Applications of ontology-based knowledge representation for comparative biology

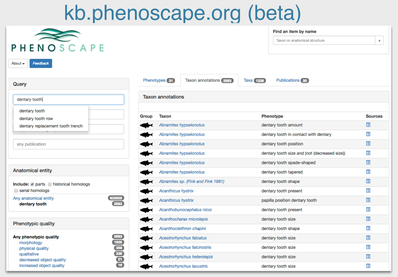
Wasila Dahdul[1], James Balhoff[2], Hilmar Lapp[3], Josef Uyeda[4], Todd Vision[5], Paula Mabee[1]
[1]University of South Dakota, Vermillion SD; [2]RENCI, Chapel Hill NC; [3]Duke University, Durham NC; [4]Virginia Tech, Blacksburg VA; [5]University of North Carolina at Chapel Hill, NC

Semantic annotation of phenotypes

Comparative descriptions of how taxa vary in phenotype are core to the study of phylogenetic relationships and trait evolution. Authors use rich, descriptive free text to describe their observations on museum specimens. Free text, however, is challenging for non-experts to parse and reuse, and on a larger scale, opaque to integrative analyses by computational methods. **The Phenoscape project (www.phenoscape.org)** has demonstrated that by annotating phenotypes with semantic terms from ontologies, links can be made between novel species phenotypes and the candidate genes that may underlie them.

Characters from the systematics literature are **annotated by attaching terms from anatomy, quality, spatial, and taxonomy ontologies to the free text descriptions** found in character-by-taxon matrices. These annotations form fully computable, logical expressions that can be queried in the **Phenoscape Knowledgebase (KB).**

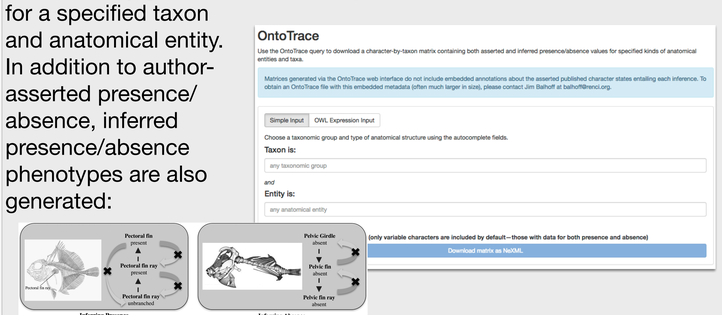
kb.phenoscape.org (beta)



The Phenoscape KB is enriched in vertebrate skeletal features, particularly for the fin and limb, and contains over 600,000 annotated phenotypes for:
- 10,497 characters (22,321 states)
- 185 publications
- 5,310 vertebrate taxa

Automated presence/absence supermatrices

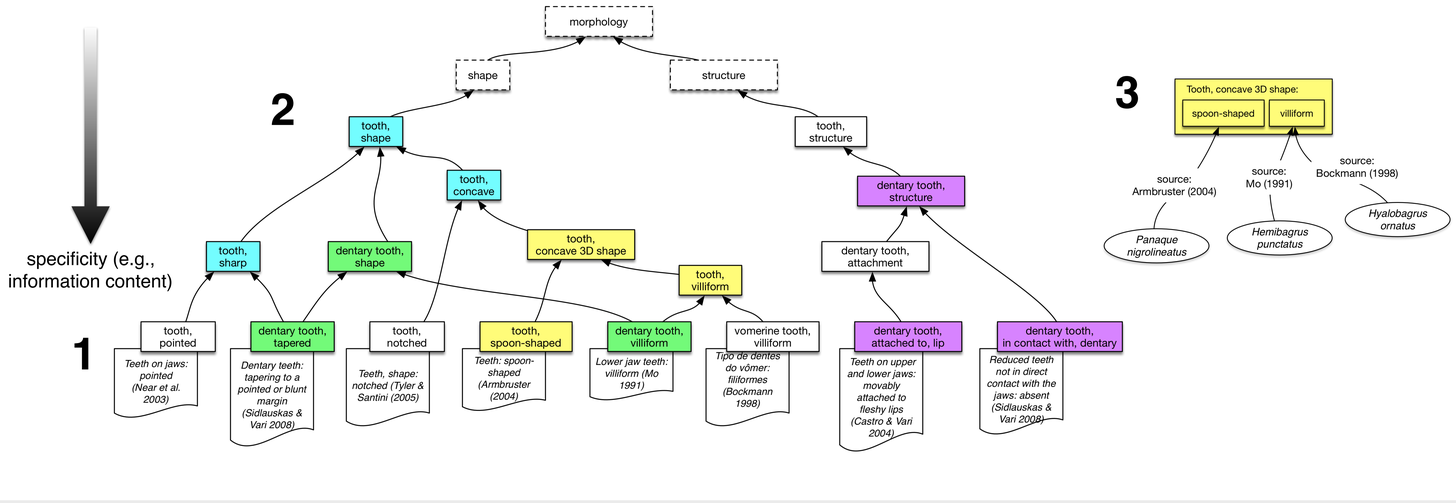
The combination of ontology-annotated data and OWL (Web Ontology Language) reasoning in the Phenoscape KB enables data aggregation and synthesis across taxa and studies. The OntoTrace tool automatically generates **synthetic presence/absence supermatrices** for a specified taxon and anatomical entity. In addition to author-asserted presence/absence, inferred presence/absence phenotypes are also generated:



Jackson et al. (2018) Sys Biol

Using machine reasoning on phenotypes to infer diverse characters and states

Phenoscape is extending the current capability of generating supermatrices for presence/absence characters to include other character types (e.g., shape, structure, position) by using **semantic similarity to aggregate phenotypes into character and states**:



1. **Published characters (text in italics) are annotated** with terms (bottom row of boxes) from anatomy and quality ontologies. These annotations are linked to the knowledge graph formed by the ontologies (boxes with dashed lines), and candidate concepts (boxes with solid lines) for characters and states are generated by an automated reasoner.

2. **Selection of candidate concepts for characters and states** (blue, green, yellow, purple boxes) can be based on whether a concept falls under an “attribute” (e.g., shape, size) in the quality ontology and evaluated based on semantic similarity metrics, such as information content.

3. **Taxa from the character matrices and materials examined lists** of the source publications are also linked to the synthetic states.

Developing tools for the semantic analysis of trait evolution

The semantic annotation of phenotypes, as shown here, can automate data synthesis across domains and individual publications and potentially generate knowledge that goes beyond that reported in the original studies.

We are now expanding the toolset in Phenoscape to enable users to **access machine reasoning in the KB for comparative trait analyses**. First, we will develop tools that incorporate the anatomical domain knowledge encoded in ontologies into **models of trait evolution** by providing metrics of trait relatedness based on their ontological semantic similarity and degree of genetic association. Second, we will adapt **concept enrichment analysis** for morphological data to allow users to address questions on the evolution of traits on trees. Specifically, new tools will be developed that identify overrepresentation of terms along the branches of a phylogeny, and incorporate semantic phenotypes in the inference of ancestral states.

Acknowledgements

We thank the many collaborators who have contributed data and expertise to Phenoscape. The Phenoscape project is supported by National Science Foundation ABI Innovation collaborative grants (1661529, 1661356, 1661456, 1661516) and an ABI Development grant (1062542).

PHENOSCAPE