GOCI: An Ontology-Driven Search and Curation Infrastructure for the NHGRI GWAS Catalog

Danielle Welter[1,], Tony Burdett[1], Lucia Hindorff[2], Heather Junkins[2], Jackie MacArthur[1] and Helen Parkinson[1]*

[1] EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, UK

[2] Office of Population Genomics, National Human Genome Research Institute, NIH, Bethesda

1 INTRODUCTION

We present the GWAS Ontology and Curation Infrastructure (GOCI), a collection of modules and features for the improvement of the curation, trait organization and querying of the NHGRI GWAS catalog (Hindorff *et al*., 2010). Augmenting the catalogue's phenotypic traits with the semantic framework of an ontology will increase the range of possible catalogue queries, facilitate the creation of a dynamic version of the iconic GWAS diagram and accelerate the integration of catalogue data with other sources. GOCI also includes a tracking system to support the complex curation process and safeguard the extremely high quality of curation the catalogue is known for.

2 ONTOLOGY

Until recently, the phenotypic traits in the GWAS catalogue were available only as an unstructured flat list partially mapped to MeSH (Rogers, 1963). In order to formalise the trait representation, the GWAS traits were integrated into the Experimental Factor Ontology (EFO) (Malone *et al.,* 2010). Representing GWAS traits in an expressive knowledge representation language like OWL will allow for much richer queries over the GWAS catalogue. By choosing an established ontology like EFO, the long-term maintenance of these terms is assured and it also provides the potential for future integration of the GWAS catalogue with other resources already consuming EFO. Much of the coverage provided by EFO meets the needs of the GWAS catalogue in describing diverse concepts ranging from diseases to measurements to complex, often context-dependent phenotypes. EFO not only contains disease categories (such as MeSH), but also phenotypic descriptions, compound treatments, and so on. EFO's policy of reuse also ensures longer term issues of integration are accounted for.

At the start of the integration process, around 20% of all GWAS traits were already described in EFO. New traits are added either by importing appropriate classes from other ontologies or, in cases where no appropriate class can be identified in a reference ontology, created directly in EFO. Full coverage of existing GWAS traits is expected to be achieved by September 2012. A provision for adding further traits to EFO in the future is under development in line with on-going development in EFO, which requires this provision for some of its other use cases as well.

3 GWAS DIAGRAM

The GWAS catalogue produces a quarterly diagram of all SNP-trait associations mapped onto their chromosomal locations. Due to the considerable increase in SNP-trait associations since the first version of the diagram and the number of different phenotypes in the colour-coded legend, it is currently almost impossible to identify traits by visually analysing the diagram, which is only available in static PDF or Powerpoint format.

A key feature of GOCI implements a novel approach for automating the creation of the GWAS diagram using an ontology and scalable vector graphics (SVG), an XML-based language for describing geometric objects. This makes it possible to create an up-to-date, dynamic diagram that can be filtered and searched at different levels of granularity and by different criteria, including trait, chromosomal region and time. It is possible to zoom in over chromosomes in order to allow users to see all SNP-trait associations for a given region. SNP-trait associations are also interactive, providing summary information on mouse-over as well as being clickable to allow the user to proceed from an association to the catalogue entry and to the publication.

4 TRACKING SYSTEM AND AUTOMATION

Finally, GOCI also contains an online tracking system to support the highly complex curation process, as well as some other support tools such as a batch loader to upload a set of SNPs to the catalog from a spreadsheet.

REFERENCES

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolio, T.A. (2009) Potential Etiologic and Functional Implications of Genome-Wide Association Loci of Human Diseases and Traits. *Proc. Natl. Acad. Sci. USA.* 106, 9362–9367.

Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A. and Parkinson, H. (2010) Modeling Sample Variables with an Experimental Factor Ontology. *Bioinformatics*. 26, 1112–1118.

Rogers, F.B. (1963) Medical Subject Headings. *Bull. Med. Libr. Assoc.* 51, 114–116.

[*] To whom correspondence should be addressed: dwelter@ebi.ac.uk