Using biological process co-annotation for ontology and annotation quality control
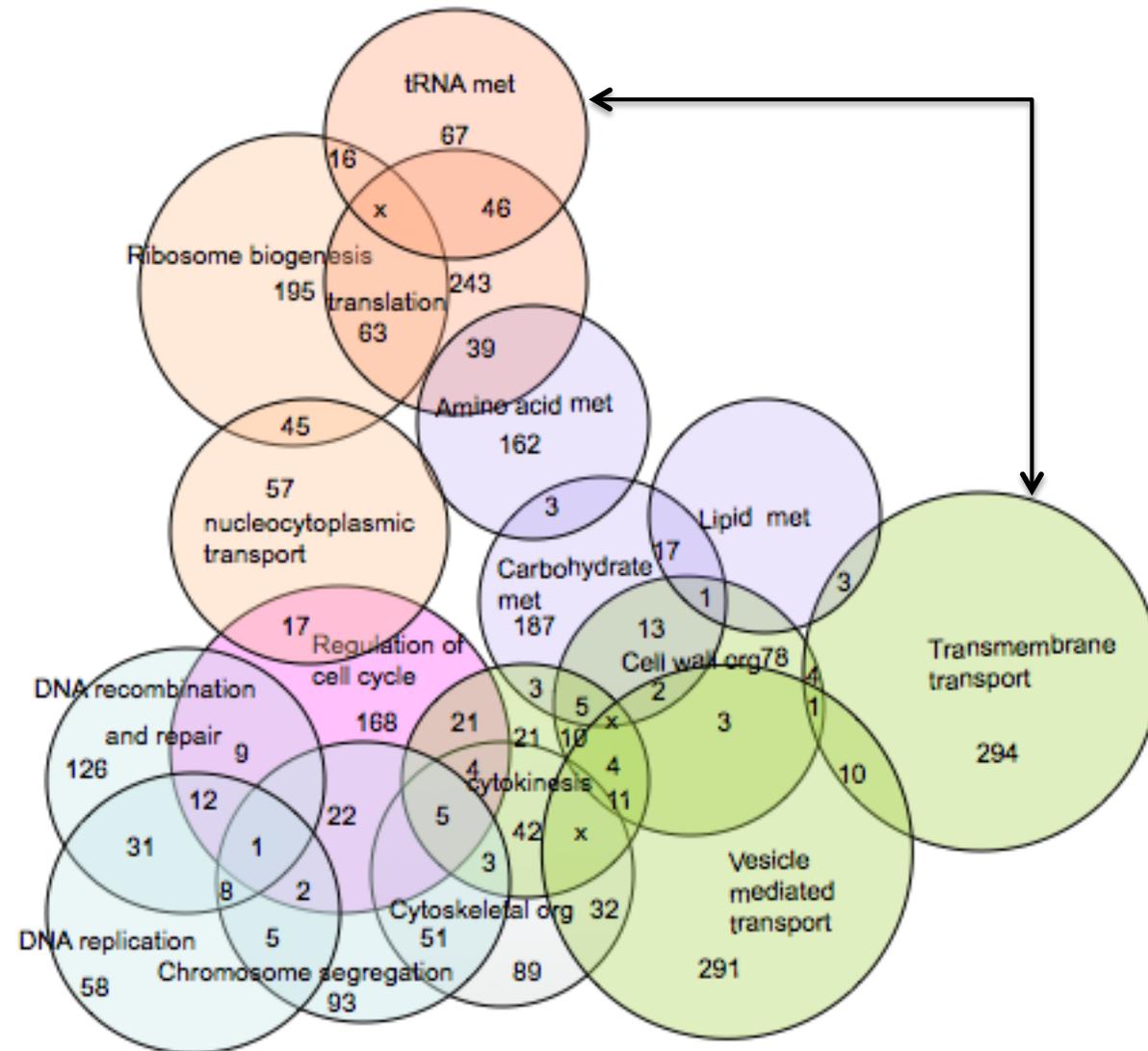
Biological processes are accomplished by the coordinated action of sets of gene products. However, some processes are rarely connected to each other because they functionally, temporally or spatially distant.

Matrix View

Matrix Inputs

Axis 1:

```
GO:0005975
GO:0009117
GO:0006807
GO:0006790
GO:0032502
```

Axis 2:

```
GO:0005975
GO:0009117
GO:0006807
GO:0006790
GO:0032502
```

Axis 3 (optional):

Select species

S. pombe ▾

Select relations

- ● include all relations
- ○ exclude regulates (waiting for database change...)

Submit Query

https://sourceforge.net/apps/trac/pombase/wiki/MatrixProject

Matrix

Information about Matrix ⊘

Input GO IDS

GO:0043473 GO:0009987 GO:0022008

http://amigo2.berkeleybop.org/cgi-bin/amigo2/matrix

(caveat, only pombe filter currently available)

Taxon

⊙ All
○ pombe

Step 1

Step 1 Co-annotated terms (annotation intersections) were identified for all pairs of "high level process" (slim) GO terms
Can see that many intersections are zero (cerise) and some are low (red)

Step 2

Step 2 :
Create rules for "zero" intersects
(based on known biology)

GO:0006520,GO:0006310, NO OVERLAP
GO:0006520,GO:0006260, NO OVERLAP
GO:0006520,GO:0030437, NO OVERLAP

"x is not usually co-annotated with y".

Step 3

GO:0006520,GO:0006310, NO OVERLAP
GO:0006520,GO:0006260, NO OVERLAP
GO:0006520,GO:0030437, NO OVERLAP
GO:0006520,GO:0006281, INTERSECT HAS GO:0006521 OR GO:0006338
GO:0006520,GO:0007155, INTERSECT HAS GO:0006521
GO:0006520,GO:0070882, INTERSECT HAS GO:0004360 OR GO:0004067

Intersections where annotations were sparse were inspected for spurious annotations, which were either corrected, or, if the annotations were validated the rules were extended to allow these exceptions.

For example the intersect between GO:0006520 (amino acid metabolic process) and GO:0006399 (tRNA metabolic process) includes gene products annotated to GO:0043039 (tRNA aminoacylation) OR GO:0031071 (cysteine desulphurase), which are involved in both processes
More examples here
:https://sourceforge.net/apps/trac/pombase/wiki/MatrixProject

Step 4

	DNA repair GO:0006281	tRNA metabolic process GO:0006399	translation GO:0006412	cellular amino acid metabolic process GO:0006520	vitamin metabolic process GO:0006766	ribosome biogenesis GO:0042254	transmembrane transport GO:0055085	nitrogen cycle metabolic process GO:0071941
DNA repair GO:0006281 GO:0006281	197	0	4	1	0	4	0	0
tRNA metabolic process GO:0006399 GO:0006399	0	154	46	37	1	32	0	0
translation GO:0006412 GO:0006412	4	46	631	41	0	67	1	0
cellular amino acid metabolic process GO:0006520 GO:0006520	1	37	41	234	14	0	0	11
vitamin metabolic process GO:0006766 GO:0006766	0	1	0	14	41	0	0	0
ribosome biogenesis GO:0042254 GO:0042254	4	32	67	0	0	327	0	0
transmembrane transport GO:0055085 GO:0055085	0	0	1	0	0	0	304	1
nitrogen cycle metabolic process GO:0071941 GO:0071941	0	0	0	11	0	0	1	17

Identify annotations which break rules.

Step 1 :
Identify intersections between pairs of GO terms, using taxonomically diverse set of model species

Step 2 :
Create rules for "zero" intersects (based on known biology)

GO:0006520,GO:0006310, NO OVERLAP
GO:0006520,GO:0006260, NO OVERLAP
GO:0006520,GO:0030437, NO OVERLAP

Step 3 :
Query annotation outliers and Fix annotations in contributing MODS

Add egs

Step 4 :
Periodically identify annotations which break rules across annotation corpus

ANNOTATION QC CYCLE

OR extend rules by allowing specific intersects

GO:0006520,GO:0006310, NO OVERLAP
GO:0006520,GO:0006260, NO OVERLAP
GO:0006520,GO:0030437, NO OVERLAP
GO:0006520,GO:0006281, INTERSECT HAS GO:0006521 OR GO:0006338
GO:0006520,GO:0007155, INTERSECT HAS GO:0006521
GO:0006520,GO:0070882, INTERSECT HAS GO:0004360 OR GO:0004067

Fix annotations….
Extend rules….
Repeat….

Rules so far

for intersections with:

- GO:0006520 cellular amino acid metabolic process
- GO:0006399 tRNA metabolic process
- GO:0042254 ribosome biogenesis
- GO:0006766 vitamin metabolic process
- GO:0007346 regulation of mitotic cell cycle
- GO:0006281 DNA repair
- GO:0006260 DNA replication

http://build.berkeleybop.org/job/check-shared-annotations/lastBuild/console
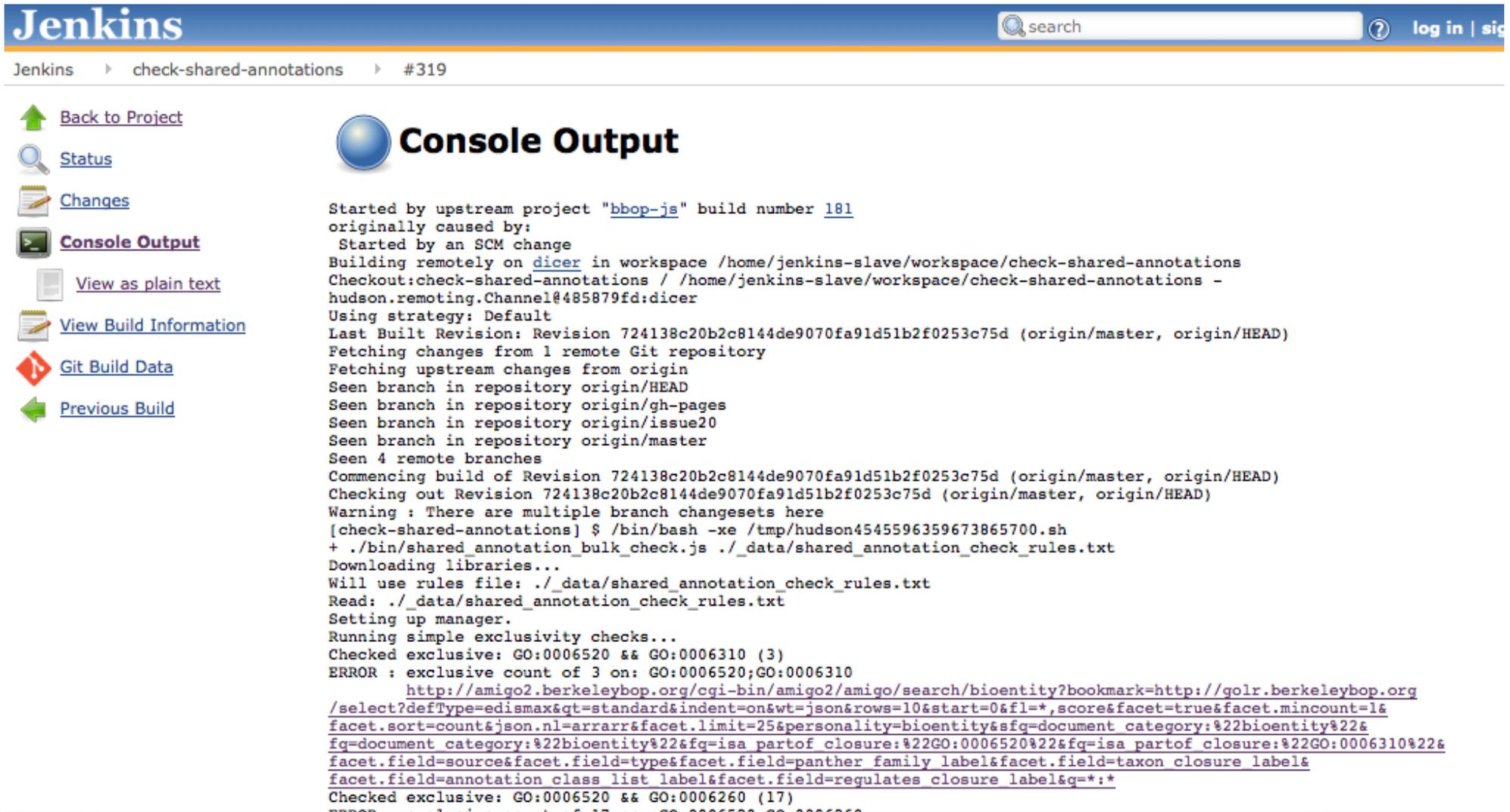
Jenkins

search | log in | si

Jenkins ▸ check-shared-annotations ▸ #319

⬆ Back to Project
🔍 Status
📝 Changes
▶ **Console Output**
　📄 View as plain text
📝 View Build Information
🔴 Git Build Data
⬅ Previous Build

🔵 **Console Output**

```
Started by upstream project "bbop-js" build number 181
originally caused by:
 Started by an SCM change
Building remotely on dicer in workspace /home/jenkins-slave/workspace/check-shared-annotations
Checkout:check-shared-annotations / /home/jenkins-slave/workspace/check-shared-annotations -
hudson.remoting.Channel@485879fd:dicer
Using strategy: Default
Last Built Revision: Revision 724138c20b2c8144de9070fa91d51b2f0253c75d (origin/master, origin/HEAD)
Fetching changes from 1 remote Git repository
Fetching upstream changes from origin
Seen branch in repository origin/HEAD
Seen branch in repository origin/gh-pages
Seen branch in repository origin/issue20
Seen branch in repository origin/master
Seen 4 remote branches
Commencing build of Revision 724138c20b2c8144de9070fa91d51b2f0253c75d (origin/master, origin/HEAD)
Checking out Revision 724138c20b2c8144de9070fa91d51b2f0253c75d (origin/master, origin/HEAD)
Warning : There are multiple branch changesets here
[check-shared-annotations] $ /bin/bash -xe /tmp/hudson4545596359673865700.sh
+ ./bin/shared_annotation_bulk_check.js ./_data/shared_annotation_check_rules.txt
Downloading libraries...
Will use rules file: ./_data/shared_annotation_check_rules.txt
Read: ./_data/shared_annotation_check_rules.txt
Setting up manager.
Running simple exclusivity checks...
Checked exclusive: GO:0006520 && GO:0006310 (3)
ERROR : exclusive count of 3 on: GO:0006520;GO:0006310
        http://amigo2.berkeleybop.org/cgi-bin/amigo2/amigo/search/bioentity?bookmark=http://golr.berkeleybop.org
/select?defType=edismax&qt=standard&indent=on&wt=json&rows=10&start=0&fl=*,score&facet=true&facet.mincount=1&
facet.sort=count&json.nl=arrarr&facet.limit=25&personality=bioentity&sfq=document_category:%22bioentity%22&
fq=document_category:%22bioentity%22&fq=isa_partof_closure:%22GO:0006520%22&fq=isa_partof_closure:%22GO:0006310%22&
facet.field=source&facet.field=type&facet.field=panther_family_label&facet.field=taxon_closure_label&
facet.field=annotation_class_list_label&facet.field=regulates_closure_label&q=*:*
Checked exclusive: GO:0006520 && GO:0006260 (17)
```

http://build.berkeleybop.org/job/check-shared-annotations/lastBuild/console

```
facet.field=annotation_class_list_label&facet.field=regulates_closure_label&q=*:*
Checked exclusion: GO:0042254 && GO:0006091 && !(GO:0023052) (108)
ERROR : bad co-annotations for: GO:0042254;GO:0006091;GO:0023052
        http://amigo2.berkeleybop.org/cgi-bin/amigo2/amigo/search/bioentity?bookmark=http://golr.berkeleybop.org
/select?defType=edismax&qt=standard&indent=on&wt=json&rows=10&start=0&fl=*,score&facet=true&facet.mincount=1&facet.sort=count&
json.nl=arrarr&facet.limit=25&personality=bioentity&sfq=document_category:%22bioentity%22&fq=document_category:%22bioentity%22&
fq=isa_partof_closure:%22GO:0042254%22&fq=isa_partof_closure:%22GO:0006091%22&fq=-isa_partof_closure:%22GO:0023052%22&
facet.field=source&facet.field=type&facet.field=panther_family_label&facet.field=taxon_closure_label&
facet.field=annotation_class_list_label&facet.field=regulates_closure_label&q=*:*
Checked exclusion: GO:0042254 && GO:0007126 && !(GO:0023052) (31)
ERROR : bad co-annotations for: GO:0042254;GO:0007126;GO:0023052
        http://amigo2.berkeleybop.org/cgi-bin/amigo2/amigo/search/bioentity?bookmark=http://golr.berkeleybop.org
/select?defType=edismax&qt=standard&indent=on&wt=json&rows=10&start=0&fl=*,score&facet=true&facet.mincount=1&facet.sort=count&
json.nl=arrarr&facet.limit=25&personality=bioentity&sfq=document_category:%22bioentity%22&fq=document_category:%22bioentity%22&
fq=isa_partof_closure:%22GO:0042254%22&fq=isa_partof_closure:%22GO:0007126%22&fq=-isa_partof_closure:%22GO:0023052%22&
facet.field=source&facet.field=type&facet.field=panther_family_label&facet.field=taxon_closure_label&
facet.field=annotation_class_list_label&facet.field=regulates_closure_label&q=*:*
Checked exclusion; trivially passed with no base overlap: GO:0055085 && GO:0000910 (0)
Checked exclusion: GO:0006766 && GO:0030163 && !(GO:0070647) (8)
ERROR : bad co-annotations for: GO:0006766;GO:0030163;GO:0070647
        http://amigo2.berkeleybop.org/cgi-bin/amigo2/amigo/search/bioentity?bookmark=http://golr.berkeleybop.org
/select?defType=edismax&qt=standard&indent=on&wt=json&rows=10&start=0&fl=*,score&facet=true&facet.mincount=1&facet.sort=count&
json.nl=arrarr&facet.limit=25&personality=bioentity&sfq=document_category:%22bioentity%22&fq=document_category:%22bioentity%22&
fq=isa_partof_closure:%22GO:0006766%22&fq=isa_partof_closure:%22GO:0030163%22&fq=-isa_partof_closure:%22GO:0070647%22&
facet.field=source&facet.field=type&facet.field=panther_family_label&facet.field=taxon_closure_label&
facet.field=annotation_class_list_label&facet.field=regulates_closure_label&q=*:*
Checked exclusion; trivially passed with no base overlap: GO:0006766 && GO:0070647 (0)
Checked exclusion: GO:0006281 && GO:0006913 && !(GO:0031080) (79)
ERROR : bad co-annotations for: GO:0006281;GO:0006913;GO:0031080
        http://amigo2.berkeleybop.org/cgi-bin/amigo2/amigo/search/bioentity?bookmark=http://golr.berkeleybop.org
/select?defType=edismax&qt=standard&indent=on&wt=json&rows=10&start=0&fl=*,score&facet=true&facet.mincount=1&facet.sort=count&
json.nl=arrarr&facet.limit=25&personality=bioentity&sfq=document_category:%22bioentity%22&fq=document_category:%22bioentity%22&
fq=isa_partof_closure:%22GO:0006281%22&fq=isa_partof_closure:%22GO:0006913%22&fq=-isa_partof_closure:%22GO:0031080%22&
facet.field=source&facet.field=type&facet.field=panther_family_label&facet.field=taxon_closure_label&
facet.field=annotation_class_list_label&facet.field=regulates_closure_label&q=*:*
Looked at 194 rules.
Completed with 145 broken rule(s).
Finished: SUCCESS
```

Next step?

pombe

All organisms

DNA recombination (GO:0006310)
DNA repair (GO:0006281)
DNA replication (GO:0006260)
ascospore formation (GO:0030437)
carbohydrate metabolic process (GO:0005975)
cell adhesion (GO:0007155)
cellular amino acid metabolic process (GO:0006)
cell wall organization or biogenesis (GO:0071555)
chromatin modification (GO:0016568)
chromosome organization (GO:0051276)
chromosome segregation (GO:0007059)
cofactor metabolic process (GO:0051186)
conjugation with cellular fusion (GO:0000747)
cytokinesis (GO:0000910)
cytoplasmic translation (GO:0002181)
cytoskeleton organization (GO:0007010)
establishment or maintenance of cell polarity (G
generation of precursor metabolites and energy)
lipid metabolic process (GO:0006629)
mRNA metabolic process (GO:0016071)
meiotic nuclear division (GO:0007126)
mitochondrion organization (GO:0007005)
nitrogen cycle metabolic process (GO:0071941
nucleobase-containing small molecule metabol
nucleocytoplasmic transport (GO:0006913)
peroxisome organization (GO:0007031)
protein catabolic process (GO:0030163)
protein complex assembly (GO:0006461)
protein folding (GO:0006457)
protein glycosylation (GO:0006486)
protein maturation (GO:0051604)
protein modification by small protein conjugatic
protein targeting (GO:0006605)
regulation of mitotic cell cycle (GO:0007346)
ribosome biogenesis (GO:0042254)
signaling (GO:0023052)
tRNA metabolic process (GO:0006399)
transcription, DNA-templated (GO:0006351)
transmembrane transport (GO:0055085)
vacuole organization (GO:0007033)
vesicle-mediated transport (GO:0016192)
vitamin metabolic process (GO:0006766)
embryo development (GO:0009790)

190 298 165   341 749 167 47 637 896 272 90   185 2 1074 340 133 553 171 365 190 15 621 206 9 278 575 50 114 608 433 228 554 100 4461 39 3433 336 68 1037 32 14997

Embryo development

Classes of errors in intersections

A   'Indirect' downstream effects
Phenotype data often doesn't distinguish direct activity and downstream effects, leading to problems with annotations, detected by this system
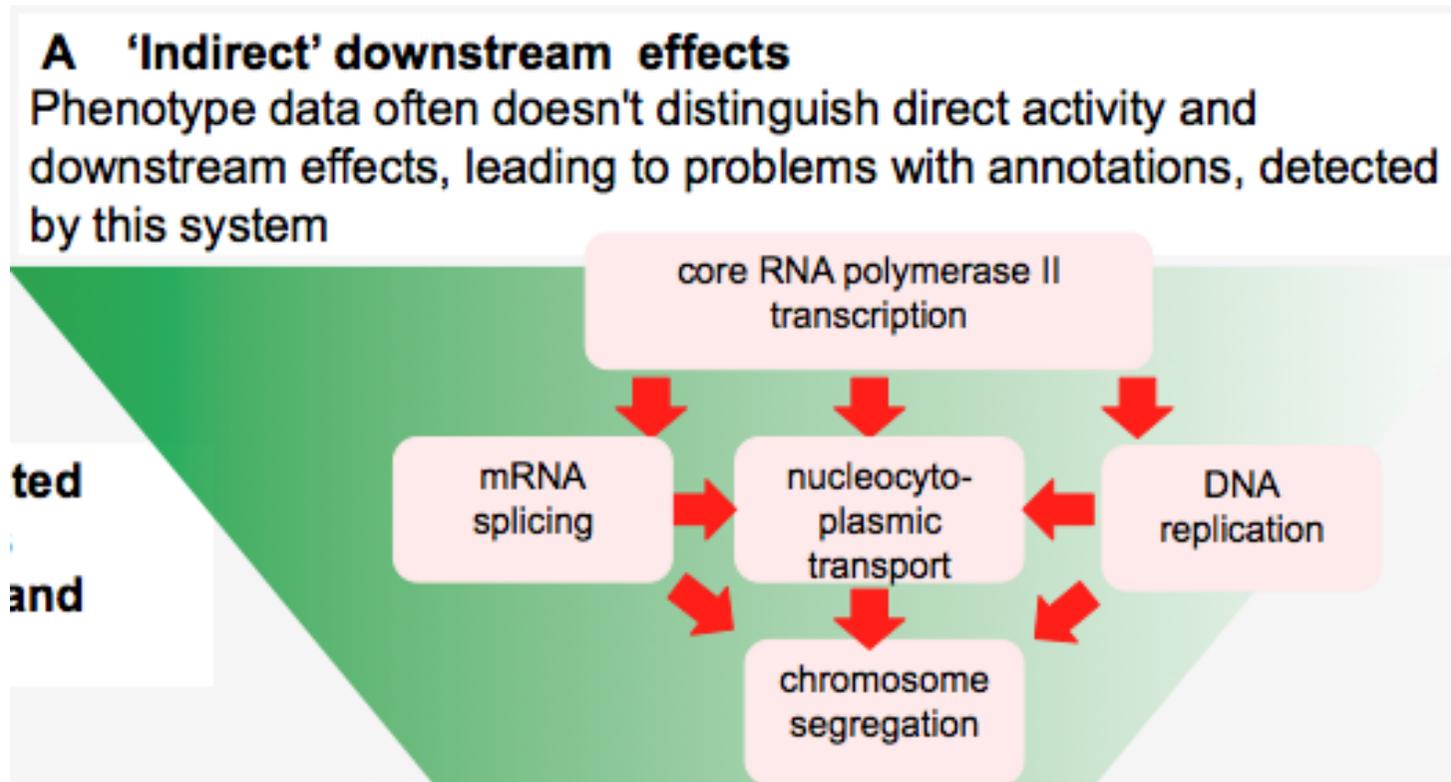
B Automated mapping errors

GO Biological Process

Ontology Graph

Contact Curators

Term ID	Term Name	Evidence	With/From	Reference
GO:0006355	regulation of transcription, DNA-dependent	IEA	KW-0805 X	GO_REF:0000004
GO:0051083	'de novo' cotranslational protein folding	ISO	SGD:S000001236	GO_REF:0000024

GO Cellular Component

Ontology Graph

Contact Curators

Term ID	Term Name	Evidence	With/From	Reference
GO:0005634	nucleus	IEA	KW-0539	GO_REF:0000004
GO:0005829	cytosol	IDA		Matsuyama A et al. (2006)
GO:0005854	nascent polypeptide-associated complex	ISO	SGD:S000001236	GO_REF:0000024
GO:0042788	polysomal ribosome	IDA		Andersen KM et al. (2007)

More example
https://sourceforge.net/tracker/?group_id=36855&atid=605890&source=navbar

C: Manual Annotation Errors

Amino acid metabolism annotation updates

Database	gene	update		reason
PomBase?	fmt1	removed methionyl-tRNA aminoacylation	→	missannotation should be methionyl-tRNA to N-formyl-methionyl-tRNA
PomBase?	SPAC823.09	waiting		
PomBase?	hst4	mitotic sister chromatid segregation		indirect
PomBase?	sla1	removed regulation of conjugation with cellular fusion		indirect
PomBase?	sla1	removed regulation of ascospore formation		indirect
SGD	MET5	removed fungal-type cell wall organization		indirect
SGD	ARG7	removed fungal-type cell wall organization		indirect
SGD	GAP1	ammonia assimilation cycle		indirect
SGD	FMT1	removed methionyl-tRNA aminoacylation	→	error, should be methionyl-tRNA to N-formyl-methionyl-tRNA
SGD	ARG81	replaced arginine metabolic process with "regulation of arginine metabolic process		upstream regulation
SGD	ARG81	removed negative regulation of calcium ion-dependent exocytosis	→	error
SGD	CDC60	removed DNA replication initiation		indirect
SGD	MET18	removed methionine metabolic process		indirect
SGD	THR4	removed endocytosis		indirect
SGD	FPR1	replaced arginine metabolic process with "regulation of arginine metabolic process		upstream regulation
SGD	CBF1	methionine biosynthetic process		indirect
MGI	Sod1	query nucleotide excision repair IMP	→	error, incorrect interpretation
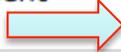
D Ontology True Path Violations

GO:0006606 protein import into nucleus
removed parent
GO:0044743 intracellular protein transmembrane import

GO:0042307 positive regulation of protein import into nucleus
remove parent
GO:0034762 regulation of transmembrane transport

GO:0010559 regulation of glycoprotein biosynthetic process
remove parent
GO:0006417 regulation of translation

E: Paradigm shifts and new biology

Allowed Intersections
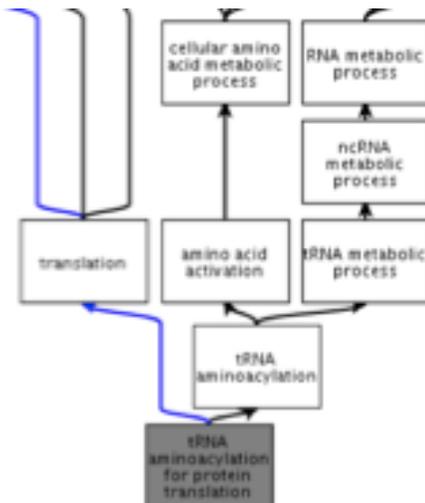
1. Multiple Parentage
Specific function/process is involved in more than one process simultaneously, Link between processes is encoded in the ontology

e.g. tRNA ligase is a child of tRNA metabolism, amino acid metabolism, translation

e.g. pentose phosphate shunt is a child of nucleotide metabolism, carbohydrate metabolism and coenzyme metabolism

2. Multifunctional Gene Product
gene product or pathway functions independently in the multiple processes

e.g. Noc3: intranuclear transport ; and DNA replication initiation

e.g. Urmylation pathway ADD

e.g. Tandem fusion ADD

3. One process directly regulating another
In most cases are probably indirect downstream effects, but are sometimes regulatory.

e.g. Meiosis specific mRNA splicing

e.g. (eg of regulated transport controlling a process?) ADD

e.g chromatin remodelleing -> DNA replication? FING EG
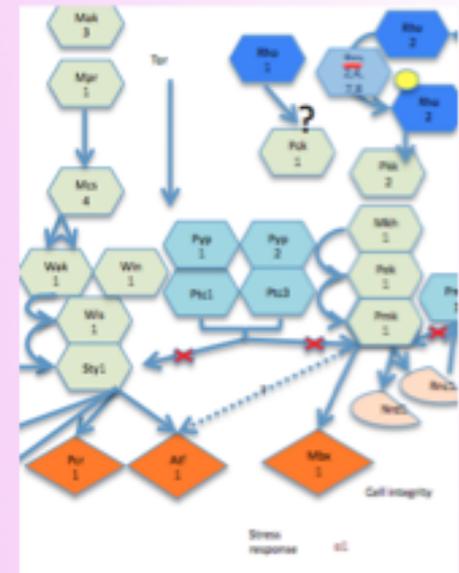
4. Upstream signalling
regulating multiple diverse processes, via protein modification/ transcription

Future work

- Implement rules into GO QC
- Implement rules for intersects with many annotations to cover entire annotation space
- Create rules for within and between other ontology annotations (process/ component, component phenotype etc…)
- Rules can be used to create a set of "negative" annotations for **** (applications? I know some users need these for reasoning…)

Acknowledgements

- Seth Carbon & Chris Mungall (BDGP)
- Antonia Lock and Midori Harris (PomBase)
- Maria Costanzo and ALL SGD curators
- David Hill, Harold Drabkin (MGI)
- InterPro (Everyone)
- GOA (Rachael Huntley, Emily Dimmer)
- UniPROT (Mark Feuermann Ivo Pedruzzi and others)