Synthesis of phylogeny and taxonomy into a comprehensive tree of life

Stephen A. Smith, Karen A. Cranston*, James F. Allman, Joseph W. Brown, Gordon Burleigh, Ruchi Chaudhary, Lyndon M. Coghill, Keith A. Crandall, Jiabin Deng, Bryan T. Drew, Romina Gazis, Karl Gude, David S. Hibbett, Cody Hinchliff, Laura A. Katz, H. Dail Laughinghouse IV, Emily Jane McTavish, Christopher L. Owen, Richard Ree, Jonathan A. Rees, Douglas E. Soltis, Tiffani Williams

*Corresponding

Abstract

Reconstructing the phylogenetic relationships that unite all biological lineages (the tree of life) is a grand challenge of biology. However, the paucity of readily available homologous character data across disparately related lineages renders direct phylogenetic inference currently untenable. Our best recourse towards realizing the tree of life is therefore the synthesis of existing collective phylogenetic knowledge available from the wealth of published primary phylogenetic hypotheses, together with taxonomic hierarchy information for unsampled taxa. We combined phylogenetic and taxonomic data to produce a draft tree of life -- the Open Tree of Life – containing 2.3 million tips. Realization of this draft tree required the assembly of two resources that should prove valuable to the community: 1) a novel comprehensive global reference taxonomy, and 2) a database of published phylogenetic trees mapped to this common taxonomy. Our open source framework facilitates community comment and contribution, enabling a continuously updatable tree when new phylogenetic and taxonomic data become digitally available. While data coverage and phylogenetic conflict across the Open Tree of Life illuminates significant gaps in both the underlying data available for phylogenetic reconstruction and the publication of trees as digital objects, the tree provides a compelling starting point from which we can continue to improve through community contributions. Having a comprehensive tree of life will fuel fundamental research on the nature of biological diversity, ultimately providing up-to-date phylogenies for downstream applications in comparative biology, ecology, conservation biology, climate change studies, agriculture, and genomics.

Significance statement

Scientists have used the characteristics of species and genetic data to construct tens of thousands of evolutionary trees that describe the evolutionary history of animals, plants and microbes. This study is the first to apply an efficient and automated process for assembling published evolutionary trees into a complete tree of life. This tree, and the underlying data, are available to browse and download from the web, allowing for use in subsequent analyses that

require evolutionary trees. It can be easily updated with newly-published data. Our analysis of coverage highlights gaps in sampling and naming biodiversity but also that the results of most published studies are not available in digital formats that can be summarized into a tree of life.

Introduction

The realization that all organisms on Earth are related by common descent (1) was one of the most profound insights in scientific history. The goal of finding the phylogenetic relationships of all species - reconstructing the tree of life - has emerged as one of the most daunting challenges in biology. The scope of the problem is immense: there are approximately 1.8 million named species, and the vast majority of taxa have yet to be described(2)(3)(4). Despite decades of effort and tens of thousands of phylogenetic studies focused on diverse clades of organisms, we still lack a comprehensive tree of life, or even a cogent summary of our current phylogenetic knowledge. One reason for this major shortcoming is the paucity of data. GenBank contains DNA sequences for roughly 411,000 named species, which is only 22% of estimated named species, and a far lower percent of estimated total species. Even if we did attempt to combine all available molecular data in a single large analysis, there are few homologous gene regions that span deep clades, making it difficult to recover monophyly even for clades thought to be well-supported(5).

While there is an extensive publication stream of new phylogenies, new data, and new inference methods, less attention has been paid towards the synthesis of collective phylogenetic knowledge. Given the data limitations noted above, we focus here on constructing a synthetic tree of life through the integration of phylogenetic information from published phylogenies. Phylogenies produced by systematists with expertise in particular taxa almost certainly represent the best estimates of evolutionary relationships for individual clades in the tree of life. By focusing on trees instead of raw data, we avoided issues of dataset assembly (6). However, there are other daunting challenges. One difficulty in building a comprehensive tree of life is that most recognized species have never been included in a phylogenetic analysis because no appropriate molecular or morphological data has been collected. Another challenge of the approach we employ is that most published phylogenies are available only as figures in journal articles, rather than electronic formats that can be integrated into databases and synthesis methods(7)(8)(9). Although there are some efforts to digitize trees from manuscript figures (10), we focused instead on synthesis of newly-published, digitally-available phylogenies. Through intensive efforts we obtained, curated, and made publicly available thousands of published trees that can be used in analyses.

Here we applied an automated and scalable approach to constructing a complete phylogeny of over 2.3 million OTUs (operational taxonomic units) by combining published phylogenetic trees with taxonomic hierarchies. When source phylogenies are absent or sparsely sampled,

taxonomic hierarchies can provide structure in the tree of life (11, 12). In fact, given the limits of data availability, synthesizing phylogeny and taxonomy is the only way to construct a tree of life that includes all recognized species. However, the primary obstacle impeding the synthesis of collective phylogenetic knowledge has been the absence of an all-inclusive taxonomy that spans traditional taxonomic codes (13). We therefore assembled a comprehensive global reference taxonomy through the alignment and merging of multiple openly-available taxonomic resources. The Open Tree Taxonomy (OTT) is an enormous contribution on its own and provides major benefits to the scientific community at large in that it is open, extensible, and updatable, and reflects the overall phylogeny of life to the best extent now possible. Moreover, with the continued updating of phylogenetic information from published studies, this framework is well-poised to update taxonomy itself in a phylogenetically-informed manner far more rapidly than has occurred historically.

We used newly developed graph methods (14) to synthesize a comprehensive tree of life from our reference taxonomy and hundreds of published phylogenies. Advantages of our approach include allowing easy storage of topological conflict among underlying source trees in a single database, the construction of alternative synthetic trees, and the ability to continuously update the tree with new phylogenetic and/or taxonomic information. Importantly, the methodology we employ also highlights the current state of knowledge for any given clade and reveals those portions of the tree that are in most need of additional study--those areas for which character-based phylogenetic analyses are lacking. We stress that although a massive undertaking in its own right, this draft tree of life represents only a first step. In addition to subsequent improvements provided by the community at large, in the future we will also explore (and invite others to explore) additional methods for constructing comprehensive trees.

Results

Input trees and taxonomy

We built a draft tree of life by vetting thousands of published trees and ultimately combining 483 representative phylogenies, connected and supplemented with the comprehensive Open Tree Taxonomy (OTT) that we constructed from multiple taxonomic sources. Taxonomy provides structure and completeness in all branches of the tree of life where phylogenetic studies have not sampled all known lineages (i.e., in the vast majority of clades); in regions covered by phylogenetic estimates, phylogenies overrule the taxonomic scaffolding. A major, underappreciated challenge to any large phylogenetic analysis involving multiple studies is the need to map organisms to a common taxonomic framework. Tips in each input phylogeny, which may represent different taxonomic levels, must be mapped to known taxonomic entities in order to align phylogenies from different sources (14). As a taxonomy sufficient for our purposes was not available, we constructed a reference taxonomy. Our reference taxonomy

OTT is an automated synthesis of taxonomies from NCBI (15), GBIF (16), Index Fungorum (17), IRMNG (18) and SILVA (19, 20). It contains both taxa with traditional Linnaean names and unnamed taxa known only from sequence data. OTT v 2.8 has 2,722,024 "tips" (OTUs without descendants) and includes 382,564 higher taxa; 585,081 of the names are classified as non-phylogenetic units (e.g., *incertae sedis*) and were therefore not included in the synthesis pipeline. As part of our tree curation workflow we developed a taxonomic resolution service accessible through web services that provides not only the ability to align trees input into our database with the OTT, but also allows other researchers to align their own trees with our taxonomy.

At the time of publication, only a fraction of studies in our phylogeny database (http://github.com/phylesystem) contain trees sufficiently well-curated and judged appropriate for incorporation into the synthesis pipeline. The tree database at the time of constructing the tree presented here contained 6753 trees from 3040 studies, where the number of trees per study ranges from 1 to 61. Although we vetted thousands of input trees, ultimately there were 483 source trees that we considered sufficiently well-curated and significant enough to be used for synthesis. Here, a "well-curated" tree is one in which the majority of OTU labels have been mapped to the taxonomic database, the root is correctly identified, and an ingroup clade has been identified. Not all of the sufficiently-curated trees available in the database were incorporated into the tree of life. For example, while we have many phylogenies spanning Angiosperms, we did not include older trees where there is a newer tree for the same clade that uses and extends the same underlying data. Additional discussion of why all trees are not included is presented in the Methods. Moreover, there are still vast areas of the tree of life where published phylogenies are not available. In these areas, the limited number of input trees highlights the need for the additional phylogenetic studies and the low rate of deposition of published tree files into data repositories.

Phylogenetic synthesis

We imported our taxonomy (OTT) and 483 prioritized and well-curated source phylogenies into a graph database, resulting in a "tree alignment graph" (14) which we refer to as the graph of life. The graph of life contains 2,339,460 leaf nodes (the same number as the Open Tree Taxonomy after excluding non-phylogenetic units), plus 227,093 internal nodes. It also stores all of the conflict among phylogenies and between phylogenies and the taxonomy. By traversing the graph and resolving conflict based on priority of inputs, we extracted a tree that represents a synthetic summary of the source information (Figure 1). The priority of source trees was determined by expert curators ranking source trees. This allows for a clear communication of how conflicts will be resolved (i.e., the ranking) as well as an easy way to communicate the source trees that support the particular resolution. This synthetic tree contains phylogenetic

structure where we have published trees and taxonomic structure where we do not. It contains 2,339,460 tips, of which 46,162 are represented in at least one phylogenetic input. However, this underestimates the phylogenetic information content of the tree, as many phylogenies use higher taxa as tips. The tree is available to browse and download, and APIs allow extraction of subtrees given lists of species (see Data and Software Availability, below).

A. Coverage

The synthetic tree of life contains tips representing the input from both taxonomy and phylogeny. Figure 2 examines the change in information content across major taxonomic groups when we compare various biodiversity data sources and the Open Tree of Life. In groups such as Bacteria, Fungi, Nematoda, and Insecta, there is a large gap between the estimated number of species and what exists in taxonomic and sequence databases. In contrast, Chordata and Embryophyta are nearly fully sampled in databases and in OTT. Clades that are not well sampled require more data collection and deposition and, in some cases, formal taxonomic codification and identification in order to be incorporated in taxonomic databases.

B. Resolution and conflicts

The tree of life we provide here is only one representation of the Open Tree of Life data. Analysis of the full graph database (the graph of life) allows us to examine conflict between the synthetic tree of life, taxonomy, and source phylogenies. Figure 3 depicts the types of alternate resolutions that exist in the graph. In total we recovered 153,109 clades in the tree of life, of which 129,778 (84.8%) are shared between the tree of life and the Open Tree Taxonomy. There are 23,331 clades that either conflict with the taxonomy (4610 clades; 3.0%) or where the taxonomy is agnostic to the presence of the clade (18721 clades; 12.2%). The average number of children for each node in the taxonomy is 19.4, indicating a poor degree of resolution. When we combine the taxonomy and phylogenies into the synthetic tree, the resolution improves to an average of 16.0 children per internal node. The average number of children in the input phylogenies is 2.1.

Examination of the alignment of edges between the taxonomy and the synthetic tree of life reveals how well taxonomy reflects current phylogenetic knowledge. Strong alignment is found in groups such as Primates and Mammalia, while our analyses reveal a wide gulf between taxonomy and phylogeny in other clades such as Fungi, Viridiplantae (green plants), Bacteria, and various microbial eukaryotes (see Table 1).

Table 1: Alignment between taxonomy and phylogeny in various clades of the tree of life.

Clade	Number tips	Edges supported by taxonomy only	Edges supported by trees only	Edges supported by trees and taxonomy
Bacteria	260323	264921 (97%)	2147 (0.8%)	4254 (1.6%)
Cyanobacteria	10581	10069 (88%)	59 (0.5%)	1230 (12.8%)
Ciliates	1497	2142 (99%)	1 (0%)	12 (0.6%)
Nematoda	31287	34593 (99%)	43 (0.1%)	152 (0.4%)
Chlorophytes	13100	14268 (99%)	13 (0.1%)	81 (0.6%)
Rhodophytes	12214	13458 (99%)	14 (0.1%)	35 (0.3%)
Fungi	296667	304295 (99%)	375 (0.1%)	636 (0.2%)
Insecta	941753	1024621 (99%)	2156 (0.2%)	3600 (0.3%)
Chordata	88434	86382 (74%)	11056 (9.6%)	17744 (15.4%)
Primates	681	282 (24%)	431 (36.7%)	460 (39.2%)
Mammals	9539	1861 (13%)	7294 (53.4%)	4501 (33.0%)
Embryophytes	284447	292376 (92%)	6437 (2.0%)	15884 (5.0%)

C. Comparison with supertree approaches

There were no pre-existing methods that scaled to phylogenetic reconstruction of the entire tree of life, meaning that our graph synthesis approach was the only option for tree-of-life-scale analyses. Therefore to compare our synthesis method against existing supertree methods, we employed a hybrid MultiLevelSupertree (MLS, (21)) + synthesis approach. We built MLS supertrees for the largest clades that were computationally feasible and then used these non-overlapping trees as input into the graph database and conducted synthesis, effectively stitching these trees together. The total number of internal nodes in the MLS tree is 151458, compared to 153109 in the graph synthesis tree, although the average number of children is the same (16.0 children / node). If we compare the source phylogenies against the MLS supertree and the draft synthetic tree, the synthesis method seems to be doing a better job at capturing the signal in the inputs. The average topological error (normalized Robinson-Foulds distance, where 0 = share all clades and 100 = share no clades (22)) of the MLS vs input trees is 31, compared to 25 for the graph synthesis tree. See Supplementary Material for additional details about conflict and support in the MLS supertree.

Discussion

Using novel methods that employ graph databases, we combine published phylogenetic data and our comprehensive Open Tree Taxonomy to produce the first draft tree of life -- the Open Tree of Life. This comprehensive tree contains over 2.3 million tips. The synthetic tree of all named species presented here is comprehensive in terms of clades across the tree of life, but it is far from complete in terms of biodiversity or phylogenetic knowledge. Many species, including many microbial eukaryotes, Bacteria, and Archaea, are not present in openly available taxonomic databases and therefore not incorporated into the synthetic tree. The vast majority of published trees have never been archived in any data repository. As a result, many published relationships are not represented in the synthetic tree because this knowledge only exists as images in journal articles, rather than tree files that can be imported into a database or used in a downstream analysis. Our infrastructure allows for the synthetic tree to be easily and continuously updated through the use of updated taxonomies and newly published phylogenies. The latter is dependent on authors making tree files available in repositories such as TreeBASE (23), Dryad (http://datadryad.org) or through direct upload to Open Tree of Life (http://tree.opentreeoflife.org/curator) and on having sufficient metadata for trees. We hope this synthetic approach will provide incentive for the phylogenetics community to fundamentally change the way we view our phylogenies - as resources to be cataloged in an appropriate public and open repositories (as is done with most sequence data) rather than as results represented as static images.

Conflicts in the tree of life

Our synthesis included only a small subset of the thousands of trees from our database. We do not include every tree for two reasons: 1) problems with data quality, and 2) issues related to conflict resolution. Data quality problems include, but are not limited to, taxonomy mapping issues (e.g., incorrect mapping, incomplete mapping), incorrect rooting, and available trees differing from their published versions.

The resolution of many clades in the Open Tree of Life may be contentious in some areas due to lack of data, while other areas contain extensive conflict (Figure 3). For example, there is debate regarding the monophyly of Archaea - some analyses indicate that eukaryotes are a lineage embedded within Archaea (24)(25) rather than a separate clade. Similarly, multiple resolutions of early diverging animal lineages have been proposed (26–29). Conflict is prevalent for many reasons, including analytical error, inadequate sampling of taxa or characters, paralogy, and lateral/endosymbiotic gene transfer (24, 25, 30). The underlying graph database allows us to store conflicting trees and examine the impact of different resolutions of the tree of life based on different hypotheses (Fig 2).

While supertree methods, including our graph method, can be helpful in pulling datasets together, they can be less satisfactory in resolving conflicts. One problem is due to the nature of supertrees, which are steps removed from the original data. When comparing two edges from two studies, there is little to no relevant information stored in each tree that can be used to resolve conflict. For example, while the number of published trees that support a synthetic edge may be considered a reasonable criterion for resolving conflict (i.e., more edges equate to higher support), the datasets used to construct each source tree may have overlapping data, making them non-independent data points. Other information, such as the number of taxa or gene regions involved, cannot be used alone without other information to assess the quality of the particular analysis. Strong conflict can be difficult to resolve without the additional metadata about the underlying genetic data and phylogenetic inference methods. The resolution of these difficult areas will be better addressed with the direct analysis of genetic data and not through synthesis of trees constructed from those data.

In some cases, source trees available in our database may not be included in synthesis because a published tree supersedes an existing study, in some cases with similar data from the same authors. We wanted to include only the most recently inferred relationships in the synthetic tree, and incorporating the superseded trees merely increases computational time. So, while the addition of trees that include edges that have not been sampled previously is important, the addition of all available trees into the synthesis pipeline is not our goal.

As a result of data availability, data quality, conflict resolution, and because of the focus on grand synthesis, we emphasize that specific portions of the tree may not always agree with the relationships that experts in that group adhere to at this point (e.g., relationships within Fabaceae, Compositae, Arthropoda). But this draft tree of life represents just the initial step. The next step in this community-driven process is for experts worldwide representing the breadth of phylogenetic expertise to contribute trees and annotate those areas of the tree they know best.

Source trees as a community resource

Although not every published tree is necessarily informative for broad synthesis, the availability of published trees for comparative analyses is nonetheless a very important resource. For example, researchers may want to calculate the increase in information content for a particular clade over time or by a particular project or lab. Alternatively, researchers may want to compare trees constructed by different phylogenetic approaches or record the reduction in conflict in clades over time. All of these analyses require that tips on trees be mapped to a common taxonomy in order to compare across trees. With the Open Tree Taxonomy we have already mapped thousands of trees to a common taxonomy and therefore created, in itself, an

extremely valuable resource. Trees that have been mapped to the Open Tree Taxonomy have added value in being immediately comparable across thousands of other trees. The data input and curation interface is publicly available (http://tree.opentreeoflife.org/curator) as is the underlying data store (http://github.com/opentreeoflife/phylesystem).

Dark parts of the tree

There are large parts of the tree of life that are not yet well represented in input taxonomies, particularly in hyperdiverse, poorly understood groups including Fungi, microbial eukaryotes, Bacteria, and Archaea. Hence, another important aspect of the tree is that it highlights where major effort is still needed to achieve a better understanding of existing biodiversity. For example, metagenomic studies routinely reveal huge numbers of Fungi, Bacteria and microbial eukaryotes that cannot be assigned to named species (31, 32). For Archaea and Bacteria, there are additional challenges created by their immense diversity, lack of clarity regarding species concepts, and the occurrence of rampant horizontal gene transfer (33)(34)(35). For microbial eukaryotes, Archaea, and Bacteria, the operational unit is often strains rather than species. Strains are not regulated by any taxonomic code, making it difficult not only to map taxa between trees and taxonomy, but also to estimate named biodiversity in these huge clades. New efforts in this area, such as the new BioProject (36) and BioSample (37) databases at NCBI, have the potential to better describe and catalog biodiversity that does not fit into traditional taxonomic workflows.

Materials and methods

Input data: taxonomy

There is no single taxonomy that is both complete - including both traditional and non-Linnaean taxa - and in which the backbone is well-informed by phylogenetic studies. We constructed such a taxonomy, the Open Tree Taxonomy (OTT), by merging Index Fungorum (17), SILVA (19, 20), NCBI (15), GBIF (16), IRMNG (18) and two clade-specific resources (38)(39) using a fully-documented, repeatable process. (See SOM). The taxonomy (v 2.8.5) consists of 2,722,024 well-named entities and 1,360,819 synonyms and is a huge biodiversity resource. There are an additional 585,081 entities with non-biological or taxonomically incomplete names, such as “environmental samples” or “incertae sedis”, that are retained in OTT, but not included in the synthetic phylogeny.

Input data: phylogenetic trees

We imported and curated phylogenetic trees using a newly-developed curation interface that saves tree data directly into a GitHub repository. We obtained published trees from TreeBASE (23) and Dryad data repositories, and by direct appeal to authors. The data retrieved are by no means a complete representation of the phylogenetic literature, as only 16% of recently published studies

have provided data in a digital format that can be input into a database (8). Even when available, input trees require a significant amount of curation to be usable for synthesis. Taxon labels in tree files often include non-standard strings such as lab codes or abbreviations that have to be mapped to taxonomic entities in OTT. Trees often need to be rooted (or re-rooted) to match figures from papers. As relationships among outgroup taxa were often problematic, we identified the ingroup / focal clade for the study. For studies that contained multiple trees, we tagged the tree that best matched the conclusions of the study as "preferred". Then, within major taxonomic groups (eukaryotic microbial clades, animals, plants and fungi) we ranked these preferred trees to generate a prioritized list. Rankings were assembled by authors with expertise in specific clades and were based on date of publication, underlying data, and methods of inference (see SOM). In general, these rankings reflect general community consensus. As we collect more metadata (or if trees are published with structured, machine-readable metadata), automated filtering / weighting trees based on metadata will be possible.

Synthesis

Although storing input data in a graph database enables the simultaneous representation of alternative hypotheses, it is also necessary to synthesize the inputs into a single tree as required by many downstream analyses as well as to summarize phylogenetic knowledge. In the absence of structured metadata about the phylogenetic methods and data used to infer the input trees, topological conflicts were arbitrated and resolved using prioritized study lists compiled by the expert curators. Where no other phylogenetic information was present, taxonomy provided resolution. This method requires intensive involvement by the curators to determine a ranking of input trees. Once complete, we pass the tree through a set of regression tests that include tests for monophyly for expected clades and inclusion of taxa in clades (see SOM for details). As an alternative to this user-intensive analysis, we also created a synthetic tree using the MultiLevelSupertree (MLS) approach (21). The MLS method allows for combining trees where the tips in the source trees represent different taxonomic hierarchies. Unfortunately, it is limited in dataset size to roughly a thousand taxa and so we could not run a single MLS analysis on all of our input data. Instead, we inferred 16 separate MLS trees of individual clades, and combined these MLS trees with taxonomy using our graph database methods to create a MLS synthesis tree. Because they were non-overlapping in terms of taxon sampling, there was no topological conflict among the individual MLS trees, and creating the final MLS supertree simply involved traversing the graph and preferring phylogeny over taxonomy.

Data and software availability

The current version of the draft tree is available to browse and download at http://tree.opentreeoflife.org. All software is open-source and available at http://github.com/opentreeoflife. Where not limited by pre-existing terms of use, all data are published with a CC-Zero copyright waiver. An archive of the versions of the Open Tree of Life and the Open Tree Taxonomy described here, as well as CC-Zero inputs, are available at the Dryad data

repository: <insert Dryad DOI here>. Non-CC-Zero input data are available at <insert location of Dataverse repo>.

Acknowledgements

We are grateful to Paul Kirk at Index Fungorum, Tony Rees at IRMNG, Markus Doering at GBIF for input data and advice on taxonomy synthesis, Pam Soltis for helpful comments on the manuscript, Mark Holder for software development related to the phylogenetic data store and for many discussions, to numerous authors who made their tree files available in TreeBASE or Dryad and authors that provided phylogenetic tree files that were not otherwise available, and finally for funding from NSF AVATOL #1208809.

Figures and captions

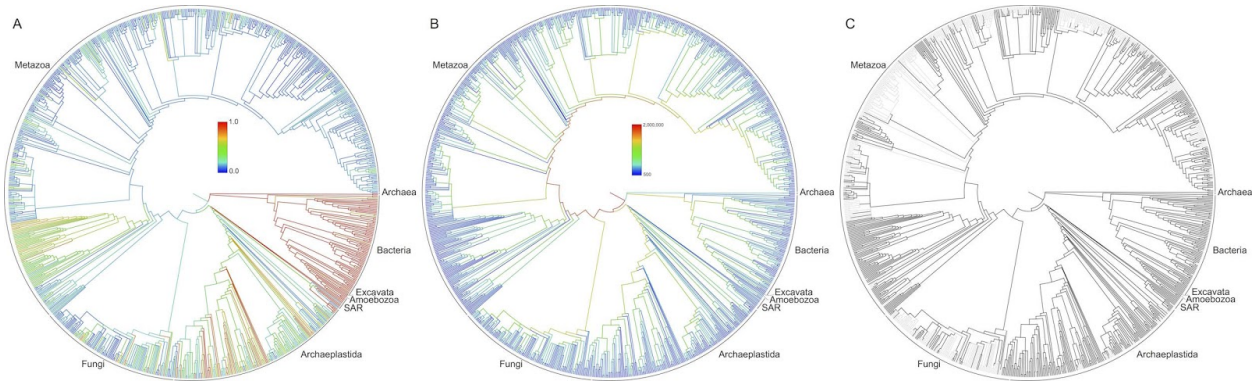

Figure 1. Phylogenies representing the synthetic tree: The depicted tree is limited to lineages containing at least 500 descendants. A. Colors represent proportion of lineages represented in NCBI databases; B. Colors represent the amount of diversity measured by number of descendant tips; C. Dark lineages have at least one representative in an input source tree.

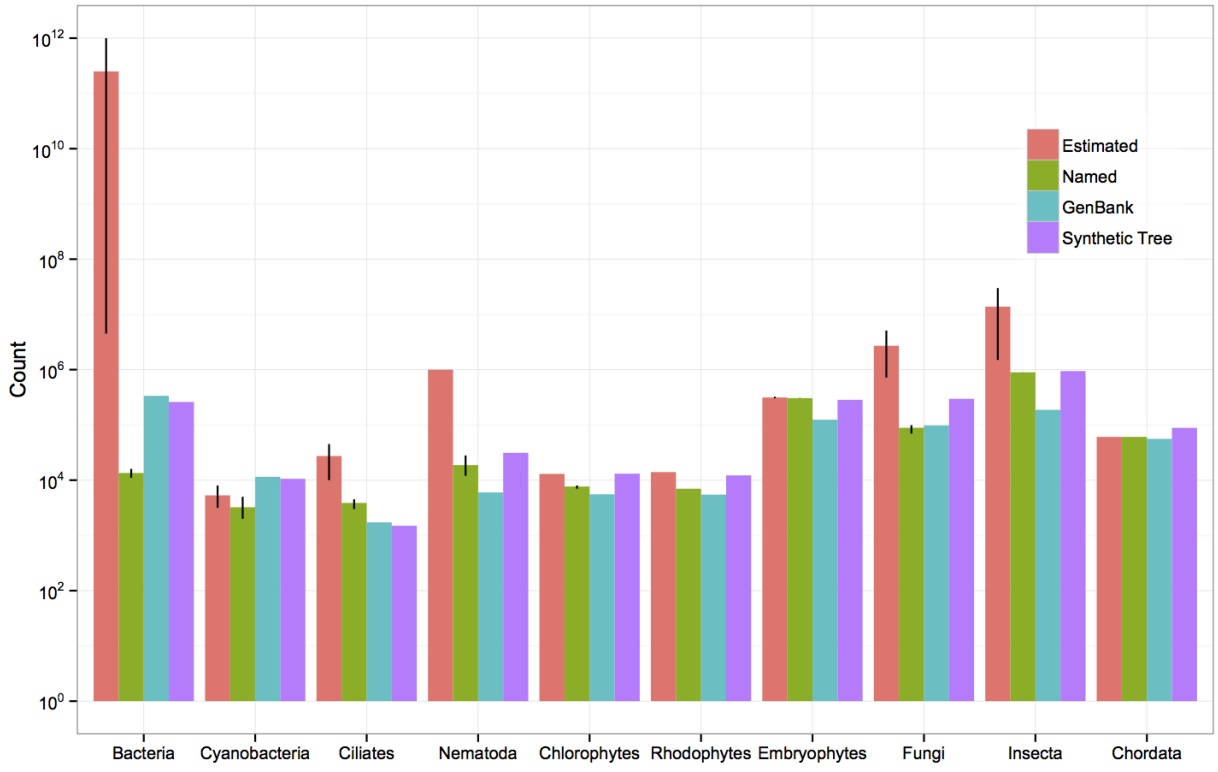

Figure 2: **The estimated total number of species, estimated number of named species in taxonomic databases, the number of OTUs with sequence data in GenBank, and the number of OTUs in the Open Tree of Life synthetic tree, for 10 major clades across the tree of life. Error bars (where present) represent the range of values across multiple sources. Underlying data and references are available in the supplementary material.**

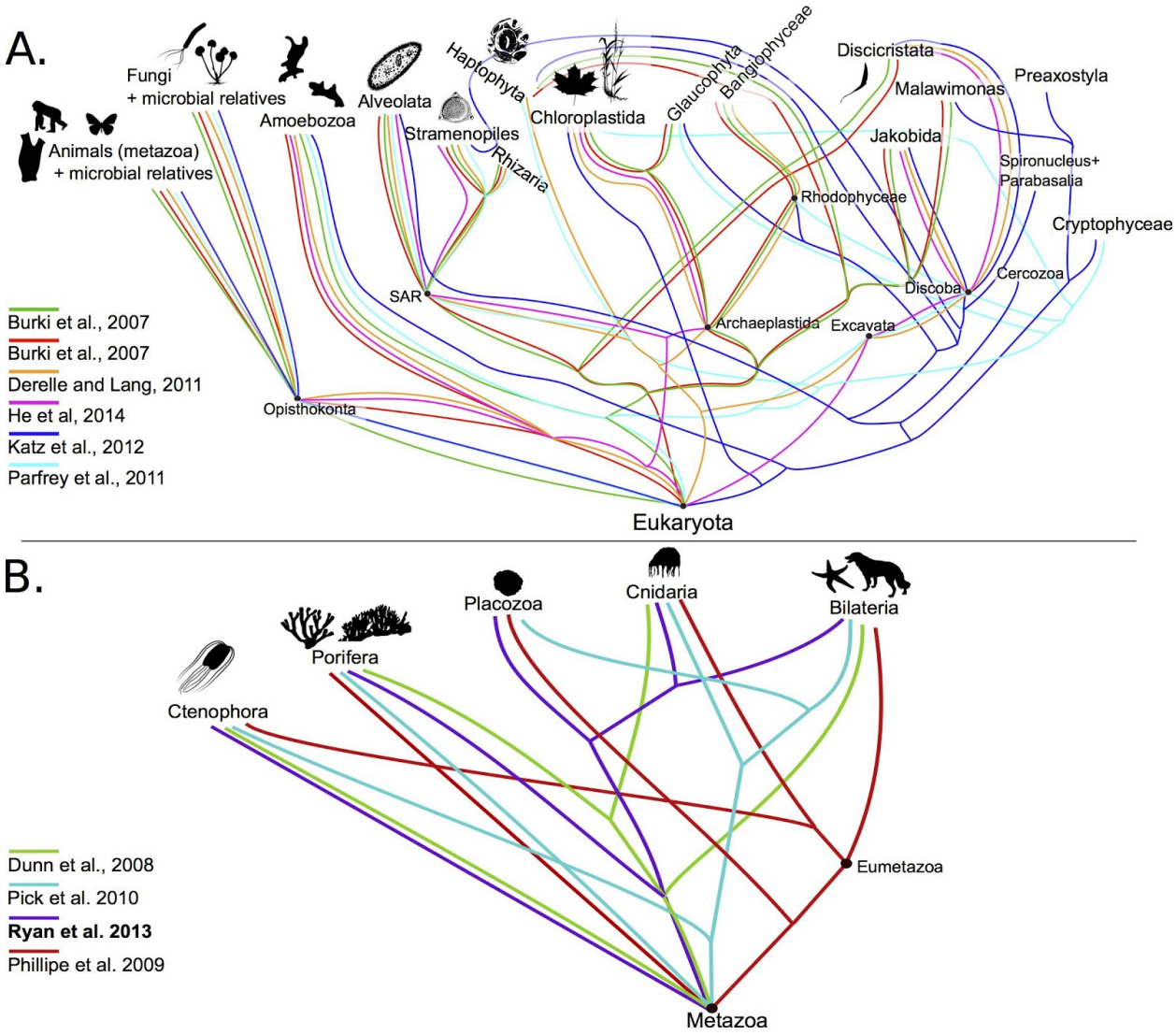

Figure 3: Conflict in the graph of life. While the tree of life can only contain one resolution at any given node, the underlying graph database contains much conflict between trees and taxonomy. These two examples from the graph of life highlight ongoing conflict near the base of Eukaryota and Metazoa. Images from PhyloPic (http://phylopic.org).

References

1. Darwin C (1859) *The Origin of Species: By Means of Natural Selection, Or the Preservation of Favoured Races in the Struggle for Life* (Cambridge University Press).

2. Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on Earth

and in the ocean? *PLoS Biol* 9:e1001127. Available at: http://dx.doi.org/10.1371/journal.pbio.1001127.

3. Costello MJ, Wilson S, Houlding B (2011) Predicting total global species richness using rates of species description and estimates of taxonomic effort. *Syst Biol*:syr080–. Available at: http://dx.doi.org/10.1093/sysbio/syr080.

4. Dykhuizen D (2005) Species Numbers in Bacteria. *Proc Calif Acad Sci* 56:62–71. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21874075.

5. Sanderson MJ (2008) Phylogenetic signal in the eukaryotic tree of life. *Science* 321:121–123. Available at: http://dx.doi.org/10.1126/science.1154449.

6. Sanderson MJ, McMahon MM, Steel M (2010) Phylogenomics with incomplete taxon coverage: the limits to inference. *BMC Evol Biol* 10:155. Available at: http://dx.doi.org/10.1186/1471-2148-10-155.

7. Stoltzfus A et al. (2012) Sharing and re-use of phylogenetic trees (and associated data) to facilitate synthesis. *BMC Res Notes* 5:574. Available at: http://dx.doi.org/10.1186/1756-0500-5-574.

8. Drew BT et al. (2013) Lost branches on the tree of life. *PLoS Biol* 11:e1001636. Available at: http://dx.doi.org/10.1371/journal.pbio.1001636.

9. Magee AF, May MR, Moore BR (2014) The Dawn of Open Access to Phylogenetic Data. *arXiv [q-bioPE]*. Available at: http://arxiv.org/abs/1405.6623.

10. Murray-Rust P, Smith-Unna R, Mounce R (2014) AMI-diagram: Mining Facts from Images. *D-Lib Magazine* 20. Available at: http://www.dlib.org/dlib/november14/murray-rust/11murray-rust.html.

11. Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO (2012) The global diversity of birds in space and time. *Nature* 491:444–448. Available at: http://dx.doi.org/10.1038/nature11631.

12. Bininda-Emonds OR, Sanderson MJ (2001) Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. *Syst Biol* 50:565–579. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12116654.

13. Polaszek A (2010) *Systema Naturae 250 - The Linnaean Ark* (Taylor & Francis) Available at: http://books.google.com/books?id=GCgEngEACAAJ.

14. Smith SA, Brown JW, Hinchliff CE (2013) Analyzing and synthesizing phylogenies using tree alignment graphs. *PLoS Comput Biol* 9:e1003223. Available at: http://dx.doi.org/10.1371/journal.pcbi.1003223.

15. Sayers EW et al. (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 37:D5–15. Available at: http://dx.doi.org/10.1093/nar/gkn741.

16. GBIF (2013) The Global Biodiversity Information Facility: GBIF Backbone Taxonomy. Available at: http://www.gbif.org/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c.

17. Kirk P Index Fungorum. Available at: http://www.indexfungorum.org/ [Accessed April 1, 2014].

18. Rees T Interim Register of Marine and Nonmarine Genera (IRMNG). Available at: http://www.obis.org.au/irmng/ [Accessed January 31, 2014].

19. Quast C et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41:D590–6. Available at: http://dx.doi.org/10.1093/nar/gks1219.

20. Yilmaz P et al. (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res* 42:D643–8. Available at: http://dx.doi.org/10.1093/nar/gkt1209.

21. Berry V, Bininda-Emonds ORP, Semple C (2013) Amalgamating source trees with different taxonomic levels. *Syst Biol* 62:231–249. Available at: http://dx.doi.org/10.1093/sysbio/sys090.

22. Kupczok A, Schmidt HA, von Haeseler A (2010) Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. *Algorithms Mol Biol* 5:37. Available at: http://dx.doi.org/10.1186/1748-7188-5-37.

23. Sanderson MJ, Donoghue MJ, Piel W, Eriksson T (1994) TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *Am J Bot* 81:183.

24. Williams TA, Foster PG, Nye TMW, Cox CJ, Martin Embley T (2012) A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc Biol Sci* 279:4870–4879. Available at: http://dx.doi.org/10.1098/rspb.2012.1795.

25. Lake JA, Henderson E, Oakes M, Clark MW (1984) Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc Natl Acad Sci U S A* 81:3786–3790. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6587394.

26. Dunn CW et al. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452:745–749. Available at: http://dx.doi.org/10.1038/nature06614.

27. Ryan JF et al. (2013) The genome of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution. *Science* 342:1242592. Available at: http://dx.doi.org/10.1126/science.1242592.

28. Pick KS et al. (2010) Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Mol Biol Evol* 27:1983–1987. Available at: http://dx.doi.org/10.1093/molbev/msq089.

29. Philippe H et al. (2009) Phylogenomics revives traditional views on deep animal relationships. *Curr Biol* 19:706–712. Available at: http://dx.doi.org/10.1016/j.cub.2009.02.052.

30. Wu D et al. (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462:1056–1060. Available at: http://dx.doi.org/10.1038/nature08656.

31. Lee CK et al. (2012) Groundtruthing next-gen sequencing for microbial ecology-biases and errors in community structure estimates from PCR amplicon pyrosequencing. *PLoS One* 7:e44224. Available at: http://dx.doi.org/10.1371/journal.pone.0044224.

32. Hibbett DS et al. (2011) Progress in molecular and morphological taxon discovery in Fungi and options for formal classification of environmental sequences. *Fungal Biol Rev* 25:38–47. Available at: http://www.sciencedirect.com/science/article/pii/S1749461311000030.

33. Syvanen M (2012) Evolutionary implications of horizontal gene transfer. *Annu Rev Genet* 46:341–358. Available at: http://dx.doi.org/10.1146/annurev-genet-110711-155529.

34. Gilbert C, Cordaux R (2013) Horizontal transfer and evolution of prokaryote transposable elements in eukaryotes. *Genome Biol Evol* 5:822–832. Available at: http://dx.doi.org/10.1093/gbe/evt057.

35. Qiu H, Yoon HS, Bhattacharya D (2013) Algal endosymbionts as vectors of horizontal gene transfer in photosynthetic eukaryotes. *Front Plant Sci* 4:366. Available at: http://dx.doi.org/10.3389/fpls.2013.00366.

36. NCBI BioProject database Available at: http://www.ncbi.nlm.nih.gov/bioproject/ [Accessed December 2014].

37. NCBI BioSample database Available at: http://www.ncbi.nlm.nih.gov/biosample [Accessed December 2014].

38. Hibbett DS et al. (2007) A higher-level phylogenetic classification of the Fungi. *Mycol Res* 111:509–547. Available at: http://dx.doi.org/10.1016/j.mycres.2007.03.004.

39. Schäferhoff B et al. (2010) Towards resolving Lamiales relationships: insights from rapidly evolving chloroplast sequences. *BMC Evol Biol* 10:352. Available at: http://dx.doi.org/10.1186/1471-2148-10-352.

Supplementary materials and methods

Overview

Constructing the Open Tree of Life involves two different types of inputs: taxonomies and published phylogenies. Figure 1 gives an overview of the process, data stores and services. We first combine multiple taxonomic hierarchies into a single taxonomy, the Open Tree Taxonomy (OTT). In a web application built specifically for this project, we input and curate published phylogenies that are then saved to a public GitHub repository. The input trees and OTT are then uploaded to a neo4j graph database. From this database, we traverse the graph starting at the root node and extract a tree by resolving conflict base on priority of inputs (phylogeny takes priority over taxonomy, and then we refer to a prioritized list of the input trees).

Constructing a composite taxonomy

The synthesis of the OpenTree Taxonomy (version 2.8 used here) is a fully automated process. The pipeline takes taxonomy database inputs, in this case, Index Fungorum (1), SILVA (2, 3), NCBI (4), GBIF (5), IRMNG (6) and two clade-specific resources (7, 8). Source taxonomies, each of which is published in its own idiosyncratic representation, are first preprocessed to convert them to a common format. Each source taxonomy in turn is then merged into developing a union taxonomy. Merging a source taxonomy into the union taxonomy consists of two steps: aligning source nodes to union nodes to resolve homonyms, followed by transferring unaligned (new) nodes into the union. A set of about 300 scripted ad hoc manipulations repair situations where automatic alignment has failed and fix errors in the input taxonomies. Because the process is scripted, it can be executed any time one of the input taxonomies is revised. The source code for this process is available at https://github.com/OpenTreeOfLife/reference-taxonomy. Version 2.8 of the OpenTree Taxonomy consists of 3,307,105 names, of which 2,722,024 are external (tips) and 585,081 are internal. The taxonomy also has 1,360,819 synonyms. Each name is given a unique id (OTT id) that is used for mapping taxa in trees. The produced taxonomy is then ingested into a neo4j graph database managed by the open source program taxomachine

(https://github.com/OpenTreeOfLife/taxomachine). Taxomachine serves the taxonomy with REST calls over a network and provides a taxonomic name resolution service that allows for disambiguation of taxonomic names within tree sets as a result of misspellings, changed classification, or homonyms. Taxomachine returns the unique OpenTree taxonomy id for each name in a taxonomic name resolution call.

Treestore

We developed a web-based curation interface for phylogenies (http://tree.opentreeoflife.org/curator) connected to a back-end datastore in GitHub, phylesystem (http://github.com/opentreeoflife/phylesystem). The phylogenies in our datastore come from automated upload from TreeBASE or from input of downloaded files from Dryad, from journal supplementary material, and from contacting authors directly for files. At time of synthesis, the datastore contained 6753 trees from 3040 studies (see Supp. Fig 2 for distribution of data). Studies contained information about the publication, trees, and the list of taxa included. In all cases, the original tree files did not contain sufficient annotation and metadata for synthesis, so there was significant curation by experts. Curation involved two major steps. First, curators mapped the tip labels in the trees to entities in taxonomic databases, assigning an OTT id and disambiguating any problems due to homonyms. Then, curators checked that the root was correct, the ingroup was identified, and that the tree matched the figure in the publication. The current public data submission systems (for trees or otherwise) do not verify the validity of the tree files, which, for a number of reasons, are often very different from the original publication. Ingroups needed to be specified because often the rooting of the outgroup is not accurate, and therefore, relationships in the outgroups may be poor. Once curated, studies are then stored in the phylesystem GitHub repository as Nexson files (NeXML (38) serialized as javascript object notation, JSON). More information about NeXSON can be found at http://purl.org/opentree/nexson.

Synthesis

The source trees used for input were determined based on the coverage and resolution of major clades, ability to confidently root the tree and map taxa in the trees to the OpenTree Taxonomy. Although more than a thousand trees had some form of curation, 335 were curated to the point of being able to be included in synthesis. These trees from the tree store were processed and added to a graph database Neo4j using the software treemachine as described by Smith et al. ((9); https://github.com/OpenTreeOfLife/treemachine). Once all trees have been added into a graph database, we conducted synthesis based on a ranking of the input source trees and taxonomy. A basic diagram can be seen in Supp. Fig. 2 but for more information, consult Smith et al., (*12*). Taxonomy was the lowest ranked input. Synthesis is managed using additional Python software called gcmdr (https://github.com/OpenTreeOfLife/gcmdr) that allows for repeatability with simple scripts. In this first version of the tree, we focused on four major groups: plants, fungi, metazoa, and microbes. Once the source trees were loaded into the graph database, we constructed the

synthetic tree by proceeding from the root of the graph database (in this case, cellular organisms) to the tips and resolving conflicts along the way. Conflicts were considered to be partially overlapping partitions with an alternative resolution. When conflicts, or alternative routes, were presented we resolved by always preferring phylogeny to taxonomy. When phylogenetic branches were in conflict, we resolved based on a priority order of input phylogenies (see Supplementary Table 1 for rankings). The software supports conflict resolution based on other metadata (for example, MIAPA metadata such as type of topology or method of phylogenetic analysis), but given that most tree files do not contain this information, this is not currently possible. Therefore, in this first version, OpenTree curators ranked the input trees for their clades of expertise with the intention to reflect general consensus in the field. In many cases, phylogenies found named clades to be non-monophyletic. In these cases, the taxa included in the original study would be resolved and the taxa that had been placed in the named clade but not found in the study, in the absence of other information, were then sunk to higher monophyletic taxa. The source trees that support each branch in the synthetic tree are then recorded and are reported in the web version of the synthetic tree. We also pass the tree through a set of regression tests that check for monophyly for expected clades and inclusion of taxa in clades. These tests are automated, and new tests can be contributed through pull requests on GitHub (https://github.com/OpenTreeOfLife/germinator/tree/master/taxa).

Conflict between trees and taxonomies

We measured support for the nodes in the supertree using an approach described by Wilkinson et al. (10). Let c be a clade in the supertree S and c' be its restriction to the leaves of an input tree T, i.e., c' contains only those leaves that are present in T. If c' contains all the leaves of T or less than 2 leaves, then T is *irrelevant* to c. T *supports* c if c' is present in T. T *conflicts* with c when the induced bipartition (of c) contradicts the relationships in T (11). T *permits* c if c' is a resolution of a polytomy in T, thus T is agnostic with respect to c. See Fig. 2 for an example. First, we compared the taxonomy tree to the Open Tree of Life. There are 153,109 clades in the Open Tree of Life, and 129,778 (84.8%) of these are supported by the taxonomy tree. There are 4,610 (3.0%) clades in the Open Tree of Life that are in conflict with the taxonomy, and 18,721 (12.2%) that are permitted.

When we compare the collection of 484 non-taxonomy input trees to the Open Tree of Life, there are 27,259 (17.8%) clades in the Open Tree of Life that are unambiguously supported (i.e., ≥1 non-taxonomy input trees support and 0 non-taxonomy input trees conflict with or permit the node) and 765 (0.5%) clades are in unambiguous conflict (i.e., ≥1 non-taxonomy input trees conflict with and 0 non-taxonomy input trees support or permit the node). However, 123,362 (80.6%) of the clades in the Open Tree of Life are irrelevant to the non-taxonomy input trees. Thus, the information for most of the clades in the Open Tree of Life is coming from the taxonomy. The remaining 1723 (1.1%) nodes in the Open Tree of Life have a combination of support, conflict, and permit, instead of complete support or conflict,

among the non-taxonomy input trees with respect to these clades. Overall, 2,684 (1.8%) clades in the Open Tree of Life are supported by at least two non-taxonomy input trees. When we compile all of the input trees together (the taxonomy tree and the 484 non-taxonomy input trees), there are 128,879 (84.2%) nodes in the Open Tree of Life that are unambiguously supported by all relevant input trees and 56 (0.04%) clades are in unambiguous conflict. None of the all assessed clades was irrelevant to the input trees. The remaining 24,174 (15.8%) nodes in the Open Tree of Life have a combination of support, conflict, and permit, instead of complete support or conflict, among the non-taxonomy input trees with respect to these clades. Overall, 7,441 (4.9%) clades in the Open Tree of Life are supported by at least two taxonomy or non-taxonomy input trees.

In contrast, when we compare the collection of 484 non-taxonomy input trees to the MLS Tree of Life, there are 24,891 (16.4%) clades in the MLS Tree of Life that are unambiguously supported (i.e., ≥1 non-taxonomy input trees support and 0 non-taxonomy input trees conflict with or permit the node) and 2,013 (1.3%) clades are in unambiguous conflict (i.e., ≥1 non-taxonomy input trees conflict with and 0 non-taxonomy input trees support or permit the node). However, 123,330 (81.4%) of the clades in the MLS Tree of Life are irrelevant to the non-taxonomy input trees. The remaining 1,224 (0.8%) nodes in the MLS Tree of Life have a combination of support, conflict, and permit, instead of complete support or conflict, among the non-taxonomy input trees with respect to these clades. Overall, 2,693 (1.8%) clades in the MLS Tree of Life are supported by at least two non-taxonomy input trees.

Supplementary Figures

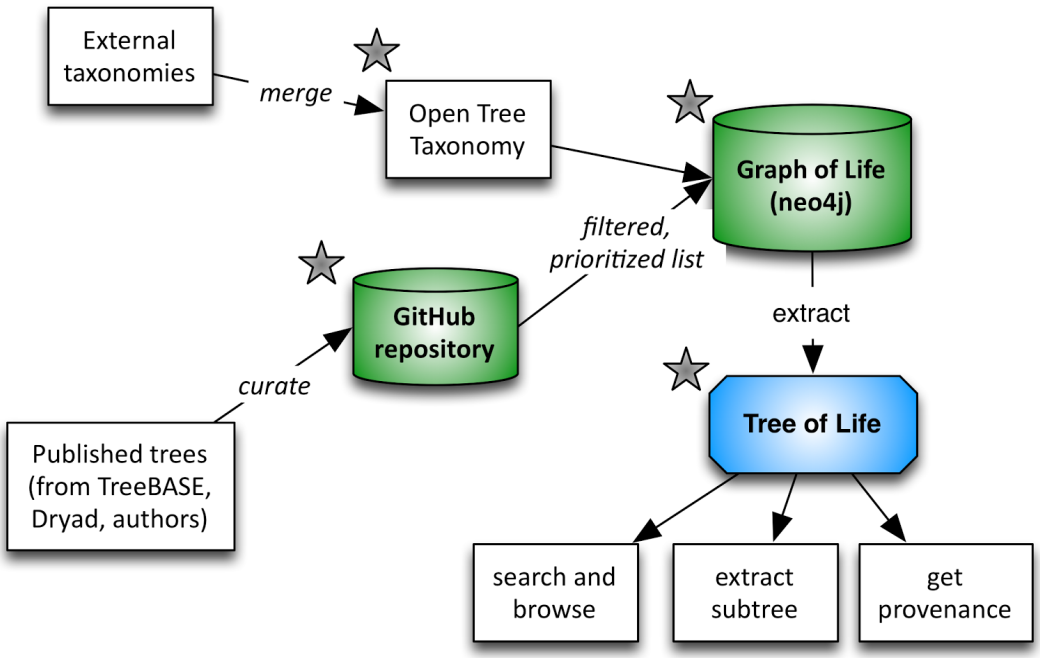

Figure 1. **The Open Tree of Life workflow**: External taxonomies (and synonym lists) are merged into the Open Tree Taxonomy, OTT. Published phylogenies are curated (rooted, and names mapped to OTT) and stored, with full edit history, in a GitHub repository. The source trees and OTT are loaded into a common graph database, and we traverse the resulting graph and extract a tree of life based on priority of inputs. Components with stars in indicate presence of application programming interfaces (APIs) to access data and services.

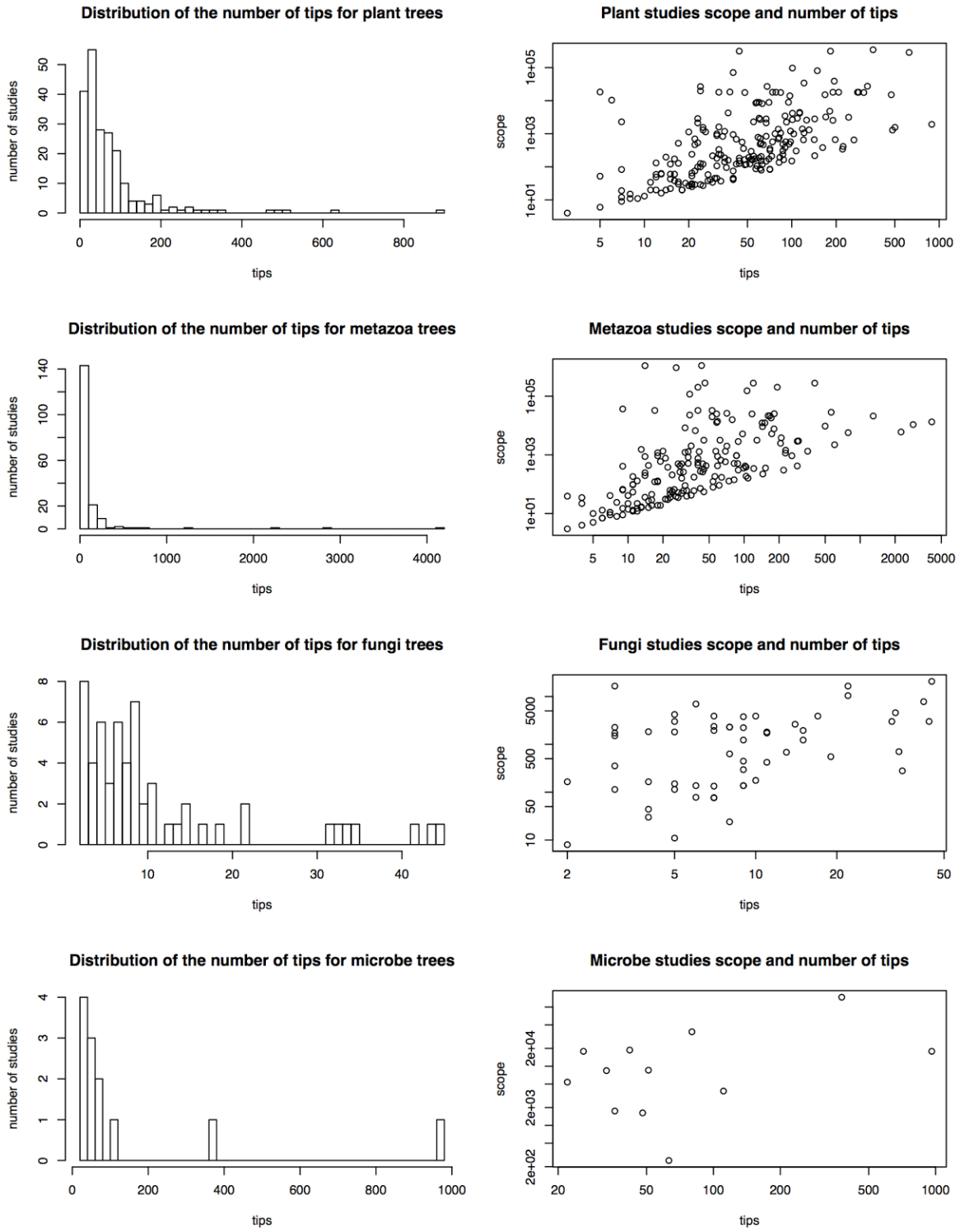

Figure 2. Size and scope of input trees: Plot of the number of tips in each of the 1188 trees with some curation in the treestore. Scope is measures as the total number of tips recognized to be descended from the inferred most recent common ancestor of the source tree.

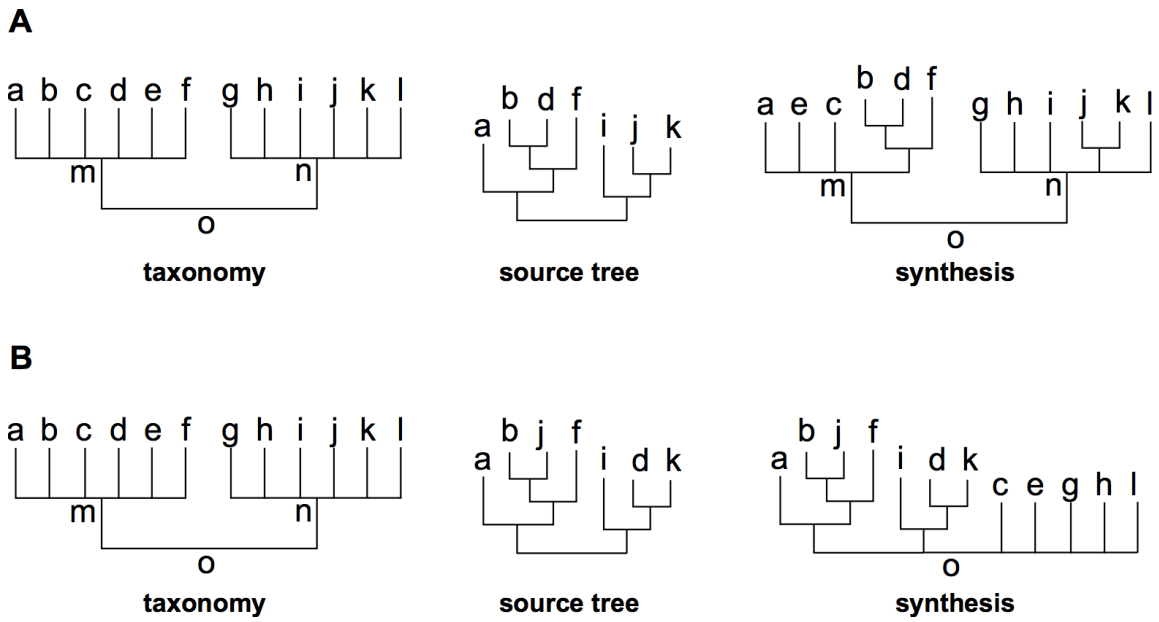

Figure 3. Diagram of synthesis of taxonomy and phylogeny: A. Taxonomy of 12 hypothetical species and a source tree combine in a synthesis with m and n monophyletic. B. Taxonomy of the same 12 species and a source tree combine in synthesis without m and n monophyletic.

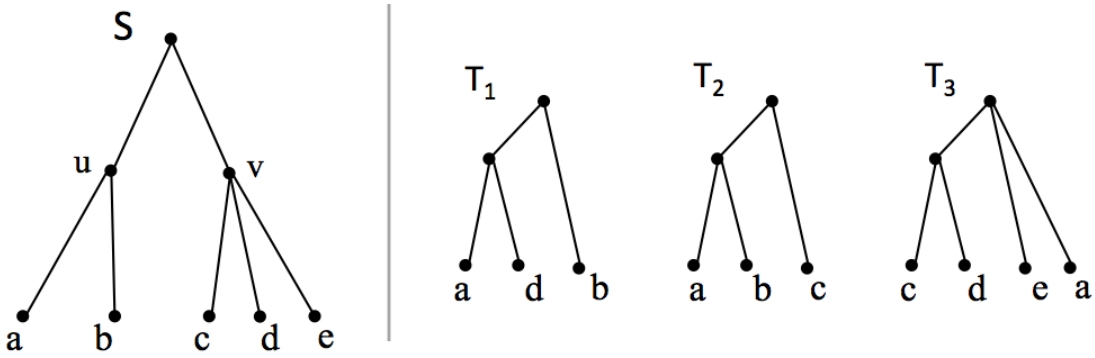

Figure 4. Conflict analysis: A supertree S with two internal nodes u and v, and three input trees T_1, T_2, and T_3. The clade u is in conflict with T_1, supported by T_2, irrelevant to T_3. The clade v is irrelevant to T_1 and T_2, and permitted by T_3, as v is a resolution of the polytomy at the root in T_3.

Supplementary Tables

Table 1. Phylogenetic source trees for synthesis in order of ranking. Includes Open Tree of Life study ID, tree ID (studies may contain multiple trees) and the reference. See SupplementaryTable1.csv.

Table 2. Data underlying Figure 2 in the main text: Estimates for number of binomials and number of species across several taxa. Where multiple rows refer to the same source, that source contains more than one estimate. Includes references for estimates. See supplementaryTable2.csv.

References

1. Kirk P Index Fungorum. Available at: http://www.indexfungorum.org/ [Accessed April 1, 2014].

2. Yilmaz P et al. (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res* 42:D643–8. Available at: http://dx.doi.org/10.1093/nar/gkt1209.

3. Quast C et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41:D590–6. Available at: http://dx.doi.org/10.1093/nar/gks1219.

4. Sayers EW et al. (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 37:D5–15. Available at: http://dx.doi.org/10.1093/nar/gkn741.

5. GBIF (2013) The Global Biodiversity Information Facility: GBIF Backbone Taxonomy. Available at: http://www.gbif.org/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c.

6. Rees T Interim Register of Marine and Nonmarine Genera (IRMNG). Available at: http://www.obis.org.au/irmng/ [Accessed January 31, 2014].

7. Hibbett DS et al. (2007) A higher-level phylogenetic classification of the Fungi. *Mycol Res* 111:509–547. Available at: http://dx.doi.org/10.1016/j.mycres.2007.03.004.

8. Schäferhoff B et al. (2010) Towards resolving Lamiales relationships: insights from rapidly evolving chloroplast sequences. *BMC Evol Biol* 10:352. Available at: http://dx.doi.org/10.1186/1471-2148-10-352.

9. Smith SA, Brown JW, Hinchliff CE (2013) Analyzing and synthesizing phylogenies using tree alignment graphs. *PLoS Comput Biol* 9:e1003223. Available at: http://dx.doi.org/10.1371/journal.pcbi.1003223.

10. Wilkinson M, Pisani D, Cotton JA, Corfe I (2005) Measuring support and finding unsupported relationships in supertrees. *Syst Biol* 54:823–831. Available at: http://dx.doi.org/10.1080/10635150590950362.

11. Semple C, Steel A (2003) *Phylogenetics* (Oxford University Press).