Developmental Biology 354 (2011) 9–17

Contents lists available at ScienceDirect

Developmental Biology

journal homepage: www.elsevier.com/developmentalbiology

The representation of heart development in the gene ontology

Varsha K. Khodiyar [a,*,1], David P. Hill [b,i,1], Doug Howe [c], Tanya Z. Berardini [d,i], Susan Tweedie [e], Philippa J. Talmud [a], Ross Breckenridge [f], Shoumo Bhattarcharya [g], Paul Riley [h], Peter Scambler [h], Ruth C. Lovering [a]

[a] Cardiovascular GO Annotation Initiative, Centre for Cardiovascular Genetics, Rayne Institute, University College London, London, UK
[b] Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME, USA
[c] The Zebrafish Information Network, 5291 University of Oregon, Eugene, OR, USA
[d] The Arabidopsis Information Resource, Department of Plant Biology, Carnegie Institute for Science, Stanford, CA, USA
[e] FlyBase, Department of Genetics, University of Cambridge, UK
[f] Centre for Metabolism and Experimental Therapeutics, Rayne Institute, University College London, London, UK
[g] Department of Cardiovascular Medicine & Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, UK
[h] University College London-Institute of Child Health, Guilford St, London, UK
[i] Gene Ontology Consortium, USA

ARTICLE INFO

Article history:
Received for publication 25 November 2010
Revised 14 February 2011
Accepted 9 March 2011
Available online 17 March 2011

Keywords:
Annotation
Cardiovascular
Development
Gene Ontology
Heart

ABSTRACT

An understanding of heart development is critical in any systems biology approach to cardiovascular disease. The interpretation of data generated from high-throughput technologies (such as microarray and proteomics) is also essential to this approach. However, characterizing the role of genes in the processes underlying heart development and cardiovascular disease involves the non-trivial task of data analysis and integration of previous knowledge. The Gene Ontology (GO) Consortium provides structured controlled biological vocabularies that are used to summarize previous functional knowledge for gene products across all species. One aspect of GO describes biological processes, such as development and signaling.

In order to support high-throughput cardiovascular research, we have initiated an effort to fully describe heart development in GO; expanding the number of GO terms describing heart development from 12 to over 280. This new ontology describes heart morphogenesis, the differentiation of specific cardiac cell types, and the involvement of signaling pathways in heart development. This work also aligns GO with the current views of the heart development research community and its representation in the literature. This extension of GO allows gene product annotators to comprehensively capture the genetic program leading to the developmental progression of the heart. This will enable users to integrate heart development data across species, resulting in the comprehensive retrieval of information about this subject.

The revised GO structure, combined with gene product annotations, should improve the interpretation of data from high-throughput methods in a variety of cardiovascular research areas, including heart development, congenital cardiac disease, and cardiac stem cell research. Additionally, we invite the heart development community to contribute to the expansion of this important dataset for the benefit of future research in this area.

Introduction

Forming as the result of an elegant coordination of integrated processes, the heart is one of the first organs to develop in a vertebrate embryo. Understanding this developmental process is critical to the understanding of cardiovascular disease (CVD), a leading cause of mortality worldwide (Batsis and Lopez-Jimenez, 2010). One aspect of CVD is damaged heart tissue, and the possibility of using stem cells to repair a heart, is an active area of research (Bollini et al., 2011). However, to fully understand how to repair a heart we must first understand the processes by which a heart is formed. Heart development is a complex process controlled by a multitude of coordinated cellular processes resulting in proper patterns of cell

* Corresponding author at: Centre for Cardiovascular Genetics, Rayne Institute, 5 University Street, University College London, London WC1E 6JF, UK.
E-mail addresses: v.khodiyar@ucl.ac.uk (V.K. Khodiyar), dph@informatics.jax.org (D.P. Hill), dhowe@zfin.org (D. Howe), tberardi@acoma.stanford.edu (T.Z. Berardini), sart2@gen.cam.ac.uk (S. Tweedie), p.talmud@ucl.ac.uk (P.J. Talmud), r.breckenridge@ucl.ac.uk (R. Breckenridge), shoumo@me.com (S. Bhattarcharya), p.riley@ich.ucl.ac.uk (P. Riley), p.scambler@ich.ucl.ac.uk (P. Scambler), r.lovering@ucl.ac.uk (R.C. Lovering).
URL's: http://www.geneontology.org (D.P. Hill), http://www.geneontology.org (T.Z. Berardini).
[1] These authors contributed equally to this work.

0012-1606/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ydbio.2011.03.011

differentiation and tissue morphogenesis (Abu-Issa and Kirby, 2007; Dyer and Kirby, 2009). The events in the process are still being identified, thus the study of genes and proteins involved in cardiovascular development is an important research area. As with any multi-genic process, there is an important role for high-throughput methods to characterize the genes and proteins involved in both developmental and disease processes. However, interpretation of high-throughput data with regard to previously established work is a non-trivial task, associated with a need for categorization of gene functions to enable data analysis. To assist with the interpretation of these data, the Gene Ontology Consortium (GOC) provides a robust hierarchical controlled vocabulary, the Gene Ontology (GO) (Ashburner et al., 2000). Individually each GO term can be applied to gene products across all species to summarize individual experiments. However, the power of GO lies in the fact that it is a categorization of gene product characteristics, rather than a categorization of gene products themselves. Thus, large numbers of gene products can be grouped on the basis of their characteristics, as defined by GO terms. Many high-throughput analysis tools have been developed for this purpose (Hendrickson et al., 2008; Malik et al., 2010; Werner, 2008). GO currently contains over 33,000 terms and is routinely used for the analysis of large datasets (Colak et al., 2009; Herbert et al., 2009; Mace et al., 2009).

The GOC comprises of GO editors who develop the ontology, and GO annotators who read the primary literature and create annotations using the ontology (Gene Ontology Consortium, 2009). The two groups of curators work closely together. As a result GO terms are continuously examined, evaluated and tested during the annotation process based on the experimental literature. The GOC has set up several mechanisms for handling ontology issues. We utilize a SourceForge tracker for ontology questions, have a 'help' resource available from the GO web site and provide contact information for curator interest groups who coordinate biological areas of ontology development (see box). There are more than a dozen model organism databases (MOD) represented in the GOC, which use GO to annotate the gene products of their specific organism. Additionally there are other annotation groups focusing on specific areas of biology. The BHF-UCL (British Heart Foundation-University College London) GO team focuses on cardiovascular biology and has the specific remit of annotating human proteins involved in cardiovascular processes and disease (Lovering et al., 2008). An advantage to having a specialized annotation group like the BHF-UCL group is that it allows for development of a whole branch of the ontology alongside the creation of gene annotations using the new ontology terms. GO annotators read the primary literature and create GO annotations using a variety of evidence codes that describe the nature of the experiments that support the annotation. In practice annotations to developmental processes are mostly derived from direct assays, mutational studies or gene interaction studies. Annotators are cautious about using expression data for GO annotations since expression may correlate with a process, but not necessarily be actively involved in the process.

In general, there are two ways in which ontology development can be coordinated with annotation: 1) terms can be requested individually as annotators find need for them; 2) large-scale development of the ontology can be undertaken in anticipation of the terms needed for annotation. In practice, we have found that the latter method is very efficient for developing focused areas of the ontology (Diehl et al., 2007; Feltrin et al., 2009; Maccagnan et al., 2010). Specifically, we find that collaboration between expert annotators, ontology developers and experimental biologists working in the relevant field is a productive way to create an accurate and complete representation of an area of biology in the shortest amount of time.

In order to establish the groundwork for an expansion of the representation of GO biological processes involved in heart development, an initial meeting was held between gene annotators (from the BHF-UCL GO team and several model-organism databases), GOC ontology developers and cardiac development experts (all authors of this paper). At the start of the work, there were 12 terms in GO that represented all of heart development. Work during the meeting, as well as subsequent revisions and discussions have resulted in the addition of 281 new terms to date (see Supplemental Table for full list of new terms). These new terms have been added to GO and are fully integrated with relationships to other developmental processes in the existing biological processes ontology, and all additional parent terms. The new terms are publically available as part of the current version of GO. GO now includes terms describing an anatomical representation of heart development (such as the valves and the heart chambers), as well as terms that describe specific types of processes that contribute to heart development (such as cell differentiation and signaling pathways). In this article we introduce the new heart development Gene Ontology, demonstrate how it is used for annotations and invite the heart development community to contribute to this important resource.

There are multiple ways in which the research community can contact curators on any aspect of GO

The GOC helpdesk can be contacted by completing the online form at www.geneontology.org/GO.contacts.shtml.

The GOC uses the SourceForge software for issues to be submitted and tracked. This software is publicly accessible and further details can be found at geneontology.sf.net.

Each MOD and curation group also has its own website with contact details for the curators working within that group. Details of how to contact the BHF-UCL team, can be found at www.cardiovasculargeneontology.com.

A gene ontology primer

Gene Ontology (GO) is a controlled vocabulary that is used to classify the biological characteristics of gene products. GO terms describe three characteristics, *Biological Process* (BP), *Molecular Function* (MF) and *Cellular Component* (CC). BP terms describe the general process a gene product is involved in, MF terms describe the specific molecular function of a gene product and CC terms describe the subcellular compartment in which a gene product is found. This work focused on the expansion of the BP part of GO, which encompasses all developmental processes.

Gene annotators use experimentally supported data from published literature to associate specific GO terms with the genes that have been shown to bear the attributes described by the GO term. Taken as a whole, the set of annotations to a given gene product aims to describe the totality of what is currently known about that gene product's role in biology, while an individual annotation describes the results of a single experiment. Annotators use their biological knowledge alongside information presented in the paper to judge the most specific term possible for each annotation. For example, individual experiments have shown that the homeobox *NKX2-5* gene product takes part in the process of cardiac muscle cell differentiation (GO:0055007; BP) (Tanaka et al., 1999), has transcription factor activity (GO:0003700; MF) (Kasahara and Izumo, 1999) and is found in the nucleus (GO:0005634; CC) (Zhu et al., 2000). Thus the *NKX2-5* gene product has been annotated to each of those terms. Additionally gene products may have multiple functions, often take part in more than one process and can be found in multiple subcellular compartments. GO allows a single gene product to be annotated to any number of terms from each of the three ontologies.

GO terms are structured in directed acyclic graphs (DAG), where each term can have multiple relationships to broader 'parent' and