Annotation QC checks

Categories of QC checks

- Goal is to identify inconsistencies and alert curators about these errors in an automatic fashion
- Incorrect or suspicious annotations
  - where annotations are **always considered incorrect**, and therefore annotations should be automatically removed.
    - Hard QC checks
  - where annotations **appear suspicious** and groups encouraged to recheck their data.
    - Soft QC checks

GAF pipeline

- Annotations are submitted by the various groups to the go/gene_association/submission directory
  - ftp://ftp.geneontology.org/pub/go/gene-associations/submission/
- A filtering script is run over these annotations, incorrect annotations are removed, groups contacted about those rows and the filtered annotations are what is available to the public/AmiGO
  - ftp://ftp.geneontology.org/pub/go/gene-associations

Error checks currently in place

- Most are carried out by an annotation control script maintained by Mike Cherry (Stanford), and run over all annotation files uploaded to the GOC. This includes checks on:
  1. **correct format of submitted files**
     Example: correct number of columns, no inappropriate white spaces etc.
  2. **appropriate use of qualifiers with terms from the different GO ontologies**
     Example: Only use the colocalizes_with qualifier with cellular component terms
  3. **use of primary GO identifiers in annotations**
  4. **correct usage of the 'with' field in annotations using certain evidence codes**

- Groups are also supported in the generation of **gp2protein** files
  – Emails are sent from the UniProtKB-GOA project highlighting problematic UniProtKB-MOD identifier mappings

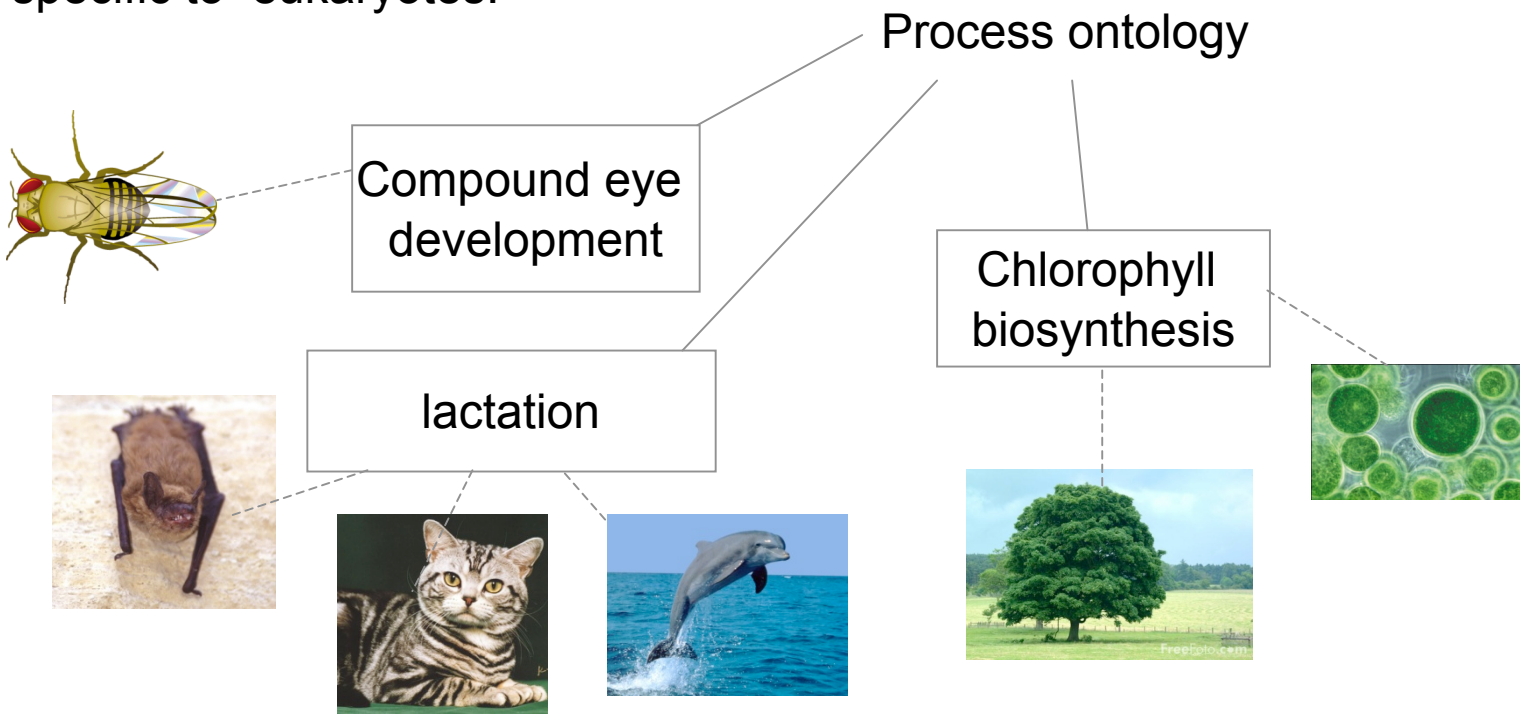
New QC checks

- New annotation checks are being formulated by annotation working groups.
- Hard QC checks (annotations that shd be removed)
  - Already agreed upon from the Binding working group
    - No use of the 'NOT' qualifier with 'protein binding'; GO:0005515.
    - Annotations to 'protein binding'; GO:0005515, should only be supplied with an evidence code where the interactor can be identified in the 'with' field
    - Annotations to 'protein binding' should not use the ISS evidence code
    - IEP evidence code with terms from the Biological Process Ontology
  - Possible new check from protein complex
    - IPI evidence code may not be used with catalytic activity molecular function terms

Taxon-based constraints to detect annotation inconsistencies

developed by Jennifer Deegan with Chris Mungall (paper submitted)

Although the ontologies are developed to be taxon neutral, and to cover all species, there are inherent taxon specificities in some branches. For example, the process ‘lactation’ is specific to mammals and the location ‘mitochondrion’ is specific to eukaryotes.



• An inference system has been developed to check for violations of these constraints in annotations.

• Helps detect and remove errors in annotations and improve the structure of the ontology.

Rules are collected in a central taxon constraint file

```
[Term]
id: GO:0007595
name: lactation
relationship: only_in_taxon
NCBITaxon:40674 ! Mammalia
```

```
[Term]
id: GO:0019684
name: photosynthesis, light
   reaction
relationship: only_in_taxon
ID:0000007 ! Viridiplantae or
   Bacteria or Euglenozoa
```

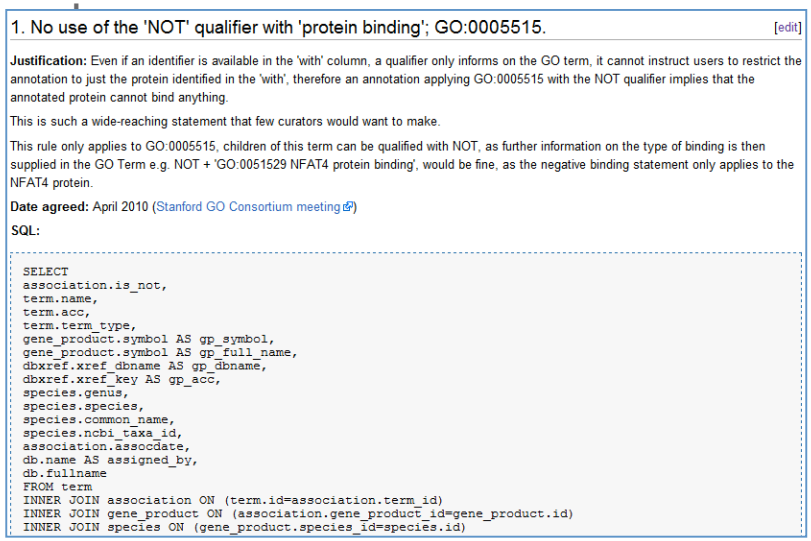
New taxon constraints should be suggested via the **Ontology SourceForge tracker.**

Annotations that need to evaluated

- Soft QC
  - Reciprocal annotations for protein binding should be made
  - Taxon constraint violations
  - Matrix project analysis
- Additional checks are now being proposed
- System will be in place to alert curators about the annotations that need to be evaluated

All annotation checks must be fully described at:

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



1. No use of the 'NOT' qualifier with 'protein binding'; GO:0005515. [edit]

Justification: Even if an identifier is available in the 'with' column, a qualifier only informs on the GO term, it cannot instruct users to restrict the annotation to just the protein identified in the 'with', therefore an annotation applying GO:0005515 with the NOT qualifier implies that the annotated protein cannot bind anything.

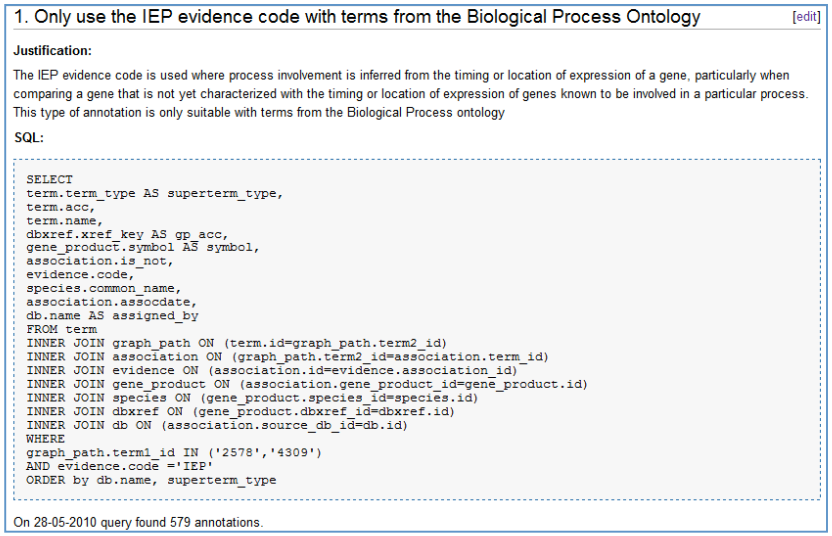
This is such a wide-reaching statement that few curators would want to make.

This rule only applies to GO:0005515, children of this term can be qualified with NOT, as further information on the type of binding is then supplied in the GO Term e.g. NOT + 'GO:0051529 NFAT4 protein binding', would be fine, as the negative binding statement only applies to the NFAT4 protein.

Date agreed: April 2010 (Stanford GO Consortium meeting)

SQL:

```
SELECT
association.is_not,
term.name,
term.acc,
term.term_type,
gene_product.symbol AS gp_symbol,
gene_product.symbol AS gp_full_name,
dbxref.xref_dbname AS gp_dbname,
dbxref.xref_key AS gp_acc,
species.genus,
species.species,
species.common_name,
species.ncbi_taxa_id,
association.assocdate,
db.name AS assigned_by,
db.fullname
FROM term
INNER JOIN association ON (term.id=association.term_id)
INNER JOIN gene_product ON (association.gene_product_id=gene_product.id)
INNER JOIN species ON (gene_product.species_id=species.id)
```



1. Only use the IEP evidence code with terms from the Biological Process Ontology [edit]

Justification:

The IEP evidence code is used where process involvement is inferred from the timing or location of expression of a gene, particularly when comparing a gene that is not yet characterized with the timing or location of expression of genes known to be involved in a particular process. This type of annotation is only suitable with terms from the Biological Process ontology

SQL:

```
SELECT
term.term_type AS superterm_type,
term.acc,
term.name,
dbxref.xref_key AS gp_acc,
gene_product.symbol AS symbol,
association.is_not,
evidence.code,
species.common_name,
association.assocdate,
db.name AS assigned_by
FROM term
INNER JOIN graph_path ON (term.id=graph_path.term2_id)
INNER JOIN association ON (graph_path.term2_id=association.term_id)
INNER JOIN evidence ON (association.id=evidence.association_id)
INNER JOIN gene_product ON (association.gene_product_id=gene_product.id)
INNER JOIN species ON (gene_product.species_id=species.id)
INNER JOIN dbxref ON (gene_product.dbxref_id=dbxref.id)
INNER JOIN db ON (association.source_db_id=db.id)
WHERE
graph_path.term1_id IN ('2578','4309')
AND evidence.code ='IEP'
ORDER by db.name, superterm_type
```

On 28-05-2010 query found 579 annotations.

And alerts to new annotation checking rules must be sent to all groups via the **go@geneontology.org** or **annotation@geneontology.org** lists.

Annotation advocacy group

- Lot going on:
  - ontology development
  - evidence code usage
  - annotation working groups
  - annotation inferences (PAINT)
- Goal of this group is to address these and other concerns
  - http://wiki.geneontology.org/index.php/Annotation_Advocacy_and_Coordination
- We encourage you to post your queries/questions to the GO annotation mailing lists.

go@geneontology.org or annotation@geneontology.org

Annotation Matrix Project

Developed by Val Wood

- A way to globally assess annotation consistency in and between organisms by using annotation intersections between biological processe or cellular component terms.
- A matrix of intersections can be generated between high level terms, where the expected intersection is close to zero.

S. pombe	DNA replication	Amino acid met	mRNA metabolism	Vesicle mediated transport	Transmembrane transport	Protein aa glycosylation
DNA replication	127	-	-	-	-	-
Amino acid metabolism		199		-	-	-
mRNA metabolism	2		219	-	-	
Vesicle mediated transport				307	-	-
Transmembrane transport				12	323	-
Protein aa glycosylation				1	4	69

- After checking, a set of annotation rules can be generated where a “zero term intersects” is approved.

This could result in curators being sent annotation suggestions:

- “Intersections of term A and term B indicates an annotation should be made instead to “regulation of a’ “
- “Intersections of term A and term B only allowed when annotated to complex C “
- “Intersections of term A and term B only allowed in taxon x “

Generation of new, inferred annotations to improve annotation consistency

- Current electronic and large-scale annotation methods to transfer annotations between orthologs:
  - Manually: PAINT and individual ISS annotations from groups
  - Electronically: Ensembl Compara, HAMAP2GO

New annotations from inference methods from improved relationships in the ontology

- GO is now able to link terms between the 3 GO ontologies.

```
[Term]
id: GO:0004842
name: ubiquitin-protein ligase activity
namespace: molecular_function
def: "Catalysis of the reaction: ATP + ubiquitin +
protein lysine = AMP + diphosphate + protein N-
ubiquityllysine." [EC:6.3.2.19, PMID:9635407]
    is_a: GO:0019787 ! small conjugating protein
ligase activity
relationship: part_of GO:0016567 ! protein
ubiquitination
```

New inference methods from improved relationships in the ontology

- Until a sufficient number of users/tools use the extended GO file, annotation sets will be supplemented with automatically inferred annotations generated from these inter-ontology links.

- Currently, links are available between **Biological Process** and **Molecular** Function terms.

- Curators are encouraged to use ontology SourceForge tracker using the category '*inter-ontology link*' to request new relationships

- Proposed relationships will be made available to curation groups to test before being added into the ontology.

Maintenance of the annotation rules

- Important that all agreed checks and annotation inference rules are:

    1. Maintained in a central GOC location
    2. Run regularly over current GO Consortium annotation sets
    3. Available to be run locally by individual GO tools or curation groups, to allow:

        1. Batch QC checking of entire annotation sets before annotations are released to the GOC.
        2. Provide immediate feedback to curators while generating a manual annotation (e.g. Possibility of JavaScript libraries available to support curation tools)