Article Type: Special Issue Article

RESEARCH ARTICLE

INVITED SPECIAL ARTICLE

For the Special Issue: Using and Navigating the Plant Tree of Life

Short Title: Smith and Brown—Constructing a broadly inclusive seed plant phylogeny

Constructing a broadly inclusive seed plant phylogeny

Stephen A. Smith[1,2] and Joseph W. Brown[1]

Manuscript received 8 August 2017; revision accepted 19 October 2017.

[1]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, 48109, USA

[2]Corresponding author (e-mail: eebsmith@umich.edu); ORCID id 0000-0003-2035-9531

Citation: Smith, S. A. and J. W. Brown. 2018. Constructing a broadly inclusive seed plant phylogeny. *American Journal of Botany* 105(3): XXX.

DOI: XXXX

PREMISE OF THE STUDY: Large phylogenies can help shed light on macroevolutionary patterns that inform our understanding of fundamental processes that shape the tree of life. These phylogenies also serve as tools that facilitate other systematic, evolutionary, and ecological analyses. Here we combine genetic data from public repositories (GenBank) with phylogenetic data (Open Tree of Life project) to construct a dated phylogeny for seed plants.

METHODS: We conducted a hierarchical clustering analysis of publicly available molecular data for major clades within the Spermatophyta. We constructed phylogenies of major clades, estimated divergence times, and incorporated data from the Open Tree of Life project, resulting in a seed plant phylogeny. We estimated diversification rates, excluding those taxa without

molecular data. We also summarized topological uncertainty and data overlap for each major clade.

KEY RESULTS: The trees constructed for Spermatophyta consisted of 79,881 and 353,185 terminal taxa; the latter included the Open Tree of Life taxa for which we could not include molecular data from GenBank. The diversification analyses demonstrated nested patterns of rate shifts throughout the phylogeny. Data overlap and inference uncertainty show significant variation throughout and demonstrate the continued need for data collection across seed plants.

CONCLUSIONS: This study demonstrates a means for combining available resources to construct a dated phylogeny for plants. However, this approach is an early step and more developments are needed to add data, better incorporating underlying uncertainty, and improve resolution. The methods discussed here can also be applied to other major clades in the tree of life.

KEY WORDS: clustering; divergence-time estimation; diversification; GenBank; Open Tree of Life; phylogenetics; phylogenetic methods; plant tree of life; seed plants

The promise of a comprehensive view of extant diversity, whether for a single clade or the entire tree of life, has been a major motivation of the systematics community for decades—arguably centuries. Not only does a more complete view of the tree of life excite the imagination of both evolutionary biologists and the public, but broader and more complete phylogenies allow the exploration of evolutionary, biogeographic, and ecological questions at a scope that cannot be achieved with smaller phylogenies (Smith and Beaulieu, 2009; Goldberg et al., 2010; Edwards et al., 2010; Smith et al., 2011; Rabosky et al., 2013; Cornwell et al., 2014; Zanne et al., 2014; Tank et al., 2015; O'Meara et al., 2016). Individual, thoroughly studied systems are fundamentally important to evolutionary research and provide unprecedented details that facilitate in-depth analyses and exploration. Large-scale phylogenetic analyses often contain more variation and error, but provide different perspectives that often address entirely different questions. Both large- and small-scale phylogenetic studies can be useful for addressing and developing evolutionary hypotheses.

The first large seed plant phylogeny (Chase et al., 1993) paved the way for what would become an important research component for plant phylogenetics and evolution. Large

phylogenetic trees have been used in plants to address rates of molecular evolution (Smith and Donoghue, 2008), ecological questions (Beaulieu et al., 2012; Cornwell et al., 2014), evolution of climate tolerance (Smith and Beaulieu, 2009; Edwards and Smith, 2010; Edwards et al., 2010; Zanne et al., 2014), flower evolution (O'Meara et al., 2016; Sauquet et al., 2017), genome duplications (Tank et al., 2015; Smith et al., 2017), and diversification (Smith et al., 2011). While these studies contributed to discussions about large-scale patterns of plant evolution, caution needs to be practiced when interpreting them and improvements in the underlying phylogenies will continue to increase their utility and accuracy (Beaulieu et al., 2012; Hinchliff and Smith, 2014; Edwards et al., 2015).

Researchers have constructed these enormous phylogenies in several ways. For example, many researchers have conducted analyses of publicly available molecular data in NCBI's GenBank (Driskell et al., 2004; McMahon and Sanderson, 2006; Smith et al., 2011; Bocak et al., 2013). While some focused on constructing data sets for specific gene regions such as 18S or *rbc*L, others have constructed data sets intended for supermatrix analysis (Hibbett et al., 2005; Goloboff et al., 2009). Tools such as PhyLoTa were developed to automate and pre-calculate clusters of data for clades in the tree of life and provide a means of browsing the results of these analyses (Sanderson et al., 2008). Smith et al. (2009) developed PHLAWD to conduct a so-called "baited" analysis where gene regions may be identified a priori thereby dramatically speeding up clustering analyses. This procedure was extended with PUmPER to allow for automatic updating when new sequences become available (Izquierdo-Carrasco et al., 2014). Several newly developed software packages have built on these methods including SUMAC (Freyman, 2015) that incorporates both "baited" analyses and single-linkage clustering methods as well as a novel means of determining when there are enough overlapping data, and SUPERSMART (Antonelli et al., 2017) that includes analyses from clustering to divergence-time estimation. Recently, analyses that can accommodate DNA barcoding sequences have also been developed (Chesters, 2017).

While methods have been developed to analyze publicly available data, molecular data are not available for all taxa. To overcome this challenge when constructing comprehensive phylogenies, researchers have synthesized other sources. Jetz et al. (2012) combined molecular data available in public databases with taxonomic information for data-deficient taxa to construct a comprehensive phylogeny of Aves. Beaulieu et al. (2012) manually synthesized phylogenies to

construct a tree that could be used for comparative ecological studies. More recently, the Open Tree of Life presented a draft tree of all life constructed from a synthetic taxonomy and a phylogenetic synthesis analysis based on sets of published phylogenies contributed and curated by the community of systematists (Smith et al., 2013; Hinchliff et al., 2015; Redelings and Holder, 2017). The taxonomy, called OT-taxonomy, was constructed through combining taxonomies from different sources (e.g., NCBI, and domain-specific resources) while excluding "nonphylogenetic" taxa (e.g., incertae sedis). OT-taxonomy attempts to be comprehensive and is updated as component taxonomies are updated or as refinement edits are contributed (Rees and Cranston, 2017). The Open Tree of Life project also has resources that allows researchers to contribute phylogenies (McTavish et al., 2015) that can then be synthesized into a comprehensive tree of life (Hinchliff et al., 2015). In addition to the updates provided by the community, synthesis methods that combine phylogenies and taxonomy also continue to improve (Redelings and Holder, 2017). In this study, we aimed to use these resources along with other molecular data to construct resolved and dated "comprehensive" phylogenies. Here, by "comprehensive", we mean that the trees include the taxa in the Open Tree of Life taxonomy, regardless of whether these taxa have molecular data available. However, many clades may still require significant taxonomic revision or examination to determine species composition. Smaller scale studies that detail the systematics within these clades will continue to improve the taxonomies and phylogenies.

One consistent limitation of the synthetic trees produced by the Open Tree of Life project is the lack of branch lengths, whether molecular or relative to divergence times (Hinchliff et al., 2015). While branching order informs how species are related, branch lengths are necessary for conducting many other downstream comparative analyses. Calibrations, in the form of fossil data or secondary calibrations, necessary for conducting divergence time analyses are available through public resources such as the paleobioDB (https://paleobiodb.org; http://fossilworks.org), Fossil Calibration Database (Ksepka et al., 2015), TimeTree (Hedges et al., 2006), and DateLife (http://datelife.org) projects. These may be useful for large comprehensive phylogenetic projects such as the Open Tree of Life, but have yet to be incorporated. There are other ways in which branch lengths, relative to time, can be incorporated into large phylogenies. For example, tools such as CONGRUIFY (available in the R package GEIGER v. 2 [Harmon et al., 2008; Pennell et

al., 2014]) generate secondary calibrations from a dated tree and apply them to an undated tree (Eastman et al., 2013).

Here, we present a draft phylogeny for seed plants that includes divergence times. We used a hierarchical divide-and-conquer approach for constructing data sets using publicly available molecular data and combine these data sets with the Open Tree of Life results. We then used existing resources to help calibrate and date the phylogenies we construct. While this phylogeny may be used for several purposes, as a preliminary exploration, we discuss patterns of diversification and areas of phylogenetic uncertainty. We also discuss significant limitations in the existing data and the need for further developments moving forward.

<H1>MATERIALS AND METHODS

<h2>*Data description*—

We used the available information in GenBank release 218 (ftp://ftp.ncbi.nlm.nih.gov/genbank) as downloaded and processed by phlawd_db_maker (available at https://github.com/blackrim/phlawd_db_maker). With some exceptions, we excluded most sequences with fewer than 600 bp because we found many of these to have incorrect species identifications or insufficient information for resolution (see Discussion for more details). This problem was often worse in more complex and speciose clades. We also excluded genomic sequences (often in the form of mitogenomes or plastomes) because the size of the large sequences precludes efficient incorporation.

We used the Open Tree of Life synthetic tree release 9.1 and taxonomy version 3, which researchers can obtain from the Open Tree of Life website (https://tree.opentreeoflife.org/about/synthesis-release/v9.1). We noted some errors in the tree that needed to be fixed for our merging procedure to work correctly (procedure described below). For example, the Sapindales in release 9.1 of the Open Tree of Life synthetic tree is nonmonophyletic, and as a result, many of the taxa in that clade were found at the base of the eudicots. We removed these taxa because many would be added by molecular data. Our edited tree can be found at https://github.com/FePhyFoFum/big_seed_plant_trees along with the other data used for the project. Improvements and issues may be contributed at this website to continue updating and refining the phylogenies. We also obtained the time-calibrated phylogeny comprising 798 Spermatophyta taxa by Magallón et al. (2015) and used the inferred divergence

times as secondary calibrations for our analyses (more details in *Divergence time estimation* below).

<h2>Data set and phylogeny construction—

To construct a comprehensive phylogeny, we conducted a hierarchical analysis with individual phylogenies constructed for major clades (listed in Appendix S1, see Supplemental Data with this article) and placed into context based on the Open Tree of Life and Magallón et al. (2015). This procedure resulted in two comprehensive seed plant phylogenies: one with the deep branches resolved according to Magallón et al. (2015) and one with the deep branches resolved according to the Open Tree of Life release 9.1. We developed a new software package, PyPHLAWD, to construct data sets for each major clade of seed plants (see Appendix S1 for a list of clades). We describe the general procedure as it relates to these analyses here (see Fig. 1).

For each individual major clade, we conducted the following analyses. First, PyPHLAWD constructed folders for each clade, identified by NCBI taxonomy, within the major clade of interest. For example, for Apiales, PyPHLAWD constructed folders for Apiineae, Griseliniaceae, Pennantiaceae, and Torricelliaceae. Within each of these, PyPHLAWD then constructed folders for their respective subclades (e.g., within Apiineae, folders for Apiaceae, Araliaceae, Myodocarpaceae, and Pittosporaceae were created) and so on. Within the folder of the most nested clade, PyPHLAWD placed all sequences of the contained taxa. For example, the folder for genus *Sanicula* within the Apiales contained a file with 178 sequences. PyPHLAWD then conducted a clustering analysis consisting of an all-by-all blastn analysis, followed by a Markov cluster algorithm (MCL) (Dongen, 2000). Here, a cluster refers to a set of sequences that are potentially homologous (usually corresponding to a gene region). For blastn analyses, we considered successful hits to overlap at least by 65%, have an e-value of at least 10^{-10}, and to have identity of at least 20%. For MCL analyses, we used the options "--abc-neg-log10 -te 12 -tf 'gq(50)' -I 2.1". Once constructed, the clusters were placed in a folder within the clade folder (e.g., 23 clusters were placed in a folder called "clusters" within the folder for *Sanicula*). Alignments using MAFFT v.7.305b were constructed for each of these clusters (Katoh and Standley, 2013). These analyses were repeated for each tipward clade.

To construct clusters for each of the rootward clades (e.g., the parent of *Sanicula*, Saniuleae), we proceeded in a postorder fashion (i.e., from tips to root). At each rootward clade,

we would conduct the following analysis. If there was one subtending clade, the clusters of the child folder would be placed in the parent folder. If there was more than one subtending clade, the clusters from the first subtending clade, chosen arbitrarily, were placed in the parent folder. For each additional subtending clade, we conducted a blastn analysis of the subtending clade clusters and the parent clusters, which could result in multiple clusters hitting each other (e.g., if a cluster was split in a subtending clade because of poor overlap but was more complete in another set of sequences). We therefore constructed graphs where nodes represented clusters and, if there was a successful hit between sequences in different clusters, an edge connecting those nodes was placed. For each connected component that consisted of more than one cluster, we merged the sequences and created a new cluster in the parent clade. This procedure is similar to the algorithm used to construct initial clusters, cliques, by PhyLoTA (Sanderson et al. 2008) but applied between clusters. We then conducted a profile alignment, merging the subtending and parent clusters using MAFFT v.7.305b (Katoh and Standley, 2013). This process was repeated until the root of the major clade was reached (e.g., Apiales).

Once clusters were constructed, we built a supermatrix data set for each major clade listed in Appendix S1. If the clade had more than 100 taxa, we included clusters that contained at least 20% of the taxa included in the NCBI taxonomy. If the clade had fewer than 100 taxa, we included the cluster if it contained at least 70% of the taxa included in the NCBI taxonomy. This difference in percentages was intended to compensate for those gene regions sampled for tipward clades that have species level sampling (and so would be expected to have a high percentage of species included). Although this initial procedure was automated, we manually examined whether major gene regions (e.g., those sampled by the angiosperm tree of life project [Soltis et al., 2011]) were present in the set of clusters but missed in the supermatrix construction given the filters above. In these cases, the missed gene regions were added.

For each supermatrix, we constructed phylogenetic trees using RAxML v. 8.2.11 (Stamatakis, 2014) using the GTR+Γ molecular model partitioned by gene region. For the first analysis, we constrained all clades, as recognized by NCBI, to be monophyletic. Many of the rootward nodes in the angiosperm phylogeny require genomic or transcriptomic data to be resolved, data that are not included in these species-centric analyses. After conducting this constrained analysis, we tested constraints by calculating a quartet proportion measure (Pease et al., 2018, in this issue) and collapsing the node if less than 30% of the quartets supported the

clade. We then reran RAxML using the previous ML result as a constraint with the taxa from unsupported clades removed in order that they may be estimated without the constraint. Generally, we assumed that the taxonomy was correct unless demonstrated otherwise. After trees were constructed, we manually inspected the phylogenies and removed outlying taxa assumed to be misidentified (based on branch length or position). Rooting was performed based on information available on the most recent systematic studies (typically as referenced in the Angiosperm Phylogeny Website version 12–13 http://www.mobot.org/MOBOT/research/APweb/).

<h2>*Calculating support and data overlap*—Because we may have a series of constraints applied to each branch, we could not conduct traditional bootstrap analyses as implemented in RAxML. To ascertain the confidence in edges, we instead employed the quartet approaches described in this volume (Pease et al., 2018, in this issue). Briefly, these analyses consisted of using the alignment and the maximum likelihood tree to, on each edge, draw a random number of quartets of sequences that represent the quartets defined by the edge in the ML tree. Then the likelihoods for that edge and the two alternative resolutions were calculated, and the resolution that has the highest likelihood was recorded. This procedure was done 200 times for each edge in each subtree. We then summarized them with the Quartet Concordance (QC) measure, which calculates the ICA (Salichos et al., 2014) based on the distribution of quartets that support or conflict with the resolution found in the focal tree.

Data overlap was measured for each of the major clades and visualized on each tree. In this case, data overlap was defined as the number of sites that had overlapping data between sister clades. To calculate overlap, we proceeded through the phylogeny in a postorder fashion and calculated the total number of sites that contained at least 1 bp of overlap between each subtending sister clade. We calculated this as a site-wise measure because each gene region may contain sequences with poor overlap.

<h2>*Divergence-time estimation*—

As with the phylogenetic construction, we also conducted a hierarchical analysis for divergence-time estimation. We conducted divergence-time analysis using the penalized likelihood approach as implemented in treePL (Sanderson, 2002; Smith and O'Meara, 2012). To apply constraints,

we examined overlap between the Magallón et al. (2015) dated tree and each individual clade tree. For every clade in the individual trees that was monophyletic in the Magallón et al. (2015) tree, we applied a constraint with a fixed age of the node height, resulting in 590 constraints. We then conducted treePL analyses with a relatively high rate smoothing penalty value (logp = 10), given the size of the phylogenies.

<h2>*Large tree construction*—

We constructed four large phylogenies: GenBank taxa with a backbone provided by Open Tree of Life version 9.1 (GBOTB), GenBank taxa with a backbone provided by Magallón et al. (2015) (GBMB), GenBank and Open Tree of Life taxa with a backbone provided by Open Tree of Life version 9.1 (ALLOTB), and GenBank and Open Tree of Life taxa with a backbone provided by Magallón et al. (2015) (ALLMB). To examine detailed differences between the two backbones, please consult the Open Tree of Life website (https://tree.opentreeoflife.org); the differences are too great to detail here. Primarily, the Magallón et al. (2015) and Open Tree of Life backbones were similar but with the Open Tree of Life backbone providing more resolution toward the tips, that can be useful when there are no molecular data. For GBOTB and GBMB, we replaced each major clade with the phylogeny constructed, as described above. For ALLOTB and ALLMB, we replaced the clade representing each major clade (constructed as described above) with the constructed clade and then added taxa that were not sampled in the phylogeny but found in the original Open Tree of Life tree with the resolution retained. Many of the taxa added back will be unresolved. The taxonomic names found in the final tree consist of those found primarily in both the Open Tree of Life and NCBI taxonomies.

We aim to continue to improve the phylogenies constructed here. To that end, we provide updated versions of this tree at https://github.com/FePhyFoFum/big_seed_plant_trees with corresponding alignments and individual clade trees linked within. We also hope to have issues discussed and noted in the issue tracking system on this website in order to continue improving the resources.

<h2>*Diversification analyses*—We conducted diversification rate-shift analyses using MEDUSA (vers. 0.951 [Harmon et al., 2008; Pennell et al., 2015]) on the GBOTB phylogeny. We chose MEDUSA, primarily, because the size and scope of the phylogenies prevented

convergence of Bayesian methods in a timely manner. MEDUSA adds piecewise clade-specific diversification models to a time-calibrated tree in a manner that best explains (using AIC) the configuration of the tree. Because birth–death models generally require phylogenies to be fully bifurcating and to have non-zero branch lengths, we randomly resolved any polytomies and set minimum branch lengths to 0.1. We arrived at this number after exploring a range of smaller and larger minimum values because this value was the smallest minimum that did not increase spurious rate shifts. Because larger trees can suffer from statistically spurious rate shifts simply from the combinatorics involved, piecewise models were only added if they improved the AIC score by more than a threshold of 15.97 units (correction calculated by MEDUSA for a tree with 79,882 tips; see Pennell et al. [2014]). The analysis was terminated when no subsequent piecewise model improved the AIC beyond the threshold.

<h1>RESULTS AND DISCUSSION

<h2>*Sampling in the large phylogenies*—

There were 119,355 species, 13,328 genera, and 2477 higher taxa of seed plants recognized by GenBank as of release 218. Of these, 79,689 species had data that were sufficiently overlapping, based on the methods discussed here, to include in the analyses presented here. Of these, 812 were not present in the OT-taxonomy, most probably due to mismatch in the version of NCBI used for taxonomy merging in OT-taxonomy and that used for our GenBank analyses. The GBOTB contained 79,881 taxa (Figs. 2–4) and GBMB contained 79,874 (Appendix S2). ALLOTB (Fig. 5) contained 353,185 taxa and ALLMB contained 356,305 taxa (Appendix S3). The discrepancy in the number of taxa was a result of clades being lost to conflict between the trees constructed of major clades and the Open Tree of Life and Magallón et al. (2015) trees. The higher number of taxa in ALLMB was the result of fewer input trees in the synthetic analysis used to create the backbone tree with OT-taxonomy and therefore, fewer potential conflicts. For example, the Open Tree of Life (version 9.1) lacked a monophyletic Sapindales, and so those taxa are removed from the Open Tree of Life backbone. We then added a monophyletic Sapindales to the backbone based on data from GenBank but were unable to add any unplaced taxa back. Because the Sapindales are monophyletic in the Magallón et al. (2015) backbone, those unsampled taxa could be placed back. The fewer input trees for Magallón et al. (2015) results in fewer conflicts, more monophyly, and therefore, more taxa that were unsampled by

GenBank being represented in the final tree. The trade-off, however, was less resolution for those unsampled taxa than would be found in the ALLOTB.

Some of the conflict found in the Open Tree of Life synthesis tree had to be removed to successfully place the major clades. The Sapindales, discussed above, is an example of this problem, which is just one of several challenges that highlight the need for human intervention in these large analyses, also discussed by Beaulieu and O'Meara (2018) in this issue. Human intervention was also necessary in removing obvious outliers (based on branch lengths or taxonomic placement) and identifying gene regions for data set construction. Until data quality issues decrease and/or data availability increases dramatically, human intervention seems to be a necessary element to construct high-quality large data sets.

Data overlap, as measured by the overlap in sites between subtending nodes is presented on the GBOTB (Fig. 3). This analysis provided an edgewise view of the distribution of data while accommodating for the fact that even if the same gene was sampled between taxa, the sites may not overlap significantly. The distribution across edges (Fig. 3B) roughly approximated an exponential distribution with a minimum overlap of 0 bp, maximum overlap of 29,229 bp, mean of 2340 bp, and median of 1792 bp. The 0-bp overlap may reflect either a constraint that has no overlapping data (one reason to use a constraint) or a resolution with no supporting data (perhaps a random resolution between equal alternatives). A median value of 1792 bp suggests that many of the edges had some overlap in data, roughly corresponding to one or two gene regions. This result is not unexpected considering previously analyzed data sets of this magnitude that found similar results (Sanderson, 2008; Hinchliff and Smith, 2014). Both the relatively low overlap between sequences and the lack of data in GenBank for roughly 200,000 taxa highlight the need for additional sequencing of molecular data to resolve more confidently most of the phylogeny of seed plants. For, despite the size of the data set presented here, there is still little overlap between species, and there are still unsampled taxa. In the analyses presented here, we largely excluded smaller gene regions because of misidentification problems (see discussion below). If we were to include those smaller gene regions (e.g., ITS), we would expect the median number of sites overlapping at each edge to decrease.

Support was measured as Quartet Concordance (Pease et al., 2018, in this issue) and plotted on the GBOTB (Fig. 4). The distribution of support (Fig. 4B) is relatively flat with spikes at −1, 0, and 1 and with a median of 0.29 and mean of 0.285. The spikes at −1, 0, and 1 reflect

significant values for the QC measure: −1 reflects complete support for an alternative, 0 reflects no support for any resolution, and 1 reflects complete support for the resolution in the ML tree presented. This analysis, along with the data overlap, highlights the relative low support throughout the tree because the median support value was only slightly higher than "no support". This finding may be the result of poor data overlap, underlying conflict due to incomplete lineage sorting or other processes, or true biological uncertainty (e.g., saturation, lack of informative substitutions, etc.). Nevertheless, more detailed analyses and additional data would likely shed more light on the details of uncertainty at specific nodes.

<h2>*Diversification results*—

The primary goal of this study was to explore a way of constructing dated phylogenies using molecular data along with resources available through the Open Tree of Life project. However, to demonstrate one way to use the resulting phylogenies, we conducted diversification analyses. Our diversification analyses found 472 distinct diversification models (471 rate shifts) that best describe the seed plant phylogeny. While every clade experienced some change in the rate of diversification, the most extensive, in terms of number and rate, were found in the Asterales (Fig. 2). Ranunculales, Gentianales, and Caryophyllales all also experienced multiple large shifts. Furthermore, many of the shifts were nested within other diversification shifts. The observation of nested diversification and lag times between major clades and diversification shifts have been noted by other authors (Donoghue, 2005; Smith et al., 2011; Donoghue and Sanderson, 2015; Tank et al. 2015). A notable pattern highlighted by the results here is that although shifts were associated with the origin of angiosperms and mesangiosperms, few large shifts occurred at the rootward internal nodes of the tree. However, we do not wish to over-interpret this result considering the uncertainty associated with large phylogenetic trees, discussed below. This demonstration serves as an example of the potential utility of these trees.

There were challenges in using this phylogenetic tree for diversification analyses that are worth examining. For example, diversification analyses were sensitive to the minimum branch lengths chosen. The penalized likelihood dating procedures can result in zero or near-zero branch lengths where there is conflict or little information from the molecular data (i.e., zero or near-zero molecular branch lengths in the chronogram). Usually, branches with very small lengths will be collapsed. However, MEDUSA analyses require bifurcating trees with non-zero branch

lengths at each edge. We set the minimum divergence time branch lengths to be 0.1 but found that different values resulted in different diversification inferences. Often, the result of smaller minimums was an increase in the estimated diversification shifts around the zero branch length edges. As a result, and to be conservative regarding our estimates, we favored the large minimum branch length value. Nevertheless, we regard these results as coarse approximations that may be refined in the future with more nuanced divergence-time estimation results and integration over the uncertainty in the phylogenetic estimation of topology and branch lengths.

In addition to branch length considerations, there are persistent concerns regarding taxon sampling and diversification analyses. Here, we used the phylogeny without the additional taxa from the Open Tree of Life because the lack of resolution in those additional taxa would require some form of either random or birth–death resolution (Kuhn et al., 2011). While the placement of taxa based on birth–death models may be useful, the large number of unplaced taxa in the ALLOTB tree led us to use the smaller GBOTB tree. Ideally, the placement of these taxa should be informed by molecular data (Rabosky, 2015). In addition to these issues, incomplete or biased taxon sampling will also influence the results. This problem, however, is not specific to large trees as it impacts all diversification studies, and very few clades of large size have been sampled completely.

The diversification analyses presented here demonstrate one way these phylogenies may be used. We highlight the potential pitfalls and caveats with these data; however, most of these apply to smaller data sets as well. In the case of large or small data sets, uncertainties and assumptions—i.e., unsupported relationships, incomplete sampling, and/or taxonomic misidentification—need to be understood when interpreting the results of diversification and other evolutionary comparative analyses.

<h2>*Comparison to other techniques*—

Over the last few years, there have been several methods developed for utilizing the publicly available data stored in GenBank. The method presented here is similar in some ways but differs in other important ways. We do not present an exhaustive comparison but instead provide a brief discussion of a few alternatives. PyPHLAWD differs from PHLAWD in that PyPHLAWD is more flexible. PyPHLAWD, unlike PHLAWD, was developed as a series of different scripts, any one of which can be modified. PyPHLAWD also does not require the user to provide

sequences a priori because clustering is part of the analysis. PyPHLAWD differs from Phylota (Sanderson et al., 2008) in that PyPHLAWD only conducts a major clustering analysis tipward, with BLAST being used to combine clusters deeper in the tree. This procedure allows for large clusters to be constructed as BLAST can be a limiting factor given that computational requirements increase dramatically as the number of taxa increases. As a result, Phylota does not report larger clusters toward the root of the tree of life. SUPERSMART takes a list of taxa or clade names and constructs dated phylogenies (Antonelli et al., 2017). SUPERSMART is, perhaps, the most similar to that which we present here with some exceptions. First, we did not construct a backbone as part of the analysis presented here and instead used an existing backbone (from the Open Tree of Life). Many of the major lineages have required genomic or transcriptomic data to resolve major clades (e.g., Wickett et al., 2014), and analyses of those data types often require different methods than those conducted in any of the aforementioned packages (e.g., Yang and Smith, 2014). Future developments could incorporate methods typically used for genomic data into PyPHLAWD to facilitate the construction of deeper edges in the tree of life. Another difference between PyPHLAWD and SUPERSMART is that we do not conduct Bayesian analyses for divergence times or phylogenetic reconstruction because of the size of the data presented precludes that possibility. This approach could be implemented within PyPHLAWD but is not currently. The clustering analyses, merging, and other aspects of the sequence analyses also differ. However, similarities between the methods suggest that both could produce similar results, with slight differences in the means of calculating similarity likely to inject some variation. This comparison should be explored further as both packages continue to develop. Finally, we integrated the information from the Open Tree of Life back into the phylogenies. Integration with the Open Tree of Life is not a goal of SUPERSMART, and so unsurprisingly that is not part of the analysis.

There are now a variety of programs that can process data from public databases to produce clusters, alignments, and trees with and without divergence times. We do not suggest that PyPHLAWD is the single, best solution for constructing molecular phylogenies using GenBank. Instead, we feel as though PyPHLAWD is another option among a set of good alternatives and hope that the flexibility of the software will allow for continued updating and extension. As new sequences are added to GenBank and other resources, and as the Open Tree of Life continues to be updated, the alignments and trees generated here can be refined.

Furthermore, we present these phylogenies through a framework (https://github.com/FePhyFoFum/big_seed_plant_trees) that allows for reporting and tracking of issues and improvements like that used for software development. We are hopeful that this framework will facilitate the enhancement of these resources because the community can communicate any problems directly. The flexibility of automated but not fully automated procedures also facilitates the ability to intervene, for example, to remove outliers, adjust gene sets, and monitor overall quality. The need for human intervention has also been highlighted by Beaulieu and O'Meara (2018, in this issue). Additionally, we hope that the connection to the Open Tree of Life will enhance those resources and those comparative analyses that benefit from more complete sampling.

<h2>*Limitations of these data sets and analyses*—The data sets and phylogenetic trees presented here, while they have many benefits, are not without limitations. Challenges are associated with all large phylogenetic data sets that must be considered and that relate to uncertainty and lack of information discussed above and by others (Hinchliff and Smith, 2014; Edwards et al., 2015). There are also issues specific to the data set and analyses presented here. One fundamental limitation specific to this data set is that many taxa in both the ALLOTB and ALLMB do not have molecular data associated and are placed based on taxonomy. Most species of seed plants have no molecular data currently in GenBank, and those that do may not have significant overlap with other sequences. While many sequencing projects focus on collecting more gene regions, there is still a great need for more species that have no data to be collected and sequenced. In the meantime, researchers may choose to conduct Bayesian analyses using a birth–death before randomly resolving polytomies (Kuhn et al., 2011), though ideally the placement of all taxa should be informed by molecular data (Rabosky, 2015).

In addition to the fact that most species only have taxonomic data, there is also significant uncertainty in the placement of many taxa that do have molecular data. Here, we measure uncertainty using the Quartet Concordance (QC) measure (Pease et al. 2018, in this issue). While this measure allowed us to record how often the concordant quartet was inferred over the two alternative quartets, the method does not generate alternative resolutions. The individual data sets are available such that users could generate a set of trees from any other set of analyses (e.g., Bayesian analyses, bootstrap analyses). Nevertheless, the QC and data overlap analyses

presented here demonstrate that uncertainty is still a concern in these large phylogenies. While data sets with a large number of taxa may present specific problems, recent transcriptomic and genomic analyses have shown that increasing the number of genes will also expose the underlying complexity of conflict inherent to genomic evolution (Salichos et al., 2014; Smith et al., 2015; Brown and Thomson, 2016; Shen et al., 2017). Finding better ways of incorporating this uncertainty in these large trees instead of hoping to resolve it, or relying on a single resolution, may prove more beneficial as we move forward. Finally, while some comparative analyses may be robust to alternative placements of taxa, uncertainty should be considered by both researchers making use of these data sets and researchers developing comparative methods.

Divergence time estimation is also a major challenge for any phylogenetic study, and the challenge only increases with data set size. Large data sets present a computational burden where constrained optimization algorithms become stuck in local optima, a problem that is exacerbated as data set size grows and heterogeneity increases (Smith and O'Meara, 2012). Many of the data sets analyzed in this study are some of the largest analyzed and so likely suffer from this problem. In addition to this problem inherent to optimization, there are well-known problems of rate heterogeneity that can significantly increase estimation error (Smith and Donoghue, 2008; Beaulieu et al., 2015). For those researchers that wish to incorporate uncertainty, data sets are made available, but how best to generate a set of trees using penalized likelihood that represent a credible interval and how best to accommodate the extensive rate heterogeneity should both be explored further.

Finally, these analyses demonstrate a means for combining molecular data with the Open Tree of Life into a "comprehensive" phylogeny. However, these trees are comprehensive only in that they include the sampled and unsampled taxa represented in the taxonomies of the Open Tree of Life. Many clades may still require significant taxonomic work and smaller, species-level, examination before there can be confidence about species composition. So, while these phylogenies contain all the taxa from the Open Tree of Life, revisions based on smaller scale studies will continue to improve these data and analyses.

<h2>*A remark on short sequences*—

We excluded most small gene regions from the data set construction in these analyses, especially in more complex and speciose clades. In some cases, this means that gene regions that have been

sampled for many taxa were excluded. Primarily, the removal was done to avoid the inclusion of misidentified and problematic sequences, many of which were collected as part of barcoding projects. DNA barcoding aims to collect a small number of specific gene regions for many species to help with identification and to address other specific questions. These efforts, though each may have different goals, result in the submission of many sequences to GenBank that other researchers can download and use in, among other things, phylogenetic analyses. One major goal for biologists is to increase the completeness of phylogenetic trees. Barcode data, which can increase sampling in undersampled geographic regions, would be a desirable resource if they could be incorporated into phylogenetic analyses.

Despite these benefits, our analyses have found many of these sequences to be a hindrance, resulting in our attempt to exclude most short sequences. While most short sequences may not suffer from any of following problems, we found that many suffered from several issues that hindered accurate reconstruction. First, many short sequences contained little to no phylogenetic information (e.g., few informative sites), which may be the result of the small size of the gene regions used, slow molecular evolution of the gene, or slow molecular evolution of the lineage. When lineages have little to no phylogenetic information, single maximum likelihood analyses can be misleading as there are many nearly equally probable placements of a specific taxon. Bayesian methods and likelihood methods that integrate over topological uncertainty can correctly report the uncertainty in the placement of uninformative sequences. However, Bayesian methods are intractable for data sets of the sizes presented here. There are ways to better incorporate phylogenetic uncertainty in maximum likelihood analyses, but the computational burden for these large data sets is quite high. Furthermore, while sequence similarity can be useful for taxonomic identification, having taxa integrate across the familial or ordinal level because of lack of phylogenetic information is not particularly useful for phylogenetic analyses. So, while we may be able to accommodate for the uncertainty in the placement of these taxa, it is unclear whether the increase in complexity and runtime is worth the inclusion of such sequences. Of course, not all barcode or small sequences have this problem, but the ability to identify which do have the problem will improve these large phylogenetic efforts enormously.

The second major problem that we found, misidentification, is more difficult to address and, without correction, negates our ability to accurately estimate phylogenies. In our analyses,

we found that many of the sequences that violated the constraints were misidentified. For example, in an analysis of the available data for Laurales, we found several sequences such as *Litsea collina* and *Alseodaphne andersonii* that are more probably *Endiandria*, *Beilschmiedia*, *Cryptocarya*, or *Neolitsea* (Appendix S4). While these genera may not in fact be monophyletic, the samples seem to fall far from their labeled taxonomic placement. Either taxonomic revision may be necessary or these sequences were misidentified. Even when there are multiple loci that represent a taxon, if one sequence is egregiously misidentified, that sequence can drive the incorrect placement of the taxon. The problem of misidentification was so egregious that we filtered out most short sequences to eliminate misidentification when possible. There may also be problems with misidentification of larger sequences, but our analyses found that the exclusion of most short sequences dramatically reduced this difficulty. While methods for correctly identifying sequences in GenBank are beyond the scope of this paper, the use of constraint trees aided in our ability to isolate misidentified sequences, and future research will expand these efforts.

Both problems highlight the need to address how we should proceed with short and misidentified sequences for large phylogenetic analyses. There are thousands of useful sequences that do not suffer from these issues that we have excluded. However, when conducting large analyses, small percentages of bad data can dramatically inhibit accurate phylogenetic estimates. With additional developments, we hope that bad sequences may be filtered so that the many good short sequences can be included. We have included with PyPHLAWD lists of sequences or taxa that may be problematic to better incorporate shorter sequences. We have also begun documenting problematic sequences in a more general resource that all software and researchers can utilize (https://github.com/FePhyFoFum/seq_filters). Uncertainty can be accommodated, but perhaps for short sequences, more constrained searches as implemented in software packages meant for barcode identification would be more appropriate (Matsen et al., 2010; Berger et al., 2011). Misidentification may be more difficult to handle, but resources that allow for the identification and correction of these sequences that could be utilized by the multitude of software packages would be preferable.

<h2>*Constraints and large phylogenies*—

One of the fundsamental challenges to constructing phylogenies are edges deep in the tree, a problem exacerbated by complex patterns of conflict and lack of information (e.g., Salichos et al., 2014; Smith et al., 2015; Brown and Thomson, 2016; Shen et al., 2017). The computational challenge of constructing phylogenetic trees scales exponentially with the addition of taxa, and so reconstructing deeper and deeper edges increases the computational burden and complexity significantly. This challenge leads to the question of whether we need to always reconstruct these deeper edges when constructing large phylogenies. In other words, can we build on the knowledge that we have accrued from other analyses? Previous studies may have successfully analyzed rootward nodes that identify major clades using genomic and/or transcriptomic data with sophisticated analyses that can incorporate more complex evolutionary models (Wickett et al., 2014; Yang et al., 2015; Givnish et al., 2015; McKain et al., 2016; Comer et al., 2016; Walker et al., 2017). Data sets with large numbers of taxa may not be able to take advantage of these large and complex data sets or the more complex evolutionary models simply due to the scale of the data set. More importantly, the large phylogenetic data sets examined here, consisting of thousands of species, are generally not adequate for reconstructing or testing hypotheses regarding deep and complex evolutionary relationships due to poor data overlap and computational complexity. The approach we take here is to leave the resolution of the deeper edges of the tree to other analyses (summarized in the Open Tree of Life) and instead focus our analyses on the finer details nested within each major clade. For example, we assume that the gymnosperms form a monophyletic group that is sister to the angiosperms. This assumption, while still discussed (Donoghue and Doyle, 2000), is not controversial and removing the assumption would not only increase the runtime significantly, but the data we use to reconstruct the tips may not be the optimal data to reconstruct the deeper edges. By making less controversial assumptions, we not only reduce runtime but focus our reconstruction efforts to more uncertain parts of the tree.

One limitation of the analyses presented here, however, is that we relied on a single resolution for these constraints. Recent genomic and transcriptomic analyses have highlighted extensive conflict across the tree of life and uncertainty associated with more than simply lack of phylogenetic information (Salichos et al., 2014; Smith et al., 2015; Brown and Thomson, 2016; Shen et al., 2017; Walker et al. 2017). Future work should explore means of incorporating the uncertainty found in these more extensive genomic data sets into the constraints used by larger

phylogenies to better reflect our knowledge of the complexity within these parts of the tree of life. There may not be one resolution for a part of the tree of seed plants, and we should look to incorporating that into our analyses of these large data sets.

In addition to the deeper edges in the tree, we extended this approach by constructing phylogenies with taxonomy-based constraints throughout, removing them when they were unsupported. This approach may be controversial given the knowledge and quality of underlying taxonomy and phylogeny. For example, for Fungi, due to complex taxonomic resources and data availability, applying more tipward constraints may be harder than for plants and mammals. Furthermore, there may be genera with questionable monophyly. However, as with the deeper nodes, we argue that these large phylogenies consisting of thousands of taxa are not ideal data sets with which to *test* these hypotheses. Instead, focused studies aimed at phylogenetic reconstruction of a particular group are the best place for taxonomic revision. These can then be incorporated into the large phylogenetic analyses where data set coverage and taxonomic sampling may not be adequate to test those taxonomic hypotheses. Nevertheless, as with the deeper nodes, these large analyses should be designed to accommodate the inherent conflict underlying the tree of life where possible. In addition to the computational benefits, as mentioned above, constraints can also be helpful when attempting to identify misidentified taxa and clades that have relatively little molecular information required for successfully resolving the clade. We suggest further work should be done to examine whether this approach would be helpful, generally, in reducing runtimes, in identifying misidentified taxa, and where researchers have reduced the resolution of major clades to a small number of well-known alternative resolutions.

<h2>*Where do we go from here?—*

Here we present a set of large phylogenies for seed plants. We do not intend this result to be *the* definitive view of seed plant evolution. Instead, we hope that these efforts underscore the challenges of these projects. The exercise has highlighted issues concerning uncertainty, data overlap, and data availability that suggest the need to continue to improve methods and generate new data. This exercise has also underscored the need for human intervention in the process that has been highlighted by others (Hinchliff and Smith, 2014), including most recently by authors in this issue (Beaulieu and O'Meara, 2018). While many of the analyses can be automated, because of the complexity of the data, problems with misidentification, and other data quality

issues, no steps can or should be fully automated. Nevertheless, we need places from which to start to measure progress and build toward the goal of an accurate and resolved seed plant phylogeny. Despite the challenges, the trees presented here will hopefully serve as resources that will continue to be updated as new data become available and as the Open Tree of Life resources are updated.

<h1>ACKNOWLEDGEMENTS

We thank Hervé Sauquet and Jeremy Beaulieu for helpful comments and reviews of the manuscript and Karen Cranston, Mark Holder, Benjamin Redelings, and Jonathan Rees and other folks from the Open Tree of Life project for helpful comments. We also thank Simon Uribe-Convers, Greg Stull, Oscar Vargas, Ning Wang, and Joseph Walker for helpful comments on the manuscript.

<h1>FUNDING

This work was supported by the National Science Foundation DEB grant 1207915 and ABI grant 1458466.

<h1>AUTHOR CONTRIBUTIONS

S.A.S and J.W.B. conducted all analyses and wrote the manuscript.

<h1>LITERATURE CITED

Antonelli, A., H. Hettling, F. L. Condamine, K. Vos, R. H. Nilsson, M. J. Sanderson, H. Sauquet, et al. 2017. Toward a self-updating platform for estimating rates of speciation and migration, ages, and relationships of taxa. *Systematic Biology* 66: 152–166.

Beaulieu, J. M., R. H. Ree, J. Cavender-Bares, G. D. Weiblen, and M. J. Donoghue. 2012. Synthesizing phylogenetic knowledge for ecological research. *Ecology* 93: S4–S13.

Beaulieu, J. M., B. C. O'Meara, P. Crane, and M. J. Donoghue. 2015. Heterogeneous rates of molecular evolution and diversification could explain the Triassic age estimate for angiosperms. *Systematic Biology* 64: 869–878.

Berger, S. A., D. Krompass, and A. Stamatakis. 2011. Performance, accuracy, and Web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic Biology* 60: 291–302.

Bocak, L., C. Barton, A. Crampton-Platt, D. Chesters, D. Ahrens, and A. P. Vogler. 2013. Building the Coleoptera tree-of-life for >8000 species: composition of public DNA data and fit with Linnaean classification. *Systematic Entomology* 39: 97–110.

Brown, J. M., and R. C. Thomson. 2016. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Systematic Biology* 66: 517–530.

Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les, B. D. Mishler, M. R. Duvall, et al. 1993. Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbcL*. *Annals of the Missouri Botanical Garden* 80: 528–580.

Chesters, D. 2017. Construction of a Species-level tree of life for the insects and utility in taxonomic profiling. *Systematic Biology* 66: 426–439.

Comer, J. R., W. B. Zomlefer, C. F. Barrett, D. W. Stevenson, K. Heyduk, and J. H. Leebens-Mack. 2016. Nuclear phylogenomics of the palm subfamily Arecoideae (Arecaceae). *Molecular Phylogenetics and Evolution* 97: 32–42

Cornwell, W. K., M. Westoby, D. S. Falster, R. G. FitzJohn, B. C. O'Meara, M. W. Pennell, D. J. McGlinn, et al. 2014. Functional distinctiveness of major plant lineages. *Journal of Ecology* 102: 345–356.

Dongen, S. M. V. 2000. Graph clustering by flow simulation. PhD dissertation, University of Utrecht, Utrecht, Netherlands.

Donoghue, M. J. 2005. Key innovations, convergence, and success: macroevolutionary lessons from plant phylogeny. *Paleobiology* 31: 77–93.

Donoghue, M. J., and J. A. Doyle. 2000. Seed plant phylogeny: Demise of the anthophyte hypothesis? *Current Biology* 10: R106–R109.

Donoghue, M. J., and M. J. Sanderson. 2015. Confluence, synnovation, and depauperons in plant diversification. *New Phytologist* 207: 260–274.

Driskell, A. C., C. Ané, 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.

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

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: 2532–2537.

Edwards, E. J., C. P. Osborne, C. A. E. Strömberg, S. A. Smith, C_4 Grasses Consortium, W. J. Bond, P.- A. Christin, et al. 2010. The origins of C_4 grasslands: integrating evolutionary and ecosystem science. *Science* 328: 587–591.

Edwards, E. J., J. M. de Vos, and M. J. Donoghue. 2015. Doubtful pathways to cold tolerance in plants. *Nature* 521: E5–E6.

Freyman, W. A. 2015. SUMAC: Constructing phylogenetic supermatrices and assessing partially decisive taxon coverage. *Evolutionary Bioinformatics Online* 11: 263–266.

Givnish, T. J., D. Spalink, M. Ames, S. P. Lyon, S. J. Hunter, A. Zuluaga, W. J. D. Iles, et al. 2015. Orchid phylogenomics and multiple drivers of their extraordinary diversification. *Proceedings of the Royal Society, B, Biological Sciences* 282: 20151553.

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

Goloboff, P. A., S. A. Catalano, J. Marcos Mirande, C. A. Szumik, J. Salvador Arias, M. Källersjö, and J. S. Farris. 2009. Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups. *Cladistics* 25: 211–230.

Harmon, L. J., J. T. Weir, C. D. Brock, R. E. Glor, and W. Challenger. 2008. GEIGER:

investigating evolutionary radiations. *Bioinformatics* 24: 129–131.

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

Hibbett, D. S., R. H. Nilsson, M. Snyder, M. Fonseca, J. Costanzo, and M. Shonfeld. 2005. Automated phylogenetic taxonomy: an example in the homobasidiomycetes (mushroom-forming fungi). *Systematic Biology* 54: 660–668.

Hinchliff, C. E., and S. A. Smith. 2014. Some limitations of public sequence data for phylogenetic inference (in plants). *PloS One* 9: e98986.

Hinchliff, C. E., S. A. Smith, J. F. Allman, J. G. Burleigh, R. Chaudhary, L. M. Coghill, K. A. Crandall, et al. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences, USA* 112: 12764–12769.

Izquierdo-Carrasco, F., J. Cazes, S. A. Smith, and A. Stamatakis. 2014. PUmPER: phylogenies updated perpetually. *Bioinformatics* 30: 1476–1477.

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

Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780.

Ksepka, D. T., J. F. Parham, J. F. Allman, M. J. Benton, M. T. Carrano, K. A. Cranston, P. C. J. Donoghue, et al. 2015. The fossil calibration database—a new resource for divergence dating. *Systematic Biology* 64: 853–859.

Kuhn, T. S., A. Ø. Mooers, and G. H. Thomas. 2011. A simple polytomy resolver for dated phylogenies. *Methods in Ecology and Evolution* 2: 427–436.

Magallón, S., S. Gómez-Acevedo, L. L. Sánchez-Reyes, and T. Hernández-Hernández. 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytologist* 207: 437–453.

Matsen, F. A., R. B. Kodner, and E. V. Armbrust. 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 11: 538.

McKain, M. R., H. Tang, J. R. McNeal, S. Ayyampalayam, J. I. Davis, C. W. dePamphilis, T. J. Givnish, et al. 2016. A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. *Genome Biology and Evolution* 8: 1150–1164.

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

McTavish, E. J., C. E. Hinchliff, J. F. Allman, J. W. Brown, K. A. Cranston, M. T. Holder, J. A. Rees, and S. A. Smith. 2015. Phylesystem: a git-based data store for community-curated phylogenetic estimates. *Bioinformatics* 31: 2794–2800.

O'Meara, B. C., S. D. Smith, W. S. Armbruster, L. D. Harder, C. R. Hardy, L. C. Hileman, L. Hufford, et al. 2016. Non-equilibrium dynamics and floral trait interactions shape extant angiosperm diversity. *Proceedings of the Royal Society, B, Biological Sciences* 283: 20152304.

Pennell, M. W., J. M. Eastman, G. J. Slater, J. W. Brown, J. C. Uyeda, R. G. FitzJohn, M. E. Alfaro, and L. J. Harmon. 2014. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics* 30: 2216–2218.

Rabosky, D. L. 2015. No substitute for real data: A cautionary note on the use of phylogenies from birth–death polytomy resolvers for downstream comparative analyses. *Evolution* 69: 3207–3216.

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

Redelings, B. D., and M. T. Holder. 2017. A supertree pipeline for summarizing phylogenetic and taxonomic information for millions of species. *PeerJ* 5: e3058.

Rees, J., and K. Cranston. 2017. Automated assembly of a reference taxonomy for phylogenetic

data synthesis. *Biodiversity Data Journal* 5: e12581.

Salichos, L., A. Stamatakis, and A. Rokas. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Molecular Biology and Evolution* 31: 1261–1271.

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

Sanderson, M. J. 2008. Phylogenetic signal in the eukaryotic tree of life. *Science* 321: 121–123.

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.

Sauquet, H., M. von Balthazar, S. Magallón, J. A. Doyle, P. K. Endress, E. J. Bailes, E. Barroso de Morais, et al. 2017. The ancestral flower of angiosperms and its early diversification. *Nature Communications* 8: 16047.

Shen, X.- X., C. T. Hittinger, and A. Rokas. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution* 1: 126.

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

Smith, S. A., J. W. Brown, and C. E. Hinchliff. 2013. Analyzing and synthesizing phylogenies using tree alignment graphs. *PLoS computational Biology* 9: e1003223.

Smith, S. A., J. W. Brown, Y. Yang, R. Bruenn, C. P. Drummond, S. F. Brockington, J. F. Walker, et al. 2017. Disparity, diversity, and duplications in the Caryophyllales. *New Phytologist* 217: 836–854.

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

Smith, S. A., M. J. Moore, J. W. Brown, and Y. Yang. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* 15: 150.

Smith, S. A., and B. C. O'Meara. 2012. treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* 28: 2689–2690.

Soltis, D. E., S. A. Smith, N. Cellinese, K. J. Wurdack, D. C. Tank, S. F. Brockington, N. F. Refulio-Rodriguez, et al. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany* 98: 704–730.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Tank, D. C., J. M. Eastman, M. W. Pennell, P. S. Soltis, D. E. Soltis, C. E. Hinchliff, J. W. Brown, et al. 2015. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytologist* 207: 454–467.

Walker, J. F., Y. Yang, M. J. Moore, J. Mikenas, A. Timoneda, S. F. Brockington, and S. A. Smith. 2017. Widespread paleopolyploidy, gene tree conflict, and recalcitrant relationships among the carnivorous Caryophyllales. *American Journal of Botany* 104: 858–867.

Wickett, N. J., S. Mirarab, N. Nguyen, T. Warnow, E. Carpenter, N. Matasci, S. Ayyampalayam, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences, USA* 111: E4859–E4868.

Yang, Y., and S. A. Smith. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for

phylogenomics. *Molecular Biology and Evolution* 31: 3081–3092.

Yang, Y., M. J. Moore, S. F. Brockington, D. E. Soltis, G. K.- S. Wong, E. J. Carpenter, Y. Zhang, et al. 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. *Molecular Biology and Evolution* 32: 2001–2014.

Zanne, A. E., D. C. Tank, W. K. Cornwell, J. M. Eastman, S. A. Smith, R. G. FitzJohn, D. J. McGlinn, et al. 2014. Three keys to the radiation of angiosperms into freezing environments. *Nature* 506: 89–92.

FIGURE 1. Workflow for analyses in constructing the trees. For each clade listed in Appendix S1, here Apiales, we clustered sequences starting at the most tipward clade (*Sanicula* highlighted here), and merged clusters as we moved rootward Apiales. We used the clusters at the most rootward clade (Apiales in this example) to construct a supermatrix where we chose clusters that had good representation for the rootward clade or good representation for a subtending clade (as shown by the gene region on the far right). Using this supermatrix, we constructed a phylogeny and estimated divergence times. We then placed this constructed phylogeny into the backbone (either Open Tree of Life or Magallón et al. [2015]) as-is or with the unsampled taxa from the OT-taxonomy placed back.

FIGURE 2. (A) GBOTB tree with colors corresponding to binned rates of diversification (see text for details on analyses). Rates were binned to make for easier visualization of rates. Red dots denote nodes with a shift in diversification rate and the size of the dot corresponds to the magnitude of the shift. (B) Lineage through time plot of the GBOTB tree. Divergence times are denoted with concentric circles.

FIGURE 3. Data overlap presented on the GBOTB tree (A) and (B) as a histogram summarizing across all data. Values on the tree are displayed as $\log_{10}$-transformed to allow for easier visual discrimination.

FIGURE 4. Support as measured by Quartet Concordance scores presented on the GBOTB tree (A) and (B) as a histogram summarizing across all data.

FIGURE 5. The ALLOTB tree with some major clades labeled and 353,185 tips. Divergence times are denoted with concentric circles.