CACAO Training
Part 1

Jim Hu and Suzi Aleksander

Fall 2015

Outline

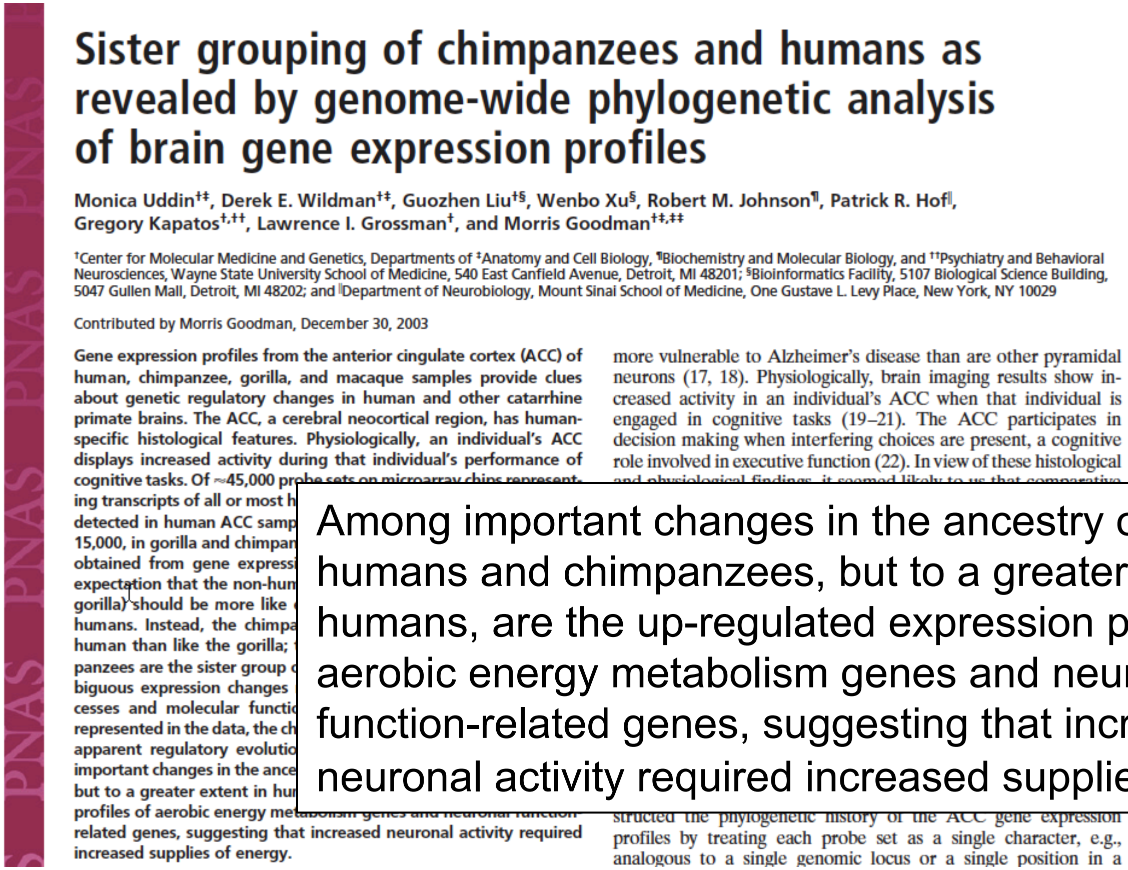
- Part 1: Gene Ontology and functional annotation
  - How known functions are used to reveal new knowledge
  - Gene Ontology
  - What is an annotation?
  - CACAO
- Part 2: Making annotations and challenges

Leveraging what we know about function

Leveraging What We Know About Function

- **Functional profiling**: For a list of genes, what functions are important?
  - Genes turned up or down together
    - Disease states
    - Environmentalresponses
    - Genotypes
    - …
  - Genes encoding proteins that physically interact
  - Genes conserved in specific taxa
  - Genes found in specific microbial communities

Functional Profiling Example



Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles

Monica Uddin[†‡], Derek E. Wildman[†‡], Guozhen Liu[†§], Wenbo Xu[§], Robert M. Johnson[¶], Patrick R. Hof[‖], Gregory Kapatos[†,††], Lawrence I. Grossman[†], and Morris Goodman[†‡,‡‡]

[†]Center for Molecular Medicine and Genetics, Departments of [‡]Anatomy and Cell Biology, [¶]Biochemistry and Molecular Biology, and [††]Psychiatry and Behavioral Neurosciences, Wayne State University School of Medicine, 540 East Canfield Avenue, Detroit, MI 48201; [§]Bioinformatics Facility, 5107 Biological Science Building, 5047 Gullen Mall, Detroit, MI 48202; and [‖]Department of Neurobiology, Mount Sinai School of Medicine, One Gustave L. Levy Place, New York, NY 10029

Contributed by Morris Goodman, December 30, 2003

Gene expression profiles from the anterior cingulate cortex (ACC) of human, chimpanzee, gorilla, and macaque samples provide clues about genetic regulatory changes in human and other catarrhine primate brains. The ACC, a cerebral neocortical region, has human-specific histological features. Physiologically, an individual's ACC displays increased activity during that individual's performance of cognitive tasks. Of ≈45,000 probe sets on microarray chips represent-
ing transcripts of all or most h
detected in human ACC samp
15,000, in gorilla and chimpan
obtained from gene expressi
expectation that the non-hum
gorilla) should be more like
humans. Instead, the chimpa
human than like the gorilla;
panzees are the sister group
biguous expression changes
cesses and molecular functi
represented in the data, the ch
apparent regulatory evolutio
important changes in the ance
but to a greater extent in hu
profiles of aerobic energy metabolism genes and neuronal function-related genes, suggesting that increased neuronal activity required increased supplies of energy.

more vulnerable to Alzheimer's disease than are other pyramidal neurons (17, 18). Physiologically, brain imaging results show increased activity in an individual's ACC when that individual is engaged in cognitive tasks (19–21). The ACC participates in decision making when interfering choices are present, a cognitive role involved in executive function (22). In view of these histological and physiological findings, it seemed likely to us that comparative

structed the phylogenetic history of the ACC gene expression profiles by treating each probe set as a single character, e.g., analogous to a single genomic locus or a single position in a

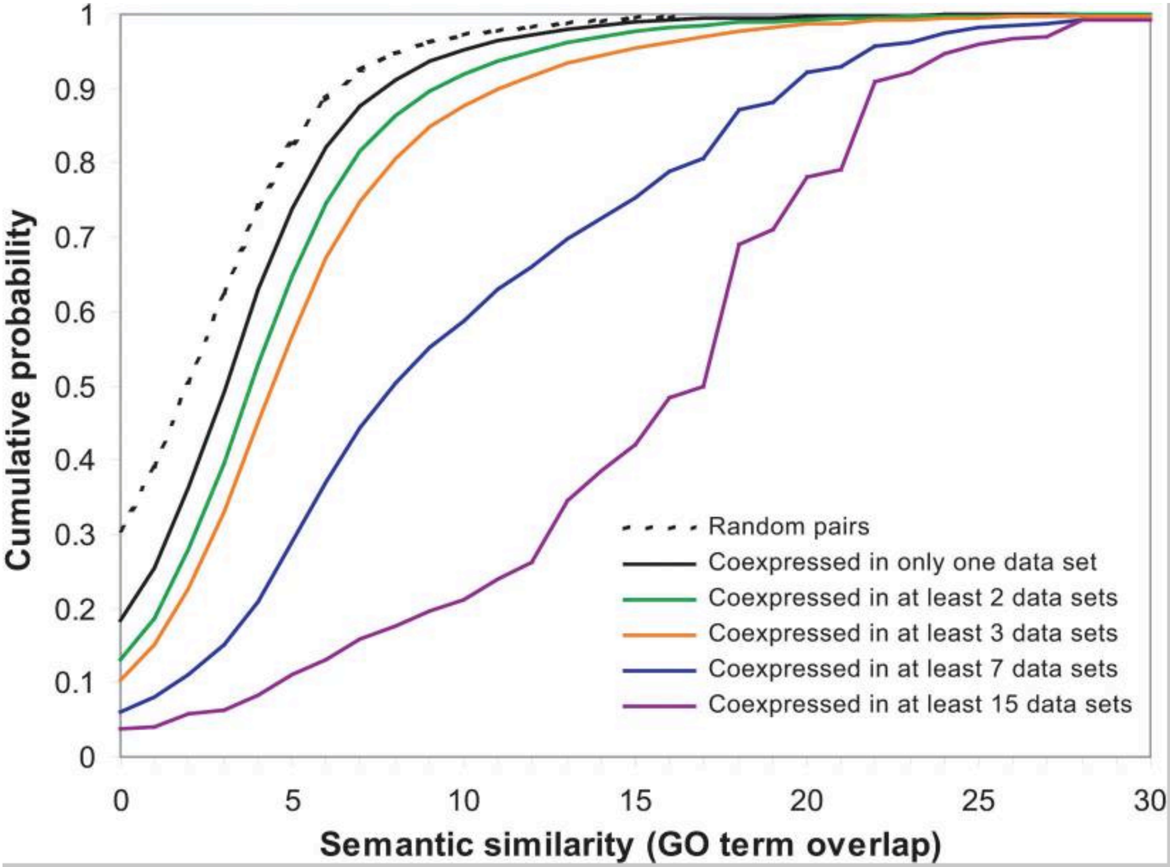
Among important changes in the ancestry of both humans and chimpanzees, but to a greater extent in humans, are the up-regulated expression profiles of aerobic energy metabolism genes and neuronal function-related genes, suggesting that increased neuronal activity required increased supplies of energy.

Uddin et al. (2004) PNAS 101:2957-2962

Leveraging What We Know About Function

- **Guilt by association**: For a gene of unknown function, can we infer its function from genes of known function:
  - that are coexpressed across many conditions
  - that are homologs
  - that are coinherited across evolution
  - that physically interact in a multiprotein complex

Coexpression Correlates with Common Function



Lee et al. (2004) Genome Research 14:1085-1094

What Do We Mean by Function?

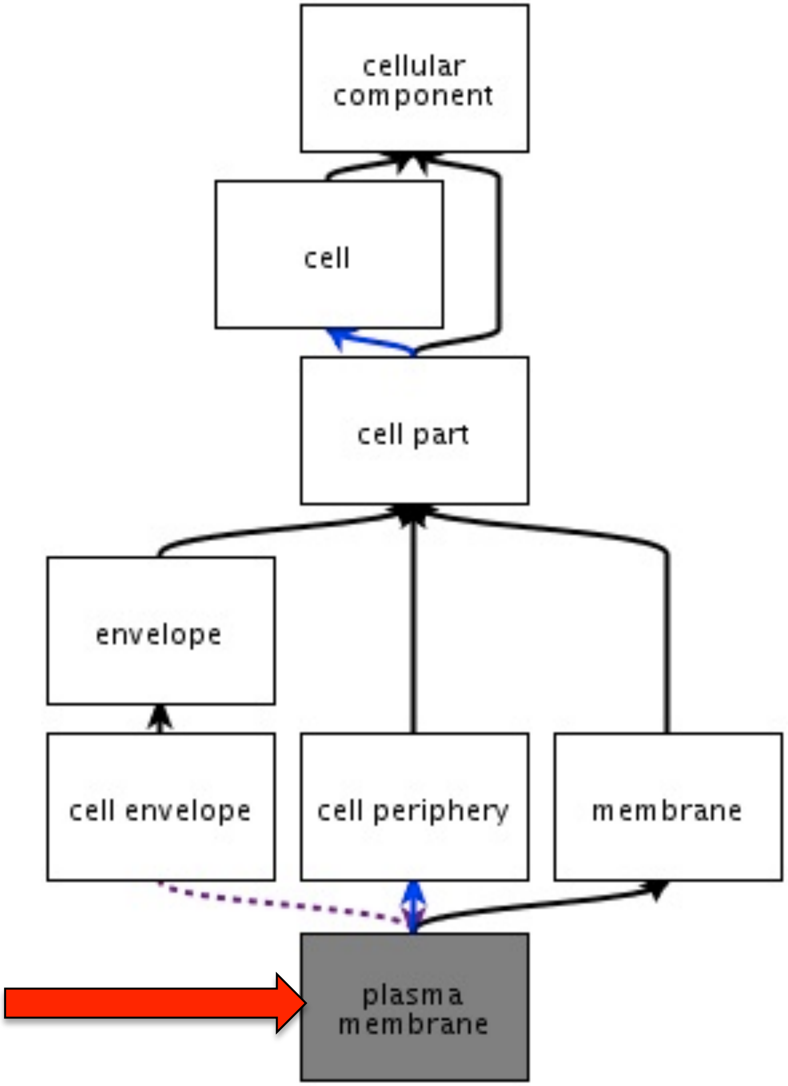
- Massive body of published knowledge
  - Almost useless by itself!!
- We need
  - Knowledge that computers can analyze
  - Common vocabulary across different organisms
  - Disambiguation of synonyms
  - Connection of related ideas that are more or less specific
    - Examples:
      - polygon – quadrilateral – rectangle – square
      - Enzyme – kinase – protein kinase – protein tyrosine kinase

Gene Ontology

Gene Ontology (GO)

- 3 aspects (ontologies) :
  - Molecular Function
  - Biological Process
  - Cellular Component
- Controlled vocabulary
  - ID number for computers
  - Name and definition for humans
- Relationships

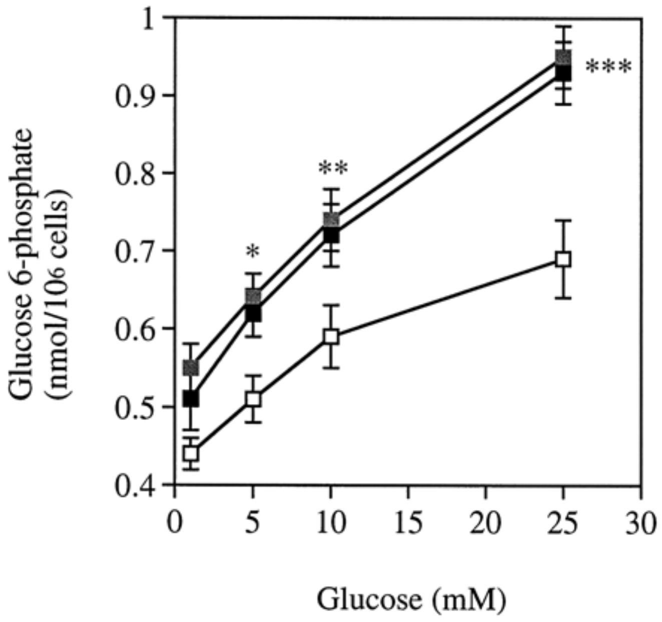
GO:0005886

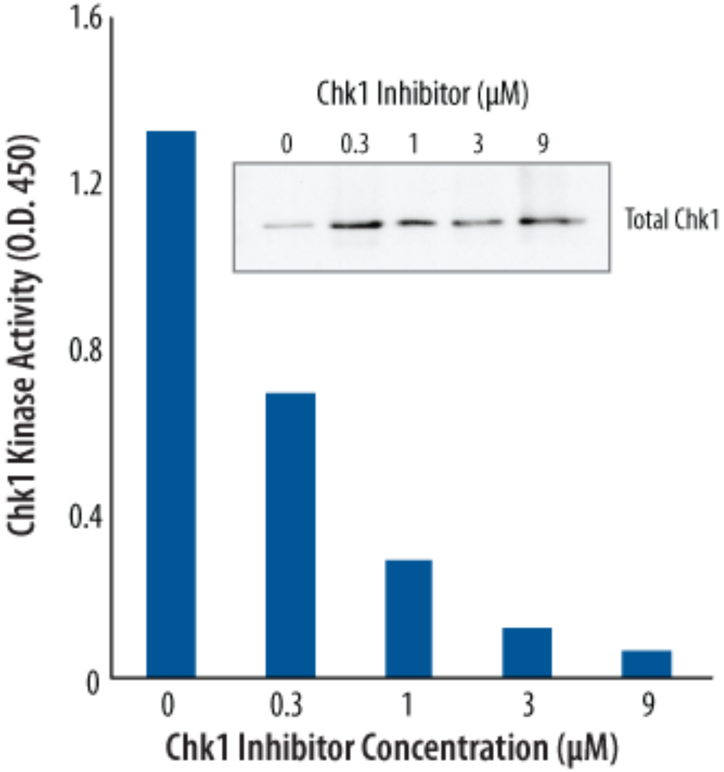


QuickGO - http://www.ebi.ac.uk/QuickGO

Molecular Function

- activities = what a protein can do by itself





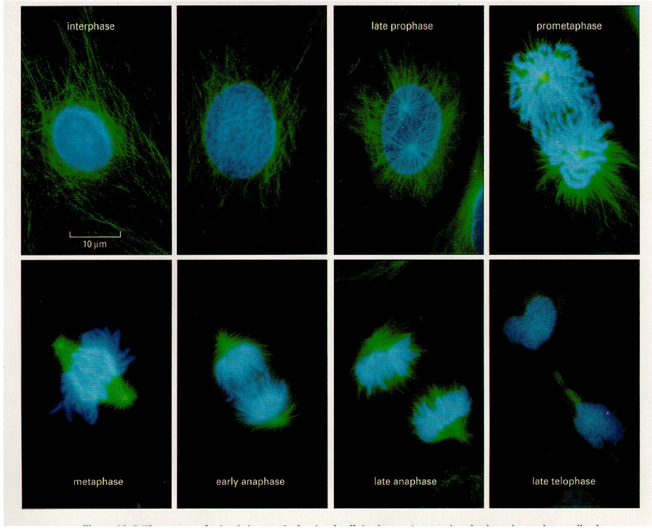
GO:0004347 hexokinase activity

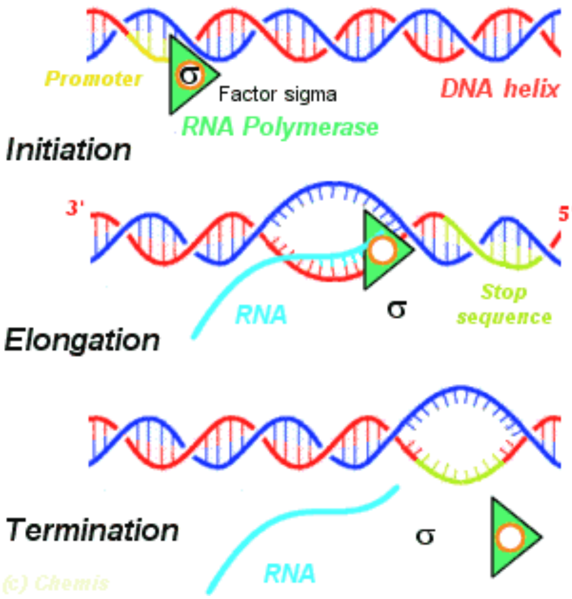
GO:0016301 Kinase activity

From PMID:9341134, rndsystems.com

Biological Process

- a commonly recognized series of events
  - Including, but not just biochemical pathways





GO:0051301
cell division

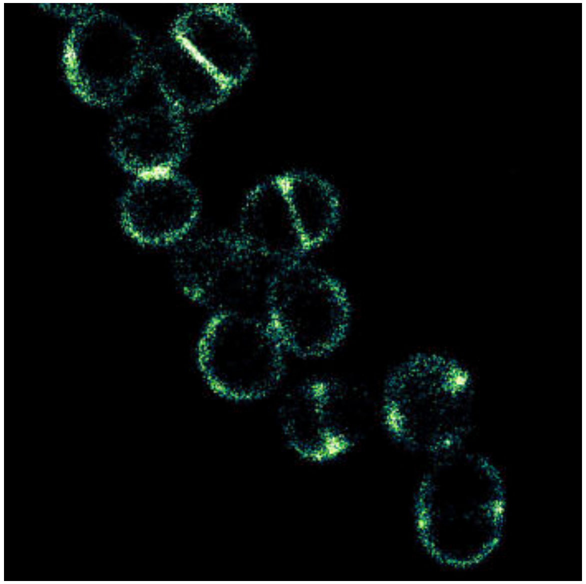
GO:0006351
transcription, DNA dependent

From ridge.icu.ac.jp,

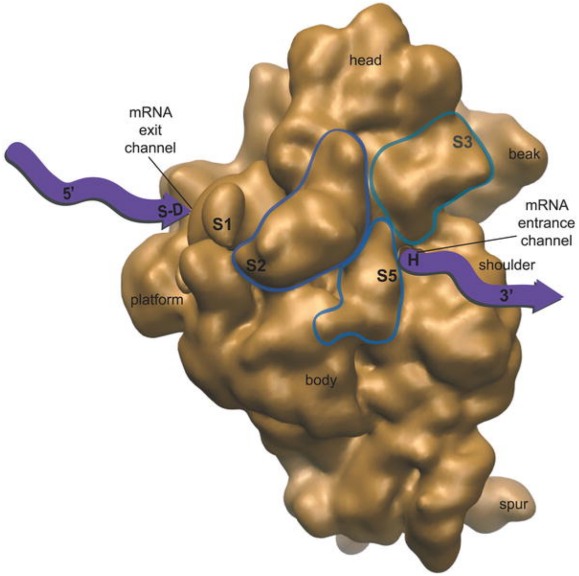
Cellular Component

- where a gene product acts
  - Subcellular location
  - Multicomponent complex

GO:0005739
mitochondrion



GO:0009274
peptidoglycan-based cell wall

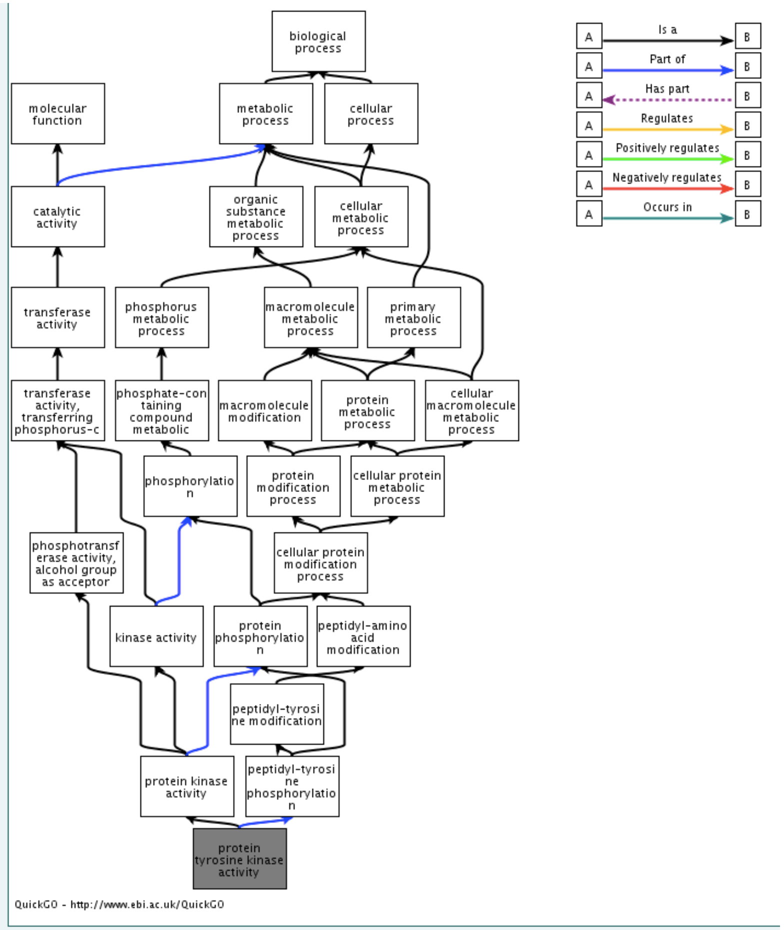


GO:0005840
ribosome

From visualphotos.com, epmm.group.shef.ac.uk, wikimedia commons

GO terms

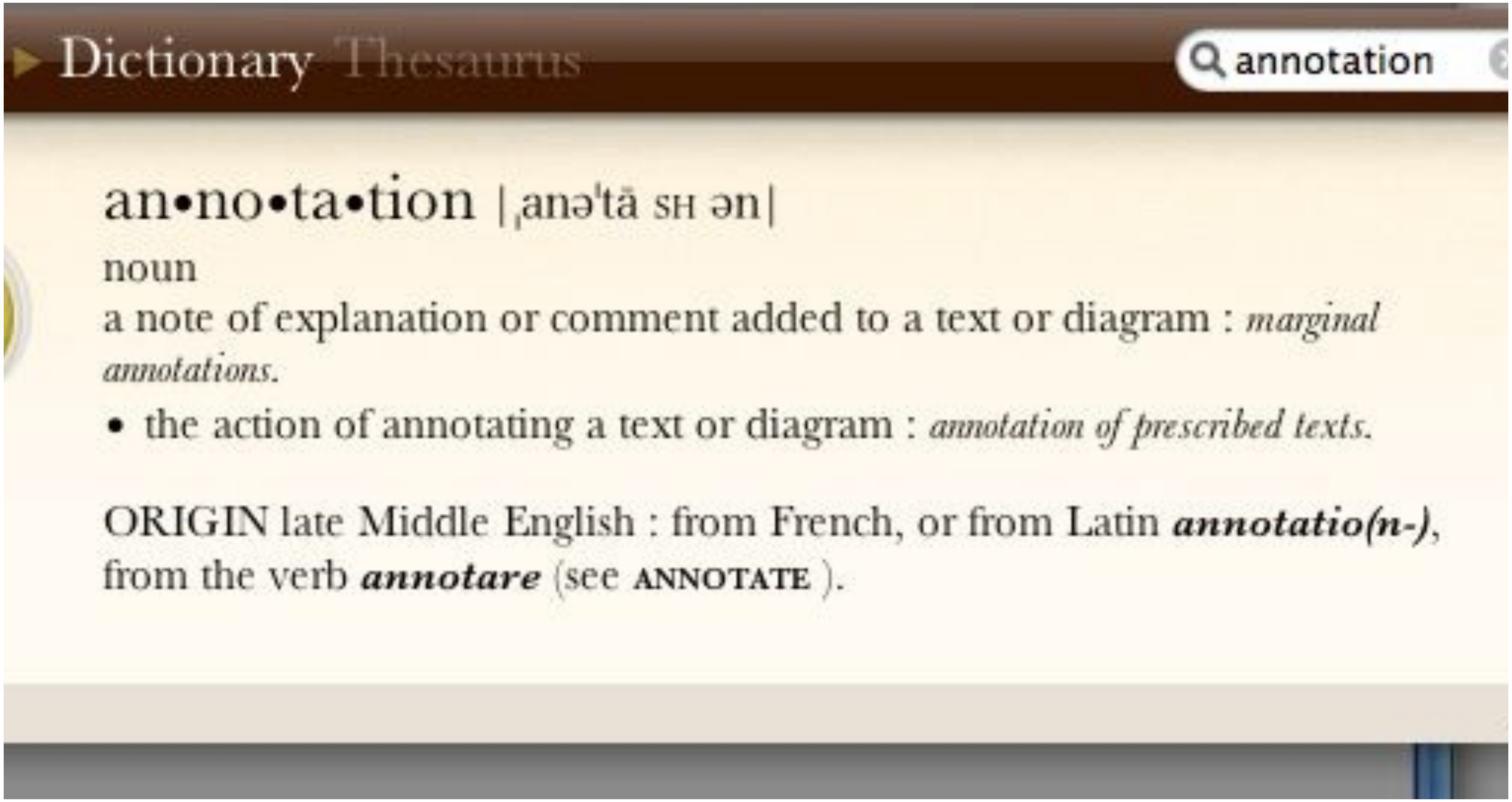
- ID numbers
- Definitions
- Relationships
  - Directed Acyclic Graph
- GO terms provide a way to describe functions, now we have to associate them with genes!
  - AKA GO annotation


biological process
molecular function
metabolic process
cellular process
catalytic activity
organic substance metabolic process
cellular metabolic process
transferase activity
phosphorus metabolic process
macromolecule metabolic process
primary metabolic process
transferase activity, transferring phosphorus-c
phosphate-containing compound metabolic
macromolecule modification
protein metabolic process
cellular macromolecule metabolic process
phosphorylation
protein modification process
cellular protein metabolic process
phosphotransferase activity, alcohol group as acceptor
cellular protein modification process
kinase activity
protein phosphorylation
peptidyl-amino acid modification
peptidyl-tyrosine modification
protein kinase activity
peptidyl-tyrosine phosphorylation
protein tyrosine kinase activity
A
Is a
B
A
Part of
B
A
Has part
B
A
Regulates
B
A
Positively regulates
B
A
Negatively regulates
B
A
Occurs in
B

QuickGO - http://www.ebi.ac.uk/QuickGO

GO Annotation

What is Annotation?



an•no•ta•tion |ˌanəˈtā sh ən|

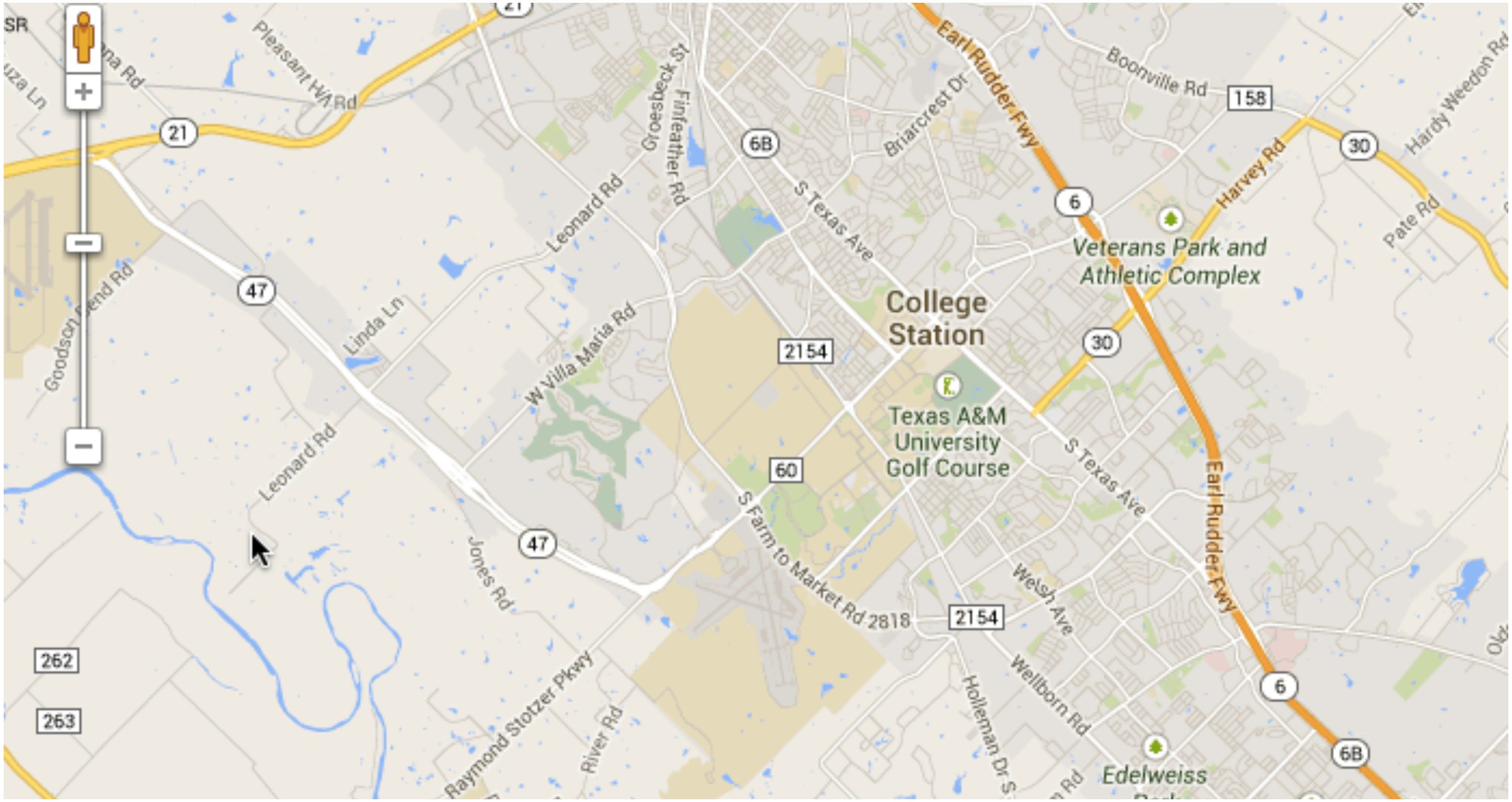
noun

a note of explanation or comment added to a text or diagram : *marginal annotations.*

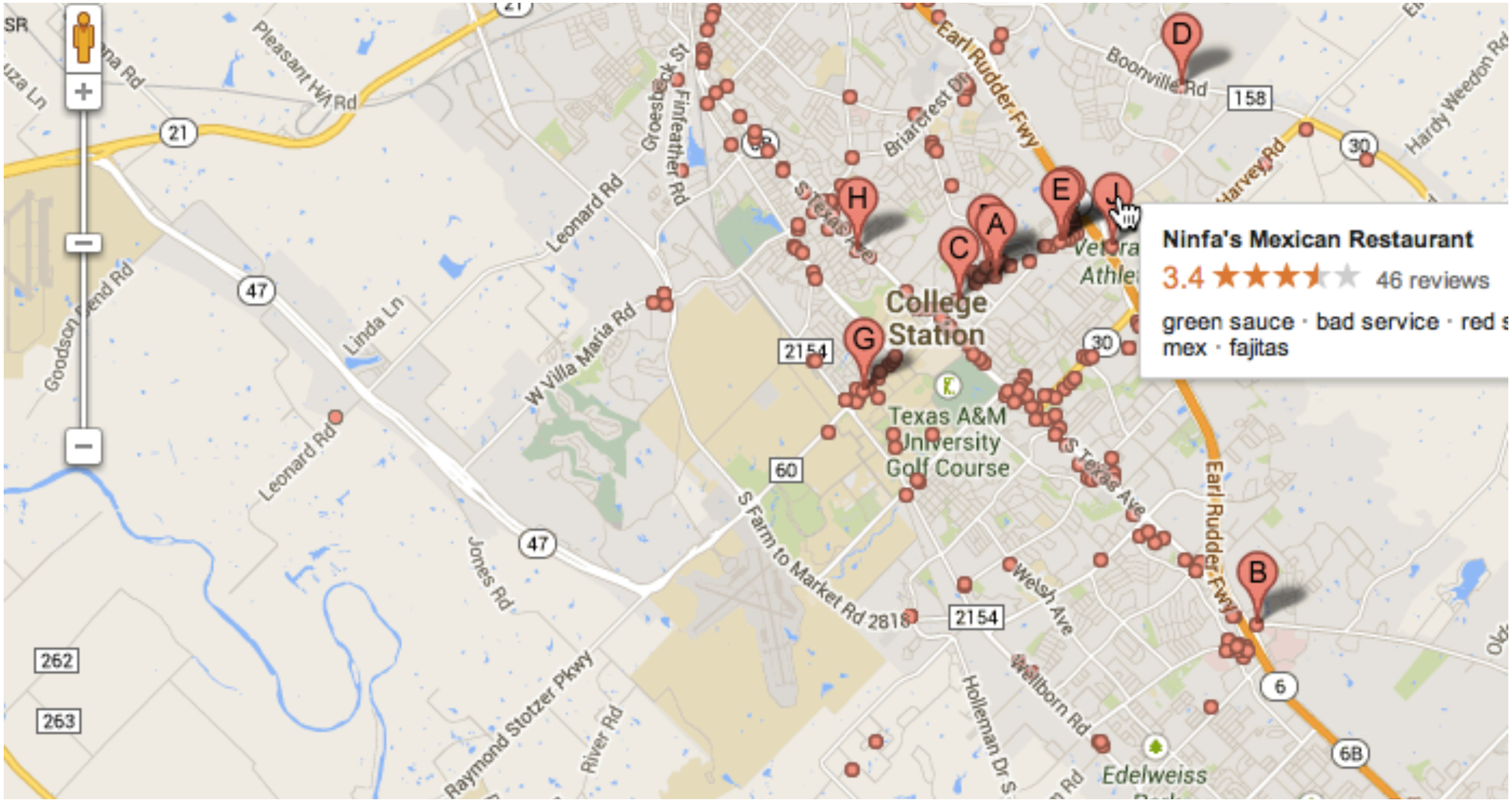
- the action of annotating a text or diagram : *annotation of prescribed texts.*

ORIGIN late Middle English : from French, or from Latin ***annotatio(n-)***, from the verb ***annotare*** (see ANNOTATE).

What is Annotation?



What is Annotation?

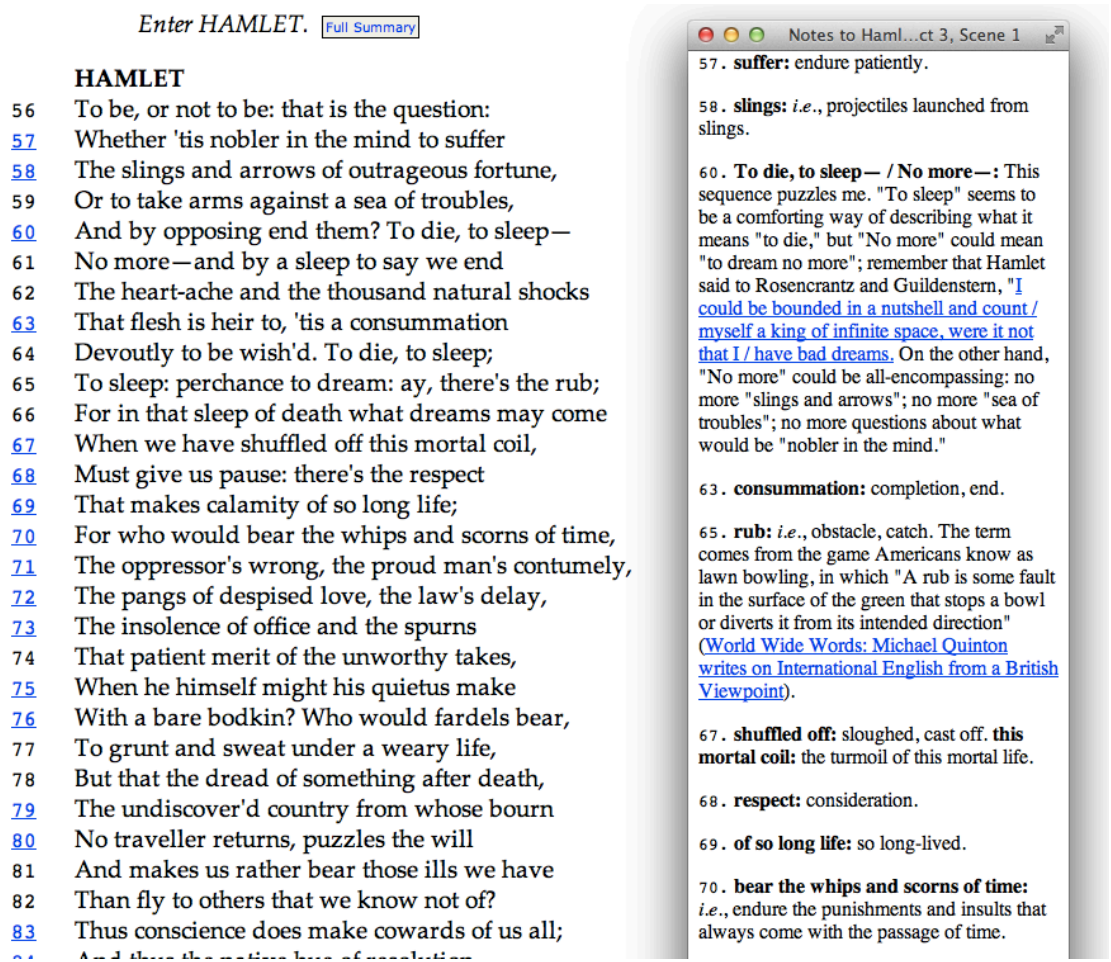


Ninfa's Mexican Restaurant

3.4 ★★★★☆ 46 reviews

green sauce · bad service · red s
mex · fajitas

What is Annotation?



Enter HAMLET. Full Summary

HAMLET

56 To be, or not to be: that is the question:
57 Whether 'tis nobler in the mind to suffer
58 The slings and arrows of outrageous fortune,
59 Or to take arms against a sea of troubles,
60 And by opposing end them? To die, to sleep—
61 No more—and by a sleep to say we end
62 The heart-ache and the thousand natural shocks
63 That flesh is heir to, 'tis a consummation
64 Devoutly to be wish'd. To die, to sleep;
65 To sleep: perchance to dream: ay, there's the rub;
66 For in that sleep of death what dreams may come
67 When we have shuffled off this mortal coil,
68 Must give us pause: there's the respect
69 That makes calamity of so long life;
70 For who would bear the whips and scorns of time,
71 The oppressor's wrong, the proud man's contumely,
72 The pangs of despised love, the law's delay,
73 The insolence of office and the spurns
74 That patient merit of the unworthy takes,
75 When he himself might his quietus make
76 With a bare bodkin? Who would fardels bear,
77 To grunt and sweat under a weary life,
78 But that the dread of something after death,
79 The undiscover'd country from whose bourn
80 No traveller returns, puzzles the will
81 And makes us rather bear those ills we have
82 Than fly to others that we know not of?
83 Thus conscience does make cowards of us all;
84 And thus the native hue of resolution

57. **suffer:** endure patiently.

58. **slings:** *i.e.*, projectiles launched from slings.

60. **To die, to sleep— / No more—:** This sequence puzzles me. "To sleep" seems to be a comforting way of describing what it means "to die," but "No more" could mean "to dream no more"; remember that Hamlet said to Rosencrantz and Guildenstern, "I could be bounded in a nutshell and count / myself a king of infinite space, were it not that I / have bad dreams. On the other hand, "No more" could be all-encompassing: no more "slings and arrows"; no more "sea of troubles"; no more questions about what would be "nobler in the mind."

63. **consummation:** completion, end.

65. **rub:** *i.e.*, obstacle, catch. The term comes from the game Americans know as lawn bowling, in which "A rub is some fault in the surface of the green that stops a bowl or diverts it from its intended direction" (World Wide Words: Michael Quinton writes on International English from a British Viewpoint).

67. **shuffled off:** sloughed, cast off. **this mortal coil:** the turmoil of this mortal life.

68. **respect:** consideration.

69. **of so long life:** so long-lived.

70. **bear the whips and scorns of time:** *i.e.*, endure the punishments and insults that always come with the passage of time.

http://www.shakespeare-navigators.com/hamlet/H31.html

Levels of Annotation for Genomes

- Metadata
  - What is this genome?
- Features
  - Where are things in the sequence?
- Products
  - What do we know about the features?
- Systems
  - What do the products do?
    - individually?
    - Working together?

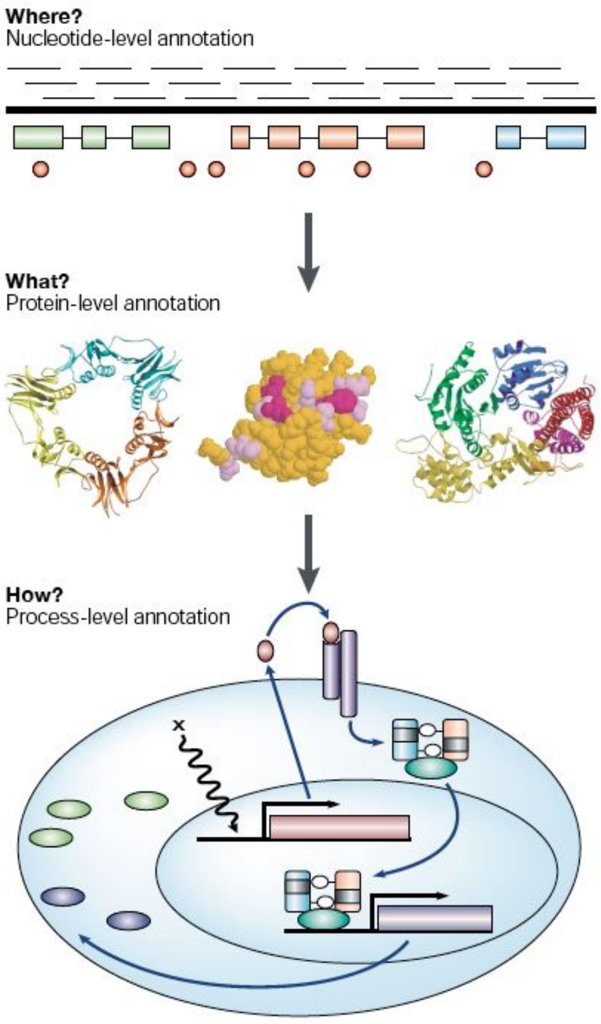


Figure 1 | **The three layers of genome annotation: where, what and how?**

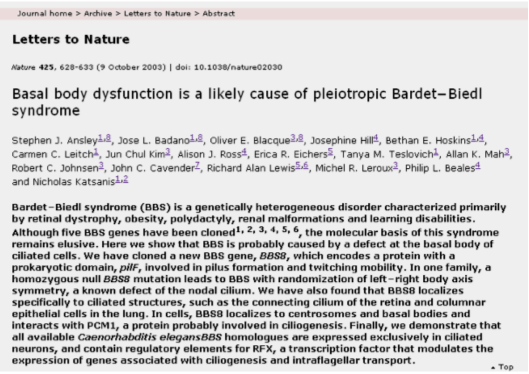
L. Stein (2001) *Nature Reviews Genetics* 2:493

Functional Annotation w/GO

- Annotation: a note that is made while reading any form of text

- GO Annotation: a database entry in a **specific format** that associates a **GO term** with a **gene product** made based on **evidence** in a peer-reviewed **paper**
  - Specific format makes the annotations readable by both computers and humans
  - GO annotations capture the chain of evidence for how functions were inferred from experiments
  - More when we talk about CACAO

Where Do Annotations Come From?



Journal home > Archive > Letters to Nature > Abstract

Letters to Nature

Nature **425**, 628-633 (9 October 2003) | doi: 10.1038/nature02030

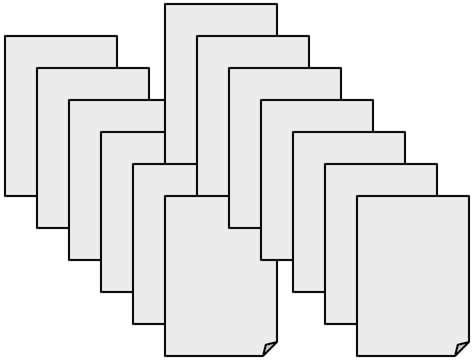
Basal body dysfunction is a likely cause of pleiotropic Bardet–Biedl syndrome

Stephen J. Ansley[1,8], Jose L. Badano[1,8], Oliver E. Blacque[3,8], Josephine Hill[4], Bethan E. Hoskins[1,4], Carmen C. Leitch[1], Jun Chul Kim[3], Alison J. Ross[4], Erica R. Eichers[5], Tanya M. Teslovich[1], Allan K. Mah[3], Robert C. Johnsen[3], John C. Cavender[7], Richard Alan Lewis[5,6], Michel R. Leroux[3], Philip L. Beales[4] and Nicholas Katsanis[1,2]

Bardet–Biedl syndrome (BBS) is a genetically heterogeneous disorder characterized primarily by retinal dystrophy, obesity, polydactyly, renal malformations and learning disabilities. Although five BBS genes have been cloned[1, 2, 3, 4, 5, 6], the molecular basis of this syndrome remains elusive. Here we show that BBS is probably caused by a defect at the basal body of ciliated cells. We have cloned a new BBS gene, *BBS8*, which encodes a protein with a prokaryotic domain, *pilF*, involved in pilus formation and twitching mobility. In one family, a homozygous null *BBS8* mutation leads to BBS with randomization of left–right body axis symmetry, a known defect of the nodal cilium. We have also found that BBS8 localizes specifically to ciliated structures, such as the connecting cilium of the retina and columnar epithelial cells in the lung. In cells, BBS8 localizes to centrosomes and basal bodies and interacts with PCM1, a protein probably involved in ciliogenesis. Finally, we demonstrate that all available *Caenorhabditis elegansBBS* homologues are expressed exclusively in ciliated neurons, and contain regulatory elements for RFX, a transcription factor that modulates the expression of genes associated with ciliogenesis and intraflagellar transport.

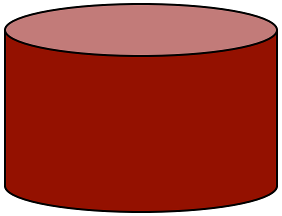
▲ Top

Literature



Datasets

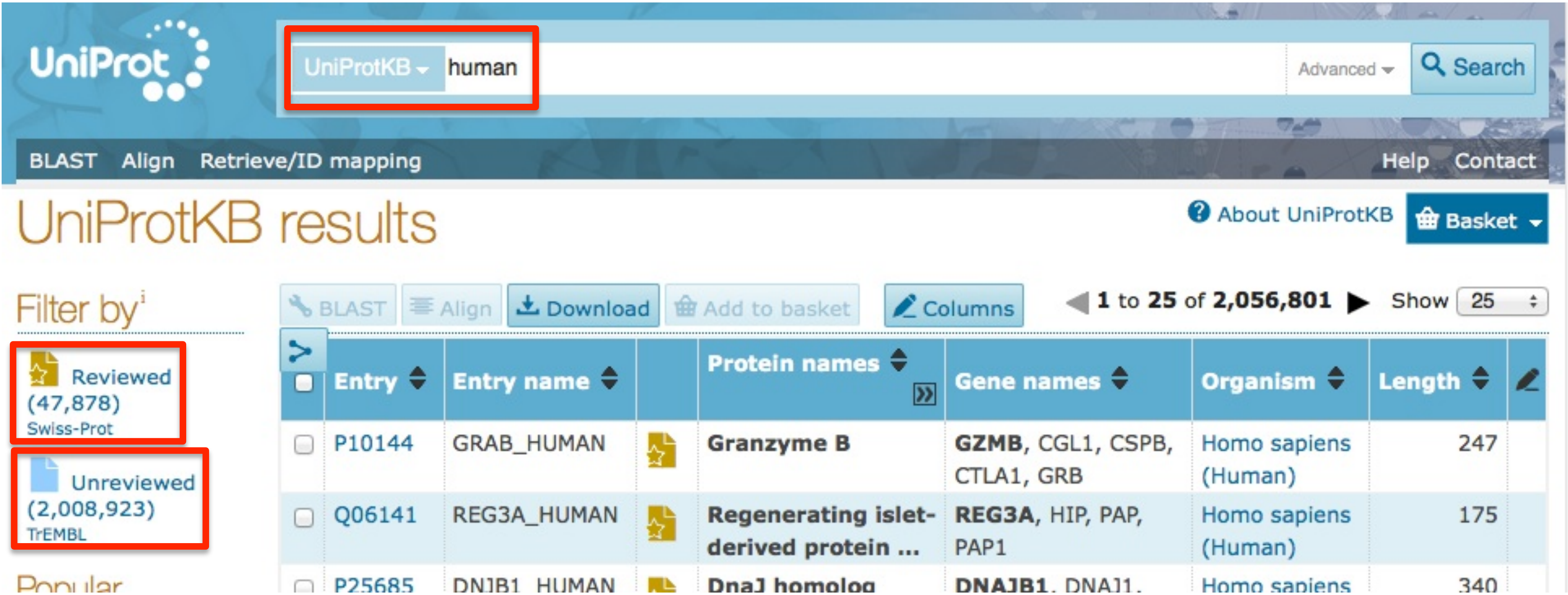
Biocurators
(rate limiting)



Database

Databases Need Help!

- >24 million peer-reviewed articles in PubMed
- Many millions of proteins recorded in UniProt



UniProtKB results

1 to 25 of 2,056,801

Entry	Entry name		Protein names	Gene names	Organism	Length
P10144	GRAB_HUMAN		**Granzyme B**	**GZMB**, CGL1, CSPB, CTLA1, GRB	Homo sapiens (Human)	247
Q06141	REG3A_HUMAN		**Regenerating islet-derived protein ...**	**REG3A**, HIP, PAP, PAP1	Homo sapiens (Human)	175
P25685	DNJB1_HUMAN		**DnaJ homolog**	**DNAJB1**, DNAJ1,	Homo sapiens	340

http://www.uniprot.org

CACAO

What is CACAO?

- **C**ommunity **A**ssessment of **C**ommunity **A**nnotation with **O**ntologies (CACAO)
  - Annotation of gene function
  - Competition
    - Within a class
    - Between teams at different schools
    - More details next week

How Does CACAO Work?

- Working in teams we will use the GONUTS website:
  - http://gowiki.tamu.edu
- Multiple innings: each is two weeks
  - Annotation week: you make annotations on the website to get points
  - Challenge week: you challenge annotations made by other teams to steal their points
- Open week
  - You can challenge AND annotate in the same week
  - RARE but help balance out holiday breaks, etc. between schools
- You can make as many annotations as you want.
  - You pick the topic
  - You have to convince us that they are correct.
    - The default is that they are wrong!!
- Your annotations could end up in databases used by researchers all over the world

How Does CACAO Work?

- Getting help is not cheating!
  - Talk to your teammates
  - Ask us questions
  - Talk to other professors
  - Email authors of papers

What To Annotate

- You can start with a paper
  - Find the proteins discussed, potential GO terms
- You can start with a GO term
  - Modify PubMed search with keywords or organism
- You can start with a protein
  - Find papers about the protein
- Don’t get stuck on what you started with
  - Your first paper may not have **experiments** about function (Reviews)
  - Reading about your initial protein may lead you to better information about other proteins

Functional Annotation w/GO

- Annotation: a note that is made while reading any form of text

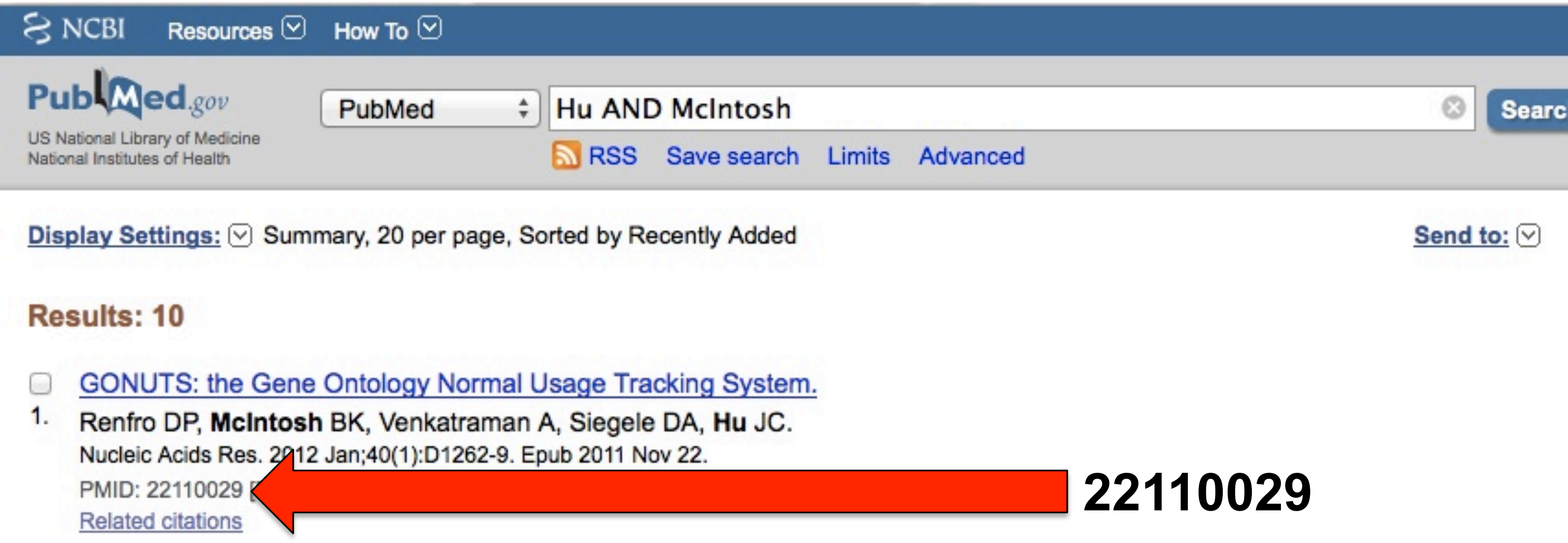
- GO Annotation: a database entry in a **specific format** that associates a **GO term** with a **gene product** made based on **evidence** in a peer-reviewed **paper**

Starting with a paper

- Need a scientific paper with experimental data
  - Use PubMed: http://www.ncbi.nlm.nih.gov/pubmed/
    - Or use an alias like http://pubmed.com
  - No review articles, no books, no textbooks, no Wikipedia articles, no class notes…
  - BUT you could start with those!
  - DON'T start with the first paper you see from a random PubMed search

Starting with a paper

- Need a scientific paper with experimental data
  - PubMed review?
  - We refer to the paper through the PMID number
    - Not the full citation



Search: Hu AND McIntosh

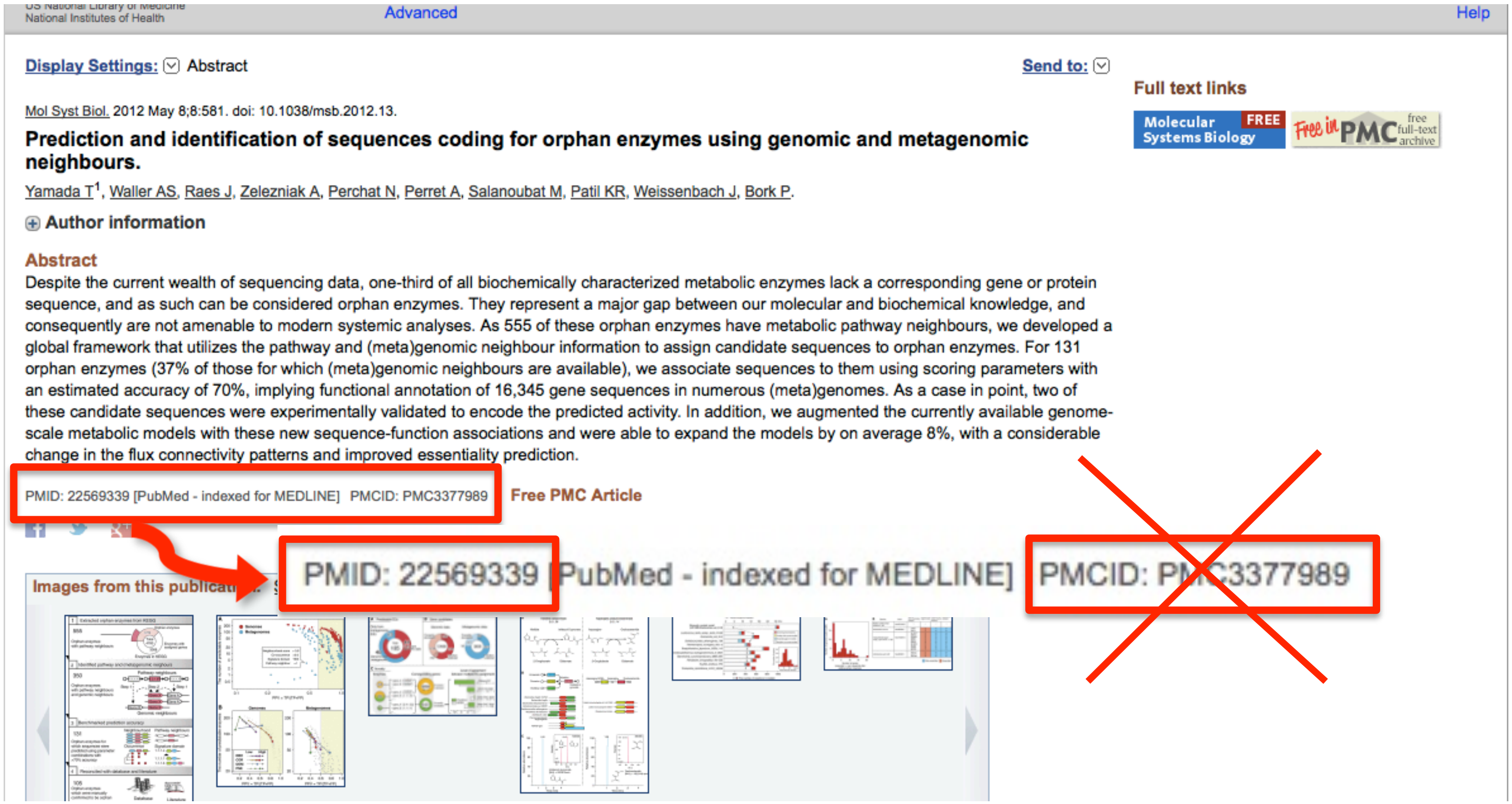
Display Settings: Summary, 20 per page, Sorted by Recently Added

Results: 10

□ 1. GONUTS: the Gene Ontology Normal Usage Tracking System.
Renfro DP, **McIntosh** BK, Venkatraman A, Siegele DA, **Hu** JC.
Nucleic Acids Res. 2012 Jan;40(1):D1262-9. Epub 2011 Nov 22.
PMID: 22110029
Related citations

22110029

PubMed Record



US National Library of Medicine
National Institutes of Health

Display Settings: Abstract

Send to:

Full text links

Molecular Systems Biology FREE
Free in PMC free full-text archive

Mol Syst Biol. 2012 May 8;8:581. doi: 10.1038/msb.2012.13.

Prediction and identification of sequences coding for orphan enzymes using genomic and metagenomic neighbours.

Yamada T[1], Waller AS, Raes J, Zelezniak A, Perchat N, Perret A, Salanoubat M, Patil KR, Weissenbach J, Bork P.

Author information

Abstract

Despite the current wealth of sequencing data, one-third of all biochemically characterized metabolic enzymes lack a corresponding gene or protein sequence, and as such can be considered orphan enzymes. They represent a major gap between our molecular and biochemical knowledge, and consequently are not amenable to modern systemic analyses. As 555 of these orphan enzymes have metabolic pathway neighbours, we developed a global framework that utilizes the pathway and (meta)genomic neighbour information to assign candidate sequences to orphan enzymes. For 131 orphan enzymes (37% of those for which (meta)genomic neighbours are available), we associate sequences to them using scoring parameters with an estimated accuracy of 70%, implying functional annotation of 16,345 gene sequences in numerous (meta)genomes. As a case in point, two of these candidate sequences were experimentally validated to encode the predicted activity. In addition, we augmented the currently available genome-scale metabolic models with these new sequence-function associations and were able to expand the models by on average 8%, with a considerable change in the flux connectivity patterns and improved essentiality prediction.

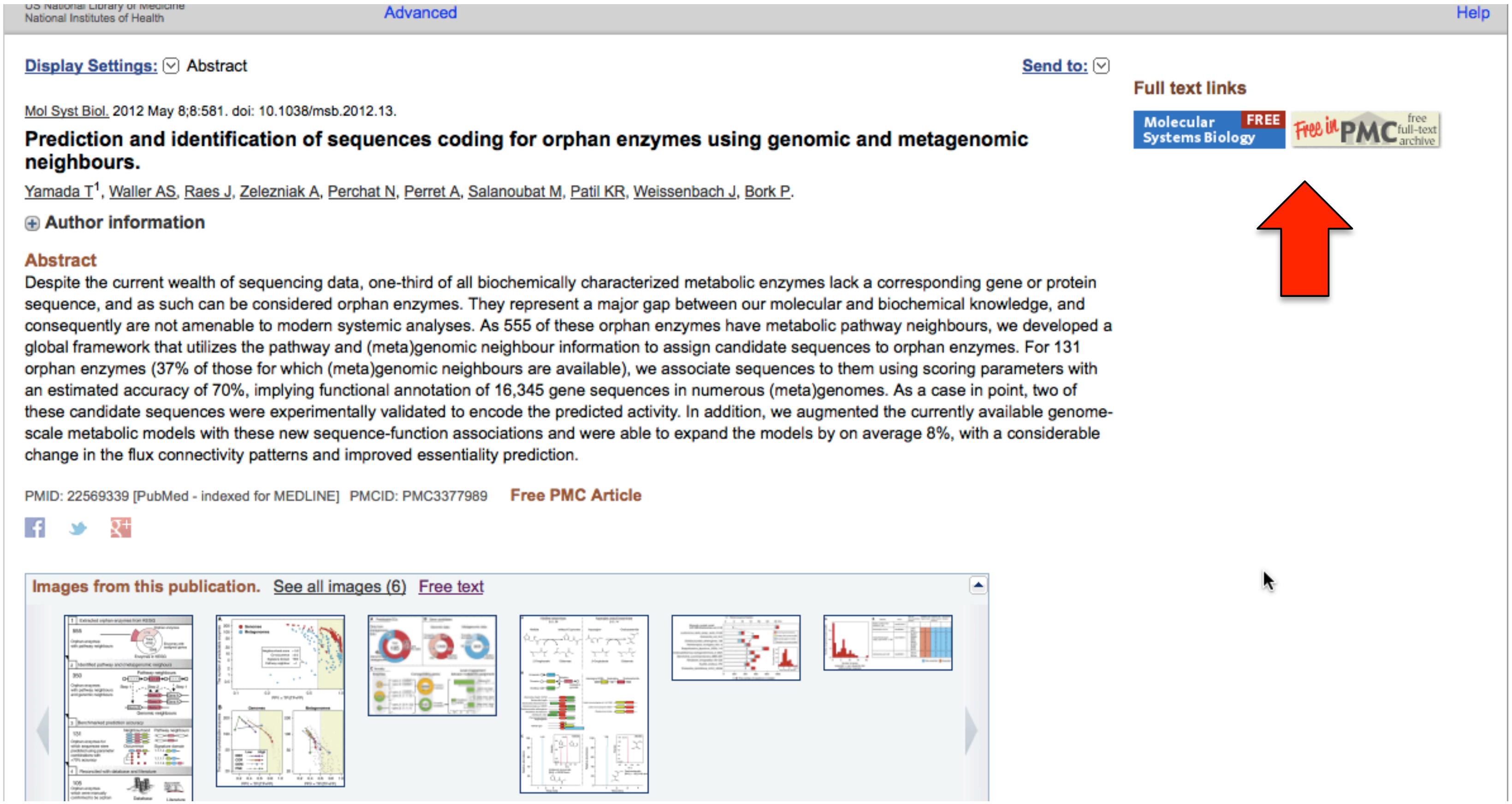
PMID: 22569339 [PubMed - indexed for MEDLINE] PMCID: PMC3377989

Free PMC Article

Images from this publication.

PMID: 22569339 [PubMed - indexed for MEDLINE] ~~PMCID: PMC3377989~~

Getting the Full Text



US National Library of Medicine
National Institutes of Health
Advanced
Help
Display Settings: Abstract
Send to:
Full text links
Molecular Systems Biology FREE
Free in PMC free full-text archive

Mol Syst Biol. 2012 May 8;8:581. doi: 10.1038/msb.2012.13.

Prediction and identification of sequences coding for orphan enzymes using genomic and metagenomic neighbours.

Yamada T[1], Waller AS, Raes J, Zelezniak A, Perchat N, Perret A, Salanoubat M, Patil KR, Weissenbach J, Bork P.

Author information

Abstract

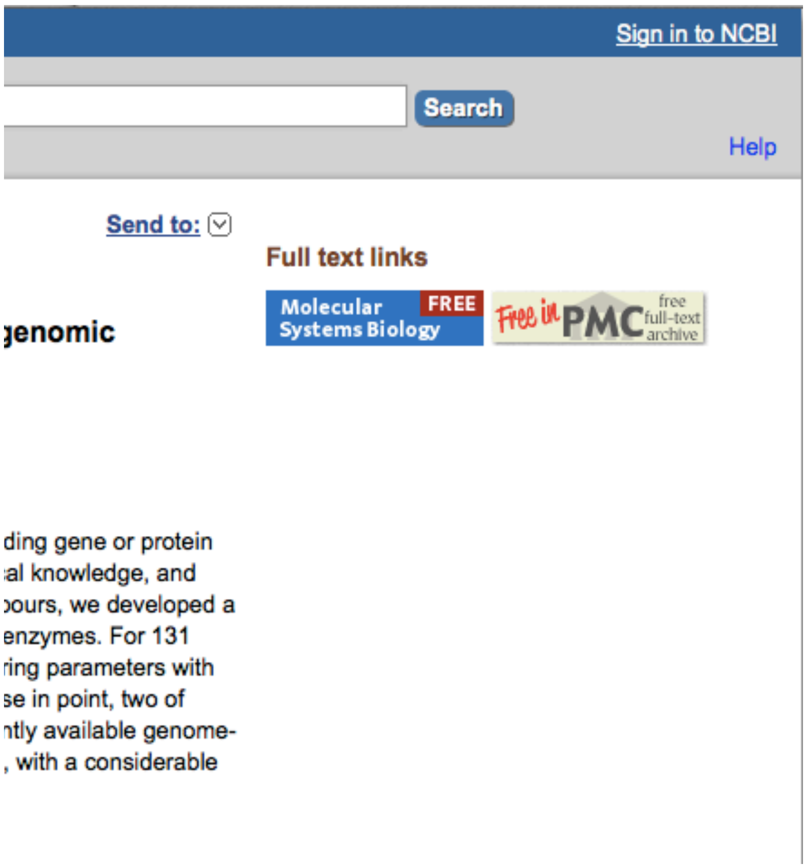
Despite the current wealth of sequencing data, one-third of all biochemically characterized metabolic enzymes lack a corresponding gene or protein sequence, and as such can be considered orphan enzymes. They represent a major gap between our molecular and biochemical knowledge, and consequently are not amenable to modern systemic analyses. As 555 of these orphan enzymes have metabolic pathway neighbours, we developed a global framework that utilizes the pathway and (meta)genomic neighbour information to assign candidate sequences to orphan enzymes. For 131 orphan enzymes (37% of those for which (meta)genomic neighbours are available), we associate sequences to them using scoring parameters with an estimated accuracy of 70%, implying functional annotation of 16,345 gene sequences in numerous (meta)genomes. As a case in point, two of these candidate sequences were experimentally validated to encode the predicted activity. In addition, we augmented the currently available genome-scale metabolic models with these new sequence-function associations and were able to expand the models by on average 8%, with a considerable change in the flux connectivity patterns and improved essentiality prediction.

PMID: 22569339 [PubMed - indexed for MEDLINE] PMCID: PMC3377989 **Free PMC Article**

- The abstract is not enough for an annotation
  - But, may be enough to reject a paper!!!

Getting the Full Text

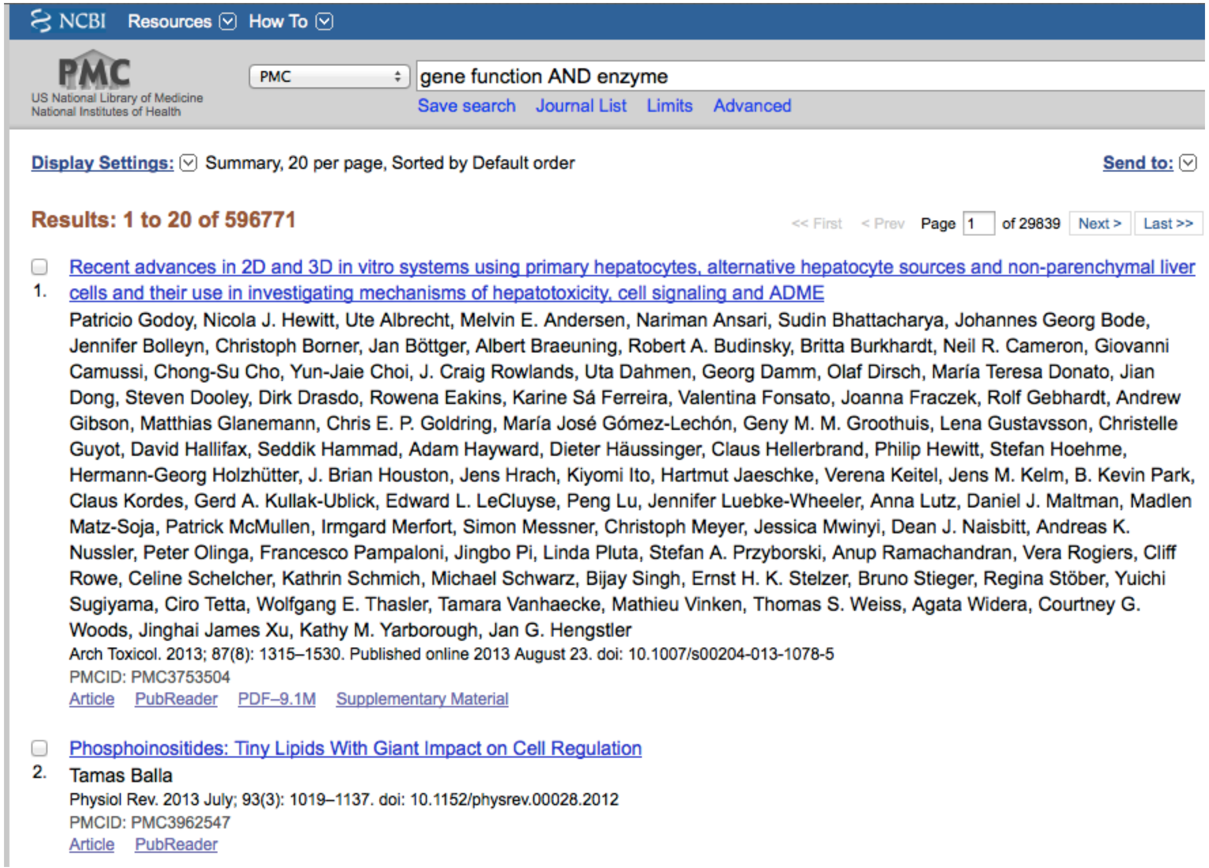
- Some papers are open access
  - Pubmed Central
  - Journal sites
- Others are pay only
  - Don’t pay real $$!
    - Your library may have subscriptions
    - Pick a different paper
    - Email the author and ask for a pdf
      - Send us a copy


Sign in to NCBI
Search
Help
Send to:
Full text links
Molecular Systems Biology FREE
Free in PMC free full-text archive

ɟenomic

ding gene or protein
ːal knowledge, and
ɔours, we developed a
enzymes. For 131
ring parameters with
se in point, two of
ntly available genome-
, with a considerable

Alternative Path: Start w/Full Text



gene function AND enzyme

Display Settings: Summary, 20 per page, Sorted by Default order

Results: 1 to 20 of 596771

Page 1 of 29839

□ 1. Recent advances in 2D and 3D in vitro systems using primary hepatocytes, alternative hepatocyte sources and non-parenchymal liver cells and their use in investigating mechanisms of hepatotoxicity, cell signaling and ADME

Patricio Godoy, Nicola J. Hewitt, Ute Albrecht, Melvin E. Andersen, Nariman Ansari, Sudin Bhattacharya, Johannes Georg Bode, Jennifer Bolleyn, Christoph Borner, Jan Böttger, Albert Braeuning, Robert A. Budinsky, Britta Burkhardt, Neil R. Cameron, Giovanni Camussi, Chong-Su Cho, Yun-Jaie Choi, J. Craig Rowlands, Uta Dahmen, Georg Damm, Olaf Dirsch, María Teresa Donato, Jian Dong, Steven Dooley, Dirk Drasdo, Rowena Eakins, Karine Sá Ferreira, Valentina Fonsato, Joanna Fraczek, Rolf Gebhardt, Andrew Gibson, Matthias Glanemann, Chris E. P. Goldring, María José Gómez-Lechón, Geny M. M. Groothuis, Lena Gustavsson, Christelle Guyot, David Hallifax, Seddik Hammad, Adam Hayward, Dieter Häussinger, Claus Hellerbrand, Philip Hewitt, Stefan Hoehme, Hermann-Georg Holzhütter, J. Brian Houston, Jens Hrach, Kiyomi Ito, Hartmut Jaeschke, Verena Keitel, Jens M. Kelm, B. Kevin Park, Claus Kordes, Gerd A. Kullak-Ublick, Edward L. LeCluyse, Peng Lu, Jennifer Luebke-Wheeler, Anna Lutz, Daniel J. Maltman, Madlen Matz-Soja, Patrick McMullen, Irmgard Merfort, Simon Messner, Christoph Meyer, Jessica Mwinyi, Dean J. Naisbitt, Andreas K. Nussler, Peter Olinga, Francesco Pampaloni, Jingbo Pi, Linda Pluta, Stefan A. Przyborski, Anup Ramachandran, Vera Rogiers, Cliff Rowe, Celine Schelcher, Kathrin Schmich, Michael Schwarz, Bijay Singh, Ernst H. K. Stelzer, Bruno Stieger, Regina Stöber, Yuichi Sugiyama, Ciro Tetta, Wolfgang E. Thasler, Tamara Vanhaecke, Mathieu Vinken, Thomas S. Weiss, Agata Widera, Courtney G. Woods, Jinghai James Xu, Kathy M. Yarborough, Jan G. Hengstler

Arch Toxicol. 2013; 87(8): 1315–1530. Published online 2013 August 23. doi: 10.1007/s00204-013-1078-5

PMCID: PMC3753504

Article PubReader PDF–9.1M Supplementary Material

□ 2. Phosphoinositides: Tiny Lipids With Giant Impact on Cell Regulation

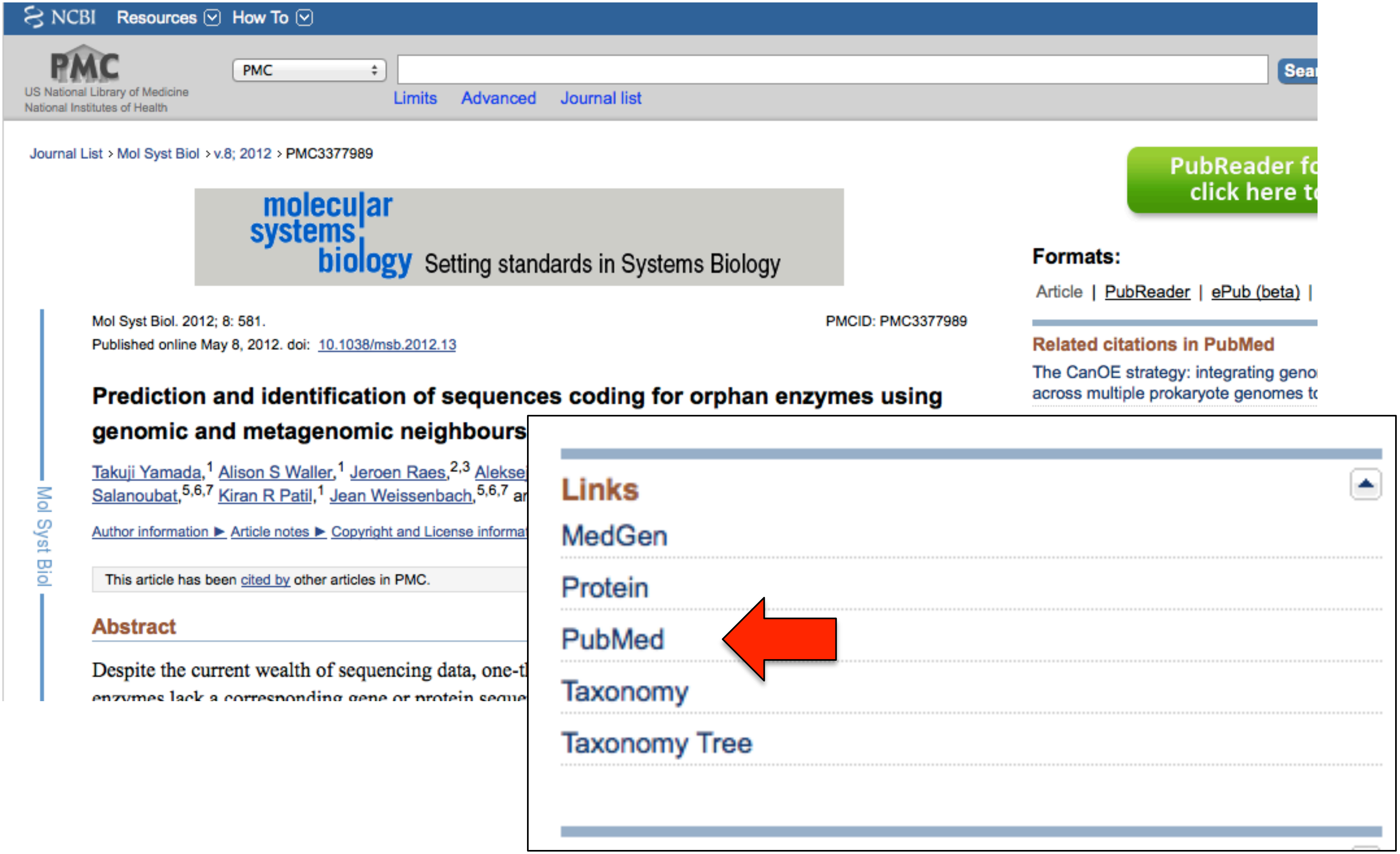
Tamas Balla

Physiol Rev. 2013 July; 93(3): 1019–1137. doi: 10.1152/physrev.00028.2012

PMCID: PMC3962547

Article PubReader

Alternative Path: Start w/Full Text



NCBI
Resources
How To
PMC
US National Library of Medicine
National Institutes of Health
PMC
Sear
Limits
Advanced
Journal list
PubReader fo
click here to
molecular systems biology
Setting standards in Systems Biology
Mol Syst Biol

Journal List > Mol Syst Biol > v.8; 2012 > PMC3377989

Mol Syst Biol. 2012; 8: 581.
Published online May 8, 2012. doi: 10.1038/msb.2012.13

PMCID: PMC3377989

Prediction and identification of sequences coding for orphan enzymes using genomic and metagenomic neighbours

Takuji Yamada,[1] Alison S Waller,[1] Jeroen Raes,[2,3] Alekse
Salanoubat,[5,6,7] Kiran R Patil,[1] Jean Weissenbach,[5,6,7] ar

Author information ▶ Article notes ▶ Copyright and License informa

This article has been cited by other articles in PMC.

Abstract

Despite the current wealth of sequencing data, one-t
enzymes lack a corresponding gene or protein seque

Formats:

Article | PubReader | ePub (beta) |

Related citations in PubMed

The CanOE strategy: integrating geno
across multiple prokaryote genomes to

Links

MedGen

Protein

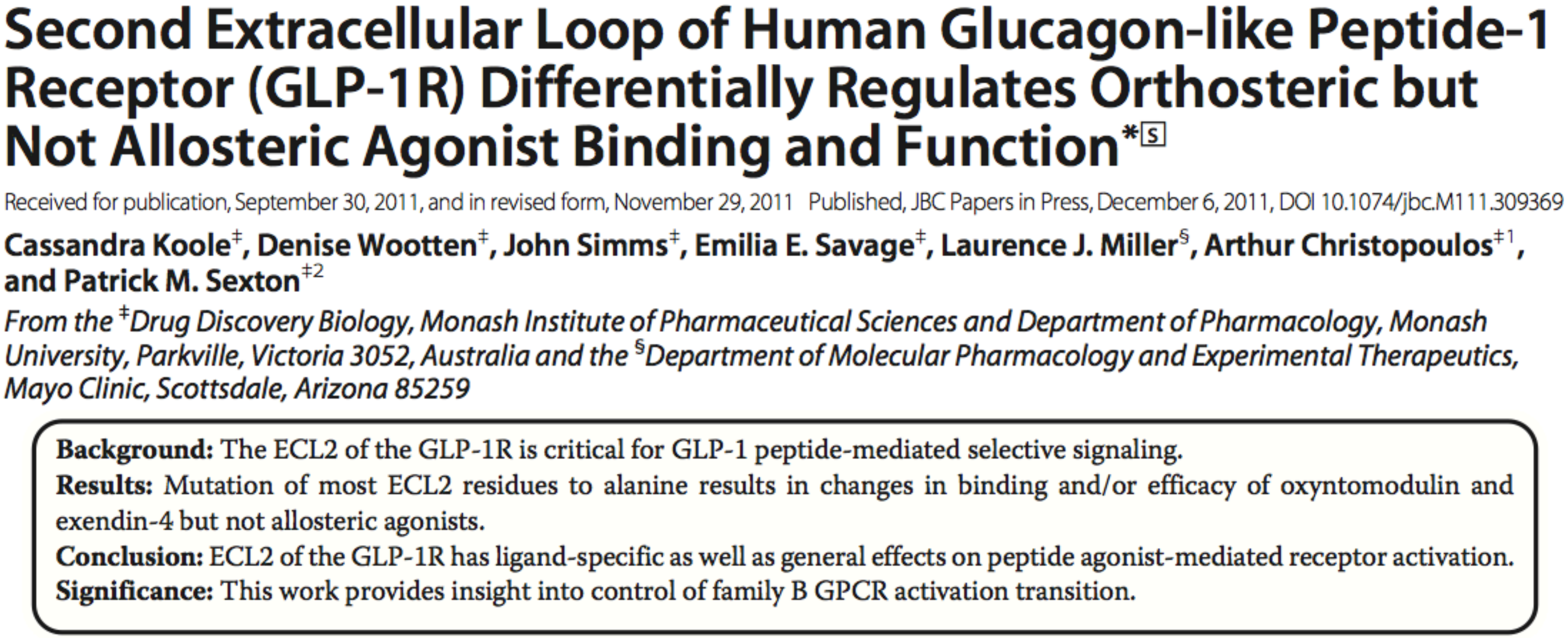
PubMed

Taxonomy

Taxonomy Tree

Beware!

- Good science ≠ good for annotation



Second Extracellular Loop of Human Glucagon-like Peptide-1 Receptor (GLP-1R) Differentially Regulates Orthosteric but Not Allosteric Agonist Binding and Function*[S]

Received for publication, September 30, 2011, and in revised form, November 29, 2011 Published, JBC Papers in Press, December 6, 2011, DOI 10.1074/jbc.M111.309369

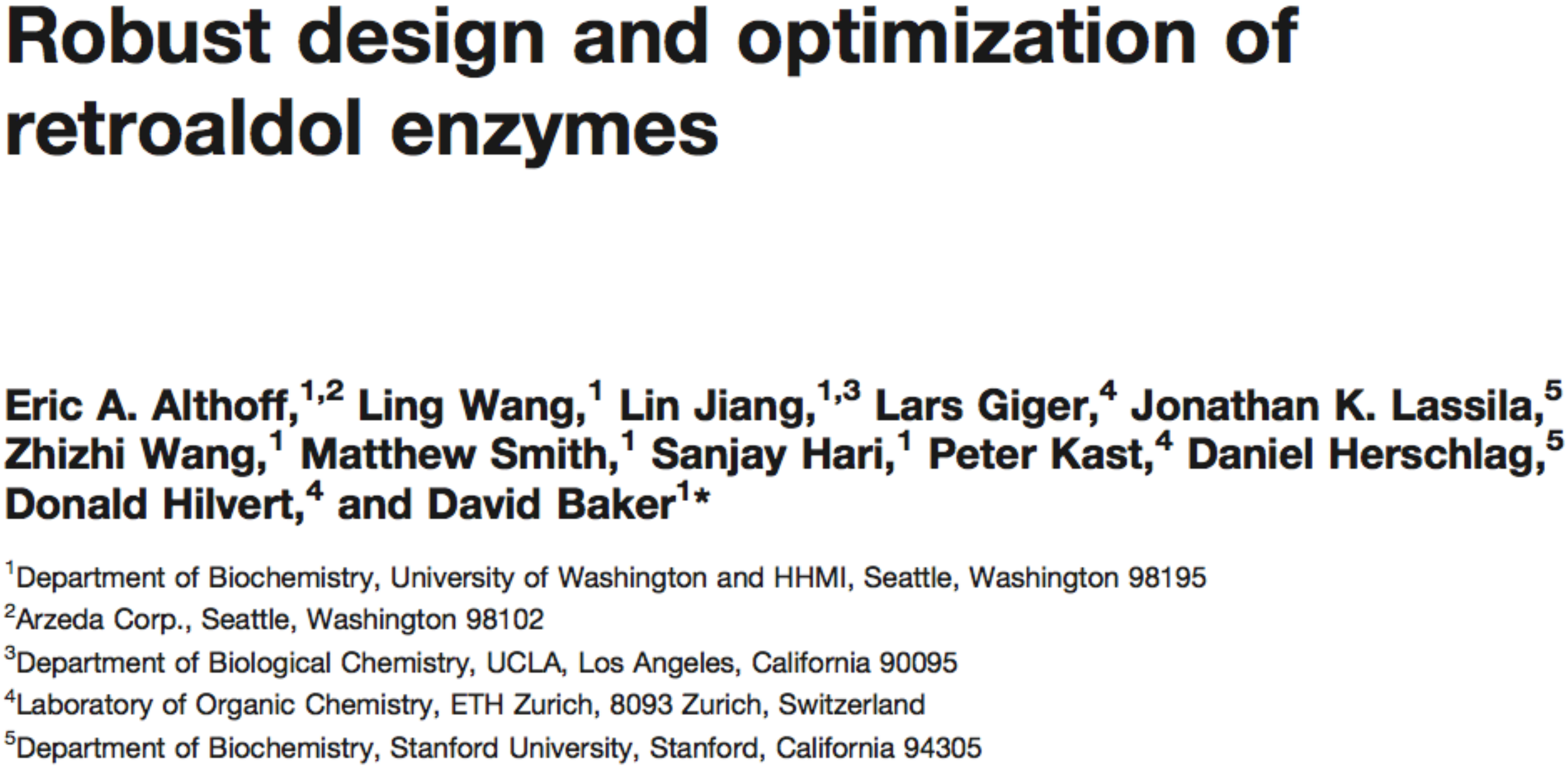
Cassandra Koole[‡], Denise Wootten[‡], John Simms[‡], Emilia E. Savage[‡], Laurence J. Miller[§], Arthur Christopoulos[‡1], and Patrick M. Sexton[‡2]

From the [‡]Drug Discovery Biology, Monash Institute of Pharmaceutical Sciences and Department of Pharmacology, Monash University, Parkville, Victoria 3052, Australia and the [§]Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Scottsdale, Arizona 85259

Background: The ECL2 of the GLP-1R is critical for GLP-1 peptide-mediated selective signaling.
Results: Mutation of most ECL2 residues to alanine results in changes in binding and/or efficacy of oxyntomodulin and exendin-4 but not allosteric agonists.
Conclusion: ECL2 of the GLP-1R has ligand-specific as well as general effects on peptide agonist-mediated receptor activation.
Significance: This work provides insight into control of family B GPCR activation transition.

Beware!

- Good science ≠ good for annotation



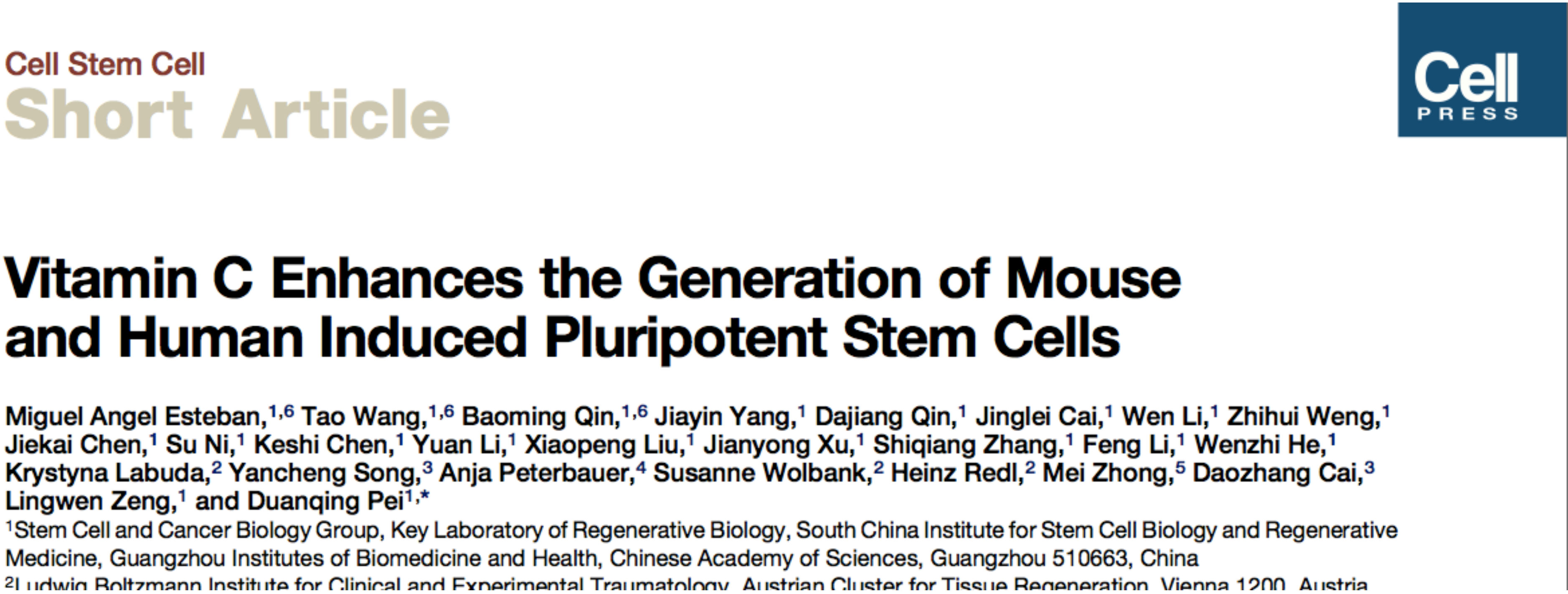
Robust design and optimization of retroaldol enzymes

Eric A. Althoff,[1,2] Ling Wang,[1] Lin Jiang,[1,3] Lars Giger,[4] Jonathan K. Lassila,[5] Zhizhi Wang,[1] Matthew Smith,[1] Sanjay Hari,[1] Peter Kast,[4] Daniel Herschlag,[5] Donald Hilvert,[4] and David Baker[1]*

[1]Department of Biochemistry, University of Washington and HHMI, Seattle, Washington 98195
[2]Arzeda Corp., Seattle, Washington 98102
[3]Department of Biological Chemistry, UCLA, Los Angeles, California 90095
[4]Laboratory of Organic Chemistry, ETH Zurich, 8093 Zurich, Switzerland
[5]Department of Biochemistry, Stanford University, Stanford, California 94305

Beware!

- Good science ≠ good for annotation



Cell Stem Cell

Short Article

Cell PRESS

Vitamin C Enhances the Generation of Mouse and Human Induced Pluripotent Stem Cells

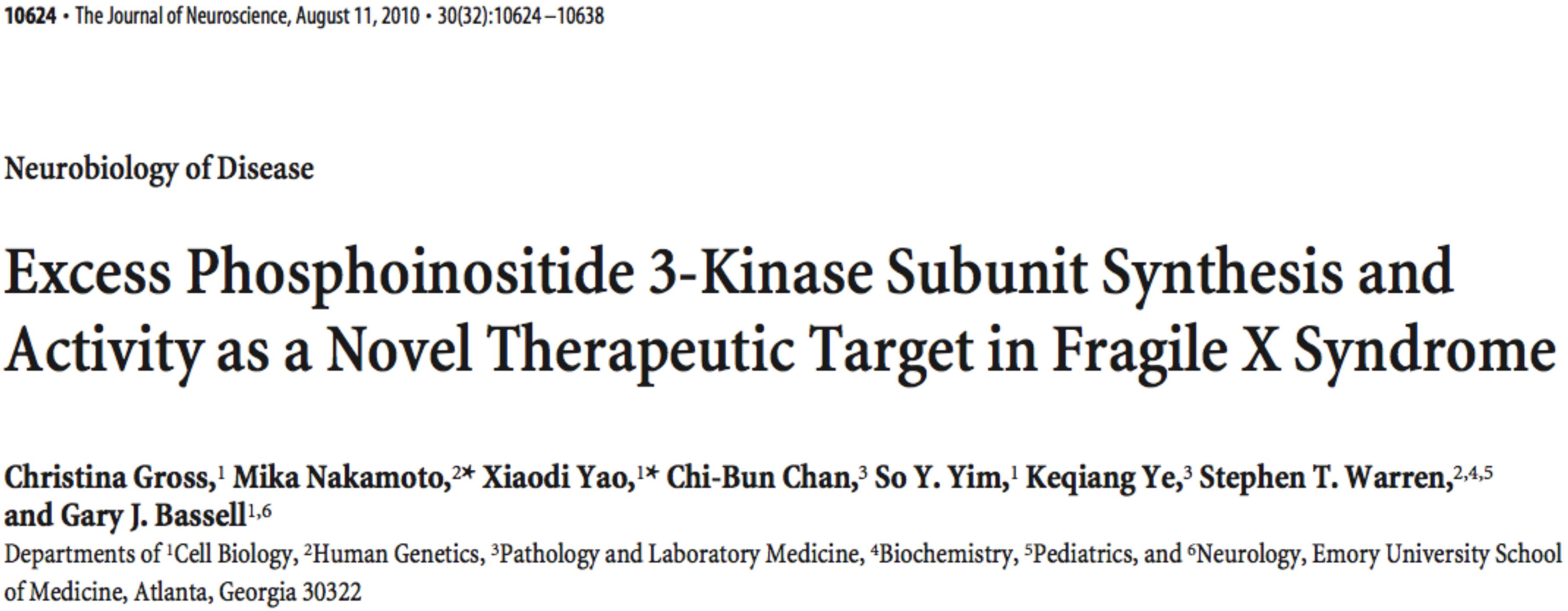
Miguel Angel Esteban,[1,6] Tao Wang,[1,6] Baoming Qin,[1,6] Jiayin Yang,[1] Dajiang Qin,[1] Jinglei Cai,[1] Wen Li,[1] Zhihui Weng,[1] Jiekai Chen,[1] Su Ni,[1] Keshi Chen,[1] Yuan Li,[1] Xiaopeng Liu,[1] Jianyong Xu,[1] Shiqiang Zhang,[1] Feng Li,[1] Wenzhi He,[1] Krystyna Labuda,[2] Yancheng Song,[3] Anja Peterbauer,[4] Susanne Wolbank,[2] Heinz Redl,[2] Mei Zhong,[5] Daozhang Cai,[3] Lingwen Zeng,[1] and Duanqing Pei[1,*]

[1]Stem Cell and Cancer Biology Group, Key Laboratory of Regenerative Biology, South China Institute for Stem Cell Biology and Regenerative Medicine, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510663, China

[2]Ludwig Boltzmann Institute for Clinical and Experimental Traumatology, Austrian Cluster for Tissue Regeneration, Vienna 1200, Austria

Beware!

- Good science ≠ good for annotation



10624 • The Journal of Neuroscience, August 11, 2010 • 30(32):10624–10638

Neurobiology of Disease

Excess Phosphoinositide 3-Kinase Subunit Synthesis and Activity as a Novel Therapeutic Target in Fragile X Syndrome

Christina Gross,[1] Mika Nakamoto,[2]* Xiaodi Yao,[1]* Chi-Bun Chan,[3] So Y. Yim,[1] Keqiang Ye,[3] Stephen T. Warren,[2,4,5] and Gary J. Bassell[1,6]
Departments of [1]Cell Biology, [2]Human Genetics, [3]Pathology and Laboratory Medicine, [4]Biochemistry, [5]Pediatrics, and [6]Neurology, Emory University School of Medicine, Atlanta, Georgia 30322

Functional Annotation w/GO

- Annotation: a note that is made while reading any form of text

- GO Annotation: a database entry in a **specific format** that associates a **GO term** with a **gene product** made based on **evidence** in a peer-reviewed **paper**

Finding Proteins

- Search UniProt for something interesting
- Look in UniProt for the protein(s) in the paper you are reading.

No matter what, you will need to find the protein's accession on UniProt (http://uniprot.org)

↓

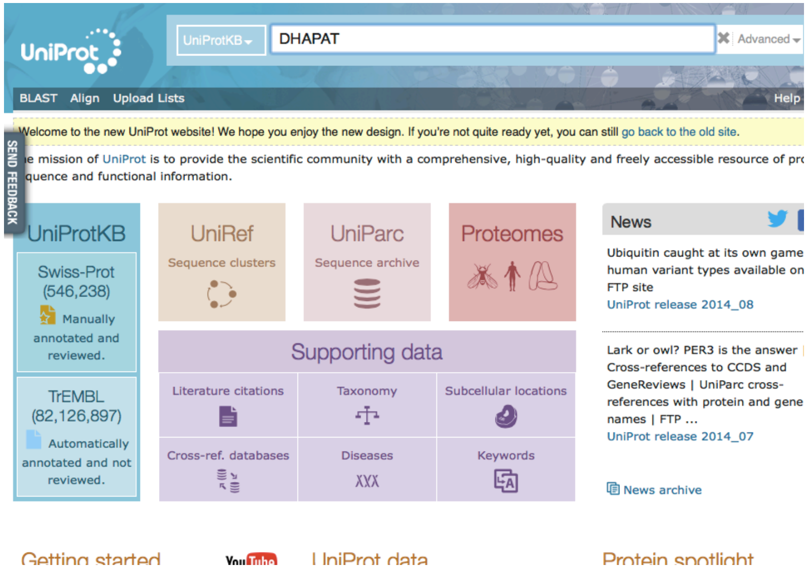
Use that accession to make a page for that protein on GONUTS (http://gowiki.tamu.edu)

↓

Add your GO annotations to the protein's page on GONUTS

UniProt (http://www.uniprot.org)

- If you have a paper, look for an accession
  - UniProt accession
  - NCBI Gene ID
- If you don't have an accession, search by name/keyword



UniProtKB ▾ DHAPAT

Welcome to the new UniProt website! We hope you enjoy the new design. If you're not quite ready yet, you can still go back to the old site.

e mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of pr
quence and functional information.

UniProtKB

Swiss-Prot (546,238) Manually annotated and reviewed.

TrEMBL (82,126,897) Automatically annotated and not reviewed.

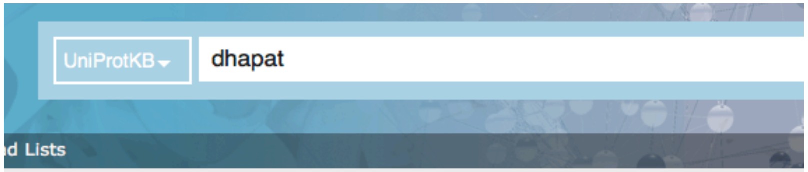
News

Ubiquitin caught at its own game
human variant types available on
FTP site
UniProt release 2014_08

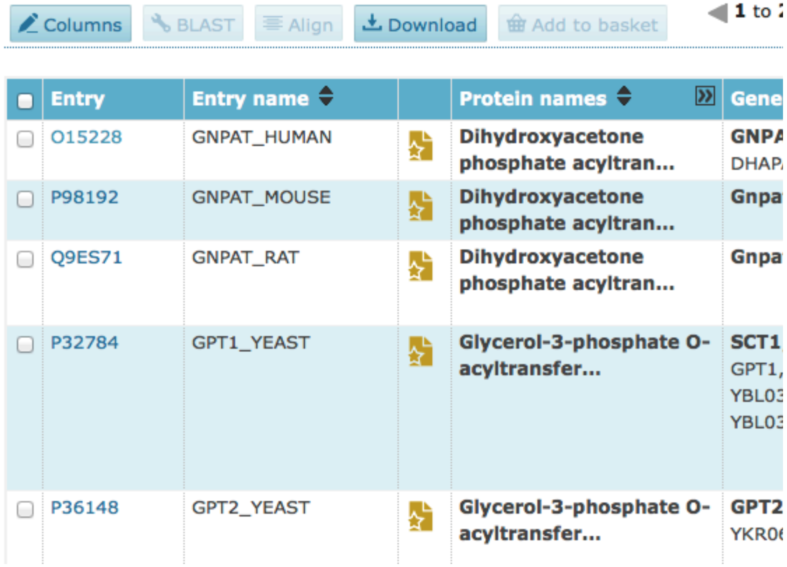
Lark or owl? PER3 is the answer |
Cross-references to CCDS and
GeneReviews | UniParc cross-
references with protein and gene
names | FTP ...
UniProt release 2014_07

UniProt Search Results

- Multiple entries
  - Find the right one
  - Icons
    - Gold = Swissprot = reviewed
    - Plain = TrEMBL = automated



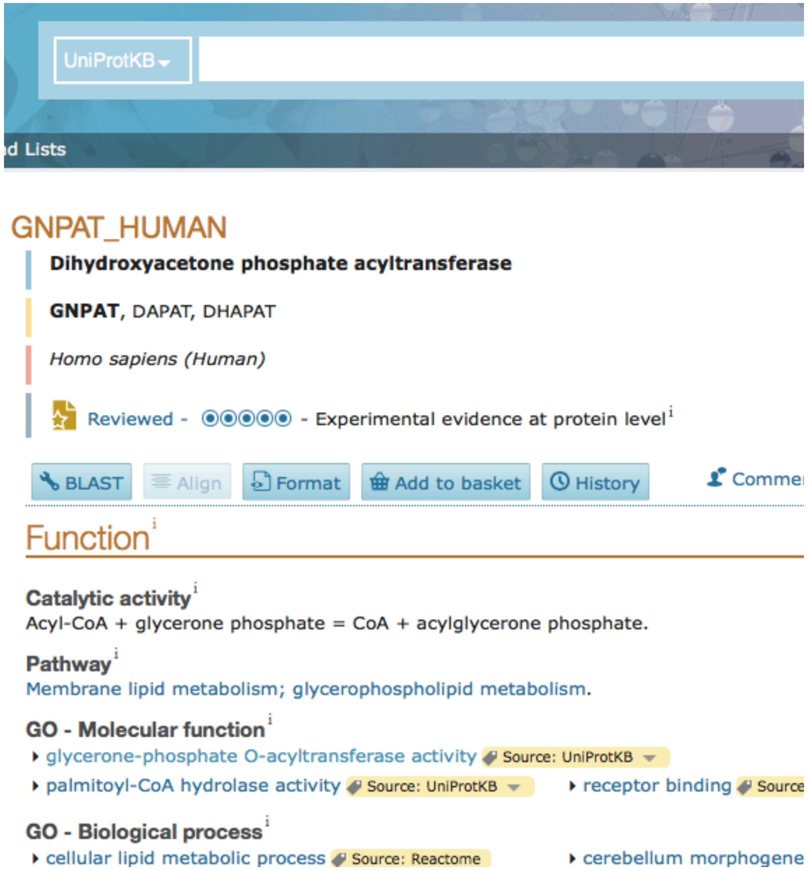
UniProtKB dhapat



□	Entry	Entry name		Protein names	Gene
□	O15228	GNPAT_HUMAN	★	Dihydroxyacetone phosphate acyltran...	GNPA DHAP
□	P98192	GNPAT_MOUSE	★	Dihydroxyacetone phosphate acyltran...	Gnpa
□	Q9ES71	GNPAT_RAT	★	Dihydroxyacetone phosphate acyltran...	Gnpa
□	P32784	GPT1_YEAST	★	Glycerol-3-phosphate O-acyltransfer...	SCT1 GPT1, YBL03 YBL03
□	P36148	GPT2_YEAST	★	Glycerol-3-phosphate O-acyltransfer...	GPT2 YKR0

UniProt Records

- Lots of information to help you
  - Summary of existing GO annotations
    - Link to QuickGO for complete set of existing annotations
  - Information about the protein



GNPAT_HUMAN

Dihydroxyacetone phosphate acyltransferase

GNPAT, DAPAT, DHAPAT

Homo sapiens (Human)

Reviewed - ◉◉◉◉◉ - Experimental evidence at protein level[i]

Function[i]

Catalytic activity[i]
Acyl-CoA + glycerone phosphate = CoA + acylglycerone phosphate.

Pathway[i]
Membrane lipid metabolism; glycerophospholipid metabolism.

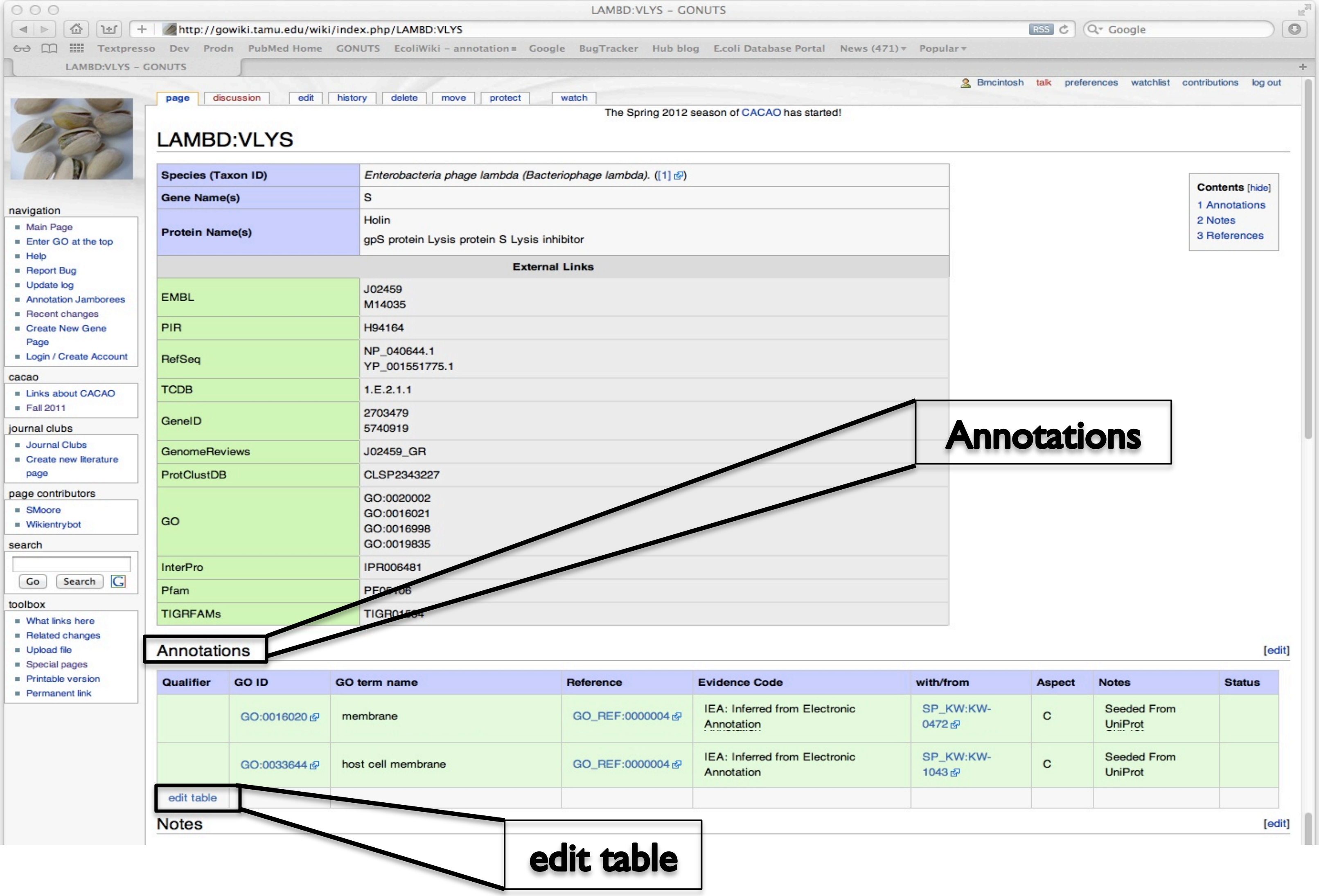
GO - Molecular function[i]
- glycerone-phosphate O-acyltransferase activity Source: UniProtKB
- palmitoyl-CoA hydrolase activity Source: UniProtKB
- receptor binding Source

GO - Biological process[i]
- cellular lipid metabolic process Source: Reactome
- cerebellum morphogene

Make Sure You Have the Right Protein

- Right species/strain
- Not a fragment
- Sometimes UniProt has multiple entries for the same protein
  - Gold star = SwissProt = reviewed
  - Blank star = TrEMBL = computational entry
- Sometimes the protein you want is not in UniProt
  - May want to find another paper/protein
- Ask for help
  - OK to email the UniProt help desk
  - check your reasoning with us!

Create a Protein Page in GONUTS



The Spring 2012 season of CACAO has started!

LAMBD:VLYS

Species (Taxon ID)	*Enterobacteria phage lambda (Bacteriophage lambda).* ([1])
Gene Name(s)	S
Protein Name(s)	Holin gpS protein Lysis protein S Lysis inhibitor
External Links	
EMBL	J02459 M14035
PIR	H94164
RefSeq	NP_040644.1 YP_001551775.1
TCDB	1.E.2.1.1
GeneID	2703479 5740919
GenomeReviews	J02459_GR
ProtClustDB	CLSP2343227
GO	GO:0020002 GO:0016021 GO:0016998 GO:0019835
InterPro	IPR006481
Pfam	PF05106
TIGRFAMs	TIGR01594

Annotations

[edit]

Qualifier	GO ID	GO term name	Reference	Evidence Code	with/from	Aspect	Notes	Status
	GO:0016020	membrane	GO_REF:0000004	IEA: Inferred from Electronic Annotation	SP_KW:KW-0472	C	Seeded From UniProt	
	GO:0033644	host cell membrane	GO_REF:0000004	IEA: Inferred from Electronic Annotation	SP_KW:KW-1043	C	Seeded From UniProt	
edit table								

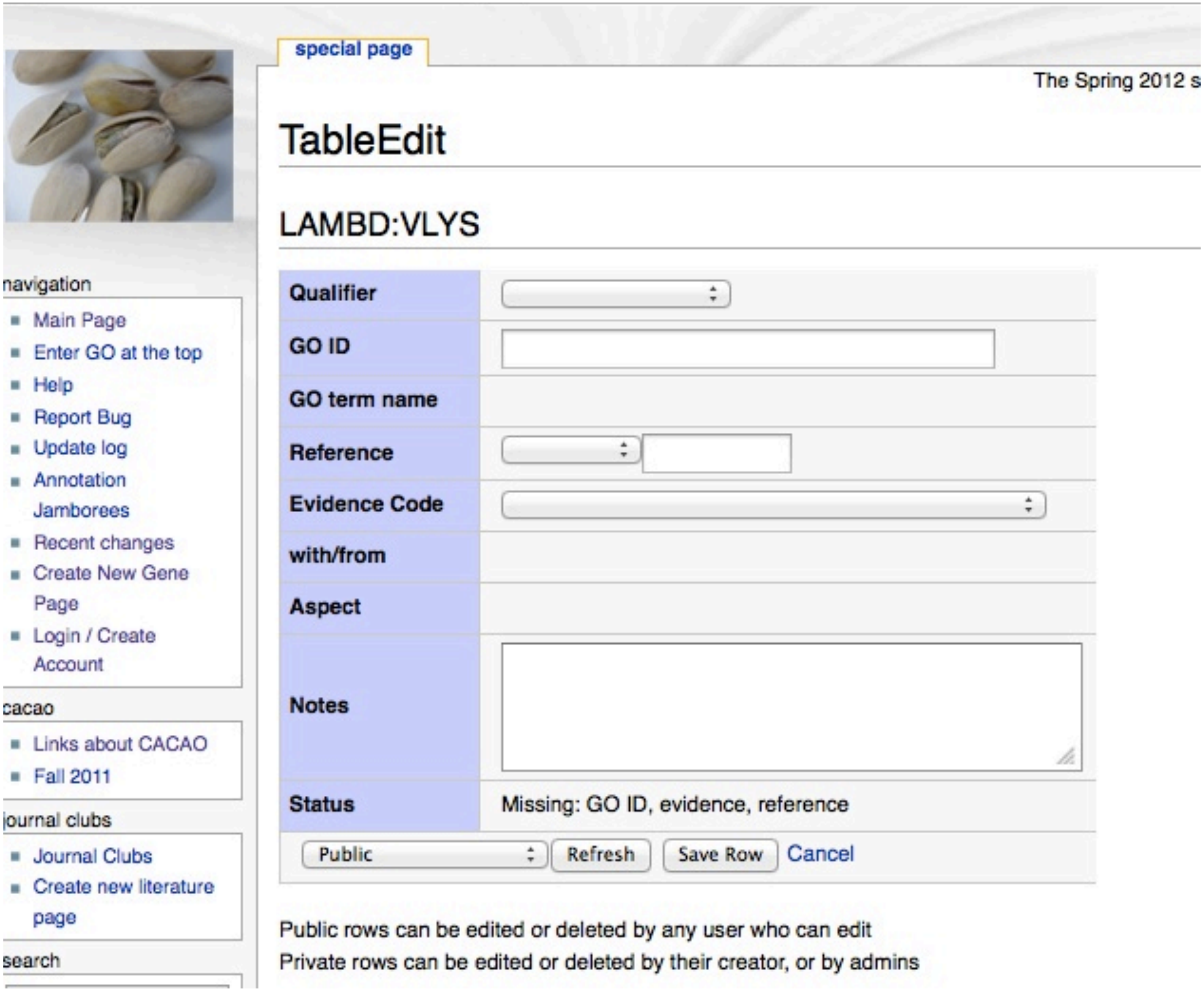
Notes

[edit]

Annotations

edit table

Entering/Editing Annotations



special page

The Spring 2012 s

TableEdit

LAMBD:VLYS

Qualifier	
GO ID	
GO term name	
Reference	
Evidence Code	
with/from	
Aspect	
Notes	
Status	Missing: GO ID, evidence, reference

Public rows can be edited or deleted by any user who can edit

Private rows can be edited or deleted by their creator, or by admins

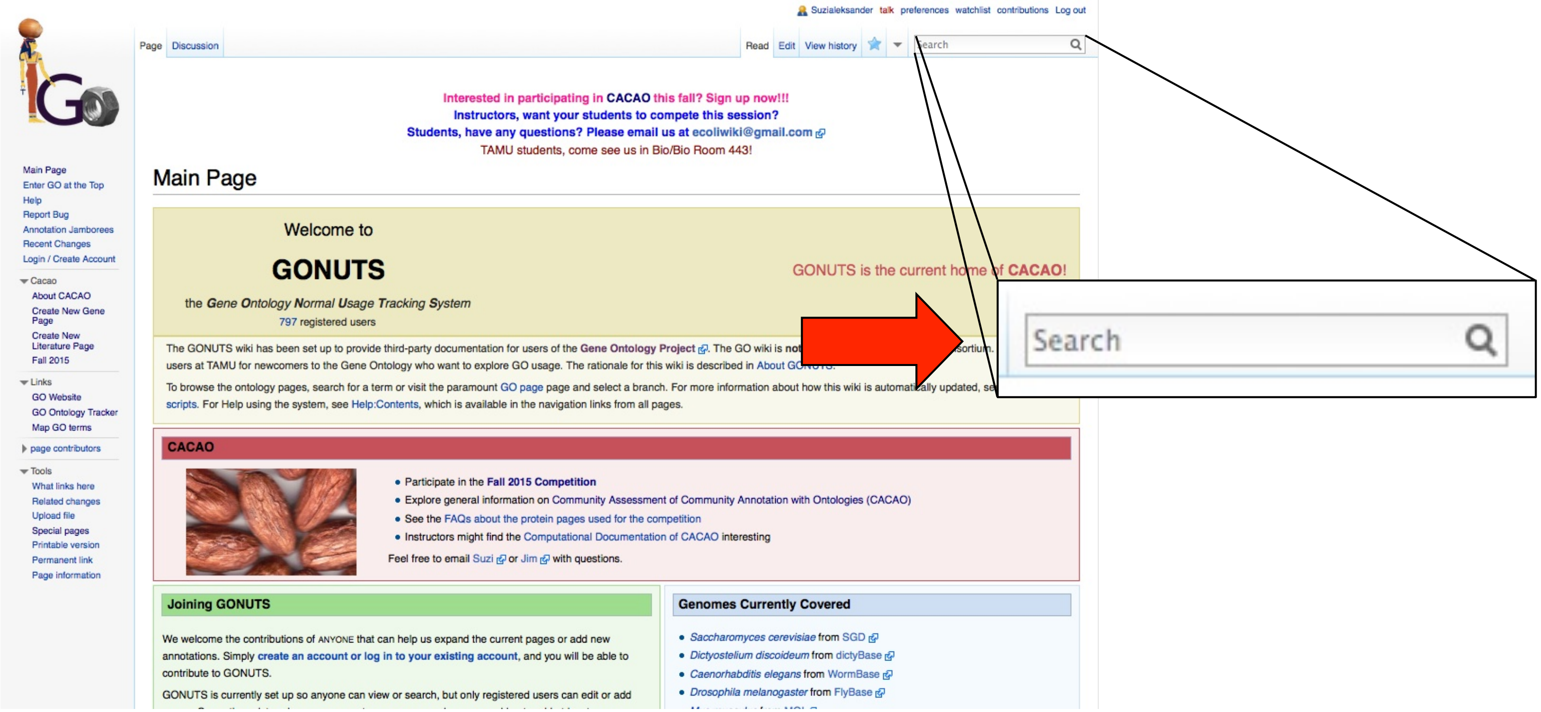
Functional Annotation w/GO

- Annotation: a note that is made while reading any form of text

- GO Annotation: a database entry in a **specific format** that associates a **GO term** with a **gene product** made based on **evidence** in a peer-reviewed **paper**

Finding GO Terms

- GONUTS: http://gowiki.tamu.edu
- QuickGO: http://www.ebi.ac.uk/QuickGO
- AmiGO2: http://amigo.geneontology.org/amigo



Suzialeksander talk preferences watchlist contributions Log out

Main Page
Enter GO at the Top
Help
Report Bug
Annotation Jamborees
Recent Changes
Login / Create Account

Cacao
About CACAO
Create New Gene Page
Create New Literature Page
Fall 2015

Links
GO Website
GO Ontology Tracker
Map GO terms

page contributors

Tools
What links here
Related changes
Upload file
Special pages
Printable version
Permanent link
Page information

Interested in participating in CACAO this fall? Sign up now!!!
Instructors, want your students to compete this session?
Students, have any questions? Please email us at ecoliwiki@gmail.com
TAMU students, come see us in Bio/Bio Room 443!

Main Page

Welcome to

GONUTS

the *Gene Ontology Normal Usage Tracking System*

797 registered users

GONUTS is the current home of **CACAO**!

The GONUTS wiki has been set up to provide third-party documentation for users of the **Gene Ontology Project**. The GO wiki is **not** sortium. users at TAMU for newcomers to the Gene Ontology who want to explore GO usage. The rationale for this wiki is described in About GONUTS.

To browse the ontology pages, search for a term or visit the paramount GO page page and select a branch. For more information about how this wiki is automatically updated, se scripts. For Help using the system, see Help:Contents, which is available in the navigation links from all pages.

CACAO

- Participate in the **Fall 2015 Competition**
- Explore general information on Community Assessment of Community Annotation with Ontologies (CACAO)
- See the FAQs about the protein pages used for the competition
- Instructors might find the Computational Documentation of CACAO interesting

Feel free to email Suzi or Jim with questions.

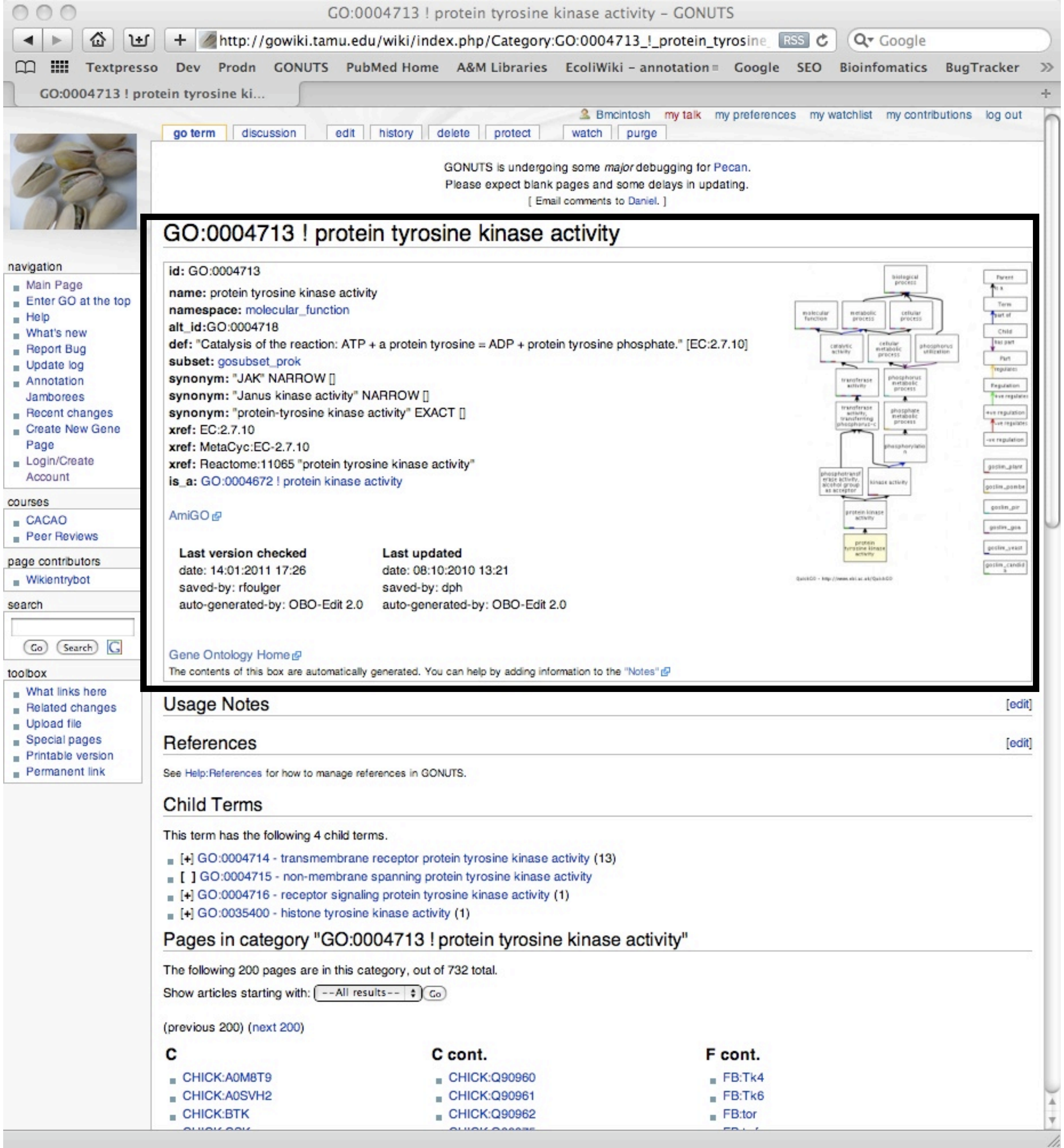
Joining GONUTS

We welcome the contributions of ANYONE that can help us expand the current pages or add new annotations. Simply **create an account or log in to your existing account**, and you will be able to contribute to GONUTS.

GONUTS is currently set up so anyone can view or search, but only registered users can edit or add

Genomes Currently Covered

- *Saccharomyces cerevisiae* from SGD
- *Dictyostelium discoideum* from dictyBase
- *Caenorhabditis elegans* from WormBase
- *Drosophila melanogaster* from FlyBase



GONUTS is undergoing some *major* debugging for Pecan.
Please expect blank pages and some delays in updating.
[Email comments to Daniel.]

GO:0004713 ! protein tyrosine kinase activity

id: GO:0004713

name: protein tyrosine kinase activity
namespace: molecular_function
alt_id:GO:0004718
def: "Catalysis of the reaction: ATP + a protein tyrosine = ADP + protein tyrosine phosphate." [EC:2.7.10]
subset: gosubset_prok
synonym: "JAK" NARROW []
synonym: "Janus kinase activity" NARROW []
synonym: "protein-tyrosine kinase activity" EXACT []
xref: EC:2.7.10
xref: MetaCyc:EC-2.7.10
xref: Reactome:11065 "protein tyrosine kinase activity"
is_a: GO:0004672 ! protein kinase activity

AmiGO

Last version checked	**Last updated**
date: 14:01:2011 17:26	date: 08:10:2010 13:21
saved-by: rfoulger	saved-by: dph
auto-generated-by: OBO-Edit 2.0	auto-generated-by: OBO-Edit 2.0

Gene Ontology Home

The contents of this box are automatically generated. You can help by adding information to the "Notes"

Usage Notes

[edit]

References

[edit]

See Help:References for how to manage references in GONUTS.

Child Terms

This term has the following 4 child terms.

- [+] GO:0004714 - transmembrane receptor protein tyrosine kinase activity (13)
- [ ] GO:0004715 - non-membrane spanning protein tyrosine kinase activity
- [+] GO:0004716 - receptor signaling protein tyrosine kinase activity (1)
- [+] GO:0035400 - histone tyrosine kinase activity (1)

Pages in category "GO:0004713 ! protein tyrosine kinase activity"

The following 200 pages are in this category, out of 732 total.

Show articles starting with: --All results-- Go

(previous 200) (next 200)

C

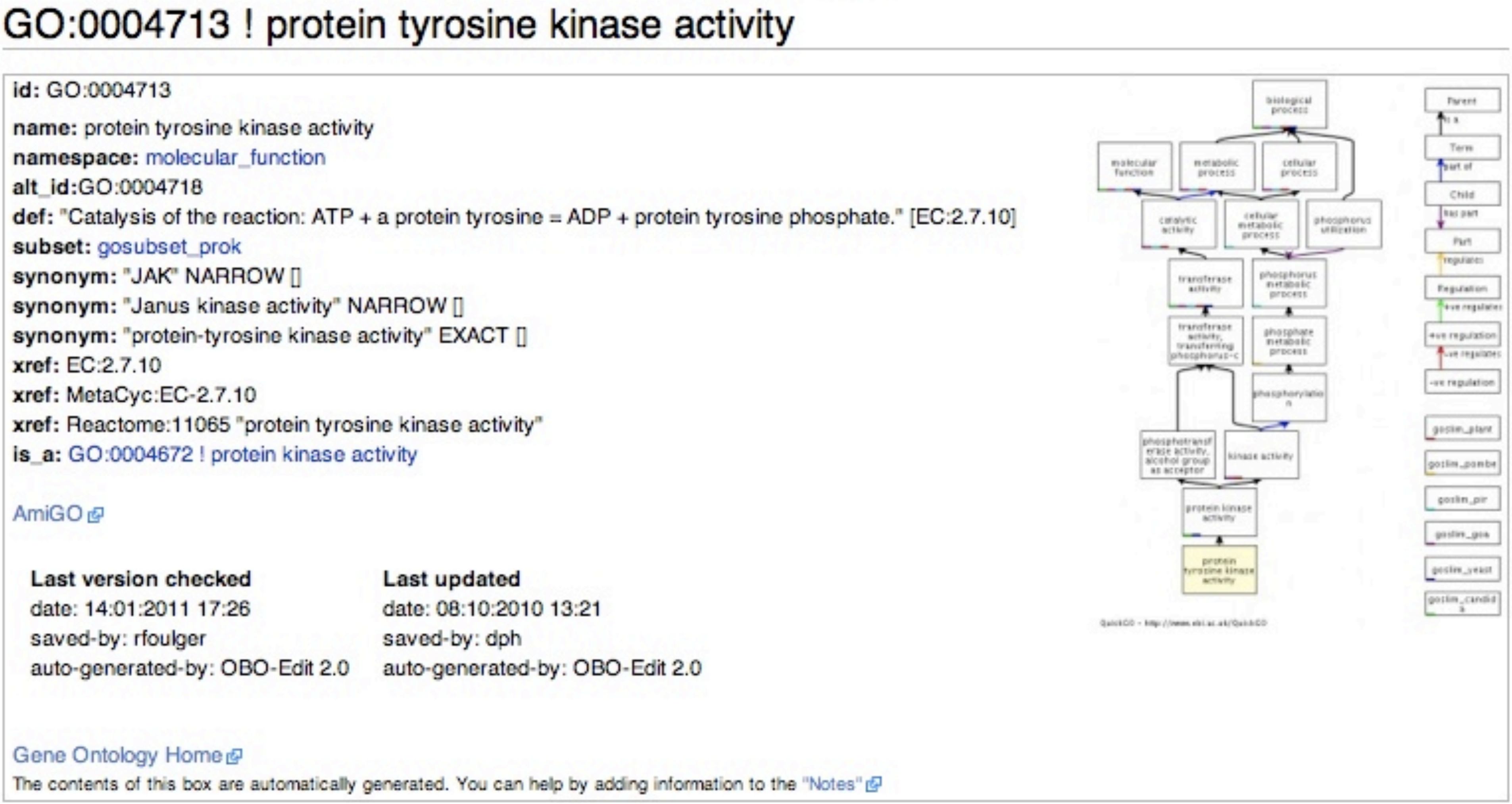
- CHICK:A0M8T9
- CHICK:A0SVH2
- CHICK:BTK

C cont.

- CHICK:Q90960
- CHICK:Q90961
- CHICK:Q90962

F cont.

- FB:Tk4
- FB:Tk6
- FB:tor



GO:0004713 ! protein tyrosine kinase activity

id: GO:0004713

name: protein tyrosine kinase activity

namespace: molecular_function

alt_id:GO:0004718

def: "Catalysis of the reaction: ATP + a protein tyrosine = ADP + protein tyrosine phosphate." [EC:2.7.10]

subset: gosubset_prok

synonym: "JAK" NARROW []

synonym: "Janus kinase activity" NARROW []

synonym: "protein-tyrosine kinase activity" EXACT []

xref: EC:2.7.10

xref: MetaCyc:EC-2.7.10

xref: Reactome:11065 "protein tyrosine kinase activity"

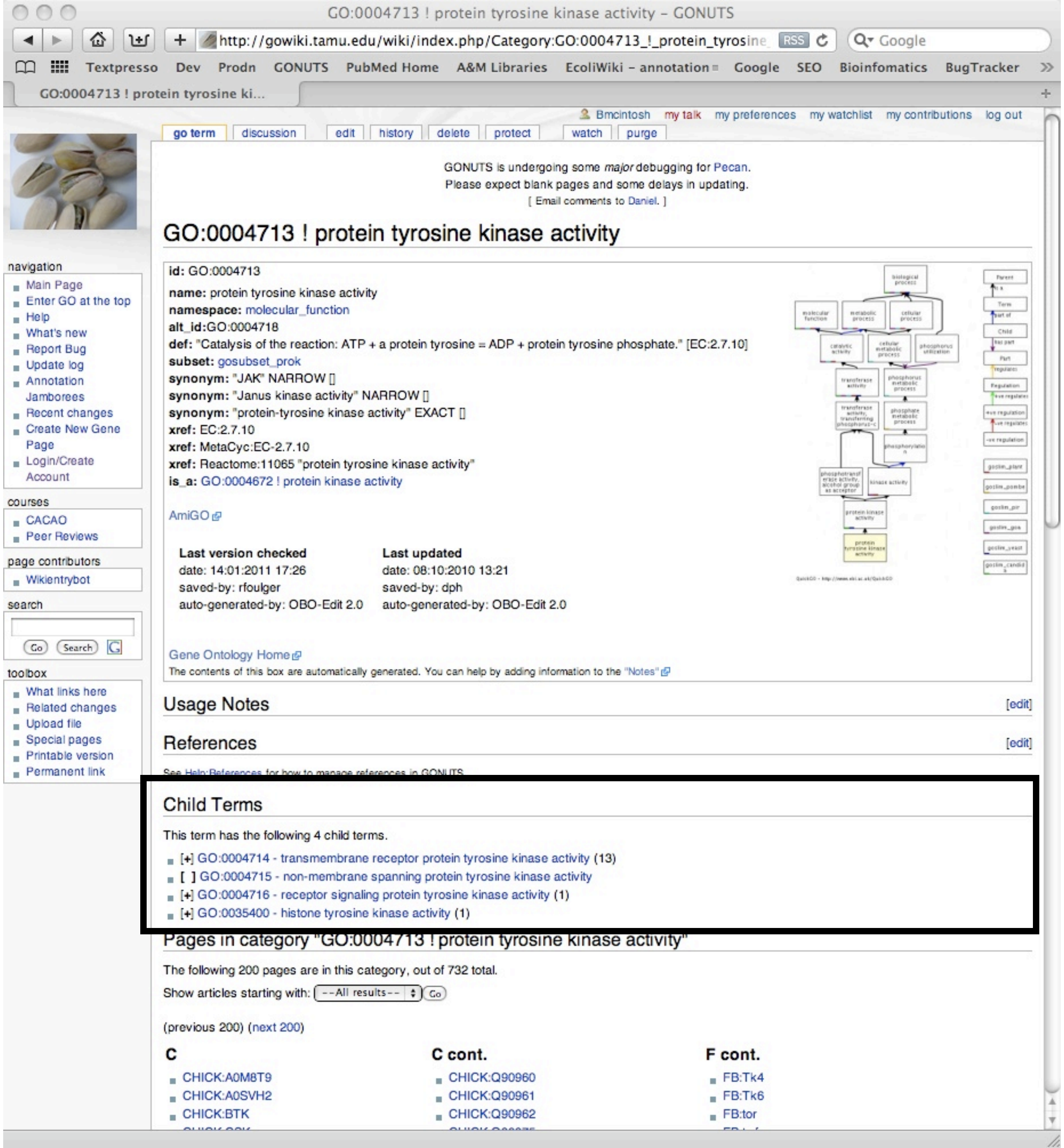
is_a: GO:0004672 ! protein kinase activity

AmiGO

Last version checked	**Last updated**
date: 14:01:2011 17:26	date: 08:10:2010 13:21
saved-by: rfoulger	saved-by: dph
auto-generated-by: OBO-Edit 2.0	auto-generated-by: OBO-Edit 2.0

Gene Ontology Home

The contents of this box are automatically generated. You can help by adding information to the "Notes"



GONUTS is undergoing some *major* debugging for Pecan.
Please expect blank pages and some delays in updating.
[Email comments to Daniel.]

GO:0004713 ! protein tyrosine kinase activity

id: GO:0004713
name: protein tyrosine kinase activity
namespace: molecular_function
alt_id:GO:0004718
def: "Catalysis of the reaction: ATP + a protein tyrosine = ADP + protein tyrosine phosphate." [EC:2.7.10]
subset: gosubset_prok
synonym: "JAK" NARROW []
synonym: "Janus kinase activity" NARROW []
synonym: "protein-tyrosine kinase activity" EXACT []
xref: EC:2.7.10
xref: MetaCyc:EC-2.7.10
xref: Reactome:11065 "protein tyrosine kinase activity"
is_a: GO:0004672 ! protein kinase activity

AmiGO

Last version checked	**Last updated**
date: 14:01:2011 17:26	date: 08:10:2010 13:21
saved-by: rfoulger	saved-by: dph
auto-generated-by: OBO-Edit 2.0	auto-generated-by: OBO-Edit 2.0

Gene Ontology Home
The contents of this box are automatically generated. You can help by adding information to the "Notes"

Usage Notes

[edit]

References

[edit]

See Help:References for how to manage references in GONUTS.

Child Terms

This term has the following 4 child terms.

- [+] GO:0004714 - transmembrane receptor protein tyrosine kinase activity (13)
- [ ] GO:0004715 - non-membrane spanning protein tyrosine kinase activity
- [+] GO:0004716 - receptor signaling protein tyrosine kinase activity (1)
- [+] GO:0035400 - histone tyrosine kinase activity (1)

Pages in category "GO:0004713 ! protein tyrosine kinase activity"

The following 200 pages are in this category, out of 732 total.

Show articles starting with: --All results-- Go

(previous 200) (next 200)

C	C cont.	F cont.
CHICK:A0M8T9	CHICK:Q90960	FB:Tk4
CHICK:A0SVH2	CHICK:Q90961	FB:Tk6
CHICK:BTK	CHICK:Q90962	FB:tor

Strategies

- Search for a keyword and browse the ontology for the right term
  - In GONUTS only search “Category” namespace if you get too many hits
  - Look at the parents, children, and relatives
  - Use Google, Wikipedia etc. to find alternative search terms
- Look at terms suggested by others for your protein
  - Computational with the IEA evidence code
  - Curators with TAS or IC
- Look at terms used for homologous proteins in model organisms

Functional Annotation w/GO

- Annotation: a note that is made while reading any form of text

- GO Annotation: a database entry in a **specific format** that associates a **GO term** with a **gene product** made based on **evidence** in a peer-reviewed **paper**

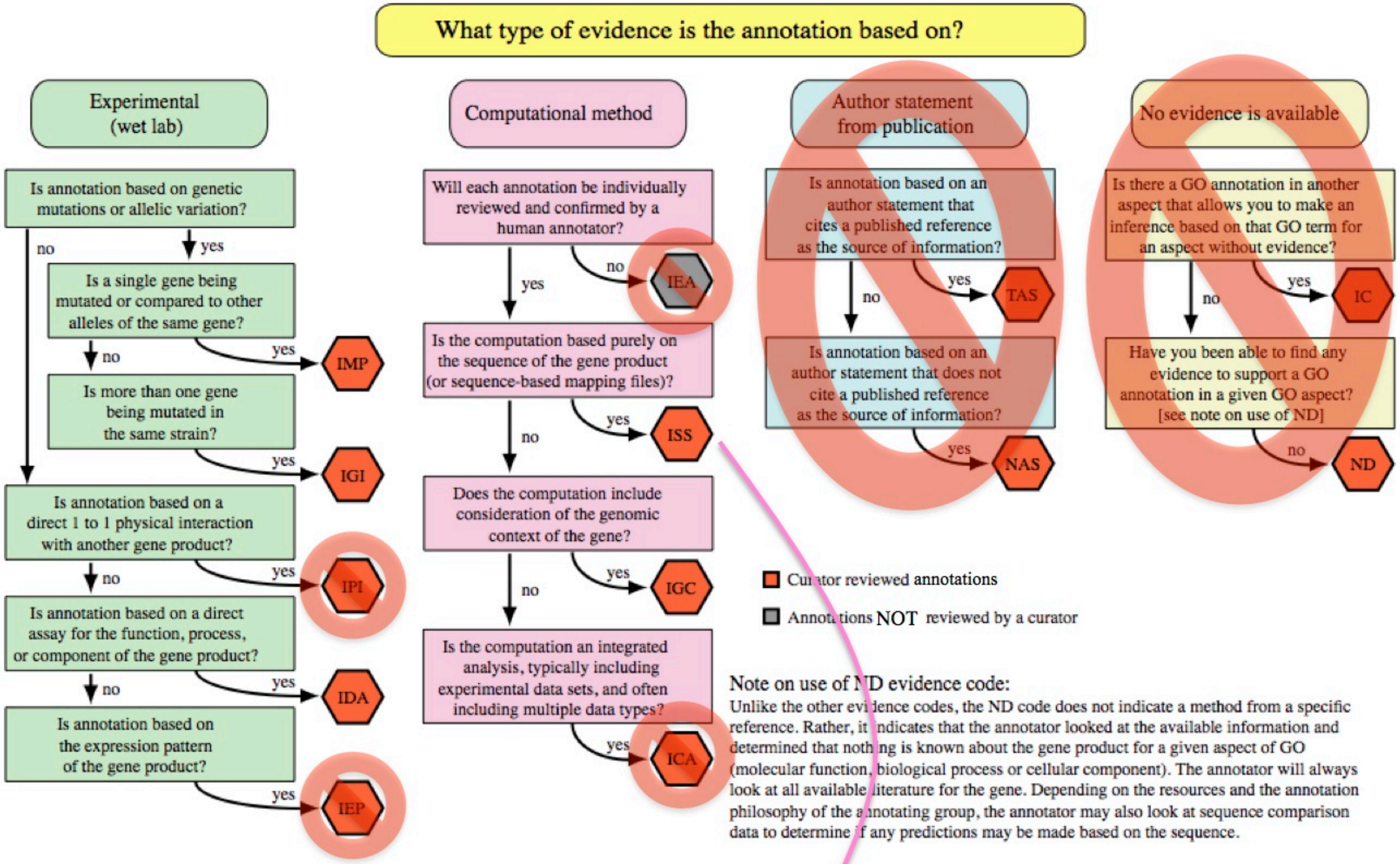
Evidence Codes for CACAO

- Evidence codes describe the type of work or analysis done by the authors
  - IDA: Inferred from Direct Assay
  - IMP: Inferred from Mutant Phenotype
    - NOT just for mutations! Includes inferred from inhibition in vivo by drugs, RNAi, etc.
  - IGI: Inferred from Genetic Interaction
  - ISO: Inferred from Sequence Orthology
  - ISA: Inferred from Sequence Alignment
  - ISM: Inferred from Sequence Model
  - IGC: Inferred from Genomic Context

- Expert biocurators get to use others, but we restrict them for CACAO. If it' s not one of these 7, your annotation is incorrect!!!

- http://gowiki.tamu.edu/wiki/index.php/evidence_codes

Decision Tree to Choose Evidence



What type of evidence is the annotation based on?

Experimental (wet lab)

Is annotation based on genetic mutations or allelic variation?

Is a single gene being mutated or compared to other alleles of the same gene?

Is more than one gene being mutated in the same strain?

Is annotation based on a direct 1 to 1 physical interaction with another gene product?

Is annotation based on a direct assay for the function, process, or component of the gene product?

Is annotation based on the expression pattern of the gene product?

Computational method

Will each annotation be individually reviewed and confirmed by a human annotator?

Is the computation based purely on the sequence of the gene product (or sequence-based mapping files)?

Does the computation include consideration of the genomic context of the gene?

Is the computation an integrated analysis, typically including experimental data sets, and often including multiple data types?

Author statement from publication

Is annotation based on an author statement that cites a published reference as the source of information?

Is annotation based on an author statement that does not cite a published reference as the source of information?

No evidence is available

Is there a GO annotation in another aspect that allows you to make an inference based on that GO term for an aspect without evidence?

Have you been able to find any evidence to support a GO annotation in a given GO aspect? [see note on use of ND]

🟥 Curator reviewed annotations

⬛ Annotations NOT reviewed by a curator

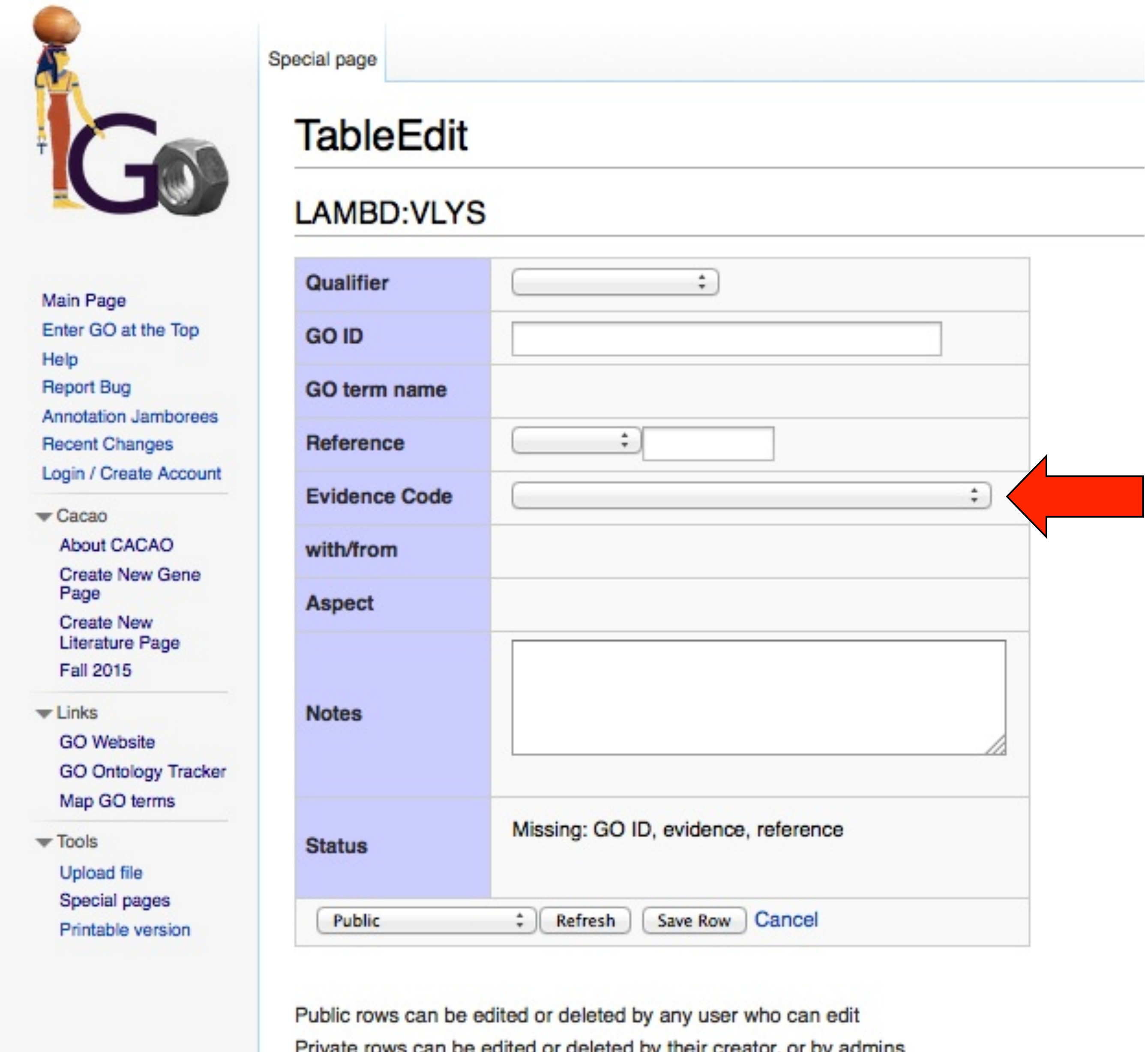
Note on use of ND evidence code:
Unlike the other evidence codes, the ND code does not indicate a method from a specific reference. Rather, it indicates that the annotator looked at the available information and determined that nothing is known about the gene product for a given aspect of GO (molecular function, biological process or cellular component). The annotator will always look at all available literature for the gene. Depending on the resources and the annotation philosophy of the annotating group, the annotator may also look at sequence comparison data to determine if any predictions may be made based on the sequence.

ALLOWED CODES FOR ALL CACAO STUDENTS:

1. **IDA: Inferred from Direct Assay**
2. **IMP: Inferred from Mutant Phenotype**
3. **IGI: Inferred from Genetic Interaction** - requires with/from field to be filled in
4. **ISO: Inferred from Sequence Orthology** - requires with/from field to be filled in
5. **ISA: Inferred from Sequence Alignment** - requires with/from field to be filled in
6. **ISM: Inferred from Sequence Model** - requires with/from field to be filled in
7. **IGC: Inferred from Genomic Context**

Use one of these three codes (ISO, ISA, ISM) if the Decision Tree points you to ISS

Evidence Pull-Down Menu



Special page

TableEdit

LAMBD:VLYS

Qualifier	
GO ID	
GO term name	
Reference	
Evidence Code	
with/from	
Aspect	
Notes	
Status	Missing: GO ID, evidence, reference

Public rows can be edited or deleted by any user who can edit

Private rows can be edited or deleted by their creator, or by admins

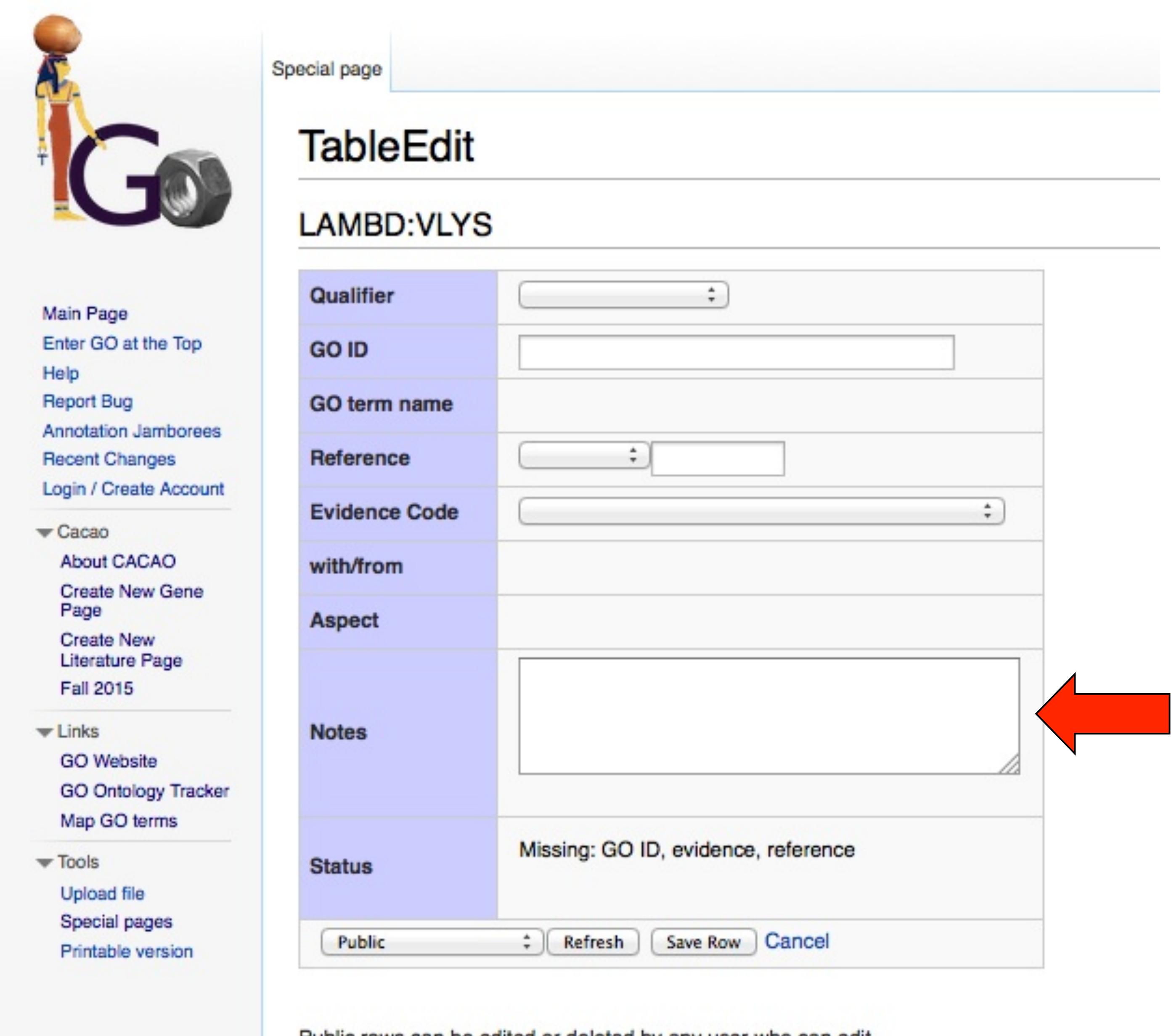
Some Evidence Types Require More Information

- With/from
- Evidence from sequence comparison
  - With the protein accession for the protein you are comparing to
    - That protein must have experimental annotation to the same GO term
- Evidence from computational analysis
  - With the reference for the analysis tool
- Evidence from genetic interaction
  - With the other gene(s) your protein is interacting with

Evidence Codes for CACAO

- Picking the right evidence code is important
- Use the evidence code decision tree
  - http://gowiki.tamu.edu/wiki/images/3/32/CACAO_decisiontree.pdf
- Use the evidence code guidelines at the GO consortium website:
  - http://www.geneontology.org/GO.evidence.shtml
- Discuss!

Note Required for CACAO



TableEdit

LAMBD:VLYS

Qualifier	
GO ID	
GO term name	
Reference	
Evidence Code	
with/from	
Aspect	
Notes	
Status	Missing: GO ID, evidence, reference

Public rows can be edited or deleted by any user who can edit

Example Paper

http://www.ncbi.nlm.nih.gov/pubmed/8227000

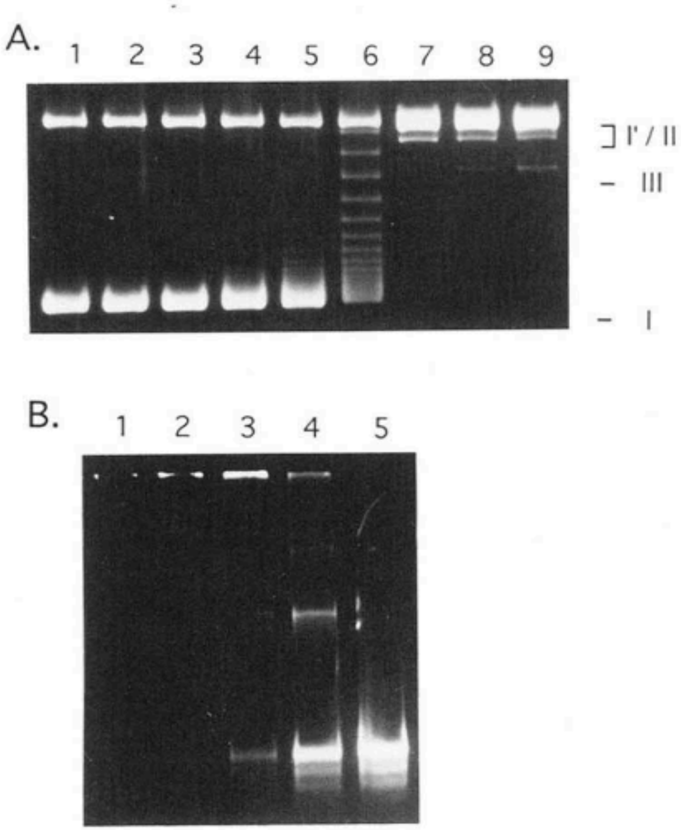
=

http://www.jbc.org/content/268/32/24481.full.pdf

Example Paper: What They Did

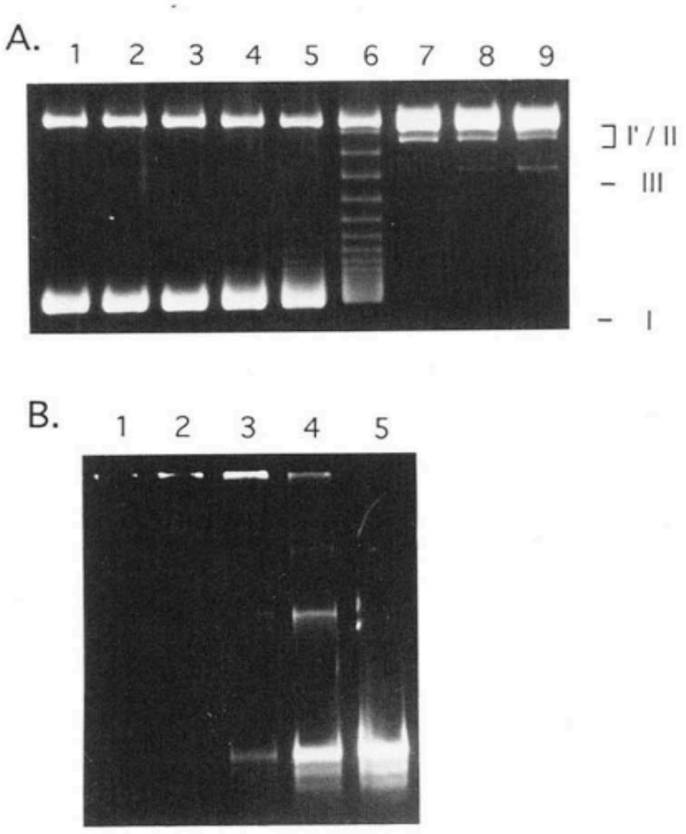
- Finding the proteins
- Do these tell us about the function?
  - Figure 1: sequenced ParC and Part of ParE
  - Figure 2: SDS page of purified proteins
  - Figure 3: Relaxation and decatenation activities of TopoIV
  - …

Example Paper: Figure 3



- Panel A: relaxation
- Panel B: decatenation
- What do these mean?

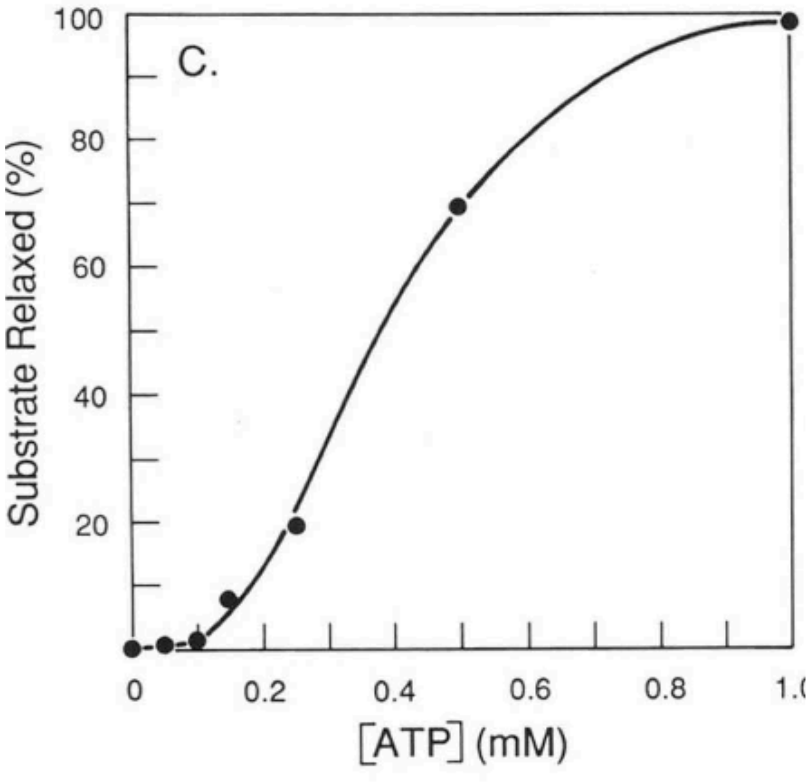
Example Paper: Figure 3



- Panel A: relaxation
- Panel B: decatenation
- What do these mean?

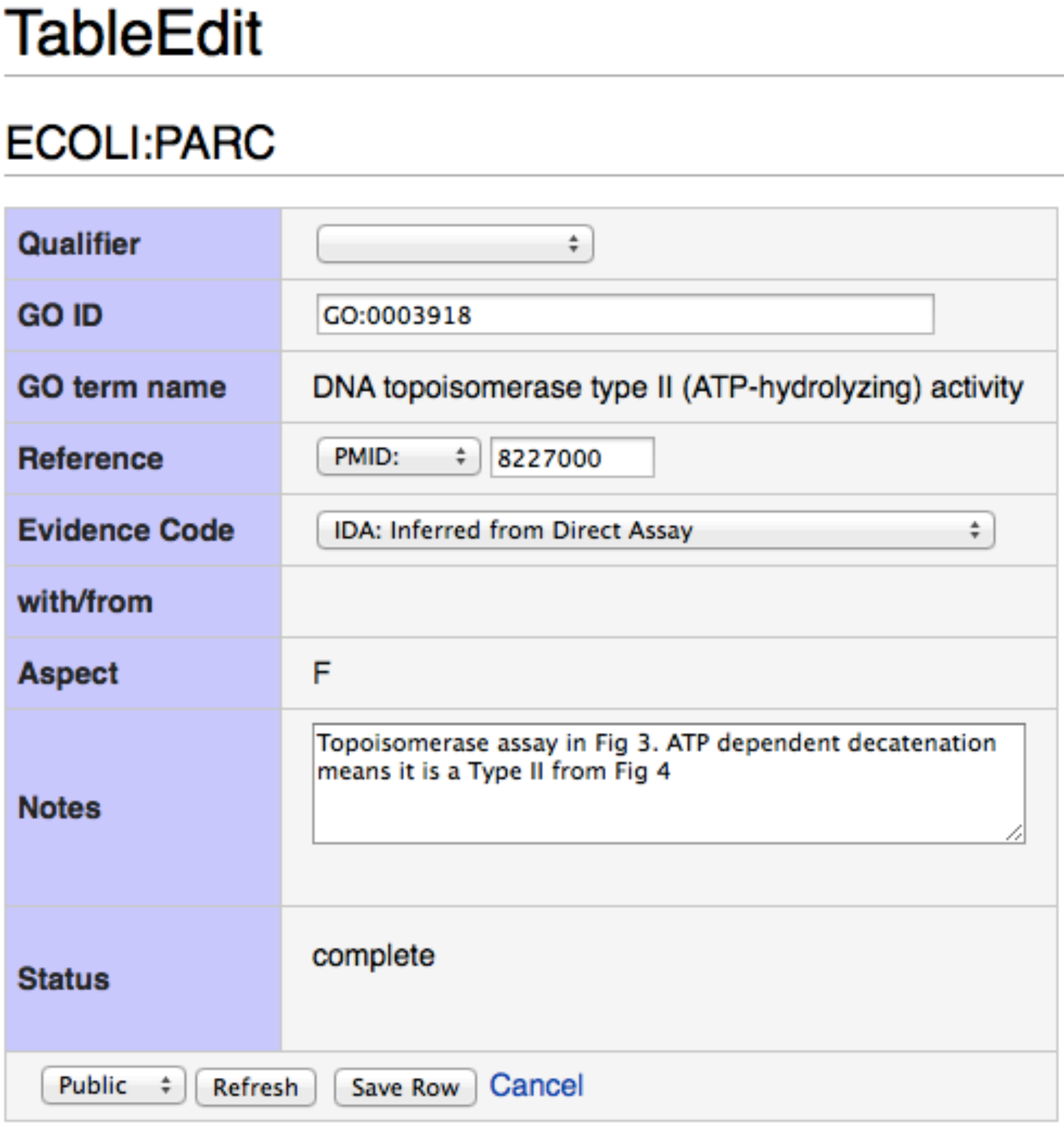
- Panel A shows GO: 0003916 ! DNA topoisomerase activity but does not show what kind
- Panel B shows GO: 0061505 ! DNA topoisomerase II activity

Example Paper: Figure 4



- Shows ATP dependence: GO: 0003918 ! DNA topoisomerase type II (ATP-hydrolyzing) activity

Example Paper: GO annotation for *E. coli* ParC



TableEdit

ECOLI:PARC

Qualifier	
GO ID	GO:0003918
GO term name	DNA topoisomerase type II (ATP-hydrolyzing) activity
Reference	PMID: 8227000
Evidence Code	IDA: Inferred from Direct Assay
with/from	
Aspect	F
Notes	Topoisomerase assay in Fig 3. ATP dependent decatenation means it is a Type II from Fig 4
Status	complete