The Phenoscape Knowledgebase: Integrating phenotypic data across taxonomy, from biodiversity to developmental genetics

James P. Balhoff[1,2], Wasila M. Dahdul[1,3], Hilmar Lapp[1], Paula M. Mabee[3], Peter E. Midford[1], Todd J. Vision[1,2], and Monte J. Westerfield[4], on behalf of the Phenoscape Project Team

[1]National Evolutionary Synthesis Center, Durham, North Carolina, USA; [2]Dept. of Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA; [3]Dept. of Biology, University of South Dakota, Vermillion, South Dakota, USA; [4]Institute of Neuroscience, University of Oregon, Eugene, Oregon, USA

Background

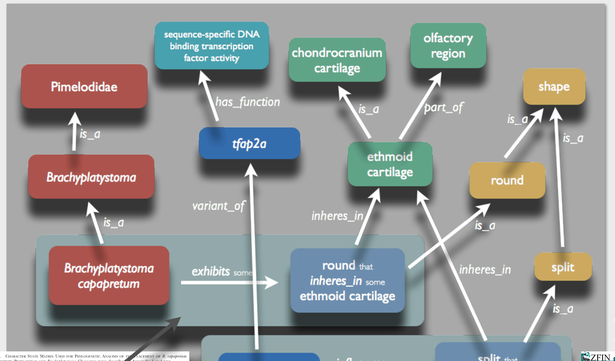
Traditionally, phenotypic differences among species have been expressed in natural language within the context of individual journal publications or monographs. As such, this rich store of phenotype data has been largely unavailable for statistical and computational comparisons across studies or integration with other biological knowledge, such as that from model organism genetics. We have created the **Phenoscape Knowledgebase**, which consists of a database and web application (http://kb.phenoscape.org/). The current version of the database integrates **ontologically annotated** phenotypic character data for a **large and diverse group of fishes** with developmental phenotypic annotations from the **ZFIN** model organism database. The web application provides query and browsing interfaces which allow users to exploit the the logical framework provided by the ontologies which underpin the data.

We are in the process of expanding the Knowledgebase taxonomically to include data from **across vertebrates**, including comparative descriptions of amphibian and archosaur taxa. We are also broadening our integration of developmental genetic data, to include mutant phenotype and gene expression information from both *Xenopus*, via **Xenbase**, and mouse, via **MGI**.

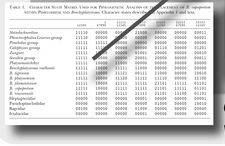
Ontology-based database and reasoner

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

Implied statements are precomputed and added to the database, facilitating more efficient runtime queries by our web application. The OBD reasoner precomputes statements implied by **class subsumption**, **property transitivity**, and **property chains**.



Mutant phenotypes from ZFIN.org



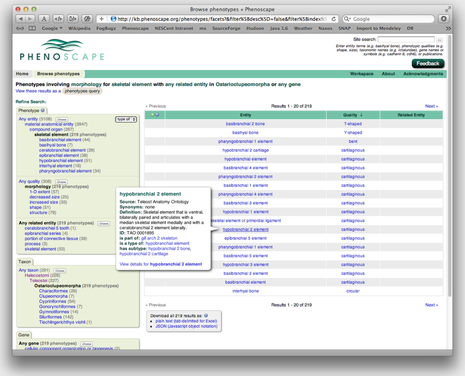
Character matrix encodes phenotypic variation for many species

Data in the KB

The Knowledgebase currently contains **561,089 phenotypic statements** about **2,510 taxa**, annotating **8390 described character states** sourced from **57 comparative systematic publications**, as well as **34,614 phenotypic statements** about **4,506 genes**, retrieved from **ZFIN**.

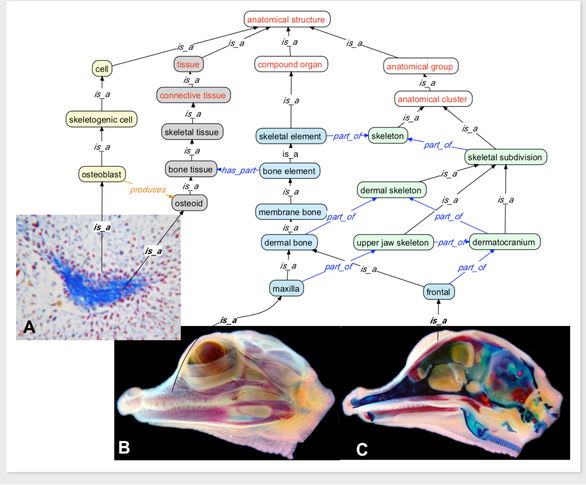
The current data focus on **ostariophysan fish** (catfish, loaches, characins, knifefishes, and minnows—including the zebrafish ***Danio rerio***).

The web interface allows users to query comparative and model organism phenotype data using reasoner-driven ontology queries.



Vertebrate skeletal system classification

Integrating data across all vertebrates will require a comprehensive ontology which presents a coherent musculoskeletal anatomical hierarchy across a wide range of taxonomic groups. Phenoscape hosted a workshop to develop the **Vertebrate Skeletal Anatomy Ontology (VSAO)**, clarifying relationships between skeletal elements, tissues, and cell types. The VSAO integrates high-level terms for cells, tissues, biological processes, skeletal elements such as bones and cartilages, and subdivisions of the musculoskeletal system across **extinct and extant** vertebrates. The VSAO is currently being integrated into the **Uberon** cross-species anatomy ontology, increasing its utility to a broad range of ontology projects.



Four high-level classes of skeletal anatomy ('cell', 'skeletal tissue', 'skeletal element', 'skeletal subdivision') and their children based on anatomical granularity. Cell terms are shown in yellow fill, tissue terms in grey fill, skeletal element terms in blue fill, and skeletal subdivision terms in green fill. Parent classes from CARO in red font. *Alligator mississippiensis* sectioned maxilla (~day 27 in ovo; Ferguson stage 19) stained with Mallory's trichrome (A); midsagittally sectioned embryonic head (day 45 in ovo; Ferguson stage 23) in lateral (B) and medial (C) view, double stained whole-mount (alizarin red and alcian blue).

In progress: migration to semantic web

In order to scale up the Knowledgebase to support a broader array of taxa for both comparative and developmental genetic data, we are transitioning to semantic web standards wherever possible. Currently in development are:

- A **reasoning pipeline** which precomputes inferences requiring **OWL DL** semantics across logically independent data subsets
- A Virtuoso **RDF triplestore backend** providing a public SPARQL endpoint and RDFS reasoning
- **Full downloads** of the entire dataset in OWL

We anticipate that adopting web standards will accelerate future developments in the Phenoscape Knowledgebase and greatly increase its potential for community reuse.

Call for collaborators

These data present a tremendous **opportunity for integration** with other data types, including large scale genomic data, etc. to address interesting questions about the evolution of phenotype. But exactly what are the approaches that will best bridge the gaps across phenotype and newly emerging data sets to lead to insights? **We are looking for participants for a small, 3-day workshop in September 2012** who are interested in engaging in creative problem-solving directed at this outstanding problem and initiating collaborations. The ideal outcome would be several collaborative projects whose goals would drive the development of the Phenoscape toolset/interface and would present new and creative ways to deepen understanding of phenotypic evolution.

We are particularly interested in a broad approach to this problem and welcome interest from scientists with backgrounds in computational and systems biology, mathematics, development, genomics, and evolution. If you are interested, please contact **Paula Mabee (pmabee@usd.edu)** or **Todd Vision (tjv@unc.edu)**.

Acknowledgments

We thank the Phenoscape team, and the many contributors to the project (http://phenoscape.org/wiki/Acknowledgments). This work has been supported by research funding from NSF (DBI-1062404 and DBI-1062542) and also the National Evolutionary Synthesis Center (NSF EF-0423641).

See also:

- http://www.phenoscape.org/
- http://kb.phenoscape.org/

NESCent PHENOSCAPE ZFIN MGI Xenbase