Article Type: Special Issue Article

RESEARCH ARTICLE

INVITED SPECIAL ARTICLE

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

Running head: Eiserhardt et al.—Roadmap for global synthesis of the plant tree of life

A roadmap for global synthesis of the plant tree of life

Wolf L. Eiserhardt[1,2,20], Alexandre Antonelli[3,4,5], Dominic J. Bennett[3,4,5], Laura R. Botigué[1], J. Gordon Burleigh[6], Steven Dodsworth[1], Brian J. Enquist[7,8], Félix Forest[1], Jan T. Kim[1], Alexey M. Kozlov[9], Ilia J. Leitch[1], Brian S. Maitner[7], Siavash Mirarab[10], William H. Piel[11], Oscar A. Pérez-Escobar[1], Lisa Pokorny[1], Carsten Rahbek[12,13], Brody Sandel[14], Stephen A. Smith[15], Alexandros Stamatakis[9,16], Rutger A. Vos[17,18], Tandy Warnow[19], and William J. Baker[1]

Manuscript received 13 October 2017; revision accepted 8 November 2017.

[1] Royal Botanic Gardens, Kew, TW9 3AE Richmond, Surrey, UK

[2] Department of Bioscience, Aarhus University, Ny Munkegade 116, 8000 Aarhus C, Denmark

[3] Gothenburg Global Biodiversity Centre, Box 461, 405 30, Gothenburg, Sweden

[4] Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, 405 30, Gothenburg, Sweden

[5] Gothenburg Botanical Garden, Carl Skottsbergs Gata 22B, SE-413 19, Gothenburg, Sweden

[6] Department of Biology, University of Florida, Florida 32611, USA

[7] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721 USA

[8] The Santa Fe Institute, Santa Fe, NM 87501 USA

[9] Scientific Computing Group, Heidelberg Institute for Theoretical Studies, 69118 Heidelberg, Germany

[10] Department of Electrical and Computer Engineering, University of California, San Diego, San Diego, CA 92093 USA

[11] Yale-NUS College, 16 College Avenue West, Singapore 138527, Republic of Singapore
[12] Center for Macroecology, Evolution and Climate, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen O, Denmark
[13] Imperial College London, Silwood Park, Buckhurst Road, Ascot, Berkshire SL5 7PY, UK
[14] Department of Biology, Santa Clara University, Santa Clara, CA 95053 USA
[15] Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109 USA
[16] Institute for Theoretical Informatics, Karlsruhe Institute of Technology, 76128, Karlsruhe, Germany
[17] Naturalis Biodiversity Center, P.O. Box 9517, 2300RA Leiden, The Netherlands
[18] Institute of Biology Leiden, P.O. Box 9505, 2300RA Leiden, The Netherlands
[19] Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA
[20] Author for correspondence (e-mail: w.eiserhardt@kew.org); ORCID id 0000-0002-8136-5233

Citation: Eiserhardt, W. L., A. Antonelli, D. J. Bennett, L. R. Botigué, J. G. Burleigh, S. Dodsworth, B. J. Enquist, et al. 2018. A roadmap for global synthesis of the plant tree of life. *American Journal of Botany* 105(3): XXX.
DOI: XXXX

<h1>ABSTRACT

Providing science and society with an integrated, up-to-date, high quality, open, reproducible and sustainable plant tree of life would be a huge service that is now coming within reach. However, synthesizing the growing body of DNA sequence data in the public domain and disseminating the trees to a diverse audience are often not straightforward due to numerous informatics barriers. While big synthetic plant phylogenies are being built, they remain static and become quickly outdated as new data are published and tree-building methods improve. Moreover, the body of existing phylogenetic evidence is hard to navigate and access for non-experts. We propose that our community of botanists, tree builders, and informaticians should converge on a modular framework for data integration and phylogenetic analysis, allowing easy collaboration, updating, data sourcing and flexible analyses. With support from major institutions, this pipeline should be re-run at regular intervals, storing trees and their metadata long-term. Providing the trees to a diverse global audience through user-friendly front ends and application development interfaces

should also be a priority. Interactive interfaces could be used to solicit user feedback and thus improve data quality and to coordinate the generation of new data. We conclude by outlining a number of steps that we suggest the scientific community should take to achieve global phylogenetic synthesis.

KEY WORDS: angiosperms; bryophytes; GenBank; cyberinfrastructure; land plant phylogeny; megaphylogenies; phylogenomics; phyloinformatics; pteridophytes; sampling.

The tree of life is a crucial reference system for the life sciences. It is a fundamental infrastructure of scientific knowledge that is as central to biology as the periodic table is to chemistry. Nevertheless, the tree of life remains incompletely known and insufficiently accessible to potential users. That phylogenies are fundamental to evolution and, thus, the life sciences has been recognized for decades (Hennig, 1950; Felsenstein, 1985; McTavish et al., 2017), and the demand for phylogenetic trees is higher than ever as the availability of data that can be analyzed in a phylogenetic framework soars. For example, trait and distribution data are now publicly available for tens to hundreds of thousands of species (e.g. Kattge et al., 2011; Enquist et al., 2016), facilitating very large comparative studies in evolutionary biology, biogeography, ecology, conservation, and other fields (e.g., Zanne et al., 2014). However, big data efforts in biodiversity science and the global change biology community are largely progressing without phylogenetic information (Jetz et al., 2016; Joppa et al., 2016; Proença et al., 2017). While the scientific community is finding ever more creative ways to utilize phylogenetic evidence (e.g., Strauss et al., 2006; Liu et al., 2012), access to the tree of life is still insufficient even after several decades of big tree building, and the huge contributions made by data synthesis projects like TimeTree (Kumar et al., 2017) and The Open Tree Of Life (Hinchliff et al., 2015). Thus, our ability to address research questions that can only be answered using very large phylogenetic trees remains limited (Folk et al., 2018, in this issue).

The plant phylogenetic community has been highly collaborative and productive over the last three decades. The major branches of the land plant tree of life are now generally well established, although some problematic nodes remain (Ruhfel et al., 2014; Wickett et al., 2014; PPG I, 2016; Angiosperm Phylogeny Group, 2016; Gitzendanner et al., 2018, in this issue). Public databases such as NCBI GenBank contain at least some DNA data from 27% of known vascular plant species and 75% of genera (Hinchliff and Smith, 2014; RBG Kew, 2016). However, the extent to which these data can resolve well-supported phylogenetic relationships has been questioned

(Hinchliff and Smith, 2014). Moreover, the most commonly sequenced loci represent a minuscule fraction of the total information in plant genomes, with land plant nuclear genomes ranging in size from ca. 61 million to 149 billion base pairs (Dodsworth et al., 2015). As of January 2017, only 225 vascular plant genomes had been published, equivalent to <0.1% of land plant diversity (RBG Kew, 2017). The gap between actually and potentially available DNA sequence data for plants is thus immense.

More insidiously, public sequence data are plagued by serious data quality concerns (e.g., Nilsson et al., 2006). For example, species names are often incorrectly spelled or, worse, taxonomically incorrect. The problem is exacerbated as listed species names often are not linked to vouchers (Gratton et al., 2017). In addition, species nomenclature does not keep pace with taxonomic updates. Together, these issues point to the fact that data quality control is a central challenge in the provision of an accurate plant tree of life.

Several new projects are now rising to the challenge of filling the data gaps through high-throughput genomic sequencing across the plants. For example, the Plant and Fungal Trees of Life Project (PAFTOL) and Genealogy of Flagellate Plant Project (GoFlag) together aim to analyze hundreds of nuclear genes and plastid genomes from all genera and many species of land plants using a gene capture approach (Weitemier et al., 2014). Large whole-genome projects such as the Open Green Genomes Project and the 10,000 Plants Project (10KP: Normile, 2017) are also underway, which build on the recent success of the 1,000 Plants Project (Wickett et al., 2014). In different ways, these initiatives promise to deliver extraordinary new resources for plant comparative biology. However, together, they will tackle less than 10% of the known species diversity of land plants, presenting a fundamental limitation to the usefulness of the phylogenies resulting from them. While complete genome sequencing of all species of life on Earth is a stated ambition of the scientific community (Pennisi, 2017), the results may not be realized for many years to come. It is essential, therefore, that all available data, whether from public databases or new genomic initiatives, are integrated to deliver the best possible estimate of the plant tree of life at any given time.

The idea to generate synthetic phylogenies that combine all available phylogenetic evidence is not new. For example, The Open Tree of Life and related AVATOL projects were herculean efforts to synthesize and facilitate the analysis of the entire tree of life (Hinchliff et al. 2015). These projects resulted in several resources that continue to be useful and will continue to be updated

(e.g., data store, taxonomy, synthetic tree, online tree viewer). For plants, important synthetic trees of life have been built through mining and compiling both public DNA sequence data (e.g., Hinchliff and Smith, 2014; Zanne et al., 2014; Maitner et al., 2018), published phylogenies (Hinchliff et al., 2015), or a combination of both (Smith and Brown, 2018, in this issue). While these trees have facilitated many analyses, each is limited in some respect. For example, despite the ever-increasing rate at which DNA sequence data are generated, these synthetic trees are not routinely updated and thus become quickly outdated. Moreover, these phylogenies often fail to capture the uncertainty and conflict underlying the data that has now been exposed by large genomic analyses (Wickett et al., 2014; Shen et al., 2017). Thus, the users of the plant tree of life are obliged either to choose an existing tree, regardless of its deficiencies, or to build their own tree by mining public repositories and reconstructing phylogenetic relationships themselves. Despite the creation of new pipelines (e.g., Antonelli et al., 2017; Smith and Brown, 2018, this issue), the latter option remains beyond the skills and desires of many potential users.

We believe that the plant phylogenetic community must find new ways to provide an integrated, up-to-date, high quality, open, reproducible and sustainable tree (Table 1) to a diverse user community. Here we propose a roadmap that outlines how our community could produce such a tree, focusing on the synthesis of all publically available DNA sequence data. We argue that we need a modular tree of life pipeline that allows distributed development of tools across research groups. We find it useful to break down this pipeline into four main parts (Fig. 1): gathering the data, phylogenetic reconstruction, data storage, and disseminating the tree of life. Below, we outline the major challenges and opportunities associated with each part and conclude with a call to action, proposing nine steps that we think would materially advance our quest for global phylogenetic synthesis in plants. We note that the case study here focuses on plants, but the principles could apply to any group of organisms or even all of life.

<h1>GATHERING THE DATA

Constructing accurate and comprehensive phylogenies for extant plants requires comprehensive molecular sampling. Despite herculean efforts by thousands of scientists over the last decades to collect molecular data across the tree of life, there are still major data gaps (Fig. 2). Not only do we lack molecular data for approximately 285,000 of the 391,000 known species of vascular plants (RBG Kew, 2016), but also there is poor genomic coverage for most species for which we do have data. Nevertheless, available molecular resources are immense and continue to grow

rapidly in size and complexity: the NCBI database currently contains almost 38 million nucleotide sequences for land plants, yet the challenge lies in the computational demand of handling these data volumes. For example, all-versus-all BLAST searching and clustering, a critical step in homology and orthology assessment, becomes computationally prohibitive as data increase. Moreover, data integration becomes more complex as the number of databases increases, bringing different schemas and interfaces. More importantly, we must now also adapt to diversifying data types, such as single loci, transcriptomes, genomes, and restriction-site-associated DNA sequencing (RADSeq) data. Despite these challenges, there have been significant advances in data set assembly that have addressed some of the complexity associated with genomic and transcriptomic data (Dunn et al., 2013; Yang and Smith, 2014; Walker et al., 2018, in this issue). Researchers can leverage these recent developments along with advances in large data set construction (Freyman, 2015; Antonelli et al., 2017; Smith and Brown, 2018, in this issue) to overcome the challenges faced by diverse and large data sources.

In addition to the computational and biological complexities that accompany diverse data, significant concerns surround data quality in public databases, such as contamination, lack of sequence validation, and a dearth of links to specimens. The identification of mislabeled or contaminant sequences is an important yet difficult cleaning step that can now be facilitated by semi-automated methods (e.g., Kozlov et al., 2016; Rulik et al., in press). In addition, a public record of questionable sequences in GenBank is starting to emerge (e.g., https://github.com/FePhyFoFum/seq_filters). Ideally, this information would be stored together with the sequence data, but such storage is not currently possible given the limitations of public databases. Community-curated reference sequence databases have been successfully implemented by other communities, e.g., for fungal ITS (Kõljalg et al., 2005), protist 18S rDNA (Berney et al., 2017), and bacterial genomes (Chen et al., 2017), and a similar resource would be invaluable for plants.

Taxonomic reconciliation is yet another significant challenge that emerges when integrating species data from multiple sources. For example, whereas molecular databases such as GenBank use the NCBI taxonomy, trait databases (e.g., BIEN) and geographical archives (e.g., GBIF) may use other taxonomies. Each of these recognizes their own sets of synonyms, alternative spellings, and taxon concepts. Taxonomic reconciliation is the process of navigating this heterogeneity for purposes of data integration. Several web services (e.g. iPlant TNRS, GlobalNames, TaxoSaurus) and “meta-taxonomies” (e.g., the Open Tree of Life taxonomy) exist to support this process (Rees

and Cranston, 2017). Nevertheless, a modular infrastructure for periodically rebuilding the plant tree of life, as proposed here, would benefit from a pre-computed taxonomic mapping of input data sources, which would be both a more efficient approach than accessing web resources each time, and a community-based product that can itself be released, critiqued, corrected, and annotated.

Looking forward, the plant phylogenetics community can partly preempt data integration problems by converging on common sets of molecular loci, thus maximizing overlap among data sets. Such convergence has happened in the past, when a small set of loci (e.g., *rbcL*, *matK*, ITS) was widely sequenced and used for phylogenetic reconstruction and barcoding (CBOL Plant Working Group, 2009). These loci facilitated large phylogenetic analyses that spanned all plants, but we now know that, for several reasons, additional data sets are needed. For example, genomic analyses have exposed the underlying complexity of phylogenetic conflict, concordance, and gene and genome duplication (Jarvis et al., 2014; Wickett et al., 2014; Shen et al., 2017). Our data collection strategies need to reflect the reality of these patterns and processes. Common loci have yet to emerge for the genomic age: for example, recently developed marker sets for Asteraceae, Arecaceae and Detarioideae (Mandel et al., 2014; Heyduk et al., 2016; M. de la Estrella, Royal Botanic Gardens, Kew, unpublished data), each containing hundreds of markers, only have five loci in common. However, initiatives like PAFTOL and GoFlag are now developing toolkits that will isolate a defined set of several hundred orthologous loci across land plants. Data generated in this way could play a similar role in the future that *rbcL* and other popular loci have done in the past, but one that reflects the lessons we have gained from analyzing genomes and transcriptomes over the last decade.

<h1>PHYLOGENETIC RECONSTRUCTION

Any phylogenetic analysis at the scale of the plant tree of life will challenge standard approaches for multiple sequence alignment and phylogenetic inference. As the number of species and/or genes increases, the accuracy of likelihood-based phylogenetic methods can decrease, in particular when more taxa but not more genes are added. Meanwhile, running times will always increase with increasing data. As a concrete example, concatenation analyses using maximum likelihood (ML) are the most common approach for species tree estimation, and existing parallel implementations (e.g., Kozlov et al., 2015; Nguyen et al., 2015) can analyse data sets comprising dozens to hundreds of whole genomes or transcriptomes (Jarvis et al., 2014; Peters et al., 2017).

However, no current ML method scales in reasonable time to enable analyses of data sets with tens of thousands of species *and* loci. For example, inferring a tree on 1600 insect transcriptomes (including bootstraps) would still take an estimated 70 million CPU hours. The development of ever more efficient and accurate methods for multiple sequence alignment and phylogeny estimation is driven by the "arms race" between the rapidly growing sequencing capacity on the one side and computational capacity and phylogenetic algorithms on the other side.

The biological realism of phylogenetic models (e.g., models of sequence evolution) is another important challenge to accurate phylogenetic reconstruction. Perhaps most importantly, recent genomic and transcriptomic studies (e.g., Wickett et al., 2014; Sun et al., 2015; Shen et al., 2017) have exposed considerable amounts of gene tree discordance that need to be modeled appropriately. Discordance had typically been considered to be the result of noise and error, but these new data suggest that widespread discordance is likely due, at least in part, to biological processes (e.g., incomplete lineage sorting, hybridization, gene duplication and loss). This challenge is being addressed by species tree methods, which is an area of rapid methodological development (e.g., Ané et al., 2007; Liu et al., 2007; Heled and Drummond, 2010; Boussau et al., 2013; Chifman and Kubatko, 2014; Mirarab et al., 2014). In spite of these promising advances, several problems remain. Most species tree methods only address a single source of discordance, and some sources remain difficult to address, such as hybridization and allopolyploid speciation (but see Yu et al., 2014; Yu and Nakhleh, 2015; Solís-Lemus and Ané, 2016), which are particularly frequent in plants (Wood et al., 2009; Van de Peer et al., 2017). In addition, it is not known how accurate species tree approaches are for large numbers of taxa, although some methods now scale to 10,000 species (Zhang et al., 2017). Also, while it may be difficult to reconstruct reliable gene trees due to lack of phylogenetic signal, techniques such as weighted statistical binning can be helpful (Bayzid et al., 2014; Mirarab et al., 2014), though additional developments that address this problem may be necessary. In addition to discordance, heterogeneity in the process of molecular evolution (e.g., lineage specific rate shifts, compositional evolution) may also complicate phylogenetic reconstruction (Li et al., 2014; De La Torre et al., 2017). Researchers continue to address this complexity and comprehensive phylogenetic reconstruction of plants should incorporate these developments where possible (Foster et al., 2009; Cox et al., 2014).

Missing data are a notorious feature of phylogenetic analyses that synthesize partly overlapping data from multiple sources, i.e., not all loci are sampled for all taxa. Such analyses may be

susceptible to errors or analytical issues associated with missing data (e.g., Sanderson et al., 2015). Projects such as PAFTOL and GoFlag that are expanding the number of orthologous regions sequenced, in addition to continuing genomic and transcriptomic efforts, will, at least in part, address this problem. However, methodological developments that tackle phylogenetic reconstruction with a "divide and conquer" approach may also overcome these issues by reducing the phylogenetic problem to data matrices that have less missing data (e.g., Smith and Brown, 2018, in this issue. These methods can then be combined with other developments in supertree construction to graft these subtrees into a comprehensive tree (Akanni et al., 2015; Lafond et al., in press; Redelings and Holder, 2017; Vachaspati and Warnow, 2017).

Many of the phylogenetic challenges that face the reconstruction of a comprehensive plant tree will require new developments in phylogenetic methods, but are common to the reconstruction of other parts of the tree of life. The alignments and data sets compiled as part of an effort to construct a comprehensive plant phylogeny would serve the phylogenetics community in driving the development of new methods. These new methods could then be used to reconstruct a more accurate and useful comprehensive plant phylogeny.

<h1>DATA STORAGE

Assembling the tree of life is fundamentally a *big data* problem: not only does it produce large quantities of results in an iterative process, but each data object produced is large and complex. Consider that if the tree of all plant species were oriented horizontally and the species labels printed in 9-point font, the tree would extend twice the height of the tallest human-made structure in the world, the Burj Khalifa in Dubai (i.e., 830 m). Thus, not only is it a challenge to manage each iteration of the pipeline, but also the trees themselves are too big for any kind of meaningful visual inspection as a whole. Furthermore, multiple sequence alignments are even larger than the trees. Also, given the wide-ranging set of techniques and data sets available for phylogenetic reconstruction, there will likely be multiple alternative resolutions for many parts of the plant tree of life. To help users of phylogenetic trees to make sense of such discordances requires effective ways of storing, comparing, and summarizing alternative resolutions. For efficient management, quality control, and data output, we require a scalable database, designed and optimized for the purpose.

Fundamentally, the database module of a tree of life pipeline is responsible for tracking the provenance of input data, alignments, metadata about the analysis, and phylogenetic results, and

is also essential for ensuring transparency and reproducibility (Leebens-Mack et al., 2006). A key challenge is to establish the appropriate balance between allowing flexibility, and thereby future-proofing the assembly pipeline, while on the other hand fully normalizing the data model to provide data integrity and query efficiency for core components (McTavish et al., 2015). The Open Tree of Life uses a git-based system for tree storage, called Phylesystem (McTavish et al., 2015). This system allows for versioning and metadata to be attached. Furthermore, it allows for easy replication by other researchers. This provides a potential model for future decentralized databasing projects.

Importantly, a database for storing phylogenetic trees must not be developed in isolation. The demand to combine phylogenetic information with additional biological and abiotic data is increasing, and any tree of life database should thus be compatible with global common data standards (Panahiazar et al., 2013), allowing links to initiatives that deliver, for example, plant distribution or trait data (e.g. Kattge et al., 2011; Enquist et al., 2016; Maitner et al., 2018).

<h1>DISSEMINATING THE TREE OF LIFE

The use of phylogenetic information is crucial for solving pure and applied problems in biology (Brooks and McLennan, 1991; Faith, 1992; Magurran, 2013) and has enormous potential for outreach and education (Jenkins, 2009; MacDonald and Wiley, 2012). Thus, a central challenge for developing a phylogenetic workflow and serving big trees is to anticipate correctly a plethora of use cases (see Box 1) and to develop a general cyberinfrastructure accordingly (Goff et al., 2011; Stoltzfus et al., 2013). As outlined above, this flexibility relies on an appropriate database structure, but the actual user interface is equally important.

Publicly depositing phylogenetic trees in an editable electronic format is largely standard practice nowadays (but see Stoltzfus et al., 2012; Drew et al., 2013), allowing researchers to access a wealth of phylogenetic information online (e.g., https://treebase.org/, https://tree.opentreeoflife.org/). Online storage would be particularly important for frequently updated trees that might not be associated with a traditional, static publication. In this instance, proper versioning is essential, and care must be taken that each version of the tree is citable (e.g., using a digital object identifier). If alternative phylogenetic methods were employed, the user should be enabled to make an informed choice about the different resulting trees. Special care must also be taken to communicate uncertainty (e.g., support values) in an understandable way. It should be noted that trees stored in databases such as TreeBASE (Piel et al., 2009) are not

necessarily readily navigated by non-expert audiences, and more accessible interfaces can greatly increase the impact (e.g., OneZoom: Rosindell and Harmon, 2012; and the Open Tree of Life).

In addition to an easily accessible means for interacting with the tree or set of trees, any associated metadata need to be available. For example, sequence metadata (e.g., voucher, reference), including both data stored in the repositories that the sequences were obtained from, and data that cannot be stored in such repositories (e.g., digital images of voucher specimens) should be linked and made available where possible. This information contributes to future-proofing the tree, as for example, taxonomic changes can be applied retrospectively, and errors can be rectified. More generally, users conducting phylogenetic analyses often discover issues with particular sequences, such as probable misidentifications, unlikely divergent sequences within species, and overly short, long, or gappy sequences. There should be a mechanism allowing users to highlight issues with the database in terms of sequences, alignments, or tree errors. The Open Tree of Life interface allows for the curation and comment of input trees and data sources as well as the synthetic tree (Hinchliff et al., 2015). This functionality could be expanded to include more specific information about alignments and sequences.

If presented in an appropriate way, a synthetic plant tree of life has the potential to make the generation of new data more efficient by highlighting clades and regions that should be prioritized to increase total phylogenetic sampling. For example, the Open Tree of Life synthetic tree browser allows users to explore which primary phylogenetic studies any edge is derived from. While currently only implemented in a supertree framework, this approach could be extended to sequence data. We envision a dynamic interface where users can easily identify clades and regions that are poorly sampled taxonomically and/or genetically. Such an interface should show where species are missing, as well as reflect the amount of data underpinning the inferred relationships (Hinchliff et al., 2015). The interface could also allow users to annotate planned sequencing efforts, i.e., which taxa and loci they plan to sequence, when, where, and contact information for the project. This way, unnecessary duplication of work could be reduced, scientific collaboration increased, and logistics associated with fieldwork and permit applications facilitated.

Besides viewing and downloading the entire tree, perhaps the most central need is to provide tools to extract custom subtrees from the plant tree of life, based on a list of taxa of relevance to a specific research context. Methods such as Phylomatic (Webb and Donoghue, 2005) and

Phylotastic (Stoltzfus et al., 2013) have already demonstrated the broad interest in such an application. Easy access to custom subtrees would require tools and algorithms to generate partial views of user-defined regions of larger trees. Importantly, such tools would need to include a service for name reconciliation (e.g., Boyle et al., 2013), allowing for taxonomic differences between the user input and the tree.

Although some generic uses are readily anticipated, perhaps the most important way of serving the plant tree of life is through flexible software interfaces. For example, integration with the R (https://www.r-project.org/) or Biopython (http://biopython.org/) software environments would allow the plant tree of life to be used in a wide range of biostatistics and bioinformatics applications. More generally, the development of application programming interfaces (APIs) is essential for ensuring a wide use of the tree, which could range from websites and educational apps to stand-alone software. APIs allow external users to formally query and download data, opening the door to an almost unlimited number of uses.

<h1>CONCLUSIONS AND CALL TO ACTION

Providing science and society with an integrated, up-to-date, high quality, open, reproducible and sustainable plant tree of life would be a huge service that is coming within reach. Technological and methodological advances have paved the way for this synthesis, but putting it into practice requires a concerted effort by the scientific community. Here, we call on the community to embrace the following actions, which would materially advance our quest for global phylogenetic synthesis in plants:

1. Unite behind the collective goal of an integrated, up-to-date, high quality, open, reproducible and sustainable tree of life for plants (Table 1).
2. Agree on an open framework for a tree of life pipeline with discrete, interchangeable modules, drawing on the wealth of existing tools (Fig. 1).
3. Encourage computer scientists and software developers to address priority analytical problems requiring innovative solutions.
4. Commit to computing trees at regular intervals (e.g., yearly, monthly), ensuring that an up-to-date plant tree of life is always available.
5. Establish a sustainable infrastructure for long-term storage and distribution of the resulting trees and associated metadata.

6. Create web tools that allow trees to be easily explored, queried, and downloaded by diverse audiences, ranging from experts to school children.
7. Create application programming interfaces (API) that allow trees to be integrated in external software.
8. Engineer a mechanism for community feedback on data quality, which also feeds back to the original public source (e.g., NCBI GenBank).
9. Provide a mechanism for identifying and prioritizing knowledge gaps through dynamic cross-matching trees with public data sets.

In this call to action, we emphasize the importance of community coordination and institutional responsibility. Building and maintaining pipelines that perform optimally at all steps discussed in this paper is beyond the skills and resources of most individual research labs. Similarly, within the constraints of standard research grants, a firm commitment to regular tree updates, indeterminate storage of trees and metadata, and actively maintained interfaces is near impossible. Thus, we need to build a collaborative, community-driven platform that allows many individuals, groups, and institutions to contribute according to their scientific strengths and resources. The recently founded PhyloSynth network (https://phylosynth.github.io/) aims to facilitate the development of such a platform, paving the way toward an integrated, up-to-date, high quality, open, reproducible and sustainable tree of life for plants. By embracing this call to action, our community would extend its impact beyond the ivory tower of pure comparative plant biology research, broadening its societal reach and bringing tree of life research to bear on the global challenges facing humanity today.

<h1>ACKNOWLEDGEMENTS

The authors thank Douglas E. Soltis and two anonymous reviewers for helpful feedback on the manuscript and Olivier Maurin, Tuula Niskanen, Beata Klejevskaja, and William Pearse for thoughtful discussion. This work was partly supported by grants from the Calleva Foundation, the Garfield Weston Foundation and the Sackler Trust to the Royal Botanic Gardens, Kew. Part of this work was funded by the Klaus Tschira Foundation to A.S.; U.S. National Science Foundation grant ABI-1458652 to T.W.; Yale-NUS grants IG15-SI101 and R-607-265-200-121 to W.H.P.

<h1>LITERATURE CITED

Akanni, W. A., M. Wilkinson, C. J. Creevey, P. G. Foster, and D. Pisani. 2015. Implementing and testing Bayesian and maximum-likelihood supertree methods in phylogenetics. *Royal Society Open Science* 2: 140436.

Ané, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. 2007. Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution* 24: 412-426.

Angiosperm Phylogeny Group. 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.

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.

Bayzid, M. S., T. Hunt, and T. Warnow. 2014. Disk covering methods improve phylogenomic analyses. *BMC Genomics* 15(supplement 6): S7.

Berney, C., A. Ciuprina, S. Bender, J. Brodie, V. Edgcomb, E. Kim, J. Rajan, et al. 2017. UniEuk: time to speak a common language in protistology! *Journal of Eukaryotic Microbiology* 64: 407-411.

Boussau, B., G. J. Szöllosi, L. Duret, M. Gouy, E. Tannier, and V. Daubin. 2013. Genome-scale coestimation of species and gene trees. *Genome Research* 23: 323-330.

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.

Brooks, D. R., and D. A. Mclennan. 1991. Phylogeny, ecology, and behavior: a research program in comparative biology. University of Chicago Press, Chicago, IL, USA.

CBOL Plant Working Group. 2009. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences, USA* 106: 12794-12797.

Chen, I. M. A., V. M. Markowitz, K. Chu, K. Palaniappan, E. Szeto, M. Pillay, A. Ratner, et al. 2017. IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Research* 45: D507-D516.

Chifman, J., and L. Kubatko. 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30: 3317-3324.

Cox, C. J., B. Li, P. G. Foster, T. M. Embley, and P. Civán. 2014. Conflicting phylogenies for early land plants are caused by composition biases among synonymous substitutions. *Systematic Biology* 63: 272-279.

De La Torre, A. R., Z. Li, Y. Van De Peer, and P. K. Ingvarsson. 2017. Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. *Molecular Biology and Evolution* 34: 1363-1377.

Dodsworth, S., A. R. Leitch, and I. J. Leitch. 2015. Genome size diversity in angiosperms and its influence on gene space. *Current Opinion in Genetics and Development* 35: 73-78.

Drew, B. T., R. Gazis, P. Cabezas, K. S. Swithers, J. Deng, R. Rodriguez, L. A. Katz, et al. 2013. Lost branches on the tree of life. *PLoS Biology* 11: e1001636.

Dunn, C. W., M. Howison, and F. Zapata. 2013. Agalma: an automated phylogenomics workflow. *BMC Bioinformatics* 14: 330.

Enquist, B. J., R. Condit, R. K. Peet, M. Schildhauer, and B. M. Thiers. 2016. Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity. *PeerJ Preprints* e2615v2.

Faith, D. P. 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation* 61: 1-10.

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

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

Foster, P. G., C. J. Cox, and T. M. Embley. 2009. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Philosophical Transactions of the Royal Society of London, B, Biological Sciences* 364: 2197-2207.

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

Goff, S. A., M. Vaughn, S. Mckay, E. Lyons, A. E. Stapleton, D. Gessler, N. Matasci, et al. 2011. The iPlant Collaborative: cyberinfrastructure for plant biology. *Frontiers in Plant Science* 2: 34.

Gratton, P., S. Marta, G. Bocksberger, M. Winter, E. Trucchi, and H. Kühl. 2017. A world of sequences: Can we use georeferenced nucleotide databases for a robust automated phylogeography? *Journal of Biogeography* 44: 475-486.

Heled, J., and A. J. Drummond. 2010. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* 27: 570-580.

Hennig, W. 1950. Grundzüge einer Theorie der phylogenetischen Systematik. Deutscher Zentralverlag, Berlin, Germany.

Heyduk, K., D. W. Trapnell, C. F. Barrett, and J. Leebens-Mack. 2016. Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. *Biological Journal of the Linnean Society of London* 117: 106-120.

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.

Jarvis, E. D., S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li, S. Y. W. Ho, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346: 1320-1331.

Jenkins, K. P. 2009. Evolution in biology education: sparking imaginations and supporting learning. *Evolution: Education and Outreach* 2: 347-348.

Jetz, W., J. Cavender-Bares, R. Pavlick, D. Schimel, F. W. Davis, G. P. Asner, R. Guralnick, et al. 2016. Monitoring plant functional diversity from space. *Nature Plants* 2: 16024.

Joppa, L. N., B. O'Connor, P. Visconti, C. Smith, J. Geldmann, M. Hoffmann, J. E. M. Watson, et al. 2016. Big data and biodiversity. Filling in biodiversity threat gaps. *Science* 352: 416-418.

Kattge, J., S. Díaz, S. Lavorel, I. C. Prentice, P. Leadley, G. Bönisch, E. Garnier, et al. 2011. TRY – a global database of plant traits. *Global Change Biology* 17: 2905-2935.

Kembel, S. W., P.D. Cowan, M.R. Helmus, W.K. Cornwell, H. Morlon, D.D. Ackerly, S.P. Blomberg, and C.O. Webb. 2010. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 26:1463-1464.

Kõljalg, U., K.-H. Larsson, K. Abarenkov, R. H. Nilsson, I. J. Alexander, U. Eberhardt, S. Erland, et al. 2005. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist* 166: 1063-1068.

Kozlov, A. M., A. J. Aberer, and A. Stamatakis. 2015. ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics* 31: 2577-2579.

Kozlov, A. M., J. Zhang, P. Yilmaz, F. O. Glöckner, and A. Stamatakis. 2016. Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Research* 44: 5022-5033.

Kumar, S., G. Stecher, M. Suleski, and S. B. Hedges. 2017. TimeTree: A resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution* 34: 1812-1819.

Laffan, S.W., E. Lubarsky, and D. F. Rosauer. 2010. Biodiverse, a tool for the spatial analysis of biological and related diversity. *Ecography*: 33: 643-647.

Lafond, M., C. Chauve, N. El-Mabrouk, and A. Ouangraoua. 2017. Gene tree construction and correction using supertree and reconciliation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, early online, doi 10.1109/TCBB.2017.2720581.

Leebens-Mack, J., T. Vision, E. Brenner, J. E. Bowers, S. Cannon, M. J. Clement, C. W. Cunningham, et al. 2006. Taking the first steps towards a standard for reporting on phylogenies: Minimum Information About a Phylogenetic Analysis (MIAPA). *OMICS* 10: 231-237.

Li, B., J. S. Lopes, P. G. Foster, T. M. Embley, and C. J. Cox. 2014. Compositional biases among synonymous substitutions cause conflict between gene and protein trees for plastid origins. *Molecular Biology and Evolution* 31: 1697-1709.

Liu, L., D. K. Pearl, and T. Buckley. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology* 56: 504-514.

Liu, X., M. Liang, R. S. Etienne, Y. Wang, C. Staehelin, and S. Yu. 2012. Experimental evidence for a phylogenetic Janzen–Connell effect in a subtropical forest. *Ecology Letters* 15: 111-118.

MacDonald, T., and E. O. Wiley. 2012. Communicating phylogeny: evolutionary tree diagrams in museums. *Evolution: Education and Outreach* 5: 14-28.

Magurran, A. E. 2013. Measuring biological diversity. John Wiley, Chichester, UK.

Maitner, B. S., B. Boyle, N. Casler, R. Condit, J. Donoghue, S. M. Durán, D. Guaderrama, et al. 2018. The bien r package: A tool to access the Botanical Information and Ecology Network (BIEN) database. *Methods in Ecology and Evolution* 9: 373-379.

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.

McTavish, E. J., B. T. Drew, B. Redelings, and K. A. Cranston. 2017. How and why to build a unified tree of life. *BioEssays* 39: 1700114.

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.

Mirarab, S., R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30: i541-548.

Nguyen, L.-T., H. A. Schmidt, A. Von Haeseler, and B. Q. Minh. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32: 268-274.

Nilsson, R. H., M. Ryberg, E. Kristiansson, K. Abarenkov, K.-H. Larsson, and U. Kõljalg. 2006. Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLoS One* 1: e59.

Normile, D. 2017. Plant scientists plan massive effort to sequence 10,000 genomes. Website http://www.sciencemag.org/news/2017/07/plant-scientists-plan-massive-effort-sequence-10000-genomes.

Panahiazar, M., A. P. Sheth, A. Ranabahu, R. A. Vos, and J. Leebens-Mack. 2013. Advancing data reuse in phyloinformatics using an ontology-driven Semantic Web approach. *BMC Medical Genomics* 6: S5.

Pennisi, E. 2017. Biologists propose to sequence the DNA of all life on Earth. Website http://www.sciencemag.org/news/2017/02/biologists-propose-sequence-dna-all-life-earth.

Peters, R. S., L. Krogmann, C. Mayer, A. Donath, S. Gunkel, K. Meusemann, A. Kozlov, et al. 2017. Evolutionary history of the hymenoptera. *Current Biology* 27: 1013-1018.

Piel, W., L. Chan, M. Dominus, J. Ruan, R. Vos, and V. Tannen. 2009. TreeBASE v. 2: a database of phylogenetic knowledge.

PPG I. 2016. A community-derived classification for extant lycophytes and ferns. *Journal of Systematics and Evolution* 54: 563-603.

Proença, V., L. J. Martin, H. M. Pereira, M. Fernandez, L. McRae, J. Belnap, M. Böhm, et al. 2017. Global biodiversity monitoring: from data sources to essential biodiversity variables. *Biological Conservation* 213: 256-263.

RBG Kew. 2016. The state of the world's plants report 2016. Royal Botanic Gardens, Kew, Richmond, Surrey, UK. Available at https://stateoftheworldsplants.com/2016/.

RBG Kew. 2017. The state of the world's plants report 2017. Royal Botanic Gardens, Kew, Richmond, Surrey, UK. Available at https://stateoftheworldsplants.com/2017/.

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. A., and K. Cranston. 2017. Automated assembly of a reference taxonomy for phylogenetic data synthesis. *Biodiversity Data Journal* 5: e12581.

Rosindell, J., and L. J. Harmon. 2012. OneZoom: a fractal explorer for the tree of life. *PLoS Biology* 10: e1001406.

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.

Rulik, B., J. Eberle, L. Von Der Mark, J. Thormann, M. Jung, F. Köhler, W. Apfel, et al. 2017. Using taxonomic consistency with semi-automated data pre-processing for high quality DNA barcodes. *Methods in Ecology and Evolution* 8: 1878-1887.

Sanderson, M. J., M. M. McMahon, A. Stamatakis, D. J. Zwickl, and M. Steel. 2015. Impacts of terraces on phylogenetic inference. *Systematic Biology* 64: 709-726.

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 and Evolution* 1: 126.

Smith, S. A., and J. W. Brown. 2018. Constructing a comprehensive seed plant phylogeny. *American Journal of Botany* 105 (in press).

Solís-Lemus, C., and C. Ané. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genetics* 12: e1005896.

Stoltzfus, A., B. O'Meara, J. Whitacre, R. Mounce, E. L. Gillespie, S. Kumar, D. F. Rosauer, and R. A. Vos. 2012. Sharing and re-use of phylogenetic trees (and associated data) to facilitate synthesis. *BMC Research Notes* 5: 574.

Stoltzfus, A., H. Lapp, N. Matasci, H. Deus, B. Sidlauskas, C. M. Zmasek, G. Vaidya, et al. 2013. Phylotastic! Making tree-of-life knowledge accessible, reusable and convenient. *BMC Bioinformatics* 14: 158.

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

Sun, M., D. E. Soltis, P. S. Soltis, X. Zhu, J. G. Burleigh, and Z. Chen. 2015. Deep phylogenetic incongruence in the angiosperm clade Rosidae. *Molecular Phylogenetics and Evolution* 83: 156-166.

Vachaspati, P., and T. Warnow. 2017. FastRFS: fast and accurate Robinson–Foulds Supertrees using constrained exact optimization. *Bioinformatics* 33: 631-639.

Van De Peer, Y., E. Mizrachi, and K. Marchal. 2017. The evolutionary significance of polyploidy. *Nature Reviews Genetics* 18: 411-424.

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

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.

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.

Wood, T. E., N. Takebayashi, M. S. Barker, I. Mayrose, P. B. Greenspoon, and L. H. Rieseberg. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences, USA* 106: 13875-13879.

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.

Yu, Y., and L. Nakhleh. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics* 16(supplement 10): S10.

Yu, Y., J. Dong, K. J. Liu, and L. Nakhleh. 2014. Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences, USA* 111: 16448-16453.

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.

Zhang, C., E. Sayyari, and S. Mirarab. 2017. ASTRAL-III: increased scalability and impacts of contracting low support branches. *In* J. Meidanis and L. Nakleh [eds.], Comparative genomics, RECOMB-CG 2017. Lecture Notes in Computer Science, vol. 10562, 53-75. Springer, Cham, Switzerland.

TABLE 1. Major desiderata, challenges, and opportunities for global plant phylogenetic synthesis.

The tree of life should be:	**Challenge**	**Opportunities**
Integrated	Synthetic trees are currently produced in an uncoordinated	Implementation of modular pipelines, common data

	way, using diverse methods with different limitations and sampling. Additionally, trees are often generated in isolation from related research communities, e.g., palaeontologists vs. neontologists.	standards and application programming interfaces (APIs) would allow multiple research groups to contribute to a central and flexible tree-building platform to serve different tree use applications and better facilitate cross-community coordination.
Up to date	Trees are usually static products that are out of date as soon as they are published since new genetic data are constantly produced. They have no specified routine for updates.	Phylogeny reconstruction can be scripted with minimal or no user interference, allowing scripts to be rerun automatically at regular intervals.
High quality	Quality controls on data in public repositories are weak, which reduces confidence in synthetic phylogenies that use the data.	New data should be generated to rigorous quality standards, supported by the major repositories. Existing data can be cleaned with automated algorithms, and problematic data should be clearly marked. User feedback can improve data quality.
Open	Not all methods and pipelines are open source, preventing the community from fully using them, limiting development potential.	Well-established platforms such as GitHub, Dryad, FigShare, and others allow sharing and customization of code, data, and pipelines.
Reproducible	Phylogeny reconstruction often involves manual editing, and not all steps are fully	Phylogeny reconstruction can be scripted to run without any user intervention. Scripts and

	documented. Thus, analyses cannot readily be verified or re-run with updated input data.	intermediate data (e.g., alignments) can be archived and provided together with trees.
Sustainable	Tree of Life research is often hampered by short project lifetimes and funding cycles. No individual or organisation has responsibility for maintaining a dynamic tree of life.	Institutions and data repositories could collaborate, pooling complementary resources to create a sustainable service to the scientific community.

BOX 1. An outline of general uses of global phylogenetic trees. The following use cases together help define and guide short and long-term goals for a phylogenetic cyberinfrastructure.

(1) Applied user. A plant breeder may ask, does a given species have the potential to be selected for certain traits (e.g., drought tolerance)? To answer this question, they will want to input a taxon name and see a list of close relatives, ideally annotated with the trait of interest.

(2) Educator: A botanic garden educator may want to make a panel showing the phylogenetic relationships among some species growing in the garden. They will want to input a short list of species (usually less than a 100) or identify a clade of interest (e.g., Rosaceae) and download a phylogeny of those species in a format that can be easily turned into a visually appealing figure.

(3) Conservationist: A conservation biologist may want to compare the phylogenetic diversity of a set of areas (e.g., forest fragments) to prioritize conservation efforts. They will want to calculate phylogenetic diversity using statistical packages such as PICANTE (Kembel et al., 2010) or Biodiverse (Laffan et al., 2010), ideally without having to choose and handle a phylogenetic tree.

(4) Comparative biologist: A comparative biologist may want to test the relationship between climate and leaf traits across a set of species. They will want to run a phylogenetic regression model that uses the most up-to-date phylogenetic relationships, ideally without having to choose and handle a phylogenetic tree (although they may have an opinion on phylogenetic methods and appreciate getting to choose among several alternative trees).

(5) Phylogeneticist: An experienced phylogeneticist may want to build a tree using a specific

combination of methods, and potentially even modify/customize some of them. They would fork the phylogenetic pipeline, modify it, and potentially run it on their own computational infrastructure.

(6) Senior biodiversity scientist: A principal investigator writing a grant application may wonder where phylogenetic knowledge gaps are, where most sequencing effort is currently focused, and where additional effort would yield the highest returns. They would want to see a tree annotated with data gaps (Fig. 2), and ideally also with planned and ongoing sequencing projects run by other groups.

FIGURE 1. Schematic representation of a pipeline for building and disseminating an integrated, up-to-date, high quality, open, reproducible, and sustainable tree of life for plants. Colors refer to the sections in the text: blue, gathering the data; yellow, phylogenetic reconstruction; purple, storing the data; green, disseminating the tree of life.

FIGURE 2. A phylogeny of seed plants, Smith and Brown (2018, this issue), where the color of each branch corresponds to the proportion of species from that clade that are represented in public sequence databases. Red branches are missing all or nearly all species, blue branches have a high proportion of species sampled, and yellow and green branches have from one to three thirds of species sampled. Labeled internal nodes show estimates of the number of species lacking sequence data in some major clades.