%, Fabaceae = 5.8%, and Orchidaceae = 4.8%. Furthermore, significant diversification shifts are not widely separated from one another, with the average minimum nodal distance between inferred events being 3 (95% CI = 2–6). Essentially, there is a frequently repeated pattern in which a pectinate series of relatively species-poor clades subtends a more species rich clade in which the two primary subclades are more nearly equal in species diversity. We also note that most of the shifts in diversification (71%) are shifts upward in diversification rate, with many fewer downward shifts in diversification. This pattern could be a result of the methods ability to detect upward vs. downward shifts, but it is also possible that there has been a general trend in angiosperms toward increasing diversification (Liow et al., 2010). As noted already, however, further analysis of these shifts will require a more detailed consideration of significance values and methods of correcting for multiple tests.

DISCUSSION

Comparing backbone tree and mega-phylogeny approaches—We contrast two approaches for examining shifts in diversification: (1) using a small backbone tree in which the terminals often represent many species and (2) using a mega-phylogeny without assigning any additional species to the terminals. In the case of the backbone tree approach, diversity estimates obtained from published sources are assigned to the relevant tips of the phylogeny (cf. Alfaro et al., 2009; Santini et al., 2009). One benefit of this approach is that it can take into account the inferred diversity of clades. The main drawback is that the backbone trees are generally small, and the terminals effectively become "black-boxes." Consequently, if a shift is inferred along the stem leading to a tip, we do not know whether that shift occurred along that branch or whether it occurred somewhere within the species rich terminal. For this reason, we have not marked these terminal shifts in the trees in Fig. 1 and have instead marked only the significant shifts that were inferred along internal branches.

In contrast, in using the mega-phylogeny approach, we have not assigned species numbers to any of the tips and instead have inferred diversification shifts based only on the species included in the tree. This approach is obviously susceptible to potential sampling biases—that is, it depends to some extent on their having been more or less random sampling throughout the tree. We know that this is not true across the angiosperms, but we were pleasantly surprised to find that many large order-level clades appear to have been sampled more or less in proportion to their inferred diversities. This seems to be the case for the major named clades that we have focused on here. On the other hand, smaller, relatively early-diverging lineages have often been oversampled, which to some extent should bias against seeing significant shifts in their vicinity. In any case, this general approach offers some distinct advantages, particularly the potential to quantitatively explore general patterns across very large clades such as the angiosperms.

Interestingly, the two approaches yielded generally similar results with respect to the location of diversification shifts in the major angiosperm clades that we have examined. We find this agreement somewhat comforting. It suggests that backbone approaches can yield meaningful results, but also that our sequence databases have matured to the point that studies using mega-phylogenies can yield similar results, despite nonrandom sampling and considerable phylogenetic uncertainty. Phylogenetic uncertainty is especially worth noting in this context and is highlighted by the two major differences we found between the two approaches, namely, for Fabaceae and the Eudicotyledonae. The placement of Quillajaceae in our mega-phylogeny in relation to the origin of Fabaceae may be incorrect based on other analyses (see Forest et al., 2002; Wojciechowski et al., 2004; Banks et al., 2008; Bello et al., 2009). Yet, this placement contributes to the finding of a significant shift in diversification along the branch subtending the Fabacaeae. By contrast, the placement of *Ceratophyllum* as sister to the rest of the mesangiosperms, as opposed to sister to the eudicots, and the placement of the core eudicots itself, contribute to not recovering a shift in diversification along the branch subtending the eudicots. We worry about the accuracy of phylogenetic inferences in conducting such large analyses and in the face of so much missing data, and we caution that these problems may compromise the down-stream use of mega-phylogenies for some purposes.

Of course, we must also note that agreements between results using these two approaches do nothing to guarantee that the results are correct or that we have identified true underlying causes. Take, for example, our finding that a shift in diversification did not occur at the base of the angiosperms, but instead subtending the Mesangiospermae. It is possible that there was an upward shift in diversification at this point, perhaps owing to a feature (an "innovation") that evolved along that branch, such as the congenital closure of the carpel (Endress and Igersheim, 2000). However, it is also possible that an upward shift did not occur at this point. Instead, there may have been an inordinate number of species extinctions within the branches represented by extant *Amborella*, Nymphaeales, and Austrobaileyales. Perhaps these branches once each contained thousands of species, and speciation occurred at roughly the same rate as in the mesangiosperm clade. If so, "key innovation" or "key opportunity" explanations would not be called for; instead, we would want to focus on whatever factors resulted in the differential extinction rates. Unfortunately, the fossil record may remain too sparse to

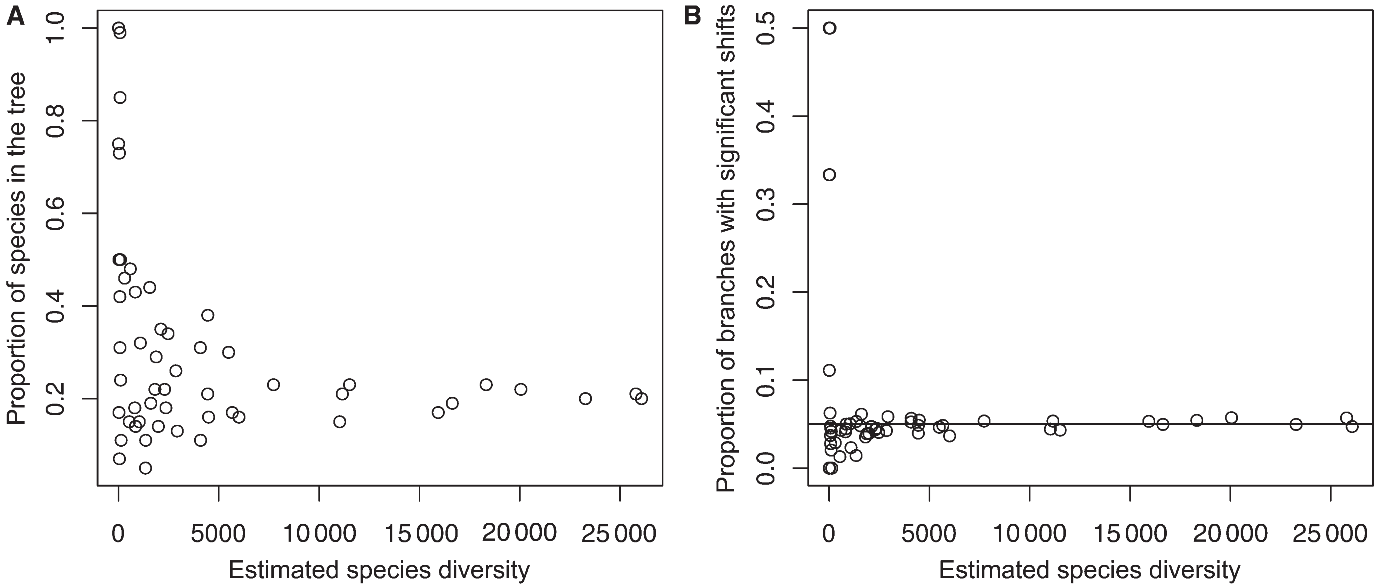

Fig. 2. (A) Scatter plot of known diversity in orders vs. the observed diversity in the large-scale phylogenetic approach. (B) Scatter plot of the known diversity in orders vs. the proportion of branches that are estimated to have significant shifts in diversification. The solid line denotes the 5% type I error rate of the Δ_i statistic.

provide an accurate estimate of the number of species that once existed in these lineages, and the estimation of extinction rates from knowledge of extant taxa remains a difficult problem (e.g., Rabosky, 2009; but see Wertheim and Sanderson, 2010). In general, it is difficult to disentangle the effects of shifts in speciation from shifts in extinction (Ricklefs, 2007; Rabosky, 2009).

Significance for angiosperm diversification—Taken at face value, our results demonstrate that shifts in diversification are not consistently associated with the origin of major named groups such as the ones examined here. Instead, shifts in diversification may often be inferred not far within these clades. These findings are consistent with the view that radiations tend to be lit by a long "fuse" (Cooper and Fortey, 1998), and also with the idea that an initial innovation enables subsequent experimentation and, eventually, the evolution of a combination of characteristics that drives a major radiation (Donoghue, 2005). For example, it is possible that congenital fusion in the mesangiosperm line of the originally unsealed carpel (Endress and Igersheim, 2000) in some way resulted in an increased rate of diversification. In other cases, the proximate drivers of diversification may not be direct modifications of the identifying characteristic(s) of the named clade. For example, in legumes, diversification may have been driven by changes in flower architecture, not by modifications of the legume fruit itself. As always, however, it will be very challenging in individual cases to determine all the factors (intrinsic or extrinsic) that have contributed to apparent radiations, and factors such as gene and genome duplication or biogeographic movement may well have been involved (e.g., De Bodt et al., 2005; Moore and Donoghue, 2007). And, as we emphasized, the real drivers may have had their impacts on rates of extinction in different lineages.

These considerations highlight that named taxonomic groups have a tendency to orient our evolutionary studies. Specifically, our views on diversification have often focused on diverse clades that have long been named. These are the groups about which we tend to make generalizations and for which we seek causal mechanisms. We have often, therefore, concentrated on the characteristics of these groups as possible drivers. Our analyses suggest that we have mis-attributed diversification rate shifts to many of these species-rich named groups. In part, we think that this is because we have, until recently, failed to properly recognize the truly diverse clades nested within the traditional groups and to properly attach names to these clades (e.g., Mesangiospermae within Angiospermae, or Eudicotyledonae). Moving forward, we expect that the accelerated naming of clades (e.g., Cantino, et al., 2007) will allow us to more precisely identify and explain patterns of diversification. It is worth noting, however, what we suspect is an important methodological asymmetry: we may generally be more confident in identifying where diversification rate shifts have not occurred (i.e., not directly subtending particular named clades) than we will be in identifying precisely where they have occurred.

Our analyses using the mega-phylogeny point to a number of patterns that could not previously have been recognized and quantified, and we believe that further studies along these lines will open up new avenues of research. As we have suggested, it would appear that shifts in diversification are quite common and perhaps more uniformly distributed throughout the angiosperm tree than we might have expected. This pattern requires further analysis, but if it holds up it could lead to a profound shift in our outlook on such problems. A key issue will become how to distinguish real shifts in diversification from noise and statistical artifacts. As it emerges, the total of ca. 2700 diversification shifts that we have inferred throughout the tree is not far from the 5% error rate and could have come about through chance alone (i.e., we expect the Δ_i statistic to incorrectly reject the null hypothesis in favor of a significant shift about 5% of the time). Although we believe that this is unlikely and that there have been genuine shifts in diversification rate, both up and down, the underlying error rate will make it difficult to confidently identify significant shifts in diversification when they indeed exist.

It seems clear that as we pursue such questions on this very large scale we will need to focus more attention on properly specifying null expectations. Perhaps it may even be necessary to abandon methods that seek to reject or confirm a hypothesis based on some underlying type I error probability in favor of methods that identify a model that best approximates the information contained in the data (Burnham and Anderson, 2002). The step-wise Akaike information criterion (AIC) framework, proposed by Alfaro et al. (2009), seems like an important step in this direction. This method measures the fit of rate shifts, added in a stepwise manner, until the addition of new parameters exhausts the information contained within the tree. So far, this method has only been applied to backbone trees and, therefore, has not fully confronted the complexity inherent in trees the size of our mega-phylogeny. In this regard, we are optimistic that with the rapid growth of sequence data in GenBank, and with the continued development of new methods, coupled with a better understanding of the timing of the origin and radiation of angiosperms (Bell et al., 2010; Magallon, 2010; Smith et al., 2010), we will soon be able to explore more parameter-rich models that may better capture the complexity of the processes that affect flowering plant diversification.

A NOTE ON THE UTILITY AND AVAILABILITY OF THE MEGA-PHYLOGENY

The mega-phylogeny featured here is currently the largest one for flowering plants. As we have emphasized, there are a variety of problems associated with the inference of a tree of this size, and these may limit its utility in some contexts. However, we believe that it may be useful for some comparative studies, and there are several other arenas in which this tree could be of immediate value. For example, it could be useful in testing methods for visualizing large trees, or in the development of methods and programs for carrying out comparative analyses on this scale. The tree and the underlying data set can be downloaded from Dryad (http://datadryad.org) and can also be obtained from the first author.

LITERATURE CITED

Agapow, P., and A. Purvis. 2002. Power of eight tree shape statistics to detect non-random diversification: A comparison by simulation of two models of cladogenesis. *Systematic Biology* 51: 866–872. doi:10.1080/10635150290102564

Alfaro, M. E., F. Santini, C. Brock, H. Alamillo, A. Dornburg, D. L. Rabosky, G. Carnevale, and L. J. Harmon. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proceedings of the National Academy of Sciences, USA* 106: 13410–13414. doi:10.1073/pnas.0811087106

Alvarez, I., and J. F. Wendel. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution* 29: 417–434. doi:10.1016/S1055-7903(03)00208-2

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.

Banks, H., B. B. Klitgaard, F. Claxton, F. Forest, and P. R. Crane. 2008. Pollen morphology of the family Polygalaceae (Fabales). *Botanical Journal of the Linnean Society* 156: 253–289. doi:10.1111/j.1095-8339.2007.00723.x

Bell, C. D., D. E. Soltis, and P. S. Soltis. 2010. The age and diversification of the angiosperm re-revisited. *American Journal of Botany* 97: 1296–1303. doi:10.3732/ajb.0900346

Bello, M. A., A. Bruneau, F. Forest, and J. A. Hawkins. 2009. Elusive relationships within order Fabales: Phylogenetic analyses using *matK* and *rbcL* sequence data. *Systematic Botany* 34: 102–114. doi:10.1600/036364409787602348

Bruneau, A., M. Mercure, G. L. Lewis, and P. S. Herendeen. 2008. Phylogenetic patterns and diversification in the caesalpinioid legumes. *Botany* 86: 697–718. doi:10.1139/B08-058

Burnham, K. P., and D. R. Anderson. [eds.]. 2002. Model selection and multimodel inference: A practical information–theoretic approach. Springer, New York, New York, USA.

Cameron, K. M. 2006. A comparison and combination of plastid *atpB* and *rbcL* gene sequences for inferring phylogenetic relationships within Orchidaceae. *Aliso* 22: 447–464.

Cantino, P. D., J. A. Doyle, S. W. Graham, W. S. Judd, R. G. Olmstead, D. E. Soltis, P. S. Soltis, and M. J. Donoghue. 2007. Toward a phylogentic nomenclature of Tracheophyta. *Taxon* 56: 822–846. doi:10.2307/25065865

Chan, K. M. A., and B. R. Moore. 2002. Whole-tree methods for detecting differential diversification rates. *Systematic Biology* 51: 855–865. doi:10.1080/10635150290102555

Chan, K. M. A., and B. R. Moore. 2005. SymmeTREE: Whole-tree analysis of differential diversification rates. *Bioinformatics (Oxford, England)* 21: 1709–1710. doi:10.1093/bioinformatics/bti175

Chase, M. W., M. F. Fay, D. Devey, O. Maurin, N. Rønsted, J. Davies, Y. Pillon, et al. 2006. Multigene analyses of monocot relationships: A summary. *Aliso* 22: 63–75.

Cooper, A., and R. Fortey. 1998. Evolutionary explosions and the phylogenetic fuse. *Trends in Ecology & Evolution* 13: 151–156. doi:10.1016/S0169-5347(97)01277-9

Crane, P. R., E. M. Friis, and K. R. Pederson. 1995. The origin and early diversification of angiosperms. *Nature* 374: 27–33.doi:10.1038/374027a0

Crepet, W. L., and K. J. Niklas. 2009. Darwin's second "abominable mystery": Why are there so many angiosperm species? *American Journal of Botany* 96: 366–381. doi:10.3732/ajb.0800126

Darwin, F., and A. C. Seward. [eds.]. 1903. More letters of Charles Darwin. John Murray, London, UK.

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 Sciences, USA* 101: 1904–1909. doi:10.1073/pnas.0308127100

Davis, J. I., D. W. Stevenson, G. Petersen, O. Seberg, L. M. Campbell, J. V. Freudenstein, D. H. Goldman, et al. 2004. A phylogeny of the monocots, as inferred from *rbc*L and *atp*A sequence variation, and a comparison of methods for calculating jackknife and bootstrap values. *Systematic Botany* 29: 467–510. doi:10.1600/0363644041744365

De Bodt, S., S. Maere, and Y. Van de Peer. 2005. Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution* 20: 591–597. doi:10.1016/j.tree.2005.07.008

Donoghue, M. J. 2005. Key innovations, convergence, and success: Macroevolutionary lessons from plant phylogeny. *Paleobiology* 31: 77–93. doi:10.1666/0094-8373(2005)031[0077:KICASM]2.0.CO;2

Doyle, J. A. 2001. Significance of molecular phylogenetic analyses for paleobotanical investigations on the origin of angiosperms. *Palaeobotanist* 50: 167–188.

Doyle, J. A., and M. J. Donoghue. 1993. Phylogenies and angiosperm diversification. *Paleobiology* 19: 141–167.

Edgar, R. C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797. doi:10.1093/nar/gkh340

Edwards, E. J., and S. A. Smith. 2010. Phylogenetic analyses reveal the shady history of C_4 grasses. *Proceedings of the National Academy of Sciences, USA* 107: 2532–2538. doi:10.1073/pnas.0909672107

Endress, P. K., and A. Igersheim. 2000. Gynoecium structure and evolution in basal angiosperms. *International Journal of Plant Sciences* 161 (supplement): S211–S223. doi:10.1086/317572

Forest, F., A. Bruneau, J. Hawkins, T. Kajita, J. J. Doyle, and P. R. Crane. 2002. The sister of the Leguminosae revealed: Phylogenetic relationships in the Fabales determined using *trn*L and *rbc*L sequences.

In Botany 2002: Botany in the Curriculum, 124 [abstract], Madison, Wisconsin, USA. Botanical Society of America, St. Louis, Missouri, USA.

Friedman, W. E. 2009. The meaning of Darwin's abominable mystery. *American Journal of Botany* 96: 5–21. doi:10.3732/ajb.0800150

Funk, V. A., R. J. Bayer, S. Keeley, R. Chan, L. Watson, B. Gemeinholzer, E. Schilling, et al. 2005. Everywhere but Antarctica: Using a supertree to understand the diversity and distribution of the Compositae. *Biologiske Skrifter* 55: 343–374.

Gaut, B. S., S. Muse, W. D. Clark, and M. T. Clegg. 1992. Relative rates of nucleotide substitution at the *rbcL* locus of monocotyledonous plants. *Journal of Molecular Evolution* 35: 292–303. doi:10.1007/BF00161167

Goldberg, E. E., J. R. Kohn, R. Lande, K. A. Robertson, S. A. Smith, and B. Igic. 2010. Species selection maintains self-incompatibility. *Science* 328: 587–591. doi:10.1126/science.1177216

Goloboff, P. A. 1999. Analyzing large data sets in reasonable times: Solutions for composite optima. *Cladistics* 15: 415–428. doi:10.1111/j.1096-0031.1999.tb00278.x

Goloboff, P. A., S. A. Catalano, J. M. Mirande, C. A. Szumik, J. S. Arias, M. Källersjö, and J. S. Farris. 2009. Phylogenetic analysis of 73060 taxa corroborates major eukaryotic groups. *Cladistics* 25: 211–230. doi:10.1111/j.1096-0031.2009.00255.x

Grass Phylogeny Working Group. 2001. Phylogeny and subfamilial classification of the grasses (Poaceae). *Annals of the Missouri Botanical Garden* 88: 373–457. doi:10.2307/3298585

Guyer, C., and J. B. Slowinski. 1993. Adaptive radiation and the topology of large phylogenies. *Evolution; International Journal of Organic Evolution* 47: 253–263.

Hartmann, S., and T. Vision. 2008. Can one accurately infer a phylogenetic tree from a gappy alignment? *BMC Evolutionary Biology* 8: 95. doi:10.1186/1471-2148-8-95

Hejnol, A., M. Obst, A. Stamatakis, M. Ott, G. W. Rouse, G. D. Edgecombe, P. Martinez, et al. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings. Biological Sciences* 276: 4261–4270. doi:10.1098/rspb.2009.0896

Holman, E. W. 2005. Nodes in phylogenetic trees: The relation between imbalance and number of descendent species. *Systematic Biology* 54: 895–899. doi:10.1080/10635150500354696

Jansen, R. K., Z. Cai, L. A. Raubeson, H. Daniell, C. W. dePamphilis, J. Leebens-Mack, and K. F. Müller. et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proceedings of the National Academy of Sciences, USA 104: 19369–19374. doi:10.1073/pnas.0709121104

Katoh, K., and H. Toh. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics* 9: 286–298. doi:10.1093/bib/bbn013

Leebens-Mack, J., L. A. Raubeson, L. Cui, J. V. Kuehl, M. H. Fourcade, T. W. Chumley, J. L. Boore, R. K. Jansen, and C. W. dePamphilis. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: Sampling one's way out of the Felsenstein zone. *Molecular Biology and Evolution* 22: 1948–1963. doi:10.1093/molbev/msi191

Lemmon, A. R., J. M. Brown, K. Stanger-Hall, and E. M. Lemmon. 2009. The effect of missing data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Systematic Biology* 58: 130–145. doi:10.1093/sysbio/syp017

Liow, L. H., T. B. Quental, and C. R. Marshall. 2010. When can decreasing diversification rates be detected with molecular phylogenies and the fossil record? *Systematic Biology* 59: 646–659. doi:10.1093/sysbio/syq052

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. doi:10.1093/sysbio/syq027

Magallón, S., and M. J. Sanderson. 2001. Absolute diversification rates in angiosperm clades. *Evolution; International Journal of Organic Evolution* 55: 1762–1780.

McMahon, M. M., and M. J. Sanderson. 2006. Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. *Systematic Biology* 55: 818–836. doi:10.1080/10635150600999150

Mooers, A. O., and S. B. Heard. 1997. Inferring evolutionary process from the phylogenetic tree shape. *The Quarterly Review of Biology* 72: 31–54. doi:10.1086/419657

Moore, B. R., K. M. A. Chan, and M. J. Donoghue. 2004. Detecting diversification rate variation in supertrees. *In* O. R. P. Bininda-Emonds [ed.], Phylogenetic supertrees: Combining information to reveal the tree of life, 487–533. Kluwer, Dordrecht, Netherlands.

Moore, M. J., C. D. Bell, P. S. Soltis, and D. E. Soltis. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proceedings of the National Academy of Sciences, USA* 104: 19363–19368. doi:10.1073/pnas.0708072104

Moore, B. R., and M. J. Donoghue. 2007. Correlates of diversification in the plant clade Dipsacales: Geographic movement and evolutionary innovations. *American Naturalist* 170: S28–S55. doi:10.1086/519460

Moore, B. A., and M. J. Donoghue. 2009. A Bayesian approach for evaluating the impact of historical events on rates of diversification. *Proceedings of the National Academy of Sciences, USA* 106: 4307–4312. doi:10.1073/pnas.0807230106

Nee, S. 2001. Inferring speciation rates from phylogenies. *Evolution; International Journal of Organic Evolution* 55: 661–668.

Nee, S., A. Mooers, and P. H. Harvey. 1992. Tempo and mode of evolution revealed from molecular phylogenies. *Proceedings of the National Academy of Sciences, USA* 89: 8322–8326. doi:10.1073/pnas.89.17.8322

Panero, J. L., and V. A. Funk. 2002. Toward a phylogenetic classification for the Compositae (Asteraceae*). Proceedings of the Biological Society of Washington* 115: 909–922.

Panero, J. L., and V. A. Funk. 2008. The value of sampling anomalous taxa in phylogenetic studies: Major clades of the Asteraceae revisited. *Molecular Phylogenetics and Evolution* 47: 757–782. doi:10.1016/j.ympev.2008.02.011

Purvis, A., A. Katzourakis, and P. Agapow. 2002. Evaluating phylogenetic tree shape: Two modifications to Fusco and Cronk's method. *Journal of Theoretical Biology* 214: 99–103. doi:10.1006/jtbi.2001.2443

Qiu, Y., L. Li, B. Wang, Z. Chen, V. Knoop, M. Groth-Malonek, and O. Dombrovska. et al. 2006. The deepest divergences in land plants inferred from phylogenomic evidence. Proceedings of the National Academy of Sciences, USA 103: 15511–15516. doi:10.1073/pnas.0603335103

Rabosky, D. L. 2009. Extinction rates should not be estimated from molecular phylogenies. *Evolution; International Journal of Organic Evolution* 64: 1816–1824.

Rabosky, D. L., and I. J. Lovette. 2008. Density-dependent diversification in North American wood warblers. *Proceedings of the Royal Society of London. Series B. Biological Sciences* 275: 2363–2371.

Raikow, R. J. 1986. Why are there so many kinds of passerine birds? *Systematic Zoology* 35: 255–259. doi:10.2307/2413436

Ramírez, S. R., B. Gravendeel, R. B. Singer, C. R. Marshall, and N. E. Pierce. 2007. Dating the origin of Orchidaceae from a fossil orchid with its pollinator. *Nature* 448: 1042–1045. doi:10.1038/nature06039

Ricklefs, R. E. 2007. Estimating diversification rates from phylogenetic information. *Trends in Ecology & Evolution* 22: 601–610. doi:10.1016/j.tree.2007.06.013

Sanderson, M. J., and M. J. Donoghue. 1994. Shifts in diversification rate with the origin of angiosperms. *Science* 264: 1590–1593. doi:10.1126/science.264.5165.1590

Sanderson, M. J., and M. J. Donoghue. 1996. Reconstructing shifts in diversification on phylogenetic trees. *Trends in Ecology & Evolution* 11: 15–20. doi:10.1016/0169-5347(96)81059-7

Sanderson, M. J., M. M. McMahon, and M. Steel. 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. *BMC Evolutionary Biology* 10: 155. doi:10.1186/1471-2148-10-155

Santini, F., L. J. Harmon, G. Carnevale, and M. Alfaro. 2009. Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. *BMC Evolutionary Biology* 9: 194. doi:10.1186/1471-2148-9-194

Scotland, R. W., and A. H. Wortley. 2003. How many species of seed plant are there? *Taxon* 52: 101–104. doi:10.2307/3647306

Slowinski, J. B., and C. Guyer. 1989. Testing the stochasticity of patterns of organismal diversity: An improved null model. *American Naturalist* 134: 907–921. doi:10.1086/285021

Smith, S. A., and J. M. Beaulieu. 2009. Life history influences rates of climatic niche evolution in flowering plants. *Proceedings of the Royal Society of London. Series B. Biological Sciences* 276: 4345–4352. doi:10.1098/rspb.2009.1176

Smith, S. A., J. M. Beaulieu, and M. J. Donoghue. 2009. Megaphylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. *BMC Evolutionary Biology* 9: 37. doi:10.1186/1471-2148-9-37

Smith, S. A., J. M. Beaulieu, and M. J. Donoghue. 2010. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. *Proceedings of the National Academy of Sciences, USA* 107: 5897–5902. doi:10.1073/pnas.1001225107

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

Soltis, D. E., A. S. Chanderbali, S. Kim, M. Buzgo, and P. S. Soltis. 2007. The ABC model its applicability to basal angiosperms. *American Journal of Botany* 100: 155–163. doi:10.1093/aob/mcm117

Soltis, D. E., S. A. Smith, N. Cellinese, M. J. Moore, C. C. Davis, K. J. Wurdack, S. F. Brockington, et al. In press. Inferring angiosperm phylogeny: 17 gene analyses. *American Journal of Botany*.

Stamatakis, A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)* 22: 2688–2690. doi:10.1093/bioinformatics/btl446

Stamatakis, A., P. Hoover, and J. Rougemont. 2008. A rapid bootstrap algorithm for the RaxML web servers. *Systematic Biology* 57: 758–771. doi:10.1080/10635150802429642

Stevens, P. F. 2010. Angiosperm Phylogeny Website, version 9, June 2008 [and more or less continuously updated since]. Website http://www.mobot.org/MOBOT/research/APweb/.

Thomson, R. C., and H. B. Shaffer. 2010. Rapid progress on the vertebrate tree of life. *BMC Biology* 8: 19. doi:10.1186/1741-7007-8-19

Wertheim, J. O., and M. J. Sanderson. 2010. Estimating diversification rates: how useful are divergence times? *Evolution; International Journal of Organic Evolution* 65: 309–320.

Wiens, J. J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. *Systematic Biology* 52: 528–538. doi:10.1080/10635150390218330

Wiens, J. J. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? *Systematic Biology* 54: 731–742. doi:10.1080/10635150500234583

Wojciechowski, M. F., M. Lavin, and M. J. Sanderson. 2004. A phylogeny of legumes (Leguminosae) based on analysis of the plastid *matK* gene resolves many well-supported subclades within the family. *American Journal of Botany* 91: 1846–1862. doi:10.3732/ajb.91.11.1846

Worberg, A., D. Quandt, A. M. Barniske, C. Löhne, K. W. Hilu, and T. Borsch. 2007. Phylogeny of basal eudicots: Insights from noncoding and rapidly evolving DNA. *Organisms, Diversity & Evolution* 7: 55–77. doi:10.1016/j.ode.2006.08.001

Zwickl, D. J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, University of Texas at Austin, Austin, Texas, USA.