# SEA-PHAGES Gene Ontology Annotation
# Student Guide

## Overview
This lab unit is devoted to genome annotation. In it you will compete in teams with students from your own and other universities to annotate different aspects of genes (their molecular function, location in a cell or their participation in specific cellular processes) using the Gene Ontology as a reference framework. Teams participating in the CACAO competition earn points by submitting correct annotations and challenging inaccurate ones made by other teams.

## Objectives
After completing this lab unit you should be able to:

- Explain to a lay audience what ontologies are what they are used for
- Discuss biocurator as a viable career path in the life sciences
- Summarize how ontologies can be applied to biology
- Describe the Gene Ontology structure and its main sub-ontologies
- Critically review and assess peer-reviewed primary literature in biology
- Generate and critique Gene Ontology annotations based on primary literature
- Utilize the CACAO interface for making GO annotations
- Navigate the QuickGO and UniProt websites
- Differentiate GO terms, evidence codes and their usage
- Explain the differences between different types of GO annotations
- Be familiar with the CACAO interface for making GO annotations
- (Optional) Leverage BLAST and other tools to infer homology

## Unit Structure
This lab unit is broadly structured in three different periods: instruction, annotation/challenge and revision. During the instruction period you will receive basic training on the concept of ontology, the overall architecture of the Gene Ontology and the main concepts behind Gene Ontology annotations and their usefulness to the scientific community. You will also be given time to register and familiarize yourselves with the CACAO web interface for Gene Ontology annotation. After completing instruction, you will be able to participate in the annotation and challenge innings defined by the CACAO competition. During annotation innings, you and your team can submit as many Gene Ontology annotations as you like, but you should bear in mind that unsubstantiated or inaccurate annotations will likely be challenged and not will earn you credit. During challenge innings, you can critique other teams' annotations, providing feedback on any errors or inaccuracies present in them. As with annotations, challenges must be substantiated to earn credit. After the last challenge inning is over, you will have the chance to address any outstanding problems raised by challengers or instructor feedback. Once this final revision period is complete, your annotations are considered final and cannot be further modified. If they are accepted, your annotations will be submitted to the Gene Ontology Consortium and incorporated into their growing knowledgebase.

## Background

### *Ontologies and the Gene Ontology*
An ontology is a formal representation of a particular real-world domain (Gruber 1993). Ontologies define entities that exist in the real world (e.g. pizzas and their ingredients) and the relationships between them (e.g.

toppings are *parts of* pizzas) (**Figure 1**). Ontologies serve two main simultaneous purposes: (1) by providing a unified, controlled vocabulary ontologies eliminate synonyms (e.g. veggie pizza and vegetarian pizza) and disambiguate homonyms (i.e. same word having two different meanings in different contexts); (2) by defining relationships among entities and mappings between entities and their real-world instances ontologies enable computers to reason over the ontology and perform inferences on real-life applications.

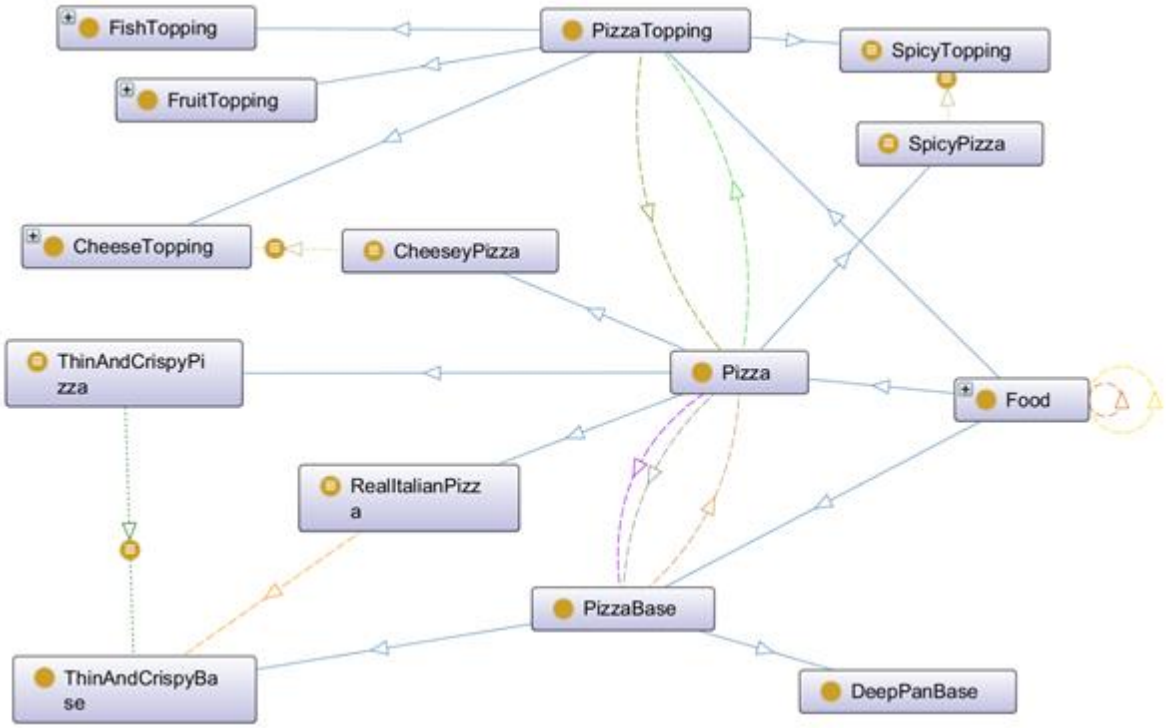



**Figure 1 –** Partial view of the Pizza Ontology developed by ontology researchers at the University of Manchester (Horridge et al. 2004). The figure shows the main entities (Food, Pizza, PizzaTopping and PizzaBase) and the different relationships (e.g. RealItalianPizza *is a* Pizza (and hence Food) that *has part* ThinAndCrispyBase, which *is a* type of PizzaBase). Image was rendered using the OntoGraph Protégé plug-in.

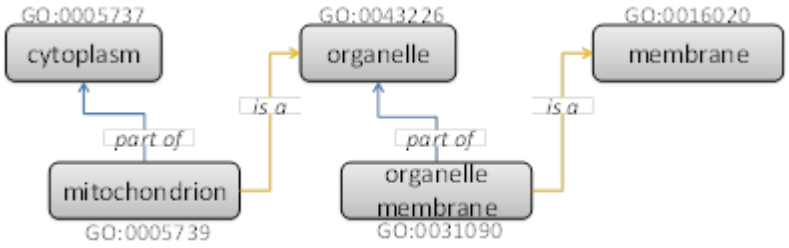
The Gene Ontology (GO) is a specialized ontology that formalizes knowledge on three key aspects of gene products (i.e. proteins, RNAs and derived biomolecules) (**Figure 2**). These three aspects make up the three GO sub-ontologies: molecular function, biological process and cellular component.

- Molecular function refers to activities that occur at the molecular level, such as "catalytic activity" or "binding activity". GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform them, and do not specify where, when, or in what context the action takes place.

- Biological process refers to a series of events accomplished by one or more organized assemblies of molecular functions. Examples of broad biological process terms are "cellular physiological process" or "signal transduction". The general rule to assist in distinguishing between a biological process and a molecular function is that a process must have more than one distinct step.

- Cellular component denotes a component of the cell that is part of a larger object, such as an

anatomical structure (e.g. rough endoplasmic reticulum) or a gene product group (e.g. a ribosome or a protein dimer)




| ID | GO:0043226 |
|---|---|
| Name | organelle |
| Ontology | Cellular Component |
| Definition | Organized structure of distinctive morphology and function. Includes the nucleus, mitochondria, plastids, vacuoles, vesicles, ribosomes and the cytoskeleton, and prokaryotic structures such as anammoxosomes and pirellulosomes. Excludes the plasma membrane. |

**Figure 2 –** (left) Schematic view of a section of the Gene Ontology, depicting the relationship between different cellular components. The mitochondrion *is a* type of organelle and is also *part of* the cytoplasm, in the same manner that an organelle membrane *is part* of an organelle but *is a* type of membrane too. (right) All terms in the Gene Ontology are defined by a unique identifier and contain the consensus name, synonyms (if any), their primary sub-ontology and a crisp definition.

### *Gene Ontology Annotations*

Beyond an exercise in modeling reality, creating ontologies is not that useful if one cannot map ontology terms to real-world entities. The Gene Ontology provides a highly structured framework to make such mappings, by means of Gene Ontology annotations. Once gene products (e.g. proteins or small regulatory RNAs) in a genome have been mapped to the Gene Ontology one can apply statistical inference and machine learning approaches to interpret data and perform genome-wide comparison. One such example is the use of the Gene Ontology in interpreting data from transcriptome analysis (du Plessis, Škunca, and Dessimoz 2011). If a genome has been mapped to Gene Ontology terms, one can interrogate sets of relevant genes (e.g. genes highly expressed in anoxic conditions) to see if they are enriched in particular subsets of the ontology (e.g. they preferentially map to stress response terms)

A Gene Ontology annotation is therefore a mapping from a given gene product to a specific Gene Ontology term (**Figure 3**). Beyond these two main components, the formalism in Gene Ontology annotations requires that the annotation contain two additional elements: a reference and an evidence code (Balakrishnan et al. 2013). The combination of these two elements is referred to as the *source* for the annotation.

### Evidence Codes

Conventional Gene Ontology annotations are typically made by professional biocurators (Howe et al. 2008). Biocurators search the literature for relevant publications containing experimental work that demonstrates the molecular function of a gene product, its involvement in a biological process and/or its location in a particular cellular component. After critically reviewing the results reported in the manuscript, biocurators identify an adequate Gene Ontology term that reflects the findings and determine what type of experimental evidence was used to demonstrate them. For instance, if the authors created a mutant of the human p53 protein and then observed that after irradiation mutant cells, compared to the wild-type, did not advance beyond the G1/S regulation point, a biocurator would use the evidence code Inferred from Mutant Phenotype (IMP) and the GO term “cell cycle arrest” (GO:0007050) to record this observation (**Figure 3**).

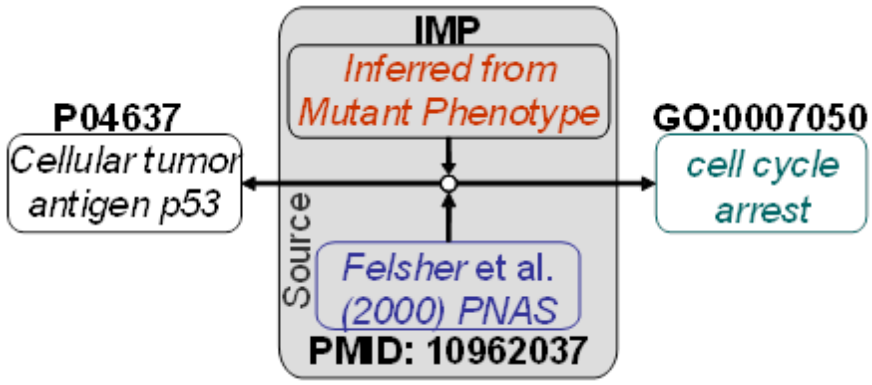



**Figure 3 –** Schematic representation of a Gene Ontology annotation. A human *cellular tumor antigen p53 gene* product (UniProt: P04637) is annotated as mapping to biological process *cell cycle arrest* (GO:0007050), based on the results reported by Felsher *et al.* in a 2000 PNAS manuscript: "Overexpression of MYC causes p53-dependent G2 arrest of normal fibroblasts" (PMID: 10962037). The experiment supporting this association in the paper is based on measurements of DNA content (as proxy for cell cycle progression) in wild-type cells and mutants expressing the human papillomavirus E6 oncogene, which facilitates the proteolytic destruction of p53. This is summarized by the evidence code *Inferred from Mutant Phenotype* (IMP).

In some cases, authors may use computational tools to determine the function of a gene product. For instance, based on sequence analysis a manuscript might report that the mouse protein P02340 is a close homolog of the human p53 protein (P04637) and that it also contains a DNA-binding motif, indicating that P02340 binds DNA in the same way as its human homolog. In such a case, the biocurator might use GO term "DNA binding" (GO:0003677) in conjunction with the evidence code Inferred from Sequence Orthology (ISO) and the identifier for the human p53 protein (P04637) that is used to make such assertion. A full list of evidence codes with usage examples is available at: http://geneontology.org/page/guide-go-evidence-codes. Gene Ontology evidence codes have now been superseded by the Evidence and Conclusion Ontology (ECO), which defines the relationships between different types of evidence (e.g. "loss-of-function mutant phenotype evidence" (ECO:0000016) *is a* type of "mutant phenotype evidence" (ECO:0000015)) (Chibucos et al. 2014). While CACAO still uses native GO evidence codes, it is often convenient to navigate ECO (http://www.evidenceontology.org/) in order to identify the proper GO evidence code to use.

## Alternative methods for Gene Ontology annotation

Even though large, the amount of available experiments determining different aspects of gene products is vanishingly small when compared to the number of genes present in sequenced organisms. Members of the Gene Ontology Consortium and others have developed tools to automatically annotate gene products in genomes using computational methods to establish homology with annotated genes or to parse manuscripts in order to extract relevant information. The reliability of these methods increases yearly, but computerized approaches are still very far from being as thorough and accurate as human biocurators. For this reason, all computer-generated annotations with no human supervision are tagged with the Inferred from Electronic Annotation (IEA) evidence code.

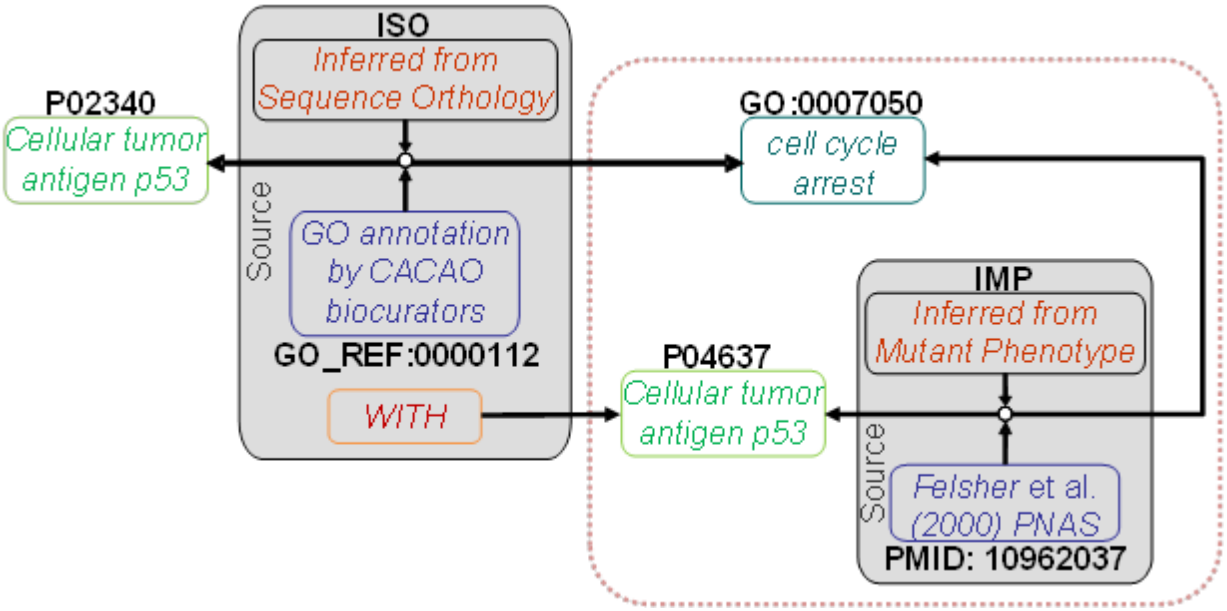



**Figure 4 –** Schematic representation of a "transfer" Gene Ontology annotation. Using the computational tools described in GO_REF:0000112, CACAO biocurators determine that the mouse p53 protein (P02340) is homologous to the human p53 protein (P04637), which has been previously annotated (**Figure 3**) as being involved in *cell cycle arrest* (GO:0007050) based on experimental (IMP) results published by Felsher *et al.* (PMID: 10962037). The assignment of the GO:0007050 term to the mouse P02340 protein is formally defined as deriving from a computational approach (Inferred from Sequence Orthology; ISO) reported in a published reference (GO annotation by CACAO biocurators; GO_REF:0000112) that establishes the homology of the mouse P2340 protein WITH the human p53 protein (P02340), allowing the biocurator to conclude that the mouse P2340 protein also participates in cell cycle arrest (GO:0007050).

Gene Ontology annotations require that a source be referenced in the annotation. Conventionally, the source is a peer-reviewed scientific manuscript reporting experiments, but there are cases in which we may want to capture results following a well-established methodology that are not published in peer-reviewed manuscripts. For instance, biocurators working on the Mouse Genome Informatics (MGI) project at the Jackson Laboratory have developed well-established computational processes to establish homology between rat and mouse genes. MGI biocurators examine, verify and contextualize these computational predictions and use them to assign GO terms to mouse genes based on experimental annotations of rat genes. When they do so, they use a special type of reference (a GO reference; GO_REF:0000008) that describes the methodology they have used in the annotation. As a sutdent participating in CACAO you can make use of a dedicated GO reference (GO_REF:0000112) to annotate gene products for which there is no available experimental literature. As in the case of MGI biocurators, you will do so through the establishment of homology with gene products containing experimental annotations using a variety of computational methods. Instead of referencing a peer-reviewed scientific manuscript, these "transfer" annotations will reference a source composed of a computational evidence code (e.g. ISO), the CACAO GO reference (GO_REF:0000112) and the identifier of the homologous protein containing the experimental annotation (**Figure 4**).

## Performing Gene Ontology Annotations

Creating a Gene Ontology annotation entails three separate steps: reading and assessment, mapping and annotating (Figure 3). The first, and most complex step, is the critical reading of a peer-reviewed scientific manuscript and the assessment of the claims made therein. Mapping refers to the identification in reference databases of the entities detailed in the manuscript (i.e. the gene product accession, the GO term and the evidence code). The last step concerns the use of CACAO to perform the annotation and submit it for review.

There are many approaches to reading scientific manuscripts, but for the purposes of Gene Ontology annotations the following procedure is recommended:

- Read the abstract carefully to get a general idea of what the paper is about and what are the main claims made by the authors. Hopefully, one of these claims will involve the function, process or location of gene product.
- Read the introduction and attempt to identify the specific species/strain the authors work on and accurate descriptions (or accession number, if provided) of relevant protein products.
- Use the NCBI RefSeq (https://www.ncbi.nlm.nih.gov/refseq/) and EBI UniProtKB (http://www.uniprot.org/) services to identify the accession numbers of the protein products referenced by the authors. If you cannot identify a valid accession number for your gene product, contact your instructor.
- Look at the *Material and Methods* section to familiarize yourself with the main experimental/computational techniques used by the authors.
- Read through the *Results* (or *Results and Discussion*) section. Most annotation-worthy claims in a scientific manuscript will be backed up by figures or tables. Identify the manuscript regions that cite a given figure to understand what the authors seek to accomplish (i.e. demonstrate) with the experiments reported in the figure. A figure reporting an experimental procedure can be the source of one or more annotations.
- Use the QuickGO (http://www.ebi.ac.uk/QuickGO/) or AmiGO (http://amigo.geneontology.org/) web services to see if the aspect the authors seek to validate through their experiments corresponds to a Gene Ontology term. The autocomplete function will suggest GO terms matching your query words. Use the *Ancestor Chart* and *Child Terms* list to navigate the ontology from any given start point. These services also provide guidelines for the annotation of specific topics (e.g. cell death). You should always aim to annotate the most specific GO term possible (i.e. if the manuscript reports the involvement of a gene in apoptosis in hepatocytes you should annotate “hepatocyte apoptotic process” and not its parent term “apoptotic process”). If you cannot find a matching Gene Ontology term, or you believe the existing ones are inadequate (e.g. too general) for the aspect you are trying to annotate, contact your instructor. CACAO has a guide on how to submit new Gene Ontology terms for approval by the Gene Ontology Consortium. CACAO students have contributed several GO terms in the past.
- Take your time analyzing the table/figure referenced in the text, and reading the figure/table legend and the text referencing it. Try to identify the type of experimental technique used in the figure (or within a figure panel) and to understand how the use of such technique allows the authors to validate the particular aspect of the gene product they identify in the main text. Ask yourself: does (do the authors claim that) the figure allows us to conclude something regarding the gene product (e.g. does it tell us that it performs a certain molecular function, that is localizes somewhere in the cell or that it participates in a specific biological process)?
- Map the experimental method to one of the Gene Ontology evidence codes. A decision tree and sampler for picking the correct experimental code are available in the CACAO webpage. The Evidence and Conclusion Ontology (ECO) is also a good resource to navigate experimental techniques and identify the relevant Gene Ontology evidence codes (which map to ECO root terms).
- Note that some evidence codes are not allowed in CACAO. In particular CACAO does not accept IPI (Inferred from Physical Interaction) and IEP (Inferred from Expression Pattern). These codes are not accepted in the competition to avoid the use of manuscripts reporting a high-throughput experiment to perform large numbers of annotations. Evidence codes based on traceable (TAS) or untraceable author statements (NAS), or inferences made by curators (IC) are also not accepted in CACAO. These terms are mostly in disuse and reserved to professional biocurators.
- Note down the GO term and evidence code, the gene product accession number and the manuscript

PubMed ID (which you can find through the NCBI Entrez interface).

- Write a concise explanation of the deductive process you have followed to determine that the annotation is possible and the terms/codes you have chosen to use. You have examples of such summaries in all previous CACAO annotations.
- Remember that a single manuscript may contain data for several annotations on one or multiple aspects of a single or multiple gene products.

The CACAO website contains example papers to train on before you perform your first annotation.

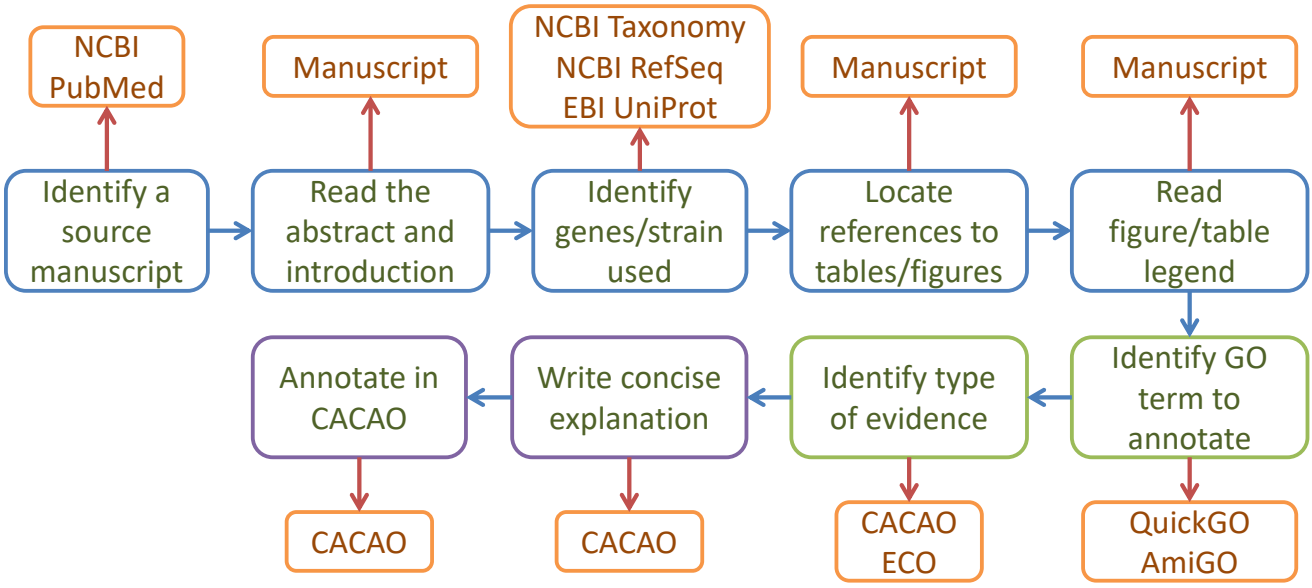



**Figure 5 –** Schematic representation of the steps in a Gene Ontology annotation and the different resources (orange boxes) used in the process. Blue boxes correspond to reading and assessment steps, green to mapping steps and purple to annotation. After completing a successful annotation, students should try to determine if further annotations can be extracted from the manuscript.

### *Performing Gene Ontology annotations with CACAO*

Performing Gene Ontology annotations in CACAO is fairly straightforward once you understand the basic elements of an annotation. CACAO provides a simple, intuitive wiki interface to generate Gene Ontology annotations. Creating a new Gene Ontology annotation in CACAO requires three distinct steps: (1) searching/creating a gene product page, (2) creating the annotation and (3) saving the changes. The following illustrates these three basic steps with the annotation example from **Figure 3**. A more detailed step-by-step annotation example is available on the CACAO website.

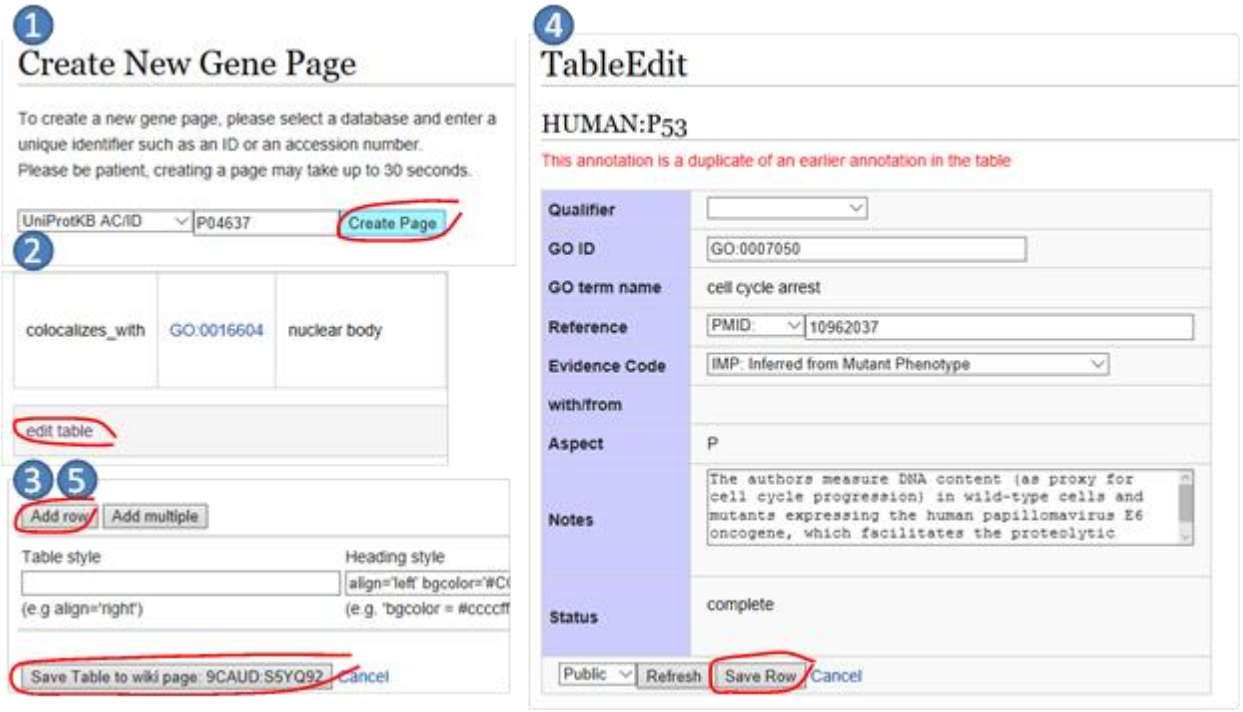



**Figure 6 –** Essential steps of a Gene Ontology annotation in CACAO. (1) If not existent, a gene product page must be created. (2) At the bottom of the annotation list, click *edit table*. Once the edit page for the table loads, click on *Add row* to create a new annotation. (4) Enter the relevant Gene Ontology annotation information, including a detailed note explaining your rationale for the annotation. Click refresh to populate GO term name and aspect and hit *Save Row* before leaving the page. (5) Once you return to the edit table page, you must also *Save the table to wiki page* for the added row (annotation) to be saved.

## Searching/creating a gene product page

The first thing to do is to search CACAO and check whether the gene product already exists in the system. If the gene product is not yet in CACAO, you can create a new gene product page by clicking on *Create New Gene Page* (**Figure 6**). When you do so, CACAO will import all relevant data for the gene, including existing Gene Ontology annotations. You should check whether annotations from the manuscript you desire to annotate from have already been made and verify that the annotation that you intend to perform has not been previously made.

## Creating an annotation

In the gene product page, at the bottom of the list of existing annotations, you will find an *edit table* link (**Figure 6**). Clicking on it will bring you to the annotations table edit page and, at the bottom of the table you will find an *Add row* button that will take you to the data entry page for the annotation (**Figure 6**). On the data entry page, you can enter all the relevant elements of a Gene Ontology annotation: the GO term, the manuscript PubMed ID, the evidence code and your rationale for the annotation.

## Saving an annotation

Once you have entered all the annotation elements, you must save the annotation. In CACAO, which is a wiki, this involves a two-step process. You must first save the row, and then save the table back to the wiki (**Figure 6**).

### Identifying manuscripts and gene products

Identifying manuscripts with reliable Gene Ontology annotations is not trivial, and in many ways it is more art than science. For starters, many manuscripts simply do not contain relevant annotations for gene products.

Some articles are reviews, which may well cite original research articles with relevant annotations but which, by themselves, cannot be used for annotation (since experiments are not carried out in the article). Many other articles, by their nature and topic, just do not contain research material for gene product annotation. For instance, an epidemiological article is unlikely to demonstrate the cellular component, molecular function or biological gene process a gene product locates, performs or participates in.

### *Finding manuscripts for annotation*

Finding manuscripts for annotation should not be too difficult (NCBI PubMed currently contains more than 27 million citations for biomedical literature), but can get a bit tricky depending on your specific assignment. NCBI PubMed (https://www.ncbi.nlm.nih.gov/pubmed) is by far the best resource for this purpose, and it has the added bonus that, once you locate the manuscript, you will have a PubMed identifier (PMID) for it (CACAO works primarily with PubMed identifiers, even though other manuscript identifiers are accepted under special circumstances).

#### Searching NCBI PubMed

The NCBI PubMed (and other NCBI databases) is accessed through a comprehensive search interface that predates Google by almost two decades. You can search with simple terms <*Escherichia coli*>, or enforcing the combination <(*Escherichia* AND *coli*)>. You can specify that you want to see the words in the title or abstract (*Escherichia*[Title/Abstract]) AND (*coli*[Title/Abstract]). You can also set up personalized Filters to see specific types of records (like those linking to a protein record. Full instructions on how to use PubMed search can be found at https://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.PubMed_Quick_Start.

#### Searching via other services

NCBI PubMed is a powerful and convenient resource, but by no means the only one. Google Scholar (https://scholar.google.com) can do a fair job at locating manuscripts that might not show up easily on PubMed. PubMedCentral (https://www.ncbi.nlm.nih.gov/pmc/) and EuropePMC (https://europepmc.org/) provide different types of search features to retrieve open-access manuscripts (which will also have a PMID and which do not depend for access on the particular journal subscriptions of your school).

### *Linking manuscripts to gene products*

In theory, an article reporting experimental work on a gene product should be an obvious source of Gene Ontology annotations. However, this is not necessarily the case. Given that performing a Gene Ontology annotation is quite time consuming, you should try to first triage any candidate manuscript before investing too much time on it. The next sections provide a few clues on what can go wrong and how to identify it (and address it if possible).

#### UniProt Identifiers

Annotations in CACAO need a unique identifier for the gene product. CACAO restricts annotations to a specific type of gene product (proteins) and uses a single source for protein identifiers: the UniProtKB database (http://www.uniprot.org/uniprot/). This means that in order to perform a Gene Ontology annotation in CACAO you will need a UniProtKB identifier. And therein lies the problem, because not all the species and strains are represented in UniProtKB. In the last few years, there has been an unprecedented surge in the number of (mostly bacterial) genomes sequenced, leading to thousands of identical protein records predicted from the genome sequences (*Escherichia coli* alone has almost 4,000 complete genomes available, most of the with identical translated protein sequences). Faced with this surge, UniProt decided to implement a redundancy reduction strategy (http://www.uniprot.org/help/proteome_redundancy) by designating some strains as reference proteomes in UniProt, and relegating other strains to the UniParc archive (with no UniProtKB

identifiers). If you cannot find a match in UniProt for the gene product reported in the manuscript, check with your instructor and/or CACAO staff (ecoliwiki@gmail.com). It is possible to annotate the gene product reported in the manuscript using the reference UniProt protein, but you should make this explicit in the annotation notes. Specifically, you should be able to locate and report in the notes the accession number of the proteome of the particular strain your organism is in and of the reference proteome you will be using (through http://www.uniprot.org/proteomes/), and detail in the notes how you have established that the protein you are annotating is a homologue of the one in the reference proteome, following the guidelines in http://gowiki.tamu.edu/wiki/index.php/Category:CACAO_GO_REF.

Undefined species/strain

Believe it or not, many scientific manuscripts reporting experimental results do not clearly identify the species/strain the work has been carried out on. Or, if they do so, they identify them in a substantially oblique manner. For instance, some manuscripts identify the strain they work on with the name of the derivative strain (e.g. an *E. coli* K-12 MG1655 strain in which a specific gene has been knocked out). The specific strain used should be named in the Abstract, the Introduction or the Materials and Methods section. In many cases, a Table with the strains used will be listed in the Materials and Methods section. If the authors use a derivative strain, they may mention at some point where it derives from, or a quick Google search with the derivative strain name may do the job. If both venues provide infructuous, NCBI Taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy) or Genomes Online (GOLD; https://gold.jgi.doe.gov/organisms) may do the trick. If you cannot easily find the parent of a derivative strain with these resources or if the authors simply do not state the strain's name, discard the manuscript and look for another one.

Undefined gene product

Gene names for which a likely annotation is possible will typically be mentioned in the abstract or the introduction (and obviously more in detail in the Results section), so scanning these two initial segments of the manuscript for a gene mention in some kind of assertive statement (e.g. "we show that") will allow us to quickly gauge whether a gene product may be annotated. As with strains, authors are sometimes not very precise about what gene or genes they are working on. This is particularly problematic in model organism (fly, worm, mouse…) and human literature, where gene names have a long history, typically multiple original naming conventions with their adherents and detractors, and where the model organism context tends to imply that the reader will know about the gene through offhand references. In many cases, a search on NCBI RefSeq or EBI UniProt with the synonym used in the manuscript will quickly resolve the issue, but in some others this may not prove easy. In such cases, as with undefined strains, it is better to discard the manuscript and move onto another.

## References Cited

Balakrishnan R, Harris MA, Huntley R, Van Auken K, Cherry JM. 2013. A guide to best practices for Gene Ontology (GO) manual annotation. Database J. Biol. Databases Curation 2013:bat054. doi:10.1093/database/bat054.

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 J. Biol. Databases Curation 2014. doi:10.1093/database/bau075.

Gruber TR. 1993. A Translation Approach to Portable Ontology Specifications. Knowl. Acquis 5:199–220. doi:10.1006/knac.1993.1008.

Horridge M, Knublauch H, Rector A, Stevens R, Wroe C. 2004. A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools Edition 1.0. The University Of Manchester.

Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S, et al. 2008. Big data: The future of biocuration. Nature 455:47–50. doi:10.1038/455047a.

du Plessis L, Škunca N, Dessimoz C. 2011. The what, where, how and why of gene ontology—a primer for bioinformaticians. Brief. Bioinform. 12:723–735. doi:10.1093/bib/bbr002.