Reference genome project

Ref genome project overview

- Goal: provide comprehensive annotations for 12 species
- Better annotations: each model organism has unique strengths for probing gene function, and bringing this information together helps to interpret experimental results, which improves the accuracy and consistency of annotations.
- More annotations: homology relationships allow accurate inference of functions for genes that have not been characterized experimentally
- Improvements in the Gene Ontology: cross-organism discussion about annotations frequently leads to new terms being added to the Gene Ontology.

Web presence: http://geneontology.org/GO.refgenome.shtml?all

Annotation targets

- 2006-2007: Disease genes from OMIM
- 2007-2009: added three categories, in addition to (1) diseases : (2) 'hot genes', (3) metabolic pathways, (4) conserved but uncharacterized genes.
- 2009 - onward: text-book style projects
  - Nov 09 - March 10: lung branching morphogenesis
  - July – October 10: wnt signaling pathway

Establishing families

- 2006-2008: Mix of YOGI, OrthoMCL, TreeFam, In paranoid – family membership established by MOD curators
- May 2008- October 2009: PPOD
- November 2009 – onwards: Panther

Defining the Reference genome biocuration projects

Co-current annotation of biological ‘modules’

- Annotation consistency, guidelines, and quality control
- Ontology development
- Enable propagation of annotations via PAINT
- Publicize the Reference Genome initiative
- Provide opportunities to involve experts

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

RefG annotation projects: Requirements & Priorities

- Impact on human biomedical research
- Each project should be feasible in 3-4 months; Split into several subprojects if necessary
- Projects will be publicized on the GO website and as publications
- Encourage external collaborations

Factors in prioritization

- Impact and significance of the pathway:
  - How conserved the process is
  - Has implications in vertebrate biology
  - Things that happen within a single cell (molecular)
- Practical aspects that will help the success of the project:
  - Need a 'project leader' that devote the time to drive it forward
  - There are external experts available and willing to provide feedback

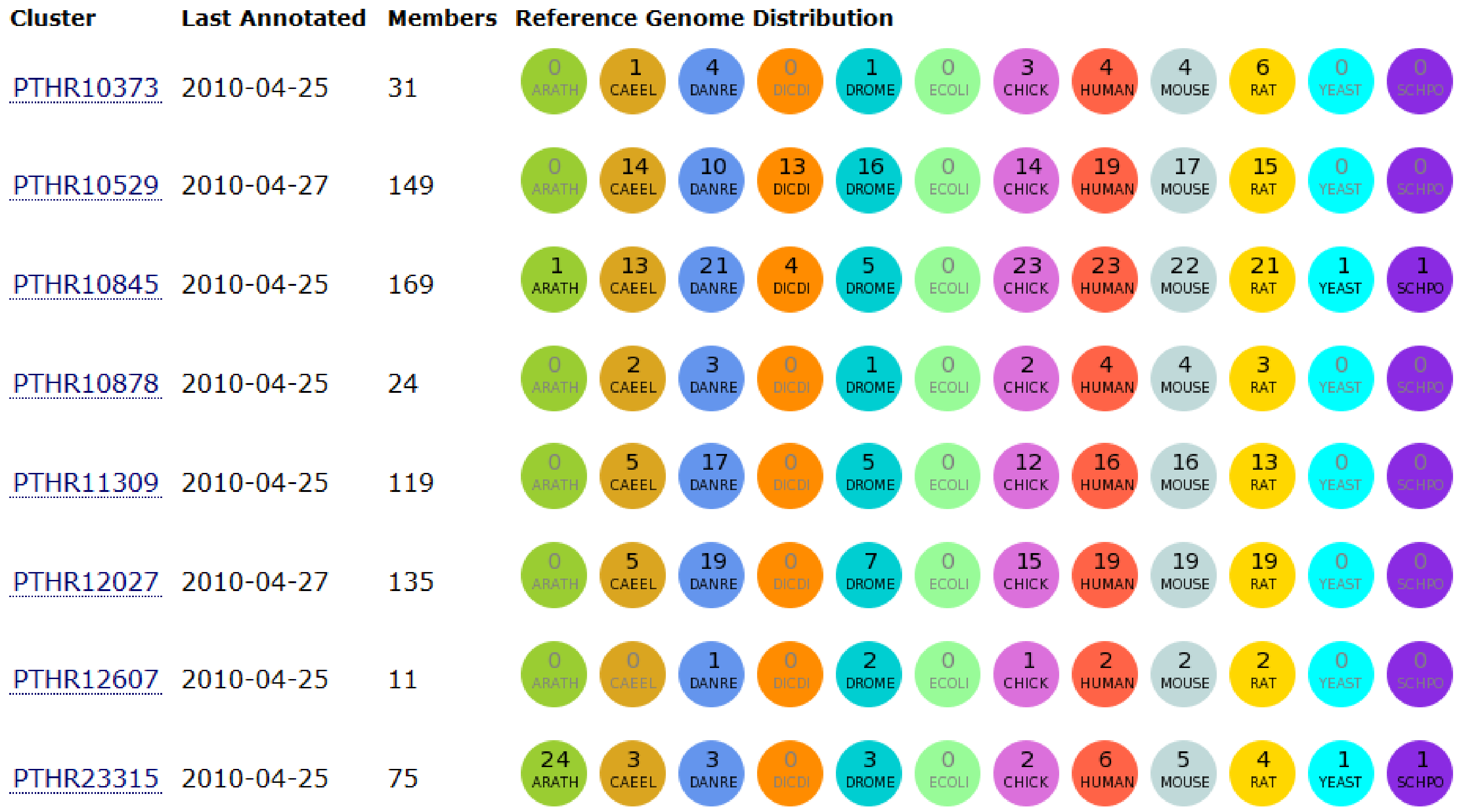
Annotation targets: numbers

- ~ 450 families
- ~ 8,000 proteins from ref genome species
- 11 families (that include annotations for up to 42 species)

Current project: canonical Wnt signalling (July – October 2010)

- Focus on gene products specific to the pathway.
  - Therefore limit to specific Panther subfamilies
- Limit of 2 weeks for the annotation of each Panther family
  - Aim is to comprehensively annotate the gene products as opposed to completely annotate. This is especially relevant for MODs with large volumes of literature
- PAINT curation 2 weeks after annotation period ends
  - Allows time for annotations to filter through to PAINT database
  - Also allows time for recently requested terms to be annotated
- MODs review PAINT annotations
  - Curators to look at PAINT annotations soon after they are available, whilst the initial literature review is still in mind

Wnt Target families: Annotation Progress



	Cluster	Last Annotated	Members	Reference Genome Distribution											
LEF1	PTHR10373	2010-04-25	31	0 ARATH	1 CAEEL	4 DANRE	0 DICDI	1 DROME	0 ECOLI	3 CHICK	4 HUMAN	4 MOUSE	6 RAT	0 YEAST	0 SCHPO
LRP	PTHR10529	2010-04-27	149	0 ARATH	14 CAEEL	10 DANRE	13 DICDI	16 DROME	0 ECOLI	14 CHICK	19 HUMAN	17 MOUSE	15 RAT	0 YEAST	0 SCHPO
Axin	PTHR10845	2010-04-25	169	1 ARATH	13 CAEEL	21 DANRE	4 DICDI	5 DROME	0 ECOLI	23 CHICK	23 HUMAN	22 MOUSE	21 RAT	1 YEAST	1 SCHPO
Dishevelled	PTHR10878	2010-04-25	24	0 ARATH	2 CAEEL	3 DANRE	0 DICDI	1 DROME	0 ECOLI	2 CHICK	4 HUMAN	4 MOUSE	3 RAT	0 YEAST	0 SCHPO
frizzled	PTHR11309	2010-04-25	119	0 ARATH	5 CAEEL	17 DANRE	0 DICDI	5 DROME	0 ECOLI	12 CHICK	16 HUMAN	16 MOUSE	13 RAT	0 YEAST	0 SCHPO
wnt	PTHR12027	2010-04-27	135	0 ARATH	5 CAEEL	19 DANRE	0 DICDI	7 DROME	0 ECOLI	15 CHICK	19 HUMAN	19 MOUSE	19 RAT	0 YEAST	0 SCHPO
APC	PTHR12607	2010-04-25	11	0 ARATH	0 CAEEL	1 DANRE	0 DICDI	2 DROME	0 ECOLI	1 CHICK	2 HUMAN	2 MOUSE	2 RAT	0 YEAST	0 SCHPO
beta-catenin	PTHR23315	2010-04-25	75	24 ARATH	3 CAEEL	3 DANRE	0 DICDI	3 DROME	0 ECOLI	2 CHICK	6 HUMAN	5 MOUSE	4 RAT	1 YEAST	1 SCHPO

http://amigo-sven.princeton.edu/cgi-bin/amigo/phylotree?mode=index&id= PTHR10373+ PTHR12607+ PTHR23315+ PTHR10529+ PTHR10845+ PTHR10878+ PTHR11309+ PTHR12027

Legend:	DONE	DONE by majority of MODs	IN PROGRESS

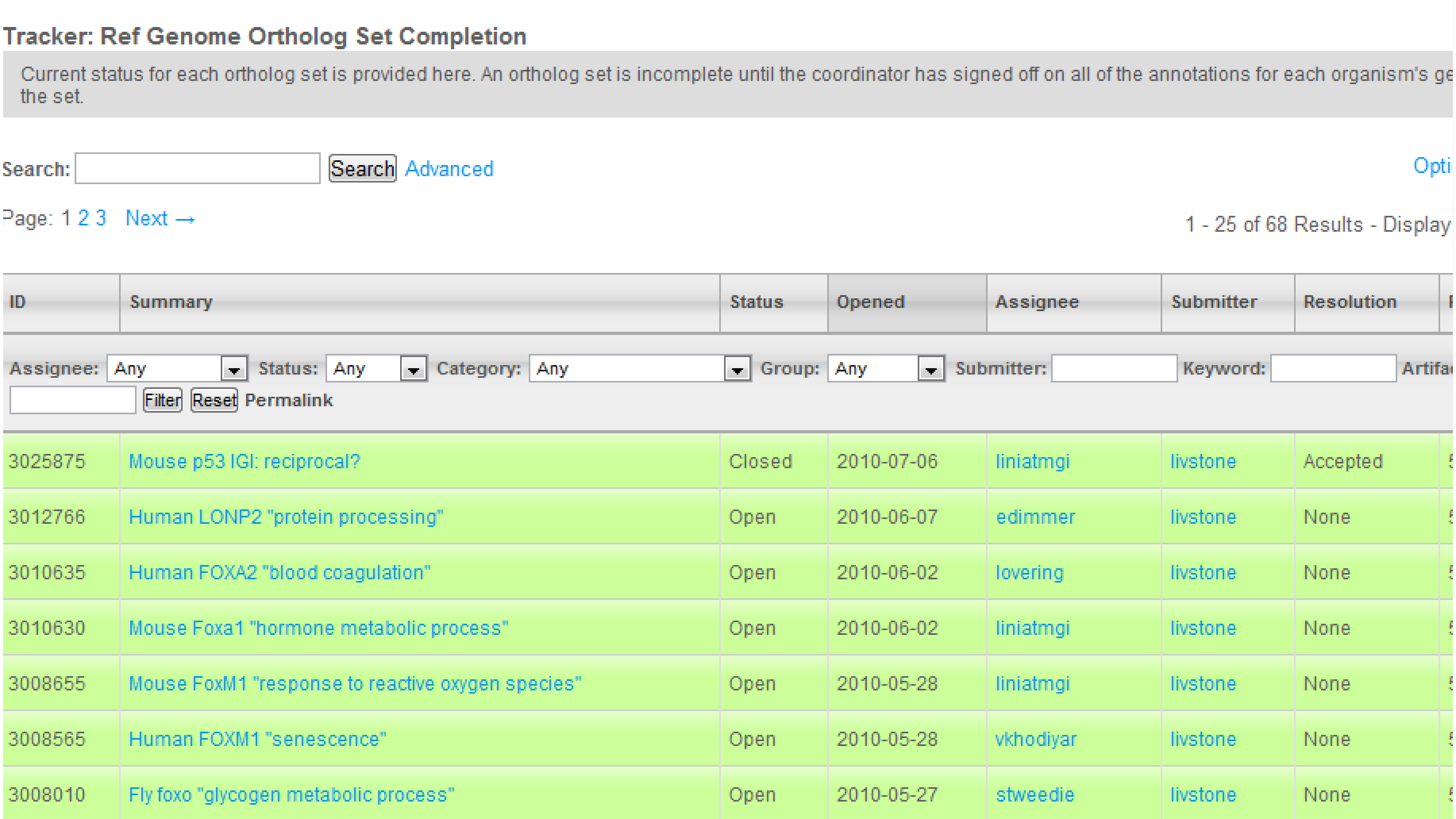
Communication

- Monthly conference calls since June 2007
- Minutes are on GO Wiki
- Email list
- GAF files are on the CVS repository
- Source forge tracker
- Questions/suggestion from tree annotations (from the evidence.txt file) are on the GOC wiki under http://wiki.geneontology.org/index.php/PTHR#####

(continued on next slides)

--- Communication ---

- Source forge tracker (not being used anymore)



Tracker: Ref Genome Ortholog Set Completion

Current status for each ortholog set is provided here. An ortholog set is incomplete until the coordinator has signed off on all of the annotations for each organism's ge
the set.

Search: Search Advanced

Opti

Page: 1 2 3 Next →

1 - 25 of 68 Results - Display

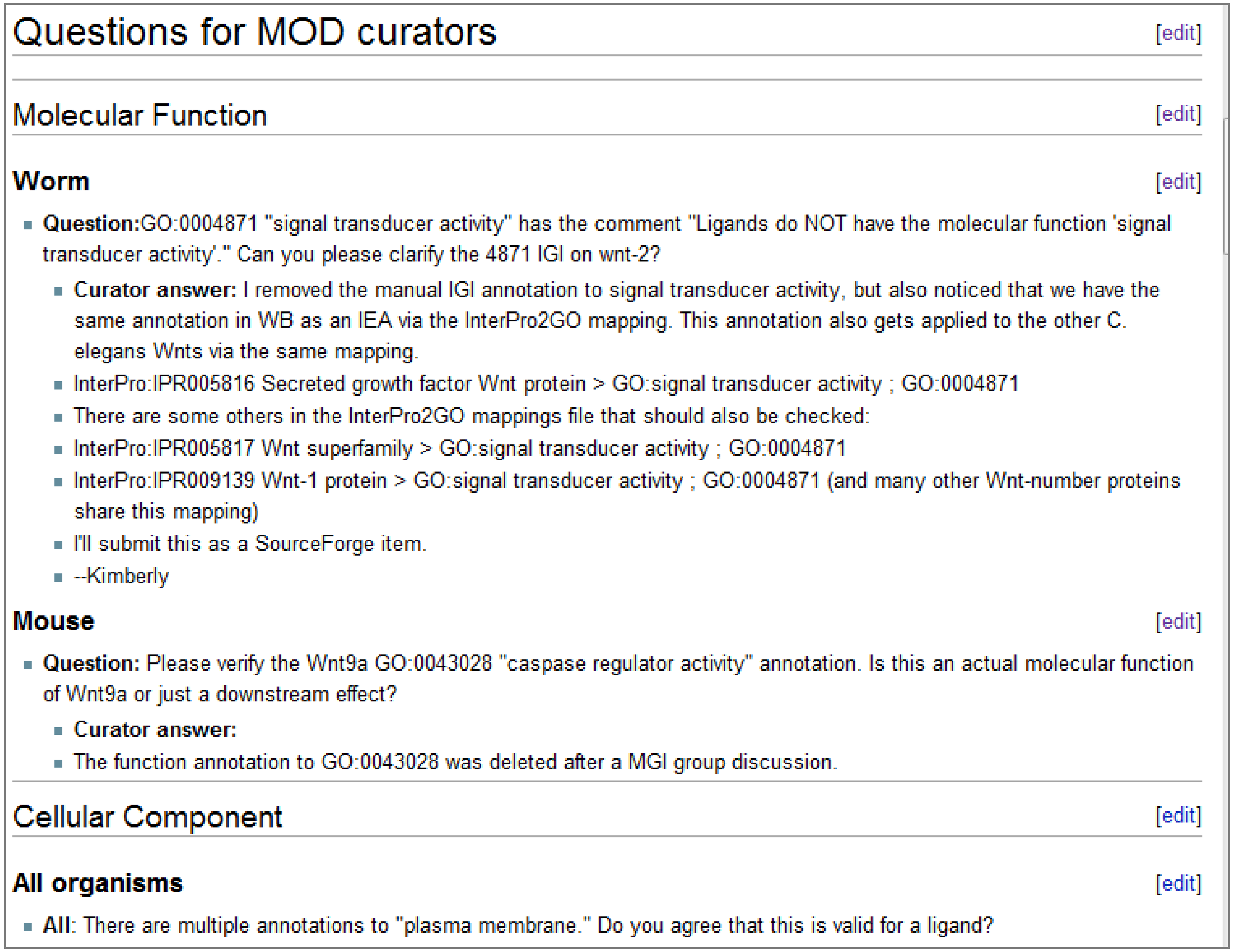
Assignee: Any ▾ Status: Any ▾ Category: Any ▾ Group: Any ▾ Submitter: Keyword: Artifa

Filter Reset Permalink

ID	Summary	Status	Opened	Assignee	Submitter	Resolution
3025875	Mouse p53 IGI: reciprocal?	Closed	2010-07-06	liniatmgi	livstone	Accepted
3012766	Human LONP2 "protein processing"	Open	2010-06-07	edimmer	livstone	None
3010635	Human FOXA2 "blood coagulation"	Open	2010-06-02	lovering	livstone	None
3010630	Mouse Foxa1 "hormone metabolic process"	Open	2010-06-02	liniatmgi	livstone	None
3008655	Mouse FoxM1 "response to reactive oxygen species"	Open	2010-05-28	liniatmgi	livstone	None
3008565	Human FOXM1 "senescence"	Open	2010-05-28	vkhodiyar	livstone	None
3008010	Fly foxo "glycogen metabolic process"	Open	2010-05-27	stweedie	livstone	None

--- Communication ---

- Wiki, see example: http://wiki.geneontology.org/index.php/PTHR12027



Questions for MOD curators [edit]

Molecular Function [edit]

Worm [edit]

- **Question:**GO:0004871 "signal transducer activity" has the comment "Ligands do NOT have the molecular function 'signal transducer activity'." Can you please clarify the 4871 IGI on wnt-2?
  - **Curator answer:** I removed the manual IGI annotation to signal transducer activity, but also noticed that we have the same annotation in WB as an IEA via the InterPro2GO mapping. This annotation also gets applied to the other C. elegans Wnts via the same mapping.
  - InterPro:IPR005816 Secreted growth factor Wnt protein > GO:signal transducer activity ; GO:0004871
  - There are some others in the InterPro2GO mappings file that should also be checked:
  - InterPro:IPR005817 Wnt superfamily > GO:signal transducer activity ; GO:0004871
  - InterPro:IPR009139 Wnt-1 protein > GO:signal transducer activity ; GO:0004871 (and many other Wnt-number proteins share this mapping)
  - I'll submit this as a SourceForge item.
  - --Kimberly

Mouse [edit]

- **Question:** Please verify the Wnt9a GO:0043028 "caspase regulator activity" annotation. Is this an actual molecular function of Wnt9a or just a downstream effect?
  - **Curator answer:**
  - The function annotation to GO:0043028 was deleted after a MGI group discussion.

Cellular Component [edit]

All organisms [edit]

- **All**: There are multiple annotations to "plasma membrane." Do you agree that this is valid for a ligand?

Ref Genome& Ontology development: Source forge requests

- 386 reference genome issues
- 25 issues open - 357 done
- For 2010: ref genome issues are about 10% of all SF term requests

Ref Genome & Annotation Coordination

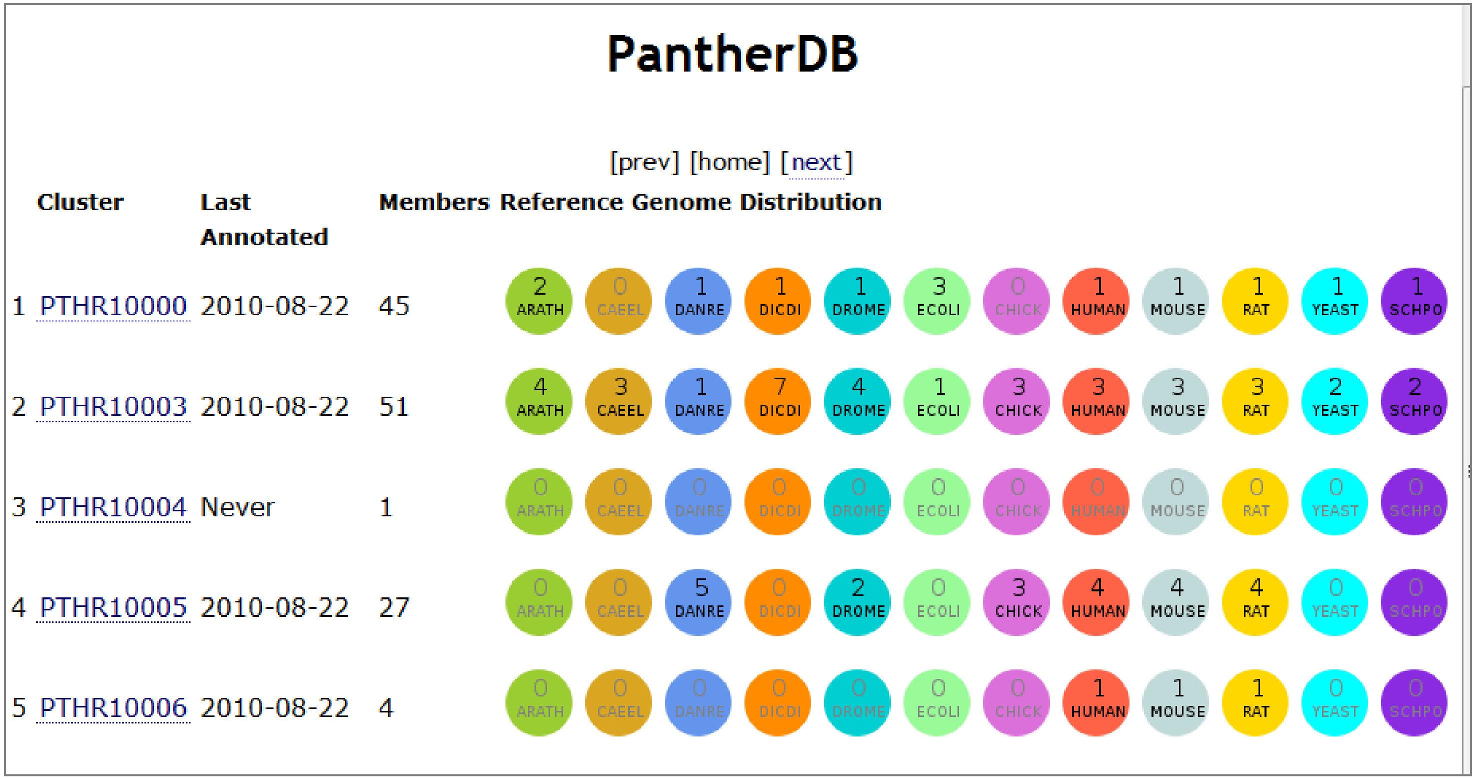
- Electronic annotation jamborees twice / year
- Annotation camp in Geneva June 2010
- 70 attendees: 23 GOC, 40 Swiss-Prot, 10 remote
- Topics:
  - protein binding
  - protein complexes
  - response to
  - regulates
  - downstream events

PAINT:
Protein **A**nnotation **In**ferencing **T**ool

- curation tool for annotating by phylogenetic relationships used by the reference genome project.
- Since 2009 a large part of the effort of the reference genome group is the development of the PAINT tool, which allows to visualize annotations and annotate multiple sequences in a single step.

Annotation tracker tool

- Sven Heinicke, Chris Mungall, Seth Carbon
- Goal: help keep track of progress and prioritize



PantherDB

[prev] [home] [next]

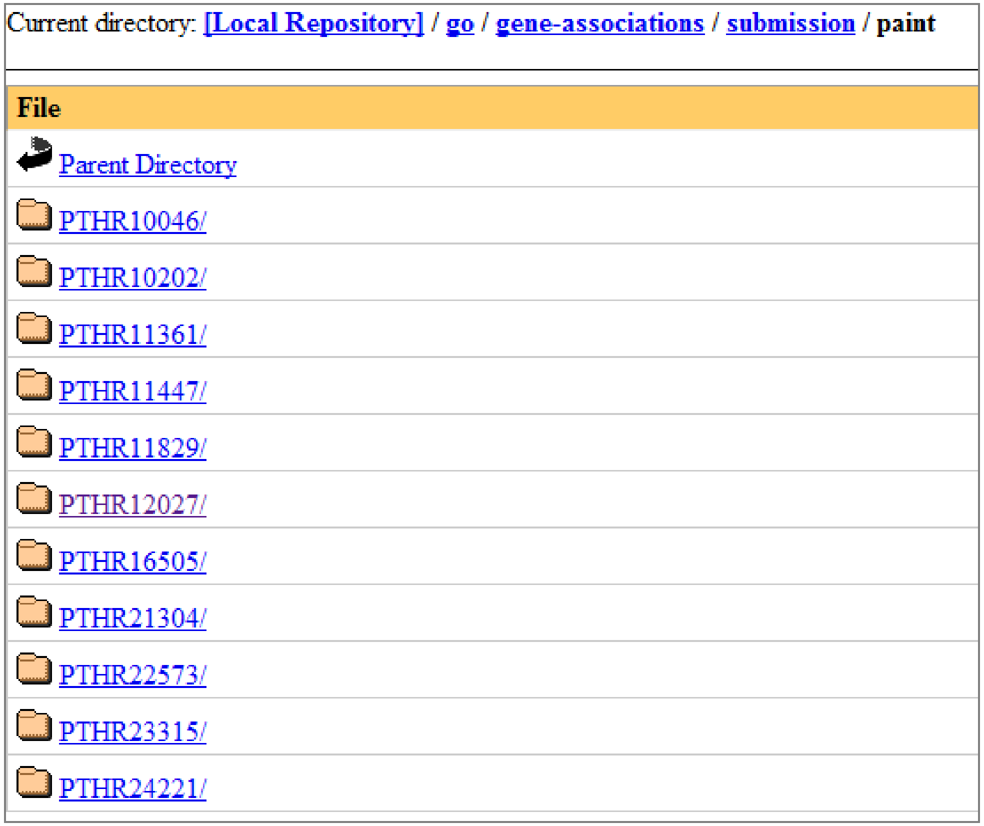
	Cluster	Last Annotated	Members	Reference Genome Distribution											
1	PTHR10000	2010-08-22	45	2 ARATH	0 CAEEL	1 DANRE	1 DICDI	1 DROME	3 ECOLI	0 CHICK	1 HUMAN	1 MOUSE	1 RAT	1 YEAST	1 SCHPO
2	PTHR10003	2010-08-22	51	4 ARATH	3 CAEEL	1 DANRE	7 DICDI	4 DROME	1 ECOLI	3 CHICK	3 HUMAN	3 MOUSE	3 RAT	2 YEAST	2 SCHPO
3	PTHR10004	Never	1	0 ARATH	0 CAEEL	0 DANRE	0 DICDI	0 DROME	0 ECOLI	0 CHICK	0 HUMAN	0 MOUSE	0 RAT	0 YEAST	0 SCHPO
4	PTHR10005	2010-08-22	27	0 ARATH	0 CAEEL	5 DANRE	0 DICDI	2 DROME	0 ECOLI	3 CHICK	4 HUMAN	4 MOUSE	4 RAT	0 YEAST	0 SCHPO
5	PTHR10006	2010-08-22	4	0 ARATH	0 CAEEL	0 DANRE	0 DICDI	0 DROME	0 ECOLI	0 CHICK	1 HUMAN	1 MOUSE	1 RAT	0 YEAST	0 SCHPO

Data availability: files

- GAF files:

http://cvsweb.geneontology.org/cgi-bin/cvsweb.cgi/go/gene-associations/submission/paint/#dirlist
http://cvsweb.geneontology.org/cgi-bin/cvsweb.cgi/go/gene-associations/submission/#dirlist

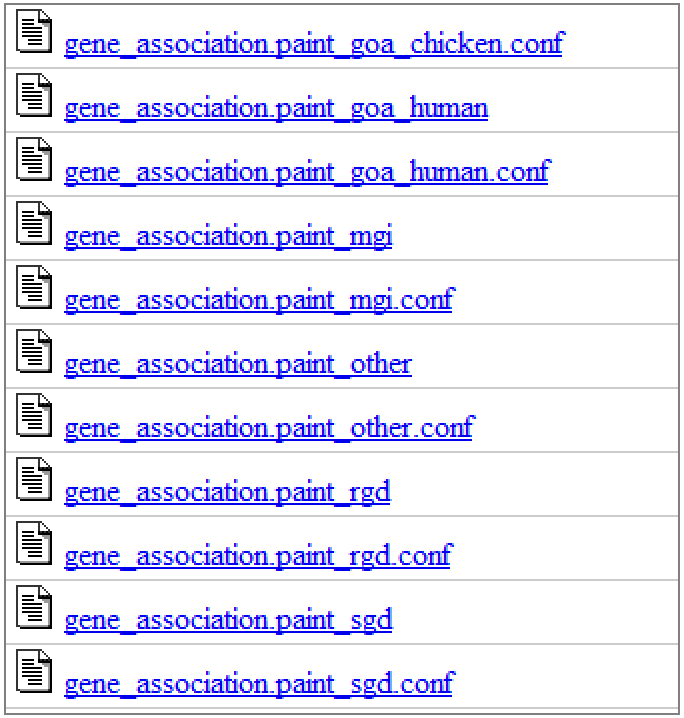
By protein family



Current directory: [Local Repository] / go / gene-associations / submission / paint

File
Parent Directory
PTHR10046/
PTHR10202/
PTHR11361/
PTHR11447/
PTHR11829/
PTHR12027/
PTHR16505/
PTHR21304/
PTHR22573/
PTHR23315/
PTHR24221/

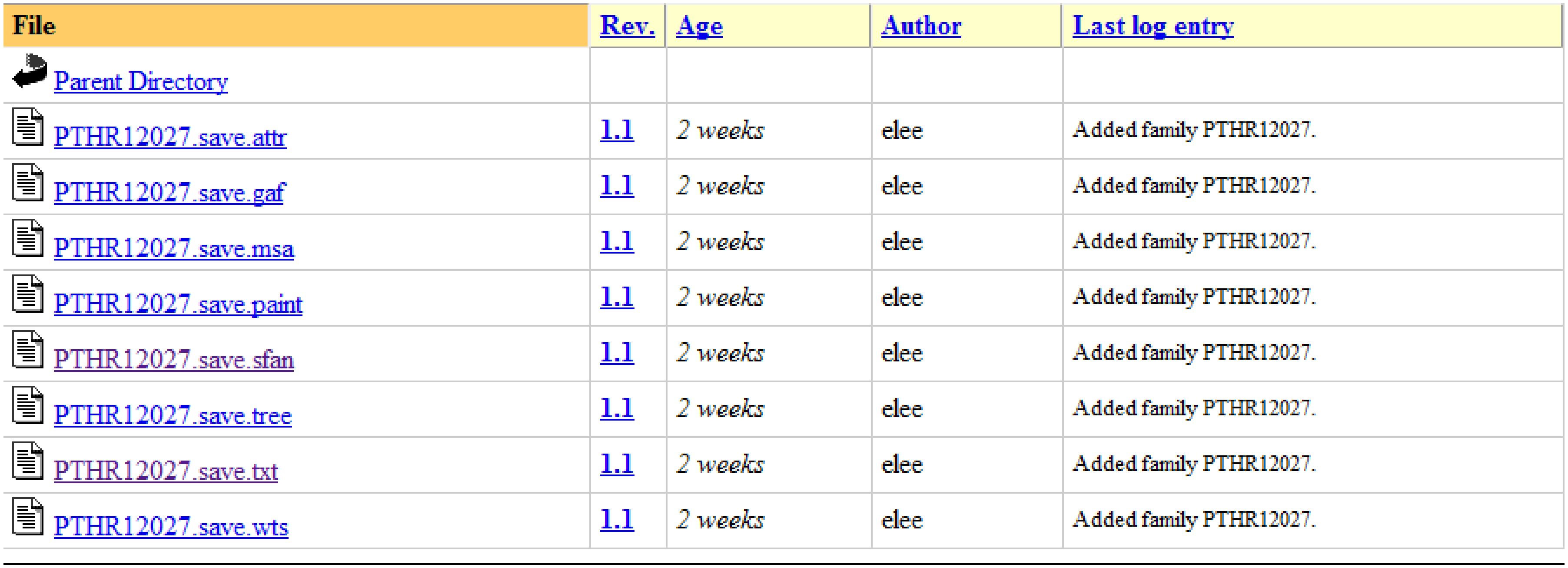
By species



gene_association.paint_goa_chicken.conf
gene_association.paint_goa_human
gene_association.paint_goa_human.conf
gene_association.paint_mgi
gene_association.paint_mgi.conf
gene_association.paint_other
gene_association.paint_other.conf
gene_association.paint_rgd
gene_association.paint_rgd.conf
gene_association.paint_sgd
gene_association.paint_sgd.conf

Contents of the Protein family directory:

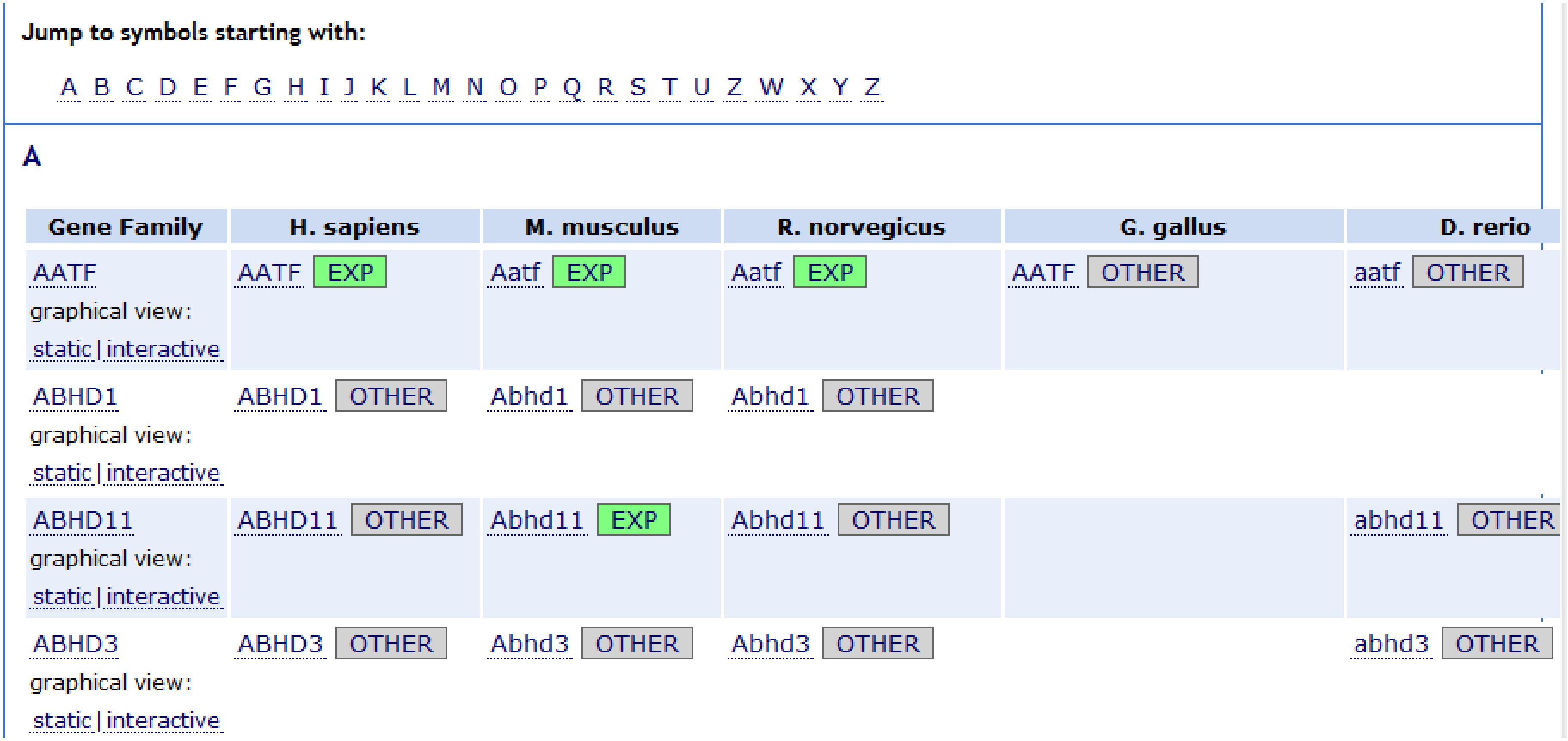
- Files to re-load in PAINT
- GAF file
- Evidence.txt file with description of the tree annotation, comments and questions



File	Rev.	Age	Author	Last log entry
Parent Directory				
PTHR12027.save.attr	1.1	*2 weeks*	elee	Added family PTHR12027.
PTHR12027.save.gaf	1.1	*2 weeks*	elee	Added family PTHR12027.
PTHR12027.save.msa	1.1	*2 weeks*	elee	Added family PTHR12027.
PTHR12027.save.paint	1.1	*2 weeks*	elee	Added family PTHR12027.
PTHR12027.save.sfan	1.1	*2 weeks*	elee	Added family PTHR12027.
PTHR12027.save.tree	1.1	*2 weeks*	elee	Added family PTHR12027.
PTHR12027.save.txt	1.1	*2 weeks*	elee	Added family PTHR12027.
PTHR12027.save.wts	1.1	*2 weeks*	elee	Added family PTHR12027.

Data availability: web view in AmiGO

http://amigo.geneontology.org/cgi-bin/amigo/amigo?mode=homolset_summary



Jump to symbols starting with:

A B C D E F G H I J K L M N O P Q R S T U Z W X Y Z

A

Gene Family	H. sapiens	M. musculus	R. norvegicus	G. gallus	D. rerio
AATF graphical view: static | interactive	AATF EXP	Aatf EXP	Aatf EXP	AATF OTHER	aatf OTHER
ABHD1 graphical view: static | interactive	ABHD1 OTHER	Abhd1 OTHER	Abhd1 OTHER		
ABHD11 graphical view: static | interactive	ABHD11 OTHER	Abhd11 EXP	Abhd11 OTHER		abhd11 OTHER
ABHD3 graphical view: static | interactive	ABHD3 OTHER	Abhd3 OTHER	Abhd3 OTHER		abhd3 OTHER

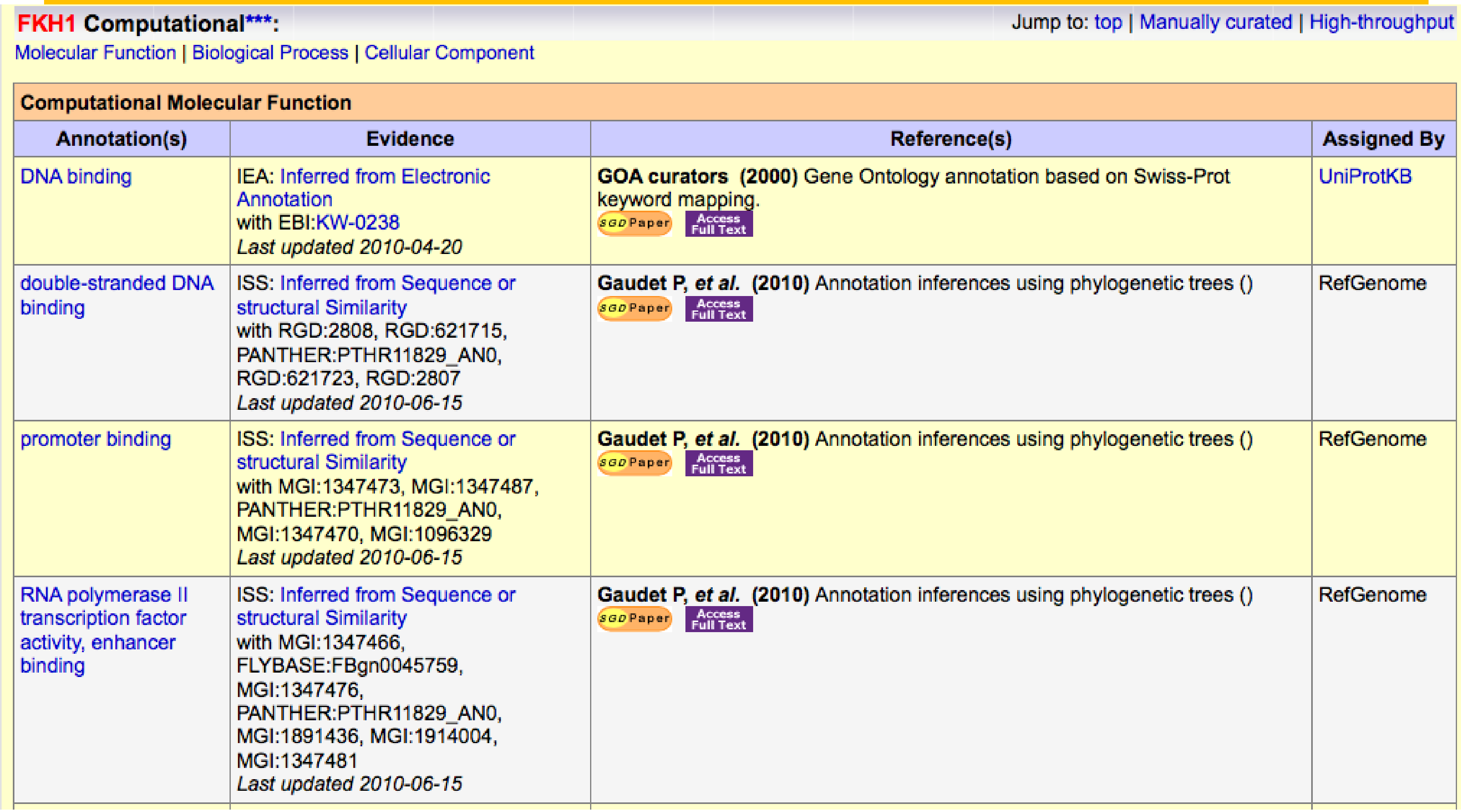
Data availability in MODs

- SGD, MGI, UniProt are uploading GAF files
- Pombe : next week
  - QuickGO has a Ref Genome filter
  http://www.ebi.ac.uk/QuickGO/GAnnotation?&protein=ReferenceGenome
- dictyBase, wormbase, ZFIN are in the process of adapting their tools

PAINT annotations in SGD

- Have a loading script to load annotations
- PANTHR_ID in the ‘with’ column
  - loading is easier
  - Display is simple

PAINT annotations for FHK1 @ SGD



FKH1 Computational***:

Jump to: top | Manually curated | High-throughput

Molecular Function | Biological Process | Cellular Component

Computational Molecular Function

Annotation(s)	Evidence	Reference(s)	Assigned By
DNA binding	IEA: Inferred from Electronic Annotation with EBI:KW-0238 *Last updated 2010-04-20*	**GOA curators (2000)** Gene Ontology annotation based on Swiss-Prot keyword mapping. SGD Paper Access Full Text	UniProtKB
double-stranded DNA binding	ISS: Inferred from Sequence or structural Similarity with RGD:2808, RGD:621715, PANTHER:PTHR11829_AN0, RGD:621723, RGD:2807 *Last updated 2010-06-15*	**Gaudet P, *et al.* (2010)** Annotation inferences using phylogenetic trees () SGD Paper Access Full Text	RefGenome
promoter binding	ISS: Inferred from Sequence or structural Similarity with MGI:1347473, MGI:1347487, PANTHER:PTHR11829_AN0, MGI:1347470, MGI:1096329 *Last updated 2010-06-15*	**Gaudet P, *et al.* (2010)** Annotation inferences using phylogenetic trees () SGD Paper Access Full Text	RefGenome
RNA polymerase II transcription factor activity, enhancer binding	ISS: Inferred from Sequence or structural Similarity with MGI:1347466, FLYBASE:FBgn0045759, MGI:1347476, PANTHER:PTHR11829_AN0, MGI:1891436, MGI:1914004, MGI:1347481 *Last updated 2010-06-15*	**Gaudet P, *et al.* (2010)** Annotation inferences using phylogenetic trees () SGD Paper Access Full Text	RefGenome

MGI

Publicizing the reference genome project

- **Newsletters**: GO, wormbase, FlyBase, ZFIN
- **MODs websites**:

 - GOA http://www.ebi.ac.uk/GOA/RGI/index.html
 - BHF-UCL http://www.ucl.ac.uk/silva/cardiovasculargeneontology
 - TAIR: http://www.arabidopsis.org/portals/genAnnotation/functional_annotation/reference.jsp
 - pombe

- **Search**: expanding our search capabilities so that an User would be able to use search terms like human disease names or human genes

Publications/communications

- PLoS Computational Biology 5(7):e1000431. PMID: 19578431
- A note about the reference genome project was published in the GO news site:http://go.berkeleybop.org/news4go/node/27
- Talks :
  - 3rd Biocurator meeting, Berlin, April 09 (Gaudet)
  - Quest for ortholog meeting, Hinxton, July 09 (Gaudet)
  - Genome Informatics conference, Cold Spring Harbor, NY, October 09 (Thomas)
- Posters
  - 2nd Biocurator meeting, San Jose, CA, October 09 (Tweedie)
  - Dicty meeting, Estes Park, CO, September 09 (Gaudet)
  - Genome Informatics conference, Cold Spring Harbor, NY, October 09 (Livstone)
  - ISMB meeting, Boston, MA, July 10 (Gaudet)

Next 6-12 months

- (PAINT: no full time developer)
- Annotation tracker (see Sven)
- MOD curation: open for discussion http://wiki.geneontology.org/index.php/RefG_annotation_priorities
- Protein family annotation: 1-2/week
- Writing papers

Kara: Princeton progress report

Grant: ideas for reference genome group

- Continue ‘project’ approach but prioritize better
-- poorly covered areas of the ontology
-- poorly annotated areas
-- be more pro-active about collaborations
- Reach out to researchers for input in ontology and annotation
- Uncouple tree annotation from MOD curation

Aim 1: Literature Curation and Annotation

- Develop literature annotation standards and best practices guidelines
- Establish ongoing reviews and quality assessments of the reference annotations, using both manual and automated approaches
- Amend the ontology as needed

Aim 2: Phylogenetic Curation and Annotation

- Develop phylogenetic annotation standards and best practices guidelines.
- Increase the efficiency of phylogenetic curation by providing closer integration with literature-based curation.
- Extend the relevant software used by curators in propagating reference genome annotations to enhance efficiency and accessibility (Web-PAINT, database, web QC services).
- Develop the infrastructure required for the ongoing updates to the phylogenetic-based protein families.

Aim 3: Extend the GO into major new communities

- Establish partnerships with researchers working in major human organ/cellular systems for comprehensive annotation efforts on specific "text-book" style modules. This approach offers a finite list of topics to be covered, and gives community experts a sense of ownership.
- Establish partnerships with bacterial research communities, including the metagenomic communities, and work with them to enable their use of GO