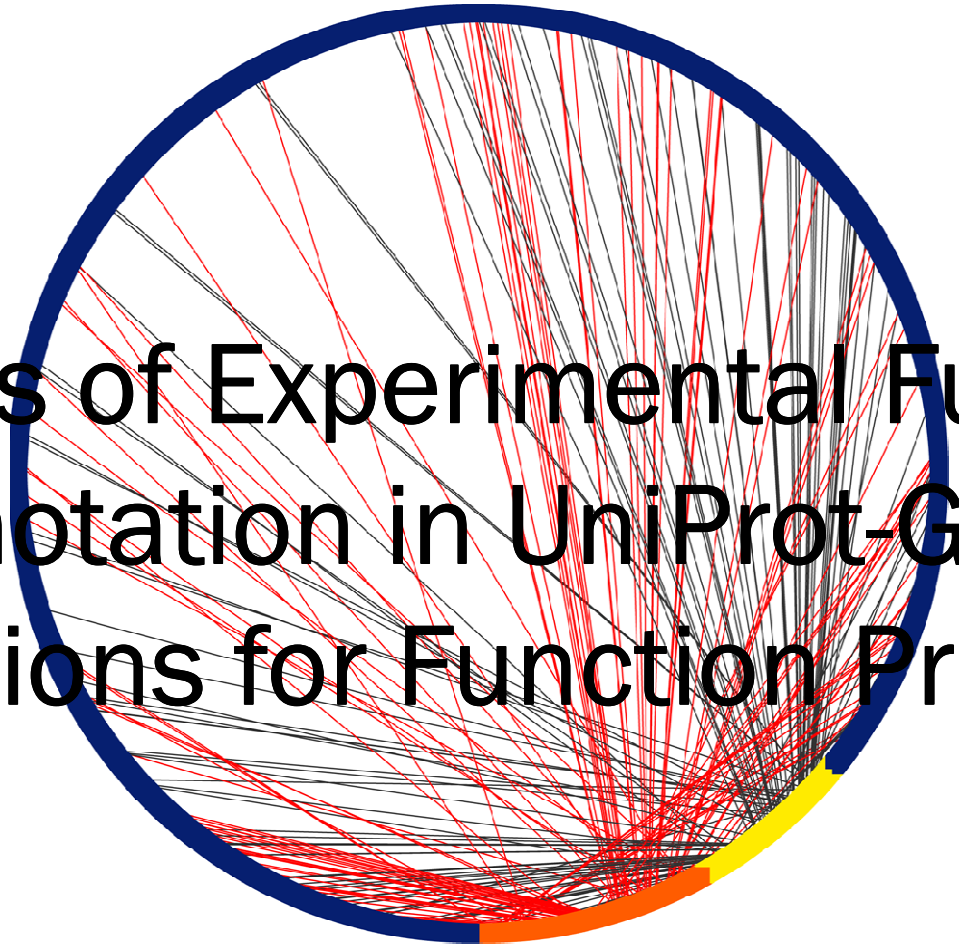


Sources of Experimental Function Annotation in UniProt-GOA, Implications for Function Prediction

Alexandra M Schnoes, Alexander W Thorman, Iddo Friedberg

UCSF
University of California
San Francisco

MIAMI UNIVERSITY
OXFORD OHIO

The Importance of Experimentally Characterized Proteins

- Furthers experimental progress in basic, biomedical and clinical science

- Computation:
  - Gold standard data → test function prediction algorithms
  - Use in function prediction

The Composition of Experimental Data Impacts Function Prediction

- Bias
  - Increases error
  - Decreases utility

- One potential source of data composition bias:
  - Overrepresentation of certain experimental methods → high-throughput protein function characterization
    - Source bias: Few papers, many proteins?
    - Predominance of certain types of annotations?

Examine the UniProt-Gene Ontology Annotation (GOA) Database

- Purpose:
  - Determine whether high-annotating papers are biasing the composition of GOA annotations
  - If a bias exists, to what level?

Examine the UniProt-Gene Ontology Annotation (GOA) Database

Experimental Evidence Codes in GOA	Count
All	522,208
IEP (Inferred from Expression Pattern)	13,972
IPI (Inferred from Physical Interaction)	63,832
IMP (Inferred from Mutant Phenotype)	184,084
EXP (Inferred from Experiment)	31
IGI (Inferred from Genetic Interaction)	26,522
IDA (Inferred from Direct Assay)	233,767

Examine the UniProt-Gene Ontology Annotation (GOA) Database

Experimental Evidence Codes in GOA	Count
All	522,208
IEP (Inferred from Expression Pattern)	13,972
IPI (Inferred from Physical Interaction)	63,832
IMP (Inferred from Mutant Phenotype)	184,084
EXP (Inferred from Experiment)	31
IGI (Inferred from Genetic Interaction)	26,522
IDA (Inferred from Direct Assay)	233,767

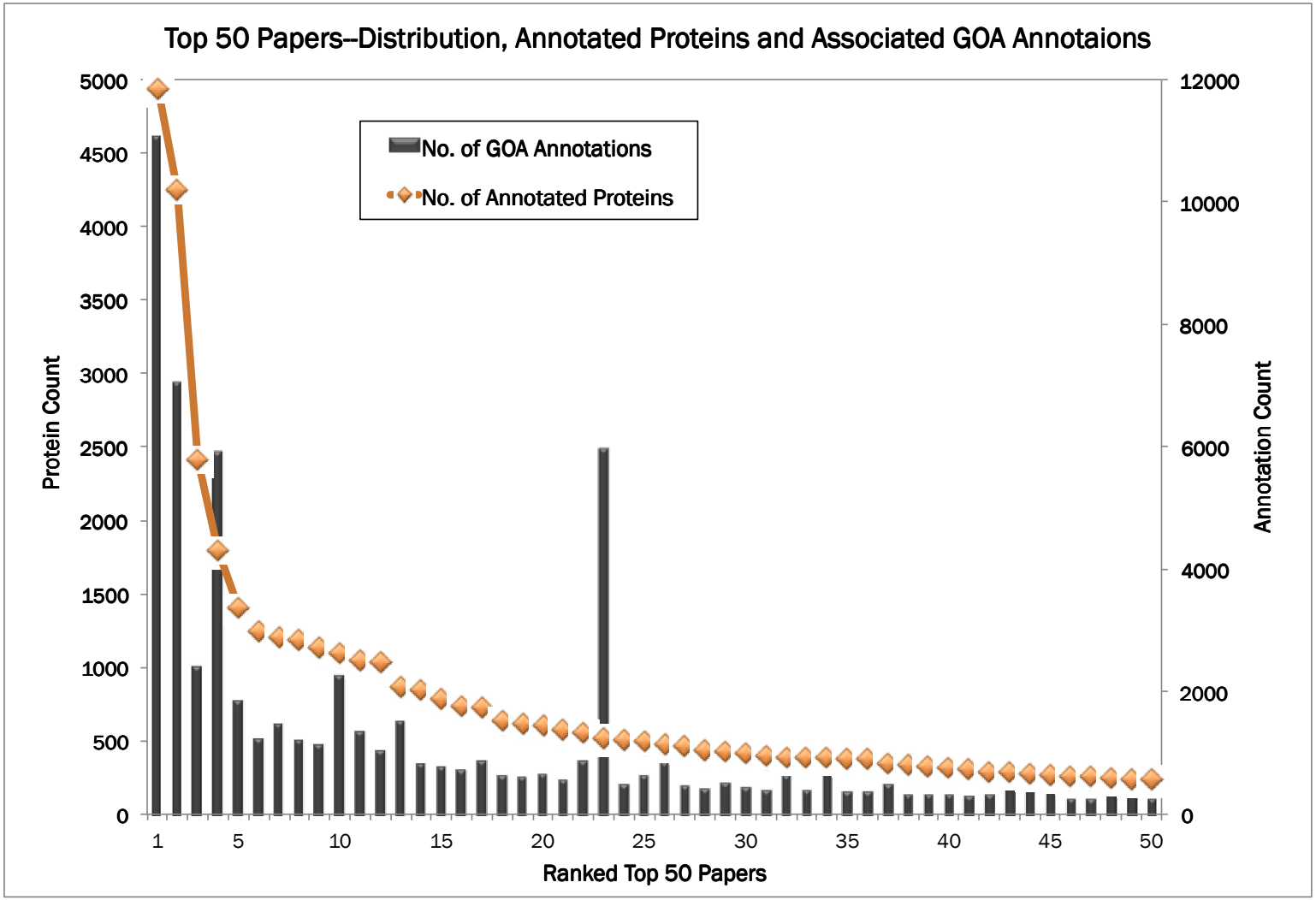
GO Ontologies	Count
Molecular Function (MFO)	115,440
Cellular Component (CCO)	131,804
Biological Process (BPO)	274,964

December 2011, http://www.ebi.ac.uk/GOA/

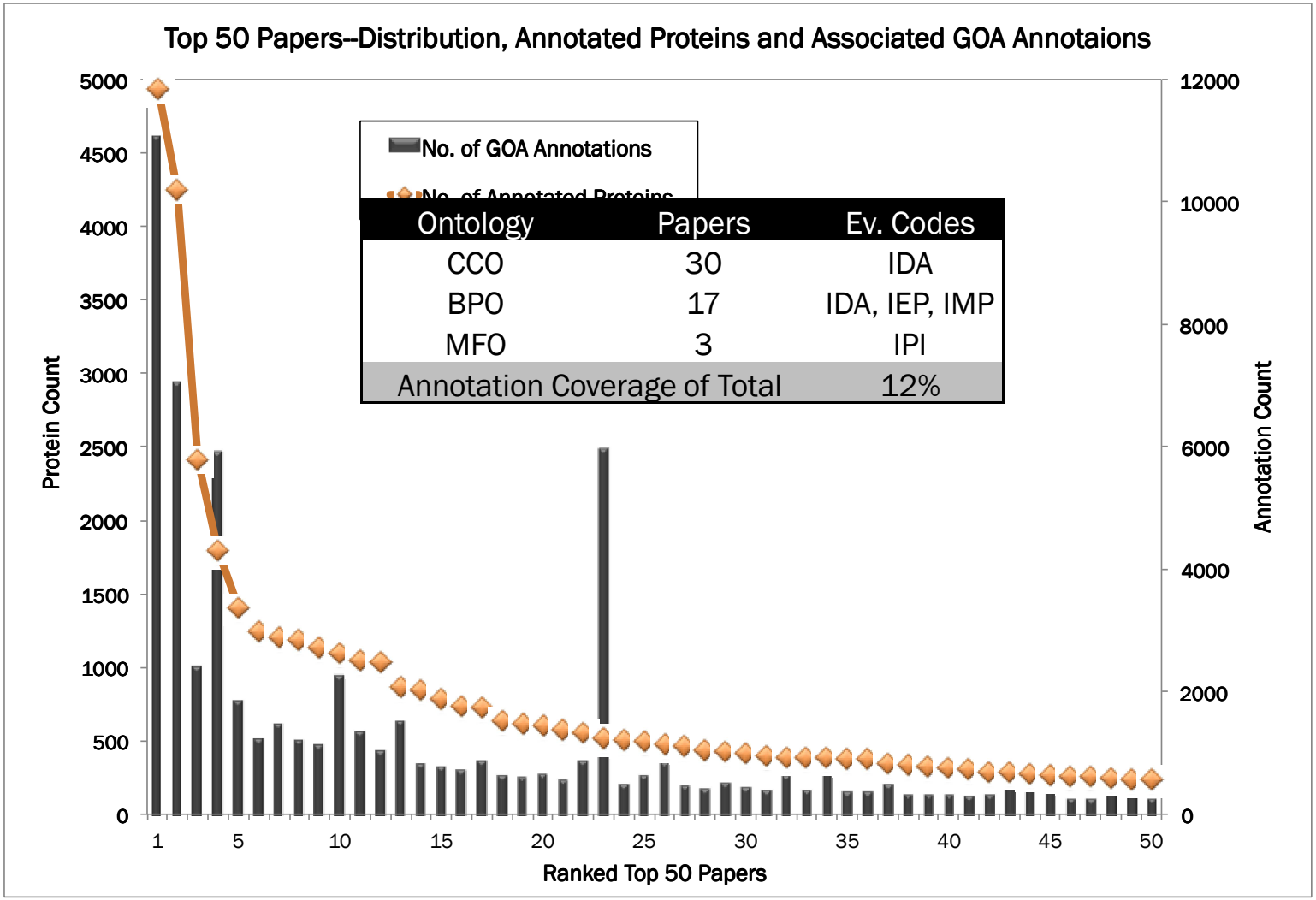
Top 50 Papers—
Annotate the Most Proteins in GOA

	No. of Annotated Proteins	No. of GOA Annotations	Species	Majority Ontology	Majority Evidence Code
1	4937	11050	*Human*	CCO	IDA
2	4247	7046	*Schizosaccharomyces pombe*	CCO	IDA
3	2412	2412	*Mouse*	CCO	IDA
4	1791	5918	*Caenorhabditis elegans*	BPO	IMP
5	1406	1863	*Saccharomyces cerevisiae S288c*	CCO	IDA
6	1251	1251	*Arabidopsis thaliana*	CCO	IDA
7	1205	1476	*Caenorhabditis elegans*	BPO	IMP
⋮	⋮	⋮	⋮	⋮	⋮
50	241	263	*Arabidopsis thaliana*	CCO	IDA

Top 50 Papers—
Annotate the Most Proteins in GOA

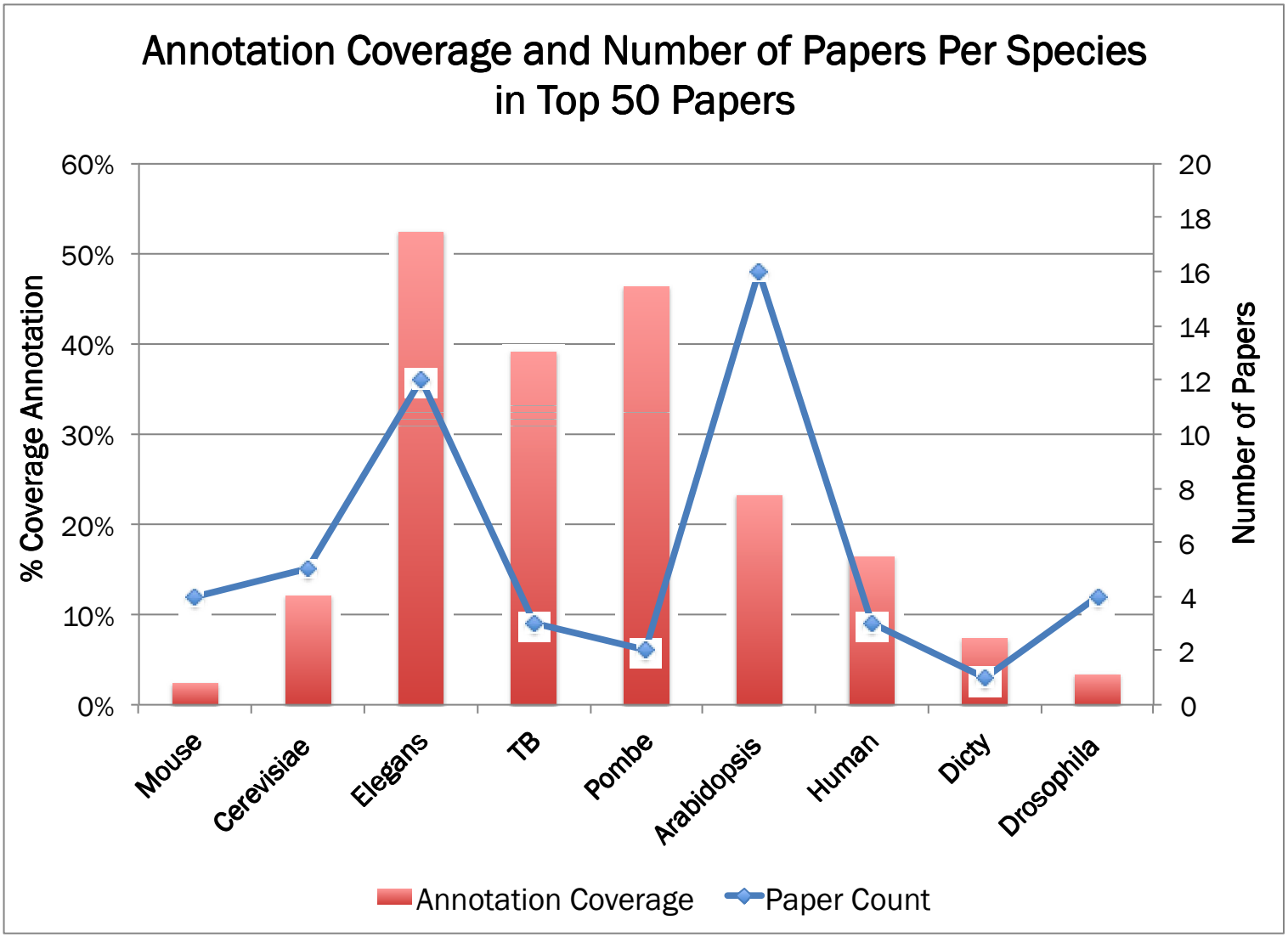


Top 50 Papers—
Annotate the Most Proteins in GOA

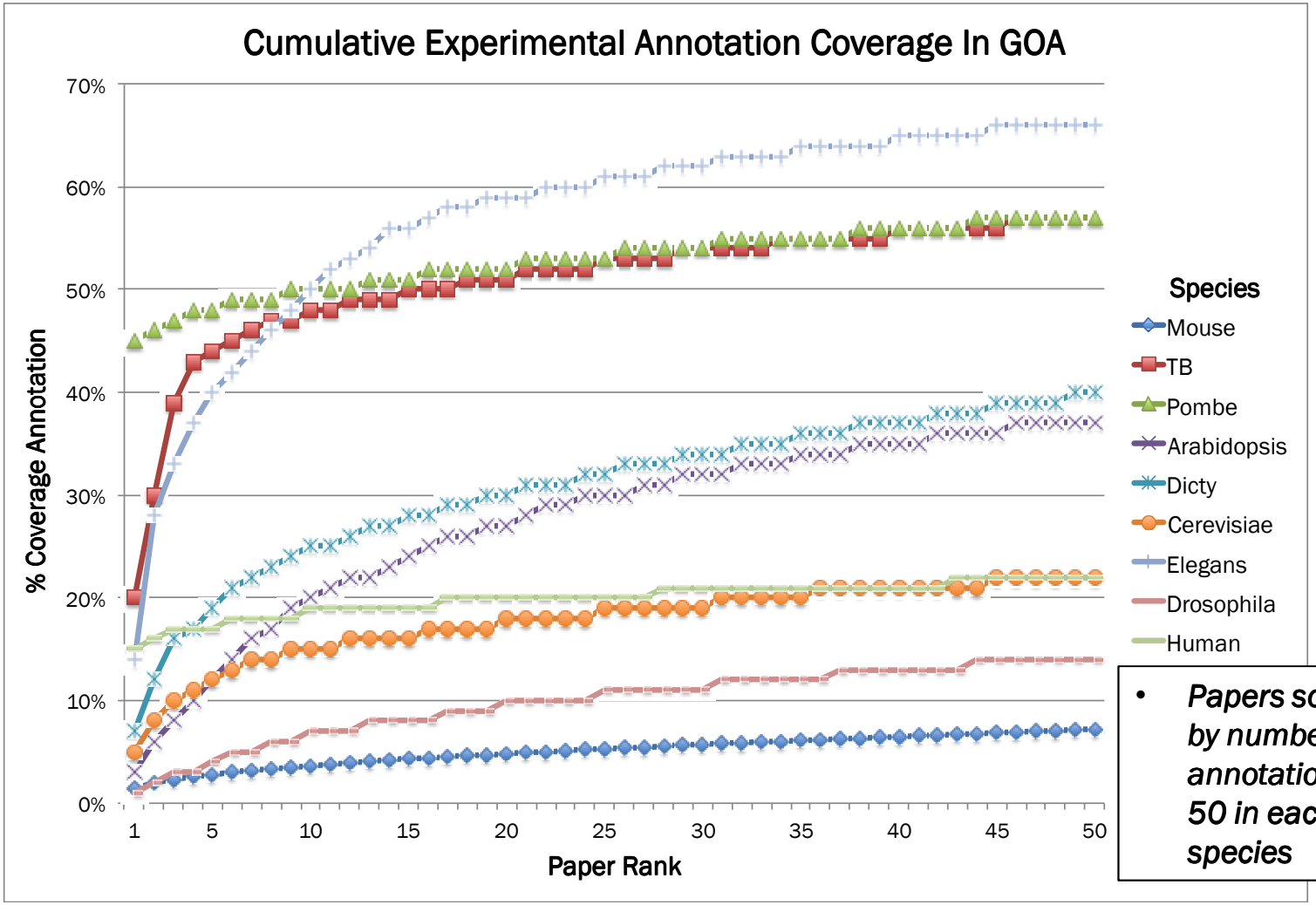


Ontology	Papers	Ev. Codes
CCO	30	IDA
BPO	17	IDA, IEP, IMP
MFO	3	IPI
Annotation Coverage of Total		12%

6/9 Species > 10% Annotation Coverage

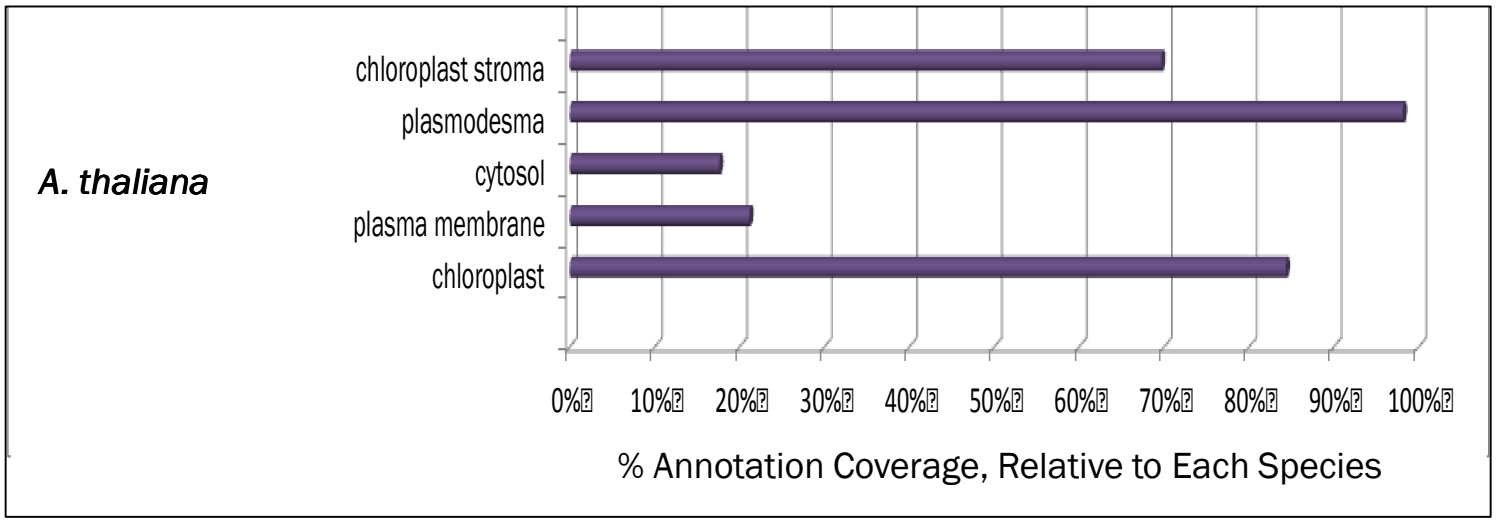


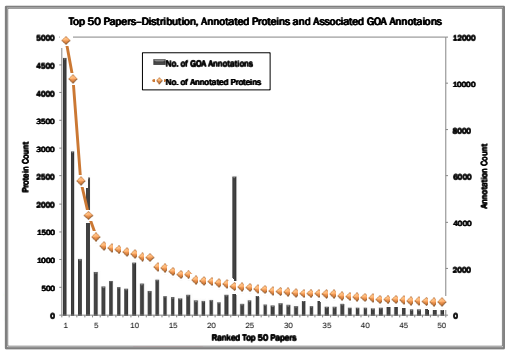
Single Papers Can Greatly Increase Coverage

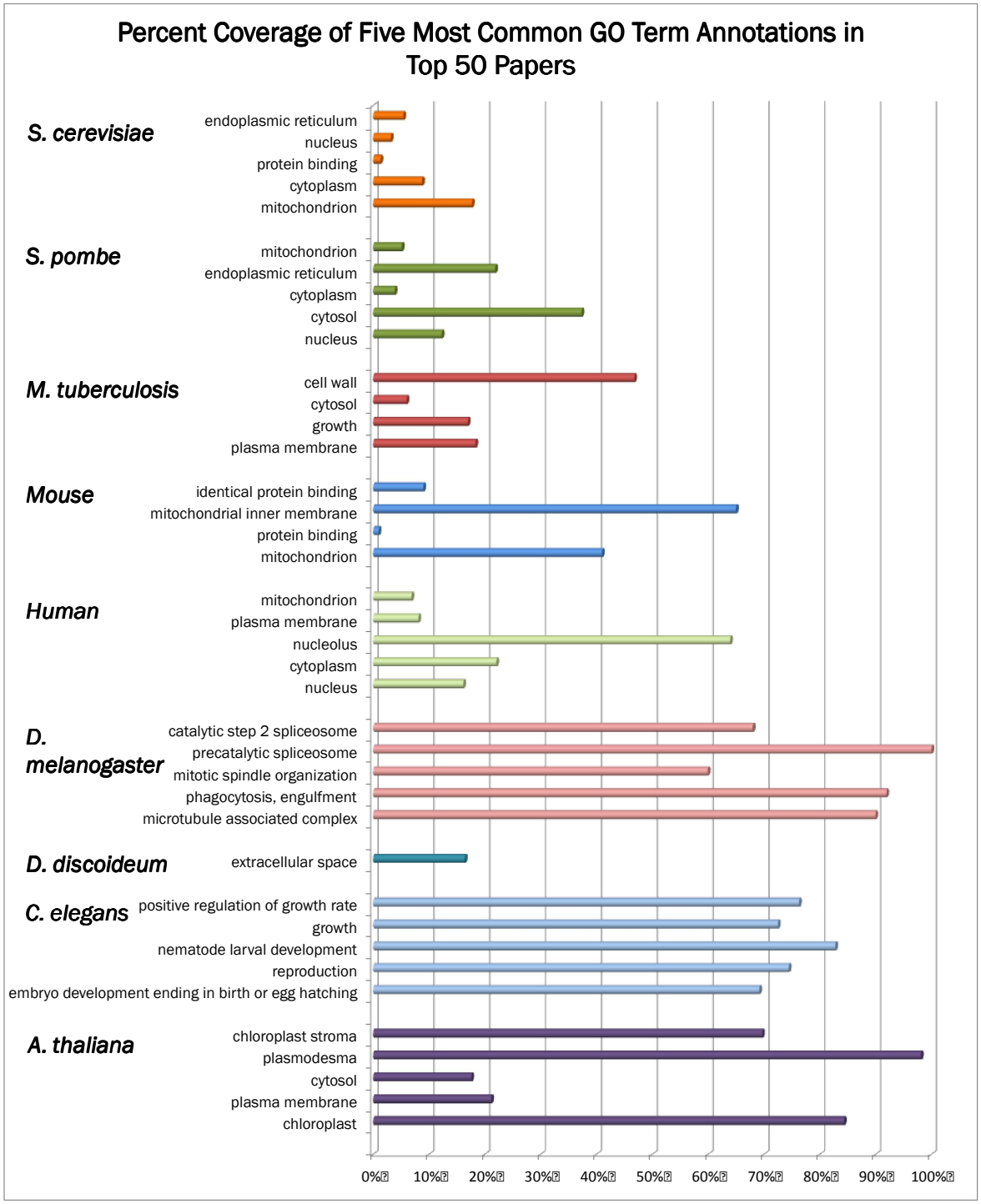


- *Papers sorted by number of annotations, top 50 in each species*

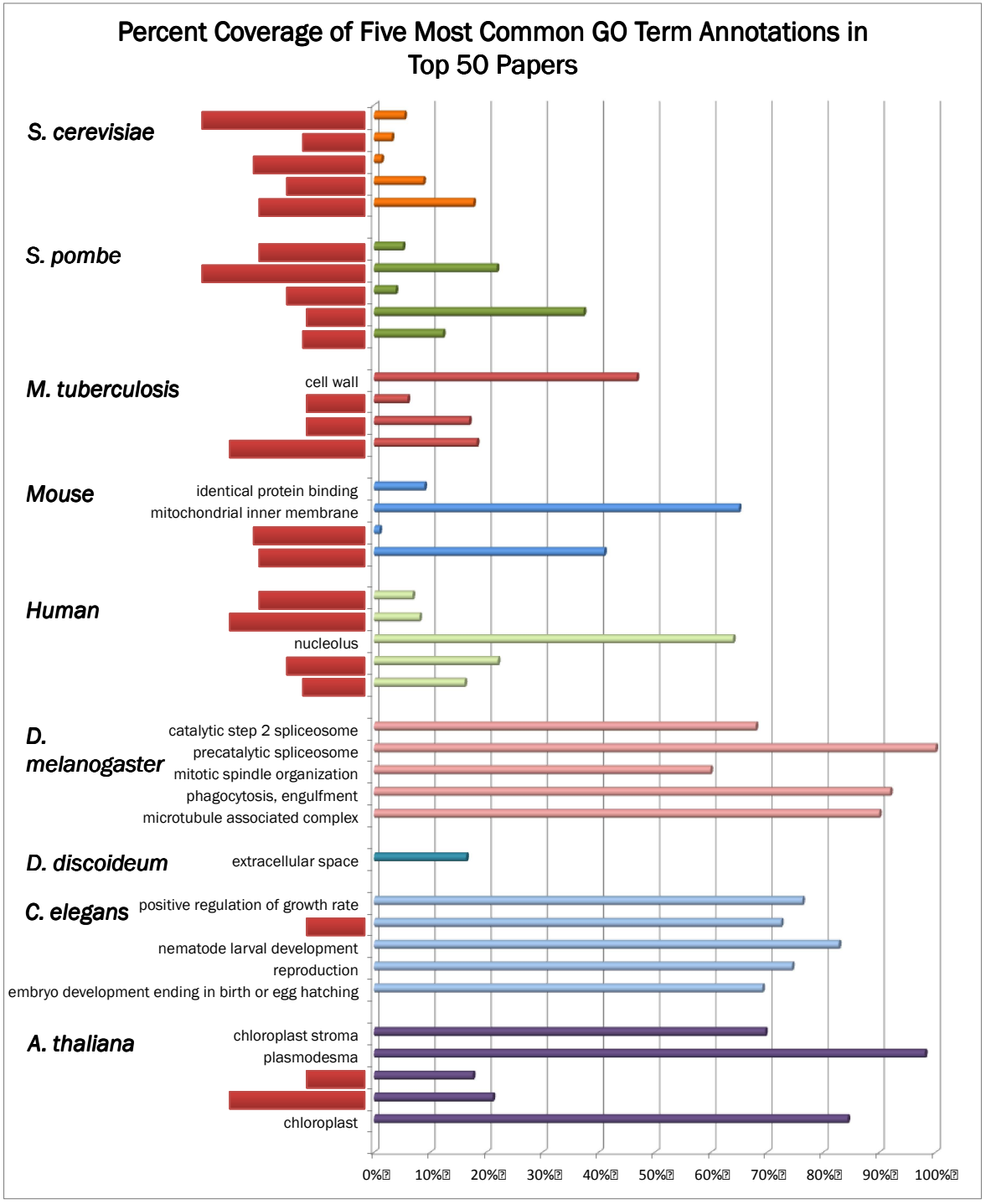
What are the coverage propensities for common terms in the global top 50 papers?







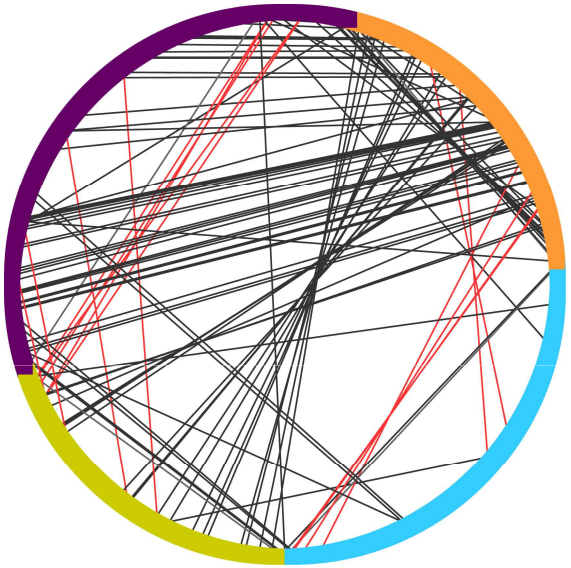
- Certain terms predominantly occur in high-annotating papers



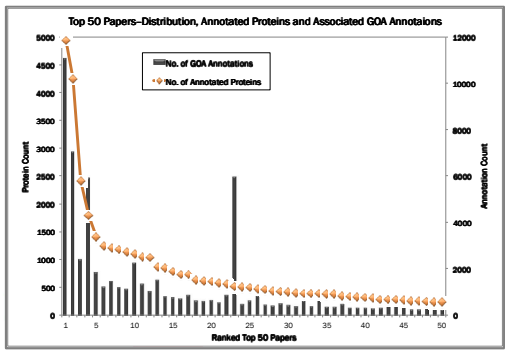
- Certain terms predominantly occur in high-annotating papers
- Terms are repeated over the different species (terms used twice or more are highlighted)

How often do we annotate the same proteins?

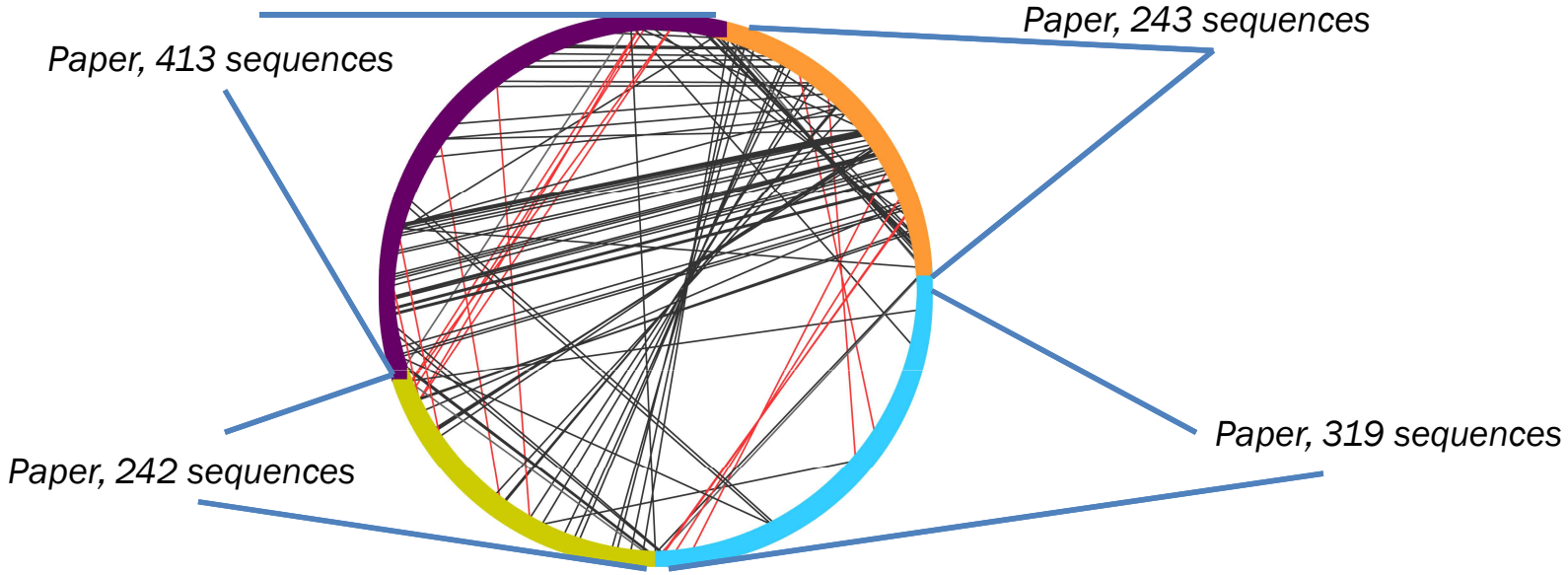
Top 50 Papers—
Sequence Identity Between Papers



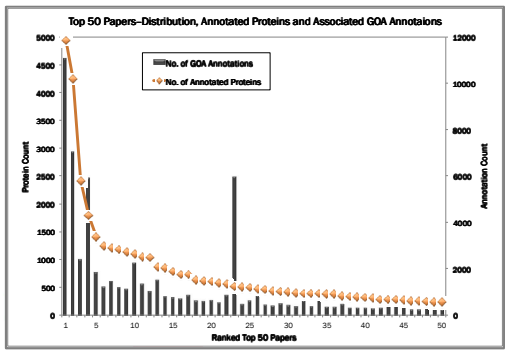
D. melanogaster (1217 sequences)



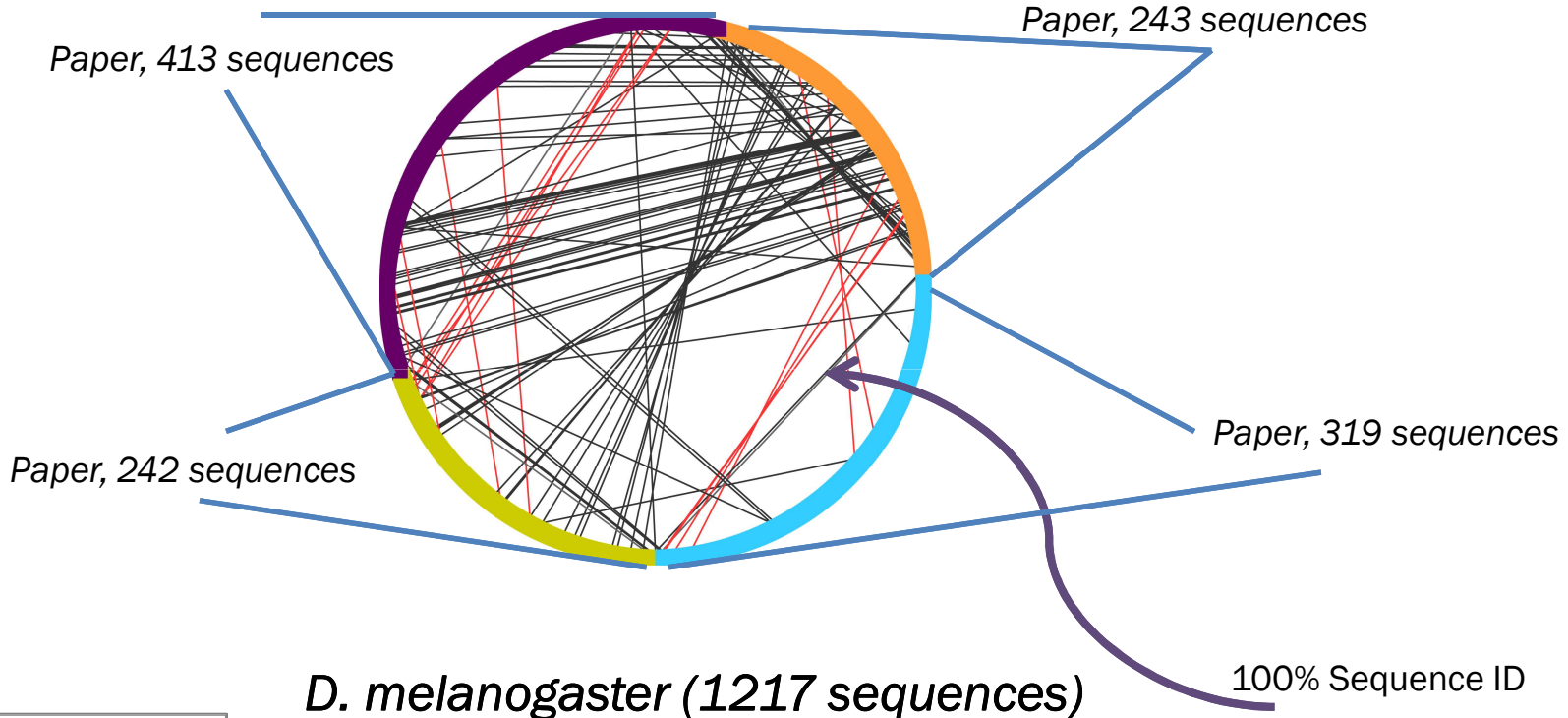
Top 50 Papers—
Sequence Identity Between Papers

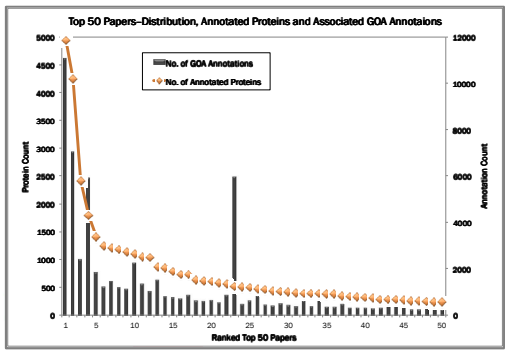


D. melanogaster (1217 sequences)



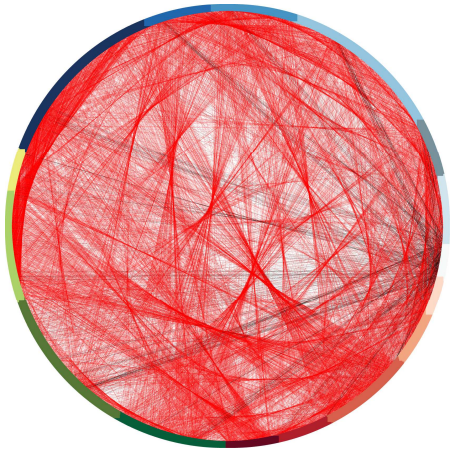
Top 50 Papers—
Sequence Identity Between Papers



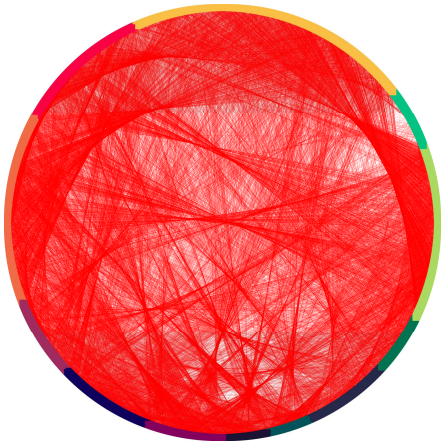


Different Ontology Same Ontology

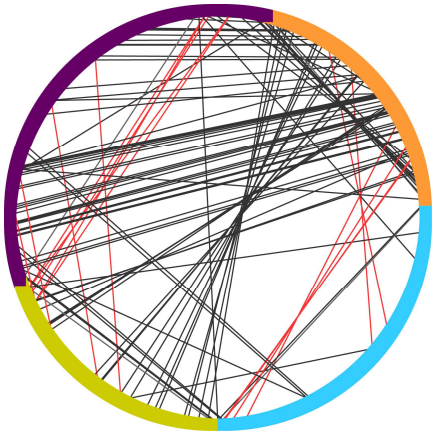
High Redundancy, Often in Same Ontology



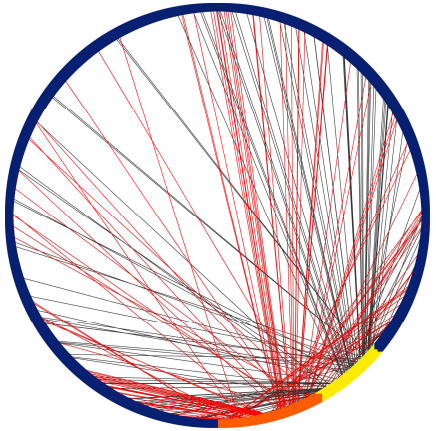
A. thaliana (8879)



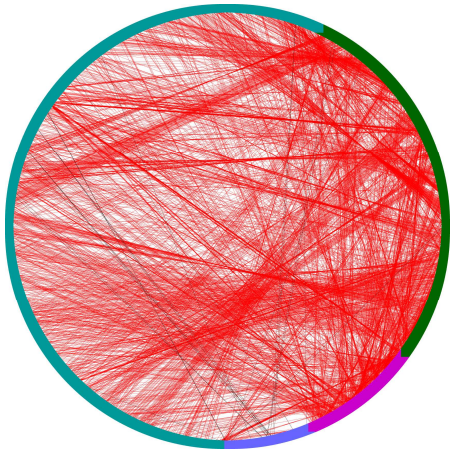
C. elegans (8416)



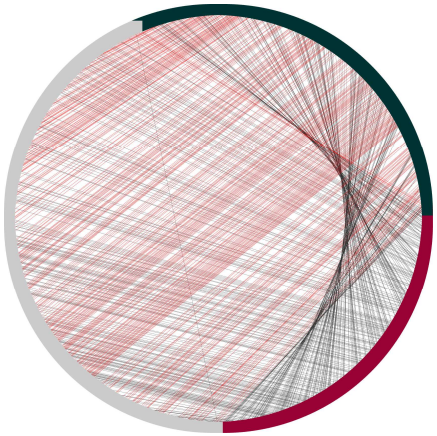
D. melanogaster (1217)



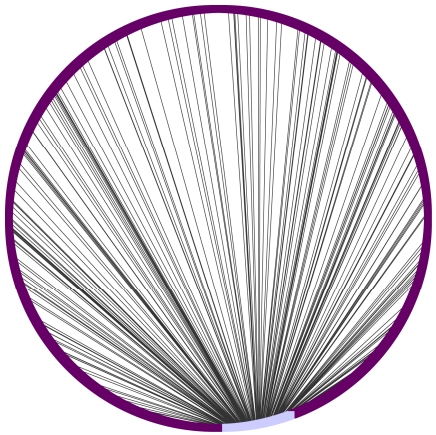
Human (5593)



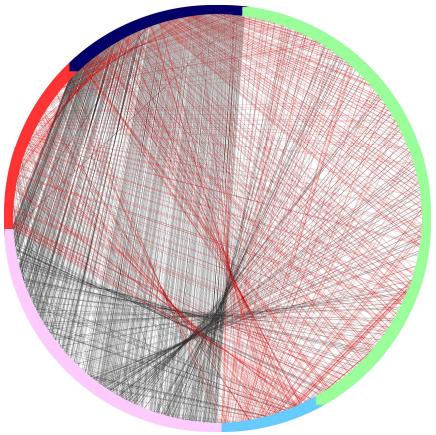
Mouse (4220)



M. tuberculosis (2351)



S. pombe (4502)



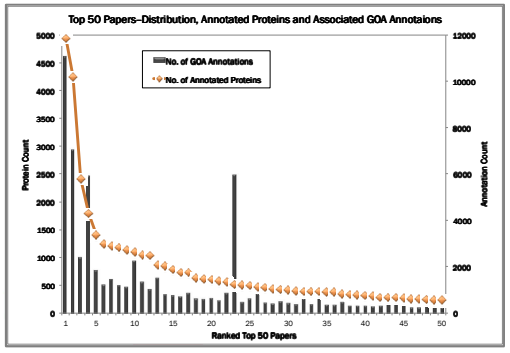
S. cerevisiae (3542)

Different Ontology

Same Ontology

Redundancy from 95-100% Sequence Identity

Species	Total Sequences	No. Clusters 100%	% Redundant	No. Clusters 95%	% Redundant
A. thaliana	8879	4694	47%	4595	48%
C. elegans	8416	3338	60%	3292	61%
D. melanogaster	1217	1003	18%	767	37%
Human	5593	4509	19%	3748	33%
M. tuberculosis	2351	1702	28%	1700	28%
Mouse	4220	2273	46%	1813	57%
S. cerevisiae	3542	2550	28%	2546	28%
S. pombe	4502	4281	5%	4253	6%



Conclusions

- In certain species, annotation composition bias
  - Single papers have large impact on annotation coverage
  - Repeated use of same GO terms
  - High-throughput papers typically annotate the same set of sequences

Implications for Function Prediction

- Not all experimental annotations in GOA are 'equal' → unexpected subtleties to the data
- Care needed for gold standard set creation
  - 'De-bias' for annotation source, annotation ontology depth, species
- Consideration of bias in relation to prediction activities
- Evaluation of where to focus database curation activities

Under Construction

- Term depth analysis for high-annotating papers vs low annotating papers
  - Term enrichment
- For identical sequences in different papers
  - Are the annotations the same?
  - Redundant data? Conflicting data? Higher confidence data?

Acknowledgements

Patsy Babbitt, UCSF

Babbitt Lab

Resource for Biocomputing, Visualization and Informatics (RBVI) at UCSF

Doug Stryke

Iddo Friedberg, MU, OH

Alexander Thorman

CAFA 2011

NSF & NIH for Funding