

Comparison of computationally- and manually-assigned Gene Ontology annotations to improve functional characterization of gene products

Maria C. Costanzo, Rama Balakrishnan, Karen R. Christie, Eurie L.
Hong, Julie Park, J. Michael Cherry, and The *Saccharomyces*
Genome Database Project. Department of Genetics, Stanford
University, Stanford, CA, USA

Biocuration 2010
October 12, 2010



The *Saccharomyces* Genome Database (SGD)

www.yeastgenome.org

SGD Search

Site Map | Search Options | Help | Home

Community Info | Submit Data | BLAST | Primers | PatMatch | Gene/Seq Resources | Advanced Search | Community Wiki

RVS167/YDR388W Summary

Summary | Locus History | Literature | Gene Ontology | Phenotype | Interactions | Expression | Protein | Wiki

RVS167 BASIC INFORMATION

Standard Name	RVS167 ^{1, 2}
Systematic Name	YDR388W
Feature Type	ORF, Verified
Description	Actin-associated protein, interacts with Rvs161p to regulate actin cytoskeleton, endocytosis, and viability following starvation or osmotic stress; homolog of mammalian amphiphysin (1, 3, 4, 5)
Name Description	Reduced Viability on Starvation ¹

GO Annotations

All RVS167 GO evidence and references

[View Computational GO annotations for RVS167](#)

Molecular Function

Manually curated

- cytoskeletal protein binding (IPI, ISS)
- contributes_to lipid binding (IDA)

Biological Process

Manually curated

- actin cortical patch localization (IMP)
- endocytosis (IMP)
- lipid tube assembly (IDA)
- vesicle-mediated transport (IGI, IPI)

Cellular Component

Manually curated

- actin cortical patch (IDA)
- mating projection tip (IDA)

High-throughput

- mating projection tip (IDA)

Mutant Phenotype

All RVS167 Phenotype details and references

Classical genetics

null

- actin cytoskeleton morphology: abnormal
- Lucifer yellow accumulation: abnormal
- resistance to sodium chloride: decreased

RVS167 RESOURCES

Click on map for expanded view

SGD ORF map | GBrowse

1249000 to 1254000

Chr IV

1250k

YDR387C

YDR388W

- Literature
 - Literature Guide | View
- Retrieve Sequences
 - Genomic DNA | View
- Sequence Analysis Tools
 - BLASTP | View
- Protein Info & Structure
 - Protein Info | View
- Localization Resources
 - YeastRC Localization (Seattle) | View
- Interactions
 - BioGRID (Toronto) | View
- Phenotype Resources
 - PROPHECY | View
- Maps & Displays
 - Chromosomal Features Map | View
- Comparison Resources
 - Fungal Alignment | View
- Functional Analysis
 - Expression Connection Summary | View

Characterization of Open Reading Frames



■ 4884 ORFs, 73.92% ■ 913 ORFs, 13.82% ■ 810 ORFs, 12.26%



Types of GO annotation in SGD

Manually curated

- assigned individually by curators based on the published literature

High-throughput

- based on published large-scale experiments; individual annotations are not necessarily reviewed by curators

Computational

- predictions assigned by an external source



Computational GO annotations in SGD are derived from several different sources

Source	Method
UniProt (InterPro)	InterPro domains in UniProt entries mapped to GO terms
UniProt (SPKW)	Swiss-Prot keywords in UniProt entries mapped to GO terms
UniProt (E.C. number)	E.C. numbers in UniProt entries mapped to GO terms
BioPIXIE	Algorithm uses a protein-protein linkage map derived from diverse genomic data to predict a process-specific network
YeastFunc	Algorithm integrates protein-protein and genetic interactions, expression patterns, protein domains, protein complex membership

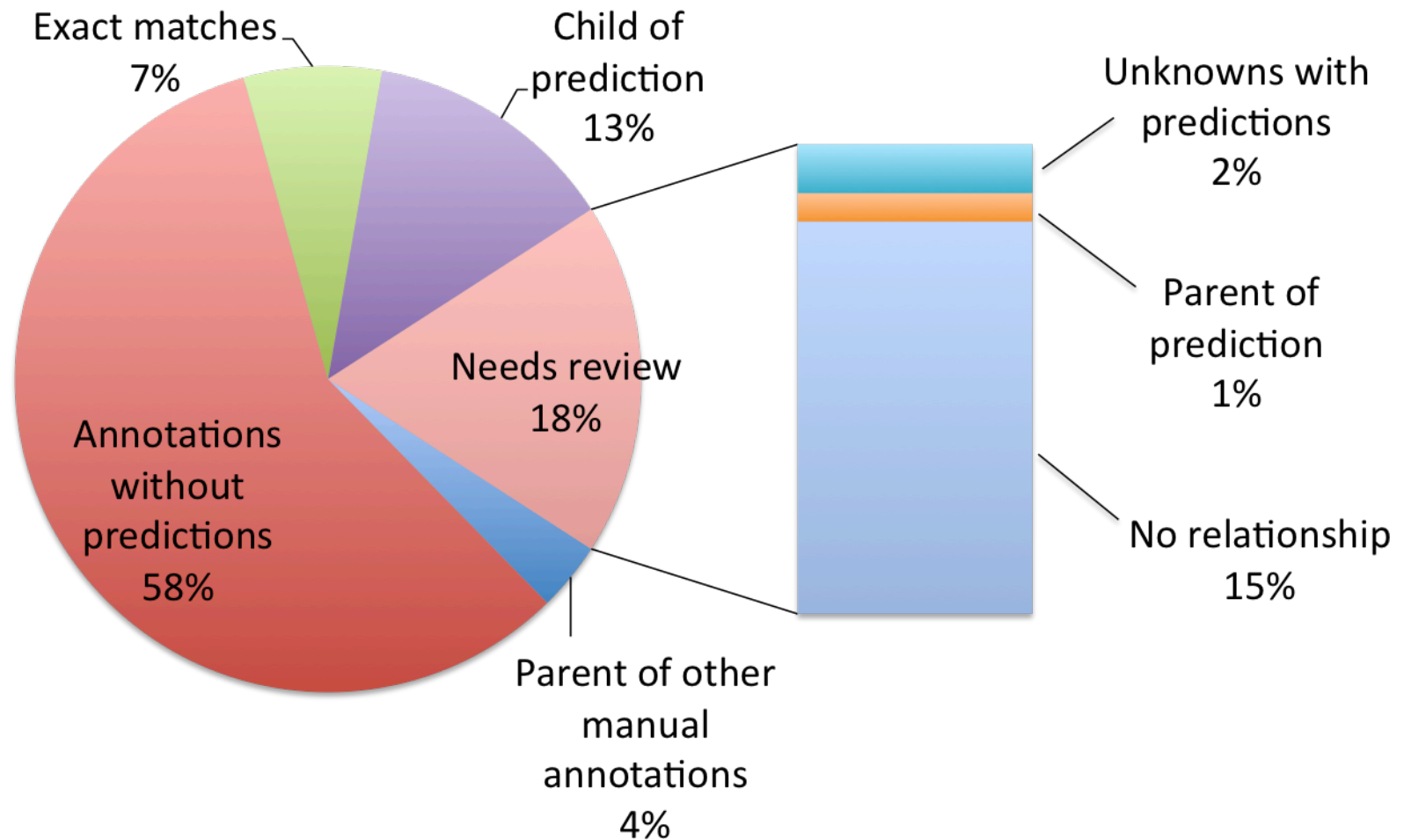


Why compare manual and computational annotations?

1. To improve manual annotation quality, finding:
 - errors
 - omissions
 - “shallow” annotations (i.e., not as granular as possible)
2. To improve computational prediction methods:
 - are certain domains incorrectly mapped to GO terms?
 - are prediction algorithms consistently generating incorrect predictions in any particular area of biology?
3. To improve the Gene Ontology content and structure:
 - do inconsistencies between manual annotations and predictions reveal issues with GO structure, such as incorrect or missing parentage, or true path violations?



All manual annotations compared to InterPro computational predictions



31977 total manual annotations; 5832 flagged as needing review



Manual annotations reviewed

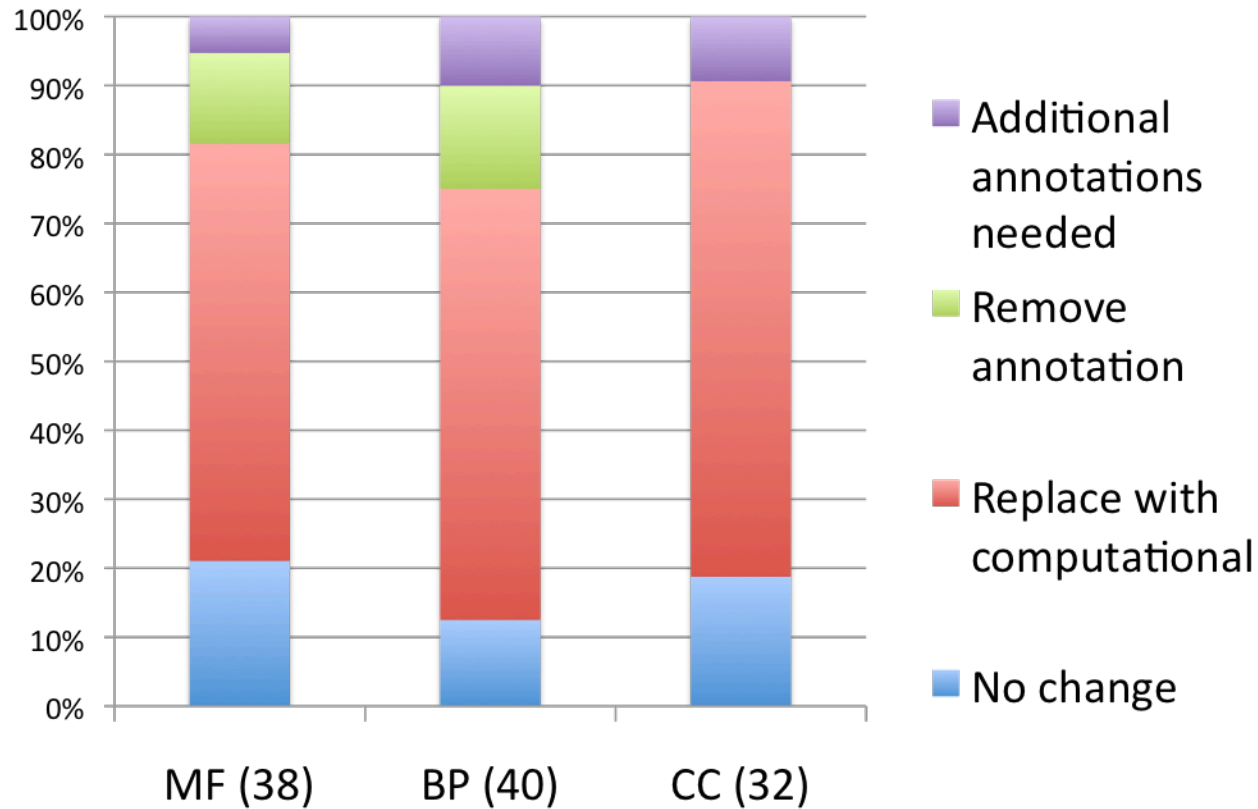
We reviewed three sets of annotations:

- “granular” – the term used for the manual annotation is a parent of (less granular than) the term used for the prediction
- “unknown” – the manual annotation is to a root term, but there is a prediction in that GO aspect
- “discrepancy” - the terms used for manual and computational annotations are not related in the ontology

We compared the manual annotations to the computational predictions and looked at the published literature to evaluate whether there is an experimental basis for the prediction.



Manual annotation is less granular than InterPro prediction



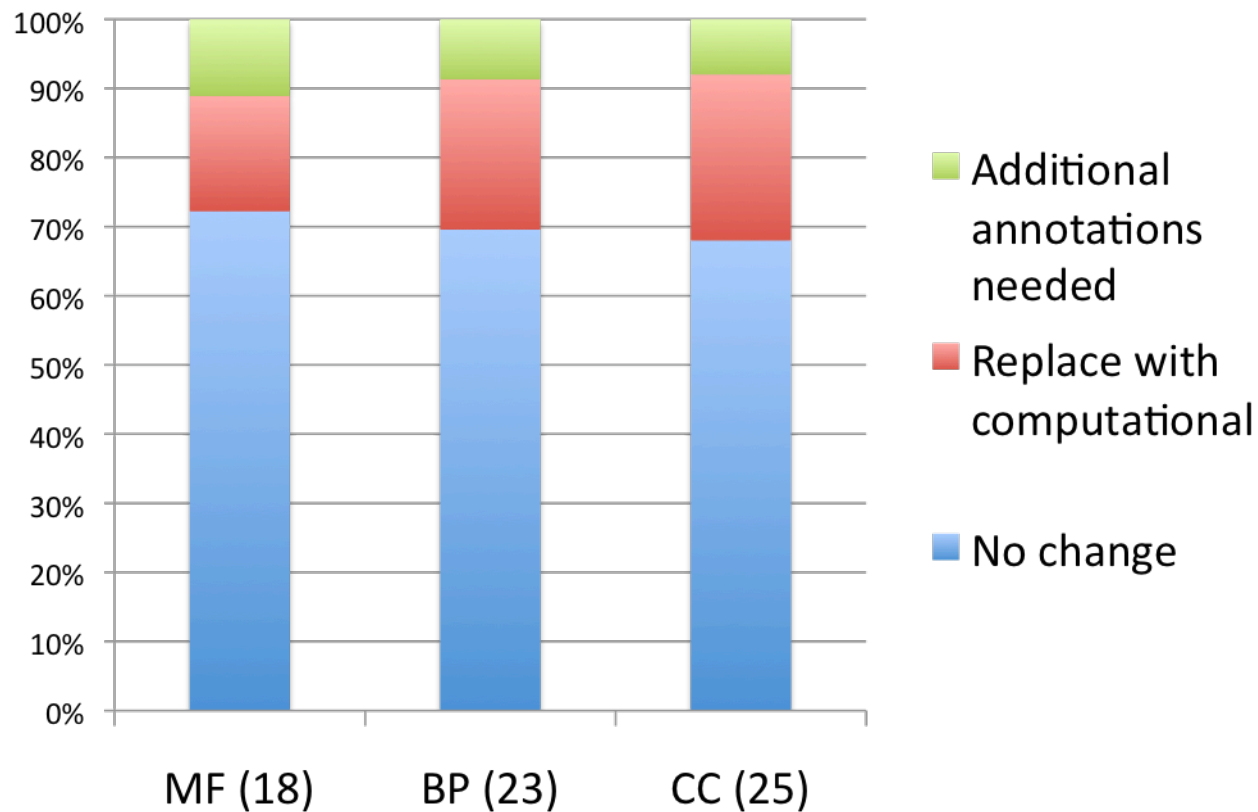
Example:

manual annotation = “metallopeptidase activity”

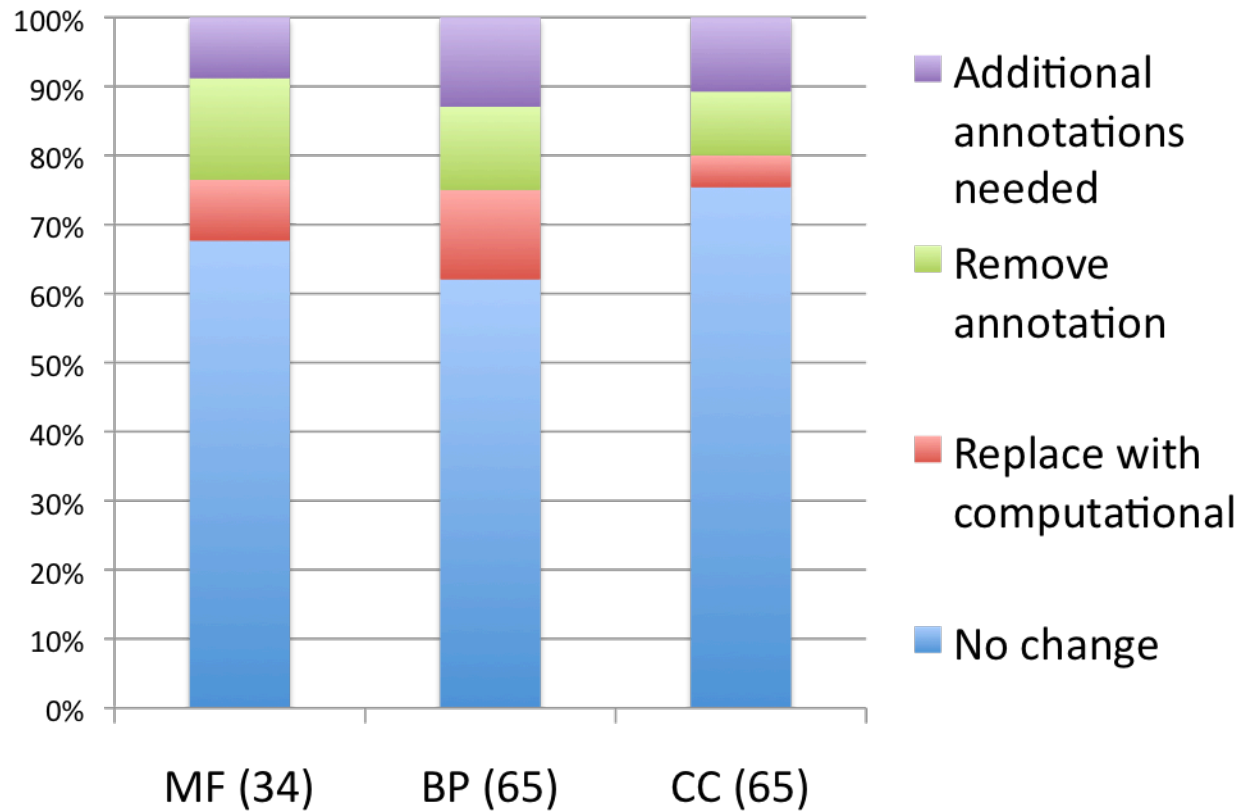
InterPro prediction = “metalloendopeptidase activity”



Manual annotation is “unknown”, but there is an InterPro prediction



Discrepancy (manual and InterPro annotations are unrelated)



What do the discrepancies tell us?

Sometimes we miss details that are revealed by the protein domains

Example: Tpo1p is a polyamine transporter

- manual annotation is to “spermidine transmembrane transporter activity”
- InterPro annotation is to “polyamine:hydrogen antiporter activity”
- reexamination of the literature confirms that it is an antiporter

Sometimes the GO structure needs to be changed

Example: Afg3p is a subunit of the m-AAA protease that is embedded in the mitochondrial inner membrane

- manual annotation is to “m-AAA complex”
- computational annotation is to “integral to membrane”
- flagged as a discrepancy because “m-AAA complex” does not have “integral to membrane” parentage

InterPro to GO mapping may (rarely!) be incorrect

Example: IPR000222 Protein phosphatase 2C, “manganese/magnesium aspartate binding site” is mapped to Cellular Component term GO:0008287, “protein serine/threonine phosphatase complex”
However, PP2Cs are described in the InterPro entry as a monomeric family of protein phosphatases



Is this an efficient way to target manual annotations for review?

The “granular” set involved 72 genes with a total of 1200 publications

87 annotations were reviewed

55 manual annotations were replaced with a more granular computational annotation

= average of 21 papers per annotation change

In other sets, even fewer manual annotations were replaced with the computational term = even more papers per annotation change



Conclusions

- This type of analysis can result in improvements to manual annotations, computational methods, and the GO ontology
- we hope to use this method to target and prioritize manual annotations that need review
- Still to do: comparison of manual annotations to computational predictions other than InterPro



Acknowledgments



Mike Cherry



Eurie Hong



Rama Balakrishnan



Karen Christie



Maria Costanzo



Selina Dwight



Stacia Engel



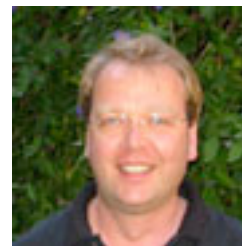
Dianna Fisk



Jodi Hirschman



Cindy Krieger



Rob Nash



Julie Park



Marek Skrzypek

SGD is funded by the U.S. National Human Genome Research Institute

