Chapter 21

The Vision and Challenges of the Gene Ontology

Suzanna E. Lewis

Abstract

The overarching goal of the Gene Ontology (GO) Consortium is to provide researchers in biology and biomedicine with all current functional information concerning genes and the cellular context under which these occur. When the GO was started in the 1990s surprisingly little attention had been given to how functional information about genes was to be uniformly captured, structured in a computable form, and made accessible to biologists. Because knowledge of gene, protein, ncRNA, and molecular complex roles is continuously accumulating and changing, the GO needed to be a dynamic resource, accurately tracking ongoing research results over time. Here I describe the progress that has been made over the years towards this goal, and the work that still remains to be done, to make of the Gene Ontology (GO) Consortium realize its goal of offering the most comprehensive and up-to-date resource for information on gene function.

Key words Gene Ontology, Gene function, Genomics, Biological modeling

1 Motivation

From their outset in the early 1990s it was obvious that biological databases demanded a methodical way of describing the function of genes. For one thing, a model system's *raison d'etre* was to gain insight into human health and, in the days before entire genomes and proteomes were available, the relevant connections to human biology were largely based on textual descriptions of biological role. In conjunction, as genomes such as yeast were being completed, new laboratory techniques were being developed for surveying the genome, such as microarray expression panels, and these data cried out for systematic description of the voluminous results. Finally, lest we forget, this period also saw the advent of the "World Wide Web." The early pioneers in biological databases were quick to take advantage of the latest technologies for data dissemination (much easier than shipping a copy of GenBank on tape or disk drive as was the norm), but exchanging data in a rational and efficient manner required concomitant syntactic and semantic

Christophe Dessimoz and Nives Škunca (eds.), *The Gene Ontology Handbook*, Methods in Molecular Biology, vol. 1446, DOI 10.1007/978-1-4939-3743-1_21, © The Author(s) 2017

agreement. Those of us building these data resources (Including Amos Bairoch, Jonathan Bard, David Botstein, Michelle Gwinn, Minoru Kanehisa, Stan Letovsky, and Monica Riley) were avidly discussing what might be done. Biologists needed a way of making some sense of the information we were so diligently collecting about genes, both to locate information and to traverse across taxa.

Specifically one slightly obsessive biologist, Michael Ashburner, wanted to classify all fly genes and have the corresponding worm, mouse, human, yeast groups use the same classification scheme (see ftp://ftp.ebi.ac.uk/pub/databases/edgp/misc/ashburner/fly_function_tree for an early example, and ftp://ftp.geneontology.org/pub/go/www/gene.ontology.discussion.shtml for the white paper as it was first publicly presented in 1998). That way, if he found a fly gene involved in a particular process, he could then ask what genes in other taxa are (thought) to be involved in the "same" process, and what insights can be gleaned from its counterpart? We needed a way to describe the attributes of gene products in a rigorous way that would enable biologists to roam the universe of genomes and biology, to explore: temporally and spatially characteristic expression patterns; the specific (often) cellular compartment localization where they acted; whether they were constitutive parts of particular cellular components and/or complexes; and their biochemical or physiological functions and activities. These are attributes of genes that are of great interest to all biologists. And in an ideal world all biological databases would agree on how such information can be made discoverable and comparable.

2 *Desiderata* (Principles) Circa 1996–1997 (Banbury & Les Treilles)

Two seminal workshops were organized in 1996 and 1997 largely devoted to discussing the need for agreement among the genomic resources on how semantic comparability should be achieved. The first of these was sponsored by the Banbury Center,[1] (organized by M. Ashburner, E. Harlow, P. Karp and J. Witkowski), and the second on building genome databases sponsored by the Fondation des Treilles[2] (organized by W.M. Gelbart, and M. Ashburner). These meetings set the stage for the Gene Ontology Consortium by defining our working definitions and essential principles.

These axiomatic working definitions, begin with "gene product": a physical object, typically associated with a gene or genes indirectly through transcription and translation (for proteins), affecting some biological process. Such things as proteins, ncRNAs, protein complexes, and so forth are all typical functional objects. These were the objects to be described. In turn the essential attributes of a gene

[1] http://www.cshl.edu/Banbury

[2] http://www.les-treilles.com

product—its function, the process(es) it participates in, and the cellular location at which these occur—were also defined: Function being a capability that a physical gene product carries as a potential, describing only what a gene product can do, without necessarily specifying where or when this usage actually occurs; Process as a transformation that has a temporal aspect to it, even if virtually instantaneous, accomplished via one or more ordered assemblies of functions; And (originally) cellular component as an anatomical structure within the cell, a location in which a function or process occurs (since expanded to include extracellular space).

Following agreement on these basic definitions came the animated discussions on the desired (and required) characteristics for actual operations.

3 Essentials for the "Ontology"

The name "Gene Ontology" was originally a jest, but the joke was on us, as it turns out GO is indeed an ontology—at least in the computational sense, with the primary operational data structure now being OWL. Every attribute mandated at the outset has proven its worth and remains at the core of the GO. Some of these essential criteria are outlined here.

3.1 Unique IDs

It was rapidly understood that unique identifiers were essential operationally. This allowed the collaborating resources to reference the ontology classes (terms) unambiguously and stably. Furthermore by using a semantically meaningless identifier, as opposed to using the label as the identifier, we were free to change the label at any time, and to display different preferred labels for different communities. At the time this was a major difference compared to other frame based systems such as "Ontolingua" or even Ontology Web Language (OWL, although OWL did not exist at the time) which used the label (name) as the identifier.

3.2 Graph Structure

It was also determined that it would be essential for the GO terms to have a graphical relationship to each other, rather than the prevalent norm in biology at the time: a flat list of keywords used for tagging. In the early, consciously simplistic, model GO began with there were only two relationship types: is_a and part_of. But it was recognized even then that more relationships would ultimately be required.

3.3 Human Readable Definitions and Labels

The decision to make numerical identifiers the stable GO "object" had implications for the human readable labels. And, in addition, rather than attempting to convey all pertinent biological information by encoding it directly into the label, human readable definitions would provide the definitive definition. Thus it is the definition, *not* the label, which defines an ontology class in GO. If a label changes it means nothing and there are no serious consequences. If a definition changes, such that the meaning of the class has changed, then this has obvious consequences for any gene product that was annotated to the original class. Thus the original class is made obsolete, with a reference to the new class as a suggestion, and the new class is given a new identifier.

Another misconception that often needs to be clarified is that GO has nothing to do with nomenclature. The confusion arises because we are using (and often have to use) exactly the same words to describe both the product and its function. For example, "alcohol dehydrogenase" can describe what you can put in an Eppendorf tube (the gene product) or it can describe the function of this protein. There is, however, a formal difference—a "gene product" has (potentially) a many-to-many relationship with a "function." That is to say there are many gene products that have the function "alcohol dehydrogenase" (and some of these may indeed be encoded by a gene with the name alcohol dehydrogenase, but many will not be). Moreover a particular gene product may have both functions "alcohol dehydrogenase" and "acetaldehyde dismutase" and possibly more. Since GO's remit is describing functions and processes, nomenclature is irrelevant to its purpose.

Finally, the labels themselves are intended to be familiar to researchers using GO. Over the years some unfortunate "standardization" efforts, have rendered terms non-user-friendly (for example what researchers call a transcription factor is "sequence-specific DNA binding RNA polymerase II transcription factor activity" in the GO). The consequence is that both annotation and searching are made more error-prone and difficult because the familiar term, that a biologist would instinctively use, cannot be quickly located. The GO Consortium continues working to rectify these labeling issues, both by an effort to use familiar labels and through the judicious use of synonyms.

3.4 Synonyms

Multiple synonyms of different flavors are essential for allowing GO to deal with: colloquialisms, community preferences, abbreviations, legacy names, the multiple ways of referring to chemical elements, capitalization, and all the possible variations that occur in natural language. Because our top priority was communication of biological knowledge, we needed GO to accommodate every individual researcher by speaking in their particular idiom.

3.5 Versioning the Ontology and Classes: The History of Changes

In 2000 we began to maintain a history of the ontology and of each term. Comprehensive snapshots of both the ontology and the annotations are taken on a monthly basis enabling progress to be quantified and retrospective analyses to be carried out. Additionally, from the outset, date stamping and authorship for each class were captured. Originally, and currently, the form is rather rudimentary: (Modified|Added|Deleted|Split from |Merged with) by firstnameorinitial,surname yymmdd. This early decision to support "micro-attribution" remains valid, but the form is gradually transitioning into a more modern approach through the development of online editing and annotation tools with authentication and authorization.

3.6 Slims

From the outset members of the community were asking for subsets of the GO containing only the major categories and subcategories, or a branch of relevance to their particular application. These "Slims" enable the users to broadly group their gene products using a very limited set of broad categories, or confine themselves to specific branches dealing with a particular biological topic, or constrain the GO by a taxonomic criterion. "Slims" are handled internally by tagging the different GO classes as members of various categories. These GO subsets are used in multiple different ways: for high-level classification; for defining sub-branches at the finest granularity; for clade specific versions; and other utility subsets.

4 Applying the GO

4.1 Evidence and Attribution

We determined that collaborating databases would be responsible for attributing any functional assignment to a source (e.g., a literature reference or computational analysis) and for indicating the type evidence used by this attribution source. The initial set of "evidence codes" was primed from this short list:

- Inferred from genetic interaction with
- Inferred from protein interaction with
- Inferred from sequence similarity with
- Inferred from direct assay

This enabled statements such as "Publication NNN" asserted that "gene A" has "function XYZ" by inference from a "direct assay." Since this time evidence codes have developed into an autonomous ontology [1] and discussed in Chap. 18 [2] but the principle remains the same: if you are asserting that something is true then you must provide the evidence—its general category and the published reference—for making this assertion.

4.2 Database Cross-Reference

GO did not arise from nothing. Like every technology it used what came before it. Furthermore, given that we wanted to give attribution to our predecessors and provide a migration path for anyone with legacy data that had utilized these prior vocabularies. This practice came out of our own need as well. As the ontology was being built up we wanted to track some of our original sources.

Thus references to Monica Riley's functional categories for *E. coli* [3], Enzyme Commission numbers (EC—http://www.chem.qmul.ac.uk/iubmb/enzyme/) and SwissProt keywords (SP—http://www.uniprot.org/docs/keywlist) helped to bootstrap GO at the outset, and additional cross-references, such as Medical Subject Headings (MESH—https://www.nlm.nih.gov/mesh/) [4], were added shortly thereafter to aid in interoperability.

4.3 Negation

Another expressivity requirement was to allow assertions stating that a given gene product does not hold for a given GO class. Experimentalists often test for an expected function, with negative results. Rather than lose this information we needed to provide a solution that could convey such negative results. Hence we provide for qualifiers on the GO annotations.

4.4 Taxon Constraints, aka Taxonomic Scope

Like most of the challenges facing the GO we recognized the need for identifying classes that are taxon specific in the very early years (1996 or earlier). The solution finally fell into place when the taxon-constraint resource and corresponding web service were implemented (e.g., http://owlservices.berkeleybop.org/isClassApplicableForTaxon?format=txt&idstyle=obo&id=GO:0005737&taxid=NCBITaxon:131567) [5].

4.5 Annotation

Following the precept of test early and often, the first annotation effort began at SGD in early 1999. Fly genes were already "annotated" because these were the seeds that GO grew from. The question was how well proto-GO, based on the needs of fly, would translate to another, very different, organism. An extremely simple tab-delimited annotation format was devised and the dialog began. Similarly the first automated pipeline "love-at-first-sight" was developed by Mark Yandell in late 1999 [6, 7] to describe the genes of the newly completed fly and human genomes. It was straightforward inference based on BLAST alignments, but it provided a reasonable overview of the landscape. The response to these first efforts was overwhelmingly positive and adoption of GO very quickly accelerated.

The GO project remains focused on providing an integrated data resource for functional information, both experimental (Chaps. 4 and 6 [8, 9]) and predicted (Chap. 5 [10]), for all known proteins, noncoding RNA sequences, and cellular components. In other words, carrying out comprehensive functional annotation is what drives the project, not the ontology itself. The ontology provides the biological model that serves as the conceptual scaffolding for the biological data. The Gene Ontology database contains currently over 5.2 million function annotations for almost 900,000 gene products (mostly proteins but also some noncoding RNAs). About 660,000 of these annotations are based on experimental results reported in the published literature, and the remainder are predictions derived from a variety of different methods. All of which are freely available for the community to use. That said, there is still considerable room for improvement. There was, and remains, a significant amount of accumulated knowledge to be captured. In particular for human, the annotation task is still more about capturing old data than capturing new data because an equivalent to a Model Organism Database does not exist. Until the day the GO catches up it will need to capture existing data in parallel with capturing more recent data to achieve the coverage it aims for.

5 Where We Stand Today

Based on the wide adoption by the community, we can claim that the project met a real need. The GO is a useful alternative to simple nomenclature, as nomenclature fails to fully convey the biology and is too limited to describe protein roles fully. There is still a long way ahead: several of the key elements that we recognized as essential in the nineties are still works in progress today.

5.1 Multiple Relationship Types and Relations Between BP, MF, and CC

In 1999 we decided at the first official GO meeting against implementing relationships across the three branches of the GO until a later time. Needless to say this drastically over-simplified the biological model, a simplification we were fully cognizant of but one that allowed us to prioritize our work. In this simplistic model with which GO began there were only two relationship types: is_a and part_of. And even here the meaning of part_of was conflated, since part_of in the cellular component branch of GO meant that that it was a sub-component while part_of in BP meant a step or subprocess. Since that time we continue to work on enriching the Relations Ontology and applying it appropriately (https://github.com/oborel/obo-relations). Currently there are eight relationships in use. Most significantly the three branches of the GO the ontologies are now being linked.

5.2 Orthogonality

We did not and do not want multiple "rival" ontologies for one domain. The initial necessity for embedding terms within other terms led to the creation of numerous implicit ontologies embedded within the GO (chemicals, anatomical parts, tissues, and cell types). In the early years, while we recognized that this might be dealt with by incorporating the unique identifier that refers to the full definition elsewhere, in practice this could not be reliably accomplished at that time and it is taking some time to remedy.

Work to rectify the situation began shortly after the turn of the century [11] and has given rise to a small set of core ontologies, which have been teased out of the GO and replaced by including the unique identifier for the new class as part of the logical definition of the GO class. The first exercise was replacing all implicit references to chemicals in the GO with explicit references to ChEBI classes [12]. Similarly the Cell Type ontology was derived from the GO [13–15] and, as an autonomous ontology, has proven its own value for other applications. Expression analyses and RNASeq experiments often draw their samples from particular cell types and projects such as ENCODE [16] and FANTOM [17] are using the cell type ontology to indicate the source cell type for their data. In addition, there are coordinated efforts connecting the cell line ontology, used in cancer studies, to the cell type ontology to indicate the original cell type [18]. There is immense benefit to constructing any ontology from its most element components because it provides a connective route across the widest possible network of projects. For example, RNA expression data from a cancer study that used a particular cell line can be automatically connected to an ENCODE RNA expression data from a normal cell type.

As regards anatomy, Jonathan Bard initially raised the question of how we might consider a common language for anatomy. It was clear that we needed a methodology for anatomical interoperability and querying data across our various organisms, not just for gene function, but ultimately for phenotypes as well. As with chemicals and cell types, a species-neutral anatomical ontology was extracted from GO, but also incorporated existing anatomical ontologies (e.g., mouse, zebrafish, fossils) thereby creating bridges between them [19–21]. Beyond its use by GO Uberon is connecting phenotype data, for example, from human (it is used for the logical definitions of the Human Phenotype Ontology) to mouse (likewise there are logical definitions underlying the Mouse Phenotype ontology) with direct applicability to human health research [22].

The challenge of comparability and interoperability can largely be overcome by community adoption of a small set of standard elemental core ontologies, from which special purpose ontologies, which meet the unique needs of a given project, can be constructed. It is hard to emphasize this enough. While the community seems to be blooming with a cacophony of idiosyncratic “ontologies” the GO is actively working to reduce the proliferation by deconstructing its terms into the elemental core set of conceptual classes needed to define its complex terms. This approach is producing enormous dividends in terms of interoperability and comparability across widely divergent data sets.

5.3 Contextual Annotation

The context in which a function is carried out was recognized from the outset as crucial. For example, the role of glucagon-mediated signal transduction in liver concerns gluconeogenesis, glycogenolysis and plasma glucose homeostasis, whereas the role of this process in adipose tissue is lipolysis. At the level of gene products, the role of cytochrome C is in oxidative phosphorylation and energy supply (when it is in the mitochondrion), and apoptosis (when it is in the cytoplasm). This has proven operationally (that is: how easy it is for someone to annotate) to be one of our biggest challenges (*see* Chap. 17 [23] on annotation extensions). While this has given curators a great deal more expressivity it still can be improved upon, and developing new annotation strategies and methods is where GO is actively working.

6 What Lies Ahead

The fundamental motivation driving the GO has remained unchanged: we are attempting to build a realistic model of biology to enable research, based on the collective evidence gathered by the research community. As originally envisioned we needed a way to describe the attributes of gene products in a rigorous way that would enable biologists to explore the universe of genomes and biology. As described above we were cognizant of them all initially and incrementally are addressing them and taking advantage of technological advances as we go.

That said, the GO is predicated upon a reliable foundation of "annotation." To gather accumulated knowledge as well as keep up with new research requires us to continue to seek new, more efficient approaches for biologists to provide their data. This is one of our current big challenges. One approach is collaborative data exchange with other annotation initiatives. For example, our collaborations with Reactome (http://www.reactome.org/) and IntAct (http://www.ebi.ac.uk/intact/) allow data from these resources to be incorporated into GO. Another key strategy is community annotation, such as described in Chap. 7 [24], which has provided GO with additional annotations. Our future plans are to provide online community annotation tools, which will also be used by GO Consortium curators—tools that will also support refinement of the GO itself in addition to providing annotations.

6.1 Phylogenetic Annotation

Providing a resource that captures functional data for every extant protein is, to say the least, a formidable challenge. One obvious reason is that most sequences are not, nor ever will be, experimentally characterized (and not just because of volume, but also because some are experimentally intractable). Therefore most annotations must necessarily be based on predictions. Furthermore, for inferences to be as accurate as possible they should be predicated on an explicit evolutionary framework. For the past several years a small group of GO curators have been using an annotation tool, Phylogenetic Annotation and INference Tool (PAINT) [25] to infer annotations among members of a protein family. PAINT allows curators to make precise assertions as to when functions were gained and lost during evolution and record the evidence (i.e., the experimentally supported GO annotations from the leaves of the tree and their phylogenetic relationship to an ancestral protein) for those assertions. PAINT is as yet a stand-alone desktop application, but work is underway to incorporate it into a suite of integrated, online annotation tools for GO curators and community contributors. Among the other tools in current development is one based on biological modules.

6.2 Modular Annotation

Biological systems are modular at many levels. For example, within a single domain a catalytic site may be coupled to an (allosteric) binding site that regulates the catalytic activity. Or, within a single protein different domains may form a module, e.g., the ligand binding domain and protein kinase domain of a transmembrane protein kinase receptor. And further up the size are functional modules composed from subunits within a macromolecular complex (e.g., the ribosome). And, at an even higher level, molecular interactions can define a pathway that can be used or reused in multiple different processes (e.g., the ubiquitin-dependent proteolysis pathway or JAK-STAT pathway). The goal of this modular approach is to define each GO term through a combination of terms, and enable extensible representation of biological modularity: how elemental molecular interactions are combined in different ways to produce compound molecular functions, how molecular functions are combined to produce processes, and how processes are combined to produce larger processes. A first release of this curation tool (dubbed "Noctua"[3]) is now being evaluated by GO curators. One notable feature of this new tool is that it combines the tasks of annotation *and* ontology construction. Historically the artificial disconnect between these two inseparable tasks created serious bottlenecks, as annotators were forced to wait for a separate group to create or modify requisite terms. With Noctua the curators will more directly describing biology, with known relationships in the ontology associated with specific instances that support this model.

7 Summary

The goal of the Gene Ontology (GO) project is to provide a uniform way to describe the functions of gene products from organisms across all kingdoms of life and thereby enable analysis of genomic data. It is an ongoing enterprise as our understanding of biology grows and is refined. It is a computational model of biological reality that we ultimately hope every researcher will happily contribute to and regard as the optimum means of sharing the knowledge they have gained from their own research with the wider community.

[3] Little owl (*Athene noctua*) is a bird that was sacred to the goddess *Athena*, the Greek goddess of wisdom.

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. Chibucos MC, Mungall CJ, Balakrishnan R, Christie KR, Huntley RP, White O, Blake JA, Lewis SE, Giglio M (2014) Standardized description of scientific evidence using the Evidence Ontology (ECO). Database (Oxford):pii:bau075. doi: 10.1093/database/bau075. Print 2014. PubMed PMID: 25052702; PubMed Central PMCID: PMC4105709
2. Chibucos MC, Siegele DA, Hu JC, Giglio M (2016) The evidence and conclusion ontology (ECO): supporting GO annotations. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 18
3. Riley M (1993) Functions of the gene products of Escherichia coli. Microbiol Rev 57(4):862–952, PubMed PMID: 7508076, PubMed Central PMCID:PMC372942, Review
4. Lomax J, McCray AT (2004) Mapping the gene ontology into the unified medical language system. Comp Funct Genomics 5(4):354–61. doi:10.1002/cfg.407, PubMed PMID: 18629164, PubMed Central PMCID: PMC2447454
5. 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. doi:10.1186/1471-2105-11-530, PubMed PMID:20973947, PubMed Central PMCID: PMC3098089
6. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, Cherry JM, Henikoff S, Skupski MP, Misra S, Ashburner M, Birney E, Boguski MS, Brody T, Brokstein P, Celniker SE, Chervitz SA, Coates D, Cravchik A, Gabrielian A, Galle RF, Gelbart WM, George RA, Goldstein LS, Gong F, Guan P, Harris NL, Hay BA, Hoskins RA, Li J, Li Z, Hynes RO, Jones SJ, Kuehl PM, Lemaitre B, Littleton JT, Morrison DK, Mungall C, O'Farrell PH, Pickeral OK, Shue C, Vosshall LB, Zhang J, Zhao Q, Zheng XH, Lewis S (2000) Comparative genomics of the eukaryotes. Science 287(5461):2204–15, PubMed PMID: 10731134; PubMed Central PMCID: PMC2754258
7. Venter JC, Adams MD, Myers EW et al (2001) The sequence of the human genome. Science 291(5507):1304–51, Erratum in: Science 2001 Jun 5;292(5523):1838
8. Poux S, Gaudet P (2016) Best practices in manual annotation with the gene ontology. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 4
9. Ruch P (2016) Text mining to support gene ontology curation and vice versa. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 6
10. Cozzetto D, Jones DT (2016) Computational methods for annotation transfers from sequence. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 5
11. Mungall CJ, Bada M, Berardini TZ, Deegan J, Ireland A, Harris MA, Hill DP, Lomax J (2011) Cross-product extensions of the Gene Ontology. J Biomed Inform 44(1):80–6. doi:10.1016/j.jbi.2010.02.002, PubMed PMID:20152934,

PubMed Central PMCID:PMC2910209, Epub 2010 Feb 10

12. Hill DP, Adams N, Bada M, Batchelor C, Berardini TZ, Dietze H, Drabkin HJ, Ennis M, Foulger RE, Harris MA, Hastings J, Kale NS, de Matos P, Mungall CJ, Owen G, Roncaglia P, Steinbeck C, Turner S, Lomax J (2013) Dovetailing biology and chemistry: integrating the Gene Ontology with the ChEBI chemical ontology. BMC Genomics 14:513. doi:10.1186/1471-2164-14-513, PubMed PMID:23895341, PubMed Central PMCID: PMC3733925
13. Masci AM, Arighi CN, Diehl AD, Lieberman AE, Mungall C, Scheuermann RH, Smith B, Cowell LG (2009) An improved ontological representation of dendritic cells as a paradigm for all cell types. BMC Bioinformatics 10:70. doi:10.1186/1471-2105-10-70, PubMed PMID: 19243617, PubMed Central PMCID: PMC2662812
14. Diehl AD, Augustine AD, Blake JA, Cowell LG, Gold ES, Gondré-Lewis TA, Masci AM, Meehan TF, Morel PA, Nijnik A, Peters B, Pulendran B, Scheuermann RH, Yao QA, Zand MS, Mungall CJ (2011) Hematopoietic cell types: prototype for a revised cell ontology. J Biomed Inform 44(1):75–9. doi:10.1016/j.jbi.2010.01.006, PubMed PMID: 20123131, PubMed Central PMCID: PMC2892030, Epub 2010 Feb 1
15. Meehan TF, Masci AM, Abdulla A, Cowell LG, Blake JA, Mungall CJ, Diehl AD (2011) Logical development of the cell ontology. BMC Bioinformatics 12:6. doi:10.1186/1471-2105-12-6, PubMed PMID: 21208450, PubMed Central PMCID:PMC3024222
16. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74. doi:10.1038/nature11247, PubMed PMID: 22955616, PubMed Central PMCID: PMC3439153
17. FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al (2014) A promoter-level mammalian expression atlas. Nature 507(7493): 462–70. doi:10.1038/nature13182, PubMed PMID: 24670764, PubMed Central PMCID: PMC4529748
18. Sarntivijai S, Lin Y, Xiang Z, Meehan TF, Diehl AD, Vempati UD, Schürer SC, Pang C, Malone J, Parkinson H, Liu Y, Takatsuki T, Saijo K, Masuya H, Nakamura Y, Brush MH, Haendel MA, Zheng J, Stoeckert CJ, Peters B, Mungall CJ, Carey TE, States DJ, Athey BD, He Y (2014) CLO: the cell line ontology. J Biomed Semantics 5:37. doi:10.1186/2041-1480-5-37, PubMed PMID:25852852, PubMed Central PMCID: PMC4387853, eCollection 2014
19. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA (2012) Uberon, an integrative multi-species anatomy ontology. Genome Biol 13(1):R5. doi:10.1186/gb-2012-13-1-r5, PubMed PMID: 22293552, PubMed Central PMCID:PMC3334586
20. Dahdul WM, Balhoff JP, Blackburn DC, Diehl AD, Haendel MA, Hall BK, Lapp H, Lundberg JG, Mungall CJ, Ringwald M, Segerdell E, Van Slyke CE, Vickaryous MK, Westerfield M, Mabee PM (2012) A unified anatomy ontology of the vertebrate skeletal system. PLoS One 7(12):e51070. doi:10.1371/journal.pone.0051070, PubMed PMID: 23251424, PubMed Central PMCID: PMC3519498, Epub 2012 Dec 10
21. Haendel MA, Balhoff JP, Bastian FB, Blackburn DC, Blake JA, Bradford Y, Comte A, Dahdul WM, Dececchi TA, Druzinsky RE, Hayamizu TF, Ibrahim N, Lewis SE, Mabee PM, Niknejad A, Robinson-Rechavi M, Sereno PC, Mungall CJ (2014) Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. J Biomed Semantics 5:21. doi:10.1186/2041-1480-5-21, PubMed PMID: 25009735, PubMed Central PMCID: PMC4089931, eCollection 2014
22. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GC, Brown DL, Brudno M, Campbell J, FitzPatrick DR, Eppig JT, Jackson AP, Freson K, Girdea M, Helbig I, Hurst JA, Jähn J, Jackson LG, Kelly AM, Ledbetter DH, Mansour S, Martin CL, Moss C, Mumford A, Ouwehand WH, Park SM, Riggs ER, Scott RH, Sisodiya S, Van Vooren S, Wapner RJ, Wilkie AO, Wright CF, Vulto-van Silfhout AT, de Leeuw N, de Vries BB, Washingthon NL, Smith CL, Westerfield M, Schofield P, Ruef BJ, Gkoutos GV, Haendel M, Smedley D, Lewis SE, Robinson PN (2014) The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res 42(Database issue):D966–74. doi:10.1093/nar/gkt1026, Epub 2013 Nov 11. PubMed PMID: 24217912; PubMed Central PMCID: PMC3965098
23. 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
24. Lovering RC (2016) How does the scientific community contribute to gene ontology? In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 7
25. Gaudet P, Livstone MS, Lewis SE, Thomas PD (2011) Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Brief Bioinform. 12(5):449-462. doi:10.1093/bib/bbr042. Epub 2011 Aug 27. PubMed PMID: 21873635; PubMed Central PMCID: PMC3178059