# BioCreative III Interactive Task: curation bottlenecks and solutions


Phoebe Roberts

15 September 2010

Pfizer

# Agenda

- Acknowledgments
- Recommendations
- Interactive Task gene normalization bottlenecks
- Possible solutions
- Possible metrics
- Questions for UAG panel

If we assume that system won't get it perfect, how can system make it easier to deal with errors and still aid curation?

# Acknowledgments

All participating systems!

*Without your participation, User Advisory Group would have nothing to respond to*

- User Advisory Group
  - Lois Maltais (MGI)
  - Pascale Gaudet (DictyBase)
  - Andrew Chatr-Aryamontri (MiNT)
  - Livia Perfetto (MiNT)
  - Donghui Li (TAIR)
  - Phoebe Roberts (Pfizer)
  - Many others not in attendance today
- IAT organizers
  - Cecilia Arighi
  - Zhiyong Lu
- BioCreative organizers

# Recommendations

- Allow curator to highlight all gene mentions at once
  - Relationships among gene mentions helps normalization
- Interactive systems a must!
  - Allow curator to delete wrong gene mentions
  - Allow curator to add missed gene mentions
    - Include lookup function to find options
- Alert curator to level of ambiguity within species and between
- Present curator with clues for ambiguity resolution
  - Description
  - Synonyms
  - Chromosomal location
  - Sub-cellular localization
  - Interacting partners
    - Less helpful:
      - Other mentions in article
      - Titles of articles with same gene mention
- Allow curator to make decision *in situ* (bring answers to them, don't make them go to answers)
- Allow curator to see decisions and change them

Computational Sciences CoE

RCoEs YouR CoEs

TOGETHER WE MAKE MEDICINES

# DIFFICULT GENE NORMALIZATION EXAMPLES

# How ambiguous is the gene mention & what information helps resolve ambiguity?

J Clin Invest. 2003 June 15; 111(12): 1933–1943. PMCID: PMC161425
doi: 10.1172/JCI200317790.

Copyright © 2003, American Society for Clinical Investigation

**AIP1 mediates TNF-α–induced ASK1 activation by facilitating dissociation of ASK1 from its inhibitor 14-3-3**

from iHOP:

| | | |
|---|---|---|
| BIRC3 | baculoviral IAP repeat-containing 3 | **AIP1** |
| WDR1 | WD repeat domain 1 | **AIP1** |
| PDCD6IP | programmed cell death 6 interacting protein | **AIP1** |
| ABI2 | abl interactor 2 | **AIP1** |
| MAGI2 | membrane associated guanylate kinase, WW and PDZ domain containing 2 | **AIP1** |
| DAB2IP | DAB2 interacting protein | **AIP1** |
| ACTRT1 | actin-related protein T1 | **AIP1** |
| ARL6IP1 | ADP-ribosylation factor-like 6 interacting protein 1 | **AIP1** |

VERY AMBIGUOUS
8 HUMAN GENES!

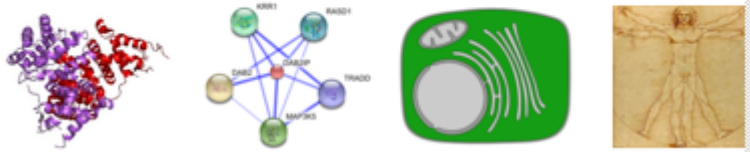
Protein | Add

DAB2IP (ENSP00000362887) H. sapiens

DIP1/2; AF9Q34; KIAA1743

DAB2P_HUMAN, Sequence, Domains, Structure, Locus, Literature

MSAGGSARKSTGRSSYYYRLLRRPRLQRQRSRSRSRTRPARESPQEP



Disabled homolog 2-interacting protein (DAB2 interacting protein) (DAB2 interaction protein) (ASK-interacting protein 1) Functions as a

ANSWER: gene description

# Another ambiguity resolution problem

PLoS ONE. 2008; 3(4): e1948. PMCID: PMC2275796
Published online 2008 April 9. doi: 10.1371/journal.pone.0001948.



**Association of Common Polymorphisms in *GLUT9* Gene with Gout but Not with Coronary Artery Disease in a Large Case-Control Study**

Klaus Stark,[#1] Wibke Reinhard,[#1] Katharina Neureuther,[1] Silke Wiedmann,[1] Kamil Sedlacek,[1] Andrea Baessler,[1] Marcus Fischer,[1] Stefan Weber,[1] Bernhard Kaess,[1] Jeanette Erdmann,[2] Heribert Schunkert,[2] and Christian Hengstenberg[1*]

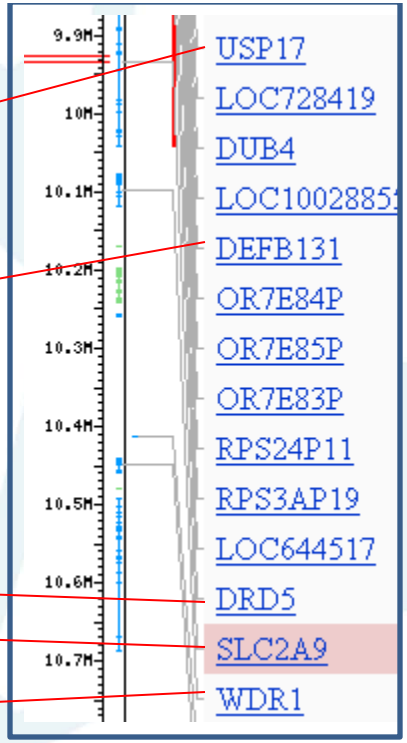
from iHOP:

| | | |
|---|---|---|
| SLC2A6 | solute carrier family 2 (facilitated glucose transporter), member 6 | GLUT9 |
| SLC2A9 | solute carrier family 2 (facilitated glucose transporter), member 9 | GLUT9 |

SORT OF AMBIGUOUS


Figure 1 from paper

answer: chromosome map



# Species ambiguity: fair or foul?

Use case determines need for species specificity

- Sometimes species is required for curating:
  - model organism database info
  - Protein-protein interactions with definitive molecules
- Sometimes species interferes:
  - Authors are making point about evolutionary conservation of finding
    - Target of thalidomide http://www.ncbi.nlm.nih.gov/pubmed/20223979
  - Most proteins not sufficiently characterized to warrant species filtering
    - E.g. at Pfizer, our protein dictionary is mammalian, not human, to increase recall of biological process and function

# Earth-shattering conclusion derived from results across species

*Science* 12 March 2010:
Vol. 327. no. 5971, pp. 1345 - 1350
DOI: 10.1126/science.1177319

RESEARCH ARTICLES

## Identification of a Primary Target of Thalidomic

Takumi Ito,[1,*] Hideki Ando,[2,*] Takayuki Suzuki,[3,4] Toshil
Yuki Yamaguchi,[2] Hiroshi Handa[1,2,†]

Half a century ago, thalidomide was widely prescribed to pregnant women as a sedative but was found to be teratogenic, causing multiple birth defects. Today, thalidomide is still used in the t of leprosy and multiple myeloma, although how it causes limb malformation and other ntal defects is unknown. Here, we identified cereblon (CRBN) as a thalidomide-binding protein. CRBN forms an E3 ubiquitin ligase complex with damaged DN ding protein 1 (DDB1) and Cul4A that is important for limb outgrowth and expression of the f factor Fgf8 in zebrafish and chicks. Thalidomide initiates its teratogenic effects by binding to CRBN and inhibiting the associated ubiquitin ligase activity. This study reveals a basis for thalidomide tagenicity and may contribute to the development of new thalidomide derivati

ZF, chick result

Human result

Conclusion from results: non-specific

# Difficult gene mentions

Synonym not found

- Synonym is not found in databases searched (FN)
  - AtHscB (PMC2764847)
- Synonym is not found in **any** databases (FN)
  - Arabidopsis examples (PMC2764847)
- Species prefix obfuscates synonym (FN)
  - AtHscB (PMC2764847)

Ambiguity

- Synonym is a common English word (FN not in dictionary/FP many hits)
  - WASp
- Synonym maps to more than one identifier (FN for missed mapping/FP for wrong mapping)
  - AIP1
- Species not clearly specified (FP)
  - AIP1/ALIX
- Species deliberately not specified (FP)
  - One of the AIP1 papers
- Synonym is adjective that modifies a non-gene (FP)
  - SufD-like protein (PMC2764847)
- Synonym refers to a protein family or an enzymatic activity (FP)
  - ATPases (PMCID 2275796)
  - Not appropriate to map to an identifier, BUT still some utility from annotating it

Computational Sciences CoE

RCoEs YouR CoEs

TOGETHER WE MAKE MEDICINES

# SOLUTIONS

# Difficult gene mentions and solutions

How can a curator more easily resolve a…

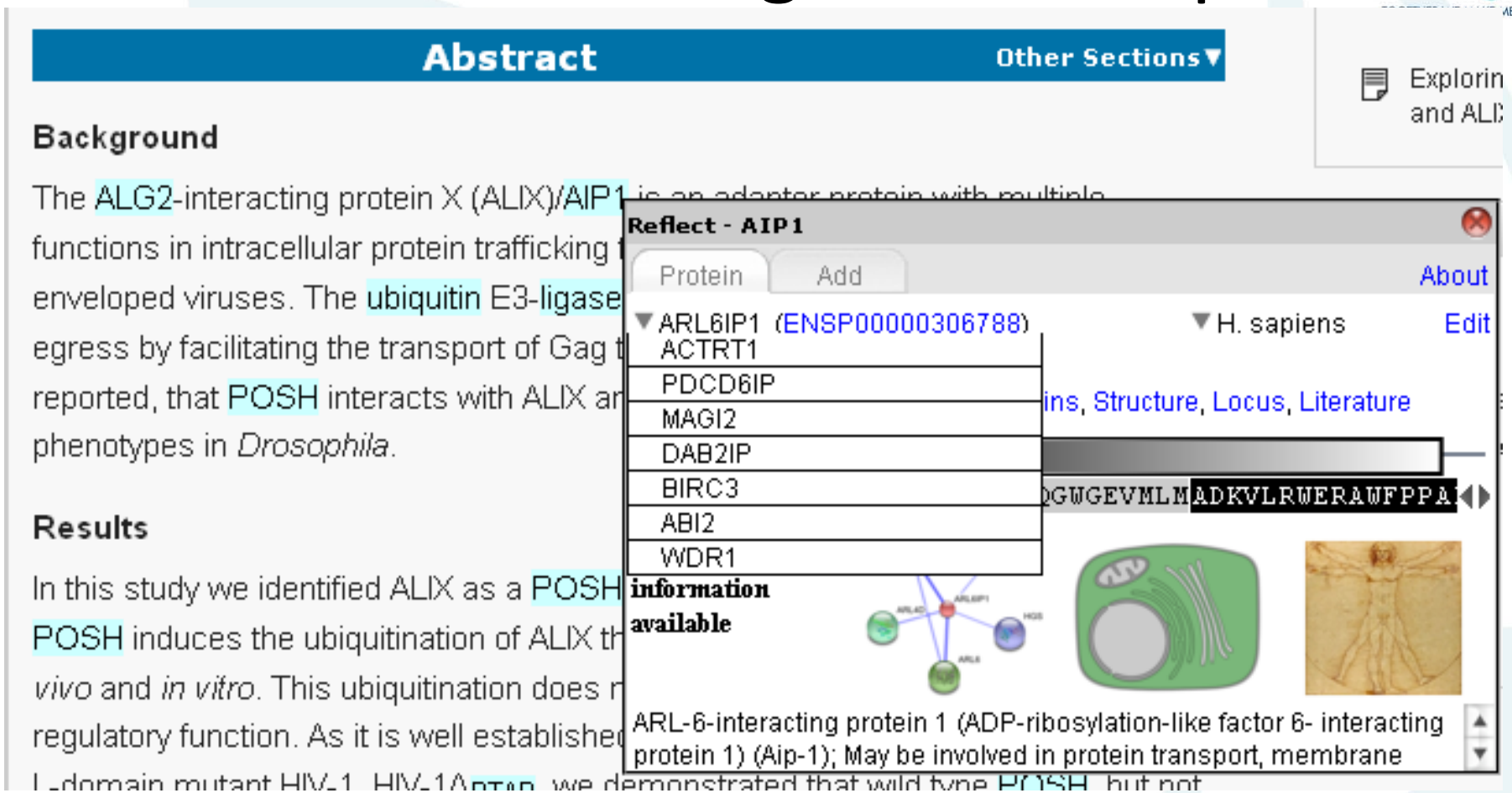
Synonym not found

- New synonym is not found in any databases (FN)
  - Increase breadth of databases searched by tool
- Synonym is not found in all databases (FN)
  - Ability to add a synonym and reprocess highlighting
- Species prefix obfuscates synonym (FN)
  - Ability to add synonym or species-specific rules for string matching

Ambiguity

- Synonym is a common English word (FN not in dictionary /FP many hits)
  - Ability to add or remove a synonym and reprocess highlighting
- Synonym maps to more than one identifier (FN for missed ID/FP for wrong ID)
  - Present matches simultaneously with clues like other synonyms and interacting partners
- Species not clearly specified
  - Be able to navigate to other sections of the paper, other papers
- Species deliberately not specified (FP????)
  - Navigate to references
- Synonym is adjective that modifies a non-gene (FP)
  - Ability to remove from list
- Synonym refers to a protein family or an enzymatic activity (FP)
  - Ability to removed from list

# Methods for dealing with multiple hits




View options, check each, don't have to commit, commitment is propogated

# Synonym lookup functions to aid GN

IscU/Isu, which is regulated by HscB/Jac1 by binding to IscU/Isu to assist [Fe-S] delivery to the chaperone [12], [44]. Yeast Jac1, Ssq1 and Isu have been confirmed to be mitochondrial proteins [12].

Here we demonstrate that Arabidopsis contains a functional AtHscA1/AtHscB/AtIscU1 protein cluster involved in [Fe-S] protein biogenesis. In contrast to yeast, the AtHscA1/AtHscB/AtIscU1 protein cluster is localized to both mitochondria and the cytosol of Arabidopsis suggesting a dual action between these two spatially separate compartments.

**Results**

**AtHscB can rescue yeast Jac1knockout mutant**

A full-length cDNA (759 nt) encoding the At5g06410 open reading frame was cloned
to *E. coli* HscB and
respectively (Figur
contains the HPD
a predicted 59 ami
according to the C
in Arabidopsis has
At5g06410 AtHscB

To confirm that At
protein, we perfor
lethal yeast knock
transformed Delta
(*URA3*) marked pla
AtHscB and positiv
dropout media SD/-Leu (minus L-leucine) or on SD/-Trp (minus L-trypothan), respectively. Once scored the wild-type *Jac1* cDNA

**Reflect - At5g06410**

Protein | Add | About

At5g06410 (AT5G06410.1) A. thaliana

No Synonyms

Sequence, Domains, Structure, Locus, Literature

**No structure information available**

No annotation available

# New evaluation metrics

- Time spent in system (please include PAUSE button!)

Utility

- Increased TPs and TNs, decreased FPs and FNs
- Novelty
- Number of URLs visited outside application
- Indicator of resources provided within application and how often curator has to leave to accomplish task
- Subjective metric
  - Did curator enjoy the user experience?

# Other questions

System specification questions

- Is any identifier better than none?
  - Entrez, UniProt, IMAGE, TAIR, etc.
- Is one application with many functions better than multiple applications (e.g. browser plus spreadsheet)
- Is finding the identifier the hardest part of GN, or determining whether an identifier is appropriate?

More research needed?

- Is there a point at which too many FPs are worse than FNs?
- Do we need to know how curators normalize gene mentions?
- Do we need to know the frequency of GN difficult tasks?
- Will someone build a first mention corpus?
  - Here we describe LINGO-1, a nervous system-specific transmembrane protein that binds NgR1 and p75 and that is an additional functional component of the NgR1/p75 signaling complex. (pmid:14966521)
  - We describe a yeast enzyme, Doa4, that is integral to the degradation of ubiquitinated proteins and is required in diverse physiological processes. PMID: 8247125

# Who is curation tool for?