GO Annotation Camp
Clark Center
Stanford University
Stanford, CA
June 1-4 2005

SUMMARY

The second GO Annotation Camp was held at Stanford University for three and a half days from June 1st through the 4th. The first GO Annotation Camp (held in Cambridge, England in June 2004) was a relatively small event (20 people) attended exclusively by people from groups that were already members of the GO Consortium (GOC) and were already familiar with using GO to annotate genes. In contrast, this second Annotation Camp was larger (55 people) and had a large proportion of attendees coming from groups that are not currently GOC members. While 19 attendees were representatives of GOC groups, there were 36 people (65% of attendees) coming from non-GOC groups. The 19 GOC attendees were either curators or PIs of established GO groups, most of which were Model Organism Databases (MODs). The background of the non-GOC attendees, with respect to GO, varied widely. Some attendees were already quite familiar with GO and had chosen to come in order to learn how to get started using GO as an annotation system. Other attendees were less familiar with GO and had come to evaluate whether GO would meet their needs. With the varied background of the attendees at this second Annotation Camp, we decided to focus on the basics to introduce the GO system and how to use it for genome annotation.

Each of the four days began with Introductions. At some point during the four days, everyone stood up and took a few minutes to introduce themselves, their background, and their current interests to the group. It was really interesting to hear what each person was doing and served to help people connect with others having interests in common. The remainder of the first day consisted largely of presentations. In the morning, there were two presentations, each given by a panel of experienced GO curators. The first presentation (given by panel 1: Karen Christie, Ceri van Slyke, Petra Fey, Harold Drabkin, Rebecca Foulger, Rachael Huntley, and Rama Balakrishnan) covered the basics of the Gene Ontologies themselves, the evidence codes, and provided a quick overview of a number of tools useful for viewing the ontologies. The second presentation (given by panel 2: Eurie Hong, Tanya Berardini, Karen Pilcher, Doug Howe, Michelle Gwinn, Russell Collins, Stacia Engel, and Carol Bastiani) described what constitutes a GO annotation and presented specific examples of making GO annotations. In the afternoon, there were two presentations (by Rebecca Foulger, formerly of FlyBase and now of UniProt, and Karen Christie, of *Saccharomyces* Genome Database) going through a single paper in detail to show what annotations could be drawn from the paper and what evidence codes were appropriate for each annotation. Ceri Van Slyke, of Zfin, also presented a specific example from a third paper. The day concluded with a time for group discussion and questions that had not yet been addressed.

On the remaining three days, there was one presentation after the Introductions. On the second day, Michelle Gwinn-Giglio of The Institute for Genomic Research (TIGR) spoke about the prokaryotic genome annotation pipeline at TIGR, which is based largely on sequence similarity based methods of comparison. This was of particular interest to many of the attendees as they will often be in a similar situation of trying to annotate a genome for which there has been relatively little experimental characterization of the genes. On the morning of the third day, Jane Lomax, of the GO Editorial Office, spoke about the activities of that office dealing with requests to add or make changes to GO terms. This was also very useful as a number of people had indicated that they wanted to know how to submit a new term. As many of these new groups will be annotating organisms not previously considered by the GOC, it is likely that there will need to be changes in existing GO terms as well as new terms to accommodate the biology of these organisms. On the fourth day, Mike Cherry spoke about creating and maintaining gene_association files and Jane Lomax spoke about the other ontologies available from the Open Biomedical Ontologies umbrella, both of which are practical topics for any group that will be using GO and other controlled vocabularies to annotate a genome. The presentations given during the GO Annotation Camp can be found here:

ftp://ftp.geneontology.org/pub/go/teaching_resources/presentations/

One of the main activities of the Annotation Camp was the Small Group Working sessions. In preparation for the Camp, all attendees were asked to submit two papers that they would be interested in talking about at the Annotation Camp. They were instructed to "choose a paper that presents a tricky issue that you want to talk about or a paper that you feel is a good sample paper from your research community." They were also asked to send some information about each paper, including a 1-2 sentence synopsis of the key issue of each paper, with respect to why they chose it for the annotation exercises. The papers submitted by attendees were used to select papers to read for Small Group Working sessions. In the course of three such sessions, each person read four papers. The group was divided into twelve small groups, of 3-5 people, where at least one person was a curator from a GOC group. Each group read a paper and discussed what annotations could be made from the paper. Having at least one GOC member per group gave all the people new to GO access to someone experienced while they got some practice in making GO annotations. This exercise worked well when the papers had been selected by GO Consortium members, with one or more specific aspect (Function, Process, or Component) or Evidence Code in mind. It was more variable when done with papers selected by non-GOC attendees; some papers were not suitable for GO annotations at all. As a consistency exercise, two different groups read each paper, with the results recorded in a table. At the end of the day, for both day two and day three, there was a large group discussion. This discussion worked best when both of the groups that had read the same paper got together in advance and compared their choices. Next time, it would probably be better to select a smaller set of papers from the list of sample annotation papers being developed by the GOC. However, despite refinements to improve the exercise, the Small Group Working sessions were found to be very useful by the majority, as evidenced both by verbal comments and the responses to our survey.

Overall the GO annotation camp went very well. Funding from the Stanford Genetics Department and an anonymous gift allowed the participation of a number of people to attend who otherwise would have been unable to come.

PARTICIPANT LIST

Consortium Members

Petra Fey	<pfey@northwestern.edu>	DictyBase
Karen Pilcher	<kpilcher@northwestern.edu>	DictyBase
Russell Collins	<r.collins@gen.cam.ac.uk>	FlyBase
Jane Lomax	<jane@ebi.ac.uk>	GO Editorial Office
Judith Blake	<jblake@informatics.jax.org>	MGI
Harold J Drabkin	<hjd@informatics.jax.org>	MGI
Jennifer R. Smith	<jrsmith@mcw.edu>	RGD
Rama Balakrishnan	<rama@genome.stanford.edu>	SGD
Mike Cherry	<cherry@stanford.edu>	SGD
Karen Christie	<kchris@genome.stanford.edu>	SGD
Stacia Engel	<stacia@genome.stanford.edu>	SGD
Eurie Hong	<eurie@genome.stanford.edu>	SGD
Tanya Berardini	<tberardi@acoma.stanford.edu>	TAIR
Rachael Huntley	<huntley@stanford.edu>	TAIR
Michelle Gwinn-Giglio	<mlgwinn@tigr.org>	TIGR
Rebecca Foulger	<rfoulger@ebi.ac.uk>	UniProt
Carol Bastiani	<bastiani@its.caltech.edu>	WormBase
Doug Howe	<dhowe@cs.uoregon.edu>	ZFIN
Ceri E. Van Slyke	<van_slyke@uoneuro.uoregon.edu>	ZFIN

Other Participants

Edward Braun	<ebraun@zoo.ufl.edu>	University of Florida
Vasily Cherepanov	<vasilyt@uvic.ca>	University of Victoria (Canada)
Candace Collmer	<ccollmer@wells.edu>	Wells College
Colleen Crangle	<crangle@converspeech.com>	ConverSpeech LLC
Peter D'Eustachio	<deustp01@med.nyu.edu>	NYU School of Medicine
Christine Elsik	<c-elsik@tamu.edu>	Texas A&M University
Kamal Gajendran	<kg@ncgr.org>	Nat. Center for Genome Resources
Jeremy D. Glasner	<glasner@svm.vetmed.wisc.edu>	University of Wisconsin
Mark Hance	<mhance@vbi.vt.edu>	Virginia Bioinformatics Institute
Mark Heiges	<mheiges@uga.edu>	University of Georgia
Heather Hood	<hmhood@ebs.ogi.edu>	Oregon Health & Science Univ.
Trupti Joshi	<joshitr@missouri.edu>	Univ. of Missouri-Columbia
Joseph Karalius	<joseph.karalius@seminis.com>	Seminis Vegetable Seeds
Varsha Khodiyar	<varsha@galton.ucl.ac.uk>	University College London
Cindy Krieger	<cindy@genome.stanford.edu>	Tetrahymena Genome Database
Kitsos Louis	<louis@imbb.forth.gr>	Inst. of Mol. Bio. & Biotech. (Greece)
Xinghua Lu	<lux@musc.edu>	Medical Univ. South Carolina
John MacMullen	<macmw@email.unc.edu>	UNC Chapel Hill
Chunhong Mao	<cmao@vbi.vt.edu>	Virginia Tech
Emma Master	<emaster@gene.concordia.ca>	Concordia University
Paul Morris	<pmorris@bgnet.bgsu.edu>	Bowling Green State University
Lukas Mueller	<lam87@cornell.edu>	Cornell
Nicole Perna	<perna@svm.vetmed.wisc.edu>	University of Wisconsin
Anjan Purkayastha	<anjan@vbi.vt.edu>	Virginia Bioinformatics Institute
Shingo Sakaniwa	<sakaniwa@chanko.lab.nig.ac.jp>	National Inst. of Genetics (Japan)
Jun Sese	<sesejun@cb.k.u-tokyo.ac.jp>	University of Tokyo
Jaswinder Singh	<jsingh@nature.berkeley.edu>	UC Berkeley
Nicholas A. Stover	<nick@genome.stanford.edu>	Tetrahymena Genome Database
Tsuyoshi Tanaka	<tstanaka@affrc.go.jp>	Natl. Inst. Agrobiol. Sci. (Japan)
Pantelis Topalis	<topalis@imbb.forth.gr>	Inst. of Mol. Biol. & Biotech. (Greece)
Sucheta Tripathy	<sutripa@vbi.vt.edu>	Virginia Polytechnic and State
Babu Valliyodan	<valliyodanb@missouri.edu>	University of Missouri
Chisato Yamasaki	<cyamasak@jbirc.aist.go.jp>	BIRC, AIST
Gongxin Yu	<gyu@vbi.vt.edu>	Virginia Tech
Jim Zheng	<zhengw@musc.edu>	Medical Univ. South Carolina

COLLECTED ACTION ITEMS

Below are the collected action items, and one potential additional item for discussion by the GO Consortium as a whole. Each action item also appears in context after the question and discussion that spurred it.

ACTION ITEM 1: We need to expand the documentation for IMP to be more explicit about the fact that IMP is appropriate for ALL cases involving changes or variation in a single gene, not just for cases where a given allele is designated "wild type" and others are designated "mutant".

ACTION ITEM 2: Question for the GO list: Should a difference in philosophy regarding use of TAS be included in each group's README file?

ACTION ITEM 3: Revisit the guidelines for TAS? The discussion at the recent Annotation Camp highlighted that it doesn't really seem reasonable to use TAS for common knowledge annotations. The newer code IC seems better suited to represent this type of annotation, especially since often there is not a traceable statement in the reference used, e.g. Stryer, that can actually be traced back to original data. Use of TAS in this situation seems a hold over from the idea that the evidence codes can provide an indication of the reliability of the annotation and the desire to use a "better" evidence code. However, we have debunked this idea every time we have discussed it a Consortium meeting, so perhaps it is time to revisit whether the evidence code TAS should be allowed for a common knowledge statement, even if it is not traceable in the reference used.

ACTION ITEM 4: Confirm/clarify with GO Consortium as a whole when it is appropriate to use an experimental evidence code for an experiment not actually shown in the paper.

ACTION ITEM 5: Ask the GO Consortium whether genomic structure and synteny can be included as acceptable evidence to include in the ISS evidence code.

ACTION ITEM 6: Can the people representing new groups here, please let us know the status of your projects, and if there is any assistance from the GO Consortium that would be useful to you. Please send this info to Jennifer Clark (jenclark@ebi.ac.uk).

POSSIBLE ACTION ITEM: Perhaps the GO group should discuss whether the question below should have a definitive answer. If diversity of practice is acceptable, then the specific practice should be described in each group's README file.

> **QUESTION:** If a paper with a sequence comparison, for example paper 13 (Keeling PJ and Palmer JD) that you wish to use for annotation does not provide the sequence accession number, should you go look it up? **DISCUSSION:** Different groups had different practices. Some felt that they must have an accession # in order to make an annotation, and they would try to identify the appropriate accession #. Other groups felt that it was asking for trouble to make assumptions about what accession # corresponded to the sequence the authors has used, that if it wasn't listed in the paper, it wasn't appropriate to fill one in.

TOPICS DISCUSSED

There were numerous times for group discussion during the Annotation Camp. Presentations were informal and often stimulated discussion. In addition, at the end of each day was a scheduled 90-minute time slot for group discussion to discuss the issues that had arisen during the day. The following is a summary of the various questions asked and resulting discussions that took place over the course of the 3 and a half day camp, organized in a Question and Discussion format, where questions about similar topics have been grouped together regardless of the actual order of questions and discussion. There are a handful of action items relating to topics that should be referred to the GO list for discussion and a decision.

EVIDENCE CODES

Cross species expression

IDA versus IGI

- **QUESTION:** What evidence code should be used if the researchers express a protein from one organism in a system based on a different organism? **DISCUSSION:** For a cross species expression experiment, e.g. human RAS protein complementing a RAS mutation in yeast, the appropriate evidence code is IGI. This is considered to be a genetic interaction because two different genes are interacting, the yeast RAS gene (mutated in this case) and the human RAS gene (functional in this case). This particular type of experiment is not considered to be a direct assay (IDA) because it is an *in vivo* genetic complementation assay. The expressed protein is not directly shown to have a specific activity, only that it is able to compensate for the lack of activity in the mutant.

NOTE: In discussion afterwards, the case of transfections and assaying a resulting extract was brought up. This was not discussed at the Annotation Camp.

IGI versus IMP

- **QUESTION:** What evidence code should be used if the paper shows rescue of a deletion in one organism with the wildtype gene from another organism? **DISCUSSION:** As decided previously, cross species complementation goes under IGI. Situations involving a mutant or allele of a single gene are considered IMP.

IDA

- **QUESTION:** How does one annotate recombinant proteins? **DISCUSSION:** When a recombinant protein is purified and assayed in an *in vitro* system, this is typically classified as IDA.

NOTE: It was brought up in discussion afterwards that there are non-*in vitro* ways to assay recombinant proteins; such experiments would not necessarily be considered to be IDA.

IMP

- **QUESTION:** What kinds of annotations can be made from phenotypes? **DISCUSSION:** GO tries to describe the normal processes and functions, not the phenotypes of mutants. However, analysis of mutant phenotypes often allows reasonable suppositions about the normal function of a given gene product. Typically, phenotypes of mutants (or allelic variants) allow annotations of process, if not of function.

- **QUESTION:** Is it reasonable to make an annotation using a gain of function mutation? Perhaps the only mutation available is a gain of function allele that is known to be constitutively active. Can a GO annotation be made from such a mutation? **DISCUSSION**: While Russ felt that the example that initiated the discussion with was not appropriate for annotation, others thought such an allele could be useful for annotation, particularly if nothing else is available. The basic conclusion was that it depends upon the situation. The basic goal of GO is to provide some indication as to the normal process, function, and location of the gene product. If a curator judges that a gain of function mutation provides insight as to the normal role of the gene product, then it is fine make an annotation based on it, with the IMP evidence code.

- **QUESTION:** How does one deal with polymorphism or allelic variation at a given locus ? **DISCUSSION:** This is within the domain of IMP. Though not considered a "mutant" by those studying these situations, GO places all situations where one is looking at a change in a single gene, whether it's considered to be a "mutant" versus "wild type" situation or "normal" allelic variation within a population, under the evidence IMP. Remember that the designation of a certain allele as "mutant" versus "wild type" is often context dependent anyway.

ACTION ITEM 1: We need to expand the documentation for IMP to be more explicit about the fact that IMP is appropriate for ALL cases involving changes or variation in a single gene, not just for cases where a given allele is designated "wild type" and others are designated "mutant".

IEP

- **QUESTION:** When is it appropriate to use IEP? **DISCUSSION:** Most groups feel that the IEP evidence code should be used very sparingly. However, occasionally, it is useful. For example, if a group performs microarray expression analysis and finds a cluster of genes with the same expression pattern, where nine of ten are known to be involved in ribosome biogenesis and one protein is uncharacterized. In a really tightly linked cluster like this, it may be reasonable to assign a process annotation to "ribosomal biogenesis".

IMP versus IPI

- **QUESTION:** How would GO annotate two-hybrid and deletion mutant experiments that determine which portion of a protein interacts? **DISCUSSION:** Use of the IPI evidence code will allow you capture the other protein that interacts with the protein being annotated. Note though that GO does NOT capture the specifics of which domain is involved in the interaction. This is beyond the scope of GO.

ND

- **QUESTION:** Is there a place to indicate the date that an annotation using the ND (no data) evidence code was made? **DISCUSSION:** Every annotation, regardless of evidence code, is supposed to be associated with a date indicating when it was made. The date column in the gene_association file is supposed to contain the date the annotation in that row was made, not the date the file was created or committed or any date that is general to the file. Each row should have its own date indicating when the association between the gene product and the term in that row was made. For annotations to the unknown terms using the ND code, the date is particularly useful because it provides the date when the literature and/or sequence databases were checked allowing users and curators to know that if there is a paper about the given more recent than that date, then it may be possible to update the unknown annotation to something more specific.

ND versus TAS/NAS

- **QUESTION:** What evidence code should be used for annotations to the unknown terms? **DISCUSSION:** There are a couple different situations for making annotations to the unknown terms that are handled with different evidence codes. For example, if an author says that nothing is known about the function of a given gene product, it is permissible to cite their paper for the annotation and use either TAS or NAS as appropriate. However, if a curator searches the literature and possibly also does sequence analysis and there is “no information” available for a given gene, then there is no research paper to cite. In this case, some groups, those that are able to create internal references, cite their own internal reference regarding the use of the ND evidence code. The GO consortium has also created a generic reference for use of the ND code so that groups that do not have the ability to create internal references may cite the GO reference. Exactly what constitutes “no information” depends slightly on the group. Some groups, like SGD, look only for available literature and do not do any sequence analysis. Other groups, like MGI, look at both literature and internally performed sequence analysis.

TAS

- **QUESTION:** When is it appropriate to use the TAS evidence code, or is it better to follow the trail back to one of the cited papers? **DISCUSSION:** The appropriate course depends on the specific situation. However, not all references cited in a review are relevant to the organism you are trying to annotate. Sometimes the reference is about a gene from a different organism that is similar to the gene in the organism you are trying to annotate. Thus, it is a good idea to make sure that the reference cited in a review is for the organism you are annotating. There is also some variability between groups whether the philosophy is to annotated from a review using TAS, or to use that review to trace back to the original research papers and annotated from those papers with evidential evidence codes. This difference in philosophy is acceptable, though perhaps this should be documented in each groups’ README file.

- **QUESTION:** Is it permissible to use information from the introduction of a research paper to annotate a gene? **DISCUSSION:** Different groups have different philosophies on this question. Some groups, including SGD, allow curators to make an annotation from the introduction with TAS, especially if this TAS annotation may facilitate curators being able to use the associated reference at a later date to trace it back to a paper with original data. Other groups, including MGI, do not allow this, but instead require the curator to trace the statement back to the original source to make the annotation from the original reference, if appropriate. RGD curators are also required to trace the statement back to the cited sources before making any annotation. They have noticed that sometimes the cited references do not actually deal with the organism of interest, but some other organism; thus the cited papers would not actually be allow any annotations to be made for the organism of interest.

- **ACTION ITEM 2:** Question for the GO list: Should a difference in philosophy regarding use of TAS be included in each group's README file?

<u>TAS versus IC</u>

- **QUESTION:** What evidence code is appropriate to use for statements of "common knowledge"? **DISCUSSION:** The current documentation states that TAS may be used as the evidence code for statements of common knowledge. For example, let's say you have a paper that says that Protein X is an xxxxx , with a direct assay for activity, so you can use IDA for this function term. Then it also makes a mutation in the gene for Protein X and shows that it is involved in process yyyy, so you can use IMP for the process term. But, the paper does not have any direct evidence about the localization of Protein X. However, everyone knows that process yyyy occurs in the cytoplasm, so you can annotate protein X to the component term "cytoplasm ; GO:5737" by TAS using a general reference like Biochemistry by Lupert Stryer. TAIR stores each chapter of a book as a separate reference. This will help in situations where the book is felt to be an appropriate reference for the annotation. However, Karen commented that SGD no longer allows use of TAS and the general textbook by Stryer for the situation represented by the component annotation in the example above, as there is not really a traceable statement in Stryer providing evidence that process yyyy occurs in this location in yeast. SGD feels that it is better to use the newer evidence code IC for these "common" knowledge types of annotations. Thus, if an SGD curator felt that it was reasonable to make the annotation "cytoplasm" based on the knowledge that Protein X the process annotation yyyy, then the curator could assign the component term "cytoplasm ; GO:5737" using IC and the GOid of the process term yyyyy. This seems a more accurate representation of the annotation, that the curator used common knowledge to assign the component term, based on the process term. Peter d'Estachio further commented that many of these "common knowledge" types of statements are often not well based in actual experiments conducted on the organism of interest, that early biochemists would often perform experiments with materials that were easy to obtain, e.g. calf thymus, and assume that this accurately represented the situation for another organism, e.g. human. This may or may not be the case.

ACTION ITEM 3: Revisit the guidelines for TAS? The discussion at the recent Annotation Camp highlighted that it doesn't really seem reasonable to use TAS for common knowledge annotations. The newer code IC seems better suited to represent this type of annotation, especially since often there is not a traceable statement in the reference used, e.g. Stryer, that can actually be traced back to original data. Use of TAS in this situation seems to stem from the idea that the evidence codes can provide an indication of the reliability of the annotation and the desire to use a "better" evidence code. However, we have debunked this idea every time we have discussed it a Consortium meeting, so perhaps it is time to revisit whether the evidence code TAS should be allowed for a common knowledge statement, even if it is not traceable in the reference used.

IC

- **QUESTION:** When is it appropriate to use IC? For example, transcription factors may also be found in the cytoplasm, but would you make an annotation like this by IC? **DISCUSSION:** The IC evidence code is basically meant to be used in those cases where you cannot find any references from which to make an annotation in typically one, but possibly two, of the ontologies, but based on the GO annotations that have already been made for the gene in the other ontologies, the curator feels that a reasonable annotation can be made in the remaining ontology. One of our classic examples for IC is that of a transcription factor. For example, let's say a curator has annotated a gene product to the function term "transcription factor activity ; GO;0003700" and to the process term "regulation of transcription from RNA polymerase II promoter ; GO:0006357", but cannot find any data that discusses the location of the gene product. However, for an annotator of a eukaryotic organism, with the knowledge that it is a transcription factor, it is reasonable to assume that it must be in the nucleus at least some of the time in order to act as a transcription factor, so the curator may make the component annotation "nucleus ; GO:0005634" with the evidence code IC from GO:0003700 and GO:0006357, rather than leaving this gene with no component annotation or assigning the term "cellular component unknown ; GO:0008372". It is true that some transcription factors are regulated at the level of localization and are thus found in the cytoplasm until they are activated. If there is experimental evidence showing that a transcription factor is found in the cytoplasm, then you can give this annotation the appropriate evidence code. However, it would not be reasonable to assign the component term "cytoplasm ; GO:0005737" to a transcription factor by IC, because it is not reasonable to assume that all transcription factors will be localized to the cytoplasm at some time. Note that it is possible for the same gene product to be annotated to both "nucleus ; GO:0005634" and to "cytoplasm ; GO:0005737" as GO allows all appropriate annotations to be made. Note also that we don't assign the term "cytoplasm ; GO:0005737" to every protein, just because it was made in the cytoplasm. This isn't really useful. Typically, the term "cytoplasm ; GO:0005737" would be assigned if there is experimental evidence showing cytoplasmic localization or it may be inferred by the curator if the cytoplasm is thought to be the normal localization of that type of gene product, but there are no experiments performed with this specific gene product in this organism.

IC versus ISS

- **QUESTION:** In paper 13 (Keeling PJ and Palmer JD), they do a sequence comparison and say that this indicates that the gene product in question is an enolase, which corresponds to the function term "phosphopyruvate hydratase activity ; GO:0004634" . Based on this information, what would you do for the process and component annotations? Would the appropriate evidence code be IC or ISS? **DISCUSSION:** The appropriate course of action depends on the specific situation. For example, there are some InterPro domains, e.g. IPR000013 shown below, whose mappings in the interpro2go file include process and component, as well as function, terms. In a case like this, it would be appropriate to use ISS with IPR000013 for all three annotations, as these come directly from the interpro2go file. However, if the mapping file gave only a function annotation and the curator was the source of any further annotations, then IC should be used.

```
InterPro:IPR000013 Peptidase M7, snapalysin > GO:metallopeptidase activity ; GO:0008237
InterPro:IPR000013 Peptidase M7, snapalysin > GO:proteolysis and peptidolysis ; GO:0006508
InterPro:IPR000013 Peptidase M7, snapalysin > GO:extracellular region ; GO:0005576
```

NAS versus experimental evidence codes

- **QUESTION:** What evidence code do you use for an experiment referred to in the paper, but not directly shown? **DISCUSSION:** At the Cambridge GO camp this was bought up, but different people seem to remember different outcomes for the discussion. Becky and David (MGI) remember that it was OK to use experimental evidence codes for this, with the theory that the authors had done the work but most likely had to remove the figures from the paper due to space restrictions imposed by the journal. In the Cambridge meeting minutes, this specific discussion doesn't seem to be covered but it does say in the 'Evidence Code Usage' section that NAS can be used for 'data not shown', and does not specify that an experimental evidence code may be used for an experiment not shown in the paper. At this Annotation camp, this issue was bought up again because different groups were annotating 'data not shown' in different ways: some were using NAS, and others were using experimental codes. The consenus from the Stanford GO camp discussion was that it is okay to use an experimental evidence code if the authors, for example, show an assay for enzymes A, B and C and then say "and D, E and F have the same activity (data not shown)", so that the Materials & Methods section shows a protocol for that experiment. But if they show an unrelated experiment with 'data not shown' then NAS should be used because you can't see any comparable experiment and it wouldn't always be clear which experimental code should be used.

NOTE: In subsequent discussion, it was apparent that this issue is not resolved. The Action Item below needs to be discussed at the next Consortium meeting.

ACTION ITEM 4: Confirm/clarify with GO Consortium as a whole when it is appropriate to use an experimental evidence code for an experiment not actually shown in the paper.

Multiple evidence codes for the same annotation

- **QUESTION:** Should I use both ISS and IDA as evidence for annotations to the same GO term? **DISCUSSION:** Sure, generally, the more different pieces of evidence, especially when of different types, you can use to support an annotation, the more confidence you can place in the annotation. There may be cases where you feel that keeping an ISS annotation, even when you have an IDA annotation to the same term, is appropriate.

Making annotations based on sequence comparisons

IEA and ISS

- **QUESTION:** When would one use ISS versus IEA for annotations made via sequence similarity comparisons? **DISCUSSION:** To use ISS, a human must have examined and verified the annotation. For example, if you make GO annotations by doing an automated scan of InterPro hits, then you must use IEA because there has not been any human judgment involved in making the annotations.

ISS

- **QUESTION:** I am trying to annotation an organism where there are two different "species": A and B, both of which are parasites, but one interacts with plants and the other with animals. I am trying to annotate the sequence of species A, however for some genes all of the data comes from species B. What annotations are reasonable to make for species A by sequence similarity comparison? **DISCUSSION:** We often find that sequence similarity comparisons are better suited to making annotations in the Molecular Function ontology than in either the Biological Process or Cellular Component ontologies. For example, pectin metabolism is relevant to the species that interacts with plants, but not to the species that interacts with animals, so a process annotation to "pectin metabolism" just does not make sense. In a case like this, it is often reasonable to make function annotations by ISS, though it is often not reasonable to make any process annotations by ISS.

- **QUESTION:** Can one use orthology to make GO annotations? **DISCUSSION:** One needs to be very careful about transferring GO annotations, particularly process, to orthologs. Even if orthologs in different species are still performing the same molecular activity (function), they might not be involved in the same processes. For example, Zebrafish has undergone a genome duplication. When there is an ortholog, we often give both copies the same annotation with ISS evidence from a gene in a different species. At a later time, when there is experimental evidence in Zebrafish, one of the orthologs may receive a NOT qualifier to this annotation if it is shown to be involved in some other process but not the one predicted by the original sequence similarity comparison.

- **QUESTION:** It seems that an annotation based on a simple pair-wise BLAST is not as good as an annotation based on meeting the threshold to match an HMM. How does GO indicate the quality of an ISS annotation? Would it be possible have a system to differentiate the quality of the annotation? **DISCUSSION:** The GO Consortium is aware that there are many different levels of quality of evidence used for annotations made with the ISS (or any other) evidence code. We have discussed, on multiple occasions, the idea of having a more granular hierarchy of evidence codes. However, each time, we have come to the conclusion that the type of method does NOT automatically provide any information about the quality of the annotation. To judge the quality of the annotation, one really needs to look at the specific details of the method used. Both for this reason, and to maintain a simple and easily used system of evidence codes, the GOC has decided against using an expanded hierarchy of evidence codes. In the example asked about, a simple pair-wise BLAST versus a match to an HMM, one can look at the methods described in the references for the annotations and make a judgment.

- **QUESTION:** Is it possible for GO to set standards for ISS annotations? **DISCUSSION:** It is not possible to establish a single cutoff score that is meaningful across-the-board. Michelle Gwinn will touch on this subject more thoroughly in her talk. What GO requires is for each ISS method to be described in an abstract that is used as the reference for all annotations made via that method. In this way, users can judge for themselves what criteria were used.

- **QUESTION:** Using sequence similarity methods, how do you decide which function is the most appropriate, which sub-function or which family? **DISCUSSION:** This question was deferred as Michelle Gwinn's talk about the prokaryotic annotation pipeline at TIGR will address this type of issue.

- **QUESTION:** In paper 1 (VanWagoner et al), they discuss an operon/gene cluster. What evidence code is appropriate? Is ISS acceptable for this **DISCUSSION:** We discussed whether genomic structure and synteny are acceptable evidence to include in the ISS evidence code. People seemed to feel that consideration of genome organization and synteny seemed reasonable to be included in the ISS evidence code. However, no one could remember this question ever being discussed at a Consortium meeting previously.

ACTION ITEM 5: Ask the GO Consortium whether genomic structure and synteny can be included as acceptable evidence to include in the ISS evidence code.

- **QUESTION:** Are you allowed to cite a paper for an annotation by ISS if they do not give an accession # for the protein, nucleic acid sequence, InterPro domain, etc. they used for the comparison? **DISCUSSION:** While it is best to have a with statement for an ISS annotation, it is not mandatory. It is up to each group to establish their own requirements here. SGD, which does not do any of its own sequence analysis, has relied exclusively upon the published literature for those ISS annotations that it has currently. If an accession number has been given, then we will put it in the with column. However, if no accession number is given, we may make the ISS annotation anyway, even though we cannot fill the with column. We do not try to track down a sequence identifier if none was given. If you did track one down, could you ever be certain that it was actually what the author used?

- **QUESTION:** If a paper with a sequence comparison, for example paper 13 (Keeling PJ and Palmer JD) that you wish to use for annotation does not provide the sequence accession number, should you go look it up? **DISCUSSION:** Different groups had different practices. Some felt that they must have an accession # in order to make an annotation, and they would try to identify the appropriate accession #. Other groups felt that it was asking for trouble to make assumptions about what accession # corresponded to the sequence the authors has used, that if it wasn't listed in the paper, it wasn't appropriate to fill one in.

POSSIBLE ACTION ITEM: Perhaps the GO group should discuss whether the above question should have a definitive answer. If diversity of practice is acceptable, then the specific practice should be described in each group's README file.

- **QUESTION:** Can you make an ISS annotation for the function of gene C by ISS with gene B, even if the only function annotation for Gene B was generated by ISS with Gene A? **DISCUSSION:** This is not recommended procedure. This type of transference of annotations, from an uncharacterized protein to another uncharacterized protein, has the potential to pollute the pool of annotations with information that is not well supported by either experimental data or by good sequence similarity to a characterized protein. Some groups, like MGI, do not allow this type of "ISS from an ISS" annotation to be made. However, MGI does do things to attempt to make it clear what is known about sequence similarity of its genes to other species and what information can be drawn from that knowledge. MGI has a "special reference", J: 73065, that means "orthology established by MGI (either by MGI itself, or by literature referencing)'. Using this internally-generated reference as the reference for the GO annotation, if MGI has determined that a given mouse protein has orthologs in other closely related species, it can make GO annotations for the mouse protein. For annotations made in this way, the evidence code is ISS, the with field contains the identifier of the sequence of the ortholog (from human, rat, whatever), and we have a notes field that carries the PMID of the paper with the human/rat /etc. experiment. The paper listed in the notes must describe an experiment using the orthologous protein listed in the with column; this can't be another sequence similarity comparison. In this way, MGI is able to make annotations for proteins for which they have determined one or more orthologs in other species, even when there is no publication about the mouse gene. Note though that in this system, the database must have a system and expertise to determine orthology.

The following questions were discussed in the context of Michelle Gwinn's talk about TIGR's pipeline for annotation of bacterial genomes:

- **QUESTION:** Where do the known proteins come from for an HMM and how is the trusted cutoff determined? **DISCUSSION:** We start with a combination of automated clustering methods and manual curation from the literature. For TIGR to include a protein in an HMM seed, it must be experimentally characterized or have really good bioinformatic evidence for function. To determine the trusted cutoff, we look at the distribution of the scores and check the literature. Trusted cutoff is different for each HMM. One HMM can have a trusted cutoff of 10 and another 100. You have to check if your protein is higher than the trusted cutoff. The trusted cutoff may even be negative, again simply check to see if your protein's score is greater than the trusted cutoff score. There are currently 2585 HMM families at TIGR; many of these are for prokaryotic functions/proteins. About 1300 of these are at the equivalog level, that is modeling one specific function.

- **QUESTION:** Where does the significance value of a hit come from: BLAST or Smith-Waterman? **DISCUSSION:** Both. We use liberal BLAST scores for collecting proteins to put into the Smith-Waterman portion of the BLAST_extend_repraze (BER) search, but then we are much more conservative in terms of assigning function. We store Blast P values for our pairwise matches as well as statistics from the Smith-Waterman alignment including percent id/similarity, gaps, and indels. At TIGR we tend to focus on percent identity and length of match as the most important indicators. Gaps or insertions in the alignment would reduce the matching regions as well as the overall quality of the match.

- **QUESTION:** How do you evaluate your sequence matches? **DISCUSSION:** We look at a combination of several different sequence based tools: BER (BLAST_extend_repraze; our tool for generating pairwise alignments), HMMs, Prosite, MHMM, SignalP, etc. Once we have all the data, TIGR is very conservative. We prefer to undercall, rather than overcall because we wish to avoid transitive annotation. We have evolved a multi-level system for naming proteins with evidence of various types. Our highest name level is the one where we believe that we know the precise function of a protein. In this case we will give the protein a precise name. In order for us to do this we need to see at least one of the following two pieces of evidence: a full-length match to an experimentally characterized protein at greater than ~40% identity (depending on the length of the two proteins, short proteins require higher identity) OR an above trusted cutoff match to an equivalog (modeling one specific function) level HMM. If these criteria are not met, the protein will get a name consistent with what evidence is available: it might be given a name based on its membership in a protein family ("xxxxx family protein"), it might be given a putative designation ("putative xxxxx protein") or it might be given the homolog designation ("xxxxx protein homolog", which indicates just that there is some level of sequence simlarity between the two proteins, but no claims on function). For example, if we had a poor pairwise match to an experimentally characterized biotin synthase and nothing else, we would likely call the protein "biotin synthase homolog" or if the match was very poor it would be demoted to "hypothetical protein". To avoid transitive annotation, we require that in order for a pairwise match to be considered as evidence for a functional annotation, the match protein must itself be experimentally characterized.

- **QUESTION:** Are the experimentally detemined proteins listed? **DISCUSSION:** We maintain a database containing accessions of proteins that have been experimentally characterized. Our annotation tools highlight these proteins in our search results to aid annotators. We are constantly adding new accessions to this database, however, it is not even remotely an exhaustive set. This is an area where we very much wish we could invest more resources.

- **QUESTION:** What do you call a homolog? **DISCUSSION:** TIGR uses the word homolog to denote that there is some sequence similarity but we are making no claims about the protein's function (because the similarity is so weak). We use the word ortholog very carefully only when there is evidence to support that a given gene is equivalent in function to a specific gene in a related organism. We have a hierarchy of terms we use based on how strong the sequence similarity is (see above). Ortholog [we believe we know the exact function of this protein based on a strong match to characterized examples from other organisms] is a very strong statement; homolog [weak but consistent similarity OR strong similarity to something not expected in this organism (photosynthetic gene in non-photosynthetic bug)] is much weaker. 25-30% of the genes in new genomes fall into the categories "hypothetical" or "conserved hypothetical" protein. It is common to find paralogous families of proteins within a species which have no homology to anything with a characterized function, but do match unknown proteins from other species (the conserved hypotheticals), and also to find paralogus families that do not match anything outside the species at all (hypothetical proteins).

- **QUESTION:** What proportion of proteins in a genome fall into frameshift and point mutations? and in these cases, what gets deposited to GenBank? **DISCUSSION:** Typically, we see around 10-100 proteins in this class in an average prokaryotic genome. For proteins where we can identify a biological translation exception (selenocysteins, programmed frameshifts) an appropriate translation is sent to GenBank, for the proteins in which we find single frameshifts or in-frame stops which have no experimental documentation for function in this organism or elsewhere, no translation is sent.

- **QUESTION:** Do you (TIGR) go back and add /update annotations? Not on a genome wide scale - we do propogate updates stemming from new HMMs to older genomes by searching the new HMMs against those genomes and then checking the annotation of genes that score well to the HMMs. But this only effects a few genes per genome - we would like to do large scale genome reannotation, but do not yet have the resouces to do so.

QUALIFIERS

NOT

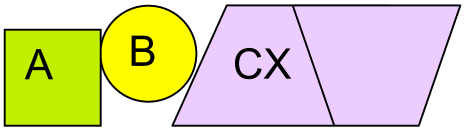
- **QUESTION:** How do you capture negative results? Can one use the NOT qualifier to indicate a negative result in an experiment? For example, in Paper 1 (VanWagoner TM et al.), they make a mutation that has no phenotype. **DISCUSSION:** Typically, GO does not attempt to capture this type of negative data. The NOT qualifier was instituted in order to be able to make statements that a particular GO term should NOT be associated with a given gene. In practice, this qualifier is used sparingly. It is particularly useful in cases where a cursory inspection might result in making a positive annotation with that term. For example, SGD has made an annotation to say that Rcl1p is NOT an RNA cyclase (Function ontology). In this paper the authors do a sequence comparison (Figure 1)and state that superficially Rcl1p is similar to other known RNA cyclases. But a closer look at the sequence conservation in the catalytic site residues revealed that it lacks key residues. A direct assay testing this protein for RNA cyclase activity was negative. [Billy E et al., EMBO J. 2000, 19(9):2115-26. PMID: 10790377]. Thus, we made a "NOT RNA cyclase activity" annotation for RCL1 because a superficial look at sequence similarity might lead someone to assume that it is an RNA cyclase. The use of the NOT qualifier allows us to make an annotation that points people to the negative evidence against this assumption.

Contributes to

- **QUESTION:** In your presentation (about the Chang et al. paper; presented by Karen Christie), you made an annotation to a complex term. Would you give RMI1 an annotation to a function term with the qualifier "contributes to"? **DISCUSSION:** That's a good question. Often, you would want to make a "contributes to 'function term'" annotation. However, I didn't feel it was warranted from this paper. They do show that Rmi1p can be isolated with the Sgs1p/Top3p complex, but they do not show that Rmi1p is present as part of a complex that has enzymatic activities associated with Sgs1p or Top3p. If they had done an assay with the whole complex and shown activity, a "contributes to: function" would definitely have been appropriate. However, from this paper, I felt that "molecular function unknown" was the appropriate function annotation for RMI1.

Colocalizes with

- **QUESTION:** In paper 18 (Masaki M et al.), protein A binds directly to protein B that binds to complex CX, as per the diagram below. Is this sufficient evidence for a component annotation? Should a term be created for the protein B-CX complex? **DISCUSSION:** Clearly, it is fine to annotate protein B to "colocalizes with 'CX complex'". It was felt that in this situation, where they are identifying protein A, protein B, and complex CX all together, that it was also reasonable to annotate protein A to "colocalizes with 'CX complex'" and also to "protein binding (with protein B)". However, GO would probably not invent a name and corresponding term for "protein B – CX complex" unless the authors describe it in such terms.



OTHER GO ANNOTATION QUESTIONS

Granularity or level of annotation

- **QUESTION:** How does one deal with downstream effects of a given mutation? For example in Paper 6 (Glise B et al.), they deal with a mutation that has downstream effects on wing disc development. Should this downstream effect get captured? **DISCUSSION:** This depends on the situation. There is no hard and fast rule. Rather the appropriate course of action depends both on the philosophy of the group and the specific situation. The fact that mutant phenotypes may reflect a downstream effect rather than the primary defect is something that the GO consortium recognized. For example, in S. cerevisiae, splicing mutations often display translation defects. We now know that this is because the majority of the few genes that contain introns in cerevisiae are ribosomal protein genes. With this knowledge, we now feel that it is not appropriate to annotate a splicing gene to the process term "translation ; GO:0043037". However, if this is the first paper characterizing a gene and its phenotype when mutated, then you will probably want to make an annotation so that you can provide some hint of what process the gene is involved in. Once more is known, you may decide that the annotation needs to be reviewed in light of current information. Another situation is illustrated with this example. Let's say you have two transcription factors. One is a general transcription factor involved in the basic process of transcription. Because it is so basic, it is involved in many processes. Clearly it should get a process annotation for "transcription from Pol II promoter", but one might not wish to annotate all the downstream effects. The second transcription factor is specifically involved in activating one specific pathway, e.g. histidine biosynthesis. In this case, in addition to annotating to "transcription from Pol II promoter", it would be very useful to also annotate to "histidine metabolism", since it is specifically involved in this process.

- **QUESTION:** In your presentation (about the Chang et al paper; presented by Karen Christie), you had a list of several possible specific terms, but you ended up selecting only "response to DNA damage" as an annotation. Why? **DISCUSSION:** I listed each of the specific phenotypes the authors mentioned in the paper as a short summary to help me when I looked for a term. Some of these phenotypes are very similar to GO terms. However, I felt that the most basic process that was consistent with and represented all of the various phenotypes, some of which may be downstream effects, was "response to DNA damage".

Specific Annotation Questions

- **QUESTION:** Relating to protein binding, how do you deal with annotation of homo-oligomers? **DISCUSSION:** There are already terms in GO that are appropriate for annotation of homo-oligomers. In the process ontology, there is "protein homooligomerization ; GO:0051260" and in the function ontology there is "protein homodimerization activity ; GO:0042803". Annotations to these terms would generally use the IPI evidence code and contain the same protein ID in the with column (#8) as is found in column 2, which contains the ID of the object being annotated. For example, in UniProt, the human protein RAD51_HUMAN (accession number Q06609) has this annotation:

  protein homooligomerization ; GO:0051260 ; IPI with Q06609

- **QUESTION:** In your presentation (about the Chang et al paper; presented by Karen Christie), you annotated RMI1 to "molecular function unknown", but not to "protein binding" even though they showed that Rmi1p binds to Sgs1p and Top3p. Why did you not annotate with the "protein binding "term? **DISCUSSION:** While a number of groups, such as MGI, choose to use the with column in conjunction with the term "protein binding" to store protein-protein interaction data, SGD does not. We have another system for storing this type of data and do not double curate it into GO as well. Use of the term "protein binding" to effectively become a method for creating a database of protein interactions is allowed by the GO Consortium, but is not required.

- **QUESTION:** Would you use experiments done in human cell lines for annotation, even if these lines are aneuploid or otherwise known to be abnormal? **DISCUSSION:** The general thought was that it is often fine and is really up to the discretion of curators annotating human genes. For human, an experiment in a cell line will often be the only experimental data available. In addition, while non-primary cell lines are often known to have abnormalities, these changes from the normal state are often not relevant to the given experiment. In this situation, it is reasonable to draw conclusions about the normal process of a gene and thus make a GO annotation.

Annotation methods and consistency

- **QUESTION:** Is there any checking about consistency of annotation, either between different groups, or within a single group? **DISCUSSION:** One of the reasons for this, and the previous, GO Annotation Camp, is to address the issues of training in annotation and consistency checking and to discuss specific annotation questions so that the GO Consortium as a whole comes to consensus on the appropriate course of annotation.

- **QUESTION:** Is there any training for new users? **DISCUSSION:** We are in the process of building a set of papers to serve as a training set. Each group is to provide 10 papers and a "key" to what annotations may be made from each paper. These can then be used both for training new people and for exercises to check for consistency between curators.

- **QUESTION:** How do the three ontologies relate to each other? How can a curator be sure that they have made all the appropriate annotations? **DISCUSSION:** QuickGO suggests possible terms that a curator may wish to consider, based on analysis of "concurrent annotations", i.e. terms that tend to be annotated to the same genes. Proteome used GO Pairs. We are assessing these types of methods to see if they can be incorporated into curation tools.

- **QUESTION:** How do you search for a GO term? How do you know to search for protein kinase for example? **DISCUSSION:** Sometimes by familiarity. As with any system, the more you use it, the better at it you get. Some people look at the annotation of a similar gene in another DB, and use its annotations as a guide. GO also has "synonyms", in addition to the primary term names, to help make it easier for someone unfamiliar with the GO system to find the right term. Though we have been calling them synonyms, they do not necessarily mean exactly the same thing. We have four types, referred to as "broad_synonym", "exact_synonym", "narrow_synonym", or "related_synonym". While it is not always possible to use a common usage phrase as the main name of the GO term, we try to add common usage phrases as synonyms to help with searching.

OTHER GO QUESTIONS

Obsolete terms

- **QUESTION:** What information is still associated with an obsolete term? **DISCUSSION:** All information that was part of a GO term (the GO id, term name, definition, aspect, any dbxrefs) remains part of that GO term when it becomes obsolete. When a term becomes obsolete, the word "OBSOLETE" is becomes the first word of the definition. If a term was defined at the time it became obsolete, the original definition is retained after the word "OBSOLETE". If it was not defined, the phrase "was not defined before being made obsolete" is used. In addition, the Comment field is used to suggest an appropriate replacement term or terms. It is worth noting that it is possible to create a new term with a term name that is identical to the term name of a term that has been obsoleted. This is because the GO id is attached to the definition of a term, not to the term name per se. Thus if it becomes clear that a GO term has been defined incorrectly, in a way that would have allowed inappropriate annotations to be made, our practice is to obsolete the original GO term and replace it with a new GO term that has a different GOid. In this way, users who have used the original GO term to annotate are alerted to the change and can use the comment on the obsoleted term to help identify the appropriate replacement term.

- **QUESTION:** How do databases handle obsolete terms? What happens to the obsolete terms? How are they stored? **DISCUSSION:** Terms that have been obsoleted are never deleted from the GO file. In the OBO format of the GO file, obsolete terms are given the attribute "is_obsolete: true" and the word "OBSOLETE" is added to the definition as the first word. In the older GO file format, terms made obsolete become children of the appropriate obsolete node. However, how individual databases handle obsolete terms depends on the database. SGD, for example, does not keep obsolete terms in our database. Our nightly script to load and update GO checks if a term that has become obsolete is used for any annotations in our database or is included in any GO-slims. If it is, curators examine the situation and fix it. However, once the obsolete term is not used for any annotations in the database, it will be deleted from SGD.

References for GO annotations

- **QUESTION:** Sometimes a reference has a Supplementary Material section that has been used for GO Annotations. The problem is that sometimes the Supplementary Data becomes unavailable if the journal or author doesn't maintain older data. Is it still OK to use Supplementary Data for GO annotations? **DISCUSSION:** SGD deals with this problem by getting a copy of the Supplementary Data file(s) at the time we make GO annotations from it and then we make the Supplementary Data file available through SGD's ftp site, as well as via the original site on the journal's website. This way, we make sure that the original data remains available to the public. Helping ensure that supplementary datasets, especially large scale ones, remain publicly available is something that our Advisory Board has said is important, so we are starting to make more of an effort to host the supplementary data files for published papers. However, every database group will need to make its own decision about how to handle this type of question depending on its own resources and priorities.

- **QUESTION:** For some genes, the only papers we have are older papers where only a phenotype, without knowing a specific gene, or perhaps a biochemical assay is shown. Do you use such papers for GO annotations? **DISCUSSION:** The answer to this question depends in part on curator discretion and group practice. For old papers describing biochemical assays of a protein, one of the complications can be if the protein may have multiple isozymes. If you can unambiguously assign the paper to a specific isozyme, or if the proposed function is general enough to apply to all isozymes, it may be reasonable to use the paper for GO annotations.

- **QUESTION:** Is it permissible to use a meeting abstract for a GO annotation? or a thesis? or a publication in a foreign language? **DISCUSSION:** Each group has its own policies on references that are not (and will not be) in PubMed. WormBase allows use of Meeting Abstracts, and maintains these abstracts in our database with WBPaper identifiers. In contrast, SGD does not allow these; the curator must be able to read the full text of a peer reviewed paper in order to be allowed to make a GO annotation from it.

Making changes to the Gene Ontologies

- **QUESTION:** If you don't have write access to the CVS repository, how do you request a new term? **DISCUSSION:** We have a SourceForge tracker for suggestions to improve the Gene Ontologies. All suggested changes, whether for new terms or changes to existing terms, structure, definition, etc., should be entered into the GO request tracker, regardless of whether the requester has CVS write access or not. The SourceForge tracker is our official tracking mechanism for all requests and their outcome.

- **QUESTION:** How many terms should be created, i.e. do we need a term that is specific for Drosophila? **DISCUSSION:** GO may have as many terms as is needed. The guiding principle for whether a species specific term is need or not is whether the process (or function or component) is unique to a given species or taxon. If the process occurs similarly across all known organisms, only a single term is needed. However, if a specific organism or taxon, exemplified by the model organism *Drosophila melanogaster* in this example does something differently than other organisms, then a specific term is needed. These terms representing taxon-specific processes are labeled with the word "sensu" and the relevant taxon, e.g. "sensu Drosophila". However, it is important to note that using the sensu designation in a term does not exclude that term from being used to annotate species outside that designation. For example, a 'sensu Drosophila' term might reasonably used to annotate a mosquito gene product.

Synonyms

- **QUESTION:** Do you have different types of synonyms? How do you display them? **DISCUSSION:** GO has four types of synonyms: related, exact, broad and narrow. The OBO format file uses the tags: "broad_synonym", "exact_synonym", "narrow_synonym", "related_synonym". Unfortunately, the synonym types are not currently displayed in AmiGO.

  NOTE: Since the Annotation Camp, a new version of AmiGO has been moved into production. The individual term pages, like the one for which the URL is given below, now indicate the type for all synonyms.

  http://www.godatabase.org/cgi-bin/amigo/go.cgi?view=details&depth=1&query=307

- **QUESTION:** Does AmiGO search for synonyms? **DISCUSSION:** Yes. AmiGO searches for a match within the synonyms as well as within the term names. Unfortunately, the results page does not indicate what the match was when it occurs within the synonym.

  NOTE: Since the Annotation Camp, a new version of AmiGO has been moved into production. If you search for something like "cyclin", you can now see whether the match was within the term name or a synonym and the specific text that matches is highlighted. Note that only synonyms that match the query are shown; other synonyms that do not match the query are not shown on the search results.

File Format, Error checking, and filtering of annotations

- **QUESTION:** What is the difference between the OBO file format and the GO file format? **DISCUSSION:** The GO file format is the original format of the Gene Ontology files. There is a separate file for each of the three ontologies, biological process, molecular function, and cellular component. Term relationships are indicated by the amount of indentation from the terms above it and the relationship type is indicated with a symbol in front of the term name. There is an additional fourth file for the term definitions. The OBO format file is somewhat like XML in format and it contains the information about each of the three ontologies, including the term definitions, in a single file. In addition, the OBO file contains the specific synonym type ("broad_synonym", "exact_synonym", "narrow_synonym", "related_synonym"). The GO format includes the synonyms, but not the synonym types.

- **QUESTION:** Currently the error checking script for the gene_association files just lists the line number. Could it print the whole line as well, so that it is easier to find the error? **DISCUSSION:** Mike Cherry "Yes, we can do that. The script is still evolving. The way to make it do what we all want is to provide feedback."

- **QUESTION:** In Mike Cherry's list of errors, why is it an error to have IEAs more than a year old? **DISCUSSION:** In order to prevent having a lot of out of date information that was electronically generated, the GO Consortium has decided to implement a limit on the currency of IEA annotations. Those more than a year old will be excluded. For groups with ongoing annotation programs that perform IEAs via scripts, this will mean that that they need to rerun their script. All annotations that are still made by the script will get the new date and can be kept in the gene_association file.

- **QUESTION:** If the GOA project has GO annotations for a given species, that the relevant MOD did not capture, can the checking script alert the MOD? **DISCUSSION:** To prevent redundant annotations being present in AmiGO, the GO Consortium will be instituting a new policy. Currently, AmiGO shows two copies of the same annotations in some circumstances. For example, MGI makes annotations that are subsumed into the GOA project. Although the gene_association files credit the appropriate source, both get loaded into the GO database and then displayed on AmiGO. It would be OK for there to be multiple annotations, if they had independent sources, but this is effectively the same annotation just duplicated. Thus, in the future, for each species (or taxon), there will be one official source. If GOA is also making annotations for that species, then the GOA project and the MOD have to communicate so that GOA can contribute its annotations to the MOD and all annotations relevant to one species will be in a single file, contributed by the official source.

- **QUESTION:** On the ftp site, is the time given in the "**Last Change**" column for each file accurate? **DISCUSSION:** Yes, but note that the ftp server uses times in Greenwich Mean Time (GMT) rather than Pacific Time, so an update that occurs when it is 11pm at Stanford, will get a time of 7am the next day.

Documentation

- **QUESTION:** Do the README files provide information about a given groups annotation philosophy, e.g. whether they use "protein binding" annotations to capture protein-protein interaction data, exactly what procedure they use to assign an unknown term with the evidence code ND? **DISCUSSION:** We have just recently decided that this information must be present in each group's README file. It is probably not all there yet, but we will be implementing expanded README files with this type of information.

Other Ontologies

- **QUESTION:** Is there a way to find out what other ontologies are being developed, even those that are not part of OBO? **DISCUSSION:** OBO is a main resource for most of the ontologies. Beyond that, Michael Ashburner and Suzanna Lewis keep an eye on a variety of other ontology projects.

- **QUESTION:** How does GO deal with expression data, for example if a gene is expressed in the leaf versus the root? **DISCUSSION:** GO does not deal with expression data at all. GO deals with the cellular level, not the tissue level. For information outside of the scope of GO, there are other ontologies. For plants, the Plant Ontology vocabularies (www.plantontology.org) for plant structure and plant growth and developmental stages can deal with information about the site of expression.

- **QUESTION:** Is there an ontology for microarray experiments? **DISCUSSION:** The Microarray Gene Expression Data Society (MGED Society) develops standards and ontologies for microarray experiments.

GO Tools

- **QUESTION:** Can I use tools like Textpresso to make GO annotations? **DISCUSSION:** Textpresso, a tool to index and search the full text of papers, is not designed make GO annotations. However, it may be useful to help find the appropriate papers from which a curator may make GO annotations. Textpresso indexes the full text of a paper, so it can be particularly useful in the identification of papers that are relevant to a given organism when the abstracts of those papers do not mention the organism and can be useful in identifying topics, based on the keywords it uses to search. However, curation itself is performed by a human curator who reads the paper and judges what annotations can be drawn from the paper.

- **QUESTION** : How do I trim down the full Gene Ontology to create a GO-slim? **DISCUSSION**: You can customize a GO-slim in whatever way you like, whether that's a subset of high level terms or any other subset of terms. To create your own GO slim, you need to use DAG-Edit (http://www.godatabase.org/dev/java/dagedit/docs/index.html) - there's tutorial on how to do this at http://www.geneontology.org/GO.teaching.resources.shtml#tut, called 'Creating a GO slim using DAG-Edit' (an online version will be available soon).

- **QUESTION** : How do I map a set of associations to a GO-slim? **DISCUSSION**: To map an association file to your own, or an existing, GO slim, you can use the web-based tool Generic GO Term Mapper (http://go.princeton.edu/cgi-bin/GOTermMapper) for any species, or GO Slim Mapper (http://db.yeastgenome.org/cgi-bin/GO/goTermMapper) for Saccharomyces annotations only. Alternatively, if you know some Perl, you can use the map2slim script (http://www.geneontology.org/GO.slims.shtml?all#script)

- **QUESTION:** In one of the small groups, someone asked if it would be possible for DAG-Edit to include the Evidence codes so that annotators could look at both the term and the scope of each evidence code in one tool, rather than having to go to the GO documentation each time for the evidence codes? **DISCUSSION:** This is a question for the designer of DAG-Edit.

- **QUESTION:** Have you thought of creating a forum on the website so that people can submit info on new ontologies, etc.? **DISCUSSION:** There was discussion about whether a web forum, or wiki, would improve the accessibility of GO to people looking for information on it or other biological ontologies. The main concern seemed to be whether a web forum would be any more effective at attracting users than the existing email list, for which the email archive is available for searching. We are always trying to figure out the best way to get the word out about the GO project, how to get regular researchers interested, not just the people who already know about GO.

- **QUESTION:** At the last GO Consortium meeting, there was discussion about a Common Annotation Tool. Has any progress been made on such a thing? **DISCUSSION:** While there is not a single annotation tool in use by all Consortium members, there are a couple tools that are freely available. The Arabidopsis Information Resource (TAIR) and Rat Genome Database (RGD) collaborate to create PubSearch (at pubsearch.org), a curation tool designed to help manage the literature. A paper about PubSearch will be coming out soon in *Current Protocols in Bioinformatics*. TIGR creates the Manatee tool. Two versions are available, eukaryotic and prokaryotic. Both PubSearch and Manatee are freely available.

Other

- **QUESTION:** Why is "ATPase" activity NOT a child of "ATP binding"? Why do we need a separate annotation to "ATP binding"? **DISCUSSION:** The term "ATPase activity" did used to be a child under "ATP binding" . There was much discussion at the time about moving it out, but no one present remembered the exact details well enough to clarify on the exact reasoning that this change occurred, though likely due to some true path violation to have "ATPase activity" always under "ATP binding". Peter D'Estachio commented that some ATPases are very rapid at catalysis and subsequent release, such that they do not exhibit stable binding to ATP, though without researching into the reasons for the change, we don't remember if this was a consideration in the decision.

- **QUESTION:** Does anyone spend time of the research aspects of GO, i.e. the AI/CS aspects of it? **DISCUSSION:** The majority of people developing GO are biologists working as curators trying to annotate genes. Their focus is typically on the biological perspective, i.e. trying to create a term or terms that represent the process or function they need to annotate. However. some members of Suzanna Lewis's group at the Berkeley Drosophila Genome Project (BDGP) are interested in the data structure and data models.

- **QUESTION:** Can PubMed enforce the requirement of GO annotations, similarly to how submission to GenBank requires use of Gene IDs? **DISCUSSION:** Use of standard gene names and IDs is a much bigger problem, so the Model Organism Database (MOD) community would probably focus on requiring standard gene names and /or gene IDs first, before requesting use of GOids. On the Biocurator mailing list, there is currently a discussion about what would we (the MODs) would want the journals to require. It was also mentioned that if there were undergraduate courses on biocuration, it might help people understand what information needs to appear in their publications for them to get properly indexed.

QUESTIONS RELATING TO DATABASE ORGANIZATION, RATHER THAN GO

- **QUESTION:** How does information about the strain background get captured? **DISCUSSION:** It depends upon the organization. Some groups do not capture this information at all. Other groups do. Whether and how to do this depends upon the needs and interest of your community.

- **QUESTION:** Can GO be used to annotate a(n) 1) microRNA, 2) mature processed protein 3) alternative transcript? **DISCUSSION:** GO can be used to annotate any gene product, regardless of whether the gene product is composed of protein or RNA. In practice , many groups annotate at the level of gene, rather than gene product (protein or RNA product). This is because most database groups generate database objects with unique IDs for genes, but do not generate separate database objects with their own unique IDs for the corresponding products of these genes. Every object to be annotated needs to exist as a DataBase object with its own unique identifier within your database. Any database object representing a gene that has a gene product or any database object representing a gene product directly can be annotated with GO. However, it is not appropriate to use GO to annotate a sequence feature that does not have a gene product, e.g. an origin of replication or a protein binding site. Annotation of such sequence-based features should be performed with the Sequence Ontology (SO), not the Gene Ontology (GO).

- **QUESTION:** Is it OK to do a set of first pass annotations with weak evidence, e.g. InterPro Scan? **DISCUSSION:** This is up to each group to decide based on the needs of the community each serves. The community needs should guide your annotation practice. Saying that gene A has similarity to XXX domain may motivate somebody to look at that gene, so this annotation may be useful to the community.

- **QUESTION:** How do you decide the priority between updating existing annotations and annotating another genome? **DISCUSSION:** Each group needs to get feedback from its own use community and assign priorities accordingly.

- **QUESTION:** Do databases keep what they have in their gene_association file in sync with what they show on their production website? **DISCUSSION:** Each database has its own procedures for updating its production website and for creating its gene_association file. For example, Zfin is not in sync: it submits its gene_association files only weekly, but database updates show up immediately. At SGD, database updates also show up immediately, but gene_association files are submitted daily. At MGI, both the database and the gene_association file are updated weekly. WormBase gene_association files had not been in synchrony with WormBase releases, but user-feedback has prompted WormBase to maintain synchrony between its gene_association file and WormBase releases.

- **QUESTION:** Is Community Annotation (annotations made by members of the research community) an effective way to get good annotations into the database? How do you do Quality Control on such annotations? **DISCUSSION:** This seems to be highly dependent on the specific community and perhaps also on specifically what sort of information is being asked for. The Phytophthera database (at Virginia Bioinformatics Institute) is using a community annotation system with good success. They have about 1600 genes. Users login and make annotations. Annotations from the "experts" are not checked. However, annotations from other sources go to a temp database to be checked before entry into the production database. So far, 25 people have provided 1600 annotations, mostly during the Joint Genomes Initiative (JGI) annotation jamboree for Phytophthera. Nicole Perna used a Community Annotation system for *E. coli* and it worked well. However, in contrast, SGD was asked to set up a system to allow users to contribute statements about key highlights in published papers. In over two years, only a handful of people have contributed less that a hundred annotations about research publications.

- **QUESTION:** How do you (asked of the Phytophthera community) deal with conflicting annotations? **DISCUSSION:** So far the Phytophthera project has not had to deal with conflicting annotations. However, it is important to note that conflicting annotations are allowed by GO. If there is dissention in the field, it is permissible to make annotations reflecting both points of view, each with an appropriate reference.

ACTION ITEM 6: Can the people representing new groups here, please let us know the status of your projects, and if there is any assistance from the GO Consortium that would be useful to you. Please send this info to Jennifer Clark (jenclark@ebi.ac.uk)

READING LIST

PAPERS PRESENTED TO THE FULL GROUP

A. Loyola A et al., Functional analysis of the subunits of the chromatin assembly factor RSF. Mol Cell Biol. 2003 Oct;23(19):6759-68. [PMID:12972596]
(full paper presented by Rebecca Foulger of UniProt)

B. Chang M et al., RMI1/NCE4, a suppressor of genome instability, encodes a member of the RecQ helicase/Topo III complex. EMBO J. 2005 Jun 1;24(11):2024-33. [PMID:15889139]
(full paper presented by Karen Christie of SGD)

C. Appelbaum L et al., Homeobox-clock protein interaction in zebrafish. A shared mechanism for pineal-specific and circadian gene expression. J Biol Chem. 2005 Mar 25;280(12):11544-51. [PMID:15657039]
(single example from the paper presented by Ceri Van Slyke of Zfin)

PAPERS DISCUSSED IN SMALL WORKING GROUPS

1. VanWagoner TM et al., Characterization of three new competence-regulated operons in Haemophilus influenzae. J Bacteriol. 2004 Oct;186(19):6409-21. [PMID:15375121]

2. Sakuragi N et al., Functional analysis of a novel gene, DD3-3, from Dictyostelium discoideum. Biochem Biophys Res Commun. 2005 Jun 17;331(4):1201-6. [PMID:15883003]

3. Iizuka K et al., Genetically linked C-type lectin-related ligands for the NKRP1 family of natural killer cell receptors. Nat Immunol. 2003 Aug;4(8):801-7. [PMID:12858173]

4. Huang S et al., Arabidopsis VILLIN1 Generates Actin Filament Cables That Are Resistant to Depolymerization. Plant Cell. 2005 Feb;17(2):486-501. [PMID:15659626]

5. Kawashima CG et al., Characterization and expression analysis of a serine acetyltransferase gene family involved in a key step of the sulfur assimilation pathway in Arabidopsis. Plant Physiol. 2005 Jan;137(1):220-30. [PMID:15579666]

6. Glise B et al., Shifted, the Drosophila ortholog of Wnt inhibitory factor-1, controls the distribution and movement of Hedgehog. Dev Cell. 2005 Feb;8(2):255-66. [PMID:15691766]

7. Abramovitch RB et al., Pseudomonas type III effector AvrPtoB induces plant disease susceptibility by inhibition of host programmed cell death. EMBO J. 2003 Jan 2;22(1):60-9. [PMID:12505984]

7a. Kumar S et al., The role of reactive oxygen species on Plasmodium melanotic encapsulation in Anopheles gambiae. Proc Natl Acad Sci U S A. 2003 Nov 25;100(24):14139-44. [PMID:14623973]

8a. Qutob D et al., Expression of a Phytophthora sojae necrosis-inducing protein occurs during transition from biotrophy to necrotrophy. Plant J. 2002 Nov;32(3):361-73. [PMID:12410814]

8b. Tyagi S et al., The ORF3 protein of hepatitis E virus interacts with liver-specific alpha1-microglobulin and its precursor alpha1-microglobulin/bikunin precursor (AMBP) and expedites their export from the hepatocyte. J Biol Chem. 2004 Jul 9;279(28):29308-19. [PMID:15037615]

9a. Imhof I et al., Phosphatidylethanolamine is the donor of the phosphorylethanolamine linked to the alpha1,4-linked mannose of yeast GPI structures. Glycobiology. 2000 Dec;10(12):1271-5. [PMID:11159918]

9b. Ito Y et al., Organ-specific alternative transcripts of KNOX family class 2 homeobox genes of rice. Gene. 2002 Apr 17;288(1-2):41-7. [PMID:12034492]

10a. Kucharski R and Maleszka R. Transcriptional profiling reveals multifunctional roles for transferrin in the honeybee, Apis mellifera. J Insect Sci. 2003;3:27. [PMID:15841243]

10b. Chen ZJ et al., Molecular cloning of a regulatory protein for membrane-bound guanylate cyclase GC-A. Biochem Biophys Res Commun. 2000 Nov 11;278(1):106-11. [PMID:11071862]

11a. Deng M et al., Mapping Gene Ontology to proteins based on protein-protein interaction data. Bioinformatics. 2004 Apr 12;20(6):895-902. [PMID:14751964]

11b. Palsson A and Gibson G. Association between nucleotide variation in Egfr and wing shape in Drosophila melanogaster. Genetics. 2004 Jul;167(3):1187-98. [PMID:15280234]

12a. Birbes H et al., A mitochondrial pool of sphingomyelin is involved in TNFalpha-induced Bax translocation to mitochondria. Biochem J. 2005 Mar 15;386(Pt 3):445-51. [PMID:15516208]

12b. Thorpe CJ et al., Wnt/beta-catenin regulation of the Sp1-related transcription factor sp5l promotes tail development in zebrafish. Development. 2005 Apr;132(8):1763-72. [PMID:15772132]

13. Keeling PJ and Palmer JD. Lateral transfer at the gene and subgenic levels in the evolution of eukaryotic enolase. Proc Natl Acad Sci U S A. 2001 Sep 11;98(19):10745-50. [PMID:11526220]

14. Henderson MJ et al., EDD, the human hyperplastic discs protein, has a role in progesterone receptor coactivation and potential involvement in DNA damage response. J Biol Chem. 2002 Jul 19;277(29):26468-78. [PMID:12011095]

15. Barloy-Hubler F et al., Smc01944, a secreted peroxidase induced by oxidative stresses in Sinorhizobium meliloti 1021. Microbiology. 2004 Mar;150(Pt 3):657-64. [PMID:14993315]

16. Koushika SP et al., A post-docking role for active zone protein Rim. Nat Neurosci. 2001 Oct;4(10):997-1005. [PMID:11559854]

17. Prigge MJ et al., Class III Homeodomain-Leucine Zipper Gene Family Members Have Overlapping, Antagonistic, and Distinct Roles in Arabidopsis Development. Plant Cell. 2005 Jan;17(1):61-76. [PMID:15598805]

18. Masaki M et al., Mixed lineage kinase LZK and antioxidant protein-1 activate NF-kappaB synergistically. Eur J Biochem. 2003 Jan;270(1):76-83. [PMID:12492477]