Running Head: AUTOMATED INTEGRATION OF TREES AND TRAITS

Automated Integration of Trees and Traits: A Case Study Using Paired Fin Loss

Across Teleost Fishes

Laura M. Jackson1\*, Pasan C. Fernando1, Josh S. Hanscom1, James P. Balhoff2,

Paula M. Mabee1\*

*1Department of Biology, University of South Dakota, Vermillion, SD 57069, USA*

*2Renaissance Computing Institute, University of North Carolina, Chapel Hill, NC 27517, USA*

\* Corresponding author

E-mail: laura.jackson@coyotes.usd.edu, pmabee@usd.edu

ABSTRACT

Data synthesis required for large-scale macroevolutionary studies is challenging with the current tools available for integration. Using a classic question regarding the frequency of paired fin loss in teleost fishes as a case study, we sought to create automated methods to facilitate the integration of broad-scale trait data with a sizable species-level phylogeny. Similar to the evolutionary pattern previously described for limbs, pelvic and pectoral fin reduction and loss are thought to have occurred independently multiple times in the evolution of fishes. We developed a bioinformatics pipeline to integrate a synthetic morphological supermatrix of phenotypic data for the pectoral and pelvic fins for teleost fishes from the Phenoscape Knowledgebase (two presence/absence characters for 3,047 taxa) with a species-level tree for teleost fishes from the Open Tree of Life project (38,419 species). The integration method detailed herein harnessed a new combined approach by utilizing data based on ontological inference, as well as phylogenetic propagation, to reduce overall data loss. Using inference enabled by ontology-based annotations, missing data were reduced from 98.0% to 85.9%, and further reduced to 34.8% by phylogenetic data propagation. These methods allowed us to extend the data to an additional 11,293 species for a total of 12,582 species with trait data. The pectoral fin appears to have been independently lost in a minimum of 19 lineages and the pelvic fin in 48. Though interpretation is limited by lack of phylogenetic resolution at the species level, it appears that following loss, both pectoral and pelvic fins were regained several (3) to many (14) times respectively. Focused investigation into putative regains of the pectoral fin, all within one clade (Anguilliformes), analyzed over 1,000 randomly resolved topologies, showed that the pectoral fin was regained at least twice following loss. Overall, this study points to specific areas within teleosts where strategic phylogenetic resolution and genetic investigation will be necessary to understand the pattern and frequency of pectoral fin reversals.

Keywords: ontology, inference, ancestral state reconstruction, paired fins, phenotype, evolutionary morphology, missing data, bioinformatics pipeline

How often—across all 38,000+ species of teleost fishes—were pectoral and pelvic fins lost? Are they ever regained? The information required to answer these straightforward questions regarding character evolution, namely trait data and a relatively resolved species-level phylogeny, has not been readily available at the scale of tens of thousands of species until recently (Dececchi et al. 2015; Hinchliff et al. 2015). Investigators who have tried to approach such macroevolutionary questions combining trait data with phylogenies at magnitudes not readily supported by current tools have been confronted with challenges and hurdles along the way (Harris and Arbuckle 2016; Hunt and Slater 2016). Alfaro et al. (2009) suggest that some of these challenges involve producing and manipulating large-scale phylogenies. In addition, the development of larger phylogenies has further exposed the lack of resources necessary to interpret results (Harmon et al. 2013). Not only is there a lack of visualization tools necessary to interpret the evolution of large trait data, but also phylogenetic programs lack the scalability necessary to approach macroevolutionary questions in a reasonable timeframe (Harmon et al. 2013; Harris and Arbuckle 2016; Hunt and Slater 2016). Even though visualization platforms to investigate trait evolution at larger scales are improving (e.g., iTOL, Letunic and Bork 2007; Arbor; Harmon et al. 2013), many still lack the ability to integrate across multiple data formats and analyses.

Using questions concerning the frequency of paired fin loss in fishes as an example, we demonstrate the use of new knowledge resources to address basic questions involving the pattern of evolution of phenotypic features. Increasing the taxonomic scope makes apparent the value of these new resources, as well as deficiencies in existing methods to integrate the data. The pipeline developed in the process of this work reflects a set of essential requirements for large-scale macroevolutionary syntheses.

The peculiar absence of paired fins in fishes such as eels (Nelson 2006) has fascinated scientists since at least the time of Aristotle (Ogle 1882; Leunissen 2010). Notably, over the past 40 years, ichthyologist Joseph S. Nelson (1976, 1984, 1994, 2006) documented the absence of pelvic fins in members of 92 teleost families (Nelson 1990), which represent nearly one-quarter of all fish families. This led him to conclude that pelvic fins have been independently lost at least 50 times (Nelson 1990). His analysis, as well as subsequent ones (Yamanoue et al. 2010), however, did not consider phylogenetic relationships, in which absence or presence at an ancestral node influences the number of putative losses. Further, Nelson (2006) summarized data at the family level, e.g., for Ophichthidae: ‘pectoral fins present or absent’, often without naming the species associated with a particular condition or citing the primary literature in this regard. In contrast to pelvic fins, the frequency of pectoral fin loss is poorly documented, with few exceptions (Britz 2010; Yamanoue et al. 2010; Brown et al. 2011). Further, whether pectoral or pelvic fins may have been regained following loss remains uncertain (Nelson 1990; Yamanoue et al. 2010). Though there are known exceptions to Dollo’s law, i.e., that the loss of a complex trait is evolutionarily irreversible (Gould 1970; Farris 1977), questions concerning the taxonomic scope, frequency of reversal, and potential genetic bases (Collin and Miglietta 2008) remain unanswered.

Addressing any broad-scale question concerning the evolution of traits requires a comprehensive source of data. Such large data sets must be readily extractable and computable, as manual aggregation from a dispersed literature is essentially intractable. Free text phenotypic descriptions from the literature that are tagged with appropriate ontology terms (via Uniform Resource Identifiers [URIs]) provide semantic information that allows for automated collection and computation of morphological data across species. For paired fin data, we turned to the Phenoscape Knowledgebase (KB; kb.phenoscape.org) for computable phenotypic biodiversity information across vertebrates. The KB contains ontology-annotated phenotypic data based primarily on published character matrices (Dececchi et al. 2015), but also some monographic treatments (e.g., Nelson 2006; Dececchi et al. 2016). It is particularly enriched in vertebrate skeletal features of fins, limbs, and their support structures. The ontology-based data uniquely allow inference of the presence or absence of a phenotypic feature based on indirect descriptions of the feature or its parts. Such inference has been shown to greatly enlarge the available data (Dececchi et al. 2015), a desirable feature here given the paucity of direct statements by authors concerning the presence/absence of paired fins. This resource can be queried using OntoTrace (Dececchi et al. 2015), a downloadable matrix generation tool enabling users to extract synthetic morphological supermatrices of features and all metadata for any taxon into a single file in NeXML format (Vos et al. 2012). Thus, instead of combing through publications relevant to 38,000+ teleost species and compiling a matrix encompassing the full scope of data relevant to our question, ontology concepts that populate the KB can be automatically exported into a data set for analysis.

We sought to obtain the most fully resolved phylogenetic tree for all 38,000+ extant and extinct teleost species on which to map the paired fin data. Recent literature includes some well-resolved and broad-scale trees based on molecular data (Near et al. 2012; Betancur et al. 2013); however, there are several impediments to their use. First, they involve only a subset of teleost species, e.g., Betancur and colleagues (2013) sampled 1,410 species. Second, when the terminal taxa are at a supra-specific level, e.g., families, orders, and superorders as in the Near et al. (2012) tree, it is difficult to know which species were included in these groupings at the time of their analysis (though actual species sampled were provided in the case of Near et al. (2012)). Further, even if authors report species inclusion, the effort to add these manually to a large tree is unsupportable. Third, differences in topologies between different studies often present an issue that could be handled manually by an investigator, but at the scale of all teleosts, a consistent and automated way to approach the many differences in topologies across published trees is preferable. Thus we turned to the Open Tree of Life project (Open Tree; http://opentreeoflife.org) for a digitally available comprehensive tree at the species level (Hinchliff et al. 2015). The Open Tree dynamically constructs a tree by synthesizing published phylogenies along with taxonomic data using the ‘propinquity’ supertree pipeline (Redelings and Holder 2017). The output includes detailed provenance reports (e.g., node support, conflicts and resolutions) associated with nodes resolved by a source other than the reference taxonomy.

An overarching goal in this study was not only to fully use all available trait data, but also to minimize its loss, so as to enable the most comprehensive ancestral state reconstruction. This was achieved by several means. First, we used inferred trait data for taxa (e.g., if a part of a fin is present, the fin is inferred to be present), thus fully utilizing the existing information. Second, we propagated to the species level, data that investigators associated with genera and families, thus extending the data as intended by the authors. Finally, we improved the method of taxonomic reconciliation between taxa to which trait data are attached and taxa included in the phylogeny. This was necessary because different sources of names are used in taxonomies referenced by the Phenoscape KB *vs.* the Open Tree of Life. The total expansion of phenotypic data through ontology-based inference and taxonomically-based propagation was significant, and a valuable model to be followed for macroevolutionary studies.

METHODS

*Large-scale Computable Phenotypic Data*

The Phenoscape Knowledgebase (KB) contains 21,569 morphological character states from 171 comparative studies, mainly phylogenetic. but also some review papers (Wiley and Johnson 2010; Yamanoue et al. 2010) and monographic treatments (Dececchi et al. 2016). In total the KB includes 526,221 phenotype annotations for 5,208 extant and fossil vertebrates (as of October 21, 2016). Phenotypic data from the KB are annotated with taxonomic names from the Vertebrate Taxonomy Ontology (VTO, http://purl.obolibrary.org/obo/vto/2016-1017/vto.owl; Midford et al. 2013). However, the KB does not provide the phylogenetic relationships necessary for evolutionary mapping of characters; the relationships in the VTO are based solely on taxonomy. The VTO is built upon the National Center for Biotechnology Information (NCBI) taxonomy, which provides the hierarchical backbone for extant taxa, the Catalog of Fishes (CoF; Eschmeyer 2013), and PaleoDB (Uhen et al. 2013), which supplements extant and extinct taxa (Midford et al. 2013). It contains 31,726 teleost species; somewhat differing from counts in the current CoF (33,008), because it has been updated since last being incorporated into the VTO (July 2012).

Twenty-two of the 171 studies were added to the KB (Fig. 1) to fully represent the distribution of pelvic and pectoral fins across teleosts (Supplementary Materials Table 1). These included, e.g., a recent review including 776 synapomorphic characters used to diagnose teleost families (Wiley and Johnson 2010). Characters were annotated using the Entity–Quality (EQ) formalism (Mungall et al. 2007; Mungall et al. 2010) with Phenex software (Balhoff et al. 2010; Balhoff et al. 2014). Specifically, ontological terms and relationships among anatomical aspects of the paired fins, girdles, their parts and developmental precursors (Fig. 2) were drawn from the multispecies anatomy ontology, Uberon (http://purl.obolibrary.org/obo/uberon/releases/2016-09-07/ uberon.owl; Mungall et al. 2012; Haendel et al. 2014). Quality terms that represent the variation in these anatomical entities, such as presence/absence, size, or shape, were drawn from the Phenotype and Trait Ontology (PATO, http://purl.obolibrary.org/obo/pato/releases/2016-09-15/ pato.owl; Gkoutos et al. 2005). Terms and relationships in Uberon were edited using the ontology editing software Protégé v4.3 (Noy et al. 2003). New terms were added to Uberon as driven by the curated literature, e.g., the term ‘pelvic intercleithral cartilage’ was used in the description of the pelvic fin for gobiiform fishes (Wiley and Johnson 2010), and thus it was added as UBERON:4300151, a type of ‘cartilage element’ (UBERON:0007844) and a type of ‘pelvic region element’ (UBERON:0005179). Character EQ annotations were annotated with names of species or higher taxa as specified in the publications, and these terms were drawn from the VTO (Midford et al. 2013).

In some phylogenetic treatments (e.g., Chapleau 1993; Johnson and Patterson 1993) and reviews (Nelson 1990; Wiley and Johnson 2010), anatomical features are described for higher-level taxa in addition to species. We applied assertions at the taxonomic level described by the author. For example, if an author states that a genus lacks a fin, we annotated absence (‘0’) to the genus. If they also indicate that the fin is present at the family level, we annotated presence (‘1’) to the family. In cases where authors describe features for supra-familial-level taxa, but do not specify the families to which the taxonomic grouping applies, it was annotated to the specific group named. For example, ‘sixteen principal branched caudal fin rays’ is stated to characterize Osteoglossomorpha (Wiley and Johnson 2010), and thus we annotated ‘Osteoglossomorpha’ with this character (lower taxonomic levels were not annotated). However, where authors indicate family membership, we applied annotations to the contained families (e.g., synapomorphies for Lampridiformes were applied to seven families, as per the authors’ description that ‘Lampridiformes comprise seven monophyletic families, Veliferidae, Lamprididae, Stylephoridae, Lophotidae, Radiicephalidae, Trachipteridae, and Regalecidae,’ (Wiley and Johnson 2010)). Authors may also describe characters for higher-level taxa not contained in the VTO (e.g., ‘Holacanthopterygii’ in Wiley and Johnson (2010)) and taxa for which included taxa were not specified; in these cases annotations were not made. Finally, in cases where the taxonomy used in a publication differed from the VTO, we annotated the anatomical features to the taxa as intended by the author. For example, some taxa included in Osmeridae by Wiley and Johnson (2010) are elevated to their own family (Plecoglossidae) in the VTO; we ensured that the characters were applied to the relevant taxa within the family Plecoglossidae.

In cases where a higher-level taxon is described as polymorphic for the presence and absence of a particular fin, we investigated the literature to determine which of its species possess or lack it. When presence/absence data were traceable to species, we curated these studies to the KB. Sometimes this information is not available; though pelvic fin absence is asserted for the catfish families Schilbidae and Siluridae (Nelson 2006), it could not be traced in the literature to particular species. Thus, we did not include data for these families in the analyses.

An additional challenge for annotation of data in review papers, particularly at the family level or above, is the frequent lack of alternative character state descriptions. For example, eels (Anguilliformes) were asserted to have the apomorphic feature ‘pelvic girdle and fins absent’ (Wiley and Johnson 2010), and although by implication one might assume that the alternative state ‘pelvic girdle and fins present’ might apply to other fishes, it is not stated by the authors. In these cases, we annotated only the asserted state and never an implied alternative state so as to not misrepresent information. The guidelines developed for annotation that reflect the above conditions were added to the Phenoscape Guide to Character Annotation (http://phenoscape.org/wiki/Guide\_to\_Character\_Annotation; Dahdul et al. 2010), and followed consistently.

*Synthetic Morphological Supermatrix*

We used the OntoTrace tool (Dececchi et al. 2015) to retrieve a presence/absence synthetic morphological supermatrix from the KB for all teleost taxa, pertaining to two characters, pectoral fin and pelvic fin (Supplementary Materials Matrix 1, 06/08/2017) with the following query (taxon: Teleostei, entities: ‘pectoral fin’ or ‘pelvic fin’). The matrix is in NeXML format and contains all associated metadata pertaining to data provenance (Fig. 1). Using inference that is enabled by the logical relationships among anatomical features (Fig. 2), ontology-annotated and thus computer-readable data can be reasoned to infer the presence or absence of entities that were not directly asserted. If an anatomical part of the fin is described, e.g., ‘pectoral fin rays are unbranched’ in *Neocyttus rhomhoidalis* (Tyler et al. 2003), here based on direct evidence from a voucher specimen at the American Museum of Natural History (AMNH 91746, Tyler 1980; Tyler et al. 2003), then the computer will infer that the pectoral fin rays and the pectoral fin are present (Fig. 3). The converse is not true, i.e., if the pectoral fin is present, it does not imply that particular parts are present. The absence of a paired fin, however, would be inferred from the absence of its girdle, as seen in *Acanthostracion quadricornis* (Fig. 3) based on multiple voucher specimens from the Academy of Natural Sciences of Philadelphia (e.g., ANSP 98614, ANSP 98615, ANSP 9816; Santini and Tyler 2003). This reflects domain knowledge that paired fins are never present without their supporting girdle structures. In the converse, however, i.e., if the girdle is present, it does not mean that the fin is present. There are several examples in fishes where the pectoral girdle is present, but the pectoral fin is absent(Nelson 2006), such as the black pomfret (*Parastromateus niger*)*,* the Parona leatherjacket (*Parnoa signata*)*,* and the fanfin *(Robia legula*). The absence of a paired fin can also be inferred from the absence of the larval fin or fin bud, but not the converse, i.e., the presence of a larval fin or fin bud does not imply the presence of a fin. For example, in the William’s tonguefish (*Symphurus williamsi*; Aceves-Medina et al. 1999), the larval pectoral fin does not persist in development; adults lack the pectoral fin.

*Teleostei Species-level Tree*

The Open Tree of Life (Hinchliff et al. 2015) uses the ‘propinquity’ supertree pipeline to generate synthetic trees from multiple input phylogenies and a reference taxonomy (Redelings and Holder 2017). The reference taxonomy includes taxonomic data for fishes from the National Center for Biotechnology Information (NCBI), The Interim Register for Marine and Non-marine Genera (IRMNG), the Global Biodiversity Information Facility (GBIF), and the World Register of Marine Species (WoRMS); none of these sources includes fossil species labeled as such. Further, it does not contain taxonomic information for fishes based on the Catalog of Fishes. The input phylogenies for Open Tree are published trees that are manually curated to align tips with Open Tree taxonomy (Rees and Cranston 2017). The propinquity supertree pipeline integrates and summarizes input phylogenies and the reference taxonomy into a single rooted synthetic supertree, which can be customized according to user preferences. Within this publicly available tree (Open Tree 2.10; see Hinchliff et al. 2015), fourteen families had species in which the pectoral and/or pelvic fin were absent but for whom species relationships were unresolved. To provide better resolution within these families, we curated available phylogenies for them to a Teleostei tree collection (https://tree.opentreeoflife.org/curator/collections/laurajackson/teleostei) in OpenTree (Fig. 1). Phylogeny curation was done for Anguilliformes ([Anguillidae, Congridae, Cyematidae, Derichthyidae, Ophichthidae, Nettastomatidae]: Santini et al. 2013; Chlopsidae: Tang and Fielitz 2013), Percomorpha ([Chaudhuriidae, Indostomidae, Mastacembelidae, Synbranchidae]: Kawahara et al. 2008), Gymnotiformes ([Apteronotidae, Sternopygidae]: Albert 2001) and Perciformes (Trichuridae: Johnson 1986) These phylogenies are queued for inclusion in the next version of the synthetic tree from Open Tree. We obtained a customized (Redelings and Holder 2017) synthetic tree (10/18/2016) from Open Tree limited to Teleostei (Supplementary Materials File 1), including the curated phylogenies in the above Teleostei tree collection, with different run parameters (i.e., excluding subspecies names and including *incertae sedis* taxa). All archive files associated with the supertree pipeline are available in Supplementary Materials File 2; see Dryad data repository (http://dx.doi.org/10.5061/dryad.v0s27).

*Bioinformatics Pipeline to Merge Synthetic Morphological Supermatrix and Teleostei Species-level Tree*

Because of the differences between the Open Tree and the VTO taxonomies, the synthetic morphological supermatrix from the KB required transformation to a version that can be mapped to the Open Tree phylogeny. This was achieved by developing a bioinformatics pipeline (Fig. 1; source code available at https://doi.org/10.5281/zenodo.804488).

*Pre-processing the input matrix.—* The input to this pipeline was the synthetic morphological supermatrix in NeXML format obtained from OntoTrace (Supplementary Materials Matrix 1), which contained data for pectoral and pelvic fins. The first step was to pre-process the NeXML matrix by converting it to a tab-delimited version. This generates a pre-processed data matrix (Supplementary Materials Matrix 2), which was used for the following steps (Fig. 1).

*Removal of apparent polymorphisms and conflicts.—*The matrix produced by OntoTrace enables isolation of cells with both presence and absence (represented by ‘0&1’) and detailed provenance reports for all cells. When a taxon is shown to have both presence and absence for one of the paired fins, it indicates a polymorphic condition, an apparent polymorphism, or a conflict in the data. When both presence and absence in a cell are asserted by supporting character states associated with a single source matrix (i.e., same publication) at the level of species, we consider it to be an ‘actual’ polymorphism. When this occurs at a level above species, (e.g., genus, family), where an author does not give specifics about which members of that group show presence and/or absence, we consider these to be ‘apparent’ polymorphisms. Because apparent polymorphisms are not traceable to particular species, they are of little value and were replaced with ‘?’ (Fig. 1).

*Distinguishing inference vs. assertion.—*The states of ‘presence’ (1) or ‘absence’ (0) of the pectoral or pelvic fin in the matrix may result from direct assertion, inference, or both. These were distinguished through an algorithm that used the associated metadata within the synthetic morphological supermatrix (Supplementary Materials Matrix 1), which identifies whether a character state comes from an author assertion, or inference. The pre-processed data matrix (Supplementary Materials Matrix 2) was modified to distinguish between these states for the purpose of generating statistics based on data coming from assertions *vs.* inference (Fig. 1). When a character state is based on a direct author assertion, even if also based on inference, this cell was considered to be asserted, and coded as ‘1’ for asserted presence or ‘0’ for asserted absence. Only those cells that did not contain an author-supported assertion were counted as inferred and represented as a ‘2’ in the matrix. Because there were no instances of only inferred absence, it was not necessary to create an alternative state for this in the matrix.

*Data propagation.—*The pre-processed data matrix (Supplementary Materials Matrix 2) included a substantial number of character states above the species level. These data, if used as ancestral states, could potentially enable a more accurate assessment of position and frequency of character state change. Data at higher-level internal nodes, however, are not considered when using current tools for ancestral state reconstruction without a workaround that requires manual editing, which is not feasible for large-scale data. For example, the R package PhyTools (Revell 2012) has not implemented the ability to do reconstruction using data at internal nodes (pers. comm. Liam Revell, 6/6/15), but it has developed a workaround method, which was not feasible at this scale. Thus, an algorithm was developed and applied to propagate the data of internal taxa at genus and family levels to their species-level descendants based on the VTO (Fig. 1). Taxonomic levels above family were not considered for propagation. Existing species-level data (asserted or inferred) were not modified; i.e., propagated data never replaced species-level data. Data at the genus was propagated first, but if no data were present at the genus level, it was propagated from the family level. Species in the VTO previously lacking data were thus automatically added to the data matrix with the propagated data. During the propagation process, all higher-level taxa were removed from the matrix, and the propagated matrix (Supplementary Materials Matrix 3) contained data only at the species level.

*Reconciliation of taxon names*.—The synthetic morphological supermatrix from OntoTrace (Supplementary Materials Matrix 1) contained taxon names from the (VTO; Dahdul et al. 2012). These names needed to be reconciled with those from Open Tree before merging the data with the tree. This was achieved using an algorithm in the pipeline, which initially matches taxon names using NCBI Taxonomy IDs (Sayers et al. 2009) as the common identifier. Then, the remaining unmatched taxa are matched using exact taxon name matching. This step generates the final output matrix of the pipeline (Supplementary Materials Matrix 4) (Fig. 1).

*Merging Final Output Matrix with Teleostei Species-level Tree*

Once names were reconciled between the Teleostei species-level tree (Supplementary Materials File 1) and the final output matrix (Supplementary Materials Matrix 4), they were merged into a single NEXUS file (see merged tree matrix; Supplementary Materials File 3) in Mesquite v3.10 (Maddison and Maddison 2016). With the tree file open in Mesquite, the final output matrix was added by merging incoming names with the taxon names in the tree (Fig. 1). Terminal taxa with no associated pectoral fin or pelvic fin data remained as unknown (?) in the final resulting merged tree matrix (Supplementary Materials File 3).

*Ancestral State Reconstruction*

Ancestral state reconstruction (Fig. 1) was done using unordered Fitch parsimony, with the cost of changing from one state to another counted as one step, to calculate the total number of steps corresponding to the instances of gain and loss across teleosts. This model was used instead of likelihood methods, as the tree lacks branch lengths and because polymorphic species are not currently supported by categorical data likelihood calculations in Mesquite. Although branch lengths could be estimated based on the topology, with such a highly unresolved synthesis phylogeny, it would not be likely to provide additional information. However, doing the ancestral reconstruction with a standard parsimony method gave us insight into where major transitions of interest occur along the branches. Mesquite was used to summarize state changes over trees to determine the minimum and maximum number of gains and losses across all Most Parsimonious Reconstruction (MPR) mappings. To determine the minimum number of regains of a trait following loss, polytomies were randomly resolved and branch lengths were computed using the R package APE (Paradis et al. 2004) and summarized using Mesquite. Tree visualizations were created using the Interactive Tree Of Life (iTOL: Letunic and Bork 2007).

RESULTS

*Synthetic Morphological Supermatrix*

The synthetic morphological supermatrix from OntoTrace (Supplementary Materials Matrix 1) was comprised of two characters (pectoral fin and pelvic fin) associated with 3,047taxa (2,663 species, 132 genera, 223 families, 29 above-family taxa) from 87 studies (Supplementary Materials Table 1). Higher-level taxa (genus, family, and order) were included as taxonomic units in 30 of the 87 studies. Of the 4,853 populated cells (of 6,094 total) in the synthetic morphological supermatrix, 616 contained only directly asserted data, 3,953 contained only inferred data, and 284 contained both asserted and inferred data. For pectoral fin, 246 taxa have only asserted data, 2,020 taxa have only inferred data, and 42 taxa have both asserted and inferred data. For the pelvic fin, 370 taxa have only asserted data, 1,933 taxa have only inferred data, and 242 taxa have both asserted and inferred data.

Apparent polymorphic character states and conflicts were identified from 74 taxa (50 families and 24 genera for pelvic fin and 4 families for pectoral fin) and removed from the matrix. Actual polymorphism, i.e., within species variation identified by a single author, was found for only the pelvic fin (in five species: a catfish [*Glanapteryx anguilla*],Nelson 2006; two hatchet herrings [*Pristigaster cayana*]*,* Di Dario 1999, [*Pristigaster sp. (DiDario 1999*)]*,* Di Dario 1999; and two priapumfishes [*Phallostethus lehi* and *Phallostethus dunkeri*], Nelson 2006). Conflicts at the species level that were automatically generated in the process of data aggregation and inference were all between asserted and inferred states. These were found in the pelvic fin for five species (the eel catfish [*Channallabes apus*], two air-breathing catfishes [*Dolichallabes microphthalmus, Gymnallabes typus*], the cobia [*Rachycentron canadum*], the three-spined stickleback [*Gasterosteus aculeatus*]), and in the pectoral fin for one species (the bobtail snipe eel [*Neocyema erythrosoma*]). Conflicts at the species level, as well as the species polymorphisms, were retained in the matrix, because they did not influence the propagation step.

*Teleostei Species-level Tree*

The Teleostei species-level tree retrieved from Open Tree (Supplementary Materials File 1) contained 38,419 species-level tips and 560 families (https://tree.OpenTreeoflife.org/about/taxonomy-version/ott2.10). Open Tree is assembled from a number of different taxonomic sources (Rees and Cranston 2017), but it does not include CoF. However, it includes a substantial number of additional species (over 5,000) compared to CoF. Also, because taxonomic inclusion varies among sources, it is not surprising that the number of families contained in Open Tree differs from those in the CoF and the VTO (families in OT: 560, CoF: 488, VTO: 526; Supplementary Materials Table 2).

*Data Propagation*

Propagation using the relationships in the VTO taxonomy hierarchy transferred asserted and inferred data from 182 families and 119 genera to the included species that otherwise lacked data. This resulted in the addition of 11,293 species to the pre-processed data matrix of 2,663 species (Supplementary Materials Matrix 2) for a total of 13,956 species in the propagated matrix (Supplementary Materials Matrix 3). A comparison of propagated data with directly asserted and inferred data revealed ten instances of conflicts with asserted and two with inferred data.

*Reconciliation of Taxon Names*

Using the combined method of name reconciliation, first with NCBI taxonomic IDs and then exact taxon name matching, an efficient method of alignment between the 13,956 species associated with the propagated matrix (Supplementary Materials Matrix 3) and 38,419 species in the Open Tree was achieved: 12,582 of the 13,956 species were matched with tree tips. This is higher than using either method alone (NCBI taxonomic IDs: 4,423 matches; exact taxon name matching: 12,500 matches). Of the unmatched species (1,374 of 13,956), 72 are fossil species which are not included in the Open Tree taxonomic sources, 362 are species with unconventional names that were added to the VTO because they are referenced in publications (e.g., ‘*Notropis sp. sawfin shiner* *Coburn and Cavender 1992*'), and 940 are unmatched for multiple reasons (e.g., taxonomic name changes, extinct species that are not marked as such in the VTO; Supplementary Materials Table 3).

Before propagation, the pre-processed data matrix (Supplementary Materials Matrix 2) contained only 2,663 species for two characters (pectoral fin and pelvic fin), with 3,538 populated cells for species (85.9% missing data; Table 1). The final output matrix (Supplementary Materials Matrix 4) contained 12,582 species with 16,408 populated cells (34.8% missing data; Table 1). When the final output matrix was merged with the Teleostei species-level tree, however, the missing percentage increases to 78.7% in relation to 76,838 total cells in the merged tree matrix (Supplementary Materials File 3; Table 1). Hypothetically, if the pre-processed matrix were merged with the Teleostei species-level tree before propagation, the percentage of missing data in this matrix would be considerably higher (95.4%; Table 1).

Of the 16,408 populated cells in the final output matrix, 494 (150 pectoral, 344 pelvic) contained only directly asserted data (Fig. 4). The presence of the pectoral fin is asserted in 123 species, and absence asserted in 30. The presence of the pelvic fin is directly asserted in 150 species and absence asserted in 194. In the remaining cells, 3,044 (1,511 pectoral, 1,533 pelvic) contained only inferred data, and 12,870 cells (8,798 pectoral, 4,072 pelvic) contained propagated data (Fig. 4). Of the 8,798 species for which pectoral fin data are propagated, 5,077 of these are propagated from asserted family and genus-level data. Of the 4,072 species for which pelvic fin data are propagated, 2,906 of these are propagated from asserted family and genus-level data.

*Ancestral State Reconstruction*

The pectoral fin is absent in 509 of the 12,582 matched species in the final output matrix (Supplementary Materials Matrix 4), and 21 of the 526 teleost VTO families (Supplementary Materials Table 4; Fig. 5). Of the 21 families, 17 families (494 species) also have pelvic fin absence, with four families (15 species) lacking only the pectoral fin (Fig. 4). Two of the 21 families with pectoral fin absence involve ontogenetic loss (the swamp eels [Synbranchidae], Nelson 2006; William’s tonguefish [*Symphurus williamsi*: Cynoglossidae], Aceves-Medina et al. 1999). The pelvic fin is absent in 2,140 of the 12,582 matched species, and 108 of the 526 teleost VTO families (Supplementary Materials Table 4). This does not include Siluridae and Schilbidae, where family-level assertions of absence were untraceable to a species. Ninety-two (1,652 species) of the 108 families lack only a pelvic fin; 17 families also lack the pectoral fin.

Based on the Open Tree phylogeny used herein, in which pectoral fin presence is the ancestral condition for Teleostei, there were 73,728 MPRs for pectoral fin evolution, each requiring 27 steps. A summary over 1,000 randomly sampled MPRs show a minimum of 19 losses and a minimum of 3 regains of the pectoral fin (Table 2a). All regains occurred in the eels (Anguilliformes), and losses occurred in Anguilliformes, Gobiesociformes, Ophidiiformes, Pleuronectiformes, Saccopharyngiformes, Siluriformes, Stomiiformes, Synbranchiformes, and Syngnathiformes (Fig. 5). For the pelvic fin, in which presence is also the ancestral condition for Teleostei, there were 99,777,458,995,200 MPRs, each requiring 80 steps. A summary over 1,000 randomly sampled MPRs required the regain of the pelvic fin following a loss a minimum of 14 times, and with a minimum of 48 loss events (Table 2b). For the pelvic fin, this occurred primarily within Perciformes, but also within 25 additional orders, such as Anguilliformes, Lophiiformes, Ophidiiformes, and Synbranchiformes (Fig. 6).

Because ancestral reconstruction across all teleost fishes suggested that the pectoral fin has been regained a minimum of three times in Anguilliformes, we investigated this in more detail. A comparison in the reconstruction was done using various topologies with the following parameters (unordered parsimony, computing branch lengths and performing Mk1 likelihood, randomly resolving polytomies, and assuming missing data as presence). This showed that random resolution of 1,000 polytomies (Supplementary Materials File 4) in Anguilliformes resulted in topologies that minimized the number of regains to two (Fig. 7; Supplementary Materials Table 5).

DISCUSSION

Addressing many questions in organismal and evolutionary biology requires knowledge of the traits that species possess or lack, the evolutionary relationships of those species, and the integration of this knowledge—i.e., a mapping of the traits to trees. Many excellent examples have demonstrated the value of this approach, primarily by acquiring the trait data through direct observation of a limited number of species and mapping it to a companion phylogeny that is often generated using molecular data. The availability of trait data, however, has remained limited by the time required for traditional data acquisition, while large phylogenies are increasingly available because of the dramatically accelerated use of molecular data. Further, the concatenation of trait data across different characters and taxa that have been sampled in dispersed and generally small studies has been necessarily manual and thus rarely accomplished on a large scale. This has resulted in a situation where trait data—even simple presence *vs.* absence—needed to inform questions relevant to the species assembled in increasingly comprehensive phylogenies, are sparse.

It was recently shown that morphological data can be readily integrated across matrices by annotating it with ontologies, i.e., by making it computable (Ramírez et al. 2007; Dahdul et al. 2010; Walls et al. 2012; Deans et al. 2015). A substantial level of missing data, however, is the inevitable result of combining morphological data matrices: different characters and taxa are sampled in dispersed and generally small studies. Ontology annotations, however, enable the extension of sparse morphological data to additional species through inference (Dececchi et al. 2015). Using a synthetic morphological supermatrix approach, these authors reduced missing data from 98.5% to 78.2% for a matrix of 639 characters and 1,052 sarcopterygian vertebrates. Here we show a similar reduction in missing data (98.0% to 85.9%) by applying the same logical reasoning to pectoral and pelvic fin traits.

Uniquely in this study, we used phylogenetic propagation, i.e., transferring data from families and genera to included species, to yet further extend the data. In doing so we increased the number of species with data and dramatically reduced missing data to 34.8% in the final output matrix (Supplementary Materials Matrix 4; Fig. 4). The use of propagation was motivated by the desire to preserve the data associated with higher-level taxa, which were included as taxonomic units in over a quarter of the phylogenetic studies mined in the assembly of the supermatrix. There were several considerations in this process. First, we propagated the data from genera and families, but not higher-level ranks, given the increasing expectation of evolutionary changes in character state with increasing divergence time. Second, we eliminated all annotations of ‘0&1’ to higher-level taxa so that they were not propagated. Authors use these annotations, which we termed ‘apparent polymorphisms’, as a shorthand to indicate that both the presence and absence of a trait are found in the species contained in the higher-level clade. In these cases the identity of which specific species possess it *vs*. which lack it, is not provided. Propagating both states to descendant species would be uninformative and misleading, and hence we removed these annotations. Third, we never propagated data from higher-level nodes to species that had existing data. That is, if a species had either asserted or inferred data, it was never ‘overwritten’ by data propagated from the higher-level node. We discovered that a relatively low number of species (only 12 of the 11,293) would have had erroneous data propagated to them, indicating a low risk, at least at the level of simple presence *vs*. absence for these fins. Fourth, we used care at the time of annotation to match the author’s intended higher-level taxon with the taxonomy used in Phenoscape (VTO). The propagation step occurs in the context of the VTO, i.e., data are propagated from a higher-level VTO taxon to all included species. However, we recognize the risk of propagating the data to species unintended by the original author because of changes in taxonomic inclusion (Franz et al. 2015). For example, Labridae, according to Nelson (2006), possess pectoral fins and consist of about 68 genera and 453 species. However, Labridae of the VTO (derived from CoF; Midford et al. 2013), consist of 70 genera and 526 species. Because ‘pectoral fin present’ was annotated to ‘Labridae’ as per the VTO, this trait was linked to each of these 526 species following propagation. These data are now applied to species not considered by the original author, thus clearly incurring a risk that they may be incorrect.

We discovered that after propagation, nearly double the number of species have data for the pectoral fin *vs.* the pelvic fin (Fig. 4), though before propagation, the number is similar between fins. This is likely because the number of families and genera from which data are propagated is higher for the pectoral fin (151 families, 97 genera) *vs.* pelvic fin (90 families, 71 genera). In addition, most of the families in the VTO to which pectoral fin data are annotated are more speciose (e.g., Loricariidae, 899 species; Labridae, 526 species) than those for which there are pelvic fin data (e.g., Congridae: 214 species; Synodontidae: 71 species).

The question of confidence arises with respect to data that are generated through inference. Clearly a high level of certainty can be associated with direct author statements concerning morphological features at the species level, particularly if they are associated with voucher specimens. For example, the asserted absence of the pelvic fin and girdle in the knifefish (*Apteronotus apurensis*,Albert 2001) comes from the author’s observations on voucher specimens that are housed in various collections, e.g., the Field Museum of Natural History (FMNH). The specimen lot numbers provide access to the specific individuals on which the observations were made (e.g., FMNH 85499, FMNH 100738). A similar level of confidence can be associated with the inferred presence or absence of morphological features, particularly if inferred from an observation on vouchered material. For instance, the presence of a pelvic fin in the armored catfish (*Acanthicus hystrix*)was inferred based on the observation by Armbruster (2004) on vouchered specimens that ‘two rows of the first pelvic-fin ray are fused’ in this species. Even though the author did not directly state that the pelvic fin is present, the logic leading to an inference of presence is based on their reproducible observation.

Though large-scale molecular phylogenies are increasingly obtainable, they are rarely available at the scale required by this study (38,000+ tips). Supertree approaches are required to synthesize previous phylogenetic trees. The Open Tree supertree pipeline yielded a phylogeny that to our knowledge is the most comprehensive synthetic tree ever assembled for teleosts, harvesting 200 source phylogenies (Supplementary Materials File 2). In contrast, a recent large supertree assembled for a study on basal vertebrates relied on 118 source trees (38 for teleosts) and had only 2,730 tips (Larouche et al. 2017). The Open Tree approach affords access to all source trees, node-based provenance, and brings in species that may not have been included in source trees through a reference taxonomy. The latter point was critical in this work, because many of the species with trait data had not previously been studied phylogenetically, thus not available in source trees.

*Opportunities, Challenges, and Future Directions in Integrating Traits and Trees*

Two developments made the opportunity ripe for this study: huge phylogenies with provenance, and computable traits with tools for aggregation into a supermatrix. Taxonomic reconciliation, i.e., matching taxa from different sources, was the primary challenge involved in merging the trait data from the KB with the phylogenies provided by Open Tree. Neither Phenoscape nor Open Tree is attempting to develop a taxonomic standard; both are harvesting a subset of the available resources. The sources of differences in the case of teleost fishes are the lack of incorporation of a source of extinct taxa and the expert fish reference taxonomy (CoF). Future unification of disparate sources by the broader taxonomic community will reduce some reconciliation issues.

Taxonomic reconciliation across multiple data sources is an active research area, and current methods frequently use the taxon name as the integrative unit (Patterson 2003). This introduces several challenges, such as resolving synonyms, abbreviations, misspellings, and handling improper naming syntax (Cranston et al. 2014; Patterson et al. 2016). Moreover, homonyms can exist when a single taxonomic name belongs to multiple tips within the same taxonomy (Rees and Cranston 2017). Available solutions include using online servers that perform name resolution, such as Taxonomic Name Resolution Service (TNRS), which act as scientific name repositories that aggregate data from different sources (Boyle et al. 2013), and the use of software like the toolkit distributed by Global Names Architecture (GNA; Patterson et al. 2016). The Open Tree integrates multiple source taxonomies to build the Open Tree taxonomy, and names are matched from individual taxonomic sources to identify the correct taxonomic name (Redelings and Holder 2017; Rees and Cranston 2017). However, these solutions do not support the VTO, which thus required us to develop another method for efficient reconciliation. Taxonomy ID can be used as an alternative for taxon name (Thomson and Shaffer 2010). However, depending solely on NCBI taxonomy IDs for reconciliation was inefficient because a large number of VTO taxa (9,522) in the propagated matrix did not have any reference to NCBI taxonomy IDs. Therefore, two reconciliation methods—based on taxon name and NCBI taxonomy IDs—were integrated in this work. Taxonomic reconciliation at large scale, however, remains a major challenge.

A second major technical challenge in this work was the lack of tools that support visualization and manipulation of trees at this large scale. Branch navigation on large-scale phylogenies is cumbersome, and manual efforts to investigate state changes along branches are difficult. The next generation of tools must facilitate navigation to specific nodes and support more complex analyses (Gruenstaeudl 2016).

Additional challenges include the substantial level of curation required to develop and maintain resources for phylogenetic trees and ontology-based trait data. Curation of trait data is manual and time-consuming (Dahdul et al. 2015), as is the curation of phylogenies for Open Tree. The addition of new natural language processing (NLP) machine-learning curation methods is critical. The approach here clearly demonstrates the value of integrating these data, and it highlights the need for automated tools.

Finally, the methods and pipeline demonstrated here to integrate a large-scale morphological supermatrix from the Phenoscape KB with a Teleostei species-level tree from the Open Tree of Life are currently limited in generalizability. Currently the pipeline functions for a matrix that includes only the pectoral and pelvic fins, though the methods and algorithms in the code can be adapted for other characters. A future goal of this work is to develop a more generic pipeline to integrate any large-scale morphological dataset coming from the Phenoscape KB with any large-scale phylogeny.

*Data Conflicts Motivate Future Studies*

One of the benefits of machine reasoning, as previously pointed out (Dececchi et al. 2015), is that conflicts in the data are automatically isolated. That is, cells where both presence and absence are indicated for a particular trait in a single taxon are made obvious and data provenance is available. Conflicts may result from differing assertions among authors, which may in turn be due to observations of different specimens or different interpretations of the same material. There may also be conflicts between asserted and inferred data, and these were the most common type in the study by Dececchi et al. (2015). In the supermatrix generated here, only 0.04% of the species-level data (6 of 16,408 populated cells) were conflicted, excluding actual polymorphisms. All six conflicts were between asserted and inferred data. For example, one author documented the thickness of the first pelvic fin ray in the airbreathing catfish (*Gymnallabes typus*, De Pinna 1993), from which the presence of a pelvic fin was inferred, whereas another author asserted that the pelvic fin was absent for this species (Nelson 2006). In another example, Poulsen (2015) observed ‘transparent pectoral fin lobes’ in the bobtail snipe eel (*Neocyema erythrosoma*), but noted the apparent absence of the pectoral skeleton. The annotation ‘pectoral girdle skeleton, absent’, however, from which the absence of a pectoral fin is inferred (Fig. 3), results in a conflict with the same author’s assertion of pectoral fin presence. The inference reflects the anatomical knowledge formalized in the ontology, that the presence of a pectoral fin is dependent on a pectoral girdle (Fig. 2). In fact, the presence of the fin without the underlying girdle has not been reported in any other species. This conflict, like all of the conflicts surfaced through such automated syntheses, leads back to an examination of the evidence. Here we discovered that only five specimens of this rare species have ever been collected (DeVaney et al. 2009; Poulsen 2015), and absence of the pectoral skeleton was described as ‘apparent’ (Poulsen 2015); these raise the possibility that the pectoral girdle may in fact be present. On the other hand, *Neocyema* is a member of a group of deep-sea fishes (Saccopharyngiformes) with many reduced skeletal features, and perhaps the loss of the pectoral girdle is another instance of reduction. This particular conflict, like the others, targets species and features for further investigation.

*How Often Were the Paired Fins Lost (and Regained)?*

The questions of how often, and in which fishes, paired fins were lost have driven many studies in ichthyology, and our work not only demonstrates how these can be answered automatically and scalably, but it also provides species-level data from disparate sources in a phylogenetic framework to provide insight into fin evolution at the macroevolutionary scale. Previous investigators have concluded that pelvic fins appear to be more readily lost or reduced across teleosts than any other fins (Nelson 1990; Yamanoue et al. 2008; Larouche et al. 2017) and possibly with greater frequency than other structures, including for example, scales (Nelson 1990) and the gas bladder (McCune and Carlson 2004). The expectation that pectoral fin loss is much rarer than pelvic fin loss (Yamanoue et al. 2010) was born out by our analysis, which indicates more than double the minimum number of independent losses of the pelvic fin (48) compared to the pectoral fin (19) for the phylogeny under consideration (Table 2).

The apparent ease of pelvic fin loss has been examined from a genetic standpoint in several species, including famously the three-spined stickleback (*Gasterosteus aculeatus*; Chan et al. 2010), the Japanese pufferfish (*Takifugu rubripes*; Tanaka et al. 2005), and the tiger tail seahorse (*Hippocampus comes*; Lin et al. 2016). These studies showed different genes or enhancers responsible for pelvic fin loss, indicating that similar phenotypes can evolve by different mechanisms. Studies focused on potential genetic differences that may underlie the difference in frequency of pectoral *vs*. pelvic fin loss, however, have not yet been done.

With respect to pectoral and pelvic fin evolution, there are contradictory expectations as to whether reversal following loss has (Yamanoue et al. 2010) or has not (Nelson 1990) occurred. There is a large literature on the irreversibility of evolution and Dollo’s law (Gould 1970; Farris 1977; Wagner 1982; Kohlsdorf and Wagner 2006; Klimov and Oconnor 2013), yet also examples of putative reversals, e.g., digit regain within multiple lineages of squamates (Kohlsdorf and Wagner 2006). Genetic studies suggest that the persistence of developmental pathways may provide a route for reversal (Collin and Miglietta 2008).

Our analyses indicate a substantial number of reversals have occurred in the evolution of the paired fins (minimum of 2 for pectoral fins; 14 for pelvic fins), though interpretation is compromised by the species-level polytomies throughout the teleost tree. We sought to determine whether polytomies could be resolved such that reversal was not required, and we focused on one monophyletic clade, the eels (Anguilliformes), in which most of the losses and all of the putative regains of the pectoral fin occurred. This was a clade, in fact, where we manually curated recent phylogenies into Open Tree (Santini et al. 2013; Tang and Fielitz 2013), and the phylogeny (Fig. 5) is up-to-date and resolved with respect to the most recently published studies for Anguilliformes. The polytomies that remain, however, challenged the notion of regain. Our analysis showed that among 1,000 different randomly resolved topologies, the pectoral fins re-evolved following loss a minimum of two times: once within *Hoplunnis* (duckbill eels with nine species; Family: Nettastomatidae) and again in *Macrocephenchelys* (conger eels with two species: Family: Congridae; Fig. 7). This points to species where regain is a very strongly supported hypothesis that could be examined further using genetic tools, for example. Additionally, it points to areas in the tree for strategic resolution, i.e., regions of the tree where evolutionary questions, e.g., understanding the phylogenetic pattern and frequency of reversals and their biological basis, could be answered through further phylogenetic analysis.

CONCLUSIONS

The opportunity to understand the broad patterns of evolution of organismal features, foundational knowledge for many types of studies, is at hand, with the tools in place to reuse and synthesize the data. The time-consuming nature of curating phylogenies and trait data into the formats and databases appropriate for their automatic aggregation into larger-scale synthetic products is an immediate issue, though culture shifts with respect to data in the life science community and advances in machine learning are progressing to ameliorate it. As our case study shows, rendering traits computable enables the extension of relatively sparse data to taxa for which the presence or absence of a trait had not been directly asserted. Further, by propagating author assertions about features for high-level taxa to species for which asserted or inferred data are missing, we show that trait data can be further and significantly extended. An additional benefit from these automated approaches lies in ready discovery of errors in data or knowledge by virtue of the conflicting presence and absence for a species. Such potential errors are not easily found through manual means. Finally, a critical benefit for future use is the provenance of species-level traits and phylogenetic nodes, which automated methods enable.

Given the increasingly broad scope of comparative questions across biology and the difficulty and expense of new collection of trait data, these means of existing data reuse and extension (i.e., inference, propagation), are critically important. Equally important is ensuring that the users of these products have access to the provenance of the data, including the traits for each taxon and the phylogenetic resolution and the means by which they were generated. This case study also revealed several important technical challenges to such integration. Current difficulties in aligning taxonomic sources impede ready integration, and new tools are needed to visualize and analyze the data on large trees.

Finally, from the standpoint of understanding the relative frequencies of pectoral and pelvic fin loss and regain, this study provides evidence that the pelvic fin is independently lost in more than twice the number of lineages as the pectoral fin, and that these fins were regained several to many times in the course of teleost evolution. The general method outlined here, of an automated mapping and extension of traits mined from dozens of studies, to trees assembled from over a hundred more, offers rapid assessments of trait distribution. These in turn set the stage for in-depth analyses of the potential underlying evolutionary mechanisms.

FUNDING

This work was supported by the National Science Foundation collaborative grants

(DBI-1062404, DBI-1062542) and the following support from the University of South Dakota (USD): Graduate Academic and Creative Research Grant to L.M.J. and P.C.F, Matthew A. Johnson Memorial Fund Foundation Award to L.M.J, and the Council on Undergraduate Research and Creative Activity Award to J.S.H.

ACKNOWLEDGEMENTS

We thank F. Santini and K. Tang for providing electronic copies of the tree files from their publications (Santini et al. 2013; Tang and Fielitz 2013); these were added to Open Tree. We thank J.C. Tyler for providing permission for images used in our manuscript. We thank W.M. Dahdul and C.J. Mungall for making necessary updates to the Uberon ontology. We thank L.J. Revell, W.P. Maddison, and D.R. Maddison for providing assistance with ancestral reconstruction methods in R and Mesquite. W.M. Dahdul, T.A. Dececchi, K.A. Cranston and three anonymous reviewers provided helpful comments on this manuscript. We also thank K.A. Cranston for providing assistance with the phylogeny used in this study obtained from Open Tree. This manuscript is based on work done by P.M.M. while serving at the U.S. National Science Foundation. The views expressed in this paper do not necessarily reflect those of the National Science Foundation or the United States Government.

REFERENCES

Aceves-Medina G, Gonzalez E, Saldierna R. 1999. Larval development of *Symphurus williamsi* (Cynoglossidae: Pleuronectiformes) from the Gulf of California. Fish. Bull., 97:738-745.

Albert JS. 2001. Species diversity and phylogenetic systematics of American Knifefishes (Gymnotiformes, Teleostei). Misc. publ. - Mus. Zool., Univ. Mich., 190:1-140.

Alfaro ME, Santini F, Brock C, Alamillo H, Dornburg A, Rabosky DL, Carnevale G, Harmon LJ. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proc. Natl. Acad. Sci. U.S.A., 106:13410-13414.

Aoyama J, Nishida M, Tsukamoto K. 2001. Molecular phylogeny and evolution of the freshwater eel, genus *Anguilla*. Mol. Phylogenet. Evol., 20:450-459.

Armbruster JW. 2004. Phylogenetic relationships of the suckermouth armoured catfishes (Loricariidae) with emphasis on the Hypostominae and the Ancistrinae. Zool. J. Linn. Soc., 141:1-80.

Balhoff JP, Dahdul WM, Dececchi TA, Lapp H, Mabee PM, Vision TJ. 2014. Annotation of phenotypic diversity: decoupling data curation and ontology curation using Phenex. J. Biomed. Semantics, 5:45.

Balhoff JP, Dahdul WM, Kothari CR, Lapp H, Lundberg JG, Mabee PM, Midford PE, Westerfield M, Vision TJ. 2010. Phenex: ontological annotation of phenotypic diversity. PLoS ONE, 5:e10500.

Betancur RR, Broughton RE, Wiley EO, Carpenter K, Lopez JA, Li C, Holcroft NI, Arcila D, Sanciangco M, Cureton Ii JC, Zhang F, Buser T, Campbell MA, Ballesteros JA, Roa-Varon A, Willis S, Borden WC, Rowley T, Reneau PC, Hough DJ, Lu G, Grande T, Arratia G, Orti G. 2013. The tree of life and a new classification of bony fishes. PLoS Curr., 5: ecurrents.tol.53ba26640df0ccaee75bb165c8c26288.

Boyle B, Hopkins N, Lu Z, Raygoza Garay JA, Mozzherin D, Rees T, Matasci N, Narro ML, Piel WH, Mckay SJ, Lowry S, Freeland C, Peet RK, Enquist BJ. 2013. The taxonomic name resolution service: an online tool for automated standardization of plant names. BMC Bioinformatics, 14:16.

Britz R. 2010. A new earthworm eel of the genus *Chaudhuria* from the Ayeyarwaddy River Drainage, Myanmar (Teleostei: Synbranchiformes: Chaudhuriidae). Zootaxa, 2571:62-68.

Brown KJ, Britz R, Bills R, Rüber L, Day JJ. 2011. Pectoral fin loss in the Mastacembelidae: a new species from Lake Tanganyika. J. Zool., 284:286-293.

Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, Myers RM, Petrov D, Jónsson B, Schluter D, Bell MA, Kingsley DM. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a *Pitx1* enhancer. Science, 327:302-305.

Chapleau F. 1993. Pleuronectiform relationships: a cladistic reassessment. Bull. Mar. Sci., 52:516-540.

Coburn MM, Cavender TM. 1992. Interrelationships of North American Cyprinid Fishes. In: Mayden RL editor. Systematics, Historical Ecology, and North American Freshwater Fishes. Stanford, California, Stanford University Press, p. 328-373.

Collin R, Miglietta MP. 2008. Reversing opinions on Dollo's Law. Trends Ecol. Evol., 23:602-609.

Crampton G. 1921. An exception to Dollo's Law of the irreversibility of evolution. Science, 54:91.

Cranston KA, Harmon LJ, O'Leary MA, Lisle C. 2014. Best practices for data sharing in phylogenetic research. PLoS Curr., 6:ecurrents.tol.bf01eff04a06b60ca4825c69293dc59645.

Dahdul WM, Balhoff JP, Blackburn DC, Diehl AD, Haendel MA, Hall BK, Lapp H, Lundberg JG, Mungall CJ, Ringwald M, Segerdell E, Van Slyke CE, Vickaryous MK, Westerfield M, Mabee PM. 2012. A unified anatomy ontology of the vertebrate skeletal system. PLoS ONE, 7:e51070.

Dahdul WM, Balhoff JP, Engeman J, Grande T, Hilton EJ, Kothari C, Lapp H, Lundberg JG, Midford PE, Vision TJ, Westerfield M, Mabee PM. 2010. Evolutionary characters, phenotypes and ontologies: curating data from the systematic biology literature. PLoS ONE, 5:e10708.

Dahdul WM, Dececchi TA, Ibrahim N, Lapp H, Mabee PM. 2015. Moving the mountain: analysis of the effort required to transform comparative anatomy into computable anatomy. Database, 2015:bav040-bav040.

De Pinna MC. 1993. Higher-level Phylogeny of Siluriformes, with a New Classification of the Order (Teleostei: Ostariophysi). Ph.D. Dissertation, New York, City University of New York.

Deans AR, Lewis SE, Huala E, Anzaldo SS, Ashburner M, Balhoff JP, Blackburn DC, Blake JA, Burleigh JG, Chanet B, Cooper LD, Courtot M, Csösz S, Cui H, Dahdul W, Das S, Dececchi TA, Dettai A, Diogo R, Druzinsky RE, Dumontier M, Franz NM, Friedrich F, Gkoutos GV, Haendel M, Harmon LJ, Hayamizu TF, He Y, Hines HM, Ibrahim N, Jackson LM, Jaiswal P, James-Zorn C, Köhler S, Lecointre G, Lapp H, Lawrence CJ, Le Novère N, Lundberg JG, Macklin J, Mast AR, Midford PE, Mikó I, Mungall CJ, Oellrich A, Osumi-Sutherland D, Parkinson H, Ramírez MJ, Richter S, Robinson PN, Ruttenberg A, Schulz KS, Segerdell E, Seltmann KC, Sharkey MJ, Smith AD, Smith B, Specht CD, Squires RB, Thacker RW, Thessen A, Fernandez-Triana J, Vihinen M, Vize PD, Vogt L, Wall CE, Walls RL, Westerfeld M, Wharton RA, Wirkner CS, Woolley JB, Yoder MJ, Zorn AM, Mabee P. 2015. Finding our way through phenotypes. PLoS Biol., 13:e1002033.

Dececchi TA, Balhoff JP, Lapp H, Mabee PM. 2015. Toward synthesizing our knowledge of morphology: using ontologies and machine reasoning to extract presence/absence evolutionary phenotypes across studies. Syst. Biol., 64:936-952.

Dececchi TA, Mabee PM, Blackburn DC. 2016. Data sources for trait databases: comparing the phenomic content of monographs and evolutionary matrices. PLoS ONE, 11:e0155680.

DeVaney SC, Hartel KE, Themelis DE. 2009. The first records of *Neocyema* (Teleostei: Saccopharyngiformes) in the Western North Atlantic with comments on its relationship to *Leptocephalus holti* Schmidt 1909. Northeast. Nat., 16:409-414.

Di Dario F. 1999. Filogenia de Pristigasteroidea (Teleostei, Clupeomorpha). Ph.D. Dissertation, Departamento de Zoologia, Universidade de São Paulo, p. 1-246.

Eschmeyer W. 2013. Catalog of Fishes: Genera, Species, References. In: Eschmeyer W, Fricke R, Van der Laan R editors, California Academy of Sciences.

Farris JS. 1977. Phylogenetic analysis under Dollo's Law. Syst. Biol., 26:77-88.

Franz NM, Chen M, Yu S, Kianmajd P, Bowers S, Ludäscher B. 2015. Reasoning over taxonomic change: exploring alignments for the *Perelleschus* use case. PLoS ONE, 10:e0118247.

Gkoutos GV, Green EC, Mallon AM, Hancock JM, Davidson D. 2005. Using ontologies to describe mouse phenotypes. Genome Biol., 6:R8.

Gould SJ. 1970. Dollo on Dollo's Law: irreversibility and the status of evolutiionary laws. J. Hist. Biol., 3:189-212.

Gruenstaeudl M. 2016. WARACS: wrappers to automate the reconstruction of ancestral character states. Appl. Plant Sci., 4:1500120.

Haendel MA, Balhoff JP, Bastian FB, Blackburn DC, Blake JA, Bradford Y, Comte A, Dahdul WM, Dececchi TA, Druzinsky RE, Hayamizu TF, Ibrahim N, Lewis SE, Mabee PM, Niknejad A, Robinson-Rechavi M, Sereno PC, Mungall CJ. 2014. Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. J. Biomed. Semantics, 5.

Harmon LJ, Baumes J, Hughes C, Soberon J, Specht CD, Turner W, Lisle C, Thacker RW. 2013. Arbor: comparative analysis workflows for the Tree of Life. PLoS Curr., 5:ecurrents.tol.099161de099165eabdee099073fd099163d099121a044518dc.

Harold A. 1999. Idiacanthidae. In: Carpenter KE, Niem V editors. The Living Marine Resources of the Western Central Atlantic, FAO, Rome, p. 1914-1915.

Harris JR, Arbuckle K. 2016. Tempo and mode of the evolution of venom and poison in tetrapods. Toxins, 8:193.

Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, Crandall KA, Deng J, Drew BT, Gazis R, Gude K, Hibbett DS, Katz LA, Laughinghouse HD, I.V., McTavish EJ, Midford PE, Owen CL, Ree RH, Rees JA, Soltis DE, Williams T, Cranston KA. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc. Natl. Acad. Sci. U.S.A., 112:12764-12769.

Hunt G, Slater G. 2016. Integrating paleontological and phylogenetic approaches to macroevolution. Annu. Rev. Ecol. Evol. Syst., 47:189-213.

Johnson GD. 1986. Scombroid phylogeny: an alternative hypothesis. Bull. Mar. Sci., 39:1-41.

Johnson GD, Patterson C. 1993. Percomorph phylogeny: a survey of acanthomorphs and a new proposal. Bull. Mar. Sci., 52:554-626.

Kawahara R, Miya M, Mabuchi K, Lavoue S, Inoue JG, Satoh TP, Kawaguchi A, Nishida M. 2008. Interrelationships of the 11 gasterosteiform families (sticklebacks, pipefishes, and their relatives): a new perspective based on whole mitogenome sequences from 75 higher teleosts. Mol. Phylogenet. Evol., 46:224-236.

Klimov PB, OConnor B. 2013. Is permanent parasitism reversible?—critical evidence from early evolution of house dust mites. Syst. Biol., 62:411-423.

Kohlsdorf T, Wagner GP. 2006. Evidence for the reversibility of digit loss: a phylogenetic study of limb evolution in *Bachia* (Gymnophthalmidae: Squamata). Evolution, 60:1896-1912.

Larouche O, Zelditch ML, Cloutier R. 2017. Fin modules: an evolutionary perspective on appendage disparity in basal vertebrates. BMC Bioinformatics, 15:32.

Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics, 23:127-128.

Leunissen M. 2010. Explanation and Teleology in Aristotle's Science of Nature. New York, Cambridge University Press, p. 1-250.

Lin Q, Fan S, Zhang Y, Xu M, Zhang H, Yang Y, Lee AP, Woltering JM, Ravi V, Gunter HM, Luo W, Gao Z, Lim ZW, Qin G, Schneider RF, Wang X, Xiong P, Li G, Wang K, Min J, Zhang C, Qiu Y, Bai J, He W, Bian C, Zhang X, Shan D, Qu H, Sun Y, Gao Q, Huang L, Shi Q, Meyer A, Venkatesh B. 2016. The seahorse genome and the evolution of its specialized morphology. Nature, 540:395-399.

Lin Y-S, Tzeng C-S, Hwang J-K. 2005. Reassessment of morphological characteristics in freshwater eels (genus *Anguilla*, Anguillidae) shows congruence with molecular phylogeny estimates. Zool. Scr., 34:225-234.

Maddison W, Maddison D. 2017. Mesquite: a modular system for evolutionary analysis. Version 3.2. http://mesquiteproject.org.

McCune AR, Carlson RL. 2004. Twenty ways to lose your bladder: common natural mutants in zebrafish and widespread convergence of swim bladder loss among teleost fishes. Evol. Dev., 6:246-259.

Midford PE, Dececchi TA, Balhoff JP, Dahdul WM, Ibrahim N, Lapp H, Lundberg JG, Mabee PM, Sereno PC, Westerfield M, Vision TJ, Blackburn DC. 2013. The vertebrate taxonomy ontology: a framework for reasoning across model organism and species phenotypes. J. Biomed. Semantics, 4:34.

Mungall CJ, Gkoutos GV, Smith CL, Haendel MA, Lewis SE, Ashburner M. 2010. Integrating phenotype ontologies across multiple species. Genome Biol., 11:R2.

Mungall CJ, Gkoutos GV, Washington N, Lewis SE. 2007. Representing phenotypes in OWL. Innsbruck, Austria: OWL: Experiences and Directions: OWLED proceedings.

Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. 2012. Uberon, an integrative multi-species anatomy ontology. Genome Biol., 13:R5.

Near TJ, Eytan RI, Dornburg A, Kuhn KL, Moore JA, Davis MP, Wainwright PC, Friedman M, Smith WL. 2012. Resolution of ray-finned fish phylogeny and timing of diversification. Proc. Natl. Acad. Sci. U.S.A., 109:13698–13703.

Nelson JS. 1971. Absence of the pelvic complex in Ninespine Sticklebacks, *Pungitius pungitius*, collected in Ireland and Wood Buffalo National Park Region, Canada, with notes on meristic variation. Copeia, 1971:707-717.

Nelson JS. 1976. Fishes of the World. 1st ed. New York, Wiley Interscience.

Nelson JS. 1977. Evidence of a genetic basis for absence of the pelvic skeleton in Brook Stickleback, *Culaea inconstans*, and notes on the geographical distribution and origin of the loss. Can. J. Fish. Aquat. Sci., 34:1314-1320.

Nelson JS. 1984. Fishes of the World. 2nd ed. New York, John Wiley & Sons, Inc.

Nelson JS. 1990. Analysis of the multiple occurrence of pelvic fin absence in extant fishes. Matsya, 15/16:21-38.

Nelson JS. 1994. Fishes of the World. 3rd ed. New York, John Wiley & Sons, Inc.

Nelson JS. 2006. Fishes of the World. 4th ed. New York, John Wiley & Sons, Inc.

Noy NF, Crubezy M, Fergerson RW, Knublauch H, Tu SW, Vendetti J, Musen MA. 2003. Protégé-2000: an open-source ontology-development and knowledge-acquisition environment. AMIA Annu. Symp. Proc., 2003:953.

Ogle W. 1882. Aristotle: On the Parts of Animals. K. Paul, French & Company.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics, 20:289-290.

Patterson DJ. 2003. Progressing towards a biological names register. Nature, 422:661.

Patterson DJ, Mozzherin D, Shorthouse D, Thessen A. 2016. Challenges with using names to link digital biodiversity information. Biodivers. Data J., 4:e8080.

Poulsen J. 2015. Fifth confirmed record and North Atlantic range expansion of the rare pelagic bobtail snipe eel genus *Neocyema* (Cyematidae, Elopomorpha). Mar. Biodivers. Rec., 8.

Python reference language, version 3.4.3. Python Software Foundation.

Ramírez MJ, Coddington JA, Maddison WP, Midford PE, Prendini L, Miller J, Griswold CE, Hormiga G, Sierwald P, Scharff N, Benjamin SP, Wheeler WC. 2007. Linking of digital images to phylogenetic data matrices using a morphological ontology. Syst. Biol., 56:283-294.

Redelings BD, Holder MT. 2017. A supertree pipeline for summarizing phylogenetic and taxonomic information for millions of species. PeerJ, 5:e3058.

Rees J, Cranston K. 2017. Automated assembly of a reference taxonomy for phylogenetic data synthesis. Biodivers. Data J., 5:e12581.

Revell LJ. 2012. Phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol., 3:217-223.

Santini F, Kong X, Sorenson L, Carnevale G, Mehta RS, Alfaro ME. 2013. A multi-locus molecular timescale for the origin and diversification of eels (Order: Anguilliformes). Mol. Phylogenet. Evol., 69:884-894.

Santini F, Tyler JC. 2003. A phylogeny of the families of fossil and extant tetraodontiform fishes (Acanthomorpha, Tetraodontiformes), Upper Cretaceous to Recent. Zool. J. Linn. Soc., 139:565-617.

Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Yaschenko E, Ye J. 2009. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res., 37:D5-15.

Stewart DJ. 2013. A new species of *Arapaima* (Osteoglossomorpha: Osteoglossidae) from the Solimões River, Amazonas State, Brazil. Copeia, 2013:470-476.

Tanaka M, Hale LA, Amores A, Yan Y-L, Cresko WA, Suzuki T, Postlethwait JH. 2005. Developmental genetic basis for the evolution of pelvic fin loss in the pufferfish *Takifugu rubripes*. Dev. Biol., 281:227-239.

Tang KL, Fielitz C. 2013. Phylogeny of moray eels (Anguilliformes: Muraenidae), with a revised classification of true eels (Teleostei: Elopomorpha: Anguilliformes). Mitochondrial DNA, 24:55-66.

Teng H-Y, Lin Y-S, Tzeng C-S. 2009. A new *Anguilla* species and a reanalysis of the phylogeny of freshwater eels. Zool. Stud., 48:808-822.

Thomson RC, Shaffer HB. 2010. Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. Syst. Biol., 59:42-58.

Tyler JC. 1980. Osteology, phylogeny, and higher classification of the fishes of the order Plectognathi (Tetraodontiformes). NOAA Technical Report NMFS Circular, 434:1-422.

Tyler JC, Toole BO, Winterbottom R. 2003. Phylogeny of the genera and families of zeiform fishes, with comments on their relationships with tetraodontiforms and caproids. Smithson. Contrib. Zool., 618:1-110.

Uhen MD, Barnosky AD, Bills B, Blois J, Carrano MT, Carrasco MA, Erickson GM, Eronen JT, Fortelius M, Graham RW, Grimm EC, O'Leary MA, Mast A, Piel WH, Polly PD, Säilä LK. 2013. From card catalogs to computers: databases in vertebrate paleontology. J. Vertebr. Paleontol., 33:13-28.

Vos RA, Balhoff JP, Caravas JA, Holder MT, Lapp H, Maddison WP, Midford PE, Priyam A, Sukumaran J, Xia X, Stoltzfus A. 2012. NeXML: rich, extensible, and verifiable representation of comparative data and metadata. Syst. Biol., 61:675-689.

Wagner GP. 1982. The logical structure of irreversible systems transformations: a theorem concerning Dollo's Law and chaotic movement. J. Theor. Biol., 96:337-346.

Walls RL, Athreya B, Cooper L, Elser J, Gandolfo MA, Jaiswal P, Mungall CJ, Preece J, Rensing S, Smith B, Stevenson DW. 2012. Ontologies as integrative tools for plant science. Am. J. Bot., 99:1263-1275.

Wiley EO, Johnson GD. 2010. A Teleost Classification Based on Monophyletic Groups. In: Nelson JS, Schultze HP, Wilson MVH editors. Origin and Phylogenetic Interrelationships of Teleosts. München, Germany, Verlag Dr. Friedrich Pfeil, p. 123-182.

Yamanoue Y, Miya M, Matsuura K, Katoh M, Sakai H, Nishida M. 2008. A new perspective on phylogeny and evolution of tetraodontiform fishes (Pisces: Acanthopterygii) based on whole mitochondrial genome sequences: basal ecological diversification? BMC Evol. Biol., 8:212.

Yamanoue Y, Setiamarga DH, Matsuura K. 2010. Pelvic fins in teleosts: structure, function and evolution. J. Fish Biol., 77:1173-1208.

FIGURE CAPTIONS

Figure 1. The general workflow for performing ancestral state reconstruction using the synthetic morphological supermatrix retrieved from Phenoscape KB and the Teleostei species-level tree obtained from Open Tree.

Figure 2. Subgraph from the Uberon anatomy ontology showing the relationships of terms associated with pectoral fin and girdle. Arrows represent logical relationships: *part\_of* (blue solid), *has\_part* (blue dashed), *develops\_from* (red), *is\_a* (black).

Figure 3. Ontology-based inference of presence and absence. Left: The presence of a structure (pectoral fin) is inferred from a quality (unbranched) of its part (pectoral fin ray), as seen in *Neocyttus rhomhoidalis* (Tyler et al. 2003). Right: The absence of a pelvic girdle implies the absence of a pelvic fin and thus of a pelvic fin ray, as seen in *Acanthostracion quadricornis* (Tyler 1980).

Figure 4. Combined usage of inference and propagation extends morphological data. The bar chart shows the number of species with asserted (light gray), inferred only (dark gray), and propagated (black) data for the pectoral fin and pelvic fin. The increase of number of species with data after inference and then propagation demonstrate the importance of these steps in reducing missing data. \*Of the 8,798 species for which pectoral fin data are propagated from family and genus-level data, 5,077 are propagated from asserted data, and 3,721 are propagated from inferred data. \*\*Of the 4,072 species for which pelvic fin data are propagated from family and genus-level data, 2,906 are propagated from asserted data, and 1,166 are propagated from inferred data.

Figure 5. Visualization of pectoral fin presence (black) and absence (red) across 38,419 species of teleost fishes using an unordered parsimony method of reconstruction requiring 27 steps. Fin loss is evident in nine orders (red balls). Arrows indicate higher-level groupings: 1 = Elopomorpha; 2 = Otomorpha; 3 = Percomorphaceae.

Figure 6. Visualization of pelvic fin presence (black) and absence (red) across 38,419 species of teleost fishes using an unordered parsimony method of reconstruction requiring 80 steps. Fin loss is evident in twenty-six orders (red balls). Arrows indicate higher-level groupings: 1 = Elopomorpha; 2 = Otomorpha; 3 = Percomorphaceae.

Figure 7. Visualization of pectoral fin presence (black) and absence (red) across the 1,073 species of eels (Anguilliformes). The fully resolved phylogeny shown here is one of the 1,000 randomly resolved topologies requiring the minimum number (2) of pectoral fin regain events. Pectoral fin presence is the ancestral state for Anguilliformes; red balls represent fin loss. Black balls show the taxa in which regain of the pectoral fin has occurred. Collapsed branches represent subfamilies or genera having five or more species sharing the same character state.

SUPPLEMENTARY MATERIALS

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.v0s27.

Supplementary Materials Table 1. List of 87 publications used in constructing the synthetic supermatrix, including the number of taxa, pectoral and pelvic fin characters, and states. Studies that were specifically curated for the purpose of more fully representing the distribution of pelvic and pectoral fins and girdles conditions across teleosts are denoted by an asterisk.

Supplementary Materials Table 2. Comparison of teleost families between the Vertebrate Taxonomy Ontology (VTO), Open Tree, and Catalog of Fishes (CoF). Dash indicates family name that is not recognized within a particular source.

Supplementary Materials Table 3. Statistics for reconciliation between Vertebrate Taxonomy Ontology (VTO) and Open Tree taxa. This file lists the species with data that were mismatched during the reconciliation step, and they are separated based on the reason for the mismatch (due to species being extinct, unconventional naming, etc.).

Supplementary Materials Table 4. List of VTO teleost families that show pectoral fin absence, pelvic fin absence, or the absence of both paired fins. Families with pelvic fin absence were compared to previously documented families (Nelson 1990), and details given in footnotes.

Supplementary Materials Table 5. Ancestral state reconstruction across Anguilliformes Open Tree phylogeny (Supplementary Materials File 4) for pectoral fin gain and loss (Fig. 7).

Supplementary Materials Matrix 1. OntoTrace generated NeXML synthetic morphological supermatrix for pectoral fin and pelvic fin presence and absence, including metadata for supporting states and publication information.

Supplementary Materials Matrix 2. The tab-delimited character matrix generated after pre-processing the OntoTrace matrix (Supplementary Materials Matrix 1) by converting from NeXML format.

Supplementary Materials Matrix 3. Resulting matrix after the propagation step. The taxon names are based on the VTO, and this matrix is the input for the taxon name reconciliation step.

Supplementary Materials Matrix 4. Final output matrix of the pipeline. This tab-delimited file was read into Mesquite v3.10 for mapping onto the Teleostei species-level tree (Supplementary Materials File 1).

Supplementary Materials File 1. Teleostei species-level tree from Open Tree. Tree description file (Newick) from Open Tree synthesis used as the input into Mesquite.

Supplementary Materials File 2. Output files of synthesis tree build from Open Tree. Archive files of the output directory for each step in the synthesis tree build.

Supplementary Materials File 3. Merged tree matrix. NEXUS-formatted translation of the final output matrix (Supplementary Materials Matrix 4) merged with the Teleostei species-level tree from Open Tree (Supplementary Materials File 1). This combined matrix and tree file was used for ancestral state reconstruction in Mesquite.

Supplementary Materials File 4. Anguilliformes species-level trees based on subset from Teleostei species-level tree from Open Tree (Supplementary Materials File 1) merged with the final output matrix (Supplementary Materials Matrix 4). File includes 1,000 trees each with randomly resolved polytomies performed using the APE package in R.