Linking rare and common disease: mapping clinical disease-phenotypes to ontologies in therapeutic target validation

Sirarat Sarntivijai[1,3,], Drashtti Vasant[1,3], Gary Saunders[1,3], Patricia Bento[1,3], Daniel Gonzalez[1,3], Joanna Betts[2,3], Samiul Hasan[2,3], Gautier Koscielny[2,3], Ian Dunham[1,3], Helen Parkinson[1] and James Malone[1,3]*

[1]European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, [2]GSK, Medicine Research Center, Stevenage, SG1 2NY, [3]Centre for Therapeutic Target Validation, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD United Kingdom

ABSTRACT

Motivation: The Centre for Therapeutic Target Validation (CTTV – http://www.targetvalidation.org) was established to generate evidence from genome-scale experiments and analysis to support the validity of therapeutic targets by integrating existing and newly-generated data, and mapping to the Experimental Factor Ontology (EFO), including mapping to diseases and phenotypes. The EFO is an application ontology that reuses and integrates parts of reference ontology resources. Here we present solutions to large-scale annotation-ontology mapping in the CTTV knowledgebase. A generic association model, 'OBAN', is proposed as a mechanism for rare-to-common disease mapping via phenotype association modeling.

1 INTRODUCTION

Drug discovery research involves many analytical activities with many sources of data. One important aspect is understanding the relationships between drug targets and phenotypes that may be modulated (Ma'ayan et al. 2014). The integration of disease and phenotypic information becomes increasingly important when considering rare diseases where research is typically fragmented across data type and disease, failing to integrate across rare and common disease information, clinical phenotypes and other -omics data (Thompson et al. 2014).

The Centre for Therapeutic Target Validation (CTTV) is a collaboration between the European Bioinformatics Institute (EMBL-EBI), GSK and the Wellcome Trust Sanger Institute (WTSI) to develop evidence for drug targets based on genomic experiments and bioinformatics analyses. One CTTV goal is to develop a better understanding of the rare and common disease relationship via shared phenotypes, genes and pathways. This requires integration of newly generated data, and data residing in EMBL-EBI, WTSI and GSK database. Data types include variants, genes, proteins, gene expression, pathways, compounds and related experimental variables such as disease and phenotype.

CTTV selected the Experimental Factor Ontology (EFO) (Malone et al, 2010) as its application ontology to ensure consistency within the data to be included in the CTTV platform. EFO reuses many parts of domain-specific ontologies such as Orphanet Rare Disease Ontology (ORDO), ChEBI, Gene Ontology and Uberon where these resources meet our data annotation and knowledge representation needs, filling in gaps where they do not. EFO classes from external ontologies are imported using MIREOT (Courtot et al, 2011).

In resources such as the European Variation Archive and ArrayExpress, clinical descriptions of disease and phenotypes aggregated from multiple remote data sources and projects with different foci at the CTTV also historically use different ontologies for disease/phenotype annotation. Multiple disease ontologies are used within EMBL-EBI, examples include Online Mendelian Inheritance in Man (OMIM) (Amberger et al. 2015), the Systematized Nomenclature of Medicine – Clinical Term (SNOMED-CT) (Cornet and de Keizer 2008), the Human Disease Ontology (DO) (Kibbe et al. 2015), and the Medical Dictionary for Regulatory Activities (MedDRA) (Brown et al. 1999). The use of multiple ontologies, particularly for disease and phenotype across resources has resulted in a mixed set of identifiers and concepts for disease and phenotypes, which led to an integration challenge. The EFO represents disease as a disposition, associated with one or more phenotypes. For example, a common disease *'ulcerative colitis'* is associated with a phenotypic quality *'bile duct inflammation'*. A rare *'chronic granulomatous disease'* is associated to the same bile duct inflammation phenotype. This Open Biomedical Ontology (OBO) compliant separation of disposition and quality allows for data integration with other domain-specific ontologies. Additionally the ontological classification of disease and phenotypes in the EFO assists the exploration of phenotypic connections between rare and common diseases in the CTTV. For example *colitis* is connected to the rare disease

* To whom correspondence should be addressed.

chronic granulomatous syndrome via a shared phenotype in the Open Biomedical AssociatioN (OBAN) representation (See Methods section). Surveys of the content of EMBL-EBI databases indicate that disease is more commonly represented than phenotype, but that the two are often mixed within a single resource e.g. EVA. Here we describe the data mapping process across multiple large bioinformatics databases and introduce the OBAN framework to represent text mined and clinically validated disease-phenotype associations and describe the use of these associations towards bridging common and rare diseases.

2 METHODS

2.1 Mapping disease and phenotype terms to EFO

In this study, phenotypes and diseases were mined from ChEMBL (Bento et al. 2014) exploiting the Anatomical Therapeutic Chemical (ATC) classification system, the European Variation Archive (EVA – http://www.ebi.ac.uk/eva/) importing pathogenic trait names from ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/), and UniProt (UniProt 2015). The EFO uses a cross-referencing mechanism via the source-specific *xxx_definition_citation* annotation property with *xxx* representing the source name (e.g. *OMIM_definition_citation*), and the OBOinOWL property *hasDbXref* listing the database cross references imported from the utilized ontologies allowing mapping of data from different resources. When Universal Resource Identifiers (URIs) are provided, disease and phenotype terms are programmatically mapped to the referenced source databases (e.g. OMIM, MedDRA, SNOMED-CT, ICD-9, NCI Thesaurus, DO, HP, or MeSH) via the annotation properties. We were able to quickly identify EFO classes that correspond with annotations cross referenced to OMIM, SNOMED, HP, and MP through this automated process.

When clinical phenotype information is not linked to a standardized URI system (e.g. EVA trait names), manual curation is applied to the data to carefully map the disease or phenotype annotation to the EFO. This process is in addition to the manual curation process used to assign disease terms when the record was initially curated. It also includes examination of OMIM entries, and Orphanet data to identify mappings. This step is coupled with literature review to ensure the accuracy of the mapping. For example, the unmapped phenotype term 'Glucose-6-phosphate transport defect' was manually mapped to 'Glycogen storage disease due to glucose-6-phosphatase deficiency type b' in Orphanet. If a term results in no available mappings to existing terms in EFO, external ontologies are examined for new import terms; failing this, an EFO class is added, and asserted into an appropriate classification location. EFO first attempts to create terms by requesting from the authoritative reference ontology (for example request of new rare disease term from ORDO), and avoids generating an EFO term when the scope of work is with a reference ontology. Occasionally EFO temporarily creates the term and later imports back from the reference ontology if and when it becomes available to avoid delays in data releases.

2.2 Building IBD disease-phenotype association knowledgebase

A challenge in modeling disease and phenotype connections in an ontological framework is that they are typically considered a 'sometimes associated' relationship. Ontologies expressed in OWL are not well suited to describe such relationships. OWL implementation with a probability value attached to the object property relation between two classes to describe this "sometimes-associated" are not ideal, especially when that probability is unknown and support for

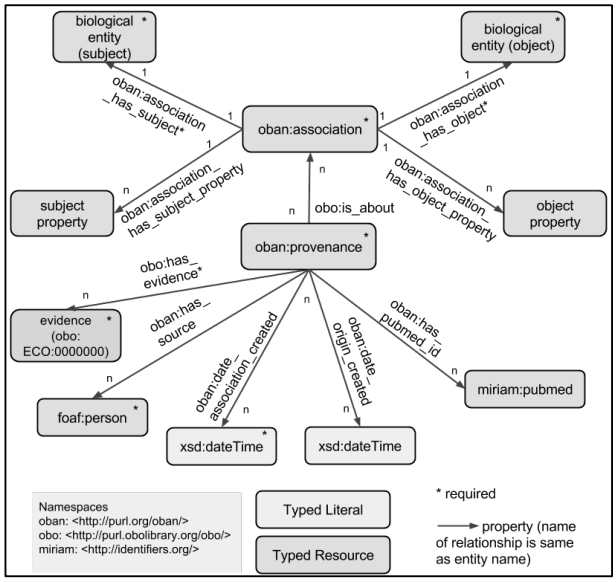

Figure 1. An OBAN association links an entity such as a disease to another such as an associated phenotype and retains the provenance (e.g. manual curation, published findings, etc). Entities marked with * are required and others are added as per association, for instance the PubMed triple in this figure. This is not intended to be an exhaustive example.

such constructs is exploratory at best. Where connections can be made existentially, they are asserted in the ontology as class descriptions. For instance, if a disease always manifests in a particular organ then a triple such as 'disease *has_disease_location* organism part' is added to the ontology. Such class descriptions are currently added by ontologists to EFO (see results for some figures).

For other 'not always true' relationships, the OBAN representation has been designed in an attempt to ease this problem. OBAN (Figure 1) decouples the relationship between the disease and phenotype classes and instead makes the relationship about an intermediate class of things – associations – true for a given disease and a phenotype. Associa-

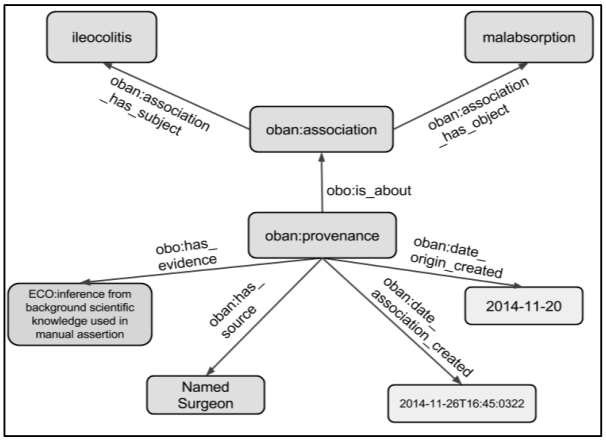

Figure 2. An example of connecting a phenotype (malabsorption) with a disease (ileocolitis) using OBAN. Provenance here is a manual curation by a named clinician (name omitted here).

tions are represented as an individual instance of class *'association*' which has one or more '*provenance*' individuals (see Figure 1). One provenance individual can also be about several associations (for instance the same paper may provide evidence for multiple disease-phenotype associations) and each association can have several provenances. In this work, diseases are typically subjects and phenotypes objects but the association is bi-directional; the subject and object relations are there to enable directionality if required.

As exemplified in Figure 2, an association between *ileocolitis* to the phenotype *malabsorption* exists where the provenance is provided via manual curation from a named surgeon (name omitted). In the OWL representation of associations the biological entities are represented using the same URI as the OWL class rather than represented as individuals – a technique known as punning (OWL 2 Web Ontology Language, 2012). Though not crucial, using punning to generate an instance identifier is preferred as it avoids the need to create many new URIs for individuals of the same diseases or phenotypes.

In addition, OBAN separates the association from the provenance, which is used in making this assertion i.e. a separation of concerns, separating the 'what is asserted' from 'who made the assertion'. A similar pattern is used in nanopublications (Patrinos et al. 2012). OBAN makes use of the Evidence Code Ontology (ECO) (Chibucos et al. 2014) and allows for extensible triples to be added, such as PubMedID, a curator name or a confidence score and methods for how it was derived as seen in Figure 1.

To populate the OBAN model with associations for Inflammatory Bowel Disease (IBD), a two-step process was undertaken. First, text mining over the European PubMed Central API was used to identify papers annotated with MeSH terms associated with IBD. A dictionary comprised of the Human Phenotype (HP) and the Mammalian Phenotyp (MP) ontology terms was then used as input to the Europe PMC hosted *Whatizit* pipeline (Rebholz-Schuhmann et al. 2008) which was applied to abstracts identified in the first stage. The process returned a list of candidate disease-phenotype associations, Term Frequency, Inverse Document Frequency, associated phenotype terms and abstract links. EBI curators performed initial cleaning of nonspecific terms – for example the HP contains the term 'All'. GSK clinicians then verified the candidate associations before the final list of IBD disease-phenotype associations was transformed into OWL format corresponding to OBAN.

3 RESULTS

3.1 Mapping Common and Rare Disease Annotations

Using the process described in the methods section, EFO has been used to annotate the data summarized in Table 1 with ontology classes. The EVA database represents an import of ClinVar where many of the trait names contained cross-references to multiple disease ontologies that required manual inspection prior to mapping. The mapping of databases which had standardized on an existing single ontology or annotation resource e.g. ATC or OMIM required less curatorial efforts expended in ensuring that mappings were valid to the meaning rather than a lexical term match.

Table 1. Summary of textual data annotations that are mapped to EFO or ORDO ontology classes following process outlined in methods section (%)

Database	% Annotated to EFO or ORDO
EVA (inc. ClinVar)	89% of annotations of frequency > 100
ArrayExpress	77%
UniProt	78%
Reactome	100%
ChEMBL	99%
GWAS Catalog	100%

3.2 Extending the ontology with disease axioms

As mentioned in section 2, as well as connecting the data to the ontology by annotaton, connections between rare and common diseases in the ontology can be formed through class descriptions. EFO has been extended to add such descriptions One such relevant description is in connecting rare and common disease to organism parts. EFO models this using a simple existential restriction: 'disease' has_disease_location some 'organism part' where has_disease_location is a subproperty of the OBO located_in object property. EFO version 2.58 (March 2015) contains 352 such relationships, connecting 4,718 diseases to the anatomical areas where they primarily manifest. A high-resolution summary is available at https://github.com/CTTV/ISMB2015/blob/master/figures/r2c.pdf.

3.3 IBD Disease-Phenotype associations

Inflammatory Bowel Disease (IBD) is one of the driving use cases for CTTV and as such has been an early focus for this work. The process pipeline in mapping and associating disease-phenotype described in this study is being expanded to cover other CTTV driving use cases (autoimmunity, cancer, and diabetes). The association file is downloadable at https://sourceforge.net/p/efo/code/HEAD/tree/trunk/src/efoassociations/ibd_2_pheno_associations.owl. The file contains 289 disease-phenotype associations for the IBD domain. Of the initial mined candidate IBD phenotype associations, 41.6% were deemed correct on manual review (precision).

To facilitate connection to rare disease, we have used and extended work already done by Orphanet and the HPO lab. We have extended their ORDO to HPO associations to 12,208 individual rare disease-phenotype associations using literature curation and clinician validation. For instance, connecting colon to Crohn's disease and similarly to Muir-Torre syndrome (a rare form of colon cancer) provides a connection between disorders which are known to share common phenotypes (Lester et al, 2009). The complete listing of these rare-to-common diseases via phenotypes are all available in the OBAN model at http://sourceforge.net/p/efo/code/HEAD/tree/trunk/src/efoassociations/. By combining the associations to phenotypes from rare diseases, or common diseases we can provide another mechanism for integrating rare and common disease. The current set of associations enables 535 connections between a phenotype and at least one common and at least one rare disease. Such connections can reveal possible new, or confirm known, findings, providing evidence for common mechanisms. Examples from our data include connections for which publications exist e.g. pruritus which connects both psoriasis and lamella ichthyosis (Stepanova et al., 2001) and also those for which publications are harder to find such as Crohn's disease and Bannayan-Riley-Ruvalcaba syndrome (via cachexia).

4 DISCUSSION

The understanding and representation of phenotype and disease is both context and domain specific. Here we operate in the translational research domain specifically to characterise drug targets and to explore phenotypic connections between rare and common disease. It is also confounded as some of the phenotype labels in HPO are primarily considered diseases, such as Crohn's disease. Medical language is also multifaceted in itself. A clinical observation is recorded in different formats and writings. Mapping to EFO, HP, or ORDO in this scenario requires a careful investigation of cross-references, and sometimes the knowledge publications on the matter. The rare-to-common disease association literature mining process provides a simplistic but rapid method to identify 'candidate' associations, which are then curated by expert clinicians. The 42% precision of this process could be improved by incorporating aspects of negation detection. Manually curating these candidate associations has also proved to be time-consuming and labour-intensive on the clinicians' side. The OBAN representation provides a means to link diseases and phenotypes via a simple association and has already been applied to Type 2 Diabetes and other disease areas of interest to CTTV. A phenotypic dissection of disease provides a concrete mechanism to translate the biological complexity to a computational representation to aid in identification and validation of therapeutic targets. The biological subject and object in the OBAN association triples exploit the ontology infrastructure provided in the EFO and provide a means to express confidence in annotations based on ECO. The EFO and phenotypic associations will be deployed in the CTTV platform, which will be freely available to the community after release in late 2015. EFO is freely available, as are the OBAN associations. In future we will include phenotypic frequencies, and disease stage subdivision of phenotypes. This will require a revision to the EFO disease hierarchy, which we hope to achieve with the wider community and the Human Disease Ontology.

ACKNOWLEDGEMENTS

This work was funded by the CTTV (SS, JM), EMBL Core Funds (HP) and the BioMedBridges project funded by the European Commission FP7 Capacities Specific Programme, grant number 284209 (DV). We thank Chris Mungall and Simon Jupp for useful discussion on OBAN, Peter Robinson for advice on the use of HPO, and GSK domain experts Jatin Patel, Soumitra Ghosh and Mei-Lun Wang.

REFERENCES

Amberger, J. S., et al. (2015), 'OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders', *Nucleic Acids Res,* 43, D789-98.

Bento, A. P., et al. (2014), 'The ChEMBL bioactivity database: an update', *Nucleic Acids Res,* 42 (Database issue), D1083-90.

Brown, E. G., et al.(1999), 'The medical dictionary for regulatory activities (MedDRA)', *Drug Saf,* 20 (2), 109-17.

Chibucos, M. C., et al. (2014), 'Standardized description of scientific evidence using the Evidence Ontology (ECO)', *Database (Oxford),* 2014.

Cornet, R. and de Keizer, N. (2008), 'Forty years of SNOMED: a literature review', *BMC Med Inform Decis Mak,* 8 Suppl 1, S2.

Courtot M, et al. (2011) MIREOT: The minimum information to reference an external ontology term. App. Ontology, 6, 23-33.

Kibbe, W. A., et al. (2015), 'Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data', *Nucleic Acids Res,* 43 (Database issue), D1071-8.

Lester, LU, Rapini, R. (2009) 'Dermatologic manifestations of colonic disorders', *Current Opinion in Gastro.*, 25(1), 66-73.

Ma'ayan, A., et al. (2014), 'Lean Big Data integration in systems biology and systems pharmacology', *Trends Pharmacol Sci,* 35 (9), 450-60.

Malone, J., et al. (2010), 'Modeling sample variables with an Experimental Factor Ontology', *Bioinformatics,* 26 (8), 1112-8.

Patrinos, G. P., et al. (2012), 'Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain', *Hum Mutat,* 33 (11), 1503-12.

OWL 2 Web Ontology Language, 2012. Available from: http://www.w3.org/TR/owl2-new-features/#F12:_Punning [12 March 2015].

Rebholz-Schuhmann, D., et al. (2008), 'Text processing through Web services: calling Whatizit', *Bioinformatics,* 24 (2), 296-8.

Stepanova, A., et al. (2001) 'Association of psoriasis and congenital lamellar ichthyosis', Hautarzt., Vol 52(8), p. 722-5.

Thompson, R., et al. (2014), 'RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research', *J Gen Intern Med,* 29 Suppl 3, S780-7.

UniProt, Consortium (2015), 'UniProt: a hub for protein information', *Nucleic Acids Res,* 43 (Database issue), D204-12.