Published online 28 November 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D331–D338***
doi: 10.1093/nar/gkw1108

Expansion of the Gene Ontology knowledgebase and resources

The Gene Ontology Consortium*

List of authors of the Gene Ontology Consortium is provided in the Appendix.

Received October 25, 2016; Editorial Decision October 26, 2016; Accepted November 16, 2016

ABSTRACT

The Gene Ontology (GO) is a comprehensive resource of computable knowledge regarding the functions of genes and gene products. As such, it is extensively used by the biomedical research community for the analysis of -omics and related data. Our continued focus is on improving the quality and utility of the GO resources, and we welcome and encourage input from researchers in all areas of biology. In this update, we summarize the current contents of the GO knowledgebase, and present several new features and improvements that have been made to the ontology, the annotations and the tools. Among the highlights are 1) developments that facilitate access to, and application of, the GO knowledgebase, and 2) extensions to the resource as well as increasing support for descriptions of causal models of biological systems and network biology. To learn more, visit http://geneontology.org/.

INTRODUCTION

Because of the staggering complexity of biological systems and the ever-increasing size of datasets to analyze, biomedical research is becoming increasingly dependent on knowledge stored in computable form. The Gene Ontology (GO) project provides the most comprehensive resource currently available for computable knowledge regarding the functions of genes and gene products. The GO knowledgebase is composed of two primary components. The first is the Gene Ontology (GO), which provides the logical structure of the biological functions ('terms') and their relationships to one another, manifested as a directed acyclic graph. The second is the corpus of GO annotations, evidence-based statements relating a specific gene product (a protein, noncoding RNA, or macromolecular complex, which we refer to hereafter as 'genes' for simplicity) to a specific ontology term. Crucially, each annotation is linked to the evidence supporting that biological conclusion, typically a specific publication from the biomedical literature. Together, the ontology and annotations aim to describe a comprehensive model of biological systems. Currently, the GO knowledgebase includes experimental findings from almost 140 000 published papers, represented as over 600 000 experimentally-supported GO annotations. These provide the core dataset for additional inference of over 6 million functional annotations for a diverse set of organisms spanning the tree of life.

In addition to this core knowledgebase, GOC resources also include software to edit and perform logical reasoning over the ontologies, web access to the ontology and annotations, and analytical tools that use the GO knowledgebase to support biomedical research.

Here, we describe new developments in the last two years, including improvements in the ontology, increases in the number of GO annotations, and enhancements to make it easier for users to obtain and properly apply the information in the GO knowledgebase. The GO and associated products are available under a Creative Commons Attribution license from http://geneontology.org.

EXPANSION OF THE GO KNOWLEDGEBASE

Ontology

The Gene Ontology defines the universe of concepts relating to gene functions ('GO terms'), and how these functions are related to each other ('relations'). It is constantly revised and expanded as biological knowledge accumulates. Table 1 shows the number of terms and relations currently comprising the GO, which continues to increase compared to our last update two years ago (1). The GO describes function with respect to three aspects: molecular function (molecular-level activities performed by gene products), cellular component (the locations relative to cellular structures in which a gene product performs a function), and biological process (the larger processes, or 'biological programs' accomplished by multiple molecular activities). Ongoing revisions to the ontology are managed by a team of senior ontology editors with extensive experience in both biology and computational knowledge representation. Ontology updates are made collaboratively between the GOC ontology team and scientists who request the updates. Most requests come from scientists making GO annotations (these typically impact only a few terms each), and from domain

*To whom correspondence should be addressed. Paul D. Thomas. Tel: +1 323 442 7975; Fax: +1 323 442 7995; Email: pdthomas@usc.edu

experts in particular areas of biology (these typically revise an entire 'branch' of the ontology comprising many terms and relations). We invite researchers and computational scientists to submit requests for either new terms or new relations in the ontology.

New ontology terms may be requested in two ways, either semi-manually via the online, templated form known as TermGenie (2) or manually, via a GitHub tracker (https://github.com/geneontology/go-ontology/issues). Using the online TermGenie interface, submitted templated terms are screened by senior ontology editors for approval. In most cases, terms are approved as-is, but the tool also allows manual editing to correct occasional typographical errors, logical construction of definitions, or obsoletion of terms when deemed inappropriate.

Requests via the GitHub tracker may also include changes to the structure of the ontology and new relationships types. As with TermGenie, requests are manually reviewed by ontology editors who conduct the appropriate survey of the literature to validate or reject the request. Often the ontology revision process includes a dialogue between the submitter and the ontology editor, and for more complex cases, within the larger ontology development group, with the members of the annotation team, and with experts in specific areas of biology. This process ensures accurate representation of the biology. Written discussions are audited through the GitHub issue tracking mechanism and guidelines for submitting new requests are posted here: https://github.com/geneontology/go-ontology/blob/master/CONTRIBUTING.md.

Logical definitions and inter-ontology links. Axioms are an important component of any ontology. They are used to define the relationships that any given class has with regards to other classes in the ontology. They are essential for supporting computational reasoning over the GO, and for maintenance of the complex logical structure of the GO. Ontology editors define new terms with axioms, and check the corresponding inferences of relations to other terms, so that the GO remains logically consistent. In our last update (1), we reported on the go-plus edition of GO, which includes OWL (Web Ontology Language) equivalence axioms connecting GO to external Open Biomedical Ontology (OBO) classes (see http://geneontology.org/page/download-ontology). These axioms allow us to automatically construct and validate large portions of the ontology, using knowledge of the relationships between classes extracted from these external ontologies (see (3) for full details). In our last update from 2014, there were 9304 inter-ontology links to eight OBO (Open Biological Ontologies) sources: CHEBI (chemicals), CL (cell types), PATO (qualities/descriptors), PO (plant anatomy), PR (proteins), SO (nucleic acid and protein sequence types), UBERON (animal anatomy) and OBA (traits). In the 2016-08-08 release of GO, the number of links has increased by over a factor of two, to 21 077 inter-ontology links. This now also includes links to an additional ontology, the Fungal Anatomy Ontology (FAO), thereby increasing GO's interconnection with descriptions of the biology of the fungal clade.

Domain-focused ontology development. We carried out coordinated ontology development and focused annotation in several domains of biological function. Neurexins and neuroligins, proteins involved in synaptogenesis and known to be associated with autism spectrum disorder, were the object of a focused annotation approach, and descriptive GO terms were created to better annotate the roles of these gene products (4). Also, the cellular component ontology was revised to increase and improve classes to represent extracellular RNA metadata, such as extracellular vesicles (5). The ExoCarta and Vesiclepedia databases (6,7) have started to use the revised ontology, and are collaborating with the GOC to include their annotations in the GO database. In another specific domain-focused ontology development and annotation effort, representation of cilia-related biology within the GO resource is currently nearing completion. While the first part of the project focused on ciliary subcomponents, more recently we worked with experts in the field towards a better representation of cilia types and biological processes relevant to the functions of these important organelles (manuscripts in preparation). In addition, a focused curation of cilia-related gene annotations was undertaken. Lastly, we have added over 300 terms describing plant enzyme molecular functions (with associated Enzyme Commission identifiers), in response to requests from a group of plant biologists. We have also begun to design a representation of biochemical pathways in GO. Beginning with glycolysis, we devised a strategy for defining pathways using combinations of necessary enzymatic activities that are executed as part of the various types of glycolytic pathways and chemicals that are used and created by them (8). We will continue to use this approach for other biochemical pathways and extend it to define signaling pathways.

GO annotations

GO annotations consist of an association between a gene and a GO term, with supporting evidence in the form of a GO 'evidence code' and either a published reference or description of the methodology used to create the annotation. All GO annotations, however, are ultimately supported by the scientific literature, either directly or indirectly. The GO evidence codes describe the evidence and roughly reflect how far removed the annotated assertion is from direct experimental evidence, and whether this evidence was reviewed by an expert biocurator. The number of GO annotations, for selected evidence codes and different aspects of the GO, is shown in Table 2.

Experimentally-supported annotations. The EXPerimental (EXP) evidence codes indicate that there is evidence from an experiment directly supporting the annotation of the gene. For example, an association between a gene product and its subcellular localization as determined by immunofluorescence would be supported by the Inferred from Direct Assay (IDA) evidence code, a subtype of EXP evidence. Annotations with direct experimental evidence are created by biocurators, PhD-level experts trained in computational knowledge representation, who read peer-reviewed literature and create GO annotations as justified by the evidence presented in those articles.

Table 1. Number of terms and relationships in the three aspects of the Gene Ontology, as of October 2016

Aspect	Terms (classes)	Relationships
Molecular function (MF)	10 417	14 039
Cellular component (CC)	4022	7854
Biological process (BP)	29 146	71 372

Table 2. Number of experimental (EXP), and phylogenetically inferred (IBA), annotations for well-studied organisms. Statistics as of October 2016

Organism	Specific protein binding EXP	Molecular function EXP	Molecular function IBA	Cellular component EXP	Cellular component IBA	Biological process EXP	Biological process IBA
Human	32 369	23 811	5892	36 555	8508	38 819	14 596
Mouse	8740	12 934	7914	22 593	11 336	59 517	18 128
Rat	4239	11 986	6704	15 047	9804	27 591	16 810
Zebrafish	392	1521	6732	937	9845	18 004	17 001
Fruit fly	1137	4965	3168	10 488	4371	30 560	5913
Nematode (*C. elegans*)	2649	2203	3386	4858	4983	11 679	7683
Slime mold (*D. discoideum*)	521	942	2386	2109	3098	3630	4637
Budding yeast	106	8264	2002	16 752	2753	17 646	3608
Fission yeast	1364	3275	1750	11 290	2526	5074	3257
A. thaliana, plant	6131	7288	5662	23 762	7375	22 595	11 167
E. coli	2290	5017	734	3911	610	5501	905

To ensure consistency and quality in expert curation practices, GOC biocurators meet regularly to discuss curation issues and participate in annotation consistency exercises. During these exercises, multiple groups of curators annotate a single paper, which leads to clarification on the use of ontology terms and GO evidence codes, and develops best practices and consistency among the distributed GO annotation groups. For example, clarifying how the results of co-transfection and functional complementation experiments should be annotated, ensures that information based on functional genetic interactions versus phenotypic rescue is unambiguously captured in the GO knowledgebase.

Until recently microRNAs were under-annotated in GO because microRNA regulation of developmental and cellular processes was a relatively new field of study. Consequently, researchers had to rely on the functional annotations of the microRNA targets as a proxy, because direct functional annotation of the microRNAs themselves did not exist. In consultation with experts in the field of microRNA research, substantial effort was dedicated to redress this situation. We created annotation guidelines for microRNA annotation (9) and following these guidelines, we have generated annotations for over 300 human microRNAs, 70 in *Drosophila melanogaster*, and almost 200 in *Arabidopsis thaliana*.

Protein binding annotations are only useful if they include the specific protein binding partner. With the addition of the IntAct database (10) as a GO annotation provider, the number of specific protein binding annotations has increased dramatically (Table 2, first column). Only high-confidence annotations are incorporated into GO from IntAct. Combined with annotations from hypothesis-driven, small-scale experiments that have been contributed to GO from multiple different annotation providers, IntAct annotations help make the GO knowledgebase a useful resource for high-confidence protein interaction network data. To create protein interaction networks, users need to utilize the 'with' field (column 8) of the GO Association Files (GAF), which contains the identifier of the interacting partner.

We ask users to be aware of annotations that state that a particular gene product has been found NOT to have a given function. The NOT annotation is generally created when a gene product with specific domain or gene family association is expected by inference to have a certain activity, but where there is explicit experimental data shows that the gene product does NOT have that activity. These annotations are relatively rare in the GO knowledgebase (currently there are ∼3300 of these, based on experimental evidence). However, we believe they may be particularly useful in some applications, such as assessing function prediction accuracy. These annotations have the qualifier 'NOT' in the qualifier field (column 4) of the GAF.

Phylogenetically-inferred annotations. Phylogenetic principles, reconstructing evolutionary events to infer relationships among genes, provide a powerful way to gain insight into gene function. The GOC has supported a dedicated Phylogenetic Annotation effort since 2008 (11), which has been expanded in the past couple of years. The Phylogenetic Annotation method is described in detail elsewhere (12). Briefly, we have developed software (PAINT, Phylogenetic Annotation Inference Tool) with which a biocurator can view all experimental annotations for genes in a gene family, and use this information to infer annotations for uncharacterized members of the family. The biocurator creates an explicit model of gain and loss of gene function at specific branches in a phylogenetic tree of the family. This model is used to infer new annotations (i.e. not overlapping with experimental annotations) for genes in the family. Phylogenetically based annotations are denoted by the IBA (Inferred from Biological Ancestry) evidence codes. Each inferred annotation can be traced to the direct experimental

annotations that were used as the basis for that assertion. The GO Phylogenetic Annotation project is now the largest source of manually reviewed annotations in the GO knowledgebase, and it has substantially increased the number of annotations even in organisms that have been well-studied experimentally (Table 2).

Computationally-inferred annotations. Finally, those furthest removed from direct experimental findings, are the 'electronic' (IEA) evidence codes, which are not individually reviewed (although an extensive manual review of a sample is generally involved). IEA-supported annotations are ultimately based on either homology and/or other experimental or sequence information, but cannot generally be traced to the experimental source. Three methods make up the bulk of these annotations. The first, and most comprehensive, method is InterPro2GO (13), which is based on the curated association of a GO term with a generalized sequence model ('signature') of a group of homologous proteins. Protein sequences with a statistically significant match to a signature are assigned the GO terms associated with the signature, a form of homology inference. A second method is the computational conversion of UniProt controlled vocabulary terms (mostly Enzyme Commission numbers describing enzymatic activities, and UniProt keywords describing subcellular locations), to associated GO terms. Lastly, annotations are made based on 1:1 orthologs inferred from Ensembl gene trees, an approach which automatically transfers annotations found experimentally in one gene, to its 1:1 orthologs in the same taxonomic clade (e.g. those within the vertebrate clade, and separately, those within the plant clade).

USABILITY ENHANCEMENTS

Gene-centric GO annotation sets

Historically, the GOC has allowed each annotation provider to decide on the set of objects (instantiated as database identifiers) that are associated with GO terms by that provider. As a result, there has been some variability between different groups, with some providers annotating genes, some annotating proteins or non-coding RNA's, some annotating protein complexes, and some annotating multiple different types. This is our intended approach, as we wish to annotate the functions for all macromolecular machines. However, if multiple different identifiers are actually referring to the same gene or protein, this can lead to possible mistakes in analyses that rely on GO annotations. For example, consumers of GO annotations might count the same gene multiple times in their analyses if identifiers are not resolved to a single non-redundant set.

To ensure that each of the GO annotations uses only a single identifier for any given protein-coding gene, we now adopt a single, standard identifier for each gene. For well-studied model organisms that have a dedicated resource, the primary gene identifier from that resource is used. This has been the standard for some time, but not consistently employed by some annotation groups. For other organisms, we use the protein identifiers from the 'gene-centric reference proteome' (GCRP) sets from the UniProt resource. We collaborate with UniProt (14) and the Quest for Orthologs Initiative (15,16) to develop and maintain a GCRP set for each organism across a wide phylogenetic spread. The team at the UniProt resource generates the GCRP set by selecting a single 'reference' protein entry for each protein-coding gene in a genome. We are also working with these groups toward complete consistency between the GCRP sets in UniProt, and the dedicated model organism resources.

As always, additional information about the annotated entity (e.g. a specific isoform or modified form), when available, is recorded in a different column of the GAF. Annotations directly to macromolecular complexes are provided in separate files to avoid confusion with annotations to genes. To ensure completeness for gene-based analyses, the GO annotations for the genes encoding individual complex members are also included in the gene-centric annotation file. In the gene-centric file, each member of a complex is annotated with the functions of the entire complex when appropriate (these are flagged with the contributes_to qualifier, see http://geneontology.org/page/go-annotation-conventions#contri for more details).

Exploring the gene ontology and annotations using AmiGO 2

Since our last update, we have implemented a number of new features and usability improvements to AmiGO 2 (http://amigo.geneontology.org/) to facilitate how the community explores and uses GO. AmiGO 2 now has an interactive ontology and annotation browser. This allows users to navigate the GO structure by drilling down from more general to more specific classes and retrieve filtered annotations to any branch of the ontology. Annotation retrieval has also been improved. Whereas in the previous version downloads were limited to 10,000 lines, one can now download up to 100,000 lines. The addition of these two functionalities allows users to download large, highly customized sets of GO annotations using the integrated faceting capabilities, including taxon and evidence subsets, and free-text searching.

Other highlights include: (i) an integrated PubMed ID search that retrieves both annotations and intermediary PubMed information pages (see below); (ii) a new Matrix Tool that allows users to explore the overlaps between gene sets annotated to different GO classes (http://amigo.geneontology.org/matrix#order); (iii) new integration capabilities to connect the GO resources into customized workflows via a public bookmarking API and Galaxy (17); a Solr (https://cwiki.apache.org/confluence/display/solr/Apache+Solr+Reference+Guide) document store search environment which offers more powerful means to query the GO data; and (v) much improved integrated and interactive statistics and graphics summarizing the entirety of the GO annotations.

Gene set enrichment analysis

The gene set enrichment analysis tool on the GO homepage now links directly to the interface at the PANTHER website (http://go.pantherdb.org/). This enables users to take advantage of the PANTHER visualization tools, such as the new hierarchical view that organizes enrichment analysis results using the relations in the GO (18). This view groups related terms together to facilitate biological interpretation

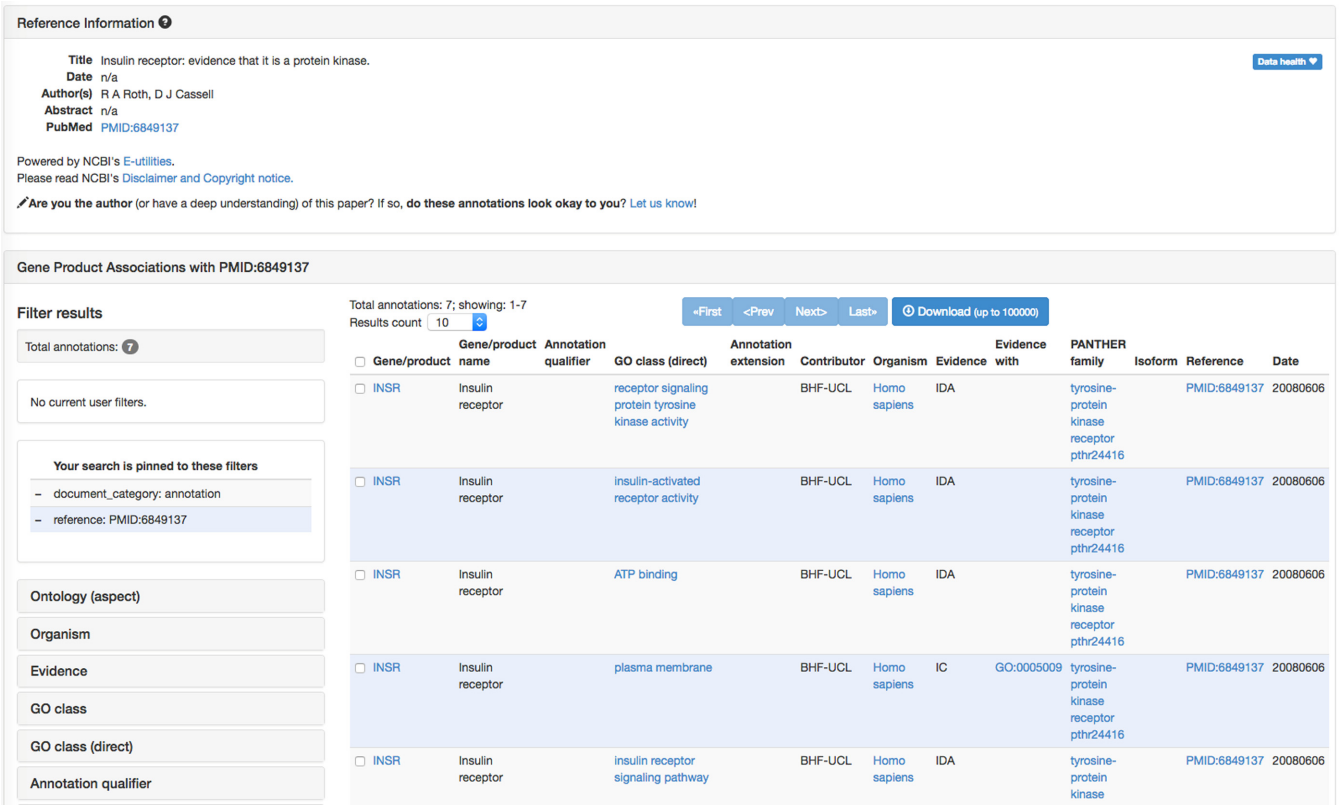

Figure 1. GO PubMed article page. All GO annotations that cite the article as evidence are shown on the page (table in lower right). Summary information on the article is obtained from NCBI web services (upper panel). Lower left panel shows general AmiGO2 filtering functionality: clicking on any of the data types (e.g. Ontology (aspect)) will allow selection of filters to apply.

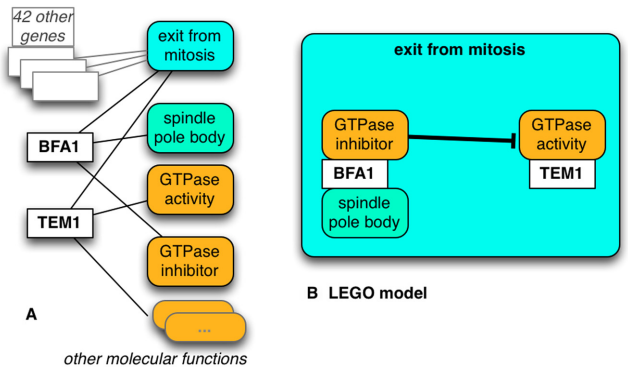

Figure 2. LEGO connects annotations. (**A**) Conventional annotations for two genes, and (**B**) the same annotations connected together in a LEGO model. This example shows 'The activity of BFA1 (in spindle pole body) inhibits the GTPase activity of TEM1, as part-of exit from mitosis.' Additional context, e.g. cell type, etc. can be added (not shown). Curated from data in (22).

of the enrichment results. In addition, recent studies have shown that many enrichment tools use outdated versions of the ontology or annotations, strongly impacting analyses (19). The GO annotations in PANTHER are updated monthly. The tool also displays key analysis parameters, such as GO annotation date/version and analysis tool version, that should be reported upon publication to aid in reproducibility.

Representing PubMed articles on the GO website

The GOC has now integrated a PubMed ID search, which generates a page for each PubMed (20) article that was used as evidence to support GO annotations (Figure 1). The page lists all GO annotations that were made using experimental evidence published in that paper. PubMed pages can be accessed from gene annotation data searches on the GO website. For example, one can enter a PubMed identifier (not including the 'PMID' prefix) in the 'Search GO data' box on the homepage, or click on the 'Filter and download' link in the 'Annotations' box. Clicking on a PubMed identifier in the 'Reference' column in the annotation results table directs users to the selected GO PubMed article page. The GO annotations comprise a high-level summary of the findings of a published paper with respect to gene functions. We expect that authors of papers may also find these pages very useful to assess how their work has been represented in the GO knowledgebase, and to provide feedback on how this representation might be improved. We have worked with the team at NCBI LinkOut (https://www.ncbi.nlm.nih.gov/books/NBK3805/) to include links from PubMed records directing users back to these GO article pages. These links allow users to access GO annotations while searching PubMed.

FUTURE DIRECTIONS

From annotations to models of biology (LEGO)

The GO annotation structure is historically quite simple, a statement consisting of one gene and one GO term (along with the evidence for that association, as described above). Because gene function is complex, and relates to larger systems and biological 'programs' carried out by multiple gene products, a typical GO annotation therefore represents just a single aspect of that function. Previously we reported on a simple extension to the GO annotation model, called 'annotation extensions' (21) that allows biocurators to capture additional contextual information using defined relations and entities to modify the selected GO term.

However, in order to allow more comprehensive, accurate statements about gene function and how multiple genes may function together, we have developed a 'grammar' for combining traditional GO annotations together into a more fully integrated representation of how gene functions relate to each other and to larger biological processes. We call this new formalism Linked Expressions using the Gene Ontology, or LEGO. An initial announcement can be found on http://geneontology.org/article/gaf-gpad-and-lego.

The LEGO formalism will be described in a separate publication, but briefly, it defines how different traditional GO annotations can be combined into a larger 'model' of gene and system function. A simple example is shown in Figure 2. Importantly, the larger model can and will be computationally decomposed into traditional GO annotations, so all the current applications of GO annotations, such as enrichment analysis, will still be supported. However, we also encourage developers of network-based analysis tools to download the native OWL (Web Ontology Language) representation of each LEGO model, that specifies how the functions of different gene products are linked into causal networks. Users may also be interested in browsing and viewing published models, which are available at: http://noctua.berkeleybop.org.

We have developed a software platform for creating and editing LEGO models, which we call Noctua. Noctua enables web-based collaborative annotation of LEGO models. Currently the GOC is in the process of transitioning to Noctua (http://noctua.berkeleybop.org) as the primary GO curation tool. Several GO annotation providers are already using the Noctua software to create LEGO models, and the GOC expects the number and utility of such models to increase rapidly in the coming period. We have conducted five annotation workshops in the past year to introduce biocurators to the Noctua annotation tool and the principles of OWL-based LEGO curation. Documentation for LEGO curation is linked from the Noctua home page, and is dynamically updated to reflect ongoing curatorial analysis and dialogue.

SUMMARY

The Gene Ontology Consortium is a growing, multidisciplinary community spanning biology, medicine and computer science. We aim to create a comprehensive, computational model of biological knowledge, that will continue to support analysis and interpretation of the ever-increasing store of molecular biomedical data. The endeavor is dependent on continued evaluation of our current understanding of biological systems, and has been strengthened and improved through the contributions of a large number of biologists and software developers.

We invite the research community to offer their input in all biological areas, as we strive to continuously improve the quality of GO knowledgebase and tools. Research groups may contribute updates to the ontology (e.g. request new terms) or provide new and updated annotations; feedback on the usability of existing tools or data, or suggestions for new features, are also welcome. Learn more about how to contribute your work to the GO resource at http://geneontology.org/page/contributing-go.

ACKNOWLEDGEMENTS

We want to acknowledge the broad community of scientists who have contributed to the GO knowledgebase, as biocurators and software developers (see http://geneontology.org/page/acknowledgments-contributors), and as authors of published papers that provide the basis for GO annotations (see http://geneontology.org/page/acknowledgments-authors).

FUNDING

National Institutes of Health/National Human Genome Research Institute [HG002273] awarded to the PI group formed by (alphabetically) Judith A. Blake, J. Michael Cherry, Suzanna E. Lewis, Paul W. Sternberg and Paul D. Thomas, as well as additional funding awarded to each participating institution. For more details please visit: http://geneontology.org/page/go-consortium-contributors-list.
Funding for open access charge: National Institutes of Health/National Human Genome Research Institute [HG002273].
Conflict of interest statement. None declared.

REFERENCES

1. The Gene Ontology Consortium. (2015) Gene ontology consortium: going forward. *Nucleic Acids Res*, **43**, D1049–D1056.
2. Dietze,H., Berardini,T.Z., Foulger,R.E., Hill,D.P., Lomax,J., Osumi-Sutherland,D., Roncaglia,P. and Mungall,C.J. (2014) TermGenie—a web-application for pattern-based ontology class generation. *J. Biomed. Semantics*, **5**, 48.
3. Mungall,C.J., Dietze,H. and Osumi-Sutherland,D. (2014) Use of OWL within the gene ontology. *BioRxiv*, 010090.
4. Patel,S., Roncaglia,P. and Lovering,R.C. (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.
5. Cheung,K.H., Keerthikumar,S., Roncaglia,P., Subramanian,S.L., Roth,M.E., Samuel,M., Anand,S., Gangoda,L., Gould,S., Alexander,R. *et al.* (2016) Extending gene ontology in the context of extracellular RNA and vesicle communication. *J. Biomed. Semantics*, **7**, 19.
6. Keerthikumar,S., Chisanga,D., Ariyaratne,D., Al Saffar,H., Anand,S., Zhao,K., Samuel,M., Pathan,M., Jois,M., Chilamkurti,N. *et al.* (2016) ExoCarta: A web-based compendium of exosomal cargo. *J. Mol. Biol.*, **428**, 688–692.
7. Kalra,H., Simpson,R.J., Ji,H., Aikawa,E., Altevogt,P., Askenase,P., Bond,V.C., Borràs,F.E., Breakefield,X., Budnik,V. *et al.* (2012) Vesiclepedia: a compendium for extracellular vesicles with continuous community annotation. *PLoS Biol.*, **10**, e1001450.

8. Hill,D.P., D'Eustachio,P., Berardini,T.Z., Mungall,C.J., Renedo,N. and Blake,J.A. (2016) Modeling biochemical pathways in the gene ontology. *Database (Oxford)*, **2016**, baw126.
9. Huntley,R.P., Sitnikov,D., Orlic-Milacic,M., Balakrishnan,R., D'Eustachio,P., Gillespie,M.E., Howe,D., Kalea,A.Z., Maegdefessel,L., Osumi-Sutherland,D. *et al.* (2016) Guidelines for the functional annotation of microrna using the gene ontology. *RNA*, **22**, 667–676.
10. Meldal,B.H., Forner-Martinez,O., Costanzo,M.C., Dana,J., Demeter,J., Dumousseau,M., Dwight,S.S., Gaulton,A., Licata,L., Melidoni,A.N. *et al.* (2015) The complex portal–an encyclopaedia of macromolecular complexes. *Nucleic Acids Res.*, **43**(Database issue), D479–D484.
11. Reference Genome Group of the Gene Ontology Consortium. (2009) The gene ontology's reference genome project: A unified framework for functional annotation across species. *PLoS Comput. Biol.*, **5**, e1000431.
12. Gaudet,P., Livstone,M.S., Lewis,S.E. and Thomas,P.D. (2011) Phylogenetic-based propagation of functional annotations within the gene ontology consortium. *Brief Bioinform.*, **12**, 449–462.
13. Mitchell,A., Chang,H.Y., Daugherty,L., Fraser,M., Hunter,S., Lopez,R., McAnulla,C., McMenamin,C., Nuka,G., Pesseat,S. *et al.* (2015) The interpro protein families database: The classification resource after 15 years. *Nucleic Acids Res.*, **43**(Database issue), D213–D221.
14. Dimmer,E.C., Huntley,R.P., Alam-Faruque,Y., Sawford,T., O'Donovan,C., Martin,M.J., Bely,B., Browne,P., Chan,W.M., Eberhardt,R. *et al.* (2012) The uniprot-go annotation database in 2011. *Nucleic Acids Res.*, **40**(Database issue), D565–D570.
15. Dessimoz,C., Gabaldón,T., Roos,D.S., Sonnhammer,E.L., Herrero,J. and Quest for Orthologs Consortium. (2012) Toward community standards in the quest for orthologs. *Bioinformatics*, **28**, 900–904.
16. Sonnhammer,E.L., Gabaldón,T., Sousa da Silva,A.W., Martin,M., Robinson-Rechavi,M., Boeckmann,B., Thomas,P.D. and Dessimoz C and quest for orthologs consortium. (2014) Big data and other challenges in the quest for orthologs. *Bioinformatics*, **30**, 2993–2998.
17. Afgan,E., Baker,D., van den Beek,M., Blankenberg,D., Bouvier,D., Čech,M., Chilton,J., Clements,D., Coraor,N., Eberhard,C. *et al.* (2016) The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.*, **44**, W3–W10.
18. Mi,H., Huang,X., Muruganujan,A., Mills,C., Tang,H., Kang,D. and Thomas,P.D. (2017) PANTHER version 11: Expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic Acid Res.*, doi:10.1093/nar/gkw1138.
19. Wadi,L., Meyer,M., Weiser,J., Stein,L.D. and Reimand,J. (2016) Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods*, **13**, 705–706.
20. NCBI Resource Coordinators. (2016) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **44**, D7–D19.
21. Huntley,R.P., Harris,M.A., Alam-Faruque,Y., Blake,J.A., Carbon,S., Dietze,H., Dimmer,E.C., Foulger,R.E., Hill,D.P., Khodiyar,V.K. *et al.* (2014) A method for increasing expressivity of gene ontology annotations using a compositional approach. *BMC Bioinformatics*, **15**, 155.
22. Geymonat,M., Spanos,A., Smith,S.J., Wheatley,E., Rittinger,K., Johnston,L.H. and Sedgwick,S.G. (2002) Control of mitotic exit in budding yeast. In vitro regulation of tem1 gtpase by bub2 and bfa1. *J. Biol. Chem.*, **277**, 28439–28445.

APPENDIX

The following is a list of the members of the Gene Ontology Consortium, who together authored this article. Authors marked with a star (*) made the largest contributions to the manuscript. **Berkeley BioinformaticsOpen-Source Projects (BBOP), Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory** (Berkeley, CA, USA): S. Carbon*, H. Dietze, S.E. Lewis, C.J. Mungall*, M.C. Munoz-Torres*; **dictyBase, Northwestern University** (Chicago, IL, USA): S. Basu, R.L. Chisholm, R.J. Dodson, P. Fey; **Division of Bioinformatics, Department of Preventive Medicine, University of Southern California** (Los Angeles, CA, USA): P.D. Thomas*, H. Mi, A. Muruganujan, X. Huang, S. Poudel; **EcoliWiki, Departments of Biology and Biochemistry and Biophysics, Texas A&M University** (College Station, TX, USA): J.C. Hu, S.A. Aleksander, B.K. McIntosh, D.P. Renfro, D.A. Siegele; **FlyBase, Department of Physiology, Development andNeuroscience, University of Cambridge** (Cambridge, UK): G. Antonazzo, H. Attrill, N.H. Brown, S.J. Marygold, P. McQuilton, L. Ponting, G.H. Millburn, A.J. Rey, R. Stefancsik, S. Tweedie; **FlyBase, The Biological Laboratories, Harvard University** (Cambridge, USA): K. Falls, A.J. Schroeder; **GO-EMBL-EBI** (Hinxton, UK): M. Courtot*, D. Osumi-Sutherland, H. Parkinson, P. Roncaglia*; **Center for Cardiovascular Genetics, University College London** (London, UK): R.C. Lovering*, R.E. Foulger, R.P. Huntley, P. Denny, N.H. Campbell, B. Kramarz, S. Patel, J.L Buxton, Z. Umrao, A.T. Deng, H. Alrohaif, K. Mitchell, F. Ratnaraj, W. Omer, M. Rodríguez-López.**; Institute for Genome Sciences, University of Maryland School of Medicine** (Baltimore, MD, USA): M. C. Chibucos, M. Giglio, S. Nadendla; **IntAct/Complex Portal, EMBL-EBI** (Hinxton, UK): M.J. Duesbury, M. Koch, B.H.M. Meldal, A. Melidoni, P. Porras, S. Orchard, A. Shrivastava; **InterPro, EMBL-EBI** (Hinxton, UK): H.Y. Chang, R.D. Finn, M. Fraser, A.L. Mitchell, G. Nuka, S. Potter, N.D. Rawlings, L. Richardson, A. Sangrador-Vegas, S.Y. Young; **MGI, The Jackson Laboratory** (Bar Harbor, ME, USA): J.A. Blake*, K.R. Christie, M.E. Dolan, H.J. Drabkin, D.P. Hill*, L. Ni, D. Sitnikov; **PomBase, University of Cambridge** (Cambridge, UK): M.A. Harris, J. Hayles, S.G. Oliver, K. Rutherford, V. Wood; **PomBase, University College London** (London UK): J. Bahler, A. Lock; **RGD, Medical College of Wisconsin** (Milwaukee, WI, USA): J. De Pons, M. Dwinell, M. Shimoyama, S. Laulederkind, G.T. Hayman, M. Tutaj, S.-J. Wang; **Reactome, Department of Biochemistry & Molecular Pharmacology, NYU School of Medicine** (New York, NY, USA): P. D'Eustachio, L. Matthews; **RTI International** (Research Triangle Park, NC, USA): J.P. Balhoff; **SGD, Department of Genetics, Stanford University** (Stanford, CA, USA): R. Balakrishnan, G. Binkley, J.M. Cherry, M.C. Costanzo, S.R. Engel, S.R. Miyasato, R.S. Nash, M. Simison, M.S. Skrzypek, S. Weng, E.D. Wong; **SIB Swiss Institute of Bioinformatics** (Geneva, Switzerland): M. Feuermann, P. Gaudet*; **TAIR, Phoenix Bioinformatics** (Redwood City, CA, USA): T.Z. Berardini, D. Li, B. Muller, L. Reiser, E. Huala; **UniProt: EMBL-EBI** (Hinxton, UK), **SIB Swiss Institute of Bioinformatics (SIB)** (Geneva, Switzerland), and **Protein Information Resource (PIR)** (Washington, DC, USA and Newark, DE, USA): J. Argasinska, C. Arighi, A. Auchincloss, K. Axelsen, G., Argoud-Puy, A. Bateman, B. Bely, M.-C. Blatter, C. Bonilla, L., Bougueleret, E. Boutet, L. Breuza, A. Bridge, R. Britto, H. Hye-A-Bye, C. Casals, E., Cibrian-Uhalte, E. Coudert, I. Cusin, P. Duek-Roggli, A. Estreicher, L., Famiglietti, P. Gane, P. Garmiri, G. Georghiou, A. Gos, N., Gruaz-Gumowski, E. Hatton-Ellis, U. Hinz, A. Holmes, C. Hulo, F. Jungo, G. Keller, K. Laiho, P. Lemercier, D. Lieberherr, A. MacDougall, M. Magrane, M.J. Martin, P. Masson, D.A. Natale, C. O'Donovan, I., Pedruzzi, K. Pichler, D. Poggioli,

S. Poux, C. Rivoire, B. Roechert, T. Sawford, M. Schneider, E. Speretta, A. Shypitsyna, A. Stutz, S. Sundaram, M., Tognolli, C. Wu, I. Xenarios, L.-S. Yeh; **WormBase, California Institute of Technology** (Pasadena, CA, USA), Wellcome Trust Sanger Institute (Hinxton, UK), **EBI** (Hinxton, UK), and Ontario Institute for Cancer Research (Toronto, Canada: J. Chan, S. Gao, K. Howe, R. Kishore, R. Lee, Y. Li, J. Lomax, H.-M. Muller, D. Raciti, K. Van Auken*, M. Berriman, L. Stein, Paul Kersey, P. W. Sternberg; **ZFIN, University of Oregon** (Eugene, OR, USA): D. Howe, M. Westerfield.