Article Type: Special Issue Article

RESEARCH ARTICLE

INVITED SPECIAL ARTICLE

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

Short Title: Running title: Wrestling with the rosids

Challenges of comprehensive taxon sampling in comparative biology: Wrestling with rosids

Ryan A. Folk[1,5], Miao Sun[1], Pamela S. Soltis[1,3], Stephen A. Smith[2], Douglas E. Soltis[1,3,4], and Robert P. Guralnick[1]

Manuscript received 16 August 2017; revision accepted 19 December 2017.

[1] Florida Museum of Natural History, Gainesville, FL 32611 USA

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

[3] Genetics Institute, University of Florida, Gainesville, FL 32610 USA

[4] Department of Biology, University of Florida, Gainesville, FL 32611 USA

[5] Author for correspondence (e-mail: ryanfolk@ufl.edu)

Citation: Folk, R. A., M. Sun, P. S. Soltis, S. A. Smith, D. E. Soltis, and R. P. Guralnick. 2018. Challenges of comprehensive taxon sampling in comparative biology: Wrestling with rosids. *American Journal of Botany* 105(3): XXX.

DOI: XXXX

Abstract

Using phylogenetic approaches to test hypotheses on a large scale, in terms of both species sampling and associated species traits and occurrence data—and doing this with rigor despite all the attendant challenges—is critical for addressing many broad questions in evolution and ecology. However, application of such approaches to empirical systems is hampered by a lingering series of theoretical and practical bottlenecks. The community is still wrestling with the challenges of how to develop species-level, comprehensively sampled phylogenies and associated geographic and phenotypic resources that enable global-scale analyses. We illustrate difficulties and opportunities using the rosids as a case study, arguing that assembly of biodiversity data that is scale-appropriate—and therefore comprehensive and global in scope—is required to test global-scale hypotheses. Synthesizing comprehensive biodiversity data sets in clades such as the rosids will be key to understanding the origin and present-day evolutionary and ecological dynamics of the angiosperms.

KEY WORDS: comparative methods; data layers; phylogeny; *Rosidae*; rosids; scientific infrastructure

Although systematists have established a robust phylogenetic framework for angiosperms, the march to the tips has proceeded at a considerably slower pace. Uncovering the basic framework of the angiosperm branch of the tree of life was a challenging, decades-long process (Chase et al., 1993; Soltis et al., 1997, 1999, 2000, 2011; Qiu et al., 1999; Ruhfel et al., 2014; Wickett et al., 2014), but, due to the sheer size of the angiosperm clade (~220,000–400,000 spp., reviewed by Scotland and Wortley, 2003), two even greater challenges will be (1) producing a comprehensive understanding of species-level relationships across flowering plants and (2) pairing this tree with phenotypic traits and geographic data. Even the best-sampled angiosperm clades have species-level coverage only slightly better than 30% (e.g., Saxifragales, 2400 species; Soltis et al., 2013; de Casas et al., 2016), with sampling typically resulting from piecemeal, focused, and typically small-scale case studies.

While studies of small exemplar clades are important for many questions in comparative biology, they neither intend to test, nor are capable of testing, the broadest evolutionary questions across flowering plants as a whole. While still uncommon, recent

investigations have used thousands or tens of thousands of taxa (e.g., Jetz et al., 2012a; Zanne et al., 2014; Werner et al., 2014; Faurby and Svenning, 2015). Tetrapods may provide the best example of progress on dense taxon sampling. Initial trees for this comparatively small clade (~35,000 species; unpublished data from the VertLife project; http://vertlife.org/), with dense coverage of extant species based on standard phylogenetic markers, are nearing completion. Comprehensive synthetic trees based on backbone phylogenies, as well as some deeply sampled supermatrices, have existed for several years in birds and other tetrapod groups (Jetz, et al., 2012a; see also citations in Title and Rabosky, 2017).

The use of comprehensive taxon sampling—up to and including complete coverage—is central to future progress in answering key questions in evolution and ecology framed at broad scales. Despite promise, progress in building comprehensive, broad-scale phylogenies and their associated data layers (i.e., biologically relevant taxon-level data linked to tips in a phylogenetic tree, such as phenotypic traits and occurrence records) for testing hypotheses has been limited by diverse challenges, such as incomplete phylogenetic coverage, lack of associated and accessible data layers, and a lack of available infrastructure to disseminate phenotypic and geographic data in ways that facilitate integration with phylogenetic information.

Collating such large-scale data sets is not trivial; thus, a set of factors converges to render macroevolutionary studies on vast scales as increasingly tractable, yet tantalizingly out of reach for many researchers. The fact that so many global-scale analyses (e.g., Jetz et al., 2012a) have focused on the rich data available for vertebrates (e.g., VertNet, http://vertnet.org/; FishBase, http://www.fishbase.org/; AmphibiaWeb, http://www.amphibiaweb.org) demonstrates how building linked biodiversity community resources spurs transformative research (for example, enabling assessment of drivers of diversification that may include phenotypic traits, geographic range, and ecological niche occupancy, among other candidates). Extending the technical and social approaches for developing such resources to other clades would lower barriers to performing macroscale comparative analyses in other groups. While the overall state of knowledge in the angiosperms generally lags well behind similar efforts in other groups (e.g., vertebrates, a more tractable target at perhaps one tenth the diversity of flowering plants), there are

angiosperm subclades well suited for realizing the vision of comprehensively assembling the large-scale picture of evolution of terrestrial ecosystems. What are the ingredients for lowering this barrier in flowering plants?

Here we provide an example of a subclade within the angiosperms that exemplifies the value of broad-scale approaches, the rosids (*Rosidae* sensu Cantino et al. 2007). Rosids are a major angiosperm clade, with ~90,000 species (Sun et al., 2016; M. Sun et al., unpublished) representing 22% of all angiosperms (assuming 400,000 species of angiosperms)—with properties that make this clade ideal for realizing the vision of global-scale hypothesis testing through a synthesis of biodiversity data.

In this paper, we ask: What are the grand challenge questions that could be addressed if a robust comparative framework—a well-resolved phylogeny linked with phenotypic and geographic data—were developed? This contribution is organized as a series of questions:

1. Why rosids? What is the case for building an exemplary comparative data set for this or any other large clade of life?
2. What challenges persist in building large-scale trees and trait layers despite progress to date, and how can these challenges be addressed?
3. Why use comprehensive approaches to analyze large clades of life? What motivations underlie large-scale analyses in ecology and evolution?

<H1>ROSIDS: AN EXEMPLAR CLADE FOR THE ANGIOSPERMS

Rosids, which capture many of the evolutionary and ecological dynamics of angiosperms as a whole, are ideal as a case study for demonstrating data-driven arguments behind building comparative resources in the flowering plants. Rosids exhibit substantial diversity in morphology, habit, reproductive strategy, and life history, and hence occupy a substantial portion of the phenotypic and ecological space that characterizes angiosperms as a whole. Near-complete phylogenetic and trait coverage would permit elucidation of the tempo and mode of global diversification of this large, ecologically dominant clade, enabling comparative analyses with other major lineages of life, and eventually global assessment and synthesis of the evolution of terrestrial landscapes. Because the rosid clade and its associated biomes constitute a major driver of terrestrial biodiversity, predicting

future biodiversity patterns for rosids based on historical diversification may likewise be key to understanding the future of other terrestrial clades of life. In short, the rosid clade provides the opportunity to link our understanding of biodiversity from the past to both present and future. We proceed by outlining key properties of the clade and how these exemplify the prerequisites for building any large-scale comparative system.

<h2>Paleo-perspectives

Rosids have a particularly good fossil record. Many families are well known for their detailed fossil histories (e.g., Fabaceae, Juglandaceae, Betulaceae, Fagaceae); overall, all major subclades (Fig. 1), have well-documented fossils (Manchester, 1988,1989,1992, 1994a, b, 2001; Crepet and Nixon, 1989; Cevallos-Ferriz and Stockey, 1991; Herendeen et al., 1992; Pigg et al., 1993; Boucher et al., 2003; Endress and Friis, 2006; Manchester et al., 2006, 2012; DeVore and Pigg, 2007; Burge and Manchester, 2008; Wing et al., 2009; Estrada-Ruiz and Martínez-Cabrera, 2011; Herrera et al., 2012, 2014; Gandolfo et al., 2011; Han et al., 2016; Jud et al., 2016; Larson-Johnson, 2016; Wang et al., 2013; Xing et al., 2014). This rich record provides a superb opportunity for integration of the fossil record with modern diversity and a critical resource for novel approaches for time-calibrating the rosid phylogeny (e.g., Gavryushkina et al., 2017).

The rosid clade originated in the Early to Late Cretaceous (115–93 million years ago [Ma]), followed by rapid diversification of two major subclades, the *Fabidae* and *Malvidae* crown groups, about 112 to 91 Ma and 109 to 83 Ma, respectively (Wang et al., 2009; Bell et al., 2010). The rosid clade is further divided into clades recognized as 17 orders and 135 families (APG IV, 2016; Fig. 1).

<h2>Rosids and terrestrial biome dynamics

Understanding rosid evolution also means characterizing the origin and diversity of major biomes. The radiation of the rosids represents the presumably rapid rise of angiosperm-dominated forests and associated co-diversification events that profoundly shaped much of current terrestrial biodiversity (Wang et al., 2009; Boyce et al., 2010). Among major clades in the land plants, perhaps only the grasses and conifers (both smaller clades that are better understood phylogenetically than rosids) could also lay claim to

building biomes covering large sections of the globe. The megadiverse rosid clade is home to most dominant forest trees (e.g., Betulaceae [alder, birch], Casuarinaceae [Australian pine], Fabaceae [legumes], Fagaceae [oak], Juglandaceae [walnut, hickory], Moraceae [fig], Salicaceae [willow], Ulmaceae [elm], Rutaceae [citrus], Meliaceae (mahogany), Sapindaceae [maple, buckeye], Malvaceae [linden], Dipterocarpaceae [dipterocarps], and Myrtaceae [eucalypts]). Rosid herbs and shrubs are also prominent components of arctic/alpine and temperate floras (e.g., Salicaceae, Rosaceae, Brassicaceae) and comprise aquatics (e.g., Podostemaceae), desert plants (e.g., Euphorbiaceae), and parasites (e.g., *Rafflesia*).

Rosid-dominated forests changed the terrestrial landscape, and this biome-shaping clade has been responsible for the concomitant diversification of other clades (e.g., ants, beetles, amphibians, and other animals; fungi; liverworts, ferns) that inhabit these forests. Accumulating evidence shows that other terrestrial lineages quite literally evolved and diversified in the shadow of rosid-dominated angiospermous forests (Farrell, 1998; Wilf, 2000; Algeo et al., 2001; Schneider et al., 2004; Moreau et al., 2006; Bininda-Emonds et al., 2007; Roelants et al., 2007; Hibbett and Matheny, 2009; Wang et al., 2009; Watkins and Cardelus, 2012; Moreau and Bell, 2013; Feldberg et al., 2014).

<h2>Applied dimensions

Rosids exhibit spectacular diversity in biological processes that may be responsible for the many practical uses of members of the clade. Foremost among these are symbioses with nitrogen-fixing bacteria in legumes and nine other families, the phylogenetic distribution of which is remarkably concentrated in one clade, the nitrogen-fixing clade (Soltis et al., 1995; Werner et al., 2014; Li et al., 2015). This symbiosis has enabled many members to thrive in resource-poor soils; thus, the functional genomics of this symbiosis is of great interest for crop improvement (Stokstad, 2016). Rosids also exhibit diverse phytochemistry, providing potent biochemical defense mechanisms, such as glucosinolate production in Brassicales (Rodman et al., 1998; Edger et al., 2015). This chemical diversity is also associated with the many economic uses of members of Brassicaceae. The plant model *Arabidopsis thaliana* (Brassicaceae) is in the rosid clade; many other rosids are also genetic models with sequenced genomes, e.g., *Brassica rapa* also of Brassicaceae (*Brassica*

rapa Genome Sequencing Project Consortium, 2011) and several legumes (Sato et al., 2008; Schmutz et al., 2010, 2014; Young et al., 2011; Varshney et al., 2012, 2013).

<H1>CHALLENGES IN THE ROSIDS

<*h2*>*State of the art*

Despite the ecological and economic importance of rosids, after decades of data accumulation, our knowledge of the clade remains remarkably limited along any metric. Rosids thus not only serve as a case study for the possibilities of large-scale biodiversity research, but also reveal the constraints on this research due to limitations in basic biodiversity knowledge. This knowledge gap is characteristic of nearly all large clades across the Tree of Life with the possible exception of vertebrates. Shedding a quantitative light on these disparities is critical to raising awareness about how little we truly know about global biodiversity and identifying priorities for future efforts in flowering plants.

Mapping DNA sequence availability onto a supertree estimate of the complete rosid clade (Fig. 2, combining both phylogenetic and taxonomic knowledge from the Open Tree of Life; Hinchliff et al., 2015) shows that current DNA sampling of rosids is highly biased toward subclades of economic interest and significant temperate diversity (e.g., legumes). Groups with the worst representation (e.g., Malpighiales) have few economically important members, yet are critical elements of tropical floras. Only a minority of rosid species—30,234 of 90,000, or 34%—have sequence data of any kind in GenBank (https://www.ncbi.nlm.nih.gov/genbank/). Many of these sequences are microsatellites, ESTs, or other sequences with low species coverage and are not usable for phylogenetics. Even well-known clades, such as Rosales (predominantly temperate), are poorly represented, with only 23% of species having usable DNA sequence data available (Table 1). Only one small group, Fagales, surpasses 50% coverage. Curating the available DNA sequence data for supermatrix phylogenetic analyses (Sun et al., 2016) results in further loss of data, leaving approximately 21% of species across the rosids represented as phylogenetic tips. The pattern of incomplete and biased taxon sampling in the rosids (cf. Fig. 2) is largely true of the angiosperms in general (see Fig. 2 of Eiserhardt et al. [2018] in this issue). Most known species still have no DNA data at all (Drew, 2013; Hinchliff et al., 2015); the vast majority of the flowering plant branch of the tree of life remains dark.

<h2>Phylogenetic bias

Large-scale phylogenetic efforts typically require integrating efforts and data sets from heterogeneous sources, including focused phylogenetic analyses, DNA barcoding data sets, genomic resources, and other data that were not purpose-built for comprehensive species-level inference at global scales. The piecemeal assembly of data sets often makes it difficult to control for uneven sampling of clades. A future need is the development of approaches to assess and correct phylogenetic bias in taxon sampling (either directly through improved sampling or indirectly through modeling taxon absence). In principle, phylogenetically even but incomplete sampling can be accounted for under many models if taxon sampling is unbiased (e.g., FitzJohn et al., 2009). Change in the overall shape of the tree due to biased sampling is not easily controlled for and will likely alter conclusions under models that make inferences from tree topology and branch lengths.

As more researchers assemble large-scale phylogenomic data sets, we see a need for identifying gaps in the coverage of the tree of life and of deploying this knowledge in sequencing efforts to fill these gaps and avoid duplication of effort (see also Eiserhardt et al., 2018, this issue). Although some general-purpose loci have been developed for the angiosperms (e.g., Léveillé-Bourret et al., 2018 and the PAFTOL project; Eiserhardt et al., 2018, this issue), custom-developed, often non-overlapping loci remain the norm (e.g., Weitemier et al., 2014; Mandel et al., 2014; Folk et al., 2015; Chamala et al., 2015; Schmickl et al., 2015), creating greater difficulties for post-hoc aggregation across these experiments.

<h2>Spatial bias

In addition to building comprehensive phylogenetic hypotheses, an ongoing trend in comparative research has been the assembly of equally comprehensive and globally scaled data layers. Recent plant contributions in this spirit include Werner et al. (2014), Zanne et al. (2014), and Díaz et al. (2016). For many clades of land plants, traits and geographic data are missing for most species in existing databases. This lack of coverage results partly from bias in the cumulative assembly of species trait and occurrence data over time, typically from aggregating a long series of small-scale or specialized projects and digitization efforts. Such data accumulation is highly correlated with sociological factors such as gross domestic

product, local funding sources, and distance to institutions performing digitization (Amano and Sutherland, 2013; Meyer et al., 2015). One hallmark of spatial bias is an inverse latitudinal gradient clearly observable in the rosids (Fig. 3A), where records are least heavily accumulated and species least completely represented in the tropics, some of the most biodiverse parts of the world for the rosids (Fig. 3B). Because major rosid clades are not evenly distributed across the globe (e.g., Malpighiales and Rosales are associated, respectively, with tropical and temperate latitudes), spatial and phylogenetic bias are likely to interact.

Spatial bias may propagate to downstream analyses that do not explicitly include spatial data, such as those focusing on potentially correlated traits and taxon coverage. Hence, spatial bias can occur at multiple levels of sampling; accumulation of phylogenetic tips, occurrences, and species traits are all influenced by availability of material and digitization efforts. Most directly, spatial bias has an enormous impact on the spatial distribution of occurrence records, such that nearly any large-scale clade in the tree of life has an occurrence density pattern matching closely that seen in the rosids (Fig. 3a; compare with global mammal GBIF records: Boitani et al., 2011: fig. 2). This strong bias is partly due to historical differences in collection effort. However, differing levels of investment in biodiversity digitization among countries also contribute to this unevenness, which is compounded by the tendency of digitization efforts to be locally focused initially, even for internationally representative collections (Amano and Sutherland, 2013; Meyer et al., 2015).

As with phylogenetic bias, we see not only challenges but opportunities. It would be a major step toward enabling research if future efforts specifically assigned digitization priorities on the basis of evidence for data gaps in current infrastructure. For most herbaria, it is not feasible in the immediate future to completely digitize all specimens, including georeferences, images, and other data. Targeting data gaps would provide an evidence-based method to direct digitization efforts and maximize downstream research impact.

<h2>Linked data

Linking data sets such as those discussed above is critical for large-scale inference. For instance, a common task is to subset a tree for the group of interest using a list of taxon names. Linked data already have a role—providing linkages between taxonomic concepts

and a phylogeny. If unusual phylogenetic placements are observed, it might be necessary to retrieve either original voucher specimen photographs or original sequence data. Finally, using the name list, linkages could allow users to subset trait data from online repositories such as the TRY Plant Trait Database (https://www.try-db.org/); both unusual trait scores and the possibility of polymorphism would warrant consulting original specimen material using online herbaria. Central to these aims are stable identifiers built around taxon concepts to facilitate linking of disparate data products. Links between genetic data, online herbaria, and phylogenetic tips are typically not explicit and need to be laboriously sought manually, although some linkages, such as that between GenBank and iDigBio (https://www.idigbio.org/), are currently being developed. For example, herbarium specimen records in iDigBio that serve as vouchers for GenBank sequences and have globally unique identifiers on GenBank are linked to their associated DNA sequences; unfortunately, globally unique identifiers are not consistently used or formatted properly (Guralnick et al., 2014), thwarting efforts to link most data directly.

Community consensus is lacking about minimal reporting standards for integrative research programs that include multiple data types. Minimally, we recommend that these projects should contain unique sample identifiers (e.g., GUIDs) as part of data deposition in standard data-specific repositories (e.g., GenBank and SRA; https://www.ncbi.nlm.nih.gov/sra). Unambiguous identifier practices will enable future researchers to scrape metadata for recognizable identifiers and retrieve matching information generated downstream from those samples, such as sequences, modeled geographic distributions, and other data and knowledge products.

<h2>Name reconciliation

Reconciling conflicting taxon identifiers is unavoidable for any project that attempts to accrue multispecies data from diverse sources yet remains a core challenge of large-scale biology (Patterson et al., 2010). Many large-scale databases have their own internal taxonomy (e.g., GBIF https://www.gbif.org/; GenBank; Open Tree, https://tree.opentreeoflife.org/), and standalone name products also exist (e.g., The Plant List, http://www.theplantlist.org; Tropicos, http://www.tropicos.org/). These taxonomies sometimes represent conflicting taxonomic opinions and often are incomplete and partially

out of date. Taxonomic mismatch results in major discrepancies in accepted genera, total species number, and other important metrics that inform sampling, analysis, and synthesis. The availability of community reconciliation services (Boyle et al., 2013) is an important step toward resolving these issues, at least for providing current assessments of valid taxon names. A much-needed area of growth is the improvement of existing databases by digitizing and incorporating major, yet largely inaccessible, natural history literature (below). While necessary for building the framework of online taxonomies, a static, centralized approach to the name reconciliation problem (generally the approach used to date) will lack permanency given the continual flux of taxon delimitation (Lepage et al., 2014), meaning that a resource that is updatable, preferably by the community and in close to real time, will be critical to improving resources beyond those available to date.

<h2>Expert and algorithmic range products

A rich heritage of geographic range products is available for tetrapods, resulting from massive data digitization that has enabled comprehensive macroecological analyses and conservation-oriented decision-making (Jetz et al., 2012a, b; Meyer et al., 2015). In addition to purely expert-drawn range maps, automated approaches based on point occurrences have also been developed recently (e.g., Merow et al., 2016, 2017), offering the potential for generating geographic range products in clades where few ranges have been expert-assessed. Range data are complementary to better-known occurrence record data, as range data have the potential to coarsely assess true species absence rather than pseudoabsence (Jetz et al., 2012a). Range products are not only useful for direct empirical analyses, but also for quality control of occurrence records for other research (Jetz et al., 2012b). Occurrence data sets too large to curate entirely by hand can be automatically checked against expert-derived range maps using a spatial join to remove data points likely to be incorrect. These maps typically require expert involvement to produce credible estimates and are themselves hypotheses open to reinterpretation with new reports of species detection (or lack thereof).

<h2>Digitization of legacy natural history data

Enormous effort has been made in increasing access to data in biological collections (e.g., VertNet, iDigBio, and GBIF). The availability of these resources has facilitated growth in macro-perspectives in ecology and evolution; the vast number of papers using repositories of occurrence records (nearly 6000 according to GBIF.org, 2017) illustrates how natural history data drive progress in biodiversity science. Despite this effort, literature containing natural history data in plants remain untapped resources that are as rich as specimen data. Rather than direct point observations, literature sources represent expert-assessed consensus values for geographic range (see above) and phenotype, as well as a consensus taxonomic product for a given region in the form of accepted taxa. For large-scale digitization strategies, large-scale floras are ideal data sources. These floras typically comprise comprehensive treatments of a specific area of the globe, covering information such as accepted species lists, partial synonymies, whole-plant trait data, coarse-scale geographic range descriptors at the country, state, or other regional level, and variable additional features including chromosome number and invasive status. Regional taxonomic treatments are rich data sets; products of broad utility that can be developed from these treatments include (1) improved taxon name resolution, which could be combined with existing name databases for an improved consensus product; (2) coarse-scale range maps such as are available for vertebrates, typically of political regions, for inferences of range evolution, invasive species status, or quality assessment of occurrence data and spatial bias; and (3) very large morphological matrices.

eFloras, such as the *Flora of North America* (Flora of North America Editorial Committee, 1993 onward) and *Flora of China* (Wu et al., 1994 onward; Brach and Song, 2008), represent low-hanging fruit for data mining. The text in these efforts does not identify descriptors (e.g., morphological terms do not have explicit metadata), so that indirect text scraping strategies are needed to match descriptors among taxa. While text scraping requires considerable effort, the pay-off is substantial for obtaining organismal information for hundreds or thousands of phylogenetic tips. Some recent efforts (e.g., *Flora of Tropical West Africa*; https://archive.org/details/FloraOfWestTropi00hutc) are partially semantically tagged, so that sub-blocks of text, such as a trait-related text block, can be obtained for further processing. Unfortunately, few other flora projects are so accessible. Although this is changing, e.g., for *Flora Malesiana* (Nooteboom et al. 2010 onwards) and

Flora of New Zealand (Breitwieser et al. 2010 onwards), many recent and ongoing floras are not available online. Addressing these gaps in flora production would facilitate significant progress towards the vision of illuminating the dark parts of the tree of life, going beyond simply populating the tree with tip taxa by adding geographic and trait data layers with the assistance of partially automated approaches (Burleigh et al., 2013; Liu et al., 2015; Cui et al., 2016; Endara et al. 2018).

<H1>WHY USE COMPREHENSIVE APPROACHES?

An obvious first step in performing large-scale analyses is identifying the motivation for what may be a costly and labor-intensive enterprise spanning years from planning to fruition. Why fill in the dark parts of the tree, for rosids or any other clade, if we already understand higher-level relationships? Why indeed "go big" in phylogenetics? Why not "go small" many times in succession on small subclades and ultimately sum these well-worked case studies up to the ecological and evolutionary whole? Discussion on this point is important because basic questions have been raised about the inherent value of large phylogenies for testing hypotheses in evolution and ecology (Donoghue and Edwards, 2014).

<h2>Exemplar clade

With respect to the rosids or any other group, the choice of taxon for addressing large-scale hypotheses should be evidence-based and targeted toward finding groups appropriate in scale and properties for a given research question. Explicitly or implicitly, much recent work in phylogenetics sets its aims more broadly than inferences solely constrained to the group of interest, such that the use of comprehensive approaches has contributed insights for decades in evolution and ecology (see an early review by Pagel, 1999). As has long been the case for small clades, large-scale phylogenetic research should explicitly provide reasons for studying *exemplar clades* embodying the prerequisites for understanding particular evolutionary or ecological dynamics. We use "exemplar clade" to denote a monophyletic group that captures generalizable ecological and evolutionary processes for the purpose of analytical inference. An exemplar clade (= "model clade"; e.g.,

Chanderbali et al., 2016) thus serves as a biodiversity "model" in a phylogenetic framework, with the aim of inference placed more broadly than the group under concern.

Selection of a study group should not be based primarily on data availability, a criterion that would likely only exacerbate existing knowledge gaps and phylogenetic biases in future investigations—away from what are already dark parts of the tree of life. If the aim is to study generalizable principles and processes across the angiosperms, or in other parts of the tree of life, developing large exemplar clades as community resources puts global-scale research into reach, the conclusions of which will be reciprocally enhanced as other comprehensive comparative data sets are developed.

<h2>***A tale of two approaches***

The comparative method has as its goal the testing of hypotheses using multispecies samples in a phylogenetic framework (Felsenstein, 1985). Recently, a dichotomy has been proposed, identifying what may be complementary or conflicting alternative approaches to such macroevolutionary questions (Donoghue and Edwards, 2014). One could either (1) use an integrative, large-scale approach to test hypotheses in a single framework (e.g., Meredith et al., 2011; Jetz et al., 2012a; Zanne et al., 2014), or (2) accrue a large number of small-scale, well-characterized clades, which investigators would follow by a qualitative synthetic review (e.g., Soltis et al., 2006; Donoghue and Edwards, 2014) or quantitative meta-analyses (e.g., Mayrose et al., 2011) to test the same large-scale hypotheses.

Large-scale studies have been criticized by some based in part on three largely accurate observations: (1) robust and comprehensive clade and trait sampling is very challenging to achieve on large scales, (2) identifying appropriate evolutionary models is difficult, in that a sample representing a long timespan is likely to capture a large number of evolutionary dynamics, and (3) individual instances substantiating broad patterns are anonymized and massaged out of the message of many such studies. These issues are more easily overcome if taxonomic sampling is intentionally placed within modest limits.

Despite these concerns, the scale of systematics research is steadily increasing, through improved sampling of both taxa and loci, generating phylogenetic matrices that are growing both "taller" and "wider." The same growth is true for trait and occurrence data sets that accompany phylogenetic matrices. But a community trend does not constitute

justification ipso facto; it is reasonable that the choice of a large-scale analytical approach should be accompanied by compelling reasons for being large, as we have outlined above. Likewise, are there also risks for intentionally small, well-circumscribed scales in biodiversity science?

<h2>Emergent processes

Perhaps the most immediate problem of integrating over large numbers of small case studies is the potential for consistently failing to recover patterns that inherently cannot appear in small data sets. This problem concerns analytical scale: how do we build data sets appropriate for the phylogenetic and temporal scales at which we are testing hypotheses? We argue that biodiversity questions posed globally across large taxonomic groups require sampling that is appropriate to global scales of inference. Synthesizing knowledge in this way across large expanses of space and time will consistently compel the analysis of large data sets. The use of small clades to answer questions at large scales leads to data sets that are well characterized but restricted in their sampling of biological diversity. We identify conditions below where such sampling scales could obscure emergent signals and impact hypothesis testing.

One core issue is statistical power. For inference of diversification and other approaches that use highly parameterized models, branches and their lengths are the data points. Hence, fairly large phylogenies, on the order of hundreds to thousands of taxa under idealized simulated conditions (e.g., diversification: Davis et al., 2013; Rabosky and Huang, 2016; phylogenetic correlation: Ackerly, 2009) are required to have sufficient sensitivity to detect shifts in diversification with high power. It is expected, therefore, that an intentionally taxon-limited approach will consistently underestimate the number of diversification shifts and the occurrence of character-associated diversification patterns. Although no quantitative studies have been performed to assess the effects of taxonomic scope beyond statistical power, we expect that the number of significant evolutionary patterns extractable from phylogenetic data will be consistently and artificially truncated by focusing on small case studies. Such a truncation is likely for the simple reason that such patterns may be present in subclades but without the context of broader sampling that would make them detectable.

Estimation error increases with increasingly deep trees (Salisbury and Kim, 2001), and even within a given tree, estimation error is expected to increase as estimated nodes approach the root (Garland et al., 1999), leading to unequal error in ancestral state reconstruction across a tree. If a particular ancestral state is of interest, it is possible that removing taxa could result in smaller estimated uncertainty by incompletely sampling evolutionary transitions (Heath et al., 2008a), thus underestimating trait evolutionary rates and decreasing the magnitude of estimated error (e.g., the confidence interval, cf. Garland et al., 1999). Hence, a smaller reported uncertainty does not necessarily imply that the "true" error of such an estimate has actually decreased due to sampling scheme alone. Building data sets appropriate to the scale of questions posed—for global-scale analyses, this often means including data for as many extant species as possible, maximizing the information behind our inferences and the estimated uncertainty thereof—is therefore preferable.

The detection of some processes may fundamentally require large phylogenies, irrespective of statistical power. This problem is subtler, in that it cannot be easily measured or controlled for by performing statistical power studies or extending models to account for potential data set biases. Such a problem is likely to occur in instances where deep-level patterns in highly diverse clades (e.g., the root of major angiosperm clades) are the object of inference, but where inferences are sensitive to taxon sampling. This situation could appear in ancestral state reconstruction, where a deep-level node is of interest, but the polarity of ancestral states is impacted by a complex distribution of states in descendant extant taxa. Some of the risks of poor taxon sampling in this case include incomplete sampling of evolutionary transitions in the clade of interest (Heath et al., 2008a) and warping of overall tree shape by dropping taxa (Heath et al., 2008b). These concerns cannot both be addressed in small test cases (in this case, sets of trees with limited taxon sampling at deep levels) if the relevant information for accurately distinguishing among possible ancestral states is not present in the data, irrespective of our ability to detect it. Simulation studies have shown increased estimation error as proportional taxon coverage decreases (Salisbury and Kim, 2001; Litsios and Salamin, 2012; but see Li et al., 2008).

A final issue with a solely small-scale focus, raised by Beaulieu and O'Meara (2018, this issue), is ascertainment bias. The choice of idealized small-scale clades to understand broad-scale patterns—often resulting in a focus on groups showing especially frequent shifts

in a biological trait—may result in overemphasis of unusual outlier taxa unrepresentative of overall variation patterns. Hence, large-scale biodiversity studies are needed to complement and contextualize focused clade-level studies. Likewise, as we have suggested for the rosids, the suitability of an exemplar clade is a testable assumption that can be directly assessed by asking how well a focal clade cross-sections broader diversity patterns.

Issues of both statistical power and levels of inference imply that questions exist that are uniquely suited to purposeful attempts at comprehensive taxon sampling, such that focusing solely on small, well-characterized case studies is neither always sufficient nor invariably necessary. Approaches in biodiversity science that use small study clades will continue to be relevant, particularly for understanding recent-scale evolutionary processes. By contrast, the application of such sampling schemes to global questions poses risks, possibly resulting in data sets with high confidence in individual data points but restricted and possibly biased coverage of the biodiversity that underlies many biological processes. Comprehensive phylogenetic approaches that span deep-time and global geographic scales are urgently needed for the kinds of grand challenges which the comparative approach to biology is poised to address, due not simply to an obsession with larger and more resolved data sets (Hahn and Nakhleh, 2016), but to their central necessity for answering questions on deep-time and global scales in highly diverse clades.

<h2>Ways forward

In our view, large- and small-scale approaches are complementary. Some questions are best addressed with small clades. Increasingly, however, phylogenetic effort is devoted to asking questions in evolution and ecology that require large trees and comprehensive taxon sampling (e.g., global patterns of diversification, deep-time ancestral state reconstruction and biogeography, correlated evolution of characters, community phylogenetics), often in a model-based or otherwise explicitly quantitative framework (e.g., Smith and Donoghue, 2008; Smith and Beaulieu, 2009). We argue that the need remains for large-scale, comprehensive approaches that are appropriate to address questions of major importance.

We stress that focused case studies on small clades remain crucial for addressing certain specific questions and serve as an important element of building comparative data

sets. Nonetheless, despite substantial progress in many domains, 30 years of effort on small focal clades in molecular systematics have resulted in uneven and incomplete coverage of rosids in particular, and angiosperm diversity as a whole, suggesting this approach alone may not suffice to eventually synthesize biodiversity knowledge across the flowering plants. Targeted and coordinated, large-scale sampling efforts at the community level are needed to complement these efforts and directly address data and knowledge gaps that have continually persisted despite intense efforts by individual researchers. Rather than continually aggregating upward in scope from focused data sets to create incomplete and biased larger sets, we can do more to collect comprehensive biodiversity data broadly for future users to disaggregate downward for focused work.

<H1>CONCLUSIONS

Much progress has been made in understanding deep-level relationships in the angiosperms (Chase et al., 1993; Qiu et al., 1999; Soltis et al., 1999, 2011) with large-scale sequencing projects (e.g., 1KP, Matasci et al., 2014) resulting in robust backbone resolution (Wickett et al., 2014) and community consensus taxonomic products (APG IV, 2016). Current efforts in plant systematics beyond the backbone have largely remained centered on localized taxonomic sampling efforts, with less consideration of how to develop more comprehensive, community-based, synthetic investigations or of whether such goals are feasible without purposeful large-scale generation of phylogenetic data to fill in gaps. Yet, it is just these kinds of efforts that can provide the most critical insights and applications in biology, particularly those posed at global or deep-time scales. The effort to develop such synthetic analyses is still enormous, and bottlenecks are multidimensional.

We make the case for an evidence-based assessment as we build comprehensive community resources for phylogenetically informed hypothesis testing, with a focus on exemplary, hyper-diverse clades such as the rosids. Such resources, to maximize enabled research, should comprehensively sample phylogenetic tips and linked phenotypic and geographic data as a community priority. This approach is complementary to focal studies on smaller clades, which may address significant problems but on different phylogenetic and temporal scales; both can help with goals geared towards broad-scale synthesis. However, we believe that purpose-built comprehensive phylogenies covering global scales

and ancient radiations are valuable resources that, when linked to other biodiversity data and knowledge products, will be an impetus for transformative research.

<H1>ACKNOWLEDGEMENTS

The authors benefitted from workshops supported by the Open Tree of Life (https://tree.opentreeoflife.org/) and FuturePhy (https://futurephy.org/) projects for discussions on building community resources. B. Thiers granted access to unpublished Amazonia records for our species richness estimates. M. Gitzendanner assisted with GenBank gap analyses. E. Edwards and an anonymous reviewer are thanked for comments on an earlier draft of this manuscript. This paper was supported in part by NSF (DBI-1523667 to R.A.F.; DBI-1458640 to D.E.S. and P.S.S.; DBI-1458466 to S.A. Smith; DEB-1442280 to P.S.S. and D.E.S.).

<h1>LITERATURE CITED

Ackerly, D.D. 2009. Taxon sampling, correlated evolution, and independent contrasts. *Evolution* 54: 1480–1492.

Algeo, T.J., S.E. Scheckler, and J.B. Maynard. 2001. Effects of the Middle to Late Devonian spread of vascular land plants on weathering regimes, marine biotas, and global climate. *In* Gensel, P. G. and D. E. (eds). Plants invade the land: evolutionary and environmental perspectives, 213–236. Columbia University Press, NY, USA.

Amano, T., and W.J. Sutherland. 2013. Four barriers to the global understanding of biodiversity conservation: Wealth, language, geographical location and security. *Proceedings of the Royal Society, B, Biological Sciences* 280: 20122649.

APG IV [Angiosperm Phylogeny Group IV]. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* 181: 1–20.

Beaulieu, J.M. and B.C. O'Meara. 2018. Can we build it? Yes we can, but should we use it? Assessing the quality and value of a very large phylogeny of campanulid angiosperms. *American Journal of Botany* 104 (in press). DOI: https://doi.org/10.1002/ajb2.1020.

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

Bininda-Emonds, O.R.P., M. Cardillo, K.E. Jones, R.D.E. MacPhee, R.M.D. Beck, R.

Grenyer, S.A. Price, et al. 2007. The delayed rise of present-day mammals. *Nature* 446: 507–512.

Boitani, L., L. Maiorano, D. Baisero, A. Falcucci, P. Visconti, and C. Rondinini. 2011. What spatial data do we need to develop global mammal conservation strategies? *Philosophical Transactions of the Royal Society, B, Biological Sciences* 366: 2623–2632.

Boyce, C.K., J.-E. Lee, T.S. Feild, T.J. Brodribb, and M.A. Zwieniecki. 2010. Angiosperms helped put the rain in the rainforests: The impact of plant physiological evolution on tropical biodiversity. *Annals of the Missouri Botanical Garden* 97: 527–540.

Boyle, B., N. Hopkins, Z. Lu, J.A. Raygoza Garay, D. Mozzherin, T. Rees, N. Matasci, et al. 2013. The taxonomic name resolution service: An online tool for automated standardization of plant names. *BMC Bioinformatics* 14: 16.

Brach, A.R., and H. Song. 2008. eFloras: New directions for online floras exemplified by the Flora of China Project. *Taxon* 55: 188–192.

The *Brassica rapa* Genome Sequencing Project Consortium. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics* 43: 1035–1039.

Breitwieser I., P.J. Brownsey, P.B. Heenan, W.A. Nelson, A.D. Wilton [eds.]. 2010. Flora of New Zealand Online. Available at www.nzflora.info.

Burge, D.O., and S.R. Manchester. 2008. Fruit morphology, fossil history, and biogeography of *Paliurus* (Rhamnaceae). *International Journal of Plant Sciences* 169: 1066–1085.

Burleigh, J.G., K. Alphonse, A.J. Alverson, H.M. Bik, C. Blank, A.L. Cirranello, H. Cui, et al. 2013. Next-generation phenomics for the Tree of Life. *PLOS Currents Tree of Life*. doi:10.1371/currents.tol.085c713acafc8711b2ff7010a4b03733.

Cevallos-Ferriz, S.R., and R.A. Stockey. 1991. Fruits and seeds from the Princeton chert (Middle Eocene) of British Columbia: Rosaceae (Prunoideae). *Botanical Gazette* 152: 369–379.

Chamala, S., N. García, G.T. Godden, V. Krishnakumar, I.E. Jordon-Thaden, R.D. Smet, W.B. Barbazuk, et al. 2015. MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. *Applications in Plant Sciences* 3: 1400115.

Chanderbali, A.S., B.A. Berger, D.G. Howarth, P.S. Soltis, and D.E. Soltis. 2016. Evolving ideas on the origin and evolution of flowers: New perspectives in the genomic era. *Genetics* 202: 1255–1265.

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.

Crepet, W.L., K.C. Nixon, and M.A. Gandolfo. 2004. Fossil evidence and phylogeny: The age of major angiosperm clades based on mesofossil and macrofossil evidence from Cretaceous deposits. *American Journal of Botany* 91:1666–1682.

Cui, H., D. Xu, S.S. Chong, M. Ramirez, T. Rodenhausen, J.A. Macklin, B. Ludäscher, et al. 2016. Introducing Explorer of Taxon Concepts with a case study on spider measurement matrix building. *BMC Bioinformatics* 17: 471.

Davis, M.P., P.E. Midford, and W. Maddison. 2013. Exploring power and parameter estimation of the BiSSE method for analyzing species diversification. *BMC Evolutionary Biology* 13: 38.

de Casas, R.R., M.E. Mort, and D.E. Soltis. 2016. The influence of habitat on the evolution of plants: A case study across Saxifragales. *Annals of Botany* 118: 1317–1328.

DeVore, M.L. and K.B. Pigg. 2007. A brief review of the fossil history of the family Rosaceae with a focus on the Eocene Okanogan Highlands of eastern Washington State, USA, and British Columbia, Canada. *Plant Systematics and Evolution* 266: 45–47.

Díaz, S., J. Kattge, J.H.C. Cornelissen, I.J. Wright, S. Lavorel, S. Dray, B. Reu, et al. 2016. The global spectrum of plant form and function. *Nature* 529: 167–171.

Donoghue, M.J., and E.J. Edwards. 2014. Biome shifts and niche evolution in plants. *Annual Review of Ecology, Evolution, and Systematics* 45: 547–572.

Drew, B.T. 2013. Data deposition: Missing data mean holes in tree of life. *Nature* 493: 305–305.

Edger, P.P., H.M. Heidel-Fischer, M. Bekaert, J. Rota, G. Glöckner, A.E. Platts, D.G. Heckel, et al. 2015. The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences, USA* 112: 8362–8366.

Eiserhardt, W.L., A. Antonelli, D.J. Bennett, L.R. Botigué, J.G. Burleigh, S. Dodsworth, B.J. Enquist, F. Forest, et al. 2018. A roadmap for global synthesis of the plant tree of

life. *American Journal of Botany* 104 (in press).

Endara, L., H. Cui, and J.G. Burleigh. 2018. Extraction of phenotypic traits from taxonomic descriptions for the tree of life using Natural Language Processing. *Applications in Plant Sciences* 6 (in press).

Endress, P.K., and E.M. Friis. 2006. Rosids—Reproductive structures, fossil and extant, and their bearing on deep relationships: Introduction. *Plant Systematics and Evolution* 260: 83–85.

Estrada-Ruiz, E., and H.I. Martínez-Cabrera. 2011. A new late Cretaceous (Coniacian-Maastrichtian) *Javelinoxylon* wood from Chihuahua, Mexico. *IAWA Journal* 32: 521–530.

Farrell, B.D. 1998. "Inordinate fondness" explained: Why are there so many beetles? *Science* 281: 555–559.

Faurby, S., and J.-C. Svenning. 2015. A species-level phylogeny of all extant and late Quaternary extinct mammals using a novel heuristic-hierarchical Bayesian approach. *Molecular Phylogenetics and Evolution* 84: 14–26.

Feldberg, K., H. Schneider, T. Stadler, A. Schäfer-Verwimp, A.R. Schmidt, and J. Heinrichs. 2014. Epiphytic leafy liverworts diversified in angiosperm-dominated forests. *Scientific Reports* 4: 5974.

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

FitzJohn, R.G., W.P. Maddison, and S.P. Otto. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Systematic Biology* 58: 595–611.

Flora of North America Editorial Committee [eds.], 1993 onward. Flora of North America North of Mexico. 20+ vols. New York and Oxford. Available at http://www.efloras.org/flora_page.aspx?flora_id=1.

Folk, R.A., J.R. Mandel, and J.V. Freudenstein. 2015. A protocol for targeted enrichment of intron-containing sequence markers for recent radiations: A phylogenomic example from *Heuchera* (Saxifragaceae). *Applications in Plant Sciences* 3: 1500039.

Gandolfo, M.A., E.J. Hermsen, M.C. Zamaloa, K.C. Nixon, C.C. González, P. Wilf, N.R. Cúneo, and K.R. Johnson. 2011. Oldest known *Eucalyptus* macrofossils are from South

America. *PLoS One* 6: e21084.

Garland, T. Jr., P.E. Midford, and A.R. Ives. 1999. An introduction to phylogenetically based statistical methods, with a new method for confidence intervals on ancestral values. *American Zoologist* 39: 374–388.

Gavryushkina, A., T.A. Heath, D.T. Ksepka, T. Stadler, D. Welch, and A.J. Drummond. 2017. Bayesian total-evidence dating reveals the recent crown radiation of penguins. *Systematic Biology* 66: 57–73.

GBIF.org. 2017. Literature tracking [online]. Available at https://www.gbif.org/literature-tracking. [Accessed 22 March 2018]

Guralnick, R., T. Conlin, J. Deck, B.J. Stucky, and N. Cellinese. 2014. The trouble with triplets in biodiversity informatics: A data-driven case against current identifier practices. *PLoS One* 9: e114069.

Hahn, M.W., and L. Nakhleh. 2016. Irrational exuberance for resolved species trees. *Evolution* 70: 7–17.

Han, M., G. Chen, X. Shi, and J. Jin. 2016. Earliest fossil fruit record of the genus *Paliurus* (Rhamnaceae) in eastern Asia. *Science China Earth Sciences* 59: 824–830.

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

Heath, T.A., D.J. Zwickl, J. Kim, and D.M. Hillis. 2008b. Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. *Systematic Biology* 57: 160–166.

Herendeen, P.S., W.L. Crepet, and D.L. Dilcher. 1992. The fossil history of the Leguminosae: Phylogenetic and biogeographical implications. *In* P. S. Herendeen and D. L. Dilcher [eds.], Advances in legume systematics 4, 303–316. The Fossil Record, Royal Botanical Gardens, Kew, UK.

Herrera, F., S.R. Manchester, and C.A. Jaramillo. 2012. Permineralized fruits from the late Eocene of Panama give clues of the composition of forests established early in the uplift of Central America. *Review of Palaeobotany and Palynology* 175: 10–24.

Herrera, F., S.R. Manchester, J. Vélez-Juarbe, and C.A. Jaramillo. 2014. Phytogeographic history of the Humiriaceae (Part 2). *International Journal of Plant Sciences* 175: 828–840.

Hibbett, D.S., and P.B. Matheny. 2009. The relative ages of ectomycorrhizal mushrooms and their plant hosts estimated using Bayesian relaxed molecular clock analyses. *BMC Biology* 7: 13.

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.

Höhna, S., M.J. Landis, T.A. Heath, B. Boussau, N. Lartillot, B.R. Moore, J.P. Huelsenbeck, and F. Ronquist. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology* 65: 726–736.

Jetz, W., J.M. McPherson, and R.P. Guralnick. 2012b. Integrating biodiversity distribution knowledge: Toward a global map of life. *Trends in Ecology & Evolution* 27: 151–159.

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

Jud, N. A., C. W. Nelson, and F. Herrera. 2016. Fruits and wood of *Parinari* from the early Miocene of Panama and the fossil record of Chrysobalanaceae. *American Journal of Botany* 103: 277–289.

Kunstler, G., D. Falster, D.A. Coomes, F. Hui, R.M. Kooyman, D.C. Laughlin, L. Poorter, et al. 2015. Plant functional traits have globally consistent effects on competition. *Nature* 529: 204–207.

Larson-Johnson, K., 2016. Phylogenetic investigation of the complex evolutionary history of dispersal mode and diversification rates across living and fossil Fagales. *New Phytologist* 209: 418–435.

Lepage, D., G. Vaidya, and R. Guralnick. 2014. Avibase – a database system for managing and organizing taxonomic concepts. *ZooKeys* 420: 117–135.

Léveillé-Bourret, É., J.R. Starr, B.A. Ford, E. Moriarty Lemmon, and A.R. Lemmon. 2018. Resolving rapid radiations within angiosperm families using anchored phylogenomics. *Systematic Biology* 67: 94–112.

Li, G., M. Steel, L. Zhang, and T. Oakley. 2008. More taxa are not necessarily better for the reconstruction of ancestral character states. *Systematic Biology* 57: 647–653.

Li, H.-L., W. Wang, P.E. Mortimer, R.-Q. Li, D.-Z. Li, K.D. Hyde, J.-C. Xu, et al. 2015.

Large-scale phylogenetic analyses reveal multiple gains of actinorhizal nitrogen-fixing symbioses in angiosperms associated with climate change. *Scientific Reports* 5: 14023.

Litsios, G., and N. Salamin. 2012. Effects of phylogenetic signal on ancestral state reconstruction. *Systematic Biology* 61: 533–538.

Liu, J., L. Endara, and J.G. Burleigh. 2015. MatrixConverter: Facilitating construction of phenomic character matrices. *Applications in Plant Sciences* 3: 1400088.

Manchester, S.R. 1988. Fruits and seeds of *Tapiscia* (Staphyleaceae) from the middle Eocene of Oregon, USA. *Tertiary Research* 9: 59–66.

Manchester, S.R. 1989. Systematics and fossil history of the Ulmaceae. *In* P. R. Crane and S. Blackmore [eds.], Evolution, systematics, and fossil history of the Hamamelidae, vol. 2, 'Higher' Hamamelidae, 221–252. Systematics Association Special Volume 40B. Clarendon Press, Oxford, UK.

Manchester, S.R. 1992. Flowers, fruits and pollen of *Florissantia*, an extinct malvalean genus from the Eocene and Oligocene of western North America. *American Journal of Botany* 79: 996–1008.

Manchester, S.R. 1994a. Fruits and seeds of the Middle Eocene Nut Beds flora, Clarno Formation, Oregon. *Palaeontographica Americana* 58: 1–205.

Manchester, S.R. 1994b. Inflorescence bracts of fossil and extant *Tilia* in North America, Europe and Asia: Patterns of morphologic divergence and biogeographic history. *American Journal of Botany* 81: 1176–1185.

Manchester, S.R. 2001. Leaves and fruits of *Aesculus* (Sapindales) from the Paleocene of North America. *International Journal of Plant Science* 162: 985–998.

Manchester, S.R., I. Chen, and T.A. Lott. 2012. Seeds of *Ampelocissus, Cissus,* and *Leea* (Vitales) from the Paleogene of western Peru and their biogeographic significance. *International Journal of Plant Sciences* 173: 933–943.

Manchester, S.R., W.S. Judd, and B. Handley. 2006. Foliage and fruits of early poplars (Salicaceae: *Populus*) from the Eocene of Utah, Colorado, and Wyoming. *International Journal of Plant Sciences* 167: 897–908.

Mandel, J.R., R.B. Dikow, V.A. Funk, R.R. Masalia, S.E. Staton, A. Kozik, R.W. Michelmore, et al. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the Compositae. *Applications in*

Plant Sciences 2: 1300085.

Matasci, N., L.-H. Hung, Z. Yan, E.J. Carpenter, N.J. Wickett, S. Mirarab, N. Nguyen, et al. 2014. Data access for the 1,000 Plants (1KP) project. *GigaScience* 3: 17.

Mayrose, I., S.H. Zhan, C.J. Rothfels, K. Magnuson-Ford, M.S. Barker, L.H. Rieseberg, and S.P. Otto. 2011. Recently formed polyploid plants diversify at lower rates. *Science* 333: 1257–1257.

Meredith, R.W., J.E. Janecka, J. Gatesy, O.A. Ryder, C.A. Fisher, E.C. Teeling, A. Goodbla, et al. 2011. Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science* 334: 521–524.

Merow, C., J.M. Allen, M. Aiello-Lammens, and J.A. Silander. 2016. Improving niche and range estimates with Maxent and point process models by integrating spatially explicit information. *Global Ecology and Biogeography* 25: 1022–1036.

Merow, C., A.M. Wilson, and W. Jetz. 2017. Integrating occurrence data and expert maps for improved species range predictions. *Global Ecology and Biogeography* 26: 243–258.

Meyer, C., H. Kreft, R. Guralnick, and W. Jetz. 2015. Global priorities for an effective information basis of biodiversity distributions. *Nature Communications* 6: 8221.

Miller, J.S., W.W. Thomas, M. Watson, D. Simpson, and P.W. Jackson. 2014. World Flora Online Council met in St. Petersburg. *Taxon* 63: 959–959.

Moreau, C.S., C.D. Bell, R. Vila, S.B. Archibald, and N.E. Pierce. 2006. Phylogeny of the ants: Diversification in the age of angiosperms. *Science* 312: 101–104.

Moreau, C.S., and C.D. Bell. 2013. Testing the museum versus cradle tropical biological diversity hypothesis: Phylogeny, diversification, and ancestral biogeographic range evolution of the ants. *Evolution* 67: 2240–2257.

Nooteboom, H.P., W.J.J.O de Wilde, D.W. Kirkup, P.F. Stevens, M.J.E. Coode, and J.G. Saw [eds.] 2010 onward. Flora Malesiana. Available at http://portal.cybertaxonomy.org/flora-malesiana/.

Page, R.D.M. 2008. Biodiversity informatics: The challenge of linking data and the role of shared identifiers. *Briefings in Bioinformatics* 9: 345–354.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401: 877–884.

Parr, C.S., R. Guralnick, N. Cellinese, and R.D.M. Page. 2012. Evolutionary informatics: Unifying knowledge about the diversity of life. *Trends in Ecology & Evolution* 27: 94–103.

Patterson, D.J., J. Cooper, P.M. Kirk, R.L. Pyle, and D.P. Remsen. 2010. Names are key to the big new biology. *Trends in Ecology & Evolution* 25: 686–691.

Pigg, K.B., R.A. Stockey, and S.L. Maxwell. 1993. *Paleomyrtinaea*, a new genus of permineralized myrtaceous fruits and seeds from the Eocene of British Columbia and Paleocene of North Dakota. *Canadian Journal of Botany* 71: 1–9.

Qiu, Y.-L., J. Lee, F. Bernasconi-Quadroni, D.E. Soltis, P.S. Soltis, M. Zanis, E.A. Zimmer, et al. 1999. The earliest angiosperms: Evidence from mitochondrial, plastid and nuclear genomes. *Nature* 402: 404–407.

Rabosky, D.L., and H. Huang. 2016. A robust semi-parametric test for detecting trait-dependent diversification. *Systematic Biology* 65: 181–193.

Rodman, J., P. Soltis, D. Soltis, K. Sytsma, and K. Karol. 1998. Parallel evolution of glucosinolate biosynthesis inferred from congruent nuclear and plastid gene phylogenies. *American Journal of Botany* 85: 997–997.

Roelants, K., D.J. Gower, M. Wilkinson, S.P. Loader, S.D. Biju, K. Guillaume, L. Moriau, and F. Bossuyt. 2007. Global patterns of diversification in the history of modern amphibians. *Proceedings of the National Academy of Sciences, USA* 104: 887–892.

Ruhfel, B.R., M.A. Gitzendanner, P.S. Soltis, D.E. Soltis, and J.G. Burleigh. 2014. From algae to angiosperms—inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evolutionary Biology* 14: 23.

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

Sato, S., Y. Nakamura, T. Kaneko, E. Asamizu, T. Kato, M. Nakao, S. Sasamoto, et al. 2008. Genome structure of the legume, *Lotus japonicus*. *DNA Research* 15: 227–239.

Schmickl, R., A. Liston, V. Zeisek, K. Oberlander, K. Weitemier, S.C.K. Straub, R.C. Cronn, et al. 2015. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae). *Molecular Ecology Resources* 16: 1124–1135.

Schmutz, J., S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D.L. Hyten, et al.

2010. Genome sequence of the palaeopolyploid soybean. *Nature* 463: 178–183.

Schmutz, J., P.E. McClean, S. Mamidi, G.A. Wu, S.B. Cannon, J. Grimwood, J. Jenkins, et al. 2014. A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genetics* 46: 707–713.

Schneider, H., E. Schuettpelz, K.M. Pryer, R. Cranfill, S. Magallón, and R. Lupia. 2004. Ferns diversified in the shadow of angiosperms. *Nature* 428: 553–557.

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

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* 19: rspb20091176–723.

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. Morris, J. McLachlan. P. Manos, and P. S. Soltis. 2006. Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology* 15: 4261–4293.

Soltis, D.E., M.E. Mort, M. Latvis, E.V. Mavrodiev, B.C. O'Meara, P.S. Soltis, J.G. Burleigh, and R.R. de Casas. 2013. Phylogenetic relationships and character evolution analysis of Saxifragales using a supermatrix approach. *American Journal of Botany* 100: 916–929.

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.

Soltis, D.E., P.S. Soltis, M.W. Chase, M.E. Mort, D.C. ALBACH, M. Zanis, V. Savolainen, et al. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. *Botanical Journal of the Linnean Society* 133: 381–461.

Soltis, D.E., P.S. Soltis, D.R. Morgan, S.M. Swensen, B.C. Mullin, J.M. Dowd, and P.G. Martin. 1995. Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms. *Proceedings of the National Academy of Sciences, USA* 92: 2647–2651.

Soltis, D.E., P.S. Soltis, D.L. Nickrent, L.A. Johnson, W.J. Hahn, S.B. Hoot, J.A. Sweere, et

al. 1997. Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. *Annals of the Missouri Botanical Garden* 84: 1–49.

Soltis, P.S., D.E. Soltis, and M.W. Chase. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* 402: 402–404.

Stevens, P. F. 2001 onward. Angiosperm Phylogeny Website, version 14, July 2017 [and more or less continuously updated since]. Available at http://www.mobot.org/MOBOT/research/APweb.

Stokstad, E. 2016. The nitrogen fix. *Science* 353: 1225–1227.

Sun, M., R. Naeem, J.X. Su, Z.Y. Cao, J.G. Burleigh, P.S. Soltis, D.E. Soltis, and Z.-D. Chen. 2016. Phylogeny of the Rosidae: A dense taxon sampling analysis. *Journal of Systematics and Evolution* 54: 363–391.

Title, P.O., and D.L. Rabosky. 2017. Do macrophylogenies yield stable macroevolutionary inferences? An example from squamate reptiles. *Systematic Biology* 66: 843–856.

Varshney, R.K., W. Chen, Y. Li, A.K. Bharti, R.K. Saxena, J.A. Schlueter, M.T.A. Donoghue, et al. 2012. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nature Biotechnology* 30: 83–89.

Varshney, R.K., C. Song, R.K. Saxena, S. Azam, S. Yu, A.G. Sharpe, S. Cannon, et al. 2013. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nature Biotechnology* 31: 240–246.

Wang, H., M.J. Moore, P.S. Soltis, C.D. Bell, S.F. Brockington, R. Alexandre, C.C. Davis, et al. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proceedings of the National Academy of Sciences, USA* 106: 3853–3858.

Wang, Q., S.R. Manchester, H.J. Gregor, S. Shen, and Z.Y. Li. 2013. Fruits of *Koelreuteria* (Sapindaceae) from the Cenozoic throughout the northern hemisphere: Their ecological, evolutionary, and biogeographic implications. *American Journal of Botany* 100: 422–449.

Watkins, J.E. Jr, and C.L. Cardelus. 2012. Ferns in an angiosperm world: Cretaceous radiation into the epiphytic niche and diversification on the forest floor. *International Journal of Plant Sciences* 173: 695–710.

Weitemier, K., S.C.K. Straub, R.C. Cronn, M. Fishbein, R. Schmickl, A. McDonnell, and A. Liston. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant

phylogenomics. *Applications in Plant Sciences* 2: 1400042.

Werner, G.D.A., W.K. Cornwell, J.I. Sprent, J. Kattge, and E.T. Kiers. 2014. A single evolutionary innovation drives the deep evolution of symbiotic N_2-fixation in angiosperms. *Nature Communications* 5: 4087.

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.

Wilf, P. 2000. Timing the radiations of leaf beetles: hispines on gingers from latest Cretaceous to Recent. *Science* 289: 291–294.

Wing, S.L., F. Herrera, C.A. Jaramillo, C. Gómez-Navarro, P. Wilf, and C.C. Labandeira. 2009. Late Paleocene fossils from the Cerrejón Formation, Colombia, are the earliest record of Neotropical rainforest. *Proceedings of the National Academy of Sciences, USA* 106: 18627–18632.

Wu, Z., P.H. Raven, and D. Hong [eds.], 1994 onward. Flora of China, 25 vols. Missouri Botanical Garden, St. Louis, MO, USA. Available at http://www.efloras.org/flora_page.aspx?flora_id=2.

Xing, Y., R.E. Onstein, R.J. Carter, T. Stadler, and H.P. Linder. 2014. Fossils and a large molecular phylogeny show that the evolution of species richness, generic diversity, and turnover rates are disconnected. *Evolution* 68: 2821–2832.

Young, N.D., F. Debellé, G.E.D. Oldroyd, R. Geurts, S.B. Cannon, M.K. Udvardi, V.A. Benedito, et al. 2011. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature* 480: 520–524.

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.

TABLE 1. Sampling statistics for DNA data (GenBank, https://www.ncbi.nlm.nih.gov/genbank/) and occurrence data (GBIF, https://www.gbif.org/) for orders of rosids.

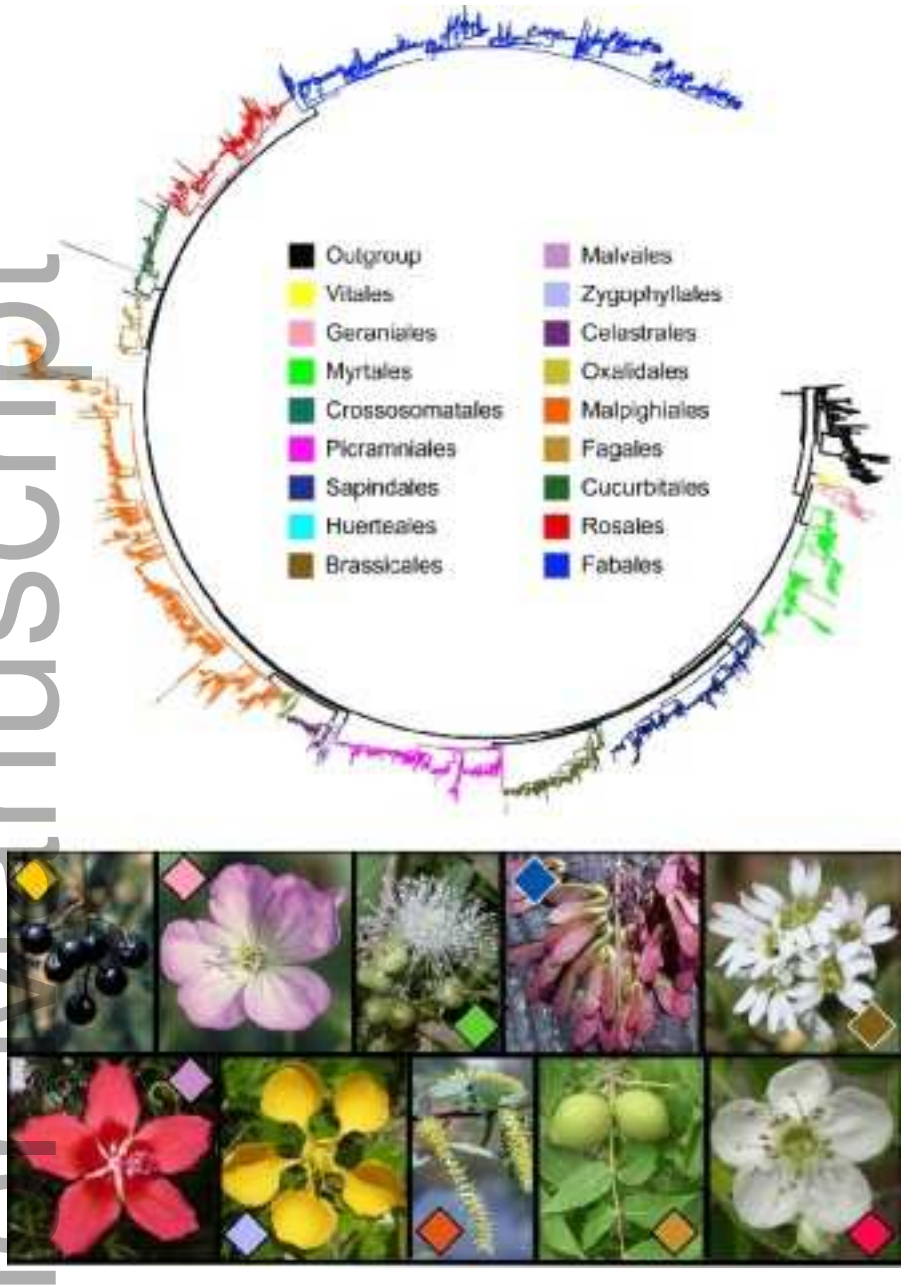
Order	GenBank DNA data		GBIF occurrence data		
	% Species sampled	% Genera sampled	Species with no records	Species with ≥10 records	Species with ≥30 records
Brassicales	38.4	92.7	19.5	55.7	35.9
Celastrales	20.4	73.4	29.4	57.3	36.0
Crossosomatales	36.4	100.0	53.7	74.2	56.1
Cucurbitales	36.8	96.9	45.2	48.4	24.5
Fabales	27.8	88.6	25.1	68.6	46.1
Fagales	50.9	100.0	11.0	97.2	68.5
Geraniales	36.4	100.0	10.7	61.4	36.8
Huerteales	29.2	100.0	53.2	16.7	12.5
Malpighiales	23.9	83.1	24.6	58.0	35.4
Malvales	21.9	76.0	30.8	50.3	30.9
Myrtales	11.5	69.2	23.4	70.7	47.6
Oxalidales	10.5	75.0	26.5	56.0	32.0
Picramniales	10.2	66.7	46.0	57.1	38.8
Rosales	22.8	87.0	15.3	73.3	48.7
Sapindales	21.0	73.7	31.5	62.1	40.2
Vitales	13.6	64.3	52.8	50.7	28.5
Zygophyllales	20.3	75.0	33.4	67.5	45.5

Notes: DNA taxon numbers were estimated by collating sequences for standard phylogenetic markers: *LEAFY*, *NIA* (nitrate reductase), ITS, ETS, 18S, 26S, *atpA*, *atpB*, *atpF*, *rbcL*, *trnL*, *matK*, *ndhF*, *ndhA*, *rpl16*, *rps16*, *ycf1*, *ycf2*, *psbA-trnH* spacer, *petB-petD* spacer, *trnC-pet1N* spacer, *trnS-trnG* spacer, *trnY-trnT* spacer, *atpB-rbcL* spacer, *trnL-trnF* intergenic spacer, *trnT-trnL* intergenic spacer. Denominators for percentage calculations come from total species estimates in Stevens (2001 onward).

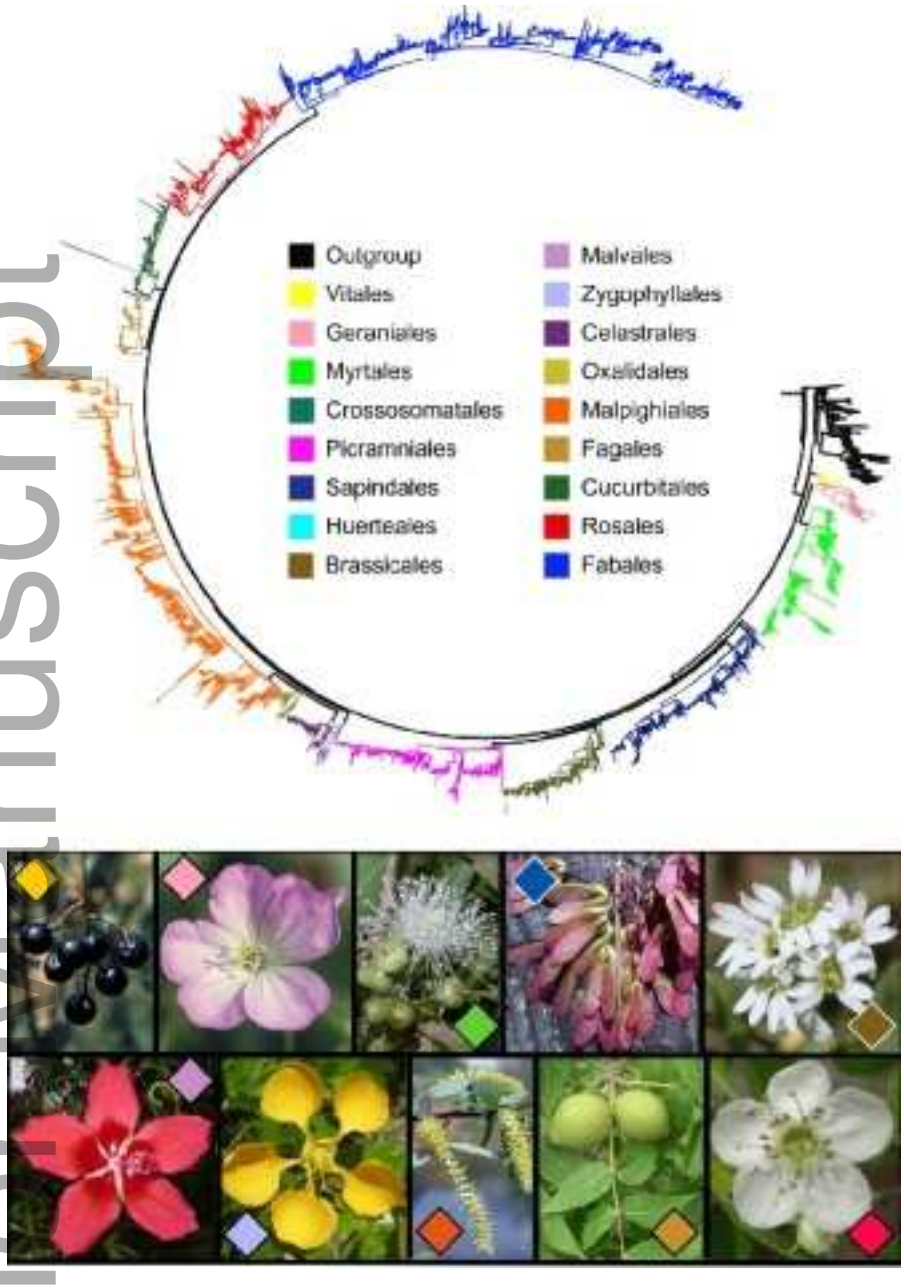
FIGURE 1. Upper panel: Summary tree for ~19,000 rosid species (four loci; Sun et al., 2016); the legend matches branch colors to recognized orders. Lower panel: Photographs of representatives of 10 familiar orders; symbols follow colors in the upper panel legend.

FIGURE 2. Phylogeny of all rosids integrating taxonomic and phylogenetic knowledge (84,153 species, from the Open Tree of Life; https://tree.opentreeoflife.org/). Branch coloration represents ordinal taxonomy and matches the legend of Fig. 1. Outer band: Species that either have (yellow) or lack (blue) phylogenetically usable data ("usable" based on taxa remaining after a series of filtering steps described by Sun et al., 2016), based on matching nomenclature with tips present in Sun et al. (2016) against the Open Tree topology (excluding Open Tree tips with labels for fossil taxa, indicating subspecific or hybrid status, etc.). Note how few taxa have data (yellow) and how phylogenetically uneven this data coverage is.

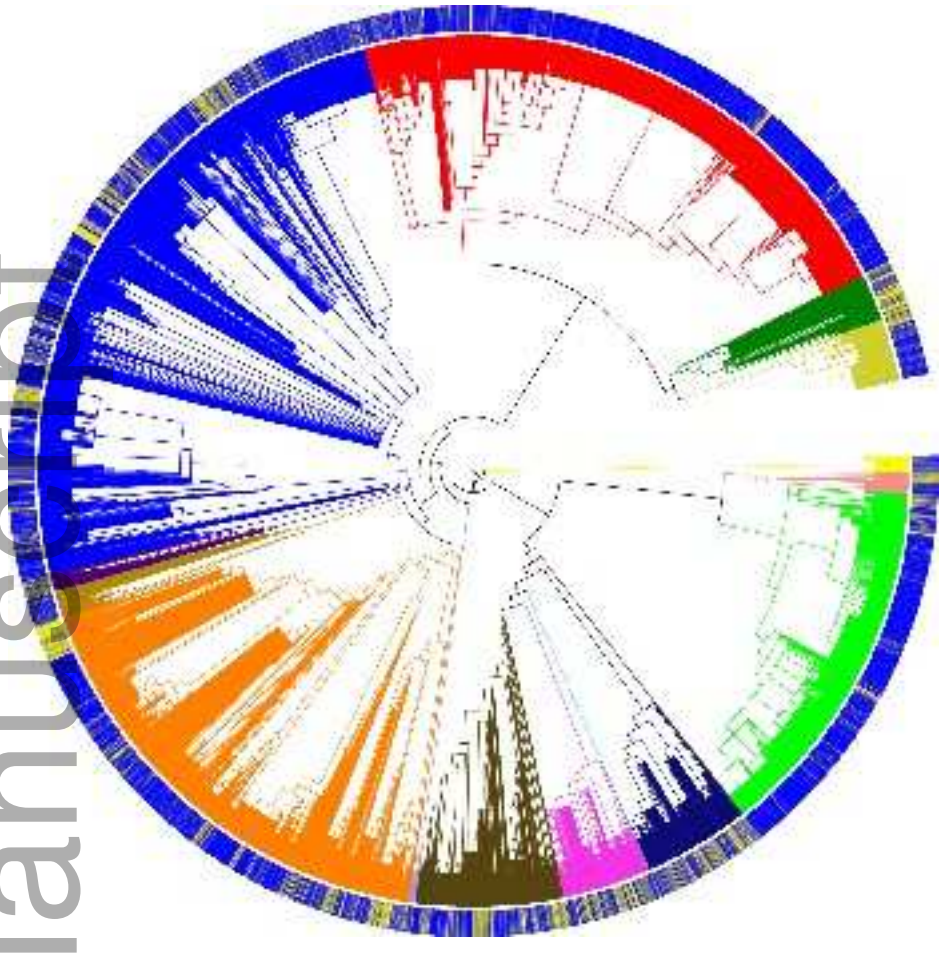
FIGURE 3. (A) Global distribution of occurrence records for the rosid clade in GBIF (https://www.gbif.org/; downloaded October 2015; 6,085,341 records), plotted on an altitude data set from R package raster. (B) Country-wise species richness, color-coded by a Jenks natural breaks classification. Species counts used country DarwinCore fields from both georeferenced and ungeoreferenced records, aggregating GBIF data with an unpublished data set of Amazonian records. The distribution of records is largely characteristic of any globally distributed clade, revealing more about global digitization effort than geographic range dynamics, while species richness estimates from available data for the rosids are close to *a priori* expectations. Projection for both maps is EPSG:4326.



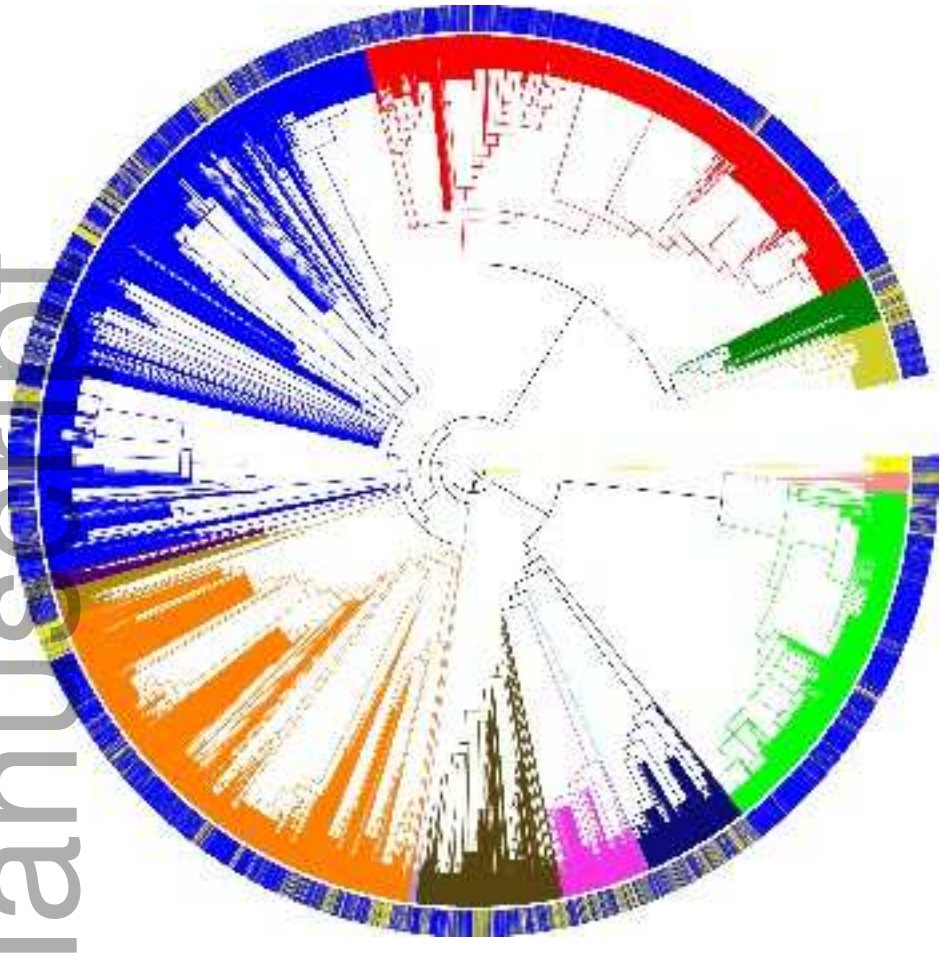
ajb2_1059_f1.png



ajb2_1059_f1.png



ajb2_1059_f2.png



ajb2_1059_f2.png