10-13-2015

Synthesis of Phylogeny and Taxonomy Into a Comprehensive Tree of Life

Cody E. Hinchliff
University of Michigan-Ann Arbor

Stephen A. Smith
University of Michigan-Ann Arbor

James F. Allman
Interrobang Corporation, Wake Forest, NC

J. Gordon Burleigh
University of Florida

Ruchi Chaudhary
University of Florida

See next page for additional authors

Authors

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

Synthesis of phylogeny and taxonomy into a comprehensive tree of life

Cody E. Hinchliff[a,1], Stephen A. Smith[a,1,2], James F. Allman[b], J. Gordon Burleigh[c], Ruchi Chaudhary[c], Lyndon M. Coghill[d], Keith A. Crandall[e], Jiabin Deng[c], Bryan T. Drew[f], Romina Gazis[g], Karl Gude[h], David S. Hibbett[g], Laura A. Katz[i], H. Dail Laughinghouse IV[i], Emily Jane McTavish[j], Peter E. Midford[d], Christopher L. Owen[c], Richard H. Ree[d], Jonathan A. Rees[k], Douglas E. Soltis[c,l], Tiffani Williams[m], and Karen A. Cranston[k,2]

[a]Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109; [b]Interrobang Corporation, Wake Forest, NC 27587; [c]Department of Biology, University of Florida, Gainesville, FL 32611; [d]Field Museum of Natural History, Chicago, IL 60605; [e]Computational Biology Institute, George Washington University, Ashburn, VA 20147; [f]Department of Biology, University of Nebraska-Kearney, Kearney, NE 68849; [g]Department of Biology, Clark University, Worcester, MA 01610; [h]School of Journalism, Michigan State University, East Lansing, MI 48824; [i]Biological Science, Clark Science Center, Smith College, Northampton, MA 01063; [j]Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS 66045; [k]National Evolutionary Synthesis Center, Duke University, Durham, NC 27705; [l]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611; and [m]Computer Science and Engineering, Texas A&M University, College Station, TX 77843

Edited by David M. Hillis, The University of Texas at Austin, Austin, TX, and approved July 28, 2015 (received for review December 3, 2014)

Reconstructing the phylogenetic relationships that unite all lineages (the tree of life) is a grand challenge. The paucity of homologous character data across disparately related lineages currently renders direct phylogenetic inference untenable. To reconstruct a comprehensive tree of life, we therefore synthesized published phylogenies, together with taxonomic classifications for taxa never incorporated into a phylogeny. We present a draft tree containing 2.3 million tips—the Open Tree of Life. Realization of this tree required the assembly of two additional community resources: (*i*) a comprehensive global reference taxonomy and (*ii*) a database of published phylogenetic trees mapped to this taxonomy. Our open source framework facilitates community comment and contribution, enabling the tree to be continuously updated when new phylogenetic and taxonomic data become digitally available. Although data coverage and phylogenetic conflict across the Open Tree of Life illuminate 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 for community contribution. This comprehensive tree 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, agriculture, and genomics.

phylogeny | taxonomy | tree of life | biodiversity | synthesis

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 reconstructing the tree of life is one of the most daunting challenges in biology. The scope of the problem is immense: there are ~1.8 million named species, and most species have yet to be described (2–4). Despite decades of effort and thousands of phylogenetic studies on diverse clades, we lack a comprehensive tree of life, or even a summary of our current knowledge. One reason for this shortcoming is lack of data. GenBank contains DNA sequences for ~411,000 species, only 22% of estimated named species. Although some gene regions (e.g., *rbcL,* 16S, COI) have been widely sequenced across some lineages, they are insufficient for resolving relationships across the entire tree (5). Most recognized species have never been included in a phylogenetic analysis because no appropriate molecular or morphological data have been collected.

There is extensive publication of new phylogenies, data, and inference methods, but little attention to synthesis. We therefore focus on constructing, to our knowledge, the first comprehensive tree of life through the integration of published phylogenies with taxonomic information. Phylogenies by systematists with expertise in particular taxa likely represent the best estimates of relationships for individual clades. By focusing on trees instead of raw data, we avoid issues of dataset assembly (6). However, most published phylogenies are available only as journal figures, rather than in electronic formats that can be integrated into databases and synthesis methods (7–9). Although there are efforts to digitize trees from figures (10), we focus instead on synthesis of published, digitally available phylogenies.

When source phylogenies are absent or sparsely sampled, taxonomic hierarchies provide structure and completeness (11, 12). Given the limits of data availability, synthesizing phylogeny and taxonomic classification is the only way to construct a tree of life that includes all recognized species. One obstacle has been the absence of a complete, phylogenetically informed taxonomy that spans traditional taxonomic codes (13). We therefore assembled a comprehensive global reference taxonomy via alignment and merging of multiple openly available taxonomic resources. The Open Tree Taxonomy (OTT) is open, extensible, and updatable, and reflects the overall phylogeny of life. With the continued updating of phylogenetic information from

Significance

Scientists have used gene sequences and morphological 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 our knowledge, to apply an efficient and automated process for assembling published trees into a complete tree of life. This tree and the underlying data are available to browse and download from the Internet, facilitating subsequent analyses that require evolutionary trees. The tree can be easily updated with newly published data. Our analysis of coverage not only reveals gaps in sampling and naming biodiversity but also further demonstrates that most published phylogenies are not available in digital formats that can be summarized into a tree of life.

Author contributions: C.E.H., S.A.S., J.G.B., R.C., K. A. Crandall, K.G., D.S.H., L.A.K., R.H.R., D.E.S., T.W., and K. A. Cranston designed research; C.E.H., S.A.S., J.G.B., R.C., L.M.C., K. A. Crandall, J.D., B.T.D., R.G., D.S.H., H.D.L., E.J.M., P.E.M., C.L.O., R.H.R., J.A.R., D.E.S., and K. A. Cranston performed research; C.E.H., S.A.S., J.F.A., J.G.B., R.C., L.M.C., E.J.M., P.E.M., R.H.R., and J.A.R. contributed new reagents/analytic tools; C.E.H., S.A.S., J.G.B., R.C., L.M.C., B.T.D., R.G., D.S.H., H.D.L., C.L.O., J.A.R., and D.E.S. analyzed data; C.E.H., S.A.S., J.F.A., J.G.B., R.C., K. A. Crandall, L.A.K., H.D.L., E.J.M., J.A.R., D.E.S., and K. A. Cranston wrote the paper; J.F.A. conducted user interface development; and K.G. provided graphic design.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The Open Tree of Life taxonomy, the synthetic tree, and processed inputs are available from the Dryad database, dx.doi.org/10.5061/dryad.8j60q.

[1]C.E.H. and S.A.S. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: karen.cranston@gmail.com or eebsmith@umich.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1423041112/-/DCSupplemental.

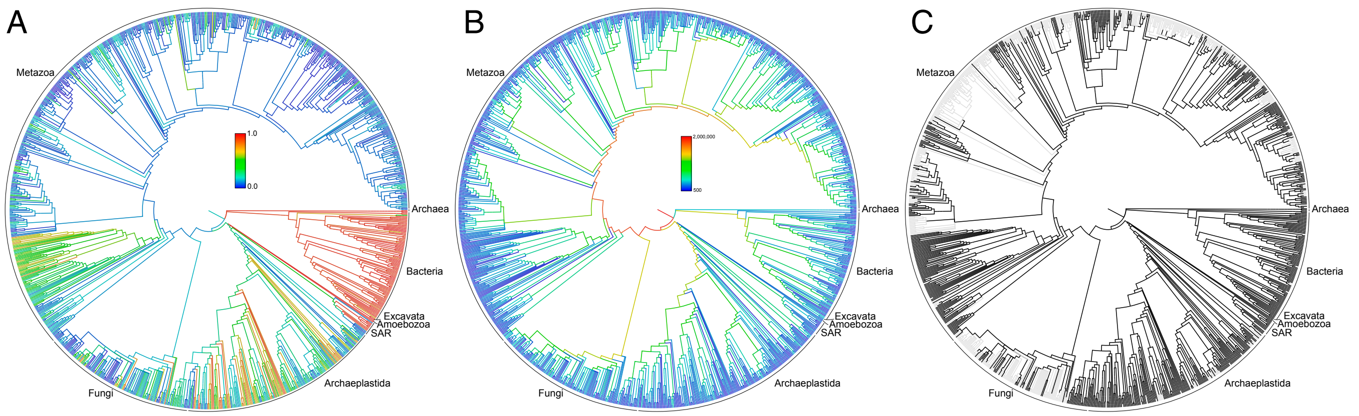

Fig. 1. Phylogenies representing the synthetic tree. The depicted tree is limited to lineages containing at least 500 descendants. (*A*) Colors represent proportion of lineages represented in NCBI databases. (*B*) Colors represent the amount of diversity measured by number of descendant tips. (*C*) Dark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

published studies, this framework is poised to update taxonomy in a phylogenetically informed manner far more rapidly than has occurred historically (see Fig. S1 for workflow).

We used recently developed graph methods (14) to synthesize a tree of life of over 2.3 million operational taxonomic units (OTUs) from the reference taxonomy and curated phylogenies. Taxonomies contribute to the structure only where we do not have phylogenetic trees. Advantages of graph methods include 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, our methodology also highlights the current state of knowledge for any given clade and reveals those portions of the tree that most require additional study. Although a massive undertaking in its own right, this draft tree of life represents only a first step. Through feedback, addition of new data, and development of new methods, the broader community can improve this tree.

Results

Open Tree Taxonomy. To align phylogenies from different sources, the tips, which may represent different taxonomic levels, must be mapped to a common taxonomic framework (14). For synthesizing phylogenetic data, taxonomy also provides completeness and structure where phylogenetic studies have not sampled all known lineages (true of most clades). Available taxonomies differ in completeness and how closely the hierarchy matches known evolutionary relationships. The Open Tree Taxonomy (OTT) is an automated synthesis of available taxonomies, maximizing the number of taxa and preferring input taxonomies that better align to phylogenetic hypotheses in various clades (*Materials and Methods*). It contains taxa with traditional Linnaean names and unnamed taxa known only from sequence data. OTT ver. 2.8 has 2,722,024 OTUs without descendants and includes 382,564 higher taxa; 585,081 of the names are classified as nonphylogenetic units (e.g., *incertae sedis*) and were therefore not included in the synthesis pipeline. The taxonomy is available for download and through online services, including a taxonomic name resolution service for aligning other trees with our taxonomy (see *Data and Software Availability*, below).

Input Phylogenies. We built a user interface for collection and curation of potential trees for synthesis (https://tree.opentreeoflife.org/curator). The complete database contains 6,810 trees from 3,062 studies. At the time of publication, 484 studies in our database are incorporated into the draft tree of life. Our goal is to generate a best estimate of phylogenetic knowledge; based on our tests, we give several reasons not to use all available trees for synthesis. First, including trees that are incorrect does not improve the synthetic estimate. In each major clade, expert curators selected and ranked input trees for inclusion based on date of publication, underlying data, and methods of inference (see *Materials and Methods* for details). These rankings generally reflect community consensus about phylogenetic hypotheses. Second, including trees that merely confirm, or are subsets of, other analyses only increases computational difficulty without significantly improving the synthetic tree. For example, although we have many framework phylogenies spanning angiosperms, we did not include older trees where a newer tree extends the same underlying data. Third, inclusion of trees requires a minimum level of curation; where most OTU labels have been mapped to the taxonomic database, the root is correctly identified, and an ingroup clade has been identified. This information is not in the input file and requires manual curation from the associated publication. Not all trees are sufficiently well-curated; at this point, we have focused curation efforts on trees that will most improve the synthetic tree. The full set of trees in the database is important for other questions such as estimating conflict or studying the history of inference in a clade, highlighting the importance of continued deposition and curation of trees into public data repositories. See Dataset S1 for a list of input trees and metadata and see Fig. S2 for size and scope of input trees.

A Draft Tree of Life. We constructed a tree alignment graph (14), the graph of life, by loading the Open Tree Taxonomy and the 484 rooted phylogenies into a neo4j database. The graph of life contains 2,339,460 leaf nodes (after excluding nonphylogenetic units from OTT), plus 229,801 internal nodes. It preserves conflict among phylogenies and between phylogenies and the taxonomy. To create the synthetic tree, we traversed the graph, resolving conflict based on the rank of inputs, and labeled accepted branches that trace a synthetic tree summarizing the source information. This method allows for clear communication of how conflicts are resolved through ranking, and of the source trees and/or taxonomies that support a particular resolution. The synthetic tree contains phylogenetic structure where we have published trees, and taxonomic structure where we do not. See the *Supporting Information*, including Figs. S3–S6, for details. The tree is available to browse and download, and online services allow extraction of subtrees given lists of species (see *Data and Software Availability*, below).

Coverage. Of the 2,339,460 tips in the synthetic tree of life, 37,525 are represented in at least one input phylogeny, with an additional 4,254 nonterminal taxa represented as tips in phylogenetic inputs (Fig. 1). In Bacteria, Fungi, Nematoda, and Insecta, there is a large gap between the estimated number of species and what exists in taxonomic and sequence databases (Fig. 2). In contrast, Chordata and Embryophyta are nearly fully sampled in

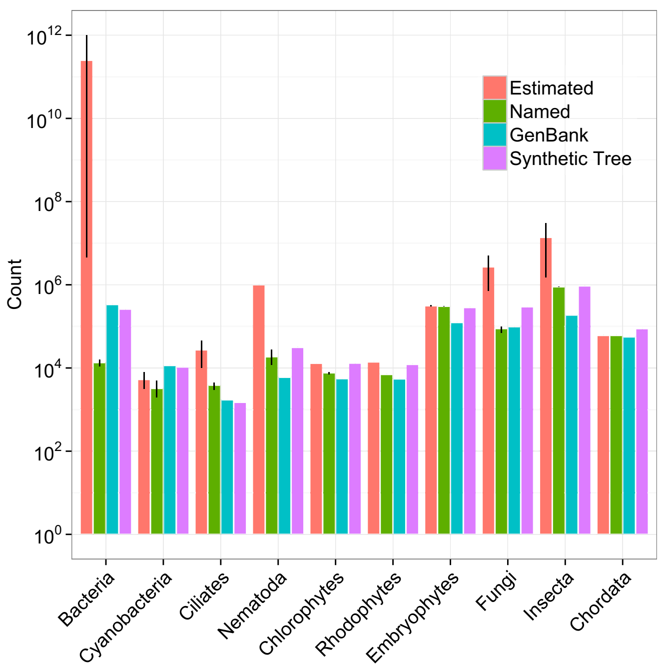

Fig. 2. The estimated total number of species, estimated number of named species in taxonomic databases, the number of OTUs with sequence data in GenBank, and the number of OTUs in the synthetic tree, for 10 major clades across the tree of life. Error bars (where present) represent the range of values across multiple sources. See Dataset S2 for the underlying data.

databases and in OTT (Fig. 2). Poorly sampled clades require more data collection and deposition and, in some cases, formal taxonomic codification and identification to be incorporated in taxonomic databases. Most tips in the synthetic tree are not represented by phylogenetic analyses. The limited number of input trees highlights the need for both new sequencing efforts, additional phylogenetic studies, and the deposition of published tree files into data repositories.

Resolution and conflicts. The tree of life that we provide 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. Fig. 3 depicts the types of alternate resolutions that exist in the graph. 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 either that conflict with the taxonomy (4,610 clades; 3.0%) or where the taxonomy is agnostic to the presence of the clade (18,721 clades; 12.2%). The average number of children for each node in the taxonomy is 19.4, indicating a poor degree of resolution compared with an average of 2.1 in the input trees. When we combine the taxonomy and phylogenies into the synthetic tree, the resolution improves to an average of 16.0 children per internal node. See the *Supporting Information*, including Fig. S7, for details.

Alignment of nodes between the synthetic tree and taxonomy reveals how well taxonomy reflects current phylogenetic knowledge. Strong alignment is found in Primates and Mammalia whereas our analyses reveal a wide gulf between taxonomy and phylogeny in Fungi, Viridiplantae (green plants), Bacteria, and various microbial eukaryotes (Table 1).

Comparison with supertree approaches. There were no supertree methods that scale to phylogenetic reconstruction of the entire tree of life, meaning that our graph synthesis method was the only option for tree of life-scale analyses. To compare our method against existing supertree methods, we used a hybrid MultiLevelSupertree (MLS) (15) plus synthesis approach (*Materials and Methods*). The total number of internal nodes in the MLS tree is 151,458, compared with 155,830 in the graph synthesis tree, although the average number of children is the same (16.0 children per node). If we compare the source phylogenies against the MLS supertree and the draft synthetic tree, the synthesis method is better 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) (16) of the MLS vs. input trees is 31, compared with 15 for the graph synthesis tree. See the *Supporting Information* for details.

Discussion

Using graph database methods, we combine published phylogenetic data and the Open Tree Taxonomy to produce a first-draft tree of life with 2.3 million tips—the Open Tree of Life. This tree is comprehensive in terms of named species, but it is far from complete in terms of biodiversity or phylogenetic knowledge. It does not aim to infer novel phylogenetic relationships, but instead is a summary of published and digitally available phylogenetic knowledge. To our knowledge, this study represents the first time a comprehensive tree of life has been available for any analyses that require a phylogeny, even if the species of interest have not been analyzed together in a single, published phylogeny.

As a result of data availability, data quality, and conflict resolution, there are many areas where relationships in the tree do not match current phylogenetic thinking (e.g., relationships within Fabaceae, Compositae, Arthropoda). This draft tree of life represents an initial step. The next step in this community-driven process is for experts to contribute trees and annotate areas of the tree they know best.

Limitations on Coverage. Many microbial eukaryotes, Bacteria, and Archaea are not present in openly available taxonomic databases and therefore were not incorporated into the Open Tree Taxonomy and the synthetic tree. Most tips in the synthetic tree (98%) come from taxonomy only, reflecting both the need to incorporate more species into phylogenies and the need to make published phylogenies available. We obtained trees from digital repositories and also by contacting authors directly, but our overall success rate was only 16% (9). Many published relationships are not represented in the synthetic tree because this knowledge exists only as journal images. Our infrastructure allows for the synthetic tree to be easily and continuously updated via updated taxonomies and newly published phylogenies. The latter are dependent on authors making tree files available in repositories, such as TreeBASE (17) and Dryad (datadryad.org) or through direct upload to Open Tree of Life (https://tree.opentreeoflife.org/curator), and on having sufficient metadata for trees. We hope this synthetic approach will provide incentive for the community to change the way we view phylogenies—as resources to be cataloged in open repositories rather than simply as static images.

Conflicts in the Tree of Life. The synthetic tree of life is a bifurcating phylogeny (with "soft" polytomies reflecting uncertainty), but some relationships are more accurately described using reticulating networks. The Open Tree of Life contains areas with conflict (Fig. 3). For example, the monophyly of Archaea is contentious—some data-store trees indicate that eukaryotes are embedded within Archaea (18, 19) rather than a separate clade. Similarly, multiple resolutions of early diverging animal (20–23) and Eukaryotic (24–28) lineages have been proposed. Reticulations help visualize competing hypotheses, gene tree/species tree conflicts, and underlying processes, such as horizontal gene transfer (HGT), recombination, and hybridization, which have had major impacts throughout the tree of life [e.g., hybridization in diverse clades of green plants (29) and animal lineages (30), including our own (31), and HGT in bacteria and archaea (32–34)]. The graphical synthesis approach used here naturally allows for storage of conflict and non–tree-like structure, enabling downstream visualization, analysis, and annotation of conflict (Fig. 3) and highlighting the need for additional work in this area.

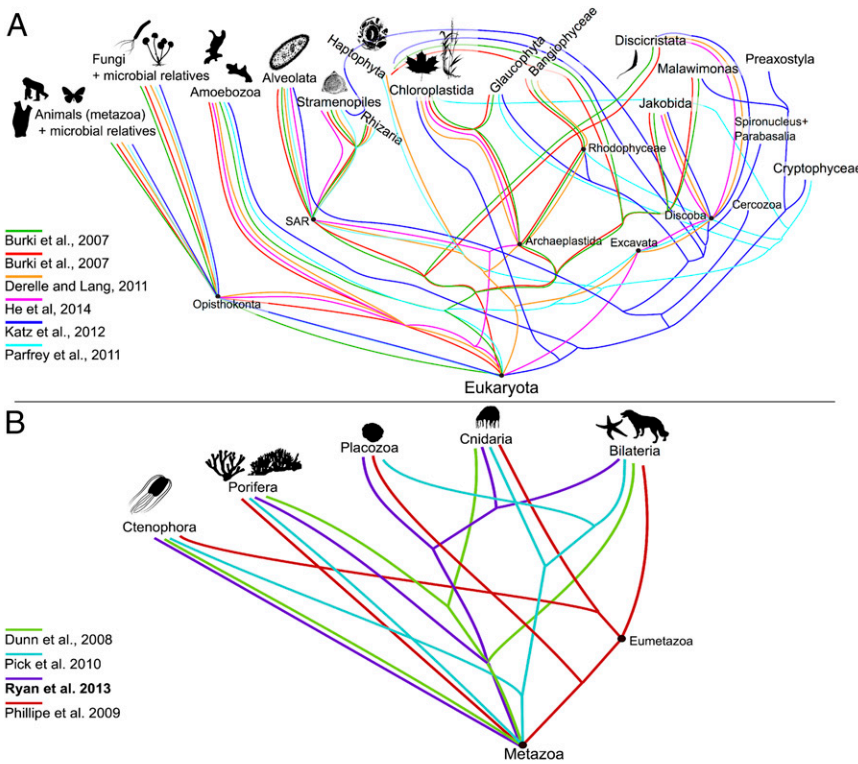

Fig. 3. Conflict in the tree of life. Although the Open Tree of Life contains only one resolution at any given node, the underlying graph database contains conflict between trees and taxonomy (noting that these figures are conceptual, not a direct visualization of the graph). These two examples highlight ongoing conflict near the base of Eukaryota (*A*) and Metazoa (*B*). Images courtesy of PhyloPic (phylopic.org).

Resolving conflict is a challenge in supertree methods, including our graph method. The number of input trees that support a synthetic edge may be considered a reasonable criterion for resolving conflict, but the datasets used to construct each source tree may have overlapping data, making them nonindependent. The number of taxa or gene regions involved cannot be used alone without other information to assess the quality of the particular analysis. Better methods for resolving conflict require additional metadata about the underlying data and phylogenetic inference methods.

Selection of Input Trees. We used only a subset of trees in the database for synthesis, filtering out trees that are redundant, are erroneous, or have insufficient metadata. Our current synthesis method relies on manual ranking of input trees by expert curators within major clades. The potential to automate this ranking, and to use metadata to resolve conflict, depends on the availability of machine-readable metadata for trees; such data currently must be entered manually by curators after reading the publication. Additional metadata would allow a comparison of synthesis trees based on, for example, morphological versus molecular data, the inference method, or the number of underlying genes. Manual curation is time-consuming and labor-intensive; scalability would improve greatly by having standardized metadata (35) encoded in the files output by inference packages (e.g., in NeXML files) (36).

Source Trees as a Community Resource. The availability of well-curated trees allows for many analyses other than synthesis, such as calculating the increase in information content for a clade over time or by a particular project or laboratory, comparing trees constructed by different approaches, or recording the reduction in conflict in clades over time. These analyses require that tips be mapped to a common taxonomy to compare across trees. Our database contains thousands of trees mapped to existing taxonomies through the Open Tree Taxonomy. The data curation interface is publicly available (https://tree.opentreeoflife.org/curator)

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

			Nodes supported by		
Clade	Tips	Internal nodes	Taxonomy	Trees	Trees + taxonomy
Bacteria	260,323	11,028	8,454 (76.7%)	2,184 (19.8%)	390 (3.5%)
Cyanobacteria	10,581	788	678 (86.0%)	83 (10.5%)	27 (3.4%)
Ciliates	1,497	657	654 (99.5%)	1 (0.2%)	2 (0.3%)
Nematoda	31,287	3,504	3,431 (97.9%)	54 (1.5%)	19 (0.5%)
Chlorophytes	13,100	1,267	1,239 (97.8%)	20 (1.6%)	8 (0.6%)
Rhodophytes	12,214	1,292	1,278 (98.9%)	14 (1.1%)	0
Fungi	296,667	8,646	8,243 (95.3%)	383 (4.4%)	20 (0.2%)
Insecta	941,753	88,666	85,936 (96.9%)	2,205 (2.5%)	525 (0.6%)
Chordata	88,434	27,315	13,374 (49.0%)	11,689 (42.8%)	2,250 (8.2%)
Primates	681	501	129 (25.7%)	294 (58.7%)	78 (15.6%)
Mammals	9,539	4,433	1,645 (37.1%)	2,194 (49.5%)	594 (13.4%)
Embryophytes	284,447	32,211	22,400 (69.5%)	8,533 (26.5%)	1,271 (3.9%)

Table 2. Tree metadata, based on the MIAPA checklist (https://github.com/miapa/miapa)

Item	Description	Typically included in tree files	Use by Open Tree of Life
Topology	The topology itself, plus the type of tree (e.g., gene tree vs. species tree, type of consensus tree)	Topology, but not tree type	Yes, topology; tree type used by curators as criteria to rank trees
Root	Whether the tree is rooted, and the location of the root	Tree in file often rooted arbitrarily; different from in manuscript figures	Yes, requires manual checking by curator to match against manuscript
OTU labels	Labels on tips of tree should include (or be mappable to) a meaningful online identifier	Yes, but often do not map to online databases	Tip labels mapped through combination of automated and manual processes
Branch lengths	The length of each branch of the tree, and the units of measurement	Branch length sometimes included; units generally not present	Imported into database when present, but not included on synthetic tree
Branch support	Support values (e.g., bootstrap proportions or Bayesian posterior probabilities)	Often in files, but support type often not specified	Not in algorithm, but curators do examine branch support
Character matrix	The data used to infer the tree, including data type and source (e.g., GenBank accession or specimen)	Sometimes included with tree file, but often without sufficient metadata	Number and type of genes used by curators as criteria to rank trees
Alignment method	Method used to align sequence data	No	No
Inference method	Method used to infer tree from data	Usually no	Inference method used by curators as criteria to rank trees

We note whether the metadata is generally available in the tree file (as opposed to in the text of the article, if at all) and how the data are used by Open Tree of Life.

as is the underlying data store (https://github.com/opentreeoflife/phylesystem).

Dark Parts of the Tree. Hyperdiverse, poorly understood groups, including Fungi, microbial eukaryotes, Bacteria, and Archaea, are not yet well-represented in input taxonomies. Our effort also highlights where major research is needed to achieve a better understanding of existing biodiversity. Metagenomic studies routinely reveal numerous OTUs that cannot be assigned to named species (37, 38). For Archaea and Bacteria, there are additional challenges created by their immense diversity, lack of clarity regarding species concepts, and rampant horizontal gene transfer (HGT) (32, 39, 40). The operational unit is often strains (not species), which are not regulated by any taxonomic code; strain collections are not available to download, making it difficult to map taxa between trees and taxonomy and estimate named biodiversity. Open databases such as BioProject at the National Center for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov/bioproject) have the potential to better catalog biodiversity that does not fit into traditional taxonomic workflows.

Materials and Methods

Input Data: Taxonomy. No single taxonomy both is complete and has a backbone well-informed by phylogenetic studies. We therefore constructed the Open Tree Taxonomy (OTT), by merging Index Fungorum (41), SILVA (42, 43), NCBI (44), Global Biodiversity Information Facility (GBIF) (45), Interim Register of Marine and Nonmarine Genera (IRMNG) (46), and two clade-specific resources (47, 48), using a fully documented, repeatable process that includes both generalized merge steps and user-defined patches (*Supporting Information*). OTT (ver. 2.8.5) consists of 2,722,024 well-named entities and 1,360,819 synonyms, with an additional 585,081 entities having nonbiological or taxonomically incomplete names ("environmental samples" or "incertae sedis"), that are not included in the synthetic phylogeny.

Input Data: Phylogenetic Trees. We designed and developed a user interface that saves phylogenetic trees directly into a GitHub repository (49) and used this interface to import and curate trees. We obtained published trees from TreeBASE (17) and Dryad and by direct appeal to authors. The data retrieved are by no means a complete representation of phylogenetic knowledge because we obtained digital phylogeny files for only 16% of recently published trees (9). Even when available (as newick, NEXUS, or NeXML files or via TreeBASE import), trees require significant curation to be usable for synthesis. We mapped taxon labels (which often include laboratory codes or abbreviations) to taxonomic entities in OTT. We rooted (or rerooted) trees to match figures from papers. Because relationships among outgroup taxa were often problematic, we identified the ingroup/focal clade for the study. For studies with 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 preferred trees to generate prioritized lists. In the absence of structured metadata about the phylogenetic methods and data used to infer the input trees, rankings were assembled by authors with expertise in specific clades and were based on date of publication, taxon sampling, the number of genes/characters in the alignment, whether the specific genomic regions are known to be problematic, support values, and phylogenetic reliability (agreement or disagreement with well-established relationships) (see Table 2 for details). In general, rankings reflect community consensus about phylogenetic hypotheses. As we collect more metadata—such as that described by the Minimum Information for a Phylogenetic Analysis (MIAPA) (35), either by manual entry into the system or by upload of tree files with structured, machine-readable metadata—automated filtering/weighting trees based on metadata will be possible.

Synthesis. The goal of the supertree (or "synthesis") operation is to summarize the ranked input trees and taxonomy (with the taxonomy given the lowest rank). We used an algorithmic approach to produce the synthetic tree rather than a search through tree space for an optimal tree. Given a set of edges labeled with the ranks of supporting trees, the algorithm is a greedy heuristic that tries to maximize the sum of the ranks of the included edges. We summarize the major steps of the method here and provide details in the *Supporting Information*.

The first steps include preprocessing the inputs. We pruned nonbiological or taxonomically incomplete names from OTT and pruned outgroups and unmapped taxa from input trees. Removal of outgroups reduces errors from unexpected relationships among outgroup taxa. Finally, we found uncontested nodes across the taxonomy plus input trees and broke the inputs at these nodes into a set of subproblems. This divide-and-conquer approach shortened running time and reduced memory requirements.

We then built a tree alignment graph (14, 50), which we refer to as the graph of life. Tree alignment graphs allow for representation of both congruence and conflict in the same data structure, allow for nonoverlapping taxon sets in the inputs (as well as tips mapped to higher taxa), and are computationally tractable at the scale of 2.3 million tips and hundreds of input trees. We loaded the taxonomy nodes and edges into the graph, and then each subproblem, creating new nodes and edges and mapping tree nodes onto compatible taxonomy nodes. We also created new nodes and edges that reflect potential paths between the inputs.

Once the graph was complete, generating the synthetic tree involved traversing the graph and preferring edges that originate from high-ranked inputs. We always preferred phylogeny edges over taxonomy edges. Given additional digitized metadata about trees, this system allows for custom synthesis procedures based on preference for inference methods, data types, or other factors.

As a comparison with this rank-based analysis, we also created a synthetic tree using MultiLevelSupertrees (MLS) (15), a supertree method where the tips in the source trees can represent different taxonomic hierarchies. We built MLS supertrees for the largest clades that were computationally feasible and then used these nonoverlapping trees as input into the graph database and conducted synthesis. Due to the lack of taxon overlap between each MLS tree, there was no topological conflict, 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 tree of life is available for browse, comment, and download at https://tree.opentreeoflife.org. All software is open source and available at https://github.com/opentreeoflife. The tree data store is available at https://github.com/opentreeoflife/phylesystem. Where not limited by preexisting terms of use, all data are published with a CC0 copyright waiver. The Open Tree of Life taxonomy, the synthetic tree and processed inputs are available from Dryad (dx.doi.org/10.5061/dryad.8j60q).

ACKNOWLEDGMENTS. We thank Paul Kirk (Index Fungorum), Tony Rees (Interim Register of Marine and Nonmarine Genera), and Markus Doering (Global Biodiversity Information Facility) for taxonomy data and advice on taxonomy synthesis; Mark Holder for discussion, feedback, and software development; Joseph Brown for data collection and curation, software development, data analysis, and writing; Pam Soltis for helpful comments on the manuscript; authors who made their tree files available in TreeBASE or Dryad and tree files that were not otherwise available; curators who imported trees and added metadata. This work was supported by National Science Foundation Assembling, Visualizing, and Analyzing the Tree of Life Grant 1208809.

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 Univ Press, Cambridge, UK).
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(8):e1001127.
3. Costello MJ, Wilson S, Houlding B (2012) Predicting total global species richness using rates of species description and estimates of taxonomic effort. *Syst Biol* 61(5):871–883.
4. Dykhuizen D (2005) Species numbers in bacteria. *Proc Calif Acad Sci* 56(6, Suppl 1): 62–71.
5. Sanderson MJ (2008) Phylogenetic signal in the eukaryotic tree of life. *Science* 321(5885):121–123.
6. Sanderson MJ, McMahon MM, Steel M (2010) Phylogenomics with incomplete taxon coverage: The limits to inference. *BMC Evol Biol* 10(1):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.
8. Magee AF, May MR, Moore BR (2014) The dawn of open access to phylogenetic data. *PLoS One* 9(10):e110268.
9. Drew BT, et al. (2013) Lost branches on the tree of life. *PLoS Biol* 11(9):e1001636.
10. Murray-Rust P, Smith-Unna R, Mounce R (2014) AMI-diagram: Mining facts from images. *D-Lib* 20(11/12). Available at 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(7424):444–448.
12. Bininda-Emonds OR, Sanderson MJ (2001) Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. *Syst Biol* 50(4):565–579.
13. Polaszek A (2010) *Systema Naturae 250: The Linnaean Ark* (CRC, Boca Raton, FL).
14. Smith SA, Brown JW, Hinchliff CE (2013) Analyzing and synthesizing phylogenies using tree alignment graphs. *PLOS Comput Biol* 9(9):e1003223.
15. Berry V, Bininda-Emonds ORP, Semple C (2013) Amalgamating source trees with different taxonomic levels. *Syst Biol* 62(2):231–249.
16. Kupczok A, Schmidt HA, von Haeseler A (2010) Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. *Algorithms Mol Biol* 5:37.
17. 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(6):183.
18. 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(1749):4870–4879.
19. 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 USA* 81(12):3786–3790.
20. Dunn CW, et al. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452(7188):745–749.
21. Ryan JF, et al.; NISC Comparative Sequencing Program (2013) The genome of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution. *Science* 342(6164):1242592.
22. Pick KS, et al. (2010) Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Mol Biol Evol* 27(9):1983–1987.
23. Philippe H, et al. (2009) Phylogenomics revives traditional views on deep animal relationships. *Curr Biol* 19(8):706–712.
24. Burki F, et al. (2007) Phylogenomics reshuffles the eukaryotic supergroups. *PLoS One* 2(8):e790.
25. Parfrey LW, Lahr DJG, Knoll AH, Katz LA (2011) Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc Natl Acad Sci USA* 108(33):13624–13629.
26. Katz LA, Grant JR, Parfrey LW, Gordon Burleigh J (2012) Turning the crown upside down: gene tree parsimony roots the eukaryotic tree of life. *Syst Biol* 61(4):653–660.
27. Derelle R, Lang BF (2012) Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Mol Biol Evol* 29(4):1277–1289.
28. He D, et al. (2014) An alternative root for the eukaryote tree of life. *Curr Biol* 24(4): 465–470.
29. Linder CR, Rieseberg LH (2004) Reconstructing patterns of reticulate evolution in plants. *Am J Bot* 91(10):1700–1708.
30. Dowling TE, Secor CL (1997) The role of hybridization and introgression in the diversification of animals. *Annu Rev Ecol Syst* 28:593–619.
31. Winder IC, Winder NP (2014) Reticulate evolution and the human past: An anthropological perspective. *Ann Hum Biol* 41(4):300–311.
32. Syvanen M (2012) Evolutionary implications of horizontal gene transfer. *Annu Rev Genet* 46:341–358.
33. Nelson-Sathi S, et al. (2015) Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* 517(7532):77–80.
34. Doolittle WF (1999) Phylogenetic classification and the universal tree. *Science* 284(5423):2124–2129.
35. Leebens-Mack J, et al. (2006) Taking the first steps towards a standard for reporting on phylogenies: Minimum Information About a Phylogenetic Analysis (MIAPA). *OMICS* 10(2):231–237.
36. Vos RA, et al. (2012) NeXML: Rich, extensible, and verifiable representation of comparative data and metadata. *Syst Biol* 61(4):675–689.
37. 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(9):e44224.
38. 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(1):38–47.
39. Gilbert C, Cordaux R (2013) Horizontal transfer and evolution of prokaryote transposable elements in eukaryotes. *Genome Biol Evol* 5(5):822–832.
40. Qiu H, Yoon HS, Bhattacharya D (2013) Algal endosymbionts as vectors of horizontal gene transfer in photosynthetic eukaryotes. *Front Plant Sci* 4:366.
41. Index Fungorum. Available at www.indexfungorum.org/. Accessed April 1, 2014.
42. Quast C, et al. (2013) The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res* 41(Database issue):D590–D596.
43. Yilmaz P, et al. (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res* 42(Database issue):D643–D648.
44. Sayers EW, et al. (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 37(Database issue):D5–D15.
45. GBIF (2013) The Global Biodiversity Information Facility: GBIF backbone taxonomy.
46. Interim Register of Marine and Nonmarine Genera (IRMNG). Available at www.cmar.csiro.au/datacentre/irmng/. Accessed January 31, 2014.
47. Hibbett DS, et al. (2007) A higher-level phylogenetic classification of the Fungi. *Mycol Res* 111(Pt 5):509–547.
48. Schäferhoff B, et al. (2010) Towards resolving Lamiales relationships: Insights from rapidly evolving chloroplast sequences. *BMC Evol Biol* 10:352.
49. McTavish EJ, et al. (2015) Phylesystem: A git-based data store for community-curated phylogenetic estimates. *Bioinformatics* 2015:btv276.
50. Chaudhary R, Fernandez-Baca D, Gordon Burleigh J (2015) Constructing and employing tree alignment graphs for phylogenetic synthesis. arXiv:1503.03877 [cs.DS].
51. Bansal MS, Burleigh JG, Eulenstein O, Fernández-Baca D (2010) Robinson-Foulds supertrees. *Algorithms Mol Biol* 5:18.
52. Semple C (2003) Reconstructing minimal rooted trees. *Discrete Appl Math* 127(3): 489–503.
53. Guillemot S, Berry V (2007) *Finding a Largest Subset of Rooted Triples Identifying a Tree Is an NP-Hard Task* (LIRMM, Montpellier, France). Research Report LIRMM-RR-07010.
54. Wilkinson M, Pisani D, Cotton JA, Corfe I (2005) Measuring support and finding unsupported relationships in supertrees. *Syst Biol* 54(5):823–831.
55. Semple C, Steel A (2003) *Phylogenetics* (Oxford Univ Press, Oxford).

EVOLUTION