

# Phenoscape: Ontologies for Large Multi-species Phenotype Datasets

Peter Midford<sup>1</sup>, Paula Mabee<sup>2</sup>, Todd Vision<sup>3,4</sup>, Hilmar Lapp<sup>3</sup>, Jim Balhoff<sup>3,4</sup>, Wasila Dahdul<sup>2,3</sup>, Cartik Kothari<sup>3,4</sup>, John Lundberg<sup>5</sup>, Monte Westerfield<sup>6</sup>

<sup>1</sup>University of Kansas; <sup>2</sup>The University of South Dakota; <sup>3</sup>National Evolutionary Synthesis Center (NESCent); <sup>4</sup>University of North Carolina at Chapel Hill; <sup>5</sup>Academy of Natural Sciences, Dept. Ichthyology; <sup>6</sup>University of Oregon, Zebrafish Information Network (ZFIN)

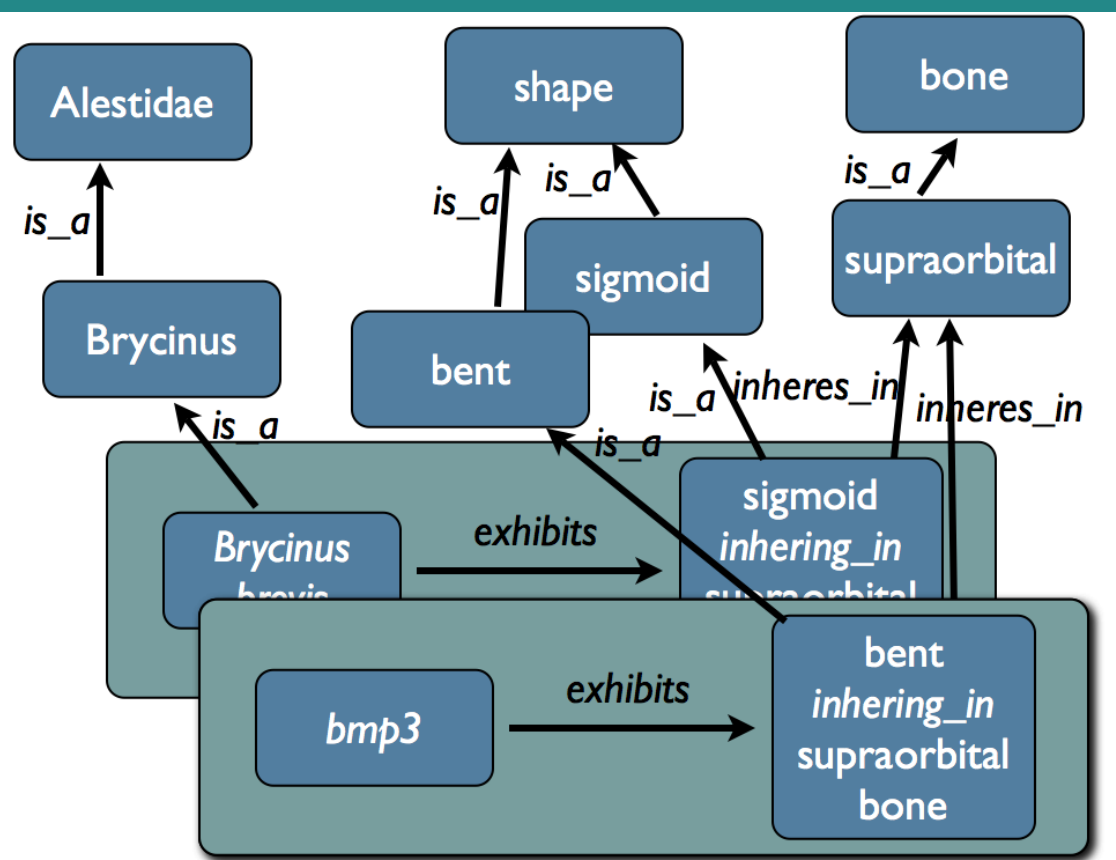


## Introduction

The Phenoscape project has developed ontologies and tools to integrate morphological and genomic data. This integration allows us to address comparative questions in evolutionary biology. We have curated 81 publications, describing nearly 6,000 phenotypic characters in 4,000 species of teleost fish. Our database of ontology-based annotations is now available via a web-based interface at <http://kb.phenoscape.org>.

Until recently, biological ontologies have either focused on single model organisms, or, like the Gene Ontology, attempted to span the tree of life. Phenoscape is a project to develop ontologies and a database to describe the phenotypes of the Osteichthyes, a group of >9,000 species of teleost fish. Among the Osteichthyes is the Zebrafish, an established model organism, frequently used in developmental biology. Because Phenoscape shares anatomical terms with the anatomy ontology maintained by the Zebrafish Information Network ([zfin.org](http://zfin.org)), we will be able to apply reasoning to queries over zebrafish mutant phenotypes and "evolutionary" phenotypes across the Osteichthyes.

Examples of such queries include finding candidate genes underlying the evolution of morphological characters and searches to discover similar phenotypes among different taxa.



Above: Graph representing similarity of phenotypes between the species *Brycinus brevius* and the *Bmp3* mutant.

## Ontologies

### Teleost Anatomy Ontology (TAO)

The Teleost Anatomy Ontology (TAO) is a multispecies anatomy ontology for teleost fishes that we initiated in September 2007. It started as a clone of the zebrafish anatomy ontology in which we still synchronize regularly. Development of the TAO is currently centered on the skeletal system, since it is often the focus of evolutionary studies in ichthyology as well as genetic studies in zebrafish. Since inception the number of skeletal terms has grown from 253 to 618.

### Phenotype Ontology (PATO)

The Phenotype and Trait Ontology (PATO) is an ontology of attributes and qualities that is used by multiple OBO projects to annotate phenotypes.

### Other OBO Ontologies

Phenoscape uses several other OBO Foundry ontologies, including the relations ontology (OBO\_REL), the spatial ontology (BSPO), the evidence codes ontology (ECO), to which we added codes for inferred homology, and the biological process portion of the Gene Ontology.

### Teleost Taxonomy Ontology (TTO)

The Teleost Taxonomy Ontology (TTO) is a tree of taxonomic terms, connected by the relation *is\_a*, and a set of rank terms (e.g., order, family, etc.), which are associated with taxonomic terms with a special *has\_rank* relation. This ontology is generated from tables exported from the Catalog of Fishes (COF), an expert database. We have also added, via manual curation, taxonomic updates from several area experts in Siluriformes, Characiformes, and Cypriniformes as well as several fossil taxa. This ontology currently contains 30,804 species and 36,785 total terms.

## Curation of evolutionary literature

### Evolutionary phenotypes (Characters and states)

The list of characters, the analysis of certain morphological characters, and the phylogenetic relationships of certain teleosts are based on the features listed below. [0] represents the phenotypic character state and [1] [2] [3] and [4] the apomorphic character states. The outgroup used to polarize characters includes *Watusianus eugrazoides*, *Amia calva*, *Lepisosteus* spp., and others in different analyses. With the exceptions indicated, characters 1 to 167 are from ARRATA (1991, 1996b, 1997) or are new characters. Because of the use of different outgroups, characters 26, 27, 28, 36, 76, 77, 78, 92, 122, 124, 125, 126, 128, 129, 130, 137, 140, and 157 changed their polarization with respect to ARRATA (1996b, 1997), and in other cases, the presentation of some characters was slightly modified (indicated below). Characters 168-199 are from GILBERT & REMS (1998); characters 176 to 191 are taken from PINNA (1996); and characters 192 to 196 are from BRUCE (1999).

- Ethmoplastine ossification in the floor of nasal capsule articulating with autopalatine: [0] absent; [1] present. (PATTERSON & ROSEN, 1977)
- Two paired endoskeletal ethmoidal ossifications: [0] absent; [1] present.
- Parietal bones fused in a median element: [0] absent; [1] present.

### Phenotype to taxon mapping (matrix)

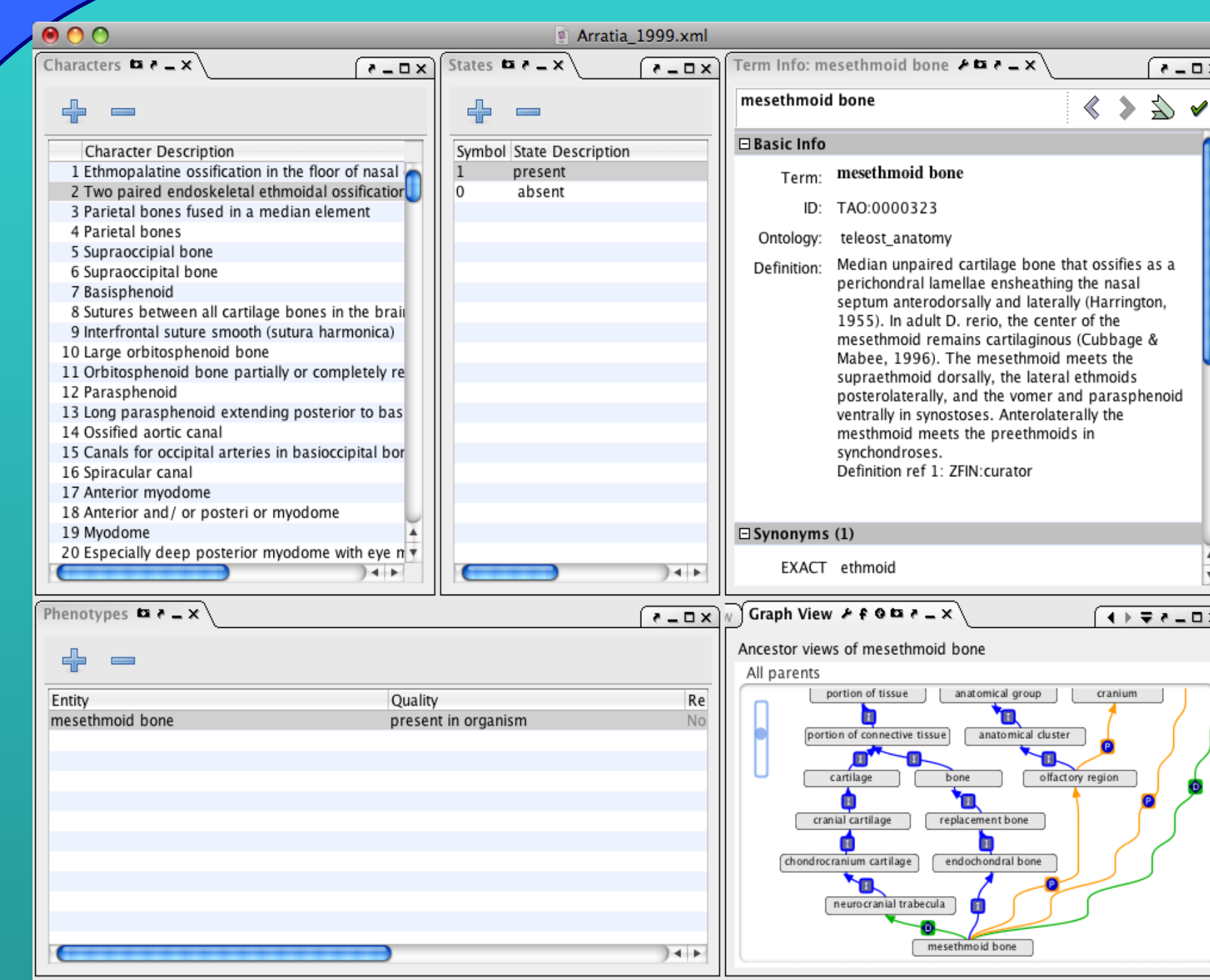
	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45
1. <i>Albula vulpes</i>	1121	2100	0007	2700	0001	0120	0001	1017	0010
2. <i>Amia calva</i>	0020	0000	0000	0000	0000	0001	0001	2000	0000
3. <i>Ameletus</i>	0000	0000	0007	0707	0007	0010	0001	2007	0000
4. <i>Ameletus</i>	0001	2107	0007	2707	0001	0110	0001	1000	0001
5. <i>Ameletus</i>	0003	2110	0107	2117	0007	0107	0700	1017	0000
6. <i>Aspidiichthys</i>	7002	0710	0707	7070	0007	0000	0000	1007	0000
7. <i>Deltentosteus</i>	0020	0711	0007	7070	0000	0000	0000	1007	0000
8. <i>Chanos</i>	0011	2100	0101	2100	0001	0101	0101	1001	1000
9. <i>Dipodomys</i>	7100	0711	1107	7170	0000	0010	0000	1007	0070

### Taxa (extant and extinct)

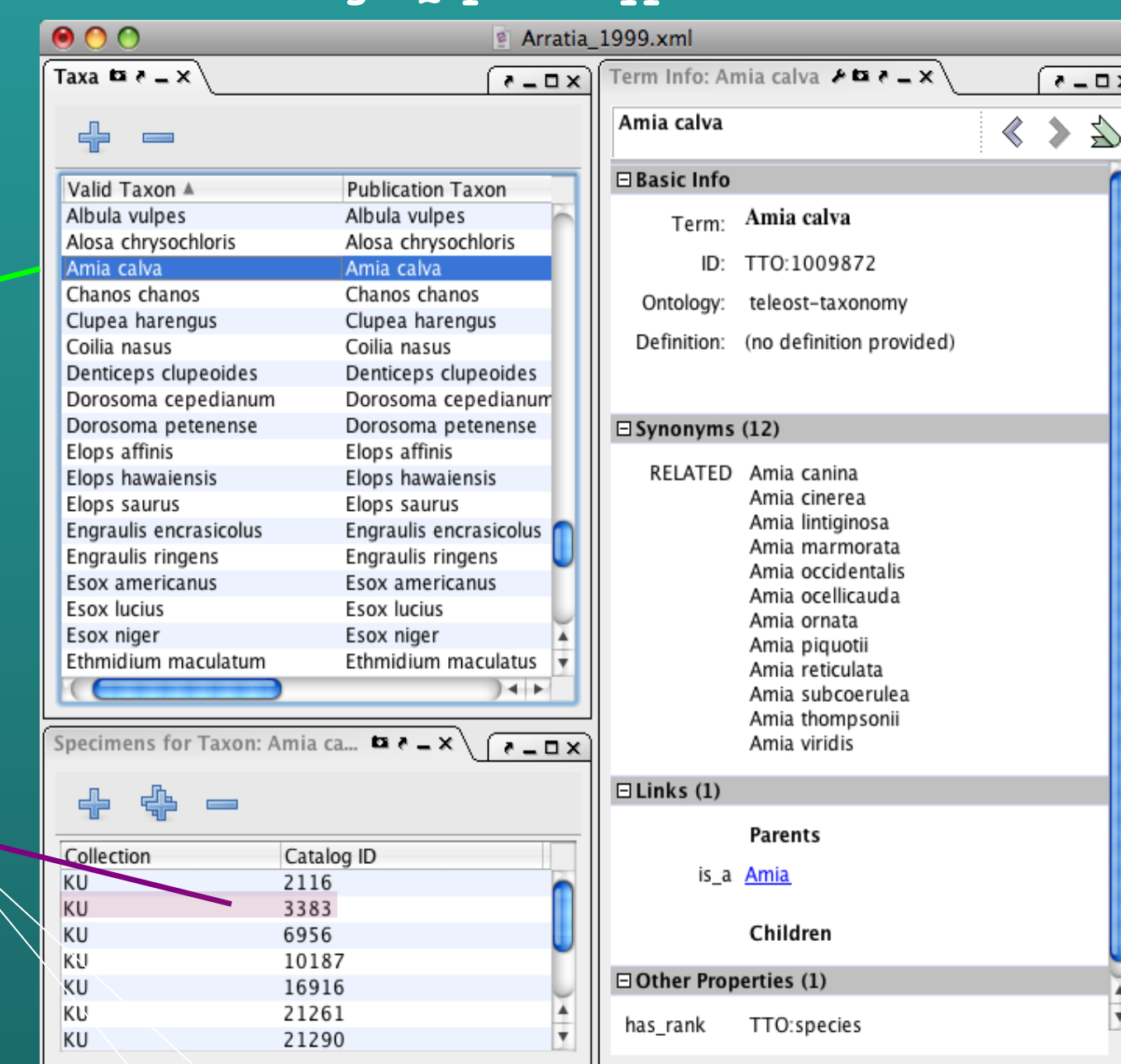
Outgroup and non-teleost taxa  
*Amia calva*: KU 2116, one alc. spec; KU 3383, one cl. & st. spec; KU 6956, one alc. spec; KU 10187, one alc. spec; KU 16916, two alc. spec; KU 21261, one dry sk; KU 21290, four cl. & st. spec; KU 21290, four cl. & st. spec; KU 21337, one dry sk; KU 21338, one dry sk; KU 21340, one cl. & st. spec; *Ameletus maculatus*: KU 21337, one dry sk; KU 21372, one dry sk; *Albula vulpes*: KU 18537, KU 18540, KU 18545, and KU 18548, dry crania.  
*Amia calva*: KU 2116, one alc. spec; KU 21261, one dry sk; KU 21290, four cl. & st. spec; KU 21337, one alc. spec; KU 16916, two alc. spec; KU 21261, one dry sk; KU 21290, four cl. & st. spec; KU 21337, one dry sk; KU 21338, one dry sk; KU 21340, one cl. & st. spec; *Amia calva*: KU 2116, one alc. spec; KU 3383, one cl. & st. spec; KU 6956, one alc. spec; KU 10187, one alc. spec; KU 16916, two alc. spec; KU 21261, one dry sk; KU 21290, four cl. & st. spec; KU 21337, one dry sk; KU 21338, one dry sk; KU 21340, one cl. & st. spec; *Ameletus maculatus*: KU 21337, one dry sk; KU 21372, one dry sk; *Albula vulpes*: KU 18537, KU 18540, KU 18545, and KU 18548, dry crania.  
*Catarrus elongatus*: MB. 13651, *Catarrus farratus*: MB. 12901, *Catarrus brevisolatus*: MB. 13849 and MB. 13850, *Catarrus* sp.: MB. 623384, *Catarrus* sp.: MB. 13848, *Depressum phylidatum*: MB. 13949, MB. 13950, and MB. 13951, *Lepisosteus oculatus*: KU 11163, one alc. spec; KU 21220, one dry sk; KU 21270, one dry sk; *Lepisosteus oculatus*: KU 11163, one alc. spec; KU 16915, one alc. spec; KU 1724, one dry sk; KU 2544, one alc. spec; KU 3651, five cl. & st. spec; KU 3677, one cl. & st. spec; KU 6958, one alc. spec; KU 8530, one alc. spec; and one cl. & st. spec; KU 12645, one cl. & st. spec; KU 16246, twelve cl. & st. spec; KU 18476, one alc. spec; KU 20597, two alc. spec; KU 22216, two cl. & st. spec; *Lepisosteus platostomus*: KU 100337, one cl. & st. spec; KU 100338, one cl. & st. spec; KU 16142, one cl. & st. spec; KU 22000, one cl. & st. spec.  
*Gyrodactylus leugnensis*: MB. 13339, *Miosaur* sp.: JM uncat. (from Westerhoff), *Neprosynsira postulata*: MB. 13774, *Phalacrolatum goodii*: JM GFR89, *Pseudocentrus*: MCSNJ0 P181, MCSNJ0 P183, MCSNJ0 P182, MCSNJ0 P184, MCSNJ0 P185, MCSNJ0 P186, MCSNJ0 P187, MCSNJ0 P188, MCSNJ0 P189, MCSNJ0 P190, MCSNJ0 P191, MCSNJ0 P192, MCSNJ0 P193, MCSNJ0 P194, MCSNJ0 P195, MCSNJ0 P196, MCSNJ0 P197, MCSNJ0 P198, MCSNJ0 P199, MCSNJ0 P200.

Arrata, 1999. Zoological Journal of the Linnean Society 151:691-757.

## Phenex: a tool for evolutionary data curation



Above: Entering EQ phenotypes to describe a character state in Phenex



Above: Editing taxa in Phenex by selecting from a taxonomy ontology

The unique challenges faced in annotating these evolutionary data have required the development of new resources, tools and extensions to existing ones. Phenex is a software application developed by Phenoscape and hosted by the OBO Project for the curation of evolutionary data using ontologies. Data annotated with Phenex is based on the Entity Quality (EQ) model for representing phenotypes. Phenex provides facilities for adding ontologies and easily selecting terms.

## Character/character state-oriented Phenex interface

- Browse the list of characters, character states and taxa directly from a character matrix file
- Compose EQ phenotype descriptions – ontological descriptions of character states – using anatomy and phenotype ontologies
- Choose taxa directly from a taxonomy ontology

## Embedding EQ into traditional character data via NeXML

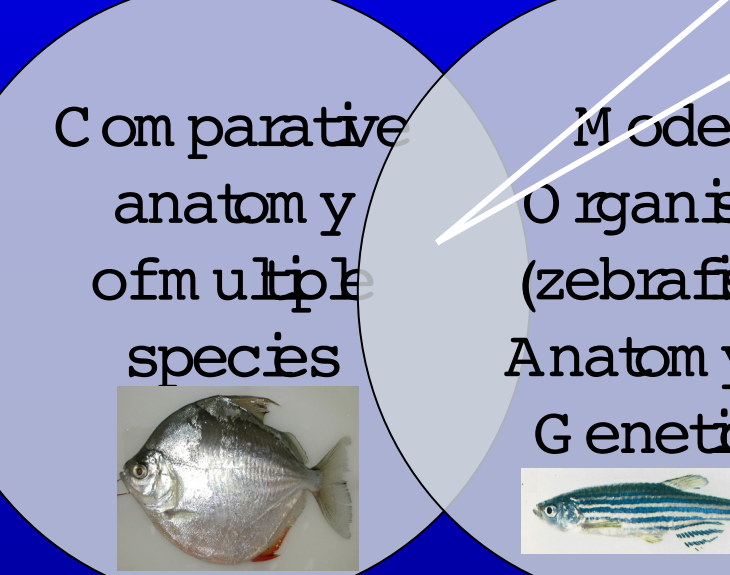
- NeXML is a new XML-based standard for the exchange of character matrices and trees. It is intended to supersede the commonly used NEXUS format. Besides following XML standards, it allows annotations which can be extracted to an emerging web standard format such as RDFa. These annotations can include custom XML extensions such as our ontology-based EQ phenotypes
- EQ data are stored using the PhenoxML format, embedded within NeXML
- Taxon TTO identifiers are stored as custom annotations to taxa in the taxa block
- Mesquite is a popular tool for editing and performing comparative analysis of data from character matrices and trees. We are testing a plug-in that allows Mesquite to view our ontology annotations to characters in NeXML files. In the future, this plug-in may support round trip editing in Mesquite as well as Phenex.



**Phenoscape knowledge-base:** This week we released our database and a web interface that supports queries to our database and enables cross-domain queries to the zebrafish (ZFIN) database

Taxon	# Papers	# Species	# Characters
Cypriniforms	8	676	794
Siluriforms	20	1724	2,110
Characiforms	10	754	1,156
Gymnotiforms	1	116	231
Gonorynchiforms	3	41	467
Clupeiforms	5	200	439
Euteleosts	3	145	582
<b>TOTAL</b>	<b>51</b>	<b>3656</b>	<b>5,779</b>

Over 400,000 phenotypes (character-taxa pairs) from the comparative morphology literature



Candidate genes for evo-devo

Over 19,000 mutant phenotypes (zfin.org)

Sharing standards across domains is critical in order to meet the needs of cross-domain queries such as those studies relating evolution and development ("evo-devo"). Such queries might, e.g., identify candidate genes for evolutionary phenotypes, candidate taxa for mutant phenotypes, and identify correlations across data types

## Phenoscape database and web services

The knowledge-base is stored in a relational database using the OBD schema – which combines the ontologies and annotations into a unified set of statements. Semantics in the ontologies allow the OBD reasoner to generate additional implied annotations.

The web user interface uses publicly available web services to access the knowledge-base.

Web services

OBD Java API

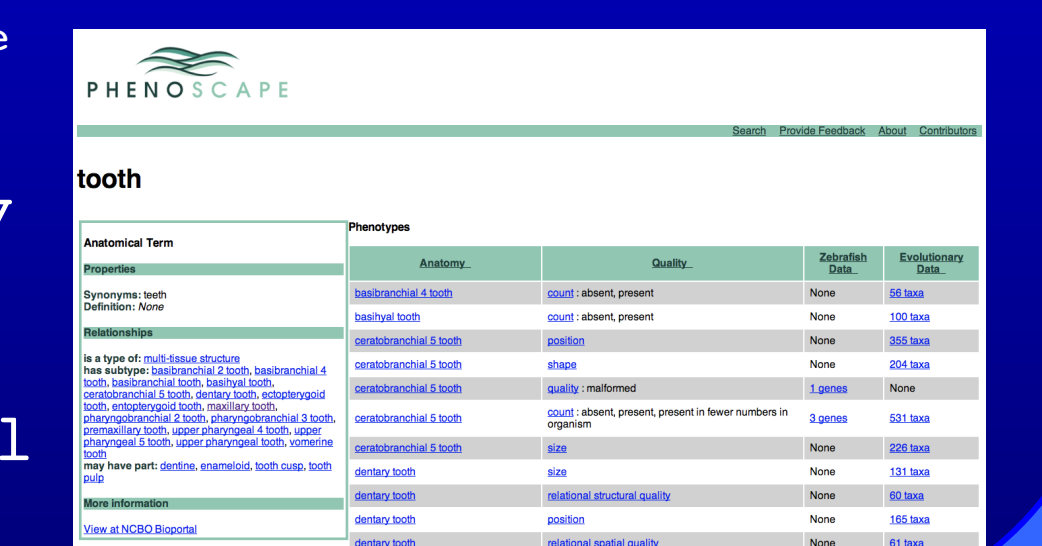
OBD database

## Phenoscape web user interface <http://kb.phenoscape.org>

The web user interface is now available!



Here you can search and browse annotations in the knowledge-base, via anatomical part, taxon, gene, or publication.



Corresponding evolutionary and model organism phenotypes are linked due to use of related anatomical and quality terms.

## Formore information:

Homepage: <http://www.phenoscape.org>  
Blog: <http://blog.phenoscape.org>  
Email list: [phenoscape-discuss@lists.sourceforge.net](mailto:phenoscape-discuss@lists.sourceforge.net)

We thank NSF DBI0641025 and the National Evolutionary Synthesis Center (NESCent) NSF #EF-0423641 for funding and the many contributors to data, character curation, ontology development and ideas (see <http://kb.phenoscape.org/contributors/>)