The NHGRI-EBI Genome-Wide Association Studies (GWAS) Catalog - imposing structure on unstructured data

Aoife McMahon[1], Annalisa Buniello[1], Tony Burdett[1], Maria Cerezo[1], Fiona Cunningham[1], Peggy Hall[2], Laura Harris[1], Emma Hastings[1], Lucia A. Hindorff[2], Heather Junkins[2], Jacqueline MacArthur[1], Cinzia Malangone[1], Annalisa Milano[1], Joannella Morales[1], Danielle Welter[1], Trish Whetzel[1], Helen Parkinson[1]

1) European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, UK; 2) Division of Genomic Medicine, NHGRI, NIH, Bethesda, MD, USA.

Genome-wide association studies (GWAS) have been enormously important in identifying links between genetic loci and complex traits. The GWAS Catalog (http://www.ebi.ac.uk/gwas/) is a comprehensive repository of all published human GWA studies. Currently the Catalog contains over 55,000 variant-trait associations from over 3,200 studies, covering a broad range of human diseases and traits in diverse ancestral backgrounds. The Catalog provides a public searchable repository and visual representation of these data.

Extensive manual curation of GWA studies by expert scientists ensures that the Catalog provides an accurate, comprehensive and structured summary of GWAS results. Phenotypic traits and ancestries are captured in a consistent and structured manner during the curation process. All traits are mapped to terms from the Experimental Factor Ontology (EFO), an application ontology that combines several biological domain-specific ontologies, allowing modelling of the diverse traits observed in GWAS, including tests, diseases, anatomy and anthropometry. Mapping of curated trait descriptions to EFO terms enables enriched ontology-driven search capabilities, visualisation and integration of GWAS Catalog data with other resources. The accurate characterization of ancestry is essential to interpret and integrate human genomics data and to facilitate these analyses we have developed a framework to systematically describe and represent detailed ancestry information. To enable the access and integration of curated ancestry data we are developing an ancestry-specific ontology based on our framework. Data is further enriched by annotating variants with genomic context (latest Ensembl release) to allow users to fully interpret results.

Both Catalog data and code are publicly available and are used by a growing community of scientists to identify causal variants, understand disease mechanisms and establish targets for treatment. Our structuring of the data, use of ontologies and integration with other resources allows our users to leverage the full potential of these powerful data.