First Gene Ontology Annotation Camp
Department of Genetics
Cambridge University
Cambridge, United Kingdom
June 13-17, 2004

Attending and research experience summary. Everyone attending introduce themselves to other camp participants and comment on their biological background and annotation camp interests.

Participant	***E-mail***	***Database & Background***
Pascale Gaudet	pgaudet@northwestern.edu	DictyBase. Slime Mould genetics
Michael Ashburner	ma11@gen.cam.uk	FlyBase, Fly genetics
Becky Foulger	ref26@gen.cam.uk	FlyBase, Fly genetics
Valerie Wood	val@sanger.ac.uk	GeneDB Pombe, sequence analysis
Midori Harris	midori@ebi.ac.uk	GO Editorial office, Yeast genetics
Jane Lomax	jane@ebi.ac.uk	GO Editorial office, Population genetics
Evelyn Camon	camon@ebi.ac.uk	UniProt-GOA, Bovine Immunology, cytokines
Emily Dimmer	edimmer@ebi.ac.uk	UniProt-GOA, Plant genetics
Michele Magrane	magrane@ebi.ac.uk	UniProt, Molecular Biology
Gill Fraser	fraser@ebi.ac.uk	UniProt, Molecular Biology
Kati Laiho	kati@ebi.ac.uk	UniProt, Molecular Biology
David Hill	dph@informatics.jax.org	MGI, Embryology & Apoptosis
Alex Diehl	adiehl@informatics.jax.org	MGI, Immunology
Li Ni	ln@informatics.jax.org	MGI
Victoria Petri	vpetri@mcw.edu	RGD
Mike Cherry	cherry@stanford.edu	SGD, Tetrahymena, ribozyme structure
Karen Christie	kchris@genome.stanford.edu	SGD, Transcription
Eurie Hong	eurie@genome.stanford.edu	SGD, Biochemistry, DNA repair
Kimberly van Auken	vanauken@caltech.edu	WormBase,
Doug Howe	doughowe@uoregon.edu	ZFIN, Myelin formation, neurodevelopment
Lisa Matthews	lisa.matthews@cshl.edu	Reactome, C. elegans genetics & dev bio

In this Draft Report to the GOC a list of annotation conclusions and action items are provided for discussion at the Chicago GOC meeting.

The following people provided text describing conclusions of the camp: David Hill, Becky Foulger, Doug Howe, Kimberly Van Auken, Evelyn Camon, Val Wood

1) **Curation examples required from all curatorial projects.** Each group will provide examples of their curation to Midori. These will be available for review from the GO website.

2) **README file required all gene association files**. We decided that every group should submit a README file along with the gene association file to summarize their current annotation strategy. For example, it should explain what genes are getting priority for annotation and whether or not groups focus on only on unique annotations or whether they include multiple annotations to the same term derived from different experiments/manuscripts. It might also include a description of the criteria that different groups use to assign terms by ISS.

3) **ANNOTATION list created.** The annotation@genome.stanford.edu email list was established to provide a venue for curators to discuss purely curation related issues.

4) **Orthology-based tool should be developed**. There was significant discussion of developing an ortholog-based tool to facilitate ortholog-based GO annotation. Much of this discussion has moved onto the GO email list.

5) **Interesting statistic**. Evidence codes listed in order of frequency of use in GO (greatest to lowest): ISS, ND, TAS, IDA, IMP, NAS, IPI, NR (no longer used), IGI, IEP, IC

6) **Protein interactions with IPI.** In the case of protein binding and its children, the evidence codes become overlapping because physical interaction is a direct assay for protein binding. We decided that if binding has been shown for a class of molecules, such as actin binding, but it is not known which specific actin molecule is involved in the interaction, then the annotation should be to the “actin binding” term and the evidence should be IDA. If we specifically know which actin molecule is being bound, then the annotation should be made to “actin binding” and the evidence should be IPI and the “with” column should be populated with the protein ID for the known actin molecule. In this case, the IPI evidence code is actually stronger than the IDA code since IPI represents a direct experiment for binding and includes information about the interacting molecules. If IPI is used in a binding annotation and the bound proteins are from the same species then it makes sense to provide reciprocal annotations. For example, if protein X binds actin isoform Y the one would expect two annotations:

   X | actin binding| IPI | UniProt:Y
   Y | protein binding | IPI | UniProt:X

   MGI allow IPI to be used when a protein from mouse is shown experimentally to interact with a protein from another species e.g. human. Then Human accession will be in the ‘with ‘ column.

7) **DNA binding evidence codes**. Assume assay of wild type protein shows DNA binding activity. If an assay is conducted where the DNA binding site(s) are lacking and the results show a lack of DNA binding activity, then it's best to use both IMP and IDA to support the “DNA binding“ term. The first is the direct assay and the second provides evidence of a mutant (the site was missing) phenotype.

8) **IGI and IPI and use of WITH column**. In general, whenever an IGI or IPI annotation is used to annotate a gene product to a process, the gene or protein in the "with" field should also be annotated to that process or one of its children. If it is not, there is little reason to believe the annotation. The specific case that started the discussion was the *C. elegans unc-29* and the annotation of unc-29 with "NOT" "TGF beta receptor signaling pathway" based on IGI with daf-4. This annotation only makes sense if daf-4 has been annotated to "TGF-beta receptor signaling pathway". ACTION ITEM: Perhaps this rule could be used in a tool to check for annotation inconsistencies using these evidence codes.

9) **Use of IGI in complimentation experiments**. We decided that IGI should be used in an experiment where a gene is transfected into a mutant cell line and is used for functional complementation. This is only the case when the transfected gene compliments the mutant. If the transfection experiment is used in a "normal" cell line to test the activity of a gene product, then the evidence code for this should be IDA.

10) **ISS with InterPro domains**. We decided to implement a rule that if an ISS annotation is made using an InterPro domain in the "WITH" field, then the annotation should be consistent with the InterPro->GO translation table. If this is not the case, then either the annotation should be changed or the translation table should be updated. ACTION ITEM: A tool that checks ISS annotations to InterPro domains for consistency with the translation table. The InterPro->GO translation table is periodically updated and thus manual annotations will drift.

11) **Component terms and IEP**. IEP should not be used to support component annotations (IDA is the correct code for e.g. antibodies or immunolocalization studies). It should be used with caution for all other annotations.

12) **Component tems using colocalizes_with qualifier**. The 'colocalizes_with' qualifier can be used for gene products that are transiently or peripherally associated with an organelle or complex. For example, a gene enriched on the surface of the polar granule, the most appropriate annotation would be to "colocalizes_with polar granule".

13) **ISS support is not appropriate for antibody assay**. Antibody cross reactivity is insufficient evidence for an ISS-supported GO annotation.

14) **Choosing the appropriate level for GO annotation**. Until it is possible to tell the direct role of a gene product, we should continue annotating to downstream processes. For example, if gene product X affects the transcription of gene product Y and gene product Y is in the Wnt signaling cascade, you can annotate gene product X to "regulation of Wnt receptor signaling pathway ; GO:0030111". Another example**,** *S. cerevisiae* Muc1p a cell wall bound protein. A *MUC1* knock-out leads to loss of invasive growth. Annotated to Process term "invasive growth : GO:0001404", with evidence IMP and IGI. Can *MUC1* be annotated to 'cell adhesion activity' using TAS code from information in the introduction? Answer: No, annotate instead to "molecular_function unknown : GO:0005554". Reason: not sure if Muc1p interacts with other cells or with other substrates. Another example, how to annotate proteins downstream of a transcription factor within a pathway. Agreed that a cell surface protein was 'output' of pathway and 'not' part of signal transduction cascade. If a protein is known to be in the middle of a transduction cascade (e.g. a scaffold protein necessary for assembly), it is reasonable to annotate protein to the process term "signal transduction : GO:0007165" using IGI but not to a function term. Reason: If the scaffold protein is not present the pathway fails.

15) **Protein dimerization annotation example**. A STAT protein can be annotated to 'protein dimerization activity ; GO:0046983' but not to “JAK-induced STAT protein dimerization ; GO:0007261” because it is the substrate of the process.

16) **When is process annotation appropriate**. There was substantial discussion of how to determine if a gene product should be annotated to a process or not. The conclusion reached acknowledges that curators must use their judgement to integrate what is actually shown in the paper (vs. author speculation in the paper), how closely a mutant phenotype is tied to a mutant gene, and whether or not a researcher would expect to see this gene in a list of genes involved in that process. In the future, a tool could be developed to scan process annotations for suspicious patterns such as annotation to GO process terms describing transcription, as well as other GO terms likely to reflect defects that are secondary to the defect in transcription. Some of these annotations will certainly reflect phenotype more closely than GO process. Future revision and updating will be required to clean these up.

17) **Points to remember for suggesting new terms**. 1) No terms that describe a mutant phenotype. 2) No terms that contain gene product names. 3) Do your homework before suggesting terms. Read as many papers as you can to make sure that the terms you're suggesting are really necessary (i.e. that the concepts don't already exist somewhere else in the existing ontology). 4) Whenever possible, include suitable references as this helps the GO Editorial office immensely with term definitions and correct placement of terms in the ontology. 5) When suggesting a term for one branch of the ontology, think about suggesting possible companion terms in other branches. Many terms still do not have definitions; suggestions for term definitions are welcome. 6) Suggest parentage, and include both the name and ID for any existing terms you mention (as proposed parents or for any other reason). 7) Please do err on the side of asking questions! If a suitable term already exists, the Editorial Office may be able to help you find it, and they can add synonyms so it's easier to find in the future.

18) **Annotation to species-specific terms**. If a general term will work, first submit that. When species-specific child terms become necessary, then a species-specific parent term would be required. Zfin has a good rule of thumb, which is to ask how often would our curators need to use this species-specific term? If the answer is a lot, then it is probably worth submitting. If it would only be used for a small number of annotations, then it's probably better to use a more general term.

19) **Annotation using meeting abstracts as references**. Although this has generally been an issue for the individual databases to decide, there seemed to be strong consensus amongst the camp attendees to not use meeting abstracts for GO curation. The main reason seems to be that it is important for users to be able to have access to the references used in any given annotation. For WormBase, meeting abstracts are used in curation, but they are trying to make sure that those used are available within the database bibliography. FlyBase curates some new genes using abstracts such as those from the annual Drosophila Research Conference. Sometimes this is the only source of information for a new *Drosophila* gene. All the newer abstracts are available online through FlyBase. When an abstract is not available online, FlyBase sends a paper copy of the abstracts when requested by a user.

20) **Policy on curation of every paper available.** Should every paper available be used for information about gene products? The feeling seems to be that ideally, we should strive to include all references for a given gene product in GO annotations. Users like this and in a sense, it provides a level of confidence to the annotations. Text mining researchers also would like to have all papers curated to aid them in the

development of Natural Language Processing techniques. In reality, this goal will be much harder for some MODs to achieve than others. Thus this is an issue for each MOD as human curation is a very resource dependent task. MGI listed ~85,000 mouse papers in their bibliography, *C. elegans* has ~7,000 and SGD has ~35,000. So, each MOD needs to decide how they are going to prioritize their annotation process, keeping the larger goal in mind. The individual approaches should then be stated clearly in an accompanying README file.

21) **Annotation of protein isoforms**. Generally, MODs are not annotating to specific protein isoforms. Yet. Provided a database has unique ID numbers for them, though, annotating to different isoforms, is fine.

22) **What to put into the DB_Object_Type column?** For IMP annotations based upon mutant alleles, gene would be the appropriate entry. Same goes for IGI. For IPI and IDA evidence codes, protein is probably the appropriate entry. However, there must be agreement between this column and column 2 (DB_ID) in the gene association file, so what has traditionally been placed in that column is really what the different MODs have available as unique identifiers. MGI and FlyBase had gene identifiers, UniProt has protein identifiers. When all object types, gene, protein, transcript, have unique identifiers, then we should retrofit our files to indicate the correct object type.

23) **Annotation with NOT**. The appropriate use of the NOT qualifier is to capture really unexpected results, not to just annotate a negative result. For example, UNC-129 is annotated as NOT "transforming growth factor beta receptor binding : GO:0005160" is okay because is annotated to the process "axon guidance : GO:0007411". Two of three HDACs not required for embryonic development is not a good use of this qualifier.

24) **Expanding GO evidence codes**. General consensus is not to increase the number of evidence codes. TAIR has increased the granularity of codes by introducing a number of subcategories for each code so that there are now 103 codes in total. They use this internally but only submit consortium-agreed codes. TAIR subcodes can be requested from TAIR. Michael Ashburner has developed a hierarchy of evidence codes but this doesn't yet incorporate the TAIR codes and it doesn't map back to the GO codes as the TAIR system does. GO codes won't be expanded but Michael's system could be used by groups who want more internal granularity. There have also been calls from some groups outside GO for better ways of determining reliability. The unintended consequence of duplicate annotations from different papers is that it increases users assurance of the annotation. More papers for 1 term increases confidence levels for that term. Each group should document how they handle this. In FlyBase, if same term is added more than once, the term is displayed only once in the summary report but each case is shown in the full report. FlyBase does not suppress IEA GO annotations from their releases even if they have manual/better GO annotation for a particular gene.

25) **Annotation of Complexes**. No need for *sensu* terms when the same complex is present in different organisms, even if subunit structure differs. Can request changes to the definition of cellular component terms describing the main complexes to increase their scope. Generic complexes: GO editors will include biological knowledge of components of complex in the definition so annotators can request that a new subunit be added. If a subunit is missing in the definition but complex has the same function then go ahead and use it. GO should capture well known stable complexes.

Evidence Code Usage Examples, plus notes on MOD specific usage: This part of the report was prepared by GOA: Evelyn, Michele, Emily, Gill and Kati.)

IDA (*inferred by direct assay*)

- Purification of recombinant protein expressed in different systems?
- For assays of mutant strains, use IMP not IDA.
- For purification of mutant protein, use IMP. But in these cases, they have generally also studied the function of the wild type too so can use IDA for wild-type protein.
- Cellular component info using antibodies/reporters: – epitope tagging usually does not relocalize protein inside a cell, so good evidence for IDA (from MGI).
- If your tagged protein ends up in vacuole, this would provide a dodgy annotation (SGD).
- Where the author does not specifically comment on e.g. location of a protein but it is clear in the figure or in the figure legend where the protein is located, MGI would assign IDA code based on that alone.
- For mammalian species where proteins have been tagged and attached to a promoter which up-regulates expression, and where authors comment that they don't find expression where expected, then this info isn't added by MGI.
- Cell localization from different organism e.g. mouse protein in HeLa cells. MGI would generally annotate this using IDA and would use a 'Note' field to store information about the cell type used (using OBO cell type ontology) . Better to capture available knowledge, even if is in a different cell type. This note field is internal at MGI only .

IPI (*inferred by physical interaction*)

- IPI – used by MGI in 2 ways - with specific process terms ('guilt by association') and with the function term 'protein binding', providing the protein ID in the 'with' column .
- FlyBase and GOA have also been annotating IPI with process terms.
- *FlyBase Example* :timeless gene, annotated with 'circadian rhythm' ; GO:0007623 | inferred from physical interaction with FLYBASE:tim; FB:FBgn0014396
- Multiple accessions are permitted in the 'with' field separated by a pipe '|' .
- MGI suggest if you have class of protein binding/can't find accession number for with field use IDA e.g. actin binding protein *but* If you have 'specific' protein binding uses IPI code. Depends on the term. For binding, IPI better evidence than IDA.
- MGI allow IPI to be used when a protein from mouse is shown experimentally to interact with a protein from another species e.g. human. Then Human accession will be in the 'with ' column. Brief discussion but not general consensus outcome on the matter.

IMP (*inferred by mutant phenotype*)

- **Penetrance:** Question raised by WormBase about is the acceptable penetrance level. Is there a cut-off for good annotation? e.g. if you see a mutant phenotype 5% of the time, is this good enough? Or is a higher level such as 50% required? Answer (MGI): Process and phenotype are 2 different things. If 5% is enough for process, it should be annotated. This can change as more information becomes available. Process could also be annotated as unknown if results seem dodgy. IMP serves as a flag that it could be a downstream effect.

- **Downstream pleiotropic effects:** Question raised by WormBase, how far to annotate? e.g. a knockout of RNA polymerase II is done and can be annotated to 'regulation of transcription from pol II promoter' but the knockout also disrupts other processes and gives different phenotypes such as those resulting from defects in gastrulation. Should all of these be added? Answer (MGI):In general, MGI try to annotate to the primary process and not everything downstream although this has to be judged on a case-by-case basis, depending on literature available and curator knowledge. Generated general discussion on how far to annotate, revisited again later (David Hill and apoptosis). If little is known and all you have is this IMP data then you should add all the processes seen to be affected. When further information is found then you could update and delete those terms known to be K/O artifacts. e.g. actin is involved in many processes - probably useful to annotate as users would want to see.
- IMP also serves as a flag that it could be a downstream effect.
- IMP for component terms: General consensus was that this usage is very rare.
- Also use for **"Non-sequence-based" mutations** (MGI):
    - Over-expression of protein (hypermorphs) is IMP.
    - Transgenes created by pronuclear injection are not annotated by MGI. This is because protein is often over-expressed in transgenic animals and creates a neomorph (tumor formation) that does not show normal protein function.
    - Abnormal functions not in realm of GO.
- Comparison of wt versus mutant strains: Use IDA, not IMP.
- If **knockout mouse** – phenotype then make transgenic in normal locus that's ok (Michael).
- For **double knockouts**, use IGI and in the 'with' field, add the second gene that was knocked out. (have to be able to rescue) For knockouts in 2 different mouse strains that give 2 different phenotypes, MGI uses IMP. Reason: When 2 genes are redundant you get no phenotype when you KO one gene, have to KO both.
- Use of IMP by MGI – MGI creates an allele record for mutant, annotates gene using IMP and uses allele record in 'with' column.

IGI (*inferred by genetic interaction)*

- Extent of annotation based on functional complementation**:** Question: What code to use if an orthologous *S. cerevisiae* gene is expressed in Candida and rescues phenotype?
- Answer: It was suggested to add term to *S. cerevisiae* gene and transfer to *Candida* by ISS. Previous discussion on GO mailing list said that IGI was acceptable. IDA also suggested as a possibility. Consensus was that IGI is fine for functional complementation studies.
- Multiple accessions are permitted in the 'with' field separated by a pipe '|' . Stating multiple accessions signifies there is a three-way (or more) way interaction taking place. If there are several two-way interactions being reported those should be listed as separate lines in the gene association file.

ISS (*inferred by sequence similarity)*

- If derived from author statement in literature but alignment is poor? Judgment call in some cases. SGD leave out 'with' column if author doesn't specify gene. Better if 'with' protein has direct assay to prove term but not necessary. Some databases such as SGD don't do any systematic sequence analysis but just use alignments from papers.
- Alignments not indicating orthology/paralogy: not ideal (MGI).

- For terms based on predications via algorithms: e.g. transmembrane domains, COG/ Pfam domains, many databases use ISS but leave 'with' column empty.
- InterPro **:** Many groups assign terms using ISS based on InterPro domains.
- Multiple accessions are permitted in the 'with' field separated by a pipe '|' .
- GOA doesn't use ISS from terms assigned by IEA but other databases do. Some databases ISS from NAS terms, some only from experimental codes.

IEP (*inferred by expression pattern)*

- Seems to be rarely used by most of the databases. Usually used with microarray .
- Most groups use it only for process terms.

NAS(*Non-traceable author statement)*

- Can be used cautiously for hypothetical statements where no experimental method described in paper, e.g. 'Data not shown' in paper:
- If an author says that something is unknown, can use NAS with pubmed of paper.
- Curators should be cautious not to over-annotate with NAS e.g. where author speculates with little evidence . Comes down to curator judgment/confidence and biological background.
- *Evelyn (GOA)* explained that the surplus of NAS and TAS codes in GOA-Human dataset was historical. The Proteome Inc. GO annotations extracted from LocusLink in 2001 did not use legal GO evidence codes. Their use of 'E' for experimental and 'P' for predicted were converted to NAS/TAS when integrated into the GOA association file.
- Also during the fast-tracking of Human GO annotation in 2001, the UniProt team curated 3000 proteins using abstracts only.
- If you see a NAS annotation from UniProt please do not assume that the experiment is not in the paper, often it is.
- As most users want more rather than less Human GO annotations removing these annotations is not an option. GOA are gradually going through Proteome Inc annotations and redoing them. All new GO annotations by UniProt curators use the full paper and all GO evidence codes.
- **Computational analysis:** What code to use for terms arising from papers with large-scale computational analysis? SGD chose TAS. Evelyn has done same with NAS. MGI agreed with TAS as curator has looked at paper. Boundary between TAS and NAS is blurred in some cases.

TAS (T*raceable author statement)*

- Generally used only if the source of the information used when the experimental details that corroborate the gene product/term association are referenced in an article/review but not described in detail. The reference for the original experiment is found in the bibliography of the article used in this annotation.
- Introductions often provide a lot of TAS statements.

IC (Inferred from curator judgement*)*

- Must be based on experimentally-assigned term. e.g. if know a protein possesses kinase activity, then could add the term ATP binding with IC code. Also note if protein has no ATP binding domain then does not have kinase activity.

ND (No Data*)*

- Don't assign 'unknown' terms until you have really looked at more than one paper.
- Can use NAS with pubmed Id if author in CURRENT paper states function is unknown.
- This code is being removed from some MOD's usage.

NR (Not Recorded*)*

- Not recorded (NR) is a legal GO evidence code only used by Proteome Inc. in 2001. *Evelyn* explained that these annotations will be updated but as some are correct can not delete all 3000 at the moment..
- Used for annotations done before curators began tracking evidence types (appears in SGD and FlyBase annotations). It should not be used for new annotations - use TAS or NAS.