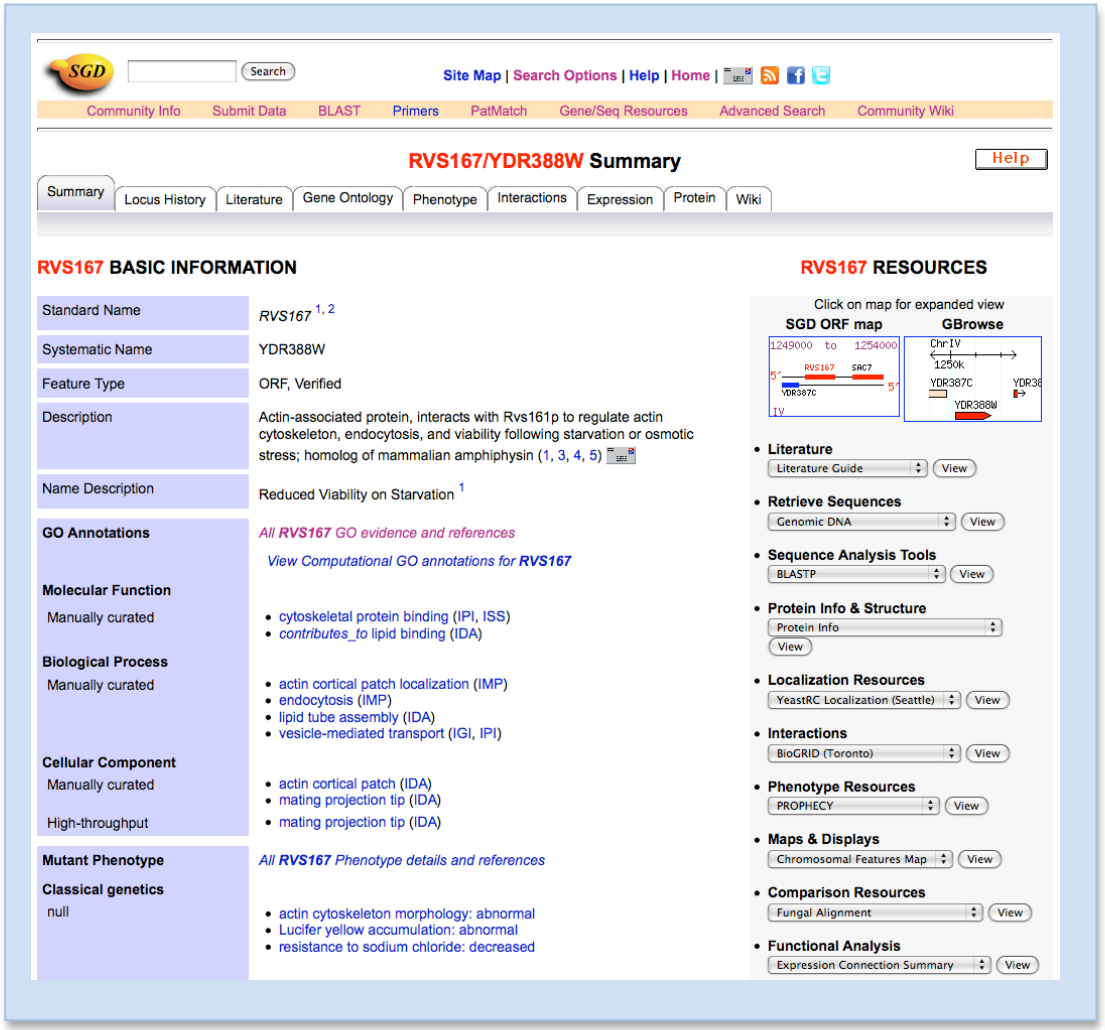
Comparison of computationally- and manually-assigned Gene Ontology annotations to improve functional characterization of gene products

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

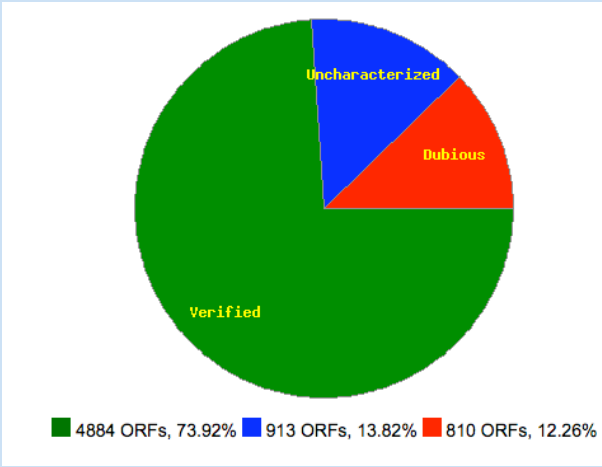
Biocuration 2010

October 12, 2010

The *Saccharomyces* Genome Database (SGD)
www.yeastgenome.org



Characterization of Open Reading Frames



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

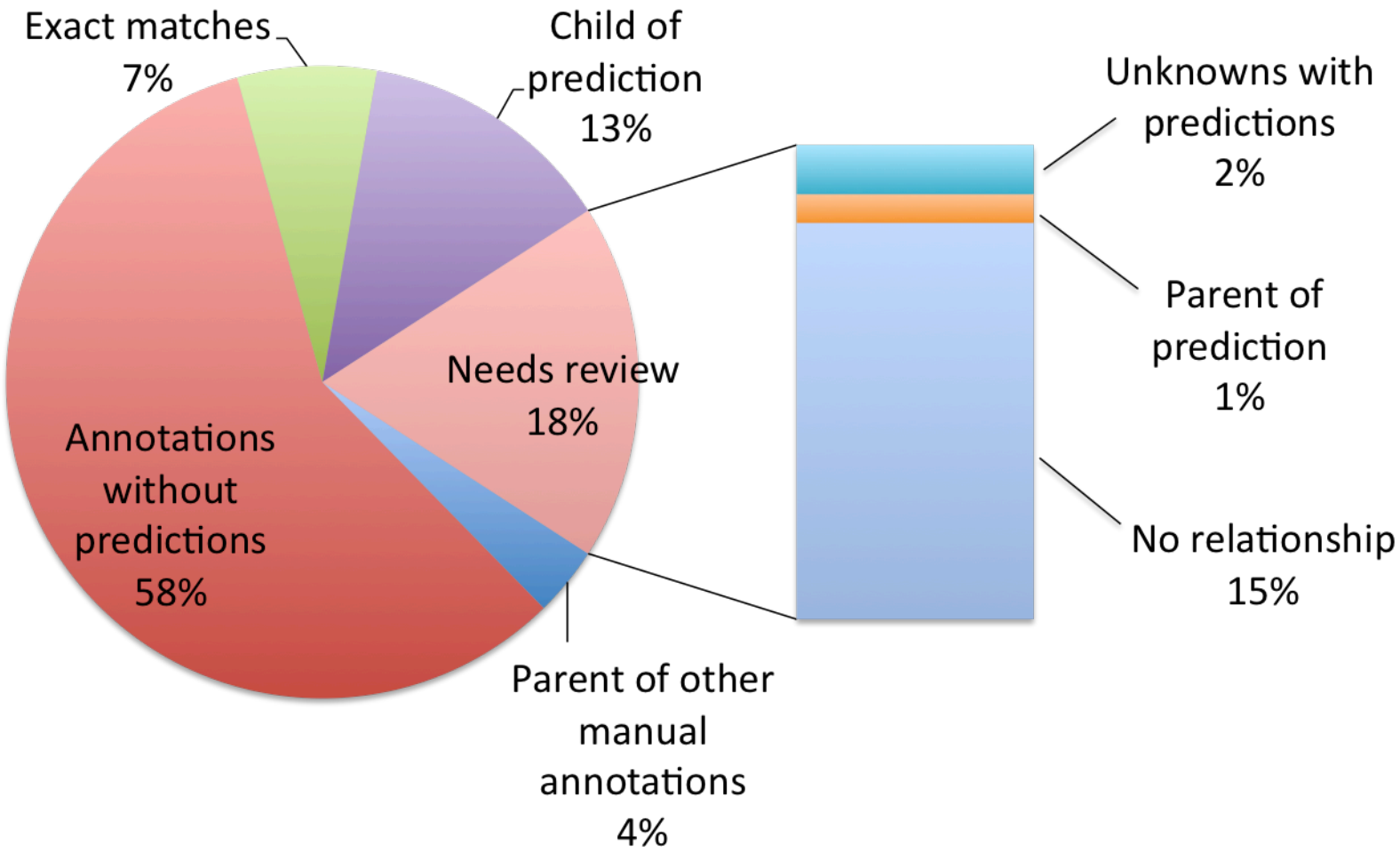
Why compare manual and computational annotations?

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

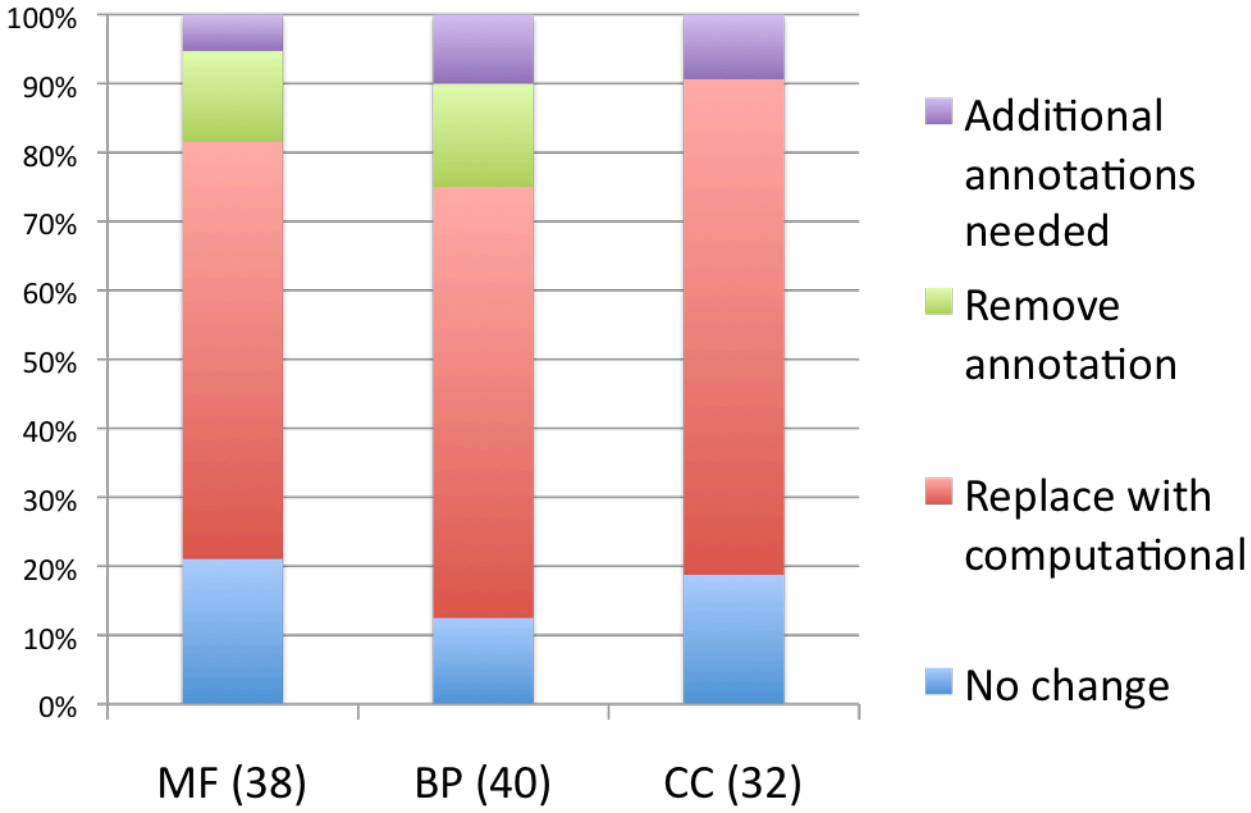
Manual annotations reviewed

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

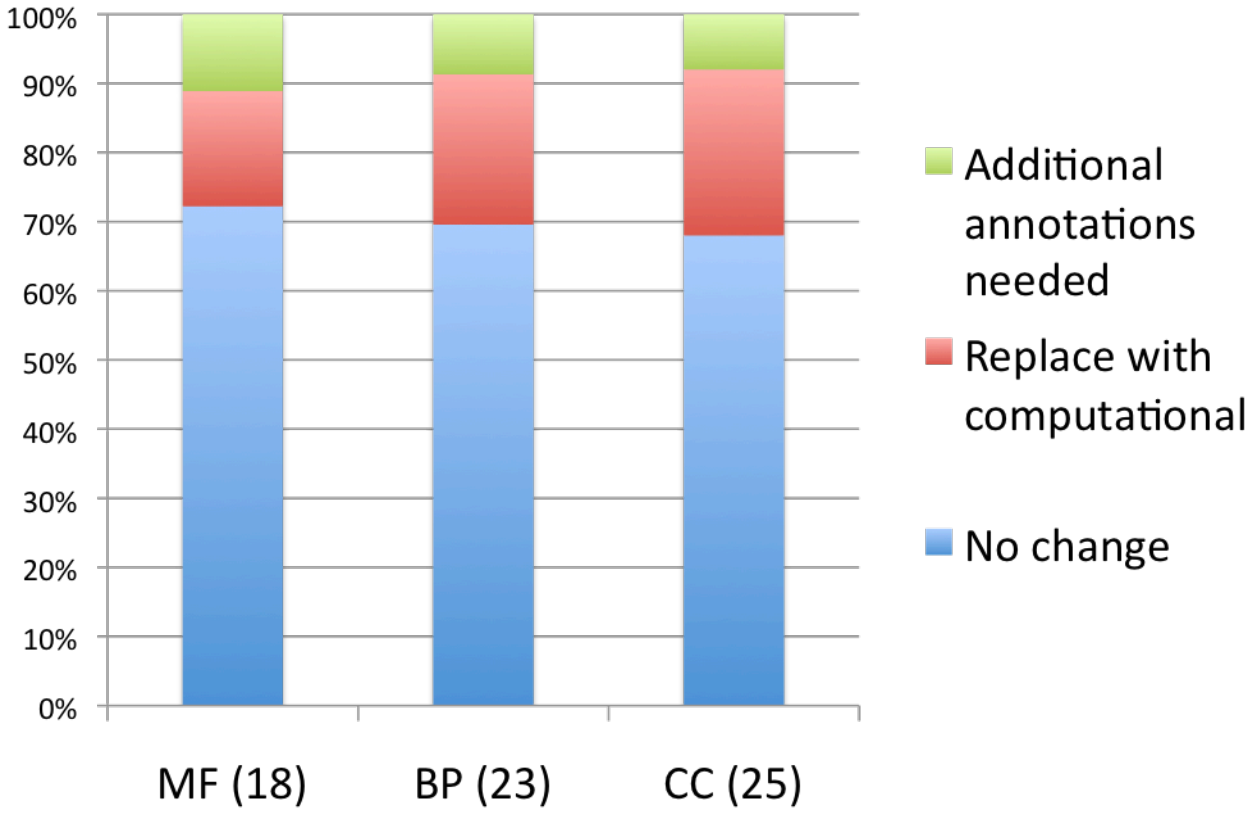
We compared the manual annotations to the computational predictions and looked at the published literature to evaluate whether there is an experimental basis for the prediction.

Manual annotation is less granular than InterPro prediction

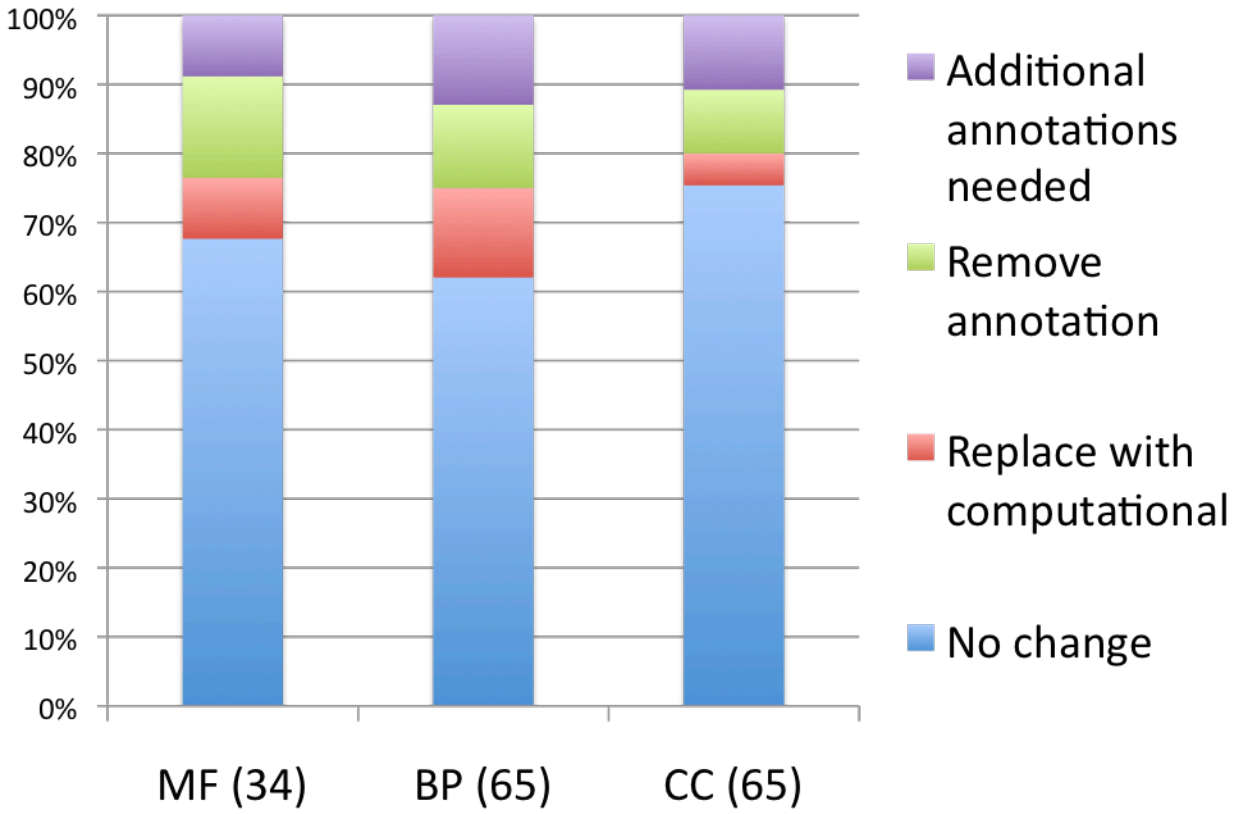


Example:
manual annotation = “metallopeptidase activity”
InterPro prediction = “metalloendopeptidase activity”

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