Biases in the Experimental Annotations of Protein Function and their Effect on Our Understanding of Protein Function Space

1 Alexandra Schnoes Dept/Program/Center, University of California, San Francisco, San Francisco, CA, USA

2 David C. Ream Department of Microbiology, Miami University, Oxford, OH, USA

3 Alexander W. Thorman, Department of Microbiology, Miami University, Oxford, OH, USA

4 Patricia C. Babbitt Dept/Program/Center, University of California, San Francisco, San Francisco, CA, USA

5 Iddo Friedberg, Department of Microbiology and Computer Science & Software engineering, Miami University, Oxford, OH, USA

∗ **E-mail: corresponding i.friedberg@muohio.edu**

Abstract

Background: Computational protein function prediction programs rely upon well-annotated databases for testing and training their algorithms. These databases, in turn, rely upon the work of curators to capture experimental findings from scientific literature and apply them to protein sequence data. However, with the increasing use of high-throughput experimental assays, a small number of experimental articles dominate the functional protein annotations collected in databases. Here we investigate just how prevalent is the "few articles – many proteins" phenomenon. We hypothesize that the dominance of high-throughput experiments in proteins annotation biases our view of the corpus of functions

Should probably not have that comment in bckground.

enabled by proteins.

Is it "a log-odd" or "log-odds"?

Results: We examine the annotation of UniProtKB by the Gene Ontology Annotation project (GOA), and show that the distribution of proteins per article is a log-odd, with 0.06% of articles dominating 20% of the annotations. Since each of the dominant articles describes the use of an assay that can find only one function or a small group of functions, this leads to substantial biases, in several aspects, in what we know about the function of many proteins.

Conclusions: Given the experimental techniques available, protein function annotation bias due to high-throughput experiments is unavoidable. Knowing that these biases exist and understanding their characteristics and extent is important for database curators, developers of function annotation programs, and anyone who uses protein function annotation data to plan experiments.

Author Summary

Introduction

Functional annotation of proteins is a primary challenge in molecular biology today [**?**, 1–3]. The ongoing improvements in sequencing technology had the emphasis shifting from realizing the $1000 genome to the 1-hour genome [**?**]. The ability to rapidly and cheaply sequence genomes is creating a flood of sequence data, which require extensive analysis and characterization before these data can be useful. A large proportion of this work involves assigning biological function to newly determined gene sequences, a process that is both complex and costly [4]. Furthermore, the ability to accurately assign function

Pmid in very old version of paper.

through computational means is challenging and open problem [5]. To aid current annotation procedures and improve computational function prediction algorithms, sources of high-quality, experimentally derived functional data are necessary. Currently, one of the few repositories of such data is the UniProt-GOA database [6], which contains both computationally derived and literature derived functional information. The literature derived information is extracted by human curators who capture functional data from publications, assign the data to its appropriate place in the Gene Ontology hierarchy [7] and label them with appropriate functional evidence codes. The UniProt-GOA database is one of only a small number of databases that explicitly connects functional data, publication references and evidence codes to specific, experimentally studied sequences. In addition, annotations captured in UniProt-GOA directly impact the annotations in the UniProt/Swiss-Prot database, widely considered to be a gold standard set of functional annotation [5].

It is therefore important to understand any trends and biases that are encapsulated by the UniProt-GOA database, as those impact well-used sister databases and therefore a large number of users worldwide. Furthermore, any biases would impact function prediction algorithms development and training.

One concern surrounding the capture of functional data from articles is the propensity for high-throughput experimental work to become a large fraction of the data in UniProt-GOA, thus having few experiments dominate the protein function landscape. In this work we analyzed the relative contribution of articles to the experimental annotations in UniProt-GOA. We found some striking biases, stemming from the fact that a small fraction of articles that describe high-throughput experiments disproportionately contribute to the pool of experimental annotations of model organisms. Consequently, we show that: 1) annotations coming from high-throughput experiments are mostly less

Only for certain ontologies though, right?

informative than those provided by low-throughput experiments; 2) annotations from high throughput experiments bias the annotations towards a limited number of functions, and, 3) many high-throughput experiments overlap in the proteins they annotate, and in the annotations assigned. Taken together, our findings offer a comprehensive picture of how the current protein function landscape is generated. Furthermore, due to the biases inherent in the current system of sequence annotations, this study serves as a caution to the producers and consumers of biological data from high-throughput experiments.

I really like the sentence, but are we overstating it?

Not so sure about the word "specific". Maybe something more like... Consequently, some studies now reveal certain types of functional characteristics for large groups of proteins as a result of the particular type of assay or assays used.

Methods and Results

Articles and Proteins

With the advent of high-throughput experiments it has become possible to conduct large-scale studies of protein functions. Consequently, some studies reveal very specific functional aspects of a large amount of proteins as a result of the particular type of assay or assays used. To understand the impact of large-scale studies on the corpus of experimentally annotated proteins, we looked at ~~the UniprotKB Gene Ontology (GO) annotation files, or UniProt-GOA. UniProt-GOA proteins are individually annotated by~~ one or more GO terms using a procedure described in [6]. Briefly, this procedure consists of six steps which include sequence curation, sequence motif analyses, literature-based curation, reciprocal BLAST [8] searches, attribution of all resources leading to the included findings, and quality assurance. If the annotation source is a research article, the attribution includes its PubMed ID. For each GO term associated with a protein, there is also an *evidence code* (EC) which is used to explain how the association between the protein and the GO term was made. Experimental evidence codes include such terms as: Inferred by Direct Assay (IDA) which indicates that "a direct assay was carried out to determine the function,

At entries in the UniprotKB Gene Ontology Annotation database, or UniProt-GOA.

Proteins in UniProt-GOA have been annotated with

process, or component indicated by the GO term" or Inferred from Physical Interaction (IPI) which "Covers physical interactions between the gene product of interest and another molecule." (All EC definitions were taken from the GO site, geneontology.org). Computational evidence codes include terms such as *Inferred from Sequence or Structural Similarity* (ISS) and *Inferred from Sequence Orthology* (ISO). Although the evidence is non-experimental, the proteins annotated with these evidence codes are still assigned by a curator, rendering a level of human oversight. There are also ~~non-~~computational and non-experimental evidence codes, the most prevalent being *Inferred from Electronic Annotation* (IEA) which is "used for annotations that depend directly on computation or automated transfer of annotations from a database". IEA evidence means that the annotation was not made or checked by a person. Different degrees of reliability are associated with the different evidence codes, with experimental codes generally considered to be of higher reliability than non-experimental codes. However, the increase in the number of high-throughput experiments used to determine protein functions may introduce biases into experimental protein annotations, due to the inherent experimental capabilities and limitations of high-throughput assays. To test the hypothesis that such biases exist, and to study their extent if they do, we compiled the details of all experimentally-annotated proteins in UniProtKB. This included all proteins whose GO annotations have the GO experimental evidence codes EXP, IDA, IPI, IMP, IGI, IEP. We first examined the distribution of articles by the number of proteins they annotate. As can be seen in Figure 1, the distribution of the number of proteins annotated per article follows a power-law distribution. $f(x) = a\dot{x}^k$. Using linear regression over the log values of the axes we obtained a fit with $p < 1.18 \times 10^8$ and $R^2 = -0.72$. We therefore conclude that there is indeed a substantial bias in experimental annotations, in which there are few articles that annotate a large number of proteins.

Switch terms, use "but" instead of "and"

Alpha order?

Should be neg, i hope

Should probably mention log odds again here.

To better understand the consequences of such a distribution, we divided the annotating articles into four cohorts, based on the number of proteins each article annotates. *Single-throughput* articles are those articles that annotate only one protein; *low throughput* articles annotate 2-9 proteins; *moderate throughput* articles annotate 10-99 proteins and *high throughput* articles annotate over 99 proteins. The results are shown in Table 1. The most striking finding is that high throughput articles are responsible for 25% of the annotations in Uniprot-GOA, even though they comprise 0.08% of the articles. 96% of the articles are single-throughput ~~and~~ Or low throughput, however those annotate only 53% of the proteins in Uniprot-GOA. So while moderate throughput and high-throughput experiments account for almost half of the annotations in Uniprot-GOA, they comprise only 4% of the experiments published.

What typifies high-throughput articles? Also, how may the log-odds distribution bias what we understand of the protein function universe? To answer these questions, we examined different aspects of the annotations in the four article cohorts. Also, we examined in higher detail the top 50 annotating articles. (Overall, 62 articles in our study annotated more than 100 proteins). An initial characterization of the top 50 high-throughput articles is shown in Table**??**. As can be seen, all of the articles are specific to a single species (typically a model organism) and assay that is used to annotate the proteins in that organism. Since a single assay was used, then typically only one ontology (MFO, BPO or CCO) was used for annotation. For some species this means that a single functional aspect (MFO, BPO or CCO) of a species will be dominated by a single experiment.

? Only 62 papers in exp uniprot-goa are over 100? Just trying to make sure i understood the sentence...

?? Not sure what this means

Category (?)

Term frequency bias

To see how much a single species– and method– specific high-throughput assay affects the entire annotation of a species, we examined the relative contribution of the top-50 articles to the entire corpus of experimentally annotated protein in each species. All the species found in the top-50 articles were either common model organisms or human. For each species, we looked at the five most frequent terms in the top 50 annotating articles. We then examined the contribution of this term by the top 50 articles to the general annotations of that species. The *contribution* is the number of annotations by any given GO term in the top 50 articles divided by the number of annotations by that GO term in all of ~~UniProtKB.~~ Uniprot-goa For example, as seen in Figure 4 in *D. melanogaster* 88% of the usage of "precatalytic splicosome" is contributed by the top-50 articles. With (?)

For most organisms in the top-50 articles, the annotations were within the cellular component ontology. The exceptions are *D. melanogaster* and *C. elegans* where the dominant terms were from the Biological Process ontology, and in mouse, where "protein binding" and "identical protein binding" are from the Molecular Function Ontology. *D. melanogaster*'s annotation for the top terms is dominated (over 50% contribution) by the top-50 articles.

The term frequency bias described here can be viewed more broadly within the ontology bias. The proteins annotated by the cohorts of single-protein articles, low-throughput articles, and moderate throughput articles have similar ratios of the fraction of proteins annotated. Twenty-two to twenty-six percent of assigned terms are in the Molecular Function Ontology, and 51-57% are in the Biological Process Ontology and the remaining 17-25% are in the Cellular Component ontology. These ratios change dramatically with high-throughput articles (over 99 terms per article). In the high-throughput articles, only 5% of assigned terms are in the Molecular Function Ontology, 38% in the Biological Pro-

cess Ontology and 57% in the Cellular Compartment Ontology, ostensibly due to a lack of high-throughput assays that can be used for generating annotations using the Molecular Function Ontology.

Reannotation Bias

Are we ok with bias here? Too negative sounding?

Another type of annotation bias is that of protein re-annotation. How many of the top-50 articles actually re-annotate the same set of proteins? And how much of an agreement is there between different experiments? To investigate the extent of repetitive annotations in different articles, we clustered all the proteins annotated by the top-50 articles using CD-HIT [**?**] at 100% sequence identity. We then examined the number of clusters containing 100% identical sequences per model species. The product of the number of proteins divided by the number of clusters is the redundancy percentage. For example, if each of the top-50 articles annotating the proteins in a given species annotated the same protein set, the redundancy percentage would be 100%. The results of the reannotation bias analysis are shown in Figure 2 and in Table 3. As can be seen, the highest redundancy (65%) is in the 12 articles annotating *C. elegans*.

We have determined therefore, that there is a varying degree of repetition between experiments in the proteins they annotate, with some overlaps being quite high. In those cases, many of the same proteins in the same organism are being annotated. However, there is still a need to determined whether this annotation is consistent or not. To do this, we looked for the proteins that are annotated by more than one article, within the same ontology.

Given a protein P, let G be the GO-terms $g_1, g_2, \ldots, g_m$ that annotate that protein in all top-50 articles for an ontology $O \in \{BPO, MFO, CCO\}$. The count of each of these go terms per protein per ontology is $n_1, n_2, \ldots, n_m$ with n_i being the number of times GO

PubMedID	UniProt ID	Ontology	GO-term	description
14562095	P36023	CCO	GO:0005634	nucleus
14562095	P36023	CCO	GO:0005737	cytoplasm
16823961	P36023	CCO	GO:0005739	mitochondrion
14576278	P36023	CCO	GO:0005739	mitochondrion

term g_i annotates protein P.

The number of total annotations for a protein in an ontology is $\sum_1^m n_i$. The *maximum annotation consistency* for protein P in ontology O $0 \leq k_{P,O} \leq 1$ is calculated as:

Semi-colon Space

$$k_{P,O} = \frac{max(n_1, n_2, \ldots, n_m)}{\sum^m n_i} for max(n_1, n_2, \ldots, n_m) \geq 2$$

For example, the protein Oleate activated transcription factor 3 (UniProtID: P36023) in *S. cerevisiae* is annotated by three articles in the CC ontology: Probably write it out here not mixed abbreviate.

The annotation consistency for P36023 is therefore the maximum count of identical GO terms (*mitochondrion*, 2), divided by the total number of annotations, 4: 0.5.

Table 4 shows the results of this analysis. In *A. thaliana*, 1941 proteins are annotated by 15 articles and 18 terms in the Cellular Component ontology. The mean maximum-consistency is 0.251. The highest mean consistency is for the annotation of 807 mouse proteins annotated in Cellular Component ontology with an annotation consistency 0.832. However, that is not surprising given that there are only three annotating articles, and two annotating terms. We omitted the ontology and organism combinations that were annotated by less than three articles or two GO terms, or both.

Says 2 articles in the figure legend.

Quantifying annotation information

A common assumption holds that while high-throughput experiments do annotate more protein functions than low-throughput experiments, the former also tend to be more

shallow in the predictions they provide. The information provided, for example, by a large-scale protein binding assay will only tell us if two proteins are binding, but will not reveal whether that binding is specific, will not provide an exact K_{bind}, will not say under what conditions binding takes place, or whether there is any enzymatic reaction or signal-transduction involved. Having on hand data from experiments with different "thorughputness" levels, we set out to investigate whether there is a difference in the information provided by high-throughput experiments vs. low-throughput ones. To answer this question, we first had to quantify the information given by GO terms. One way to do so, is to use the depth of the term in the ontology: the term "enzyme activity" would be less informative than "dehalogenase" and the latter will be less informative than "haloalkane dehalogenase". We therefore counted edges from the ontology root term to the GO-term to determine term information. The larger the number of edges, the more specific – and therefore informative – the annotation. In cases where several paths lead from the root to the examined GO-term, we used the minimal path. We did so for all the annotating articles split into groups by the number of proteins each article annotates.

Just to be clear, i suggest including level numbers here

Edge counting provides a measure of term-specificity. It is, however, imperfect. The reason is that different areas of the GO DAG have different connectivities, and terms may have different depths unrelated to the intuitive specificity of a term. For example "high-affinity Tryptophan transporter", (GO:0005300) is 14 terms deep, while "anticoagulant", (GO:0008435) is only three terms deep. For this reason, information content, the logarithm of the inverse of the GO term frequency in the corpus is generally accepted as a measure of GO-term information content [**?**]. To account for the possible bias created by the GO-DAG structure, we also used the log-frequency of the terms in the experimentally annotated proteins in Uniprot-GOA. However, it should be noted that the log-frequency measure is also imperfect because, as we see throughout this study, a GO-term's frequency

may be heavily influenced by the top annotating articles, injecting a circularity problem into the use of this metric. Since no single metric for measuring the information conveyed by a GO term is wholly satisfactory, we used both in this study.

The results of both analyses are shown in Figure **??** In general, the results from the depth-based analysis and the log-frequency based analysis are in agreement, when compared across groupings based on the number of proteins annotated by the articles. For the Molecular Function ontology, the distribution of edge counts and log-frequency scores decreases as the number of annotated proteins per-article increases. For the Biological Process ontology, the decrease is significant. However the contributer to the decrease are the high-throughput articles while there is little change in the first three article cohorts. Finally, there is no significant trend of GO-depth decrease in the Cellular Component Ontology. However, using the information content metric, there is also a significant decrease in information content in the high-throughput article cohort.

Evidence and Assertion

Have (?)

There are two complementary ways by which we come to knowledge about a protein's function. The twenty evidence codes, discussed above, encapsulate the type results by which the function was inferred, but they do not capture all the necessary information. For example, "Inferred by Direct Assay (IDA)" informs that experimental evidence was used, but does not say which type of experiment was performed. This information is often needed, since knowing which experiments were performed can help the researcher establish the reliability and scope of the data. For example, RNA used in an RNAi experiment does not traverse the blood-brain-barrier, meaning that no data from the central nervous system can be drawn from an RNAi experiment. The Evidence Code Ontology, or ECO, seeks to improve upon the GO-attached evidence codes. ECO provides more elaborate

terms than "Inferred by Direct Assay": ECO also conveys which assay was used, i.e. "microscopy", "RNA interference" etc. In addition to evidence terms, the ECO ontology provides *assertion terms* in in which the nature of the assay is given. For example, an enzyme-linked immunosorbent assay (ELISA) provides quantitative protein data *in vitro* while an immunogold assay may provide the same information, and cellular localization information *in vivo*. It is therefore important to know both the assertion and the evidence to understand what sort of information may be gleaned from the assay. ~~However,~~ to understand which types of assertions are made in the top-50 high throughput articles, we manually assigned Evidence Codes Ontology (ECO) assertion and evidence terms to the top-50 articles. The ECO ontology is more elaborate than the evidence codes used by Uniprot-GOA. Although there are plans to insert ECO terms into Uniprot-GOA in the near future, those will probably not be done manually for proteins already existing in Uniprot-GOA, but by automatic mapping EC terms to ECO ontology terms using a preset table (Rachael Huntley, Chris Mungall and Tony Sawford, personal communication). Thus, the ECO-based annotations we provide here to the top 50 articles is probably more informative than a future annotation may provide. Necessary?

The results are shown in Figure **??** and in Table 5.

Interestingly, the most third most-frequently used assertion in the top experimental articles was not an experimental assertion, but rather a computational one: the term ECO:00053 "computational combinatorial evidence" is defined as "A type of combinatorial analysis where data are combined and evaluated by an algorithm." This is not a computational prediction per-se, but rather a combination of several experimental lines of evidence used in a article.

The most used experimental assertion term was ECO:000160 "protein separation followed by fragment identification evidence", which encompasses different types of mass-

spectrometry experiments. The next ranking assertion terms were computational: "motif similarity evidence" and "sequence similarity evidence used in automatic assertion". Those were generally combined with the mass-spectrometry experiments to identify protein sequence fragments reconstructed from the mass-spectrometry. Another frequently used experimental techniques was "RNAi experimental evidence". This type of experiment was mostly with the articles that used RNA interference in studying *C. elegans*, whose study comprised 12 of the top-50 articles.

Discussion

We have identified several annotation biases in UniProt-GOA. These biases stem from the uneven number of annotations produced by different types of experiments. It is clear that results from high-throughput experiments contribute substantially to the function annotation landscape, as up to 20% of experimentally annotated proteins are annotated by high-throughput assays, with most of them not being annotated by medium– or low–t̃hroughput experiments.

At the same time, high throughput experiments produce less information per protein than moderate–, low– and single–t̃hroughput experiments as evidenced by the type of terms produced in the Molecular Function and Biological Process ontologies. Furthermore, the number of total GO terms used in the high-throughput experiments is much lower than that used in low and medium throughput experiments. Therefore, while high throughput experiments provide a high coverage of protein function space, it is the low throughput experiments that provide more specific information, as well as a larger diversity of terms.

We have also identified several types of biases that are contributed by high throughput experiments. First, there is the enrichment of low-information content GO-terms, which

means that our understanding of the protein function as provided by high throughput experiments is limited. Second, there is the small number of terms used, when considering the large number of proteins that are being annotated. Third is the general term bias towards the cellular component ontology and, to a lesser extent, the Biological Process ontology; at the same time, there are very few articles that deal with the Molecular Function ontology. These biases all stem from the inherent capabilities and limitations of the hight-throughput experiments. A fourth related bias is the organism studied: taken together, *C. elegans* and *A. thaliana* studies comprise 36 (72%) of the top-50 annotating articles.

The most frequent experiment performed is cell fractionation and mass-spectrometry to assign a Cellular Component ontology terms (citations). Consequently this means that the assignment procedure is limited to the cellular compartments that can be identified with the fractionation methods used [**?**]. So while Cellular Component is the most frequent annotation used, mass-spectrometry is the most common method used to localize proteins in subcellular compartments. A notable exception to the use of MS for protein localization is in the top annotating article [9] which uses microscopy for subcellular localization. The only MS ~~experiment~~ Study (?) in the top-50 articles whose proteins were not annotated with cellular localization was "Proteome survey reveals modularity of the yeast cell machinery" [9]. The resulting annotation was "protein binding" form the Molecular Function ontology. A more detailed discussion on this study follows in the section **Information Capture** below.

Cut? How is it different from the previous sentence?

The second most frequent type of experiments used (Experiment was) RNA Interference (RNAi) whole-genome gene knockdowns in *C. elegans*, *D. melanogaster* and one in *C. albicans*. RNAi experiments typically use targeted dsRNA which is delivered to the organism and silences specific genes. Typically the experiments here used libraries of RNAi targeted to the whole exome. The phenotypes searched for were mostly associated with embryonic and

post-embryonic development [?]. Some studies focused on mitotic spindle assembly [10], lipid storage [10] and endocytic traffic [10]. One study used RNAi and MP to identify mitochondrial protein localization [11]. ?

These two types of assays (mass-spectrometry and RNAi) were strongly linked to the other frequently used experimental ECO terms, by the nature of the methodology used. Thus, "protein separation followed by fragment identification evidence" is usually accompanied with "cell fractionation evidence" and "Western blot evidence". "RNAi experimental evidence" is generally associated with "mutant phenotype evidence used in manual assertion". All experiments are associated with computational ECO terms, which describe sequence similarity and motif recognition techniques used to identify the sequences found. Thus, a strong reliance on computational annotation is an integral part of high throughput experiments. It should be noted that computational annotation here is not necessarily used directly for functional annotation, but rather for identifying the protein by a sequence or motif similarity search.

Information Capture and Scope of GO

We have discussed the information loss that is characteristic of high-throughput experiments, due to the nature of these experiments. However, another reason for information loss is the inability to capture certain types of information using the Gene Ontology. GO is knowingly limited to three aspects (MF, BP and CC) of biological function, which are assigned per protein. However, other aspects of function may emerge from experiments that cannot be captured by GO. Of note is the study mentioned earlier, "Proteome survey reveals modularity of the yeast cell machinery" [9]. In this study, the information produced was primarily of protein complexes, which proteins are binding which proteins, and the relationship to cellular compartmentalization and biological networks. At the same

time, the only GO-term captured in the curation of this study was "protein binding". Some, but not all of this information can be captured more specifically using the children of the term "protein binding", but such a process is arguably laborious by manual curation of a high throughput article. Furthermore, the main information conveyed by this article, namely the types of protein complexes discovered and how they relate to cellular networks, is outside the scope of GO. It is important to realize that while high-throughput experiments do convey less information per protein within the functional scope as defined by GO, they still convey composite information such as possible pathway mappings – information which needs to be captured into annotation databases by means other than GO. In the example above, the information can be captured by a protein interaction database, but not by GO annotation.

So binding partners arent supposed to be captured in go?

I think we are possibly missing a paragraph on identical vs non identical annotaions for 100% id sequences. What do you think?

Conclusions

Taken together, the annotation biases noted in this study affect our understanding of protein function space. This, in turn, affects out ability to properly understand the connection between predictors of protein function and the actual function – the hallmark of computational function annotation. As a dramatic example, during the Critical Assessment of Function Annotation experiment (Radivojac *et al* in review) we have noticed that about 20% of the proteins participating in the challenge and annotated with the Molecular Function Ontology were annotated as "protein binding", a GO-term that conveys little information. Furthermore, it was shown that the major contribution of "protein binding" term to the CAFA challenge data set was due to high-throughput assays. This illustrates how the concentration of a large number of annotations in a small number of studies provides only a partial picture of the function of these proteins. As we have seen,

the picture provided from high throughput experiments is mainly of: 1. subcellular localization cell fractionation and MS based localization and 2. developmental phenotypes. While these data are important, we should be mindful of this bias when examining protein function in the database, even those annotations deemed to be of high quality, i.e. with experimental verification. Furthermore, such a large bias in prior probabilities can adversely affect programs employing prior probabilities, as most machine-learning programs do. Many researchers use programs based on machine-learning algorithms to predict the function of unknown proteins. If the training set for these programs has included a disproportional number of annotations by thigh-throughput experiments, the results these programs provide will be strongly biased towards a few frequent and shallow GO-terms.

Several steps can be taken to remedy this situation. Annotations are derived from high-throughput experiments can be flagged as such in the database. The flagging can then be read by sequence similarity or other search software, and flagged proteins removed or otherwise marked in the search. In a typical scenario, a researcher will BLAST their query protein to determine its function by sequence similarity. If a target protein is tagged as annotated by a high throughput assay, it would be removed form the search if asked to do so by the user. This filtering can also be done by assay type, number of proteins annotated per experiment, or a combination of the above. This requires that GO-annotated proteins should also be annotated with assertion codes in addition to the evidence codes and GO term-codes; but given the large volume of data in UniprotKB is it hard to expect such massive reannotation with assertion terms undertaken. (Any other ideas?)

We call upon the communities of annotators, computational biologists and experimental biologists to be mindful of the phenomenon of the biases described in this study, and to work together to understand its implications and mitigate its impact.

Isnt that happening, though, with goa moving to eco?

Things that could be recommended...
1. Adjust/ be careful of composition of training sets.
2. Use caution when transfering annotations from highthroughput sequences (because of unspecificity?)
3. Refrain from annotating sequence dbs with highthroughput paper info.
4. Use assertions/eco
5. Support standards & practices in publications to make finding function in a paper "easy" so that the temptation wont always be to add highthroughput papers because you "get more bang for your buck"

Conclusions:
1. Highthroughput experiments, currently, will not particularly help our protein function problem

Note on Methods

We used the Uniprot-GOA database from December 2011. Data analyses were performed using Python scripts. ECO terms classifying the proteins in the top 50 experiments were assigned to the proteins manually after reading the articles. All data and scripts are available on http://github.com/FriedbergLab/DataBias/

Acknowledgments

We thank Predrag Radivojac, members of the Friedberg and Babbitt labs for insightful discussions. This research was funded, in part by NSF / ABI XXXXXXX-XXX award to IF and XXXXXXXXXXXX to PCB.

References

References

1. Friedberg I (2006) Automated protein function prediction–the genomic challenge. Brief Bioinform 7: 225–242.

2. Erdin S, Lisewski AM, Lichtarge O (2011) Protein function prediction: towards integration of similarity metrics. Current Opinion in Structural Biology 21: 180 - 188.

3. Rentzsch R, Orengo CA (2009) Protein function prediction the power of multiplicity. Trends in Biotechnology 27: 210 - 219.

4. Sboner A, Mu X, Greenbaum D, Auerbach R, Gerstein M (2011) The real cost of sequencing: higher than you think! Genome Biology 12: 125+.

5. Schnoes AM, Brown SD, Dodevski I, Babbitt PC (2009) Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 5: e1000605+.

6. Dimmer EC, Huntley RP, Alam-Faruque Y, Sawford T, O'Donovan C, et al. (2012) The uniprot-go annotation database in 2011. Nucleic Acids Research 40: D565–D570.

7. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. Nature Genetics 25: 25–29.

8. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic acids research 25: 3389–3402.

9. Barbe L, Lundberg E, Oksvold P, Stenius A, Lewin E, et al. (2008) Toward a confocal subcellular atlas of the human proteome. Mol Cell Proteomics 7: 499–508.

10. Goshima G, Wollman R, Goodwin SS, Zhang N, Scholey JM, et al. (2007) Genes required for mitotic spindle assembly in Drosophila S2 cells. Science 316: 417–421.

11. Hughes JR, Meireles AM, Fisher KH, Garcia A, Antrobus PR, et al. (2008) A microtubule interactome: complexes with roles in cell cycle and mitosis. PLoS Biol 6: e98.

12. Matsuyama A, Arai R, Yashiroda Y, Shirai A, Kamata A, et al. (2006) ORFeome cloning and global analysis of protein localization in the fission yeast Schizosaccharomyces pombe. Nat Biotechnol 24: 841–847.

13. Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, et al. (2008) A mitochondrial protein compendium elucidates complex I disease biology. Cell 134: 112–123.

14. Simmer F, Moorman C, van der Linden AM, Kuijk E, van den Berghe PV, et al. (2003) Genome-wide RNAi of C. elegans using the hypersensitive rrf-3 strain reveals novel gene functions. PLoS Biol 1: E12.

15. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, et al. (2003) Global analysis of protein localization in budding yeast. Nature 425: 686–691.

16. Zybailov B, Rutschow H, Friso G, Rudella A, Emanuelsson O, et al. (2008) Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. PLoS ONE 3: e1994.

17. Sonnichsen B, Koski LB, Walsh A, Marschall P, Neumann B, et al. (2005) Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans. Nature 434: 462–469.

18. Mootha VK, Bunkenborg J, Olsen JV, Hjerrild M, Wisniewski JR, et al. (2003) Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. Cell 115: 629–640.

19. Benschop JJ, Mohammed S, O'Flaherty M, Heck AJ, Slijper M, et al. (2007) Quantitative phosphoproteomics of early elicitor signaling in Arabidopsis. Mol Cell Proteomics 6: 1198–1214.

20. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, et al. (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 421: 231–237.

21. Mawuenyega KG, Forst CV, Dobos KM, Belisle JT, Chen J, et al. (2005) Mycobacterium tuberculosis functional network analysis by global subcellular protein profiling. Mol Biol Cell 16: 396–404.

22. Ito J, Batth TS, Petzold CJ, Redding-Johanson AM, Mukhopadhyay A, et al. (2011) Analysis of the Arabidopsis cytosolic proteome highlights subcellular partitioning of central plant metabolism. J Proteome Res 10: 1571–1582.

23. Rual JF, Ceron J, Koreth J, Hao T, Nicot AS, et al. (2004) Toward improving Caenorhabditis elegans phenome mapping with an ORFeome-based RNAi library. Genome Res 14: 2162–2168.

24. Reinders J, Zahedi RP, Pfanner N, Meisinger C, Sickmann A (2006) Toward the complete yeast mitochondrial proteome: multidimensional separation techniques for mitochondrial proteomics. J Proteome Res 5: 1543–1554.

25. Fernandez-Calvino L, Faulkner C, Walshaw J, Saalbach G, Bayer E, et al. (2011) Arabidopsis plasmodesmal proteome. PLoS ONE 6: e18880.

26. Gu S, Chen J, Dobos KM, Bradbury EM, Belisle JT, et al. (2003) Comprehensive proteomic profiling of the membrane constituents of a Mycobacterium tuberculosis strain. Mol Cell Proteomics 2: 1284–1296.

27. Ferro M, Brugiere S, Salvi D, Seigneurin-Berny D, Court M, et al. (2010) ATCHLORO, a comprehensive chloroplast proteome database with subplastidial

localization and curated information on envelope proteins. Mol Cell Proteomics 9: 1063–1084.

28. Kleffmann T, Russenberger D, von Zychlinski A, Christopher W, Sjolander K, et al. (2004) The Arabidopsis thaliana chloroplast proteome reveals pathway abundance and novel protein functions. Curr Biol 14: 354–362.

29. Sassetti CM, Boyd DH, Rubin EJ (2003) Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol 48: 77–84.

30. Balklava Z, Pant S, Fares H, Grant BD (2007) Genome-wide analysis identifies a general requirement for polarity proteins in endocytic traffic. Nat Cell Biol 9: 1066–1073.

31. Mitra SK, Gantt JA, Ruby JF, Clouse SD, Goshe MB (2007) Membrane proteomic analysis of Arabidopsis thaliana using alternative solubilization techniques. J Proteome Res 6: 1933–1950.

32. Maeda I, Kohara Y, Yamamoto M, Sugimoto A (2001) Large-scale analysis of gene function in Caenorhabditis elegans by high-throughput RNAi. Curr Biol 11: 171–176.

33. Ceron J, Rual JF, Chandra A, Dupuy D, Vidal M, et al. (2007) Large-scale RNAi screens identify novel genes that interact with the C. elegans retinoblastoma pathway as well as splicing-related components with synMuv B activity. BMC Dev Biol 7: 30.

34. Sickmann A, Reinders J, Wagner Y, Joppich C, Zahedi R, et al. (2003) The proteome of Saccharomyces cerevisiae mitochondria. Proc Natl Acad Sci USA 100: 13207–13212.

35. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440: 631–636.

36. Green RA, Kao HL, Audhya A, Arur S, Mayers JR, et al. (2011) A high-resolution C. elegans essential gene network based on phenotypic profiling of a complex tissue. Cell 145: 470–482.

37. Simpson JC, Wellenreuther R, Poustka A, Pepperkok R, Wiemann S (2000) Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. EMBO Rep 1: 287–292.

38. Marmagne A, Ferro M, Meinnel T, Bruley C, Kuhn L, et al. (2007) A high content in lipid-modified peripheral proteins and integral receptor kinases features in the arabidopsis plasma membrane proteome. Mol Cell Proteomics 6: 1980–1996.

39. Dunkley TP, Hester S, Shadforth IP, Runions J, Weimar T, et al. (2006) Mapping the Arabidopsis organelle proteome. Proc Natl Acad Sci USA 103: 6518–6523.

40. Jaquinod M, Villiers F, Kieffer-Jaquinod S, Hugouvieux V, Bruley C, et al. (2007) A proteomics dissection of Arabidopsis thaliana vacuoles isolated from cell culture. Mol Cell Proteomics 6: 394–412.

41. Heazlewood JL, Tonti-Filippini JS, Gout AM, Day DA, Whelan J, et al. (2004) Experimental analysis of the Arabidopsis mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins. Plant Cell 16: 241–256.

42. Ashrafi K, Chang FY, Watts JL, Fraser AG, Kamath RS, et al. (2003) Genome-wide RNAi analysis of Caenorhabditis elegans fat regulatory genes. Nature 421: 268–272.

43. Piano F, Schetter AJ, Morton DG, Gunsalus KC, Reinke V, et al. (2002) Gene clustering based on RNAi phenotypes of ovary-enriched genes in C. elegans. Curr Biol 12: 1959–1964.

44. Carter C, Pan S, Zouhar J, Avila EL, Girke T, et al. (2004) The vegetative vacuole proteome of Arabidopsis thaliana reveals predicted and unexpected proteins. Plant Cell 16: 3285–3303.

45. Da Cruz S, Xenarios I, Langridge J, Vilbois F, Parone PA, et al. (2003) Proteomic analysis of the mouse liver mitochondrial inner membrane. J Biol Chem 278: 41566–41571.

46. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al. (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437: 1173–1178.

47. Bakthavatsalam D, Gomer RH (2010) The secreted proteome profile of developing Dictyostelium discoideum cells. Proteomics 10: 2556–2559.

48. Froehlich JE, Wilkerson CG, Ray WK, McAndrew RS, Osteryoung KW, et al. (2003) Proteomic study of the Arabidopsis thaliana chloroplastic envelope membrane utilizing alternatives to traditional two-dimensional electrophoresis. J Proteome Res 2: 413–425.

49. Stroschein-Stevenson SL, Foley E, O'Farrell PH, Johnson AD (2006) Identification of Drosophila gene products required for phagocytosis of Candida albicans. PLoS Biol 4: e4.

50. Rutschow H, Ytterberg AJ, Friso G, Nilsson R, van Wijk KJ (2008) Quantitative proteomics of a chloroplast SRP54 sorting mutant and its genetic interactions with CLPC1 in Arabidopsis. Plant Physiol 148: 156–175.

51. Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, et al. (2002) Subcellular localization of the yeast proteome. Genes Dev 16: 707–719.

52. Fraser AG, Kamath RS, Zipperlen P, Martinez-Campos M, Sohrmann M, et al. (2000) Functional genomic analysis of C. elegans chromosome I by systematic RNA interference. Nature 408: 325–330.

53. Gonczy P, Echeverri C, Oegema K, Coulson A, Jones SJ, et al. (2000) Functional genomic analysis of cell division in C. elegans using RNAi of genes on chromosome III. Nature 408: 331–336.

54. Suzuki H, Fukunishi Y, Kagawa I, Saito R, Oda H, et al. (2001) Protein-protein interaction panel using mouse full-length cDNAs. Genome Res 11: 1758–1765.

55. Sarry JE, Kuhn L, Ducruix C, Lafaye A, Junot C, et al. (2006) The early responses of Arabidopsis thaliana cells to cadmium exposure explored by protein and metabolite profiling analyses. Proteomics 6: 2180–2198.

56. Chen D, Toone WM, Mata J, Lyne R, Burns G, et al. (2003) Global transcriptional responses of fission yeast to environmental stress. Mol Biol Cell 14: 214–229.

57. Herold N, Will CL, Wolf E, Kastner B, Urlaub H, et al. (2009) Conservation of the protein composition and electron microscopy structure of Drosophila melanogaster and human spliceosomal complexes. Mol Cell Biol 29: 281–301.

58. Bayer EM, Bottrill AR, Walshaw J, Vigouroux M, Naldrett MJ, et al. (2006) Arabidopsis cell wall proteome defined using multidimensional protein identification technology. Proteomics 6: 301–311.

Figures

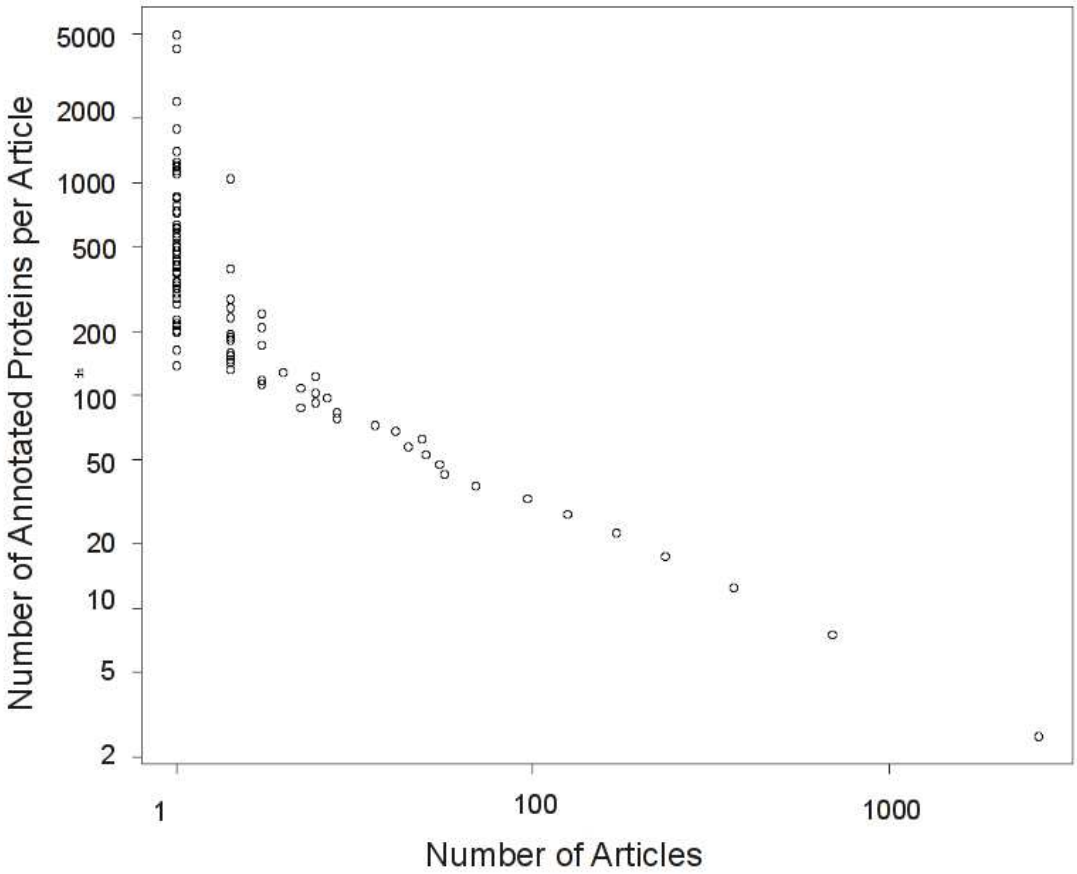

Figure 1. Distribution of the number of proteins annotated per article. X-axis: number of annotating articles. Y-axis: number of annotated proteins. The distribution was found to be logarithmic with a significant ($R^2 = 0.72$, $p < 1.10 \times 10^{18}$) linear fit to the log-log plot. The data came from 76137 articles annotating 256033 proteins with GO experimental evidence codes, in Uniprot-GOA 12/2011.

Not the same p value as in the paper

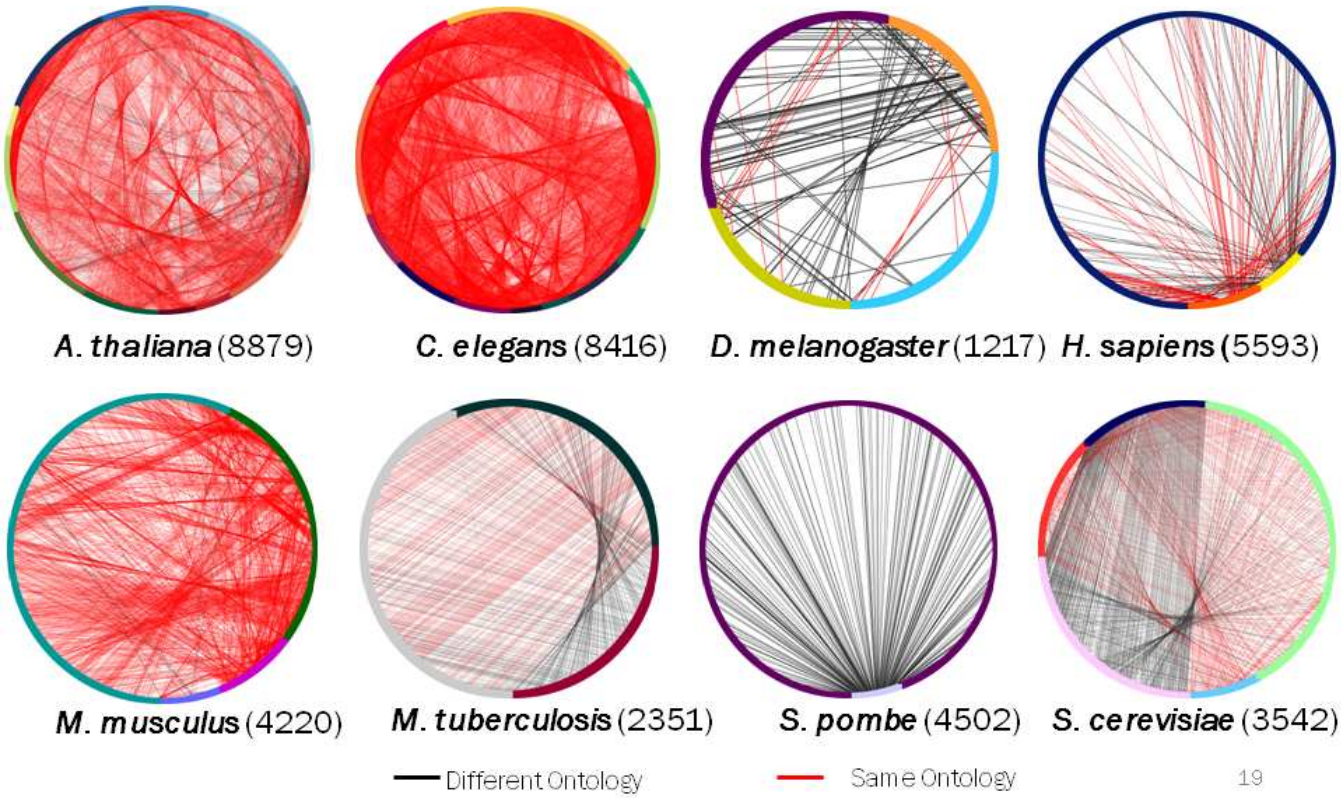

Figure 2. Redundancy in proteins described by the top-50 articles. A circle represents the sum total of articles annotating each organism. Each colored arch is composed of all the proteins in a single article. A line is drawn between any two points on the circle if the proteins they represent have 100% sequence identity. A black line is drawn if they are annotated with a different ontology (e.g. in one article the protein is annotated with the MFO, and in another article with BPO); a red line if they are annotated in the same ontology. Example: *S. pombe* is described by two articles, one with few protein (light arch on bottom) and one with many (dark arch encompassing most of circle). Many of the same proteins are annotated by both articles. See table 3 for numbers.

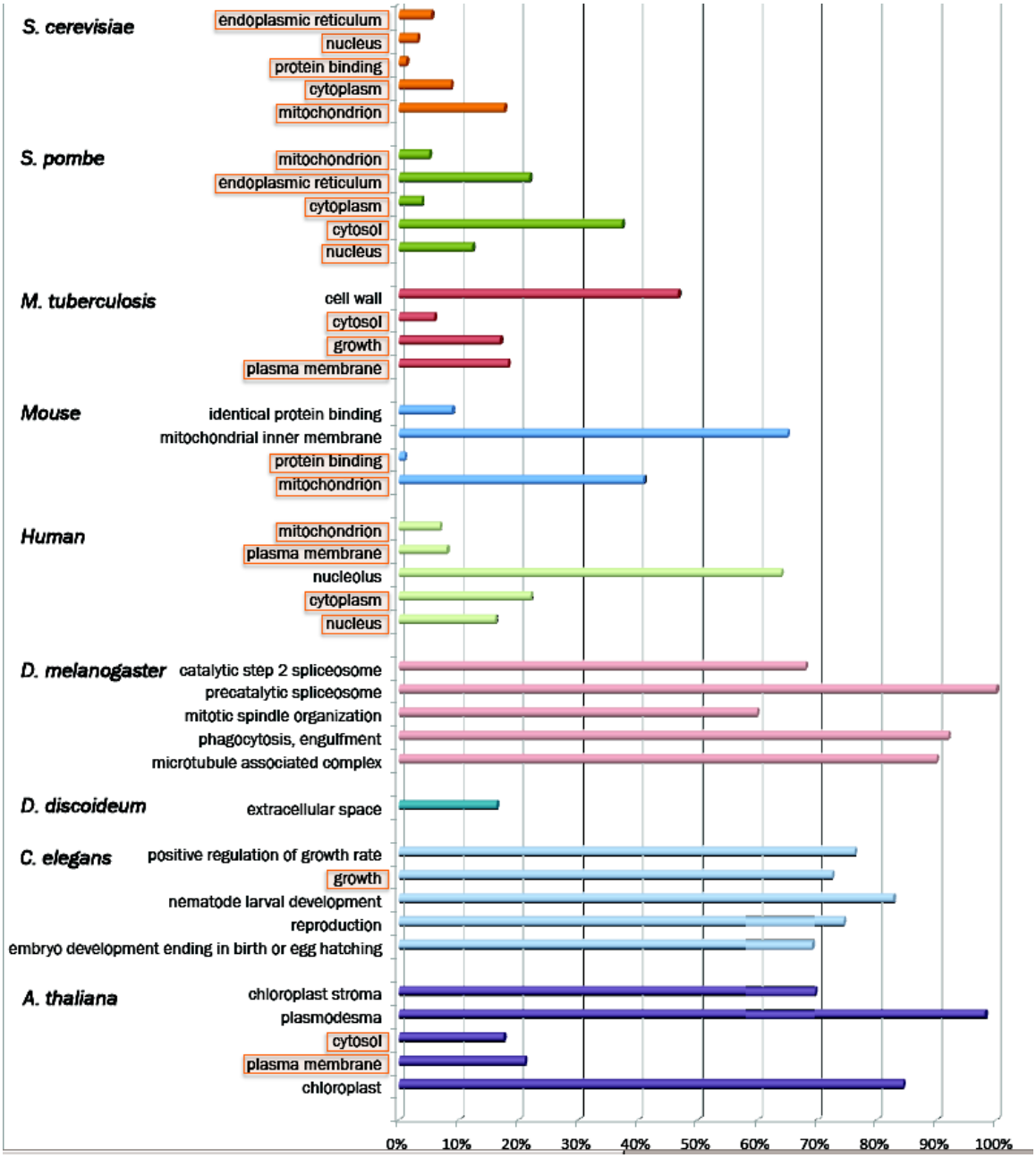

Figure 3. Relative contribution of top-50 articles to the annotation of major model organisms. The length of each bar represents the percentage of proteins annotated by the top-50 articles in a given organism by a given GO term.

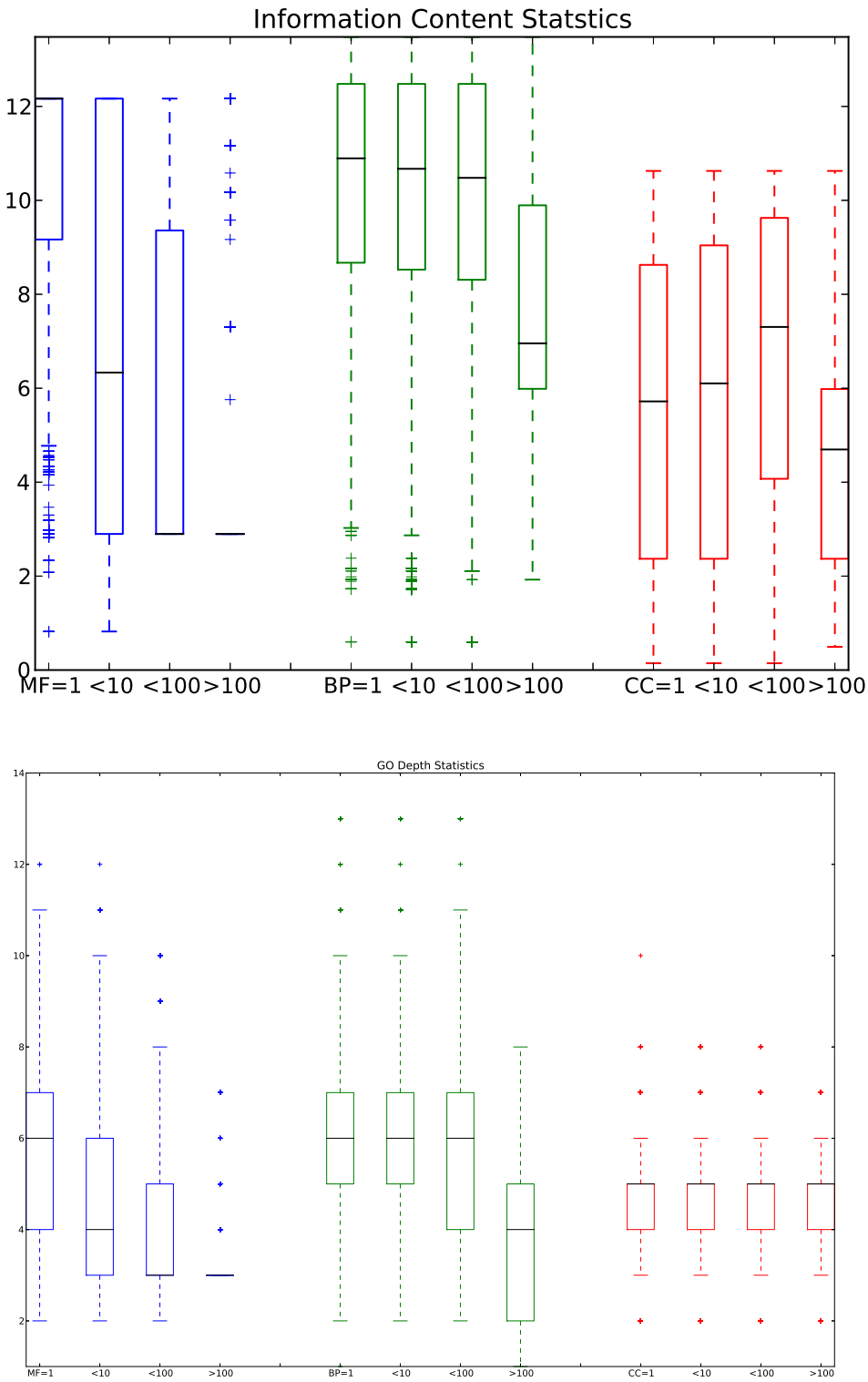

Figure 4. Information provided by articles depending on the number of proteins the articles annotate. Articles are grouped into cohorts: 1: one protein annotated by article; ¡10: more than 1, less than 10 annotated; ¡100: more than 10, less than 100 annotated; ≥ 100: more than 100 proteins annotated per article. Blue bars: Molecular Function ontology; Green bars: Biological Process ontology; Red bars: Cellular Component ontology. Information is gauged by **A**: Information Content and**B**: GO depth. See text for details.

Tables

Table 1. Annotation Cohorts

Articles annotating the following number of proteins	1	$1 < n \leq 10$	$10 < n \leq 100$	$n > 100$	SUM
Number of proteins annotated	20699	46383	26485	31411	124978
Number of annotating articles	41156	32201	2668	62	76087
Percent of proteins annotated	16.56	37.11	21.19	25.13	100
Percent of annotating articles	54.09	42.32	3.51	0.08	100

Table caption

Table 2. Top 50 Annotating Articles

N	Proteins	Annotations	Species	ref.	MFO	BPO	CCO
1	4937	11050	*H. sapiens*	[9]	0	0	11050
2	4247	7046	*S. pombe*	[12]	0	0	7046
3	2412	2412	*H. sapiens*	[13]	0	0	2412
4	1791	5918	*C. elegans*	[14]	0	5918	0
5	1406	1863	*S. cerevisiae*	[15]	0	0	1863
6	1251	1251	*A. thaliana*	[16]	0	0	1251
7	1205	1476	*C. elegans*	[17]	0	1476	0
8	1186	1213	*M. musculus*	[18]	0	0	1213
9	1136	1136	*A. thaliana*	[19]	0	0	1136
10	1101	2269	*C. elegans*	[20]	0	2269	0
11	1043	1365	*M. tuberculosis*	[21]	0	0	1365
12	1041	1041	*A. thaliana*	[22]	0	0	1041
13	865	1533	*C. elegans*	[23]	0	1533	0
14	845	845	*S. cerevisiae*	[24]	0	0	845
15	784	784	*A. thaliana*	[25]	0	0	784
16	735	735	*M. tuberculosis*	[26]	0	0	735
17	724	882	*A. thaliana*	[27]	0	0	882
18	634	634	*A. thaliana*	[28]	0	0	634
19	613	613	Mycobacter sp.	[29]	0	613	0
20	607	661	*C. elegans*	[30]	0	659	2
21	577	577	*A. thaliana*	[31]	0	0	577

Continued on next page

N	Proteins	Annotations	Species	ref.	MFO	BPO	CCO
22	553	884	*C. elegans*	[32]	0	884	0
23	516	5972	*C. elegans*	[33]	0	5972	0
24	503	503	*S. cerevisiae*	[34]	0	0	503
25	498	638	*S. cerevisiae*	[35]	638	0	0
26	479	848	*C. elegans*	[36]	0	848	0
27	465	468	*H. sapiens*	[37]	0	0	468
28	436	436	*A. thaliana*	[38]	0	0	436
29	430	513	*A. thaliana*	[39]	0	0	513
30	413	456	*D. melanogaster*	[11]	0	39	417
31	401	401	*A. thaliana*	[40]	0	0	401
32	392	392	*A. thaliana*	[41]	0	0	392
33	392	639	*C. elegans*	[42]	0	639	0
34	383	917	*C. elegans*	[43]	0	917	0
35	380	380	*A. thaliana*	[44]	0	0	380
36	375	375	*M. musculus*	[45]	0	0	375
37	343	509	*H. sapiens*	[46]	509	0	0
38	338	338	Ddiscoideum	[47]	0	0	338
39	328	328	*A. thaliana*	[48]	0	0	328
40	319	329	*C. albicans*	[49]	1	328	0
41	305	312	*A. thaliana*	[50]	0	0	312
42	290	331	*S. cerevisiae*	[51]	0	0	331
43	285	761	*C. elegans*	[52]	0	761	0

Continued on next page

N	Proteins	Annotations	Species	ref.	MFO	BPO	CCO
44	283	499	*C. elegans*	[53]	0	499	0
45	266	433	*M. musculus*	[54]	433	0	0
46	260	260	*A. thaliana*	[55]	0	260	0
47	258	259	*S. pombe*	[56]	0	259	0
48	244	397	*D. melanogaster*	[10]	0	367	30
49	242	397	*D. melanogaster*	[57]	0	0	397
50	241	263	*A. thaliana*	[58]	0	0	263

The top 50 annotating articles. N: article rank; **Proteins**: number of proteins annotated in this article; **Annotations**: number of annotating GO terms; **Species**: annotated species; **ref.** annotating article; **MFO/BPO/CCO**: number of proteins annotated in the Molecular Function, Biological Process and Cellular Component ontologies, respectively.

Table 3. Sequence Redundancy in Top-50 Annotating Articles

Species	**num. articles**	**num. prot**	**Clusters at 100%**	**% redundancy**	**Mean genes/ cluster**
C. elegans	12	8416	3338	60	3.74
A. thaliana	16	8879	4694	47	3.92
M. musculus	3	4220	2273	46	2.75
M. tuberculosis	2	2351	1702	28	2.22
S. cerevisiae	5	3542	2550	28	2.33
H. sapiens	4	5593	4509	19	2.36
D. melanogaster	3	1217	1003	18	2.17
S. pombe	2	4502	4281	5	2.00

Species: annotated species; **num. articles** number of annotating articles; **num. prot**: number of proteins annotated by top-50 articles for that species; **Clusters at 100%**: number of clusters of 100% identical proteins; **% redundancy**: the ratio between column 3 and column 2: this is the percentage of proteins annotated more than once for a given species in the top 50 articles; **Mean genes/cluster**: the mean number of genes per cluster, for clusters having more than a single gene.

Table 4. Annotation Consistency in Top 50 articles

Species	Ontology	num prot	mean $k_{P,O}$	stdv	stderr	num articles	num t
A. thaliana	cco	1941	0.251	0.328	0.007	15	18
C. elegans	bpo	1847	0.388	0.239	0.006	12	41
D. melanogaster	bpo	76	0.086	0.22	0.025	3	8
D. melanogaster	cco	81	0.068	0.234	0.026	3	5
H. sapiens	cco	167	0.285	0.365	0.028	2	20
M. musculus	cco	807	0.832	0.291	0.01	3	2
S. cerevisiae	cco	744	0.759	0.379	0.014	4	15
B. tuberculosis	cco	532	0.309	0.41	0.018	2	3

Species: annotated species; **Ontology**: annotating GO ontology; **num prot**: number of annotated proteins in that species & ontology that are annotated by more than one paper. **mean, stdv, stderr**: mean number of consistent annotations for a protein in that species and ontology. Standard deviation form the mean and standard error are also provided. **num articles**: number of annotating articles **num terms** number of annotating terms. Annotations by less than 2 articles or two terms (or both) for the same protein/ongology combination have been omitted. Different in paper text

Table 5. Assertion codes used in top-50 papers

N	ECO id	ECO term	Articles
1	ECO:0000160	protein separation followed by fragment identification evidence	25
2	ECO:0000004	cell fractionation evidence	21
3	ECO:0000053	**computational combinatorial evidence**	18
4	ECO:0000249	**sequence similarity evidence used in automatic assertion**	18
5	ECO:0000315	mutant phenotype evidence used in manual assertion	16
6	ECO:0000019	RNAi experimental evidence	15
7	ECO:0000028	**motif similarity evidence**	14
8	ECO:0000112	Western blot evidence	9
9	ECO:0000081	targeting sequence prediction evidence	7
10	ECO:0000083	transmembrane domain prediction evidence	5
11	ECO:0000126	GFP fusion protein localization evidence	5
12	ECO:0000250	**sequence similarity evidence used in manual assertion**	4
13	ECO:0000031	**protein BLAST evidence used in manual assertion**	4
14	ECO:0000044	**sequence similarity evidence**	4
15	ECO:0000104	microarray RNA expression level evidence	3
16	ECO:0000245	**computational combinatorial evidence used in manual assertion**	3
17	ECO:0000015	transposon integration	2
18	ECO:0000128	YFP fusion protein localization evidence	2
19	ECO:0000092	epitope-tagged protein immunolocalization evidence	2
20	ECO:0000007	immunofluorescence evidence	2
21	ECO:0000248	**sequence alignment evidence used in automatic assertion**	1
22	ECO:0000010	protein expression evidence	1
23	ECO:0000231	qRT-PCR evidence	1
24	ECO:0000122	protein localization evidence	1
25	ECO:0000181	*in-vitro* assay evidence	1
26	ECO:0000208	**protein BLAST evidence**	1
27	ECO:0000108	reverse transcription polymerase chain reaction transcription evidence	1
28	ECO:0000062	genomic microarray evidence	1
29	ECO:0000106	Northern assay evidence	1
30	ECO:0000026	nucleic acid hybridization evidence	1
31	ECO-0000068	yeast 2-hybrid evidence	1
32	ECO:0000176	mutant visible phenotype evidence	1
33	ECO:0000324	imaging assay evidence	1
34	ECO:0000079	affinity chromatography evidence	1
35	ECO:0000022	co-purification evidence	1
36	ECO:0000266	**sequence orthology evidence used in manual assertion**	1
37	ECO:0000025	hybrid interaction evidence	1
38	ECO:0000124	RFP fusion protein localization	1

Table caption