# Using computational predictions to improve literature-based Gene Ontology annotations

Julie Park, Ph.D.

*Saccharomyces* Genome Database • http://www.yeastgenome.org/
Department of Genetics • Stanford University School of Medicine

# Attaining curation nirvana…

- Curation efficiency
- Annotation consistency
- Data accuracy

# …is not easy!

## Annotation errors

1. Mistakes in capturing the annotation

2. Outdated information

3. Missing annotations

*How can you find these errors?*

# Flavors of GO annotations

1. Literature-based – “Manual”

   Individually assigned by biocurators based on the published literature

2. Computationally-predicted – “Computational”

   Automatically generated by *in silico* methods such as protein signatures or computational algorithms

Sources of computational predictions in SGD

InterPro [1]

Swiss-Prot Keywords (SPKW) [2]

YeastFunc [3]

BioPixie [4]

[1] Camon, et al (2003) *Genome Res.* 13:662-72

[2] http://www.ebi.ac.uk/GOA/Swiss-ProtKeyword2GO.html

[3] Tian, et al (2008) *Genome Biol.* 9 Suppl 1:S7

[4] Huttenhower and Troyanskaya (2008) *Bioinformatics*. 24:i330-8

# Flavors of GO annotations

1. Literature-based – "Manual"

   Individually assigned by biocurators based on the published literature

2. Computationally-predicted – "Computational"

   Automatically generated by *in silico* methods such as protein signatures or computational algorithms

Sources of computational predictions in SGD

InterPro [1]

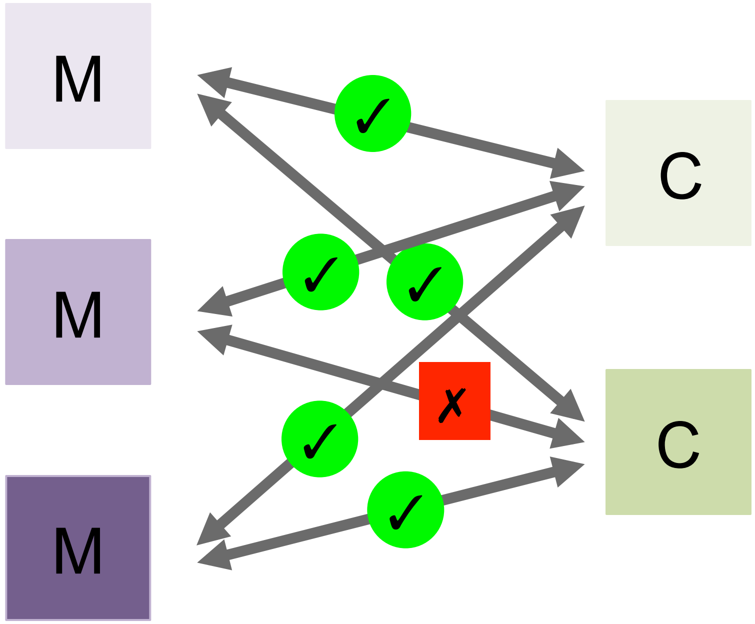
Swiss-Prot Keywords (SPKW) [2]

YeastFunc [3]

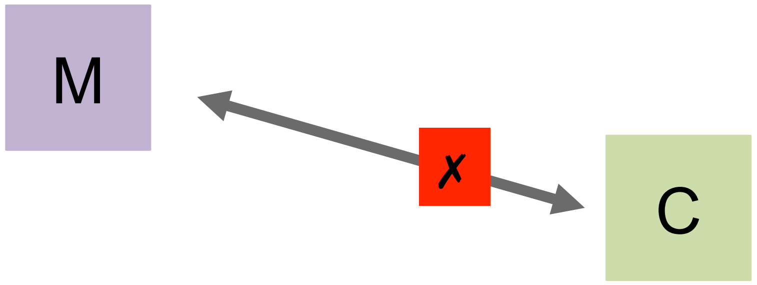
BioPixie [4]

*Is it possible to take advantage of the strengths of computational predictions and leverage these annotations to improve manual ones?*

# CvManGO:
# Computational vs. Manual GO Annotations



# CvManGO:
# Computational vs. Manual GO Annotations

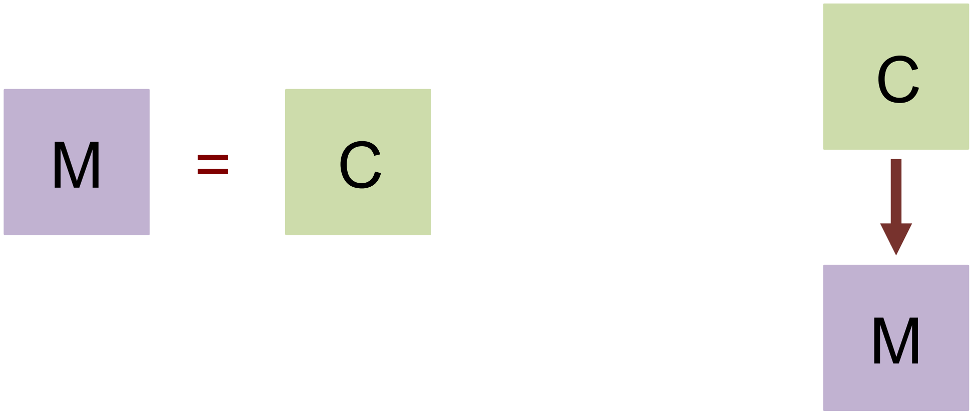


Do discrepancies between a literature-based annotation and a computational prediction indicate that the manual annotation needs to be updated?

# CvManGO:
# Computational vs. Manual GO Annotations

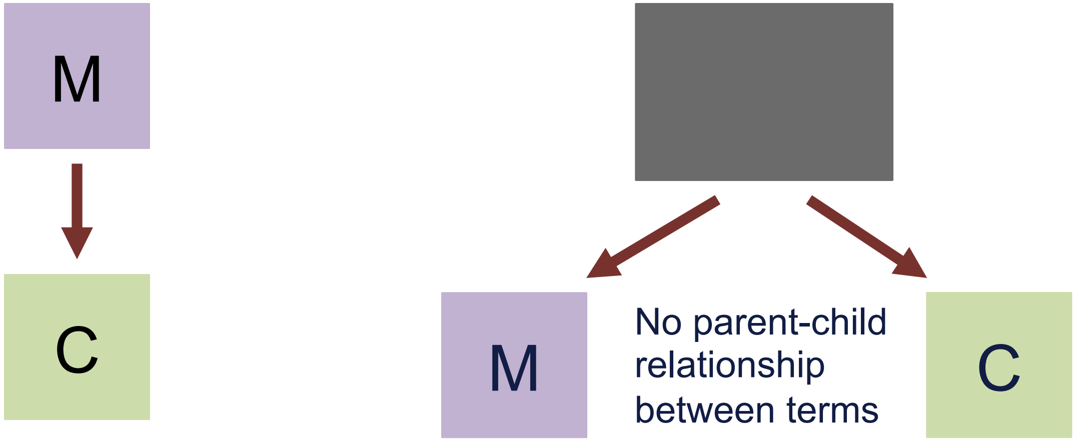
OK

✓

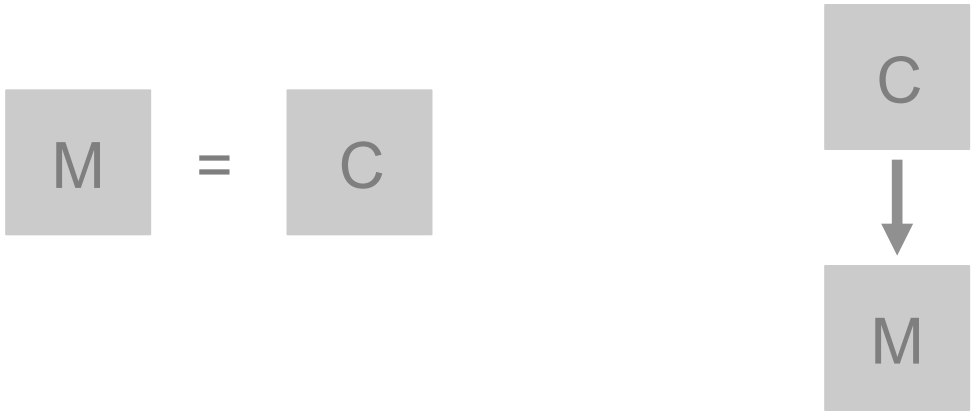


Discrepancies

✗



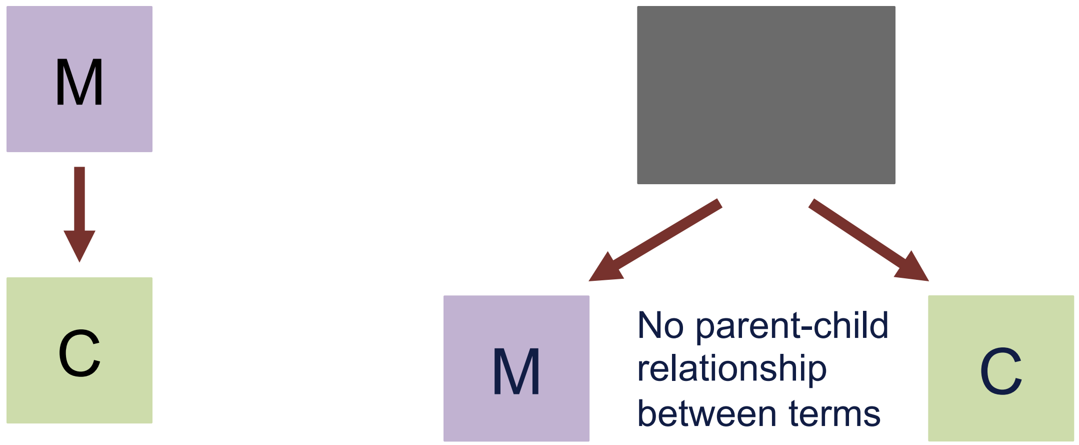
# CvManGO:
# Computational vs. Manual GO Annotations



4379 genes

↓

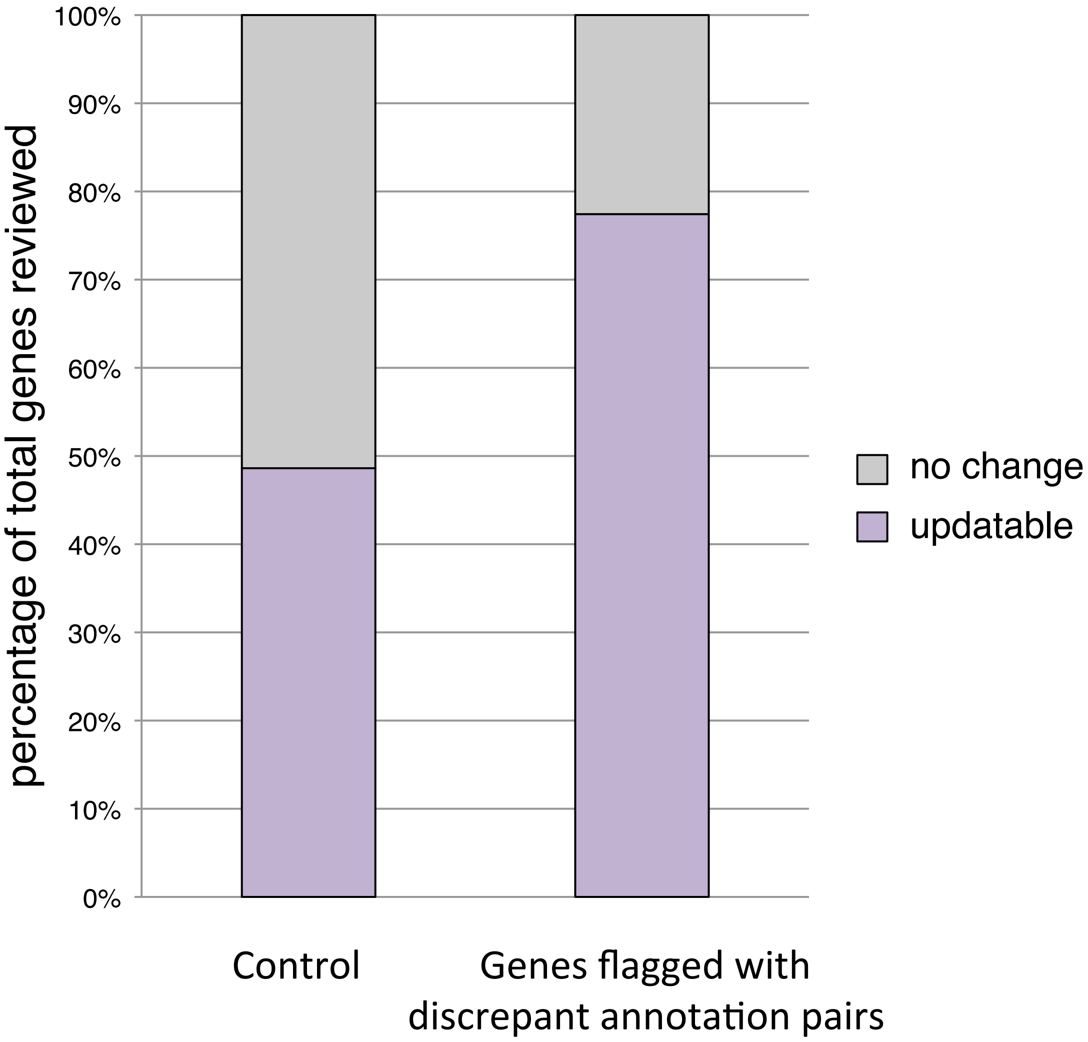
Reviewed
336 genes



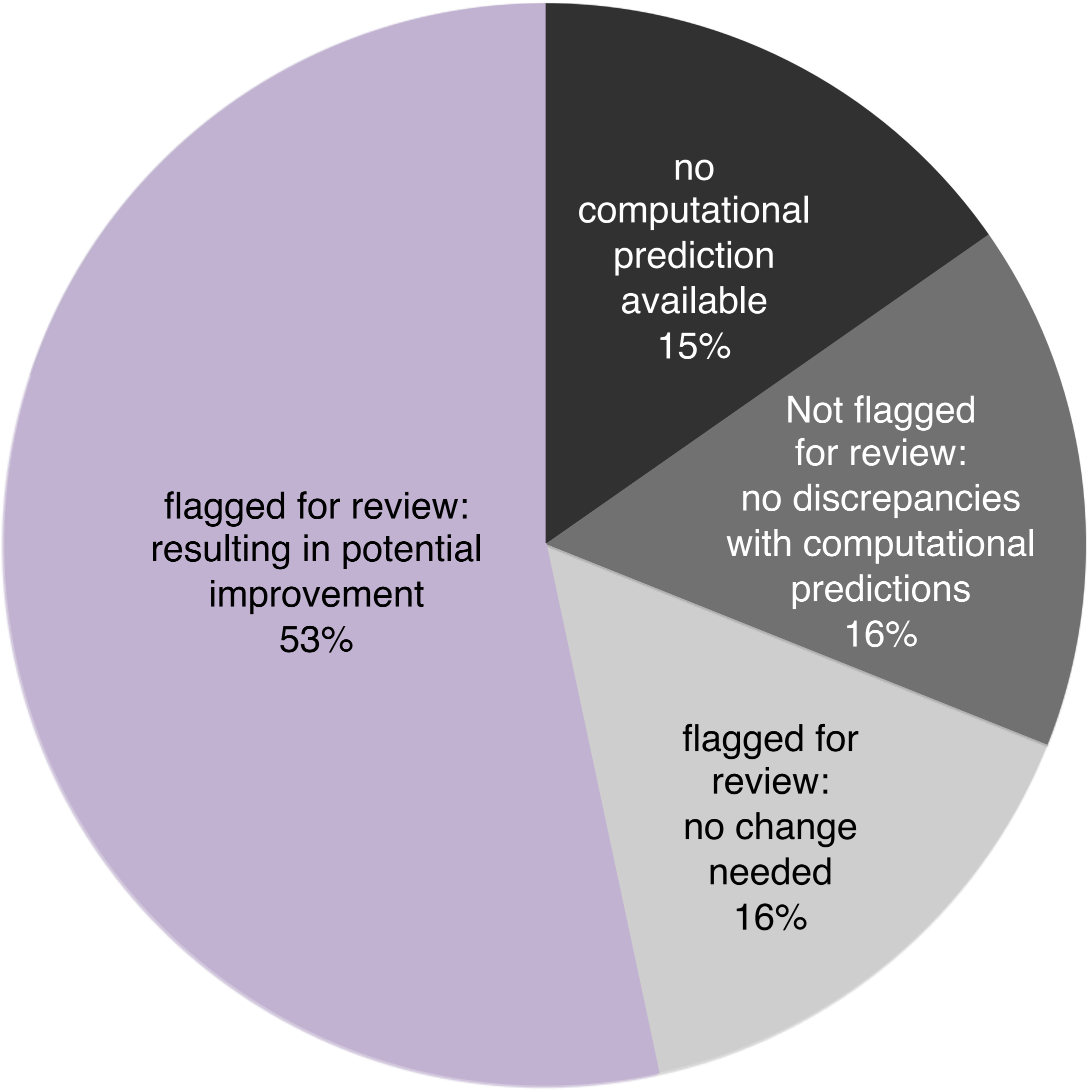
No parent-child relationship between terms

*from October 2009 gene_association.sgd file*
*(6353 total genes)*

# Discrepancies can identify genes that need updating

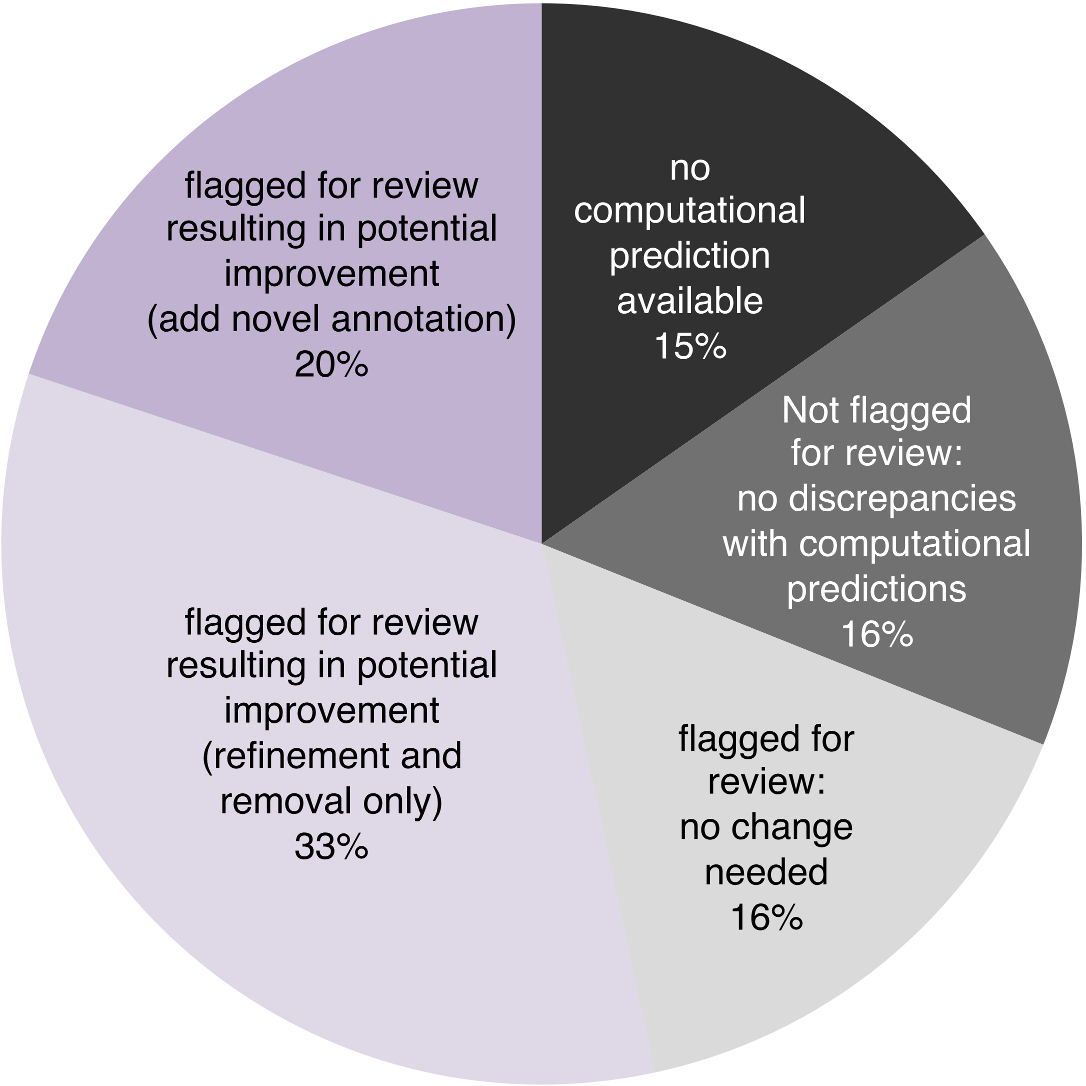


# Extrapolating to the entire genome

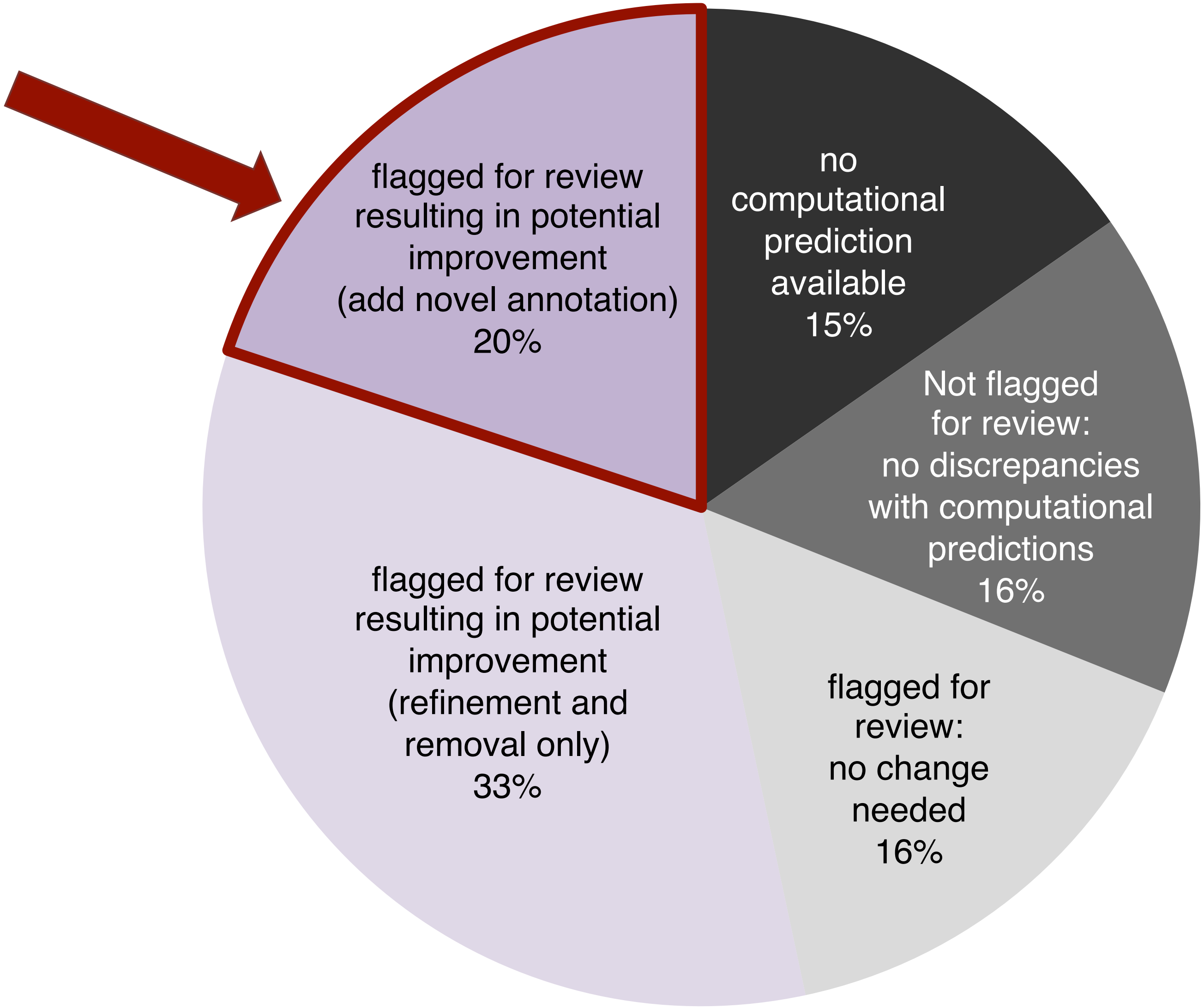


Still requires reviewing 4379/6353 genes—can we narrow this down further?

# Factoring in the type of update



# Factoring in the type of update



# Attributes of flagged genes

What are factors that enrich for genes missing annotations?

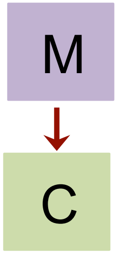
- Type of discrepancy
- GO aspect
- Amount of literature for a gene
- Source of the computational prediction
- Number of computational sources with discrepancies

# Attributes of flagged genes

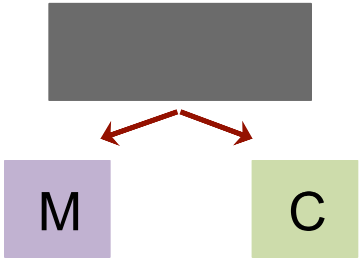
What are factors that enrich for genes missing annotations?

- Type of discrepancy
- ~~GO aspect~~
- ~~Amount of literature for a gene~~
- ~~Source of the computational prediction~~
- ~~Number of computational sources with discrepancies~~

# Analysis by Class of Discrepancies

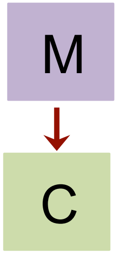
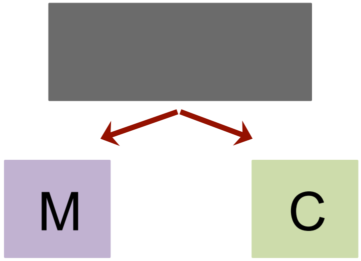


Shallow class

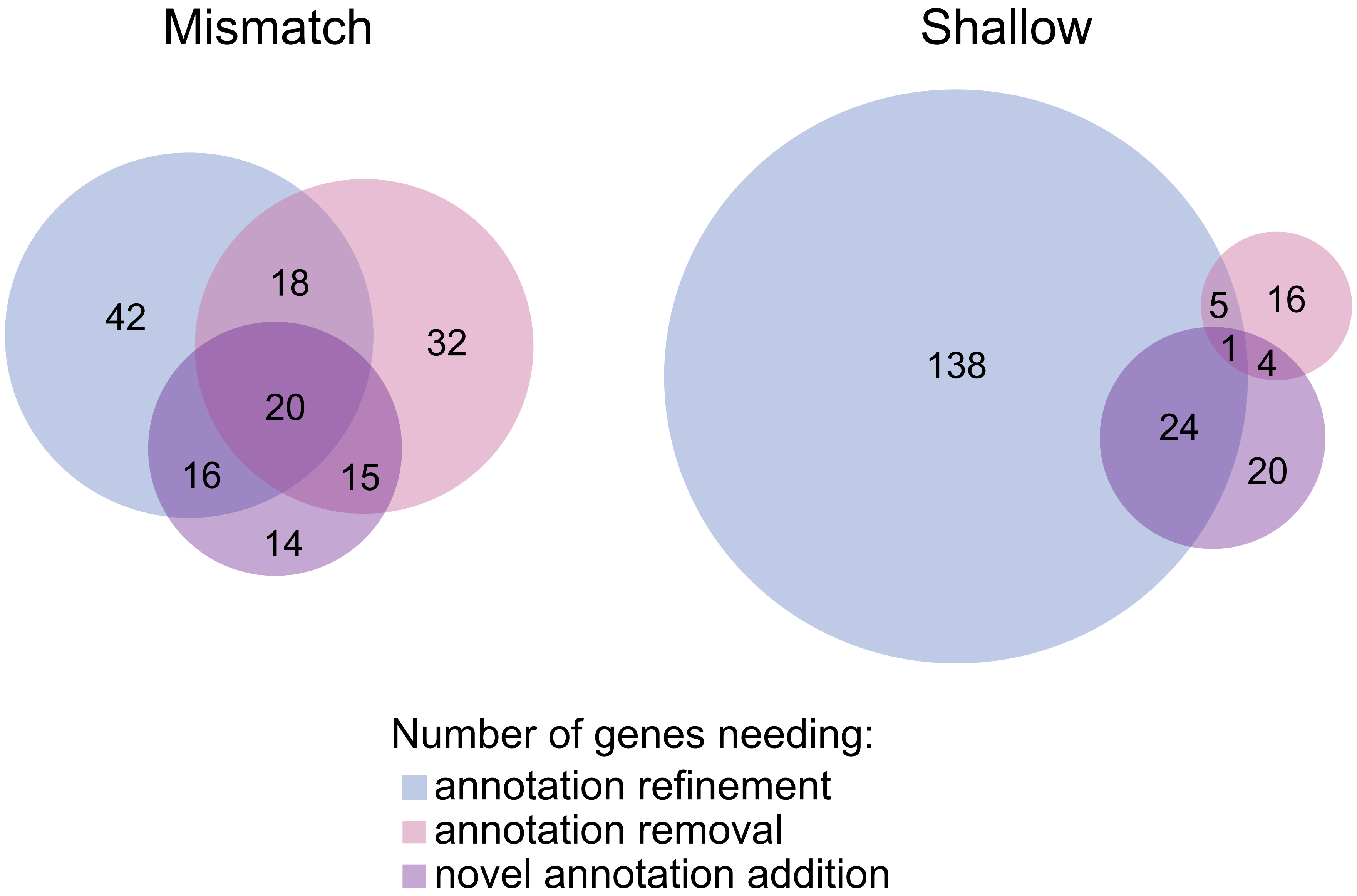


Mismatch class

# Analysis by Class of Discrepancies

| | | % updatable |
|---|---|---|
|  | Shallow class | 78.8% |
|  | Mismatch class | 59.2% |

# Types of annotation updates by class



# Summary & Conclusions

- Majority of *S. cerevisiae* literature-based GO annotations are good

- Comparing manual vs. computational prediction can identify genes whose annotations need updating

- Additional work needs to be done to pinpoint these annotations and genes

# Summary & Conclusions

- Majority of *S. cerevisiae* literature-based GO annotations are good
- Comparing manual vs. computational prediction can identify genes whose annotations need updating
- Additional work needs to be done to pinpoint these annotations and genes

It works but
there is still work to do!

# Future plans

- Identify predictive features of genes that need updating
  - Are there specific GO terms used for manual curation more likely to be updated?
  - Do specific computational predictions indicate a GO term should be updated?
  - Examine node distance between GO terms used for computational and literature-based annotations
  - Examine contribution of annotation date and new publications
  - A combination of or all of the above?

- Evaluate the accuracy of computational predictions for *S. cerevisiae*

- Expand to evaluate annotations made based on orthology
  - Annotations from GOC PAINT project

- Develop a pipeline for curation prioritization at SGD

- Extend to other annotation projects

# GO Consortium

RGD
MGI
GRAMENE
J. Craig Venter
INSTITUTE
tair
BDGP
SGD
GeneDB
IGS
GO
WormBase
The Biology and Genome of C. elegans.
FlyBase
wellcome trust
sanger
institute
dictyBase
UniProt
EMBL-EBI
GO
Annotation@EBI
ZFIN
REACTOME
InterPro
AstraZeneca
TGD
CGD
PAMGO
PLANT-ASSOCIATED MICROBE
GENE ONTOLOGY

http://www.geneontology.org/

# *Saccharomyces* Genome Database staff

@yeastgenome

http://on.fb.me/ksgskb

sgd-helpdesk@lists.stanford.edu

Supported by NIH, National Human Genome Research Institute