Phenoscape Data Roundup - Final Report

Location: Higher Education Center, Rapid City, South Dakota (Sept. 27-29) and Sylvan Lake Lodge, South Dakota (Sept. 29-Oct.1)

Hosted by the University of South Dakota and the National Evolutionary Synthesis Center (NESCent). Supported by NSF Phenoscape funding (NSF BDI-0641025)

1. Participants

Project leaders
• Paula Mabee, pmabee@usd.edu, University of South Dakota
• Monte Westerfield, monte@uoneuro.uoregon.edu, University of Oregon
• Todd Vision, tjv@bio.unc.edu, NESCent and UNC Chapel Hill

Phenoscape personnel
• Jim Balhoff, balhoff@nescent.org, NESCent
• Wasila Dahdul, dahdul@acnatsci.org, NESCent & University of South Dakota
• Cartik R. Kothari, cartik@nescent.org, NESCent
• Hilmar Lapp, hlapp@nescent.org, NESCent
• John Lundberg, lundberg@acnatsci.org, Academy of Natural Sciences
• Peter Midford, peteremidford@yahoo.com, University of Kansas

Guest data curators
• Eric Hilton, ehilton@vims.edu, Virginia Institute of Marine Science
• Richard Mayden, cypriniformes@gmail.com, St. Louis University
• Terry Grande, TGRANDE@luc.edu, Loyola University, Chicago
• Mark Sabaj Perez, sabaj@acnatsci.org, Academy of Natural Sciences, Philadelphia

Advisors
• Judith Blake, jblake@informatics.jax.org, The Jackson Laboratory
• Suzanna Lewis, suzi@berkeleybop.org, Berkeley Bioinformatics and Ontology Project, Lawrence Berkeley Labs

Students
• Jeffrey Engeman, jeffrey.engeman@usd.edu, University of South Dakota (Biology)
• Michael Wallinga, mwalling@nwciowa.edu, University of South Dakota (Computer Science)

2. Workshop goals

Phenoscape (http://phenoscape.org) is a project (funded by the NSF Biological Databases & Informatics program) that arose from the NESCent Working Group "Towards an Integrated Database for Fish Evolution", led by Paula Mabee (a PI on the NSF Cypriniformes Tree of Life grant) and Monte Westerfield (head of the Zebrafish Information Network, zfin.org). The aim of Phenoscape is to develop tools for machine-reasoning on phenotype data from evolutionary

morphology and model organism developmental genetics (Mabee et al. 2007a, Mabee et al. 2007b), using ostariophysan fishes as a proof of principle. In its first year, the project developed customized data curation software (using Phenote as a framework, http://phenote.org), developed ontology resources (most importantly a new multi-species Teleost Anatomy Ontology and a Teleost Taxonomy Ontology), and established a curatorial workflow for annotating systematic character data using the same entity-quality syntax being used by genetic model organism databases, in particular the zebrafish database, ZFIN. By the end of the first quarter of the second year (currently), the project has finished development of a stable curation tool (Phenex), is actively curating systematic character data, is developing the database, and has prototyped one high priority use case.

The goals of the workshop were to 1) assemble and train a group of guest data curators who are expert fish morphologists, to curate the anatomical character data from the high priority ostariophysan publications; 2) test curation consistency; 3) further refine the curation workflow, and the interface of the *Phenex* tool; 4) evaluate the prototypes for the web-based user interface to the database being built; 5) solicit advice pertaining to the above from outside advisors; and 6) hold our semi-annual project personnel meeting. Phenoscape project personnel, together with two external advisors and two graduate students, were on hand to help acquaint the guest data curators with the Phenex software, concepts, workflow and tools, and to record and discuss issues arising as the guest curators curated real character data. An experiment that tested consistency of character representation among curators was conducted. This was followed by discussion and wrap-up. Concurrently, prototypes of the web user interface and workflow were introduced to pairs of workshop participants and feedback recorded. A project personnel meeting in which priorities for the short and long term were established or refined and in which advice from outside advisors was incorporated, concluded the meeting.

3. Summary of activities and discussion

Documentation for guest data curators and outside advisors was compiled on the Phenoscape wiki in preparation for the jamboree (http://phenoscape.org/wiki/Data_Jamboree_2, section "Resources"). The workshop began with introductions of participants and several presentations focused on the utility of synthetic research databases (Monte Westerfield) and the way that phenotype-genotype relationships are handled in model organism databases (Suzanna Lewis, Judy Blake). The presentations are available at http://phenoscape.org/wiki/Data_Jamboree_2/Agenda. After a brief description of the Phenoscape data policy (Todd Vision), the curation software tool, Phenex, was introduced by Jim Balhoff and followed by a hands-on curation exercise led by Wasila Dahdul involving use of Phenex and associated ontologies to annotate phenotypes (characters) using EQ syntax. Guest data curators were then paired with project personnel and began curating one of the three papers that they were assigned. These publications were either authored by the guest data curators themselves or within their area of specialty (see references). These papers had been “pre-curated” such that the taxon lists, matrices, and free-text character and state descriptions were already entered. Thus experts’ efforts were focused on Entity-Quality curation and ontology development.

Issues and suggestions pertaining to curation and the Phenex interface were collected on the project wiki.

A significant portion of curator time was spent on ontology development. Specifically, new anatomical and quality entities, together with their relationships and definitions, needed to be added to either the Teleost Anatomy Ontology or PATO before characters could be curated using EQ syntax. Curators were trained to submit new entities to the trackers for these ontologies. It was suggested that meetings of small groups of experts focused on developing certain areas of the ontology (e.g. lateral line and muscle subontologies within TAO; shape within PATO) would be a productive way to extend the ontologies and expedite EQ curation. We are now planning one or two meetings in the next three months.

The curation workflow is significantly different in Phenex, in response to issues brought up at our first data jamboree. We noted a significant improvement in curatorial efficiency. One frequent topic of discussion was the depth of annotation/granularity to which characters should be curated. One of our outside advisors (Blake) pointed out that there is a continuum between use of a structured vocabulary and free text. A reasonable guideline is to say that data specific to an individual study should be left as free-text while data that can be compared across studies should be annotated with ontologies. Many systematic characters pertain to shape, frequently with complex descriptors. These can be curated to a high level (e.g. fin:shape) or to a more granular level (e.g. fin: anterior margin rounded). The higher the level of granularity, the more often post-composition is required. The mechanics in using multiple post-composition windows in Phenex was somewhat confusing to guest curators. Another important topic involved the difficulty of not having a universal standard across systematic studies – e.g.. descriptors pertaining to size and shape cannot easily be extended across systematic studies. This issue also came up at our previous data jamboree. For comparisons of size within a study, an internal grading of character states (numbered beginning wth 1 for smallest and with higher numbers corresponding to larger sizes) was suggested for trial by John Lundberg.

Following several days of curation, we conducted an experiment to assess curation consistency among this group of curators, and to identify areas of improvement in curator training, ontology development, and software improvement. We wanted to determine how often, and for what reasons, curators choose divergent EQ conceptualizations for the same character and character states. Five curators (Engeman, Grande, Hilton, Mayden, Sabaj) used Phenex to encode EQ annotations for the same 10 character/state descriptions, and the results were compiled and reviewed immediately afterward with the group. Only two of the 10 characters was annotated identically among all curators. The reasons why the other annotations differed among curators revealed different interpretations of shape descriptors, inexperience and unfamiliarity with the ontologies and software, and lack of adequate terms in the ontologies (shape) as major hurdles towards consistency between curators. These results were discussed with the advisors in order to prioritize effort on visualization tools, ontology development and workflow and Phenex development.

Project personnel and advisors met to discuss the taxonomy ontology, taxon concepts, and intermediate synonyms with Peter Midford (Teleost Taxonomy Ontology developer). Peter has added the intermediate synonyms from the Catalog of Fishes (CoF), but many additional synonyms are present in the literature that must be added (and are on the Tracker). We discussed the need to associate synonyms to their references/publications, and this will require an OBO

request for one or more database identifiers. We would like to use the CoF publication database to generate dbxrefs rather than hunting for DOIs or generating our own, but we need to see whether CoF contains our publications.

Over a two and a half day period, pairs of participants, including most project personnel, met with Jim and Cartik for demonstrations of the web user interface. The goal is to allow users to enter the database by gene, anatomy, taxon, or publication, and to make queries such as: “find evolutionary phenotypes that match a mutant ZFIN phenotype” or “find ZFIN mutants for a set of phenotypes that differ between taxa”. Feedback and suggestions on the proposed UI were compiled on the wiki. The first demonstration of the interface is planned for the educational outreach workshop at SICB.

The project personnel conducted an all-hands meeting (with one advisor, Suzi Lewis) to review progress and prioritize and plan for activities over the coming months. Major items for discussion included database and interface development and usability testing, possibility of using OBD as the underlying database, curation priorities, community engagement activities, future project meetings, participation of project members in upcoming meetings and workshops, and planned publications. A discussion of handling homology through the "uberon" approach proposed by OBD vs. the annotation approach that we have used to date was initiated and will be continued online in the next few weeks.

4. Strategy and plans for follow-up activities

The meeting suggested major changes to the web user interface prototypes and these will be implemented in the coming months. We are planning the first usability testing of this interface at the January SICB meeting (preceding our January 6, 2009 “Evolution and Ontologies” outreach workshop). We will additionally hold a project personnel meeting at SICB. Relatively minor, but high priority, Phenex interface changes suggested by this workshop will be implemented by Balhoff in the next few months. We are planning ontology development workshops in small groups this coming year for the Teleost Anatomy Ontology (in February in concert with DeepFin, CToL, EToL, and AllCat meeting) and for the PATO (with Suzi Lewis’ group). We are planning our next outreach workshop for ASIH 2010. Curation priorities for the next 6 months will be determined in relation to our high priority use cases and experiments required for demonstration of successful queries across ZFIN and Phenoscape. The Phenoscape database requirements and possible shared use of OBD will be evaluated in relation to performance and efficiency and in relation to an agreement of how homology is represented. We plan to engage the community in a broader ontology meeting at NESCent in Spring 2009.

5. Anticipated outcomes and products

The outcomes of this workshop will be posted to the blog and sent to the Friends of Phenoscape mailing list for wider community engagement. We have introduced Phenote to several other group of evolutionary morphologists who are using ontologies (Amphib Anat and Spider Tree of Life) and we will be aiding them in adaptation of Phenex. The web interface will be released in the coming year, and we anticipate significant changes to this as it is refined by the broader community. The database will hold all the evolutionary morphology data entered by our curators

and will serve as the backend to the web interface that will enable the queries and visualization tools. We anticipate entering data into this database shortly; the curation of the 13 highest priority papers will be complete by the end of 2008. The Teleost Taxonomy Ontology is stable, though Midford is refining attribution of synonyms and continuing to add synonyms in addition to intermediate synonyms. Papers describing the highly productive past year and three months of accomplishments in terms of tools, software, ontologies and concepts are in process.

6. References

Papers for Annotation: for more information on the papers below see: http://spreadsheets.google.com/ccc?key=pTeXfTnVPxC-P1URVHbI4Qg&hl=en

1. Terry Grande (Gonorynchiformes and Ostariophysi)

De Pinna, M. C. C., and T. Grande. 2003. Ontogeny of the accessory neural arch in pristigasteroid clupeomorphs and its bearing on the homology of the otophysan claustrum (Teleostei). Copeia 2003:838-845.

Grande, T., and F. J. Poyato-Ariza. 1999. Phylogenetic relationships of fossil and Recent gonorynchiform fishes (Teleostei: Ostariophysi). Zoological Journal of the Linnean Society 125:197-238.

Poyato-Ariza, F. J. 1996. A revision of the ostariophysan fish family Chanidae, with special reference to the Mesozoic forms. Palaeo Ichthyologica 6:1-52.

Grande, T., and L. Grande. 2008. Reevaluation of the gonorynchiform genera Ramallichthys, Judeichthys and Notogoneus, with comments on the families Charitosomidae and Gonorynchidae. Mesozoic Fishes 4:295-310.

2. Richard Mayden (Cypriniformes)

Mayden, R. L. 1989. Phylogenetic studies of North American minnows, with emphasis on the genus Cyprinella (Teleostei: Cypriniformes). University of Kansas Museum of Natural History Miscellaneous Publications 80:1-189.

Cavender, T. M., and M. M. Coburn. 1992. Phylogenetic relationships of North American Cyprinidae. Pp. 293-327 in R. L. Mayden, ed. Systematics, Historical Ecology, and North American Freshwater Fishes. Stanford University Press, Stanford.

Conway, K. W., and R. L. Mayden. 2007. The gill arches of Psilorhynchus (Ostariophysi: Psilorhynchidae). Copeia 2007: 267–280

3. Eric Hilton (Clupeiformes and Ostariophysi)

Chang, Miman; Maisey, J. G. 2003. Redescription of †Ellimma branneri and †Diplomystus shengliensis, and Relationships of Some Basal Clupeomorphs. American Museum Novitates 3404:1-35.

Arratia, G. 1999. The monophyly of Teleostei and stem-group teleosts. Consensus and disagreements. Pp. 265-334 in G. Arratia, and H.-P. Schultze, eds. Mesozoic Fishes 2-Systematics and Fossil Record. Verlag Dr. F. Pfeil, Munchen.

4. Mark Sabaj Pérez (Siluriformes)

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

Mo, T. 1991. Anatomy, Relationships and Systematics of the Bagridae (Teleostei: Siluroidei) with a Hypothesis of Siluroid Phylogeny, Theses Zoologicae, 17, Koeltz, Koenigstein.

De Pinna, M. C. C. 1996. A phylogenetic analysis of the Asian catfish families Sisoridae, Akysidae, and Amblycipitidae, with a hypothesis on the relationships of the neotropical Aspredinidae (Teleostei, Ostariophysi). Fieldiana: Zoology (New Series) 84:i–iv + 1-83.

Background Reading

Mabee PM, Ashburner M, Cronk Q, Gkoutos GV, Haendel M, Segerdell E, Mungall C, and Westerfield M. Phenotype ontologies: the bridge between genomics and evolution. Trends Ecol Evol 2007 Jul; 22(7) 345-50. doi:10.1016/j.tree.2007.03.013

Mabee PM, Arratia G, Coburn M, Haendel M, Hilton EJ, Lundberg JG, Mayden RL, Rios N, and Westerfield M. Connecting evolutionary morphology to genomics using ontologies: a case study from Cypriniformes including zebrafish. J Exp Zoolog B Mol Dev Evol 2007 Jun 28. doi:10.1002/jez.b.21181

Haendel, M.A., Neuhaus, F., Osumi-Sutherland, D.S., Mabee, P.M., Mejino J.L.V., Mungall, C.J., and Smith, B. (2008) CARO - The Common Anatomy Reference Ontology. In: Albert Burger, Duncan Davidson and Richard Baldock (Editors): Anatomy Ontologies for Bioinformatics: Principles and Practice. ISBN 978-1-84628-884-5.