31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Entity/Quality-Based Logical Definitions for the Human Skeletal Phenome using PATO

Georgios V. Gkoutos, Chris Mungall, Sandra Dölken, Michael Ashburner, Suzanna Lewis, John Hancock, Paul Schofield, Sebastian Köhler, and Peter N. Robinson

Abstract— **This paper describes an approach to providing computer-interpretable logical definitions for the terms of the Human Phenotype Ontology (HPO) using PATO, the ontology of phenotypic qualities, to link terms of the HPO to the anatomic and other entities that are affected by abnormal phenotypic qualities. This approach will allow improved computerized reasoning as well as a facility to compare phenotypes between different species. The PATO mapping will also provide direct links from phenotypic abnormalities and underlying anatomic structures encoded using the Foundational Model of Anatomy, which will be a valuable resource for computational investigations of the links between anatomical components and concepts representing diseases with abnormal phenotypes and associated genes.**

Phenotypic analysis plays a fundamental role in human genetics diagnostics and research. Phenotypic descriptions in publications describing new disease genes or genotype-phenotype correlations in known syndromes have generally relied on free text descriptions of phenotypic features. In recent years, computational analysis of the spectrum of phenotypic features found in human disease, the so-called "phenome" [9], has been taken up by a number of groups for a number of goals including prioritizing candidate disease genes [8] and investigating modularity of the genetic disease-phenotype network [7]. In order to fully realize the potential of computational phenome analysis and to optimally use phenotypic information for clinical purposes, there is a clear need for a standardized way of communicating phenotypic information in medical reports, publications, and databases. It is also desirable to be able to compare phenotypic data accross species. In this report, we describe the application of the entity/quality decomposition methodology to the Human Phenotype Ontology in the realm of musculoskeletal phenotypic abnormalities.

I. The Human Phenotype Ontology

Currently, the Online Mendelian Inheritance in Man (OMIM) database [1] is the most important source of information about Mendelian genetic diseases in humans. However, the fact that OMIM does not employ a controlled vocabulary has made computational analysis of this data difficult. The Human Phenotype Ontology (HPO) was initially constructed to comprise terms corresponding to all descriptions used two or more times in OMIM as well as many descriptions used only once. In the process of constructing the HPO, synonyms were merged and domain knowledge was used to create links between the terms. In contrast to the shallow hierarchical structure of the OMIM clinical synopsis section, the leaf nodes of the HPO are typically located 5 to 10 levels deep and are organized as a directed acyclic graph (DAG) in which terms may have multiple parent terms. Terms are related to parent terms by `is_a` relationships. The HPO currently provides over 8000 terms, each of which describes a single human phenotypic abnormality. Approximately 50,000 annotations to nearly 5,000 mainly Mendelian diseases listed in OMIM are provided, and annotations to other classes of human disease such as chromosomal disorders are currently in preparation (Fig. 1) [10].

Terms in the HPO all possess names that correspond to current medical usage. The syntax of terms in the HPO follow two basic patterns. The first type contains words or terms reflecting the qualities of an entity, for example *Mitral valve prolapse* (HP:0001634). This term can be broken down into the anatomical entity term *Mitral valve* and the quality term *prolapsed*.

The second type of term in HPO contains intrinsic anatomical predicates as well as qualities, often pre-composed into a single canonical term. For instance, the term *Arachnodactyly* (HP:0001166) consists of a single word which is familiar to most physicians rather than naming a physical entity that has an abnormal quality. However, looking to the definition of the term, we find that *Arachnodactyly* refers to abnormally long and slender fingers (the word was coined from Greek words meaning spider fingers), that is, there is an abnormal quality (long and slender) of an entity (fingers).

II. PATO: An ontology of phenotypic qualities

PATO is an ontology of phenotypic qualities, intended for use in a number of applications, primarily phenotype annotation. PATO consists of a single hierarchy of qualities and currently offers 2014 terms. PATO is designed to be used in conjunction with ontologies of "quality-bearing entities", prominently including ontologies of anatomical entities such as the Foundational Model of Anatomy (FMA)

This work was supported by the DFG and the BMBF

Georgios V. Gkoutos and Michael Ashburner are with the Department of Genetics, University of Cambridge , Downing Street, Cambridge, CB2 3EH, England gg295@gen.cam.ac.uk

John Hancock is with the MRC Mammalian Genetics Unit, Harwell, England

Paul Schofield is with the Department of Anatomy, University of Cambridge, Cambridge, CB2 3EH, England

Chris Mungall and Suzanna Lewis are with the Lawrence Berkeley National Laboratory, Berkeley, California, USA

Sandra Dölken, Sebastian Köhler and Peter Robinson are with the Institute for Medical Genetics, Charité Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin Germany peter.robinson@charite.de

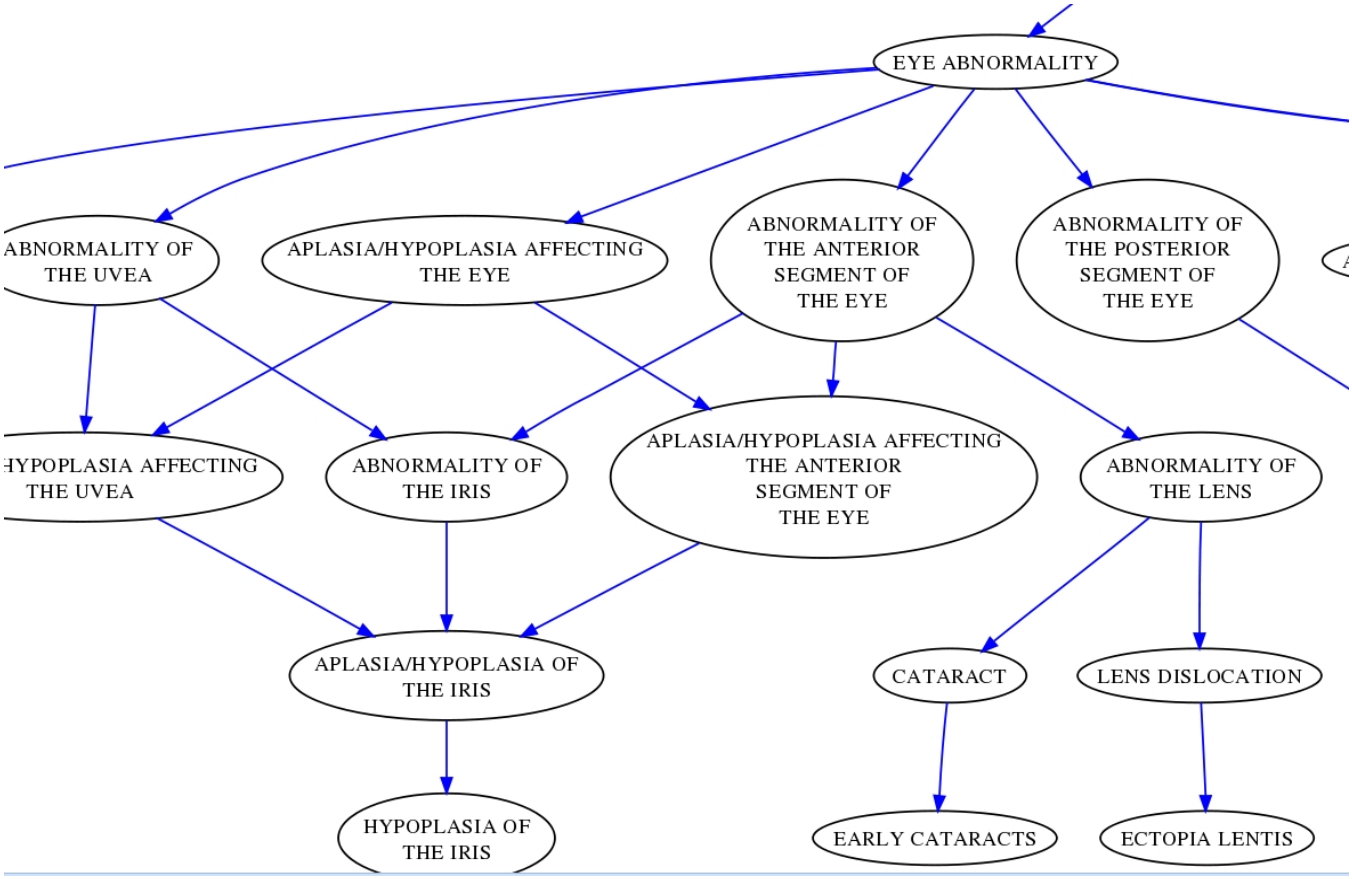

Fig. 1. An excerpt of the HPO subgraph of terms used to annotate Marfan syndrome.

ontology [12], GO [3], or the cell type ontology [4]. We can use PATO in combination with one of these ontologies to create composite terms. For instance, to describe a "red-eye" phenotype in Drosophila, we can combine the PATO term *red* with the Drosophila gross anatomy (FBbt) term "eye".

We say that we are "composing" (or "coordinating") the description by using existing terms as elements of the description. Sometimes the composition is done at the time of annotation, in which case it is referred to as "post-composition".

There are a number of advantages to associating ontology terms with EQ descriptions, as we shall see in the following sections.

III. Decomposing the Human Skeletal Phenome

There are a number of advantages to composing phenotype descriptions including the fact that it is easier to use inference algorithms if the "meanings" of the terms of an ontology are broken down into components that allow computerized reasoning as well as the ability to compare phenotypes between different species if a mapping between the respective anatomy ontologies is available. However, a serious disadvantage of this approach is the fact that a post-composed terminology does not always reflect the vocabulary of physicians and scientists involved in medical research, as was mentioned above. Therefore, the HPO has adopted a policy of retaining the medical terminology in common use for the term names, while also providing equivalence mappings to EQ-based compositional descriptions of the terms that can be used for inference or cross-species comparisons. Because of the difficulties in providing exact definitions for medical terms and defining pathophysiologically relevant semantic relationships between them, such a decomposition of HPO terms is a difficult but intellectually rewarding task that requires manual expert curation. The HPO and PATO teams are collaborating on a decomposition of the musculoskeletal subontology of the HPO as a pilot project to identify the curation and annotation strategies that will best work for decomposing the entire human phenome, which will be an ongoing project. In the following, we provide some example decompositions to illustrate our approach.

The majority of terms in the HPO describe abnormalities of anatomic structures. For these HPO terms, therefore, terms from the FMA human anatomy ontology are used to identify the affected entity (bearer of the abnormal quality). An appropriate PATO term is then chosen to describe the abnormal quality that the anatomic structure possesses, which can be described either in qualitative or quantitative terms, and are said to `inhere_in` the bearer. For instance, for the HPO term *Reduced bone mineral density* (HP:0004349),

we combine the FMA term for *bone* with the PATO term for *decreased density*. This combination is expressed in obo format notation[1] as follows:

```
[Term]
id: HP:0004349 ! Reduced bone mineral density
intersection_of: PATO:0001790 ! decreased density
intersection_of: inheres_in FMA:30317 ! bone
```

This states that the phenotype denoted by HP:0004349 is equivalent to all instances of decreased density occurring in bones.

Another example is the decomposition of the HPO term *Amyotrophy* (HP:0003202) (which is a medical term meaning "muscular atrophy" or wasting of muscles):

```
[Term]
id: HP:0003202 ! Amyotrophy
intersection_of: PATO:0001623 ! atrophied
intersection_of: inheres_in FMA:30316 ! muscle
```

The annotation model can additionally include further qualifiers concerning the developmental stage at which a phenotype holds or qualifiers indicating the expressivity of the phenotype with respect to some baseline (e.g., "abnormal"). For instance, the HPO term *Osteosclerosis* refers to an abnormal increase in bone density. The intersection of the FMA term for *bone* with the PATO term *increased density* does not faithfully reflect the meaning of the HPO term *Osteosclerosis*, because not all increased density is *abnormal*. We therefore add a further intersection with the PATO term *pathological*:

```
[Term]
id: HP:0002796 ! Osteosclerosis
intersection_of: PATO:0001788 ! increased density
intersection_of: has_quality PATO:0001869 !
        pathological
intersection_of: inheres_in FMA:30317 ! Bone
```

Once an HPO term has been logically defined in this way, other decompositions can make reference to it. For instance, the HPO term *Patchy osteosclerosis* refers to irregular areas with increased in bone density such as can be seen in some hereditary disorders such as Hypothyroidism-retardatino-dysmorphism syndrome caused by mutations in the gene encoding tubulin-specific chaperone E [MIM 241410]. Here we represent this as a "quality aggregate", using the `has_part` relation to collect the sub-components of the phenotype together:

```
[Term]
id: HP:0005686  ! Patchy osteosclerosis
intersection_of: PATO:0000001 ! quality
intersection_of: has_part PATO:0001608 ! patchy
intersection_of: has_part HP:0002796 !
        Osteosclerosis
```

There are many cases in which medical terminology does not reflect modern notions about pathophysiology or etiology of disease, meaning that any computational inference techniques that rely on natural language processing techniques to infer the meaning of a term would be doomed to failure. One example is the HPO term *Spinal muscular atrophy* (HP:0007269), which, despite the name, does not refer to atrophy affecting specifically the muscles of the spine. Rather, spinal muscular atrophy refers to muscular weakness and atrophy related to loss of the motor neurons of the spinal cord and brainstem. This has been decomposed using a *relational quality* using the FMA terms *Motor neuron* and *Spinal cord*. In some cases, this part of the "phenotype" will be observed by neuropathological analysis [2], although it may be merely inferred in other patients based on clinical findings. The inevitable result of a reduction of the number of motor neurons is amyotrophy of the innervated muscles, and this is encoded by the further intersection of the HPO term *Amyotrophy*. The full decomposition is:

```
[Term]
id: HP:0007269 ! Spinal muscular atrophy
intersection_of: PATO:0002001 ! has fewer parts of type
intersection_of: towards FMA:83617 ! motor neuron
intersection_of: inheres_in FMA:13478 ! Vertebral column
intersection_of: results_in HP:0003202 ! Amyotrophy
```

IV. Cross-Species Comparisons

The ability to manipulate the mouse genome has rendered the mouse one of the most important model organisms for studying human disease. In the gene driven approach to model discovery targeted mutants are made in specific genes associated with disease in man. In the phenotype driven approach the phenotypes of known or unknown mutations are screened for similarities with human diseases thereby gaining insight into their pathogenesis and genetic aetiology. In both cases establishment of the accurate relationship of mouse phenotypes to human diseases is essential. Although many mouse models display phenotypes that are reminiscent of the phenotypes of humans with inherited mutations at the same genes, important differences between human and mouse phenotypes resulting from mutation in homologous genes are frequently observed [11], [13]. Bridging the gap between mouse phenotypes and human diseases is therefore problematical; partly because formal disease nomenclature differs between mouse and man, but more importantly because not all of the aspects of cognate diseases will be manifested in both species. One approach to discovering a disease model is therefore to break down the summative (precomposed) diagnosis into its component parts and to search for matches within the resulting pool of constituent phenotype elements across both species. The type of decomposition presented in this paper will provide an important impetus in this direction. As an example, consider the MPO term *thymus hypoplasia* (MP:0001823) [14], which is defined as "underdevelopment or reduced size, usually due to a reduced cell number, in the thymus". This can be decomposed using PATO as the cross product of the mouse anatomy ontology term *thymus* (MA:0000142) and the PATO term *hypoplastic* (PATO:0000645), which is defined as "Underdevelopment or incomplete development of a tissue or organ". Likewise, the HPO term *Thymus hypoplasia* (HP:0000778) can be decomposed using the FMA term for *thymus* (FMA:9607)

[1] http://www.geneontology.org/GO.format.obo-1_2.shtml

and the PATO term *hypoplastic* (PATO:0000645). E/Q decompositions are being developed for mouse phenotype data [5]. Therefore, together with mappings between mouse and human anatomy (e.g., [6], [16]), the PATO E/Q decompositions developed in our project will provide a means to search for similar phenotypic abnormalities shared by mouse and human on a systematic level.

V. Conclusions

In this report, we have presented our methodology for providing an E/Q decomposition of the HPO. Currently, approximately 1000 HPO terms have been decomposed, and work on the remaining musculoskeletal terms is expected to be finished shortly. This will be an important step towards linking the HPO to well-established ontologies in the anatomy and molecular biology research communities. This will provide a consistent mapping of anatomical components to diseases and abnormal phenotypes, which will enable the use of clinical data for basic and translational computational biology research. In addition to the links to the FMA described above, we are working on linking the appropriate HPO terms to other ontologies. For instance, the HPO term *Glucosephosphate isomerase deficiency* (HP:0003290) has been decomposed as the intersection of the PATO term *decreased* (PATO:0001997) and the GO term *glucose-6-phosphate isomerase activity* (GO:0004347). Similar decompositions are being made using other ontologies, and other links to pathology ontologies [15] are planned. It is hoped that these refinements will make the HPO useful not only for human geneticists and other physicians interested in phenotypic analysis, but also to molecular biologists and bioinformaticians who are interested in incorporating the human phenotype into investigations on cellular networks and related topics. The HPO is freely available at http://www.human-phenotype-ontology.org.

VI. Acknowledgments

The authors gratefully acknowledge the contributions of Denise Horn and Stefan Mundlos towards improving the structure and terms of the musculoskeletal portion of the HPO. This work was supported by the Deutsche Forschungsgemeinschaft (DFG RO 2005/4-1, SFB 760) and the Berlin-Brandenburg Center for Regenerative Therapies (BCRT) (Bundesministerium für Bildung und Forschung, project number 0313911).

References

[1] J. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh, "McKusick's Online Mendelian Inheritance in Man (OMIM)." *Nucleic Acids Res*, vol. 37, no. Database issue, pp. D793–D796, Jan 2009. [Online]. Available: http://dx.doi.org/10.1093/nar/gkn665

[2] S. Araki, M. Hayashi, K. Tamagawa, M. Saito, S. Kato, T. Komori, Y. Sakakihara, T. Mizutani, and M. Oda, "Neuropathological analysis in spinal muscular atrophy type II." *Acta Neuropathol*, vol. 106, no. 5, pp. 441–448, Nov 2003. [Online]. Available: http://dx.doi.org/10.1007/s00401-003-0743-9

[3] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." *Nat Genet*, vol. 25, no. 1, pp. 25–29, May 2000. [Online]. Available: http://dx.doi.org/10.1038/75556

[4] J. Bard, S. Y. Rhee, and M. Ashburner, "An ontology for cell types." *Genome Biol*, vol. 6, no. 2, p. R21, 2005. [Online]. Available: http://dx.doi.org/10.1186/gb-2005-6-2-r21

[5] T. Beck, H. Morgan, A. Blake, S. Wells, J. M. Hancock, and A.-M. Mallon, "Practical application of ontologies to annotate and analyse large scale raw mouse phenotype data." *BMC Bioinformatics*, vol. 10 Suppl 5, p. S2, 2009. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-10-S5-S2

[6] O. Bodenreider, T. F. Hayamizu, M. Ringwald, S. D. Coronado, and S. Zhang, "Of mice and men: aligning mouse and human anatomies." *AMIA Annu Symp Proc*, pp. 61–65, 2005.

[7] X. Jiang, B. Liu, J. Jiang, H. Zhao, M. Fan, J. Zhang, Z. Fan, and T. Jiang, "Modularity in the genetic disease-phenotype network." *FEBS Lett*, vol. 582, no. 17, pp. 2549–2554, Jul 2008. [Online]. Available: http://dx.doi.org/10.1016/j.febslet.2008.06.023

[8] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes." *Am J Hum Genet*, vol. 82, no. 4, pp. 949–958, Apr 2008. [Online]. Available: http://dx.doi.org/10.1016/j.ajhg.2008.02.013

[9] M. Oti, M. A. Huynen, and H. G. Brunner, "Phenome connections." *Trends Genet*, vol. 24, no. 3, pp. 103–106, Mar 2008. [Online]. Available: http://dx.doi.org/10.1016/j.tig.2007.12.005

[10] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos, "The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease." *Am J Hum Genet*, vol. 83, no. 5, pp. 610–615, Nov 2008. [Online]. Available: http://dx.doi.org/10.1016/j.ajhg.2008.09.017

[11] N. Rosenthal and S. Brown, "The mouse ascending: perspectives for human-disease models." *Nat Cell Biol*, vol. 9, no. 9, pp. 993–999, Sep 2007. [Online]. Available: http://dx.doi.org/10.1038/ncb437

[12] C. Rosse and J. L. V. Mejino, "A reference ontology for biomedical informatics: the Foundational Model of Anatomy." *J Biomed Inform*, vol. 36, no. 6, pp. 478–500, Dec 2003. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2003.11.007

[13] P. Schofield and J. Sundberg, "One medicine, one pathology and the one health concept," *J Am Vet Med Ass*, vol. In press, 2009.

[14] C. L. Smith, C.-A. W. Goldsmith, and J. T. Eppig, "The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information." *Genome Biol*, vol. 6, no. 1, p. R7, 2005. [Online]. Available: http://dx.doi.org/10.1186/gb-2004-6-1-r7

[15] J. P. Sundberg, B. A. Sundberg, and P. Schofield, "Integrating mouse anatomy and pathology ontologies into a phenotyping database: tools for data capture and training." *Mamm Genome*, vol. 19, no. 6, pp. 413–419, Jun 2008. [Online]. Available: http://dx.doi.org/10.1007/s00335-008-9123-z

[16] S. Zhang and O. Bodenreider, "Experience in aligning anatomical ontologies." *Int J Semant Web Inf Syst*, vol. 3, no. 2, pp. 1–26, 2007.