3. Here we seek to understand your process for GO annotation (ideally as a detailed flowchart). This will help us effectively organize the flow of standard annotation streams that will facilitate the adoption of a common GO annotation process among contributing groups.

AgBase

flowchart

BHF/UCL

blank

CGD/AspGD

Step 1. Manual curation of GO from the literature paper identification (automated and manual screening)-->full curation on a paper-by-paper basis for all species in our DBs: GO curation is one component of our curation process (described in more detail above). We curate GO for every gene described, as warranted by the evidence presented.Curators may choose to do a comprehensive review of all curation for a given gene if it appears to be warranted, and our tools allow us to mark the date that the entire gene's curation was last reviewed, as appropriate.

Step 2. Automated GO predictions

We supplement the manual, literature-based curation with inferences based on orthology and protein characteristics (domains/motifs), using the criteria described here:

- CGD Prediction of Gene Ontology (GO) annotations based on orthology: http://www.candidagenome.org/cgi-bin/reference/reference.pl?dbid=CAL0121033
- CGD Prediction of Gene Ontology (GO) annotations based on protein characteristics (e.g., domains and motifs): http://www.candidagenome.org/cgi-bin/reference/reference.pl?dbid=CAL0142013
- AspGD Prediction of Gene Ontology (GO) annotations based on protein characteristics (e.g., domains and motifs): http://www.aspergillusgenome.org/cgi-bin/reference/reference.pl?dbid=ASPL0000166200
- AspGD (2011) Prediction of Gene Ontology (GO) annotations based on orthology: http://www.aspergillusgenome.org/cgi-bin/reference/reference.pl?dbid=ASPL0000000005

Annotations are inferred where the ortholog has informative annotations with evidence code IDA, IPI, IGI or IMP, and where these annotations are nonredundant with (more granular than) existing manual annotations, and not in conflict with existing "NOT" annotations. The annotations made by this procedure are given evidence code IEA. We also use these inferences in text descriptions for genes that lack experimental characterization. We update the analysis quarterly or so.

Step 3. GO-completeness

For species for which we have completed curation of the literature (C. albicans, C. glabrata, A. nidulans, A. fumigatus, A. niger): after making the automated orthology and IPR-based annotations, where genes still lack an annotation in one or more of the GO aspects, we assign the root term with ND evidence to indicate that, to the best of our knowledge, we have reviewed all of the information that exists, and

there is no evidence make a GO term assignment for this gene in this GO aspect (in contrast to a situation where we have not yet finished review of all of the existing evidence in the corpus of scientific literature).
Step 4. Re-review of annotations affected by changes to the structure of the Gene Ontology itself. We conduct an automated routine review of the GO for changes to the ontology that affect terms to which we have annotations in our databases, and any such annotations are flagged for immediate re-evaluation. (For example, if a new child term is added to a term to which we have an annotation, this annotation is flagged for review to see if assignment to the new, more granular terms warranted.)

Dictybase

Reading paper ->manual markup for GO -> check terms in GO tool, or, if more complex in AmiGO -> Annotate each gene mentioned in paper for GO -> request new terms when necessary via Sourceforge -> go back to paper/annotation when new term is available

EcoCyc

Blank

EcoliHub

Blank

Flybase

Blank

GeneDB

Blank

InterPro

A recently published paper describing our GO annotation process is available here: http://database.oxfordjournals.org/content/2012/bar068.short

Maize DB

Blank

MTBbase

Basic chain: every year in July, go through all papers that are listed in Pubmed as having appeared up to 1.5 years before with the keyword "Mycobacterium tuberculosis". Annotate those using the Phenote editor. Process resulting file through a Ruby script, producing the correct format, and the filter-gene-association.pl script among other checks. Zip resulting file together with Changelog etc, upload to site, where it is picked up by the EBI Uniprot-GOA group.

MGI

Flowchart

Pombase

Does this mean which data sources we use (i.e KW to GO, PAINT, function process mappings etc. and how there are integrated with the manual annotation?)

Manual annotation from the literature provides our primary annotation source.
This is supplemented by:
1. IEA data from InterPRO to GO and SPKW to GO (filtered to remove redundancy)
2. Annotation from function process links encoded in the ontology
3. PAINT annotations

Reactome

blank

RGD

Flowchart. Gene list (usually based on disease category or pathway) - manual PubMed search for all rat literature associated with target gene - reading of abstract/full paper - manual entering of data into RGD curation tool (handles multiple ontologies and multiple data types) - submission of annotation from RGD curation tool to curation database - weekly release to RGD public database and upload to GOC

SGD

Flowchart

Everyone at SGD has a slightly different method and this differs depending on whether we are doing paper-based annotation or if we are doing a complete gene review to provide the most accurate summary of GO annotations for a gene. Recently SGD has shifted from doing less of the former and more of the latter. An example process for reviewing a gene for GO is outlined below. The individual steps are performed by every curator but the order may vary.

Complete gene review:
-Read a couple of recent reviews to familiarize oneself with the field of study and the current synthesized knowledge of a gene product's functional characterization.
-Search the literature for experimental evidence and the ontologies for terms to best represent the current biological knowledge and annotate using experimental evidence codes (IDA/IMP/IPI/IGI/IEP) or sequence based codes (e.g. ISS) when that evidence is presented in the literature.
-Add additional annotations that accurately represent the biology of the gene product using the IC evidence code, when possible, by inferencing from one or more other GO terms
-Peruse the entire body of literature associated with that gene to ensure no important aspects of function/process/component were missed.
-Review existing annotations to ensure they accurately reflect the biology of the gene product and still meet current annotation standards.
-Remove annotations that are no longer supported by the literature or do not meet current annotation standards.
-Request new ontology terms when necessary to better represent the biology of the gene product. Report inaccuracies in the ontology in the relevant branches.
-Update the "Date-last-reviewed" date for the annotations.

Karen Christie 2/17/12 2:05 PM
Comment [1]: Not an experimental code

SGN(Tomato)

We import new papers every few months from pubmed. The gene names/synonyms are matched automatically and curators can go through the list on line. A curator makes basic updates to the gene page and then writes to one of the authors to assign as a locus editor.

The locus editor then takes over with the updates.

From a user interface perspective, the process is pretty much going from our gene annotation part from the top down. GO annotation is one of the last sections, and some selections are being pre-populated based on previous inputs (such as reference is based on articles associated with the locus).

SoyBase

Spasmotically, is the best term

TAIR

Flowchart

UniProt

flow chart

Wormbase

Flowchart

Zfin

blank

3A. What software tools do you use for GO annotation? Include any curation interface tools as well as paper triaging, text mining and sequence analysis tools.

Agbase

Gene Prioritization (GP) Interface. AgBase biocurators target manual biocuration using a Gene Prioritization interface that ranks genes based upon user requests or presence on microarrays.

Biocuration Interface (BI). The main BI may be accessed directly or via the GP interface, enabling biocurators to also add GO annotations for gene products that do not have a specific GP list.

eGIFT Integration. Another novel tool that we use to focus our manual biocuration effort is the extracting Genic Information From Text tool (eGIFT). eGIFT searches PubMed to identify literature associated with specific genes and associates informative terms, called iTerms, and sentences containing them, with these genes. We link iTerms to GO terms and display these in the AgBase BI. Integrating eGIFT with our BI enables AgBase biocurators to rapidly identify publications for GO annotation.

Journal Database (JDB). Like many other biocuration projects we also track the literature that we annotate. We developed the AgBase Journal Database (JDB), which we have subsequently integrated into our biocuration pipeline. A key feature for the JDB, which differentiates it from other JDBs for biocuration projects for which we are aware, is that we not only record articles which have GO annotation (which may also be parsed directly from the GO gaf) but also articles that do not have GO.

BHF/UCL

Protein2GO is the curation interface tool we use. SourceForge tracker and TermGenie for requesting new GO terms. For sequence analysis we use UniProtKB's Align and Blast tools. For orthology predictions we use the HGNC's HCOP tool as well as Homologene. We do not use paper triaging or text mining tools.

CGD/AspGD

Custom literature triage interface: We identify highest-priority papers for immediate based on species and gene name searches of PubMed; however, we manually scan the abstract of all papers that match our species-name searches and are up-to-date with curation of all but the most-recently added species in our databases. We read all relevant papers in full as part of our manual curation process, and make GO assignments manually from the full text. We use the GO curation tools from SGD.

Dictybase

- For literature topic curation, which is usually done first to categorize the paper, we have our own literature curation tool. This tool also allows linking and unlinking genes and adding papers in PubMed that were missed by our script.
- For GO annotation we use Protein2GO from the EBI. We then import these annotations biweekly into dictyBase.
- In collaboration with WormBase we have been developing a process to use Textpresso for GO component annotations. For this we will use the semi-automatic annotation tool provided by WormBase.

EcoCyc

Curation interface is Pathway Tools; no other specialized tools are used.

EcoliHub

GONUTS, UniProt, PubMed, AmiGO, QuickGO.
Skype.

Flybase

Our web-based 'Fast Track Your Paper' tool (http://flybase.org/submission/publication/) is used by authors to curate basic information from a paper (e.g. the genes studied) and add triage flags. The tool outputs a text file in the same format as our standard curation pipeline that is loaded into our database weekly with the curator generated files.

PubMed, PMC, Textpresso for fly (http://www.textpresso.org/fly/) to find relevant literature. Adobe Reader or preview to view pdfs. QuickGO or OBO-Edit to find GO terms.

Curators currently submit data via text files. The most common editor used is Textwrangler. We have a suite of custom perl scripts that are used to support curation and check curation records. This includes some GO annotation QC (e.g. to ensure that the with column is completed where appropriate and use of qualifier with correct aspect).

BLAST for ISS annotation.

We are developing a new curation interface and the first module to be developed is for GO annotation. The current version is essentially functional but requires more work before is it adopted as part of our standard curation pipeline.

We have started working with Michael Muller at WormBase to explore the use of Textpresso to make cellular component annotations. We are at stage of updating the fly Textpresso corpus and deciding on the appropriate search terms (there is a lot of ambiguity in fly gene symbols).

GeneDB

Artemis for GO annotation, interpro2go mappings, sometime pfamscan2go mappings. Pubcrawler and Zotero for literature management.

InterPro

We use our own curation tool that allows us to browse manually annotated GO terms for UniProt sequences matching an InterPro entry. We also draw upon communal resources, specifically the Quick-GO website to visualise ancestor/child terms, and OBO-Edit.

MaizeDB

The main tools we will use are Textpresso and Pathway tools. We are currently working on some internal tools for GO annotations.

MTBBase

Phenote to edit annotations.
We used Bibdesk on MacOS, and now use JabRef on Linux to keep track of annotated papers.

MGI

Quosa (www.quosa.com) for primary literature triage
OBO-Edit for ontology development and as a search tool during annotation.
Dedicated MGI GO Editorial Interface (GO-EI).

Pombase

CANTO
http://oliver0.sysbiol.cam.ac.uk/ (test)
http://oliver0.sysbiol.cam.ac.uk/pombe/ (live)
For triage and curation

Reactome

None

RGD

Currently, RGD uses only the data entry tool in the manual GO annotation process. This data entry tool was developed in-house at RGD to facilitate curation across species, data types and ontologies.

SGD

in-house curation interface, SGD web pages, PubMed literature searches and list management, Textpresso, AmiGO, OBO-Edit, QuickGO

SGN (Tomato)

We use in house developed community curation system. It matches gene names and synonyms automatically to abstracts.

Soybase

None, manual paper inspection

TAIR

We use PubSearch as our main curation system for GO annotation. PubSearch is a web-based literature curation tool that allows curators to search and annotate genes to keywords from articles. It is based on a MySQL relational database for the back-end, and Java Servlet and Java Server Pages running in a Tomcat container for the API and front-end applications. PubSearch was initially developed by the GMOD project (http://gmod.org/wiki/PubSearch); we have made extensive improvement to this tool recently with new features such as community annotation processing.

For community annotation, we developed an online author submission tool with built in term-suggestion capability:
http://www.arabidopsis.org/doc/submit/functional_annotation/123

For paper triage, we use PaperGrab, another in-house developed tool.

Text mining: 1) In our current literature curation workflow, we use an algorithm to automatically extract gene names from the abstracts. A gene-reference link is also automatically generated during this process. This is followed by a manual verification step to confirm that the gene-reference link is valid. 2) Recently, in collaboration with the Wormbase team, we have developed a procedure to automatically extract protein subcellular localization information from full text literature using the Textpresso text mining tool. In this approach, the entire Arabidopsis full text literature corpus is processed by Textpresso, sentences that contain Arabidopsis gene names, protein subcellular localization data, as well as assay related words are extracted and GO terms are suggested. A curator then manually validates each suggested annotation.

Sequence analysis tools: We use standard tools such as BLAST for sequence analysis: e.g. http://www.arabidopsis.org/Blast/index.jsp

UniProt

Protein2GO (annotation tool)
QuickGO (GO browser)
UniProt.org (sequence analysis and paper location)
PubMed
iHOP

Wormbase

Support Vector Machine (SVM) algorithm for document classification.
Textpresso (http://textpresso.org) for fact extraction.
Ontology Annotator for curation.

Zfin

Our paper triage process is a manually executed by a single dedicated literature analyst using no text mining or other assistance.
Our GO curation interface is custom built in Java on one tab of a tabbed curation interface.
Enforcement of many business rules starts in the UI of the curation interface. Many options in the curation interface are context sensitive to only offer valid curation options, like only allowing the 'contributes_to' qualifier with a GO process term.
Different curators use different sites to gather more info about terms of interest or annotation policy. These include AmiGO, QuickGO, and the OLS for more detailed examination of a potential GO term for an annotation, and the GO evidence code documentation page for help determining the proper evidence code. Robust term searching, ontology navigation, graph views etc in these sites are very helpful.

3B. How do you identify and prioritize which papers, genes, etc. are targeted for GO annotation?

Agbase

Our policy is to identify all literature for a target gene product and then triage for papers that contain functional information. We are now switching to using the eGIFT text-mining tool to help us identify & triage papers.

BHF/UCL

We have a list of 4000 genes that are considered relevant to cardiovascular processes. At the start of this initiative we prioritized 250 genes and worked through these on a gene-by-gene basis. Since then we have recognized the value of annotating a specific biological process, these processes (and often the genes involved) are chosen following discussions with scientific experts.
We also receive requests to annotate specific genes which are located close to specific risk associated SNPs, and in these cases we will provide a brief summary of the gene and how it's function may explain it's association with the risk trait/phenotype.
Our current annotation targets are heart jogging (11 proteins identified) and cardiac conduction (100 proteins identified).
Our annotation of each protein will vary according to the volume of literature describing the gene. If there are 10-40 papers describing mentioning a specific gene we will annotate all of those that we can. However, it will be obvious from the title of many of these papers that there will not be experimental evidence in the paper, for example they will be discussing SNPs/risk, expression data, eg association of gene expression with specific cancer type. If there are more than 40 papers, we will filter using specific keywords to limit the papers retrieved to those describing the process we are focused on annotating.
PubMed searches for papers; sometimes filtering on 'human' or the specific biological process we are interested in annotating.
When annotating on a gene-by-gene basis we occasional use iHOP,or GeneRifs to check all

key functions/processes captured, if there is a very large volume of publications.

CGD/AspGD

We read and curate all of the papers that contain gene-specific information for each of our organisms of interest. For each species, PubMed records that match gene names that we already have in the database are flagged as highest priority and curated first. We carry no significant literature backlog for all but the most-recently-added species that we curate.

Dictybase

Typically newest papers are then selected to annotate, unless a gene is underannotated, in which case older papers will also get annotated. See Question 3D for more info.

EcoCyc

Genes, papers aren't targeted for GO annotation per se. However, gene products with outdated "regular" annotation are targeted for updates; these usually coincide with gene products with no or inadequate GO annotation.

EcoliHub

a) Students choose any protein in UniProt for which they can find published experimental or sequence analysis information. BM does not control what proteins the students choose.
b) Some instructors focus students on a pathway (i.e. nitrogen cycling, retinal development, etc)
c) Some students choose to focus on a particular topic (i.e. viruses, jellyfish toxins, etc).

Flybase

As described in 2E, the 5 genetic literature curators work strictly on a paper-by-paper based and prioritisze papers based on the number of triage flags they receive - all papers curated are used to make GO annotations were appropriate.

The specialist GO curator prioritizes genes by a variety of criteria and a mixture of these methods is used in parallel. Where practical, each gene looked at is also signed off as 'completely annotated'.

1. Current annotation status
   Genes that have no GO annotation for any aspect of GO
   Genes that are missing GO annotation for one or two aspect of GO
   Genes that are only associated with root terms but have publications linked to them in FlyBase
2. New gene annotations, splits, renames
   Each release the newly annotated genes are given GO annotation and genes that are split into new gene models have GO annotation reviewed. Genes that have been renamed in the previous month are reviewed - gene symbol changes from a numerical placeholder to a real meaningful name often indicate a new functional information and it is a good opportunity to check this has been fully captured by the curator that recorded the symbol change.

3. Reference genome targets
4. Genes associated with human diseases - this is in our MRC grant but very little effort has been devoted to this to date.
5. Genes with errors or deficient GO annotation brought to our attention by users

Susan also prioritizes GO annotation by process - e.g. following an ad hoc new GO term request she tries to assign the terms to the relevant set of genes rather than just the gene that prompted the request. Similarly after the kidney term development workshop we tried to find genes to apply the new terms to.

Having identified a gene of interest, suitable papers are identified by a number of strategies:
By review of papers already associated with the gene in FlyBase - these may have been curated in the 10 years of FlyBase before GO annotation was introduced, or GO annotation may have been overlooked.
Search PubMed - review title and/or abstracts for good curation candidates - typically start with the most recent relevant paper and follow up key references cited.
Full-text searches of PMC and fly Textpresso to see the context of the hits within the full text - good for the more obscure references to uncharacterized genes.

GeneDB

We triage publications according to how much time we have, and user comments are considered of higher importance than publications for the most part.

InterPro

Our priorities are determined by the signatures we receive from our member databases. Where there are a large number of signatures, we prioritise the ones that match well-defined families where there is published information available to aid our annotation. We also try to prioritise signatures that match sequences not already covered by InterPro, or that add increased functional specificity (eg, signatures that represent functionally specific subfamilies of large, functionally diverse protein families).

MaizeDB

We currently have an editorial board which chooses a set of papers each month. These papers get annotated into our database. Genes are done on an individual basis or as bulk if it is provided to us. In the very near future we will be adding papers to support BioCyc maize annotations for specific pathways.

MTBbase

Everything up to the PMID number xx000000 has top priority, where xx is at the moment equal to "20". From July this year, this will change to "21". Since microarrays have proven deceptive for M.tb. only such results are included that are confirmed through other experimental means. Modelling papers are excluded too.

MGI

Paper are manually selected for GO during the MGI triage process; the relevant genes are manually associated to the papers with some software aid

(ProMiner).

Pombase

Papers are prioritised during triage but procedures will change when community curation is launched. At this point curators will concentrate on older papers and new papers will be assigned to the community.

Reactome

We have a prioritized list of biological processes developed in collaboration with outside expert biologists. That drives our selections of human proteins and then of GO terms.

RGD

See flowchart overview under #3. In short, RGD's GO curation is gene-centric. In general, genes are prioritized on the basis of suspected association with a targeted disease category or known involvement in a targeted pathway. Papers are selected for curation at the discretion of the individual curator, based on targeted PubMed searches.

SGD

Curator triage and internal flags. We also use a combination of Date-last-reviewed date and presence of TAS annotations to prioritize which genes should be targeted. Please also refer back to the curation pipeline flowchart.

SGN

We have a list of prioritized journals that we go through first.

Soybase

We do not discriminate.

TAIR

Published articles with Arabidopsis in the title, abstract or keywords are collected on a monthly basis. A subset of these articles are chosen for curation based on whether the article includes the first characterization of a novel gene.

UniProt

UniProt curators annotate on a protein-by-protein, or protein family basis.
Relevant papers are frequently identified by carrying out a PubMed search.
UniProt concentrates its manual annotation effort on entries from model organisms. We aim to provide high quality annotation for representative members of all protein families across diverse taxonomic groups.

Wormbase

Genes are picked for annotation from the following groups:
1) Reference genome project
2) Human disease gene orthologs
3) Newly published papers describing significant advances in *C. elegans* biology
4) Semi-automated pipelines for Cellular Component Curation (CCC) and Molecular

Function annotation, specifically physical interactions, enzymatic activities, and transporter activities.
Note that the semi-automated pipelines are designed to capture specific types of information from papers and therefore all GO annotations from these papers may not be captured at the same time.
5) Genes curated for phenotype by WormBase phenotype curators may also receive Biological Process annotations by a semi-automated Phenotyp2GO pipeline that maps WormBase Phenotype Ontology (WPO) terms to GO Biological Process terms.

Zfin

We generally do no prioritize papers OR genes for curation because of GO. The exception is when the GOC has asked groups to focus curation effort on a specific set of genes. In that case I would get a listing of all the papers in ZFIN associated with the target gene, and start with the oldest papers first typically. Based on the title and abstract, I would decide if the paper was likely to have GO. If it seemed likely, I would then scan the figures and methods. If it still looked promising I would curate the GO from the paper. Barring a few literature heavy genes, I was able to deal with all the papers associated with the target gene in this way. Many of the newest papers would be curated as part of our regular curation process in time.

3C. Do you regularly make both literature and inferred (e.g. ISS, IEA) annotations?

Agbase

Yes. If there are no papers (happens for ag species!) we will sue ISS before we make any NDs and call the annotation complete.

BHF/UCL

We never make IEA annotations.
We often make ISS annotations, as many of the key experiments for cardiovascular processes are carried out in model organisms. We also make TAS and NAS annotations to capture biological concepts that are known in the literature, but where we are unable to annotate all the papers, due to time constraints and the vast amount of literature that human genes often have.

CGD/AspGD

Yes, we do.

Dictybase

Currently curators usually do literature-based annotations, and ISS only from paper reference. We have legacy ISS annotations inferred by sequence similarity, but are currently not adding new annotations of this type. We will import IEAs on a biweekly schedule from GOA.

EcoCyc

Usually only experimental annotations are made.

EcoliHub

Strictly literature and only a subset of literature-based annotations are permitted (no EXP, IPI, IC, TAS, or NAS). Students may only make annotations using IDA, IMP, IGI, IEP for experimental evidence or ISA, ISO, ISM or IGC for published sequence analyses. They are never allowed to make annotations using IC, TAS, NAS. They do not ever make IEA annotations.

Flybase

Yes but only the specialist GO curator carries out sequence analysis to make ISS annotations; the regular genetic literature curators only make ISS annotations supported based on data in research papers.
We incorporate IEA annotations via InterPro2GO mappings - these are updated every FlyBase release (~6-10 times per year)

GeneDB

We try to avoid inferred annotations for individual published results, but will use it in annotation transfers. We include IEA mappings as part of the interpro2go mapping process.

InterPro

All our annotations require at least some level of experimental evidence (ie, literature based), although we use inferred annotations to guide our curation (when we're mapping GO terms to InterPro entries that cover a large number of sequences, only a subset of which are experimentally characterised, for example).

MaizeDB

Yes, we will.

MTBbase

We don't do ISS/IEA except the author of the experimental paper does.

MGI

We focus primarily on literature-based annotation. Inferred annotation is mostly done through automated loads. These loads include IEA generation during UniProt loads, and the bulk of our ISO annotations (generated automatically from rat and human experimental annotation). A very small amount of manually curated annotation using ISO, ISS, or ISA does occur.

Pombase

Curators can make ISS annotations to experimentally characterised orthologs for unstudied genes.

RGD

RGD imports IEA and some ISS GO annotations via automated pipelines from GOA. ISO annotations are assigned to RGD's rat genes during import of (manually curated, experimentally derived) mouse and human GO annotations based on known orthology between rat and mouse or human genes.

Reactome

Yes – the former as a part of curation; the latter as a script-driven part of our quarterly release process.

SGD

SGD primarily makes experiment-based annotations. We will use ISS if the evidence is in a peer-reviewed publication but do not do any in-house alignments. We do not make any electronic annotations ourselves. We integrate computational annotations from UniProt and other external groups.

SGN

blank

Soybase

We do not infer annotations by ISS/IEA currently, we use community/published annotations.

TAIR

yes

UniProt

Yes. Both ISS and IEA.

Wormbase

Yes, WormBase makes both ISS and IEA annotations. ISS annotations are made manually, while IEA annotations are newly generated during each WormBase database build by in-house use of InterProScan.

Zfin

Our IEA annotations come from local application of interpro2go, ec2go, spkw2go, and the inferred annotation GAF file produced by GOC. ISS annotations are made rarely by curators, and generally based on a statement made by authors.

3D. Do you conduct gene level review to support comprehensive annotation for each gene, in contrast to proceeding paper by paper without overall assessment of status for the gene?

AgBase

I don't understand what you mean by 'gene level review' and 'assessment of status for gene' – I guess that must mean that we don't do this.

BHF/UCL

Yes, we often work in a gene-centric manner to ensure that all the concepts known about a particular gene are captured. We rarely add more than 2-3 of the same GO term to a protein record, (unless we are capturing protein binding, and then the target in the 'with' column will be different between each annotation). This means that a paper annotated late in the annotation process for a gene may not have all experimental data captured, with only new information annotated.

If the introduction to a paper suggests that a gene is involved in a specific process/function at a specific location, and this information is not already captured in the GO annotations for this gene, then TAS/NAS annotations will be created if literature searches do not easily

provide experimental data to support this (e.g. species in referenced paper not mentioned and not traceable). Ruth may also create these TAS/NAS annotations if the gene is not on our list of cardiovascular relevant genes, or is not on our list of genes to annotate to a specific process.

CGD/AspGD

We curate paper-by-paper, but the process is manual and curators do assess the annotations in the context of the entire body of curation (not just GO) that is associated with each gene.

Dictybase

We curate papers with a balance of curating new papers, but also working on a backlog by curating older papers attached to the gene we have a new paper for and that have good experimental data. When all of the papers relevant to a gene have been reviewed we consider the gene “comprehensively” annotated.

EcoCyc

The usual mode of curation is gene-level review.

EcoliHub

No. CACAO focuses on peer reviewed literature for proteins only.

Flybase

Yes but only the specialist GO curator takes this approach; all other GO annotation is carried out on a paper-by-paper basis without overall assessment of the gene.

GeneDB

blank

InterPro

This isn't really applicable for InterPro annotation.

MaizeDB

We currently support gene level review

MTBbase

Not within GO annotation worktime. However, such review is part of reactome work.

MGI

In general, outside of consortium assigned workflows, we proceed on a paper by paper basis, sometimes guided by various QC reports (i.e., "genes with no GO annotation but have papers selected for GO but not used", etc.).

Pombase

We implement/ will implement gene level review.

RGD

No systematic gene level review is used to assess comprehensive annotation. However, every gene that is curated for disease, pathway, or another project is "reviewed" for comprehensive GO annotation by the individual curator "on the fly".

Reactome

Gene-level review. Literature references are drawn in only as needed to provide evidence – providing comprehensive coverage of literature is explicitly not a curation goal.

SGD- Yes

Soybase

currently no

SGN

Usually, the initial focus is a paper. However, this changes when we associate a locus to a locus editor, who will then usually fill in on a locus basis.

TAIR

For genes selected for Reference Genome project-related annotation, a gene level review of literature is conducted. For the rest of our annotation targets, we work strictly paper by paper with no gene level review.

UniProt

The specialized UniProt-GOA curators do review the comprehensive annotation level for proteins that they have prioritized for annotation.

Wormbase

During fully manual curation, we mostly try to be comprehensive or at least curate the key non-redundant information for each gene. However, when curating with the semi-automated pipelines, or when curating papers describing significant new results, we may only curate the specific information described in a particular paper.

Zfin

No, we do the later, paper by paper approach. Each curator curates the entire paper, including GO. No one specializes specifically and exclusively in GO. It is part of my job as the lead GO curator to keep us up to date with GO policies and assist with GO annotation questions at ZFIN. Not having a dedicated person or team for GO creates a difficulty when we begin to talk about prioritizing GO curation in any way. In the past, I have taken on that role and done the gene-by-gene curation for gene-centric GO curation efforts that came out of the GOC. It is noteworthy that our grant through Spring 2016 does not include any major push or focus in the area of gene functional annotation specifically. As you are aware, when the GO asks MODS to meet certain GOC milestones or curation goals, there is a conflict of resource distribution when those GOC goals are not also part of the MODs primary aims as stated in their grants. I don't have a good suggestion for how to address that issue.

3E. What is your process for creating and maintaining an updated gp2protein file?

AgBase

We are using a combination of UniProtKB & NCBI proteins for chicken. I expect that having a reference set will help enormously.

BHF/UCL

We leave this to UniProt-GOA

CGD/AspGD

The CGD gp2protein file was last updated in November 2011 (with updates to some accession ID's), but the actual mapping is several years old and currently contains C. albicans mappings only. We plan to update the analysis but have not determined a timeframe for doing so, and we have not yet generated a gp2protein mapping for AspGD.

Some procedural details from the README:
http://www.candidagenome.org/download/External_id_mappings/README
"BLAST analysis was performed to map sequences from each of the external database resources to CGD features. We first downloaded all sequences from Uniprot by querying the database with an organism specific query for 'Candida albicans'. Similarly, we downloaded all sequences from Entrez Nucleotide database with an organism specific query for 'Candida albicans SC5314'. Then, we performed BLAST comparisons for each of these sets of sequences against the haploid set of Assembly 19 sequences.
The following strict thresholds were used to ensure good quality matches: i) E-value threshold < 1E-5;
ii) Percent of query sequence in the alignment = 100%; iii) Percent of matching sequence in the alignment = 100%; iv) Percent identity of best HSP = 100%.
For the sequences that could not be mapped to CGD genes using the first step as explained above,we ran another BLAST analysis of those sequences against the diploid set of Assembly 19 sequences. At this step, we were able to find additional mappings to allelic genes. The same strict thresholds were used for this run of BLAST analysis as well. Using the above procedure, we were able to map 7481 Uniprot sequences and 11517 RefSeq Nucleotide sequences to CGD genes."

Dictybase

The dictyBase gp2protein file contains mapping between dictyBase GeneID and UniProt and if absent a GenBank protein id is supplied instead. The identifier correspondences are generally updated en masse from GenBank and UniProt after every GenBank release. Between every GenBank submission, weekly batch scripts refresh the UniProt mapping in our database and create a gp2protein file for GOC submission.

EcoCyc

EcoliWiki handles that aspect of the work.

EcoliHub

The *gp2protein.ecocyc* file is created each month (or when submitting a new gene_assocation file) by mining the accession tables on Gene Product pages in EcoliWiki for EcoCyc identifiers as well as UniProt entry-names/accessions. (An entry-name points is a biologically relevant identifier, an accession is a stable id that points to a particular version of a file/entry.) In cases where one is not known or missing, the product is skipped. Differences in database layout and nomenclature have previously led to problems with the *gp2protein.ecocyc* file. Mapping EcoCyc identifiers onto EcoliWiki gene products has not always been straightforward. More recently, the file was created with BioPerl using identifiers directly from EcoCyc and RefSeq downloads, but this also proved to be an problematic due to problems with the RefSeq file.

The gp2protein file was typically maintained by Daniel Renfro.

Flybase

The specialist GO curator uses a perl script to parse the info from our precomputed file of IDs. Unfortunately we have had long-standing problems with our pipeline to keep UniProtKB accessions in sync with FlyBase and this has resulted in neglect of this file. However this issue has very recently been resolved and the next release of FlyBase (March 2012) will reflect this. An updated gp2protein file will be submitted along with the GAF each new release of FlyBase.

GeneDB

I'm not sure we have gp2protein mappings for T. brucei or L. major. For P. falciparum, it was updated when the genome was last updated in EMBL and the UniProt IDs also changed.

InterPro

N/A

MaizeDB

We have not discussed this yet.

MTBbase

We don’t

MGI

Our gp2protien file is automatically generated daily using the identifiers for "representative protein" as determined by the MGI sequence group workflow.

Pombase

gp2protein pipeline will change over to new PomBase shortly and regular (monthly) updates resumed

RGD

RGD has an automated software pipeline which extracts the applicable gene and protein ID information from the RGD database and writes the gp2protein file on a weekly basis.

Reactome

A script that is run at release time.

SGD

We routinely update the mappings of SGDIDs to UniProt and other external IDs and we use these mappings to produce the gp2protein file.

SGN

We don't provide a gp2protein file.

Soybase

We have none

TAIR

A Clover ETL script reads data from the PubSearch database and reformats it to GAF specs. We run this with each genome release.

UniProt

This is not such an issue as we directly annotate to UniProtKB entries. We do however support other groups in maintaining gp2protein files that reference primary UniProtKB accessions.

Wormbase

Currently, the gp2protein file is generated by a script run manually by a GO curator. We would like to changes this pipeline such that the gp2protein file is generated as part of the WormBase database build, and expand it to include other nematode species for which WormBase has genomic data, so we can begin to send at least IEA annotations for other nematodes to the GOC.

Zfin

Approximately monthly we run a "UniProt load" script which updates data we obtain from UniProt, including UniProt protein IDs, and associates them with gene records in ZFIN. Each weekend, our system generates a new gp2protein file containing a single UniProt ID for each gene in ZFIN. Gene records that have no UniProt ID show up in the file as well, but they lack any protein identifier. Between uniProt loads, we see typically a small number of secondary or invalid UniProt IDs cropping up. We let the UniProt load correct those. So at any time, our gp2protein file may contain a small number of these secondary or invalid UniProt IDs.

3F What is your process for creating a GAF file for submission to the GOC?

AgBase

We don't submit directly. Don't have access and need to submit via EBI Protein2GO, so we are really behind here particularly now that we have our own QC in place.
However we periodically (approx. every 2 months) pull all our annotations form the AgBase Biocuration Interface, do QC and generate gaf 2.0 files.

BHF/UCL

We leave this to UniProt-GOA

CGD/AspGD

We update our GAF's weekly, for both CGD and AspGD. We submit them to the GOC via GO CVS.

Dictybase

In our existing pipeline, a weekly batch script writes the GO annotations to a GAF2.0 file from our database. We are working toautomate the biweekly import of our annotations from GOA. In that process, our GAF export script will run after our GOA import to assure up-to-date information in our GAF submission. Our export script also dumps our entire annotation set even those which are not handled by protein2go, such as ncRNA annotations.

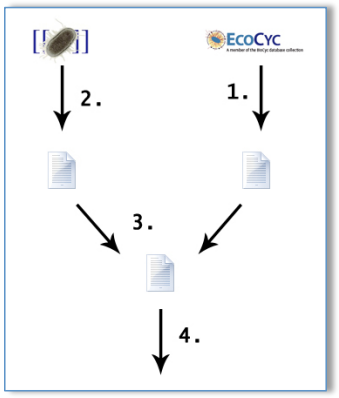
EcoCyc

EcoliWiki handles that aspect of the work.

EcoliHub

The *gene_association.ecocyc* file is prepared by combining the annotations from EcoCyc and from EcoliWiki.

EcoCyc releases their data quarterly, EcoliWiki does not have releases since it is a wiki and is (conceptually) always in the newest revision. Each month the annotations from latest EcoCyc release on record are parsed and validated using Mike Cherry's scripts (step 1.) These are then combined with the validated EcoliWiki annotations (step 2.), and validated once again to resolve nomenclature issues that arise between the two databases (step 3.) This file is then submitted via CVS to the GOC (step 4.) Although this process sounds straightforward, it is anything but.[i]



Flybase

Custom written script is run by developers at our Harvard site and then submitted by the specialist GO curator. Typically a new GAF is submitted to GO every time a new release of

FlyBase is made public (~6-10 times per year) but we now make a new GAF on a weekly basis in case we need to update more regularly (e.g. concurrent annotation projects).

GeneDB

In theory at least, we dump out the gene association file from Chado using a custom script of ours, use Mike Cherry's filter-gene-association.pl script to validate them then submit them to GO's cvs repository. For a variety of reasons, this hasn't happened for a while!

InterPro

N/A

MaizeDB

We have not discussed this yet. But we do supply files to PO which are similar format.

MTBbase

Conversion from Phenote TAB output, consistency and correctness checks using Ruby and Perl scripts.

MGI

Our GAF file is automatically generated daily using the data contained in the MGI database. During this process, column 7, if not populated directly, is automatically populated with the representative protein, transcript (for functional RNAs) or gene ids. Cell type data if present is added to column 16.

Pombase

GAF file export will change over to PomBase pipeline shortly and regular (monthly) updates resumed.

RGD

RGD has an automated software pipeline which extracts all GO annotations for rat genes from the RGD database and writes the GAF on a weekly basis.

Reactome

A script that is run at release time.

SGD

This is generated by an in-house script.

SGN

A dump script dumps the info from the database in the right format and is then submitted to the repository.

Soybase

We have none

TAIR

A Clover ETL script reads data from the PubSearch database and reformats it to GAF specs. We run this every week.

UniProt

1. Integration of external annotations (involves syntax checking, ID mapping to UniProtKB accessions, filtering of redundant or undesirable annotation statements)
2. Running or newly importing annotations from IEA pipelines
3. Creation of inferred annotations from inter-ontology links
4. Creation of a UniProt GAF, GPAD and GPI files, as well as species-specific subsets, using the UniProt Complete Proteome sets.
5. All files created monthly
6. Human, chicken species-specific files created fortnightly

Wormbase

This is a manual procedure that consists of a manual concatenation of files from three different pipelines--our manual annotation file, and 2 other files generated by automated/semi-automated methods—Phenotype2GO annotations and IEA annotations.

ZFin

Each weekend, our system generates a new GAF file. My local GO CVS is updated and the new GAF file is checked with the GOC filter script. Errors are addressed in our database until a GAF is generated that passes the GOC filter script. The clean GAF is then checked in to the GO CVS.

3G What types of references/sources do you cite for your annotations? Are there other types of information source that you would like to use for GO annotation?

AgBAse

PubMed predominantly with some Agricola & DOI. DOI are flagged during our QC so biocurators can check for PMIDs.

BHF/UCL

We currently use PMIDs. We cannot think of any other information sources that we would want to be able to annotate.

CGD/AspGD

We cite the primary literature from which we make the annotation.
For inferences made by us at CGD/AspGD, we cite an internal reference that describes the procedure that we use in detail:
- CGD Prediction of Gene Ontology (GO) annotations based on orthology:
http://www.candidagenome.org/cgi-bin/reference/reference.pl?dbid=CAL0121033
- CGD Prediction of Gene Ontology (GO) annotations based on protein characteristics (e.g., domains and motifs):
http://www.candidagenome.org/cgi-bin/reference/reference.pl?dbid=CAL0142013

- AspGD Prediction of Gene Ontology (GO) annotations based on protein characteristics (e.g., domains and motifs): http://www.aspergillusgenome.org/cgi-bin/reference/reference.pl?dbid=ASPL0000166200
- AspGD (2011) Prediction of Gene Ontology (GO) annotations based on orthology: http://www.aspergillusgenome.org/cgi-bin/reference/reference.pl?dbid=ASPL0000000005

Dictybase

Most manual dictyBase annotations cite PubMed references (70%). We rarely need to add a non PubMed reference, because some older Dictyostelium papers are not in PubMed. GO references are cited for ND annotations and for ISS annotations (orthologs with manual GO in the 'with')

EcoCyc

We usually cite publications that have been indexed in PubMed. On occasion, some older paper does not have a PubMed ID, though. Could there be a way to send such references to PubMed, so they can be indexed and more easily cited?

EcoliHub

PMIDs only. No other sources are permitted for CACAO GO annotations.

Flybase

PMIDs, GO_REFS. We have legacy annotations to source such as meeting abstracts and DNA/protein sequence entries that we are gradually replacing. UniProtKB/InterPro GO annotations are also used as sources.

GeneDB

blank

InterPro

N/A

MaizeDB

We will cite literature and other sources on an individual basis.

MTBbase

Full papers are the main source in the annotation process. Additionally, we use abstracts where both gene/product and its characterization plus the method are clearly stated. Finally, BRENDA summaries of enzyme function serve where the paper is not free after 1.5 years. See the GO mailing lists discussion on usage of early microbiology enzyme characterizations for an unused but needed source of annotations.

MGI

We supply PMID for all literature-based annotations. Annotations made from

data loads (IEAs and most ISOs) are referenced to specific MGI procedural references (which have their equivalents in the GO Reference Collection).

Pombase

PMID and GO_refs

RGD

RGD uses references from PubMed to make GO annotations. We are not aware of any other sources of information which would give us additional usable data.

Reactome

Research publications from PubMed (preferred source; used in almost all cases), books and web resources (URLs) are used in a small number of cases when they provide high-quality data that are not otherwise available.

SGD

All of our manual and high-throughput annotations are from published literature.We integrate PAINT and IEA annotations with GO_REF.

SGN

mostly pubmed, but we find a lot of articles are not in pubmed. We load them into SGN manually.These are difficult to standardize though.

Soybase

Since we have not inferred our own annotations, we cite the contributors

TAIR

published literature

- GOC reference genome project (PAINT), e.g.
http://www.arabidopsis.org/servlets/Search?action=search&type=annotation&tair_object_id=2173862&locus_name=AT5G55310
reference:
http://www.arabidopsis.org/servlets/TairObject?type=communication&id=501741973

- TIGR Arabidopsis annotation team, e.g.
http://www.arabidopsis.org/servlets/Search?action=search&type=annotation&tair_object_id=2200960&locus_name=AT1G01040
reference:
http://www.arabidopsis.org/servlets/TairObject?type=communication&id=501714663

- For annotations coming from UniProtKB, IntAct and BioGRID, the annotations are attributed to these annotation groups but the references are the publications on which the annotations are based.

UniProt

PMIDs, DOIs, GO_REFs

Wormbase

We primarily cite published papers and also some of the references created by the GO consortium (GO_Refs) for annotation.

Zfin

We cite published literature as well as internal publications that describe various annotation methods. The internal pubs would be in the GO_REF set.

3H. Do you currently provide annotations that include the ANNOTATION_EXTENSION or GENE_PRODUCT_FORM_ID information? If not, have you any plans to start releasing such information in the next 12 months?

AgBase

Yes. Most particularly the extension.

BHF/UCL

We have been adding notes to a free text field in preparation for the availability of this facility. Majority of these so far have been cell/tissue types and protein/ChEBI targets of a molecular function.

CGD/AspGD

No, we do not, and we do not currently have plans to do this.

Dictybase

yes, Recently started ANNOTATION_EXTENSION as available in protein2GO.

EcoCyc

No, and probably not.

EcoliHub

No and probably not.

Flybase

We do not provide annotations that include either of these types of information. We do not have any plans to start releasing such information in the next 12 months.

GeneDB- Blank

InterPro

No, No

MaizeDB

We currently do not. We will consider this. We have done this for Plant Ontology

MTBbase

No, and not yet.

MGI

As indicated above, we currently supply gene_product form when known, as either UniProtKB or Protein Ontology (PRO). In some cases a RefSeq or other protein id is used, depending upon the paper. Although we tract anatomy (adult (MA) and embryonic (EMAP)), cell type (CL), modification (PSI-MOD), target, dual taxon, and extended evidence (ECO), we currently only supply cell type data in column 16.

Pombase

ANNOTATION_EXTENSION already included, GENE_PRODUCT_FORM_ID will be included shortly (within 12 months)

RGD

RGD has no current plans to add that extra information to GO annotations.

Reactome

No

SGD

We do not provide these two types of information in our GAFs. We haven't started curating col-16, 17 data and hence very unlikely that we will have this data in the next 12 months.

SGN

NO

Soybase

Don't have the slightest idea of what you are talking about here

TAIR

ANNOTATION_EXTENSION: no (none planned in the next 12 months)
GENE_PRODUCT_FORM_ID: yes

UniProt

Yes, we currently supply both annotation_extension and gene_product_form_id information.

Wormbase

We have begun to collect data to use in the Annotation_Extension column and will begin to use it, and the Gene_Product_Form_ID, in 2012.

We are still in the process of deciding what type of information we would like to collect for

Annotation_Extension. For WormBase, some of the information that could potentially be collected as an Annotation_Extension is redundant with other curation pipelines. For example, we have a separate curation pipeline for expression pattern data that includes curation of anatomy terms, cell types, and life stages, all of which could be included as part of Annotation_Extensions. We will thus proceed by assessing what additional information Annotation_Extension could provide that is not yet represented in WormBase (e.g., enzyme-substrate relationships) and perhaps how we could mine existing WormBase data to populate Annotation_Extension. We will carefully consider balancing the time/effort of this work with the potential benefit this would bring to the user community.

ZFin

We do not currently curate this data, and we have no plans to expand in that direction at this time.

3I. Has your group decided not to produce or release any of the supported types of GO annotations? (e.g. IEA or IEP-evidenced annotations, protein binding, column 16) and if so, why?

AgBase

No

BHF/UCL

We try to keep the number of IEP annotations to a minimum, these are only used if there are no other papers with direct experimental evidence to support the annotation and the involvement of the gene product in the process, of for example development of specific tissue, is considered by experts in the field as highly likely.

CGD/AspGD

We use IEP sparingly, but we do annotate with it on occasion. We use IEA routinely, for orthology-based and domain-based inferences made via our automated pipeline (described above). We have not populated Column 16, primarily because of limited resources and other priorities. We are likely to reduce the protein binding annotations in the future, and use the IntAct editorial tool to create protein binding annotations instead.

Dictybase

No. We started annotating column 16 recently and will release those annotations soon.

EcoCyc

The Pathway Tools curation interface limits the kind of GO information that can be captured. As far as I know, there are currently no plans for extending the interface.

EcoliHub

We do not produce annotations to the following evidence codes because our focus is entirely on curating peer-reviewed literature:

a) EXP - too broad & students must identify HOW the experiment was done & annotate to a child of EXP (IDA, IMP, IGI or IEP only).

b) IPI - too easy to "game" our system with binding annotations (i.e zinc binding). Goes with the rule that students cannot annotate high throughput papers.
c) ISS - they have to pick ISA or ISO for a published sequence analysis
d) IBA, IBD, IKR, IRD, RCA - too hard to teach to use correctly in limited time when Bren hasn't really worked with these anyways.

TAS, NAS, IC, ND – for obvious reasons having to do with the objectives of CACAO, students cannot annotate to these. They have to find the papers that demonstrate the attributes of the protein.

Flybase

Annotations directly to protein binding are now added rarely (but we do still use the specific child terms). This is because the information content is relatively low compared to other terms and there is a redundancy of effort with our new curation of physical interaction data (this is carried out by our Harvard curators).

We have decided not to produce any column 16 data at present. While we would like to be able to add this refining information, any such additions to the data we capture involves a significant amount of developer time and also puts extra burden on already stretched curators. At present we cannot justify prioritizing an extension to GO annotation over other additional data types that have been requested by our users and SAB (e.g. protein interactions, modENCODE data, links to human diseases) but this decision maybe subject to review pending our next user survey. We have certainly not ruled out making annotation extensions in future. Susan has been making internal notes that could be used to make col16 entries and has participated in the discussion regarding the contents of this column. Once there is formal documentation and final decisions on the appropriate relationships it will become feasible to encourage the other curators to record this data (where time permits).

Similarly we do not have plans to include GENE_PRODUCT_FORM_ID in the near future.

GeneDB

Blank

InterPro

No, although we are working with the GOA project to remap our protein binding annotations, making them more specific where possible.

MaizeDB

No, we currently have not decided to exclude any supported types of GO annotations.

MTBbase

No pressing need (UniProt does IEA for M.tb).

MGI

MGI does not support the use of the IEP evidence code. We feel that it is very difficult to support a functional or process assertion for a gene product based on differential expression of that gene product, as direct cause and effect is hard to establish.

Pombase

We do not make new annotations using TAS or NAS but legacy data is included (why, all experiments should be experimentally or ISS supported)
We filter IEA annotation and submit non redundant IEA annotation and hope that this component will move to zero within the next couple of years.

RGD

RGD has made no decisions to discontinue release of any supported types of GO annotations.

Reactome

No

SGD

SGD only captures annotations from published literature.
Protein binding: We capture our interactions through a separate curation system specific for interactions (bioGRID)
Column 16: We do not currently have the infrastructure to capture column 16 data
We use IEP for BP annotations only.

SGN

We show everything on the website, but we only submit non IEA annotation to GO.

SoyBase

No

TAIR

We do not have the infrastructure (database, curation software features) nor the resources to modify our existing infrastructure to accommodate ANNOTATION_EXTENSION information.

UniProt

We do not integrate or create ISM-style annotations. We rely on InterPro2GO for predictions from sequence models. Using both methods could result in different versions of the same domain model being used.

Wormbase

No, we do not exclude any particular type of annotation.
On a related note, we have a limited number (104) of IEP annotations. These annotations are likely good candidates for 'chain of evidence' annotations, as we typically consider the nature of the gene product, as well as the process affected, before making an IEP annotation.

Zfin

We do not currently support the "colocalizes with" qualifier. It was never clear to me precisely when to use it, and it seemed like a low usage item in our domain. No one here has been asking for it, so for the sake of simplicity in our interface it has not been included. We also currently only use the original basic set of evidence codes. Though more evidence codes may be more specific, in this case less is more I think. Who wants to spend time looking up yet another bit of data in yet another large ontology to be sure they've got the right one?
TAS and NAS evidence are not used at ZFIN, though we may have some older annotations with these evidence codes.

3J. What are the obstacles to (or rate-limiting steps in) your GO curation and submission pipeline, aside from curator resources?

Agbase

Moving the annotations from AgBase to GOC – nobody's fault but my own as I haven't had time to make this a priority. Bad Fiona.

BHF/UCL

Finding papers with species specific information, often tracing a series of references to identify the species of the gene, occasionally deciding that the papers with the 'evidence' are so far from the paper being annotated that it is not appropriate to annotate with EXP evidence codes. Notes about the trail followed to identify the source of the species are usually made in the Protein2GO notes field.

It would be useful for TermGenie to send an email to each annotator summarizing the requests they have made. This could be along the lines of SourceForge, where each term generates an email, or as a monthly summary of the requests each curator has made. Due to the time lag between making the request in TermGenie to being able to select the term in our curation tool Protein2GO, this would help to speed up the annotation process.

An additional notes field in TermGenie allowing the curator to included the gene product ID and the PMID they are requesting the term for would be great. As often 3 regulation terms are requested in one submission the text field would also need room to indicate which of the requested terms would be associated with the gene/PMID.

At present we use the Protein2GO specific SourceForge request free text field which enables the curator to note down the annotations that need to be added to the protein once the term is created. However, if multiple proteins will be annotated with the new term then this is not the most efficient approach. Protein2GO will also generate a report when requested so that the curator can go back to these comments and check if the term is available to annotate to, which links to the protein record which needs to be annotated. Which is very useful.

Q9ULV1	SourceForge Pending	24 Jun 2011	2 updates: 1. update GO:0016055 Wnt receptor signaling pathway with norrin signaling pathway IDA PMID: 17955262 and repeat for NDP Q00604 and LRP5 O75197 2. update GO:0060070 canonical Wnt receptor signaling pathway with norrin canonical signaling pathway IDA PMID: 15035989 and repeat update to NDP Q00604 and LRP5 O75197
Q61088	SourceForge Pending	24 Jun 2011	update GO:0060070 canonical Wnt receptor signaling pathway with norrin canonical signaling pathway IDA PMID: 15035989 and repeat for mouse NDP P48744 and LRP5 Q91VN0
Q6P3W7	SourceForge Pending	24 Jun 2011	add New term GO:0038012 receptor internalization involved in negatively regulating canonical Wnt receptor signaling pathway IDA PMID:19643732

In addition we often add something like: For the annotation of human MAPK14 Q16539 and MAPKAPK2 P49137 based on PMID: 18440775
To each sourceforge request for a new term so that we are reminded what we are planning to use the term for.

CGD/AspGD

Curator resources, funding.

Dictybase

blank

EcoCyc

GO curation: training of curators. A subset of EcoCyc curators are adding GO terms very occasionally, and thus make more mistakes than one would hope for, or else forget to add them entirely.
Submission pipeline: done by EcoliWiki

EcoliHub

a) Students make GO annotations faster than experienced curators can check them. Thus, the new bottleneck is assessing GO annotations. Using the online system and by requiring students to identify the exact figure or table that supports the annotation, this process is still significantly faster than making annotations de novo, but it is difficult to keep up with >150 students.
b) We delete incorrect annotations and fix every annotation flagged for changes after each semester ends (and the competition therein).
c) Students generally do not identify all of the potential GO annotations in papers. They tend to find "easy" annotations and move on to the next.

Flybase

The general consensus is that selecting the most appropriate GO term is the rate-limiting step. There are various reasons for this including:

a. The perfect term is often not present so time is wasted weighing up the time and effort associated with requesting a new one v choosing a compromise

b. Time spent deciding whether existing terms do represent what you are curating – especially for things like protein complexes.

c. Term names used by author often don't match with the GO definition so choosing between cell differentiation, specification, commitment, development etc can be tricky

d. Missing synonyms

e. Deciding between terms with subtle differences in meaning.

f. Definitions that include words that are not themselves defined elsewhere in GO

g. Ambiguous definitions – e.g. where start and end of a process are not defined

Finding the best papers for GO is also rate-limiting so a lesser extent.

GeneDB

Developer Time

InterPro

Curator inexperience and understanding the complexity of GO: it can sometimes take a long time for curators to decide on the most applicable GO term(s) for an InterPro entry.

MaizeDB

Setting up the infrastructure to support this.

MTBbase

Top is paper/experiment shortage. Second is reading/understanding the paper thoroughly. Close second is annotating high-throughput experiments where conversion of available data almost always takes some time. Similarly, any software effort for optimization of the annotation process.

MGI

At this point, we feel the main problem IS curator resources.

Pombase

None aside from upheaval from migration to new system and curator resources.

RGD

False positives and false negatives in PubMed search results are the main obstacles to manual GO curation at RGD.

Reactome

None really

SGD

Software resources to make changes to curation interfaces and GO tools.

SGN

Since we are a community based system, the training of community curators in GO is a limitation.

Soybase

That would be it!

TAIR

In the face of the need for more curators, any other obstacles are truly minor.

UniProt

The wait associated with requesting new GO terms, where TermGenie cannot be used. The process of identifying relevant papers to annotate can be time consuming. Could intelligent text mining efforts help to identify relevant lists of papers for curators?

Wormbase

For GO curation, ontology development can still be rate-limiting, especially if we find that a particular branch of the ontology may need significant expansion before it can accurately represent the biology. Curators always need to assess whether they have the time and/or the expertise to take on major ontology development projects.

For the submission process, continually changing GO annotation requirements and/or file format changes require significant time and effort in terms of updating curation forms and related scripts. The development of the common annotation framework tool along with curating directly into the GO database, will make the submission of annotations easier.

ZFin

In the past year or so we have spent some effort bringing GO into our database and integrating it into our curation and search interface more completely. This has reduced the need for curators to use tools outside ZFIN to locate terms for example, though many still do so for their more robust features. Our current GO annotation process is relatively efficient I think. Remaining efficient while taking advantage of more expressivity in our GO annotations will be problematic though due to resource limitations. Another aspect that GO curators (curators in general really) suffer with is decision fatigue. There are a LOT of decisions to make in the course of annotation of any kind. Anything that can lower the number of decisions or guide the curator to a smaller set of items to choose from would lower decision fatigue in the curation process. GO seems to have a high "decision barrier" with it's complex and large term set.

---

[i] **Availability**

EcoCyc does not traditionally validate their annotations before sending them to EcoliWiki, which is sometimes problematic. Previously, the annotations are sometimes not provided in any easily-used format (such as GAF1/2.), but EcoCyc made this significantly easier last year with a patch to PathwayTools was written to dump annotations.

Round-tripping

EcoliWiki imports the EcoCyc annotations regularly for users to view and change. Early imports of these annotations did not have any associated metadata, making it difficult to determine which annotations were seeded from EcoCyc and which were manually entered. This also effectively duplicated EcoCyc's annotations in step 3 above (merging the two sets of annotations together.) While a simple `sort` and `uniq` of the file takes care of this, attribution (column 15 in the GAF format) needs to be considered in order to keep from mis-attributing all of EcoCyc's annotations to EcoliWiki.

In addition to EcoliWiki importing EcoCyc annotations, EcoCyc imports annotations from the GOC. Any annotations that get removed from EcoCyc usually end up back in EcoCyc due to their lack of being removed in EcoliWiki. Similarly, and annotations that get changed in one database usually fail to get changed in the other. This is a central problem in maintaining the *gene_association.ecocyc file*, but we haven't really figure out the best solution. |

Taxons

This has been a bit of a problem making our GAF. EcoCyc uses the taxon 511145 (*Escherichia coli str. K-12 substr. MG1655)* when making annotations. However, EcoliWiki annotates the *E. coli* pangenome and includes annotations to many lab strains, not just MG1655. Currently EcoliWiki uses a complex algorithm to determine which genomes the annotation gene-product is in, and find the largest taxonomic group that encompasses just those genomes. If this can be computed, it is used, otherwise, the taxon 83333 (*Escherichia coli str. K-12*) is used.