Minutes of the GO Annotation Workshop Part 1
- Annotation Standards
Stanford University
July 10-11, 2006

Attendees:

dictyBase: Rex Chisholm, Pascale Gaudet, Karen Pilcher
E. coli Hub: Jim Hu, Barry Wanner, Sarah Ess, Yang Yu
FlyBase: Susan Tweedie
GOA: Emily Dimmer
GO Editorial Office: Midori Harris
Gramene: Pankaj Jaiswal, Dean Ravenscroft
HGNC: Ruth Lovering
LBNL: Seth Carbon
MGI: Judy Blake, David Hill
Mississippi State University: Fiona McCarthy
PAMGO: Trudy Torto-Alalibo
PombeBase: Val Wood
RGD: Victoria Petri, Jennifer Smith
SGD: Mike Cherry, Ben Hitz, Eurie Hong, Karen Christie, Rama Balakrishnan, Julie Park, Stacia Engel
TAIR: Tanya Berardini, Chris Tissier
UNC Chapel Hill: John MacMullen
WormBase: Kimberley Van Auken, Ranjana Kishore
ZFIN: Doug Howe, Leyla Bayraktaroglu

Minutes by Tanya Berardini, Pascale Gaudet, Susan Tweedie, Victoria Petri, and Karen Christie

Conclusions and Recommendations are in the first portion of the minutes. Transcripts of Discussions are included separately at the end.

TABLE OF CONTENTS

Items in red contain proposals from the annotation group that the GO Consortium as a whole should discuss and come to consensus upon.

Reference Genomes (Rex Chisholm)

Overview of purpose of group
Document with details sent out prior to meeting, can refer to that later, also see the wiki for up-to-date info: http://wiki.geneontology.org/index.php/Reference_Genome_Focus

BACKGROUND

More and more genomes are being sequenced, but few of these are going to have well funded databases or curators.

GOALS:

- GO provides the set of reference genome (RG) annotations: 9 genomes, 9 organization of which E.coli Hub is the newest, selected for various reasons
- Aim for 'fully curated' genomes, broad and deep annotation

What does broad and deep annotation mean?

- **Broad**: Every gene has a functional annotation* in each of the 3 ontologies.
- However, even SGD with ~6K genes has ~25% with no experimental data for making a MF annotation – focusing specifically on experiments done in yeast and NOT comparative annotations
- Deep: ??? maybe something related to # papers selected for GO annotation versus # papers linked to the gene

*Functional annotation = any GO annotation (MF/BP/CC)

A. Metrics: breadth

GOAL: Want to track numbers of genes with experimental data and present that as a table.

Conclusions – Metrics to monitor breadth:

1. How many genes does your genome have?
   Best estimate of protein or functional RNA genes. Pseudogenes do NOT count.
2. How many genes have any GO annotation? Genes with ND annotations will be counted for this metric since the presence of unknown annotations indicates that they have been looked at and annotated by a curator.
   Break down by MF/BP/CC. Give totals for any evidence code.
3. How many genes have GO annotations based on experiments in your organism?
4. How many genes **only** have GO annotations inferred from some sort of sequence analysis?

The GOC decides how to use evidence codes and the MODs decides how to apply them. The MODs may choose not to use certain codes but they agree to follow the set rules for the ones that are put to use. The GOC does not dictate how to capture functional information beyond providing baseline rules. Standard example: use of 'protein binding' to capture protein-protein interactions

Core metrics (#3): Only use experimental evidence codes (IDA, IEP, IMP, IGI, IPI). (Converse: Do not use ISS, IEA, TAS, NAS, RCA, IGC, ND, IC.)

Rationale for not using IC for core metrics: That same gene should have an experimental based annotation associated to it that was used the annotation made by the IC code, hence that gene will already be counted in the experimental annotation set. Adding IC to the core metric set would not increase gene numbers.

Don't double count genes. The numbers reported for 3 and 4 above should be mutually exclusive sets. If a gene has an IDA annotation and an IEA annotation, it goes into the experimental bin.

Trying to track increase in # genes with experimentally derived GO annotation.

Total number of genes with GO annotations =

- Genes with any experimental evidence code annotation (#3 above)
- Genes with no experimental evidence code annotation but with an ISS annotation (#4 above)
- Genes with no experimental evidence code annotation and no ISS annotation (some combination of TAS, NAS, ND, IEA, RCA, IC, IGC annotations)

See additional discussion on page 17

B. Metrics: depth

Concept: Use all information from all papers about a single gene to annotate.
Reality: A subset of the total papers associated with a gene has GO information, other papers have redundant or back-up information (from a GO perspective).

Some MODs have ways of associating genes to papers (SGD, TAIR, MGI…), some don't (*S. pombe*, *E. coli*, Gramene). (Broader question: What constitutes a valid gene-paper association?)

Conclusions? - Metrics - How to measure depth?:

1. Number of papers associated to each gene
   (MODs need to know this number, may not have to transmit this info to GO.)
2. % of papers curated per gene
3. Number of papers used for each gene's GO annotations
4. Number of genes completely annotated (based on curator tag and date)

Not sure if we came to firm conclusions on exactly what is needed here; Rex and the reference genomes group may work on this further.

See additional discussion on page 18

C. Primary focus of annotation: genes involved in human disease

Annotation of genes involved in human disease, and their orthologs in other species, is one of the GOC's priorities as we have received NIH funding specifically for this task.

How?

1. Generate a list of human genes involved in human diseases
   - OMIM, other sources
2. Translation of this list to orthologous genes in MODs
   - Starting point: InParanoid, Homologene, TreeFam
   - Get intersection of genes from three approaches and use these
   - This is not necessarily a complete set but is a good starting point (a starting set of a few hundred genes would be good.)
   - See how big the resulting gene sets are and go from there.
   - Single ortholog per human gene, 1:1, best hit only.
   - Try to group genes because annotating one gene usually leads to annotating several other related ones. (Related either by sequence or by process.)
3. Prioritization of gene list
   - Have an existing set of 180 genes overlap between human, ZFIN, Drosophila
   - Are human disease genes, have literature, have homologs in human, fly, fish.
   - 32 have mouse mutants.
   - Good starting point.
4. Annotation by MODs.
   - Annotation may be done by one or more curators, up to MODs to decide how to distribute the work.

Q: How often will the gene lists be updated?
A: Unknown.

Q: What will the process be like?
A: Each week, the reference genomes will get a list of n (initially 5) human genes and their respective orthologs for annotation.

Benefits of the group effort:

- Any annotation issues that arise can be tackled by the whole group because everyone is focused on the same group of genes.
- Drives ontology development, if new terms are needed, they can be added right away.
- Robust discussions can arise.
- Annotation consistency.

Other points:

- Rough start date: mid-August 2006 for first set of genes, Rex will send these out

- Suggestion: Let this process go, don't mess with it too much, settle in and discuss results at the next GO meeting (Jan. 2007)

- Updates: via periodic phone conference and RG mailing list

- Need a conduit for rapid ontology development, tie in via David/Midori. Some terms may be added right away, some may need to be deferred for some time.

Final points for Reference Genomes Discussion:

- Submit numbers of genes to RG mailing list in next two weeks if you haven't already. Could include this information in the header of the gene association file.

- Other numbers will be generated centrally by script.

- Individual MODs need to have a way to monitor these metrics in house.

GO Consistency Study (John MacMullen)

Note: Use of the term 'consistency' in this context does not refer solely to evidence codes but the formalized annotation as a whole.

Goals:

- Try to understand where variation in annotation comes from, what pieces introduce variation (e.g. curator background), correlate with variation in output annots (for same paper)
- Test out measures of annot quality from co-curated data + contextual data, define what annot quality facets might be

Study 1: (At SGD.) What factors influence output in a single MOD?

- 4 curators, several papers
- Look at different outputs
- Relatively homogenous group of curators

Study 2: (GOC) Same papers, different curator backgrounds, different MODs.
Study 3: (GOC vs. annotation camp trainees) Novice vs. expert

Aspects investigated:

- Consistency: At the level of the individual annotation, how similar/different are annotations between different curators.
- Reliability: Same paper, same curator, different time points.
    - 2 curators (Eurie, KarenC), 2 papers each
    - 1 curated 6 months ago, 1 curated a year ago
    - Issues related to which term to use (ontology developed in interim), also which evidence code to use (new guidelines)
- Specificity: Relative granularity of terms utilized across annotations.
    - How are terms used related to each other? Parent/child?
    - Are the terms an exact match?
    - Are the terms on the same or different branch?
    - If different, how far apart are these branches?
    - Note: Changes in the GO tree are not taken into account as the study was done over a relatively short amount of time, less than a month.
- Accuracy: Based on consensus annotation vs. individual annotations, measure variation. Where did the consensus come from?
- Completeness: Compare individual annotations vs. consensus annotation, looking for presence/absence not correctness.
- Validity: Comparison of instance annotation to GO standard file format. Are the fields filled in correctly?

JM will have 1-1 interaction with each curator participant (10 total) to get contextual information, i.e. curator's background and experience.

AmiGO discussion (Rama Balakrishnan)

Goals
Future
All info is on wiki page, can be accessed by all.

http://wiki.geneontology.org/index.php/AmiGO_Release_schedule

Problem: Google doesn't index AmiGO because of session id in URL. (Can this be remedied?

Need to get opinions from actual bench scientists, this is the user group we want to target (not necessarily us, the curators).

Why are you NOT using AmiGO?/ Suggestions to make AmiGO better.

- Numbers are confusing.
- OBO-Edit better (faster) for browsing
- Batch download
- Val: use QuickGO for quick graph view, prefers that view for browsing; Would be good to have AmiGO be more flexible and easy to understand at first glance.
- Desire for iterative queries, ways to narrow searches.
- Allow Boolean queries: AND, OR, NOT
- Desire for a complex query and then a way to download the results.
- Multiple hits are easier to view as a list in OBO-Edit vs. AmiGO
- Sorting problem
- Desire for intermediate results page with a limited amount of information (maybe just name, id, synonyms). Currently, defs take up a lot of real estate. Give option for return all results but default to show just 10 results. Select some, all.
- On first results page, show number of species and number of annotations to each term.
- Ignore word order in input (like the OBO-Edit keyword search does).
- Pankaj: Provide slim terms as home page (vs. only the root terms + obsolete).
- Improve home page to make it more user-friendly. Add some text to explain, add legends.
- OBO-Edit is good. Can provide ideas for where to take AmiGO.
- On term detail pages, display reference genome annotations as default? Expand if desired.
- "plant", "animal", "eubacteria" as grouping selections in addition to the scientific names

GO Users Meeting

Midori Harris announced that there is a GO Users meeting in Seattle in September and invited anyone to submit abstracts. She added that although the meeting is in conjunction with MGED9, it is not exclusively for microarray data and that it is a good way to connect with the wider research community.

Mike Cherry pointed out that these meeting were now 'quite respectable' and encouraged GO people to attend as a means of advertising what GO does to the wider community.

General Annotation Issues

1. Determination of orthologs for reference genomes

CONCLUSION - Determination of orthologs for reference genomes: This was in the agenda for the afternoon but was covered this morning. Each database will be given a list of genes to curate in their organism. Although other tools were proposed, the plan is to go with the tools we have (InParanoid, TreeFam, HomoloGene), and reevaluate later. The *E. coli* genome should be added to the Inparanoid set.

See additional discussion on page 18

2. Allowed IDs for references in the association files

RECOMMENDATION - IDs for references in the association files: Currently, to refer to references in the gene_association files, we allow the use of PubMed IDs for published papers and internal database identifiers for internal, unpublished references. We recommend expanding the allowed ID types for published papers to include these four: PubMed, AGRICOLA, BIOSIS, and ISBN. Allowing these additional IDs for published references will help us identify annotations made from published papers.

See additional discussion on page 19

3. Supplementary data from references

CONCLUSION - Supplementary data from references: Consensus is to treat the supplementary data from a paper as an integral part of the paper, using the same PMID. Good idea to save the information locally in case the journal removes the supplementary material from their website

4. Filtering ISS annotations without "with" information

RECOMMENDATION - Removing ISS annotations without "with" information:
ACTION ITEM: As of October 1st, 2006, "with" will be mandatory for ISS annotations made on this date or later. Starting on October 1st, annotations using the ISS evidence that do not contain a sequence identifier in the with column will be filtered out of the gene-association files. This rule does not affect annotations made prior to October 1st, 2006.

5. Internal GO References

Emily had a reference suggestion: for groups that use the same reference, like the one for InterPro2GO mapping – consolidate them into one reference.

ACTION ITEM: Midori and Karen are already have an Action Item from the St. Croix meeting to go through the GO references collection to consolidate. Once this is done, a proposal will be sent around for each group to confirm that a given consolidated abstract is a suitable description of their method. Once the GO references are consolidated, each can be associated with synonymous IDs to link a GO reference with all appropriate MOD internal references.

6. Analysis of GO co-annnotation

Another suggestion from Emily – they use tool/statistics to see which GO terms tend to be used together; the terms could be from different vocabularies. Then they look at the annotations being done and see if term(s) were perhaps missed. The approach could potentially improve curation consistency. Other groups could use it too.
David suggested it be put in AmiGO.
Val mentioned that they use the tool as well.
Karen C said the proposal should be placed on the AmiGO list for things to consider/to do.

7. Capturing common knowledge in the ontology

CONCLUSION: none

See discussion on page 19

Evidence Codes Issues

1. ISS

1A. General Use of the ISS evidence code by reference genomes

RECOMMENDATION 1. For Reference Genomes, gene product in the "with" field of an ISS annotation must have been annotated to one of the 5 experimental evidence codes (IDA, IMP, IGI, IPI, IEA). If a paper describes a sequence similarity to an uncharacterized gene, then no annotation can be made.

RECOMMENDATION 2. In cases where we need to refer in the with column to genes that have not yet been annotated, curators should contact the MOD or GOA (if there is no MOD for that organism) and make the appropriate annotations for these genes, then do the ISS for the gene product in their organism.

ACTION ITEM: ISS documentation (or reference?) needs to reflect that the annotations are being done over a significant part of the protein.

Question? (Pankaj Jaiswal): What happens when annotations of the "with" protein [geneB] changes? (No conclusion on that; expected to be infrequent).

See additional discussion on page 20

1B. Making ISS annotations from an IC?

(question from Ruth Lovering)

No. The reference genomes will only be making ISS annotations when the sequence ID in the with column can be annotated using one of the 5 experimental evidence codes: IDA, IMP, IGI, IPI, IEP.

There shouldn't really be any need to make an ISS from an IC anyway. For example: geneA is annotated to "MF: transcription factor activity, IDA" and "CC: nucleus, IC from 'transcription factor activity'". geneB (similar to geneA) can be annotated to "MF: transcription factor activity by ISS with geneA" and to "CC: nucleus by IC from 'transcription factor activity'", but *not* to "nucleus by ISS with geneA".

1C. ISS annotations for which there is conflicting experimental data

(question from Susan Tweedie)

If a gene product A is predicted to have some activity by sequence similarity (Gene product A - some activity - ISS - with Gene product B), and further analysis of the sequence shows that this gene product is missing critical residues required for that activity, curators should add the "NOT" annotation (Gene product A - NOT some activity - RCA) and remove the other (ISS) annotation.

If an alternative paper describes an experiment that shows gene product A does not have activity then this gets the annotation: "Gene product A - NOT some activity - IDA" and remove the ISS annotation.

Of course, if it is obvious from the literature that there is ongoing controversy, it may be apppropriate to keep conflicting annotations from the various papers.

1D. NOT annotations by ISS

(question from Rama Balakrishnan & Karen Christie)

The example is that of a gene family where some members have an activity while others don't (specific example is RCL1 in *S. cerevisiae*). You can tell by certain residues missing. How do you annotate these?

Two possibilities were agreed to be acceptable

1. If there is experimental evidence for members of both groups:
group I: Gene product A - some activity - ISS with gene B (which also has the activity by IDA)
group II: Gene product A - NOT some activity - ISS with gene C (which also DOES NOT have the activity by IDA)

2. If there is experimental evidence for one group only
group I: Gene product A - some activity - ISS with gene B (which also has the activity by IDA)
group II: Gene product A - NOT some activity - ISS with gene B (which also DOES NOT have the activity when the key residue is mutated by IMP)

see additional discussion on page 22

2. ISS, RCA, or IEA? e.g. InterPro2GO, TMHMM, tRNA scan, snoRNAs

(Guy Plunkett III, Doug Howe, Karen Christie)

What evidence code should we use for these tools?

RECOMMENDATION – ISS, RCA, or IEA? e.g. InterPro2GO, TMHMM, tRNA scan, snoRNAs

Consensus is to use IEA; if a curator reviews the annotation, it becomes a RCA with InterPro domain or tool name in the "with" column.
Annotations made using InterPro and other computational tools that use HMM-based algorithms (TMHMM, SignalP, etc) should have RCA as the evidence code. The rationale is that the HMM is generated is through comparison of a large number of sequences having that function (SwissProt set), therefore not a strict sequence comparison. The conservation of a residue is assumed to imply its importance for function, but there is not necessarily experimental data supporting this.

NOTE added in summation: There seems to be a lack of clarity on the proposed new boundaries between ISS, RCA, and IEA, particularly RCA and IEA. Even just the above two paragraphs leave me confused as to where one would use IEA versus RCA for an HMM-based method. The group as a whole may need to discuss this further.

ACTION ITEM: (Midori Harris): the documentation for ISS, RCA, and IEA must be updated to reflect these changes, including updating documentation on use of RCA to include reviewed sequence data.

See additional discussion on page 23

3. IPI and Process *(David Hill)*

CONCLUSION - IPI and Process
IPI can be used with care for process annotation. The use of the 'with' field is strongly recommended for IPI. In cases where interactions are with partial proteins or domains the full length protein ID should be used. Any cases where a suitable target ID cannot be found should be referred to the reference genome list. For reference genomes, whatever is put in the 'with' field should be experimentally characterised and should be annotated to that process. Non-reference genomes may have to accept less stringent standards about the state of characterization of the gene referred to in the with column.

See additional discussion on page 25

4. IC – what should this really mean?

CONCLUSION: What should IC really mean?: No change. IC should be reserved for curator statements; if the annotation comes from an author statement then NAS or TAS should be used. We should try to include a GO term ID in the 'with' column for NAS. In some cases the curator may go outside the paper that is being read to make an IC annotation. In such cases, where the annotation is based on a combination of information the reference for the annotation will usually be the initial paper that prompted it and inspired the curator to look for supporting knowledge.

See additional discussion on page 26

ACTION ITEM: review documentation for IC

5. Use of NAS

RECOMMENDATION – Use of NAS: Although the use of NAS is not encouraged for reference genomes, the group suggested that in the future it may be used similarly IC to capture process annotations inferred by the author. In these cases, try to include a 'with' GO term ID with the NAS code. There may be cases where some groups may choose to make an NAS annotation where it is not relevant to use a GO id in the with column. NAS is still excluded from certain reference genome metrics.

ACTION ITEM: Send cases where NAS target cannot be identified to reference genome list to discuss if is possible to always require NAS to have a GOid in the with column.

See additional discussions on page 27

6. Use of TAS

RECOMMENDATION – Use of TAS: The group agreed that the use of TAS should be limited to annotations which are made on the basis of a statement the author makes where they refer to experimental data from another paper. Annotations based on author statements which are drawn from things within that paper should use the NAS evidence code, if they are annotated at all. Note added in summary: I recall coming to this agreement, but it doesn't seem to have made it into the discussion transcripts and wasn't particularly contentious, but may reflect a slight change in the usage of this evidence code, so the GOC as a whole should come to agreement on this recommended usage.

See additional discussions on page 27

7. IPI versus IDA

RECOMMENDATION: IPI versus IDA: IPI is preferred over IDA in all cases where an interaction target ID can be identified. This includes binding to self and binding to specific protein groups. For example, a protein binding actin where the specific isoform of actin is identified the annotation should be IPI 'with' that protein ID rather than simply actin binding IDA. This preference for IPI is based on the fact that it captures the more information ('with' is not allowed for IDA) and allows us to distinguish between isoforms. If there no clear direct interaction then IDA should be used.

See additional discussion on page 28

8. IMP versus IDA

IMP versus IDA controversy – mainly due to variability in the use of IMP as an evidence code.

RECOMMENDATION - IMP versus IDA: After much discussion, there was general agreement that the current guidelines for IMP, particularly the phrase "Overexpression/ectopic expression of wild-type or mutant genes" were too broad and were leading to the use of the IMP code for types of experiments that authors would consider to be a direct assay of function.

Particularly in mammalian systems, there are a large number of experiments performed where a gene from one organism is transfected into a cell line that might be from a different organism, often with another reporter plasmid and along with a series of control plasmids. Typically these experiments are interpreted by the authors as direct evidence that the gene transfected in has the function that appeared in the cell line. As the authors consider their experiment to be a direct assay of function, the appropriate evidence code should be IDA.

In contrast, in the example Karen brought up, Study Paper 1 on wybutosine biosynthesis (PMID 16642040), the authors made mutant strains and characterized differences in which biochemical intermediates were present in various mutant strains. Despite the complexity of their biochemical assay, clearly these authors are characterizing mutant strains. Thus the evidence code for these experiments is IMP.

Basically, the recommendation is to follow the author's lead/thinking as to whether they are making inferences on the basis of having made mutations (or on the basis of comparing normal allelic variation) versus making inferences on the basis of some experiment that tries to directly address the function of a gene product and to change the guidelines accordingly.

See additional discussion on page 29

9. Evidence code to indicate large scale experiments

RECOMMENDATION: Evidence code for large scale experiments: There is qualified support for the introduction of new evidence codes that distinguish between small scale and large scale experiments. The fact that this has been requested by the user community was generally accepted as a strong argument in favour of this plan. It was agreed that the best solution would be to introduce five new evidence codes corresponding to the experimental evidence codes appended with HTP (high-through-put) e.g. IMP HTP (the exact format of the codes is still open to debate).

See additional discussion on page 34

ACTION ITEM: Need to decide how to identify small versus large data set. We should seek examples of such data, particularly at the boundary between the two classes and post them to the GO list. [Need to consider whether any of the other non-experimental evidence codes also need the HTP qualification?]

TRANSCRIPTS OF DISCUSSIONS

Discussion - Metrics to monitor breadth:

Suggestion from Val: How about doing the counts centrally by script from the gene_association files and/or GO database?

Judy: The point is not "comparative" or a competition, who/what database has more annotations, but rather, for your particular organism, what was known before vs. what is known now.

Mike: Goal is to show progress and have this progress well-documented. This is not a contest.

Judy: RG groups should work towards common usage of evidence codes across MODs because the outside world wants consistency

Mike: Some algorithms use GO annotations for training sets but may not take evidence codes into consideration, which can lead to errors. We need to educate and keep educating our user community.

Note: Issue of annotations resulting from large-scale/high throughput analyses remain, for example, whole proteome protein-protein interaction sets. (More discussion on HTP data sets followed later in the meeting.) MODs should decide on which data can and cannot be used for GO annotation as they (the MODs in conjunction with the community members that authored the dataset) know their datasets best.

Point made: Using only experimentally derived annotations for measuring annotation progress is meant to capture what is meaningful and what we are confident is correct. We may be underrepresenting the extent of knowledge, but that is ok.

Q about ISS: Is there a difference between making an ISS to a gene in another organism vs. an ISS to a gene in the same organism?
A: No, as long as the evidence_with sequence has been experimentally verified to have that function or be involved in that biological process.

ISS is better than TAS because it provides evidence with information.

Discussion - Metrics - How to measure depth?:

David: Some kind of flag that marks a gene as done, per aspect of GO, at a particular point in time.
Q: what about new lit? how long is 'done' flag good?
KarenC: SGD 'last reviewed' date (David: MGI has similar); annots keep date created and date reviewed

Val: In specific, are we going to send GO numbers of papers per gene on a gene by gene basis?

Rex: Probably not, probably just send percentages (papers used/total number).

Q to group: Can all MODs implement a way to track whether a gene is done or not?
GOC answer: Yes. Those who do not currently have ways of tracking this will be starting from zero. This number will increase as time goes by and fluctuate with the influx of new literature.

4. Number of papers that are read but not used for GO annotation
   What does 'read' mean? Abstract only? Full text? Depends on MOD?
   Least important/critical metric
   How could each DB provide this information?

5. Distance of term used to leaf (and to root?)
   GO can calculate this (Suzi has some ideas)
   Aim is to use leaf terms whenever possible.
   Ontology may expand over time so that a term that was a leaf at one point is no longer one. This is a recognized fact but the aggregate annotations should move over time to the leaf terms.

Discussion - Determination of orthologs for reference genomes

This was in the agenda but was covered this morning.
-Pascale Gaudet: who establishes the orthology (do curators do that?) Rex Chisholm/Mike: this will come from a table generated by Rex Chisholm and others. Each database will be given a list of genes to curate in their organism.

Other possible tools to establish orthology:
- HCOP: (Ruth) Currently has mouse, human and chicken. Could probably be adapted to include other reference genomes.
Produces a table that establishes the orthology and give a confidence level.
Objections are that this tool doesn't include Inparanoid.

- YOGY - eukaryotic genomes (Valerie Wood)
http://www.sanger.ac.uk/PostGenomics/S_pombe/YOGY/
Judith Blake: we cannot use it, since prokaryotes are not included

- Rex Chisholm (Conclusion): The plan is to go with the tools we have (Inparanoid, treefam, homologene), reevaluate later. *E. coli* genome should be run through Inparanoid.

Allowed IDs for references in the association files

Trying to find the corpus of literature. Distinction between peer-reviewed vs. not peer reviewed. Currently, a reference must have a PubMed ID to be counted as published, but there are some groups where significant numbers of papers are published in journals that are not indexed by PubMed.

How many are there? How do we distinguish between internal references (for ND) and internal references for articles that are not indexed by PubMed/Agricola/Biosis? Maybe we don't need to worry about this for reporting purposes because there really aren't that many references that don't get an ID from one of these:
PubMed, Agricola, BIOSIS, ISBN

NOTE: Though counting up IDs from these sources may slightly overestimate the number of peer-reviewed references are used for GO annotations because not all articles with PubMed IDs are peer-reviewed, we don't think this is a major concern.

Discussion – capturing common knowledge in the ontology:

During the discussion of IPI process, David H brought up his recent experience of looking at the CNS. There are lots of papers about homeobox mutations that result in patterning defects in the CNS - therefore transcription is necessary for patterning but the papers don't show that these proteins are transcription factors even though the knowledge is out there. This could be captured by IC or by the ontology itself.

Discussion followed about where annotations stop and the ontology begins...

David H is in favour of capturing as much common knowledge as possible in the ontology e.g. biochemical pathway data. He thinks users should be able to ask what genes are important in patterning. Val W supports the idea of including very specific terms.

Karen C pointed out that the ontology needs to be generally true and that capturing all of this knowledge often leads to true-path violations. Mammalian and fungal pathways are not necessarily the same. She feels there may be better ways to

address this problem e.g. less granularity and links between ontologies although it was acknowledged that this presents problems of what goes with what. The possibility of developing a suitable tool was suggested by Rex C and Jennifer S.

There is general debate/concern about the level of granularity desirable. Pascale G is worried this could lead to lots of examples and a gene product ontology. Victoria P is concerned about endless branching.

Discussion - Evidence Codes Issues: ISS

[NOTE: I will use "geneA" as the gene being annotated by ISS, and "geneB" as the annotated gene that geneA is compared to.]

Introduction by Judith Blake: It's interesting to consider as a group what we are doing with ISS. For example, orthologs versus homologs. In the mammalian genomes they do use orthologs, therefore the degree of similarity when annotations are transferred is always very high. This probably doesn't apply to every organism. One issue is the determination of the robustness of the annotation.

Judith Blake/Karen: "with" gene product [geneB] must have been annotated to one of the 5 experimental evidence codes (IDA, IMP, IGI, IPI, IEA)

Valerie Wood: there are cases where we annotate to genes that have not yet been annotated

Pankaj Jaiswal: What happens when annotations of the "with" protein [geneB] changes?
Other issue is sequence used to assess similarity might change. This is not necessarily kept track of; some IDs might get dropped.
Karen: reference genomes use comparisons to existing sequences.
At St. Croix meeting, we decided that the "with" column is mandatory for ISS annotations.
Until then, SGD was not systematically using the "with" when annotating by ISS from papers, which sometimes are impossible to track. This will mean that some annotations will have to get dropped because the "with" gene was not directly experimentally characterized.

What goes in the with for an annotation made from a paper? (Kimberley)
Summary of the discussion: There are two criteria: a) What the authors show and b) for the annotation to be acceptable, "geneB" to must be experimentally characterized. Otherwise no annotation is made. Reference genomes would be a core dataset highly dependent on experimental data-- therefore higher quality

Judith Blake: The quality of the annotation depends on the power of the orthology determination tools, as well as the evidence in the "with" protein
Rex Chisholm: if you have to go too far from the gene you are curating, you potentially get misled. This is why it is forbidden.

Karen ?: orthology is important here

David Hill: the "with" field must be an object that has experimental evidence

Rex Chisholm: an issue might be that the similarity to the closest characterized gene is too low to make an annotation --- then no annotation is made.

Judith Blake: maybe IC would be a better code. (to which GO term?).

Emily Dimmer: At GOA orthology annotated with both IEA and ISS. IEA being done with Compara. Advantage is be that it allows to update IEAs. Allows to *have* some orthology information.
ISS annotations will be done using a program provided by Panther
Rex Chisholm: that's fine, as long as you are able to define "high quality"

Susan T: Issue about reference: ISS annotations are made with an internal reference. What about if you need to read another paper? Can you cite that ?
David Hill: they curate that paper and send the information to the correct database or GOA and then import these ISS
Karen Christie: that's probably the right thing to do
David Hill: they have a file on a ftp site that contains all their annotations to human genes, and GOA takes it periodically and import the information in GOA.
Rex Chisholm: to summarize: if you identify an ISS in your genome, without an identifier to another MOD, then we could provide GOA with a file. If it's a reference genome the data should go to that MOD.

-Emily Dimmer/Judith Blake: only use the 5 experimental evidence codes

-What about IC? (people can't seem to agree)

-MGI also provides rat data, but that gets filtered out in AmiGO

-David Hill: GOA could act as the clearing house for all annotations available

- Rex Chisholm: quality of alignments: vary from organism to organism. In Dicty, we use 35% identity over 75% of the proteins. This number will vary between different organisms

- Judith Blake: ISS documentation needs to reflect that the annotations are being done over a significant part of the protein

Discussion - NOT annotations by ISS

(not sure about the issue here, but I'll give it a try) This was the RNA cyclase question.
The example is that of a gene family where some members have an activity while others don't. You can tell by certain residues missing. How do you annotate these? Two possibilities (?)

1. If there are IDA for members of both groups:
group I: Gene product A - some activity - ISS with gene B (which also has the activity by IDA)
group II: Gene product A - NOT some activity - ISS with gene C (which also DOES NOT have the activity by IDA)

2. If there is an IDA and an IMP for one group only
group I: Gene product A - some activity - ISS with gene B (which also has the activity by IDA)
group II: Gene product A - NOT some activity - ISS with gene B (which also DOES NOT have the activity by IMP)

Discussion:

Eurie Hong: the paper mutated the *E. coli* gene and lost activity; therefore can annotate to "NOT" with ISS

David Hill: gene B (the "with" gene) is annotated to the function by IDA and IMP; therefore if gene A is missing the important residue =NOT

Jennifer Smith: therefore you are comparing to a mutant protein
(David Hill: yes)

Karen Pilcher: isn't the annotation weird? you are saying "NOT" to a sequence. Shouldn't we then put in the with the mutant sequence?
Karen Christie: this is not really feasible. Plus there are very few NOT annotations; we probably do not need to expand that.

Rex Chisholm: you are missing a link: there is no evidence in geneA is missing the activity. This is only based on a comparison
Pascale Gaudet: but this is only an ISS, so it's fine
Judith Blake agrees with both

Judith Blake: we are going into too much detail.
Karen Christie: we need to know if we can make a rule

Ruth Lovering: it is possible to be missing a residue and you are not sure whether you have activity or not

ISS, RCA, or IEA

2. InterPro2GO/TMHMM

Issues brought up by Doug Howe on annotation mailing list:
I. If a paper says the gene product they are working on is a kinase because it has the proper InterPro domains consistent with a kinase, but they do not further characterize it as such, should that gene [product] be annotated to a reasonable GO term like "protein kinase activity" by ISS?

II. Hydrophobicity plots: what evidence code?

Conclusion to this discussion was that annotations made using InterPro and other computational tools that use HMM-based algorithms (TMHMM, SignalP, etc) should have RCA as the evidence code. The rationale is that the HMM is generated is through comparison of a large number of sequences having that function (SwissProt set), therefore not a strict sequence comparison. The conservation of a residue is assumed to imply its importance for function, but there is not necessarily experimental data supporting this.

ACTION ITEM: (Midori Harris): the ISS/RCA documentation must be updated to reflect this.

Discussion:

Emily Dimmer: Nicky Mulder is not very confident that this is right. It is a prediction.

-Pascale Gaudet: RCA? Judith Blake, David Hill seems to agree

-Emily Dimmer: InterPro2GO mappings are done by looking at all proteins that have these domains, from both SwissProt/TrEMBL. For a GO term to be added, ALL Swiss-Prot proteins must have been annotated to that function/process/component.

-Rex Chisholm: that is support for NOT using ISS

-Pankaj Jaiswal: if the annotation is reviewed by a curator, it should be valid

-Valerie Wood: for example: if a protein had a RNA binding domain, she annotates to RNA binding ISS.

-Judith Blake: reference genomes want high quality annotations so InterPro cannot be used to make ISS. InterPro annotations should be IEA

- Midori Harris: ISS documentation must be updated to reflect this.

-Rex Chisholm: but this has been a bad usage anyway.

-Jim Hu: for example: trp operon gene
Karen Christie: this is fine; this will be allowed; there will be a new reference code: inferred from genomic context

- Jennifer Smith: another issue is changing the definition of IEA- IEA now will also be used for reviewed information

-Pascale Gaudet: if you use IEA, how does one know that the information has been looked at?

- Judith Blake: evidence codes: in GO, purposely general. The "reference" should allow to make the difference. The rationale for not adding new evidence codes is that this could be endless and the information can be captured by using a different reference.

- Pankaj Jaiswal: we are making more work for the curators
Judith Blake/Karen Christie: that is right! but we are removing bad information. The reference genomes will just have higher standards. We want to provide data for all other genomes

- David Hill: InterPro is a consensus sequence. So it is a IEA or a RCA.

- Emily Dimmer: if you see problems with InterPro domains, write to them and they will fix it. The mappings should be highly reliable.

- DECISION: Karen Christie: so for reference genomes, what evidence code do we use? RCA with InterPro domain in with column.

3. tRNA scan/snoRNAs
-What evidence to use? RCA

Discussion - IPI and Process:

As presented in the agenda issues document, David H thinks that there is a distinct difference between using an IPI evidence code for an annotation to a molecular function vs. one for a biological process. Namely, when IPI is used for a function annotation, the functioning event has actually been observed. In most cases, this is some type of binding or it is used as supportive evidence in an assay for a multimer. When IPI is used for biological process annotations, the biological process is never actually observed as a result of the functioning event of the gene product under investigation. Instead, the process is associated with the gene product via "guilt by association"

David proposes we should not use IPI for process annotations and instead use IC. In the "with" field we should use GO:0005515 or the annotation to a complex with a known function. In addition to this, we should capture the binding information in a molecular function or cellular component annotation to the appropriate gene product.

For example beta-catenin binding protein binds beta-catenin IPI. Given the curator knows that beta-catenin is involved in Wnt signalling it is reasonable to annotate beta-catenin binding protein to that process. However, since the paper does not actually show that the beta-catenin binding protein is involved in Wnt signalling, it is not ok to use the evidence code IPI because there is no evidence to show that is true in the paper; it should be IC. Note that it is important to know for sure that beta-catenin is involved in Wnt signalling to avoid a chain of inferences.

Victoria P was uneasy about making process from IPI at all. Eurie H felt that if our aim is to capture the experiment then IPI protein binding is the only annotation that can be made. Jim H was concerned that the evidence for process can be very thin and users aren't going to figure this out; too many false inferences could make the analysis of process terms useless. Emily D was particularly concerned about inferring process from large scale experiments.

There was general concern that not all IPI data is meaningful (e.g. seemingly pointless binding of actin to DNaseI, David H). It was agreed that false positives are an unavoidable caveat all IPI data and we still have to annotate this data in the absence of better evidence.

Emily was concerned that if the author infers a process from IPI then we shouldn't be using IC - it is not inferred by the curator. She suggested using NAS for this situation - there is no actual evidence other than the binding. Karen C suggested changing the IC code to inferred by GO term as a way round this problem This idea found some support; a GO term would be required in the with field (see more discussion of these issues below).

Mike C and Val W pointed out that these inferences are often in the discussion so should not necessarily be trusted. According to current documentation, we don't capture GO data from the discussion.

Val W objected to using IC for process annotation; if the inference is based on the protein interaction data then it should be IPI. Otherwise we lose the data in the with column that shows what interaction the annotation is based on. Stacia agreed that it is not good to lose this information and thinks it should be IPI to capture the type of experiment the inference is based on. David H also came round to the to conclusion that since the interaction is the basis of the conclusion it is should be IPI. The authors make the conclusions so it is legitimate for us to represent them. The onus is on us to annotate everything.

It was agreed that IPI could be used for process annotation but that caution was required. The question was raised whether it would be mandatory to have something in the with column. Stacia E thought it should be mandatory and whatever is in the with column should be annotated to that process with an experimental evidence code. It was agreed that 'with' is strongly recommended. It was also acknowledged that there may be situations where there is only ISS evidence for target so this may have to be used.

Tanya B asked if the target may also be a synthetic seq. or protein domain. David H recommended using IDA for synthetic sequences e.g. for a transcription factor binding a hox gene promoter and the ID for the whole protein for domain interactions. However some clarification may still be required as to the preference for IDA v IPI for DNA protein interactions. Emily D and David H highlighted the value of using UniProt IDs as these allow you to identify specific isoforms if known or the generic isoform if not known.

Discussion: **What should IC really mean?**

It was generally agreed to keep use as it is for cases where there are no author statements that can be used as the basis of an annotation; IC should be reserved for curator statements. If the author makes the statement then it should either be based on experimental evidence or NAS/TAS. In case where an author makes the statement in the paper the curator can use NAS but you should try to include a GO ID whenever possible.

Pascale asked whether the references for a IC annotation should be an internal reference. It was generally agreed no - most MODs use the reference that they used to make the annotation on which the IC annotation is drawn from, even if this involved additional knowledge. For example, the curator makes an annotation to transcription factor from a reference with the evidence code IDA. If they want to make an annotation to nucleus by IC, you would use the same reference as was

used to make the annotation to “transcription factor”. Internal references are generally used for situation such as annotation to unknown (ND).

Discussion - NAS and TAS are sometimes used in odd ways

Judith Blake: aren't we moving forward? We don’t really want to use these anymore, so there is not much point using that.

Evelyn Camon question in list: sometimes we don’t have the choice, journal not available, etc
Pascale Gaudet: if the database wants to keep it, is it doing any harm?
Consensus is to not allow it. It's not adding useful information either, and sometimes it adds confusion, as the gene mentioned in the paper might actually be from another organism.

Emily Dimmer: what about a statement not really supported but that makes sense?
Karen Christie: that information is not so useful then
Judith Blake: maybe there is a place for NAS?

David Hill: kinase and ATP binding: this should be in the ontology
Karen Christie: this was removed because it caused TPVs across different organisms

Discussion – Use of NAS (and TAS):

The use of NAS was raised again during discussion of both IPI process annotation and what should IC mean. The discussion is summarised here.

Emily D asked why we favour the use of IC in situations where the author makes a statement. She feels uncomfortable using IC for author statements and thinks we should trust author and acknowledge that the annotation comes from the author not the curator.

Rex C agreed and was concerned that we are placing a higher value on the judgement of curators over authors. He pointed out that the author statements are based on additional knowledge and that the information is also peer reviewed. It is not our job to review the paper.

Emily D had suggested we use NAS for these cases instead of IC. Karen C pointed out that TAS/NAS were never meant to represent author statements contained within the paper being annotated but - more for capturing information from introductions and reviews. She agreed that it would be good to clarify the use.

Mike C pointed out that the current documentation for NAS would not conflict with this usage. David H said that we would still need to a GO term in the with column for these annotations.

Ruth was concerned that using NAS in this context didn't square with the advice to avoid NAS for reference genomes and asked for confirmation that we can use NAS with a GO ID. Karen C reiterated the fact that NAS is not banned for RGs just not recommended; it is still not experimental so it won't be counted in the metrics. David H pointed out that newer databases with fewer staff will still need to use NAS/TAS codes - they are just trying to get some annotations made. Karen is not sure that there will always be something to put in the 'with' column for NAS and suggested that we need examples.

There was further debate about whether IC should be changed to another term to take some of the onus off the curator. Suggestions were: Inferred from GO Term (Rama B) and Inferred from Curation (David H). Pascale G supported the idea of a more neutral term. It was agreed to keep IC as is and to allow the use of NAS to capture statements made by the author.

David H provided an example: If I draw the conclusion that a transcription factor is in the nucleus then it is IC; if the author draws that conclusion then it is NAS. The with field would contain the GOid for “transcription factor activity” in each of these cases. Note that this is an expansion of theuse of the with field for the NAS evidence code.

Discussion - IPI versus IDA:

SGD have an experiment where a spliceosome protein was tagged and the tag used to pull down the whole spliceosome. What evidence code should be used to annotate the association of these proteins with the cellular component term spliceosome? Karen C said that they have switched from IPI to IDA for this situation as you can't tell which of the 80 proteins identified in the spliceosome physically interact with each other so you wouldn't know what should legitimately go in the 'with' column.

This was generally agreed to be a correct usage of IDA. It is generally applicable to multicomponent complexes. Karen C pointed out that IPI would still be appropriate for complexes with fewer components and where cross-wise interactions are known.

Jennifer S. asked about whether the preference for IPI would apply to specific child terms of protein binding such as actin binding. She felt actin binding IDA was better in this case. David H pointed out that these protein A actin binding IDA is not the same as protein A actin binding IPI with actin isoform. In the second case you can capture the specific isoform of actin used in the experiment whereas the term actin

binding could involve any member of the actin protein family. It is better to include the specific isoform data where known.

Pascale G uses IDA for a protein binding to itself. Even in this case it was felt that IPI with the same protein in the with column is better than IDA. Val W thought this was good for computational reasons - information is lost with IDA and it is more complete to put it in.

Discussion - IMP versus IDA

The first example Karen gave was from the first paper used in the consistency study – the article on wybutosine biosynthesis. There was discussion on whether the evidence code should be IDA or IMP. The authors used mutant strains and then did a lot of experiments characterizing complicated biochemical phenotypes of those strains. In this case, there was general agreement that these experiments are characterization of mutants strains and thus should be given the IMP code.

David's recollection of a discussion from a long time ago regarding expression of the gene of interest in another organism and looking at localization – evidence code used/accepted was IDA.
Karen commented that our current annotation guidelines on IMP indicate that expressing anything in another system requires IMP.
David's question: "what's the mutation?"
RNAi is IMP (Ranjana), antibody blocking is IMP.
David's example of luciferase and co-transfection to assess something is IDA.
Ruth commenting on any change – see transfection, any assay that alters something - overexpression is like a phenotype and should be IMP.
Yet, David says that altering to the point that is like a dominant gain of function, which may not be the regular function of the gene, does not even warrant annotation to GO. According to David, this case is not about IMP or evidence code(s), is about not doing the annotation altogether.

Julie gave the example of opposite function(s) from cotransfection, overexpression, doing or not doing the experiment in the reference organism and the need/call for being more conservative.
Pascale says this is not the point.
Transcription factor expressed in the other organism is not the approach for the annotations to GO for the reference genome. If the gene is placed in a mouse cell, the annotations cannot be done for the organism from which the gene was generated – not sure who said it or if there were several people talking at the same time.
Pascale gave the example of HeLa cells which are not the normal cells but people do experiments in HeLa cells to infer function/draw conclusions.

Julie has a background in disease and genetics and she wants/needs to know the real function of the gene and the fly gene in the mouse cell, in her opinion, is not the real function of the gene.
Ranjana – on authors, do they make the distinction?
David - they put the gene in the other organism precisely to find out the function.
Jennifer – cannot do experiments in humans.
Julie – Oh yes, in clinical trials.
David – you're going to miss/lose info.
Julie – people are lazy and do not do the experiments in the proper cell lines.
WE do not police how authors do experiments, which cell type – several people.

Karen remembered the discussion from last year's annotation camp, in which Peter D'Eustachio of Reactome participated –we do have to annotate these types of transfection experiments because this is what the authors can do to find out about mammalian systems.
Rex – go back to what the authors say.
Karen – while there is a lot of curation judgment that goes on – from the annotation perspective we report what is written in the literature. How we should and should not represent, but clearly we should represent what is in the literature.
David did annotate using IGI because the authors put the mouse gene in some other cell type.
Julie argued that in the case discussed the issue is over-expression, not simply putting the gene in another cell type.
Rex – is it not legitimate to say that the gene product has that function and is involved in that process? That's why they [authors] did the experiment.
Julie/Ruth – difference between expression versus over and/or under-expression.

Karen - can we go back to IDA versus IMP?
Doug – he consistently uses IMP when the gene is placed in another organism.
David/Pascale – think that if it is complementing the function in the experiment, the evidence code should not be IMP but IDA.
David – the cell is used as a machine, it des not matter which cell type is.
Julie – it does.

Karen – remember that we're not talking about doing/not doing the annotations, but what evidence code should be used.
Pascale - assaying the function should get IDA; the process should get IMP.

Karen brought up the GO home page on the screen, to show the evidence codes entry and the definition for IMP.
Ruth pointed to the phrase 'causing a mutant phenotype'.
Karen – causing overexpression, etc., for the mutant, if you ignore the title.

David – what does the experiment tell us? Authors are in fact testing the normal function. In the case of overexpression of a mutant phenotype, that is a different case.
Karen – mentioned ectopic expression.
Rex – a protein from mouse expressed in *E. coli* is assayed for enzymatic activity – IDA? - he asked.
Karen - it depends on whether you look at the entire cell, then you don't know what is really going on.
Julie – but aren't there interactions with the host?
David – Of course.

Ruth – if you put a mouse gene in another cell type and look for localization, is that IMP?
Karen – No, it is IDA.
Pascale - if you do this as IMP, we no longer know what IMP is.
Midori – some mutations alter only the amount, not the sequence, structure, etc. of the product
Karen – can we come up with something that is clearer?
Doug – we use IMP used a lot, even if another gene is placed in Zebrafish cell.
David – Wow!

Kimberly – we are having different views of mutant phenotype.
[not me, but I don't remember who] – reads from? This was to test constructs, expression, etc in yeast. It is possible to say whether the gene 'mocks up' the actual/real function.
Pascale supports David's example/view.
Cell line from another species that are characterized cell lines for particular functions – are systems for instance differentiation. Authors don't say: I'm studying the mutation. They say they study the actual function.
Karen – do we have a guideline for IMP/IDA?
Function – you study the normal function – maybe the assay does not allow you to be that specific.
Pascale – supports the use of IDA more than the use of IMP.

Doug – an example: the zebrafish gene was isolated and introduced in some cell line (other organism) –the gene product is a kinase involved in cell division. Then, cell extracts, direct assay for activity and kinase activity was found – the activity term was annotated with IDA. However, he annotated the process 'cell division' with IMP because cells were dividing at half the speed.
Ruth – that's why I see apoptosis as a phenotype to be used with IMP. She says everybody interprets IDA and IMP. The documentation at the GO site has examples but not enough examples to make t clear. Give more examples before we can conclude.

Karen – we eliminated hierarchies for evidence codes. Yes, we should decide what experiments require what evidence codes.
Eurie – she sees arguments for both IDA and IMP – the cell line as a test tube, the environment as mutational because is artificial; for the transfection example discussed in Doug's example – probably IMP.
David – nailing – does the author think of the environment as mutational or does he think is valid?
Ectopic expression – Jennifer – another tissue, but to her placing the gene in another cell type is not ectopic per se. Ectopic expression is where is not expected.
Rex – you need to revise your thinking, revise the definition. It's not about IMP, it's about the definition and what it brings you to do, as Doug said. True ectopic is not about another cell line.
Jim Hu – you put the gene where you can.
Expression of luciferase is not phenotype.
Pascale/Rex/Jim Hu – back and forth on the issue of going back and going to the paper – authors do not look at what the abnormal function is, they look at the normal function and use the cell line as a test tube.

Karen – looks as if what we need to do is to change the documentation.
Jennifer – how do you deal with minor differences?
David – read the paper.

Karen – using a purified protein calls for IDA, if is not purified for IMP.
Rex – what's the rationale of purified/not purified?
Karen – if I look at the entire cell, something is overexpressed and I see something, I don't know who's contributing to what I see.
Rex/David – but it's not about what you think, what do the authors think?
Pascale – activity and triple mutant – different situation since it should have been IGI.
Ranjana – supports Rex, mutant phenotype – most people would think about an amino acid change as being a mutation.
Jennifer – rat genes are placed in many cell lines – most people don't care.
Julia disagrees.
Mike – is not for us to peer review the paper, whether they correctly used/chose the experiment. We report the results.

Back to the GO page and definitions

General - cell lines to study the function is fine for IDA.
IMP as it is, is misleading.
Ruth – we're going in circle – reduce IMP increase IDA.
Rex - back to the design of the experiment which was done to assess the actual function [IDA], to use IMP is to pervert its meaning.

Fiona – I agree, but I do as Doug does, because of the definition as it is now which mentions abnormal environment
Rex – if definition forces us to use IMP for things such as those discussed we need to change/revise the definition.

Karen proposed to vote – who's in favor for changing the definition? Everybody is in favor that transfection experiments, i.e. "ectopic" expression in a cell line to figure out function and use of IDA is fine. The evidence code that is appropriate depends on the intention of the paper.
Rex – we assay the normal function – IDA.
Karen's example – two alleles, one normal one mutant, how do we handle this?
Rex - change IMP documentation, look at the entire population, one cannot say what's abnormal; rather, what's more dominant and it will be different from population to population .

Proposal to have IMP – inferred from phenotypic variation rather than inferred from mutant phenotype (current title).
David thinks it is still confusing and pointed to Doug's case where decrease in division is phenotypic variation.
David – it is a phenotype, but it is testing the normal function – so IDA
Ruth – where do you draw the line – mutant and phenotype?
David – everything is a phenotype.
David – the intention of the author is important; in Doug's case he would have had the cell division with IDA as well.
Karen – if the real goal of evidence codes is to give a measure of the type of experiments then we may not be able to draw the line. Are the systems used to test the normal function? Then IDA.

IMP versus IDA distinction?
Val, David – what do the authors want to show, how do they do it

Pascale/David – David understands the rationale for choosing IMP in Doug's case.

Rex – this is not a perfect world. Some cases would be clearer cut than others, some on the edge.

David - very strict about sticking with what the authors say – one protein is phosphorylated at the right place as assayed from gels and a signal transduction cascade is concluded.

Rex – the KEY is the intent of authors .

Discussion - Evidence code to indicate large scale experiments:

Eurie H. raised the question of whether we should we flag annotations from large scale approaches versus traditional experiments so that users can separate them. Annotations from these experiments have different caveats to small scale experiments so we should try to be more specific about the evidence. However, in the last discussion we had about this at St. Croix people were worried about making value judgements.

Several people felt that while high throughput experiments can affect the quality of the data this didn't necessarily merit new evidence codes. Kim VA felt that some big experiments are better than others so they shouldn't be grouped together. David H agreed that users would want to screen out data from a specific dodgy experiment not all large scale data. He felt that the distinction can be made based on the reference rather than the evidence code. Pascale G objected to differentiating between experiments using the same method simply scaled-up.

Emily D recounted that Evelyn has had authors contact GOA to ask for only a subset of their published data to be annotated. All of the data had been annotated and the author complained even though all data in the paper was presented without qualification about what was good or bad.

Karen C and Mike C stressed that this is an issue of caveats not quality. There are differences between these experimental approaches that are not based on data quality and these have been documented in the literature (e.g. paper by Mike Tyers). Mike C also pointed out that users don't 'get' evidence codes and this would provide a useful division.

In support, Rex C argued that there is value in highlighting these experiments because they are subject to different caveats. Small scale experiments have a hypothesis whereas large scale experiments are not hypothesis driven. High throughput experiments may have completeness but each single data points has not been given same attention as small scale experiments. HTP experiments may not include replicates. If you do 1000 experiments with 0.5 cut-off you expect some false positives - this is different from a small scale experiment where you are testing one by one.

Eurie H said another reason to distinguish these experiments is that the curator isn't going to review the data in the way you would normally do. Special codes would reflect that.

Mike C pointed out that further reason for flagging these experiments is that the user community has requested it. David H, Ruth L and others agreed that we should try to serve our user community so this was a good reason for marking these experiments.

Having agreed in principle to mark large scale experiments there was discussion about how this should be done.

David H. suggested tagging the paper as high throughput rather than using new evidence codes. This was rejected by Mike C on the grounds that the paper may also contain small scale experiments.

Val W suggested using qualifier field would be a good option but Karen C pointed out that the qualifier field qualifies the GO term not the evidence code. It was considered bad practice to alter the use of the qualifier column for this purpose.

Val W suggested using a single new evidence code HTP for high throughput experiments and that these codes could be changed later to IDA etc if the data is confirmed by experiment. While it was agreed that HTP is a good abbreviation for this code, there was concern that a single evidence code would not reflect the type of experiment used. Also, Karen C thought it is unlikely that curators would to go back to update/check these annotations.

Mike C proposed using all codes with HTP appended. Ruth L supported separate codes on the grounds that you retain information about the original experiment and don't have to go back to them. David H suggested five such codes corresponding to the experimental evidence codes. Stacia E thought we should have one for IEA too - especially since we are moving some things from ISS to IEA/RCA.

There was further debate about whether all of the evidence would need a HTP version. Karen C and David H argued that RCA HTP was redundant; if it is reviewed then it is not HTP. This led to discussion about how HTP annotations would be reviewed; would you have to review every data point to promote HTP annotations to RCA (Emily D)? Rex C acknowledged that large data sets are reviewed in a different way. MGI uses RCA only if individual annotations are reviewed (David H). Karen C agreed that to change IEA to ISS you must look at each sequence so this is a hard question for HTP data. In reviewing HTP data, it was suggested that false positives should be removed but that conflicting data should be left in (Ruth L). Eurie H agreed that conflicting data can help spur new research.

As part of this discussion Karen C confirmed that RCA is not necessarily limited to non-sequence data (i.e. can be used to review sequence data) tRNA scan is RCA not ISS, snRNAs RCA not ISS and hydrophobicity plots are RCA not ISS. Guidelines for RCA will have to be updated to reflect this change in use.

Tanya B said that TAIR have ISS annotations without something in the with column which are based on many lines of evidence rather than a single piece of evidence.

(Manatee). She asked if there should they be promoted to RCA rather than leaving them as ISS without anything in the with column. It was agreed - yes.

David H said RCA should be used in cases where sequence based evidence in used in addition to other evidence. This raised the question of whether a combination of evidence such as TMHMM and sequence similarity should this be given a different evidence code. Susan T has no record of any response to this question but thinks the general consensus was no.

Returning to HTP...

Val W has had problems with data sets inferred from orthologs - got lots of false positives that conflicted with other annotations and all came from bulk uploaded HTP experiments. It was suggested that it would be necessary for data to be spot-checked before inferring process terms from an HTP experiment.

The remainder of the discussion focussed on how to identify HTP experiments. There was general concern that it was not clear what distinguished a large data set from a small one. Jim H wondered if the number of annotations associated with a single PubMed ID is a useful metric for determining this.

Rex C suggested that a large scale experiment would be a genome-wide approach. However, Mike C pointed out that a microarray experiments with 100 versus 10,000 genes are subject to the same caveats so should have the same evidence code.

As an aside, Emily D was concerned about annotating microarray data; GOA don't do it. Others agreed that they did not routinely annotate microarray data.

It was agreed that we need to find examples of data sets to discuss - particularly cases at the boundary between large and small.

Brief discussion(s) on the papers for the consistency set in order to be prepared for the actual workshop

Julie (SGD), first paper - yeast
Consensus – All genes with four annotations – did not follow this very well.
Ruth had a lot of annotations with IC evidence code – generated some discussion.
General: how should we curate for the consistency set?
Stacia – curate the way you usually do.
David on processes and the difficulty of dealing with them [using for annotations] - where to they begin and end?
Jennifer – that was exactly our ‘fight’ over terms and uses [at RGD].

Ruth reads from the paper to make it clear that [frame shifting(?)], on the basis of which she chose the GO term with the IC evidence code, was there.
Was in the paper, in the introduction, but the intro would have sent us to another paper and/or TAS which we were not supposed to do and use.

Annotation to parent term – tRNA modification [parent of wybutosine biosynthesis] – RGD used it only because the more granular term is so little known. They very seldom do this, but if the granular term seems to be so rare as to have very few people knowing what it is, they allow for the parent term to also be present to help the user.
SGD never does both parent and child from the same paper.
Ranjana – the paper was confusing on component information; the authors used component term.
Iron-sulfur cluster binding with IMP.
Karen - if that was the only paper on this issue, it would be ok.

Second paper - yeast
Stacia said was straightforward .
Pascale does not think they really showed ergosterol biosynthesis.
Karen pointed to a paragraph that warrants the use/choice.
We – Jennifer – had various stress responses and the parent term 'response to stress' because GO does not have all the terms in the vocabulary [new term needed – placed on the bottom of the consistency set].
Karen said just have the parent term [response to stress].
David thought we need the granular terms.
RGD - we put the parent term in addition to the more granular ones because GO didn't have all the granular ones. Ask for terms.
David and the use of IGI - Jennifer mentioned that we [RGD] used IGI for three of the terms.
Kimberly – response to cation stress was annotated with IEP.
Euries disagreed on the choice of the evidence code.

Third paper – worm - Kimberly presented.
Co-localization qualifier and arguments on its use – Pascale - should we always use it?
Karen – the historic perspective on the use of the qualifier(s).
Ruth - the use of NOT qualifier.
Karen - you don't expect [the gene product] to be in the membrane because it is a splice variant that is missing the membrane domain. Here, the use of NOT is not warranted; NOT is to be used when something is truly expected and is not borne out by experiment.
Tanya – did anyone use IGI instead of IMP.
RGD – we used the function term 'protein anchor', maybe…
Choice of term 'embryonic development' is realistic for C. elegans

David, Rex, others on two localizations – two localizations for the same gene products are possible, but right now in GO the only way to capture this information is by doing the double annotation although [caveat] the gene may be active in one localization but not in the other.

It is getting late and we are not going to be able to go over all the papers before the workshop.

Questions – were there any tricky paper?
David – Yes, paper eight – mouse.

The authors mention cloning the mouse gene and finding homologs based on GenBank accession number. Then the paper goes on describing a number of experiments and results. One [two curators?] at MGI noticed a discrepancy in the amino acid number between the figure in the paper and the gene in MGI. They believed was a typo and wrote a letter to the authors asking them what gene [species] they used in the experiments. It turned out that they used the human not the mouse gene. In the response they pointed to Figure 1A showing the sequence of mouse and human genes.

General – authors are many times ambiguous. If they are, so are our annotations. It is beyond our scope to call/e-mail the authors. If it is believed, based on the authors say that the gene in the experiment is from species A, we assume it is indeed from species A.