BI Journal Club 04.11.2013

Quality of computationally inferred gene ontology annotations

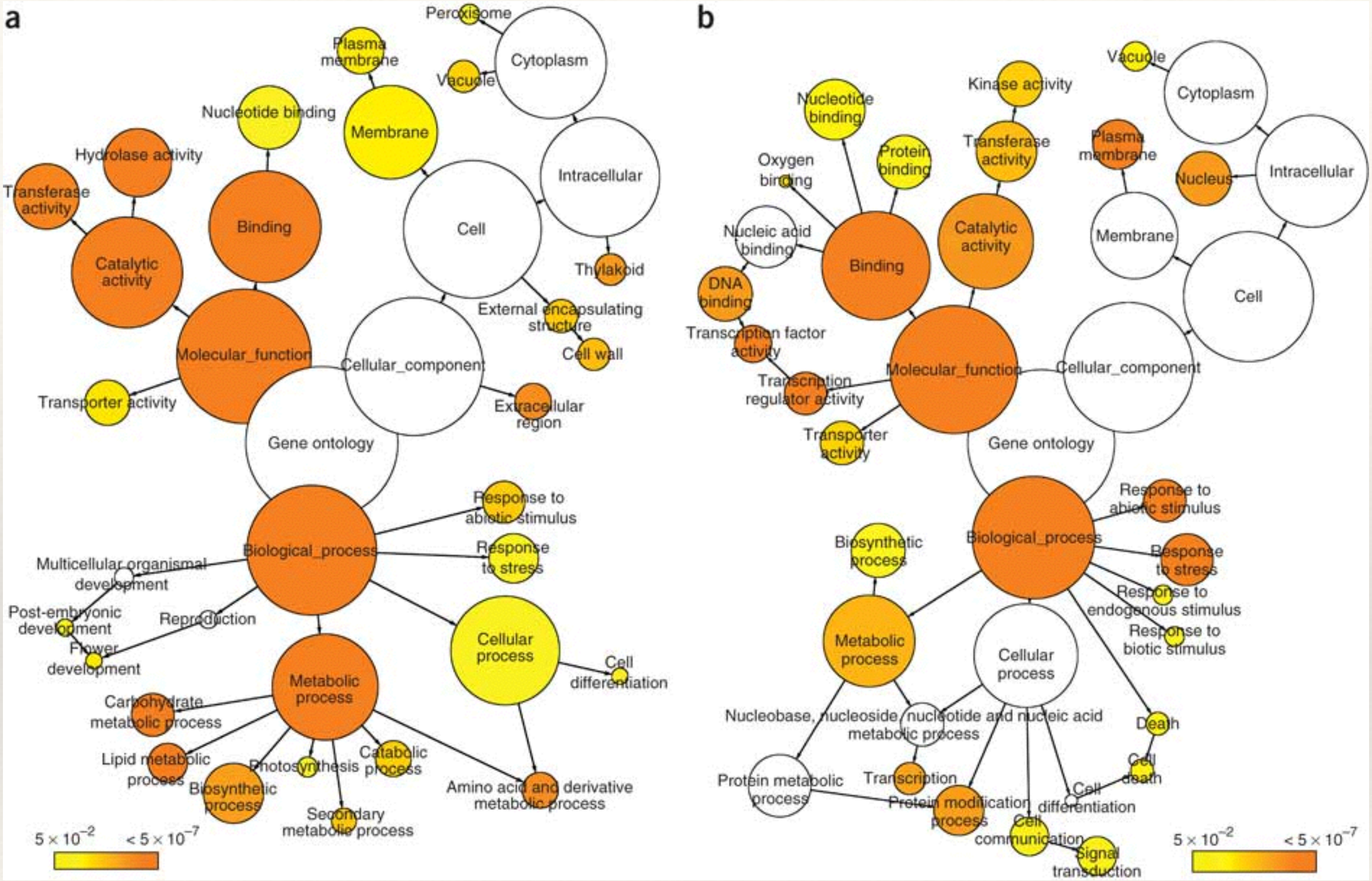
Škunca N, Altenhoff A, Dessimoz C

PLoS Computational Biology, May 2012

Gene ontology

Controlled vocabulary to describe gene products in terms of associated:

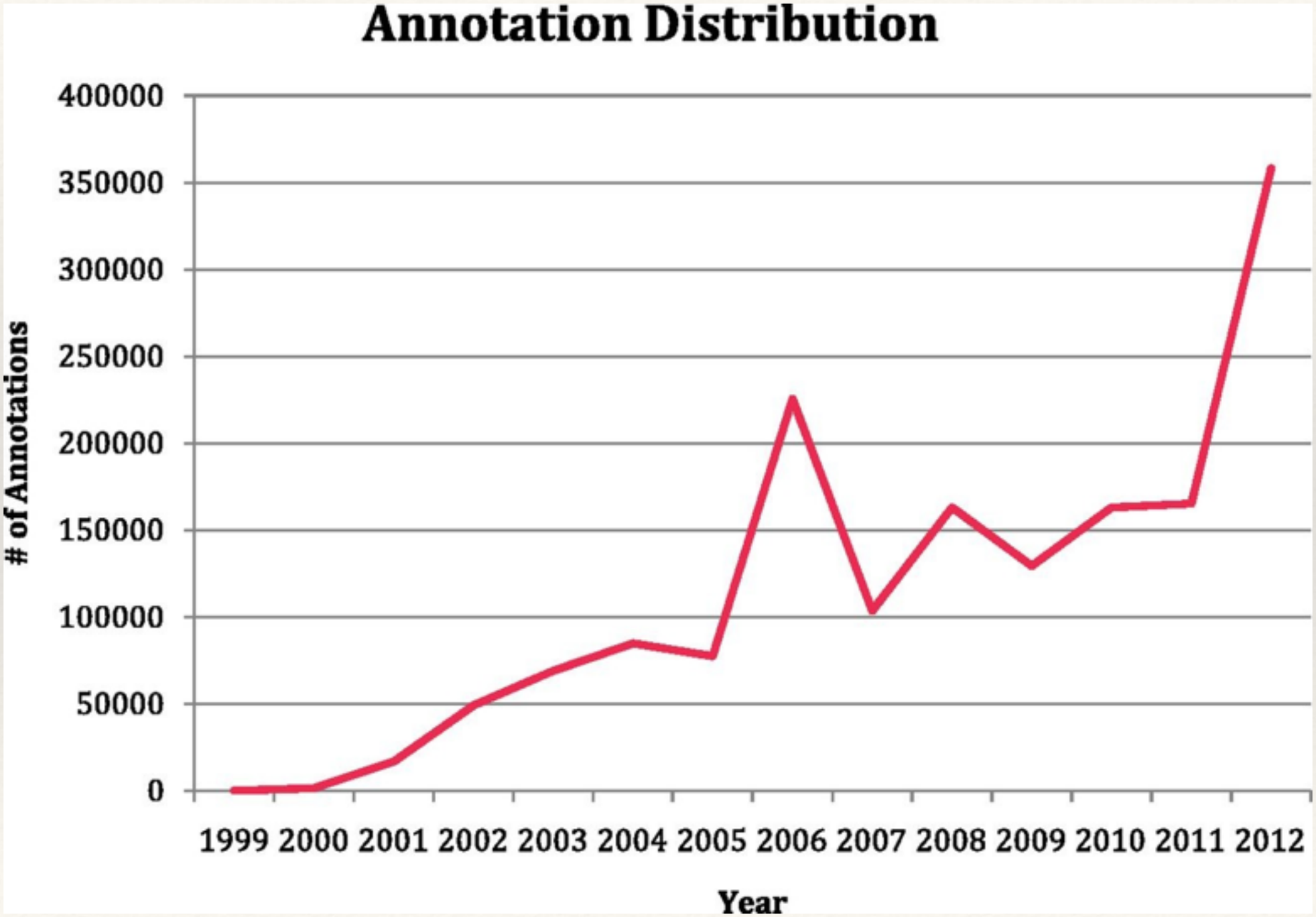
- biological processes
- cellular components
- molecular functions



Gene ontology mapping and functional annotation of strawberry genes.

Nature Genetics **43**, 109–116 (2011) doi:10.1038/ng.740

Increase in the number of manual GO annotations since 1999



Nucleic Acids Res. 2013 January; 41(D1): D530–D535.

Status of GO as of Sept. 2012

Biological process terms	23 907
Molecular function terms	9459
Cellular component terms	3050
Species with annotation (includes strains)	347 778
Total annotated gene products	96 602 850
Manually annotated gene products	358 319

Nucleic Acids Res. 2013 January; 41(D1): D530–D535.

UniProt GOA

- **Manual annotation** by curators using published literature. Each is given an evidence code that describes what evidence supports the annotation
- **Electronic annotation** - use existing information within database entries which are manually mapped. Another mehthod uses orthology data from Ensembl Compara to project GO annotations from a source species onto one or more target species. Evidence code IEA

Experimental annotations

Experimental evidence codes (EXP)

- **IDA**: Inferred from Direct Assay
- **IPI**: Inferred from Physical Interaction
- **IMP**: Inferred from Mutant Phenotype
- **IGI**: Inferred from Genetic Interaction
- **IEP**: Inferred from Expression Pattern

Curated non-experimental annotations

Computational analysis evidence codes

- **ISS**: Inferred from Sequence or Structural Similarity
  - ISO: Inferred from Sequence Orthology
  - ISA: Inferred from Sequence Alignment
  - ISM: Inferred from Sequence Model
- **IGC**: Inferred from Genomic Context
- **IBA**: Inferred from Biological aspect of Ancestor
- **IBD**: Inferred from Biological aspect of Descendant
- **IKR**: Inferred from Key Residues
- **IRD**: Inferred from Rapid Divergence
- **RCA**: inferred from Reviewed Computational Analysis

Author statement evidence codes

- **NAS**: Non-traceable Author Statement
- **TAS**: Traceable Author Statement

Curator statement evidence codes

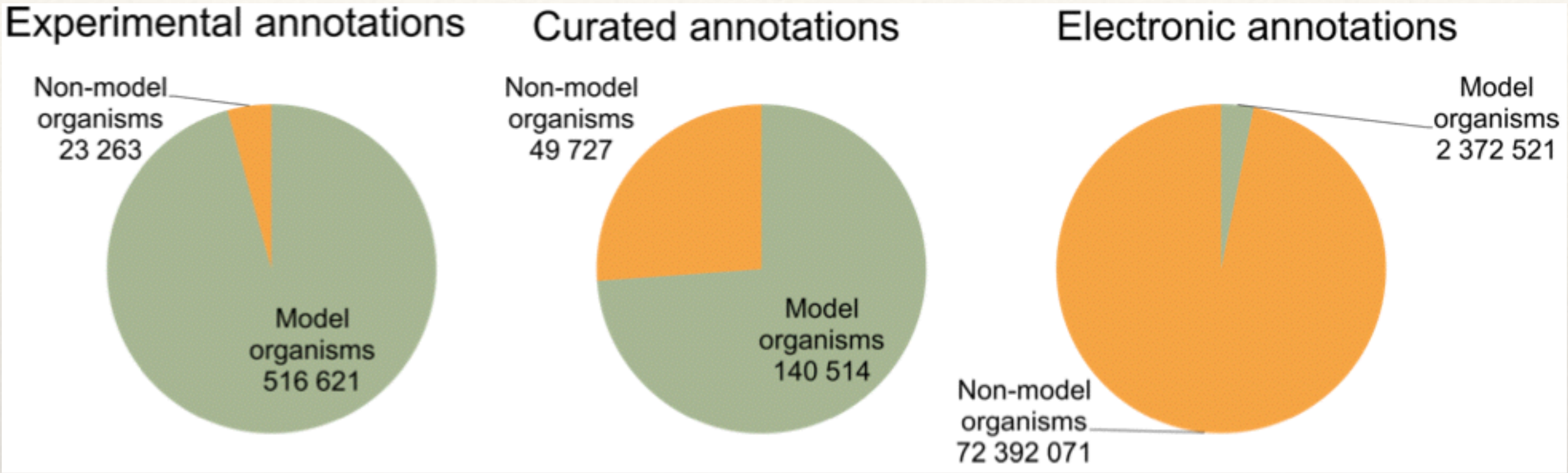
- **IC**: Inferred by Curator
- **ND**: No biological Data available

Electronic annotations

Automatically-assigned evidence code (IEA)

- Inferred from Enzyme Commission
- Inferred from InterPro
- Inferred from UniProt-GOA
- Inferred from UniProt-GOA (subcellular)
- Inferred from Ensembl Compara
- Inferred from HAMAP

Distribution of annotations



>98% of available GO annotations are electronic

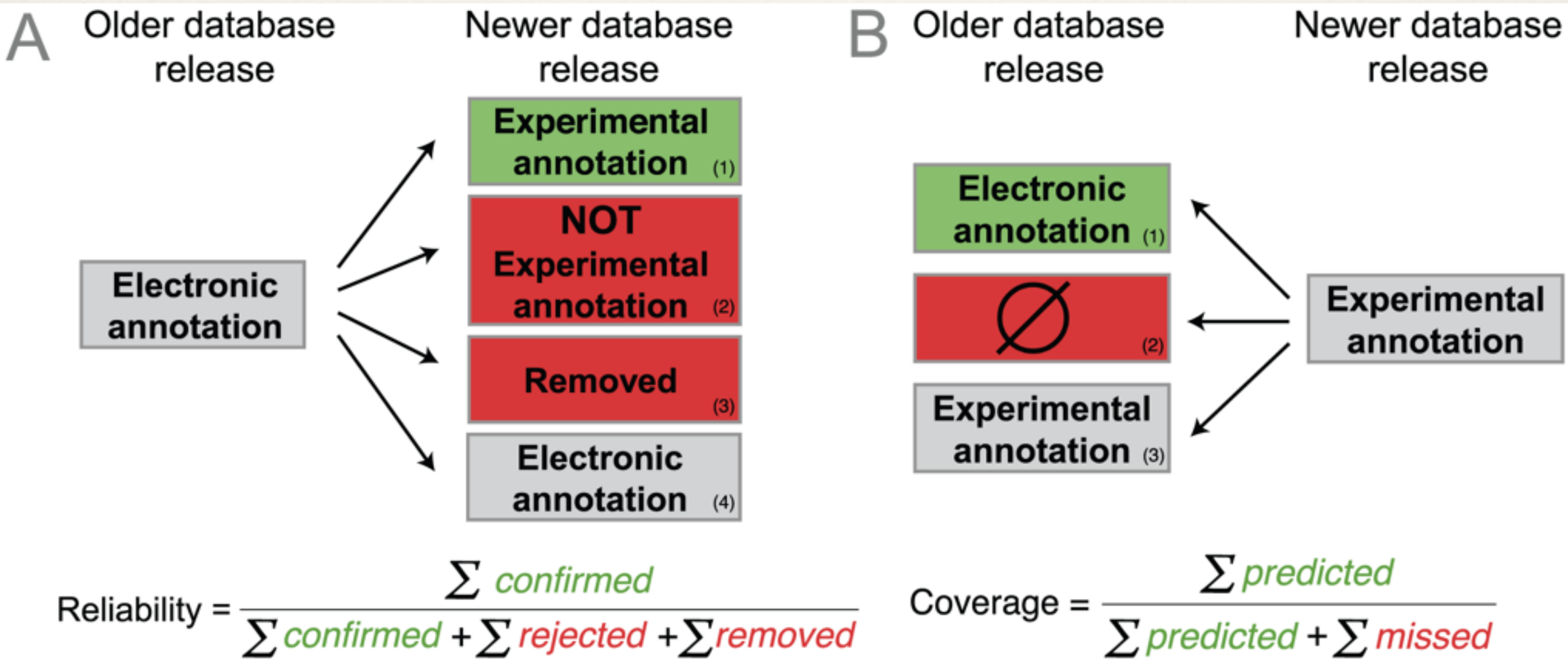
Evaluation of electronic GO annotation quality

- Analysed successive releases of UniProt-GOA
- Experimental annotations added in newer releases were used to confirm or reject earlier electronic annotations
- Only model organisms genomes were used in this analysis

Measures of quality

- **reliability** - proportion of electronic annotations confirmed by experiments
- **coverage** - power of electronic annotations to predict experimental annotations
- **specificity** - how informative the predicted GO terms are

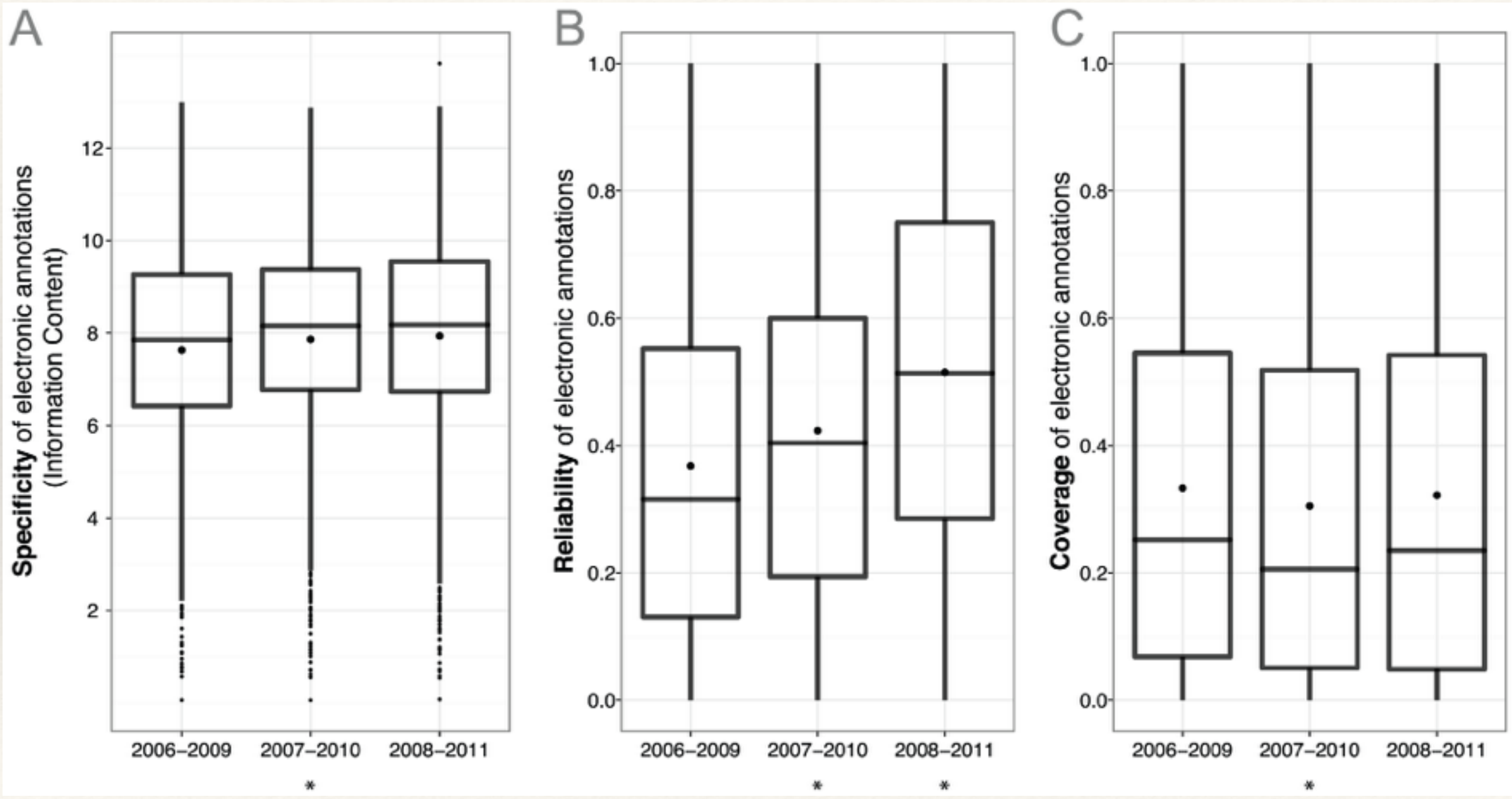
Measures of quality

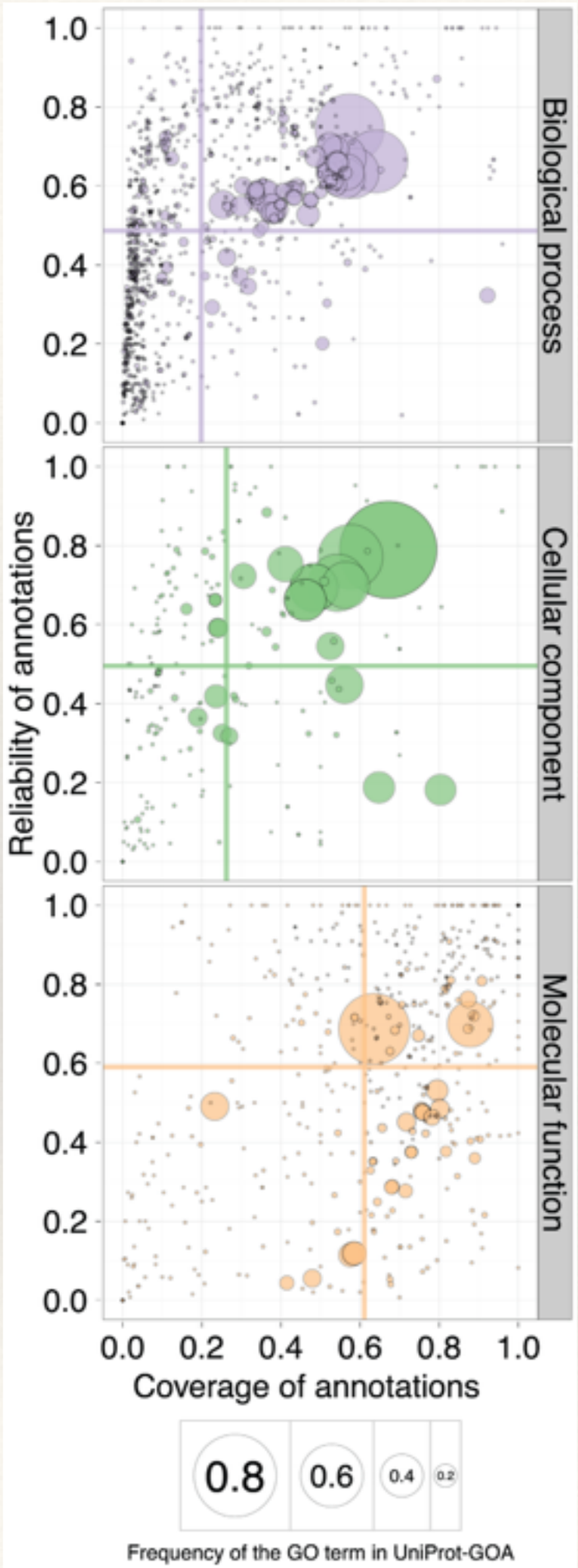


$$\text{Reliability} = \frac{\sum confirmed}{\sum confirmed + \sum rejected + \sum removed}$$

$$\text{Coverage} = \frac{\sum predicted}{\sum predicted + \sum missed}$$

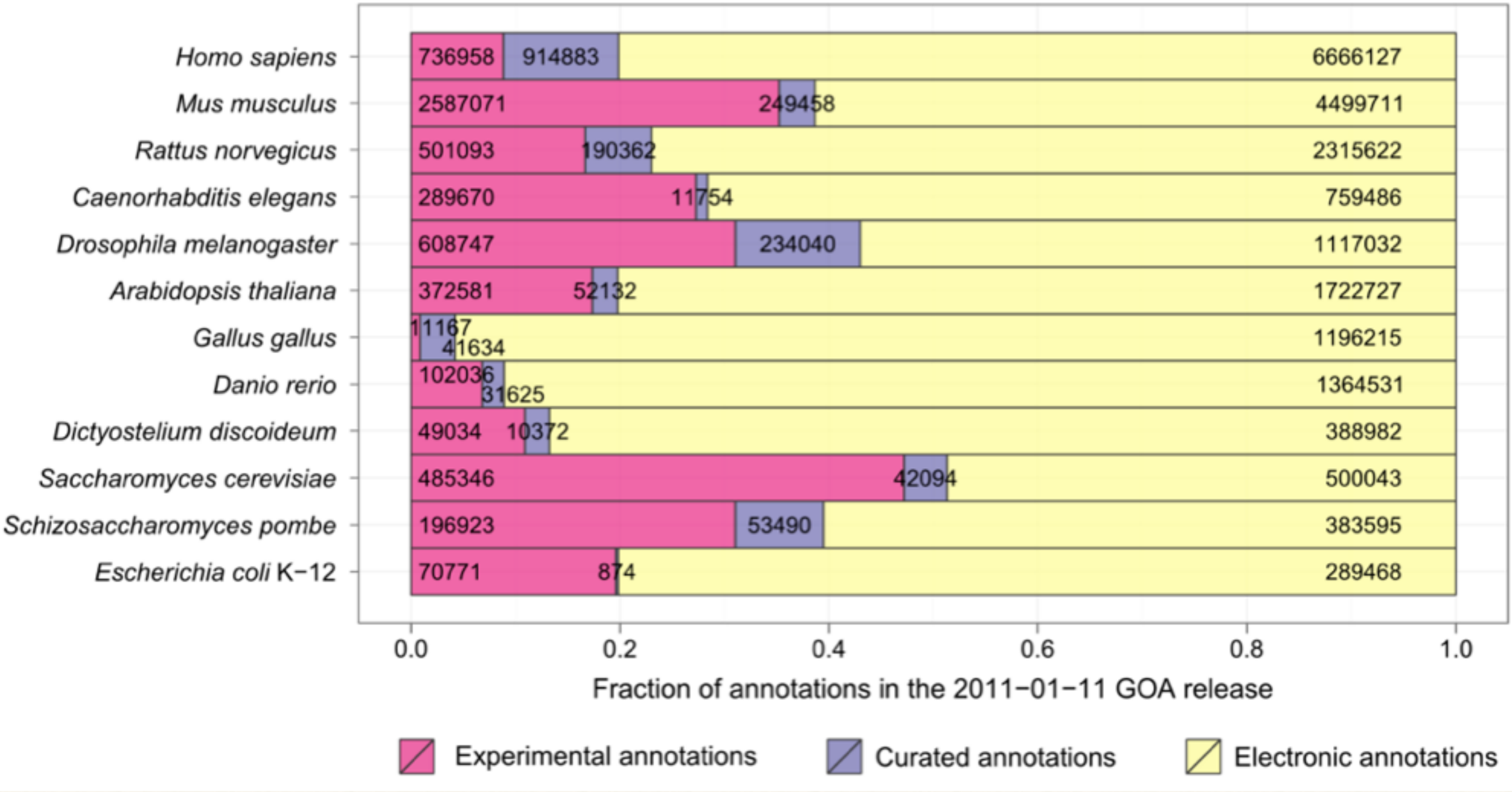
Subsequent UniProt-GOA releases





- Molecular function terms had highest coverage
- Biological process terms had lowest coverage
- Similar reliability
- General GO terms have higher reliability than specific terms.

Different model organisms



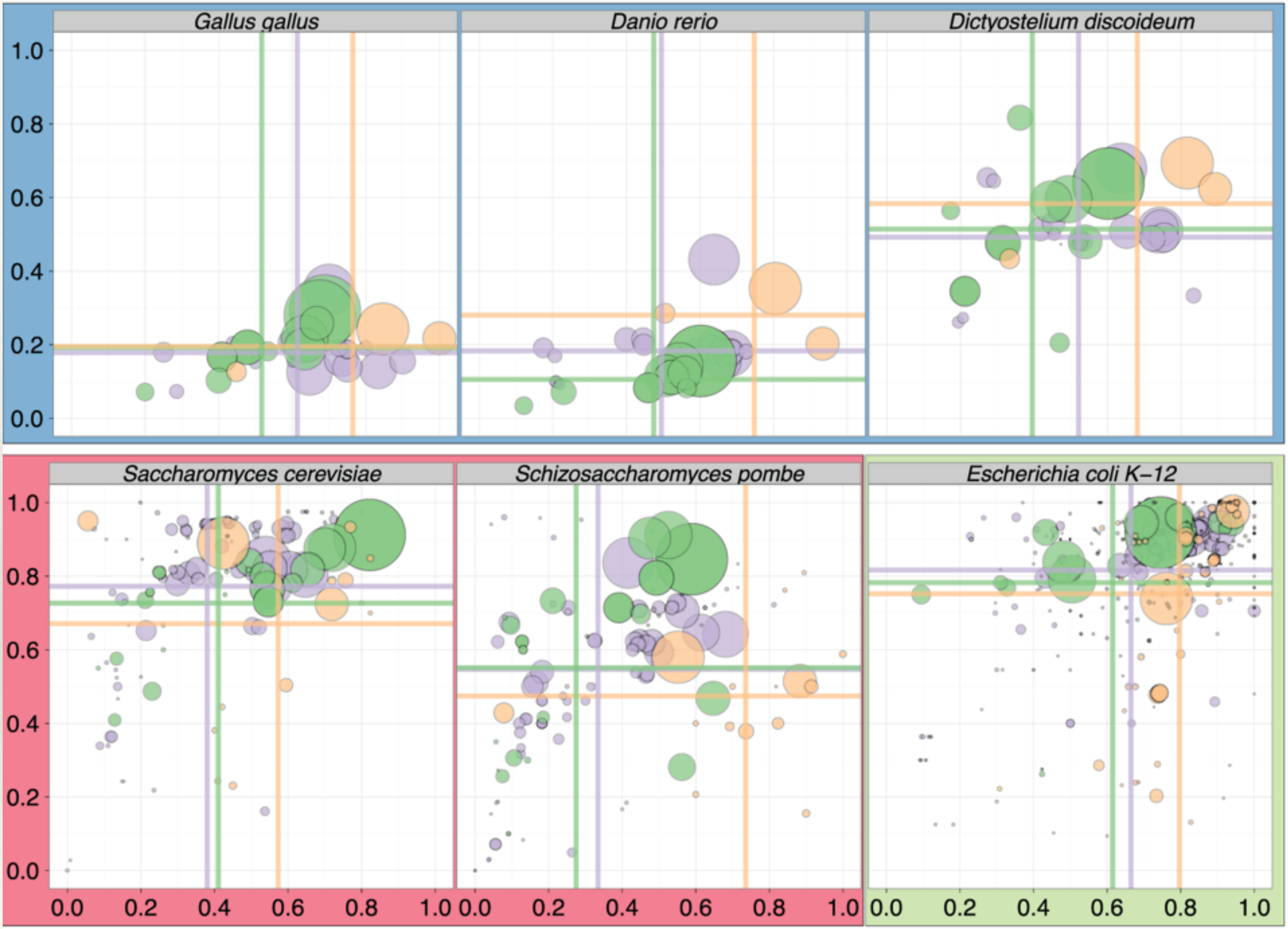
Biological process
Cellular component
Molecular function

Homo sapiens
Mus musculus
Rattus norvegicus
Caenorhabditis elegans
Drosophila melanogaster
Arabidopsis thaliana

1.0
0.8
0.6
0.4
0.2
0.0

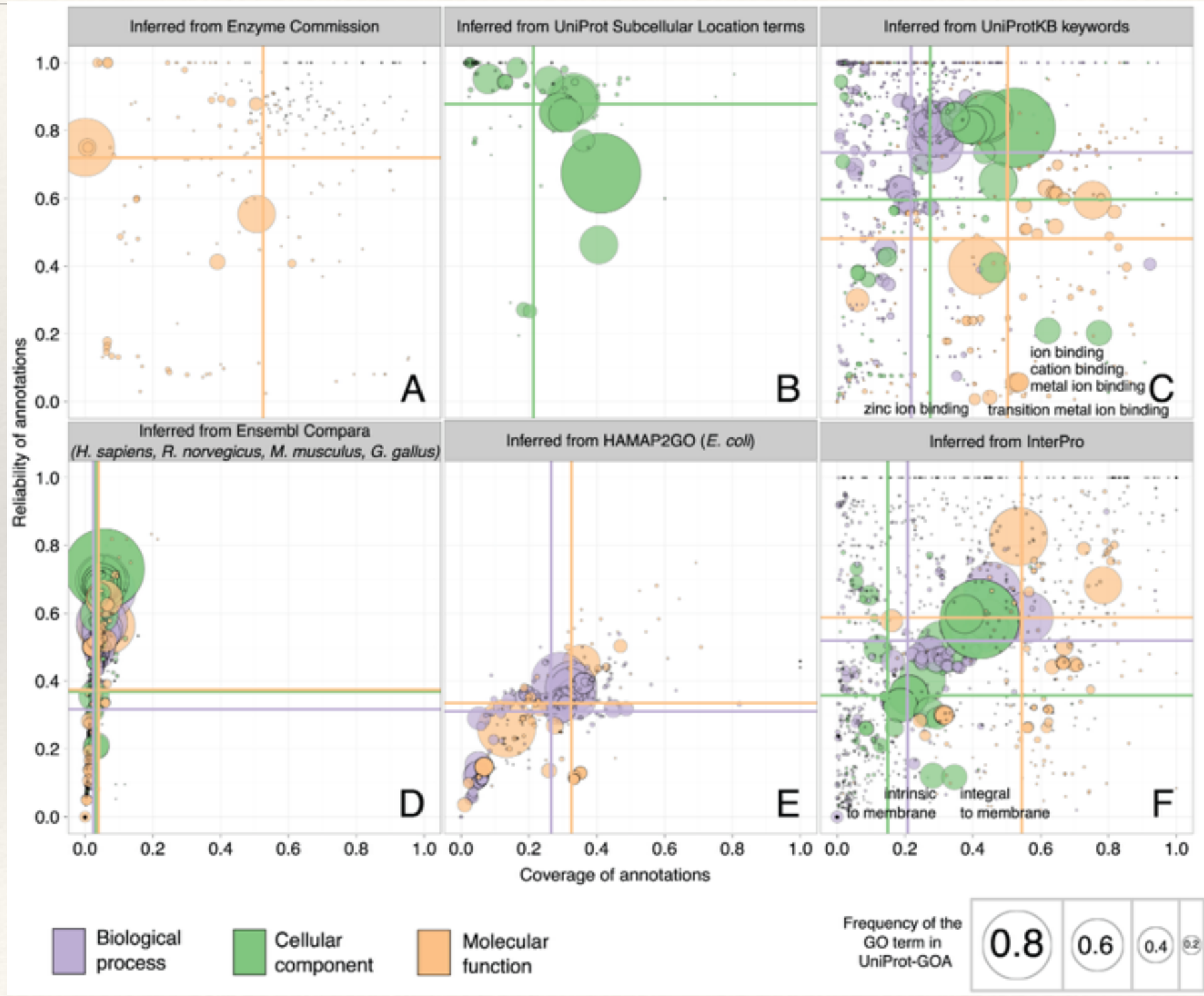
Relicability

Coverage


Gallus gallus
Danio rerio
Dictyostelium discoideum
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Escherichia coli K−12
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Relicability
Coverage

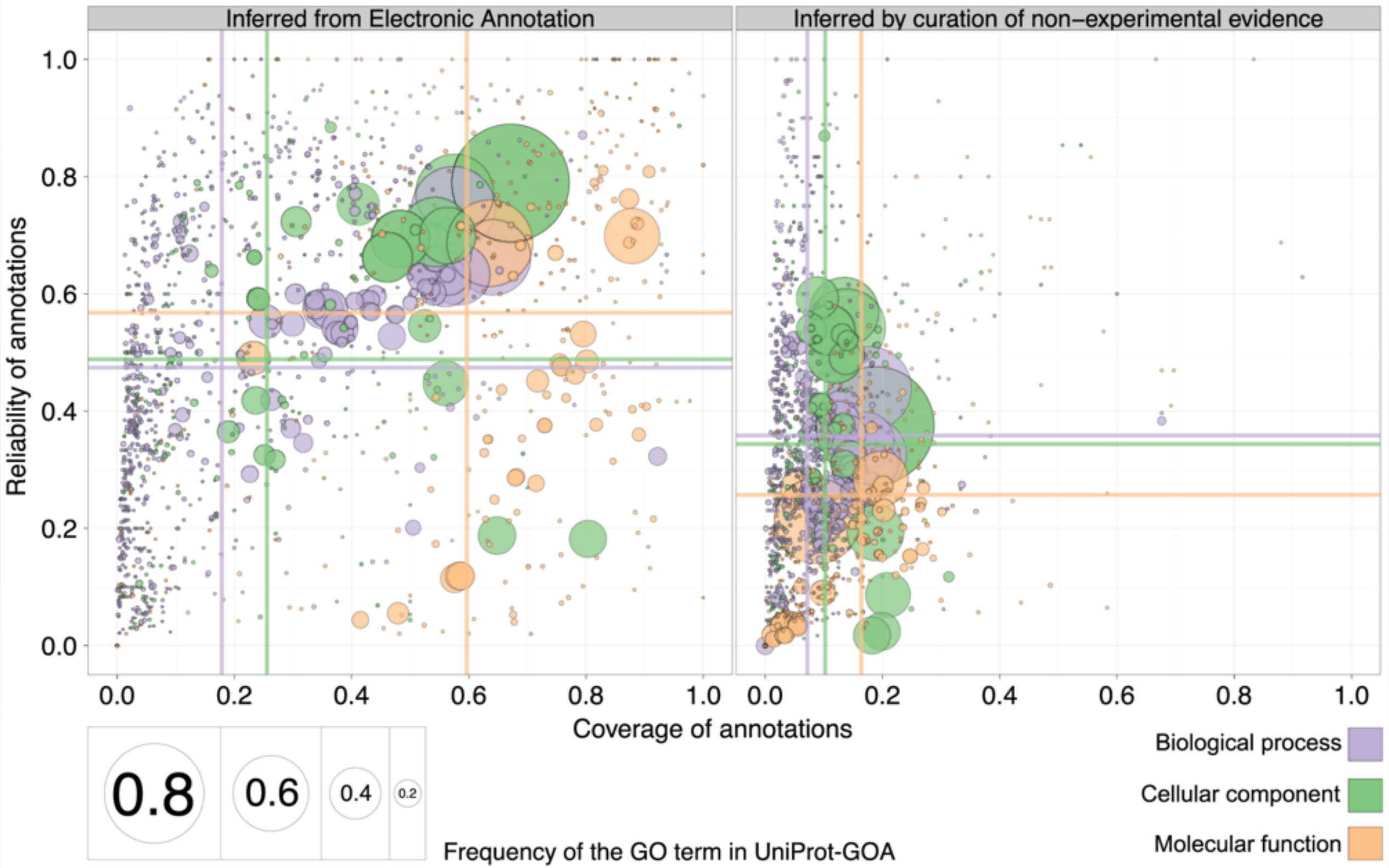
Organisms with largest number of changes have the highest quality of annotation

Different sources of electronic annotation

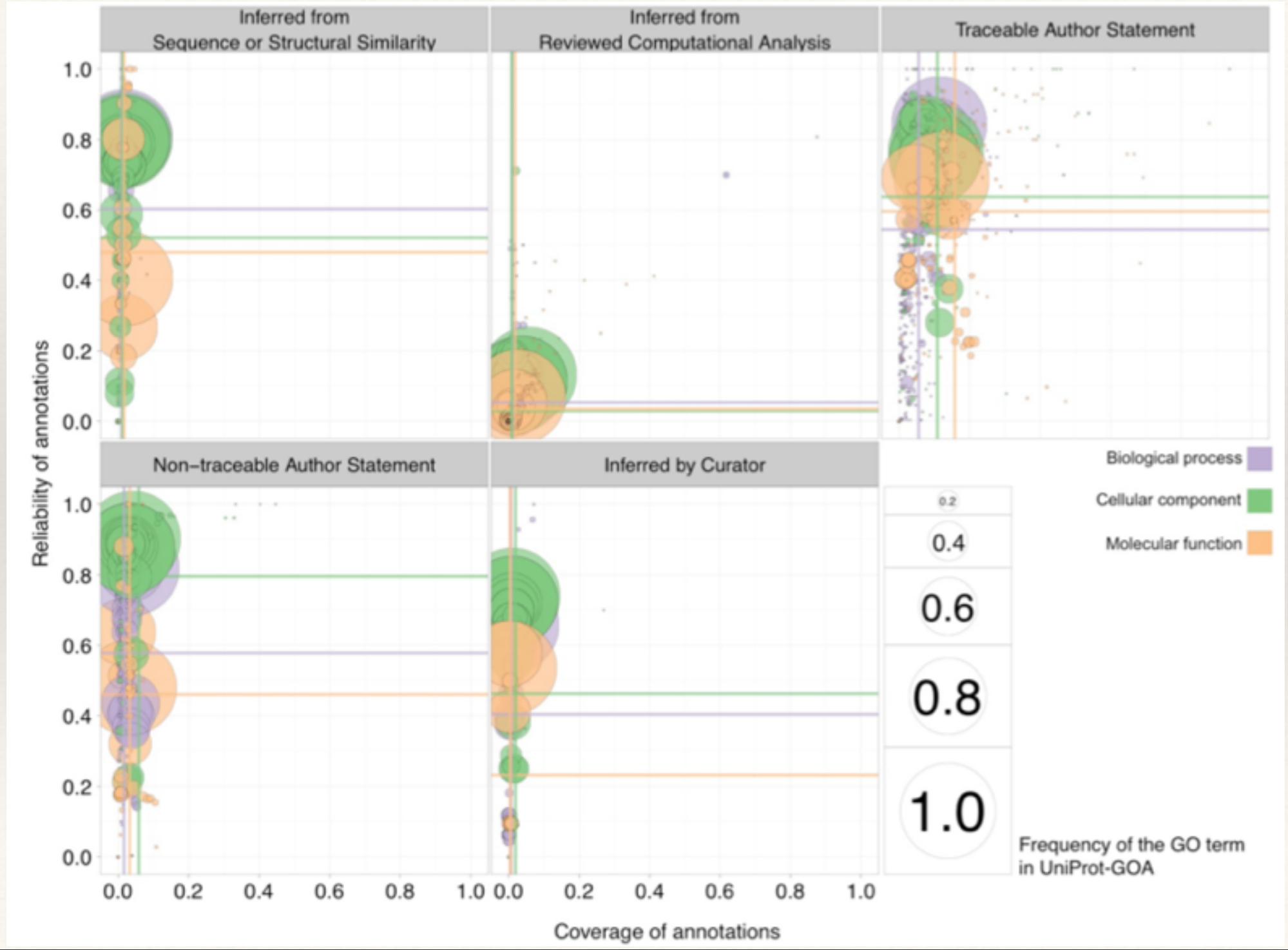


GOA strategies based on comparative genomics are currently less reliable than approaches based on sequences features

Quality of electronic and curated annotations



Quality of curated non-experimental annotations



Electronic annotations are as reliable as curated non-experimental annotations

- Coverage of electronic annotations considerably larger
- Reliability of electronic annotations 0.52, reliability of curated non-experimental annotations 0.33
- If RCA annotations were excluded, the reliability of curated annotations 0.58

Conclusions

- Reliability and specificity of annotations has improved in recent years even despite the exponential growth of databases
- Most specialised sources of annotation are most reliable. UniProt Subcellular location and EC numbers.
- Strategies based on comparative genomics are least reliable.

Curators are not redundant as the best electronic annotations rely heavily on manually curated database entries