Transcription Overhaul project report

Ontology Developers: Karen Christie (SGD), David Hill (MGI)
Annotation Effort: All groups

Annotation Effort

- It was not concurrent with the ontology development.
- Genes identified from the heart development targets were part of a coordinated annotation effort (60 total, 36 human), to use the newly-created terms.
- Otherwise, mostly involved reannotating gene products annotated to obsolete terms
- Annotations to merged terms were transferred to the term it was merged with
- First set of terms were obsoleted in May/June 2011

- **First Annotation Jamboree on July 2011**
  - Karen's presentation on ontology overhaul
  - Specific annotation questions from papers/examples
  - http://gocwiki.geneontology.org/index.php/Transcription_jamboree
- Some of the new terms could not be used because evidence needed to annotate to that term is not available in one paper.
  - Example: http://bit.ly/xvrdOP
  - Annotation format development: Extended the scope of IC evidence code to be able to annotate to a granular term form multiple pieces of evidence
- Karen created/generated an annotation manual
  - The manual suggests terms for gene products and offers PMIDs for reviews
  - http://gocwiki.geneontology.org/index.php/File:TxnOHreannotationGuide.xls.pdf

- There are 25K Manual annotations made to the new terms (MF and BP and by all groups) created during this overhaul (excludes IEAs).
- Annotations made to New Terms is charted at the end of this document (here I am just counting the # of unique geneproducts to GO term associations)
- There are 150 terms from the GOC:txnOH project (out of a total of 283) that have annotations made to them. About 74 of these terms have less than 9 gps annotated to them and about 34 have less than 2 gene products annotated to them (shown in table at the end)
- Terms that were not used are higher level terms, turns out that there are two terms that were 2 levels from root, the rest of the 133 are reasonably granular (distance to root range from 2 to 9)

General Observations

Pros

- Overall the MF ontology got more accurate for this branch of biology
- Relationships were used efficiently/correctly to represent the biology
- infrastructure is in place in the ontology which makes it easy to add new terms and expand the ontology
    - in fact this is already happening and curators are able to request terms in this area via TermGenie
- SGD was able to request lot of terms in areas unrelated to transcription because we did a complete overview of annotations.

Cons

- Defining the scope of Reannotation: The ontology overhaul project was undertaken to address several inconsistencies in the MF ontology and to some extent the BP ontology. These changes in the ontology had mixed effect on the annotations. Annotations for transcription factors involved in heart development got an overall review while for many other gene products, annotations to just the obsolete terms got fixed. A clear goal for reannotation was not put in place (a goal to review the MF annotations for all/some subset of the several hundred transcription factors, a consistent depth etc)

- Efficiency of annotation: The ontology overhaul resulted in adding new terms, obsoleting incorrect ones, merging synonymous ones and clarifying definitions in some cases.
    - When terms were obsoleted some groups tried to fix their annotation errors by visiting the paper that was originally used to make the annotation and replaced the term. Some groups moved annotations to a close BP term
    - Some groups did not go in to the literature to see if additional annotations to all the new/granular terms are possible
    - Many terms were created from Reviews. Some were not GO annotatable from a single paper. In some cases a curator knew a term was appropriate for a gene product and spent hours trying to look for an evidence and couldn't find evidence and many of the granular terms got IC annotations. This may be a caveat of the annotation system.
    - There are transcription factors unaffected by the overhaul (i.e did not have annotations to obsolete terms). Annotations to these proteins could have been improved but were not part of the review. Curators reviewed annotations only to the obsoleted terms.
- Consistency: Groups have annotated at different levels. Also when terms are so granular and long, consistent interpretation can be an issue. Vague terms can lead to vague/misleading annotations
- Depth of Annotations: Some of the most granular terms were used to annotate just a handful of gps. Was this the desired output?

- If the granular terms were annotated using IC, then they cannot be propagated

Suggestions

- Setting the expectation or defining a method for reannotation for all the annotating groups ahead of time might have helped. The expectation should be to annotate to a certain depth (set some standard) and this will ensure everybody is consistent across the board. (*Annotation Advocacy's job*)
- Targets for reannotation should be picked ahead of time and this should ensure wide coverage and this should be in addition to fixing annotations to obsolete terms.
- Involve a select set of curators representing different taxons to give feedback from the beginning (*Annotation advocacy's job*)
- Accurate vs Appropriate, perfect vs better ontology terms
    - At what level of ontology can we expect consistent annotations?
    - Should ontology developers be making terms from Reviews (*ontology dev. Process*)
    - Could we have created the basic ontology structure with some medium level terms, and let the curators request terms as they looked at papers? (*ontology dev. Process*)
    - Ontology development should be done in a Modular fashion with more frequent feedback/discussions with the curation groups so that reannotation efforts are more manageable (*ontology dev. Process*)
    - Users are consuming annotations. We should make sure the terms and annotations are appropriate for their use: http://biostar.stackexchange.com/questions/41/how-much-do-you-trust-geneontology-annotations
- Ontology developers have fixed something that was inaccurate. How do we know it is better now? How to best use the ontology? How do we measure this effort?

Reaction from various groups:

SGD

- SGD did complete review of all annotations in all aspects for those transcription factors that threw an error. We reviewed the literature completely to look for evidence to annotate to the most granular terms since more terms were added to the ontology.
- # of genes that were reannotated: 260, 9FTEs working on it **part time** for 9 months. For GO we pretty much focused on Txn factors during these months.
- But there are other transcription factors that we did not look at at all because they did not throw an error (i.e. unaffected by the overhaul).
- Some annotations to granular terms could be made only using IC
    - http://www.yeastgenome.org/cgi-bin/locus.fpl?locus=spt15

- Overall we can say that our annotations are improved from a curator's perspective. This also spawned the addition of terms in other parts of the ontology since we did a complete review
- But the whole process took several rounds of discussions and months to complete
- To ensure consistency it would be better if a small group of curators worked on the reannotation
- using the terms "DNA binding txn factor activity" and "protein binding txn factor activity" would have been sufficient for function. Additional information gained by annotation to more granular terms was often redundant as the same information was already captured by existing process/complex terms. Also for longer compound terms, it is not obvious we did our users any favors by annotating to the children of "DNA binding txn factor activity" and "protein binding txn factor activity". Especially since we often have to use IC to annotate to these terms.
- The types of terms in the ontology for the different polymerases is inconsistent. For pol II the term names are laid out in as the individual components of what is involved (as far as I can tell, all of the steps performed by a given txn factor complex) while for pol III the terms are named after the txn factor complex.
    - For example: sequence-specific core promoter binding RNA polymerase II transcription factor activity involved in preinitiation complex assembly
    - TFIIIB-type transcription factor activity
- Terms are suboptimal, meaning they are difficult to use for annotation, and they must look silly to users (some of the term names are very long and repetitive in their use of words). New terms are still being requested in spite of the overhaul which seems suboptimal.

RGD (from Stan)

It seems to me that I linked over to AmiGO or GOnuts to look for suggested terms to replace the obsolete terms. Most of the time though, I think we just went into the ontology to look for the appropriate term after glancing at the associated abstract/paper. As for general effect on the RGD curation effort, it hasn't really had a noticeable impact. However, from the other RGD curators I have received complaints about overly long transcription-related terms.

FlyBase (from Susan)

I was lucky to have some help from a temporary curator. She revisited all of the papers for which a term was made obsolete and no direct replacement was available - in the process she was able to improve some of the original annotation e.g. making more specific annotations to RNA pol II terms rather than the more general terms that had been added originally added. Our main focus was where we lost molecular function annotations because they had been replaced by a process e.g. transcription regulator activity. In cases where the original paper had no evidence to support a specific molecular function and the gene had no remaining molecular function

annotations as a result of the obsoletion she searched for new papers to make an MF annotation.

However we have not been able to revisit all transcription related GO annotations to make them completely consistent with the transcription overhaul. This will happen gradually as genes are reviewed.

The work we did took around 40 days of full-time effort and on reflection it is doubtful that this was the best use of the time given the amount of new information/improvement was modest compared to curating papers about genes that lack GO annotation. I appreciate that a lot of effort when into the transcription overhaul and that it was good to make a more consistent set of terms but it is unlikely that we will ever be able to devote this much effort to revisiting annotations in future.

TAIR (from Tanya)

I reviewed all of the annotations in question manually. I don't know how many were affected because I didn't keep track. I ended up obsoleting a bunch of annotations because they were only IEA or ISS by TIGR and I do not reassign to a new term for those annotations. I also removed several non-IEA/non-ISS annotations because the original annotations were wrong. For the remaining annotations, I moved them to better terms that were created during the overhaul. For the latter three sets, I revisited the papers on which the original annotations were based to figure out the appropriate term to use.

In all, annotation quality of the Arabidopsis set was improved by the transcription overhaul.

WormBase (Kimberly and Ranjana)

In general, when a term has been obsoleted and there is no direct replacement we would typically go back to the original paper and, looking at the experiments performed and the new version of the ontology, try to find the most appropriate valid term.

I've been trying to update transcription factor annotations as part of the Ref Genome project. Outside of that, I think we've both probably just updated annotations as we've come across the genes wrt new papers we're curating. For transcription factors, as well as other gene families, we are both trying to check annotation consistency as we curate, to bring annotations to different genes in the same family more in line with one another (as experiments permit).

On the whole, our transcription factor annotations are somewhat limited by the types of experiments performed in C. elegans. Typically, we might have an IDA sequence-specific DNA binding annotation, and perhaps a complex or protein binding-type IPI annotation. For process annotation,

much of what we curate is based on in vivo experiments that demonstrate a change in expression of a reporter gene in the mutant background of the transcription factor. These we would annotate to a process term using IMP.

I don't know if we will ever have the types of biochemical experiments in C. elegans that will allow use to make more granular function and process IDA annotations. That said, however, at some point I'd like to explore using GO_REF:0000036 to make more granular IC annotations, if I can. Unfortunately, that's not a high priority right now, though, so I don't know how soon I'll be able to do that.

Zfin (Doug)

For the most part, the level of detail in the transcription overhaul didn't affect us too much. People don't study transcription per se really in zebrafish. They do examine gene expression a lot, but not the mechanism of transcription itself. We didn't have too many changes to make, and if I remember correctly, the ones I did have to make were mostly distinguishing between protein binding transcription factor activity and DNA binding transcription factor activity. I typically went to the paper involved and gave it a look, examined the other annotations on the gene, then made a judgement and updated the existing annotations.

Val (Pombase)

From an obsoletion point of view, this wasn't so painful.... this was because I had used the existing terms very specifically (over interpreting them), for example all sequence specific transcription factors used http://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0003704#term=info and a DNA binding term)

I think, many of the other obsoleted terms, I never used as it wasn't clear what they meant, or for those I did the remapping was clear. I wasn't aware of any major problems anyway....

However, we are left with a bit of a mish-mash because the annotations of the 90 known DNA binding transcription factors, 90 have ended up annotated to the DNA binding term: RNA polymerase II core promoter proximal region sequence-specific DNA binding, but only 60 have ended up mapped to the transcription factor term sequence-specific DNA binding RNA polymerase II transcription factor activity (this is likely because not all of the annotations came from manual, and the remainder was made up of IEAs which could have come from other places.....

So mainly our work will involve bringing back the consistency which was present before (and this may apply to other types of transcription apparatus). For instance I know that complexes like RNA pol II, SAGA, mediator etc, had consistent annotation, but this was based on the totality of ,manual and inferred annotations. These will need to be checked over time. I was waiting for Karen and David's list of which

function terms the common transcriptional machinery should have.

For instance,
mediator = RNA polymerase II transcription cofactor activity
etc

UniProt
We found the overhaul difficult to cope with, as we had so many manual annotations that had applied an obsolete molecular function term. It was not feasible to reannotate every single annotation .

We wanted to carry out an automatic update of our annotations to correct biological process terms as a first round correct. Although the comments in the obsolete terms pointed to a biological process that would have been correct in all instances, it was tagged with a 'consider_by' rather than 'replaced_by' tag. Meaning we could automatically update obsolete terms. There was a long exchange on SourceForge, trying to convince the editors to change this tag, without success:

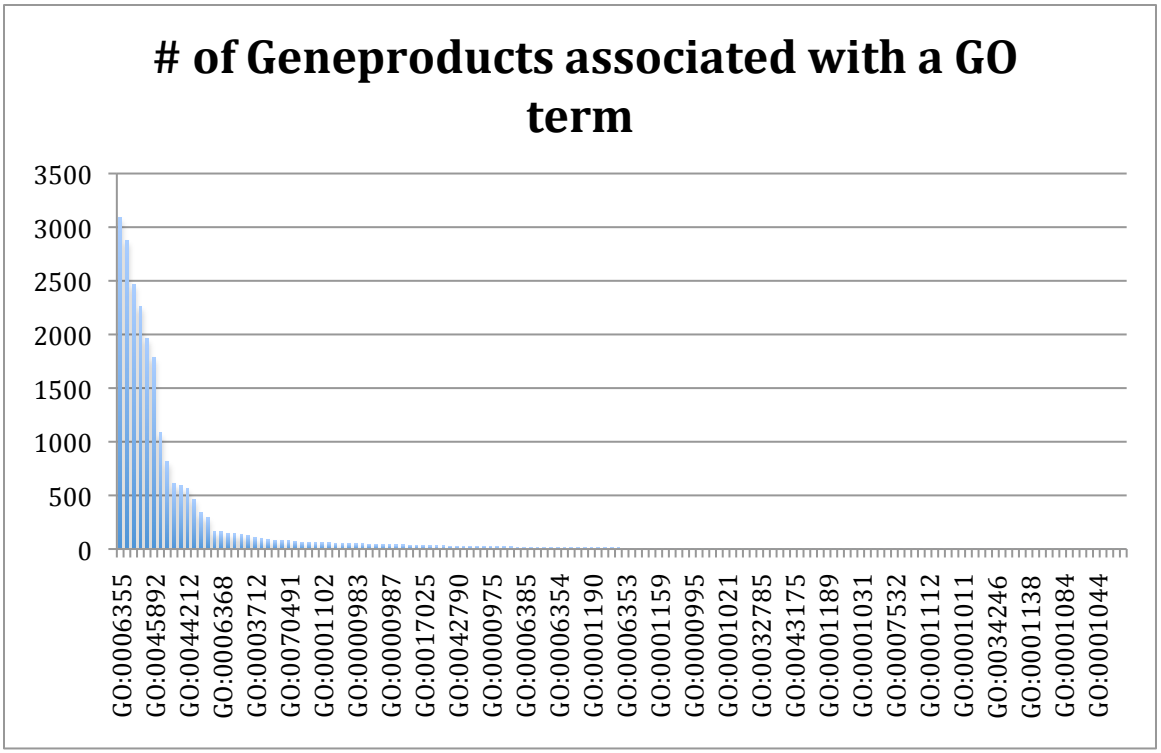
http://sourceforge.net/tracker/?func=detail&atid=440764&aid=3309586&group_id=36855

This meant that even the first-round update of our annotation set was unnecessarily problematic. While some MODs have a small enough set of annotations to be able to review each one manually, it would be helpful for further consideration to be made for groups which are so badly affected by an ontology change that they do not have the luxury of carrying out a complete manual revision of annotations, at least in the short-term, and need to make bulk changes so not to lose annotations in their next file release.

This change causes our dataset to change considerably in annotation characteristics, with so many molecular function annotations disappearing and changing to biological process terms. I think that we need to have a way of communicating with our users when the GO Consortium annotation set changes so much.

It also seems a pity that there is no grouping term called transcription factor activity – when this is such a well-known, and well-used term by the scientific community. However I do understand the reasoning behind the changes to the molecular function terms.

Finally, it might be useful to have a regular reassessment on the annotation impact after each large ontology revision goes live that affects many annotations – we can then get a more detailed understanding of everyone's experience and use it to review and improve these processes.



Distribution of Annotations to various Txn GO terms

# of associations	# of terms	Path to root (indicates granularity)
>1000	7	7 or 8
500 - 999	5	4 to 7
100 - 499	10	8 to 3
50 - 99	16	8 to 1
25 - 49	19	9 to 1
24 - 10	19	8 to 2
<9	74	10 to 3
<3	34	9 to 3

Annotation by Group

8883	MGI
7818	UniProtKB
1828	FlyBase
1567	GOC
1467	BHF-UCL
1443	SGD
1072	TAIR
957	RefGenome
941	PINC
543	Reactome
526	RGD
484	PomBase
222	ZFIN
176	CGD
168	WormBase
165	AgBase
162	dictyBase
126	HGNC
83	MTBBASE
76	EcoCyc
27	DFLAT
11	EcoliWiki
10	JCVI
8	GDB
7	GR
4	TIGR
2	Roslin_Institute