Chapter 4

Best Practices in Manual Annotation with the Gene Ontology

Sylvain Poux and Pascale Gaudet

Abstract

The Gene Ontology (GO) is a framework designed to represent biological knowledge about gene products' biological roles and the cellular location in which they act. Biocuration is a complex process: the body of scientific literature is large and selection of appropriate GO terms can be challenging. Both these issues are compounded by the fact that our understanding of biology is still incomplete; hence it is important to appreciate that GO is inherently an evolving model. In this chapter, we describe how biocurators create GO annotations from experimental findings from research articles. We describe the current best practices for high-quality literature curation and how GO curators succeed in modeling biology using a relatively simple framework. We also highlight a number of difficulties when translating experimental assays into GO annotations.

Key words Gene ontology, Expert curation, Biocuration, Protein annotation

1 Background

Biological databases have become an integral part of the tools researchers use on a daily basis for their work. GO is a controlled vocabulary for the description of biological function, and is used to annotate genes in a large number of genome and protein databases. Its computable structure makes it one of the most widely used resources. Manual annotation with GO involves biocurators, who are trained to reading, extracting, and translating experimental findings from publications into GO terms. Since both the scientific literature and the GO are complex, novice biocurators can make errors or misinterpretations when doing annotation. Here, we present guidelines and recommendations for best practices in manual annotation, to help curators avoid the most common pitfalls. These recommendations should be useful not only to biocurators, but also to users of the GO, since the understanding of the curation process should help understand the meaning of the annotations.

Christophe Dessimoz and Nives Škunca (eds.), *The Gene Ontology Handbook*, Methods in Molecular Biology, vol. 1446, DOI 10.1007/978-1-4939-3743-1_4, © The Author(s) 2017

1.1 Knowledge Inference: General Principles

Our understanding of the world is built by observation and experimentation. The overall process of the scientific method involves making hypotheses, deriving predictions from them, and then carrying out experiments to test the validity of these predictions. The results of the experiments are then used to *infer* whether the prediction was true or not [1]. Hypotheses are tested, validated, or rejected, and the combination of all the experiments contributes to uncovering the mechanism underlying the process being studied (Fig. 1).

Examples of experiments include testing an enzymatic activity in vitro using purified reagents, measuring the expression level of a protein upon a given stimulus, or observing the phenotypes of an organism in which a gene has been deleted by molecular genetics techniques. Different inferences can be made from the same experimental setup depending on the hypothesis being tested. Thus, the conclusions that can be derived from individual experiments may vary, depending on a number of factors: they depend on the current state of knowledge, on how well controlled the experiment is, on the experimental conditions, etc. It also happens that the conclusions from a low-resolution experiment are partially or completely refuted when better techniques become available. These factors are inherent to empirical studies and must be taken into account to ensure correct interpretation of experimental results.

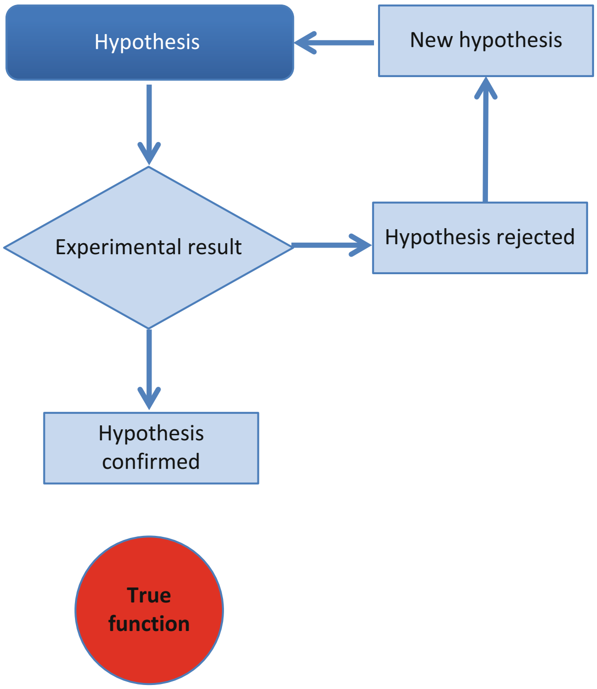

Fig. 1 How the scientific method is used to test and validate hypotheses

1.2 Knowledge Representation Using Ontologies

GO is a framework to describe the roles of gene products across all living organisms [2] (*see* also Chap. 2, [3]). The ontology is divided into three branches, or aspects: Molecular Function (MF) that captures the biochemical or molecular activity of the gene product; Biological Process (BP), corresponding to the wider biological module in which the gene product's MF acts; and Cellular Component (CC), which is the specific cellular localization in which the gene product is active.

The association of a GO term and a gene product is not explicitly defined, but implicitly means that the gene product *has* an activity or a molecular role (MF term), *directly participates* in a process (BP), and the function takes place *in a specific cellular localization* (CC) [2]. Therefore, transient localizations such as endoplasmic reticulum and Golgi apparatus for secreted proteins are not in the scope of GO. Biological process is the most challenging aspect of the GO to capture, in part because it models two categories of processes: *subtypes*: "mitotic DNA replication" (GO:1902969) is a particular type of "nuclear DNA replication" (GO:0033260), and *sub-processes*: mitotic DNA replication is a step of the "cell cycle" (GO:0000278). These two classification axes are distinguished by "is a" and "part of" relations with their parents, respectively. Gene products can be annotated using as many GO terms as necessary to completely describe its function, and the GO terms can be at varying levels in the hierarchy, depending on the evidence available. If a gene product is annotated to any particular term, then the annotations also hold for all the is-a and part-of parent terms. Annotations to more granular terms carry more information; however the annotation cannot be any deeper than what is supported by the evidence.

The complexity of biology is reflected in the GO: with 40,000 different terms [4], learning to use the GO can be compared to learning a new language. As when learning a language, there are terms that are closely related to those we are familiar with, and others that have subtle but important differences in meaning. The GO defines each term in two complementary ways: first by a textual definition intended to be human readable. Secondly, the structure of the ontology as determined by relationships of terms between each other is also a way by which terms are defined these can be utilized for computational reasoning.

1.3 Methods for Assigning GO Annotations

There are two general methods for assigning GO terms to gene products. The first is based on *experimental evidence*, and involves detailed reading of scientific publications to capture knowledge about gene products. Biocurators browse the GO ontologies to associate appropriate GO term(s) whose definition is consistent with the data published for the gene product. *See* Chaps. 3 [5] and 17 [6], for a description of the elements of an annotation. Expert curation based on experiments is considered the gold standard of functional annotation. It is the most reliable and provides strong support for the association of a GO term with a gene product.

The second method involves making *predictions* on the protein's function and subcellular localization, most often with methods relying on sequence similarity. Although not detailed in this chapter, prediction methods are highly dependent on annotations based on experiments. Indeed, all methods to assign annotations based on sequence similarity are more or less directly derived from knowledge that has been acquired experimentally; that is, at least one related protein must have been tested and shown to have a given function for that information to be propagated to other proteins. Hence, the accurate assignment of GO classes to gene products based on experimental results is crucial, since many further annotations depend on their accuracy.

2 Best Practices for High-Quality Manual Curation

2.1 GO Inference Process

Similar to the process by which experimental results get translated into a model of the biological phenomenon being investigated, biocurators take the conclusions from the investigation and convert it into the GO framework. Thus, the same assay may lead to different interpretations depending on the question being tested.

As shown in Table 1, an assay must be interpreted in the wider context of the known roles of the protein, and how directly the assay assesses the protein's role in the process under investigation. Here, several experiments are described in which the readout is DNA fragmentation upon apoptotic stimulation, but that lead to different annotations. DFFB (UniProtKB O76075) is annotated to "apoptotic DNA fragmentation" (GO:0006309) because the protein is also known to be a nuclease. CYCS (UniProtKB P99999) is annotated to caspase activation ("activation of cysteine-type endopeptidase activity involved in apoptotic process" (GO:0006919)) because a direct role has been shown using an in vitro assay. However CYCS is not annotated to "apoptotic DNA fragmentation" (GO:0006309) despite the observation that removing it from cells prevents DNA fragmentation, since the activity of CYCS occurs before DNA fragmentation. Any step that takes place afterwards will inevitably fail to happen, but this does not imply participation in this downstream sequence of molecular events. Finally, the FOXL2 (UniProtKB P58012) transcription factor has a positive effect on the occurrence of apoptosis, by an unknown mechanism, so it is annotated to "positive regulation of apoptotic process" (GO:0043065). This is where the curator's knowledge is critical and provides most added value over, e.g., machine learning and text mining

Table 1
GO inference process, from the hypothesis in the paper to the assay and result, and to the inference of a GO function or role

Protein	Known roles	Hypothesis	Assay → Result	Conclusion → GO	Reference
DDFB (O76075)	DNase	The nuclease activity of DDFB is required for nuclear DNA fragmentation during apoptosis	Apoptotic DNA fragmentation →Increased in the presence of DDFB	DDFB *mediates* nuclear DNA fragmentation during apoptosis →Apoptotic DNA fragmentation (GO:0006309)	[7]
CYCS (P99999)	Cytochrome C; electron transport	CYCS triggers the activation of caspase-3	Apoptotic DNA fragmentation →Decreased upon immunodepletion of CYCS 7 Purified CYCS →Stimulates the auto-proteolytic activity of caspase-3	CYCS *directly activates* caspase-3 →Activation of cysteine-type endopeptidase activity involved in apoptotic process (GO:0006919)	[8]
FOXL2 (P58012)	Transcription factor	Mutations in FOXL2 are known to cause premature ovarian failure, which may be due to increased apoptosis	Apoptotic DNA fragmentation →Increased in the presence of FOXL2	FOXL2 increases the rate of apoptosis →Positive regulation of apoptotic process (GO:0043065)	[9]

2.2 Needles and Haystacks

With more than 500,000 records indexed yearly in PubMed, it is not possible for the GO to comprehensively represent all the available data on every protein. To address this, a careful prioritization of both articles and proteins to annotate is done. The publications from which information is drawn are selected to accurately represent the current state of knowledge. Accessory findings and non-replicated data are not systematically annotated; confirmation or at least consistency with findings from several publications is invaluable to accurately describe the function of a gene product.

Focusing on a topic allows the curator to construct a clear picture of the protein's role and makes it easier to make the best decisions when capturing biological knowledge as annotations. Reading different publications in the field helps to resolve issues and select terms with more confidence. Existing GO annotation in proteins that participate in the same biological process is also helpful to decide on how best to represent the experimental data with the GO. On the other hand, without the broader context of the research domain, some papers may be misleading: first, as more data accumulate, a growing number of contradictory or even incorrect results are found in the scientific literature. Second, the way knowledge evolves occasionally obsoletes previous findings. Curators use their expertise to assess the scientific content of articles and avoid these pitfalls [10].

2.3 How Low Can You Go: Deciding on the Level of Granularity of an Annotation

The level of granularity of an annotation is dictated by the evidence supporting it. A good illustration is provided by ADCK3 protein in human (UniProtKB Q8NI60), an atypical kinase containing a protein kinase domain involved in the biosynthesis of ubiquinone, and an essential lipid-soluble electron transporter. Although it contains a protein kinase domain, it is unclear whether it acts as a protein kinase that phosphorylates other proteins in the CoQ complex or acts as a lipid kinase that phosphorylates a prenyl lipid in the ubiquinone biosynthesis pathway [11]. While it would be tempting to conclude that the protein has "protein kinase activity" (GO:0004672) from the presence of the protein kinase domain, the more general term "kinase activity" (GO:0016301) with no specification of the potential substrate class (lipid or protein) is more appropriate.

2.4 Less Is More: Avoiding Over-Interpretation

2.4.1 Biological Relevance of Experiments

Annotations focus on capturing experiments that are biologically relevant. Thus, substrates, tissue, or cell-type specificity are annotated only when the data indicates the physiological importance of these parameters. One difficulty is that it is not always possible to distinguish between *experimental* context and *biological* context, which can potentially result in GO terms being assigned as if they represented a specific role or under specific conditions, while in fact this only reflects the experimental setup and does not have real biological significance. For example, the activity of E3 ubiquitin protein ligases is commonly tested by an in vitro autoubiquitination assay. While convenient, the assay is not conclusive with respect to the "protein autoubiquitination" (GO:0051865) in vivo. In the absence of additional data, only the term "ubiquitin protein ligase activity" (GO:0061630) should be used. Similarly, the cell type in which a function was tested does not imply that the cell type is relevant for the function; any hint that the protein is studied outside its normal physiological context (such as overexpression) should be carefully taken into consideration.

2.4.2 Downstream Effects

Downstream effects, as well as readouts (discussed above in Subheading 2.1), can lead to incorrect annotations if they are directly assigned to a gene product playing a role many steps further. Here we use downstream as "occurring after," with no implication on the *direct* sequentiality of the events.

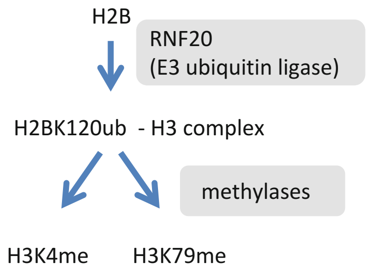

Fig. 2 Monoubiquitination of histone H2B (H2BK120ub) promotes methylation of histone H3 (H3K4me and H3K79me)

Gene products that play housekeeping functions or function upstream of important signaling pathways have many indirect effects and pose a challenge for annotation. This can be illustrated by proteins that mediate chromatin modification. Histone tails are posttranslationally modified by a complex set of interdependent modifications. For instance, histone H2B monoubiquitination at Lys-120 (H2BK120ub) is a prerequisite for the methylation of histone H3 at Lys-4 and Lys-79 (H3K4me and H3K79me, respectively) (Fig. 2). RNF20 (UniProtKB Q5VTR2), an E3 ubiquitin ligase that mediates H2BK120ub, therefore indirectly promotes H3K4me and H3K79me methylation [12]. Thus, the annotation of enzymes that modify histone tails is limited to the primary function of the enzyme ("ubiquitin-protein ligase activity" (GO:0004842) and "histone H2B ubiquitination" (GO:0033523), in this case), while the further histone modifications are only annotated to the proteins mediating these modifications.

A similar approach is taken for cases where the experimental readout is also a GO term. Examples of this include DNA fragmentation assays to measure apoptosis, and MAPK cascade to measure the activation of an upstream pathway. Proteins that are involved in signaling leading to apoptosis do not mediate or *participate* in DNA fragmentation, but their addition or removal causes changes in the amount of DNA fragmentation upon apoptotic stimulation. In other words, the effect of a protein on a specific readout can be very indirect. Whenever possible, annotation of these very specific terms ("apoptotic DNA fragmentation" (GO:0006309), "MAPK cascade" (GO:0000165)) is limited to cases where there is evidence of a molecular function supporting a direct implication in the process. If that information is not available, the annotation is made to a more general term, such as "apoptotic process" (GO:0006915) or "intracellular signal transduction" (GO:0035556), for instance.

2.4.3 Phenotypes

One common method to determine the function or process of a gene is mutagenesis. However, interpreting the results from mutant phenotypes is very difficult, as the effects caused by the absence or disruption of a gene can be very indirect. Any kind of knockout/

knockdown or "add back" experiments (in which proteins are either overexpressed or added to a cellular extract) cannot demonstrate the *participation* of a protein in a process, only its requirement for the process to occur. Inferring a participatory role would be an over-interpretation of the results. A striking illustration of this can be made with housekeeping genes, such as those involved in transcription and translation: knockouts in these proteins (when not lethal) can be pleiotropic and affect essentially all cellular processes. It would be both inaccurate and overwhelming for curators to annotate these gene products to every cellular process impacted. The more prior knowledge we have about a protein's function, in particular its biochemical activity, the more accurate we can be when interpreting a phenotype.

Phenotypes caused by gene mutations are of great interest, not only to try to understand the function of proteins, but also to provide insights into mechanisms leading to disease. The scope of the GO, though, is to capture the *normal* function of proteins. There are phenotype ontologies for human—HPO [13], mouse—MP [14] and other species that allow capturing phenotype in a structure that is more relevant to this type of data.

2.5 Main Functions and Secondary Roles

One limitation of the GO is that main functions and secondary roles are not explicitly encoded, so that this information is difficult to find. For example, enzymes may have different substrates: in some cases, the substrate specificity is driven by the biological context, but in other cases by the experimental conditions. While some activities represent the main function of the enzyme, others are secondary or can be limited to very specific conditions.

A good example is provided by the CYP4F2 enzyme (UniProtKB Q9UIU8), a member of the cytochrome P450 family that oxidizes a variety of structurally unrelated compounds, including steroids, fatty acids, and xenobiotics. In vivo, the enzyme plays a key role in vitamin K catabolism by mediating omega-hydroxylation of vitamin K1 (phylloquinone), and menaquinone-4 (MK-4), a form of vitamin K2 [15, 16]. While hydroxylation of phylloquinone and MK-4 probably constitutes the main activity of this enzyme since this activity has been confirmed by several in vivo assays, CYP4F2 also shows activity towards other related substrates, such as arachidonic acid omega and leukotriene-B [10] omega [17–21]. Clearly vitamin K1 and MK-4 are the main physiological substrates of CYP4F2, but since it is plausible that the enzyme also acts on other molecules, these different activities are also annotated. In the absence of additional evidence, it is currently impossible to highlight which GO term describes the in vivo function of the enzyme. For the reactions known to be implicated in vitamin K catabolism, adding this information as an annotation extension helps clarify the main role of that specific reaction (*see* Chap. 17, [6]).

2.6 Hindsight Is 20/20: Dealing with Evolving Knowledge

Our understanding of biology is dynamic, and evolves as new experiments confirm or contradict previous results. It is therefore essential to read several, preferably recent publications on a subject to make sure that prior working hypotheses, that have subsequently been invalidated, are not annotated. That is, sometimes it is necessary to remove annotations in order to limit the number of false positives. A number of mechanisms exist in GO to capture evolution of knowledge. New GO terms are added to the ontology when knowledge is not covered by existing GO terms. Curators work in collaboration with the GO editors, defining new terms or correcting the definitions of existing terms when required. Conflicting results can be dealt by using the "NOT" qualifier, which states that a gene product is not associated with a GO term. This qualifier is used when a positive association to this term could otherwise be expected from previous literature or automated methods (for more information read www.geneontology.org/GO.annotation.conventions.shtml#not).

A good example of how GO deals with evolving knowledge as new papers are published on a protein is provided by the recent characterization of the NOTUM protein in human and *Drosophila melanogaster*. Notum was first characterized in *D. melanogaster* (UniProtKB Q9VUX3) as an inhibitor of Wnt signaling [22, 23]. Based on its sequence similarity with pectin acetylesterase family members, it was initially thought to hydrolyze glycosaminoglycan (GAG) chains of glypicans by mediating cleavage of their GPI anchor in vitro [24]. Two different articles published recently contradict these previous results, showing that the substrate of human NOTUM (UniProtKB Q6P988) and *D. melanogaster* Notum is not glypicans, and that human NOTUM specifically mediates a palmitoleic acid modification on WNT proteins [25, 26]. This new data confirms the role of NOTUM as an inhibitor of Wnt signaling, but with a mechanism completely different from what the initial studies had suggested. To correctly capture these findings in GO, new terms describing protein depalmitoleylation were added in GO: "palmitoleyl hydrolase activity" (GO:1990699) and "protein depalmitoleylation" (GO:1990697). In addition, NOTUM proteins received negative annotations for "GPI anchor release" (GO:0006507) and "phospholipase activity" (GO:0004620) to indicate that these findings had been disproven.

Although relatively infrequent, this type of situation is critical because it may affect the accuracy of the GO. Ideally, when new findings invalidate previous ones, old annotations are revisited in the light of new knowledge and annotation from previous papers reevaluated to ensure that annotation was not the result of overinterpretation of data.

The most widely used manual protein annotation editor for GO, Protein2GO, has a mechanism to dispute questionable or outdated annotations that sends a request for reevaluation of annotations [27]. Users who notice incorrect or missing annotations are strongly encouraged to notify the GO helpdesk (http://geneontology.org/form/contact-go) so that corrections can be made.

3 Importance of Annotation Consistency: Toward a Quality Control Approach

The goal of the GO project is to provide a uniform schema to describe biological processes mediated by gene products in all cellular organisms [2]. Annotation involves translating conclusions from biological experiments into this schema, such that we are making inferences of inferences. To avoid deriving too much from the biologically relevant conclusions of experiments, consistent annotation within the GO framework is essential.

The GO curators make every effort to ensure that annotations reflect the current state of knowledge. As new findings are made that invalidate or refine existing models there is a need for course correction; otherwise both the ontology and the annotations may drift.

Over 20 groups contribute to manual annotations to the GO project (http://geneontology.org/page/download-annotations). The number of annotations by species, broken down into experimental versus non-experimental, is shown in Fig. 3. Since manual annotations are so critical to the overall quality of the entire corpus of GO data, it is important that each biocurator from every contributing group interprets experiments consistently.

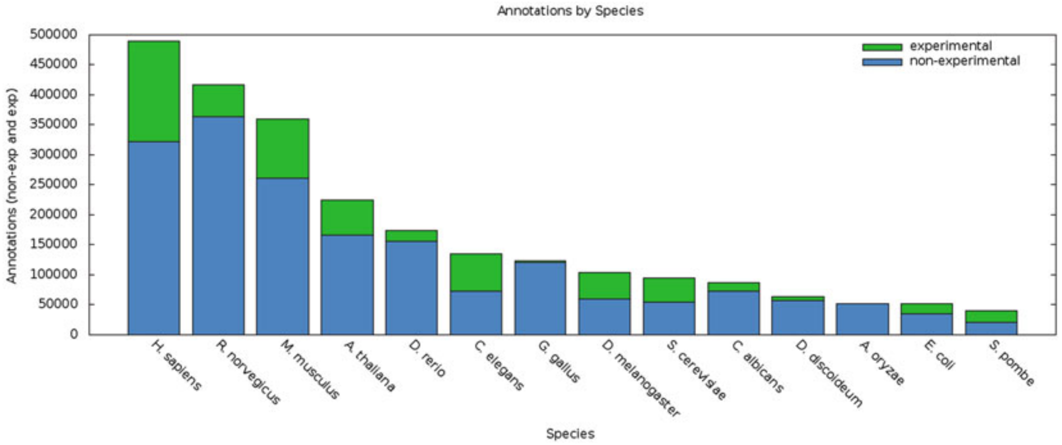

Fig. 3 Number of annotations in 12 species annotated by the GO consortium. Source: http://geneontology.org/page/current-go-statistics

While the GO Consortium does not possess sufficient resources to review all annotations individually on an ongoing basis, several approaches are in place to ensure consistency:

- GO uses automated procedures for validating GO annotations. An automated checker runs through the GO annotation rulebase (http://geneontology.org/page/annotation-quality-control-checks), which validates the syntactic and biological content of the annotation database, and verifies that correct procedures are followed. Examples include taxon checks [28] and checks to ensure that the correct object type is used with different types of evidence.
- The annotation team of the GO consortium also has regular annotation consistency exercises, where participating annotators independently annotate the same paper to ensure that guidelines are applied in a uniform manner, discuss any discrepancy, and update guidelines when these are lacking or need clarification.
- Finally, the Reference Genome Project [29] has proven to be a very useful resource to improve annotation coherence across the GO (Feuermann et al., *in preparation*). The project uses PAINT, a Phylogenetic Annotation and INference Tool, to annotate protein families from the PantherDB resource [30]. PAINT integrates phylogenetic trees, multiple sequence alignments, experimental GO annotations, as well as references pointing to the original data. PAINT curators select the high-confidence data that can be propagated across either the entire tree or specific clades. By displaying different GO annotations for all members of a family, PAINT makes it easy to detect inconsistencies, thus improving the overall quality of the set of GO annotations. It also gives a mean of identifying consistent biases that usually indicate a problem in the ontology or in the annotation guidelines.

4 Summary

Expert curation of GO terms based on experimental data is a complex process that requires a number of skills from biocurators. In this chapter, we describe a number of guidelines to warn curators on common annotation mistakes and provide clues on how to avoid them. These simple rules, summarized in Table 2, can be used as a checklist to ensure that GO annotations are in line with GO consortium guidelines.

Table 2
Summary of annotation guidelines

Carefully select publications. Only annotate papers that provide the most added value.
Read recent publications. Research is not a straightforward process and reading recent publications helps resolving conflicts and detecting experimental discrepancies.
Check annotation consistency. Review the existing annotations for related proteins to see whether the annotations you are adding are consistent.
Look for confirmation for unusual findings with multiple papers, if possible. Avoid entering annotations based on experiments that do not directly implicate the protein with the GO term you annotate.
Annotate the conclusion of the experiment. Keep in mind that this may be different from the results presented. Be especially careful of interpreting the function of proteins based on mutant phenotypes.
Remove obsolete annotations. If you encounter an annotation that is based on an interpretation of an experiment that is no longer valid, use the Challenge mechanism or GO helpdesk to ask to have the annotation removed.

5 Perspective

The guidelines presented here are easy to follow and reinforce curation quality without reducing curation efficiency, which is a serious and valid challenge in the era of big data. In view of the amount of data to be dealt with, it has often been argued that manual curation "just doesn't scale," and an ongoing search for alternative methods is under way in the world of biocuration and bioinformatics. However, examples described in this chapter show that most publications describe complex knowledge that cannot be captured by machine learning or text mining technologies. To continue having an acceptable throughput, manual curation should be able to cope with the increasing corpus of scientific data. From this perspective, PAINT constitutes an excellent example of a propagation tool based on experimental GO annotations, which ensures maximum consistency and efficiency without compromising the quality of the annotations produced. Such system provides one possible answer to the concerns addressed on scalability of expert curation.

Funding Open Access charges were funded by the University College London Library, the Swiss Institute of Bioinformatics, the Agassiz Foundation, and the Foundation for the University of Lausanne.

References

1. Popper KR (2002) Conjectures and refutations: the growth of scientific knowledge. Routledge, New York
2. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29
3. Thomas PD (2016) The gene ontology and the meaning of biological function. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 2
4. Gene Consortium (2015) Gene Ontology Consortium: going forward. Nucleic Acids Res 43:D1049–D1056
5. Gaudet P, Škunca N, Hu JC, Dessimoz C (2016) Primer on the gene ontology. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 3
6. Huntley RP, Lovering RC (2016) Annotation extensions. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 17
7. Korn C, Scholz SR, Gimadutdinow O, Lurz R, Pingoud A, Meiss G (2005) Interaction of DNA fragmentation factor (DFF) with DNA reveals an unprecedented mechanism for nuclease inhibition and suggests that DFF can be activated in a DNA-bound state. J Biol Chem 280:6005–6015
8. Liu X, Kim CN, Yang J, Jemmerson R, Wang X (1996) Induction of apoptotic program in cell-free extracts: requirement for dATP and cytochrome c. Cell 86:147–157
9. Lee K, Pisarska MD, Ko JJ, Kang Y, Yoon S, Ryou SM, Cha KY, Bae J (2005) Transcriptional factor FOXL2 interacts with DP103 and induces apoptosis. Biochem Biophys Res Commun 336:876–881
10. Poux S, Magrane M, Arighi CN, Bridge A, O'Donovan C, Laiho K (2014) Expert curation in UniProtKB: a case study on dealing with conflicting and erroneous data. Database (Oxford):bau016
11. Stefely JA, Reidenbach AG, Ulbrich A, Oruganty K, Floyd BJ, Jochem A, Saunders JM, Johnson IE, Minogue CE, Wrobel RL et al (2015) Mitochondrial ADCK3 employs an atypical protein kinase-like fold to enable coenzyme Q biosynthesis. Mol Cell 57:83–94
12. Kim J, Hake SB, Roeder RG (2005) The human homolog of yeast BRE1 functions as a transcriptional coactivator through direct activator interactions. Mol Cell 20:759–770
13. Groza T, Kohler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T, Schriml LM, Kibbe WA, Schofield PN, Beck T et al (2015) The human phenotype ontology: semantic unification of common and rare disease. Am J Hum Genet 97(1):111–124
14. Smith CL, Eppig JT (2015) Expanding the mammalian phenotype ontology to support automated exchange of high throughput mouse phenotyping data generated by large-scale mouse knockout screens. J Biomed Semantics 6:11
15. Edson KZ, Prasad B, Unadkat JD, Suhara Y, Okano T, Guengerich FP, Rettie AE (2013) Cytochrome P450-dependent catabolism of vitamin K: omega-hydroxylation catalyzed by human CYP4F2 and CYP4F11. Biochemistry 52:8276–8285

16. McDonald MG, Rieder MJ, Nakano M, Hsia CK, Rettie AE (2009) CYP4F2 is a vitamin K1 oxidase: an explanation for altered warfarin dose in carriers of the V433M variant. Mol Pharmacol 75:1337–1346

17. Fava C, Montagnana M, Almgren P, Rosberg L, Lippi G, Hedblad B, Engstrom G, Berglund G, Minuz P, Melander O (2008) The V433M variant of the CYP4F2 is associated with ischemic stroke in male Swedes beyond its effect on blood pressure. Hypertension 52:373–380

18. Jin R, Koop DR, Raucy JL, Lasker JM (1998) Role of human CYP4F2 in hepatic catabolism of the proinflammatory agent leukotriene B4. Arch Biochem Biophys 359:89–98

19. Kikuta Y, Kusunose E, Kondo T, Yamamoto S, Kinoshita H, Kusunose M (1994) Cloning and expression of a novel form of leukotriene B4 omega-hydroxylase from human liver. FEBS Lett 348:70–74

20. Lasker JM, Chen WB, Wolf I, Bloswick BP, Wilson PD, Powell PK (2000) Formation of 20-hydroxyeicosatetraenoic acid, a vasoactive and natriuretic eicosanoid, in human kidney. Role of Cyp4F2 and Cyp4A11. J Biol Chem 275:4118–4126

21. Stec DE, Roman RJ, Flasch A, Rieder MJ (2007) Functional polymorphism in human CYP4F2 decreases 20-HETE production. Physiol Genomics 30:74–81

22. Gerlitz O, Basler K (2002) Wingful, an extracellular feedback inhibitor of Wingless. Genes Dev 16:1055–1059

23. Giraldez AJ, Copley RR, Cohen SM (2002) HSPG modification by the secreted enzyme Notum shapes the Wingless morphogen gradient. Dev Cell 2:667–676

24. Kreuger J, Perez L, Giraldez AJ, Cohen SM (2004) Opposing activities of Dally-like glypican at high and low levels of Wingless morphogen activity. Dev Cell 7:503–512

25. Kakugawa S, Langton PF, Zebisch M, Howell SA, Chang TH, Liu Y, Feizi T, Bineva G, O'Reilly N, Snijders AP et al (2015) Notum deacylates Wnt proteins to suppress signalling activity. Nature 519:187–192

26. Zhang X, Cheong SM, Amado NG, Reis AH, MacDonald BT, Zebisch M, Jones EY, Abreu JG, He X (2015) Notum is required for neural and head induction via Wnt deacylation, oxidation, and inactivation. Dev Cell 32: 719–730

27. Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, O'Donovan C (2015) The GOA database: gene ontology annotation updates for 2015. Nucleic Acids Res 43:D1057–D1063

28. Deegan (née Clark) JI, Dimmer EC, Mungall CJ (2010) Formalization of taxon-based constraints to detect inconsistencies in annotation and ontology development. BMC Bioinformatics 11:530

29. Gaudet P, Livstone MS, Lewis SE, Thomas PD (2011) Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Brief Bioinform 12:449–462

30. Mi H, Muruganujan A, Thomas PD (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res 41:D377–D386