How and Why to Build a Unified Tree of Life

Emily Jane McTavish,* Bryan T. Drew, Ben Redelings, and Karen A. Cranston

Phylogenetic trees are a crucial backbone for a wide breadth of biological research spanning systematics, organismal biology, ecology, and medicine. In 2015, the Open Tree of Life project published a first draft of a comprehensive tree of life, summarizing digitally available taxonomic and phylogenetic knowledge. This paper reviews, investigates, and addresses the following questions as a follow-up to that paper, from the perspective of researchers involved in building this summary of the tree of life: Is there a tree of life and should we reconstruct it? Is available data sufficient to reconstruct the tree of life? Do we have access to phylogenetic inferences in usable form? Can we combine different phylogenetic estimates across the tree of life? And finally, what is the future of understanding the tree of life?

1. Introduction

Since Linnaeus first set out to categorize life on earth more than 250 years ago,[1,2] biologists have sought to understand species by grouping them into bins and hierarchies. Following Darwin’s publication of “The Origin of Species,”[3] shared common ancestry has been used as the basis for understanding species relationships. Now phylogenetic trees capturing evolutionary relationships among species are essential for biological research. Accurate phylogenetic information is required for analyses such as comparative genomics, measuring selection, biogeography, and epidemiology, among many other applications. As researchers uncover evolutionary history, taxonomic hierarchies and nomenclature are updated to reflect hypotheses of shared ancestry. It is not possible to fully understand evolutionary processes among species or lineages without controlling for their shared histories. With the advent of ever more sophisticated and inexpensive molecular sequencing technologies during the past three decades, researchers have access to unprecedented insights regarding organismal relationships. Despite this revolution in sequence data availability, the evolutionary relationships among many of the species on earth are unstudied, and it is challenging to access accurate phylogenetic relationships even for well-studied species of interest.

We are researchers interested in improving the availability of phylogenetic knowledge across the tree of life. In pursuit of that goal, we collaborate on the Open Tree of Life project. Open Tree of Life (OpenTree) has developed an open, accessible platform for summarizing knowledge about phylogenetic relationships of all species across the tree of life. In 2015 OpenTree published a summary tree comprised of 2.3 million tip taxa (Figure 1).[4] Of these tips, the relationships of 37,525 were gathered from existing evolutionary phylogenetic estimates.[5] The relationships among the rest of the species were based on taxonomic relationships from a combination of source taxonomies.[6,7] Since that publication, nine new drafts of the tree have been published online at tree.opentreeoflife.org, incorporating more phylogenetic information, improvements to the combined taxonomy, and methodological developments in automated assembling of trees and taxonomies. The current draft (v9.1, March 2017) includes 2,617,204 taxa with 55,226 tips from 819 input phylogenies.

Our research in pursuit of a unified “Open Tree of Life” has generated discussion about several key questions:

- Is there a tree of life and should we reconstruct it?
- Can we reconstruct the tree of life from available data?
- Do we have access to inferences about relationships?
- Can we summarize information and conflict about relationships across the tree of life?
- What is the future of understanding the tree of life?

Here we address these questions from our perspective as researchers involved in building the Open Tree of Life.

2. Is There a Tree of Life? Should We Reconstruct It?

When estimating the tree of life, it is worthwhile to first ponder whether there is such a tree to be inferred. While there are good reasons to doubt the notion that a single bifurcating phylogeny is the best way to represent species relationships, we argue that
estimating phylogenetic relationships is still essential to understanding biological processes.

2.1. A Tree of Life Is Not the Only Model

The question of whether a “tree of life” exists has been discussed for decades,[7–10] and the recent publication of genome level phylogenies across many taxa brings enormous quantities of empirical data to bear on this question. These data demonstrate that a single genome often contains regions with divergent evolutionary histories. Processes such as incomplete lineage sorting, horizontal gene transfer, endosymbioses, and the incorporation of viruses into genomes can drive differences in the shared evolutionary history in different parts of genomes.

In some regions of tree space, such as the “anomaly zone,” the most common gene tree topology differs from the true species tree topology due to incomplete lineage sorting.[11] Across three closely related species of *Drosophila*, nearly a third of nucleotide substitutions support each of the three possible topologies of shared ancestry.[12] Gene transfer early in the evolution of eukaryotes, bacteria, and archaea makes determining the root of the tree of life very challenging.[13] Independent of the biological processes causing discordance, gene histories which are not concordant with species level relationships may underly traits of interest, creating issues for estimating convergence. For example, recent multi-locus studies of bat evolutionary history have confidently reconstructed a species tree in which two groups of echolocating bats are not sister to one another.[14] Therefore, based on this species tree, echolocation, a key innovation in bats, appears convergent. However, even if this species level inference is correct, 23% of gene trees support the sister group relationship of the two echolocating groups of bats. Therefore, if echolocation has a single genetic basis, it may have a single origin, which is not concordant with the species tree.[14]

Although this gene tree discordance in bats is consistent with incomplete lineage sorting in the ancestral population, hybridization, and horizontal gene transfer (HGT) can also drive complex patterns of gene tree variation. Up to 18% of *Escherichia coli*’s genome is estimated to have arisen from horizontal gene transfer events from other species.[9] Not only gene ancestry but even gene content can vary widely across closely related bacterial lineages[15] (but see Box 1). Doolittle and Brunet[16] argued that gene transfer and cellular fusion are so common as to invalidate the tree of life concept entirely. Is it therefore better to infer a “graph of life” or “web of life”? Does the variety of evolutionary
Box 1. Vertical relationships are traceable despite HGT in bacteria. Horizontal gene transfer (HGT) can play very important roles in bacterial adaptation, such as allowing antibiotic resistance and virulence genes to spread rapidly across diverse lineages.\(^{[17,48]}\) This lateral transfer of genes can present difficulties for accurate phylogenetic estimation.\(^{[16]}\) However, HGT can be constrained by divergence among lineages, which limits the extent to which it blurs phylogenetic boundaries.\(^{[49]}\) In addition, genes that are gained seem to be labile through evolutionary time, and are readily lost, while mutations tracing vertical phylogenetic signal continue to accumulate in the core genome.\(^{[50]}\)

Recent analyses of HGT across genomes suggest that long term successful transfer of genes across lineages is rare,\(^{[49]}\) leaving phylogenetic structure intact.\(^{[51]}\) These results are consistent with the observation that bacteria do form biologically coherent groups, which would not be expected if horizontal gene flow were rampant and unconstrained.\(^{[52]}\) Although biologists are still struggling to develop a cohesive classification system for bacterial lineages (reviewed in Achtman and Wagner\(^{[52]}\)), some form of phylogenetic relationships can capture key aspects of lineages’ histories.

histories across each genome mean that the idea of a bifurcating species tree should be retired? Is researchers’ “exuberance” for inferring resolved species trees “irrational,” as suggested by Hahn and Nakleh?\(^{[14]}\) We argue that the answer is no.

2.2. A Tree of Life Model Is Useful

Species level relationships provide key evolutionary information that is greater than the sum of individual gene tree relationships. Without a framework of vertical shared ancestry, it is not possible to recognize the importance of other overlaid evolutionary processes. Most traits of interest involve interactions among multiple genes. Understanding species and speciation requires recognizing that species are more than clusters of genes. Therefore, a tree estimate capturing this ancestry is necessary to answer many biological questions.

Evolutionary estimates about lineages provide an essential backbone for recognizing contrasting evolutionary patterns, which can signal processes of interest. Discordant gene trees provide valuable information about the process of evolution, and these discordances are only recognizable in the context of vertical relationships. As Puigbò et al.\(^{[17]}\) eloquently pointed out, “the concept of ‘horizontal genomics’ involves an internal contradiction because the notion of horizontal gene transfer inherently implies the existence of a standard of vertical, tree-like evolution.” And indeed, using this tree-like framework to test alternative hypotheses is essential to the practice of biology.\(^{[18]}\) When Mallet et al.\(^{[10]}\) argue that “species tree signal . . . may only be represented by a small fraction of genes,” they highlight the interesting and potentially very important introgression dynamics of gene lineages by contrasting these histories with the species level history of bifurcations. Even in microbial lineages, where HGT is particularly prevalent, there is still a true underlying bifurcating process, as cells divide, and give rise to descendants. That evolutionary history is often still traceable despite ongoing HGT (Box 1).

Although deep interrogation of individual gene lineages is essential for understanding the trajectory of certain traits with simple genetic bases, there are many other evolutionary dynamics which can only be understood by tracing the emergent properties of populations and species through time, that is, a species tree. Species trees explain the distribution of gene trees across genes, even when the majority of gene trees differ from the species tree. Most traits of interest to biologists are complex, and involve many interacting genes. In these cases, the trajectory of individual gene lineages is insufficient to capture evolutionary history. For sexually reproducing organisms, the species tree captures which lineages could be found within a single individual. As epistasis across genes drives the evolution of many complex traits,\(^{[19]}\) treating gene lineages individually cannot capture these evolutionary trajectories.

Without tracing species and group level relationships we cannot make inferences about diversity and divergence. Although across many species it is possible that no single gene accurately captures the process of diversification, a tree capturing this vertical process describes the evolutionary context in which genes across the genome evolved. The processes of diversification and divergence are species level processes. Even in cases where there is no conflict between gene tree and species tree relationships, splits in gene trees will necessarily pre-date species divergence times.\(^{[20,21]}\) Therefore, divergence times will be underestimated if species level processes are not considered. Speciation itself cannot be recognized on a genic basis, as it is an emergent property of groups of organisms which can only be recognized across groups of genes and individuals. Conservation and land management decisions need to be made for evolving populations or species. It is not feasible to conserve or manage individual gene lineages. Conservation decisions are made for populations and the individuals that comprise them. Identifying groups of individuals using species level trees for taxonomic naming allows biologists to describe, recognize, and manage these evolving populations.

Abandoning the search for these group level relationships in favor of acquiescence of a general tangle of gene trees about which no higher order statements may be made is precipitous, and would make recognition of novel biological processes very challenging. The concept of “Heliconious” butterflies, which Mallet et al.\(^{[10]}\) point out have an interesting and complex history of introgression dynamics, relies on the idea that groups of organisms may, at some level, be differentiated form one another. While these authors describe regions of the tree where introgression is so common as to make inference of true bifurcating relationships challenging, they do not disregard the concept of describing lineages based on deeper level phylogenetic divergences.

Given that inferences of evolutionary relationships are necessary to understand key biological processes, we must then ask the pragmatic question of whether making these inferences accurately is possible.
3. Can We Accurately Estimate Phylogenetic Relationships?

While some studies seek to make inferences bridging the deepest relationships in the tree of life (e.g., Raymann et al.[22]), most systematic studies are taxonomically constrained. The recent genomic revolution has provided a surfeit of data with which to analyze the evolutionary relationships within these groups of interest. Despite these massive quantities of data major questions remain unresolved. Where data are available, unresolved relationships may be due to limitations of signal in data or limitations in our models' ability to capture the signal that exists.[23] In addition, wide reaches of the diversity of life we wish to understand remain un-sampled. Nonetheless, most relationships for sampled taxa can easily be reconstructed, and our sampling is improving rapidly.

3.1. Phylogenetic Inference Can Be a Hard Problem

Some phylogenetic questions have proven difficult to conclusively answer even when whole genome data is available across the taxa of interest. For example, the question of whether ctenophores or sponges are the sister lineage to all other extant animals has been confidently answered in several different and conflicting ways.[24–26] even with the availability of genomic data for these lineages. The source of these discrepancies remains unresolved. It is possible that the data required to resolve these relationships no longer exists. The combination of ancient incomplete lineage sorting and millions of years of subsequent mutation may have obscured any signal of species relationships. If so, even if the relationships are tree like, we may not even be able to reconstruct them. Alternatively, genomic data may hold sufficient information, but this may be a region of the tree in which better models, not more data, will be required for confidence in a single reconstruction. Pisani et al.[27] demonstrated that models that take into account heterogeneity in the process of evolution across sites in an alignment can improve accuracy of phylogenetic estimation, even without additional data. However, the best approach to discovering the relationships at the root of metazoans continues to be debated.[28,29] Even if this specific question is a solvable problem, there will always be regions of the tree of life where divergences occurred too rapidly, or key signatures of vertical processes have been overwritten by HGT or subsequent mutation. In regions of the tree where network-like processes overwhelm the signal of vertical relationships, bifurcating trees cannot be accurately resolved. However, as data quality and quantity increase, and as we develop appropriate models to fully exploit genomic data, these recalcitrant regions will comprise a diminishing proportion of relationships of interest. Molecular phylogenetics has created upheaval in many regions of the tree of life, but with time and further analyses consensus begins to emerge.

While the branches most challenging to correctly estimate in the tree of life, such as the sister group relationships at the root of metazoa,[27] the deep relationships among major arthropod groups,[30] or the root of angiosperms,[31] are often of great interest to phylogeneticists most of the relationships across the tree of life are readily revealed. Despite the recent genomic revolution, most phylogenetic estimates are based on one or a few loci, or on morphological datasets. Genomic data will overturn these estimates in some cases, but in most we can anticipate that relationships reflected by those few loci will be supported by additional data, as branches with high levels of incongruence are rare.

3.2. Many Taxa Are Un-Sampled or Un-Named

In contrast to these well-studied metazoan lineages where genomic data is improving inferences and we are getting closer to revealing evolutionary relationships, there are vast swaths of the microbial world about which we know virtually nothing. Even after many years of biodiversity sampling and phylogenetic inference, we still lack information about not only the phylogenetic positions of most taxa, but the very existence of most taxa. This knowledge gap is particularly pronounced for microbial taxa which cannot be cultured in the lab. Hug et al.[32] recently published a study suggesting the existence of a vast candidate phylum of bacteria that have been captured only by environmental meta-genomic sequencing. Environmental sequencing is opening windows through which it is possible to glimpse the enormous diversity of previously unrecognized lineages, and much more is still unknown. Across undescribed microbial taxa several issues interact to make summarizing these relationships extremely challenging. In many cases when describing species from environmental sampling, it is difficult to reconstruct which genes are found in the same individuals, although developments in single cell genomics may ameliorate this problem in future.[33] In addition, it is not straightforward to recognize when individual representatives of the same lineages are found across studies, as only sequence similarity can provide a diagnosis. This makes merging inferences across studies difficult to impossible, unless the sequence data sets themselves can be combined and jointly analyzed. Improved phylogenetic methods and models will not be sufficient to capture the evolutionary relationships among these diverse lineages. Revolutions in sequencing technology and data processing, as well as improvements in breadth of sampling and taxonomy are required to begin to grasp the diversity and evolutionary relationships of these groups.

However, in contrast to these areas of ongoing research, there are well-sampled and well-studied parts of the tree of life where species relationships are readily described by a tree, and many, many publications have described these phylogenetic relationships. In order to summarize these relationships, and to capture concordance and conflict across these published studies on evolutionary relationships, we need to be able to access these prior inferences.

4. Do We Have Access to Inferences About Relationships?

Great progress has been made on phylogenetic inference in recent decades, however access to these inferences is limited. Taxonomy is often the only information that is readily available about a species’ evolutionary context. Taxonomy attempts to
capture broad strokes of evolutionary relatedness, but these groupings are highly unresolved and often do not reflect a full evolutionary analysis. The problem is not only a lack of data required for phylogenetic inference, it is difficult to access information even from phylogenetic trees that have been inferred and published. Many phylogenies are only available in the form of an image in a pdf, which is not easily reusable for further analyses. In addition, labeling of taxa is often inconsistent across studies. For phylogenies to be compared, each taxon must be given the same name in different published phylogenies, or names representing the same groups must be somehow associated with one another. Neither this phylogenetic data availability problem, nor the taxonomic name resolution problem are trivial to resolve.

4.1. Existing Phylogenetic Data Is Not Available

Although sequence data is available for \( \approx 22\% \) of named species in GenBank, these sequences are scattered across genomes (nuclear, mitochondrial, and plastid) and cannot be directly used to produce a representative phylogenetic tree. Therefore, a critical component for developing a comprehensive synthetic phylogeny of all life is access to published phylogenetic trees. In Hinchliff et al. we used a synthesis technique that combined information from hundreds of existing phylogenetic trees with several taxonomic databases to create a single synthetic tree summarizing relationships across 2.3 million taxa. However, less than 10% (\( \approx 37,500 \) out of \( \approx 411,000 \)) of sequenced species available in GenBank were included in this synthetic tree. While many or most of the other 90% of sequenced species may have been included in phylogenetic estimates, most previously inferred phylogenetic trees are not publicly available. This lack of data availability precludes both estimates of how many species have been included in phylogenies, as well as the incorporation of those estimates into a unified phylogenetic summary.

Previously, it has been shown that less than 17% of published phylogenies are available via online repositories and that much of the data that is available publicly is archived in a manner that makes reuse difficult or impossible. Requesting phylogenetic tree files directly from corresponding authors is successful only about 16% of the time. However, since the above publications, there has been an increased push in the scientific community to make published data more available via publicly available archives. Nonetheless, deposition rates have not improved (see Box 2).

The result of evolutionary estimates not being readily available is that taxonomy is often used as a proxy for evolutionary relatedness (e.g., Tree of Sex Consortium). Taxonomy often captures the broad strokes of relationships, but it does not provide fine scale resolution. This is one of the key problems that we are seeking to solve with the OpenTree project.

4.2. Reconciling Names Is Difficult

Even when trees are publicly archived in some digital format, for the phylogenetic data to be reusable tip labels must be associated with standardized taxon identifiers. In order to recognize representatives of lineages across studies a unified naming scheme is needed. This unified naming structure allows us to make general statements about lineages, despite having sampled only a subset of individuals from those groups. The concept that data gathered from a subset of individuals may be used to represent a broader group is essential to the practice of biology.

Assigning samples included in phylogenies to recognizable taxonomic units can be challenging even in well circumscribed species, due to long histories of species name changes and revisions, combined with typological and spelling errors. Genus and species names are repeated across regions of the tree governed by different taxonomic codes. These repeated names need to be disambiguated when tracking identifiers across these deep evolutionary divisions. Nonetheless, for clearly delineated species, taxonomic name resolution is possible. To reliably reconcile names across the tree of life OpenTree has merged several large taxonomy resources available online to create a unified taxonomy. We have also developed a user friendly web-browser curation interface where using a combination of label matching and human curation, taxon names in published trees are mapped to unique identifiers. This curated data store serves as the back end for the OpenTree synthetic summary tree analyses, and provides open access to well curated phylogenetic estimates.

However, the mapping of names across studies breaks down almost entirely when considering new lineages discovered from environmental sequencing. The discovery of novel bacterial and other microbial taxa is already straining traditional taxonomic resources, and the explosion of sampling of new lineages will require updates to the procedures for naming species and their inclusion in taxonomies. Synthesis of phylogenetic knowledge does not require that each new taxon has a traditional Linnaean name, but does require that we know how to map tips of trees from one publication to tips in another publication, and whether tips represent individual organisms or distinct lineages. Taxa found in multiple studies can be recognized through sequence similarity only when homologous sequence data is available. Capturing the diversity of microbial life is an area of active research by many groups, and the next decades are certain to provide a revolution in our understanding of the tree of life. In the context of the Open Tree of Life, tracing lineages that do not fit neatly into existing taxonomies is still problematic, and we are searching for better ways to synthesize this novel information as it is published.

5. Can We Combine Different Phylogenetic Estimates Across the Tree of Life?

The OpenTree Project is working to capture consensus and conflict by summarizing phylogenetic information across the tree of life. Generating a combined summary of estimates of species relationships across the tree of life requires a two-part process. The first step is reconciling names across different publications, as described above, and the next step is to merge phylogenetic estimates. We have generated an automated procedure to rapidly merge many phylogenetic estimates and a unified taxonomy into a summary tree. This summary captures
Box 2. Case study: The results of phylogeny inference have not become more accessible. Why not? To investigate whether data deposition has improved since 2013, when Drew et al.\cite{36} found that only \( \approx 17\% \) of published phylogenetic data was digitally available, we examined 100 phylogenetic studies published in 2015. We used the search term “phylogeny” in Google Scholar and chose the first 100 studies that contained original phylogenetic data for examination. These 100 studies spanned all major lineages from the tree of life. Our search found that only 20% of the surveyed studies had deposited their alignment or tree data in publicly available repository such as Dryad, Github, Treebase, or Zenodo. This number of 20% is only slightly higher than the 16.7% (1262 out of 7539) reported by Drew et al.\cite{36} which examined phylogenetic publications from 2000 to 2012 and included animals, fungi, seed plants, microbial eukaryotes, archaea, and bacteria. Our results indicate that the rate of publicly archiving published phylogenetic data has not appreciably changed since 2012. Indeed, the 20% rate we recovered here is actually not much different than in 2000.\cite{36} Even more concerning, Google Scholar search results are based on citations, so the 100 publications we surveyed were relatively well cited and typically from high impact factor journals, which have a higher tendency to require the deposition of data into a public repository.\cite{38} It is worth considering potential reasons why researchers do not archive their phylogenetic data in an accessible way.

Since 2011 National Science Foundation has required a “data management plan” as part of grant applications. The standard in the field has long been to publish and share sequence data. This sequence data, at least historically, was the most hard-won data product of the study. Often the alignments from analyses are available online, but the trees are not. Are trees themselves considered “data” from a study? The NSF’s data management plan requirements state that “what constitutes such data will be determined by the community of interest.”\cite{53} While some members of the phylogenetic community agree that phylogenies are key data products,\cite{34} the continued lack of inclusion of phylogenetic inference in accessible data stores suggests that not all practitioners do.

Why, after so much expertise and computational effort to generate these estimates of evolutionary relatedness, would researchers not also share that information in a reusable way? One reason is that researchers generating trees may not see a need for re-use of trees. This is potentially connected to a common mindset in the field of systematics: that if you need a tree to answer a question, you should build it yourself so as to be sure it is correct. Indeed, there has been push-back against several innovations devoted to making generating or acquiring phylogenetic trees in a “black box” manner. Building phylogenetic trees as needed may serve to provide researchers with the inferences and taxon sampling most appropriate for their questions, but it is a major hurdle to research progress to require that any ecologist wishing to consider the relationships among species in their study area must either become proficient in phylogenetics, or convince a phylogeneticist to collaborate on their project.

information about concordance and conflict across the input trees.

In the past, within taxonomic groups, synthetic summary trees have been created by stitching together multiple trees; for example, taking a backbone tree where each tip is a genus, and grafting on separately estimated trees for each genus. However, this kind of ad hoc approach is not easily automated and cannot be applied to general collections of trees. For the Open Tree of Life we have developed a synthesis procedure to merge and combine information from any collection of phylogenetic trees.\cite{44} While much discussion of supertrees has been in the context of the supertree-versus-supermatrix debate,\cite{45} our strategy for summarizing information across the tree of life differs from these earlier approaches. Our method is referred to as a supertree because the inputs are phylogenies as opposed to data matrices, but it does not treat those input phylogenies as a substitute for data matrices. Our supertree is not designed to construct a phylogeny that is more accurate than the input trees, but instead summarizes and merges information from input phylogenies in a transparent fashion. This supertree provides a summary of phylogenetic knowledge at a broader scale than any individual phylogeny, and more accurately than the taxonomy. Our procedure constructs an annotation file that records conflicts between each input tree edge and the summary supertree. These annotations are used in the online tree browser, where conflict between the input trees and the supertree is visually represented.

The approach is designed to generate a summary supertree with four basic properties.\cite{44} First, every edge of the supertree must be supported by an edge in one of the input phylogenies or the taxonomy. Second, our summary supertree should have no unnecessary polytomies. Third, conflict is handled by ranking the input phylogenies. A branch in a higher rank tree will be included in the supertree preferentially to a branch in a lower ranked input. These rankings allow curators with domain specific knowledge to adjust problematic regions of the supertree by altering the ranking of input trees. The taxonomy is always ranked last, and thus never overrides input phylogenies. The input phylogenies can overrule or refine the relationships suggested by the taxonomy. Fourth, we seek to represent as many relationships from input-phylogenies as possible. However, this final criterion is not fully optimized by our current approach.

Version 3.0 (March 2017) of the OpenTree Taxonomy has approximately 2.6 million tips that are suitable for use in supertree construction. To make computation tractable we...
construct a pruned taxonomy tree by removing taxa that are not found in any input trees. This shrinks the problem size from 2.6 million tips to about 55 000 tips. We use the taxonomic relationships to add taxa that are not found in any input phylogeny back to the grafted solution. This procedure decreases computational running time of creating a tree at this scale from several hours to eight minutes. The completed tree is available at tree.opentreeoflife.org. This online resource allows anyone to easily access an evolutionary estimate of relationships for any set of taxa at any time.

6. What Is the Future of Understanding the Tree of Life?

Continued infrastructure development and improved community engagement will be necessary to achieve our goal of providing easy access to information about evolutionary relationships across the tree of life.

6.1. Bioinformatic Infrastructure Development is Still Needed

The Open Tree of Life project continues to move forward in the curation and synthesis of phylogenetic estimates into a unified tree of life. We are developing several important additional components which were not included in the original synthesis tree including conflict visualization, automated updating of phylogenies, and branch length inferences on the summary supertree topology. The rate of sequencing has outpaced the ability of researchers to analyze sequence data. This means that many species are not included in any phylogeny, despite available sequence data. Researchers in many fields require phylogenies in order to address their questions of interest. However, these researchers are not specialists in phylogenetics, which means that they need to either devote time to understanding and applying complex phylogenetic methods, or they apply outdated and potentially misleading approaches. There is a real opportunity to take the next step with this infrastructure to provide active and automated extension of the tree of life with inferences made directly from sequence data. Integrating automatic updating of phylogenies into the Open Tree of Life will serve to shorten the lag time between sequencing and phylogenetic inference.

Currently the OpenTree summary supertree contains information only about topology, and does not include branch lengths. Many downstream users of phylogenetic estimates require not only inference of evolutionary relatedness, but also estimates of timing of divergence between groups. Integrating this information across disparate taxonomic groups with stark differences in fossil records, and combining dating estimates across trees built using different loci, or even different data types is a significant challenge. Nonetheless, it is a challenge that must be resolved for the Open Tree of Life to serve the needs of downstream users. Overall, improving the machine readability of trees and metadata output from phylogenetic inference software would help greatly in re-usability by OpenTree and other summary and analysis programs.

6.2. Community Engagement is Essential

While there has been widespread adoption of sequence databases and sharing of the raw data on which phylogenies are built, the products of these analyses are seldom available in a reusable format. This results in widespread duplication of effort, and researchers performing analyses without access to the most up to date phylogenetic hypotheses. Going forward, the biological community should strive to be more efficient in compiling and sharing both raw data and inferences derived from the data. In 2013, Drew et al. suggested that phylogenetic journals should adopt policies requiring that phylogenetic data be publicly archived prior to publication (Box 2). The comprehensive synthetic tree of life presented by Hinchliff et al. was an achievement, but only 2% of the species in the tree were represented by phylogenetic data. One approach to improving data deposition would be for all phylogenetic journals to implement strict data archiving policies and for reviewers and editors to ensure data publication as part of the review process.

These phylogenetic inferences are not only a result, but also an essential input in future analysis. Mandates by journals and/or funding organizations is one approach that has worked in past to improve data sharing and data deposition. Another option is to provide services in exchange for data sharing to generate an incentive for researchers to curate and upload their data. We have developed and are developing these incentives in the context of the OpenTree curation tool. As part of capturing conflict in phylogenetic estimates, we are also rolling out the infrastructure for users to generate their own synthetic phylogenetic estimates dynamically, based on their own rankings of input trees. By adding the functionality for researchers to build domain specific trees, we will motivate tree deposition and curation, and provide an additional service to users. This, in addition to the conflict visualization that provides a straightforward mechanism to explore how a new phylogeny fits into existing taxonomic and phylogenetic knowledge will encourage users to share their data. Hopefully these incentives will help to improve the low rates of data deposition.

7. Conclusions and Outlook

Evolutionary relationships are complex, and a single bifurcating tree cannot fully capture the intricacy of the evolutionary process. However, from a pragmatic perspective, a tree does capture the major relationships among organisms. Biologists have been successful at resolving many of the important branches, and are improving in their abilities to make accurate inferences about recalcitrant relationships. Biological progress cannot proceed without at least a hypothesis of the underlying bifurcating relationships, against which to test and understand alternative processes. As Dobzhansky explained in 1973, "Nothing in biology makes sense except in the light of evolution. From an applied perspective that translates to the idea that nothing in biology makes sense except in the context of phylogenetics. Without that overall evolutionary perspective, and the phylogenetic inferences
which describe the ancestry of organisms, biological research cannot proceed. Creating and openly sharing a unified tree of all life provides researchers access to evolutionary context across the diversity of life.

Acknowledgements

We thank Jessica Blois, Shilpa Khatri, and Mark Holder for discussion and ideas, and Rebecca McTavish and Dan Edwards for detailed comments on the manuscript.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

open science, evolution, data deposition, phylogeny, tree of life

Received: June 30, 2017
Revised: August 27, 2017
Published online: October 5, 2017

[1] C. von Linné, Species Plantarum Exhistentes Platas Rite Cognitas ad Genera Relatas Cum Differentis Specificis, Nominibus Trivialibus (Etc.). Vol. 2, Ed: Aucta Vi, Nauck, Berolfini, Nauck, 1753.
[2] C. von Linné, Systema Naturae Per Regna Tria Naturae Secundum Classes, Ordines, Genera, Species., ., Vol. 1, Impensis Georg, Emanuel Beer, 1788.
[3] C. Darwin, The Origin of Species by Means of Natural Selection; The Descent of Man and Selection in Relation to Sex: Or The Preservation of Favoured Races in the Struggle for Life, Ed. Murray J., John Murray, London, 1859.
[4] C. E. Hinichiff, S. A. Smith, J. F. Allman, J. G. Burleigh, R. Chaudhary, L. M. Coghill, K. A. Crandall, J. Deng, B. T. Drew, R. Gazis, K. Gude, D. S. Hibbett, L. A. Katz, H. D. Laughinghouse, E. J. McTavish, P. E. Midford, C. L. Owen, R. H. Ree, J. A. Rees, D. E. Soltis, T. Williams, K. A. Cranston, Proc. Natl. Acad. Sci. 2015, 112, 12747.
[5] E. J. McTavish, C. E. Hinichiff, J. F. Allman, J. W. Brown, K. A. Cranston, M. T. Holder, J. A. Rees, S. A. Smith, Bioinformatics 2015, 31, 2794.
[6] J. Rees, K. Cranston, Biodivers. Data J. 2017, 5, e12581.
[7] C. R. Woese, Proc. Natl. Acad. Sci. 2000, 97, 8392.
[8] E. Pennisi, Science 1999, 284, 1305.
[9] J. G. Lawrence, H. Ochman, Proc. Natl. Acad. Sci. 1998, 95, 9413.
[10] J. Mallet, N. Besansky, M. W. Hahn, BioEssays 2016, 38, 140.
[11] J. H. Degnan, N. A. Rosenberg, PLoS Genet. 2006, 2, e68.
[12] D. A. Pollard, V. N. Iyer, A. M. Moses, M. B. Eisen, PLoS Genet. 2006, 2, e173.
[13] F. Forterre, H. Philippe, BioEssays 1999, 21, 871.
[14] M. W. Hahn, L. Nakhlke, Evolution 2016, 70, 7.
[15] F. Baumdicker, W. R. Hess, P. Pfaffhuber, Genome Biol. Evol. 2012, 4, 443.
[16] F. W. Doolittle, T. D. P. Brunet, PLoS Genet. 2016, 12, e1005912.
[17] P. Puigbò, Y. I. Wolf, E. V. Koonin, BMC Biol. 2013, 11, 46.
[18] K. E. Omrland, L. G. Cook, M. D. Crisp, BioEssays 2008, 30, 854.
[19] A. L. Tyler, F. W. Asselferbs, S. M. Williams, J. H. Moore, BioEssays 2009, 31, 220.
[20] J. H. Gillespie, C. H. Langley, J. Mol. Evol. 1979, 13, 27.
[21] J. E. McCormack, J. Heled, K. S. Delaney, A. T. Peterson, L. L. Knowles, Evolution 2011, 65, 184.
[22] K. Raymann, C. Brocher-Armanet, S. Gribalbo, Proc. Natl. Acad. Sci. 2015, 112, 6670.
[23] J. M. Brown, Syst. Biol. 2014, 63, 334.
[24] E. S. Chang, M. Neuhof, N. D. Rubinstein, A. Diamant, H. Philippe, D. Huschon, P. Cartwright, Proc. Natl. Acad. Sci. 2015, 112, 14912.
[25] G. Jéky, J. Paps, C. Nielsen, EvoDevo 2015, 6, 1.
[26] N. V. Whelan, K. M. Kocot, L. L. Moroz, K. M. Halanych, Proc. Natl. Acad. Sci. 2015, 112, 5773.
[27] D. Pisani, W. Pett, M. Dohrmann, R. Feuda, O. Rota-Stabelli, H. Philippe, N. Lartillot, G. Wörheide, Proc. Natl. Acad. Sci. 2015, 112, 15402.
[28] M. L. Borowiec, E. K. Lee, J. C. Chiu, D. C. Plachetzki, BMC Genom. 2015, 16, 987.
[29] P. Simion, H. Philippe, D. Baurain, M. Jager, E. Quinonez, A. Ereskovsky, P. Läppebä, E. Corre, F. Delsuc, N. King, G. Wörheide, M. Manuel. Curr. Biol. 2017, 27, 958.
[30] J. Caravas, M. Friedrich, BioEssays 2010, 32, 48.
[31] A. Stoltzfus, B. O’Meara, J. Whitacre, R. Mounce, E. L. Gillespie, S. Kumar, D. F. Rosauer, R. A. Vos, BMC Res. Notes 2012, 5, 574.
[32] D. G. Roche, R. Lanfear, S. A. Binning, T. M. Haff, L. E. Schweitz, K. E. Cain, H. Kokko, M. D. Jennions, L. E. B. Kruuk, PLoS Biol. 2014, 12, e1001779.
[33] T. H. Vines, R. L. Andrew, D. G. Bock, M. T. Franklin, K. J. Gilbert, N. C. Kane, J.-S. Moore, B. T. Moyners, S. Renaut, D. J. Rennison, T. Veen, S. Yeaman, FASEB J. 2013, 27, 1304.
[34] B. Fecher, S. Friesike, M. Hebing, PLoS ONE 2015, 10, e0118053.
[35] L. C. Pope, L. Liggins, J. Keyse, S. B. Carvalho, C. Riginos, Mol. Ecol. 2015, 24, 3802.
[36] A. F. Magee, R. M. May, R. Moore, PLoS ONE 2014, 9, e110268.
[37] B. D. Redelings, M. T. Holder, PeerJ 2017, 5, e3058.
[38] J. Gatesy, M. S. Springer, Phylogenetic Supertrees, Number 4 in Computational Biology, Ed: Bininda-Emonds O. R. P., Springer, Netherlands, 2004, pp. 369–388.
[39] T. Dobzhansky, Am. Biol. Teach. 1973, 35, 125.
[40] M. A. O. Meara, J. Whitacre, R. Mounce, E. L. Gillespie, S. Kumar, D. F. Rosauer, R. A. Vos, BMC Res. Notes 2012, 5, 574.