GO Consortium annotation activities Progress Report

Rama and Emily

General Goals

Coordinate annotation efforts

- Educate curators
- Improve/refine documentation
- Expansion of Annotation paradigm
- Outreach, educate users
- Connect ontology development, annotation and work with software developers to bring all this to light

Contents

- RoadMap
  - http://wiki.geneontology.org/index.php/Annotation_Advocacy_Roadmap_2011
- Work in Progress
  - Outreach (Emily)
  - Annotation guidelines for Transcription, Signaling
  - Col-16
  - Annotation QC checks
- What is in the pipeline
  - Apoptosis
  - Mapping of ECO codes, GPAD format
  - Community Annotation tool
  - contributes_to qualifier, annotating to complexes
  - Relationship between annotation Object and GO term

External and Collaborative GO annotation activities

- **MTBbase; Ralf Stephan**
- paper-centric manual GO annotation effort
- Status: Annotated nearly all papers indexed in Pubmed on *Mycobacterium tuberculosis* H37Rv strain up to and including publication date 2009
  - 1,056 papers annotated
  - 6,731 annotations for 2,320 proteins (58% of genome)

External and Collaborative GO annotation activities

- **AGOA (Auditory GO annotation initiative)**
- Guvanch Ovezmyradov, University of Göttingen and supported by deafness researchers from a number of labs.
- *Aim: Review and update gene products that are assigned to the GO:0007605 term (sensory perception of sound)*
- In contact with ZFIN, FlyBase, MGI and UniProtKB regarding manual GO annotation of ~236 gene products involved in deafness.

External and Collaborative GO annotation activities

Continued collaborations:

Tufts University; Human Fetal Development annotation

- Heather Wick, a curator from Tufts University, working as a part of an NIH grant investigating proteins implicated in human fetal development (PI: Donna Slonim).
- (supported by MGI and UniProt-GOA groups)

Trondheim, Norwegian University of Science and Technology – Gastrin gene product annotation

- Annotations submitted by the Martin Kuipers systems biology group at NTNU
- (supported by UniProt-GOA)

GO annotation outreach by the Consortium

November 2011

PomBase: undergraduate GO annotation practical for first year biochemists at Cambridge University

- involves GO annotation of a set of fission yeast papers.
- Hoping to roll this out as a departmental mini-workshop for people to annotate their own papers when the generic community annotation tool is available.

October 2011

Jim Hu and Brenley McIntosh: GO and CACO talk; whether high school students from these high-performing science schools could participate in CACAO Anual professional meeting of the National Consortium of Specialized Secondary Schools of Mathematics, Science, and Technology.

September 2011

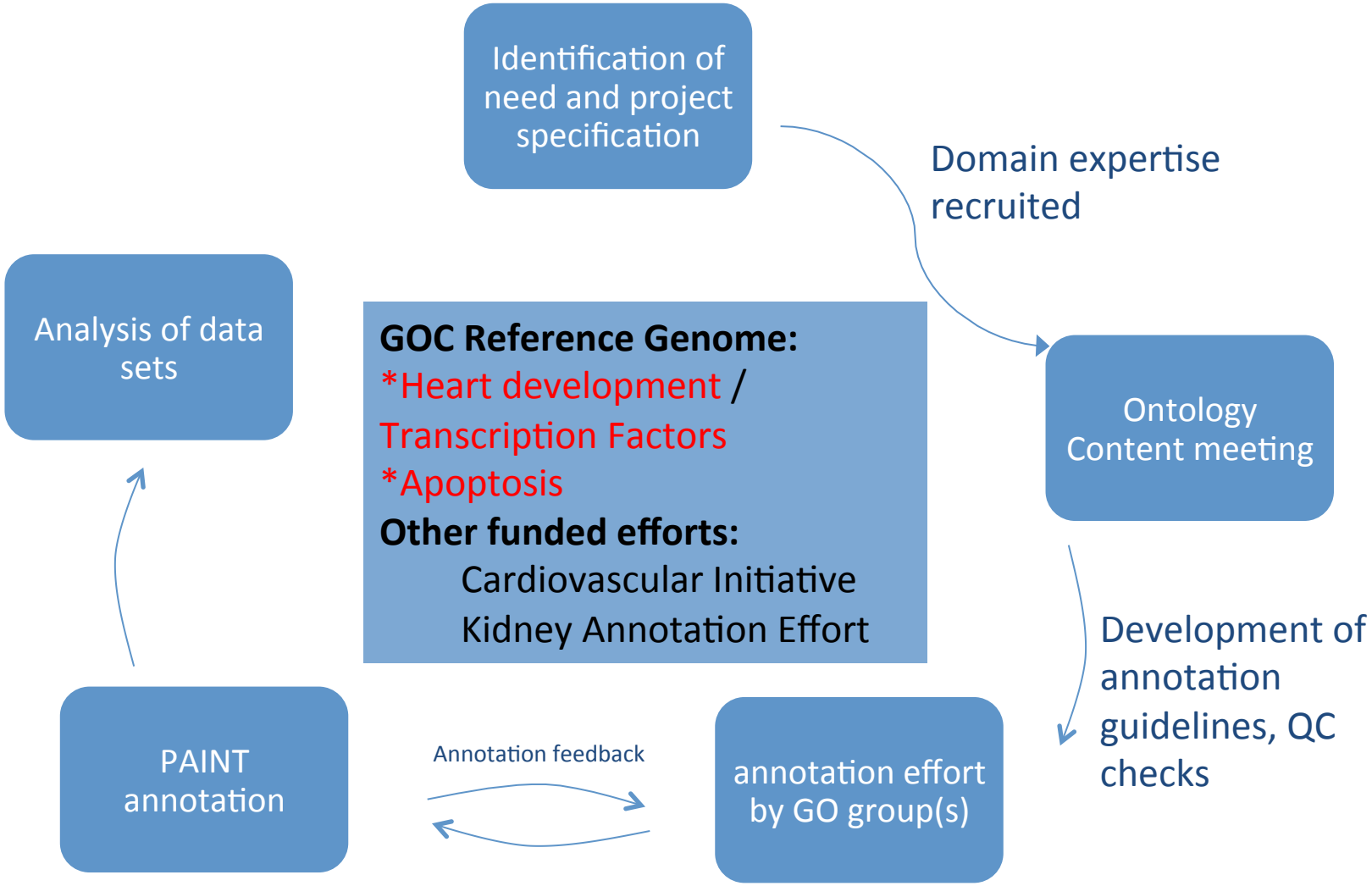
BHF-UCL:

- 2 day annotation workshop for 20 UCL staff and postgraduate students
- 4 of last years MSc students to participate in the CACAO competition
- Taught a module of the MSc genetics of Human Disease (attended by 18 students), the students are required to annotate 3 papers using GO terms during the course.

July 2011

Jim Hu: GO and CACO talk at the ASM-JGI Functional Genomics Workshop at Hiram College in Hiram, Ohio.

Annotation and Ontology development



Transcription factors

- Several conference calls in the last 6 months dedicated to transcription factors
- Outcome: Curation manual
  - Plan to make it available via the GOC website
  - Documentation on the wiki
  - Extended the scope of the IC reference
  - TermGenie has been very useful

Documentation of Annotation Guidelines

- Annotation guidelines from the Geneva Camp are on the GOC website
- Updates to Evidence Code Documentation
  - IBA, IBD, IKR, IRD codes available in the Evidence Code web page on the GO Consortium website (and the codes are accepted by Mike Cherry's filtering script) http://www.geneontology.org/GO.evidence.shtml
- Extending the scope of the IC evidence code (new GO_REF was added and Evidence code documentation was updated)

Annotation Quality Control Checks

Implemented:

- Illegal to use of the 'NOT' qualifier with 'protein binding ; GO:0005515‘ **(annotations removed)**
- 'binding ; GO:0005488', and its child 'protein binding ; GO:0005515', should be made with IPI and interactor should be in the 'with' field **(annotations removed)**
- Only use the IEP evidence code with terms from the Biological Process Ontology **(annotations removed)**

Approved

- Reciprocal protein binding annotations should be made
- Annotations to 'protein binding ; GO:0005515' should not use the ISS evidence code or any flavor of ISS
- Illegal to use the IPI evidence code along with catalytic activity molecular function terms

Proposed

- All IC annotations should include a GO id in column 8 (with)
- Should the binding restriction (see implemented rule above) include the term ‘GO:0019904 protein domain specific binding’
- All IDA annotations should not include an identifier in column 8 (with)
- All identifiers in the GAFs must use the correct DB abbreviation
- Annotation to high level terms should be discouraged (e.g.
- Correct format for PubMed refs
- All gene/protein/chemical identifiers used in GO annotations should conform to RegExps supplied in the GO.xref.abbs file
- ND Annotations to root nodes only and using the same GO_REF?
- Should be illegal to use secondary or obsolete GO Ids
- Dual species taxon check **(**multiple taxon ids in an annotation are used to capture information about multi-organism interactions, should be used only in conjunction with 'GO:0051704 : multi-organism process' 'GO: 0044215 : other organism' as an ancestor.)

Discussion: Are we handling annotation exceptions correctly?

Taxon Checks

- Checks on submitted GAFs are run weekly, results are here:

ftp://ftp.geneontology.org/pub/go/quality_control/annotation_checks/taxon_checks/

- Make these checks hard QC checks

IEA improvements using Taxon Checks

Currently ~250,000 IEA annotations are failing the taxon rules

- many annotations fail due to the selection of a non-host CC term for viral sequences
- Post-processing of IEA methods will start shortly to correct GO term will mean changes to the annotation set not represented via the external2go mapping files

In the pipeline…

Annotation projects

- Sort out correct use of contributes_to qualifier
- Annotation of Complexes
- We are looking into forming working groups for areas in the ontology that are hard to annotate

Community Curation Tool

- Development of a community annotation tool, to support community GO Consortium annotation contributors, by adapting one developed by Kim Rutherford (PomBase)

Advantages:

- PomBase have already developed this tool
- Little development needed before it can be used
- Greatly lowers the threshold for community annotation contributions
- Will be ready much sooner than the Common Annotation Framework

Progress:

- Wiki page of requests reviewed by PomBase software developer:

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

- Development to start this month.

Inclusion of external annotations in primary GOC files

- Annotations have been created for many species by a number of GO annotation projects:

- Manual: Reactome, RefGenome, GOC-inferred from inter-ontology links, MGI, PAMGO, BHF-UCL, UniProt GO annotation

- Automatic: InterPro, Ensembl, EnsemblGenomes, UniProtKB

  Some annotations for certain species are not available from the main GO /gene-association ftp directory or AmiGO.

  It is the database's choice as to what annotations they display on their website, **however** shouldn't all GO Consortium annotations should be made available to all users from the GOC website?

Database	Species	Integrating annotations?	Frequency?
CGD	*C. albicans*	None. No intent to integrate external annotations	n/a
dictyBase	*Dictyostelium spp.*	None. Intentions to start, but limited by developer issues	n/a
FlyBase	*D. melanogaster*	InterPro2GO, old UniProt annotations	InterPro2GO updated 10/ year.
GeneDB	*L. Major, P. Falciparum, T. Brucei, G. morsitans*	None	n/a
PomBase	*S. pombe*	RefGenome, GOC, UniProt-GOA	Restarting monthly update shortly, intending increase frequency
UniProt-GOA	*Gallus spp., B. taurus, H. sapiens*	All manual GOC annotations	Nightly updates . Weekly QuickGO releases, files released fortnightly.
Gramene	*Oryza spp.*	None	n/a
MGI	*M musculus.*	RefGenome, GOC, UniProt-GOA	Weekly
PAMGO	*A. tumefaciens*	None	n/a
RGD	*R. Norvegicus*	UniProtKB, RefGenome, Reactome, MGI, IntAct, HGNC, GOC, BHF-UCL, Ensembl	?
SGD	*S. Cerevisiae strains*	GOC, HGNC, InterPro, MGI, RefGenome, UniProt-GOA	4/year
TAIR	*A. thaliana*	RefGenome, GOC, UniProt-GOA	weekly
JCVI	*A. Phagocytophilum, B. Anthracis, C. Burnetii, C. Hydrogenoformans, C. Jejuni, C. Perfringens, D. Ethenogenes, E. Chaffeensis, G. Sulfurreducens, H. Neptunium, L. Monocytogenes, M. Capsulatus, N. Sennesu, P. Fluorecens, F. Syringae, S. Oneidensis, S. Pomeroyi, T. Brucei, V. Cholerae.*	PAMGO,	?
WormBase	*C. elegans*	Non-IEA annotations from UniProt	Approx. Every 2 months
ZFIN	*D. rerio*	RefGenome, GOC, UniProt-GOA	monthly

How could the GOC could assist groups add in external annotations to GOC submitted files?

1. Provide one central location with information describing location and format of all files groups need to integrate.

2. Add in annotations centrally where annotation group unable to assist

   - Annotations could be regularly appended to submitted GAFs where data is lacking

   - Files needing a high-level of intervention could be renamed so the annotation group would not be implied as being responsible for the altered contents.

Resources for Annotating groups

- Lot of discussion on Col-16, ECO evidence codes, complexes etc.
- MF-BP, PAINT inferences not incorporated
- Turns out **not** all groups have resources to extend their curation practice
- How can the Annotation team/GOC help?