Collaborative Research: Automated and community synthesis of the tree of life

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 elucidating the phylogenetic relationships of all species - building the complete tree of life - has since emerged as one of the grandest and most daunting scientific challenges ever undertaken. The scope of the problem is immense: current estimates of the number of species range from 1.8 million [2] to 8.7 million [3], but these largely ignore microbial diversity, discovery of which is increasing rapidly with environmental sequencing [4]. In recent decades, much progress has been made on resolving the tree, and today a vibrant community of systematists continues to generate new phylogenetic knowledge at all depths of ancestry. However, despite 150 years of effort, 55 AToL projects, and numerous other projects, we lack a comprehensive tree of life. In this proposal, we describe a plan to establish a community-driven, continually updated estimate of the entire tree, and develop new software tools and methods for merging and sharing data. By engaging the systematics community with these resources, our overarching goal is to cultivate ongoing synthesis on a large scale, in a manner that will transform current cultural norms in the field.

Our motivation stems from two primary observations about current practices in systematics that inhibit synthesis. First, research efforts are disjoint, with grants and publications generally focusing on isolated branches. More importantly, the results of such research - new knowledge about the tree of life - are overwhelmingly disseminated solely as static images and text in the pages of journal articles, despite the availability of public data repositories such as TreeBASE [5] and Dryad [6]. In short, from a veritable fire hose of systematic knowledge, almost none is captured in digital reservoirs, and very little is synthesized across clades, notably because the systematics community lacks compelling means and incentives for sharing and re-using data for the purpose of large-scale synthesis.

The time is ripe to overcome these limitations. The broader scientific community has increasingly come to appreciate the value of phylogeny - that it provides a mechanism through which to interpret the patterns and processes of evolution and to predict the responses of life to rapid environmental change. Just as sequencing the human genome provided numerous, largely unanticipated new insights, reconstructing the entire tree of life will similarly fuel fundamental research on the nature of biological diversity and its relationship to our well-being. Phylogenies and phylogenetic methods are now being used to enhance agriculture [7], identify and combat diseases (in humans and crops [8, 9], conserve biodiversity [10, 11], and predict responses to global climate change [12, 13] and to biological invasions [14].

The novelty of this proposal is the focus on synthesis of published phylogenetic data. It rests on the premise that producing a comprehensive, continually updated tree of life must be a community-based endeavor. To this end, we have four primary objectives:

1. Build and make publicly available the first complete draft tree of life
2. Engage the community in refinement and annotation of the draft tree
3. Promote a culture of data sharing among systematists
4. Develop, disseminate, and evaluate novel methods for phylogenetic synthesis

We will meet these objectives through a combination of data synthesis, software development, community outreach, and methodological research. Our philosophy throughout this project is to be as open as possible, developing software under open-source licenses, providing the same access to data and

methods to users as to members of the project team, allowing for both browser-based and programmatic access to data and maintaining a constantly-updated online presence of the project status.

A. Building a comprehensive draft tree of life

In the first year of the project we will assemble a first-draft version of a comprehensive tree for all named species (referred to here as the "Big Bang Tree"). This tree will capture the depth of knowledge we have about biodiversity on Earth, including that there has been a single origin of life that gave rise to three major domains (i.e. eukaryotes, archaea, bacteria). The Big Bang Tree will both highlight the considerable advances in reconstructing phylogenies within relatively well-sampled clades such as plants, animals and fungi, and illustrate the tremendous diversity of lineages, many of which are microbial, that are poorly represented in phylogenetic analyses. Finally, the Big Bang Tree will provide a tool to visualize our knowledge of both vertical and lateral phylogenetic events in the history of life, allowing users to conceptualize the high levels of exchange that have been estimated among microbial lineages.

Building the Big Bang Tree will involve four main lines of activity: (1) identifying and aggregating source phylogenies, (2) supplementing phylogenetic data with taxonomic data for completeness; (3) developing software to host the tree on the Internet; and (4) providing a highly intuitive, visually engaging online interface for users to browse, search, and contribute improvements to the tree.

We will first exploit existing databases for those parts of the tree of life that have been the focus of phylogenetic analyses. Table 1 lists the resources from which we will obtain the first set of data for synthesis.

Resource	*Taxonomic Scope*	*Scale*
TreeBASE, treebase.org	Any	6500 published trees
ToLWeb, tolweb.org	Eukaryotes	7,957 clade pages 5,094 species pages
PhyLoTA, phylota.net	Eukaryotes	~125,000 gene trees
GreenGenes, http://greengenes.lbl.gov/	Archaea and Bacteria	16S rRNA tree ~83,000 OTUs

Table 1: Existing phylogenetic resources with digital data that we will incorporate into the Big Bang Tree.

Next, AToL grants and other systematics endeavors will provide community-validated subtrees and data sets for our Big Bang Tree effort. These projects are assembling densely-sampled trees for many clades in the tree of life and also working to resolve key polytomies, such as the base of modern birds. PIs Crandall (arthropods), Katz (microbial eukaryotes), Hibbett (fungi, nucleariids, microsporidia), and Soltis (green plants, red algae) will supervise post-docs who will curate matrices and trees for major subclades (see letters of collaboration with commitments to share data). We plan to incorporate phylogenies that include both extant and extinct species. These outreach efforts will continue to refine and improve the tree in years 2 and 3.

Due to the fact that many species have yet to be included in modern phylogenetic analyses, we will fill in gaps with taxonomic data. Classifications provided by the Global Names Initiative [15] provide a flexible, powerful starting point, and their GNITE interface enables efficient taxonomic editing. For example, during preparation of this proposal, the entire fungal section was updated in a matter of hours using information from Index Fungorum. Collaboration with this initiative (see letter from David Patterson) will avoid duplicated effort toward producing scalable, community-editable taxonomies. This will allow both projects to benefit from lessons learned about community editing. The incorporation of taxonomic data is the only way to provide a draft tree that contains all named species, and one of our early goals is to provide quantitative and visual exploration of the state of our phylogenetic knowledge relative to overall biodiversity.

Synthesis of this Big Bang Tree will use a combination of new and existing analytical approaches (supertree and similar methods [16, 17]) and simple merging and grafting techniques [18, 19]. There will be continuing research conducted on this topic throughout the three years (see section "Novel Methods for Phylogenetic Synthesis"). However, for the first year we will encourage plurality and accommodate multiple methods of integration. The availability of a large collection of input trees, including large, well-accepted and well-supported single phylogenies from AToL and similar projects, provides an excellent set of empirical data against which to test our methods for synthesis. At least for the first year, we will largely focus on existing supertree, grafting, and supermatrix implementations with the understanding that other methods will be employed as they are developed either within the group or by other researchers.

Once we have constructed the initial Big Bang Tree, we will explore methods for assessing its quality, including analysis of clade support values, quantifying phylogenetic vs. taxonomic coverage and highlighting areas of conflict, including detecting instances of reticulation. For the purposes of constructing and presenting the first version of the Big Bang Tree, we do not propose to develop new methods for incorporating reticulation, but we will instead provide a means to visualize these events on the Big Bang Tree, highlighting their greater frequency in microbial lineages compared to plants and animals). The development of new methods in this area will be the subject of additional research in tree building (see below).

We stress that our goal is to build an initial, maximally comprehensive tree hosted on an interactive software platform. Our subsequent emphasis will be to inspire the community of systematists to undertake synthesis on the scale of the entire tree of life and to change the cultural norms for sharing phylogenetic data in digital, reusable formats.

Publishing the Big Bang Tree

To engage the public, educational, and scientific communities, we will launch a broadly focused website called **OpenPhylo**, accompanied by an extensive traditional and social media campaign. Developed under the oversight of PI Gude, a professional information designer and artist, the site will exhibit a wide variety of visual content for media outreach, including an interactive infographic that explains the tree of life, a video that takes the visitor on a narrated tour of the tree, and a compelling logo that we hope will become the new icon for evolution. The home page will be dynamic and visually exciting, with continually updated photo slideshows, curation updates, news and Twitter feeds. Primary sections will target different audiences, e.g., content for novice visitors will explain basic concepts of phylogenetics in simple terms.

A central goal of OpenPhylo is to enable the systematics community to annotate and revise the Big Bang Tree. For researchers, the ability to display and determine conflicting relationships and alternative topologies is very important. OpenPhylo will accommodate the discovery of such conflict and provide a means to visualize and access alternate hypotheses. However, many users of the tree such as ecologists, medical professionals, and environmental scientists will want a single tree on which to conduct analyses. The Big Bang Tree will be the tree that represents the most accepted and frequently recovered relationships.

To provide the necessary infrastructure for phylogenetic synthesis and for hosting the Big Bang tree, we will build an open-source database-driven web application, named **PhyloShare**. This software will serve as a 'synthesis engine' capable of storing, displaying and analyzing trees with millions of leaves while retaining provenance of the underlying data sets. The software will be implemented using Web2py [20], a framework for rapid development of database-driven web applications. A prototype Web2py application implementing some of the desired features (e.g., tree and data storage, large tree navigation, clade grafting, rudimentary interaction with TreeBASE and taxonomic reconciliation services through web APIs) is Phylografter, currently in development by Ree [21]. PhyloShare will use this codebase as a launching point for developing the following interoperable components:

1) **Graphical User Interface for Navigation and Curation:** The graphical user interface (GUI) will allow users to store, query, and display phylogenetic trees, data, and metadata such as analysis methods, GenBank accession numbers, and taxonomic names. It will provide features for creating synthetic trees by simple grafting (i.e., replacing a representative OTU in a backbone tree with a more comprehensively sampled tree from a more focused study), as well as assembling merged data sets for new supermatrix or supertree analyses.

The GUI will also support **visualizing the tree of life**. For interactive, visual navigation of large trees, we will further develop the iPlant Collaborative tree visualization software [22]. Similar to Google Maps, this software uses level-of-detail rendering, limiting data requests to the information required for the current view. This alleviates the computational burdens and memory requirements of displaying millions of nodes on screen at once. In addition to viewing the Big Bang Tree (our "best estimate") we will highlight areas of phylogenetic conflict or reticulate evolution and provide ways for users to discover other phylogenetic hypotheses.

Users will be able to **annotate the tree with relevant data**. These may include synapomorphies, divergence times, genome evolution events, character state transitions or citations to the primary literature. Using the Google Maps analogy again, any user – researchers, students or the general public – but we will allow fine-grained control over the display of layers. For common data types, we will capture semantics to allow for query and re-use. These annotation layers can provide links to other online resources, such as Encyclopedia of Life (EOL), Global Biodiversity Information Facility (GBIF) or Tree of Life Web Project (ToLWeb).

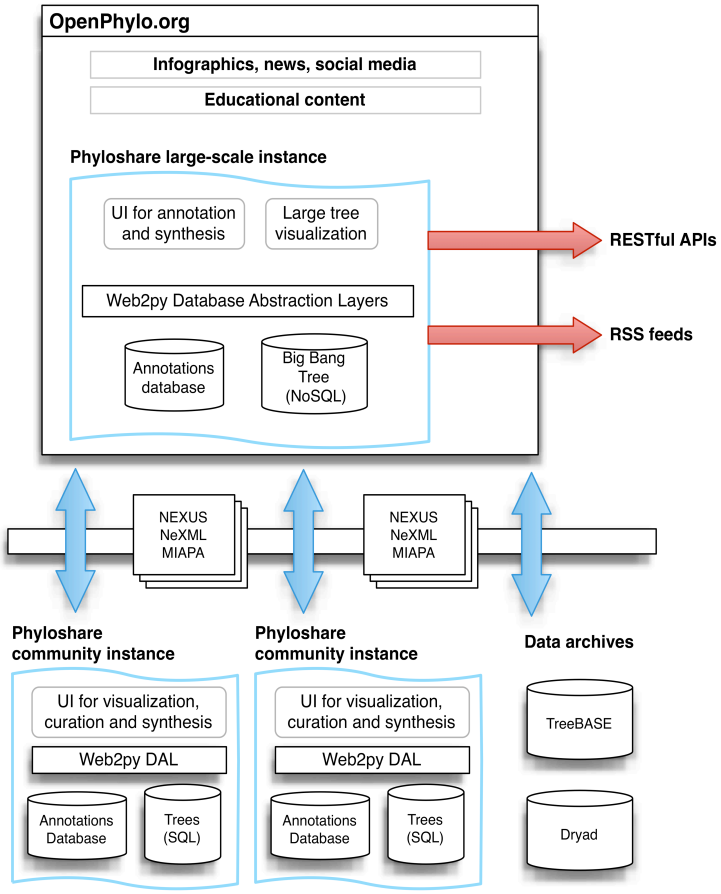

Figure 1: Architecture of OpenPhylo and software components

2) **Phylogenetic databases:** Due to the scale of the complete tree, we will use a NoSQL database engine optimized for large graphs, such as OrientDB or Neo4j, and build associated database abstraction layers for Web2py. These support graphs as a native datatype and are optimized for rapid traversal of linked nodes, overcoming the performance problems associated with recursive queries and table indexing that are required with relational databases. Graph databases are thus better suited for phylogenetic queries of very large trees, as confirmed by preliminary benchmarks by PI Smith.

3) **Annotations database:** To stably store annotations while allowing the underlying topology to be dynamic, we will develop a database schema of branch and clade specific identifiers based on phyloreferences [23, 24]. The database will store annotations by node definition (i.e., 'most recent common ancestor of A and B') rather than solely by attachment to a particular taxonomic name or internal node identifier. These phyloreferences will enable a system of specific keys for annotations of the tree, allow biologists to subscribe to RSS feeds for particular parts of the tree, and allow for a variety of topological queries.

4) **Taxonomic reconciliation library**: We will use existing web services provided by NameBank [25] and iPlant [26] to align and reconcile taxonomic names of OTUs and clades across data sets, cross-referenced to universal identifiers (NameBank LSIDs, NCBI ids). We will collaborate with the Global Names Initiative {http://gnaclr.globalnames.org}. GNI has already developed algorithms to find lexical variants of names and to discover homotypic synonyms, and is now working to draw in heterotypic synonyms from classifications (see letter from Patterson).

5) **Application programming interfaces (APIs)**: We will provide a RESTful interface based on the PhyloWS standard [27] for programmatic access to the Big Bang Tree and other data. For initial release, we will support a limited set of topology-based queries, for example, searching based on a list of input taxa or a clade defined as the most recent common ancestor of two other taxa. We will expand these queries in years 2 and 3 based on both our requirements for access to data, requirements of other AVAToL proposals (see letter from Harmon) and community input.

B. Engaging the community to refine and annotate the tree

Engagement of the broader community is critical for construction of the initial Big Bang Tree, for continued refinement and annotation, and for adoption of these resources and tools. It will take the entire systematics community to create a trusted phylogenetic resource, and participation by a much larger group of scientists, developers, teachers and students to ensure that the platform meets the needs of the broad community of tree-users.

Reward-based curation

The dynamic nature of the tree hosted on OpenPhylo will directly encourage participation. The site will let visitors add annotations, incorporate new studies, and directly curate versions of the topology. It will allow users to discover and view such activity: subscribe to clade notifications, view changes over time, and see the activity of other curators. To incentivize the contributions of researchers, we will provide curation metrics, for example the number of annotations, number of published phylogenies incorporated into the tree, number of downloads for curated clades, number of students that contributed. When users download trees, we will provide citations for the data or publications that contributed to the downloaded data to encourage users to cite authors of source data. By tying OpenPhylo data to contributors, the system will be able to provide quantitative metrics for the purpose of career advancement for the scientific community [28]. Finally, OpenPhylo and the proposed links to data repositories such as TreeBASE and Dryad (see below) will provide a user-friendly mechanism for researchers to comply with the data archiving requirements of journals and funding agencies.

Workshops and conferences

We will directly engage both the systematics (tree-builders) and tree-user communities through workshops and via our presence at major conferences. We have contacted the SSE/SSB officers, who are interested in exploring the possibility of an add-on symposium/workshop at the 2012 meeting. Such an event would include speakers who would provide a status report for the tree of life, and then run workshops to get early input from potential users on creating a community-driven, updatable tree of life. In years 2 and 3 we have planned community workshops to obtain feedback on the current state of the synthesized tree, identify data that needs to be incorporated and perform usability testing on software products. We have already contacted the PIs of recent AToLs (see letters of collaboration), which indicate commitment to share data and participate in our workshops. In addition to workshops, we will also use symposia or presentations at major meetings (e.g., Society of Systematic Biologists, Botanical Society America, Mycologial Society of America, International Crustacean Congress, World Conference on Marine Biodiversity) at which we present an outline of the OpenPhylo project, a software demonstration, presentation of the Big Bang Tree and the ways to get involved. We have one such symposium budgeted in year 2 of the project. In many cases connectivity may be best accomplished by direct invitation and active interaction and we have therefore included funding for co-PIs to travel to international meetings and engage the world-wide systematics community (e.g., Soltis has connected with ongoing green plant efforts in Europe and China).

We also recognize the special need to engage researchers who focus on bacterial and archaeal phylogenetics due to unique features of these clades, including high levels of reticulate evolution and lack of Linnaean taxonomic names. Our synthesis and software development efforts are designed to incorporate such data. We will invite members of this community to our workshops (e.g., see letters from Konstantinidis and Gogarten) and attend relevant meetings to solicit their input.

Community-driven software development

The success of our software infrastructure will depend on building links with other developers of phylogenetic inference software, related analytical tools, and online biodiversity resources. Our software components will be modular by design, to allow for re-use in other projects (e.g., by developing a custom interface to our data using the APIs, or re-using the TreeBASE submission library in another data management tool). All software will be open-source and hosted on GitHub [29], a social coding platform that allows other developers to follow, comment and contribute. Being open-source allows us to participate in community hackathons and conferences such as iEvoBio [30]. The goals of iEvoBio are to catalyze the development of new open source tools and to increase awareness of existing technologies. We will provide partial sponsorship of the iEvoBio meeting in each year of the grant, supporting either an Informatics Challenge or special themed session relevant to phylogenetic synthesis (see letter from H. Lapp).

In year 2, we will host a hackathon at NESCent. This hands-on workshop will bring together developers from various projects to identify short programming targets and implement solutions. While projects are not identified in advance, other hackathons at NESCent have produced extended data import / export capabilities, developed programming interfaces that allow for interoperability between resources, or created novel visualization tools. We will complement this effort with a contest in year 2 to solicit ideas for software extensions from our community of users. In year 3, we will identify the most promising hackathon outcomes and contest entries and use our developer resources to bring them to production-ready components. Components of the OpenPhylo platform would also provide excellent projects for NESCent internships through the Google Summer Code program [31], which provides summer funding for undergraduate and graduate students to work on open-source software projects.

Distributed resources for undergraduate education and research experiences

We will build resources for undergraduate education that will employ the new tools developed in this project, and we will couple curriculum development with research experiences for undergraduates. The ultimate product is a web-based, distributed course that allows students to identify and explore a clade; understand the data used to reconstruct the phylogeny; add new data; perform phylogenetic analyses; annotate the tree with images, comments, and references; create publication-quality graphics; and generate public web pages about the tree of life. The underlying tools will be the same as those used by the research community, but the educational environment will include explanatory material geared to undergraduates and a modular course plan that educators can use in diverse classes focused on phylogenetics or biodiversity.

C. Promoting a culture of data sharing

In addition to the community-editing features of the OpenPhylo site, we will also develop software products that target key bottlenecks in phylogenetic workflows. These products include collaborative, web-based tools for storing, annotating, and merging trees and data sets, and easy-to-use, incentive-driven tools for submitting published material to digital repositories. Our goal is to encourage data sharing through incentives (“carrots”) rather than requirements (“sticks”), although our products will make it easy for producers of phylogenetic data to comply with data sharing policies of various journals [32, 33] and new NSF data management plans. We propose to change behaviors by providing tools and services that provide value to individual researchers, and which facilitate data sharing as a side effect. We will design and implement the following open-source software products, and promote their use in the community.

PhyloShare synthesis engine

This web application, described in section A (Building a Comprehensive Draft Tree of Life) will be used to assemble and host the Big Bang Tree, and will also be available as a downloadable application that individual systematists or labs may deploy to host their own versions of particular clades. Our goal is to provide easy and intuitive methods for managing, merging and sharing trees and data sets, and linking to and from the Big Bang Tree. Enabling these activities will be a compelling incentive for community members to share their knowledge on the web. These community-run instances will be able to "push" their trees to the OpenPhylo database for inclusion in the Big Bang Tree, on which conflicts will be identified and displayed. The phylogenetic scale of these local installations will generally be considerably less than the entire tree of life, and so will not require a NoSQL database back-end; rather, they can use ubiquitous and portable relational databases such as SQLite and MySQL, which Web2py supports by default. As part of our incentivization strategy, we can provide no-charge hosting for a limited number of community-run instances.

Phylogenetic tree illustrator

Figures in systematic publications typically include not only tree topologies, but also branch lengths, support values, synapomorphies and other annotations. An impediment to synthesis is that these static images contain no re-usable record of the phylogenetic knowledge that they depict. To alleviate this problem, we will develop a "Tree Illustrator" tool for creating publication-quality figures while capturing the phylogeny and metadata in a structured XML file format. The user will have full control over the styling (colors, line widths, etc.) as well as the semantic meaning of all phylogenetic content in a figure (e.g., for a set of text labels above branches, specifying the font as well as the fact they represent bootstrap support values). Styling preferences can be saved and re-applied to different trees or updated versions of the same tree, making the time-consuming process of creating figures much more efficient.

The Tree Illustrator represents a novel approach to synthesis that taps directly into the universal incentive to publish in peer-reviewed journals. It will simultaneously increase the productivity of individual researchers and direct the flow of published phylogenetic knowledge into digital repositories, as the semantically enriched phylogenetic tree figures produced will be easily transformed into Dryad or TreeBASE submissions (see below) or submitted to OpenPhylo for synthesis. We will implement it as an extension to Inkscape [34], a mature open-source, cross-platform vector-graphics editor similar in style to Adobe Illustrator. Inkscape's native file format is SVG, an open standard based on XML. Inkscape extensions currently exist to create diagrams from structured data (e.g., flowcharts from decision trees), which we will use as templates for development.

Version control for phylogenetic data

Biologists conducting phylogenetic analyses have to manage numerous files, transform their formats, share them with collaborators, and conduct a variety of analyses, often repeating these tasks after obtaining new data. These workflows could be significantly improved by using the distributed version control systems commonly used by software developers (e.g., git and mercurial), but the complexity and learning curve associated with the default command-line interfaces of these systems inhibit adoption by non-technical users. In the 2nd year we will extend the PhyloShare system to provide a phylogenetically-aware graphical interface to git-based version control functions. This will allow biologists to archive, rewind, merge, and share their analyses while maintaining a complete revision history of each file. Editing trees or data in PhyloShare, adding files, or adding the products of a new analysis will trigger commits to the version control system, and the user will be prompted for contextual information that will

provide provenance for the final results. The graphical interface will wrap file format conversion tools, taxonomic name resolution, and other common steps in phylogenetic workflows, so that provenance information can be automatically generated for many common operations. The history of an analysis can then be summarized, e.g. for inclusion in a published article, and analyses and data can be made publicly available through the PhyloShare web-services API.

Dryad/TreeBASE submission library

The products described above will provide a means for the community at large to manage and synthesize phylogenies and associated semantic data. For long-term archival of published results, we will leverage existing repositories: TreeBASE and Dryad. We will develop a software library that enables the smooth transfer of data from OpenPhylo, other instances of PhyloShare, and Tree Illustrator to these repositories (for example, using OAI-ORE [35], see letters of collaboration from the Phyloinformatics Research Foundation and from T. Vision). This type of 'handshaking' protocol has already been adopted in Dryad, allowing sequence data deposited in Dryad to be passed directly to GenBank. This library will allow users to bypass the current TreeBASE interface in favor of one that is simpler and easier to use. Internally, the software will use NeXML [36], an XML-based standard for phylogenies.

D. Novel methods for phylogenetic synthesis

Despite advances in algorithm development, software engineering, and computational resources, there remain at least two orders of magnitude difference between the number of species included in the largest phylogenetic analyses completed to date [37-39] and the size of the tree of life. A diverse methodological approach is necessary to synthesize trees at this scale. In addition, we will characterize the performance of different methods, assess the quality of tree estimates, and identify the methods that best balance computational effort and accuracy. A unique aspect of this proposal is the availability of our databases of trees and matrices, many with overlapping taxon sets, from multiple sources. Rather than relying largely on simulated data for methods development, we can validate our methods with vast empirical data sets.

The OpenPhylo platform will be a key component of our own research into new methods for phylogenetic synthesis. Auto-updating gene trees, supertrees, and trees from supermatrices will all be published via PhyloShare instances; imputed gene duplication, LGT and deep coalescence events will be available as annotations to the relevant trees. Postdocs will be involved in developing PhyloShare in order to support the Big Bang Tree in the first year of the grant, so they will be very familiar with the software architecture when they transition to research on synthesis methods. Using OpenPhylo web-services to enable collaboration *within* our project will make our results are immediately available to the broader community. It will also assure that the implementations of our new methodologies are fully interoperable and that the API for interacting with OpenPhylo is heavily used and fully supported. Just as the Big Bang Tree will serve to kick start the process of community-directed synthesis of the tree of life, our suite of synthesis tools centered around OpenPhylo will provide an incentive for the bioinformatics community to become involved in our effort to build the whole tree of life.

Analytical merging vs. grafting approaches

A pivotal question in building maximally comprehensive and accurate trees from existing trees and data sets concerns the decision to either (1) merge those that overlap in some or all taxa, and build larger trees by performing new supermatrix or supertree analyses, or (2) exploit the inherent nestedness of phylogenetic research that resolves localized regions on the tree of life, and recursively graft the results of

more focused studies (e.g., of a single genus) onto backbone trees that contain placeholder OTUs (e.g., of the genera within a family) [18]. We will explore analysis parameters and properties of merged data sets that influence phylogenetic accuracy [40, 41] using simulations and empirical data in OpenPhylo. For example, most research on this topic to date has assumed random patterns of missing data in merged matrices; however, in reality such patterns are often phylogenetically structured, and it remains unclear how this affects the accuracy of the inferred tree. The goal of this research topic is to gain a richer theoretical understanding of the conditions that favor analytical merging or grafting, and to develop guidelines for diagnosing those conditions in practice.

Tree Comparisons

Assembling and hosting the Big Bang Tree will require methods for comparing input trees, to identify areas of phylogenetic conflict and to assess if new data are congruent with existing hypotheses. These methods must be able to operate quickly over large, heterogeneous sets of trees (e.g., rooted and unrooted trees, and trees with partially overlapping leaf sets and varying levels of resolution).

We will initially use high-performance algorithms such as HashRF [42, 43] and MrsRF [44] for computing the Robinson-Foulds (RF) distance [45]. However, the scope of the Big Bang Tree represents a unique opportunity to explore new algorithms for comparing trees across diverse leaf sets. We will also explore quartet distances [46] [47, 48] for trees with only partially overlapping leaf sets. This approach may provide a different and more detailed view of tree distance than the RF. We will also develop a new weighted quartet-based distance metric. In addition to quartets, we will explore triplets and agreement subtrees to diagnose reticulate events such as lateral gene transfer (LGT). Some instances of LGT can be detected using different tree distance statistics to compare the best estimates of the gene trees and species tree. For example, an LGT event across distantly related taxa may drastically increase the RF distance among trees, but it may have little effect on the maximum agreement subtree [9].

Once we can efficiently compare input trees with non-identical taxon sets, we can then use that mechanism to change the question from "What is the distance from tree X from tree Y?" to "What are the most similar trees to tree X?" We will work to implement such queries through OpenPhylo, leading to novel and interesting studies of the Big Bang Tree.

Auto-updating Gene Trees

Given an existing data matrix and tree estimate, we will explore efficient methods to update a tree with new sequences. Recasting tree inference as a problem of updating previous estimates can radically alter feasibility of large maximum likelihood (ML) tree estimation [49]. We will extend methods that place a new sequence on existing trees [50, 51] by implementing algorithms for evolutionary placement (based on [52]) followed localized searching to escape local optima. Initial explorations with these methods have been explored by PI Smith in collaboration with iPlant. These initial experiments serve as a necessary proof of concept that will be integrated and extended. We will also adapt other recently developed memory and time saving techniques: for example localizing branch swaps [53, 54] [55]) and allowing for alignment uncertainty [56]. We will provide branch support for the updated trees by modifying rapid bootstrapping and ratcheting techniques [57, 58]. Because heterogeneity in the substitution process can mislead tree inference (see [59]), we will develop and fit rich, locus-specific models (e.g., [60, 61]) for large and densely-sampled trees.

We will provide automatically updated tree estimates for several genes that are of crucial importance for synthetic efforts, such as ribosomal RNA gene [62, 63], the mitochondrial genome [64] (including loci of particular interest to barcode of life initiatives [65]), and the chloroplast genome (including proposed barcoding genes such as *rbcL*). During the 2nd and 3rd year of the grant, we will detect new sequences in GenBank for these loci, filter chimeric sequences, and update the gene trees. We will conduct intensive searching on densely sampled subsets of the data to verify our updating approaches. As a test case, we will work with the Chondrichthyan AToL group (see letter from Naylor) to compare automatically updated estimates to other search strategies for single-copy loci used in Gnathostome phylogenetics. In year 3, we will provide auto-updating gene tree estimates for additional loci; the selection of additional genes will be based on community feedback in year 2. Automatically-updated trees will feed directly into other approaches for phylogenetic synthesis (see below), and will provide a powerful platform for discovering new species from sequence data.

Phylogenetic Estimation via Supermatrices

Mega-phylogenetic approaches [17, 39, 66] integrate public sequence databases with phylogenetic searches on very large supermatrices. Supermatrices can overcome the sampling error associated with single-locus analyses. The gene tree updating effort described above will make it feasible to reanalyze specific aspects of the supermatrices that contribute to the backbone of the Big Bang tree. The alignments used in the gene tree updating effort can be used to update the supermatrix.

The integration of supermatrix construction into OpenPhylo will provide feedback for such algorithms, and permit us to test the effect of persistent problems like matrix completeness, effects of missing data, and orthology and paralogy estimates. Unlike gene tree estimation, supermatrix analyses require filtering loci for orthologous sequences [67, 68]. The availability of continually-updated gene trees in OpenPhylo will allow us to revisit supermatrix paralogy and orthology estimates based on new analyses. Newly-detected paralogs can then be recoded in the supermatrix, triggering reanalysis of the appropriate parts of the tree. Hypothesized gene duplication events will be stored for each gene tree to make those inferences reusable by others. As demonstrated by PI Smith [68], supermatrix construction methods now have the ability to utilize next generation data. With the integration of these methods into OpenPhylo, we will be able to incorporate next generation data into the Big Bang Tree. As is true for the rest of this project, objects created at each step (such as supermatrices) will be available to the rest of the community to download to analyze in their own ways.

Phylogenetic Estimation via Supertrees

Supertree methods take collections of input trees with partially overlapping species and construct a comprehensive tree, or *supertree*, that contains all the species from the input trees [69-71]. Supertree methods can combine trees built from any data sets, and thus can potentially incorporate all the disparate phylogenetic data collected across the tree of life. We will refine existing supertree methods to assemble the Big Bang Tree and later expand to address specific supertree problems that have impeded broad scale reconstructions.

First, we will implement algorithms for combining compatible trees [72-74] that provide polynomial-time solutions to construct supertrees. We also will develop further approaches for combining trees with nested leaf sets. Most supertree algorithms assume that the leaves of the input trees represent a single taxon; however, leaves of many published trees represent entire clades of taxa. The AncestralBuild algorithm [75] is a polynomial time algorithm for combining trees with nested leaves, but it only handles compatible

input trees. We will work to adapt algorithms for MRP and Robinson-Foulds (RF) supertree problems to handle trees with nested leaf sets. Finally, we will integrate numerous recent heuristic advances [76, 77] for supertree analyses into OpenPhylo. Additionally, we will focus on new approaches, like Superfine [16], that construct an unresolved, backbone tree representing nodes without conflict and then resolve the remaining portions of the tree with independent local searches. These heuristic approaches may be applied to any supertree problem, and they are extremely amenable to parallel computing solutions.

Reconciling conflicting gene trees

Evolutionary processes such as incomplete lineage sorting, gene duplication and loss, recombination, hybridization, and lateral gene transfer can cause gene trees to differ from the species phylogeny [78]. We will implement a series of gene-tree reconciliation heuristics to allow gene trees, and ultimately large-scale genomic data, to be incorporated into estimates of the tree of life. Gene tree parsimony (GTP; [79-83] represents a computationally tractable approach for reconciling large sets of gene trees. It infers the species tree that implies the minimum number of conflict-causing events implied by a collection of gene trees. There now exist effective and efficient heuristics to infer the species tree when the reconciliation cost is defined in terms of gene duplications [84], duplications and losses [85], or deep coalescence events [85]. These enable reliable GTP analyses with thousands of taxa and genes [86, 87]. We will expand upon existing heuristics in much the same way as we propose for standard supertree methods. Numerous probabilistic approaches also exist for estimating species trees from gene trees based on coalescence processes (e.g. [88-90]). While many of these are not feasible for the analysis of thousands of taxa, some (e.g. STAR and STEAC [91]) will be practical for species that have potentially thousands of loci sampled. OpenPhylo annotations will allow labeling of gene trees nodes thought to be affected by duplications or deep coalescence.

Assessing Lateral Gene Transfer (LGT)

Synthesizing the tree of all life requires that we grapple with lateral gene transfer (LGT), as substantial movement of genes and sometimes whole genomes (e.g., acquisition of mitochondria and chloroplasts) is now well documented [92-99]. The prevalence of LGT led Doolittle [100] to propose the 'web of life' to capture the combination of vertical and lateral events among lineages.

To explore the impact of LGT on reconstructions of the tree of life, we will collect data by modifying a pipeline built for eukaryotes in the Katz lab to include bacteria and archaea. The pipeline: 1) starts with homologs compiled in OrthoMCL [101]; 2) creates alignments and gene trees through an iterative process that allows removal of alleles/recent duplicates; 3) improves alignments using Guidance [102], which automates assessment of both robust taxa and sites; and 4) constructs gene trees using RAxML ([103], [104]). We will use new approaches available for estimating LGT events in larger trees (e.g. [105-108]). We will also explore several approaches to calculate the minimum number of LGT events from a given gene and species tree (e.g. [109-112]). Finally, we will employ available simulation tools, including HGT_simul [113], and EvolSimulator [114, 115], to assess our approach. We will exploit the semantically-rich annotation scheme of OpenPhylo to publish any detected LGT events as annotations of the gene trees.

Estimating Branch Lengths

Branch lengths are necessary for many uses of phylogenetic trees, but the Big Bang Tree will lack estimates for many of its branches. Molecular data exist for only a small fraction of the tree of life, and some methods for synthesis that we will employ do not directly estimate branch lengths (e.g., supertrees;

and grafting approaches). Where possible, we will provide branch lengths that come directly from input trees. While a database for divergence times exists, (TimeTree [116]), it does not provide large-scale or programmatic access to such data. We therefore expect divergence times to be common community-provided annotations on the Big Bang Tree. When dates and/or branch lengths are only available for a subset of nodes, we can extrapolate estimates across the rest of the topology, using the simple algorithm in Phylocom [19] or the methods planned for implementation in PhyloFlow (concurrent AVAToL proposal, see letter from L. Harmon). Fossil taxa in the Big Bang Tree are another source of minimum age constraints that we can use in this context. Finally, we can identify cases where sufficient sequence data exist in GenBank, or aligned sequences exist in the PhyLoTA database [86], that would allow for independent assessment of branch lengths by users.

Transformative potential

With the four primary efforts described above, we hope to initiate a fundamental shift in how systematics is practiced and perceived. Our overarching goal is to begin a transformation of the field that will cultivate sustained, active, integrated assembly of the tree of life as its branches continue to be discovered and resolved. This will enable systematics to fulfill its most basic purpose in the life sciences: to provide a comprehensive account of the evolutionary placement and origins of all species. The impact will extend well beyond systematics. It will allow non-phylogeneticists, such as ecologists, to focus on their specific hypotheses (e.g., is there phylogenetic signal in species interactions?) rather than building trees. Our OpenPhylo project has the potential to significantly accelerate the identification and evolutionary placement of newly discovered biodiversity, facilitate the organization of data in biodiversity informatics resources, provide benchmark data for development of new methods, and foster better understanding of evolution by the public.

In many ways, our effort represents the first attempt to construct a unified, all-encompassing view of life on Earth since Linnaeus' (1735) *Systema Naturae*. Creating a community-driven resource for large-scale phylogenetic synthesis represents a transformative contribution on the scale of GenBank. However, while GenBank merely archives data, the goals of OpenPhylo are much more challenging, in that we need to accommodate and convey conflict and uncertainty in phylogenetic hypotheses. The enormity of the task requires correspondingly large-scale community participation, so we have set our goals on providing compelling tools and stimulating incentives. This undertaking will break new ground in building a social platform for science - a platform not merely for summarizing current knowledge, but for stimulating and enabling new discovery.

Broader impacts

The tree of life is a highly compelling metaphor for non-scientists and scientists alike. Research and education across all fields of biology will benefit in fundamental ways from a tree that is easily explored, queried, and downloaded for study. Such a resource will provide a new lens through which to identify and assess global biodiversity and interpret broad-scale patterns and processes of evolution. In fields such as ecology, where phylogeny is being increasingly integrated into community studies, this comprehensive tree will be a central resource for determining evolutionary relationships and interpreting impacts of changing climate on Earth's biodiversity. Further, by highlighting 'dark' areas of the tree – those with only limited morphological and molecular data – a synthetic tree of life may profoundly accelerate the pace of species discovery by providing a common framework in which to place new taxa.

Integral to this proposal is a series of workshops and outreach at meetings that will ensure engagement of both systematists and the wider scientific community. The three graduate students, ten postdocs, and numerous undergraduates involved in the project will receive diverse research experiences across systematics, bioinformatics, software development, and phylogenetic analysis. Our undergraduate course development will engage an even larger number of students in critical thinking about evolution, biodiversity, and the tree of life. Through programs at our various institutions, we will recruit students from under-served communities. Our public website and social media outreach will engage the general public and K-12 educators. Finally, our software tools and community engagement activities will initiate a transformation of the culture in systematics to one in which data sharing practices are ingrained and broad-scale synthesis is actively pursued.

Our group is committed to providing opportunities for individuals from groups that are traditionally underrepresented in biology, particularly in bioinformatics and phylogenetics. For example, we will advertise postdocs and graduate positions in venues targeted to underrepresented groups (e.g. SACNAS: http://sacnas.org/), work with the multicultural offices (or equivalents) when recruiting undergraduates for participation in both the educational and research component, and pay close attention to gender balance in informatics opportunities. Our team has proven track record for involvement of underrepresented groups (see results from prior support).

References

1. Darwin, C., *The Origin of Species*. Screen, ed. J. Murray. Vol. 6. 1859: John Murray. 22-79.
2. Costello, M.J., S. Wilson, and B. Houlding, *Predicting total global species richness using rates of species description and estimates of taxonomic effort.* Systematic biology, 2011: p. syr080-.
3. Mora, C., et al., *How many species are there on Earth and in the ocean?* PLoS Biology, 2011. **9**: p. e1001127.
4. Venter, J.C., et al., *Environmental genome shotgun sequencing of the Sargasso Sea.* Science (New York, N.Y.), 2004. **304**: p. 66-74.
5. Piel, W.H., M. Donoghue, and M.J. Sanderson, *TreeBASE: A Database of Phylogenetic Information.*, 2002: Japan. p. 41-47.
6. Vision, T., *The Dryad Digital Repository: Published evolutionary data as a part of the greater data ecosystem.* Nature Precedings, 2010: p. 1-1.
7. Yates, T.L., J. Salazar-Bravo, and J.W. Dragoo, *The importance of the tree of life to society*, in *Assembling the Tree of Life*, J. Cracraft and M.J. Donoghue, Editors. 2004, Oxford University Press. p. 7-17.
8. Davies, T.J. and A.B. Pedersen, *Phylogeny and geography predict pathogen community similarity in wild primates and humans.* Proceedings. Biological sciences / The Royal Society, 2008. **275**: p. 1695-701.
9. Bardel, C., et al., *On the use of haplotype phylogeny to detect disease susceptibility loci.* BMC genetics, 2005. **6**: p. 24.
10. Forest, F., et al., *Preserving the evolutionary potential of floras in biodiversity hotspots.* Nature, 2007. **445**: p. 757-60.
11. Purvis, A., J.L. Gittleman, and T. Brooks, *Phylogeny and conservation*. Conflict, ed. A. Purvis, J.L. Gittleman, and T. Brooks. Vol. 56. 2005: Cambridge University Press. 431.
12. Willis, C.G., et al., *Phylogenetic patterns of species loss in Thoreau's woods are driven by climate change.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**: p. 17029-17033.
13. Thuiller, W., et al., *Consequences of climate change on the tree of life in Europe.* Nature, 2011. **470**: p. 531-4.
14. Procheş, Ş., et al., *Searching for phylogenetic pattern in biological invasions.* Global Ecology and Biogeography, 2007. **17**: p. 070909153804002-???
15. *Global Names Initiative*. Available from: http://gnaclr.globalnames.org.
16. Swenson, M.S., et al., *SuperFine: Fast and Accurate Supertree Estimation.* Systematic Biology, 2011.
17. Smith, S.A., J.M. Beaulieu, and M.J. Donoghue, *Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches.* BMC evolutionary biology, 2009. **9**: p. 37.
18. Beaulieu, J.M., et al., *Synthesizing phylogenetic knowledge for ecological research.* Ecology, in press.
19. Webb, C.O., D.D. Ackerly, and S.W. Kembel, *Phylocom: software for the analysis of phylogenetic community structure and trait evolution.* Bioinformatics (Oxford, England), 2008. **24**: p. 2098-100.
20. *web2py Web Framework*. Available from: http://www.web2py.com.
21. *Phylografter*. Available from: http://phylografter.googlecode.com.
22. *iPlant Collaborative Large Tree Viewer*. Available from: http://portnoy.iplantcollaborative.org/.
23. *EvoIO phyloreferencing subgroup*. Available from: http://www.evoio.org/wiki/Phyloreferencing_subgroup.
24. Keesey, T.M., *Toward a Complete Phyloreferencing Language.* Nature Precedings, 2011.

25. *uBio NameBank web services*. Available from: http://www.ubio.org/index.php?pagename=services_overview.
26. *iPlant taxonomic name resolution service (TNRS)*. Available from: http://www.iplantcollaborative.org/discover/taxonomic-name-resolution-service-tnrs.
27. *PhyloWS web serices standard for phylogenetic trees*. Available from: https://http://www.nescent.org/wg/evoinfo/index.php?title=PhyloWS.
28. McDade, L.A., et al., *Biology Needs a Modern Assessment System for Professional Productivity.* BioScience, 2011. **61**: p. 619-625.
29. *GitHub*. Available from: http://github.com.
30. *iEvoBio: Informatics for Phyogenetics, Evolution, and Biodiversity Conference*. Available from: http://ievobio.org/.
31. *Phyloinformatics Summer of Code*. Available from: http://informatics.nescent.org/wiki/Phyloinformatics_Summer_of_Code_2011.
32. Whitlock, M.C., et al., *Data archiving.* The American naturalist, 2010. **175**: p. 145-6.
33. *Joint Data Archiving Policy*. Available from: http://datadryad.org/jdap.
34. *Inkscape*. Available from: http://www.inkscape.org.
35. *Open Archives Initiative: Object Reuse and Exchange*. Available from: http://www.openarchives.org/ore/.
36. *NeXML - phylogenetic data as xml*. Available from: http://nexml.org/.
37. Goloboff, P.A., et al., *Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups.* Cladistics, 2009. **25**: p. 211-230.
38. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree: computing large minimum evolution trees with profiles instead of a distance matrix.* Molecular biology and evolution, 2009. **26**: p. 1641-50.
39. Smith, S.A., et al., *Understanding angiosperm diversification using small and large phylogenetic trees.* American journal of botany, 2011. **98**: p. 404-14.
40. Sanderson, M.J., M.M. McMahon, and M. Steel, *Terraces in phylogenetic tree space.* Science, 2011. **333**: p. 448-450.
41. Wiens, J.J. and M.C. Morrill, *Missing Data in Phylogenetic Analysis: Reconciling Results from Simulations and Empirical Data.* Systematic biology, 2011.
42. Sul, S.-J. and T.L. Williams. *A Randomized Algorithm for Comparing Sets of Phylogenetic Trees*. in *Proc. Fifth Asia Pacific Bioinformatics Conference (APBC'07)*. 2007.
43. Sul, S.-J. and T.L. Williams. *An Experimental Analysis of {Robinson-Foulds} Distance Matrix Algorithms*. in *European Symposium of Algorithms (ESA'08)*. 2008. Springer-Verlag.
44. Matthews, S. and T. Williams, *MrsRF: an efficient MapReduce algorithm for analyzing large collections of evolutionary trees.* BMC Bioinformatics, 2010. **11**: p. S15.
45. Robinson, D.F. and L.R. Foulds, *Comparison of phylogenetic trees.* Mathematical Biosciences, 1981. **53**(1-2): p. 131-147.
46. Brodal, G.S., R. Fagerberg, and C.N.S. Pedersen, *Computing the Quartet Distance between Evolutionary Trees in Time {O(n\log^2 n)}*. Lecture Notes in Computer Science, 2001. **2223**: p. 731-742.
47. Mailund, T. and C.N.S. Pedersen, *QDist--Quartet Distance Between Evolutionary Trees.* Bioinformatics, 2004. **20**: p. 1636-1637.
48. Stissing, M., et al. *Computing the All-Pairs Quartet Distance on a Set of Evolutionary Trees*. in *APBC*. 2007.
49. Wu, D., et al., *An automated phylogenetic tree-based small subunit rRNA taxonomy and alignment pipeline (STAP).* PloS one, 2008. **3**: p. e2566.
50. Matsen, F.A., R.B. Kodner, and E.V. Armbrust, *pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree.* BMC Bioinformatics, 2010. **11**: p. 538.

51. Berger, S.A., D. Krompass, and A. Stamatakis, *Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood.* Systematic Biology, 2011. **60**: p. 291-302.
52. Pearl, J. and M. Tarsi, *Structuring causal trees.* Journal of Complexity, 1986. **2**: p. 60-77.
53. Zwickl, D.J., *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*, 2006.
54. Stamatakis, A., *RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.* Bioinformatics, 2006. **22**: p. 2688-2690.
55. Guindon, S., et al., *New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0.* Systematic Biology, 2010. **59**: p. 307-321.
56. Westesson, O., et al., *An alignment-free generalization to indels of Felsenstein's phylogenetic pruning algorithm.* Transformation, 2011: p. 1-36.
57. Nixon, K.C., *The Parsimony Ratchet, a New Method for Rapid Parsimony Analysis.* Cladistics, 1999. **15**: p. 407-414.
58. Vos, R.A., *Accelerated Likelihood Surface Exploration: The Likelihood Ratchet.* Systematic Biology, 2003. **52**: p. 368-373.
59. Holder, M.T., D.J. Zwickl, and C. Dessimoz, *Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes.* Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences, 2008. **363**: p. 4013-4021.
60. Halpern, A.L. and W.J. Bruno, *Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies.* Molecular Biology and Evolution, 1998. **15**: p. 910-917.
61. Rodrigue, N., H. Philippe, and N. Lartillot, *Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**: p. 4629-4634.
62. DeSantis, T.Z., et al., *Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB.* Applied and environmental microbiology, 2006. **72**: p. 5069-72.
63. Ludwig, W., et al., *ARB: a software environment for sequence data.* Nucleic acids research, 2004. **32**: p. 1363-71.
64. O'Brien, E.a., et al., *GOBASE: an organelle genome database.* Nucleic acids research, 2009. **37**: p. D946-50.
65. Hebert, P.D.N. and T.R. Gregory, *The Promise of DNA Barcoding for Taxonomy.* Systematic Biology, 2005. **54**: p. 852-859.
66. Edwards, E.J. and S.A. Smith, *Phylogenetic analyses reveal the shady history of C4 grasses.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**: p. 2532-2537.
67. Boussau, B. and V. Daubin, *Genomes as documents of evolutionary history.* Trends in Ecology & Evolution, 2010. **25**: p. 224-232.
68. Smith, S.A., et al., *Resolving the evolutionary relationships of molluscs with phylogenomic tools.* Nature, 2011. **advance online publication**.
69. Sanderson, M.J., A. Purvis, and C. Henze, *Phylogenetic supertrees: Assembling the trees of life.* Trends in ecology & evolution, 1998. **13**: p. 105-9.
70. Bininda-Emonds, O.R.P., J.L. Gittleman, and M.A. Steel, *THE (SUPER)TREE OF LIFE: Procedures, Problems, and Prospects.* Annual Review of Ecology and Systematics, 2002. **33**: p. 265-289.
71. Bininda-Emonds, O.R.P., *Trees versus characters and the supertree/supermatrix "paradox".* Systematic Biology, 2004. **53**: p. 356-359.
72. Aho, A.V., et al., *Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions.* SIAM Journal on Computing, 1981. **10**: p. 405-421.
73. Constantinescu, M. and D. Sankoff, *An efficient algorithm for supertrees.* Journal of Classification, 1995. **12**: p. 101-112.

74. Ng, M., *Reconstruction of rooted trees from subtrees.* Discrete Applied Mathematics, 1996. **69**: p. 19-31.
75. Berry, V. and C. Semple, *Fast computation of supertrees for compatible phylogenies with nested taxa.* Systematic Biology, 2006. **55**: p. 270-288.
76. Goloboff, P.A., J.S. Farris, and K.C. Nixon, *TNT, a free program for phylogenetic analysis.* Cladistics, 2008. **24**: p. 774-786.
77. Bansal, M.S., et al., *Robinson-Foulds supertrees.* Algorithms for molecular biology : AMB, 2010. **5**: p. 18.
78. Maddison, W., *Gene trees in species trees.* Systematic biology, 1997.
79. Goodman, M., et al., *Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences.* Systematic Zoology, 1979. **28**: p. 132-163.
80. Guigó, R., I. Muchnik, and T.F. Smith, *Reconstruction of ancient molecular phylogeny.* Molecular phylogenetics and evolution, 1996. **6**: p. 189-213.
81. Maddison, D., D. Swofford, and W. Maddison, *NEXUS: an extensible file format for systematic information.* Systematic Biology, 1997. **46**: p. 590-621.
82. Page, R.D.M. and M.A. Charleston, *From gene to organismal phylogeny: Reconciled trees and the gene tree species tree problem.* Molecular Phylogenetics and Evolution, 1997. **7**: p. 231-40.
83. Slowinski, J.B. and R.D.M. Page, *How should species phylogenies be inferred from sequence data?* Systematic Biology, 1999. **48**: p. 814-825.
84. Bansal, M.S., et al., *Heuristics for the gene-duplication problem: A (n) speed-up for the local search.* Lecture Notes in Computer Science, 2007. **4453**: p. 238.
85. Bansal, M.S., J.G. Burleigh, and O. Eulenstein, *Efficient genome-scale phylogenetic analysis under the duplication-loss and deep coalescence cost models.* BMC Bioinformatics, 2010. **11**: p. S42.
86. Sanderson, M.J., et al., *The PhyLoTA Browser: processing GenBank for molecular phylogenetics research.* Systematic biology, 2008. **57**: p. 335-46.
87. Chaudhary, R., et al., *iGTP: A software package for large-scale gene tree parsimony analysis.* BMC Bioinformatics, 2010. **11**: p. 574.
88. Heled, J. and A. Drummond, *Bayesian Inference of Species Trees from Multilocus Data.* Molecular Biology and Evolution, 2010. **27**: p. 570-580.
89. Kubatko, L.S., B.C. Carstens, and L.L. Knowles, *STEM: species tree estimation using maximum likelihood for gene trees under coalescence.* Bioinformatics, 2009. **25**: p. 971-973.
90. Liu, L. and D.K. Pearl, *Species trees from gene trees: reconstruction Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions.* Systematic Biology, 2007. **56**: p. 504-514.
91. Liu, L., et al., *Estimating species phylogenies using coalescence times among sequences.* Systematic Biology, 2009. **58**: p. 468-477.
92. Dagan, T., Y. Artzy-Randrup, and W. Martin, *Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**: p. 10039-44.
93. Galtier, N. and V. Daubin, *Dealing with incongruence in phylogenomic analyses.* Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 2008. **363**: p. 4023-9.
94. Bapteste, E., et al., *Prokaryotic evolution and the tree of life are two different things.* Biology direct, 2009. **4**: p. 34.
95. Doolittle, W.F., *The practice of classification and the theory of evolution, and what the demise of Charles Darwin's tree of life hypothesis means for both of them.* Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 2009. **364**: p. 2221-8.
96. Gribaldo, S. and C. Brochier, *Phylogeny of prokaryotes: does it exist and why should we care?* Research in microbiology, 2009. **160**: p. 513-21.

97. Koonin, E.V., *Darwinian evolution in the light of genomics.* Nucleic acids research, 2009. **37**: p. 1011-34.
98. Lake, J.a., et al., *Genome beginnings: rooting the tree of life.* Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 2009. **364**: p. 2177-85.
99. Schleifer, K.H., *Classification of Bacteria and Archaea: past, present and future.* Systematic and applied microbiology, 2009. **32**: p. 533-42.
100. Doolittle, W.F., *Phylogenetic classification and the universal tree.* Science (New York, N.Y.), 1999. **284**: p. 2124-9.
101. Chen, F., et al., *Assessing performance of orthology detection strategies applied to eukaryotic genomes.* PloS one, 2007. **2**: p. e383.
102. Penn, O., et al., *An alignment confidence score capturing robustness to guide tree uncertainty.* Molecular Biology and Evolution, 2010. **27**: p. 1759-1767.
103. Stamatakis, A., T. Ludwig, and H. Meier, *RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees.* Bioinformatics (Oxford, England), 2005. **21**: p. 456-63.
104. Stamatakis, A., P. Hoover, and J. Rougemont, *A rapid bootstrap algorithm for the RAxML Web servers.* Systematic biology, 2008. **57**: p. 758-71.
105. Nakhleh, L., D. Ruths, and L.-S. Wang, *RIATA-HGT: A fast and accurate heuristic for reconstrucing horizontal gene transfer*, in *Computer*2005, Proceedings of the Eleventh International Computing and Combinatorics Conference (COCOON 05). LNCS #3595. p. 84-93.
106. Tofigh, A., M. Hallett, and J. Lagergren, *Simultaneous identification of duplications and lateral gene transfers.* IEEEACM transactions on computational biology and bioinformatics IEEE ACM, 2010. **8**: p. 517-535.
107. Doyon, J.-P., et al., *An Efficient Algorithm for Gene/Species Trees Parsimonious Reconciliation with Losses, Duplications and Transfers*, in *COMPARATIVE GENOMICS Lecture Notes in Computer Science*, E. Tannier, Editor 2011, Springer Berlin Heidelberg. p. 93-108.
108. Lienau, E.K., et al., *Cladistics The mega-matrix tree of life : using genome-scale horizontal gene transfer and sequence evolution data as information about the vertical history of life.* New York, 2011. **27**: p. 417-427.
109. Hallett, M.T. and J. Lagergren, *Efficient algorithms for lateral gene transfer problems.* Proceedings of the fifth annual international conference on Computational biology RECOMB 01, 2001: p. 149-156.
110. Addario-Berry, L., M. Hallett, and J. Lagergren, *Towards identifying lateral gene transfer events.* Pacific Symposium On Biocomputing, 2003. **8**: p. 279-290.
111. Górecki, P., *Reconciliation problems for duplication, loss and horizontal gene transfer*, in *Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*2004, ACM: San Diego, California, USA. p. 316-325.
112. Hallett, M., J. Lagergren, and A. Tofigh, *Simultaneous identification of duplications and lateral transfers.* Proceedings of the eighth annual international conference on Computational molecular biology RECOMB 04, 2004: p. 347-356.
113. Galtier, N., *A model of horizontal gene transfer and the bacterial phylogeny problem.* Systematic Biology, 2007. **56**: p. 633-642.
114. Beiko, R.G. and R.L. Charlebois, *A simulation test bed for hypotheses of genome evolution.* Bioinformatics (Oxford, England), 2007. **23**: p. 825-31.
115. Beiko, R.G., W.F. Doolittle, and R.L. Charlebois, *The impact of reticulate evolution on genome phylogeny.* Systematic biology, 2008. **57**: p. 844-56.
116. Hedges, S.B., J. Dudley, and S. Kumar, *TimeTree: a public knowledge-base of divergence times among organisms.* Bioinformatics (Oxford, England), 2006. **22**: p. 2971-2.
117. James, T.Y., et al., *Reconstructing the early evolution of Fungi using a six-gene phylogeny.* Nature, 2006. **443**: p. 818-22.

118. Hibbett, D.S., et al., *A higher-level phylogenetic classification of the Fungi.* Mycological research, 2007. **111**: p. 509-47.