The Gene Ontology Knowledgebase in 2023

The Gene Ontology Consortium*

Corresponding author: Paul D Thomas, Department of Population and Public Health Sciences, University of Southern California, Los Angeles, CA USA, pdthomas@usc.edu

Authors

GO Central: Suzi A. Aleksander (ORCID:0000-0001-6787-2901), James Balhoff (ORCID:0000-0002-8688-6599), Seth Carbon (ORCID:0000-0001-8244-1536), J.Michael Cherry (ORCID:0000-0001-9163-5180), Harold J Drabkin (ORCID 0000-0003-2689-5511), Dustin Ebert (ORCID 0000-0002-6659-0416), Marc Feuermann (ORCID: 0000-0002-4187-2863), Pascale Gaudet (ORCID:0000-0003-1813-6857), Nomi L Harris (ORCID:0000-0001-6315-3707), David P Hill (ORCID 0000-0001-7476-6306), Raymond Lee (ORCID 0000-0002-8151-7479), Huaiyu Mi (ORCID:0000-0001-8721-202X), Sierra Moxon (ORCID: 0000-0002-8719-7760), Christopher J Mungall (ORCID 0000-0002-6601-2165), Anushya Muruganugan (ORCID 0000-0001-7169-5864), Tremayne Mushayahama (ORCID 0000-0002-2874-6934), Paul W. Sternberg (ORCID 0000-0002-7699-0173), Paul D Thomas (ORCID 0000-0002-9074-3507), Kimberly Van Auken (ORCID 0000-0002-8151-747).

CACAO, EcoliWiki: Jolene Ramsey (ORCID: 0000-0002-3774-5896), Deborah A. Siegele (ORCID: 0000-0001-8935-0696)

dictyBase: Rex L. Chisholm (ORCID 0000-0002-5638-3990), Petra Fey (ORCID 0000-0002-4532-2703)

DisProt, Dept. of Biomedical Sciences, University of Padova, Italy: Maria Cristina Aspromonte (ORCiD: 0000-0002-4937-6952), Maria Victoria Nugnes (ORCiD: 0000-0001-8399-7907), Federica Quaglia (ORCiD: 0000-0002-0341-4888), Silvio Tosatto (ORCiD: 0000-0003-4525-7793)

ECO: Michelle Giglio (ORCID: 0000-0001-7628-5565), Suvarna Nadendla (ORCID: 0000-0003-3643-281X)

FlyBase: Giulia Antonazzo (ORCID 0000-0003-0086-5621), Helen Attrill (ORCID 0000-0003-3212-6364), Gil dos Santos (ORCID 0000-0003-3507-8273), Steven Marygold (ORCID 0000-0003-2759-266X), Victor Strelets (ORCID 0000-0001-6556-9335), Christopher J. Tabone (ORCID 0000-0001-8746-0680), Jim Thurmond (ORCID 0000-0002-5142-2583), Pinglei Zhou

(ORCID 0000-0002-3012-1044).

Functional Gene Annotation, Institute of Cardiovascular Science, University College London (London, UK): Saadullah H. Ahmed, Praoparn Asanitthong (ORCID:0000-0002-6286-7959), Diana Luna Buitrago (0000-0003-2010-7174), Meltem N. Erdol (ORCID: 0000-0002-6574-5519), Matthew C. Gage (ORCID: 0000-0002-6668-6573), Mohamed Ali Kadhum, Kan Yan Chloe Li (ORCID:0000-0001-7737-2118), Miao Long, Aleksandra Michalak, Angeline Pesala (ORCID: 0000-0002-4639-2295), Armalya Pritazahra, Shirin C.C. Saverimuttu (ORCID:0000-0003-1191-2681), Renzhi Su (0000-0002-6819-5356), Kate E. Thurlow (ORCID: 0000-0001-9985-3684), Ruth C. Lovering (ORCID: 0000-0002-9791-0064)

GREEKC: Colin Logie (ORCID: 0000-0002-8534-6582)

JaponicusDB (London, UK): Snezhana Oliferenko (ORCID 0000-0002-8138-6851)

MGI: Judith Blake (ORCID 0000-0001-8522-334X), Karen Christie (ORCID 0000-0001-5501-853X), Lori Corbani (ORCID 0000-0002-2366-557X), Mary E Dolan (ORCID 0000-0001-7732-3295), Harold J Drabkin (ORCID 0000-0003-2689-5511), David P Hill (ORCID 0000-0001-7476-6306), Li Ni (ORCID 0000-0002-9796-7693), Dmitry Sitnikov (ORCID 0000-0003-3394-9805), Cynthia Smith (ORCID 0000-0003-3691-0324)

PHI-base: Alayne Cuzick (ORCID 0000-0001-8941-3984), James Seager (ORCID 0000-0001-7487-610X)

Planteome: Laurel Cooper (ORCID 0000-0002-6379-8932, Justin Elser (ORCID 0000-0003-0921-1982), Pankaj Jaiswal (ORCID 0000-0002-1005-8383),

Plant Reactome: Parul Gupta (ORCID 0000-0002-0190-8753), Pankaj Jaiswal (ORCID 0000-0002-1005-8383), Sushma Naithani (ORCID 0000-0001-7819-4552),

PomBase: Manuel Lera-Ramirez (ORCID 0000-0002-8666-9746), Kim Rutherford (ORCID 0000-0001-6277-726X), Valerie Wood (ORCID 0000-0001-6330-7526)

RGD: Jeffrey L. De Pons, Melinda R. Dwinell (ORCID 0000-0002-9528-3618), G. Thomas Hayman (ORCID 0000-0002-9553-7227), Mary L. Kaldunski (ORCID 0000-0003-3645-6803), Anne E. Kwitek (ORCID 0000-0003-1024-4116), Stanley J. F. Laulederkind (ORCID 0000-0001-5356-4174), Marek A. Tutaj (ORCID 0000-0002-1025-101X), Mahima Vedi (ORCID 0000-0001-5361-6739), Shur-Jen Wang (ORCID 0000-0001-5256-8683)

Reactome: Peter D'Eustachio (ORCID 0000-0002-5494-626X)

Rhea, Swiss-Prot group, SIB Swiss Institute of Bioinformatics (SIB) (Geneva, Switzerland): Lucila Aimo (ORCID: 0000-0003-0943-6401), Kristian Axelsen (ORCID: 0000-0003-3889-2879), Alan Bridge (ORCID: 0000-0003-2148-9135, Nevila Hyka-Nouspikel (ORCID: 0000-0001-7855-209X), Anne Morgat (ORCID: 0000-0002-1216-2969)

SGD, Department of Genetics, Stanford University (Stanford, CA): Suzi A. Aleksander (ORCID:0000-0001-6787-2901), J.Michael Cherry (ORCID:0000-0001-9163-5180), Stacia R. Engel (https://orcid.org/0000-0001-5472-917X), Kalpana Karra (ORCID:0000-0002-6816-2458), Stuart R. Miyasato (ORCID:0000-0001-5250-8920), Robert S. Nash (ORCID:0000-0002-3726-7441), Marek S. Skrzypek (ORCID:0000-0001-6749-615X), Shuai Weng (ORCID:0000-0003-4233-0772), Edith D. Wong (ORCID:0000-0001-9799-5523)

The Arabidopsis Information Resource (TAIR), Phoenix Bioinformatics (Newark, CA, USA): Erika Bakker (ORCID: 0000-0001-7933-3817), Tanya Z. Berardini (ORCID: 0000-0002-3837-8864), Leonore Reiser (ORCID: 0000-0003-0073-0858)

UniProt, Swiss-Prot group, SIB Swiss Institute of Bioinformatics (SIB) (Geneva, Switzerland): Andrea Auchincloss (ORCID: 0000-0002-5297-5390), Kristian Axelsen (ORCID: 0000-0003-3889-2879), Ghislaine Argoud-Puy (ORCID: 0000-0002-2979-8613), Marie-Claude Blatter (ORCID: 0000-0002-7474-1499), Emmanuel Boutet (ORCID: 0000-0002-3743-4203), Lionel Breuza (ORCID: 0000-0002-8075-8625), Alan Bridge (ORCID: 0000-0003-2148-9135), Cristina Casals-Casas (ORCID: 0000-0001-8769-177X), Elisabeth Coudert (ORCID: 0000-0001-8314-404X), Anne Estreicher (ORCID: 0000-0001-6828-2508), Maria Livia Famiglietti (ORCID: 0000-0002-5283-6593), Marc Feuermann (ORCID: 0000-0002-4187-2863), Arnaud Gos (ORCID: 0000-0002-5018-1378), Nadine Gruaz-Gumowski (ORCID: 0000-0002-4699-4907), Chantal Hulo (ORCID: 0000-0001-8176-7999), Nevila Hyka-Nouspikel (ORCID: 0000-0001-7855-209X), Florence Jungo (ORCID: 0000-0002-7456-8390), Philippe Le Mercier (ORCID: 0000-0001-8528-090X), Damien Lieberherr (ORCID: 0000-0002-9724-1710), Patrick Masson (ORCID: 0000-0001-7646-0052), Anne Morgat (ORCID: 0000-0002-1216-2969), Ivo Pedruzzi (ORCID: 0000-0001-8561-7170), Lucille Pourcel (ORCID: 0000-0003-1522-9900), Sylvain Poux (ORCID: 0000-0001-7299-6685), Catherine Rivoire (ORCID: 0000-0002-5979-8382), Shyamala Sundaram (ORCID: 0000-0003-4209-460X).

UniProt, EMBL-EBI, (Hinxton, UK): Alex Bateman (ORCID: 0000-0002-6982-4660), Emily Bowler-Barnett (ORCID: 0000-0003-4785-7231), Hema Bye-A-Jee (ORCID: 0000-0003-2464-7688), Paul Denny (ORCID: 0000-0003-4659-6893), Alexandr Ignatchenko (ORCID: 0000-0002-6083-941X), Rizwan Ishtiaq (ORCID: 0000-0001-8041-7321), Antonia Lock (ORCID: 0000-0003-1179-5999), Yvonne Lussi (ORCID: 0000-0002-5753-0235), Michele Magrane (ORCID: 0000-0003-3544-996X), Maria J. Martin (ORCID: 0000-0001-5454-2815), Sandra Orchard (ORCID: 0000-0002-8878-3972), Pedro Raposo (ORCID: 0000-0001-6149-9456), Elena Speretta (ORCID: 0000-0003-1506-7438), Nidhi Tyagi (ORCID: 0000-0002-2065-9051), Kate Warner (ORCID: 0000-0001-8705-181X), Rossana Zaru (ORCID: 0000-0002-3358-4423)

University at Buffalo, Department of Biomedical Informatics (Buffalo, NY, USA): Alexander D. Diehl (ORCID: 0000-0001-9990-8331)

WormBase: Raymond Lee (ORCID 0000-0002-8151-7479), Juancarlos Chan (ORCID:0000-0002-7259-8107), Stavros Diamantakis (ORCID:0000-0002-0273-3406), Daniela Raciti (ORCID:0000-0002-4945-5837), Magdalena Zarowiecki (ORCID:0000-0001-6102-7731)

Xenbase: Malcolm Fisher (ORCID:0000-0003-1074-8103), Christina James-Zorn (ORCID:0000-0001-5495-4588), Virgilio Ponferrada (ORCID:0000-0002-8590-7183), Aaron Zorn (ORCID:0000-0003-3217-3590)

ZFIN: Sridhar Ramachandran (ORCID:0000-0002-2246-3722), Leyla Ruzicka (ORCID:0000-0002-1009-339X), Monte Westerfield (ORCID:0000-0003-1187-7839)

1. **Keywords:** gene ontology, gene annotation, gene function, knowledgebase, knowledge graphs

Abstract

The Gene Ontology (GO) knowledgebase (http://geneontology.org) is a comprehensive resource concerning the functions of genes and gene products (proteins and non-coding RNAs). GO annotations cover genes from organisms across the tree of life as well as viruses, though most gene function knowledge currently derives from experiments carried out in a relatively small number of model organisms. Here, we provide an updated overview of the GO knowledgebase, as well as the efforts of the broad, international consortium of scientists that develops, maintains and updates the GO knowledgebase. The GO knowledgebase consists of three components: 1) the Gene Ontology – a computational knowledge structure describing functional characteristics of genes; 2) GO annotations – evidence-supported statements asserting that a specific gene product has a particular functional characteristic; and 3) GO Causal Activity Models (GO-CAMs) – mechanistic models of molecular "pathways" (GO biological processes) created by linking multiple GO annotations using defined relations. Each of these components is continually expanded, revised and updated in response to newly published discoveries, and receives extensive QA checks, reviews and user feedback. For each of these components, we provide a description of the current contents, recent developments to keep the knowledgebase up to date with new discoveries, as well as guidance on how users can best make use of the data we provide. We conclude with future directions for the project.

Introduction

Genes encode gene products, often proteins but also non-coding RNA molecules (ncRNAs), that perform functions at the molecular, cellular, and organismal levels. The GO knowledgebase provides a comprehensive, structured, computer-accessible representation of gene function, for genes from any cellular organism or virus. The GO knowledgebase has become a critical component of life science research, supporting analysis of large-scale experiments and biological systems (Duck *et al.* 2016). It is designed to make expert knowledge of gene function accessible for bench scientists as well as computational analyses. The basic model underlying GO is the "molecular biology paradigm" (Ashburner *et al.* 2000; Thomas 2017), in which there are three types ("aspects") of functional characteristics used to describe gene function:

- Molecular function (MF): the activities performed by a gene product at the molecular level
- Cellular component (CC): the locations, relative to cellular structures, where molecular functions are performed
- Biological process (BP): a "biological program" comprising molecular activities acting in concert to achieve a particular outcome; this program can be at the cellular level or at the organism level of multicellular organisms.

The GO knowledgebase consists of three components: the Gene Ontology, GO annotations and GO Causal Activity models (GO-CAMs) (Figure 1). The **Gene Ontology** (Figure 1A) structures our current knowledge of the types of functional characteristics a gene product may possess into a connected graph-based representation. Each ontology term (called "class" in the field of

ontologies) represents a functional characteristic that can be attributed to a gene product. Terms can have relationships between them, such as one term being more specific than another term (also called "subclass"); e.g. *DNA-binding transcription factor activity* is a subclass of *transcription regulator activity*. A **GO annotation** (Figure 1B) is an association between a specific gene (or gene product) and a GO term, and should be interpreted as a statement that the specified gene product possesses the specified functional characteristic represented by the GO term. Each GO annotation includes the evidence upon which it is based. Because each GO annotation covers only a single characteristic of gene function, multiple GO annotations are generally required to completely describe the function of a gene product. **GO-CAMs** (Figure 1C) link multiple GO annotations together to create models of biological processes by 1) connecting the activities of more than one gene product together into causal networks, and 2) allowing the specification of the biological context (e.g. cell type, tissue type) in which the processes occur.

The GO knowledgebase is large and dynamic

For applications that use the components of the GO knowledgebase, it is crucial that the ontology and associated annotations represent the current state of knowledge, and are not just an archive of all public data. Therefore, all aspects of the GO knowledgebase are dynamic (ontology, annotations, GO-CAMs, links to external ontologies, etc.), and citable, versioned updates are released on a monthly basis. Below, we describe each component of the knowledgebase, focusing on recent changes made to improve the resource during the past two years. Statistics and descriptions given here are based on the GO release 2022-11-03 (http://release.geneontology.org/2022-11-03, doi:10.5281/zenodo.7407024).

Ontology

The ontology component of the GO knowledgebase consists of the terms used to describe functional characteristics of gene products, which are linked together by relations into a labeled directed acyclic graph (like a hierarchy but with multiple parentages allowed). It also includes term definitions, synonyms, and relations to terms from external ontologies. The GO is available in different editions, including (i) the “basic” edition, which includes only core relationship types; (ii) the core ontology, including additional relationship types; and (iii) the “go-plus” edition which includes relationships to terms in other ontologies. These editions are explained on the GO downloads page http://geneontology.org/docs/download-ontology/. The ontology contains 43,303 terms (Table 2), linked together by 88,099 relationships in the basic edition. When relationships to external terms are included, there are 121698 relationships; release statistics can be viewed at: http://geneontology.org/stats.

The GO is subject to constant review and revision to most accurately model current biological knowledge. Revision of the ontology includes the addition or obsoletion of terms and re-organization of the relationship structure. New GO terms are added to represent concepts

previously missing from the GO in response to published findings or when a branch of GO is revised. Terms may be obsoleted when unused or inconsistently used in annotation, when they are redundant with other terms, or during revision of specific branches of the ontology.

Most of the revisions in the structure of GO are in response to advances in biological knowledge, as well as improvements in the precision of newer experimental approaches. In addition, because many branches of the ontology have grown organically in a bottom-up fashion by accumulating specific individual term requests, we also perform systematic review aimed at improving consistency and clarity while reducing redundancy. Additional revisions are initiated by internal review and consistency and quality assurance checks. Revisions are also made following feedback from users. Whenever possible, changes are performed in collaboration with expert biocurators or domain specialists; recent examples include blood-brain barrier-related functions (Saverimuttu *et al.* 2021) and transcription factors (Gaudet *et al.* 2021).

At each release, we track all changes, and report on our website the number of added, obsoleted, and merged terms in the ontology. Table 1 shows the number of GO terms added and removed (merged or obsoleted) over the past two-year period, for each aspect of GO. In the molecular function and cellular component aspects of the ontology, term creation versus term obsoletion have approximately balanced each other, such that the number of terms in these two branches has remained roughly constant. The most significant changes have been in the biological process aspect of GO, with a net decrease of over 800 terms.

Many of these revisions result from global reviews of the ontology to address clear inconsistencies in usage, and changes in annotation practices. Terms that have been removed from the ontology over the last two years fall into several different categories, including:

- **terms that correspond to phenotypes**, and for which the understanding of the process was previously too incomplete to annotate to a different term. Examples include: *regulation of spindle density* (GO:0090225); *age-dependent general metabolic decline* (GO:0007571).
- terms that are combinations of multiple GO terms, that **can now be represented more precisely using GO-CAM models**. Examples: *chromatin remodeling in response to cation stress* (GO:0043156) and *regulation of cyclin-dependent protein serine/threonine kinase activity involved in G2/M transition of mitotic cell cycle* (GO:0031660).
- **revisions based on updated knowledge**, either by GO editors, by authoritative databases, or in the literature. For example alpha-taxilin (UniProt:P40222) was originally thought to be the high molecular weight interleukin-14 (Ambrus *et al.* 1993); an erratum was later published (Ambrus *et al.* 1996) indicating that the open reading frame had been incorrectly predicted. Hence all terms mentioning interleukin-14 have been obsoleted (*interleukin-14 binding* (GO:0019974), *interleukin-14 production* (GO:0032617), and 5 other terms). GO terms created from articles that have been retracted and for which no other supporting evidence exists are obsoleted, for example *CDP-acylglycerol O-arachidonoyltransferase activity* (GO:0047193) (Thompson and Zuk 1983). Some terms have been obsoleted from authoritative databases, for example

EC:1.3.1.59 was removed from the Enzyme Commission database, and the corresponding GO term *1,6-dihydroxy-5-methylcyclohexa-2,4-dienecarboxylate dehydrogenase activity* (GO:0018512) was obsoleted in GO.

- **single step reactions in the biological process aspect of the ontology:** There were many instances in the GO where a molecular function could be represented as both a molecular function and a biological process, for example 'histone kinase activity' and 'histone phosphorylation'. This was useful when fewer activities were characterized at the molecular level, and the best level of resolution for many experiments was that the gene has some uncharacterized role that led to histone phosphorylation, for example. However, with increasingly detailed molecular data, the redundancy between MF and BP annotations became unnecessary and the value of having a similar term in both aspects of the ontology led to inconsistency. This is an ongoing project and many BP terms still need to be obsoleted for this reason.
- **terms that conflate more than one ontological aspect**: *ubiquinone biosynthetic process monooxygenase activity* (GO:0015997) included a biological process within a molecular function; *MAP kinase phosphatase activity involved in regulation of innate immune response* (GO:0038078) included a molecular function within a biological process; and *histone deacetylation at centromere* (GO:0031059) represented all three aspects: a molecular function in the biological process branch of the ontology (histone acetylation) that also included cellular component information (centromere).
- **misclassified terms**, for example *urea homeostasis* (GO:0097274) and *creatinine homeostasis* (GO:0097273): while these compounds are important medical biomarkers, the normal process they measure is proper renal function, therefore these terms have been obsoleted. Annotations have been re-housed under *renal tubular secretion* (GO:0097254) (or one of its children) or removed if the paper supporting the annotation did not allow one to infer the process that affected the circulating levels of urea or creatinine.
- **reaction mechanisms:** *primary charge separation* (GO:0009766) and *enzyme active site formation* (GO:0018307) were obsoleted because they represent substeps of reactions which are beyond the scope of GO.
- **protein modifying activity terms that mention specific substrates**, for example *[cytochrome c]-arginine N-methyltransferase activity* (GO:0016275), which is captured by the more general *arginine N-methyltransferase activity* (GO:0016274). Substrates can be captured with the 'has input' relationship in GO-CAM models and in annotation extensions. The exception to this is the histone code: for GO to represent this important mechanism of gene expression and chromatin structure mechanism, specific activities are created for known histone modifications, for example: *histone H2AR3 methyltransferase activity* (GO:0070612) and *histone H3T3 kinase activity* (GO:0072354).
- **experimental assays and non-physiological substrates**: some experiments are easier to perform using analogs of physiological substrates. GO used to capture this information, but is now moving away from this and removing any term that represents an experiment rather than its biological conclusion. An example is *rubidium ion transport* (GO:0035826): rubidium is used as a tracer for potassium ions (Gill *et al.* 2004), but has

no physiological role in itself. Another example is *regulation of nucleosome density* (GO:0060303), which measures the degree of compaction of chromatin, and is a readout for heterochromatin assembly or disassembly.

Concomitantly with these term obsoletions, many new terms have been added to the ontology in the past two years. An example is *molecular condensate scaffold activity* (GO:0140693) for proteins that nucleate condensates that mediate liquid phase transition. This latter term represents a recent advance in the understanding of the organization of cellular biochemistry (Banani *et al.* 2017).

We have also clarified the level of specificity at which molecular function terms should be represented in GO. For example, we now strive to create GO terms that represent the range of *in vivo* substrate specificity of an enzyme or transporter. This is in contrast to earlier guidelines, in which a GO term was created for each separate molecular substrate tested in a single, isolated experimental assay or result, which could include non-physiological substrates. With recent improvements in experimental technologies and practices, it is now often possible to annotate with a concept that more closely matches the biological substrate specificity range of a protein. Therefore, while GO makes cross-references to EC (Enzyme Commission) (McDonald and Tipton 2014), Rhea (Bansal *et al.* 2022), KEGG (Kanehisa *et al.* 2022), and MetaCyc (Altman *et al.* 2013), GO does not necessarily create a different term for each of the reactions represented in these resources for each substrate on which a molecular function acts. For example, the GO term *3-oxoacyl-[acyl-carrier-protein] reductase (NADPH) activity* (GO:0004316) represents the fact that the same gene product has a broad specificity toward 3-oxo-acyl groups, and therefore we have obsoleted the more specific GO terms that refer to only one specific substrate, such as *3-oxo-cis-Delta9-hexadecenoyl-[acp] reductase activity* (GO:0102072), *3-oxo-glutaryl-[acp] methyl ester reductase activity* (GO:0102131), and *3-oxo-pimeloyl-[acp] methyl ester reductase activity* (GO:0102132). For broad-specificity enzymes and transporters, the activity on a specific substrate in a specific pathway can be captured by biocurators in a GO-CAM (Thomas *et al.* 2019) or an annotation extension (Gene Ontology Consortium 2010) rather than in a GO term.

Annotations

A GO annotation is a statement asserting that a particular gene or gene product has a particular functional characteristic (GO term), examples are shown in Figure 1B. New annotations are continually added to the knowledgebase. In the past two years, experimentally-supported gene function annotations have been added from over 10,000 scientific papers. As of November 2022, the GO knowledgebase contains experimental knowledge from almost 173,000 papers. GO annotations derived from experimental data are added primarily by the annotation groups in the GO Consortium, which typically curate biological knowledge by organism (Table 2).

GO annotations are also regularly reviewed, and may be edited or removed from the knowledgebase for various reasons, particularly when ontology terms are revised (see

"Ontology" section above) or when annotations are invalidated by later experimental data. Annotations to terms that will be obsoleted are manually reviewed and annotations are made to a different term whenever possible. For example, when we edited the ontology for histone modifications, over 2,000 annotations to the obsoleted terms were manually reviewed, and histone modifying enzymes were re-annotated to the appropriate molecular function term, while annotations from indirect effects were either removed or re-annotated to different, appropriate GO terms. More minor annotation reviews occur regularly.

The Phylogenetic Annotation with GO project (see below) involves an integrated biocurator review of annotations that has provided additional quality control. The GO user community also plays an important role in identifying incorrect annotations. Because each annotation can be traced to the published paper containing the underlying evidence or describing a method used to infer the annotation, users can quickly verify the accuracy of a given annotation. Potential errors can be reported by clicking on the “Help” link at the top of the GO homepage (http://geneontology.org). In addition, authors of a paper used to create GO annotations can easily retrieve and review all annotations from a given paper and suggest changes; this can be done from the PubMed abstract page (e.g. the PubMed page for PMID:20516198 (Lydeard *et al.* 2010)) by clicking on LinkOut and then the “Gene Ontology” link.

Phylogenetic Annotations as a source of highly reviewed annotations. The Phylogenetic Annotation using GO (PAN-GO) project creates a set of biocurator-reviewed, selected GO annotations. This creates sets of improved, augmented GO annotations for genes in the reviewed families. The PAN-GO process is described in detail in (Gaudet *et al.* 2011). Briefly, using the PAINT software tool, a biocurator reviews all experimentally-supported GO annotations collected for all members of a protein family, in the context of a phylogenetic tree from the PANTHER resource (Thomas *et al.* 2022). They then select the most informative and non-redundant GO terms that represent the gene's functional characteristics. Biocurators then model the evolution of these characteristics in the tree by specifying branches along which the GO terms were gained or lost, taking into account events such as duplications, mutations, horizontal gene transfers, as well as taxonomic specificity. This allows for different members of the same family to be annotated with different GO terms when justified by the experimental data. All PAN-GO annotations can be traced to experimental evidence in one or more related genes. To date, a total of 8,196 protein families (out of 11,719 families with experimental data) have been curated. The PAN-GO curation effort has prioritized human gene-containing families, though many other families have also been curated. As a result, annotation coverage of a genome generally depends on how closely related it is to humans. PAN-GO annotations are available for 82% of human genes (compared to 68% with experimental evidence alone). Other vertebrate genomes have similarly high coverage, with genomes from other taxa covered at lower but still substantial levels (Tables 3 and 4). PAN-GO annotations are updated at each GO release, and are included in the standard, downloadable GO annotation files. These annotations can be identified by the “IBA” (inferred from biological ancestor) evidence code, and are available for the 142 organisms included in PANTHER gene families (http://pantherdb.org/panther/summaryStats.jsp).

Protein binding and protein-containing complex annotations. We suggest that users should be particularly cautious when using GO annotations directly to the term *protein binding* (GO:0005515; see Table 2). These are highly specific annotations that include the protein binding partner in another field of the annotation (not in the GO term itself), and should not be used in applications such as gene set enrichment analysis. Instead, they are recommended for applications such as protein-protein interaction network construction for human proteins (which represent the vast majority of direct protein binding annotations in the knowledgebase). Since all protein functions encompass some type of binding (to a substrate, or to another protein), GO strives to describe the molecular activity of proteins using at least one term that is not only under the *binding* branch of GO; see also "non-catalytic molecular functions" section above. Therefore, *binding* (GO:0005488) in isolation can be considered a limited functional description and is represented as a distinct branch of GO molecular function.

Annotation evidence. All annotations are supported by evidence, comprising two fields in the annotation file (Figure 1B): an *evidence code* that describes the type of evidence and a *reference* that lists a persistent identifier for tracing the source (provenance) of the original data. It has often been asserted that the most reliable annotations are those made using an experimental evidence code. However, we suggest that users take into account **the type of experimental evidence** and the **level of review of the annotation** (Table 5). Some types of experimental evidence, such as inference from a gene expression pattern (IEP), mutant phenotype (IMP) or genetic interaction (IGI) can often be suggestive of function but not definitive when considered in isolation; other annotations for the same gene are often useful to help interpret these annotations. “High-throughput” evidence codes should be treated with particular care. These codes (beginning with the letter H) denote experiments in which many genes are analyzed at the same time, and these annotations are not individually reviewed by either the paper’s authors or GO Consortium biocurator (Attrill *et al.* 2019). Conversely, many non-experimental evidence types are carefully reviewed by experts. Phylogenetic annotations (IBA evidence code) are based on integration and expert assessment of experimental annotations, and thus are *individually reviewed* twice: once in making the annotation from published experimental results, and once in the context of all annotations for related genes (Gaudet *et al.* 2011). While annotations using the Inferred from electronic annotation (IEA) evidence code are considered automated, most implement expert review of a subset of annotations to minimize false positives (for example, UniRule (MacDougall *et al.* 2021) and InterPro2GO(Paysan-Lafosse *et al.* 2022)). The GOC considers these annotations to be accurate though they are often less specific than other annotations

GO evidence codes correspond to a subset of the terms found in the Evidence and Conclusion Ontology (ECO) (Nadendla *et al.* 2022). Combinations of particular GO internal references (GO_REFs) and evidence codes are also mapped to specific ECO terms (https://github.com/evidenceontology/evidenceontology/blob/master/gaf-eco-mapping.txt). Users needing to map granular ECO terms to GO evidence code abbreviations can use the mapping file provided by ECO (https://github.com/evidenceontology/evidenceontology/blob/master/gaf-eco-mapping-derived.txt).

GO Causal Activity Models (GO-CAMs)

GO-CAMs are models of causal influences between gene products (Thomas *et al.* 2019), or pathways. More precisely, a GO-CAM links the **activities** (GO molecular functions) of gene products together by **causal relations** that specify the effect of one activity on the other. Each element of a GO-CAM is an instance of an ontology class or other standard database identifier, so GO-CAMs are highly structured and amenable to computational analysis. The basic unit of a GO-CAM is a "gene product activity unit", which combines a GO molecular function annotation (molecular activity), together with GO cellular component (location) and GO biological process (larger functional module) annotations that provide the biological context of the activity. The context can be further specified with other ontologies to capture cell type (using the Cell Type Ontology (Diehl *et al.* 2016)), tissue/anatomical location (using several different ontologies depending on the species, e.g. Uberon (Mungall et al. 2012) for most vertebrates, other metazoan ontologies such as the *Drosophila* anatomy ontology (Costa et al. 2013), *C. elegans* anatomy ontology (Lee and Sternberg 2003), or non-animal ontologies as the Plant Ontology (Cooper et al. 2018)), or a temporal period (e.g. GO biological phase). Activity units are linked together by causal relationships from the Relations Ontology (Smith *et al.* 2005) to capture how they interact to impact larger pathways, modules or processes.

As of November 2022, GO Consortium annotation groups have created over 300 GO-CAM models that describe molecular pathways (defined as containing at least three distinct gene product activities linked into a causal chain). These models reflect curation priorities of the contributing groups. Most of the available GO-CAMs are for processes in human or mouse, with a limited number in zebrafish, *D. melanogaster* and *C. elegans*. Many of the human GO-CAMs describe chromatin-mediated regulation of gene expression and immune response pathways, while the mouse GO-CAMs focus on metabolic and signaling pathways. GO-CAMs are accessible from the GO website homepage, by clicking on the "Browse GO-CAMs" link. GO-CAMs can be viewed as pathway diagrams (Figure 2) and are currently available on GitHub at https://github.com/geneontology/noctua-models.

Community Collaborations

The GO Consortium collaborates with experts in specific areas of molecular and cellular biology to systematically update and improve their representation in the ontology and the corresponding GO annotations and GO-CAMs. We recently revised the representation for transcription factors and transcriptional regulation in collaboration with the GREEKC Consortium (Kuiper *et al.* 2022). Additional collaborative projects include working with the DisProt project (Quaglia *et al.* 2022) on improving the ontology and annotations for intrinsically disordered proteins; revising processes that involve molecular pathways between interacting species, such as viral infection processes; and integrating the gene ontology and annotations with external biochemical databases.

In 2021 the Gene Ontology started a collaboration with DisProt (https://disprot.org/) – the gold standard database of manually curated annotations from the literature for Intrinsically Disordered Proteins (IDPs). IDPs lack a stable three-dimensional structure and are characterized by highly flexible and unstructured segments, i.e. intrinsically disordered regions (IDRs). DisProt has developed a custom ontology, the Intrinsically Disordered Proteins Ontology (IDPO), and used it to annotate the structural states of IDPs. The GO Consortium and DisProt have collaborated to refactor IDPO and map the IDPO terms to GO terms whenever possible (those related to functions and interactions of IDPs). The collaboration between the GO Consortium and the DisProt database included the creation and addition of new GO terms to align with already existing IDPO terms that were not yet available in GO. These newly created terms also include the *molecular function activator* (GO:0140677) and *molecular function inhibitor* (GO:0140678) terms, used to annotate molecular function regulators that activate/inhibit or increase/decrease the activities of their targets via non-covalent binding that does not result in covalent modification to the target. This collaboration resulted in more accurate and detailed annotation of the modes of action of IDPs, e.g. *localization* (GO:0051179, IDPO:00010) and *DNA binding* (GO:0003677, IDPO:00065), as well as providing GO annotations. Currently more than 1,000 expert-curated annotations from DisProt are available in the GO knowledgebase, comprising more than 860 molecular functions, 200 biological processes and 10 cellular component annotations. The only terms in IDPO that could not be mapped to GO were those describing self-regulatory (e.g. *self-activation* and *self-inhibition*) and intrinsic disorder-specific functions (i.e. *entropic chains*), so these annotations are available only in DisProt.

Multiorganism interactions

A group that includes experts from within and outside the GO Consortium have been working together to improve and simplify the representation of interactions between organisms, including medically and agriculturally important host-pathogen interactions. Examples of these interactions include how a symbiont such as a virus enters its host, how the host's immune response recognizes and defends the body against a potentially harmful organism, and also beneficial interactions such as how plants form a symbiosis with nitrogen-fixing bacteria. The goal of this project is to revise the host-symbiont branch of GO *biological process* to reflect the current scientific knowledge in the field, and to ensure that genes are properly annotated to the new ontology terms and structure, building on previous work undertaken as part of the PAMGO consortium (Tyler *et al.* 2009). Symbionts in GO are broadly defined to include pathogens that infect a host organism. We expect that this revision will improve GO-based analyses of molecular studies of pathogens, the mechanisms by which they infect host cells, and host response processes. A major change is that the branch of GO under *biological process involved in interspecies interaction between organisms* (GO:0044419) has been reorganized. It now reflects important concepts such as the types of biological programs used by symbionts to enable infection, and by hosts to prevent or manage infection, such as *disruption of cellular component of another organism* (GO:0140975), *formation of structure involved in a symbiotic process* (GO:0044111), *killing of cells of another organism* (GO:0031640), and *modulation of*

process of another organism (GO:0035821). Each of these terms has multiple, more specific subclass terms.

One challenge in this area was that some previous GO annotations for pathogen genes used terms that apply to normal host processes, such as regulation of defense response processes. Thus, it was not clear whether the pathogen gene was regulating its own defense process or that of a host. With the new ontology terms and structure, these distinctions are clear for both GO biocurators and users of GO. In general, it was important to clearly represent that certain symbiont-initiated processes hijack various host cellular processes. This includes mechanisms to enter and exit the cell, either by binding to host membrane proteins or using the intracellular transport machinery, and using the host cellular machinery for genome replication, as well as transcription and translation. We have obsoleted terms that do not clearly distinguish hijacking with the functions that a host gene performs for the host organism, such as *dissemination or transmission of symbiont from host by vector* (GO:0044008) and *positive regulation of viral release from host cell* (GO:1902188). Conversely, a pathogenic symbiont triggers innate responses in the host that are not the evolved role of these symbiont proteins, such as *induction by symbiont of host cytokine production* (GO:0036523) and *pathogen-associated molecular pattern dependent induction by symbiont of host innate immune response* (GO:0052033) - these are not functions that a symbiont protein performs to enable its own survival and reproduction.

Integration with biochemical knowledgebases

GO also works closely with specialized databases and knowledgebases to ensure knowledge is both complete and consistent. For accurate representation of biochemical aspects of gene function, we work closely with the Rhea database of reactions (Bansal *et al.* 2022) and the ChEBI ontology of chemical entities (Hastings *et al.* 2016). Rhea provides precise representations of *in vivo* biochemical reactions, including precise chemical entity participants and their stoichiometry. Rhea uses ChEBI terms to represent chemical entities in a standardized, consistent manner. The Rhea database overlaps in content with the catalytic activity branch of the Gene Ontology, but provides additional detailed reaction information, and in some cases provides additional specificity. We have improved GO mappings to Rhea, which now covers 4399 GO catalytic activities (in the MF branch of GO). These mappings allow for non-exact matches when the chemical specificity differs between GO and Rhea. For example, Rhea has two reactions, each referring to a different type of beta glucoside (RHEA:69647 and RHEA:69655, narrow match), whereas GO:0008422, beta-glucosidase activity, covers both substrates, as no known enzyme is specific for just one of them. We have recently used the Rhea-GO mappings to include additional linkages between GO molecular function terms and ChEBI terms in the go-plus release (see below). Previously ChEBI terms were linked only to general terms in the GO biological process branch (e.g. between folate transport and folate), but the additional Rhea linkages have added a total of 4334 distinct chemical entities linked via 20,307 relationships. The extensive linkage to chemical entities opens opportunities for using GO in other applications, e.g. metabolomics analyses.

Accessing and downloading GO data

Browsing GO and its annotations

GO and associated annotations can be searched directly from the Gene Ontology home page (http://geneontology.org/), queried using the AmiGO browser (http://amigo.geneontology.org/amigo) or the QuickGO tool (https://www.ebi.ac.uk/QuickGO/) (Munoz-Torres and Carbon 2017). Gene set enrichment analysis is also directly accessible from the Gene Ontology home page, which runs the PANTHER gene analysis tool at http://pantherdb.org/webservices/go/overrep.jsp (Mi *et al.* 2019).

Ontology downloads

GO provides three editions of the ontology on its download site (http://geneontology.org/docs/download-ontology/) to accommodate various applications: go-basic, go, and go-plus (Table 6). All GO terms, including obsolete terms and term metadata such as definitions, cross references, synonyms, are available in all three editions. These editions differ in the set of relations they contain:

- **go-basic** contains the types of information that has been available for GO from the beginning of the project, hence only contains is a, part of, regulates, negatively regulates and positively regulates relationships and excludes relationships that cross different aspects (BP, MF or CC) of the ontology. This edition of the ontology is guaranteed to be acyclic and can safely be used to selectively propagate annotations across any relation. It is recommended for most GO-based annotation tools.
- **go** additionally includes has part and occurs in relationships that link terms across different aspects of the ontology (for example, a biological process can have a *has part* relation to a molecular function term, or an *occurs in* relation to a cellular component). This edition is not acyclic and annotations should not be propagated across all the relationship types it contains. This edition should not be used in most software tools that rely on the Gene Ontology.
- **go-plus** is the fully axiomatized edition of the ontology, and includes cross-ontology relationships to external ontologies including ChEBI, Cell Ontology and Uberon.

Ontology subsets (GO slims)

GO subsets are condensed versions of the GO containing a portion of the terms, which are specified by tags within the ontology that indicate if a given term is a member of a particular subset. GO subsets are particularly useful for providing a global overview of the functions of all the genes in a genome, and even for all the functions of a single gene. range of functions and processes found in a given clade or organism’s genome. We have recently revised the "GO Generic subset", a subset maintained by the GO consortium that aims to be general and applicable to any species. We have tested that the subset covers as many gene products as possible in various organisms (human, *D. melanogaster*, fission yeast, *A. thaliana*, *E. coli*) with

as little redundancy as possible. This new GO Generic Subset contains 75 biological process terms, 40 molecular function terms, and 29 cellular component terms.The GO generic subset can be accessed at http://current.geneontology.org/ontology/subsets/goslim_generic.obo. Versions in .owl, .json and .tsv are also available from http://current.geneontology.org/ontology/subsets/index.html.

As part of the Alliance of Genome Resources, we have developed a widget that provides a graphical visualization of a gene's function in a 'ribbon'-like display (Figure 3). The widget can be customized to use any GO subset, and uses the goslim_agr subset by default. This widget is implemented in the Alliance gene pages and in the UniProt entry pages. It accesses GO annotations using the GO API (application programming interface) and can be easily added to any webpage.

GO Annotations

The two major sites for downloading GO annotations are geneontology.org and UniProt-GOA. **geneontology.org** is the website developed by the GO Consortium. This downloads site (http://current.geneontology.org/products/pages/downloads.html) provides a total of 7.5 million human and model organism annotations contributed by multiple groups. It contains all manually reviewed GO annotations, and electronic (computationally predicted) annotations for the most commonly used organisms. For model organisms, all annotations use gene identifiers from the authoritative database (for example, FlyBase (FBgn), WormBase (WBGene), and SGD (S)). Human and other organisms without an authoritative dedicated database are represented by UniProtKB accession numbers. For these organisms, the GO website provides annotations to UniProt reference proteomes (https://www.uniprot.org/help/reference_proteome), which are generally one entry per gene, thus limiting redundancy in annotations. **UniProt-GOA** (https://www.ebi.ac.uk/GOA/uniprot_release) contains 1 billion annotations for all entries in UniProt (1,264,340 taxa), covering both reviewed (Swiss-Prot) entries of UniProt, and unreviewed (TrEMBL) entries. All annotations for model organism genes are converted to UniProt protein identifiers. For most organisms, all annotations are electronic annotations generated via various pipelines (see above for evidence codes and references for different methods). In addition to these resources, GO annotations are also viewable in a number of biological databases, including model organism databases, UniProt (UniProt: the universal protein knowledgebase 2017), NCBI (Sayers *et al.* 2020), and The Alliance of Genome Resources (Alliance of Genome Resources Consortium 2022). These sites show GO annotations in the broader context of a gene product's expression pattern, phenotypes, metabolic and signaling pathways, etc.

Conclusions and future directions

The extensive and wide-ranging use of the GO knowledgebase, evidenced by its recent, peer-reviewed designation as a Global Core Biodata Resource (https://globalbiodata.org/scientific-activities/global-core-biodata-resources/), demands its continued development and expansion. We are focusing on several high priority areas of development for the near future. For pathways, we will continue to accumulate GO-CAM models. The UniProt/Swiss-Prot curation team has ramped up their production of GO-CAM models and we expect to add models at a rapid rate. In parallel, we have started converting Reactome pathways into GO-CAMs (Good *et al.* 2021) and expect to release GO-CAM representations of most Reactome metabolic pathways in the near future. This will provide a complementary, causal flow representation of the chemical reaction-centered representation in Reactome. Conversion of Reactome signaling pathways is more challenging, and will be released somewhat later. We are also working on converting the YeastPathways resource (https://pathways.yeastgenome.org) into GO-CAMs, making a large number of yeast metabolic pathways available. The increasing number of GO-CAM models will allow us to expand on the utility of these highly structured pathway and process representations. Some potential areas are automated pathway visualization; using the causal links and more granular gene sets to enhance enrichment analysis; and better generation of automated descriptions of gene function (e.g., (Kishore *et al.* 2020)).

With respect to ontology development, in addition to continuing to revise the ontology in response to recent discoveries, we see an immediate need for clearly delineating the level of biological organization at which a function is described. This includes distinguishing molecular functions from biological processes, and distinguishing biological processes that occur at the level of individual cells, versus those that occur at the level of multicellular organisms. For example, the term homeostasis–the maintenance of a roughly steady level of a molecule or ion–is used very broadly in the literature to refer to both processes that maintain a steady-state level within a cell, and processes that maintain a steady state in blood or other fluid that is transported within a multicellular organism. Even in some publications, it is difficult to know which type of homeostasis is being tested.

We will continue to make the GO knowledgebase easier to use, and more community-driven. One near-term priority is to make annotations available for download by species, with a single identifier for each distinct gene. We are also planning to create quick-start guides for common GO use cases, in both written and video form. The immense user base of the GO and the need for much improvement and extension drives us to consider how to expand the number of people that contribute to the GO. From its inception the GO has been a large, open, community project. However, we are planning additional routes through which the broader GO user community can contribute their expert feedback and knowledge to GO, improving the resource for all users. For now, users are encouraged to contact the GO Helpdesk (http://help.geneontology.org/) with any questions, or to report any GO ontology terms or annotations that may be inaccurate or difficult to interpret.

Data Availability

All Gene Ontology code and resources are freely available for download and reuse. Software (https://github.com/geneontology/) is under the BSD 3-Clause open-source license. Downloads are available under the CC BY 4.0 license from http://geneontology.org/docs/downloads/

Funding

The core funding for the GOC is from the National Human Genome Research Institute (U41HG002273, U24HG012212). Curation activities supported by National Human Genome Research Institute grants U24HG002659 (ZFIN), U24HG002223 (WormBase), U41HG000739 (FlyBase), U24HG001315 (SGD), U24HG000330 (MGD), U24HG012198 (Reactome curation), U24HG011851 (Reactome - GO harmonization) and grant R01HL064541 from the National Heart, Lung and Blood Institute (RGD). Additional funding for GO curation at FlyBase is provided by UK Medical Research Council Award MR/W024233/1. PomBase is supported by Wellcome Trust 218236/Z/19/Z. Xenbase is supported by grant P41 HD064556 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. Functional Gene Annotation, University College London is supported by National Institute for Health Research University College London Hospitals Biomedical Research Centre**.** Planteome and Plant Reactome are supported by National Science Foundation awards #1340112, #1127112, and USDA-ARS.Some software development was funded by U24HG010859 (Alliance of Genome Resources Central). The TAIR project is funded by academic, institutional, corporate, and individual subscriptions; TAIR is administered by the 501(c)(3) non-profit Phoenix Bioinformatics. Chris Mungall, Seth Carbon, and Sierra Moxon were supported in part by Director, Office of Science, Office of Basic Energy Sciences of the U.S. Department of Energy Contract No. DE-AC02-05CH11231

UniProt is funded by National Human Genome Research Institute (NHGRI), Office of Director [OD/DPCPSI/ODSS]; National Institute of Allergy and Infectious Diseases (NIAID), National Institute on Aging (NIA), National Institute of General Medical Sciences (NIGMS), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Eye Institute (NEI), National Cancer Institute (NCI), National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health [U24HG007822]; Biotechnology and Biological Sciences Research Council [BB/T010541/1, BB/S01781X/1]; Open Targets; Swiss Federal Government through the State Secretariat for Education, Research and Innovation SERI; European Molecular Biology Laboratory core funds.

Summary

The Gene Ontology (GO) knowledgebase has broad applications in genetic and genomic research, and has been continually updated and improved for more than 20 years. We describe

the latest improvements to the GO resource, as well as giving an overview of the current contents of the knowledgebase. We also include guidance on the use of GO, annotations and the growing number of pathways represented as GO-Causal Activity Models (GO-CAMs).

References

Alliance of Genome Resources Consortium. Harmonizing model organism data in the Alliance of Genome Resources. *Genetics* 2022;**220**, DOI: 10.1093/genetics/iyac022.

Altman T, Travers M, Kothari A *et al.* A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics* 2013;**14**:112.

Ambrus JL Jr, Pippin J, Joseph A *et al.* Identification of a cDNA for a human high-molecular-weight B-cell growth factor. *Proc Natl Acad Sci U S A* 1993;**90**:6330–4.

Ambrus JL Jr, Pippin J, Joseph A *et al.* Identification of a cDNA for a human high molecular-weight B-cell growth factor. *Proc Natl Acad Sci U S A* 1996;**93**:8154.

Ashburner M, Ball CA, Blake JA *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9.

Attrill H, Gaudet P, Huntley RP *et al.* Annotation of gene product function from high-throughput studies using the Gene Ontology. *Database* 2019;**2019**, DOI: 10.1093/database/baz007.

Banani SF, Lee HO, Hyman AA *et al.* Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol* 2017;**18**:285–98.

Bansal P, Morgat A, Axelsen KB *et al.* Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res* 2022;**50**:D693–700.

Basu S, Fey P, Jimenez-Morales D *et al.* dictyBase 2015: Expanding data and annotations in a new software environment. *Genesis* 2015;**53**:523–34.

Bult CJ, Blake JA, Smith CL *et al.* Mouse Genome Database (MGD) 2019. *Nucleic Acids Res* 2019;**47**:D801–6.

Cerqueira GC, Arnaud MB, Inglis DO *et al.* The Aspergillus Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. *Nucleic Acids Res* 2014;**42**:D705–10.

Davis P, Zarowiecki M, Arnaboldi V *et al.* WormBase in 2022-data, processes, and tools for analyzing Caenorhabditis elegans. *Genetics* 2022;**220**, DOI: 10.1093/genetics/iyac003.

Del Toro N, Shrivastava A, Ragueneau E *et al.* The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res* 2022;**50**:D648–53.

Diehl AD, Meehan TF, Bradford YM *et al.* The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J Biomed Semantics* 2016;**7**:44.

Duck G, Nenadic G, Filannino M *et al.* A Survey of Bioinformatics Database and Software Usage through Mining the Literature. *PLoS One* 2016;**11**:e0157989.

Fabregat A, Jupe S, Matthews L *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res* 2018;**46**:D649–55.

Fortriede JD, Pells TJ, Chu S *et al.* Xenbase: deep integration of GEO & SRA RNA-seq and ChIP-seq data in a model organism database. *Nucleic Acids Res* 2020;**48**:D776–82.

Fungal-Anatomy-Ontology: A Structured Controlled Vocabulary for the Anatomy of Fungi. Github

Gaudet P, Livstone MS, Lewis SE *et al.* Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief Bioinform* 2011;**12**:449–62.

Gaudet P, Logie C, Lovering RC *et al.* Gene Ontology representation for transcription factor functions. *Biochim Biophys Acta Gene Regul Mech* 2021:194752.

Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res* 2010;**38**:D331–5.

Gill S, Gill R, Wicks D *et al.* Development of an HTS assay for Na+, K+-ATPase using nonradioactive rubidium ion uptake. *Assay Drug Dev Technol* 2004;**2**:535–42.

Gkoutos GV, Schofield PN, Hoehndorf R. The anatomy of phenotype ontologies: principles, properties and applications. *Brief Bioinform* 2018;**19**:1008–21.

Good BM, Van Auken K, Hill DP *et al.* Reactome and the Gene Ontology: Digital convergence of data resources. *Bioinformatics* 2021, DOI: 10.1093/bioinformatics/btab325.

Haendel MA, Balhoff JP, Bastian FB *et al.* Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *J Biomed Semantics* 2014;**5**:21.

Haendel MA, Neuhaus F, Osumi-Sutherland D *et al.* CARO – The Common Anatomy Reference Ontology. In: Burger A, Davidson D, Baldock R (eds.). *Anatomy Ontologies for Bioinformatics: Principles and Practice*. London: Springer London, 2008, 327–49.

Harris MA, Rutherford KM, Hayles J *et al.* Fission stories: Using PomBase to understand *Schizosaccharomyces pombe* biology. , DOI: 10.1101/2021.09.07.459264.

Hastings J, Owen G, Dekker A *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res* 2016;**44**:D1214–9.

Howe DG, Ramachandran S, Bradford YM *et al.* The Zebrafish Information Network: major gene page and home page updates. *Nucleic Acids Res* 2021;**49**:D1058–64.

Kanehisa M, Furumichi M, Sato Y *et al.* KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res* 2022, DOI: 10.1093/nar/gkac963.

Keseler IM, Mackie A, Santos-Zavaleta A *et al.* The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Res* 2017;**45**:D543–50.

Kishore R, Arnaboldi V, Van Slyke CE *et al.* Automated generation of gene summaries at the Alliance of Genome Resources. *Database* 2020;**2020**, DOI: 10.1093/database/baaa037.

Koopmans F, van Nierop P, Andres-Alonso M *et al.* SynGO: An Evidence-Based, Expert-

Curated Knowledge Base for the Synapse. *Neuron* 2019;**103**:217–34.e4.

Kuiper M, Bonello J, Fernández-Breis JT *et al.* The gene regulation knowledge commons: the action area of GREEKC. *Biochim Biophys Acta Gene Regul Mech* 2022;**1865**:194768.

Lamesch P, Berardini TZ, Li D *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* 2012;**40**:D1202–10.

Lang OW, Nash RS, Hellerstedt ST *et al.* An Introduction to the Saccharomyces Genome Database (SGD). *Methods Mol Biol* 2018;**1757**:21–30.

Lydeard JR, Lipkin-Moore Z, Sheu Y-J *et al.* Break-induced replication requires all essential DNA replication factors except those specific for pre-RC assembly. *Genes Dev* 2010;**24**:1133–44.

MacDougall A, Volynkin V, Saidi R *et al.* UniRule: a unified rule resource for automatic annotation in the UniProt Knowledgebase. *Bioinformatics* 2021;**36**:5562.

McDonald AG, Tipton KF. Fifty-five years of enzyme classification: advances and difficulties. *FEBS J* 2014;**281**:583–92.

McIntosh BK, Renfro DP, Knapp GS *et al.* EcoliWiki: a wiki-based community resource for Escherichia coli. *Nucleic Acids Res* 2012;**40**:D1270–7.

Meldal BHM, Bye-A-Jee H, Gajdoš L *et al.* Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes. *Nucleic Acids Res* 2019;**47**:D550–8.

Mi H, Muruganujan A, Huang X *et al.* Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat Protoc* 2019;**14**:703–21.

Mungall CJ, Batchelor C, Eilbeck K. Evolution of the Sequence Ontology terms and relationships. *J Biomed Inform* 2011;**44**:87–93.

Munoz-Torres M, Carbon S. Get GO! Retrieving GO Data Using AmiGO, QuickGO, API, Files, and Tools. *Methods Mol Biol* 2017;**1446**:149–60.

Nadendla S, Jackson R, Munro J *et al.* ECO: the Evidence and Conclusion Ontology, an update for 2022. *Nucleic Acids Res* 2022;**50**:D1515–21.

Natale DA, Arighi CN, Blake JA *et al.* Protein Ontology (PRO): enhancing and scaling up the representation of protein entities. *Nucleic Acids Research* 2017;**45**:D339–46.

Paysan-Lafosse T, Blum M, Chuguransky S *et al.* InterPro in 2022. *Nucleic Acids Res* 2022, DOI: 10.1093/nar/gkac993.

Quaglia F, Mészáros B, Salladini E *et al.* DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Res* 2022;**50**:D480–7.

Radivojac P, Clark WT, Oron TR *et al.* A large-scale evaluation of computational protein function prediction. *Nat Methods* 2013;**10**:221–7.

Ramsey J, McIntosh B, Renfro D *et al.* Crowdsourcing biocuration: The Community Assessment

of Community Annotation with Ontologies (CACAO). *PLoS Comput Biol* 2021;**17**:e1009463.

Saverimuttu SCC, Kramarz B, Rodríguez-López M *et al.* Gene Ontology curation of the blood-brain barrier to improve the analysis of Alzheimer's and other neurological diseases. *Database* 2021;**2021**, DOI: 10.1093/database/baab067.

Sayers EW, Beck J, Brister JR *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2020;**48**:D9–16.

Sian L, Agapite J, Attrill H *et al.* FlyBase: a guided tour of highlighted features. *Genetics* 2022;**220**:iyac035.

Skrzypek MS, Binkley J, Binkley G *et al.* The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. *Nucleic Acids Res* 2017;**45**:D592–6.

Smith B, Ceusters W, Klagges B *et al.* Relations in biomedical ontologies. *Genome Biol* 2005;**6**:R46.

Smith JR, Hayman GT, Wang S-J *et al.* The Year of the Rat: The Rat Genome Database at 20: a multi-species knowledgebase and analysis platform. *Nucleic Acids Res* 2020;**48**:D731–42.

Thomas PD. The Gene Ontology and the Meaning of Biological Function. *Methods Mol Biol* 2017;**1446**:15–24.

Thomas PD, Ebert D, Muruganujan A *et al.* PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci* 2022;**31**:8–22.

Thomas PD, Hill DP, Mi H *et al.* Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. *Nat Genet* 2019;**51**:1429–33.

Thompson W, Zuk RT. Acylation of CDP-monoacylglycerol cannot be confirmed. *J Biol Chem* 1983;**258**:9623.

Tyler BM, Collmer A, Collmer CW *et al.* The PAMGO consortium: unifying themes in microbe-host associations identified through the gene ontology. *BMC Microbiol* 2009;**9**.

UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2017;**45**:D158–69.

Walls RL, Cooper L, Elser J *et al.* The Plant Ontology Facilitates Comparisons of Plant Development Stages Across Species. *Front Plant Sci* 2019;**10**:631.

Winsor GL, Lo R, Ho Sui SJ *et al.* Pseudomonas aeruginosa Genome Database and PseudoCAP: facilitating community-based, continually updated, genome annotation. *Nucleic Acids Res* 2005;**33**:D338–43.

Tables

Table 1. Changes to GO terms in the past two year period. The ontology has undergone substantial revision and improvement, with nearly 2,000 terms added or removed.

GO aspect	Total number of terms	Added terms	Obsoleted terms	Merged terms[1]
Molecular Function	11,271	315	65	143
Cellular Component	4,039	34	19	162
Biological Process	27,993	217	782	254

[1]Also includes obsoleted terms that have been replaced by another term.

Table 2. Group contributing literature-based annotations. Includes all annotations traceable to the literature (EXP, including HTP, TAS, NAS, IC, see http://geneontology.org/docs/guide-go-evidence-codes; see below for information). Direct annotations to the term "protein binding" are listed separately, since without information about interacting partner(s), protein binding represents an activity that most proteins possess and therefore the GO class itself provides little information (see text for further description). The statistics for groups that have contributed more than 700 manual annotations. Other contributing groups include: HGNC, JaponicusDB, PHI-base, PAMGO, JCVI, MENGO, and GDB. Current GO Consortium members are labeled with an asterisk. See http://geneontology.org/docs/annotation-contributors/ for more details.

Group	Organism or area of focus	Number of literature-based annotations, excluding direct protein binding	Number of literature-based annotations directly to protein binding
UniProt* (UniProt: the universal protein knowledgebase 2017)	human, and also a wide variety of organisms not covered by other GOC members	185121	30927
MGI* (Bult et al. 2019)	mouse	106435	8051
Reactome* (Fabregat et al. 2018)	human pathways	92178	6
TAIR* (Lamesch et al. 2012)	A. thaliana (model plant)	64633	4695
FlyBase* (Sian et al. 2022)	D. melanogaster (fruit fly)	55203	892
UCL*	human	54595	2935
RGD* (Smith et al. 2020)	rat	47694	1894

SGD* (Lang et al. 2018)	S. cerevisiae (Baker's yeast)	48811	165
ZFIN* (Howe et al. 2021)	zebrafish	28261	488
PomBase* (Harris et al.)	S. pombe (fission yeast)	26128	2201
GeneDB	microbial pathogens	23884	756
ComplexPortal* (Meldal et al. 2019)	protein complexes	18343	0
WormBase* (Davis et al. 2022)	C. elegans (nematode)	17171	560
CGD* (Skrzypek et al. 2017)	C. albicans (yeast pathogen)	17113	0
EcoCyc* (Keseler et al. 2017)	E. coli (bacterium)	13372	829
AgBase	Agricultural animals, primarily chicken	11198	1110
dictyBase* (Basu et al. 2015)	D. discoideum (slime mold)	9615	844
HPA	human protein subcellular localization	9963	0
SynGO (Koopmans et al. 2019)	neuron-neuron synapses	9552	0
PINC	human and mouse	6746	0
MTBBASE	M. tuberculosis (bacterial pathogen)	6160	463
IntAct* (Del Toro et al. 2022)	protein-protein interactions	4849	216488
CAFA (Radivojac et al. 2013)	various	4818	371
CACAO* (Ramsey et al. 2021)	various	4382	0
AspGD (Cerqueira et al. 2014)	A. niger (fungal pathogen)	4099	0
PseudoCAP (Winsor et al. 2005)	P. aeroginosa (bacterium)	2323	0
EcoliWiki* (McIntosh et al. 2012)	E. coli (bacterium)	2123	55
TIGR	bacteria	2150	0
GO_Central*	various	3643	160

CollecTF	bacterial transcription factors	1850	0
NTNU_SB	human, mouse, rat transcription factors	1733	0
GR	rice	1260	0
SGN	tomato	1255	0
DisProt	disordered proteins	933	156
Xenbase* (Fortriede et al. 2020)	Xenopus (frog)	731	0

Table 3. Genome coverage of PAN-GO annotations. Percentage of protein-coding genes with at least one PAN-GO reviewed annotation, for different taxonomic groups

Taxonomic group	Number of individually reviewed, annotated genomes	Gene coverage of annotations
vertebrates	19	66% - 83%
invertebrates	15	40% - 68%
fungi	14	32% - 76%
plants	40	28% - 51%
protists, alveolates, amoebae	11	19% - 46%
archaea	8	23% - 34%
bacteria	35	20% - 57%

Table 4. Number of PAN-GO annotations for selected genomes.

Genome	Total IBA annotations	MF annotations	BP annotations	CC annotations
Danio rerio	82855	23049	33888	25918
Mus musculus	72554	19832	29947	22775
Rattus norvegicus	71276	19792	28738	22746
Homo sapiens	68695	18537	28177	21981
Gallus gallus	59293	15847	24010	19436
Xenopus tropicalis	43232	12687	16735	13810
Arabidopsis thaliana	37509	12106	12922	12481
Caenorhabditis elegans	31093	9021	11728	10344
Drosophila melanogaster	30331	8628	11202	10501
Dictyostelium discoideum	18966	5556	6815	6595
Saccharomyces cerevisiae	15675	4286	5763	5626

Schizosaccharomyces pombe	13723	3719	4953	5051
Escherichia coli	5681	2171	1916	1594

Table 5. GO evidence codes+reference combinations. Users should consider both the type of evidence and the review level. GO internal references (starting with GO_REF:) describe specific annotation methods, and are available at https://github.com/geneontology/go-site/tree/master/metadata/gorefs/README.md.

Evidence code	Reference	Evidence type description	Review level description
IDA Inferred from direct assay	Scientific publication providing experimental data	Experimental, most direct evidence for function	Individually expert-reviewed
IMP Inferred from mutant phenotype	Scientific publication providing experimental data	Experimental, from a perturbation in the normal function	Individually expert-reviewed
IGI Inferred from genetic interaction	Scientific publication providing experimental data	Experimental, from perturbations in normal functions of more than one gene	Individually expert-reviewed
IEP Inferred from expression pattern	Scientific publication providing experimental data	Experimental, used only for biological process annotations, from comparison with genes of known function	Individually expert-reviewed
IPI Inferred from protein interaction	Scientific publication providing experimental data	Experimental, used only for annotations to protein binding terms	Individually expert-reviewed
HDA Inferred from high throughput direct assay	Scientific publication providing experimental data	Experimental (high throughput), direct	Not individually reviewed; expert review methodology to exclude high false positive rate observations
HMP Inferred from high throughput mutant phenotype	Scientific publication providing experimental data	Experimental (high throughput), mutant phenotype	Not individually reviewed; expert review to exclude high false positive rate observations
HGI Inferred from high throughput genetic interaction	Scientific publication providing experimental data	Experimental (high throughput), genetic interaction	Not individually reviewed; expert review to exclude high false positive rate observations

HEP Inferred from high throughput expression pattern	Scientific publication providing experimental data	Experimental (high throughput), expression pattern	Not individually reviewed; expert review to exclude high false positive rate observations
IBA Inferred from biological ancestor	(Gaudet et al. 2011)	Homology, from experimental evidence propagated through a phylogenetic tree and/or from direct experimental evidence	Individually expert-reviewed in the context of all experimental annotations for related genes
ISS Inferred from sequence similarity	Scientific publication providing sequence similarity evidence	Homology, from experimental evidence propagated from one gene to one related gene, asserted in the publication	Individually expert-reviewed in the context of all experimental annotations for related genes
ISS Inferred from sequence similarity	GO_REF:0000024	Homology, from experimental evidence propagated from one gene to one related gene, asserted by a biocurator	Individually expert-reviewed
ISO Inferred from sequence orthology	Scientific publication providing orthology evidence	Homology, from experimental evidence propagated from one gene to one orthologous gene	Individually expert-reviewed
ISO Inferred from sequence orthology	GO_REF:0000008	Homology, from experimental evidence propagated from one mammalian gene to one orthologous mouse gene	Individually expert-reviewed
ISO Inferred from sequence orthology	GO_REF:0000024	Homology, from experimental evidence propagated from one gene to one orthologous gene	Individually expert-reviewed
ISO Inferred from sequence orthology	GO_REF:0000096	Homology, from experimental evidence propagated from one	Not individually reviewed; orthology manually reviewed

		gene to one orthologous gene among human, mouse, rat orthologs	
ISO Inferred from sequence orthology	GO_REF:0000101	Homology, from experimental evidence propagated from one gene to one orthologous gene	Not individually reviewed; orthology computed using OrthoMCL
IEA Inferred from electronic annotation	GO_REF:0000107	Homology, from experimental evidence propagated from one gene to one orthologous gene	Not individually reviewed; 1-to-1 orthology computed using Ensembl Compara phylogenetic trees
IEA Inferred from electronic annotation	GO_REF:0000002	Homology, from a hit to an InterPro signature	Not individually reviewed; expert review of annotations of signatures to ensure low or no false positives
IEA Inferred from electronic annotation	GO_REF:0000003	Imported from another resource, from mapping an EC number assigned in UniProt	Not individually reviewed; expert review of mappings; EC assignments are manually reviewed for Swiss-Prot and computationally inferred for TrEMBL
IEA Inferred from electronic annotation	GO_REF:0000004	Imported from another resource, from mapping a manually assigned Swiss-Prot keyword	Not individually reviewed; expert review of keywords and mappings
IEA Inferred from electronic annotation	GO_REF:0000104	Homology, from manually curated UniRule	Not individually reviewed; expert curation of UniRules to ensure low or no false positives
IEA Inferred from electronic annotation	GO_REF:0000108	Logical assertion using the ontology, from asserted relation between different aspects of GO	Not individually reviewed; expert curation of ontology links
IEA Inferred from electronic annotation	GO_REF:0000117	Computational, from machine learning	Not individually reviewed; assigned by machine learning from curated training sets

TAS Traceable author statement	Scientific publication citing original data	From a published statement referencing experimental evidence in a different paper	Individually reviewed
NAS Nontraceable author statement	Scientific publication with general biological knowledge statement	From an unreferenced published statement	Individually reviewed

Table 6. GO ontology editions. Editions are distinguished by the relations and metadata they include. All editions are updated at each GO release. External ontologies used in GO include: ChEBI, Uberon (Haendel *et al.* 2014), Relation Ontology (Smith *et al.* 2005), Cell Ontology (Diehl *et al.* 2016), Sequence Ontology (Mungall, Batchelor and Eilbeck 2011), Dicty Anatomy, CARO (Haendel *et al.* 2008), Fungal Anatomy Ontology (*Fungal-Anatomy-Ontology: A Structured Controlled Vocabulary for the Anatomy of Fungi*), Plant Ontology (Walls *et al.* 2019), PATO (Gkoutos, Schofield and Hoehndorf 2018), Protein Ontology (Natale *et al.* 2017).

GO edition	Format(s)	Relations included	Links to other ontologies
go-basic	OBO	is a, part of, regulates, negatively regulates and positively regulates	Not available
go	OBO and OWL-RDF/XML	Same as go-basic, plus has part and occurs in	Not available
go-plus	OWL-RDF/XML		ChEBI, Uberon, Cell Ontology, Sequence Ontology, Dicty Anatomy, CARO, Fungal Anatomy Ontology, Plant Ontology, PATO, Protein Ontology

Figure legends

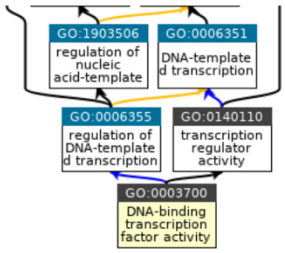
Figure 1. Examples of the three components of the GO knowledgebase. A) The GO ontology consists of terms, e.g. *DNA binding transcription factor activity*, and relationships between the terms (arrows; black=*is a*, blue=*part of*, orange=*regulates*). B) GO annotations associate a specific gene product (here, human ZNF410) with GO terms asserting its functional aspects ("GO Class" column, e.g. sequence-specific double stranded DNA binding) and the evidence for each assertion with its traceable source ("Evidence" and "Reference" columns). C) The GO-CAM model combines individual GO annotations into a model, in this case a very

simple model describing how human ZNF410 acts as a transcription factor to positively regulate (denoted by the green arrow) transcription of the *CHD4* gene, which in turn acts as a co-repressor to repress (denoted by red dashed lines) transcription of fetal hemoglobin genes (*HBG1* and *HBG2*) in erythroid lineage cells. In this view, each box in the GO-CAM is labeled with the gene product and species abbreviation for simplicity.

Figure 2. GO-CAM model of the SARS-CoV2 - host interactions as displayed using the GO-CAM Pathway Widget (code available at https://github.com/geneontology/wc-gocam-viz) on the Alliance of Genome Resources gene pages (https://www.alliancegenome.org/gene/HGNC:20144#pathways). The model includes proteins from both humans (Hsap) and the SARS-CoV-2 virus (Scov2). A simplified representation of the causal model is shown on the main figure, which is simplified by labeling with the gene and organism. The model includes many additional details, which are displayed as “cards”; the information for MAVS activity (inset) which normally acts as a signaling adaptor located in the mitochondrial membrane. MAVS activity is suppressed directly by the SARS-CoV-2 M protein, and indirectly by other SARS-CoV-2 proteins. Each of the “E” symbols on the right-hand side can be clicked to see the evidence for each assertion in the model.

Figure 3. Alliance ribbon view for the yeast RPB7 gene. High-level GO categories annotated are shown in blue squares (https://www.alliancegenome.org/gene/SGD:S000002812).

A



B

Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence	Evidence with	PANTHER family	Type	Isoform	Reference	Date
ZNF410	Zinc finger protein 410		RNA polymerase II cis-regulatory region sequence-specific DNA binding	has input UniProtKB:Q14839 occurs in erythroid lineage cell	UniProt	Homo sapiens	IMP		zinc finger protein pthr46179	protein		PMID:33301730	20210629
ZNF410	Zinc finger protein 410		sequence-specific double-stranded DNA binding		ARUK-UCL	Homo sapiens	IDA		zinc finger protein pthr46179	protein		PMID:28473536	20200608
ZNF410	Zinc finger protein 410		sequence-specific double-stranded DNA binding	has input UniProtKB:Q14839 occurs in erythroid lineage cell	UniProt	Homo sapiens	IMP		zinc finger protein pthr46179	protein		PMID:33301730	20210629

C

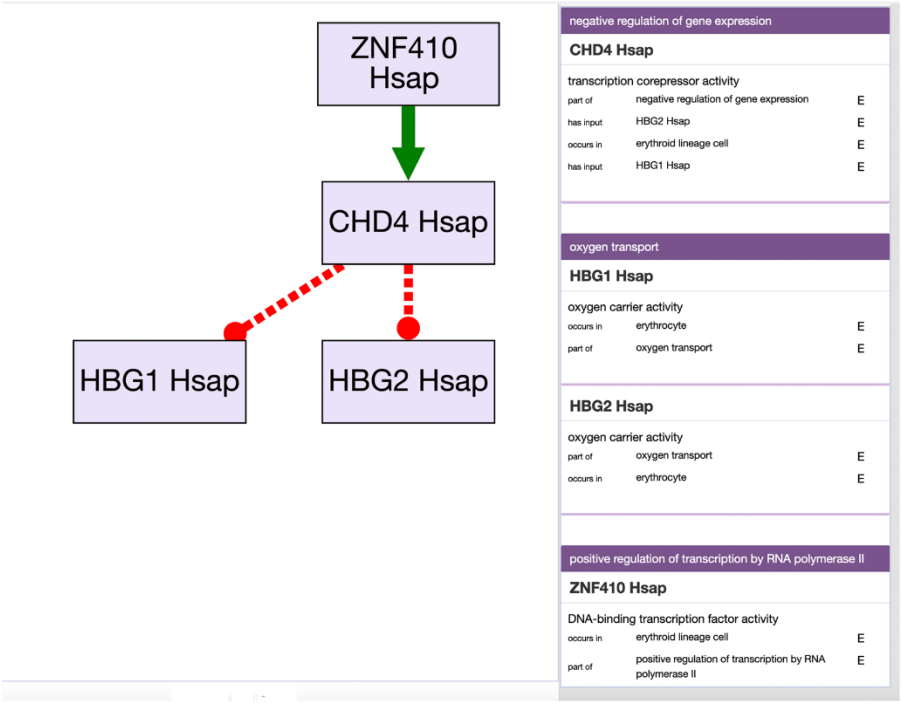


Figure 1
152x188 mm (x DPI)

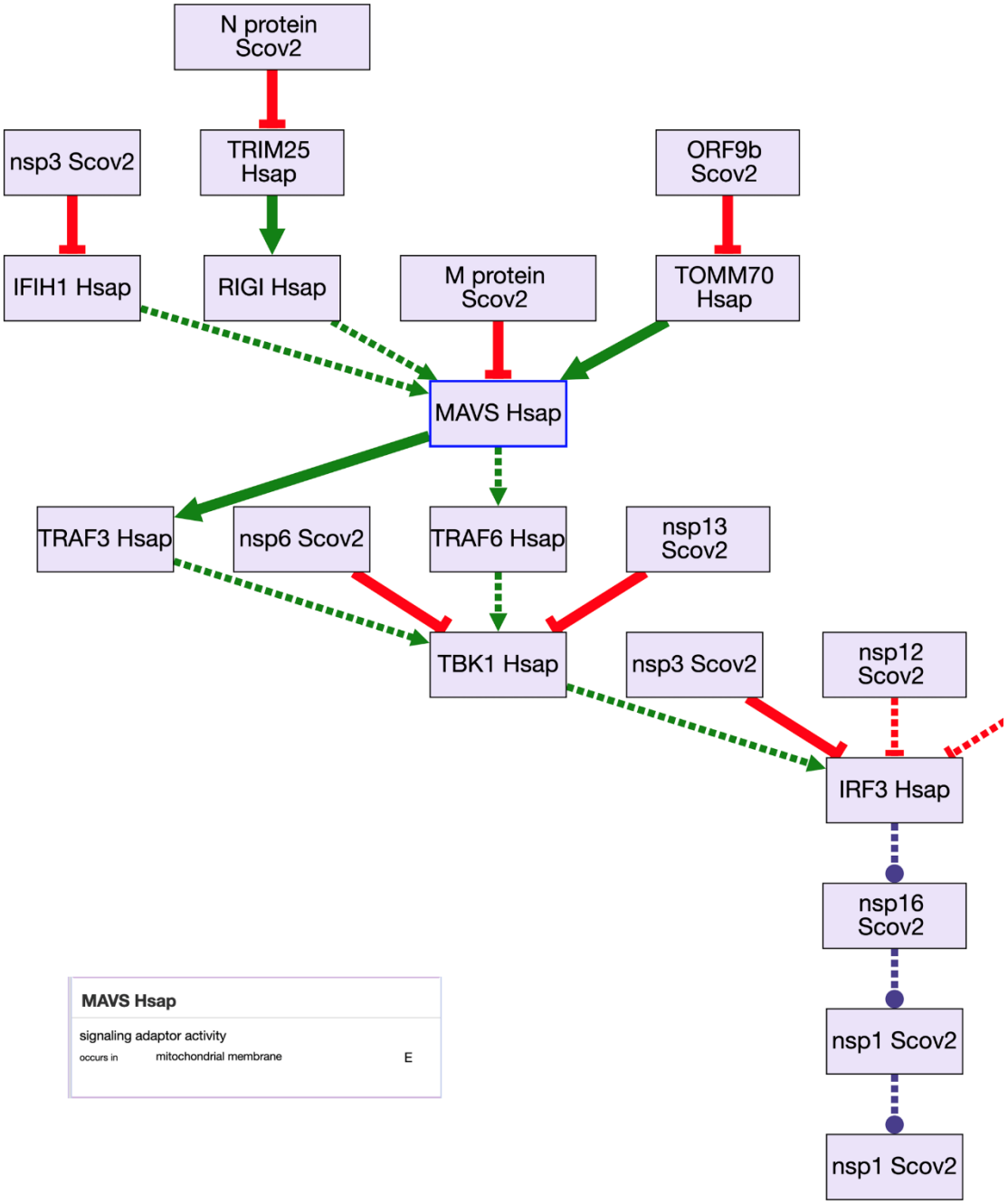


Figure 2
152x178 mm (x DPI)

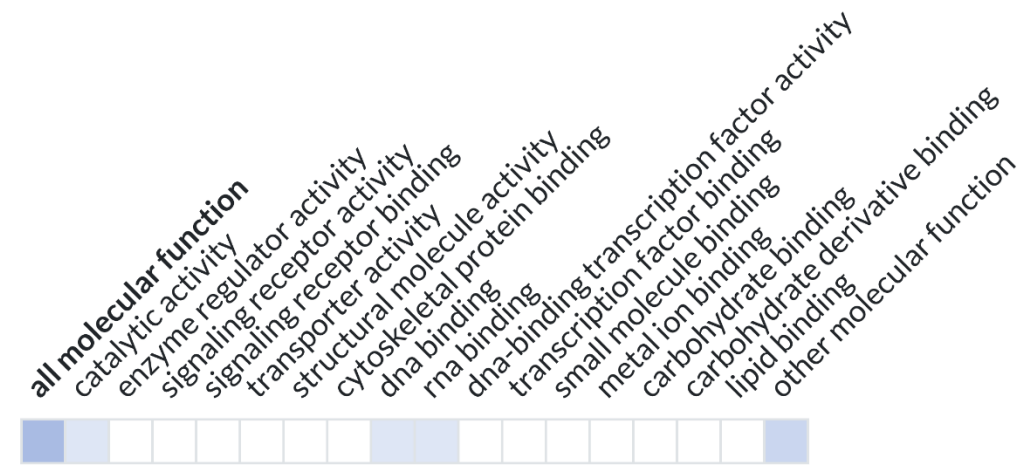


Figure 3
138x67 mm (x DPI)