Summary

Software:

Biocuration Interface/ associated curation tools

- *Biocuration interface that allows annotation to a wide gene/protein set and range of evidence codes (AgBase)*
- *Create a common annotation engine, using web-services to support creation of curation interfaces in different groups (ZFIN)*
- *Make available a community curation tool to encourage expert submissions(BHF-UCL)*
- *Paper triaging (RGD)*
- *Creation of a centralized annotation request service to support annotation in different organisms to encourage creation of ISS statements (FlyBase)*
- RNA equivalent of the InterProScan pipeline to give researchers that first pass annotation of ncRNAs. (AgBase)
- *Text mining tools (BHF-UCL)*
- *Provision of an interspecies ortholog mapping file for all GOC groups (AspGD/CGD)*
- *Mechanisms for checking validity of ISS annotations (FlyBase)*
- *Provide an alternative to submitting annotations via CVS/SVN (dictyBase)*

GO term service

- *Support to update annotations when GO terms when obsolete, or improve when more granular terms created (MaizeGDB, WormBase)*
- *Co-annotation suggestions for GO terms, possibly using text mining or using pre-used annotation combinations, included in an annotation tool (SGD, FlyBase, PomBase, TAIR)*

AmiGO development:

AmiGO should display column 16 (SGD)

High-qulaity graphical representations of data (e.g. a ancestor chart slim) (MTBBASE)

Term autocomplete feature (TAIR)

Communication

More discussions regarding optimal annotation displays (MGI)

Increased documentation, including formal documentation on final decisions made by the GOC in a better-organised wiki structure (MTBBASE, dictyBase, FlyBase, WormBase)

Increased curator communication and discussions (BHF-UCL) including high-level discussions on the 'big picture' in GO (WormBase)

Would like further iInformation on work done to integrate GO annotations from other sources (e.g. Reactome, KEGG) (BHF-UCL)

Improved monitoring of annotation problems in different groups would be useful (BHF-UCL)

Increased feedback from curators/users of InterPro2GO (InterPro)

Annotation Format development

Develop the chain of evidence annotation format (SGD)

Procedures to determine the value/impact of new annotation formats to ensure efficient usage of time (SGD)

Further descriptiveness in the 'with/from' column (UniProt)

Annotation Quality Control

More QC checks implemented in the filter script (ZFIN, FlyBase, MTBBASE)

Checks on annotation consistency using Val's Matrix and PAINT (SGD)

Improved gp2protein files supplied by other GOC groups (UniProt)

Re-annotation assistance needed when terms are obsoleted and impact on annotation sets (FlyBase)

Selection of Curation Targets

Reduce interference in MODs setting their own annotation priorities (MGI)

Re-evaluate annotation project set-up. Allow special-interest groups to annotate across species? Reach-out more to domain-expert communities (WormBase)

Greater support for groups generating target lists for the ReferenceGenome effort (UniProt)

The annotation_extension field is useful to capture needed specificity (BHF-UCL)

Increased information on current focused annotation efforts (InterPro)

Ontology Development

More care is needed when considering the removal of biologist-friendly, or highly-descriptive GO terms (BHF-UCL)

*Finding that requesting new '*regulation of xxx process" terms is frustrating and the need is not clear (EcoCyc)

Additional support for term requests – (dictyBase)

Increased links between GO and other ontologies (AgBase)

Improved monitoring of term requests by specific groups (BHF-UCL)

Improved communication on the high-level changes being worked upon in the ontologies and their effects on individual terms (EcoCyc)

Continue developing TermGenie (FlyBase)

Additional Vocabularies:

Anatomy (AgBase)

tissue-specific gene-expression (AgBase)

Phenotype (dictyBase, GeneDB, TAIR)

Strain curation(dictyBase)

Terms to describe discrete functional contributions by protein domains (InterPro)

Bacterial growth phases (MTBBASE)

Enzyme kinetics values (MTBBASE)

Inhibitor substances (MTBBASE)

Pathway information (RGD)

Regulatory networks (SGD)

Temporal changes in protein localization or function (SGD)

Plant Ontology (TAIR)

Regulation targets/substrates (TAIR)

Biological context to localization/function/process (WormBase)

Increased expressivity in GO annotations (ZFIN)

Scientific Community

Improve free text access of publications (MGI)

Further user education (SGN)

4. Please provide your thoughts on how the GO Consortium can aid you in your curation process.

AgBase

Our biggest problem was getting a biocuration interface – it took me some time to eke out the money for someone to do this but we needed it because the wonderful Protein2GO interface only includes proteins and only UniProtKB proteins (and many chicken & cow proteins were not in UniProtKB). Also, does not support all the GOC evidence codes (esp. ISA, ISO) that we are using.

MaizeGDB

We would like to talk to other Model Organism Databases and learn how they handle GO annotations. One problem is the updating of GO terms, when they are obsoleted, or more granular terms become available.

MGI

For the most part, allow us to set our own priorities based on the mouse literature.

Finally, use the influence of the GO consortium and its community of users to press the publishing community for better access to complete free text.

MTBBASE

Drown me in documentation.

SGD

Tools that help guide the curator to find the right terms could be very useful (the "Curators who chose this term also chose these terms"…concept).

For example, when you choose an MF term for an enzyme in a pathway, a window displaying related BP terms for the process in which that enzyme functions would be helpful. Similarly for CC… It would also be helpful to have tools that display an entire branch (from least to most granular) in one click. I spend a lot of time clicking down a hierarchy to get to the right level of granularity.

It would be nice if AmiGO can keep up with the ontology and annotation developments. Col-16 for example is not displayed in AmiGO and if curators want to see some sample annotations they have to mine the GAF files which is not always efficient.

SGN

Closer collaboration with the GO would be really important for our database.

UniProt

More support for generating a target list for Reference Genome annotation. This should make it more attractive for specialized curators to suggest a particular biological process as a topic for the focused GO Consortium annotation.

ZFIN

Keeping the GO filter script up to date and as comprehensive as possible as a way to enforce GO curation policies is a good thing. Everyone should be running that on their GAF before they submit it,

and it acts as a safety net to catch policies that have been overlooked by groups or for which groups have not yet developed internal enforcement.

Producing a web service based common annotation system that groups could generate their own curation interface around would be nice. That way groups could create a curation interface application or web based that looks the way they are familiar with but that has as much functionality and quality control enforcement behind it as the GOC can build into it. Also everyone is operating at the same level of policy enforcement if all the annotations go through a common pipe produced by the GOC. To me, this local understanding and enforcement of GOC policies at every site of curation is a big expenditure of redundant effort. Centralizing GO annotation QC especially would be a great help.

If the common annotation tool kept pace with newer developments, like the use of column 16, then many groups that would never develop that functionality due to resource limitations, might use it. Of course, when those annotations come back to the MOD the col.16 data would be stripped out unless the MOD updates their schema and display to support thatâ€¦but that's another story.

4A. What gene function-related information do you currently, or do you want to, capture using a controlled vocabulary that you currently CANNOT capture with GO terms?

AgBase

anatomy

Gene expression (tissue specific)

BHF-UCL

We aim to capture annotations in all 3 ontology domains, however at times the 'function' of a protein is considered to be a 'process' or not a 'function' for example cell adhesion molecule is considered just 'sticky' rather than having the 'function' of being required to ensure two cells adhere which can be a necessary step in the process of differentiation. There are more specific 'process' terms available now, and the use of column 16 may also mean that this direct involvement of a cell adhesion molecule in cell adhesion may now be captured. A similar problem may arise if structural constituent activities are removed see SF request**:** structural constituent of cardiac muscle - ID: 2174293.

In addition, some of the annotations that we felt unable to capture were lost due to the loss of specific enzyme terms, which is now addressed by the column 16 facility.

However, we do not have a procedure in place to track specific examples of where annotation is limited by the annotation approaches available. As a way of really tracking/monitoring these problems it would be sensible to set up an online form where these problems can be logged. Some of these problems are discussed in sourceforge, where a request for a term is not granted, perhaps a new SF dropdown choice could be made available, so that there is the option of selecting something that implies 'ontology requested not in line with current ontology practice, but which means that it is not possible to fully annotate the gene', instead of 'rejected'.

dictyBase

A major part of our literature annotation is phenotype annotation. We are continually developing our own phenotype ontology for that, and controlled vocabularies for assays and environment. We also use controlled vocabularies for strain curation, such as mutagenesis type and strain characteristics.

EcoCyc

Off hand, I can't think of things I can not capture. However, what I have begun to find cumbersome is the need for new "regulation of xxx process" type terms. I know there must have been a discussion on this topic, and I am probably missing the arguments for/against. However, it seems to me that the large majority of biological processes will somewhere/in some organism be regulated by some gene product. So this will essentially double (or rather, multiply by a much larger factor) the size of the ontology with all the corresponding "regulation of" terms. I wonder if there isn't a simpler way to do this.

FlyBase

None

GeneDB

Phenotype data – but we have started writing a parasite specific phenotype ontology to capture this (this is a collaboration with Midori Harris at PomBase)

InterPro

Difficult to say, since I don't believe we really grasp the full complexity and sophistication of the GO. The ongoing effort to apply terms to complexes should be beneficial to our curation efforts.

We would also be interested in the creation of terms that describe the discrete functional contributions made by protein domains, since we plan to add this type of annotation to our signatures (although in many cases existing GO terms are sufficient for this).

MaizeGB

There are likely terms without granularity for all the described gene products in maize.

MGI

None.

MTBBASE

Bacterial growth phases. Inhibitor substances of a speci_c gene product's function. Enzyme kinetics values.

PomBase

None.

RGD

RGD captures pathway information for genes using the pathway ontology (PW). The PW provides an additional list of pathway terms, some of which fall outside the domain(s) covered by GO, that can supplement pathway terms found in GO.

SGD

We are looking into capturing regulatory networks using controlled vocabulary; along those lines, we would to like to be able to capture temporal changes in protein localization or function.

TAIR

currently capture: expression (Plant Ontology)

want to: phenotype. other: regulation target/substrate? (e.g. protein A phosphorylates protein B)

WormBase

One type of information that we currently cannot capture in GO is related to the biological conditions under which certain annotations can be made. For example, the DAF-16 FOXO transcription factor is regulated, in part, by its localization. DAF-16 is sequestered in the cytoplasm when insulin signaling is active, but translocates to the nucleus to regulate transcription when insulin signaling is downregulated. We have DAF-16 annotations to both cytoplasm and nucleus, but we cannot convey the circumstances under which this localization occurs.

ZFIN

We currently can post compose entities in our phenotype annotations that effectively allow us to be more expressive in our phenotype annotation than we can be in our GO annotation. Essentially, I think column 16 would effectively address this difference, but we don't currently have plans to develop support for column 16.

4B. Are there any tools or procedures that you feel could improve the quality of your GO annotation set?

AgBase

One of my concerns is that with the transcriptome becoming more important, we have no RNA equivalent of the InterProScan pipeline to give researchers that first pass annotation of ncRNAs.

AspGD/CGD

Actually, we wonder if other groups might benefit from our procedure for flagging annotations that are affected by changes in the structure of the term ontology itself. We are happy to share.

BHF-UCL

We would be interested in a text mining tool that was able to help with the full annotation of specific processes and which was able to filter papers so that only papers where the species of the gene product discussed is available. But this is probably difficult to do!

Reactome and KEGG provide annotations, some of which are imported into GO. However, we don't know how much about progress on the pipeline from Reactome to GO, are there plans to really capture all of the annotations here, is anyone actively trying to get new GO terms created to enable more effective mapping of Reactome data, and is there a method in place to ensure that each new Reactome pathway is mapped to highly specific GO terms as soon as the pathway is released? It would be really good to feel that if someone did their microarray analysis with Reactome data and then with GO data that GO would provide just as good, if not better, interpretations of their data.

Possibly being able to capture the protein modifications which influence the activity of the protein, this was discussed at EBI last week and it was suggested that this could be captured in column 17.

We are very interested in using a community annotation tool. We currently teach MSc students to annotate papers and ask them to complete a submission form which we mark and return to them (eg attached). From this experience there are several things that we would consider very important facilities in a community annotation tool (in addition to the expected annotation columns). Mostly because when people do something wrong it takes a long time to read through a paper and reach the

conclusion they have done something wrong.

1. For the annotator: Field to provide information about the source of the species of the gene product being annotated, ideally linked to the gene product/and or the gene annotation. Advantages/disadvantages placement of links, it would be tedious to have to keep adding the source to each annotation row, but some papers use different species in different experiments.
2. For the annotator: Field to provide information about the figure/text which supports each annotation
3. For the reviewer: Feedback fields for each of the fields listed above
4. For the reviewer: ability to revise the annotations made, but keep the submitters annotations visible so that they can see the changes made.
5. For the reviewer: check box to mark each annotation as accurate and suitable to export to the GOC database
6. Exportable data, ie to have the annotations exported to the groups own tool so that future modifications can be made if necessary.

dictyBase

Have an easier way to submit new terms than SourceForge. Would be nice if requesting a new term, with fitting it into hierarchy and seeing definitions, was integrated in a GO curation tool.

Viewing other usages of GO terms, especially 'GOC-approved' uses.

EcoCyc

We need better coordination with EcoliWiki; communication between the groups is good, but I believe that limited programmer resources on both ends have held us back.

FlyBase

If we ignore the fact that it will be difficult to develop software that would integrate with our existing pipeline/database… it would be good to have additional QC checking at the point of data entry – e.g. suggesting additional terms in other aspect of GO, ensuring reciprocal annotations are made.

Checking validity of ISS annotations.

Software that checks that 'with' dbxrf entries still exist and, in the case of ISS, are valid for use.

InterPro

Increased feedback where our GO mappings are incorrect. Currently we receive some feedback through the Sourceforge Gene Ontology tracker, but we could encourage more of this.

We would also be interested in receiving feedback when an area of the GO has undergone a focused annotation effort, so that InterPro2GO mappings could be revised and improved based on improved manual GO annotations. Tools to semi-automatically review such revised mappings would be very useful.

MaizeGB

Not at this time

MGI

At this time we fell that the quality of our annotation is very good. We keep thinking about better visual displays.

MTBBASE

More diagnostics through the abovem. Perl script

PomBase

We hope to include ways to make suggestions for GO terms from other annotations and report inconsistent annotations, within CANTO

RGD

An automated, pre-sorted PubMed abstract pipeline for each rat gene would improve the speed of the curation process and thus, improve the functional annotation coverage of the rat genome.

SGD

1) It would be nice to have some consistency checking tool (like the matrix) and perhaps use PAINT to assess how coherent (or not) annotations are for a given family.

2) It would be nice to have a system where evidence from multiple papers can be used to annotate to a term (the annotation file format is limiting how we annotate). System to utilize Chain of Evidence and be able to build complex annotations (lego) should be considered. For example, for the new transcription factor activity terms in MF, it would be really nice to be able to link annotations together to indicate which pieces, e.g. a DNA binding term, support the "major" MF annotation, e.g a transcription factor activity term. It would also be nice to be able to link annotations across GO aspects. For example, there is a group of 8 LSM proteins that form 2 different 7 membered rings, where 6 members are the same and 1 subunit differs between the two. One acts in splicing of the U6 snRNA in the nucleus; the other acts in mRNA degradation in the cytoplasm. The two subunits that are present only in one complex are very frequently the target of incorrect computationally predicted annotations due to their association with the 6 subunits that actually are involved in both processes.

3) Sometimes GOC introduces more layers to the curation which makes it hard for the users and for curators to pick up and run. For example, while the idea of integral_to qualifier is great, the return on investment (ROI) is not obvious (plus as always there is a concern about how users would use it), and if various groups annotate to different levels, consistency is at stake. It would be nice to if there is a procedure in place to decide what is absolutely required/critical for an annotation.

SGN

I find that the biologist out there is still not very familiar with GO. Maybe more efforts should be directed at educating the researchers.

TAIR

I always thought it would be nice to have a tool that could offer a 'second opinion' on which GO term to use so that I could compare which is better. Essentially this is a tool that can suggest GO terms by analysing the uploaded text passage (result section from the full text).

UniProt

Improved gp2protein files from other GO Consortium groups

It would be useful to be able to include relationships for 'with/from' identifiers so that users could understand their relationship to the rest of the annotation.

WormBase

We would like to have a better way to keep up with changes to the ontology, particularly when new child terms may be added for terms to which we already have annotations. This is a feature that would be good for the common annotation tool; perhaps flagging annotations that were made prior to the date at which child terms were added to that branch of the ontology would help.

Visualization of other annotations from different groups in the context of annotation may help curation since different groups may bring expertise to different areas of biology. Again, this is a feature that would be great to have in the common annotation tool.

ZFIN

Centralized GO QC mechanisms would be helpful.

One idea might be to offer a web service to which all GO annotations could be passed as they are created. I hit the "submit" button in my curation interface to send a GO annotation into ZFINâ€¦this triggers a web service call to the GOC annotation checker service. The call sends the proposed annotation to the service where it is validated for all current GO annotation policies, and a response is sent back. If it passes, commit to ZFIN..if it fails, then the message sent from the service would alert the curator to the problem. This puts the QC centrally in the hands of the GOC.

4C. Are there other ways the GO Consortium could help your GO annotation and/or submission process?

AgBase

I would like to see links developed between the GO and other ontologies – but I feel that this is happening already (or starting to).

AspGD/CGD

Calculation of ortholog mappings is a large and time-consuming task. If the GOC could provide an interspecies orthology mapping file, that would be helpful. We realize that no single method of computing orthologs is going to be perfect, but if the GOC chose one reasonable approach and used it to create a standard (even just pairwise) mapping, it would be a great and very useful starting point.

BHF-UCL

For project reporting it would be useful to be able to list how many requests Varsha and I have submitted to SourceForge/TermGenie as part of a specific funding stream. i.e. to have a way of filtering the returned number of SF results by date. Potentially also enabling an annual number of SF requests to be given. But this is not that important. Having the ability to confirm the number of GO terms supported by the BHF in AmiGO has really helped, although if we get a renewal it might be good to filter this on date!!! However, I am sure I can manage just to subtract the number of BHF funded terms in the final grant report from the total in the future!

Scheduling more annotation-focused camps/jamborees/meetings to test thoroughly major ontology revisions of basic processes such as transcription, signalling, apoptosis etc. would be very useful. Each MOD/group works in isolation much of the time, and these meetings help identify areas where

different approaches to annotation are developing and ensure that everyone is interpreting the ontology and the terms in a similar way. Funding to attend such meetings is always an issue, but it would be great if meetings like the upcoming Stanford meeting could be scheduled once every 12-18 months. In between face-to-face meetings, more frequent electronic jamborees, could be a productive approach to this, (most recently Emily and Varsha put one together for the heart development transcription factor co-curation project), but the format still needs working on so that attendees fully participate.

EcoCyc

Is there a web site with an easily accessible overview over "these are the higher-level changes we have been making"? For example, I know that the part of the ontology dealing with transcriptional regulation has been overhauled. But I can't find out easily what has been done, and what the rationale was.

dictyBase

It would be nice to have a HTTP endpoint for GAF file submission instead of going through cvs/subversion repositories.

FlyBase

Make sure that all decisions relating to annotation from meetings end up formally documented in the annotations guideline on the official web site. The wiki is great but it can be very difficult to distinguish between what is a final decision from a work in progress from an abandoned idea. The new links from QuickGO to relevant annotation guidelines are very helpful.

By continued improvement of the ontology - especially by adding comments and removing ambiguity from definitions.

By continuing to develop term genie so that it is easier to request terms.

By helping out with reannotation when term overhaul/obsoletion significantly impacts existing annotation sets.

Via a centralized annotation request service? Making ISS annotations from papers can be frustrating because the annotation does always exist in the other species. Generally there is no time to check the cited paper for the other species to ensure it has the data, check the species and contact the relevant database (this type of work often leads to a lengthy paper-trail. It would be great if you could submit a request to GO along the lines of – PMID V cites gene product W from species X as having function Y based on PMID Z. Then a GO funded curator could check it out and either make the appropriate EXP and ISS annotations or forward the request to the relevant MOD and feedback to the submitting group.

InterPro

Training support and collaboration on projects aimed at improving manual GO mappings that are subsequently used for automatic annotation. We currently provide a significant number of GO terms for proteins, but have no dedicated funding or curator resource specifically related to our GO mapping activities.

We would also like to work with the Consortium to identify areas of protein space space that are poorly covered by the GO to target their annotation.

MGI

We would also welcome a better annotation interface than we currently have.

MTBBASE

1-Provide me with high-qual highly detailed graphical representations of my data, or of e.g. the slim derived from my data with respect to the general prokaryote slim. A picture is worth a thousand words.
2There is no graphical repr. of annotations on the GO web site. Someone asked on Wikipedia-DE about GO applications: is it for text search?" No one outside the lab (and perhaps not many inside) can imagine what it is for. I cannot do this on the MTBBASE web site but it would be useful to link to.

TAIR

I hope we add term autocomplete feature to AMIGO.

WormBase

The wiki is a great source of information, but is still poorly organized. It seems there are good top-level categories, but then, there are lots of individual items attached to them that are in need of better organization into sub-categories. This seems to be a general problem with wikis (WormBase has the same issue). It's easy to create one-off wiki pages, but a lot harder to organize them once they've been made.

Ongoing communication from the GO consortium to curators about the overall direction of the project would be helpful. The bi-weekly annotation calls, although they help address specific curatorial needs, should devote some time, perhaps at the beginning of the call, to allow GO managers or GO top to briefly update curators on what 'big picture' ideas are being discussed and/or planned. There are a lot of potential changes coming to how GO curation will be performed, and it's not clear that curators are fully aware of them. The semi-annual meetings help with this, but ongoing communication from managers and GO top about these changes is absolutely essential.

Re-examine the role and types of special curation projects that GO undertakes. For example, would it be better to have curators keenly interested in a given topic curate genes relevant to that topic across all organisms? Can consortium members do more to reach out to their user communities to find out what types of gene products (e.g., drug targets) or topics (e.g., metabolic diseases) are of greatest interest?

As you've done with this survey, continue to involve, inform and invite suggestions from curators on different aspects of the project, including ontology development, annotation tools, software development, and organization and communication of project information.

ZFIN

Robust support for centralized quality control is the big one from my perspective. That would free up developer time for everyone to focus on annotation and it would simplify the code at every site and prevent everyone from having to interpret GO annotation policies the same way, which seems to be harder than it sounds.