Overview

The power of morphological data for evolutionary science is fundamentally rooted in its connection to what we know about morphology itself. Morphological data is not distinct in this regard—as is well known, molecular data, for example, yield the strongest signals when analysed using evolutionary models most consistent with what we know about DNA. However, in contrast to molecular data, the infrastructure of tools available for managing and analyzing morphological data does so without reference to the large amount of available domain knowledge. For these tools, the length of a femur is as unrelated to the size of a humerus as to the shape of the maxilla, even though our domain knowledge tells us that humerus and femur are serially homologous structures that are parts of the appendages in vertebrates. It also tells us that in contrast the maxilla and femur are parts of very different regions of the skeleton.

These and a host of other facts we know about morphological phenotypes have been codified in ontologies [1–4], formal knowledge representation systems that enable machines to understand the meaning (semantics) of concepts in a knowledge domain. Computers can use these to determine how similar in semantics morphological structures and types of phenotypic changes are [5, 6]. Morphological data can also be linked to the myriad of gene phenotypes recorded in various model organism databases [7, 8], to determine, for example, to what extent semantically very similar changes of two closely related structures are more likely to be controlled by the same genes or developmental pathways, compared to unrelated types of changes of unrelated structures. This approach can be just as powerful for studying the morphologies of extinct organisms known only from the fossil record as for those of the model organisms reared in the laboratory.

At present, computing with formal knowledge representation and discovery technologies is quite challenging. It requires considerable investment in computational infrastructure, software development, and a workforce with specialized skills [9, 10]. As a result, the gap between the state of the art in computational knowledge engineering and what comparative phylogenetics tools for morphological data are equipped to tap into continues to widen. We are dedicated to work toward closing this gap. Our approach takes inspiration from the engineering principles through which similarly challenging machine learning and artificial intelligence-based capabilities, such as voice control, image classification, and natural language processing, have found their way into everyday homes and small mobile devices. Specifically, we will create centralized services that do all the computational heavy-lifting, enabling client tools to act on machine-interpretable semantics of morphological data through lightweight application programming interfaces (APIs) and widely supported protocols.

To have a well-defined scope for developing the computable semantics backend and API services, we will focus on addressing the following three long-standing challenges, all of which are attributable to the lack of machine access to morphological domain knowledge. First, reusing, synthesizing, and objectively assessing data matrices at scale through computation, which is routine for molecular data, is very difficult for morphological data, and thus depends entirely on domain experts. Second, the evolutionary models in use for morphological data treat characters and character states as independent and distinct, respectively. Although a researcher can use their domain expertise to assign character-specific weights and state transition probabilities, tools to do so reproducibly through computable domain knowledge are lacking. Third, comparative phylogenetic analyses of discrete characters are limited in interpretation by the patterns that can be recognized by researchers and the domain expertise they bring to bear. By employing explicit semantics, it would be possible to make novel statements about morphological evolution on trees, such as quantifying the support for general classes of morphological changes being enriched or depauperate in particular clades.

Building such an infrastructure may seem too daunting a task. However, most of the foundational technologies, and many of the needed computational pieces, have come into place in recent years. These include formal ontologies that allow rich inferences about anatomical domain knowledge across species [3, 11, 12]; fast and scalable machine reasoners [13, 14]; ontological models for linking natural language text descriptions of phenotype observations to nodes in a fully computable knowledge graph [15, 16]; and using these linkages to calculate quantitative metrics for the semantic similarity between phenotype descriptions [5, 8]. Spearheaded by the Phenoscape project [17], these technologies have since been adapted for and applied to evolutionary phenotypes published as phylogenetic character state matrices [17–19]. With this approach, Phenoscape has demonstrated the feasibility of linking evolutionary character transitions to model organism genes by the similarity of their phenotypes [8], and it has shown how to use ontologies and machine inference to automatically synthesize a morphological matrix for presence/absence characters across studies while minimizing missing data [20]. At the infrastructure level, we will adapt this large body of technology development work and engineering know-how to transform the capabilities available to comparative phylogenetics tools.

Major objectives and deliverables

Our development plan is designed to demonstrate, for each of our three chosen focus challenges, how the gap between the existing ecosystem of tools and machine-accessible domain knowledge can be closed to solve long standing needs and to enable new analytic capabilities. Specifically, our work will result in the following deliverables, grouped by the major objectives they are designed to accomplish.

1. **Enable computational cross-study synthesis and calibration of character matrices**. We will develop an objective function that quantitatively assesses a character matrix on the basis of its semantic information properties. Using this function, we will develop algorithms to synthesize and consolidate characters and character states across studies from those originally published, calibrated to the semantic information properties chosen by the user.
2. **Enable incorporation of domain knowledge into models of trait evolution**. We will develop methods that allow tools for comparative analysis of discrete phenotypic traits to score evidence of the non-independence of traits, as derived from ontology-enabled links to computable domain knowledge and to model organism genetics. We will incorporate such evidence in Bayesian models of trait evolution by informing the prior probabilities for trait associations, which we will then apply to a study system with substantial homoplasy.
3. **Enable semantically-aware phylogenetic comparative analyses of morphology.** We will develop methods for performing ancestral state reconstruction using semantic phenotypes and methods for identifying enriched or depauperate classes of morphological concepts on a phylogenetic tree.

Each capability to be delivered as part of this work will be available in the form of online API endpoints adhering to the de-facto and thus widely supported standards for such APIs (HTTP/REST-style design, JSON or XML response formats). To drive usability and performance evaluation of the APIs, as well as to demonstrate their use, we will also develop reference applications which use the APIs in the statistical programming environment R, which features a particularly rich ecosystem of tools for comparative phylogenetics [21].

Biological Applications

The research and development for each deliverable will be validated and driven by two biological research applications. One is a recently published morphological supermatrix study [22] that will be used to validate and benchmark methods we develop. The second, trait

correlation in the repeated evolution of miniaturization in ostariophysan fishes, will allow us to apply, and thereby refine, the developed methods to an open research question.

Dillman et al. supermatrix. Supermatrices consolidate original character data from disparate studies into a larger matrix, in order to allow inference of more comprehensive phylogenies or to understand patterns of morphological diversification on a broader scale. Dillman et al. [22] recently evaluated the utility of the supermatrix approach for morphological data in a phylogenetic analysis of the Anostomoidea, a morphologically and ecologically diverse lineage of characiform fishes [23]. The supermatrix consists of 463 primarily skeletal characters synthesized from 14 studies and 6 species descriptions, pertaining to a total of 176 species in four families. Despite a large proportion (>60%) of missing data in the supermatrix, the resulting phylogenetic analysis showed strong support for relationships congruent with previous studies. It also revealed differences in the distribution of character state changes across regions of the anatomy, suggesting which regions diversified earlier or later in the evolution of the group. Thus, the study both corroborated previously obtained results, and allowed for new insight. This, and the fact that it was created in painstakingly manual work by some of the renowned experts for the group of interest, make it an excellent benchmark for comparison with the results of computational supermatrix synthesis methods. About one third of the characters in the 14 source studies are already annotated with ontology terms in the Phenoscape KB, requiring only modest effort to annotate those remaining.

Miniaturization in Ostariophysi. Miniaturization refers to the evolution of extremely small body size, which for fishes is somewhat arbitrarily defined as species with standard length <26 mm [24]. Across the 26,000+ species of fishes, miniaturization has occurred at least 40 times [24–27]. Although instances of loss or reduction of anatomical structures, such as reduced development of the lateral line canals of the head and body, reduced number of fin rays and scales, and absence of many cartilages and bones, have been documented for some miniatures, it is not known whether such changes characterize miniaturization more generally [28]. For example, in a comparison of the presence, absence, and number of 20 bones in 18 miniature species of cypriniforms, some features were found absent in all but one species, whereas others are lost in only a few [29]. In contrast, some miniatures are “proportioned dwarfs” with few reductive features [29]. Reconstructing a comprehensive phylogeny from morphological data covering some groups of miniaturized fishes is challenging due to the abundance of homoplastic phenotype changes [30]. Comparative analysis of the phenotypes of miniatures and their non-miniaturized relatives requires aggregating and consolidating characters from many disparate studies, a difficult task to conduct manually. The comparative analysis of miniaturization traits in fishes thus offers an excellent study system for applying the computable semantics-based capabilities developed in this project to an open research problem, which will be important for ongoing validation of the practical utility of these capabilities. We will computationally generate a synthetic supermatrix of morphological characters for miniatures and their non-miniature relatives; test the hypothesis that miniaturized taxa are generally characterized by loss or reduction of features in comparison to non-miniaturized relatives; and if so, identify which evolutionary losses or reductions are specifically correlated with decreasing body size. By focusing our study system on ostariophysan fishes (characins, minnows, catfishes, knifefishes), which include the majority of known instances of miniaturization (at least 25) and of known miniature fish species (over 60% of 273 species) [25, 29, 31], we take advantage of the ontology-annotated characters of their non-miniature relatives that are already in the Phenoscape KB, leaving only the studies for the miniature groups to be annotated. This will also provide a more generalizable comparison of workflow and required effort between hand-crafting a supermatrix for a research objective, and generating a computable semantics-driven synthetic supermatrix.

Significance

Everyone who uses a smart mobile device benefits firsthand from the transformative effect of putting powerful yet very demanding technologies in easy reach of vibrant developer communities. For example, performing a sentiment analysis on a piece of text, a task that once required deploying, mastering, and training sophisticated machine learning algorithms, today takes no more than a call to an online API offered by a variety of providers. The ground is well prepared for a similar transformation in studies of the Tree of Life. The body of morphological data accumulated in the literature over more than 300 years is vast, and it continues to grow. The importance of combining morphological with molecular data in evolutionary analyses has been powerfully demonstrated [32–34], and for the many extinct organismal lineages in Earth's history morphological data is all we can ever hope to collect. A rich landscape of tools, and a thriving community of scientists developing them, support a myriad of ways to assemble, manage, and analyze comparative trait data sets. Calls for the wider adoption of computable semantics have appeared repeatedly in the systematics and morphology literature [4, 35–37], but have remained largely aspirational. Yet, the components needed to go from aspiration to practical applications have matured tremendously in recent years, thanks to technological breakthroughs such as fast reasoners and highly scalable query engines for knowledge graphs, and to the painstaking work invested into building community ontologies representing relevant domain knowledge [1, 38, 39] and in annotating published data with these ontologies [18, 40–42]. Enabling new research applications for these technologies facilitates a virtuous cycle in which broader scientific use of computable semantics motivates wider contribution of semantic annotations by the scientific community.

Computable semantics enables not only new analytic capabilities for data recorded as free text. It also has the potential to make traditionally manual elements of data assembly and analysis protocols computationally reproducible, repeatable, and reusable. Although we focus here on morphological trait data, there are many other scientific fields, for example biomedicine, environmental science, or geology, in which observations are recorded in natural language text descriptions rather than quantitatively. The engineering approaches we will develop in this project for removing barriers to computing with domain knowledge should prove valuable in informing similarly-minded initiatives in other fields as well.

Results From Prior NSF Support

DBI-1062542, $1,851,057, 7/01/11-6/30/16 (no-cost extension granted through 6/30/17). "Collaborative Research: ABI Development: Ontology-enabled reasoning across phenotypes from evolution and model organisms". PIs W.M. Dahdul, T.J. Vision; subawards to H. Lapp, M. Westerfield, J. Blake, A. Zorn, D. Blackburn, P. Sereno, H. Cui and C. Mungall. We have developed a system, "Phenoscape", that facilitates synthesis across evolutionary phenotypic data and genetic data; developed umbrella anatomy and taxonomy ontologies, and software for data annotation. We expanded the taxonomic scope from fishes to vertebrates, tying ontologies and software tools together with phenotypes extracted from the vertebrate systematic literature into a knowledgebase that is integrated with genetic and phenotype data from three vertebrate model organisms: zebrafish (*Danio rerio*), frog (*Xenopus laevis*), mouse, and human. **Intellectual Merit:** To reduce time and cost of manual phenotype annotation, we have developed and worked with machine learning and improved annotation software to allow for on-demand augmentation of community ontologies. We have developed a semantic similarity engine to search the KB for taxa bearing a profile of phenotypes that are similar to a query profile from a gene. As a capstone in our final year of funding, we are assessing its performance in retrieving candidate genes for the well-studied vertebrate fin–limb transition. **Broader Impacts:** We have hosted a workshop facilitated by KnowInnovation at California Academy of Sciences, trained undergraduates, graduate students, and postdoctoral researchers, and involved over 70 scientists in workshops in previous funding, many of whom have continued to contribute to ontologies and data annotation. **Publications and Other Products:** 25 papers [1, 3, 7, 8, 10, 17–20, 38–53] citing support from this grant have been published to date, and 5 others are

in preparation. All software products are available from the Phenoscape GitHub repository [54], including the Phenoscape KB data integration and reasoning system, web service API, and web user interface; the Phenex annotation tool; and additional domain-independent semantic software tools.

Development Plan

The computable semantics methods to be delivered as part of this work will be available primarily in the form of server-side HTTP/REST-style API endpoints that return their results whenever possible in JSON or other data exchange standards (in particular NeXML [55] for synthetic and other data matrices). Both the style of API endpoints and the response format have become de-facto standards for online APIs, and are therefore very well supported in most programming languages, including those (R, Python, Java, C, C++) in which many comparative phylogenetics tools are written. For methods expected to be also used interactively, such as matrix synthesis and term enrichment analysis, we will also implement online user-interfaces that allow downloading the results for further use.

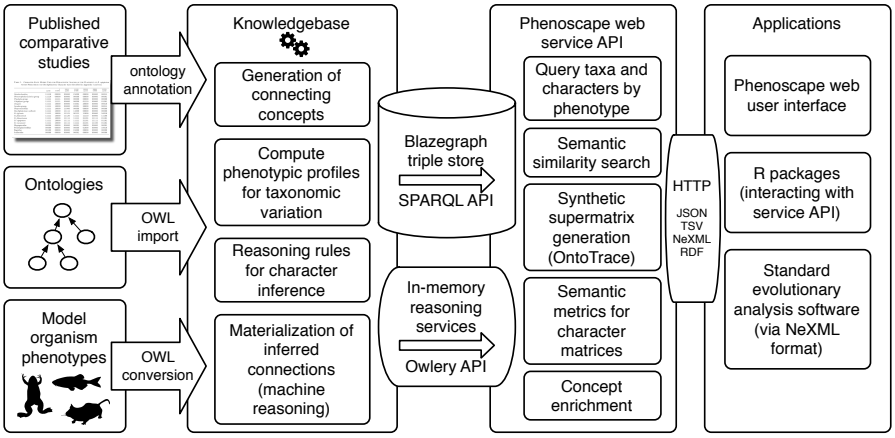

Figure 1. Architecture diagram of the computable semantics infrastructure from data ingest to API and user-interface access.

The API services, their backing algorithms, and the online user-interfaces will be built into the informatics infrastructure for computable semantics already developed by the Phenoscape project. This infrastructure currently consists of the following components, shown also in Fig. 1 alongside the flow of information from data ingest to client tools. (A) a Knowledgebase (KB), implemented as an RDF triplestore run by Blazegraph, for querying a knowledge graph of more than 300 million semantic connections between terms from 11 ontologies, tens of thousands of evolutionary characters and character states from more than 160 published studies, and thousands of model organism genes and their phenotypes [9, 17]; (B) a KB build pipeline which ingests requisite OWL ontologies, ontology-annotated (using Phenex [40, 41]) character matrix files, and gene phenotype annotations from model organism databases into a unified data model expressed in OWL (Web Ontology Language [56]), and then runs a series of automated pre-reasoning steps using industry standard machine reasoners (in particular ELK [14]); (C) an online query engine, named OntoTrace, for inferring presence/absence of anatomical structures from disparate fine-grained evolutionary character descriptions [19, 20]; (D) a semantic similarity engine that allows finding taxa that vary in suites of phenotypes similar to those associated with disruption of a particular query gene [8, 10]; and (E) a public web service API, and an online user-interface powered by it, that enables reasoner-driven queries of the KB by ontology term, taxon, gene, and phenotype, as well for performing OntoTrace and semantic similarity-based queries.

Our development plan is organized along the stages that a comparative trait analysis research study will typically go through, from data matrix construction (Aim I), to phylogenetic

reconstruction and/or analyzing trait evolution (Aim II), to generating and corroborating hypotheses about tempo and mode of trait evolution (Aim III). Details on each aim follow.

I. Enable computational cross-study synthesis and calibration of character matrices

As described above in Biological Applications, constructing supermatrices from smaller, often narrower-scoped and disparate studies can enable insight into broader evolutionary patterns, including phylogenetic relationships and patterns of convergent trait evolution. In contrast to molecular data, for which supermatrix assembly can be fully automated, their use for morphological phylogenetics remains rare, even though the studies that have been published underscore their value [22]. Assembling a morphological supermatrix is an entirely manual and labor-intensive process that requires domain experts, and even for experts it is often challenging to consolidate characters from different studies, in large part because doing so requires reconciling the homology and anatomical concepts implicitly used, but often not expressly recorded, by different authors working in different communities of practice [57]. Even where such reconciliation has been successful and its resulting characters published, it will typically have to be carried out again when creating a different supermatrix.

A number of us have recently published a method, named OntoTrace [20], that demonstrates how linking published morphological characters and states to computable domain knowledge in the form of ontologies enables the fully automatic and computational construction of synthetic supermatrices for presence/absence traits. The generated supermatrices consist of characters synthesized from the original character data and their ontology annotations by means of machine reasoning. The method differs from the traditional hand-crafting of supermatrices in several important ways: (1) Although domain experts still play crucial roles, such as in ontology development and data annotation, the results of their work, ontology-annotated characters, can be computationally reused. (2) Machine reasoning can infer a substantial amount of missing data from what, based on domain knowledge, is implied by but not expressly asserted in the original data. For example, in a synthetic supermatrix of 1051 sarcopterygian taxa and 639 characters [20], inferred data account for 93.2% of the populated cells, reducing the overall proportion of missing data from 98.5% to 78%. (3) The supermatrix synthesis is repeatable and fully reproducible, given the same ontologies, original data, and annotations. The chain of reasoning through which the synthetic character and its states were inferred can be fully traced back through entailments of the ontology to the original data annotation(s). As a corollary, matrices can immediately benefit when ontologies or data annotations are updated or corrected.

Ia. Generalizing OntoTrace. OntoTrace is already available as part of the Phenoscape KB infrastructure, both as an API service and through a web-based user-interface. However, the method can currently only synthesize presence/absence characters, in essence because this predefines the possible state values for every inferred synthetic character. Lifting this limitation is non-trivial but necessary for the method to become broadly useful for matrix synthesis. Initial experiments using a naïve, purely inference-based implementation yielded synthetic characters with tens or even hundreds of states. Furthermore, without the restriction the number of inferrable characters quickly becomes intractably large due to the rich reasoning supported by the requisite ontologies, especially when many studies with diverse types of characters are synthesized. Also, clusters of perfectly correlated (yet variable) characters [20] as a result of logical inference chains would be exacerbated for inferring characters for any phenotypic quality.

We will therefore research and develop methods for consolidating synthetic character states, and characters, according to the properties of individual characters, the collection of characters comprising a matrix, and the collection of states comprising a character, that would optimize the suitability of a synthetic supermatrix for research applications. Some properties will reflect

expectations of common downstream phylogenetic analysis tools, for example constraining the number of states per character. However, most criteria will take the role of bringing morphological domain knowledge to bear. Specifically, we will develop an objective function for scoring characters and character matrices, based on quantitatively assessing the semantic information content and similarity of the character states comprising a character, and the characters comprising a matrix, respectively. A host of algorithms exists already for determining semantic information content and similarity between nodes in a knowledge graph of domain ontologies and data annotated with them [58], and several of them have been successfully applied to and evaluated for evolutionary phenotypes by members of our team [8, 10, 50].

Ib. Objective function design. The function will be designed to score synthetic (inferred) characters, individually and as a collection forming a matrix, for a variety of possible objectives. We expect these objectives to include at least the following:

1. High and non-redundant semantic information content of a matrix. A matrix of characters covering a broader semantic diversity of anatomy and phenotypic quality, with less semantic redundancy between characters, can be expected to yield less biased phylogenetic inferences.
2. High information content of characters and character states. When unconstrained, machine reasoning will also infer trivial characters and character states that combine the most generic anatomical and phenotypic quality concepts (such as the ontologies' root terms). More specific characters have higher information content, but may yield more missing data.
3. Reduction of missing data. Machine reasoning can fill in a considerable fraction of data that would otherwise be missing, and reducing missing data improves power for comparative analyses.

Once in hand, the objective function(s) will allow us to develop the necessary methods for computationally and reproducibly consolidating character states and characters such that its score is maximized. For this, we anticipate to develop a semantic clustering algorithm; products of Phenoscape include semantic similarity (and thus distance) metrics adapted for character states and groups (i.e., clusters). Clusters of states can be consolidated to a synthetic and more general state, by finding, through machine reasoning, the least common subsuming phenotype (Fig. 2). (Clusters of characters are in essence clusters of the characters' states.) The objective function, or components of it, allows constraining the character state and character consolidation by maximizing the score of the resulting matrix. It can also be used to computationally and reproducibly filter out characters whose semantic properties make them outliers in a synthesized matrix, for example because their information content is too low.

Ic. Validation and benchmarking. A key issue for the development of these algorithms will be how to validate and benchmark them. Also, the objective function will have local maxima, and to achieve computational tractability we will need to develop heuristics instead of an exhaustive search. I.e., a synthesized matrix may not correspond to the global maximum of the objective function. To continuously validate the objective function and its components, and to benchmark the heuristics used for matrix synthesis, we will use the Dillman et al [22] supermatrix as a gold standard (see Biological Applications). Specifically, the higher a function or a set of parameters scores the gold standard, in comparison to random aggregations of characters from the underlying studies, the more suitable it should be. Similarly, we can use the tree inferred from a synthesized matrix to compare against the published tree.

Once validated using the Dillman et al. supermatrix, we will apply the matrix synthesis method to the designated open research question, the repeated evolution of miniaturization in ostariophysan fishes. The goal is to obtain a supermatrix of characters for miniatures and their

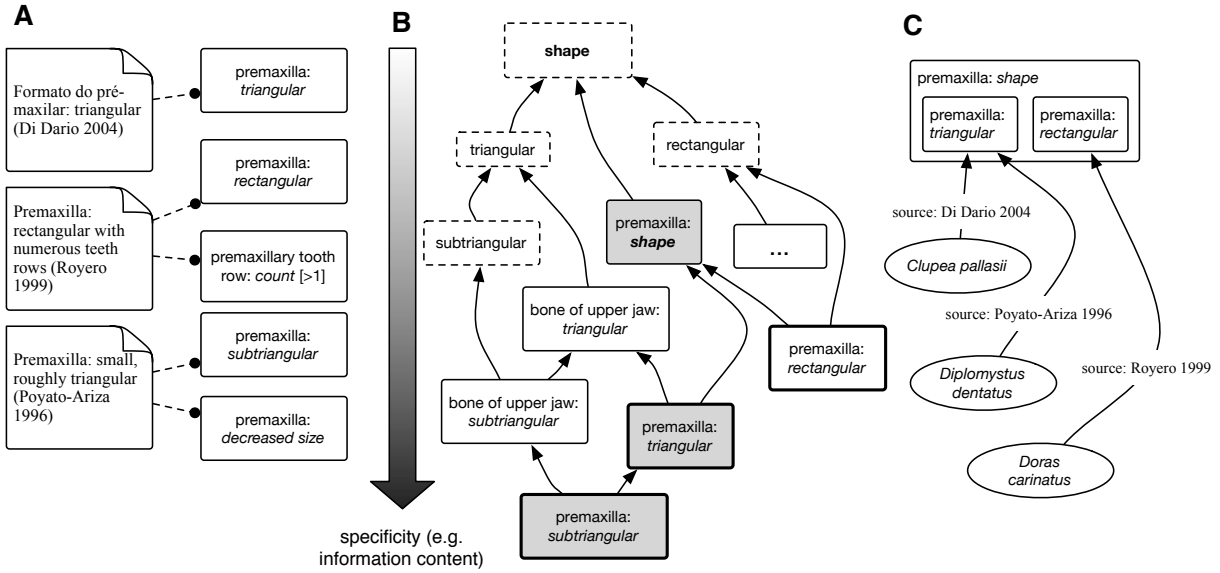

Figure 2. Inference of a synthetic character and states. (A) The KB contains links from published character states (left) to ontology-based annotations (right). (B) An automated reasoner can classify the annotations (bold solid outline) within an ontology containing existing concepts (dashed outline) along with generated candidate concepts (solid outline) for synthetic characters and states. Only selected concepts and annotations are shown. Other annotations linked to a state may contribute to entirely different synthetic states (e.g. 'premaxillary tooth row: count [>1]'). Based on various metrics, such as information content scores and whether a concept has been designated an "attribute" within the phenotypic quality ontology, unifying concepts will be selected as synthetic characters and states (gray fill). (C) The links between taxa and character states in the published matrices allow taxa to be linked to the synthetic states.

non-miniature relatives. This application provides both a more realistic scenario in terms of expected future uses, and a more challenging one. For example, there is no reference phylogeny for comparison, although we will alleviate this by using Open Tree [59] to synthesize a tree from recently published smaller phylogenies for ostariophysan clades that contain miniatures (e.g., Cyprinidae [60], which contains 33 miniatures, and Characidae, with nearly 70 miniatures [61]). Some of the morphological knowledge about miniatures has been published in monographs, and the sources of natural phenotype data are thus more mixed [52].

Id. Data annotation and ontology development. Both of the biological applications used in this and subsequent aims take much advantage of data already linked to (i.e., annotated with) ontology terms as a result of the Phenoscape project. However, both applications require additional data annotation, which typically will also require augmenting requisite ontologies [42]. Specifically, we will annotate the data from the source studies used by Dillman et al. that are not yet in the Phenoscape KB. These amount to about 400 characters (over two thirds of the total) in 13 studies (152 characters from one study [62] are already in the KB). Annotating these studies will use tools (Phenex) and workflows as described previously [18, 40, 41], and is expected to be accomplishable in 4-5 days of effort. For the miniaturization study, we will annotate phenotype descriptions for 174 species from approximately 30 matrix publications, containing ~515 characters, and 50 species descriptions or monographs with ~500 descriptions. In contrast to character descriptions, natural phenotypes in monographs are typically less structured descriptions within longer paragraphs of text. Annotating these will be aided significantly by a feature recently added to Phenex, which runs character and state descriptions through an online API providing ontology-backed named entity tagging. To further streamline the

annotation of monographs, we will also adapt the ways in which Phenex can load phenotype descriptions. Anatomical entities and terms describing phenotypic qualities that are not yet represented in Uberon [1] and other pertinent community ontologies will be submitted for vetting and addition through the ontology term request broker already built into Phenex [41]. Overall, we expect that the data annotation for the miniaturization study can be accomplished in about 1 week of effort.

Ie. Interoperability with commonly used tools. The R platform for statistical computing [63] has emerged as highly popular among evolutionary scientists. To ensure that the capabilities we create interoperate well with the existing ecosystem of R packages for comparative trait analysis [21], we will extend the RPhenoscape package [64], which was developed by the Phenoscape project as a pilot for bridging the impedance mismatch between the ontology-driven data models used by the Phenoscape KB and its API, and the simple tabular formats in which most of the pertinent R packages expect data. RPhenoscape allows R users to exercise several prominent capabilities of the Phenoscape API, including presence/absence trait matrix synthesis through the OntoTrace service, without having to be aware of ontologies. We will add to this package similarly interoperable access to the matrix synthesis and semantic information content-scoring functions developed as part of this aim.

II. Enable incorporation of domain knowledge into models of trait evolution

Phylogenetic analysis of discrete phenotypic traits for tree reconstruction or patterns of diversification rely on essentially the same models used for molecular sequence data [65, 66]. As discrete characters, both types of data face common challenges; phylogenetic information in the data is often limited because of a small state space and the potential for homoplasy, which in turn makes it difficult (in comparison to continuous traits) to attain sufficient power to distinguish between models and to estimate parameters. Furthermore, phenotypic traits often have strong correlations and violate assumptions of independence that are commonly applied to homologous sites in a DNA sequence. In contrast to molecular data, discrete phenotypic traits rarely have enough data to effectively estimate or even detect such correlations. For molecular data many of these challenges have been overcome by incorporating knowledge of physical characteristics of DNA, protein residues and processes of mutation into models of evolutionary change, but analogous approaches for discrete phenotypic traits have remained lacking.

Here we aim to enable similar advances for discrete phenotypic trait data by allowing tools used for comparative analysis to incorporate computable morphological knowledge and genetic evidence from model organisms into their models. One of the most natural targets for this capability is to enable tools to inform prior probabilities for Bayesian models, which are used, for example, to assign traits to different categories which may differ in evolutionary rates. Our goal is to provide, through programmable APIs, quantitative metrics of relatedness between traits, based both on their semantic similarity as determined by proximity in a knowledge graph, and on the evidence for a shared genetic architecture, as determined by model organism genes with similar location of expression or similar phenotype when mutated. Tools can then use these metrics to inform a Bayesian prior probability with what in essence represents prior knowledge and data.

IIa. Correlated phenotypes in multi-trait datasets. A common goal and major challenge for comparative analyses is estimating correlations between discrete traits (see [67] for a review). Furthermore, scaling to large multivariate datasets is a largely unsolved problem. For example, the widely-used Pagel model [68] estimates correlations by combining pairs of binary traits into a single character with four states (e.g. for traits A and B, AB can be 00, 01, 10, or 11). If transition rates for trait A depend on the state of trait B (for example), then the two traits are considered correlated. However, 8 transition rates must be estimated for per trait combination--

resulting in exponentially increasing numbers of parameters with traits. Yet, discrete traits contain little information in phylogenetic comparative datasets for estimating covariances between evolutionary rates across traits, and thus often require such large numbers of species that assumptions of homogenous evolutionary process are likely to break down.

To address this, we will develop a novel method that integrates computable knowledge-based evidence as prior information for detecting associations between discrete traits evolving on a phylogeny. Specifically, we will use Bayesian mixture modeling to cluster sets of discrete traits into bins with common evolutionary dynamics. For example, we will consider heterogenous rate Markov models where shifts occur on the phylogeny. These shifts can occur during epochs across the phylogeny [69], or in individual clades [70]. We will then apply a Dirichlet Process Prior on the assignment of traits into categories. This prior allows us to estimate both the number of groups as well as the assignment of traits into these groups. The result describes the posterior probability for each trait for being found in each distinct evolutionary rate class. For example, if a set of traits suggests a shift in transition rates in a particular clade, then these will be binned together into a group that has a shared parameter for the timing of the shift. Similar models have been implemented previously for combining phylogenies from many different species to estimate common timing of divergence events in phylogeographic studies (msbayes [71]), for assigning branches on the phylogeny into molecular rate categories [72] or for partitioning molecular datasets into different substitution models [73, 74].

To demonstrate and validate how models of trait evolution can integrate computable semantics-derived metrics and metrics derived from model organism genetics-based evidence, we will implement in the Bayesian phylogenetic programming language RevBayes [75] a model that uses these metrics as informative priors for trait associations [76]. RevBayes allows users to specify what parameters are shared across discrete traits (e.g. location of shifts, rate multipliers, rate classes, complexity of the Markov model), enabling the testing of a wide variety of relationships between traits. In our approach we will essentially place a prior on the network of associations between traits based on their semantic and potential genetic relationships, giving connections between traits with prior evidence of association a different distribution than those connecting traits expected to be unrelated. Leveraging knowledge in this way not only strengthens comparative methods by helping to overcome the difficulty of limited information inherent in discrete datasets. Traits that show co-diversification on the phylogeny despite low prior probability of being associated may reveal common underlying genetic or selective factors. In this way, by using prior knowledge comparative analyses can also discover relationships between traits that otherwise may lack an apparent correlation. In essence, our approach changes the question from whether or not traits are correlated, to whether or not traits share a common evolutionary tempo and/or mode.

IIb. Methods for semantic and genetic associations of traits. To enable the computational semantics and model organism genetics-based capabilities for informing trait evolution models, we will develop the following methods as extensions of the online Phenoscape API.

1. *Semantic similarity matrix for a set of phenotypic traits*. This will use the ontology term links of the traits and machine reasoning to quantify semantic similarity between traits. A variety of generic metrics exist, and are implemented in an actively developed software library. Phenoscape researchers are evaluating the merits of different types of metrics for natural phenotypic traits, and the Phenoscape KB already includes scoring the semantic similarity between model organism gene phenotypes (which number only in the thousands) and evolutionary phenotypes. For expanding this to scoring similarity between natural phenotypes, we will determine and implement a metric that is accurate while still sufficiently scalable to compute in real time.

2. *Genetic overlap for a set of characters*. Model organism databases link their expression and gene phenotype data to ontologies (e.g. [6, 77–79]) that are either shared or logically interoperable with those used by character data annotations in the Phenoscape KB. This enables computationally linking natural phenotypic traits to model organism genes by similarity of phenotype [7] or location of expression. The KB currently imports and integrates genes and their function, phenotypes, and location from several model organism databases [6, 77–79], which is maintenance-prone. We will simplify this by transitioning to services provided by the Monarch Initiative, a recently established biomedical knowledgebase [80] that pre-consolidates data across model organisms into a single model. The KB currently also does not use gene network data derived from known pathways and experimental evidence, one of the key sources for assessing potential gene associations. We will research and implement a mechanism for including such data, and integrating them into the evidence scoring.

To make our approach as widely accessible to comparative biologists as possible, we will also create an API endpoint, and an online user-interface powered by it (built into the existing Phenoscape user-interface), that allows selecting a set of phenotypic traits from the KB and outputs blocks of RevBayes code implementing the model priors as described above.

IIc. Validation and biological application. Our Bayesian approach to detecting character correlations will be validated by extensive simulation studies to ascertain that the method has appropriate statistical behavior. We will then apply our approach to the research question of whether there are sets of correlated phenotypes in the repeated evolution of miniaturization in ostariophysan fishes. Although a well-resolved phylogeny encompassing all miniaturized ostariophysans, which is necessary for this analysis, is not currently available, we will use high level phylogenies available within the group [61, 81, 82] and for smaller clades [83–85] to synthesize a tree for all Ostariophysi using Open Tree. To avoid any potential circularity from analysing trait evolution for characters also used to infer the phylogeny, we will generate two synthesized trees: one synthesized from both molecular (e.g., [81, 82, 86]) and morphological (e.g., [61, 87]) data; and the second synthesized only from molecular data.

One of the challenges here is that each step from matrix synthesis, to scoring evidence of trait association, to setting priors, can use different parameters, semantic similarity metrics, and choices for source data. To evaluate the potential multitude of different resulting sets of trait associations, we will assess Bayesian model behavior and MCMC performance, such as time to convergence and how narrow posterior probabilities are for parameter estimates. Even if much of this application would necessarily be exploratory, it would allow insights into trait correlations linked to decreasing body size that have so far not been obtainable at this scale.

III. Enable concept enrichment analysis for morphological data

The use of explicit semantics for morphological characters provides opportunities for novel classes of comparative phylogenetic analyses. Here we focus on two such opportunities: semantically-aware ancestral character state reconstruction and measuring enrichment of concepts on a phylogenetic tree

IIIa. Inference of ancestral traits. We frequently wish to make inferences about phenotypes in unobserved ancestors based on direct observations of phenotypes in extant, or fortuitously preserved extinct, lineages. Reconstructed ancestral phenotypes are used to test hypotheses about the timing and evolutionary causes of phenotypic changes, and have even been used to resurrect and test the biochemical properties of extinct protein sequences [88]. Here we are interested in questions such as: On which branches did miniaturization phenotypes arise in the history of the Ostariophysi? Do some miniaturization-associated phenotypes repeatedly precede

others? Do the suite of characters affected in miniatures covary generally throughout the phylogeny or only in cases of dramatic size reduction?

The classical way to solve the Ancestral Character Reconstruction (ACR) problem is to assign discrete character states to the nodes of a phylogeny according to a Maximum Parsimony objective function [89]; unfortunately, this is not so reliable in practice [90]. Contemporary ACR methods employ explicit statistical models. Maximum Likelihood (ML) methods can quantify support for an evolutionary hypothesis while treating the set of state assignments on a given phylogeny as a nuisance variable that is not known with certainty [91]. With Bayesian approaches, one can go further by quantifying support for a hypothesis over the joint probability distribution of state assignments and phylogenies [92].

To date, the Phenoscape KB has used a simple approach to ACR based on the parsimony algorithm of Fitch [93]. First, the set of character states for a taxon is defined to include the character states annotated to all descendant taxa. Then, characters with sets of states that differ among the daughter lineages of a particular taxon are noted as being variable for that taxon [8]. While this succeeds in identifying those characters that vary in state among the immediate descendants of a taxon, it has some of the same disadvantages as classical ACR methods, such as a bias toward the recent relative to the true position of the change, a failure to account for uncertainty in the phylogeny, and a failure to provide an estimate of uncertainty even with a fixed tree.

Here, we propose to adapt statistical approaches to the reconstruction of ancestral semantic phenotypes. A starting point for research will be modifying the multilayer probabilistic model used by Bauer et al [94], which takes as input a suite of semantically defined clinical phenotypes and estimates the probabilities of different possible causative diseases. In our application, one layer could model the probability of observing semantic phenotypes when they are present, a second layer could represent the comprehensive set of true phenotypes (which is only imperfectly observed), and a third layer would map the true phenotypes to the hierarchy of the ontology. Markov Chain Monte Carlo would be used to compute the marginal probabilities for the parameters of interest in such models. This probabilistic framework can easily handle the statistical consequences of phylogenetic relationships among nodes (at least for a fixed phylogeny), the effects of missing data, and the possibility of using phenotype annotations that are more or less specific in a given instance. We anticipate that development of this approach will be most straightforward for a fixed phylogenetic tree.

We will evaluate the performance of our probabilistic method compared to a simple parsimony approach primarily using simulations of character evolution (where the true changes on the phylogeny are known). We will also compare results to ML and Bayesian methods with non-semantic discrete phenotypes. Finally, we will apply these methods to address the questions posed above regarding suites of characters involved in miniaturization.

IIIb. Enrichment of traits within a phylogeny. Many genomic studies in the past decade have employed Gene Set Enrichment Analysis (GSEA) [95], the goal of which is to determine if there are terms, or branches, within the Gene Ontology that are over-represented among the functional annotations to some set of genes. The gene set may be from an experiment that catalogued differential gene expression in response to some treatment, or it may be from the set of candidate SNPs uncovered by a genome-wide association study.

A very similar type of question can be asked in a phylogenetic context. But here, rather than gene lists, we would want to ask if there are terms within an anatomy ontology (like Uberon [1]) that are over-represented among the evolutionary changes observed in a given region of a phylogeny (e.g. along a branch, or within a clade). Such phenotypic trends within clades are routinely identified and discussed in the evolutionary literature. A famous example is the

adaptive radiation in feeding ecology among cichlid fish that is thought to be due to the evolution of functional decoupling between the upper and the lower pharyngeal jaws [96]. But because there is a limited toolkit for quantifying and testing such trends, there is a risk that some patterns would not stand up to statistical scrutiny and, conversely, that some subtle but important patterns may be being missed.

Thus, the second opportunity we wish to pursue in this aim is to adapt the ideas of GSEA to phylogenetic enrichment analyses. A wide variety of different GSEA methods have been developed; Tarca et al [97] evaluated 16 methods in a recent study. For this application, we wish to identify the most specific level within the ontology for which enrichment is significant. That implies we cannot use naïve methods that consider only some small set of mid-level ontology terms as independent classes in a contingency test, and must instead take ontology structure into account in our statistic. We are also interested in putting enrichment tests into a more general phylogenetic comparative framework, in order to be able to test questions such as whether miniaturization in the Ostariophysi repeatedly enriches for certain classes of phenotype. We will evaluate the accuracy of the methods to be developed using simulated character evolution on a tree (as above), and by comparison to published reports of phenotypic trends within the Ostariophysi.

We will provide access to ancestral character reconstruction and enrichment analysis within the Phenoscape API as well as R reference implementations for accessing these services.

Broader Impacts

The capabilities for exploiting computable domain knowledge that are enabled by this project in the form of easy to access online APIs will transform how tools, whether large or small, can derive insight from morphological big data. We will undertake 3 complementary activities designed to promote adoption, sustainability, and workforce training. **(1) Training our target audience.** We will train our primary audience of potential tool users and developers in the foundational concepts and technologies we bring to bear, and in how the capabilities we provide can be exploited for new research methods. Specifically, we will develop the curriculum for a short-course on requisite knowledge representation and computational knowledge inference technologies, and teach the course at the Evolution Meetings from 2018 to 2020. To professionally assess the effectiveness of the training so it can be successively improved, we will contract with Data Carpentry, a non-profit organization that develops curriculum and teaches workshops for data literacy for researchers. Data Carpentry has recently hired an assessment expert to better understand and improve the effectiveness of its own workshops. **(2) Engaging our audience hands-on in adopting and influencing our products.** To promote innovative applications adopting the capabilities created by the project, we will engage comparative method developers and users through two community-oriented hackathon events, held in Years 2 and 3. Each event will be 4 days long with 20-25 participants, and will include users whose research questions and/or datasets stand to particularly benefit from the computational semantics capabilities. Rich ontologies, and comparative data annotated with them, have started to emerge from the research community at large, for example for sponges, spiders, and for plants. To afford these groups some of the same computable semantics-based capabilities, one of the hackathon events will be designated as “bring your own data”, with the goal of enabling direct submission of an ontology-annotated character matrix to the Phenoscape KB as input for user-interface and API methods that compute semantic similarity and other metrics on the fly. **(3) Engaging undergraduate students in cyberinfrastructure for treating text as data.** Despite a high demand for a workforce trained in ontologies and associated data analytics algorithms, these subjects are not taught in undergraduate computer science or bioinformatics courses at most US universities. A programming internship program run by Phenoscape through PI Lapp for undergraduates at Duke University was highly successful in exposing computer science majors to resources and methods for computing with the semantics of natural language

text while solving real-world programming tasks. In this project we will continue this program, and expand it to RTI through PI Balhoff. Summer interns at RTI participate in a structured series of workshops including a Pitch Bootcamp and a final poster session, gaining exposure to research opportunities in a non-academic but non-profit institution. RTI's program also has a track record of recruiting students from several local historically minority-serving universities.

Management Plan

	Responsibility and Timeline					
Specific Aim	**Year 1**		**Year 2**		**Year 3**	
Ia. Generalizing OntoTrace	L, B	L, B	L, B	L, B		
Ib. Objective function design	L, B	L, B	L, B	L, B		
Ic. Validation and benchmarking		L, B, D	L, B, D	L, B, D	L, B, D	L, B
Id. Data annotation and ontology development	D, B	D	D			
Ie. Interoperability with commonly used tools			U, L	U, L	U, L	U, L
IIa. Correlated phenotypes in multi-trait datasets	U	U	U	U		
IIb. Methods for semantic and genetic associations of traits		B	B, L	B, L	B, L	
IIc. Validation and biological application				B, D	B, D, U	B, D, U
IIIa. Inference of ancestral traits				V, B, D	V, B, D, U	
IIIb. Enrichment of traits within a phylogeny					V, B, D, U	V, B, D, U
Hackathons				All PIs		All PIs
Short course on knowledge representation and inference		All PIs		All PIs		All PIs

B=Balhoff, D=Dahdul, L=Lapp, U=Uyeda, V=Vision

A. Responsibilities and timelines

The project team includes 5 PIs that combine a tremendous breadth of biological, computational, and software engineering expertise, and are enthusiastic about enabling a community to incorporate computing with domain knowledge into its rich arsenal of methods. Lead-PI Dahdul (University of South Dakota) is an ichthyologist with deep expertise in ostariophysan fish morphology and evolution, and a veteran of the Phenoscape project. She will be responsible for all data and ontology curation efforts, and will serve as the expert stakeholder overseeing both biological research applications. PI Lapp (Duke University) has extensive expertise in evolutionary bioinformatics, and in using machine reasoning and ontology-driven data analytics for biological applications. He is a co-developer of several interoperability-oriented R packages (RMesquite, RNeXML, RPhenoscape), and has co-organized hackathons promoting interoperability in the R package ecosystem for evolutionary informatics. Lapp will have primary responsibility for the design, development, and testing of all APIs and R packages, and he will also lead the organization of the two planned hackathon events for community engagement and adoption (see Broader Impacts). PI Uyeda (University of Idaho) has deep expertise in comparative phylogenetics, population genetics, and Bayesian models for

sequence and trait evolution. He is author and co-developer of R packages used frequently in comparative analyses, and is a collaborator with the Arbor Workflows project for comparative methods. Uyeda will be responsible for developing, implementing, and validating the computable knowledge-informed model for trait evolution in RevBayes, and for using it to address the biological research applications. PI Vision (University of North Carolina at Chapel Hill) is a computational evolutionary biologist who was one of the co-PIs of Phenoscape. He will be responsible for work on semantic ancestral character reconstruction and phylogenetic character enrichment. PI Balhoff (RTI) originally trained as an evolutionary biologist, and combines deep expertise in knowledge modeling and formal semantics with extensive experience in scientific software engineering. He will be responsible for all software development on the Phenoscape KB, including data ingest, build pipeline, automated reasoning approaches, APIs, and user-interface.

Each of the 3 major aims involves major research efforts for algorithm and method development, as well as for the biological applications. These will all be carried out by 3 graduate students, one each at Duke (Aim I), Idaho (Aim II), and UNC Chapel Hill (Aim III), with PIs Lapp, Uyeda, and Vision as their respective primary supervisors. In addition, the project's PIs will co-mentor the graduate students as a team, giving them access to a much broader training and expertise than they would have in a single lab. We also plan to employ a total of 12 undergraduate student in internship projects, 7 at Duke, 3 at UNC, and 2 at RTI, who will be supervised by PIs Lapp, Vision and Balhoff, respectively. The internships aim to provide training experiences in bioinformatics and scientific cyberinfrastructure development (see Broader Impacts), and to assist with designated software development tasks, in particular for R packages at Duke and UNC and Phenoscape KB development at RTI.

B. Project coordination plan

The key to effective collaboration among participants in different locations is frequent and regular communication. Most of the PIs in the team have productively and successfully collaborated remotely for over 10 years. We will continue to use the varied communication channels already in place among the PIs through their collaboration in Phenoscape. These include twice monthly conference calls, a project mailing list, real-time team messaging software (Slack), web-based project management tools (Trello, GitHub), and web-based file sharing. Members of the project team will also meet face-to-face several times each year through the annual all-hands project team meeting, training workshops given at the Evolution Meetings, and two community-involving hackathons (see Broader Impacts).

C. Advisory board

We will assemble an Advisory Board with diverse expertise to evaluate progress, help minimize risks, and to provide guidance for research as well as for outreach and promoting adoption of computable semantics to enable new research frontiers. The Board will meet with the project team face-to-face once every year. Two board members already committed include Brian Sidlauskas (Oregon State University), fish morphology and evolution, phylogenetic morphospace methods; and Scott Chamberlain (rOpenSci), interoperable API design and R programming best practices (see letters of commitment). We will recruit four additional members with complementary areas of expertise once funded.

D. Dissemination plan

As for the Phenoscape project, all software source code, data annotations, and ontology contributions developed as part of this project will be available on public version control repositories, in particular GitHub, from the start of development under OSI-compliant open-source and Creative Commons Attribution (CC-BY) licenses, respectively. In addition to traditional peer-reviewed journal publications, project results will be disseminated early on at relevant scientific meetings for evolutionary biology (in particular Evolution Meetings and iEvoBio) and computational biology (in particular Pacific Symposium for Biocomputing).

References Cited

1. Haendel, M. A., Balhoff, J. P., Bastian, F. B., Blackburn, D. C., Blake, J. A., Bradford, Y., Comte, A., Dahdul, W. M., Dececchi, T. A., Druzinsky, R. E., Hayamizu, T. F., Ibrahim, N., Lewis, S. E., Mabee, P. M., Niknejad, A., Robinson-Rechavi, M., Sereno, P. C., and Mungall, C. J. "**Unification of Multi-Species Vertebrate Anatomy Ontologies for Comparative Biology in Uberon**" *Journal of biomedical semantics* 5, (2014): 21. doi:10.1186/2041-1480-5-21, Available at http://dx.doi.org/10.1186/2041-1480-5-21

2. Schofield, P. N., Sundberg, J. P., Hoehndorf, R., and Gkoutos, G. V. "**New Approaches to the Representation and Analysis of Phenotype Knowledge in Human Diseases and Their Animal Models**" *Briefings in functional genomics* 10, no. 5 (2011): 258–265. doi:10.1093/bfgp/elr031, Available at http://dx.doi.org/10.1093/bfgp/elr031

3. Dahdul, W. M., Cui, H., Mabee, P. M., Mungall, C. J., Osumi-Sutherland, D., Walls, R. L., and Haendel, M. A. "**Nose to Tail, Roots to Shoots: Spatial Descriptors for Phenotypic Diversity in the Biological Spatial Ontology**" *Journal of biomedical semantics* 5, (2014): 34. doi:10.1186/2041-1480-5-34, Available at http://dx.doi.org/10.1186/2041-1480-5-34

4. Deans, A. R., Lewis, S. E., Huala, E., Anzaldo, S. S., Ashburner, M., Balhoff, J. P., Blackburn, D. C., Blake, J. A., Burleigh, J. G., Chanet, B., Cooper, L. D., Courtot, M., Csösz, S., Cui, H., Dahdul, W., Das, S., Dececchi, T. A., Dettai, A., Diogo, R., Druzinsky, R. E., Dumontier, M., Franz, N. M., Friedrich, F., Gkoutos, G. V., Haendel, M., Harmon, L. J., Hayamizu, T. F., He, Y., Hines, H. M., Ibrahim, N., Jackson, L. M., Jaiswal, P., James-Zorn, C., Köhler, S., Lecointre, G., Lapp, H., Lawrence, C. J., Le Novère, N., Lundberg, J. G., Macklin, J., Mast, A. R., Midford, P. E., Mikó, I., Mungall, C. J., Oellrich, A., Osumi-Sutherland, D., Parkinson, H., Ramírez, M. J., Richter, S., Robinson, P. N., Ruttenberg, A., Schulz, K. S., Segerdell, E., Seltmann, K. C., Sharkey, M. J., Smith, A. D., Smith, B., Specht, C. D., Squires, R. B., Thacker, R. W., Thessen, A., Fernandez-Triana, J., Vihinen, M., Vize, P. D., Vogt, L., Wall, C. E., Walls, R. L., Westerfeld, M., Wharton, R. A., Wirkner, C. S., Woolley, J. B., Yoder, M. J., Zorn, A. M., and Mabee, P. "**Finding Our Way through Phenotypes**" *PLoS biology* 13, no. 1 (2015): e1002033. doi:10.1371/journal.pbio.1002033, Available at http://dx.doi.org/10.1371/journal.pbio.1002033

5. Washington, N. L., Haendel, M. A., Mungall, C. J., Ashburner, M., Westerfield, M., and Lewis, S. E. "**Linking Human Diseases to Animal Models Using Ontology-Based Phenotype Annotation**" *PLoS biology* 7, no. 11 (2009): e1000247. doi:10.1371/journal.pbio.1000247, Available at http://dx.doi.org/10.1371/journal.pbio.1000247

6. Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., Black, G. C. M., Brown, D. L., Brudno, M., Campbell, J., FitzPatrick, D. R., Eppig, J. T., Jackson, A. P., Freson, K., Girdea, M., Helbig, I., Hurst, J. A., Jähn, J., Jackson, L. G., Kelly, A. M., Ledbetter, D. H., Mansour, S., Martin, C. L., Moss, C., Mumford, A., Ouwehand, W. H., Park, S.-M., Riggs, E. R., Scott, R. H., Sisodiya, S., Van Vooren, S., Wapner, R. J., Wilkie, A. O. M., Wright, C. F., Vulto-van Silfhout, A. T., Leeuw, N. de, Vries, B. B. A. de, Washington, N. L., Smith, C. L., Westerfield, M., Schofield, P., Ruef, B. J., Gkoutos, G. V., Haendel, M., Smedley, D., Lewis, S. E., and Robinson, P. N. "**The Human Phenotype Ontology Project: Linking Molecular Biology and Disease through Phenotype Data**" *Nucleic acids research* 42, no. Database issue (2014): D966–74. doi:10.1093/nar/gkt1026, Available at http://dx.doi.org/10.1093/nar/gkt1026

7. Edmunds, R. C., Su, B., Balhoff, J. P., Eames, B. F., Dahdul, W. M., Lapp, H., Lundberg, J.

G., Vision, T. J., Dunham, R. A., Mabee, P. M., and Westerfield, M. "**Phenoscape: Identifying Candidate Genes for Evolutionary Phenotypes**" *Molecular biology and evolution* 33, no. 1 (2016): 13–24. doi:10.1093/molbev/msv223, Available at http://dx.doi.org/10.1093/molbev/msv223

8. Manda, P., Balhoff, J. P., Lapp, H., Mabee, P., and Vision, T. J. "**Using the Phenoscape Knowledgebase to Relate Genetic Perturbations to Phenotypic Evolution**" *Genesis* 53, no. 8 (2015): 561–571. doi:10.1002/dvg.22878, Available at http://dx.doi.org/10.1002/dvg.22878

9. Balhoff, J. P. and Phenoscape project team. "**The Phenoscape Knowledgebase: Tools and APIs for Computing across Phenotypes from Evolutionary Diversity and Model Organisms**" *bioRxiv* (2016): 071951. doi:10.1101/071951, Available at http://biorxiv.org/cgi/content/short/071951

10. Balhoff, J. P., Manda, P., Lapp, H., Mabee, P., and Vision, T. J. "**Connecting Genes with Evolutionary Knowledge Using Semantic Similarity and Ancestral Profiles of Variation**" *Bio-ontologies SIG at ISMB 2015* (2015):

11. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E., and Haendel, M. a. "**Uberon, an Integrative Multi-Species Anatomy Ontology**" *Genome biology* 13, no. 1 (2012): R5. doi:10.1186/gb-2012-13-1-r5, Available at http://dx.doi.org/10.1186/gb-2012-13-1-r5

12. Gkoutos, G. V., Green, E. C. J., Mallon, A.-M., Hancock, J. M., and Davidson, D. "**Using Ontologies to Describe Mouse Phenotypes**" *Genome biology* 6, no. 1 (2005): R8. doi:10.1186/gb-2004-6-1-r8, Available at http://dx.doi.org/10.1186/gb-2004-6-1-r8

13. Shearer, R., Motik, B., and Horrocks, I. "**HermiT: A Highly-Efficient OWL Reasoner**" *Proceedings of the 5th International Workshop on OWL: Experiences and Directions (OWLED 2008)* (2008): Available at http://www.mendeley.com//research/hermit-highly-efficient-owl-reasoner-3//

14. Kazakov, Y., Krötzsch, M., and Simančík, F. "**The Incredible ELK**" *Journal of Automated Reasoning* 53, no. 1 (2013): 1–61. doi:10.1007/s10817-013-9296-3, Available at http://link.springer.com/article/10.1007/s10817-013-9296-3

15. Mungall, C., Gkoutos, G., Washington, N., and Lewis, S. "**Representing Phenotypes in OWL**" *OWL: Experiences and Directions (OWLED 2007), Innsbruck, Austria* (2007): Available at http://www.mendeley.com//research/representing-phenotypes-owl//

16. Mungall, C. J., Gkoutos, G. V., Smith, C. L., Haendel, M. A., Lewis, S. E., and Ashburner, M. "**Integrating Phenotype Ontologies across Multiple Species**" *Genome biology* 11, no. 1 (2010): R2. doi:10.1186/gb-2010-11-1-r2, Available at http://dx.doi.org/10.1186/gb-2010-11-1-r2

17. Mabee, P. M., Balhoff, J. P., Dahdul, W. M., Lapp, H., Midford, P. E., Vision, T. J., and Westerfield, M. "**500,000 Fish Phenotypes: The New Informatics Landscape for Evolutionary and Developmental Biology of the Vertebrate Skeleton**" *Journal of Applied Ichthyology* 28, no. 3 (2012): 300–305. doi:10.1111/j.1439-0426.2012.01985.x, Available at http://dx.doi.org/10.1111/j.1439-0426.2012.01985.x

18. 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. "**Evolutionary Characters, Phenotypes and Ontologies: Curating Data from the Systematic Biology**

Literature" *PloS one* 5, no. 5 (2010): e10708. doi:10.1371/journal.pone.0010708, Available at http://dx.doi.org/10.1371/journal.pone.0010708

19. Balhoff, J. P., Dececchi, T. A., Mabee, P. M., and Lapp, H. "**Presence-Absence Reasoning for Evolutionary Phenotypes**" *Proceedings of Phenotype Day of the Bio-ontologies SIG at ISMB 2014* (2014): 1–4. Available at http://phenoday2014.bio-lark.org/pdf/11.pdf

20. Dececchi, T. A., Balhoff, J. P., Lapp, H., and Mabee, P. M. "**Toward Synthesizing Our Knowledge of Morphology: Using Ontologies and Machine Reasoning to Extract Presence/Absence Evolutionary Phenotypes across Studies**" *Systematic biology* (2015): doi:10.1093/sysbio/syv031, Available at http://dx.doi.org/10.1093/sysbio/syv031

21. O'Meara, B. "**CRAN Task View: Phylogenetics, Especially Comparative Methods**" (2016): Available at http://CRAN.R-project.org/view=Phylogenetics

22. Dillman, C. B., Sidlauskas, B. L., and Vari, R. P. "**A Morphological Supermatrix-Based Phylogeny for the Neotropical Fish Superfamily Anostomoidea (Ostariophysi: Characiformes): Phylogeny, Missing Data and Homoplasy**" *Cladistics: the international journal of the Willi Hennig Society* 32, no. 3 (2016): 276–296. doi:10.1111/cla.12127, Available at http://dx.doi.org/10.1111/cla.12127

23. Sidlauskas, B. "**Continuous and Arrested Morphological Diversification in Sister Clades of Characiform Fishes: A Phylomorphospace Approach**" *Evolution; international journal of organic evolution* 62, no. 12 (2008): 3135–3156. doi:10.1111/j.1558-5646.2008.00519.x, Available at http://dx.doi.org/10.1111/j.1558-5646.2008.00519.x

24. Weitzman, S. H. and Vari, R. P. "**Miniaturization in South American Freshwater Fishes; an Overview and Discussion**" *Proceedings of the Biological Society of Washington* 101, no. 2 (1988): 444–465. Available at http://www.sidalc.net/cgi-bin/wxis.exe/?IsisScript=LIBRI.xis&method=post&formato=2&cantidad=1&expresion=mfn=015980

25. Toledo-Piza, M., Mattox, G. M. T., and Britz, R. "**Priocharax Nanus, a New Miniature Characid from the Rio Negro, Amazon Basin (Ostariophysi: Characiformes), with an Updated List of Miniature Neotropical Freshwater Fishes**" *Neotropical ichthyology: official journal of the Sociedade Brasileira de Ictiologia* 12, no. 2 229–246. doi:10.1590/1982-0224-20130171, Available at http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1679-62252014000200229&lng=en&nrm=iso&tlng=en

26. Conway, K. W. and Moritz, T. "**Barboides Britzi, a New Species of Miniature Cyprinid from Benin (Ostariophysi: Cyprinidae), with a Neotype Designation for B. Gracilis**" *Ichthyological exploration of freshwaters* 17, no. 1 (2006): 73. Available at http://www.pfeilbook.de/04biol/pdf/ief17_1_07.pdf

27. Kottelat, M. and Vidthayanon, C. "**Boraras Micros, a New Genus and Species of Minute Freshwater Fish from Thailand (Teleostei: Cyprinidae)**" *Ichthyological Explorations Freshwaters* 4, (1993): 161–176.

28. Hanken, J. and Wake, D. B. "**Miniaturization of Body Size: Organismal Consequences and Evolutionary Significance**" *Annual review of ecology and systematics* 24, no. 1 (1993): 501–519. doi:10.1146/annurev.es.24.110193.002441, Available at http://dx.doi.org/10.1146/annurev.es.24.110193.002441

29. Britz, R. and Conway, K. W. "**Osteology of Paedocypris, a Miniature and Highly Developmentally Truncated Fish (Teleostei: Ostariophysi: Cyprinidae)**" *Journal of morphology* 270, no. 4 (2009): 389–412. doi:10.1002/jmor.10698, Available at http://dx.doi.org/10.1002/jmor.10698

30. Britz, R., Conway, K. W., and Rüber, L. "**Miniatures, Morphology and Molecules: Paedocypris and Its Phylogenetic Position (Teleostei, Cypriniformes)**" *Zoological journal of the Linnean Society* 172, no. 3 (2014): 556–615. doi:10.1111/zoj.12184, Available at http://dx.doi.org/10.1111/zoj.12184

31. Kottelat, M., Britz, R., Hui, T. H., and Witte, K.-E. "**Paedocypris, a New Genus of Southeast Asian Cyprinid Fish with a Remarkable Sexual Dimorphism, Comprises the World's Smallest Vertebrate**" *Proceedings. Biological sciences / The Royal Society* 273, no. 1589 (2006): 895–899. doi:10.1098/rspb.2005.3419, Available at http://dx.doi.org/10.1098/rspb.2005.3419

32. O'Leary, M. A., Bloch, J. I., Flynn, J. J., Gaudin, T. J., Giallombardo, A., Giannini, N. P., Goldberg, S. L., Kraatz, B. P., Luo, Z.-X., Meng, J., Ni, X., Novacek, M. J., Perini, F. A., Randall, Z. S., Rougier, G. W., Sargis, E. J., Silcox, M. T., Simmons, N. B., Spaulding, M., Velazco, P. M., Weksler, M., Wible, J. R., and Cirranello, A. L. "**The Placental Mammal Ancestor and the Post-K-Pg Radiation of Placentals**" *Science* 339, no. 6120 (2013): 662–667. doi:10.1126/science.1229237, Available at http://dx.doi.org/10.1126/science.1229237

33. Pyron, R. A. "**Novel Approaches for Phylogenetic Inference from Morphological Data and Total-Evidence Dating in Squamate Reptiles (Lizards, Snakes, and Amphisbaenians)**" *Systematic biology* (2016): doi:10.1093/sysbio/syw068, Available at http://sysbio.oxfordjournals.org/content/early/2016/08/06/sysbio.syw068.abstract

34. Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D. L., and Rasnitsyn, A. P. "**A Total-Evidence Approach to Dating with Fossils, Applied to the Early Radiation of the Hymenoptera**" *Systematic biology* 0, no. 0 (2012): 1–27. doi:10.1093/sysbio/sys058, Available at http://dx.doi.org/10.1093/sysbio/sys058

35. Vogt, L. "**The Future Role of Bio-Ontologies for Developing a General Data Standard in Biology: Chance and Challenge for Zoo-Morphology**" *Zoomorphology* (2008): doi:10.1007/s00435-008-0081-5, Available at http://dx.doi.org/10.1007/s00435-008-0081-5

36. Vogt, L., Bartolomaeus, T., and Giribet, G. "**The Linguistic Problem of Morphology: Structure versus Homology and the Standardization of Morphological Data**" *Cladistics: the international journal of the Willi Hennig Society* 26, no. 3 (2009): 301–325. doi:10.1111/j.1096-0031.2009.00286.x, Available at http://dx.doi.org/10.1111/j.1096-0031.2009.00286.x

37. Göpel, T. and Richter, S. "**The Word Is Not Enough: On Morphemes, Characters and Ontological Concepts**" *Cladistics: the international journal of the Willi Hennig Society* (2016): doi:10.1111/cla.12145, Available at http://dx.doi.org/10.1111/cla.12145

38. 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. "**The Teleost Anatomy Ontology: Anatomical Representation for the Genomics Age**" *Systematic biology* 59, no. 4 (2010): 369–383. doi:10.1093/sysbio/syq013, Available at http://dx.doi.org/10.1093/sysbio/syq013

39. Dahdul, W. M., Balhoff, J. P., Blackburn, D. C., Diehl, A. D., Haendel, M. A., Hall, B. K.,

Lapp, H., Lundberg, J. G., Mungall, C. J., Ringwald, M., Segerdell, E., Van Slyke, C. E., Vickaryous, M. K., Westerfield, M., and Mabee, P. M. "**A Unified Anatomy Ontology of the Vertebrate Skeletal System**" *PloS one* 7, no. 12 (2012): e51070. doi:10.1371/journal.pone.0051070, Available at http://dx.doi.org/10.1371/journal.pone.0051070

40. 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. "**Phenex: Ontological Annotation of Phenotypic Diversity**" *PloS one* 5, no. 5 (2010): e10500. doi:10.1371/journal.pone.0010500, Available at http://dx.doi.org/10.1371/journal.pone.0010500

41. Balhoff, J. P., Dahdul, W. M., Dececchi, T. A., Lapp, H., Mabee, P. M., and Vision, T. J. "**Annotation of Phenotypic Diversity: Decoupling Data Curation and Ontology Curation Using Phenex**" *Journal of biomedical semantics* 5, no. 1 (2014): 45. doi:10.1186/2041-1480-5-45, Available at http://dx.doi.org/10.1186/2041-1480-5-45

42. Dahdul, W., Dececchi, T. A., Ibrahim, N., Lapp, H., and Mabee, P. "**Moving the Mountain: Analysis of the Effort Required to Transform Comparative Anatomy into Computable Anatomy**" *Database: the journal of biological databases and curation* 2015, (2015): bav040. doi:10.1093/database/bav040, Available at http://dx.doi.org/10.1093/database/bav040

43. Midford, P. E., Dececchi, T. A., Balhoff, J. P., Dahdul, W. M., Ibrahim, N., Lapp, H., Lundberg, J. G., Mabee, P. M., Sereno, P. C., Westerfield, M., Vision, T. J., and Blackburn, D. C. "**The Vertebrate Taxonomy Ontology: A Framework for Reasoning across Model Organism and Species Phenotypes**" *Journal of biomedical semantics* 4, no. 1 (2013): 34. doi:10.1186/2041-1480-4-34, Available at http://dx.doi.org/10.1186/2041-1480-4-34

44. Arighi, C. N., Carterette, B., Cohen, K. B., Krallinger, M., Wilbur, W. J., Fey, P., Dodson, R., Cooper, L., Van Slyke, C. E., Dahdul, W., Mabee, P., Li, D., Harris, B., Gillespie, M., Jimenez, S., Roberts, P., Matthews, L., Becker, K., Drabkin, H., Bello, S., Licata, L., Chatr-aryamontri, A., Schaeffer, M. L., Park, J., Haendel, M., Van Auken, K., Li, Y., Chan, J., Muller, H.-M., Cui, H., Balhoff, J. P., Chi-Yang Wu, J., Lu, Z., Wei, C.-H., Tudor, C. O., Raja, K., Subramani, S., Natarajan, J., Cejuela, J. M., Dubey, P., and Wu, C. "**An Overview of the BioCreative 2012 Workshop Track III: Interactive Text Mining Task**" *Database: the journal of biological databases and curation* 2013, (2013): bas056. doi:10.1093/database/bas056, Available at http://dx.doi.org/10.1093/database/bas056

45. Deans, A. R., Yoder, M. J., and Balhoff, J. P. "**Time to Change How We Describe Biodiversity**" *Trends in ecology & evolution* 27, no. 2 (2012): 78–84. doi:10.1016/j.tree.2011.11.007, Available at http://dx.doi.org/10.1016/j.tree.2011.11.007

46. Cui, H., Dahdul, W., Dececchi, A. T., Ibrahim, N., Mabee, P., Balhoff, J. P., and Gopalakrishnan, H. "**CharaParser+EQ: Performance Evaluation without Gold Standard**" *Proceedings of the Association for Information Science and Technology* 52, no. 1 (2015): 1–10. doi:10.1002/pra2.2015.145052010020, Available at http://dx.doi.org/10.1002/pra2.2015.145052010020

47. Druzinsky, R., Mungall, C., Haendel, M., Lapp, H., and Mabee, P. "**What Is an Anatomy Ontology?**" *Anatomical record* 296, no. 12 (2013): 1797–1799. doi:10.1002/ar.22805, Available at http://dx.doi.org/10.1002/ar.22805

48. Segerdell, E., Ponferrada, V. G., James-Zorn, C., Burns, K. A., Fortriede, J. D., Dahdul, W. M., Vize, P. D., and Zorn, A. M. "**Enhanced XAO: The Ontology of Xenopus Anatomy and**

Development Underpins More Accurate Annotation of Gene Expression and Queries on Xenbase" *Journal of biomedical semantics* 4, no. 1 (2013): 31. doi:10.1186/2041-1480-4-31, Available at http://dx.doi.org/10.1186/2041-1480-4-31

49. Diehl, A. D., Meehan, T. F., Bradford, Y. M., Brush, M. H., Dahdul, W. M., Dougall, D. S., He, Y., Osumi-Sutherland, D., Ruttenberg, A., Sarntivijai, S., Van Slyke, C. E., Vasilevsky, N. A., Haendel, M. A., Blake, J. A., and Mungall, C. J. "**The Cell Ontology 2016: Enhanced Content, Modularization, and Ontology Interoperability**" *Journal of biomedical semantics* 7, no. 1 (2016): 44. doi:10.1186/s13326-016-0088-7, Available at http://dx.doi.org/10.1186/s13326-016-0088-7

50. Manda, P., Mungall, C. J., Balhoff, J. P., Lapp, H., and Vision, T. J. "**Investigating the Importance of Anatomical Homology for Cross-Species Phenotype Comparisons Using Semantic Similarity**" *Biocomputing 2016* (2016): 132–143. doi:10.1142/9789814749411_0013, Available at http://www.worldscientific.com/doi/10.1142/9789814749411_0013

51. Druzinsky, R. E., Balhoff, J. P., Crompton, A. W., Done, J., German, R. Z., Haendel, M. A., Herrel, A., Herring, S. W., Lapp, H., Mabee, P. M., Muller, H.-M., Mungall, C. J., Sternberg, P. W., Van Auken, K., Vinyard, C. J., Williams, S. H., and Wall, C. E. "**Muscle Logic: New Knowledge Resource for Anatomy Enables Comprehensive Searches of the Literature on the Feeding Muscles of Mammals**" *PloS one* 11, no. 2 (2016): e0149102. doi:10.1371/journal.pone.0149102, Available at http://dx.doi.org/10.1371/journal.pone.0149102

52. Dececchi, T. A., Mabee, P. M., and Blackburn, D. C. "**Data Sources for Trait Databases: Comparing the Phenomic Content of Monographs and Evolutionary Matrices**" *PloS one* 11, no. 5 (2016): e0155680. doi:10.1371/journal.pone.0155680, Available at http://dx.doi.org/10.1371/journal.pone.0155680

53. Balhoff, J. P. "**Scowl: A Scala DSL for Programming with the OWL API**" *The Journal of Open Source Software* 1, no. 1 (2016): doi:10.21105/joss.00023, Available at http://dx.doi.org/10.21105/joss.00023

54. "**Phenoscape**" *Phenoscape GitHub site* Available at https://github.com/phenoscape

55. Vos, R. A., Balhoff, J. P., Caravas, J. A., Holder, M. T., Lapp, H., Maddison, W. P., Midford, P. E., Priyam, A., Sukumaran, J., Xia, X., and Stoltzfus, A. "**NeXML: Rich, Extensible, and Verifiable Representation of Comparative Data and Metadata**" *Systematic Biology* 61, no. 4 (2012): 675–689. doi:10.1093/sysbio/sys025, Available at http://dx.doi.org/10.1093/sysbio/sys025

56. W3C OWL Working Group. "**OWL 2 Web Ontology Language Document Overview**" (2009): Available at https://www.w3.org/TR/owl2-overview/

57. O'Leary, M. A. and Kaufman, S. "**MorphoBank: Phylophenomics in the 'cloud'**" *Cladistics: the international journal of the Willi Hennig Society* 27, no. 5 (2011): 529–537. doi:10.1111/j.1096-0031.2011.00355.x, Available at http://dx.doi.org/10.1111/j.1096-0031.2011.00355.x

58. Pesquita, C., Faria, D., Falcão, A. O., Lord, P., and Couto, F. M. "**Semantic Similarity in Biomedical Ontologies**" *PLoS computational biology* 5, no. 7 (2009): e1000443. doi:10.1371/journal.pcbi.1000443, Available at http://dx.doi.org/10.1371/journal.pcbi.1000443

59. Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R., Coghill, L. M., Crandall, K. A., Deng, J., Drew, B. T., Gazis, R., Gude, K., Hibbett, D. S., Katz, L. A., Laughinghouse, H. D., 4th, McTavish, E. J., Midford, P. E., Owen, C. L., Ree, R. H., Rees, J. A., Soltis, D. E., Williams, T., and Cranston, K. A. "**Synthesis of Phylogeny and Taxonomy into a Comprehensive Tree of Life**" *Proceedings of the National Academy of Sciences of the United States of America* (2015): doi:10.1073/pnas.1423041112, Available at http://dx.doi.org/10.1073/pnas.1423041112

60. Rüber, L., Kottelat, M., Tan, H. H., Ng, P. K. L., and Britz, R. "**Evolution of Miniaturization and the Phylogenetic Position of Paedocypris, Comprising the World's Smallest Vertebrate**" *BMC evolutionary biology* 7, (2007): 38. doi:10.1186/1471-2148-7-38, Available at http://dx.doi.org/10.1186/1471-2148-7-38

61. Mirande, J. M. "**Phylogeny of the Family Characidae (Teleostei: Characiformes): From Characters to Taxonomy**" *Neotropical ichthyology: official journal of the Sociedade Brasileira de Ictiologia* 8, no. 3 (2010): 385–568. doi:10.1590/s1679-62252010000300001, Available at http://dx.doi.org/10.1590/s1679-62252010000300001

62. Sidlauskas, B. L. and Vari, R. P. "**Phylogenetic Relationships within the South American Fish Family Anostomidae (Teleostei, Ostariophysi, Characiformes)**" *Zoological journal of the Linnean Society* 154, no. 1 (2008): 70–210. doi:10.1111/j.1096-3642.2008.00407.x, Available at http://dx.doi.org/10.1111/j.1096-3642.2008.00407.x

63. R Core Team. "**R: A Language and Environment for Statistical Computing**" (2015): Available at https://www.R-project.org

64. Xu, H. and Lapp, H. "**Rphenoscape**" Available at https://github.com/xu-hong/rphenoscape

65. O'Meara, B. C. "**Evolutionary Inferences from Phylogenies: A Review of Methods**" *Annual review of ecology, evolution, and systematics* 43, no. 1 (2012): 267–285. doi:10.1146/annurev-ecolsys-110411-160331, Available at http://dx.doi.org/10.1146/annurev-ecolsys-110411-160331

66. Lewis, P. O. "**A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data**" *Systematic biology* 50, no. 6 (2001): 913–925. Available at http://www.ncbi.nlm.nih.gov/pubmed/12116640

67. Maddison, W. P. and FitzJohn, R. G. "**The Unsolved Challenge to Phylogenetic Correlation Tests for Categorical Characters**" *Systematic biology* 64, no. 1 (2015): 127–136. doi:10.1093/sysbio/syu070, Available at http://dx.doi.org/10.1093/sysbio/syu070

68. Pagel, M. "**Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters**" *Proceedings of the Royal Society of London B: Biological Sciences* 255, no. 1342 (1994): 37–45. doi:10.1098/rspb.1994.0006, Available at http://rspb.royalsocietypublishing.org/content/255/1342/37

69. Bielejec, F., Lemey, P., Baele, G., Rambaut, A., and Suchard, M. A. "**Inferring Heterogeneous Evolutionary Processes through Time: From Sequence Substitution to Phylogeography**" *Systematic biology* 63, no. 4 (2014): 493–504. doi:10.1093/sysbio/syu015, Available at http://dx.doi.org/10.1093/sysbio/syu015

70. Beaulieu, J. M., O'Meara, B. C., and Donoghue, M. J. "**Identifying Hidden Rate Changes**

in the Evolution of a Binary Morphological Character: The Evolution of Plant Habit in Campanulid Angiosperms" *Systematic biology* 62, no. 5 (2013): 725–737. doi:10.1093/sysbio/syt034, Available at http://dx.doi.org/10.1093/sysbio/syt034

71. Hickerson, M. J., Stahl, E., and Takebayashi, N. "**msBayes: Pipeline for Testing Comparative Phylogeographic Histories Using Hierarchical Approximate Bayesian Computation**" *BMC bioinformatics* 8, (2007): 268. doi:10.1186/1471-2105-8-268, Available at http://dx.doi.org/10.1186/1471-2105-8-268

72. Heath, T. A., Holder, M. T., and Huelsenbeck, J. P. "**A Dirichlet Process Prior for Estimating Lineage-Specific Substitution Rates**" *Molecular biology and evolution* 29, no. 3 (2012): 939–955. doi:10.1093/molbev/msr255, Available at http://dx.doi.org/10.1093/molbev/msr255

73. Lartillot, N. and Philippe, H. "**A Bayesian Mixture Model for across-Site Heterogeneities in the Amino-Acid Replacement Process**" *Molecular biology and evolution* 21, no. 6 (2004): 1095–1109. doi:10.1093/molbev/msh112, Available at http://dx.doi.org/10.1093/molbev/msh112

74. Moore, B. R., McGuire, J., Ronquist, F., and Huelsenbeck, J. P. "**Bayesian Analysis of Partitioned Data**" *arXiv [q-bio.PE]* (2014): Available at http://arxiv.org/abs/1409.0906

75. Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., and Ronquist, F. "**RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language**" *Systematic biology* 65, no. 4 (2016): 726–736. doi:10.1093/sysbio/syw021, Available at http://dx.doi.org/10.1093/sysbio/syw021

76. Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. "**Hierarchical Dirichlet Processes**" *Journal of the American Statistical Association* 101, no. 476 (2006): 1566–1581. doi:10.1198/016214506000000302, Available at http://dx.doi.org/10.1198/016214506000000302

77. Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A., Richardson, J. E., and Mouse Genome Database Group. "**The Mouse Genome Database (MGD): Facilitating Mouse as a Model for Human Biology and Disease**" *Nucleic acids research* 43, no. Database issue (2015): D726–36. doi:10.1093/nar/gku967, Available at http://dx.doi.org/10.1093/nar/gku967

78. Howe, D. G., Bradford, Y. M., Conlin, T., Eagle, A. E., Fashena, D., Frazer, K., Knight, J., Mani, P., Martin, R., Moxon, S. A. T., Paddock, H., Pich, C., Ramachandran, S., Ruef, B. J., Ruzicka, L., Schaper, K., Shao, X., Singer, A., Sprunger, B., Van Slyke, C. E., and Westerfield, M. "**ZFIN, the Zebrafish Model Organism Database: Increased Support for Mutants and Transgenics**" *Nucleic acids research* 41, no. Database issue (2013): D854–60. doi:10.1093/nar/gks938, Available at http://dx.doi.org/10.1093/nar/gks938

79. Karpinka, J. B., Fortriede, J. D., Burns, K. A., James-Zorn, C., Ponferrada, V. G., Lee, J., Karimi, K., Zorn, A. M., and Vize, P. D. "**Xenbase, the Xenopus Model Organism Database; New Virtualized System, Data Types and Genomes**" *Nucleic acids research* 43, no. Database issue (2015): D756–63. doi:10.1093/nar/gku956, Available at http://dx.doi.org/10.1093/nar/gku956

80. McMurry, J. A., Köhler, S., Washington, N. L., Balhoff, J. P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., Foster, E., Gourdine, J.-P., Jacobsen, J. O. B.,

Keith, D., Laraway, B., Xuan, J. N., Shefchek, K., Vasilevsky, N. A., Yuan, Z., Lewis, S. E., Hochheiser, H., Groza, T., Smedley, D., Robinson, P. N., Mungall, C. J., and Haendel, M. A. "**Navigating the Phenotype Frontier: The Monarch Initiative**" *Genetics* 203, no. 4 (2016): 1491–1495. doi:10.1534/genetics.116.188870, Available at http://dx.doi.org/10.1534/genetics.116.188870

81. Sullivan, J. P., Lundberg, J. G., and Hardman, M. "**A Phylogenetic Analysis of the Major Groups of Catfishes (Teleostei: Siluriformes) Using rag1 and rag2 Nuclear Gene Sequences**" *Molecular phylogenetics and evolution* 41, no. 3 (2006): 636–662. doi:10.1016/j.ympev.2006.05.044, Available at http://dx.doi.org/10.1016/j.ympev.2006.05.044

82. Calcagnotto, D., Schaefer, S. A., and DeSalle, R. "**Relationships among Characiform Fishes Inferred from Analysis of Nuclear and Mitochondrial Gene Sequences**" *Molecular phylogenetics and evolution* 36, no. 1 (2005): 135–153. doi:10.1016/j.ympev.2005.01.004, Available at http://dx.doi.org/10.1016/j.ympev.2005.01.004

83. Bührnheim, C. M., Carvalho, T. P., and Malabarba, L. R. "**A New Genus and Species of Characid Fish from the Amazon Basin: The Recognition of a Relictual Lineage of Characid Fishes (Ostariophysi: Cheirodontinae: …**" *Neotropical* (2008): Available at http://www.scielo.br/scielo.php?pid=S1679-62252008000400016&script=sci_arttext

84. Netto-Ferreira, A. L., Birindelli, J. L. O., Sousa, L. M. de, Mariguela, T. C., and Oliveira, C. "**A New Miniature Characid (Ostariophysi: Characiformes: Characidae), with Phylogenetic Position Inferred from Morphological and Molecular Data**" *PloS one* 8, no. 1 (2013): e52098. doi:10.1371/journal.pone.0052098, Available at http://dx.doi.org/10.1371/journal.pone.0052098

85. Conway, K. W., Chen, W. J., and Mayden, R. L. "**The 'Celestial Pearl Danio' Is a Miniature Danio (ss)(Ostariophysi: Cyprinidae): Evidence from Morphology and Molecules**" *Zootaxa* no. 1686 (2008): 1–28. Available at http://specifyassets.nhm.ku.edu/Ichthyology/originals/sp68665765193421653212.att.pdf

86. Oliveira, C., Avelino, G. S., Abe, K. T., Mariguela, T. C., Benine, R. C., Ort'i, G., Vari, R. P., and Corrêa e Castro, R. M. "**Phylogenetic Relationships within the Speciose Family Characidae (Teleostei: Ostariophysi: Characiformes) Based on Multilocus Analysis and Extensive Ingroup Sampling**" *BMC evolutionary biology* 11, no. 1 (2011): 1–25. doi:10.1186/1471-2148-11-275, Available at http://dx.doi.org/10.1186/1471-2148-11-275

87. Mattox, G. M. T. and Toledo-Piza, M. "**Phylogenetic Study of the Characinae (Teleostei: Characiformes: Characidae)**" *Zoological journal of the Linnean Society* 165, no. 4 (2012): 809–915. doi:10.1111/j.1096-3642.2012.00830.x, Available at http://dx.doi.org/10.1111/j.1096-3642.2012.00830.x

88. Harms, M. J. and Thornton, J. W. "**Analyzing Protein Structure and Function Using Ancestral Gene Reconstruction**" *Current opinion in structural biology* 20, no. 3 (2010): 360–366. doi:10.1016/j.sbi.2010.03.005, Available at http://dx.doi.org/10.1016/j.sbi.2010.03.005

89. Swofford, D. L. and Maddison, W. P. "**Reconstructing Ancestral Character States under Wagner Parsimony**" *Mathematical biosciences* 87, no. 2 (1987): 199–229. doi:10.1016/0025-5564(87)90074-5, Available at http://dx.doi.org/10.1016/0025-5564(87)90074-5

90. Omland, K. E. "**The Assumptions and Challenges of Ancestral State Reconstructions**"

Systematic biology 48, no. 3 (1999): 604–611. doi:10.1080/106351599260175, Available at http://dx.doi.org/10.1080/106351599260175

91. Pagel, M. "**The Maximum Likelihood Approach to Reconstructing Ancestral Character States of Discrete Characters on Phylogenies**" *Systematic biology* 48, no. 3 (1999): 612–622. doi:10.1080/106351599260184, Available at http://dx.doi.org/10.1080/106351599260184

92. Pagel, M., Meade, A., and Barker, D. "**Bayesian Estimation of Ancestral Character States on Phylogenies**" *Systematic biology* 53, no. 5 (2004): 673–684. doi:10.1080/10635150490522232, Available at http://dx.doi.org/10.1080/10635150490522232

93. Fitch, W. M. "**Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology**" *Systematic biology* 20, no. 4 (1971): 406–416. doi:10.1093/sysbio/20.4.406, Available at http://dx.doi.org/10.1093/sysbio/20.4.406

94. Bauer, S., Köhler, S., Schulz, M. H., and Robinson, P. N. "**Bayesian Ontology Querying for Accurate and Noise-Tolerant Semantic Searches**" *Bioinformatics* 28, no. 19 (2012): 2502–2508. doi:10.1093/bioinformatics/bts471, Available at http://doi.org/10.1093/bioinformatics/bts471

95. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. "**Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles**" *Proceedings of the National Academy of Sciences* 102, no. 43 (2005): 15545–15550. doi:10.1073/pnas.0506580102, Available at http://dx.doi.org/10.1073/pnas.0506580102

96. Galis, F. and Metz, J. A. J. "**Why Are There so Many Cichlid Species?**" *Trends in ecology & evolution* 13, no. 1 (1998): 1–2. doi:10.1016/s0169-5347(97)01239-1, Available at http://dx.doi.org/10.1016/s0169-5347(97)01239-1

97. Tarca, A. L., Bhatti, G., and Romero, R. "**A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity**" *PloS one* 8, no. 11 (2013): e79217. doi:10.1371/journal.pone.0079217, Available at http://dx.doi.org/10.1371/journal.pone.0079217

98. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. "**The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration**" *Nature biotechnology* 25, no. 11 (2007): 1251–1255. doi:10.1038/nbt1346, Available at http://dx.doi.org/10.1038/nbt1346

99. Mattox, G. M. T., Britz, R., and Toledo-Piza, M. "**Osteology of Priocharax and Remarkable Developmental Truncation in a Miniature Amazonian Fish (Teleostei: Characiformes: Characidae)**" *Journal of morphology* 277, no. 1 (2016): 65–85. doi:10.1002/jmor.20477, Available at http://dx.doi.org/10.1002/jmor.20477

100. Maddison, W. P. and Maddison, D. R. "**Mesquite: A Modular System for Evolutionary Analysis**" (2015): Available at http://mesquiteproject.org/