Workshop Report
Where to Next with The Tree of Life?

April 3 - 5, 2008
Washington DC
USA
Workshop Report - *Where to Next with The Tree of Life?*

Workshop Summary  
Emerging research areas addressed by the workshop:  
- Consider the fundamental question: is there a single tree of life (ToL)?  
- Phylogenomics and its role in inferring the ToL.  
- The ToL as a tool for understanding development and the origin of evolutionary innovations.  
- Using phylogenetic approaches to understand speciation, biodiversity, and biogeography.  

Introduction  
Significance of the AToL Program  
Objectives  

Workshop Breakout Sessions and the Recommendations  
- Consider the fundamental question: is there a single tree of life?  
  Summary and Recommendations  
- Phylogenomics and its role in inferring the ToL.  
  Summary and Recommendations  
- The ToL as a tool for understanding development and the origin of evolutionary innovations.  
  Summary and Recommendations  
- Using phylogenetic approaches to understand speciation, biodiversity, and biogeography.  
  Summary and Recommendations  

Final Discussion  
Vertical and Horizontal Integration of the AToL Program  

The Future of AToL  
Workshop Summary and Recommendations  
Major Workshop Outcome  
Summary  
Immediate Next Steps for the AToL Program  

Acknowledgements  
References
Organizers:
Debashish Bhattacharya University of Iowa
Susan Weller University of Minnesota

Organizing committee:
Andy Anderson Southern Illinois University, Carbondale
Josep Comeron University of Iowa
Doug Eernisse California State University, Fullerton

Meeting Coordinator:
Joe Miller University of Iowa

Workshop Participants:
Bob Andersen Bigelow Lab for Ocean Sciences
John Archibald Dalhousie University
Jeff Boore SymBio Corporotaiion
Seth Bordenstein Marine Biological Laboratory
Peter Crane University of Chicago
Charles Delwiche University of Maryland
Michael Donoghue Yale University
Jeremiah Hackett University of Arizona
John Heidelberg University of Southern California
David S. Hibbett Clark University
Bob Jansen University of Texas, Austin
Scott Lanyon University of Minnesota
John Logsdon University of Iowa
Paula Mabee University of South Dakota
Sarah Matthews Harvard University
Monica Medina University of California, Merced
Howard Ochman University of Arizona
Jan Pawlowski University of Geneva
Hervé Philippe University of Montréal
Gwenaël Piganeau Banyuls sur Mer, France
Yin-Long Qiu University of Michigan
Antonis Rokas Vanderbilt University
Pam Soltis University of Florida
John Wiens SUNY, Stony Brook
Linda Amaral Zettler Marine Biological Laboratory

Student rapporteur:
William Lanier University of Iowa
Ahmed Moustafa University of Iowa
Heather Tyra University of Iowa
Workshop Report
Where to Next with the Tree of Life?

Workshop Summary
The National Science Foundation (NSF) “Assembling the Tree of Life” (AToL) Program has been a great success, reawakening public interest in systematics and evolution and spurring significant research work in these fields. The last NSF-funded workshop to discuss the aims of this program was held in 2004. Now after completion of the 2007 funding cycle it was an appropriate time to revisit the issue of Program aims in light of significant advances that have occurred in related fields of enquiry such as genomics, developmental biology, informatics, and biodiversity research. These advances have led the scientific community to ask key questions such as: Do fundamental issues such as Horizontal Gene Transfer (HGT) stand in the way of generating a single, well-resolved tree of eukaryotes? How can emerging results from the ToL program inform and facilitate synergistic activities with other biological disciplines? An assessment of the “state of the field” was needed to incorporate emerging genomic tools, new areas of enquiry, and HGT concerns into the AToL Program while maintaining its trademark of a strong commitment to phylogenetic systematics.

To this end, we received funding from the NSF (EF-0742952) to organize a workshop to explore key areas of contemporary research on the ToL. An international group of invited scientists presented their work, explaining the potential relationship and contribution of their work to the AToL Program. Each presentation was followed by facilitated discussions in small breakout groups.

The specific issues tackled by the workshop were:

Consider the fundamental question: is there a single tree of life (ToL)?
This discussion focused on comparing and contrasting horizontal gene transfer (HGT) in prokaryotes versus eukaryotes. We asked the question: Can we infer the eukaryotic tree given forces such as HGT, endosymbiotic gene transfer (EGT), and pervasive gene duplication and loss?

Phylogenomics and its role in inferring the ToL.
This discussion focused on choosing appropriate methods for (a) generating data for phylogenomic analysis, (b) bioinformatic approaches for constructing genome-wide alignments, and (c) assessing the power of phylogenomic datasets (e.g., how much data are enough?).

The ToL as a tool for understanding development and the origin of evolutionary innovations.
This discussion focused on the prospects of using a conserved toolkit of developmentally important loci (e.g., homeobox, MADS) to understand character evolution over “deep” time in the ToL. We considered the questions of how evolutionary innovations/radiations have shaped the topology of the ToL and the importance of population genetic approaches in the ToL.

Using phylogenetic approaches to understand speciation, biodiversity, and biogeography.
This discussion focused on highlighting the importance of the TOL project as a means for
studying (a) species diversity, (b) radiation patterns, and (c) the importance of including taxonomy in the process of building the tree and describing new species. This workshop and report are seen as critical steps in focusing attention on the NSF AToL Program, both in the US and worldwide. The workshop venue explored many facets of biological and bioinformatic research that can be beneficial to this program. We envision that a judicious expansion of the breadth of the Program mission will result in a larger and more diverse community of ATOL researchers, which in turn will reaffirm the central position of phylogenetic and evolutionary approaches within the life sciences. Diversity of disciplines and an expanded research community will be needed to harness the potential of rapidly evolving technologies, e.g., genomic methods, “virtual” inventories of life forms (e.g., DNA barcoding) and applications that explain and document the biological diversity on Earth (e.g., Encyclopedia of Life).

Introduction
To truly understand biological diversity on Earth, all organisms – especially microorganisms - must be correctly placed on the tree of life. Eukaryotes were exclusively microbial for about one billion years before the evolution of the more familiar and often well-studied macroscopic eukaryotic groups: plants, animals, and fungi (Parfrey et al. 2007). Microbial eukaryotes, or protists, are characterized by a tremendous cellular diversity; e.g., photosynthetic, heterotrophic, and parasitic/symbiotic flagellates. They play an essential role in ecosystems (e.g., carbon fixation in marine systems), and some are causative agents of important infectious diseases (e.g., malaria) that impact the social and economic fortunes of many countries. In the preceding rounds of the ATOL Program much emphasis was (correctly) placed on realizing Darwin’s dream of a “very fairly true genealogical trees of each great kingdom of Nature” (Darwin 1903). The funded research groups have spurred a renaissance in the field of eukaryotic systematics and provided comprehensive and often remarkable phylogenetic frameworks for different animal groups, fungi, plants, and many protist lineages. Additionally these projects have spurred development of more sophisticated analytical tools (e.g., Hibbett et al. 2007). The last NSF-funded workshop on formulating aims for the ATOL Program was in 2004. Now is an ideal juncture to assess our progress and ask the same questions posed in 2004: “Where to next with ATOL? Which grand, unifying goals will comprise the future of this world-leading NSF program?”

Significance of the ATOL Program
It was once famously said that, “nothing in biology makes sense except in light of evolution” (Dobzhansky 1973). It can equally well be said that it is impossible to fully understand the biology of any living or fossil organism without also understanding the specifics of its evolutionary background dictated by its phylogenetic history. The evolutionary process is contingent on the interaction of ecology, development, and phylogeny. Reconstructing the tree of life that summarizes the evolutionary history and interrelationships of all organisms (extant and extinct) on our planet is a very great challenge (e.g., see Dagan and Martin 2006). This challenge is two-fold. Scientists need to address the huge biodiversity on our planet (e.g., ca. 1.7 million plant and animal species have been identified and named) and address the peculiarities of gene and genome evolution.
Reconstructing “deep” phylogenies, the earliest splits among lineages, is a very difficult venture because genes and genomes evolve at different rates within and among lineages and some genes ‘jump’ (i.e., are horizontally transferred) into unrelated lineages (e.g., Ricard et al. 2006, Carlton et al. 2007, Hackett et al. 2007). To obtain a robust estimate of the tree of life, we need to understand how genes evolve both vertically (within taxa) and horizontally (among taxa), otherwise molecular trees may potentially convey gene evolution rather than organismal history. The AToL Program offers an ideal platform to use phylogenetic systematics to understand gene and genome history on a large-scale and out of this understanding the tree of life will emerge.

To summarize, the tree of life enterprise is particularly significant because the result will be a framework for the evolutionary history of all life that will enable comparative studies addressing a diversity of questions that is only limited by our imagination and creativity. Paradoxically, perhaps the most significant challenge at present is communicating the power, utility and importance of a universal tree of life to the broadest possible audience in biology, geology, geography, medicine, chemistry, and other disciplines that can use the hierarchical evolutionary history inherent in phylogenies to frame important questions and hypotheses. Most systematists are good at using their phylogenies to frame questions, but many scientists in other disciplines are not aware of the power that such a comparative framework can provide. Therefore a key goal of the workshop was to highlight the power of “tree thinking” on addressing an array of important biological questions.

Why now? Phylogenetic systematics is now a mature discipline. We have a theoretically and methodologically robust basis for understanding the diversity of life. Phylogenetic systematics is also fundamentally integrative in its approach and therefore provides the basis for drawing together and analyzing comparative life science data of all kinds. A highly robust phylogenetic hypothesis that connects all living organisms and places them in their appropriate genealogical context is within our grasp.

The maturation of the discipline has been accompanied by significant technological advances that have recently accrued in fields such as genomics (e.g., pyrosequencing, Solexa sequencing technology), functional genomics (e.g., Affymetrix chips, SNP arrays, sequence capture methods using micro-arrays), and bioinformatics (e.g., The Encyclopedia of Life, DNA barcodes, ontologies). The rich tapestry of approaches both informs us of the basic biology of organisms and leads us towards Darwin’s dream. Given these remarkable technological and intellectual advances, the community needs to assess how best to incorporate these advances into the AToL Program while maintaining the existing strong commitment to phylogenetic systematics that has been its trademark.

Objectives
The overarching objective of the workshop was to discuss how best to find the right balance between further efforts to construct the tree of life and complementary initiatives to connect AToL program results to other areas of science. Achieving the right balance between building the Tree of Life and using the Tree of Life will ensure that the intellectual benefits of investments in the AToL program are fully realized.
Consider the fundamental question: is there a single tree of life? This discussion focused on comparing and contrasting horizontal gene transfer (HGT) in prokaryotes versus eukaryotes. The discussion centered on how we can infer the eukaryotic tree given forces such as HGT, endosymbiotic gene transfer (EGT), and pervasive gene duplication and loss.

Plenary talks by specialists in bacterial (Howard Ochman) and eukaryotic (Hervé Philippe) genomics demonstrated that trees of life are fundamental and achievable goals for both domains of life. The attendees recognized however that horizontal gene transfer (HGT) is quite common in prokaryotes and understanding this process was only achievable with large well-sampled data sets. In eukaryotes, HGT may also occur but probably varies in prevalence across different kingdoms. A significant impact of HGT probably occurs in the early history of lineages and in taxa that have undergone photosynthetic organelle (i.e., plastid) endosymbiosis. Most participants believed that a separate confounding issue in eukaryotic phylogenetics is the existence of duplicate genes and paralogy. To address these issues in eukaryotes likely requires genomic sequencing but there was some disagreement regarding this approach. The general feeling was that the tree of life could move forward by sequencing members of “core” genes; i.e., those conserved across different lineages and useful as vertical markers of phylogeny. Yet, it was unclear how to identify the core genes in the absence of genomic data. Additional caution was expressed in the ability to find a single set of core genes capable of defining all the major kingdoms of eukaryotes. Given this point, the groups agreed that a real need exists for generating genome-level data from the major supergroups of eukaryotes to facilitate the ToL initiative both for prokaryotes and eukaryotes. However, participants also agreed that complete genome sequencing on a large scale - while appropriate for smaller prokaryotic genomes - would be a waste of precious resources for defining the tree of eukaryotic life. Therefore, participants recommended that an initial set of targeted complete genome projects for the larger, and more highly complex eukaryotic genomes should be complemented with the use of high-throughput transcriptomic methods (e.g., de novo pyrosequencing and assembly of sheared cDNAs) to generate data for other eukaryotes needed to infer the ToL. Clearly, complete genome data for many eukaryote species is highly desirable, however, the range of issues that are brought up by these data (e.g., gene prediction, validation, and annotation) are beyond current requirements to establish a tree of life.

During this first breakout session, participants reviewed and discussed the goals of the current AToL Program. They debated the appropriate taxonomic focus for future phylogenetic coverage. Clearly under-studied groups exist (e.g., red algae, excavates, Rhizaria, many prokaryotic groups) and the community needs to be encouraged to submit proposals for these to fill in gaps of knowledge in the tree of life. But, even well studied groups (many vertebrates, some invertebrates) deserve continued support because resulting phylogenies are critical to comparative studies in ecology, molecular biology (evolution of genomes), physiology, and development. Synergistic, comparative questions can be asked and answered with model organisms that figure prominently in other disciplines (e.g., butterflies) that cannot be asked with poorly known taxa (many prokaryotes). Participants
reaffirmed the need to focus attention on neglected groups and supported requiring synergistic studies for well-studied organisms (i.e., “just a phylogeny” is insufficient).

During the discussion of phylogenetic breadth, a dichotomy developed between defining realistic achievable short-term goals (e.g., phylogenies of critical lineages to resolving the tree of life) and advancing the long-term goals of the AToL Program. Participants articulated a strong desire to establish the relevance of the AToL Program to all life sciences but also felt strongly that program relevance and integration with other research areas should not be dictated from the top but rather should reflect an “organic” scientist-driven initiative. The AToL program should continue to encourage synergy with other disciplines with its “call for proposals.” This discussion thread was picked up and developed in subsequent break-out sessions.

Concern was expressed over how different parts of the tree of life would be grafted onto one another. For many projects, gene sets are unique and non-overlapping with other projects. Some participants suggested that all future proposals be required to include a common gene sequencing initiative. For example, an enforced gene set could be used as a bar coding system, i.e., 16S rDNA (which is present in both prokaryotic and eukaryotic organisms). A common gene set requirement would ultimately unite all the projects irrespective of their individual goals and directly contribute to assembling the tree of life. However, several participants objected strongly to the common gene requirement because such a requirement would not necessarily contribute to the success of the individual project and would cost precious time and money. Rather, these participants favored a core group whose task was to focus on the entire tree of life and sequence the appropriate taxa needed to address this question. Individual projects would then be encouraged as a “broader impact” to facilitate this goal by providing critical taxa or additional sequences for this initiative. Participants saw the merit of each approach; the majority preferred the latter solution as more in keeping with the NSF model of science.

Summary and Recommendations

1) Analysis of the available EST and complete genome data suggest that a single tree of life exists for eukaryotes but this issue is controversial for prokaryotes due to the prevalence of HGT. Much more needs to be done however to fully understand the impact of HGT on eukaryotic genomes, although the role of endosymbiotic gene transfer (EGT) is better established. A genomics approach is needed to address the HGT issue but the AToL Program should explore transcriptomic as well as complete genome sequencing approaches for eukaryotes. In contrast, whole genome sequencing is already feasible for prokaryotes and is highly recommended as the most powerful tool for uncovering the evolutionary history of these taxa. Genome data will also be very useful for understanding the dynamics of gene family duplication across the entire tree of life.

2) We cannot successfully assemble the ToL with the current investigator-driven multi-gene approach. Collection of current sequence data are not coordinated across different ToL lineages, leading to non-overlapping sets of genes among projects. More centralized coordination is needed to share data and information about useful markers. We need to include more explicit statements about how collaborations will be fostered between different grantees. One solution is to set aside funding specifically for an organization or a PI(s) to develop a common molecular marker website to coordinate group efforts. These markers (e.g., rDNA, actin, heat-shock proteins) could form the core set that are
required for each AToL project so that they can later be assembled into a “supermatrix” that covers all studied branches of the ToL. In addition, individual proposals should be highly encouraged to contribute to a common data set as a broader impact.

3) Missing lineages of the tree of life should be aggressively targeted. The AToL program should continue to encourage proposals to address under-studied groups (e.g., red algae, excavates, Rhizaria, many prokaryotic groups) through the call for proposals mechanism. This initiative can be significantly accelerated by the advent of single cell genomic methods, novel sequencing technologies, and metagenomics. These new technologies allow researchers to work effectively with species that cannot be lab-cultured.

Phylogenomics and its role in inferring the ToL.

This discussion focused on choosing appropriate methods for generating data for phylogenomic analysis, bioinformatic approaches for constructing genome-wide alignments, and assessing the power of phylogenomic datasets (e.g., how much data are enough).

The plenary presentation of Dr. Philippe addressed the importance of genomic data to the ToL project. Phylogenomics is a particularly powerful approach for the automated analysis of genome data from multiple taxa. This new discipline is the convergence of genomics science (the study of the function and structure of genes and genomes) and molecular phylogenetics (the study of the hierarchical evolutionary relationships among organisms, their genes, and genomes). The use of phylogenetics to drive comparative genome analyses has facilitated the reconstruction of the evolutionary history of genes, gene families, and organisms. Phylogenomics promises to provide a large set of “core” genes (~120 genes) conserved across eukaryotes that can be used to infer the eukaryotic ToL. The ability to link eukaryotic and prokaryotic branches together will depend upon the level of HGT.

Genomic data are accumulating in public database at an unprecedented rate due to technological advances. Pyrosequencing and potentially short-read sequencing (e.g., Illumina or ABI platforms) are the leading approaches for generating high quantities of transcriptome data. Public databases are currently dominated by the sequences of prokaryotic, metazoan, plant, parasitic, and picoeukaryotic taxa. However, both expressed sequence tag (EST) and complete genomes of free-living eukaryotes are now rapidly appearing. This wealth of information offers the opportunity for using the tools of comparative genomics and multi-gene phylogenetics to clarify many long-standing issues in prokaryotic and eukaryotic evolution such as the interrelationships of the major lineages of microbes and of multicellular eukaryotes and the contribution of the plastid endosymbiont to nuclear genome evolution. Hervé Philippe provided clear evidence in his lecture that a mixture of expressed sequence tag (EST) and complete genome data cross the different lineages of eukaryotes was sufficient to infer a robust (albeit, currently taxonomically poorly sampled) ToL.

Within the break-out groups, the debate began anew concerning breadth of taxonomic sampling and quantity of DNA sequence needed to reconstruct the ToL. Specifically, the question was raised as to the level of data generation that should be undertaken (i.e., cDNA or complete genome) to facilitate phylogenomic analyses. Participants agreed that the answer to this question was dependent upon the scientific aim being pursued. If the focus is the ToL for all organisms, then a strong consensus emerged that taxon sampling is more
important than whole genome sequencing. If the focus is “How do genomes evolve?” then more genomes from fewer carefully selected representative taxa are needed.

Participants also focused on the logistical challenges of sequence data generation and analysis. The question remains whether we can handle 1,000 – 10,000 eukaryotic genomes using current computational tools. Are we in the stages of outpacing our capacity to analyze the data? Do we need to aid in the development of computational methods as part of AToL? The answer to these questions is clear – better algorithms and faster computers are direly needed to support broad-scale phylogenomic analyses using hundreds of genes from hundreds of taxa. Another concern was voiced that trees generated by phylogenomics projects are not being placed in public repositories. Publicly accessible repositories for raw sequences, alignments, and phylogenomic trees are essential to compare results of different studies. The scientific community (e.g., journal policies, grant agencies) has not consistently required deposition of these tree metadata and technology to store and retrieve trees easily needs to be improved. Only recently has it become possible to deposit trees in a centralized repository (e.g., TreeBase, a new repository in NCBI) and some researchers find the procedures in place to be burdensome.

To address technological hurdles, the AToL Program should actively encourage cross-disciplinary interactions with computer scientists, statisticians, and theoreticians. This computational challenge represents a tremendous opportunity for the AToL Program to increase its relevance to other scientific disciplines and to significantly broaden its participation. Many of these links can be made across the different NSF programs to place the AToL Program in the center of future funding initiatives.

Workshop participants then discussed the source of AToL funding and generally agreed that these monies should not come from existing resources for the Systematics Program. Large-scale AToL projects form a distinct initiative within the NSF, despite taxonomic overlap that occurs between AToL and systematics projects. Consistent differentiation between these two programs is desirable, and ATOL should aim to address issues that are either difficult to address or beyond conventional systematic approaches such as the role of HGT in prokaryotes or in the evolution of different eukaryotic lineages. These issues generally require a combination of phylogenetic and genomic methods to be successfully investigated.

Lastly, workshop participants recognized the necessary and vital role of national and international partnerships in elucidating the ToL. Coordination between AToL and other US organizations with strong interests in phylogenomics is strongly recommended (e.g., NESCent, CIPRES). International collaboration was suggested to be a key ingredient in managing the next level of growth in the AToL Program. AToL participants should be actively encouraged to partner with rapidly developing international expertise and similar initiatives in the European Union, Canada, and elsewhere.

**Summary and Recommendations**

1) Phylogenomics promises to provide a large set of “core” genes conserved cross eukaryotes that can be used to infer the eukaryotic ToL. This number may be in the range of 120 genes that have already been identified from EST sequencing projects. Identifying core genes from the available eukaryotic genomes will be a highly useful exercise to answer the essential question of the availability or existence of eukaryotic core genes.
A set of core genes combined with dense taxon sampling will be needed to address the eukaryotic component of the ToL.

2) Phylogenomics will continue to have a large impact on our understanding of prokaryote evolution. However, the concept of relying on core genes is far more controversial due to high rates of intra- and interphylum HGT. Here, a focus on understanding the evolution of genomes may be a necessary precursor to assembling the prokaryotic components of the ToL.

3) Technological advances are needed to handle the voluminous sequence data being generated. Better algorithms and faster computers are direly needed to support broad-scale phylogenomic analyses using hundreds of genes from hundreds of taxa. Another concern was voiced that trees generated by phylogenomics projects are not being placed in public repositories. The community needs to require the deposition of metadata and computational advances are needed to manipulate and store these data more efficiently.

4) International collaboration was suggested to be a key ingredient in managing this level of growth in the AToL Program so that expertise that is rapidly developing beyond the borders of the US are brought to bear on phylogenomic methods and tree interpretation. Coordination between AToL and other US organizations with strong interests in phylogenomics is also strongly recommended (e.g., NESCent, CIPRES).

The ToL as a tool for understanding development and the origin of evolutionary innovations.

This discussion focused on the prospects of using a conserved toolkit of developmentally important loci (e.g., homeobox, MADS) to understand character evolution over “deep” time in the ToL, the question of how evolutionary innovations/radiations have shaped the topology of the ToL, and the importance of population genetic approaches in the ToL.

In her plenary lecture, Paula Mabee addressed how bioinformatics was a new engine driving the integration of phylogenetics, comparative morphology, and developmental data. Body plan innovations underlie major taxonomic radiations and fuel the subsequent rapid evolution of animal and plants species. A growing understanding of the genetic toolkit and the role of regulatory developmental genes in generating phenotypes can be coupled with the historical discipline of comparative morphology. This coupling will lead to a more mature understanding of how different body plans (i.e., different phenotypes) can develop in evolutionary time through small shifts in timing and interaction of regulatory genes. Comparative morphology illustrates the change in phenotypes over evolutionary time. Developmental biology studies the change in phenotypes over an organism’s lifetime and the role of mutation in changing individual phenotypes. By linking the two, we gain a new perspective on generation of phenotypic variation and processes affecting macromutation (generation of novel body plans) and microevolution (species-level differences). In other words, if we can understand the regulation of developmental genes
involved in constructing a zebrafish and phenotypes resulting from its various mutations, we can search the comparative literature for “mutational” phenotypes seen in other fish species and detect the underlying developmental pathways. This approach permits us to identify candidate pathways through which novel phenotypes in non-model organisms evolved. The discipline of “evo-devo” is at once new – re-energized with the advent of molecular biology and informatics – and at the same time quite old – studies of body plan evolution and generation of novelties date to Darwin and his contemporaries.

The general approach is to take advantage of the power of the semantic WEB to connect databases of gene ontologies (a relational database of defined terms) to a comparative evolutionary morphology ontology database. The morphological ontology database has been constructed to adhere to these ontological language “rules.” Paula Mabee discussed the first such initiative, the PhenosCape database (https://phenoscape.org) to illustrate potential insights into macroevolutionary processes that can result from this approach. Phenoscape is a growing compilation of evolutionarily variable morphological characters for fishes that is linked to the developmental ontology database ZFIN (ZFIN.org). Paula discussed the challenges of linking a dynamic genetics and development database like ZFIN to one based on comparative morphology and evolution. A large hurdle is that historically, morphological terminology is non-dynamic, often idiosyncratic and may (or may not) map to homologous features. To be successful, such an initiative relies on interdisciplinary partnerships of developmental biologists, comparative morphologists, and phylogeneticists.

In the break-out session, discussion focused on whether this interdisciplinary partnership and ontology framework was feasible for other eukaryotic taxa and whether it was even applicable to prokaryotes. PhenosCape represents our first model for the integration of phylogeny, evolution, and development in the era of bioinformatics. This model has a great potential to be applied to other groups. General agreement emerged, however, that this highly developed system of developmental and bioinformatics tools was best suited for clades in which existing genetic models exist (e.g., zebrafish, Arabidopsis, Drosophila, nematode). Extensive genetic, developmental, and comparative morphological data are already available to be linked within a bioinformatics framework and lead to an enhanced understanding of the basis for evolution of body forms through developmental regulatory innovations. Evo-devo applications under the AToL framework are not a broadly applicable field of study across many taxa, however, use of existing evo-devo insights from model species to other groups of interest will benefit the AToL initiative. Conceptually “phenotype” needs to be expanded to include physiological and cellular processes that will broaden inquiry to include prokaryotic and non-model eukaryotic organisms. For example, the evolution of cell cycles or metabolic processes across the ToL could be informed by the study of the evolution of shared homeotic loci and regulatory gene cascades.

The types of comparative morphological data needed to fulfill the promise of this integrated ontology approach to development and phenotype were also discussed. Natural experimentation with animal body plans is recorded only in the fossil record; discussants recommended that fossil studies be better integrated with future ToL studies. The entire diversity of the ToL, extinct and extant species, needs to be included as much as possible. On the other end of the scale, if we are to understand microevolutionary processes (tips of the ToL), population genetic studies of modern species are needed to bridge our understanding of phenotypic selection pressures to macroevolutionary patterns.
Concerns were voiced over whether evo-devo and the integration model of Phenoscape should be considered a core function of AToL program. Phenoscape was viewed as an excellent working model and as a desirable future outcome of the ToL program. Phenoscape itself is an outcome of a NESCent working group between the Cypriniform Tree of Life morphologists (including Mabee) and the Zebrafish Information Network biologists and informaticians. However, integration with developmental biology was not considered a required early goal for this phase of the Program. In summary, the AToL initiative can catalyze the integration of different approaches (phylogeny, ontogeny, and comparative morphology), and this integration is an important broader impact. Furthermore, the AToL initiative can help integrate population level studies with development and ecology, an issue addressed in greater detail by John Wiens, the subsequent plenary speaker. The group recommended that whenever possible, current investigators collect morphological, fossil, and/or metabolic data in parallel to tree generation to allow extrapolation of these data as a broader impact.

Summary and Recommendations
1) Although the AToL Program needs to focus on building a robust and densely sampled Tree of Life, it should not be limited to just this goal. Important synergies exist with emerging (ontology studies) and reinvigorated disciplines (development). This program should be flexible and include projects with broader aims (e.g., the study of development, gene duplications, cellular functions, co-evolution), particularly when groups include well-studied model organisms.
2) Comparative morphology whether derived from fossil or living taxa should be included in AToL projects, either in the data matrices (higher eukaryotes) or as value added to the final tree (typically prokaryotes, protists). These data will lay the foundation for future initiatives like PhenoScape and permit a much richer understanding of the mode and tempo of evolutionary processes responsible for diversification of species (microevolution; gradualistic type processes) and body plans (macroevolution; punctuated type processes).
3) Population genetic techniques can provide useful information regarding incipient speciation and lineage sorting. An obvious application is to rapid radiations that have made it difficult to resolve phylogenetic trees. Where rapid radiations are known to occur, then sampling will need to address these events and could include genes involved in development to gain knowledge about both the gene tree and genetic innovations; e.g., perspectives on polymorphisms within important gene families, estimates of population size and its relation to lineage sorting, bottlenecks and their relation to the tree, and the potential for hybridization and introgression. This approach represents a horizontal integration across the ToL that is further developed in the next section.

Using phylogenetic approaches to understand speciation, biodiversity, and biogeography.
This discussion focused on highlighting the importance of the TOL project as a means for studying species diversity and radiation patterns and the importance of including taxonomy in the process of building the tree and describing new species.
Our final plenary speaker, John Wiens, focused on how phylogenetic approaches when combined with an understanding of species ecology and historical biogeography could provide important insights on mechanisms of speciation and biological diversity. His presentation also underscored the importance of the AToL program maintaining strong ties to the traditional systematic activities of documenting and describing biodiversity through collection and vouchering and taxonomy. To date, the AToL program has not encouraged funding projects aimed at the “tips” of the tree (species, species relationships) to gain insights on patterns of speciation. Wiens’ presentation showcased how diversification of lineages could be hypothesized through a new modern synthesis of phylogeny, ecology, and geology. This new frontier in systematics, historical ecology, combines knowledge of current species’ ecological niches (as estimated from distributions documented by museum collections), historical environments (global environment modeling and the geological record), and species phylogenies. This interdisciplinary synergy leads to insights as to the historical drivers of speciation.

One emerging pattern from crown-group species (higher animals and plants) is isolation by microniche (abiotic and biotic factors). Separation of species across the landscape can be due to evolved species tolerances of temperature extremes and modes, rainfall, and other abiotic factors. Narrow physiological tolerances have evolved in stable predictable environments (e.g., mountain-sides in the tropics) and may explain the different speciation patterns observed in some temperate and tropical clades of organisms. These studies have direct relevance to conservation biology and global climate change. Some microenvironments are in danger of disappearing altogether as the climate warms; others will shift location to higher altitudes or latitudes. Organisms will need to have dispersal capabilities or protected locations (parks, reserves) to shift into, to follow their preferred microniche. These studies have direct relevance to documenting global climate change and managing its effects on the biota. Wiens emphasized the critical importance of collections, their maintenance and growth, and the importance of sound taxonomy to communicate results to non-systematists.

In the final break-out groups, discussants divided their time between issues raised by the final speaker and synthesizing ideas from the entire workshop. Wiens’ seminar sparked two major areas of discussion. The first was the importance of emphasizing the TOL project as a means for studying radiation patterns and mechanisms that drive speciation. The second issue was the importance of supporting more traditional systematic activities of documentation (vouchering in research collections) and taxonomy (species discovery) while building the ToL. ToL projects can be used to discover new species and to study radiation patterns but this emphasis requires broadly surveying the shallow nodes of the tree, and emphasis that has often been missing in previous discussions of the mission of the Program and would be a very welcome addition to Program goals. Understanding the biogeography and microniches based on phylogeny will aid conservation efforts worldwide. Although conservation issues are more easily addressed in less species rich group (e.g., gymnosperms and birds), metagenomic analysis (such as sea sediment or soil 18S rDNA surveys) can lead to major discoveries of many previously unknown microbial species; TOL projects assist with this discovery and documentation of biodiversity.
PIs should consider emphasizing the need for increased taxon sampling both in numbers of species and numbers of individuals of a species across the geographic range of the species. In understudied, diverse groups, much species discovery and description has yet to be accomplished and mixed species samples is a real concern in large ToL projects. To be able to trace these possible problems, species sampling needs to be appropriately vouchered in permanent and accessible collections; private collections of organisms should be actively discouraged. All ToL voucher material should be deposited in the relevant public collection for future scientists. The discussions underscored the need for understanding the role of biogeography and challenge of cryptic (unrecognized) species when sample collection and species selection is done. The ToL program is synergistic with several ongoing systematic initiatives such as PEET (Partnership for Enhancing Expertise in Taxonomy), PBI (Planetary Biodiversity Initiative), and the Systematics and Biodiversity programs. Although it is not the goal of ToL to produce monographs and enhance collections, the Program should support these systematic activities through thoughtful broader impacts and inclusion of these scientists in the design of species sampling, vouchering and implementation of ToL projects.

Summary and Recommendations
1) To date, the AToL program has not encouraged proposals that examine the “tips” of the ToL, its extant diversity, and speciation patterns. The synergy of historical biogeography, species ecology, and climate change provides the Program with an opportunity to capitalize on this emerging area of inquiry. Recognition of the need for horizontal integration (across the tips of the tree of life) is key to the next phase of this program.
2) The AToL program has not encouraged explicitly the infrastructure and traditional taxonomic activities (museum vouchering, monographs) that is required for filling in the gaps in our ToL. Through broader impacts and relevant collaborations, these activities should be encouraged for microbes as well as multicellular life.

Final Discussion
After a short break, all participants assembled for a review of the entire workshop and discussion of “next steps.” Debashish gave a brief presentation highlighting the need for two directions in the ToL program – vertical integration (phylogenetic from root to shoot) and horizontal integration (across species, comparative studies, see Figure 1). The vertical integration represents the continuation of the Program’s goal for discovering and documenting the Tree of Life, and the horizontal integration represents building relevance and synergy of this program to other disciplines within Biology and within NSF (Figure 2). Given that evolution is the theory that unites life, horizontal integration is required to provide tree-thinking support to other biological disciplines.

For the vertical integration, the ultimate goal of the Program needs to be more clearly articulated; i.e., is it to have all 1.7 million described plant and animal taxa in the tree? If so, how should this be accomplished effectively - which gene or which approach (i.e., single or multigene phylogenetics)? Given our desire to “fill” all of the tips of the ToL, we reaffirm the need to actively encourage AToL proposals that address missing clades from the tree (e.g., red algae) and species that would offer benefit in multiple directions. Participants re-emphasized that microbes should be seriously considered when setting goals for how many and which species should be surveyed.
Ultimately, the AToL Program can have its greatest impact by facilitating the vertical and horizontal integration of NSF-funded research by providing a set of conceptual approaches and tools rooted in systematics that can organize and unify life-science research at the Foundation. This view is represented by the summary figure shown below.

Horizontal and vertical integration of scientific pursuits of the AToL Program and its impact across the life sciences. Only a subset of the applications are shown and their specific phylogenetic ‘level’ of use is flexible (i.e., developmental biology applications are also clearly important at the tips of the ToL).

The integration of AToL Program activities across existing NSF funding initiatives and public outreach.

**Figure 1.** Vertical integration of the AToL Program. Different types of data are collected to address different outstanding issues in resolving the eukaryotic ToL. For prokaryotes, data needs primarily involve complete genome sequences that are relatively less expensive to determine and computationally less challenging to analyze than from eukaryotes.

**Figure 2.** The integration of the AToL Program with complementary programs at the NSF and other existing initiatives.
The Future of AToL
Workshop Summary and Recommendations

The proposed workshop had an ambitious agenda that addressed several major and distinct areas of research impinging on the AToL Program. The active discussions that resulted at the venue provided some key insights that can be used to plan the future of the Program.

Major Workshop Outcome:
Is There a Single Tree of Life and How Should We Pursue It?
Attendees agreed that upcoming complete genome and transcriptome analyses will ultimately allow us to gain the broad understanding of eukaryotic genome evolution required to forge a single eukaryotic ToL. The AToL Program can be instrumental in making this happen via funding targeted genome-level projects. The ToL for prokaryotes is far more controversial due to the confounding issues that arise when the gene tree = species tree relationship is broken. HGT is however a biologically fascinating and fundamental process in its own right and the Program can play a leading role in this area by funding projects that clarify the role of HGT in prokaryote cell and genome evolution. It was agreed that for both eukaryotes and prokaryotes phylogenomics provides a robust and powerful tool for identifying “core” genes appropriate for reconstructing a single ToL. Phylogenomics will also play a major role in identifying HGT. Much thought needs to be given however to the computational needs/limitations that will arise when researchers are faced with the prospect of including hundreds of genomes and millions of genes as input for automated phylogenetic pipelines.

Greater international collaboration and integration of tools, projects, and organismal expertise are seen as critical to managing this level of growth for the AToL program. The goal of generating a single taxonomically broadly sampled ToL was considered of great value and the means to approach this grand goal need to be thoroughly discussed. This scaffold will by definition provide many more tips on the ToL where aspects of development, gene duplication, cellular function, and co-evolution can be most effectively pursued. These insights impinge on and inform the processes of speciation and adaptation.

Summary:
Given the exciting, emerging frontiers in systematics, as presented by our plenary speakers, many different roads could be pursued by the AToL Program. AToL projects have significantly advanced our understanding of phylogenetic relationships within many of the major lineages of life and revitalized systematics research. However, Terry Yates’ vision of a single relatively well-resolved and stable tree containing >1 million species that are more or less uniformly distributed across all biodiversity will not happen with the current AToL program model.

Most attendees strongly recommended the creation of an international steering committee to identify the gaps in the Program and to provide guidance on strategies to close these gaps. Major issues to be resolved include:

1. Identifying the approach and level of effort to be spent on prokaryotic and eukaryotic genomics to address the role of HGT in gene/genome evolution.
2. Assessing the need for infrastructural support for large-scale phylogenomic projects and the archiving of these (and other) phylogenetic data sets.
3. Appraising the virtues of a sequencing center that is devoted to generating target gene sequences (identified via empirical science) for generating an unified ToL as opposed to this project being coordinated and led by individual PIs.

4. Assess how to continue encouraging collaborations between ToL projects and other disciplines such that tree-thinking, the core theory of biology, and ToL results are rapidly and effectively integrated into other research areas.

Most critically, the AToL steering committee should be charged with determining short-term and long-term milestones for the program. As each milestone is reached, it can be celebrated publicly. Progress towards understanding the ToL can be communicated to the public. We envision announcements and media attention similar to those surrounding the completion of the human genome sequence and more recent model organism genomes.

The four areas highlighted by our plenary speakers provide a robust beginning for the steering committee to begin this gap identification and remediation work. From each, the community’s first attempt to provide guidance on these gap priorities and possible strategies to address them is provided in this report. Their main points are summarized here.

**Vertical Integration**

1. Addressing the single tree of life.
   - The program should include a strong focus on the ancestry of life starting from prokaryotes and including both microbial and multicellular eukaryotes.
   - The program needs to specify a goal for taxon-sampling. There could be several different taxon-sampling schemes (see figure below) and priority setting by an independent group is needed. Taxon discovery will need to be accommodated because novel, ancient lineages could be critical in resolving phylogenetic relationships near the base of the tree.

2. Phylogenomics.
   - Phylogenomics promises to provide a large set of “core” genes conserved cross eukaryotes that can be used to infer the eukaryotic ToL. A set of core genes combined with dense taxon sampling will be needed to address eukaryote relationships of the ToL.
   - To date, little coordination of gene selection among projects has occurred. Consequently, assembly of the tree will be difficult if the current course remains unchanged. But, the mechanism for ensuring a coordinated approach to data collection is unresolved and contentious. The steering committee will have to examine whether there is a role for a specialized AToL center (or centers) to specialize on understudied taxa or whether another approach (yet to be fully identified) is more in keeping with NSF’s commitment to question-driven science.

**Horizontal Integration**

3. Origin of evolutionary innovations.
   - The AToL Program should not be limited to just tree assembly. Important synergies exist with emerging (ontology studies) and reinvigorated disciplines (development). Program flexibility is needed to include projects with broader aims (e.g., the study of development, gene duplications, cellular functions, co-evolution), particularly when groups include well-studied model organisms.
Comparative morphology whether derived from fossil or living taxa should be included in AToL projects, either in the data matrices (higher eukaryotes) or as value added to the final tree (typically prokaryotes, protists). These data permit a much richer understanding of the mode and tempo of evolutionary processes responsible for diversification of species (microevolution; gradualistic type processes) and body plans (macroevolution; punctuated type processes).

4. Speciation, biological diversity, and biogeography.
   - The AToL program should encourage proposals that examine the “tips” of the ToL - particularly extant diversity, and speciation patterns. The synergy of historical biogeography, species ecology, and climate change provides the Program with an opportunity to capitalize on this emerging area of inquiry.
   - The AToL program has not encouraged explicitly the infrastructure and traditional taxonomic activities (museum vouchering, monographs) that is required for filling in the gaps in our ToL. Through broader impacts and relevant collaborations, these activities should be encouraged for microbes as well as multicellular life.

**Immediate Next Steps for the AToL Program**

- Establish an international Steering Committee to coordinate the AToL community. This committee will develop specific milestones and strategies for the AToL program. They will also develop and maintain ties to the international community involved in ToL research.
- Plan for public announcement of Tree of Life completion of milestones – recognizing that the completion of the initial Tree, with all major lineages placed, will be one of science’s major accomplishments.

**Overall, the workshop participants felt strongly the use of “tree thinking” (rooted and originated in systematics) was critical to allow the integration of intellectual approaches and scientific areas of enquiry to answer “big” questions.**
Acknowledgements

This workshop could not have taken place without the help and dedication of many individuals. The organizers wish to express their sincere thanks to the workshop participants for taking time out from their busy schedules to attend this event and the following individuals for their key roles in ensuring the success of the undertaking: Pat Herendeen (Division of Environmental Biology, National Science Foundation), Maureen Kearney (Division of Environmental Biology, National Science Foundation), Joe Miller (University of Iowa, Workshop Coordinator), Susanne Ruemmele (University of Iowa, Web and Report Graphic Designer), Heather Tyra, William Lanier, Ahmed Moustafa (University of Iowa, Student Rapporteurs), Andy Anderson (Southern Illinois University, Organizing Committee), Josep Comeron (University of Iowa, Organizing Committee), Doug Eernisse (California State University, Fullerton, Organizing Committee), and Matt Kane (Division of Environmental Biology, National Science Foundation). Special thanks also go to Peter Crane (University of Chicago) for providing a first draft of the workshop proposal, Scott Lanyon (University of Minnesota) for helping summarize the workshop proceedings, and Tom Koeppel (Biology, University of Iowa) for logistic support.
References


Dagan T, Martin W. 2006. The tree of one percent. Genome Biol. 7:118.


This workshop was supported by the National Science Foundation. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the participants, and do not necessarily represent the official views, opinions, or policy of the national Science Foundation.