Chapter 7

How Does the Scientific Community Contribute to Gene Ontology?

Ruth C. Lovering

Abstract

Collaborations between the scientific community and members of the Gene Ontology (GO) Consortium have led to an increase in the number and specificity of GO terms, as well as increasing the number of GO annotations. A variety of approaches have been taken to encourage research scientists to contribute to the GO, but the success of these approaches has been variable. This chapter reviews both the successes and failures of engaging the scientific community in GO development and annotation, as well as, providing motivation and advice to encourage individual researchers to contribute to GO.

Key words Clinical and basic research, Gene Ontology, Proteomics, Transcriptomics, Community, Community annotation, Community curation, Genomics, Bioinformatics, Curation, Annotation, Biocuration

1 Introduction

The overarching vision of the Gene Ontology Consortium (GOC) is to describe gene products across species—their temporally and spatially characteristic expression and localization, their contribution to multicomponent complexes, and their biochemical, physiological, or structural functions—and thus enable biologists to easily explore the universe of genomes [1]. In practical terms, this makes providing an accessible, navigable resource of gene products, rigorously described according a structured ontology, the GOC's key objective. The referenced links, between the identifiers for Gene Ontology (GO) terms and the identifiers for specific gene products, are the elemental GO annotations.

With Next Generation Sequencing technologies increasing the rate at which genomic and transcriptomic data are accumulating, the need for highly informative annotation data for the human genome is paramount. Community annotation has the potential to improve the information provided by the GO resource. Consequently, the GOC actively encourages contributions from

Christophe Dessimoz and Nives Škunca (eds.), *The Gene Ontology Handbook*, Methods in Molecular Biology, vol. 1446,
DOI 10.1007/978-1-4939-3743-1_7,

the scientific community, to ensure that the ontology appropriately reflects the current understanding of biology and to supply gene product annotations [2–4]. There are many online resources that encourage community annotation [5–7]; however, annotations created in the majority of these are not submitted to the GO database. This chapter, therefore, only discusses the progress of community contributions to the GO database.

2 Ontology Development Workshops

The success of GO is dependent on its ability to represent the research communities' interpretation of biological processes and individual gene product functions and cellular locations. This is achieved through the use of descriptive GO terms, with detailed definitions, and appropriate placement of GO terms within the ontology hierarchy. The majority of GO terms are created by GO editors, following a review of the current scientific literature, often, without the need of discussions with experts in the relevant field [8–9].

Major revisions or expansions of a specific GO domain are usually undertaken in consultation with experts working in that biological field. Notable successful ontology development projects include that of the immune system [10], heart development [2], kidney development [11], muscle processes and cellular components [12], cell cycle, and transcription [13]. The expansion of the heart development domain provides a good example of how experts in the field can guide the GO editors to create very descriptive terms. The GO heart development domain describes heart morphogenesis, the differentiation of specific cardiac cell types, and the involvement of signaling pathways in heart development. This was achieved following a 1½ day meeting with four heart development experts, as well as considerable email exchanges both before and after the meeting [2]. The result of this effort was an increase in the number of GO terms describing heart development from 12 to over 280, and the creation of highly expressive terms such as secondary heart field specification (GO:0003139) and canonical Wnt signaling in cardiac neural crest cell differentiation (GO:0061310).

3 Community Contributions to the GO Annotation Database

Lincoln Stein suggested that there are four organizational models to genome annotation: the factory (reliant on a high degree of automation), the museum (requiring expert curators), the cottage industry (scientists working out of their laboratories), and the party (or jamboree—a short intensive annotation workshop) [14]. To this, list

needs to be added "the school," where people are encouraged to annotate as part of a bioinformatics training program.

Currently, there are two major approaches taken to associate GO terms with gene products: manual curation of the literature and automated pipelines based on manually created rules (the "factory") [15]. The majority of manual annotation follows the "museum" model, relying on highly trained curators reading the published literature, evaluating the experimental evidence, and applying the appropriate GO terms to the gene record [8, 16]. The majority of these curators are associated with specific model organism databases, such as FlyBase [17], PomBase [18] and ZFIN [19], or proteomic databases, such as UniProt [20]. In general, these curators will be annotating gene products across a whole genome. In contrast, there have been a few annotation projects funded to improve the representation of specific biological domains, such as cardiovascular [3], kidney [21] and neurological [22]. Two of these projects are being undertaken by the UCL functional annotation team and provide an example of an expert curation team embedded within a scientific research group.

3.1 GO Annotation Within a Bioinformatics Course

In the "school" model, bioinformatics courses, which include an introduction to GO, provide an opportunity for attendees to contribute GO annotations. However, providing timely feedback to degree students is very labor intensive. Texas A&M University has circumvented this problem through the use of competitive peer review. A biannual multinational student competition has been established to undertake large-scale manual annotation of gene function using GO. In this competition, known as the Community Assessment of Community Annotation with Ontologies (CACAO),[1] teams of students get points for making annotations, but can also take points from competitors by correcting their annotations. A professional curator then reviews these and annotations that are judged to be correct are submitted to the GO database. This highly successful crowdsource project uses the online GONUTs wiki [23] to submit annotations and has supplied 3700 annotations to the GO database. The CACAO attribution identifies the resultant annotations, associated with over 2500 proteins. This competition has given over 700 students the opportunity not only to learn how to use some of the essential online biological knowledgebases, but to reinforce this knowledge over a 3-month period, connecting their curriculum to research applications. An MSc literature review project, at University College London (UCL), also provides an opportunity to supply GO annotations to the GO database. Four projects, to date, have resulted in annotations for proteins involved in autism [24], heart development, folic acid metabolism, and hereditary hemochromatosis, creating over 1000 annotations. A limitation of student annotations is that they do not draw on the expertise of the scientific community.

[1] http://gowiki.tamu.edu/wiki/index.php/Category:CACAO

For the past 5 years, the UCL functional annotation team has run a 2-day introduction to bioinformatics and GO course. This course has been attended by over 200 scientists, who have been given the opportunity to use the UniProt GO annotation tool, Protein2GO [20], to annotate their own papers or those published in their field of expertise. However, on average only 50 annotations are submitted during the entire course and very few scientists continue to contribute annotations after the end of the course. A similar problem has been identified in many other annotation workshops.

3.2 Annotation Workshops

The first workshop to submit GO annotations to the GO database focused on the annotation of the *Drosophila* genome [25]. Following on from this, the Pathema group ran several annotation-training workshops, in 2007, with the idea that trained scientists would continue to provide annotation updates thereafter [26]. Unfortunately, this approach had limited success. Although 150 scientists attended, in general they provided guidance to the curators, rather than creating annotations themselves.

3.3 GO Annotation by Specific Scientific Communities

One of the most successful community annotation projects is that run by PomBase [18]. During pilot projects, PomBase encouraged 80 scientists from the fission yeast community to submit a variety of annotations, including 226 GO annotations,[2] using their curation tool, CANTO [4]. Following on from this success the PomBase team now receives regular annotations from the *Schizosaccharomyces pombe* community.

Another successful community annotation project has a transcription focus and was initiated by a group at the Norwegian University of Science and Technology. To ensure a consistent annotation approach is undertaken, the Norwegian research group, with members of the GOC, has created a set of transcription factor annotation guidelines [13]. These provide details of the ideal GO terms to associate with a transcription factor, with a list of experimental conditions that would support these annotations. By using these standardized conventions, the literature-curated data (currently including annotations for 400 proteins) is imported directly into the GO database, with only minimal quality checking required. Working with the GOC, the SYSCILIA consortium may prove to be just as effective. This group has already contributed to the development of GO terms to describe ciliary components and processes and started to submit GO annotations [27].

The outstanding contributions of Ralf Stephan, demonstrates what can be achieved through dedication.[3] Stephan singlehandedly annotated 60% of the *Mycobacterium tuberculosis* genome, through the review of over 1000 papers. Furthermore, the resultant 7700

[2] http://www.pombase.org/community/fission-yeast-community-curation-pilot-project
[3] http://www.ark.in-berlin.de/Site/MTB-GOA.html

annotations associated with 2500 proteins were checked by the UniProt-GOA team [15] and needed very few edits, before incorporation into the GO database.

The success of PomBase may reflect the small size of the research community and that an early visionary investment has had a significant impact on the quality of data available at PomBase, achieved through the contributions of individual scientists and curators. In contrast, the Norwegian transcription factor project, formed to address the deficit of transcription factor annotations and in response to a need for comprehensive annotation of these proteins. The creation of a comprehensive and detailed annotation guide is key to the achievements of this project [13]. However, the GO database would also benefit from a few more "cottage industry" contributions, such as those provided for the *Mycobacterium tuberculosis* genome.

4 Why Contribute to GO?

The motivation behind "community annotation" is varied. Some scientists are contributing GO annotations purely to ensure their research area or gene product(s) of interest are well curated. Others may want to ensure data from their own papers is curated and, therefore, promoted in popular knowledgebases; potentially increasing the citation rate of these papers. Others still are motivated by peer competition! Regardless of the motivation, the GOC is always appreciative of input from the scientific community. Despite the success of some community annotation projects, taken as a whole, very few scientists suggest annotations, or papers for annotation. Consequently, the GOC continues to search for new ways to encourage the research community to contribute to curation activities. For example, the inclusion of data from gene wikis [5–7] could help take community annotation forwards. Considerable funding is being invested in NGS, proteomic and transcriptomic technologies and sequencing of population genomes. However, comprehensive gene annotation is likely to be a limiting factor in the identification of genes involved in polygenic diseases and disease-associated disregulated pathways. Many groups are turning to proprietary resources to provide these annotations [28], which also include freely available annotation data. A more sustainable approach, and one that will also support genomic research in developing countries, is to invest in improving the freely available annotation resources. All groups working with high-throughput datasets should consider working with the GOC and including in grant applications a component that would fund the submission of gene annotation data describing their area of interest, by expert curators, rather than requesting funding to enable access to proprietary software. The majority of members of the GOC do provide facilities to enable researchers to contribute to GO, the question is whether the scientific community will acknowledge that their input is required.

5 Resources Supporting Expert Contributions to GO

It is unrealistic to expect a limited number of GO curators and editors to understand all areas of biological and medical research. Consequently, a range of online facilities have been put in place to encourage scientists to review the ontology, to comment on the annotations, and to suggest papers for curation. In addition, several GO annotation tools, enable scientists to contribute annotation data [4, 8, 20]. Furthermore, the Protein2GO curation tool, automatically emails authors when one of their papers has been annotated, giving the authors an opportunity to comment on the curator's interpretation of their data [20].

Scientists interested in helping to improve the GO annotation resource can either contact the group providing annotations to their species or area of interest (see GOC contributors webpage geneontology.org/page/go-consortium-contributors-list) or submit enquires or information through the GOC webform geneontology.org/form/contact-go, which will be forwarded to the relevant database or group. Useful information to provide would be: details of key experimental publications for curation; a review of a particular annotation set (associated with a specific gene product or GO term), pointing out GO annotations that are missing, wrong, or controversial; comments on the ontology structure or definitions of GO terms, with a reference to support the changes required (Fig. 1). This would ensure that any erroneous annotations are removed promptly from the GO database, and that information from seminal papers is included. Scientists who are confident in using online resources may prefer to submit GO annotations, for any species, using the PomBase curation tool, CANTO curation.pombase.org/pombe [4]. Information provided by any of these means will be forwarded to the appropriate curation or editorial team and contributors will be notified when their suggestions have been incorporated. Full details about contributing to GO are available on the GOC website http://geneontology.org/page/contributing-go. Professional GO curators review all submitted annotations to ensure the annotations follow GO annotation rules and a consistent annotation approach is taken.

6 Following GO Developments

Scientists interested in finding out more about current GOC annotation and ontology development projects should sign up to the go-friends mailing list.[4] Alternatively, GO-relevant tweets can be followed via #geneontology, or @news4GO.

[4] http://mailman.stanford.edu/mailman/listinfo/go-friends

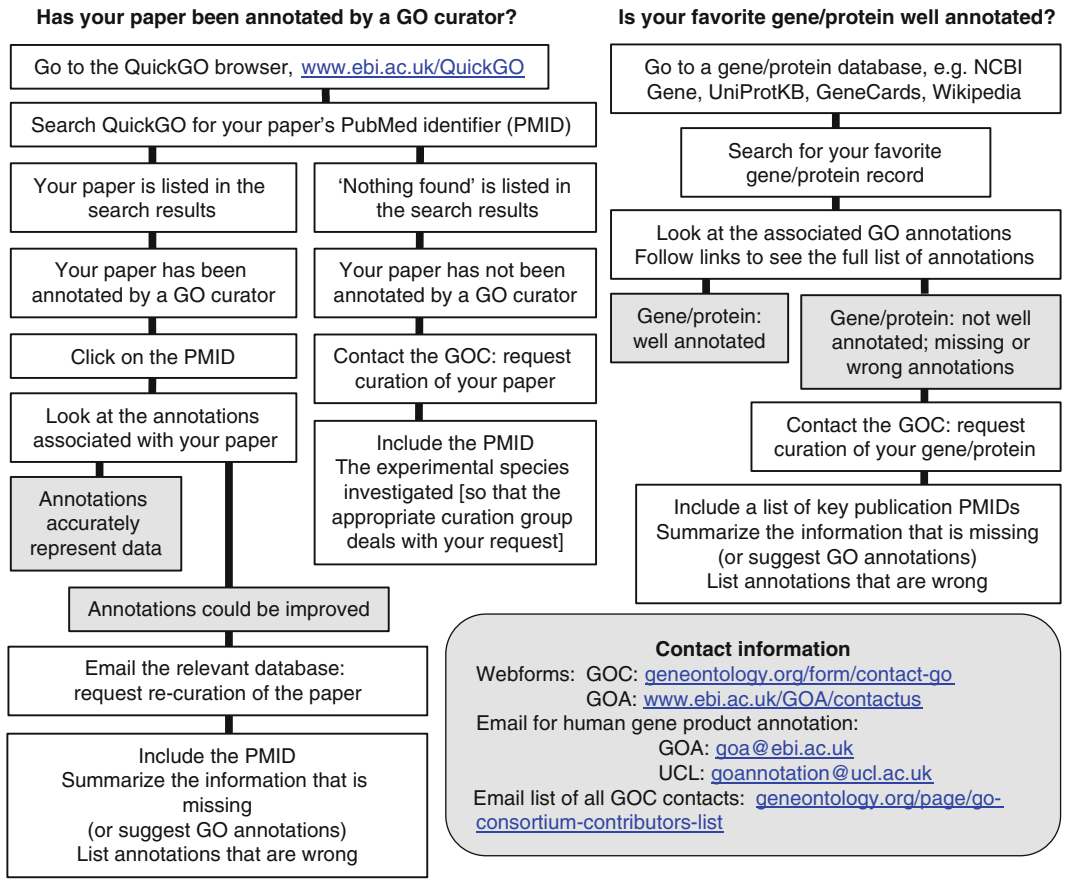

Fig. 1 How research scientists can help to improve the annotation content of GO

Acknowledgments

Supported by the British Heart Foundation (RG/13/5/30112), Parkinson's UK (G-1307), and the National Institute for Health Research University College London Hospitals Biomedical Research Centre. Many thanks to Dr. Rachael Huntley and Professor Suzanna Lewis for their reviews of this manuscript and to Doug Howe and Tanya Berardini for the information they provided. Open Access charges were funded by the University College London Library, the Swiss Institute of Bioinformatics, the Agassiz Foundation, and the Foundation for the University of Lausanne.

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29
2. Khodiyar VK, Hill DP, Howe D, Berardini TZ, Tweedie S et al (2011) The representation of heart development in the gene ontology. Dev Biol 354:9–17
3. Lovering RC, Dimmer EC, Talmud PJ (2009) Improvements to cardiovascular gene ontology. Atherosclerosis 205:9–14
4. Rutherford KM, Harris MA, Lock A, Oliver SG, Wood V (2014) Canto: an online tool for community literature curation. Bioinformatics 30:1791–1792
5. Singh M, Bhartiya D, Maini J, Sharma M, Singh AR et al (2014) The Zebrafish GenomeWiki: a crowdsourcing approach to connect the long tail for zebrafish gene annotation. Database (Oxford) 2014:bau011
6. Huss JW 3rd, Orozco C, Goodale J, Wu C, Batalov S et al (2008) A gene wiki for community annotation of gene function. PLoS Biol 6:e175
7. Menda N, Buels RM, Tecle I, Mueller LA (2008) A community-based annotation framework for linking solanaceae genomes with phenomes. Plant Physiol 147:1788–1799
8. Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. Nucleic Acids Res 43:D1049–1056
9. Leonelli S, Diehl AD, Christie KR, Harris MA, Lomax J (2011) How the gene ontology evolves. BMC Bioinformatics 12:325
10. Diehl AD, Lee JA, Scheuermann RH, Blake JA (2007) Ontology development for biological systems: immunology. Bioinformatics 23:913–915
11. Alam-Faruque Y, Hill DP, Dimmer EC, Harris MA, Foulger RE et al (2014) Representing kidney development using the gene ontology. PLoS One 9:e99864
12. Feltrin E, Campanaro S, Diehl AD, Ehler E, Faulkner G et al (2009) Muscle research and gene ontology: new standards for improved data integration. BMC Med Genomics 2:6
13. Tripathi S, Christie KR, Balakrishnan R, Huntley R, Hill DP et al (2013) Gene Ontology annotation of sequence-specific DNA binding transcription factors: setting the stage for a large-scale curation effort. Database (Oxford):bat062
14. Stein L (2001) Genome annotation: from sequence to biology. Nat Rev Genet 2:493–503
15. Camon E, Magrane M, Barrell D, Lee V, Dimmer E et al (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res 32:D262–266
16. Balakrishnan R, Harris MA, Huntley R, Van Auken K, Cherry JM (2013) A guide to best practices for Gene Ontology (GO) manual annotation. Database (Oxford):bat054
17. Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P et al (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Res 37:D555–559
18. McDowall MD, Harris MA, Lock A, Rutherford K, Staines DM et al (2015) PomBase 2015: updates to the fission yeast database. Nucleic Acids Res 43:D656–661
19. Bradford Y, Conlin T, Dunn N, Fashena D, Frazer K et al (2011) ZFIN: enhancements and updates to the Zebrafish Model Organism Database. Nucleic Acids Res 39:D822–829
20. Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C et al (2015) The GOA database: gene Ontology annotation updates for 2015. Nucleic Acids Res 43: D1057–1063
21. Alam-Faruque Y, Dimmer EC, Huntley RP, O'Donovan C, Scambler P et al (2010) The Renal Gene Ontology Annotation Initiative. Organogenesis 6:71–75

22. Foulger RE, Denny P, Hardy J, Martin MJ, Sawford T, Lovering RC (2016) Using the gene ontology to annotate key players in Parkinson's disease. Neuroinformatics
23. Renfro DP, McIntosh BK, Venkatraman A, Siegele DA, Hu JC (2012) GONUTS: the Gene Ontology Normal Usage Tracking System. Nucleic Acids Res 40:D1262–1269
24. Patel S, Roncaglia P, Lovering RC (2015) Using Gene Ontology to describe the role of the neurexin-neuroligin-SHANK complex in human, mouse and rat and its relevance to autism. BMC Bioinformatics 16:186
25. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD et al (2000) The genome sequence of Drosophila melanogaster. Science 287:2185–2195
26. Brinkac L, Madupu R, Caler E, Harkins D, Lorenzi H, Thiagarajan M, Sutton G (2009) Expert assertions through community annotation Jamborees. Nature Precedings
27. van Dam TJ, Wheway G, Slaats GG, Group SS, Huynen MA et al (2013) The SYSCILIA gold standard (SCGSv1) of known ciliary components and its applications within a systems biology consortium. Cilia 2:7
28. Stables MJ, Shah S, Camon EB, Lovering RC, Newson J et al (2011) Transcriptomic analyses of murine resolution-phase macrophages. Blood 118:e192–208