Collaborative research: ABI Development: Ontology-enabled reasoning across phenotypes from evolution and model organisms

PROJECT SUMMARY

Despite the centrality of phenotypes to biology, the standards used to record and communicate information about phenotypes are discipline-specific and typically limited by the constraints of natural language. Consequently, phenotypes are refractory to attempts at data integration across studies and inaccessible to the larger network of linked machine-readable biological data. Here, we propose development of ontology-driven tools for machine reasoning over phenotype data, building on widely adopted Entity-Quality (EQ) syntax for phenotype descriptions. We aim to develop tools that are adequate to enable machine reasoning over the extensive volume, and diverse nature, of skeletal phenotypes in fossil and modern vertebrates.

Specifically, we propose the following. One, we will develop a fast semantic similarity engine to search a large EQ database for biological taxa or genotypes bearing a profile of EQ phenotypes that are similar, but not necessarily identical, to a query profile. Two, we will develop an ontological framework for reasoning over homology that can be scaled to a large number of anatomically diverse evolutionary lineages. Three, we will work to reduce the time and cost of obtaining EQ statements from the literature, while at the same time improving the quality and consistency of those statements, by incorporation of proven natural language processing tools and by improving curation software to allow for on-demand augmentation of community ontologies. Four, we will build umbrella taxonomic and anatomical ontologies for the vertebrates, the latter to be supplemented by explicit homology relations among anatomical structures. Five, we will tie the ontologies and software tools together with information extracted from the vertebrate systematic literature into a knowledgebase integrated with genetic and phenotype data from three vertebrate model organisms: zebrafish (*Danio rerio*), frog (*Xenopus laevis*), and mouse (*Mus musculus*). We will expose this knowledgebase to generic reasoners using Ontology Web Language standards. As a capstone, we will assess how well we can apply machine reasoning to retrieve candidate genes for the well-studied vertebrate fin-limb transition and other major events in skeletal evolution.

Intellectual Merit: The proposed work will enable a diversity of applications that depend upon the interoperability and computability of phenotypic knowledge across biological domains (including developmental biology, genetics, systematics, and paleontology). It does so by strategically leveraging the tools and standards developed to support the machine-readability of genetically characterized phenotype data from model organisms, and the extraction of information from the biodiversity literature. Our development plan addresses the specific challenges of provisioning, and reasoning over, a large knowledgebase of phenotypes from anatomically diverse organisms, including the efficiency of data curation, the scalability of semantic similarity search, and the logical implications of biological homology. The knowledgebase will open up decades of research on the comparative anatomy, homology, and evolution of vertebrates and, in so doing, enable the research community to undertake open-ended investigation of the diversification of vertebrate phenotypes, and its genetic basis, over the last half billion years.

Broader impacts: The collaborators on this project represent diverse stakeholder communities, but we are unified in our commitment to the development of community standards and resources for the interoperability, and computability of phenotypic knowledge. We are also committed to the rapid and open release of the variety of products that we anticipate to be of immediate value to the greater biology community, including tools for streamlining data curation and performing large-scale semantic similarity searches, high quality vertebrate taxonomy and anatomy ontologies, and standards for reasoning over homology. The project will actively engage both biological and informatics stakeholders through workshops and coordination with key research networks. We will provide a unique training environment for students, postdocs and summer interns, including Native Americans through outreach at the University of South Dakota and minority and female students though a collaboration with Project Exploration at the University of Chicago.

TABLE OF CONTENTS

For font size and page formatting specifications, see GPG section II.B.2.

Appendix Items:

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

TABLE OF CONTENTS

For font size and page formatting specifications, see GPG section II.B.2.

Appendix Items:

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

PROJECT DESCRIPTION

OVERVIEW

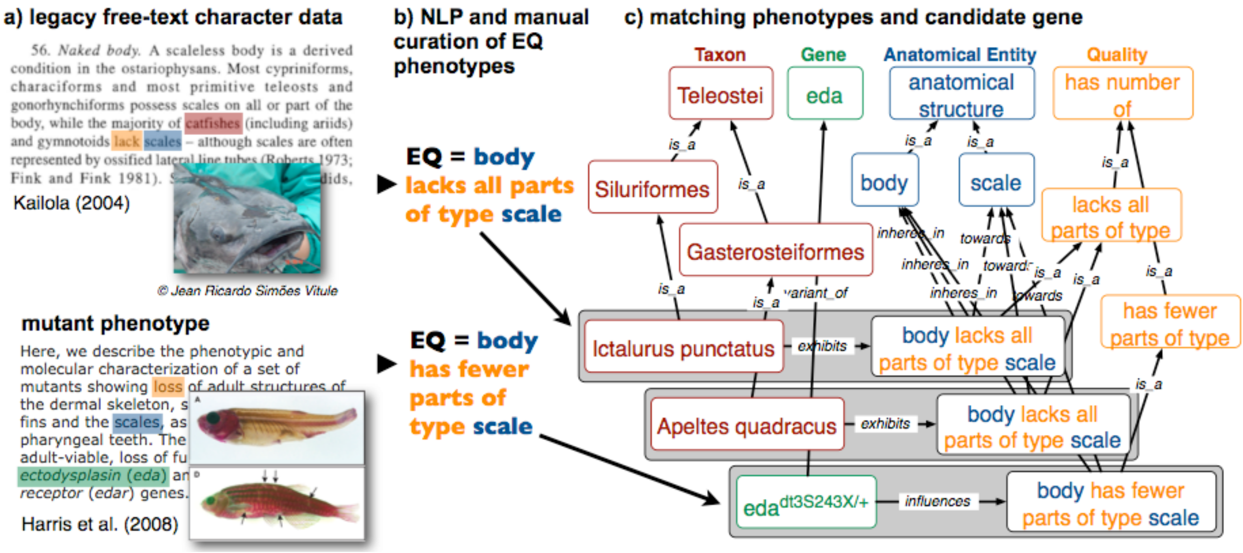
The product of evolution at the organismal level is the phenotype, or the set of observable traits present in an individual organism as a result of the interaction of heredity, environmental influences, and the developmental process. The spectacular variation in phenotype – from fishes with taste buds on their fins to clawed and feathered dinosaurs – captivates public interest and motivates scientists to understand its causes and consequences. This is evident in the many different ways that biologists investigate phenotype: developmental biologists study the unfolding of the phenotype from the moment of fertilization; geneticists infer gene function through the phenotypic effects of allelic differences; evolutionary biologists use phenotypes to inform and interpret phylogenies in living and fossil organisms; taxonomists identify and describe the characteristic phenotypes that distinguish taxa from one another; ecologists study phenotypic adaptations to the environment.

Despite the centrality of phenotypes to so much of biology, the methods used to record and communicate information about them are discipline-specific and limited by the constraints of natural language. Consequently, phenotypes are mostly inaccessible to the larger network of linked machine-readable biological data [3] and unyielding to attempts at data integration that would allow computational analyses across studies [6].

Here we propose to represent phenotypes in a way that can overcome these barriers to communication and computation by using an approach that is grounded in ontology-based knowledge representation [7-9]. Briefly, the “Entity-Quality (EQ) formalism” decomposes phenotype descriptions into three main components: a phenotypic quality (Q), such as an 'elongated' shape; the entity that is its bearer (E), such as an anatomical structure; and the organismal entity that exhibits the phenotype, either members of a taxon or the carriers of an allele. Phenotypes in EQ-format consist of terms from requisite ontologies for each component, and well-defined relationships that render them formal logic expressions. Such expressions are interoperable regardless of their source. Moreover, given the ontologies from which the terms are drawn, and a formal specification of rules about combinations and chains of relationships, domain-agnostic machine reasoners can then infer facts that are implied, but not asserted, among a set of phenotype descriptions. Such inferences may combine facts from different studies, and even from different kinds of studies (e.g., evolutionary character transitions and genetic mutants).

The EQ approach was first developed by the model organism community to integrate mutant phenotypes across different model organism species in a way that enables knowledge discovery and hypothesis generation [8,12]. Through the Phenoscape project (see Results from Prior Support), we have adopted this approach for evolutionary phenotypes described as characters and character states in phylogenetic studies published in systematics literature [15]. The fruits of that labor are published [16-18] and available online in a knowledgebase that integrates genetic phenotypes of zebrafish with evolutionary phenotypes of teleost fishes [5]. This knowledgebase has already generated hypotheses about the genetic basis of phenotypic evolution within freshwater fishes that are being tested in PI Westerfield’s lab (Fig. 1).

Here, to help realize the vision of computable phenotypic information across the tree of life, we propose to develop tools and approaches that will overcome some of the limitations in researchers’ ability to assemble, reason over, and search large collections of semantically encoded phenotypes from diverse organisms. Washington et al. [12] have shown that genetic and biochemical processes relevant to human clinical phenotypes can be discovered through phenotype similarity comparisons with mutant model organisms; however, the computational sophistication and time needed to use this approach prohibit its application by the biological research community at large. Furthermore, reasoning across EQ statements, particularly from distantly related organisms, requires a more rigorous treatment of the evolutionary relatedness among anatomical features from different organisms (i.e. homology) than currently available. There are also bottlenecks in the human effort, specifically domain ontology building and manual curation of EQ phenotypes, that currently prohibit scaling the approach to more diverse groups of organisms [17].


a) legacy free-text character data

56. *Naked body.* A scaleless body is a derived condition in the ostariophysans. Most cypriniforms, characiforms and most primitive teleosts and gonorhynchiforms possess scales on all or part of the body, while the majority of catfishes (including ariids) and gymnotoids lack scales – although scales are often represented by ossified lateral line tubes (Roberts 1973; Fink and Fink 1981). S… …dids,

Kailola (2004)

© Jean Ricardo Simões Vitule

mutant phenotype

Here, we describe the phenotypic and molecular characterization of a set of mutants showing loss of adult structures of the dermal skeleton, s… fins and the scales, as… pharyngeal teeth. The… adult-viable, loss of fu… *ectodysplasin (eda)* an… *receptor (edar)* genes.

Harris et al. (2008)

Fig. 1. Scale loss is common in fish evolution, and species that lack scales, such as all catfishes [2] and sticklebacks in the genus *Apeltes* [4], can be quickly discovered in the KB [5]. The KB also shows that 6 genes are associated with scale loss (from zebrafish mutant data), one of which is eda [10]. Experimental work already has shown that *eda* is, in fact, involved in the loss of scales within *Apeltes* [11]— perhaps no surprise, given the high level of conservation of genetic pathways across phylogenetically separated taxa [13,14]. Nevertheless, computational mining of these two disparate stores of phenotype data for such connections had not previously been possible. PI Westerfield's lab is following up on this by determining the tissue-specific expression of *eda* and other candidate loci for major evolutionary novelties unique to catfishes.

Making this vision feasible requires strategic focus on a biological domain that is diverse enough to expose the underlying difficulties, yet small enough to comprehensively annotate. Additionally, there should be ontology resources in place to leverage semantically compatible data on phenotypes from model organism genetic studies, and a body of developmental genetic research to validate the test results of machine reasoning. To meet these criteria, we have chosen to focus on the skeletal phenotypes of living and fossil vertebrates (Fig. 2). Vertebrates include the most well-characterized organisms on earth including humans, and their 500+ million year radiation and diversification provide some of the most spectacular examples of anatomical transformation in evolutionary biology. Prior investment in the Phenoscape knowledgebase for fishes, as well as the amphibian anatomy and taxonomy ontologies from the AmphibAnat project, already provides coverage for the majority of vertebrate species. We can also leverage the extensive knowledge and ontology-based resources of the major vertebrate model organisms (mouse, frog, and zebrafish) to map phenotype to genotype at a scale never before attempted.

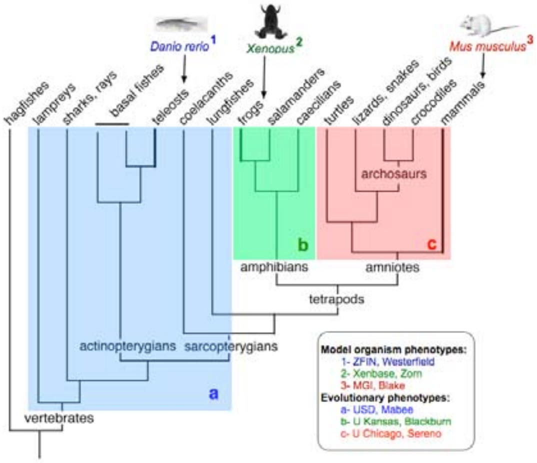


Fig. 2. Evolutionary relationships of verte– brates [1]. Correspondence of taxonomic pheno–type coverage with model organisms.

We further focus our initial knowledge acquisition on the evolutionary transition from fish fins to tetrapod limbs, one of the most prominent and well-studied phenotypic changes as vertebrates moved from water to land ~360 million years ago [19-24]. We will supplement the knowledgebase with high-level skeletal phenotypes, i.e., those that characterize major taxonomic clades, for two reasons. One, this will allow hypotheses to be generated and tested for phenotypes that may be correlated with the fin-limb transition.

Second, because phenotypes in higher taxa can, as a rule, be propagated to large numbers of individual species, this will allow us to test the scalability of the knowledgebase.

The specific development goals are designed to collectively address the challenges described above, and apply these solutions to the test case of the vertebrate fin-limb transition. The goals and deliverables are as follows:

I) **Fast semantic similarity search toolkit for phenotypes.** We will develop fast, rigorous, and user-friendly software that can take an EQ phenotype profile for a particular organism or genotype and retrieve statistically similar profiles from a large EQ store, using the structure of the ontologies to enable approximate term matches.

II) **A framework to capture and reason over homology.** We will create a logically rigorous and generic methodology to reason over assertions of homology and annotate homology assertions from the literature for the vertebrate skeleton.

III) **Scalable workflow for data annotation and ontology development.** We will integrate Natural Language Processing (NLP) into the annotation workflow to automatically generate candidate ontology terms and EQ statements from text, and we will create software that decouples annotation from ontology building to remove a major curation bottleneck.

IV) **Vertebrate ontologies and phenotype annotations.** We will generate >2 billion evolutionary EQ phenotypes for extinct and extant vertebrates, with a focus on skeletal phenotypes and those assignable to higher taxonomic units. To enable this, we will build the vertebrate anatomy and taxonomy ontologies.

V) **Integrated knowledgebase for Phenoblast and web-based exploration.** We will integrate the ontologies, homology assertions, EQ phenotype annotations, model organism genes and phenotypes, and wild-type gene expression, and semantic search tool into a knowledgebase (KB). The KB will enable discovery through multiple machine and human interfaces.

As a **capstone**, we will validate the capabilities of this suite of tools by testing how well it identifies known developmental pathways for well-studied skeletal transitions in vertebrate evolution and how well it scales to a datastore containing billions of phenotypes.

BACKGROUND

Ontologies have become a foundational technology for capturing and computing with biological knowledge. Ontologies are formal, machine-interpretable representations of knowledge domains in the form of well-defined terms and the relationships that hold between them. When used to annotate biological observations otherwise described in free text, the definition of terms helps ensure that they are used with consistent semantics that reflect the state of knowledge in the field. The relations between terms, together with rules for relation properties (such as symmetry or transitivity) and chains of relations (such as A *is_a* B and B *part_of* C yielding A *part_of* C), allow machines to understand the semantics and infer facts implied by a given set of assertions.

The most widely-used bio-ontology is the Gene Ontology (GO), which has been hugely successful in making gene annotations mutually comprehensible among many databases, and has opened up gene function to computational analysis [25]. Inspired by this success, several model organism databases have pioneered the application of ontologies to phenotype data [8], specifically for describing the extraordinary diversity of shapes, sizes, positions, compositions, functions, etc. arising from mutant genotypes [26,27]. Notwithstanding the success of this approach for making model organism phenotype data comparable [12], its potential will be fully realized only when all types of phenotype data are annotated and exchanged using shared conventions for computability. Many fundamental applications that leverage machine-readable phenotype data to synthesize knowledge across disciplines remain to be developed [28], and especially when applied to the vast store of information about phenotypic differences among species in nature.

In the Phenoscape project (see Results from Prior Support), we extended the EQ syntax to map phylogenetically informative anatomical characters as described in the systematics literature to genes via mutant phenotypes [7,15,17]. As of August 2010, the publicly available beta version of the Phenoscape Knowledgebase contains over 500,000 evolutionary phenotype assertions for taxa that were manually curated in EQ format from free text descriptions in the ichthyological literature (specifically the

Ostariophysi) [5]. The evolutionary data are integrated with over 26,000 zebrafish mutant phenotypes and associated genes, described using the same EQ format. The KB can be used to generate tens of thousands of hypotheses about the genetic basis for evolutionary change in fish (Box 1), hypotheses that can be tested experimentally to elucidate the genetics of phenotypic diversification (e.g., [29]).

SIGNIFICANCE

The proposed work will synthesize for the first time the large store of knowledge about phenotypes across the breadth of fossil and living vertebrates, to generate novel hypotheses about the genetic basis of evolutionary transitions in vertebrate form. Vertebrates include some of the most familiar and intensely studied animals on the planet, including, of course, humans. The vertebrate skeleton, in particular, is at the center of research in studies of developmental and evolutionary biology, paleontology, functional biology, and biomechanics, as well as medicine.

Along the way, we will generate a number of valuable deliverables, including a semantic similarity engine for large EQ datastores that is designed for biologists to use, two badly needed community ontologies, improved software for data and ontology curation, improved database tools for ontologically annotated data, and of course the knowledgebase itself, containing billions of annotations from the vertebrate literature. Together, these deliverables, and the open, community-driven manner in which they are to be developed and disseminated, will expand the capabilities of multiple user communities and build bridges between them via shared tools and standards.

The proposed work will also lay a foundation for efficiently constructing a phenotype knowledgebase in other diverse taxa, such as plants and insects, and for other aspects of phenotype, such as behavior and biomechanics. In fact, the EQ-formalism is not inherently limited to computing over phenotypes, because the semantics of what an EQ expression denotes lie not in the syntax, but in the ontologies from which the component terms are drawn. This is but an early example of the potential applications for formal knowledge representation of descriptive biological data.

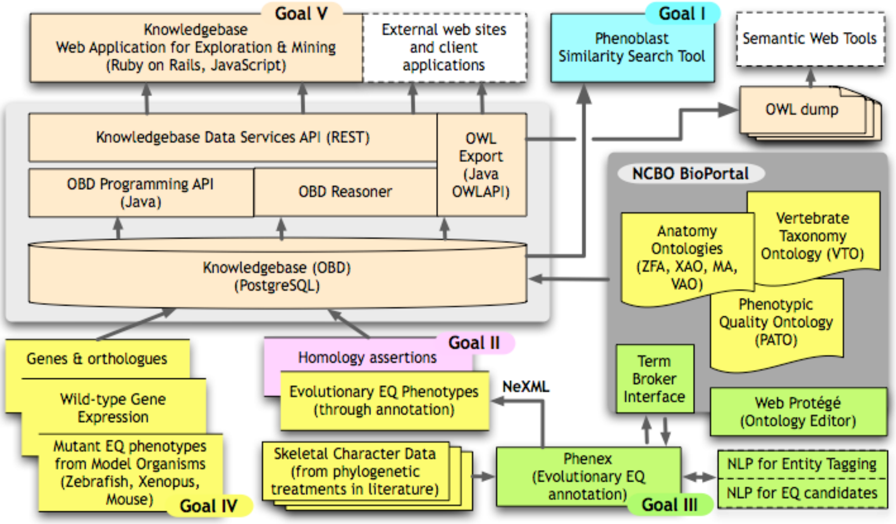


Fig. 3. Diagram of the architecture, with components color-coded by goals I-V.

DEVELOPMENT PLAN

I. Fast semantic similarity search toolkit for phenotypes

Comparing, describing, and grouping organisms according to their phenotypes is one of the oldest scientific activities, dating at least from Aristotle. Representing phenotypes in a computable manner, however, is recent. In contrast to the multitude of tools available for DNA sequence comparison [13,30-33], and the many databases to which they can be applied [34,35], the resources for computationally comparing phenotypes are rudimentary. One of the most badly needed tools is one that allows a user with a set of phenotypes in hand (i.e. a profile) to find the most similar profiles from within a large collection. Such a tool would enable a biologist to take the set of phenotypes that mark a particular evolutionary transition (e.g. from fins to limbs as vertebrates colonized the land), query a large library of model organism mutant phenotypes, and receive a ranked list of genes which, when mutated, have similar phenotypic profiles. Because similar phenotypes may be annotated with different, but related terms, and because rare phenotypes provide greater evidence for a useful match, it is helpful to use both the structure of the ontology and relative abundance of annotations in the library in discriminating among potential matches.

It has recently been shown that metrics developed for assessing similarity among other ontology annotated data [36], such as GO-annotated gene products [36-40], are well suited for phenotypes [12,41]. However, there are still major impediments to the wide use of semantic similarity search for EQ phenotypes by the wider research community. One is speed and scalability. Current algorithm implementations are slow; the implementation by Washington et al. [12] requires more than 10 hours of runtime on a relatively small database of phenotype annotations for 11 genes in three taxa (Mungall, pers. comm.). To analyze thousands of genes and taxa interactively, a speedup of at least 3 orders of magnitude would be desirable. The other impediment is usability. Complex dependencies on external software such as a database or reasoner program, the need to convert the format of source ontologies or data, and hard-coded rather than runtime-changeable parameter choices all present obstacles for non-specialists, and require dedicated software engineering to solve. Challenges to usability thus include ease of tool installation, set-up of custom target databases, and configuration of program parameters.

Here, we propose to create an open-source software package that addresses these challenges, tentatively called Phenoblast. It must be fast, simple for users to deploy on a large library of EQ statements, and succeed in ranking similarities in a biologically meaningful way. We envision the program to return an ordered list of approximately matching phenotype profiles (and their associated biological entities), along with statistically calibrated similarity scores. It will allow users to control parameters, such as the choice of metric and the score cut-off value. An advanced feature could allow filtering the results by user-specified properties that are specific to the EQ library.

We propose to use well-established distance metrics for ontology terms and ontology annotated data [9,36]. To achieve the necessary speed-up of the distance calculation, we will first investigate which calculations provide the best gain in speed if they are calculated once per library and subsequently reused. All terms in the query EQ expressions must necessarily be known to the library, and as Ovaska et al. [42] show, albeit for a much smaller problem, there is in principle much room for improving speed in this way. We will also research algorithms for fast lookup of so-called “common subsumers”, which are the terms in the ontologies of which both a query EQ term and a library EQ term are subtypes, and which are key in calculating many of the metrics. If necessary, we will evaluate heuristic criteria that accept or reject potential matches before the distance is fully calculated. We will also allow calculations that are independent of each other to execute in parallel on hardware that supports it, such as the increasingly common fast multicore processors.

II. A framework to capture and reason over homology

Anatomical features, biological processes, and biochemical pathways in different species may be similar, and hence given the same name, because they evolved from a common ancestor (homology) or because they evolved independently, and possibly fulfill a similar function (convergence, parallelism). Hence, similarity in term name does not imply that the referenced entity is homologous across species. Distinguishing between homology and convergence is fundamental to understanding the patterns of phenotypic, developmental, and genetic evolution. Although it is possible to state homology (or lack thereof) in term definitions, doing so does not enable reasoning. Therefore, most ontologies neither state homology in term definitions, nor imply it if a term is applied to different species.

Several approaches have been suggested to support reasoning by capturing homology outside of term definitions. Most of these have arisen from the need to connect terms across species-specific model organism anatomy ontologies, and use lexical similarity-based ontology alignment (see [43,44]) to create a bridge ontology that connects corresponding terms in different ontologies. For example, the Uberon ontology used by Washington et al. [12] for linking phenotypes across model organisms originated from such an alignment, although Uberon does not attempt to distinguish homology from convergence. The ontology alignment algorithm devised by Parmentier et al. [45], called Homolonto, explicitly defines homologous organ groups (HOGs), which form classes in the bridge ontology. Another bridge ontology approach, that is not dependent on ontology alignment, defines groups of homologous structures corresponding to a hypothetical structure in a Most Recent Common Ancestor (MRCA). Terms in anatomy ontologies can then be connected to MRCA terms through a proposed "*derived_by_descent_from*" relationship (see [46]), from which homologous pairs can be inferred. The MRCA method is being developed by the ArrayExpress [47,48] gene expression database team at the European Bioinformatics Institute (EBI) to create a Vertebrate Bridging Ontology (VBO) for integrating ontology-annotated gene expression data across species. Although the VBO supports reasoning across species, it does not distinguish among different kinds of homology or deal with conflicting homology assertions. Moreover, the method assumes that the MRCA is known and uncontroversial, which is not always the case.

A reasoning model that is viable for evolutionary applications must meet several requirements: because homology is a hypothesis, an assertion should include evidence and attribution; conflicting homology assertions should be supported; different kinds of homology (such as phylogenetic and iterative [49,50]) need to be distinguished and have different reasoning rules; and because the MRCA for homologous features are not always known, hypothetical MRCAs should be allowed. A recent preliminary study by Phenoscape in collaboration with C. Mungall (Lawrence Berkeley Laboratory) developed such a model and an accompanying OWL-compatible entailment framework for reasoning. We proposed to treat homology as a ternary relationship of taxon-constrained entity terms: an entity in a taxon is homologous to an entity in another taxon as an ancestral entity in an ancestral taxon. For example, the teleost pectoral fin is homologous to the tetrapod forelimb as a vertebrate limb. In addition, the "homologous as ancestral" entity or taxon need not be the most recent, but can be any ancestor that the entity and taxa share, and can thus be inferred from the entity and taxonomy ontologies. This logical framework is compatible with the HOG [45] and MRCA-based bridge ontology approaches.

Carrying this work forward requires substantial effort. None of the methods for representing homology have been thoroughly evaluated yet for criteria important for a robust implementation, including efficiency of maintenance, scalability, soundness (all inferable assertions are also correct), and recall (all correct assertions are also inferable). The phylogenetic homology relation holding between two entity classes is defined on the basis of two instance-level *derived_by_descent_from* relations between each descendant entity and the common hypothetical ancestor structure. This relation is therefore central to any approach of formal reasoning over common descent. Full definition of what the homology relation means biologically, and including it in the Relation Ontology [51] have lingered for years. Rectifying this situation, as well as testing and evaluation in an efficient manner, is best accomplished in face-to-face workshops with ontologists, homology experts, and comparative biologists. The recently NSF-funded Phenotype Ontology Research Coordination Network (RCN) and the EBI VBO group mentioned above, have committed to co-sponsoring such a workshop and collaborating on a solution. The vertebrate fin-limb transition enables a full exploration of both phylogenetic and iterative homology, and will therefore provide an ideal test bed. Once fully formalized and vetted, we will integrate the homology relation into the existing rule sets for reasoning over genes and phenotypes [52] in Phenoblast and the Phenoscape KB (see V). The logical model that we develop for reasoning over homology between entities is, to our knowledge, the first of its kind in terms of comprehensiveness and generality. Importantly, it is independent of ontology alignment, and it is not constrained to anatomical entities.

Statements of homology for vertebrate fins, limbs, and skeleton in general are scattered across a disparate and voluminous literature (development (e.g., [53-55]); comparative developmental morphology (e.g., [56,57]); phylogenetics (e.g., [58,59]); paleontology (e.g., [60-64]). A concerted effort will be required to find and represent homology statements accurately and comprehensively. We will hold a small 'homology jamboree' with fin-limb experts where participants evaluate and correct, as required, the curated homology statements, point to missing literature, and assess the user-interface and tools. Crucial in this effort is the structured and comprehensive annotation of evidence of whether or not two anatomical structures are judged to be homologues. Our goal is to annotate all of these assertions, with their sources and evidence, rather than including some and excluding others based on judgment. Standard lines of evidence for homology in the field of comparative anatomy include similarity of development, composition, morphology, gene expression and position [65], as well as computational evidence from phylogenetic analyses. Each type of evidence has a corresponding term ("code") in the Evidence Code Ontology (ECO, [66]), and we will use these codes to annotate homology assertions. If a source does not specify the evidence, we will use the ECO term 'Traceable author statement' if a citation is given, and 'Non-traceable author statement' otherwise. The concept and activity of annotating anatomical homology with structured and well-defined evidence codes is in itself potentially groundbreaking in the field of comparative morphology. Data in support of homology are dispersed across the primary literature and reviews, disorganized, and not readily available; it is nearly impossible even for domain experts to find, let alone weigh, all of the different lines of evidence supporting or opposing a homology relationship between anatomical structures. Our proposal to annotate logically computable homology assertions across a set of taxa using formalized evidence types will make these assertions broadly reusable and testable.

III. Scalable workflow for data annotation and ontology development

Developing a workflow for creating formal ontology-based phenotype annotations with sufficient throughput, accuracy, and consistency, is a major and non-trivial undertaking. Several of the challenges are specific and unique in scale for evolutionary phenotype annotation—millions of species and thousands of features that are frequently described at a high level of detail. The Phenoscape project prototyped a workflow that, albeit manual, efficiently streamlines tasks requiring different levels of domain expertise [17], and created a software tool, Phenex [16], that manages and integrates those tasks. Even so, annotating phenotypes manually is time-consuming and costly [17]. Bottlenecks in this workflow prohibit scaling to increased phenotypic diversity, such as among vertebrates. We will address these by investigating and implementing select semi-automated tools to increase annotation throughput while at the same time improving quality and consistency.

III.1. Semi-automated generation of ontology terms and phenotype annotations. In the literature the entity and quality terms describing a phenotype are buried within natural language text. In a fully manual annotation workflow, human experts parse the text and choose the appropriate ontology terms. However, the time of experts is both costly and limited. Over the past decade, the ability of automatic Natural Language Processing (NLP) tools to accurately identify parts of speech as well as terms of interest has improved tremendously, and it has been increasingly applied to literature-based data annotation. For example, co-PI Blake has recently evaluated text mining tools for incorporation into the Mouse Genome Informatics (MGI) curation system [67]. Blake is also a collaborator in the NSF-funded SciKnowMine project [68], which specifically aims to use NLP technology to accelerate knowledge extraction from the biological literature.

Our goals for using NLP to accelerate our annotation workflow are twofold. In a first step, we aim to use NLP to identify candidate entity and quality ontology terms in free text character and character state descriptions. This so-called ‘entity tagging’ is one of the most common uses of NLP and text-mining approaches. It is also a major use-case for the SciKnowMine [68] and Textpresso projects, and we will consult with the respective PIs Burns and Sternberg on the most suitable tools (see letters, P. Sternberg, G. Burns). The tagged candidate terms would be subjected to review by a human, who can remove false positives and triage the rest into new term proposals, missing synonyms, and terms already known to the ontology. Together with the facility for asynchronous ontology building (see III.2), this alone will already result in a substantial gain in efficiency. In a second step, we will try to use NLP for constructing candidate EQ expressions from the entity and quality terms that have been extracted from a given character and state. Promisingly, Hong Cui [69] has developed an NLP-based pipeline that can automatically identify character statements within free-text taxonomic treatments, and tag the morphological structure, the character attribute, and its value within them. Although phylogenetic treatments are linguistically different from taxonomic ones, an initial test of Cui’s pipeline on some of the papers curated by Phenoscape yielded encouraging results (H. Cui, pers. comm.). Cui has agreed to make her software available to us before release (see letter, H. Cui), and she is committed to advise us on tailoring it to phylogenetic treatments, including extending it so that it will propose EQ expressions, or at least the EQ component terms. We recognize that despite all optimizations the error rate may still be significant, especially for complex characters [16]; nonetheless, the method would still notably gain in scalability if it allows an anatomy expert to focus on those EQ expressions that truly require deep domain expertise, and to start from a set of proposed terms rather than from scratch.

Once an effective combination of NLP tools and pipeline is determined, we will integrate their user interaction as well as input and output into Phenex to retain a workflow that is as seamless as possible for the human curator.

III.2. Make ontology building asynchronous from data curation. Ontology building that honors best practices and includes community expertise is time consuming: terms are vetted by the community (~1 week) and then included in a released version of the ontology (~1 mo) before they can be used in EQ assignments [17,18]. This can lead curators to make ‘second-best’ EQ annotations in lieu of suggesting a better term. We will address this by making the process of ontology building asynchronous from data annotation, yet integrating them closely. Specifically, we will extend Phenex, our annotation tool, to support direct submission of new terms to an ontology management and repository platform that supports a programmable term broker interface. Through this interface, a proposed new term will immediately receive a temporary identifier, which allows the curator to create EQ expressions at the same time as the proposed term is undergoing vetting and approval within the ontology management system. Once the ontology gatekeeper makes a decision on how the proposed term is to be resolved, Phenex will programmatically obtain the resolution in the form of the approved ontology term, and subsequently replace the temporary term with the recommended one in EQ annotations created while the term proposal underwent vetting. The recommended term may be new, previously existing, or defined as a cross-product of existing terms.

The term broker interface features that are required to support this asynchronous workflow are, to our knowledge, at present not supported by an already existing ontology management and repository platform. To work seamlessly, the platform would also need to integrate the necessary API functions with an ontology-editing environment. Among the possible combinations of existing tools that could be used, we chose to collaborate with the National Center for Biomedical Ontology (NCBO) BioPortal [70] as the ontology repository and WebProtege as the ontology editor. These tools have not only received substantial resource investment from a large community, but they also already provide much of the needed functionality. The project teams are committed to implement the remaining pieces so that they meet our requirements (see letter, M. Musen), and their support for a term broker interface is synergistic with, and hence would be reusable for the needs of several other projects (such as BIRN, M. Musen, pers. comm.). The NCBO BioPortal already supports submission of new term proposals (as a type of ontology 'note') by any BioPortal user through a web-based user interface or a programmable API. Users can attach further comments to such proposals, and term proposal notes can be searched programmatically for look-up or status checking. WebProtege as the ontology editor has several added benefits for us in that it is an online tool, obviating the need for curators to install 3rd party software, and it supports collaborative editing, including the ability to grant branch-specific privileges to users. This feature will allow different branches of our ontologies (e.g., amphibians vs. fishes) to be maintained by their respective domain experts. WebProtege is being extended to integrate with the BioPortal such that a curator can pull an OWL or OBO ontology hosted by BioPortal into WebProtege, edit it there, and submit it back to BioPortal, where it would replace the previous version. This includes ontology notes, and WebProtege will allow the curator to convert a term proposal into a new term without the need to copy and paste. When fully integrated, the curator can record the identifier of either the newly added term or of a recommended existing term so that a client application (such as Phenex) can look it up. The BioPortal is also designing a notification system for changes to the ontology, including status changes to ontology notes, which will enable client tools to act upon those notifications and make the necessary replacements automatically.

IV. Vertebrate ontologies and phenotype annotations

The components of EQ expressions are ontology terms and relations between them. Hence developing the necessary ontologies is a prerequisite for the ability to annotate, and ultimately reason over, the phenotypes that use them. In practice, ontology development and phenotype annotation go hand in hand, because annotation efforts drive ontology development so that the latter meets the needs of the former. The other critical role of these activities is to continuously test the efficiency of the annotation workflow and the scalability and soundness of the reasoning tools. The end result will be the store of ontologically annotated vertebrate skeletal phenotypes, both evolutionary and from genetic mutants, that will be used in the Capstone (p. 12) to validate our reasoning tools as well as for discovering new candidate genes for phenotype changes in vertebrate evolution.

IV.1. Ontologies for Vertebrates. Applying the EQ approach to evolutionary phenotypes requires sufficiently comprehensive ontologies for organism names (taxonomy), morphological structures (anatomy), and qualities (such as shape, composition). Although the latter simply involves contributing terms to the existing cross-species quality ontology (PATO, [26,71]), two ontologies that encompass all vertebrates need to be developed: the Vertebrate Taxonomy Ontology (VTO) and the Vertebrate Anatomy Ontology (VAO).

IV.1.a. Vertebrate Taxonomy Ontology (VTO): To reason about phenotypes across the species that exhibit them, their names and taxonomic hierarchy must be formalized as an ontology. As an added benefit, a taxonomy ontology is also well suited to record various kinds of synonyms, which helps in standardizing the use of names across studies. Although some relatively comprehensive taxonomies, such as ITIS, or the National Center for Biotechnology Information (NCBI) taxonomy, exist electronically and could simply be converted into an ontology, in practice none of these taxonomies fully meets our needs if used directly. The main issues are that they are not fully comprehensive for extant vertebrate taxa; not all of the taxa we require coverage for are in their scope (e.g., extinct taxa, or for NCBI all taxa without a DNA sequence); and names used in legacy publications that have since become invalid due to taxonomic revision are not referenced as synonyms. Nonetheless, these taxonomies and other resources available for vertebrate taxon names and hierarchies [72-76], including the Paleobiology Database [77] for extinct taxa, will allow us build up the VTO in a short amount of time. We will keep the VTO updated with revisions in the source taxonomies by largely reusing the tools developed by Phenoscape project for continuously updating the Teleost Taxonomy Ontology (TTO) [78] from its source, the Catalog of Fishes [79]. Similar to TTO, VTO will also be extensively cross-referenced to the original taxonomies, including NCBI, and the Global Names Index (GNI), which ensures that our data can be cross-linked from external sequence and biodiversity databases.

IV.1.b. Vertebrate Anatomy Ontology (VAO): The anatomy ontology will provide the entity terms that bear the phenotypic qualities in the EQ formalism. This ontology does not exist yet in a well-developed form for the scope of all vertebrates, but there are several resources that we will utilize for developing the initial version of the VAO. Specifically, two multispecies anatomy ontologies exist already within vertebrates, the Teleost Anatomy Ontology (TAO) developed by Phenoscape [18], and the Amphibian Anatomy Ontology (AAO) [80]. Both cover primarily skeletal terms, which makes them well provisioned for the needs of VAO. The comparative anatomy of reptiles, including birds, and mammals, however, is not yet represented by ontologies. To begin filling this gap, the Phenoscape group recently convened some of the preeminent experts in vertebrate morphology and development to initiate the VAO [81]. This first version consists of 140 well-defined and vetted terms that form an overarching skeletal system hierarchy that applies across all vertebrates at cell, tissue, element, and system levels (see [82]). We will merge this ontology with the TAO and AAO, and extend it with skeletal terms (initially fins and limbs) as they arise from the annotation workflow (see III). The lead curators of the zebrafish, *Xenopus*, and mouse anatomy ontologies, the teleost and amphibian multi-species ontologies, and the Cell Ontology have agreed to cross-reference the VAO or import it as an upper-level skeletal system branch. The VAO will be a valuable asset for the growing number of biological data repositories that need to find datasets by anatomical location. The EBI ArrayExpress [47,48] team is developing a homology-guided vertebrate bridging ontology (VBO) for this purpose (see II). The VBO team is committed to coordinate further development with the VAO (letter, H. Parkinson).

IV.2. Phenotype annotation. The annotation effort proposed here for evolutionary and model organism phenotypes will produce to our knowledge the largest phenotype database in the world, with a projected 2.5 billion direct or inferred taxon to phenotype annotations. We are working against a backdrop of large-scale international experimental phenotyping efforts that are newly underway for the mouse [83] and planned for the zebrafish, where each of the 20-30K genes in the genome will be knocked out (deactivated), and the resulting phenotypes systematically screened and recorded. Representing these phenotypes using ontologies is considered key to analyzing the explosion of raw phenotyping data [84], and because of shared ontologies, these data will be interoperable across species [9], including humans [85]. We are well positioned to work with the international community on the challenges that will arise from these high-throughput phenotype databases: co-PI Westerfield developed and maintains the zebrafish database (ZFIN) and helps coordinate the zebrafish phenome effort. Co-PI Blake helps coordinate phenotype descriptions in the mouse database (MGI). However, the number of phenotypes in zebrafish and mouse databases projected over the next 4 years from these efforts is orders of magnitude lower (100,000's) than what we project for the KB (several billion). The forward-looking computational approaches from our work are likely to impact significantly the framework of what will someday be massive integrated phenomic and genomic data repositories.

IV.2.a. Evolutionary phenotypes: Similarities and dramatic differences in skeletal anatomy across species are recorded in the rich legacy literature of comparative vertebrate anatomy. Phenotypic descriptions are most structured in the phylogenetic systematic literature (publications from approximately 1980-present), with lists of characters and states (features and their variants in different taxa), and they comprise the most easily annotated body of literature. The domain experts and postdocs on the project will coordinate the specific selection of papers, and community input will be solicited through the Phenotype RCN. The PI and postdocs will lead the phenotype annotation (Mabee, fishes; Blackburn, amphibians; Sereno, archosaurs and amniotes, Fig. 2). These personnel will coordinate their efforts with the individuals curating fin-limb mutant phenotypes in the model organism databases.

Our annotation strategy has two phases, each with a concrete goal. First, we will curate the skeletal phenotypes for fins and limbs and their parts, for the proof-of-concept case study (see Capstone, p.12). Second, we will curate the skeletal features that characterize high-level taxonomic groups, for the purpose of creating a database that will drive advances in computing as well as enable open-ended data mining and discovery. We estimate, based in part on the previous experience of Phenoscape [17], that there will be ~300 priority papers for vertebrates with approximately 50,000 distinct skeletal phenotypes. Of these, based on the current proportion of fin-limb skeletal phenotypes in the Phenoscape KB and the ZFIN database, we estimate that approximately 10% are part of the fin-limb skeleton. We project completion of this curation midway through year two or possibly earlier, depending on the efficiencies gained from the tools developed in section III, thus providing a rich data set for the Capstone.

A second phase of curation is required to provision the KB for open-ended discovery across the vertebrate skeleton, and it also serves to create a database large enough to test the scalability of our computational components. To maximize the number of phenotypes in the database and at the same time provide a reasoning scaffold of evolutionarily significant features, we will focus annotation on skeletal phenotypes that pertain to entire taxonomic groups of species, rather than only individual species. For example, limbs evolved first in the common ancestor of tetrapods (amphibians + amniotes). Using the VTO, and given appropriate rules, the reasoner (built into Phenoblast and the KB, see Section V) can propagate an EQ of 'limb present' to all 30,000 species of tetrapods, and it can exclude from propagation all those tetrapod species that descend from a taxon, such as snakes, annotated with a mutually exclusive phenotype, such as 'limb absent'. In the user-interface, propagated annotations will be distinguished as 'inferred' from those that are 'directly observed'. We estimate that the vertebrate literature contains approximately 5,000 phenotypes pertaining to 500 high level taxonomic nodes. If each node corresponds to 1,000 species, 2.5 billion direct or inferred taxon to phenotype annotations would result, several orders of magnitude more than the current Phenoscape KB.

IV.2.b. Model organism phenotypes: Using the same standards and curation methods as for evolutionary phenotypes, we will annotate the skeletal phenotypes for genetic mutants of zebrafish, *Xenopus*, and mouse (Fig. 2). MOD curators will initially prioritize comprehensive annotation of skeletal phenotypes for the fin and limb, and subsequently of skeletal phenotypes in general. Much of these data are already routinely annotated and kept up to date by the model organism databases, but given the deluge of molecular data [86], it is critical to have dedicated attention to the skeletal phenotypes. In the zebrafish database (ZFIN, [27,87-89]), curators read all zebrafish research publications and annotate a wide range of data about zebrafish development, genetics, genomics, phenotype, including normal gene expression patterns and mutant phenotypes. ZFIN contains >24K mutant phenotypes for 9.6K genotypes, and >50K gene expression patterns of >10K genes in >2.6K anatomical structures. These numbers increase daily as more experimental results are curated. Xenbase [90] is a model organism database for *Xenopus laevis*, the African clawed frog, and *X. tropicalis*. There are presently >12K gene pages in Xenbase that contain gene expression data, including >26K annotated images. Gene expression data from transcriptome and *in situ* hybridization data are also available [91]. Creating EQ expressions for phenotypes is the next major goal for the database. The type of genetic manipulation experiments performed in frogs are different from those performed in zebrafish, which makes much of the phenotype data from *Xenopus* complementary to that obtained from mutant studies in mice and in zebrafish. Xenbase also plans to annotate skeletal phenotypes from the increasing number of *X. tropicalis* mutant screens. Mouse Genome Informatics (MGI, [92]) is the model organism database for the laboratory mouse. MGI curators identify and annotate >12K publications each year reporting mouse experimental research in the areas of genome structure, functional and comparative analysis, gene expression, phenotype, and tumor biology. Presently, MGI contains 24.9K mutant alleles for >8.6K genes. Over 36.2K genotypes are associated with phenotype annotations using the Mouse Phenotype Ontology, and >435K expression results reference >9.3K genes. New annotations for phenotype, function and expression data are posted nightly as part of regular updates of experimental data reported in MGI [93].

V. Integrated knowledgebase (KB) for Phenoblast and web-based exploration

We will combine the ontologies, homology assertions, and EQ phenotype annotations from above, together with critical auxiliary information such as the spatiotemporal pattern of gene expression in the model organisms, to create an integrated Knowledgebase (KB). We will build upon the existing Phenoscape KB [5], incorporating key enhancements such as the ability to perform semantic similarity search using Phenoblast, and exposing the datastore to generic reasoners using open standards. This will provide a community resource of lasting value for linking and mining knowledge about evolutionary changes to the skeleton and the genetic basis of development in model vertebrates.

The Phenoscape KB is based on the open-source Ontology-Based Database (OBD), originally developed by C. Mungall, and also used as the integration, query, and reasoning platform for the Washington et al. [12] study. Phenoscape added custom views and functions, a database warehouse component to simplify and accelerate complex data queries, several data loader tools (e.g., for standard NeXML files), and a REST-based web service API. Work is ongoing on the web-based user-interface. We will make a number of key improvements to the Phenoscape KB in keeping with the goals of this proposal and to increase its value as a community resource.

V.1. Supporting semantic similarity search and homology reasoning. We will integrate Phenoblast (see I) into the web-interface, including support for filters based on gene expression location. We will also revise the OBD reasoner, the data loader tools, and the data services to fully implement the homology reasoning framework (see II).

V.2. Integrating gene expression. To generate hypotheses about the genetic bases of evolutionary transitions, we need to know where genes are expressed. The three vertebrate model organism databases (Fig. 2) curate gene expression data that are precisely annotated to anatomical positions (and developmental stages) using anatomy ontology terms. We will develop data loaders and reasoning rules within OBD to import these data and integrate them into the KB, and allow the expression data to be queried. Users searching the KB for candidate genes using Phenoblast will be able to filter the list of candidates by location of gene expression. Conversely, location of gene expression can also be used to broaden a set of genes from those associated with the shared phenotypes in mutants to all those expressed in any or all of the anatomical structures of interest. A researcher can then verify whether the putative lack of phenotype has indeed been observed in experiments, or whether such an experiment has not yet been done. We will also include mappings of related genes (i.e. orthologs and paralogs) between the model organisms, so that results can be grouped by relatedness.

V.3. Support for correlated phenotypes. Interdisciplinary scientific discovery in an *in silico* environment is dependent on an interface that delivers results to multiple communities in an intuitive manner. From the outset, PhenoscapeKB has been user-driven and highly responsive to community input. The Phenoscape project team has heavily invested in user interface development during the past two years, including several formalized rounds of user-testing with the target audience that includes genetic, developmental, and evolutionary biologists [94,95]. This has resulted in a basic set of user-driven interfaces, including faceted browsing and tree visualization, that are currently being put into place and that enable queries traversing phenotypes to genes. Needs-analysis feedback from potential users consistently points to the nontrivial nature of providing an intuitive interface for the search and navigation of a large store of phenotype annotations [96-98]. Recognizing this, we plan to continue involving users in the design of the KB interface and formal evaluation of its usability. One example of a high priority feature that is not currently supported is support for exploring correlated phenotypes. A correlation between phenotypes among multiple evolutionary transitions, and the co-occurrence of the same correlated phenotypes in a genetic mutant, is suggestive evidence for a candidate gene. Until now, despite its importance, assessing the correlation or independence of phenotypes across evolutionary studies has been extremely difficult [99], but this is easy to achieve within the KB. We will develop interface components that can identify phenotypes in the KB that are exhibited together more or less frequently than expected by chance. At a minimum, this will use the structure of taxonomic hierarchy. Optionally, it will be based upon a user-selected phylogeny, which will generally have greater resolution of evolutionary relationships. In addition to being in demand by users, this feature will be important for the Capstone goal (p. 12).

V.4. Enabling open-ended mining and repurposing of the Knowledgebase. The principle language for representing knowledge, annotating data, and reasoning on the semantic web [100] is OWL, the web ontology language [101]. A data-driven and domain-neutral semantic web [102] is rapidly becoming reality, and it is our aim to maximize the opportunities for all of science to repurpose our data in ways we have not conceived by making our EQ-based phenotype data fully reusable by generic reasoning tools that use OWL. We will work with OBO and NCBO to ensure standards compliance, and to that end have included Alan Ruttenberg, of the W3C OWL Working Group, on our Advisory Board. For proof-of-concept, we will deploy the open-source edition of the OpenLink Virtuoso semantic web software platform [103] and explore its potential for supporting both basic and complex use-case queries on the OWL dumps generated from OBD.

Capstone: Testing performance and scalability.

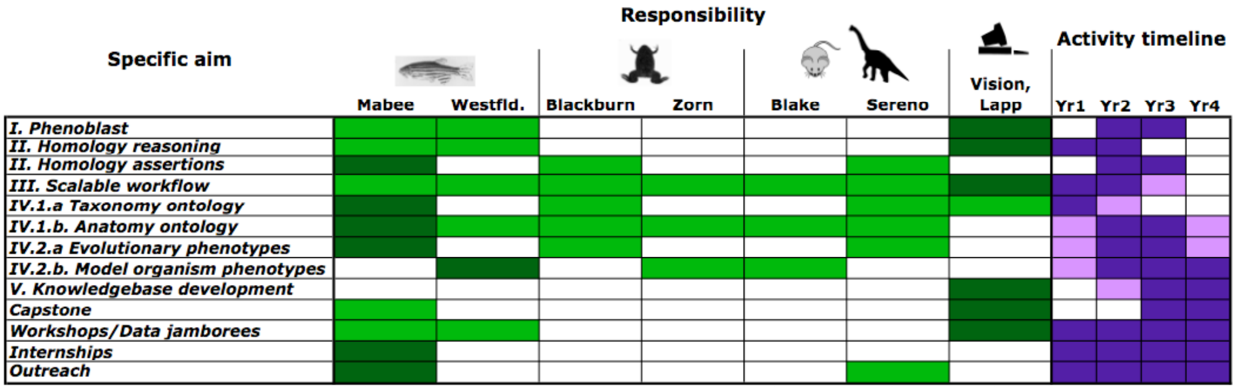
Our aim is for the KB to generate hypotheses for candidate genes that stand up to experimental scrutiny. In the absence of wet-lab work to test the validity of each candidate, success can still be judged by determining how often genes known to be involved in a well-studied trait are retrieved by the system [12,104]. We propose to use transitions such as the well-studied evolution of limbs from fins as a test system [19-24,105]. There are extensive and detailed legacy phenotype data and homology assertions for the fin-limb skeleton, including assertions of both phylogenetic and iterative (serial) homology [50], and the genes involved in growth and patterning of the limb at various stages, e.g., *Bmps, Fgfs, Gdf5, Sox9* are also well known, e.g., [20]. A search of the KB using a profile of the phenotypes associated with the fin-limb transition should retrieve a list of high scoring genes that is enriched for genes that had been preselected based on independent and external evidence (e.g., experimental studies, membership in an implicated developmental pathway). Ground-truthing the KB in this way will allow us to judge the reliability of candidate gene predictions for the many striking evolutionary transitions in vertebrate skeletal phenotypes where relatively little is known about the genetics (e.g., elongation of the snout, enlargement of the eyes, eyes set closer to the top of the head, ear region smaller, lower jaw larger, reduction in size of hyomandibula bone from part of the jaw to part of the ear, a mobile neck, loss of lepidotrichia in the fins, etc. [19,21]).

One risk is that the tools we develop would be adequate for a limited test case such as the vertebrate fin-limb skeleton, but fail to scale well when applied to other systems. This can be thought of as the 'beetle problem'. There are an order of magnitude more beetle species than all vertebrates, and a much larger knowledgebase would be needed to capture diversity even within this one order of insects. To help ensure that the tools developed will not be of restricted utility, in Yrs 3 and 4 of the project, our approach in IV.2.a (above) will be to focus on annotating phenotypes inhering primarily in higher vertebrate taxa. This will have the effect of generating large numbers of inferred phenotypes for individual species, genera, etc. The resultant datastore, which we expect to contain several billion annotations, will provide an excellent testing ground for improving the scalability of the OBD reasoner, Phenoblast, and the other components of the toolkit, and will force us to address this issue directly.

MANAGEMENT PLAN

The team brings together a tremendous breadth of biological and computational expertise and a shared vision to transform biology by enabling computers to process phenotypic knowledge.

A. Responsibilities and timelines. Mabee and Vision will oversee the co-PIs, senior personnel, and contractors. They will coordinate project communication, reports, outreach activities, collaborations, and publications. NESCent staff, under the direction of PIs Vision and Lapp, have primary responsibility for software development and will maintain hardware and software resources. NESCent will employ Chris Mungall at Lawrence Berkeley Laboratories (see letter), who brings extensive experience in knowledge representation; he will contribute to the homology reasoning framework, asynchronous ontology development tools, and Phenoblast implementation. A bioinformatics graduate student at UNC under the supervision of Vision will also contribute to Phenoblast. All co-PIs and senior personnel will serve as software development “clients” to ensure it meets project needs. Mabee will coordinate annotation of phenotypes and homology among the vertebrate groups and will work with PI Vision on the Capstone. Wasila Dahdul is currently a USD (Mabee) postdoc in residence at NESCent who has three years of experience with Phenoscape; she will be responsible for curator training, consistency evaluation, and guidance for ontology development. Curation of skeletal phenotypes and associated ontology alignment and database work for zebrafish (Westerfield), *Xenopus* (Zorn), and mouse (Blake) will be done at the U. Oregon, Cincinnati, and Jackson Labs respectively. Peter Midford (see letter), an independent researcher in residence at NESCent with over 20 years of experience in NLP, ontologies, and comparative biology will be employed by NESCent to (a) assemble the taxonomy ontology and (b) the implementation of NLP. Workshop logistics will be coordinated by NESCent staff.



Specific aim	Responsibility							Activity timeline			
	Mabee	Westfld.	Blackburn	Zorn	Blake	Sereno	Vision, Lapp	Yr1	Yr2	Yr3	Yr4
I. Phenoblast	▨	▨					■		■	■	
II. Homology reasoning	▨	▨					■	■	■		
II. Homology assertions	■		▨			▨			■	■	
III. Scalable workflow	▨	▨	▨	▨	▨	▨	■	■	■	▨	
IV.1.a Taxonomy ontology	■		▨			▨	▨	■	▨		
IV.1.b. Anatomy ontology	■	▨	▨	▨	▨	▨		▨	■	■	▨
IV.2.a Evolutionary phenotypes	■		▨			▨		▨	■	■	▨
IV.2.b. Model organism phenotypes		■		▨	▨			▨	■	■	■
V. Knowledgebase development							■		▨	■	■
Capstone	▨						■			■	■
Workshops/Data jamborees	▨	▨					■	■	■	■	■
Internships	■							■	■	■	■
Outreach	■					▨		■	■	■	■

dark shading: primary responsibility or activity; light shading: secondary responsibility or activity

B. Advisory board. An advisory board with diverse expertise has been partially assembled; an additional member will be invited as per suggestions. Committed members include: Alan Ruttenberg (OWL, OBO Foundry, W3C, and Science Commons); Brian K. Hall (Dalhousie University, vertebrate skeleton expert); Hong Cui (University of Arizona, expert in NLP applications to biodiversity literature) and Peter Vize (University of Calgary, developer of Xenbase, developmental biologist). The board will attend annual project meetings to evaluate progress and provide guidance.

C. Coordination. The key to effective collaboration among participants in different locations is frequent and regular communication. Through Phenoscape, a core of the PIs have already developed a robust communication framework. Mabee, Vision, Westerfield, Lapp, and other Phenoscape personnel have monthly conference calls, an active set of mailing lists, multiple face-to-face all-hands meetings per year, web-based file sharing, version control for software and data, and extensive documentation of progress and news through the Phenoscape wiki and blog [106]. We propose to continue using these communication outlets. All the senior personnel on the present project have already interacted through workshops over the past several years, and there has been regular communication through OBO listserves. Additionally, the project goals have been strongly informed by an April 2009 workshop on coordination of phenotype ontologies (jointly organized by Phenoscape and AmphibAnat), in which 18 researchers in fields as diverse as developmental genetics, molecular evolution, systematics, biomechanics, paleontology, and behavior identified the needs of their communities [107]. For this project, we plan to hold annual all-personnel project meetings, in the presence of the advisory board. This meeting will typically be held in conjunction with a thematic workshop. Separate annual data jamborees (5-7 people) will bring experts together to work out particular issues. We expect other opportunities to meet in smaller groups at meetings organized by other projects.

D. Collaboration and coordination with other projects. Due to the commitment of the PIs to the development of community resources and standards, the proposed work interfaces with many other projects, which we can only briefly describe here. Critical morphological expertise for the vertebrates will come from the newly funded Phenotype RCN and will be solicited from relevant NSF-Assembling the Tree of Life (**ATOL**) projects. Through the Phenoscape internship program, in summer 2009 we trained Sandrine Tercerie with the fish systematics and informatics group at the **Muséum National d'Histoire Naturelle, Paris** [108]. Tercerie has continued curation under the auspices of this group and has contributed nearly 20% of the total phenotypes to the PhenoscapeKB to date. This group is committed to continuing its work with Phenoscape as funds permit (see letter, Lecointre). **DeepFin** is a NSF-funded Research Coordination Network for Ichthyology that has worked closely with Phenoscape during the past two years in co-sponsoring exchange students, disseminating information about ontologies, holding workshops and helping to bring in community expertise to build the anatomy ontology for fishes (TAO). The **National Center for Biomedical Ontologies (NCBO)** is an NIH-funded virtual center that provides services such as the Open Biomedical Ontologies (OBO) Foundry, the BioPortal ontology browser, and WebProtege (see letter, M. Musen). San Diego Supercomputer Center's **Data Central** provides long-term hosting of important scientific data collections (see Sustainability Plan below). **FishBase** is a global information system and relational database on fishes that has provided the common names of fishes for the Teleost Taxonomy Ontology. **Encyclopedia of Life** is creating an online reference source and database for all 1.8 million extant named species. We have obtained PDFs of literature to be curated from their **Biodiversity Heritage Library**. Pending funding for **VertNet** [109], we will exchange relevant taxonomic and associated organismal data. S. Federhen of the **National Center for Biomedical Information** has agreed to work with us to align higher level taxonomy based on expert consensus. We will continue to work with **TDWG** to promote standards for taxonomic ontologies. The anatomy ontologies and ontology curation tools produced by this project will be of immediate benefit **Morphbank** and **Morphster**, who are themselves collaborating to develop methods for image annotation. The NSF-funded **Scientific Observations Network (SONet)** and a Joint **DataONE / DataConservancy** Working Group on Observational Data Models and Semantics have identified the EQ model as one of only three candidates for a unifying semantic standard for scientific observations.

E. Dissemination plan. We will disseminate the results of our work early and often. Phenoscape has and will continue to make all software source code, ontologies, and curated data publicly available prior to publication, under OSI-compliant and Creative Commons attribution licenses [110], respectively, and the KB has an explicit data use policy on its website [111]. Progress updates and other documentation will continue to be posted regularly to the web and mailing lists. To enable data reuse by semantic web agents, we will regularly publish data dumps in OWL format. All ontologies developed or extended through this grant will be deposited in the OBO Foundry and will be available through the NCBO BioPortal [70] for searching, browsing, visualization, and download.

BROADER IMPACTS

This proposed work will provide a variety of high-value resources (software tools, ontologies, data) to researchers and students. The KB will enable open-ended investigations of an extremely valuable store of phenotype data. The project will serve as a model for how to break down traditional data silos and allow phenotypic knowledge to be integrated online with the scientific literature, image databases, biodiversity and environmental data, etc.

The uniquely interdisciplinary nature of the project will serve as an excellent training environment in which to engage biology and bioinformatics students in research. Several of the personnel (Vision, Blake) teach bioinformatics courses, and will incorporate the KB into their classrooms. A summer internship will be competitively awarded annually to at least one graduate student or postdoc interested in gaining expertise with biological ontologies; training will be based in the laboratory of the appropriate senior personnel. The DeepFin RCN has committed to supporting one or more additional summer interns to our program in yr1 and beyond pending grant extension (see letter, G. Orti). Training will be also be provided to at least two graduate students (UNC, USD), two postdoctoral level personnel (USD, U Chicago) and multiple undergraduates, who will participate in data collection and annotation. Additional students will be involved through our outreach programs (see below). The institutions represented in this proposal have successful mechanisms for recruiting individuals from underrepresented groups. We have requested dedicated funds to work with two of these. One, in partnership with Lakota people of the Rosebud Indian Reservation and USGS-EROS, the University of South Dakota has NSF funding (CNS-0739320) to engage Native Americans in computing through the Getting American Indians to Information Technology (GAIN-IT) project. We will support an undergraduate student in association with this effort, and Mabee will present as part of Career Discussions in the community-based summer camps. A "Junior Biocurator" program for advanced Chicago public high school students will be implemented by Project Exploration [112], under the direction of PI Sereno. We will sponsor other undergraduate students to both augment the KB and use it as a research tool.

We will continue to disseminate information to students and colleagues through presentations at national and international meetings in our disciplines and seek opportunities for hands-on demonstrations while there. The PIs collectively attend a very wide variety of meetings, and so will have the opportunity to inform many different audiences. In addition, Mabee is a PI on the recently funded RCN proposal (DBI-RCN:0956049 Phenotype Ontology Research Coordination Network"). In concert with this group, the personnel on this grant will participate in an annual summer course, and in other training and outreach activities, including development of a handbook of phenotype ontology standards and best practices. Through this RCN, we will support *ad hoc* cross-training exchanges with personnel on other projects. We plan to hold one informatics workshop and one data jamboree annually, providing opportunities for the community to contribute to the design and content of the resources being developed.

SUSTAINABILITY PLAN

It is our intent that the KB, and particularly its unique datastore, be a persistent and stable resource. We view it as a critical node in the emerging semantic web for life sciences, nearly unique in its position at the juncture of genetic, phenotypic, and biodiversity data. Our outreach efforts to demonstrate the value of this approach to the research community, which will be enhanced through the RCN, are an important aspect of our sustainability plan since it is ultimately the community of users who must be willing to support the resource. Provided that we are successful in demonstrating that the approach can be scaled up, we anticipate that funding of further data curation and ontology development efforts can be secured from a coalition of museums, scientific societies, and private foundations interested in documenting phenotypic biodiversity in a way that allows full utilization of the investment. We will take the first steps towards this by approaching the organismal societies for vertebrates (e.g., American Society of Ichthyology and Herpetology) and those involved in synthetic biology (e.g., Society for the Study of Evolution). Additionally, we are exploring publisher-based models of sustainability and outreach with Nature Publishing Group, modeled on the Signaling Gateway [113]. Finally, our associations with major NSF cyberinfrastructure initiatives (Vision is a co-PI on DataONE [114]; Mabee is a member of the Data Conservancy [115] Life Sciences working group) provide connections to other sustainability models. The proposed work will make phenotype knowledge acquisition more efficient, and thus more easily sustained by future projects. However, even if no further funded work is possible at the conclusion of this project, our accomplishments will remain available for reuse and repurposing. All software will be OSI-licensed, and the source code available through persistent repositories (e.g., SourceForge [116]). The ontologies are maintained in the OBO ontology repository (also at SourceForge) for the indefinite future. NESCent commits to host the KB at least through 2014. Should NESCent cease to have the resources after that point, we would apply to Data Central for hosting (see letter, A. Ferbert). Finally, the KB datastore will be made available in OWL, for reuse by general-purpose visualization, reasoning, and storage tools.

RESULTS FROM PRIOR NSF SUPPORT

Paula Mabee, Todd Vision, and Monte Westerfield are PIs on NSF DBI-0641025 'Linking evolution to genomics using phenotype ontologies' 6/07 – 5/11. $1,769,501. We have developed a system, 'Phenoscape', that facilitates synthesis across evolutionary phenotypic data and genetic data by developing new and extending existing ontologies that represent expert knowledge (Teleost Anatomy Ontology [18]; Teleost Taxonomy Ontology; significant contributions to PATO, Spatial Ontology, and Evidence Code Ontology); developing Phenex software for data curation [16]; and developing the Phenoscape Knowledgebase [5]. In its last year, we have focused on enhancements to the KB user interface, including formal usability testing. We have hosted 3 outreach symposia at national meetings, 5 workshops, sponsored 3 internships, trained undergraduates, graduate students, and postdoctoral researchers, and involved over 70 scientists in workshops, many of whom have contributed to ontologies and data curation. Three papers citing support from this grant have been published so far [16-18], and two others are in preparation. **Paul Sereno** is the PI on NSF DEB-0417163 ‘New Cretaceous theropods and microvertebrates from Africa (2005-06) $155,567 and NSF DEB-0523008 suppl. REU for undergraduate participation, $5,545. Principal objectives were to prepare and describe new theropod dinosaurs and sort sediment for microvertebrates; 8 papers were published [75,117-123]. **David Blackburn** is the PI on newly funded NSF DEB-1019444 ‘Collaborative Research: Biotic Surveys of Central Saharan Oases’ (08/10–07/13; $709,161). Goals are to survey and understand fauna of isolated desert oases in Libya and Egypt, to map in detail the boundary between Mediterranean and sub-Saharan faunas, and to test long-standing biogeographic hypotheses using genetic data.

REFERENCES

1. http://tolweb.org.
2. Kailola, P.J. (2004). A phylogenetic exploration of the catfish family Ariidae (Otophysi; Siluriformes). The Beagle, Records of the Museums and Art Galleries of the Northern Territory *20*, 87-166
3. Galperin, M.Y., and Cochrane, G.R. (2009). Nucleic Acids Research annual Database issue and the NAR online Molecular Biology Database Collection in 2009. Nucleic Acids Research *37*, D1-D4. http://dx.doi.org/10.1093/nar/gkn942
4. Bowne, P.S. (1994). Systematics and morphology of the Gasterosteiformes. In The Evolutionary Biology of the Threespine Stickleback, M.A. Bell and S.A. Foster, eds. (New York: Oxford), pp. 28-60.
5. http://kb.phenoscape.org.
6. Goble, C., and Stevens, R. (2008). State of the nation in data integration bioinformatics. Journal of Biomedical Informatics *41*, 687-693. http://dx.doi.org/10.1016/j.jbi.2008.01.008
7. Mabee, P., Ashburner, M., Cronk, Q., Gkoutos, G., Haendel, M., Segerdell, E., Mungall, C., and Westerfield, M. (2007). Phenotype ontologies: the bridge between genomics and evolution. Trends in Ecology & Evolution *22*, 345-350. http://dx.doi.org/10.1016/j.tree.2007.03.013
8. Mungall, C., Gkoutos, G., Washington, N., and Lewis, S. (2007). Representing phenotypes in OWL. Proceedings of the OWLED 2007 Workshop on OWL: Experience and Directions: June 6-7, 2007; Innsbruck, Austria 2007 [http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-258/paper29.pdf].
9. Mungall, C.J., Gkoutos, G.V., Smith, C.L., Haendel, M.A., Lewis, S.E., and Ashburner, M. (2010). Integrating phenotype ontologies across multiple species. Genome Biology *11*, R2. http://dx.doi.org/10.1186/gb-2010-11-1-r2
10. Harris, M.P., Rohner, N., Schwarz, H., Perathoner, S., Konstantinidis, P., and Nüsslein-Volhard, C. (2008). Zebrafish eda and edar mutants reveal conserved and ancestral roles of ectodysplasin signaling in vertebrates. PLoS Genetics *4*, e1000206. http://dx.doi.org/10.1371/journal.pgen.1000206
11. Colosimo, P.F., Hosemann, K.E., Balabhadra, S., Villarreal, G.J., Dickson, M., Grimwood, J., Schmutz, J., Myers, R.M., Schluter, D., and Kingsley, D.M. (2005). Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. Science *307*, 1928-1933. http://dx.doi.org/10.1126/science.1107239
12. Washington, N.L., Haendel, M.A., Mungall, C.J., Ashburner, M., Westerfield, M., and Lewis, S.E. (2009). Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biology *7*, 1-20. http://dx.doi.org/10.1371/journal.pbio.1000247
13. Carroll, S.B., Grenier, J.K., and Weatherbee, S.D. (2005). From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design, (Oxford: Blackwell Publishing).
14. Shubin, N., Tabin, C.J., and Carroll, S.B. (1997). Fossils, genes and the evolution of animal limbs. Nature *388*, 638-648
15. Mabee, P.M., Arratia, G., Coburn, M., Haendel, M., Hilton, E.J., Lundberg, J.G., Mayden, R.L., Rios, N., and Westerfield, M. (2007). Connecting evolutionary morphology to genomics using ontologies: A case study from Cypriniformes including zebrafish. Journal of Experimental Zoology Part B-Molecular and Developmental Evolution *308B*, 655-668. http://dx.doi.org/10.1002/jez.b.21181
16. Balhoff, J.P., Dahdul, W.M., Kothari, C.R., Lapp, H., Lundberg, J.G., Mabee, P., Midford, P.E., Westerfield, M., and Vision, T.J. (2010). Phenex: ontological annotation of phenotypic diversity. PLoS ONE *5*, e10500. http://dx.doi.org/10.1371/journal.pone.0010500
17. Dahdul, W.M., Balhoff, J.P., Engeman, J., Grande, T., Hilton, E.J., Kothari, C., Lapp, H., Lundberg, J.G., Midford, P.E., Vision, T.J., Westerfield, M., and Mabee, P.M. (2010). Evolutionary characters, phenotypes and ontologies: curating data from the systematic biology literature. PLoS ONE *5*, e10708. http://dx.doi.org/10.1371/journal.pone.0010708
18. Dahdul, W.M., Lundberg, J.G., Midford, P.E., Balhoff, J.P., Lapp, H., Vision, T.J., Haendel, M.A., Westerfield, M., and Mabee, P.M. (2010). The Teleost Anatomy Ontology: anatomical

representation for the genomics age. Systematic Biology *59*, 369-383. http://dx.doi.org/10.1093/sysbio/syq013

19. Clack, J.A. (2009). The Fin to Limb Transition: New Data, Interpretations, and Hypotheses from Paleontology and Developmental Biology. Annual Review of Earth and Planetary Sciences *37*, 163-179. http://dx.doi.org/10.1146/annurev.earth.36.031207.124146
20. Hall, B.K. ed. (2007). Fins into Limbs: Evolution, Development, and Transformation (Chicago: University of Chicago Press).
21. Shubin, N., Tabin, C., and Carroll, S. (2009). Deep homology and the origins of evolutionary novelty. Nature *457*, 818-823. http://dx.doi.org/10.1038/nature07891
22. Towers, M., and Tickle, C. (2009). Growing models of vertebrate limb development. Development *136*, 179-190. http://dx.doi.org/10.1242/dev.024158
23. Wilkins, A.S. (2002). The evolution of developmental pathways, (Sunderland, MA: Sinauer Associates).
24. Zeller, R., López-Ríos, J., and Zuniga, A. (2009). Vertebrate limb bud development: moving towards integrative analysis of organogenesis. Nat Rev Genet *10*, 845-858. http://dx.doi.org/10.1038/nrg2681
25. Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nature Genetics *25*, 25-29. http://dx.doi.org/10.1038/75556
26. Gkoutos, G.V., Green, E.C., Mallon, A.M., Hancock, J.M., and Davidson, D. (2004). Using ontologies to describe mouse phenotypes. Genome Biology *6*, R8. http://dx.doi.org/10.1186/gb-2004-6-1-r8
27. Sprague, J., Bayraktaroglu, L., Bradford, Y., Conlin, T., Dunn, N., Fashena, D., Frazer, K., Haendel, M., Howe, D., Knight, J., Mani, P., Moxon, S., Pich, C., Ramachandran, S., Schaper, K., Segerdell, E., Shao, X., Singer, A., Song, P., Sprunger, B., Van Slyke, C., and Westerfield, M. (2008). The Zebrafish Information Network: the zebrafish model organism database provides expanded support for genotypes and phenotypes. Nucleic Acids Research *36*, D768-D772. http://dx.doi.org/10.1093/nar/gkm956
28. Jensen, L.J., and Bork, P. (2010). Ontologies in quantitative biology: a basis for comparison, integration, and discovery. PLoS Biol *8*, e1000374. http://dx.doi.org/10.1371/journal.pbio.1000374
29. Zhang, J., Wag, P., Guay, D., Sanchez-Pulido, L., Padhi, B.K., Korzh, V., Andrade-Navarro, M.A., and Akimenko, M.-A. (2010). Loss of fish actinotrichia proteins and the fin-to-limb transition. Nature *466*, 234-237. http://dx.doi.org/10.1038/nature09137
30. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of Molecular Biology *215*, 403-410. http://dx.doi.org/10.1006/jmbi.1990.9999
31. Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G.J. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, (Cambridge UK: Cambridge University Press).
32. Pearson, W.R., and Lipman, D.J. (1988). Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA *85*, 2444–2448
33. Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. J. Mol. Biol. *147*, 195–197
34. Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., and Bateman, A. (2010). The Pfam protein families database. Nucleic Acids Research *38*, D211-D222. http://dx.doi.org/10.1093/nar/gkp985
35. Galperin, G.R., and Cochrane, M.Y. (2010). The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. Nucleic Acids Research *38*, D1-D4. http://dx.doi.org/doi:10.1093/nar/gkp1077
36. Pesquita, C., Faria, D., Bastos, H., Ferreira, A.E., Falcão, A.O., and Couto, F.M. (2008). Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatics *9(Suppl 5)*, S4. http://dx.doi.org/10.1186/1471-2105-9-S5-S4
37. Fröhlich, H., Speer, N., Poustka, A., and Beißbarth, T. (2007). GOSim – an R-package for computation of information theoretic GO similarities between terms and gene products. BMC Bioinformatics *8*, 166. http://dx.doi.org/10.1186/1471-2105-8-166
38. Lord, P.W., Stevens, R.D., Brass, A., and Goble, C.A. (2007). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics *19*, 1275-1283. http://dx.doi.org/10.1093/bioinformatics/btg153

39. Schlicker, A., and Albrecht, M. (2008). FunSimMat: a comprehensive functional similarity database. Nucleic Acids Research *36*, D434-D439. http://dx.doi.org/10.1093/nar/gkm806
40. Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics *26*, 976-978. http://dx.doi.org/10.1093/bioinformatics/btq064
41. Köhler S, S.M., Krawitz P, Bauer S, Dölken S, Ott CE, Mundlos C, Horn D, Mundlos S, Robinson PN (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet *85*, 457-464. http://dx.doi.org/10.1016/j.ajhg.2009.09.003
42. Ovaska, K., Laakso, M., and Hautaniemi, S. (2008). Fast Gene Ontology based clustering for microarray experiments. BioData Mining *1*, 11. http://dx.doi.org/10.1186/1756-0381-1-11
43. Aitken, S., Korf, R., Webber, B., and Bard, J. (2005). COBrA: a bio-ontology editor. Bioinformatics *21*, 825-826. http://dx.doi.org/10.1093/bioinformatics/bti097
44. Bastian, F., Parmentier, G., Roux, J., Moretti, S., Laudet, V., and Robinson-Rechavi, M. (2008). Bgee: integrating and comparing heterogeneous transcriptome data among species. Lecture Notes in Computer Science *5109*, 124-131. http://dx.doi.org/10.1007/978-3-540-69828-9_12
45. Parmentier, G., Bastian, F.B., and Robinson-Rechavi, M. (2010). Homolonto: generating homology relationships by pairwise alignment of ontologies and application to vertebrate anatomy. Bioinformatics *26*, 1766-1771. http://dx.doi.org/10.1093/bioinformatics/btq283
46. http://bioontology.org/wiki/index.php/RO:Main_Page#Proposed_homologous_to_relation.
47. Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U., and Brazma, A. (2006). ArrayExpress--a public database of microarray experiments and gene expression profiles. Nucleic Acids Research *35*, D747-750. http://dx.doi.org/10.1093/nar/gkl995
48. http://www.ebi.ac.uk/microarray-as/ae.
49. Roux, J., and Robinson-Rechavi, M. (2010). An ontology to clarify homology-related concepts. Trends in Genetics *26*, 99-102. http://dx.doi.org/10.1016/j.tig.2009.12.012
50. Roth, V.L. (1984). On homology. Biological Journal of the Linnean Society *22*, 13-29. http://dx.doi.org/10.1111/j.1095-8312.1984.tb00796.x
51. Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A., and Rosse, C. (2005). Relations in biomedical ontologies. Genome Biology *6*, R46-R46.15. http://genomebiology.com/2005/6/5/R46
52. http://phenoscape.org/wiki/OBD_Reasoner.
53. Dahn, R.D., Davis, M.C., Pappano, W.N., and Shubin, N.H. (2007). Sonic hedgehog function in chondrichthyan fins and the evolution of appendage patterning. Nature *445*, 311-314. http://dx.doi.org/10.1038/nature05436
54. Davis, M.C., Dahn, R.D., and Shubin, N.H. (2007). An autopodial-like pattern of Hox expression in the fins of a basal actinopterygian fish. Nature *447*, 473-476. http://dx.doi.org/10.1038/nature05838
55. Vargas, A.O., and Fallon, J.F. (2005). The digits of the wing of a bird are 1, 2, and 3. A review. Journal of Experimental Zoology (Molecular and Developmental Evolution) *304B*, 1-14. http://dx.doi.org/10.1002/jez.b.21051
56. Mabee, P.M. (2000). Developmental data and phylogenetic systematics: evolution of the vertebrate limb. American Zoologist *40*, 789-800. http://dx.doi.org/10.1093/icb/40.5.789
57. Mabee, P.M., Crotwell, P.L., Burke, A.C., and Bird, N.C. (2002). Evolution of median fin modules in the axial skeleton of fishes. Journal of Experimental Zoology: Molecular and Developmental Evolution *294*, 77-90. http://dx.doi.org/10.1002/jez.10076
58. Coates, M.I., Jeffery, J.E., and Ruta, M. (2002). Fins to limbs: what the fossils say. Evolution and Development *4*, 390-401
59. Maisey, J.G. (1986). Heads and tails: A chordate phylogeny. Cladistics *2*, 201-256
60. Janvier, P. (1986). Early Vertebrates, (Oxford: Claredon Press).
61. Johanson, Z., Sutija, M., and Joss, J.M.P. (2005). Regionalization of the axial skeleton in the lungfish *Neoceratodus forsteri* (Dipnoi) Journal of Experimental Zoology Part B, Molecular and Developmental Evolution *304*, 229-237. http://dx.doi.org/10.1002/jez.b.21048

62. Long, J.A., Young, G.C., Holland, T., Senden, T.J., and Fitzgerald, E.M.G. (2006). An exceptional Devonian fish from Australia sheds light on tetrapod origins. Nature *444*, 199-202. http://dx.doi.org/10.1038/nature05243

63. Ostrom, J.H. (1976). *Archaeopteryx* and the origin of birds. Biological Journal of the Linnean Society *8*. http://dx.doi.org/10.1111/j.1095-8312.1976.tb00244.x

64. Shubin, N.H., Daeschler, E.B., and Jenkins, F.A.J. (2006). The pectoral fin of *Tiktaalik roseae* and the origin of the tetrapod limb. Nature *440*, 764-761. http://dx.doi.org/10.1038/nature04637

65. Remane, A. (1952). Die Grundlagen des naturlichen Systems der vergleichenden Anatomie und der Phylogenetick, (Leipzig: Geest und Portig).

66. http://purl.bioontology.org/ontology/ECO.

67. Dowell, K.G., McAndrews-Hill, M.S., Hill, D.P., Drabkin, H.J., and Blake, J.A. (2009). Integrating text mining into the MGI biocuration workflow. Database *2009*, bap019. http://dx.doi.org/10.1093/database/bap019

68. https://wiki.birncommunity.org:8443/display/NEWBIRNCC/SciKnowMine.

69. Cui, H. (2010). Semantic annotation of morphological descriptions: an overall strategy. BMC Bioinformatics *11*, 278. http://dx.doi.org/10.1186/1471-2105-11-278

70. http://bioportal.bioontology.org.

71. http://obofoundry.org/wiki/index.php/PATO:Main_Page.

72. http://amphibiaweb.org/.

73. reptile-database.org.

74. Gill, F., and Donsker (Eds), D. (2010). IOC World Bird Names (version 2.5). Available at http://www.worldbirdnames.org/.

75. Sereno, P.C. (2005). The logical basis of phylogenetic taxonomy. Systematic Biology *54*, 595-619. http://dx.doi.org/10.1080/106351591007453

76. Wilson, D.E., and Reeder, D.M. (2005). Mammal Species of the World: A Taxonomic and Geographic Reference (3rd ed), (Johns Hopkins University Press).

77. http://paleodb.org/cgi-bin/bridge.pl.

78. http://phenoscape.org/wiki/Teleost_Taxonomy_Ontology.

79. Eschmeyer, W.N. (1998). Catalog of Fishes, (San Francisco: Special Publication, California Academy of Sciences).

80. Maglia, A.M., Leopold, J.L., Pugener, L.A., and Gauch, S. (2007). An anatomical ontology of amphibians Proc. of the Pacific Symposium on Biocomputing *12*, 367-378. http://dx.doi.org/10.1142/9789812772435_0035

81. https://www.phenoscape.org/wiki/Skeletal_Anatomy_Jamboree.

82. https://phenoscape.svn.sourceforge.net/svnroot/phenoscape/trunk/vocab/skeletal/obo/skeletalsummary.obo.

83. Abbott, A. (2010). Mouse project to find each gene's role. Nature *465*, Published online 25 May 2010. http://dx.doi.org/10.1038/465410a

84. Beck, T., Morgan, H., Blake, A., Wells, S., Hancock, J., and Mallon, A.-M. (2009). Practical application of ontologies to annotate and analyse large scale raw mouse phenotype data BMC Bioinformatics *10*, (Suppl 5):S2 http://dx.doi.org/10.1186/1471-2105-10-S5-S2

85. Robinson, P.N., and Mundlos, S. (2010). The Human Phenotype Ontology. Clinical Genetics *77*, 525 – 534. http://dx.doi.org/10.1111/j.1399-0004.2010.01436.x

86. Hey, T., Tansley, S., and Tolle, K. eds. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery (Microsoft Research).

87. Sprague, J., Bayraktaroglu, L., Clements, D., Conlin, T., Fashena, D., Frazer, K., Haendel, M., Howe, D., Mani, P., Ramachandran, S., Schaper, K., Segerdell, E., Song, P., Sprunger, B., Taylor, S., Van Slyke, C., and Westerfield, M. (2006). The Zebrafish Information Network: the zebrafish model organism database. Nucleic Acids Research *34*, D581-D585. http://dx.doi.org/10.1093/nar/gkj086

88. Sprague, J., Clements, D., Conlin, T., Edwards, P., Frazer, K., Schaper, K., Segerdell, E., Song, P., Sprunger, B., and Westerfield, M. (2003). The Zebrafish Information Network (ZFIN): the zebrafish model organism database. Nucleic Acids Research *31*, 241-243. http://dx.doi.org/10.1093/nar/gkg027

89. Sprague, J., Doerry, E., Douglas, S., and Westerfield, M. (2001). The Zebrafish Information Network (ZFIN): a resource for genetic, genomic and developmental research. Nucleic Acids Research *29*, 87-90. http://dx.doi.org/10.1093/nar/29.1.87

90. www.xenbase.org.

91. Bowes, J.B., Snyder, K.A., Segerdell, E., Jarabek, C.J., Azam, K., Zorn, A.M., and Vize, P.D. (2009). Xenbase: gene expression and improved integration. Nucleic Acids Research *38*, D607-D612. http://dx.doi.org/doi:10.1093/nar/gkp953

92. www.informatics.jax.org.

93. Blake, J., Bult, C., Eppig, J., Kadin, J., Richardson, J., and Genome, T. (2009). The Mouse Genome Database genotypes::phenotypes. Nucleic Acids Research *37*. http://dx.doi.org/10.1093/nar/gkn886

94. http://blog.phenoscape.org/2010/03/24/revising-the-knowledgebase-interface/.

95. http://blog.phenoscape.org/2010/05/04/phenoscape-outreach-from-chicago/.

96. https://www.nescent.org/phenoscape/Needs_Analysis_Workshop.

97. https://www.nescent.org/phenoscape/Data_Jamboree_1.

98. https://www.nescent.org/phenoscape/Data_Jamboree_2.

99. Sereno, P.C. (2007). Logical basis for morphological characters in phylogenetics. Cladistics *23*, 565-587. http://dx.doi.org/10.1111/j.1096-0031.2007.00161.x

100. http://www.w3.org/2001/sw/.

101. http://www.w3.org/2004/OWL/.

102. Berners-Lee, T., Lassila, O., and Hendler, J. (2001). The Semantic Web: A New Form of Web Content that is Meaningful to Computers will unleash a Revolution of New Possibilities. Scientific American

103. http://virtuoso.openlinksw.com.

104. Gaulton, K., Mohlke, K., and Vision, T. (2007). A computational system to select candidate genes for complex human traits. Bioinformatics *23*, 1132-1140. http://dx.doi.org/10.1093/bioinformatics/btm001

105. Owen, R. (1849). On the Nature of Limbs: A Discourse. In On the Nature of Limbs: A Discourse, R. Amundson, ed. (Chicago: University of Chicago Press).

106. https://www.phenoscape.org/wiki/Guide_to_Character_Annotation.

107. https://www.phenoscape.org/wiki/Phenotype_Ontology_Coordination_Workshop.

108. Dettai, A., Bailly, N., Vignes-Lebbe, R., and Lecointre, G. (2004). Metacanthomorpha: Essay on a Phylogeny-Oriented Database for Morphology--The Acanthomorph (Teleostei) Example. Syst Biol *53*, 822-834. http://dx.doi.org/10.1080/10635150490522313

109. Constable, H., Guralnick, R., Wieczorek, J., Spencer, C., Townsend Peterson, A., and Committee¶, T.V.S. (2010). VertNet: A new model for biodiversity data sharing. PLoS Biology *8*, e1000309. 10.1371/journal.pbio.1000309

110. http://creativecommons.org/licenses/by/3.0/.

111. http://phenoscape.org/wiki/Phenoscape_data_policy.

112. http://projectexploration.org.

113. http://www.signaling-gateway.org/.

114. https://dataone.org/.

115. http://dataconservancy.org/home.

116. http://sf.net.

117. Brusatte, S., and Sereno, P.C. (2005). A new species of *Carcharodontosaurus* (Dinosauria: Theropoda) from the Cenomanian of Niger and its implications for allosauroid phylogeny. Journal of Vertebrate Paleontology *Supplement 24:42A*

118. Brusatte, S.L., Benson, R.B.J., Carr, T.D., Williamson, T.E., and Sereno, P.C. (2007). The systematic utility of theropod enamel wrinkles. Journal of Vertebrate Paleontology *27*, 1052-1056. http://dx.doi.org/10.1671/0272-4634(2007)27[1052:TSUOTE]2.0.CO;2

119. Brusatte, S.L., and Sereno, P.C. (2007). A new species of *Carcharodontosaurus* (Dinosauria: Theropoda) from the Cenomanian of Niger and a revision of the genus. Journal of Vertebrate Paleontology *27*, 902-916. http://dx.doi.org/10.1671/0272-4634(2007)27[902:ANSOCD]2.0.CO;2

120. Brusatte, S.L., and Sereno, P.C. (2008). Phylogeny of the Allosauroidea (Dinosauria: Theropoda): Comparative analysis and resolution. Journal of Systematic Palaeontology *6*, 155-182. http://dx.doi.org/10.1017/S1477201907002404

121. Lipkin, C., Sereno, P.C., and Horner, J.R. (2007). The furcula in *Suchomimus tenerensis* and *Tyrannosaurus rex* (Dinosauria: Theropoda: Tetanurae). Journal of Paleontology *81*, 1532-1536. http://dx.doi.org/10.1666/06-024.1

122. Sereno, P.C., and Brusatte, S.L. (2008). Basal abelisaurid and carcharodontosaurid theropods from the Lower Cretaceous Elrhaz Formation of Niger. Acta Palaeontologica Polonica *53*, 15-46. http://dx.doi.org/10.4202/app.2008.0102

123. Sereno, P.C., Wilson, J.A., Alcober, O.A., Varricchio, D.J., Martinez, R.N., and Larsson, H.C.E. (2008). Evidence for avian intrathoracic air sacs in a new predatory dinosaur from Argentina. PLoS ONE *3*, 1-19. http://dx.doi.org/10.1371/journal.pone.0003303