Contextual Analysis of Variation and Quality in Human-curated Gene Ontology Annotations

W. John MacMullen

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Information and Library Science.

Chapel Hill
2007

Approved by:

Gary Marchionini

Catherine Blake

J. Michael Cherry

Bradley Hemminger

Barbara Wildemuth

Abstract

W. JOHN MACMULLEN: Contextual Analysis of Variation and Quality in
Human-curated Gene Ontology Annotations
(Under the direction of Gary Marchionini)

Two prospective randomized controlled studies of scientific curators of model organism databases (MODs) were conducted using common document collections to investigate the origins, nature, and extent of variation in curators' Gene Ontology (GO) annotations. Additional contextual data about curators' backgrounds, experience, personal annotation behaviors, and work practices were also collected to provide additional means of explaining variation. A corpus of nearly 4,000 new GO annotations covering 5 organisms was generated by 31 curators and analyzed at the paper, instance, and GO element levels. Variation was observed by organism expertise, by group assignment, and between individual and consensus annotations. Years of GO curation experience was found to not be a predictor of annotation instance quantities. Five facets of GO annotation quality (Consistency, Specificity, Completeness, Validity, and Reliability) were evaluated for utility, and showed promise for use in training novice curators. Pairwise matching and comparison of instances was found to be difficult and atypical, limiting the usefulness of the quality measures. Content analysis was performed on more than 600 pages of curators' hand-annotated paper journal articles used in GO annotation, yielding six types of common notations.

Acknowledgements

"As always, I was driven on by wild expectations."

– Max Perutz[1]

"Somewhere, something incredible is waiting to be known."

– Carl Sagan[2]

"There's nothing as practical as a good theory."

– Kurt Lewin[3]

"We used to think that if we knew one, we knew two, because one and one are two. We are finding that we must learn a great deal more about 'and.'"

– Sir Arthur Eddington[4]

Many people have provided support and guidance to me during the completion of this work, especially my committee: Gary Marchionini, Cathy Blake, Mike Cherry, Brad Hemminger, and Barbara Wildemuth, who helped make the optimistic realistic. This project would have been impossible without the willing participation of the more than thirty scientific curators from many model organism databases who were not just research subjects, but true participants. I am grateful to each and every one of you for investing your time, effort, and expertise in this project. Special thanks are due to Karen Christie and Eurie Hong for their tremendous help in coordinating the annotation studies. Thanks to Sarah Meadows and Sarah Stokes for creating outstanding transcripts. The Annotation of Structured Data research group at UNC SILS (2004-2006) was a stimulating intellectual environment in which many of these ideas were born; thanks to Cathy Marshall, Gary Marchionini, Paul Solomon, Cathy Blake, Tom Ciszek, Xin Fu, Lili Luo, Mary Ruvane, David West, and Megan Winget. For earlier collaborations that led to this work (and leading still to others), thanks to Jean Blackwell, Sheila Denn, David

Fenstermacher, Morgan Giddings, Stephanie Haas, Carol Jenkins, Margaret Moore, Mary Parmelee, Alex Tropsha, and K.T.L. Vaughan. Thanks to my students for helping to keep teaching and research linked. The SILS administrative staff is a tireless force to be reckoned with; special thanks are due to Susan Thomas for grant management.

Research takes time, and time requires money. For supporting this work and the work that led to it, I thank the National Library of Medicine (under Individual Biomedical Informatics Fellowship F37 LM009194); Cathy Marshall and Microsoft Research; the UNC School of Information and Library Science (for the Lester Asheim Scholarship, the Margaret Ellen Kalp Memorial Fellowship, and several tuition, research, and teaching awards); the UNC Program in Bioinformatics and Computational Biology for a one-year fellowship; the UNC Health Sciences Library for a one-year fellowship; and, for travel support, the Stanford University School of Medicine, Department of Genetics; the *Saccharomyces* Genome Database; the Gene Ontology Consortium; the Medical Library Association (MLA); the i-Schools Project; and the American Society for Information Science and Technology (ASIS&T) Science and Technology Information (STI) SIG.

For many other kinds of support, I am unquantifiably grateful to Kevin Borek, Ron T. Brown, Amanda Burgess, David Carr, Melissa Cragin, Stefanie Hollmichel, Haven Kimmel, Sharron MacMullen, Rikki Marcin, Rob Mather, Sarah Meadows, Sarah Messer, Susan Pardue, Tim Shearer, Kristina Spurgin, Cassidy Sugimoto, John Svara, Karin Thogersen, K.T. Vaughan, Donna Vincent, and Megan Winget.

1. Quoted in Judson (1996: 534)
2-4. Attributed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND TERMS

EC	Evidence Code
Evidence source	An entity from which claims are drawn in order to create GO annotations; these are most often individual scientific journal articles.
GO	Gene Ontology
GO annotation	A set of one or more individual GO annotation instances created in relation to a particular evidence source.
GO annotation element	One of the 15 components of a formal GO annotation instance, currently defined by the GO Consortium documentation as containing 11 required and 4 optional elements; a database-independent subset of 9 of these elements was used for the two studies in this project.
GO annotation instance	The set of GO annotation elements created in relation to a particular gene product from an evidence source.
MMS	Multi-MOD study
MOD	Model organism database
SMS	Single-MOD study

1. Introduction

1.1. Problem statement

The wide variety of information in databases for biological model organisms has great potential utility for biomedical researchers, but only if it is of high quality. Human-curated annotations made in model organism databases are viewed as important links between an organism's underlying genomic data and the experimental literature. Gene Ontology (GO) annotations in particular are expected to provide significant value in reducing disciplinary fragmentation, and facilitating cross-organism information integration. As an organism-independent set of controlled vocabularies, scientists could use GO to cut across disciplinary boundaries to discover information relevant to their specialty. However, a common question associated with manually-curated knowledgebases (such as GO, model organism databases (MODs), and even PubMed/MEDLINE) is to what degree variation in human curators' annotations affects the overall quality of information in the resource. Information quality is a complex concept, with many facets that have interdependencies. While annotation quality is of great importance to the genomics community, very little is known about the types and amounts of GO annotation variation in MODs. Dolan, et al. (2005) state that:

> Since differences in application of annotation standards would dilute the effectiveness of comparative analysis, methods for assessing annotation consistency are essential. The development of methodologies that are broadly applicable for the assessment of GO annotation consistency is an important issue for the comparative genomics community (i136).

1.2. Research questions

A broad question the opening problem statement implies is, “How and why do human curators differ in their GO annotation processes and outcomes?” This project investigated this through three more specific questions:

1. *How significantly do curators differ in annotation outcomes?* This question was investigated by conducting individual and consensus annotation experiments and comparing differences in curators’ formal GO annotations against a parameterized model of annotation quality, and through the use of statistical analysis.

2. *Do differences in curators’ educational-, training-, and research backgrounds influence their GO annotation performance?* This question was investigated by comparing the formal GO annotation data from the individual and consensus annotation experiments with demographic data collected from questionnaires, and in individual semi-structured interviews.

3. *Do differences in curators’ personal annotation behaviors influence their GO annotation performance?* This question was investigated by comparing formal GO annotation data from individual and consensus annotation experiments with data on personal annotation behaviors collected in individual interviews, observations, and artifact analyses.

While this work focuses on individual differences in human curation, many other related questions are not addressed by this project. Lack of variation should not necessarily be viewed as equating with high-quality annotations (or vice-versa); for example, curators could be in perfect agreement but be incorrect. We may be interested in comparing not only individual curators' annotations against human-created reference annotations, but also comparing them to the performance of automated and semi-automated approaches to curation.

1.3. Project context

Biomedical database curation is a quickly-evolving role within the biomedical research community. 'Curation' is the process of collecting, organizing, synthesizing, and presenting biomedical data and information in online environments in order to increase their utility for end users. The underlying data and information are generally a combination of biological and biochemical data, such as gene- and genome sequences or protein structures, and information extracted from the scientific literature. Through the use of controlled vocabularies such as the Gene Ontology (GO), curators integrate these forms of information to bring together what is known about particular organisms or diseases. Curators are typically Ph.D-trained scientists who use their subject-matter expertise in conjunction with additional training in controlled vocabularies and information systems to make large quantities of data and information more accessible by, and useful for, researchers. Sections 3.6 and 3.7 describe in more detail the characteristics of the curators involved in this study.

1.4. Prior work

This section reviews the existing literature on annotation-related work relevant to this study. Since no studies of the process and context of GO annotation had been performed prior to this work, other work is reviewed from different domains that used methods similar to those employed here.

1.4.1. Introduction

Social and historical studies of biological science and scientists often focus on the culture and practice of science as socially-constructed epistemic activities (e.g., Knorr-Cetina, 1999; Latour, 1987; Latour and Woolgar, 1986), on the history of specific scientists and trends in science (e.g., Angier, 1999; Judson, 1996; Kay, 1996 & 2000; Lindee, 2005), and, increasingly, on particular organisms and the specific tools and materials of experimentation used to understand them (e.g., Ankeny, 1997; Bowker, 1995; Burian, 1993; de Chadarevian, 1998; Clause, 1993; Creager, 2002; Fujimura, 1987; Hilgartner, 1995; Hine, 2006; Kellog & Shaffer, 1993; Kohler, 1994; Leonelli, forthcoming; Rader, 2004). In biomedical informatics and information and library science, the scope is often much narrower, frequently focused on end-user information seeking and use behaviors (e.g., Bartlett & Toms, 2005; Brown, 2003; Detlefsen, 1998; Wildemuth, 2004), including interdisciplinary collaboration (e.g., Palmer, 1999), as well as such technical facets as interface design and usability (e.g., Rose et al., 2005), vocabulary and ontological issues (e.g., Whetzel, et al., 2006), retrieval performance (e.g., Camon et al., 2005; Cohen & Hersh, 2006), data- and text mining, and knowledge discovery approaches (e.g., Shatkay & Feldman, 2003; Srinivasan, 2004), and visualization of large data sets (e.g., Wong & Cartwright, 2005). Bibliometrics and scientometrics have been a longstanding but often separate approach to the measurement of the growth, productivity, and impact of the knowledge produced by science in the form of publications, and the relationships among them. (e.g., Price, 1986; White & McCain, 1989). Those approaches can be useful in understanding the structures of, and relationships between, multiple biomedical databases that are linked both physically and conceptually (e.g., MacMullen 2005b).

This project takes a different approach, and focuses on people as actors in an information management context, in an attempt to understand how human information interactions affect the development and use of information systems for biomedical research, particularly the phenomenon of annotation in model organism databases, and the unique features and challenges human scientific curators face when creating Gene Ontology (GO) annotations.

1.4.2. Structured data annotation in biomedicine

While personal or informal annotations similar to those in other domains are made in many contexts in biomedical research, formal annotation in the specific context of biomedical research databases is manifested in very different forms, as illustrated below. The phenomenon, however, is conceptually the same: assertions are extracted from some evidence source (e.g., a scientific article) and linked to an underlying object. To use an example relevant to the work proposed below, scientific curators making Gene Ontology annotations to a particular organism examine evidence sources (e.g., scientific articles) to identify entities (e.g., genes) and assertions (e.g., ‘gene *x* was found to be of type *t*’) for use in creating formal annotations (e.g., ‘gene *x* has molecular function *f*, participates in biological process *p*, and is found in cellular component *c*’). As with the earlier concepts, this process creates both intellectual linkages and linkages between physical artifacts.

The ‘annotation’ that results from this process is quite different, however, from the colloquial instance of textual marginalia. While scientific curators may in fact make those types of personal annotations on an evidence source when extracting assertions, the ultimate annotation entity that is created is a highly structured, distilled, formalized, unambiguous artifact that is intended not only for public consumption, but is expected to be computationally manipulatable as well. Use of the term ‘annotation’ varies significantly even within biological sub-domains, such as molecular biology, and research that on its face appears to be about the types of annotation quality this project is interested in evaluating are often quite different. Ouzounis and Karp (2002), for example, define annotation as “the process by which structural or functional information is inferred for genes or proteins”. Ussery and Hallin (2004) implicitly define annotation quality as gene-finding accuracy, in which factors such as the methods used to infer homology (such as sequence similarity algorithms) can affect the quality of the annotation.

An additional difference is that, in many cases, controlled vocabularies of various types are employed by annotators rather than free text. In the case of GO annotations, the large, complex three-aspect system necessitates software tools to search and browse the vocabularies to locate appropriate terms to use when creating annotations.

The term ‘annotation’ frequently occurs in reference to nucleic or amino acid sequence data, as any information that is overlaid to provide context, disambiguation, and linkage of entities. In MEDLINE there were 834 papers with the term ‘annotation’ in the title, and 3,389 with the term in the title or abstract, as of 2007-03-09.

While paper-based exist in biomedical contexts (especially in the clinical environment; see, e.g., Chapman & Dowling, 2005), this project is focused primarily on annotations that occur in structured online information resources, particularly model organism databases (MODs) such as those for the mouse, yeast, fruit fly and roundworm.

Although a substantial body of research exists on annotation in the humanities and everyday life contexts (see, e.g., Marshall 1997, 1998), very little research has characterized annotation in biomedical contexts. Apart from some basic statistics and annual reports from the databases themselves, almost no work has independently explored annotations in any of the three facets of biomedical annotation in which we are interested: as a process, as an artifact, and as a unit of knowledge. As noted in MacMullen (2005a), “[t]he majority of annotation-oriented research in the biomedical domain is focused on the problem of automatically deriving and assigning high-quality annotations to large databases of gene and protein sequences in order to understand single genes or organisms, and to aid in the recognition of cross-organism similarities of multiple molecules.”

The concept of annotation quality in biomedical databases, and particularly MODs, has been studied very little. The research on which this project builds includes Stein’s (2001) review that described the concept of biomedical annotation and its goals, and proposed four generalized examples of annotation workflow models. As part of the BioCreAtIve project (Hirschman, et al., 2005), Camon, et al. (2005) and Colosimo, et al. (2005) investigated the feasibility of automated classification for the functional annotation of proteins in the UniProt database using GO terms, and performed the first inter-annotator agreement study of manual GO curation. Camon, et al. provided a general description of the manual GO curation process, but not from the perspective of individual model organism databases, or how that process relates to other curation activities. Colosimo, et al. (2005) performed a very small inter-annotator reliability test (3 annotators, 89 documents), and seemed to equate agreement with how ‘hard’ (difficult) particular organisms are to annotate. Their task involved the extraction of genes from articles, not GO annotation consistency specifically. Dolan, et al. (2005) developed a method for assessing the consistency of GO annotations made to orthologous genes across manually-curated joint mouse-human GO annotations. However, this work focused on the ultimate inconsistencies, not on the processes that led to them, or on the human components of annotation. The results were limited by the small set of annotations available for comparative purposes, the use of 'Slim' versions of the GO ontologies (discussed in section 4, below), and minimal comparisons at very granular levels of detail. They did not discuss how inconsistencies in gene identification or Evidence Code assignment by different curators may impact their measures, nor how to rationalize cases where the numbers of annotation instances differ across curators, a major issue in evaluating consistency, as discussed in the Results section below. Apart from resource-level overviews (such as Chen et al., 2005, or Christie et al, 2004), no other detailed characterizations of annotation in biomedical databases have been found. Rother et al. (2005) compared annotations in fifteen protein databases to assess coverage, linkages, and overlap, using the unique approach of a phylogenetic tree model to assess similarity of content, but while that method could possibly be useful to assess GO annotation quality at a high level, it is not granular enough to be useful to this project. Other research on annotation quality has focused on the ontological consistency of the underlying GO infrastructure (e.g., Yeh, et al., 2003), or on the general concern about the quality of automatically-assigned annotations in biomedical databases (see, e.g., Karp, 1998; Devos & Valencia, 2001; Green & Karp, 2005), rather than on the content or the creation of the annotations themselves. While documentation and reports of individual resources are valuable, such as those published in the annual database issue of Nucleic Acids Research, they do not provide cross-resource comparisons of facets such as overlap or level of coverage, or linkages within and between resources.

1.4.3. Gene Ontology annotations

Gene Ontology annotations are standardized formal knowledge structures derived from claims extracted by human scientific curators from evidence sources (typically in the form of experimental journal articles). Each evidence source can contain zero to many claims relevant to the underlying organism, or to GO, or to both. Figure 1 provides a high-level illustration of a generic GO annotation process. The process and actual workflows may vary by model organism, and by individual curator.

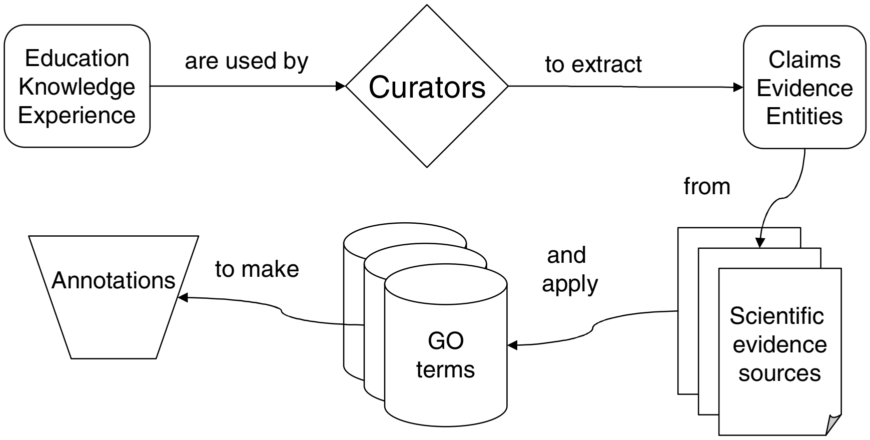


Figure 1. Generic, high-level schematic of the GO annotation process

Formal GO annotations follow a common structure across organisms, in the form of:

```
object | qualifier | ontology | GOID | GO term | evidence code | reference | with/from
```

where:

- *object* is the gene or gene product;
- *qualifier* a is modifier that affects the interpretation of an annotation depending upon which of its values are used: 'NOT', 'contributes_to', and 'colocalizes_with';
- *ontology* is which of the three GO ontologies (or 'aspects') a particular claim in an evidence source is associated with (molecular function, biological process, cellular component);
- *GO term* and *GOID* are the most specific applicable term from one of the three GO ontologies, and its unique identifier;
- *evidence code* is the type of evidence supporting the claim (e.g., a direct assay or genetic interaction);
- *reference* is the specific source of the claim (article, database entry, etc.); and
- *with/from* is a contextual field that relates the *object* to other objects.

The *object*, *ontology*, *GO ID/term*, *evidence code* and *reference* elements are mandatory for a valid GO annotation (GO, 2007). As noted above, one evidence source (and one gene) may have multiple annotatable claims, and a single claim may have multiple instances of annotations for the function and process aspects (e.g., one gene may participate in multiple biological processes). When this project uses the term 'GO annotation' (or simply 'annotation') in reference to the data and measures described below, it refers to the collection of instances associated with a particular evidence source, while an 'annotation instance' or simply 'instance' refers to one stand-alone component of an annotation, namely one set of values for the mandatory GO elements. That is, if an evidence source yields a claim about a gene that a curator annotates as having a particular molecular function, as participating in a particular biological process, and localized in a particular cellular component, this project views that annotation as being comprised of three individual instances. Table 1 provides an example annotation. From this paper, four genes are each annotated once to both the Process and Component aspects, for a total of 8 annotation instances. While this is a straightforward example, section 3.2.1 describes the complexity encountered while attempting to quantify the annotation instances generated during the studies from this project.

Table 1. GO annotations from Takayama Y, et al. (2003) [PMID: 12730134]

Object	Ontology	GO ID	GO Term	Evidence	With/From
PSF1	Process	GO:0006261	DNA-dependent DNA replication	IMP	
PSF1	Component	GO:0000811	GINS complex	IPI	PSF2, PSF3, SLD5
PSF2	Process	GO:0006261	DNA-dependent DNA replication	IGI	SLD5
PSF2	Component	GO:0000811	GINS complex	IPI	PSF1, PSF3, SLD5
PSF3	Process	GO:0006261	DNA-dependent DNA replication	IGI	PSF1
PSF3	Component	GO:0000811	GINS complex	IPI	PSF1, PSF2, SLD5
SLD5	Process	GO:0006261	DNA-dependent DNA replication	IMP	
SLD5	Component	GO:0000811	GINS complex	IPI	PSF1, PSF2, PSF3

Source: GO sample yeast annotations (GO Consortium, 2006).

The human curatorial staffs of MODs have created large collections of literature annotated to GO. For example, Table 2 shows descriptive statistics for the quantities of GO annotations in the *Saccharomyces* Genome Database (SGD) as of 2006-01-29.

Table 2. Descriptive statistics for SGD GO annotations

	Qty.	%
Total annotations in orf_geneontology.tab	5,806	-
Total unique GO annotations	1,907	100.00
- Unique GO Molecular Function annotations	986	51.70
- Unique GO Biological Process annotations	672	35.24
- Unique GO Cellular Component annotations	249	13.06

Source: MacMullen (2006a)

1.4.4. Indexer consistency

In information and library science, indexing consistency has been a topic of interest for many decades, particularly the question of whether consistency influences information retrieval performance. Rolling (1981), following Cooper (1969), differentiates indexing *quality* ("whether the information content of an indexed document is accurately represented"), and *effectiveness* ("whether an indexed document is correctly retrieved every time it is relevant to a query"), from *consistency*, asserting that the third is used as an imperfect surrogate for the first two, because they are both difficult and costly to measure (69). (Optimal consistency here would seem to need to be between the indexer and the searcher, not between indexers, but this does not seem promising; Iivonen (1995), for example, found significant inconsistency between searchers' queries for the same topics, and Murphy, et al. (2003) found inconsistency between authors, indexers, and searchers.) Funk and Reid (1983) note that although some challenges have historically been raised to the idea that consistency is an informative measure, more recent work has demonstrated a positive correlation between high inter-indexer consistency and high information retrieval effectiveness.

Funk and Reid's study assessed indexing consistency in MEDLINE. They computed the inter-indexer consistency of 760 articles that had been indexed twice, and found that headings "were applied more consistently to central concepts than to peripheral points", and that "[w]hen subheadings were added to a main heading, consistency was lowered" (176). They also found that terms from certain MeSH categories appeared more frequently in high-consistency articles than terms from other categories. Hurwitz (1969) found in part that very granular indexing vocabularies contributed to inter-indexer inconsistency because different but near-synonymous terms were often chosen by different indexers of the same text. Sievert and Andrews (1991) found that in a database of document abstracts indexed by a very small vocabulary, “indexing matched completely almost half of the time, and did not match at all almost half of the time”, with consistency decreasing as the number of terms assigned per article increased. However, Schultz, Schultz, and Orr (1965) compared indicia supplied by authors for their own papers with those assigned to the same papers by a group of readers and found substantial agreement compared to the control group of terms extracted from document titles. In a small study, Boles (1989) found that the presence or lack of domain expertise on the part of indexers did not necessarily lead to better or worse indexing performance.

Both Funk and Reid (1983) and Olson and Wolfram (2006) note that consistency studies have shown that indexers often have higher agreement about what the main or “core” topics of a text are than those that are more peripheral. Olson and Wolfram argue that this suggests that the choices might adhere to a power law that can predict the distribution of topics. If the distribution of topics indexed is predictable it might be possible to minimize the variation or to develop interfaces for searching databases that would take that distribution into account. It could also be that the variation is a positive factor [1-2].

Cooper (1969) was a critic of the idea that indexer consistency resulted in positive effects on retrieval performance, arguing that in fact the opposite could occur: as indexers became more consistent with each other, retrieval performance could be diminished if indexer-requester consistency was low (this was prior to IR systems with thesauri or full-text searching). While commonalities clearly exist between document indexing and GO annotation, it is useful here to distinguish their differences. As an example, consider the document from which Table 1 is drawn in section 1.4.3 above. The MeSH terms assigned to the article are shown below in Table 3. Consider the levels of conceptual granularity between index terms such as ‘Amino Acid Sequence’, or ‘Fungal Proteins/*metabolism’ and GO terms such as ‘DNA-dependent DNA replication’ and ‘GINS complex’. While natural language has significant variation and redundancy, information systems have features such as thesauri that are used to compensate for mismatches between indexers’ term selections and searchers’ query terms in order to return sets of relevant results. But, with the exception of some adjacent terms, the process and vocabulary of GO annotations is less tolerant of both vagueness and error. If a GO term is too vague, it is not useful, and if it is incorrect, it could actually be harmful, or a least wasteful, in terms of resources such as time, money, and material, that are mis-allocated due to an incorrect annotation. In other words, while most believe that there is not one ‘true’ set of index terms for a document, we are much closer to that in the case of GO annotations.

Table 3. MeSH terms assigned to Takayama et al. (2003) [PMID: 12730134]

- Amino Acid Sequence
- Cell Cycle Proteins/*metabolism
- DNA Replication/*physiology
- DNA, Fungal/genetics
- DNA-Binding Proteins/*genetics/metabolism
- Fungal Proteins/*metabolism
- Molecular Sequence Data
- Saccharomyces cerevisiae/ *genetics/metabolism
- Saccharomyces cerevisiae Proteins/*genetics/metabolism
- Sequence Alignment

* = MeSH Major Topic

The process of creating GO annotations is quite different from topical indexing, both in concept and in practice. The goals of GO annotation include the extraction and formalization of specific claims from the literature, and the ability to link annotations of biological data back to the primary literature, while indexing is focused instead on providing topical access points into the literature – a focus on ‘aboutness’ rather than specific claims. Thus, while this project was informed by prior work on indexer consistency, different approaches were needed, as described in the next section, to address quality evaluation in the unique environment of GO annotation.

2. Research Design and Methods

This project used multiple research methods and data collection points in an attempt to ‘triangulate’ (‘surround’ is perhaps a more accurate term in this case) the issues of curation and annotation behavior described above. The study designs described below included both experimental approaches (empirical individual and consensus annotation experiments), and contextual approaches (individual interviews, focus groups, observations, and task- and artifact analyses). The experimental approaches were prospective, in that they generated new GO annotations from evidence sources that had not been curated until that time. Prior to the studies, a set of five GO annotation quality facets were identified, and corresponding metrics were created (as described in detail in section 2.9). The newly-generated GO annotations were used to assess the utility of these metrics.

Two separate but related studies were conducted: the first in June 2006 with 10 curators from a single model organism database (MOD). The second study occurred in July 2006 with 23 curators from 11 databases prior to and during the Gene Ontology Consortium meeting and Annotation Camp (GO Camp, 2006). These studies are hereafter referred to as the ‘single-MOD study’, or ‘SMS’, and the ‘multi-MOD study’, or ‘MMS’, respectively.

To generate the core GO annotation data for the quality measurement studies, two types of annotation activities were performed by scientific curators: individual experiments, and co-curator (or consensus) experiments. Individual experiments to generate GO annotations for evaluation consisted of expert database curators individually annotating previously uncurated scientific articles, in natural work environments, with a predefined level of redundancy for each article to enable multiple pairwise comparisons and the subsequent creation of consensus annotations by curator pairs who annotated the same articles. The variations of this model for the two studies are described below.

For the individual experiments, curators were instructed to perform their normal annotation work processes, but to not seek assistance from other curators, if doing so is a normal action, since that could potentially influence the final annotation and obscure differences in annotations of papers shared by curators, especially if curators annotating shared papers discussed their annotations prior to completing them. The consensus-formulation process that followed the individual annotation step enabled curators to discuss their annotations with fellow curators and make adjustments that are reflected in the resulting sets of reference annotations.

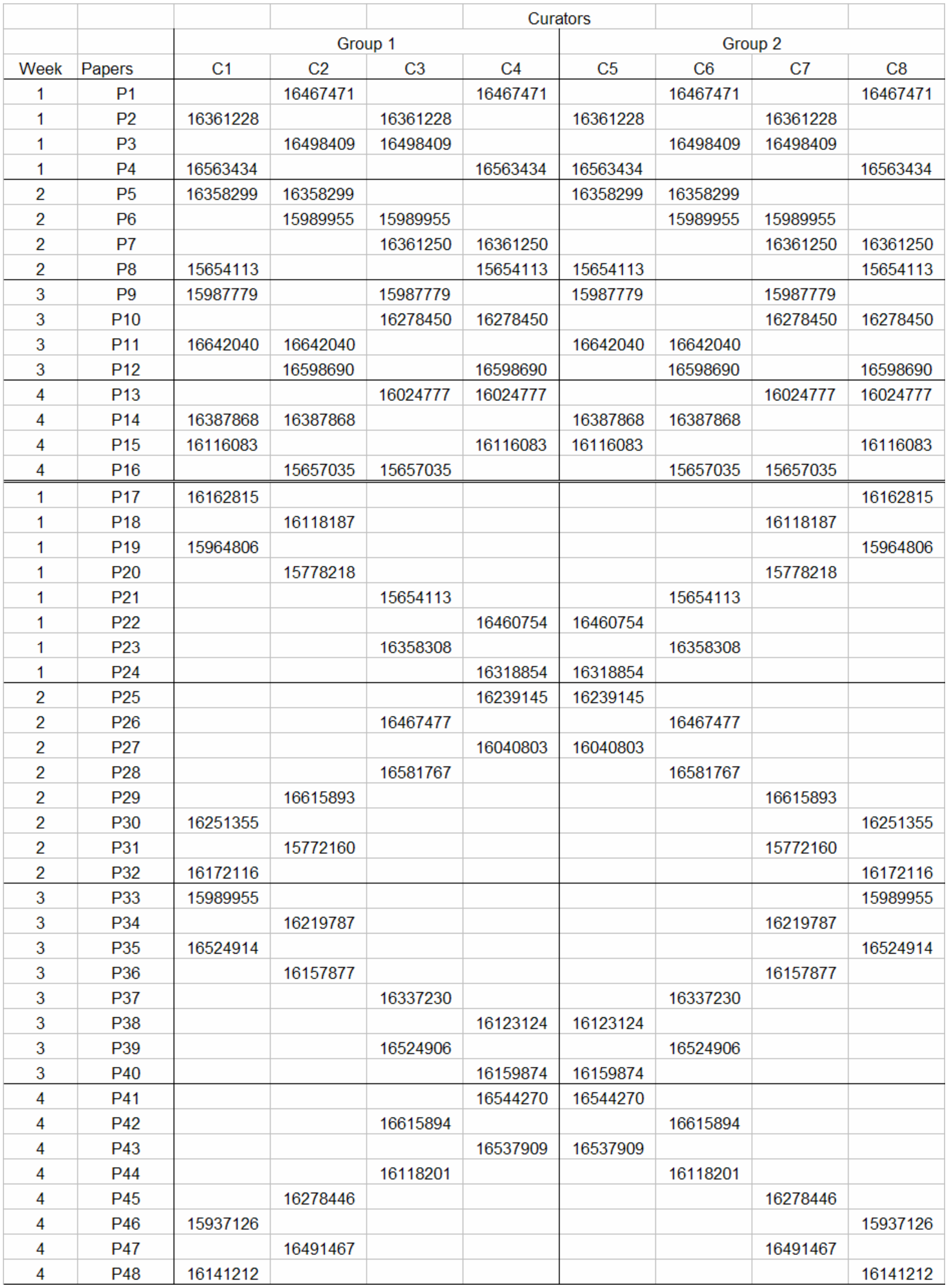
2.1. Individual annotation experiments

2.1.1. Single-MOD study

A corpus of 48 scientific articles related to the organism on which the MOD is focused was identified, with at least 16 that were predetermined to contain one or more assertions amenable to GO annotation (see section 2.4.2 for corpus construction details). The articles are papers that would normally be annotated by the curators, but were uncurated prior to this study. Eight curators from the MOD were randomly assigned to two equal groups of four, and then assigned 16 articles from the corpus. Each curator was assigned four articles per week for four weeks. The curators were instructed to fully curate each paper as they normally would. ‘Full’ curation in this MOD’s context includes other kinds of annotations in addition to GO, such as phenotypic characterization, and the selection of broad index terms selected from a small controlled vocabulary and applied to each paper (see MacMullen 2006a). Since each paper was annotated in full, follow-on studies should be able to use this additional data to examine variation in annotations of types other than GO.

Figure 2 provides a graphical view of the design for this part of the study, with the articles listed as PubMed unique identifiers (PMIDs). As described in more detail in section 2.4.2, sixteen of the articles were annotated by four curators each, and the remaining 32 articles were annotated by 2 curators each. Section 3.1.1 describes in detail the observations obtained from this design.

The two-group schema facilitated pairing for the production of consensus annotations, as described in the following section. The double-blind assignment process for the SMS study (see section 2.3.2) should have minimized experimenter bias. Progressive effects or carryover effects were not expected for multiple reasons. First, the participants performed normal work tasks, not tasks that were new to them. These are the same tasks they have performed tens, hundreds, and even thousands of times, so learning effects are unlikely. Second, the duration of the experiment (about four weeks) should have been long enough (and a low enough workload per week) to avoid performance degradation due to fatigue, yet brief enough to not have learning effects that might be expected to arise if the study occurred over a period of months.



		Curators							
		Group 1				Group 2			
Week	Papers	C1	C2	C3	C4	C5	C6	C7	C8
1	P1		16467471		16467471		16467471		16467471
1	P2	16361228		16361228		16361228		16361228	
1	P3		16498409	16498409			16498409	16498409	
1	P4	16563434			16563434	16563434			16563434
2	P5	16358299	16358299			16358299	16358299		
2	P6		15989955	15989955			15989955	15989955	
2	P7			16361250	16361250			16361250	16361250
2	P8	15654113			15654113	15654113			15654113
3	P9	15987779		15987779		15987779		15987779	
3	P10			16278450	16278450			16278450	16278450
3	P11	16642040	16642040			16642040	16642040		
3	P12		16598690		16598690		16598690		16598690
4	P13			16024777	16024777			16024777	16024777
4	P14	16387868	16387868			16387868	16387868		
4	P15	16116083			16116083	16116083			16116083
4	P16		15657035	15657035			15657035	15657035	
1	P17	16162815							16162815
1	P18		16118187					16118187	
1	P19	15964806							15964806
1	P20		15778218					15778218	
1	P21			15654113			15654113		
1	P22				16460754	16460754			
1	P23			16358308			16358308		
1	P24				16318854	16318854			
2	P25				16239145	16239145			
2	P26			16467477			16467477		
2	P27				16040803	16040803			
2	P28			16581767			16581767		
2	P29		16615893					16615893	
2	P30	16251355							16251355
2	P31		15772160					15772160	
2	P32	16172116							16172116
3	P33	15989955							15989955
3	P34		16219787					16219787	
3	P35	16524914							16524914
3	P36		16157877					16157877	
3	P37			16337230			16337230		
3	P38				16123124	16123124			
3	P39			16524906			16524906		
3	P40				16159874	16159874			
4	P41				16544270	16544270			
4	P42			16615894			16615894		
4	P43				16537909	16537909			
4	P44			16118201			16118201		
4	P45		16278446					16278446	
4	P46	15937126							15937126
4	P47		16491467					16491467	
4	P48	16141212							16141212

Figure 2. Experimental design schema for SMS study with paper assignments

Reliability pilot study

During the single-MOD study, two curators who did not participate in the main study were observed re-annotating two papers each had curated 6- and 12 months earlier in order to pilot-test an approach to evaluating the Reliability facet of annotation quality as defined in section 2.9. Reliability is essentially a measure of 'intra-annotator' consistency. That is, rather than comparing two or more curators' annotations of the same paper, we were interested in whether a single curator makes different annotations to the same paper when he or she annotates it a second time (i.e., a repeated-measures approach with time as the independent variable). The concurrent verbal report (or "think-aloud") protocol was used to elicit tacit knowledge and unobserved facets of the annotation process by having a curator verbalize decisions and actions while he or she performed an actual annotation task. Each participating curator selected two papers that the other had previously curated. This approach was intended to minimize bias that could be introduced by having curators select which of their prior papers to annotate, especially the risk of seeing the specific prior annotations. For the same reason, the curators were also instructed to not use the SMS database while creating their annotations. Appendix A6 provides the complete citations for the four papers re-annotated in the Reliability studies. Additional data were collected during the individual interviews with these curators about any personal or workflow changes or experiences in the time between original annotation and re-annotation that may have influenced the outcome (e.g., training received, different work process or tools used). The curators' original and repeated GO annotations for the papers are compared in section 3.3.4 as a part of the annotation quality facet analysis.

2.1.2. Multi-MOD study

A corpus of 10 scientific articles was selected by expert curators, as described in section 2.4.3. Two papers related to each of five different MODs were chosen by representatives from each MOD. All 10 articles were curated for GO annotations by two curators from each represented MOD (a total of 10 people), and by 13 additional curators from other databases who agreed to participate in the study.

While this study had a similar overall design as the single-MOD study, here both the curator pool and article corpus are heterogeneous in terms of model organism representation. The main research question in this case was: Does a curator's knowledge of an organism significantly influence the quality of the annotations s/he produces?

2.2. Consensus annotation experiments

Subsequent to the individual annotation task in both the SMS and MMS, curators were paired and asked to arrive at a single consensus annotation, based on their two individual annotations. In the SMS, the pairings were determined by the groups to which the curators had been assigned, while in the MMS, the two representatives from each of the participating MODs were paired. For the consensus annotation process, curators were not given any guidance or training as to how the rationalization process should be carried out, other than the restriction that each of their individual annotations must have been completed prior to the occurrence of any consensus discussions. Since in practice, particularly in the larger curation community, a formal GO annotation training component may not be available, the decision was made to keep this task as natural as possible. However, the study PI observed a number of the consensus discussions during both studies, and collected data on how the curators may have differed in their approaches to consensus formation. (See section 2.7 for details of the observation protocol.)

The consensus annotations function as reference standards, and could possibly be used for evaluating the performance of automated annotation tools. The consensus process employed was similar to Figure 1 in Hripcsak & Wilcox (2001), as modified below in Figure 3. Rather than treating either the individual or consensus annotation processes as black boxes, the data collection and analysis methods described in the following sections were used to acquire additional contextual information to assist in understanding how annotations are made, and why variation exists.

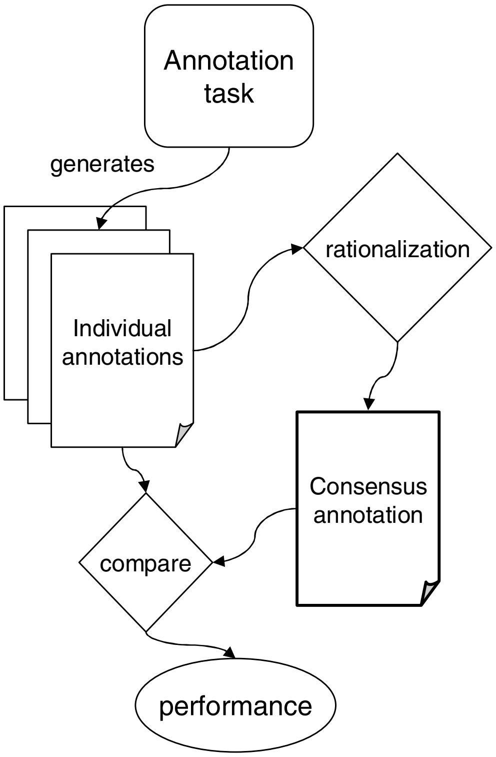


Figure 3. Construction of a reference annotation from individual annotations

2.2.1. Single-MOD study

For the consensus annotation component of this study, the 16 papers with four-fold coverage were used to generate two consensus annotations per paper, with the pairings determined by group assignment. For example, in Figure 2 above, paper P1 was annotated by curators C2 and C4 from Group 1, and C6 and C8 from Group 2, yielding curator pairs {C2,C4} and {C6,C8}. This is one reason two groups were defined in the design. The consensus annotations provide additional data points for comparison with the individual annotations, and served as the ‘gold standards’ to evaluate individual performance on the Consistency quality facet. The single-MOD leadership plans for the two curator pairs for each paper to eventually create a joint consensus annotation from the two pairwise consensus annotations, which will be made available to the larger GO community as ‘reference annotations’. As of the date of this work, those rationalizations had not been finalized.

2.2.2. Multi-MOD study

The consensus pairs for the MMS study consisted of the two representatives from each participating MOD, for a total of 9 consensus annotations per paper. (While there were 23 participants, in one case a MOD’s senior curator and three intermediate-level curators participated, and rationalized all four of their individual annotations into a single consensus annotation. There was also one instance where one participant was the only representative of a MOD, and so no consensus annotations were made for that database.) The consensus annotations generated by the MOD representatives served as the ‘expert’ reference annotations against which the other individual and consensus annotations were compared.

2.3. Participant selection and group assignment

2.3.1. General

Model organism database curators are typically current or former scientists with Ph.D-level training in molecular biology, biochemistry, or in a related area. The majority of their work time is spent curating literature and the different resources contained in their respective MODs. While this is an accepted and growing role in biomedicine, the total cross-MOD population of scientific curators numbers perhaps fewer than 100. Not all curators do GO annotation as a part of their job role. Some MODs have curators who specialize in certain types of annotation, while in other MODs, each curator may perform a greater variety of tasks. Taking a random sample of GO curators, particularly from one MOD, is not feasible, so essentially ‘convenience samples’ were used for the studies described here, although in both cases, the sample represented much, if not most, of the overall population from each MOD. Demographic data on education, training, and experience were collected from curators, as described in later sections, to check for attributes that may influence annotation behavior and outcomes. However, the creations of a uniform pool of curators was not the goal; the project is interested in investigating individual differences, not removing them.

2.3.2. Single-MOD study

The total population of the MOD’s curators is 14; the sample that participated in this study consisted of the 10 who were able to participate in the MOD's bi-annual group meeting in June 2006. This group included 4 off-site curators and 4 local curators. Of the four off-site curators, three are located on the east coast and one is in California. All work from home full-time. Two additional local curators participated in a separate arm of the study to gather data to test the Reliability quality facet (described in section 2.1.1).

The study was designed as two groups of four curators each. Curators were assigned randomly to groups, and papers were assigned randomly to curators. The two curators who participated in the Reliability facet study were involved in discussions of the research design for the main arm of the study, in selection of papers for annotation, and in assignment of curators and papers. While the characteristics of the four non-participating curators is not known, there is no reason to expect that the outcomes from the experiments would be significantly different had they participated.

When the list of participating curators was finalized by the MOD leadership, each curator was randomly assigned a curator ID by MOD staff so that the study PI would not know the curators' identities prior to the completion of the data collection and analysis. The study PI randomly assigned the papers to the curator IDs before sending them to the MOD staff, to avoid any influence by the staff on which curators received which papers. The linkage of curator IDs with names, and the subsequent joint analysis of the annotation data with the contextual data, did not occur until the individual analyses of each of those data were complete. This ensured that the study PI had no knowledge of which curator produced which data.

Participants were instructed to avoid discussing their assigned papers or their annotations with other curators prior to completing the individual and consensus annotations to avoid influencing the results or converging on the same annotations.

2.3.3. Multi-MOD study

The core pool of 10 curators for the MMS was composed of the expert curators who participated in the Gene Ontology Consortium meeting, and the GO Annotation Camp. The meeting organizers invited their participation based upon the availability of two curators from each participating MOD (to enable the creation of consensus annotations), and the ability to attend the meetings. The five organisms represented in the corpus of papers for the experiment were *Arabidopsis thaliana* (a flowering plant) *Caenorhabditis elegans* (a nematode worm), *Mus musculus* (common mouse), *Saccharomyces cerevisiae* (yeast), and human. An additional 13 curators participated from databases other than the 5 that contributed papers for curation, and represented organisms such as *Dictyostelium discoideum* (mold), *Drosophila melanogaster* (fruit fly), and *Rattus norvegicus* (rat). Two curators who participated in the single-MOD study also participated in the multi-MOD study as the experts representing their MOD.

2.4. Document corpus construction and article assignment

2.4.1. General

Since the two studies were intended to be naturalistic, it was important that the scientific articles annotated by the curators were systematically selected, in order to be representative of the range of papers they typically encounter in their normal work. In selecting the articles for the two studies as described below, some considerations included document type (e.g., experimental article, review article, book chapter), the domain of the paper (e.g., genetics, biochemistry), and the existence of content amenable to GO annotation.

2.4.2. Single-MOD study

For the main portion of the SMS, 48 journal articles were selected. The inclusion / exclusion criteria were:

- Experimental papers only, no review articles, book chapters, or other non-experimental papers ('experimental' here refers to articles which describe one or more experiments carried out by the authors);
- Each article in the four-fold subset (defined below) must contain 1 or more assertions that are annotatable to GO;
- Papers must not have been previously curated by a MOD curator.

Two expert curators who did not participate in the individual or consensus annotation arm of the study selected the 48 papers from a pool of several hundred recent uncurated papers, using the criteria above.

Sixteen articles deemed to have one or more annotatable assertions were annotated by four different curators ('four-fold coverage'), two from each group. The remaining 32 articles were each annotated by two curators ('two-fold coverage'), one from each group. Of the four papers each curator annotated each week, two were from the four-fold coverage set, and two from the two-fold coverage set. The two-fold coverage set provided additional data for pairwise comparisons, while improving the productivity of the curators. (Normally there is a 1:1 relationship between papers and curators, so any redundancy in curation results in a significant reduction in productivity.)

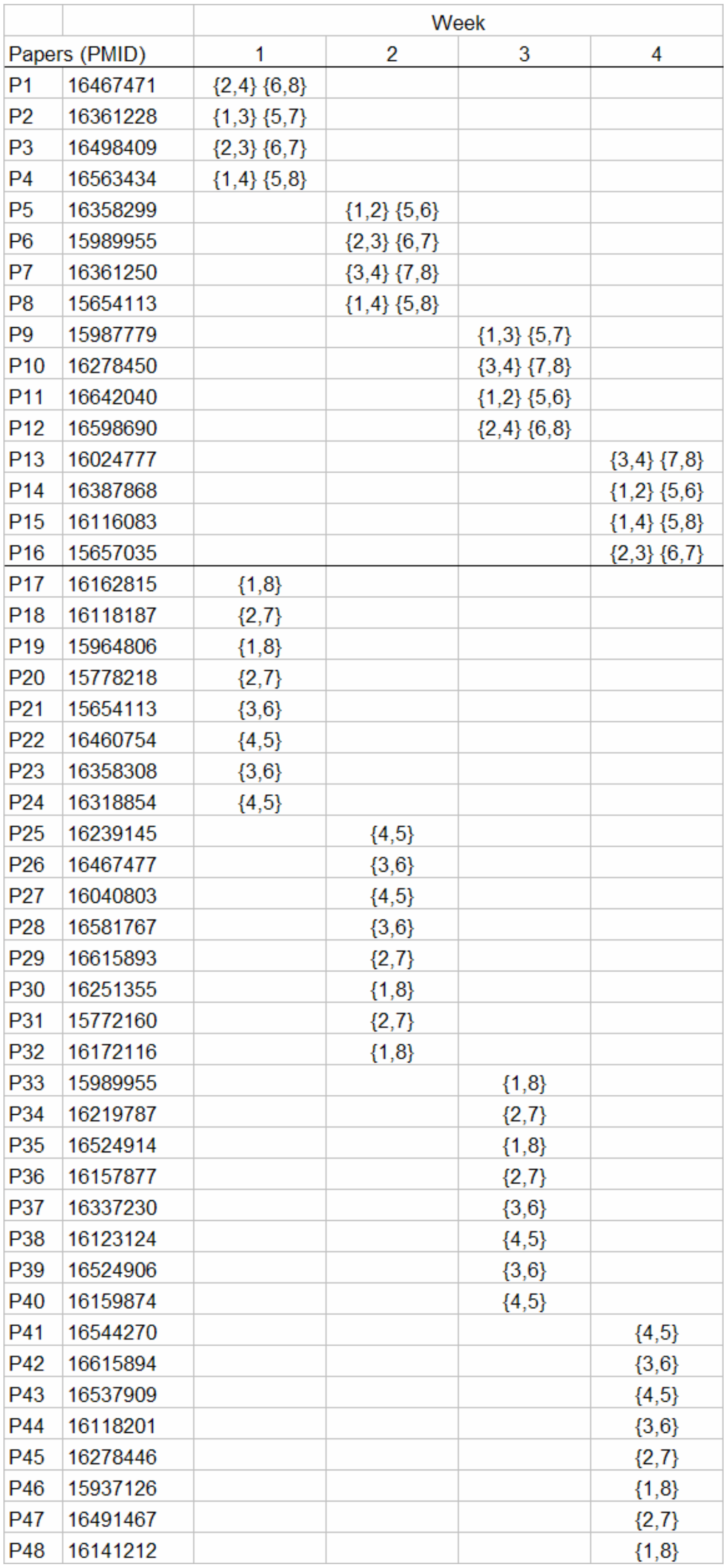
Paper assignment was completed as follows, using random numbers generated by random.org (2006). First, each paper in the 4-fold set was randomly assigned a number from 1-16, and each 2-fold paper was randomly assigned a number from 1-32. The papers were assigned by group using a partially counterbalanced design. The goals were that (a) each four-fold paper be annotated by four curators per week, and 2 curators per group; (b) that curators be paired for maximum variation (i.e., maximizing the number of unique pairings, and minimizing the frequency with which pairings occur). Groups one and two had the same pattern of assignment. Paper assignments were provided to the curators weekly, rather than at once, but curators chose the order in which they annotated the papers each week. Annotation was performed by curators in their normal work settings. The SMS curators do not all work in a centralized facility; some are based at universities, and several work at home or in other locations. Figure 4 shows the final paper assignments as consensus pairings used in the study.

This approach was designed to minimize assignment and order effects, and to create one pair of curators per paper in each group who would create consensus annotations, as described in the following section. Assignment effects should have been minimized because (a) the papers were assigned by the PI to Curator IDs, thus blinding him to which curator received which papers; (b) the Curator IDs were assigned to curators by the non-participating senior curators, so they did not assign papers directly to curators; (c) the curators did not know which of the papers each week had four-fold or two-fold coverage, and did not know for which papers they were to create consensus annotations until after they had completed their individual annotations.

Appendix A6 provides the complete citations for each of the papers in the corpus.

2.4.3. Multi-MOD study

Because the main purpose of the GO Camp meeting was to train novice GO curators in GO annotation, representatives from each participating MOD were instructed to select



		Week			
Papers (PMID)		1	2	3	4
P1	16467471	{2,4} {6,8}			
P2	16361228	{1,3} {5,7}			
P3	16498409	{2,3} {6,7}			
P4	16563434	{1,4} {5,8}			
P5	16358299		{1,2} {5,6}		
P6	15989955		{2,3} {6,7}		
P7	16361250		{3,4} {7,8}		
P8	15654113		{1,4} {5,8}		
P9	15987779			{1,3} {5,7}	
P10	16278450			{3,4} {7,8}	
P11	16642040			{1,2} {5,6}	
P12	16598690			{2,4} {6,8}	
P13	16024777				{3,4} {7,8}
P14	16387868				{1,2} {5,6}
P15	16116083				{1,4} {5,8}
P16	15657035				{2,3} {6,7}
P17	16162815	{1,8}			
P18	16118187	{2,7}			
P19	15964806	{1,8}			
P20	15778218	{2,7}			
P21	15654113	{3,6}			
P22	16460754	{4,5}			
P23	16358308	{3,6}			
P24	16318854	{4,5}			
P25	16239145		{4,5}		
P26	16467477		{3,6}		
P27	16040803		{4,5}		
P28	16581767		{3,6}		
P29	16615893		{2,7}		
P30	16251355		{1,8}		
P31	15772160		{2,7}		
P32	16172116		{1,8}		
P33	15989955			{1,8}	
P34	16219787			{2,7}	
P35	16524914			{1,8}	
P36	16157877			{2,7}	
P37	16337230			{3,6}	
P38	16123124			{4,5}	
P39	16524906			{3,6}	
P40	16159874			{4,5}	
P41	16544270				{4,5}
P42	16615894				{3,6}
P43	16537909				{4,5}
P44	16118201				{3,6}
P45	16278446				{2,7}
P46	15937126				{1,8}
P47	16491467				{2,7}
P48	16141212				{1,8}

Figure 4. Paper assignments and consensus pairing for SMS study (derived from Fig 2.)

two recent articles of average length and complexity that were related to their organism. These articles were to be used for the study, as well as the training camp portion of the meeting, during which novice curators would read and annotate them prior to the meeting, and then discuss them during the meeting in small group settings with the expert curators. Since all 23 participating curators annotated each paper, there was a resulting increase in coverage from 4-fold, as in the SMS study, to 23-fold. Annotations made by novice curators who participated in the GO Camp were not collected for this study.

Since the single-MOD curators who participated in the MMS had previously participated in the single study, they selected two articles they had already annotated in that study. As a result, for PMIDs 16642040 and 16598690 (SMS papers P11 and P12, and MMS papers P1 and P2), a total of 25 individual annotations exist.

Appendix A7 provides the complete citations for each of the papers.

2.5. Individual curator interviews

Curators who participated in the individual and consensus annotation tasks were interviewed using a semi-structured protocol to gather multiple types of data, including details about their personal education, training, and lab- and curatorial experience; their work roles; and their descriptions of their personal annotation processes and tasks. Of particular interest are curators' individual differences in their performance of the 'standard' GO process: what workflows they follow, which information tools they use, their approaches to claim extraction from journal articles, their personal annotation (note-taking) behavior, and other facets. The approach to the work- and task-related questions can be viewed as a variant of contextual inquiry, and as similar to the critical incident technique, since curators were asked for a retrospective self-report of how a task was performed in relation to one or more specific evidence sources.

Appendix A1 lists the semi-structured interview questions, and related prompts and probes.

2.5.1. Single-MOD study

Each of the ten participating curators was interviewed on-site during the SMS. Curators were asked first about their backgrounds and experiences, and then about their personal approaches to GO annotation.

2.5.2. Multi-MOD study

The interview protocol used for the MMS differed slightly from that which was used for the SMS. In the MMS, participants were provided with a spreadsheet template in which to record their annotations. The spreadsheet also included a brief questionnaire sheet (attached as Appendix A2) that collected data about their experience with biology, curation, and GO annotation. In the SMS, this information was collected during the interviews.

2.6. Curator focus groups

2.6.1. Single-MOD study

During the SMS an hour-long focus group discussion was conducted with all participating curators in an attempt to elicit additional information about their GO annotation experiences, and to enable comparisons of these statements with those made in the individual interviews, and with the observation data. This meeting was also used to gather feedback about the study itself and how similar studies might be conducted more effectively in the future. This meeting was audio recorded and transcribed. Appendix A3 lists the focus group question guide.

2.6.2. Multi-MOD study

A formal focus group was not held during the MMS due to the presence of a large number of unconsented participants. However, during the GO Consortium meeting prior to the camp, a large group discussion of the MMS occurred with the study participants, which included discussion of differences in annotation outcomes on specific papers. This meeting was not audio recorded, but the study PI took field notes of the discussions involving variation in curators' annotations and approaches to curation.

2.7. Curator observations

2.7.1. Single-MOD study

The two senior curators involved in the Reliability facet assessment were observed performing individual annotations. The general approach was a combination of observation, during which the researcher recorded curator behavior for later task decomposition, and what Beyer & Holtzblatt (1998) call "contextual inquiry", where participants were asked about the tasks and their interactions with the artifacts used while performing them. The concurrent verbal report (or "think-aloud") protocol was used to elicit tacit knowledge and unobserved facets of the annotation process by having a curator verbalize decisions and actions while he or she was performing an actual annotation process.

Thirteen of the consensus pairs were observed to gather contextual data. One area of interest is effects of strong or weak personalities, and different levels of knowledge and experience, that may artificially pressure one curator to change his/her annotation to conform to the other. The consensus observation protocol is detailed in Appendix A4.

2.7.2. Multi-MOD study

Formal observations were not conducted during the MMS, but notes were made in four of the expert/novice discussions about attributes such as areas of agreement and disagreement, and curators' descriptions of their individual workflows.

2.8. Artifact collection

2.8.1. Single-MOD study

One surprising early finding of this study is that the vast majority of SMS curators read and curate articles on their computers rather than printing paper copies and making hand-written notes. All of the local curators, and some of the off-site curators, have high-resolution 23-inch LCD flat panel screens that they are very comfortable using to read text. They frequently have the articles and curator web interface side-by-side on the screen to allow curation directly from the papers into the database with no intermediate steps, such as manual annotation. In addition, many commented that they need the papers for such a short period of time (and will very rarely ever refer back to them) that it seems pointless to print and/or save copies of the papers. Only one curator had made printed copies of the articles, which were given to the study PI for content analysis.

2.8.2. Multi-MOD study

For this study, most participants printed paper copies of the articles, because they anticipated referring to them during the discussion portions of the GO Camp. Six complete sets of hand-annotated articles were collected from participating curators, and the personal annotations were analyzed as described in section 3.4.

Appendix A5 lists the details of the artifact protocol.

2.9. GO annotation quality metrics

2.9.1. Purpose and goals

The preceding sections described the processes by which structured annotation data was generated and collected, but only implicitly described the end uses of the data. The two primary purposes for this project are to quantitatively compare the degree of variation in GO annotations, and explore the contextual reasons variation exists. This section describes the quantitative measures used to analyze the annotation data. The data, in turn, provide a way to assess the utility of the draft metrics defined below, to enable their future refinement.

Camon, et al. (2005:6) assert that "[v]ariation is acceptable between curators but inaccuracy is not". It is unclear what the authors meant by this statement, and neither that article or the small number of others that have investigated variation in GO annotations have characterized the kinds and degrees of variation. As noted in the Introduction, variation should not necessarily be equated with quality, as quality is a multifaceted concept. The six quality facets defined below attempt to deconstruct that concept into mutually exclusive elements, but even those do not permit assessments of absolute accuracy or truth about which annotations are correct, only in relation to reference standards created in the consensus process.

Two or more curators annotating a set of the same papers are likely to differ at certain times and in certain ways, which is often referred to as measurement error (see, e.g., Friedman & Wyatt, 2006:120). A frequent goal of measurement studies is the identification and minimization of measurement error through some means prior to conducting an evaluation study. In this study, we were interested in understanding why curators have different annotation outcomes. Only then can we decide whether it is feasible to discuss error reduction. In other words, the purpose of this project was not to eliminate curator measurement error, because that variation is in fact the object of study.

In this section, the concept 'annotation quality' is operationalized as five facets: Consistency, Reliability, Specificity, Completeness and Validity (revised from MacMullen 2006b). Each of these facets can be further decomposed into multiple attributes, which are parameterized below. Cumulative values for facets (and certain combinations of facets) are described synonymously here as 'measures' or 'metrics', which are derived from the values or 'scores' of their underlying attributes.

One goal of this approach was to have the ability to view and compare the annotation quality of two or more arbitrary GO annotations at varying levels of granularity, including:

- An overall quality score (all facets and underlying elements);
- individual facet-level scores (all elements of a single facet); and,
- individual element-level scores (each element of each facet).

Secondly, it was desirable that the measures be unambiguously parameterized so that the process could be automated (i.e., no human judgment required to determine scores).

A broader goal was that the facets, elements, and parameters be organism-independent to permit the use of the same measures for studies of annotation quality within and across all organisms that use GO annotation.

2.9.2. Definitions

Let P_i denote a scientific paper used as an evidence source, and let A_i be an annotation instance extracted from that paper. A particular paper may have 0-*n* annotation instances, and each may be made to any one (but only one) of the three GO aspects. For each A_i, two or more curators, denoted as C_{ij}, may have created annotation instances. If we take the Takayama annotation shown above in Table 1 as paper P1, annotated by curator C1, we can represent the annotation instance in the third row as '$P_1A_3C_1$'.

For each annotation $P_iA_iC_{ij}$, the following measures are defined below: $P_iA_iC_{ij\ \mathrm{CONSISTENCY}}$, $P_iA_iC_{ij\ \mathrm{RELIABILITY}}$, $P_iA_iC_{ij\ \mathrm{SPECIFICITY}}$, $P_iA_iC_{ij\ \mathrm{COMPLETENESS}}$, and $P_iA_iC_{i\ \mathrm{VALIDITY}}$. While the Consistency facet applies to multi-annotator cases, and the Reliability facet applies a similar measure to single-annotator cases, the Specificity, Completeness, and Validity facets are applicable to both cases.

2.9.3. Consistency

The Consistency facet measures simple pairwise and cumulative agreement among the elements of individual GO annotation instances, for two or more annotators. This is a dichotomous measure of the agreement between the element values (i.e., whether they are the same or different). The attributes of which Consistency is composed are the mandatory GO annotation elements 'Gene product', 'GO ID', 'GO term', and 'Evidence code', and the optional elements 'Qualifier', 'and With/from'. For each evidence source or each annotation instance, Consistency may be measured at the element level or the facet level (i.e., as an overall consistency score). The Consistency facet is defined as:

$P_i A_i C_{ij\ \text{CONSISTENCY}} =$

Attributes	Values
Gene	{0,1}
GO aspect	{0,1}
GO ID	{0,1}
GO term	{0,1}
Evidence	{0,1}
With/from	{0,1,-}
Qualifier	{0,1,-}
Overall	{0,1}

where '0' is a mismatch, '1' is a match, and '-' means 'not used'. 'Overall', both here and in the facets below, is a binary metric that summarizes the other attributes for each calculation: if the values for all attributes agree, 'Overall' will equal '1'; if not, it will equal '0'.

2.9.4. Reliability

The Reliability facet measures simple pairwise and cumulative agreement among the elements of the original and one or more repeated annotation instances, as made by the same annotator at different time points. Reliability is thus essentially an 'intra-annotator' version of Consistency. Both Reliability and Consistency require reference to GO to determine whether any changes made in the GO vocabularies over the time intervals between compared annotations had material effects on term selection. The Reliability facet is defined as:

$P_iA_jC_{ij\ \text{RELIABITY}}$ =

Attributes	Values
Gene	{0,1}
GO aspect	{0,1}
GO ID	{0,1}
GO term	{0,1}
Evidence	{0,1}
With/from	{0,1,-}
Qualifier	{0,1,-}
Overall	{0,1}

2.9.5. Specificity

The Specificity facet measures the degree of agreement between the GO IDs in two or more annotation instances, by GO aspect. Degree of agreement is measured using three types of granularity: first, categorically, following Camon, et al. (2005), as being 1) an exact match; 2) different terms, but with the same lineage; or 3) different terms, not in the same lineage; second, as a binary distinction of one term being broader or narrower than the other; and third, as the nodal distance(s) between the compared GO IDs in relation to the depth of the branch(es) on which they reside. Since GO is a polyhierarchical vocabulary, the shortest path between the compared terms is used when calculating distance. The nearest common ancestor of the two terms is also identified.

$P_iA_iC_{ij\ \text{SPECIFICITY}}$ =

Attributes	Values
Exact match	{0,1}
Mismatch, same lineage	{0,1,-}
Mismatch, different lineage	{0,1,-}
Broader	{0,1,-}
Narrower	{0,1,-}
Difference, in nodes	{0,*n*}
Nearest common ancestor	{GO_ID, -}

where ‘same lineage’ means the terms are descended from the same ancestor closest to the root, while ‘different lineage’ means the terms have different ancestors. Nodal difference is calculated by counting the nodes in between the two terms, but not the terms themselves, and counting their nearest common ancestor only once. A zero node difference means the terms are adjacent on the tree. ‘Broader’ or ‘narrower’ is calculated only when the terms are located on the same branch.

2.9.6. Completeness

The Completeness facet measures simple presence or absence of instances in annotations in relation to a consensus or reference annotation. Measurement of this facet requires a paper to have been curated by at least two curators, who then rationalize their individual annotations to create a reference standard against which their (and other curators’) individual annotations can be compared. The semantic content of the annotation instances is not evaluated by this facet. Completeness is defined as:

$P_iA_iC_{ij\,\mathrm{COMPLETENESS}}$ =

Attributes	Values
False positives	{0, +1}
False negatives	{0, -1}
Overall	{0,1}

where ‘Overall’ means the annotation being compared to the reference annotation is complete or not complete; ‘false positives’ means the compared annotation has instances the reference annotation lacks; and ‘false negatives’ means the compared annotation lacks instances the reference annotation has. In practice, Completeness is evaluated by examining elements of the annotation instances to determine which instances match and which do not, starting with the Gene Name element and considering Aspect.

2.9.7. Validity

The Validity facet is a basic error-checking test that evaluates whether the values supplied for each element of an annotation instance fall within the range of allowed values. This is a dichotomous measure that evaluates an annotation against the GO annotation definition, not against other individual or consensus annotations. Some MODs perform this check as an internal function of their databases, but this measure is MOD- and database-independent.

$P_iA_iC_{i\ \text{VALIDITY}}$ =

Attributes	**Values**
Gene	{0,1}
GO aspect	{0,1}
GO ID	{0,1}
GO term	{0,1}
Evidence	{0,1}
Qualifier	{0,1,-}
With/from	{0,1,-}
Overall	{0,1}

3. Results

3.1. Summary of data obtained

3.1.1. GO annotations

In both the single-MOD and multi-MOD studies, the formal GO annotations were collected using a spreadsheet template distributed to all participants. The forms had a separate sheet for each article with predefined columns for each of the GO annotation elements. Each of the participants sent the completed forms to a non-participating curator, who then compiled them, ensured they were de-identified, and sent them to the study PI. In the single-MOD study, this person was a supervisory member of the MOD staff; in the multi-MOD study, this was a curator who was coordinating the meeting at which the study was based. The participants in both studies also used the same template to record their consensus annotations, which were collected in the same manner. The observations resulting from the studies are shown in Table 4. A total of 480 annotations composed of 3,915 instances were generated by the curators during the two studies. On average, there were about 2 instances per annotation in the SMS, and 8 per annotation in the MMS.

Section 3.3.4 below assesses incomplete instances during the evaluation of the Validity quality aspect. In cases where curators made no annotations, these observations were counted as having zero instances rather than as missing data, because it is a valid annotation outcome to conclude that an evidence source contains no annotatable information. In the SMS there were 60 cases of zero annotations across the individual and

Table 4. Aggregate GO annotation data points by study

	Single-MOD	Multi-MOD	Total
Individual annotations per paper	*2/4	23	-
Total individual annotations	**128	230	358
Total consensus annotations	32	90	122
Overall annotations (individual + consensus)	160	320	480
Total individual annotation instances	265	2,289	2,554
Total consensus annotation instances	103	1,258	1,361
Overall instances (individual + consensus)	368	3,547	3,915

*4 each from the 16-paper 4x coverage set, and 2 each from the 32-paper 2x coverage set
**64 from the 4x coverage set, and 64 from the 2x set

consensus annotations, and 11 in the MMS. In the latter, most (n=8) of the cases were related to one curator, C23, while in the SMS the cases were distributed across all 8 curators; every curator had at least four papers where zero annotation instances were made, and 6 of the 12 consensus pairs had at least one case of zero instances. Thirty-four of the cases were from the 2-fold coverage set of 32 papers, which was not assessed in detail by the MOD leadership for the presence of putative annotations, while the 4-fold set was.

Within the spreadsheet templates, some curators made notes that describe why they made certain annotations, and the decisions underlying them. In some cases these notes provide additional contextual detail, but many of them relate to current database policy or other issues that do not directly affect the GO annotations, and so they were not considered in the quality evaluation. In some cases, curators chose existing GO terms for their annotations, but also noted that in a real-life situation they might request a new term instead of using the existing terms.

Because curator IDs were assigned randomly to participants, and the contextual data were not linked to the annotation data, the study PI had no knowledge of the identities of the curators until after the quantitative analysis had been performed. Likewise, the data from the personal experience questionnaire pages of the MMS template was not compiled until after the analysis of the GO annotations, since it could have been possible to identify curators based on their attributes.

3.1.2. Individual interviews (both studies)

Fifteen interviews with individual curators were conducted over the course of the two studies, ten with participants in the single-MOD study, and five with those in the multi-MOD study. Digital audio recordings were made of all interviews, which were then transcribed, corrected, and verified. Table 5 lists each interview's acquisition date, duration in minutes, and transcribed length in pages and words.

Table 5. Audio recordings of individual curator interviews

ID*	Date	Duration	Transcribed pages / words
SI01	2006-06-05	00:40:12	21 / 6,300
SI02	2006-06-05	00:37:34	21 / 7,200
SI03	2006-06-05	00:30:05	19 / 5,300
SI04	2006-06-06	00:30:35	15 / 4,700
SI05	2006-06-06	00:22:33	13 / 3,800
SI06	2006-06-07	00:32:21	18 / 4,300
SI07	2006-06-09	00:34:00	21 / 5,900
SI08	2006-06-09	00:44:48	29 / 8,900
SI09	2006-06-09	00:29:56	26 / 5,500
SI10	2006-06-09	00:23:20	16 / 4,000
MI01	2006-07-14	00:19:18	18 / 3,500
MI02	2006-07-14	00:25:00	18 / 4,200
MI03	2006-07-11	00:21:22	20 / 4,100
MI04	2006-07-12	00:27:26	14 / 4,500
MI05	2006-08-03	00:41:08	20 / 5,800
	Total	**07:39:38**	**289 / 78,000**

*'S' prefix = single-MOD study, 'M' prefix = multi-MOD study

The interviews contained data on curators' workflows and individual work processes. These data were coded as described in section 3.2.2 and used for the workflow analysis in section 3.5

3.1.3. Curator background and experience data (both studies)

In the multi-MOD study, the template file in which curators recorded their annotations also contained a four-item self-administered questionnaire which was used to gather information about each curator's backgrounds and experience. All 23 curators completed the questionnaire, and the data are described in section 3.6 below. The questionnaire is attached as Appendix A2. Subsequent to the completion of the study, curators were asked a follow-up question by email about the number of years or months experience each had with GO annotation at the time of the study.

In the single-MOD study, similar questions were asked of the 10 participating curators during the individual interviews that were conducted. This information was extracted from the transcribed interviews, recorded, and analyzed as described in section 3.7 below.

3.1.4. Evidence sources and other artifacts (both studies)

Several types of additional artifacts were collected during the two studies. Six participants in the multi-MOD study provided complete sets of their paper copies of the 10 articles used in the study, for analysis of their intermediate manual notes and annotations. These total approximately 630 pages, which included about 15 separate pages of associated notes in addition to the marked-up primary articles. Other documents and data included one curator's manual notes and article annotations from the single-MOD study, and a chain of emails from one of the consensus pairs from the single-MOD study to document their process of consensus formation. Also available were the PI's notes from both studies of observations and interviews, and screenshots of a MOD's curator web interface, collected during a portion of the single-MOD study.

3.2. Data coding

3.2.1. GO annotation coding and analysis

Annotations and annotation instances

For this project, an ‘annotation’ is defined as the set of individual instances of GO annotations that is made at the level of individual associations between gene products and GO terms, in the context of a single experimental journal article. An annotation can be composed of one or more annotation instances, which in turn are composed of the set of required and optional elements defined by the GO Consortium, as described above in section 1.4.3.

Instance-counting rubrics

The ways in which individual curators entered their annotation elements into the spreadsheet template used for data collection created some challenges in counting annotation instances. Significant space is devoted here to these issues because they are relevant to future work on GO annotation analysis, both in studies of this type as well as others that do not require a common article corpus or multiple curators annotating the same documents. These challenges range in complexity from minor formatting issues to greater ambiguity about the nature of what an instance is.

In some cases, curators used the spreadsheet application’s ‘merge’ function to combine cells that would otherwise have had their contents duplicated exactly across multiple instances. For example, if creating three instances using the same gene product, a curator might merge the three cells so that the gene product name appears only once, but spans the three rows. In this type of case, the count is straightforward: three instances were counted. A similar situation results when curators made several annotation instances to the same gene product, and omitted the gene, gene ID, and species information in subsequent instances, presumably for readability. These and other formatting irregularities were normalized for consistency in an intermediate spreadsheet that combined all curators' and consensus pairs' annotations. Incomplete or non-canonical instances were not counted or analyzed. There were very few of these, which involved cases where required GO elements were not populated with values, or in one case where the Gene Product element had the value 'various'. Missing values are discussed above in section 3.1.1.

A different situation occurs when an instance has two (or more) GO Evidence Codes recorded in the Evidence Code element. Should this be coded as one instance, or more? The rubric used in this project is that a case with two evidence codes in one instance is coded as two instances, since presumably each could stand alone as an instance in the absence of the other, given that evidence codes are mutually exclusive. For example, an instance provided by Curator 15 on Paper 4 in the multi-MOD study is (in part):

Gene product	GO aspect	GOID	GO term	**Evidence Code**
ced-10	P	GO:0043652	engulfment of apoptotic cell	**IMP, IGI**

Since apparently an experiment using mutant phenotypes (IMP) was conducted, as well as an experiment using genetic interactions (IGI), this is counted as two instances:

Gene product	GO aspect	GOID	GO term	**Evidence Code**
ced-10	P	GO:0043652	engulfment of apoptotic cell	**IMP**
ced-10	P	GO:0043652	engulfment of apoptotic cell	**IGI**

It is important to this study that this distinction be made, in order to partially address the problem of matching instances across curators, discussed further below. When calculating the similarity of the above instances with other curators' annotation instances, for example, each of those instances may be present separately (or both, or neither).

There were 56 cases of multiple evidence codes in the SMS, and 43 cases in the MMS (SMS: P3: 2; P7: 12; P11: 14; P12: 1; P13: 1; P14: 9; P15: 8; P32: 1; P33: 1; P35: 1; P39: 2; P40: 3; P43: 3; P45: 3; MMS: P1: 10; P2: 2; P3: 1; P4: 6; P5: 1; P6: 12; P7: 3; P8: 5; P9: 3). Additionally, 6 papers in the SMS had a total of 16 cases of instances with multiple gene names with the same GO annotations (P3: 2; P11: 3; P14: 6; P40: 1; P43: 3; P45: 1); these were counted in the same manner as the cases with multiple evidence codes.

Instance counts below also include instances with proposed GO terms (if complete – some lack gene referents), NOT annotations, and those with evidence codes such as TAS or ISS that are sometimes deprecated by individual MODs or the GO Consortium as a whole. Local database policy decisions were ignored in calculations (e.g., if a curator made an annotation, but then stated in the Notes column that s/he would not have made it in real life due to database policy, it was still counted).

Issues with matching instances across curators

As seen below in section 3.3.1, there were large variances in the numbers of instances created by different curators for the same paper. As a result, it is often difficult to match instances for the purposes of pairwise comparison of variation of annotation element values. For example, for Paper 4 in the multi-MOD study, the instance shown above for Curator 15 is the only one that curator made, while Curator 2 had the following:

Gene product	GO aspect	GOID	GO term	Evidence Code
ced-10	P	GO:0043652	engulfment of apoptotic cell	IMP
ced-10	P	GO:0035262	gonad morphogenesis	IGI

and Curator 10 had:

Gene product	GO aspect	GOID	GO term	Evidence Code
ced-10	P	GO:0043277	apoptotic cell clearance	IGI
ced-10	P	GO:0035262	gonad morphogenesis	IGI
ced-10	P	GO:0040039	inductive cell migration	IGI

and Curator 21 had:

Gene product	GO aspect	GOID	GO term	Evidence Code
ced-10	P	GO:0012501	programmed cell death	IMP
ced-10	P	GO:0008354	germ cell migration	IMP
ced-10	P	GO:0035262	gonad morphogenesis	IMP
ced-10	F	GO:0005515	protein binding	IPI

Since the numbers of instances vary, as well as the terms and evidence codes, it is difficult to perform pairwise comparisons of the annotations at the instance level. In this case, the curators are at least annotating to the same gene product. This is not always true, and in situations where there are multiple gene products with varying numbers of instances, or mutually exclusive gene products for each curator's instances, it becomes even more difficult to make comparisons at the instance level.

There are multiple approaches to addressing this problem, but each requires some level of manual evaluation. At a very high level, exact match comparisons can be performed at the annotation level, as shown below in section 3.3.2. At a finer grain, the numbers of individual instances per paper by curator can be compared (also shown below).

At the element level, one approach is to focus on the gene product as the unit of analysis. For each gene product that is shared across curators, we can manually review the instances to attempt to determine whether they are capturing the same claim from the paper. For instance, in the examples above, Curator 2 and Curator 15 both annotate the term 'engulfment of apoptotic cell' to ced-10. Similarly, Curators 2, 10, and 21 annotated 'gonad morphogenesis' to ced-10. Matching those instances enables pairwise evaluation of similarity. For other cases, in which term-based matching is not possible, an approach is to match instances first by gene, GO aspect, and evidence code, and then use the Specificity quality facet to assess term distance. In cases with large numbers of instances across annotations, doing this in a semi-automated fashion with matrices is more feasible. Another approach is to calculate the intersection of the corresponding GO term sets, including the number of total terms, and the number of shared terms. In the example above, the four term sets are:

C2	C10	C15	C21
- engulfment of apoptotic cell - gonad morphogenesis	- apoptotic cell clearance - gonad morphogenesis - inductive cell migration	- engulfment of apoptotic cell - engulfment of apoptotic cell	- programmed cell death - germ cell migration - gonad morphogenesis - protein binding

There is a total of 8 distinct terms, but there are no terms in common across all four sets; the closest is 'gonad morphogenesis', which was selected by three curators: C2, C10, and C21. Curators C2 and C15 share 'engulfment of apoptotic cell', but that is the only other term shared by two or more curators. For large-scale analysis, this step could be performed by loading the term sets in a database and performing intersections or inner joins.

Issues with counting genes

In multi-organism papers where gene names are identical or similar (e.g., in MMS P1, ‘Tyw1’ in mouse, ‘Tyw1’ in yeast, and ‘TYW1’ in human), counting distinct genes is unclear. Does the above example count as one gene, or more? For the analysis below, genes in different species are counted as distinct genes.

3.2.2. Interviews

For this project, the interview transcripts were coded primarily for two types of information: 1) curators’ backgrounds and experiences, and 2) individual annotation behavior and personal workflows. The former was primarily extracted from the single-MOD study interviews, as the multi-MOD study participants self-reported that information in the curator questionnaire portion of the annotation worksheet (see Appendix A2 for the questionnaire instrument). Curators provided varying levels of detail about their workflows and annotation behavior, so they were not comparable at granular levels. Questions where specific answers were available included whether the curator normally reads and curates paper or electronic versions of the articles; which GO ontology searching and browsing tools are most often used, and which additional resources curators use to supplement their knowledge. The results are reported in section 3.5 below as a combination of quantified findings and more traditional narrative text extracts from the curators’ actual words. The Future Work section (5.2) addresses additional analysis that could be performed on these data.

3.2.3. Curators' personal annotations

The 636 pages of manually-annotated paper articles were digitized, coded and analyzed. A content-analytic approach was employed using emergent coding (Marshall 1997, 1998), with an emphasis on characterizing:

- The numbers of annotations made per curator by physical type (underline, circle, highlight, text, etc.).
- The physical location of annotations in documents at different levels of granularity (e.g., page, paragraph, sentence).
- The conceptual or intellectual location of annotations in documents (e.g., section, table, figure).
- A comparison of the above features across documents by curator, and curators by document.

Coding was performed at the sentence level. 'Paragraphs' were counted by including partial paragraphs at the top and bottom of each page, so a partial paragraph at the top of Page 2 would be coded as 'P2G1', a partial paragraph at the bottom of Page 2 as 'P2G*n*', and then the continuation of G*n* on the following page is counted as 'P3G1'. This coding scheme was employed because it was observed that curators often marked up the fractional paragraph on a following page without marking up the text on the preceding page. Section titles are coded as 'G*n* section', except for claims in sentence form, which are counted as S1. A multi-paragraph section with a section title in sentence form was counted as the section title being S1 of the first paragraph, and subsequent paragraphs in that section counted normally.

Annotations were coded as distinct instances when significant white space occurred between them. For example, a case where two phrases in a long sentence were underlined, but they were distinct claims separated by other phrases would be coded as two annotations, but a case where an entire sentence was underlined with the exception of article citations enclosed in parentheses would be coded as one annotation. Text that was both highlighted and underlined was counted as separate instances. Marks were quantified by both type and location in order to assess not only similarity of notation form, but also the placement of marks within the documents. Marks were localized at the section level (Abstract, Introduction, Methods, Results, Discussion, References), at the page level, and at the sentence level. Sentence-level coding enabled granular comparisons of precisely which claims, phrases, and words the curators thought were important. This analysis was performed to enable relationships between manual annotations and GO annotations to be assessed.

3.3. GO annotation data

3.3.1. Annotation counts

The claims in this section examine the differences observed in the *quantitie*s of annotation instances created by curators and consensus pairs. The *content* of the instances is explored in the two sections that follow.

1. *Ranges of individual annotation instances per paper vary widely.*

In the single-MOD study, the 64 annotations made on the 16 4-fold coverage papers were composed of a total of 185 instances, or nearly 3 instances per annotation, with an aggregate range of zero to 13 annotation instances per paper. Two papers (P8 and P9) had zero instances from all curators, and 10 of the remaining 14 papers had ranges between 1-6 instances. Eight of the papers had at least one curator who supplied zero annotations. The instances have a per-paper mean of 2.8 and standard deviation of 3.3. Figure 5 shows a boxplot of the individual instance counts for the 16 4-fold coverage papers. While we might expect that the numbers of annotation instances would vary across papers due to differences in content, we can see also that several papers have large ranges of instance counts. The origins of these are explored below.

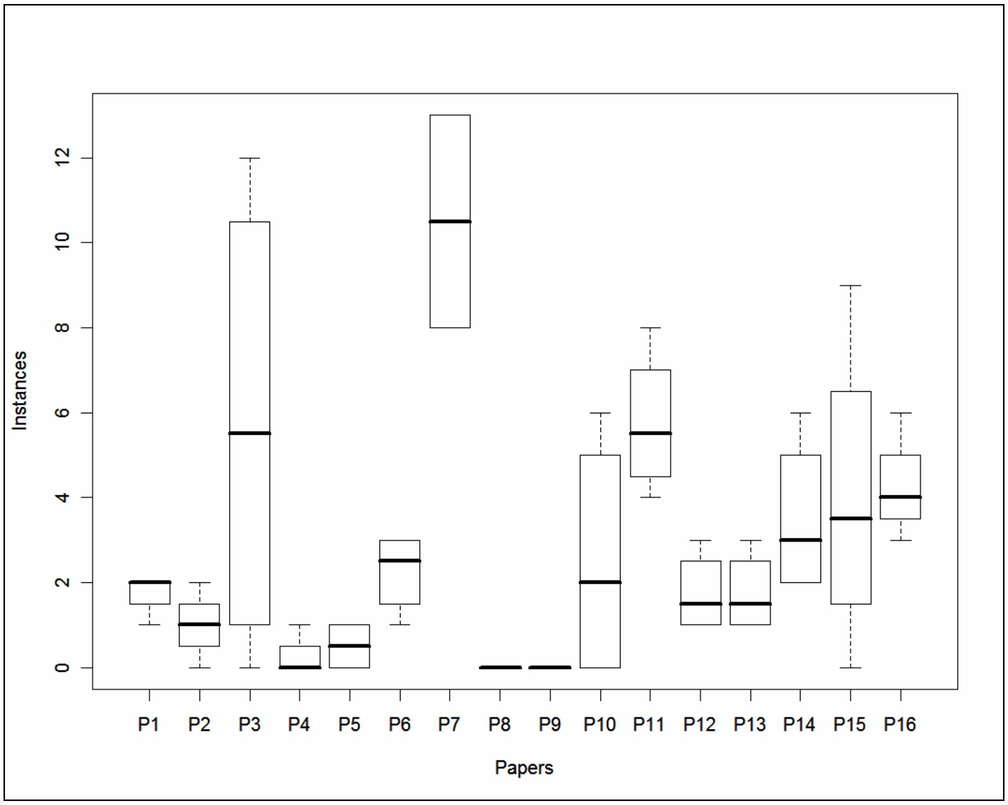


Figure 5. Boxplot of individual SMS instances by paper (4x set)

The 64 annotations made on the 32 2-fold coverage papers totaled 80 instances, and had a range of zero to 6 annotation instances, with 11 papers having zero instances from all curators, and 13 having 1-2 instances. The 65 instances have a per-paper mean of 1.3 and standard deviation of 1.9.

The preceding addressed aggregate instance variation across all papers in the study. At the individual paper level, the papers with the smallest variation in instance counts (excluding those with zero annotations by all curators) were P1, P4, and P5 (1 instance difference). The largest ranges were observed in papers P3 (11, 0-11 instances), P3 (12, 0-12 instances), P15 (9, 0-9 instances) and P10 (6, 0-6 instances). Figure 6 shows the distributions of the instances for the 16 4-fold papers. In only two cases were all four curators consistent: P8 and P9, which had zero instances from all curators.

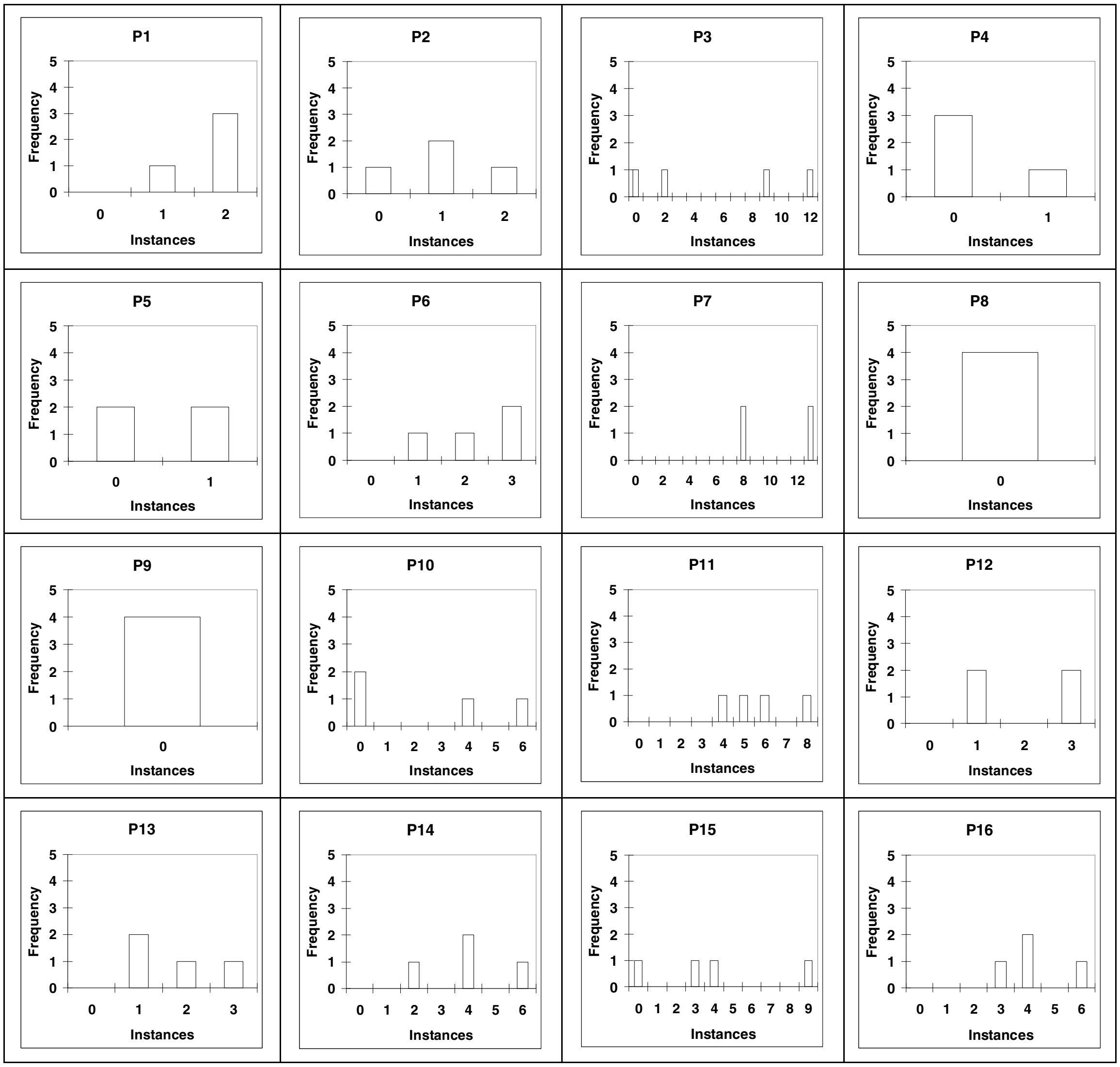


Figure 6. Frequency distributions of instances from 16 4-fold papers (SMS)

In the multi-MOD study, the quantities and ranges of instances were larger. The 23 curators made a total of 230 annotations to the set of 10 papers. These were composed of 2,289 annotation instances, which have a per-paper mean of 9.9 and standard deviation of 8.1. The number of instances ranges from zero to 46. Figure 7 shows a boxplot of the MMS instances, which illustrates the wide ranges per paper.

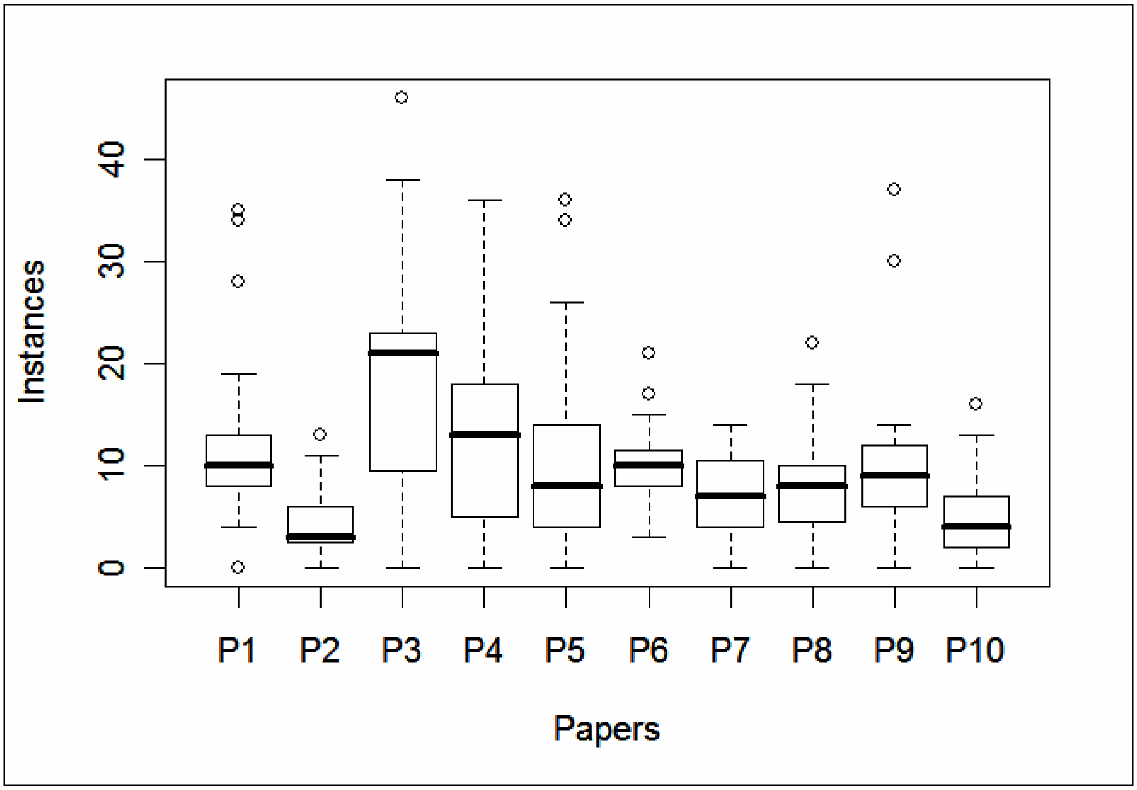


Figure 7. Boxplot of individual MMS instances by paper

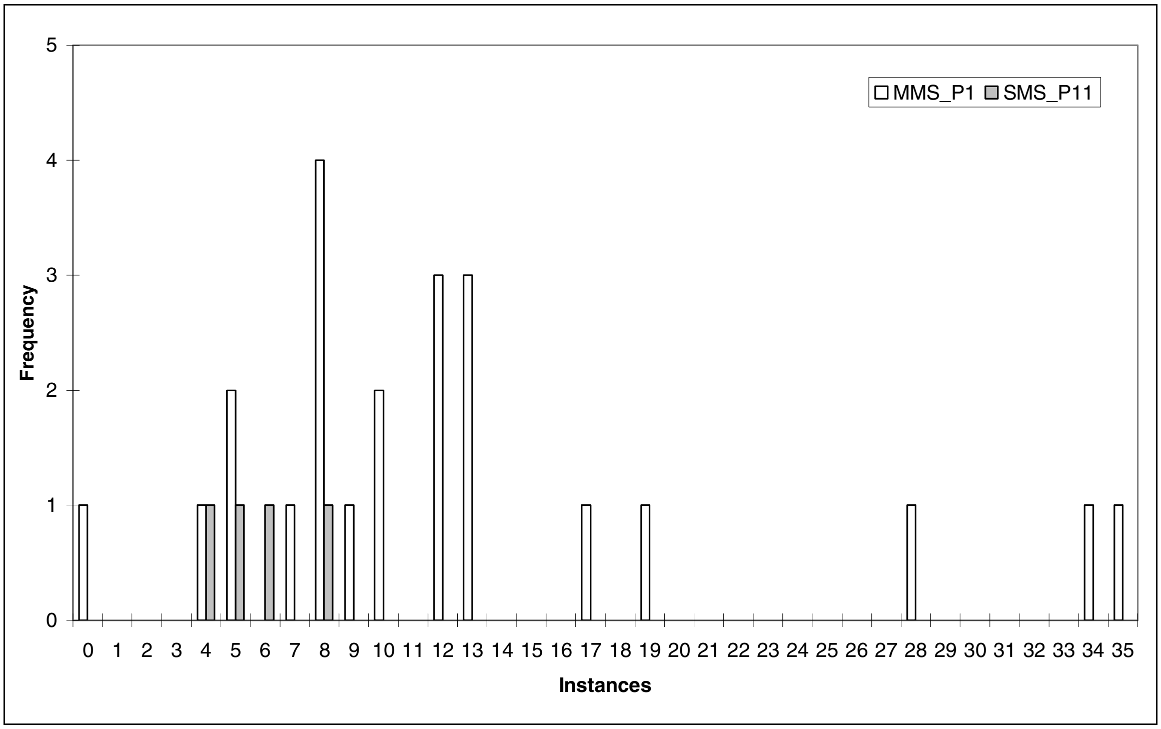
At the individual paper level, the papers with the smallest variation in instance counts were P2 (13, 0-13 instances), P7 (14, 0-14 instances), and P10 (16, 0-16 instances). The largest variation was observed in papers P3 (46, 0-46 instances), P9 (37, 0-37 instances), P4 and P5, (both 36, 0-36 instances), and P1 (35, 0-35 instances). Curator 23 made no annotations for 8 of the 10 papers, and those zero values tend to skew the ranges of some of the papers, such as P3 and P4, where the lowest non-zero instance counts are 5.

However, because of the large means (18, 13) and medians (21, 13) for those papers, the variation is still significant. Interestingly, the two papers from the SMS study that were included in the MMS study have significantly higher instance ranges, as shown in Table 6.

Table 6. Differences in instance count ranges, means, and standard deviations for papers included in both SMS and MMS

	n	Range (min-max)	Mean	SD
SMS P11	4	4 (4-8)	5.8	1.7
MMS P1	23	35 (0-35)	12.6	8.9
SMS P12	4	2 (1-3)	1.8	1.0
MMS P2	23	13 (0-13)	4.5	3.2

The SMS distributions are much lower and within tighter ranges, while the MMS distributions are more spread out, as shown in Figure 8a-b.



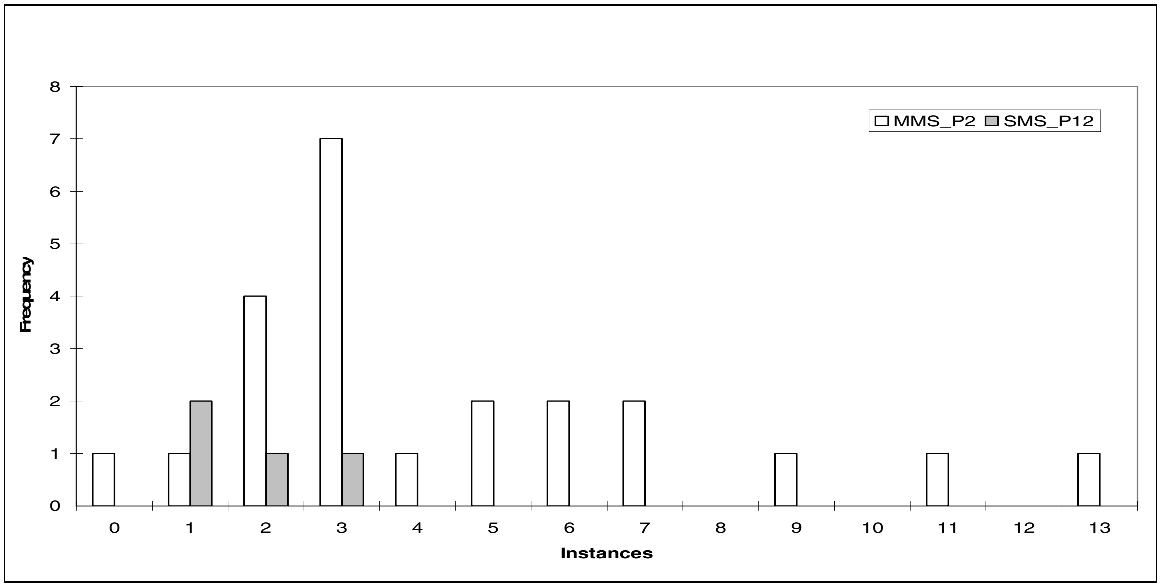


Figure 8a-b. Frequency distributions of instances in the overlapping papers from the SMS & MMS

Figure 9 shows the frequency distributions of the instances for the 10 MMS papers. With 10 data points per paper rather than the four of the SMS, we might expect that the shapes of the distributions would become more coherent, but extreme values make the ranges so wide that there are often large gaps between data points.

2. *Consensus instance counts varied from individual annotations by study.*

The 32 consensus annotations made on the 16 4-fold coverage papers in the single-MOD study were composed of 103 instances, and had a range of zero instances (papers P8 and P9) to 13 (P7). Seven cases had zero variance (i.e., matching instance counts), while the paper with the largest mismatch was P15, with a range of 0-9 instances. The mean and standard deviation of the consensus instances are 3.2 and 3.6, which are higher than those measures for either the 4-fold or 2-fold individual annotations. Figure 10 shows a boxplot of the consensus instance counts by paper, which illustrates that several papers have very small absolute ranges (e.g., P1, P12-P14, P16), while others' are large.

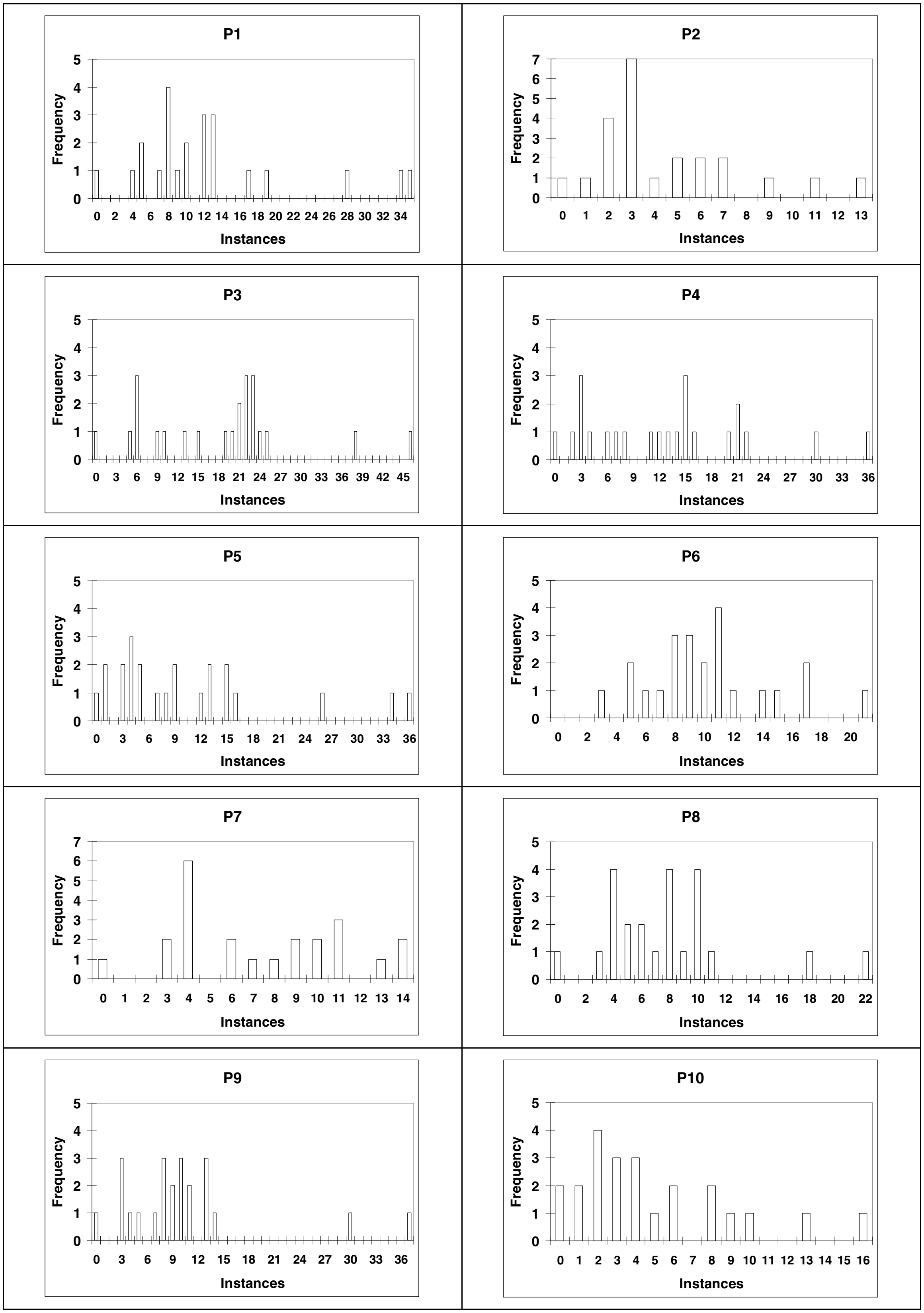


Figure 9. Frequency distributions of MMS instances by paper

An interesting difference observed here compared with the individual instances is that the wide range of P3 has been collapsed from 0-12 to 9-12; the ranges of other papers diminished as well.

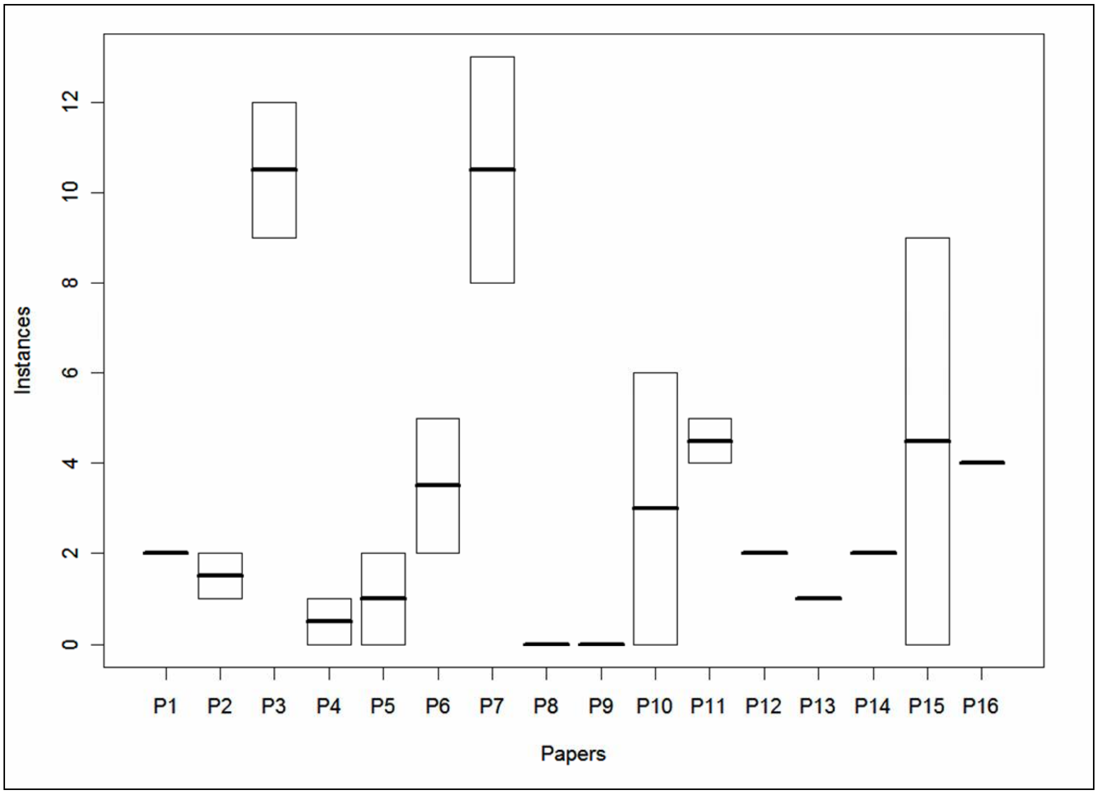


Figure 10. Boxplot of SMS consensus instances by paper

As Figure 11 shows, in only two of the 16 cases were the consensus ranges higher than those of the individual annotations: P5 and P6. All others had fewer instances (n=8), or the same number (n=6).

P8 and P9 remained unannotated, but 4 papers had matching non-zero instance counts (P1, P12, P13, and P16). P15 and P10 remained the papers with the largest variances.

The 90 consensus annotations made on the MMS papers yielded 1,258 instances, and had a range of 0-45, with the smallest non-zero value being 3. The mean and standard deviation of the consensus instances are 13.9 and 9.6, which are higher than those

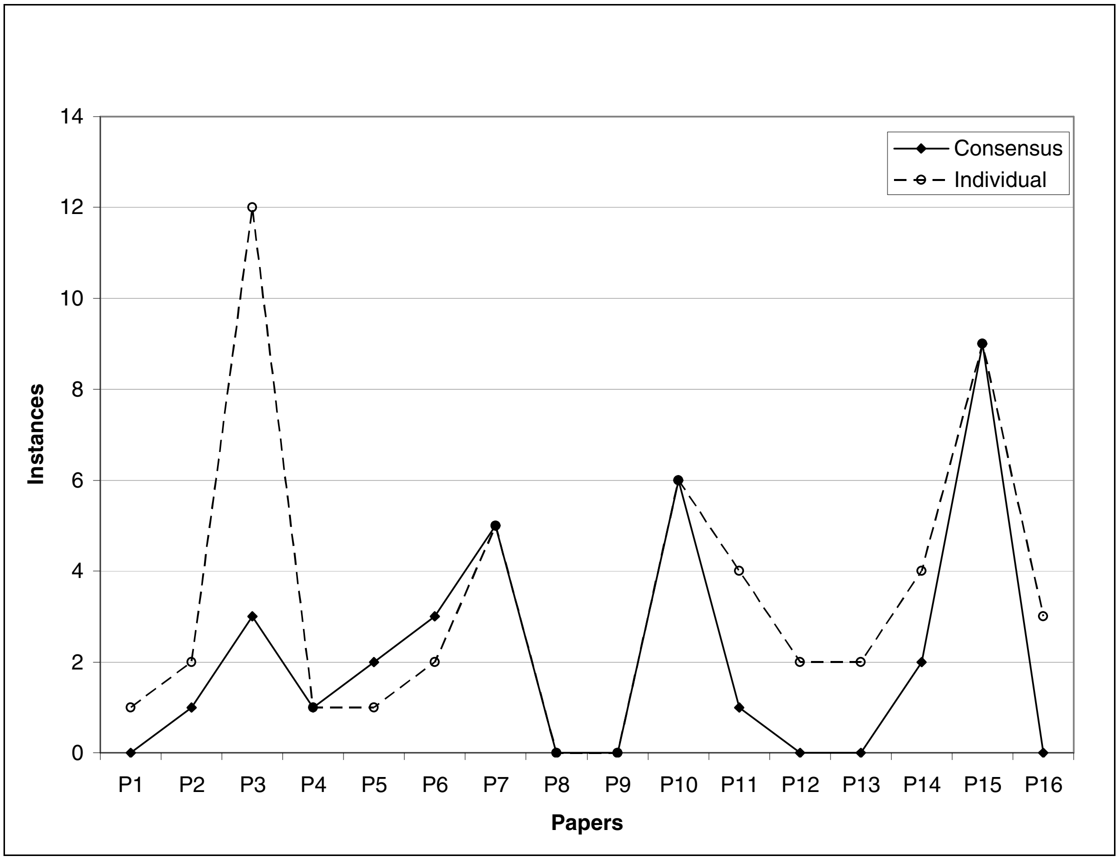


Figure 11. Comparison of instance ranges (individual and consensus, SMS)

measures for the individual annotations. For those papers with the smallest variation in instance counts (discussed above), two (P2 and P7) had slight increases in instances, while the third (P10) decreased slightly. Of the papers with the largest variation, P1 and P5 had slight increases in instance counts, while P3, P4, and P9 decreased slightly. Figure 12 shows a boxplot of the instances by paper. While the compression of ranges is not as large here as in the SMS, we can see that the number of cases with zero annotation instances has gone from nine in the individual instances (Figure 7) to one in the consensus annotations (P10).

Figure 13 shows that the nine consensus annotations per paper in the MMS had larger instance means in every case than did the counts for individual curators.

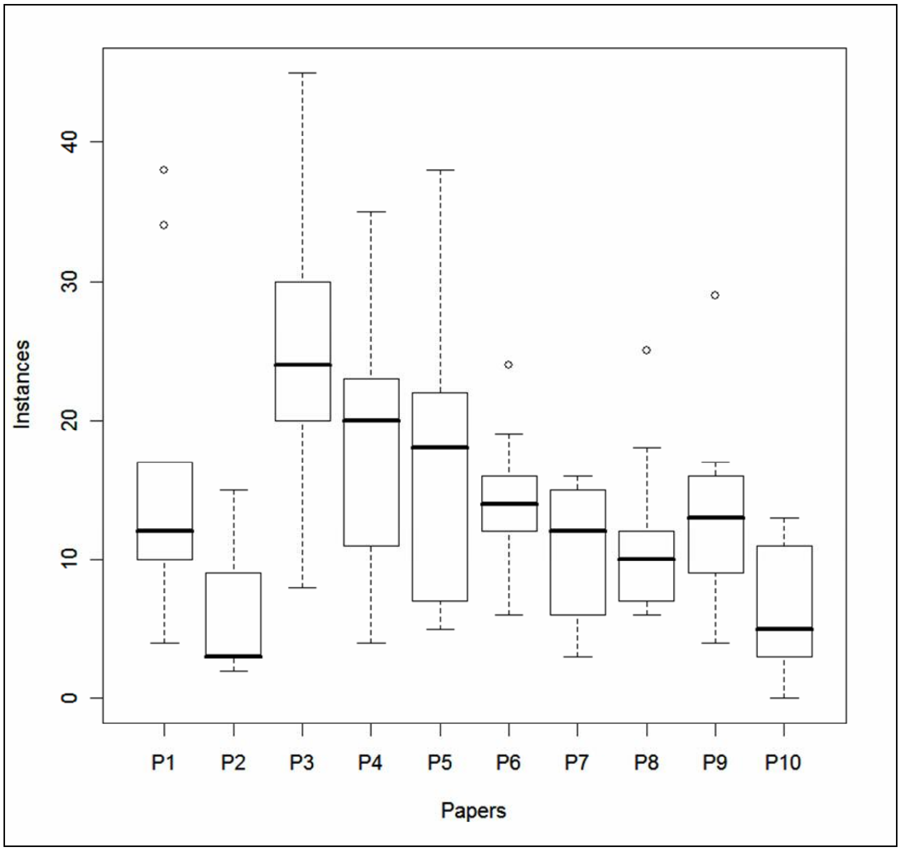


Figure 12. Boxplot of consensus instances by paper (MMS)

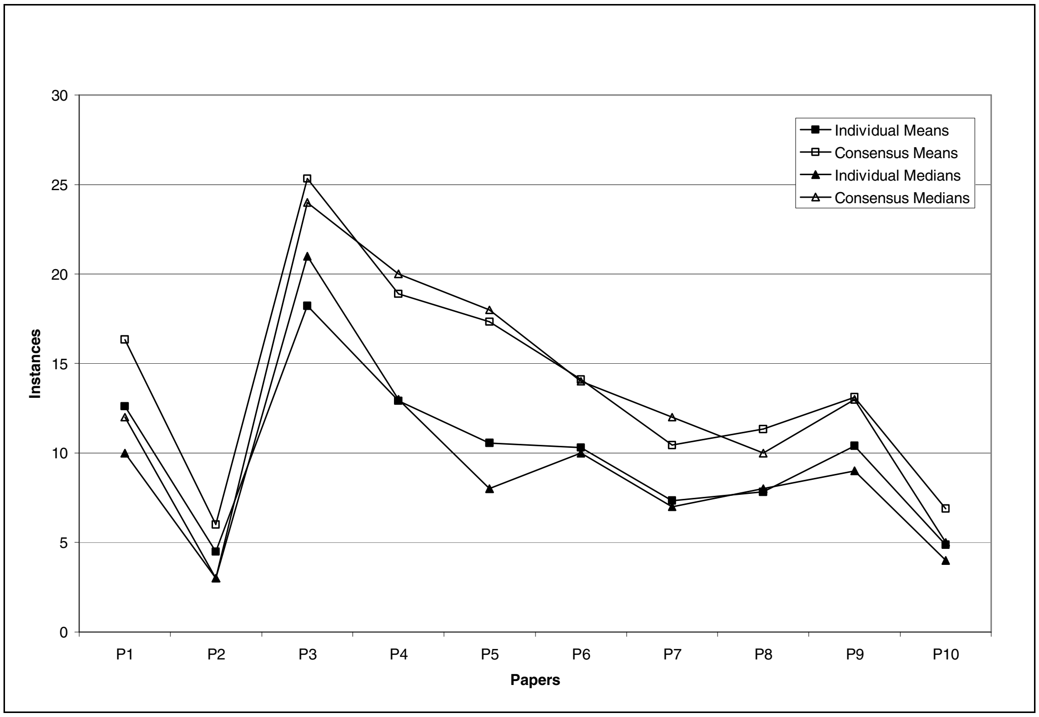


Figure 13. Mean and Median Values for Individual and Consensus Instances (MMS)

3. *Variation in annotation instances was observed by group assignment (SMS).*

As described in section 2.3, the single-MOD study participants were randomly assigned to one of two groups, and were then assigned papers. These assignments enabled analysis of differences in annotation instance quantities by group. We would expect that with random assignment, attributes such as years of experience would be distributed equally across groups. To verify this, Table 7 was constructed, which shows the differences in instance counts across groups, for the two- and four-fold coverage paper sets, combined values for both sets, and the consensus values.

Table 7. Instance count summary measures by group and coverage

	Total instances	Range (min-max)	Mean	SD
G1 individual 4x	107	13 (0-13)	3.3	3.4
G2 individual 4x	78	13 (0-13)	2.4	3.2
G1 individual 2x	45	7 (0-7)	1.4	1.7
G2 individual 2x	35	8 (0-8)	1.1	2.1
G1 individual combined	152	13 (0-13)	2.4	2.9
G2 individual combined	113	9 (0-9)	1.8	2.7
G1 consensus	59	11 (0-11)	3.7	3.6
G2 consensus	44	9 (0-9)	2.7	3.6
All 4x	185	11 (0-11)	2.9	3.3
All 2x	80	9 (0-9)	1.3	1.9
All consensus	103	11 (0-11)	3.2	3.6
All individual & consensus	368	9 (0-9)	2.3	3.0

The differences between groups appears large, particularly in the instance counts. To investigate whether one group had curators with more experience than the other, Table 8 was computed from the data extracted from individual curator interviews on years of GO experience, as described in section 3.7. The curators in Group 1 had more cumulative years of experience than those in Group 2, which could explain part of the observed inter-group variation. However, claim 4 below would seem to contradict this.

Table 8. Distribution of curators' years of GO annotation experience by group (SMS)

Yrs exp	Group 1	Group 2
0.5	0	1
1	0	0
2	0	1
3	1	0
4	2	1
5	1	1
Total yrs	16	11.5

4. Years of GO curation experience is not a predictor of instance counts.

Curators' instance counts were plotted against their years of GO annotation experience to determine whether more experience led to the creation of more or fewer instances. The plot of the eight SMS curators and their instance counts is shown in Figure 14. An R^2 value of 0.008 means that less than 1% of the variation in instances is explained by years of GO experience (p=0.31).

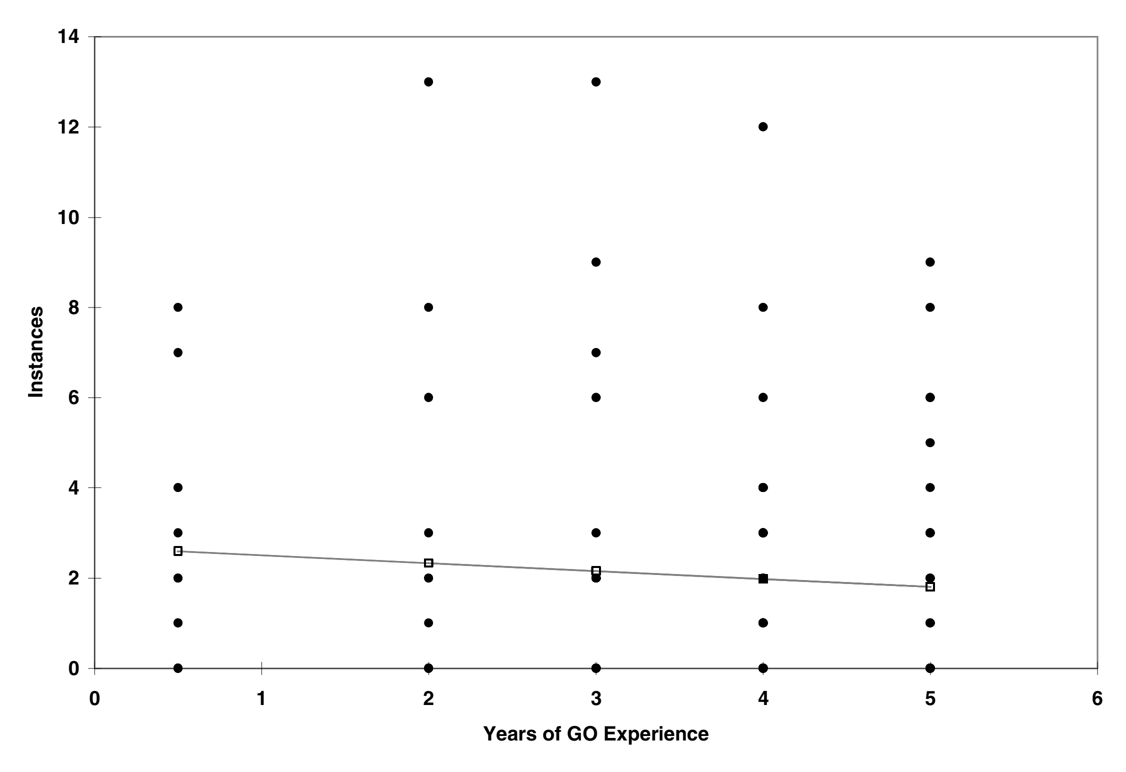


Figure 14. SMS individual instances by years of GO experience

A plot of 20 MMS curators and their instance counts against years of GO experience is shown in Figure 15 (the experience data is from Section 3.6). Data on experience was not available for two curators, and curator C23 was excluded from this analysis since he or she made annotations on only two of the ten papers. An R^2 value of 0.036 means that less than 4% of the variation in instances is explained by years of GO experience (*p*=0.007). As with the SMS findings, something other than experience is influencing instance variation.

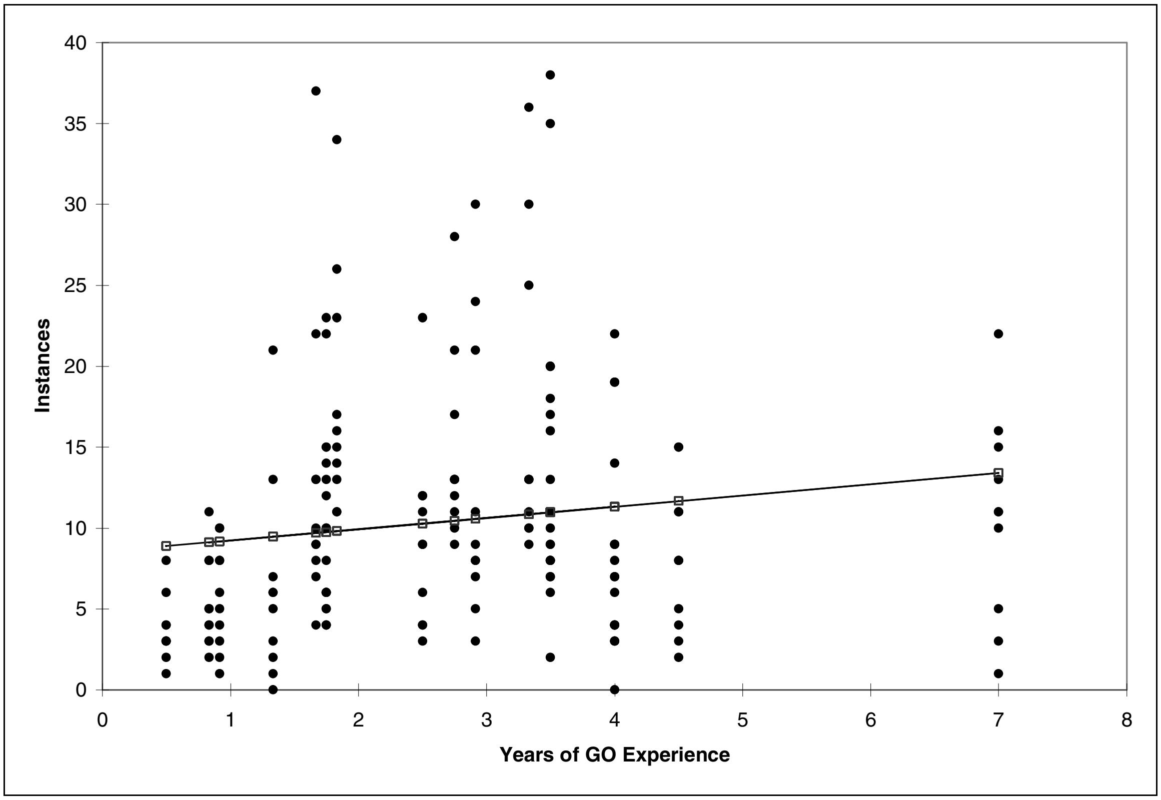


Figure 15. MMS individual instances by years of GO experience

5. Variation in annotation instances was observed by organism expertise (MMS).

In analysis of the MMS data, those curators representing the MODs whose organism was the subject of the paper were characterized as ‘organism experts’, while those who are not affiliated with that organism were characterized as ‘organism novices’. These labels vary for each paper as the subject organism of the paper changes. Table 9 is a crosstabulation of annotation instances by organism and curator expertise. Each row contains the cumulative instances for the two papers associated with each organism, the numbers of experts and novices providing annotations, and the mean number of instances per curator. The largest expert/novice differences are with human (more than twice as many instances made by experts), and mouse (60% more by experts).

Table 9. MMS individual instances by subject organism and curator expertise

	Experts			Novices		
Organism	Instances	# Experts	Instances / # Experts	Instances	# Novices	Instances / # Novices
Arabidopsis	75	4	18.8	405	19	21.3
human	141	5	28.2	210	18	11.7
mouse	141	7	20.1	208	16	13.0
worm	132	3	44.0	584	17	34.4
yeast	86	5	17.2	307	17	18.1

The individual and consensus annotation instance counts of the organism experts were plotted with the minimum, maximum, and mean instances per paper, as shown in Figure 16. In 5 of the 10 cases, the experts' consensus counts are greater than the mean.

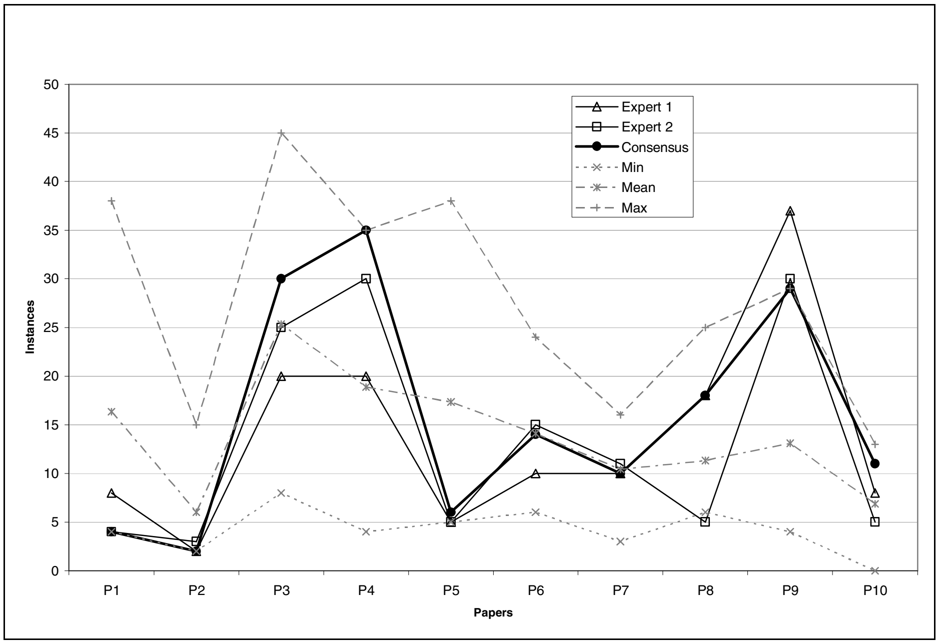


Figure 16. Organism experts' consensus annotation instances vs. mean instances (MMS)

6. *Curators exhibit high variability in the number of annotation instances they make, across papers.*

   The preceding claims clearly show that the numbers of instances per paper vary by curator. As we try to apprehend the origin of this variation, we are also interested in whether individual curators tend to have a consistent range of total annotation instances that they make, regardless of organism or paper. Figures 17 and 18 show plots of the SMS and MMS instances by curator to illustrate the intra-curator ranges across papers.

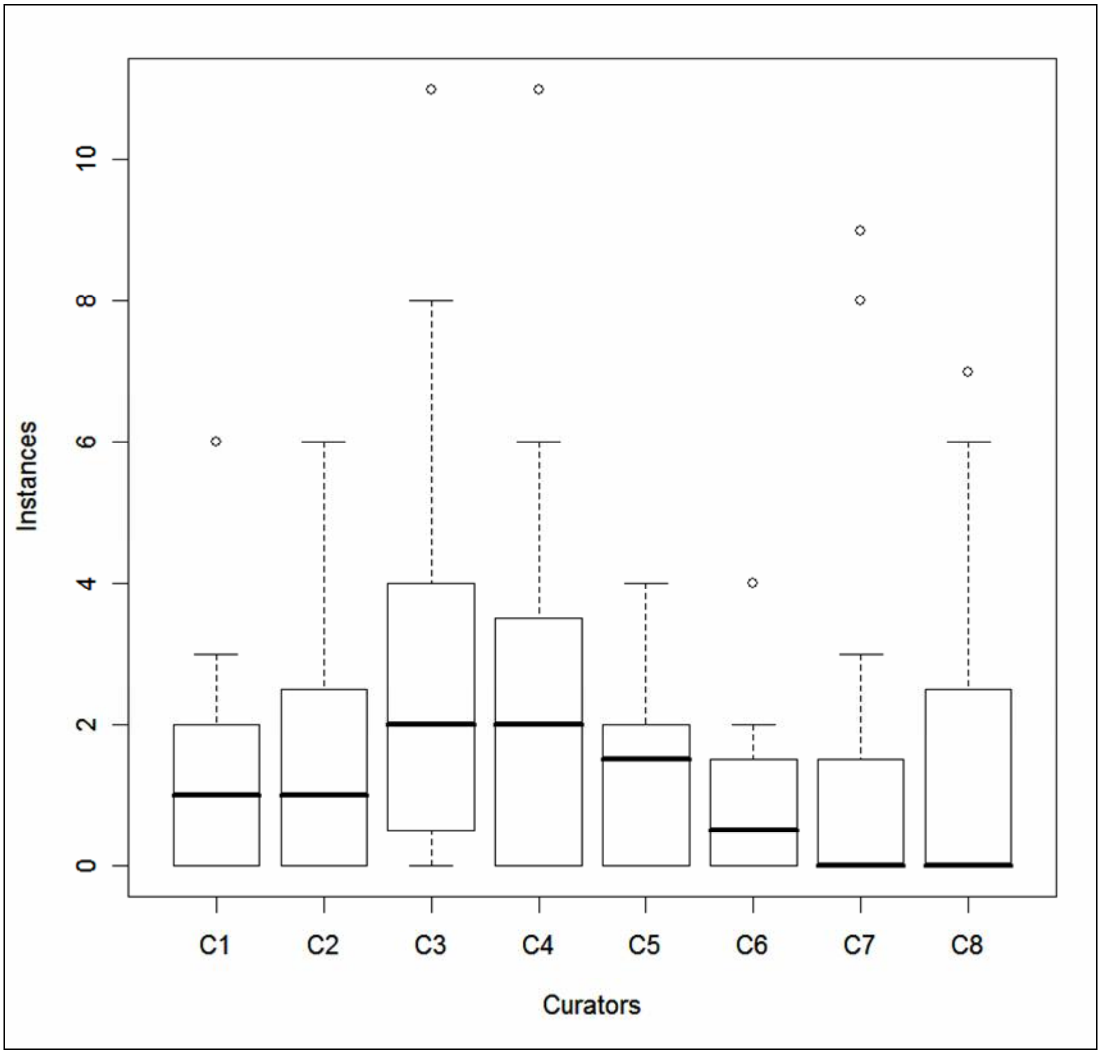


Figure 17. Boxplot of instances by curator (SMS)

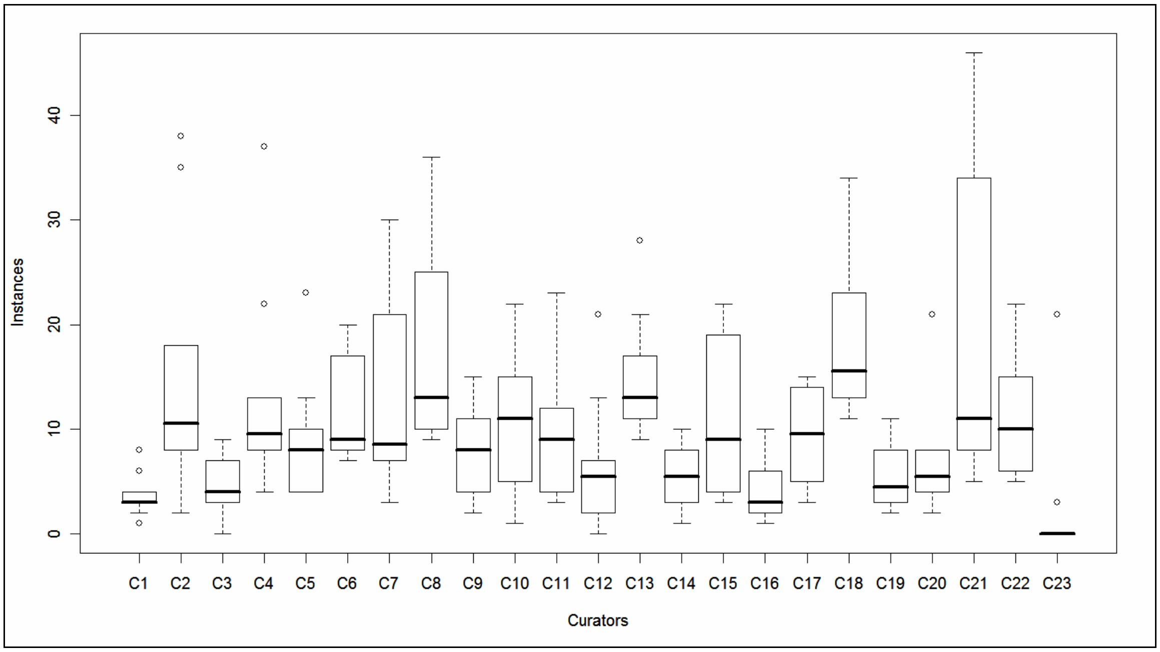


Figure 18. Boxplot of instances by curator (MMS)

3.3.2. Paper-level comparisons

7. Paper-level exact-match comparisons exhibit very low consistency.

At the highest level of pairwise comparison – the comparison of annotations at the paper level – very few exact matches across individual curators or consensus pairs were observed, in either study. 'Exact match' in this case means that 1) the numbers of annotation instances for all curators and consensus pairs were the same, and 2) all values for all elements of the GO annotation were the same.

In the single-MOD study, 3 out of the 16 4-fold coverage papers had exact matches at the consensus level, but in two of those cases, consistency came from no annotations being made. The 32 2-fold coverage papers had 11 exact matches, all of which came from no annotations being made.

In the multi-MOD study, 5 of the 10 papers had at least one pair of exact matches from the pool of 32 annotations (23 individual and 9 consensus). In the P1 and P5 sets, one individual annotation matched the subsequent consensus annotation to which the curator contributed. For P2, two curators' individual annotations matched that of their consensus pair, as did a third curator's. For P10, two curators did not make annotations, and one of their consensus pairs also declined to make an annotation, based on a technical aspect of the paper. (The other curator in the pair had originally created a single annotation instance, but that did not carry over to the consensus.)

8. *Consensus annotations exhibit low consistency with each other.*

In the single-MOD study, only two of the 16 pairs of consensus annotations matched across groups at the paper level; papers P1 and P13 had exact matches across all elements. Papers P8 and P9 each had zero instances from both consensus pairs. The pairs for P12, P14, and P16 each had the same number of instances, but differences in term selection, differing genes identified, and different evidence codes.

In the MMS, none of the consensus annotations had exact matches at the instance level.

3.3.3. Element-level comparisons

The claims in this section examine differences between the annotations from both studies at the GO element level.

9. *Variation is greatest in the GO Term element.*

Since the Ontology (aspect), Evidence Code, and Object (gene) elements of the formal GO annotation have relatively small possible values, and the numbers of terms in the three GO vocabularies are so large, it is not surprising that the GO Term element had the most observed variation. Figure 19 shows distributions by paper of the frequency of distinct terms assigned by the 23 curators in the MMS.

The variation in the numbers of total and distinct GO terms assigned per annotation reflect the variation seen above at the higher levels of comparison. Figure 20 shows mean values for the total and distinct term counts per paper, by individual, consensus, and combined groups. The consensus term counts are higher than the individuals for both total and distinct terms for all papers, suggesting that curators may be adding their non-overlapping terms rather than rationalizing them.

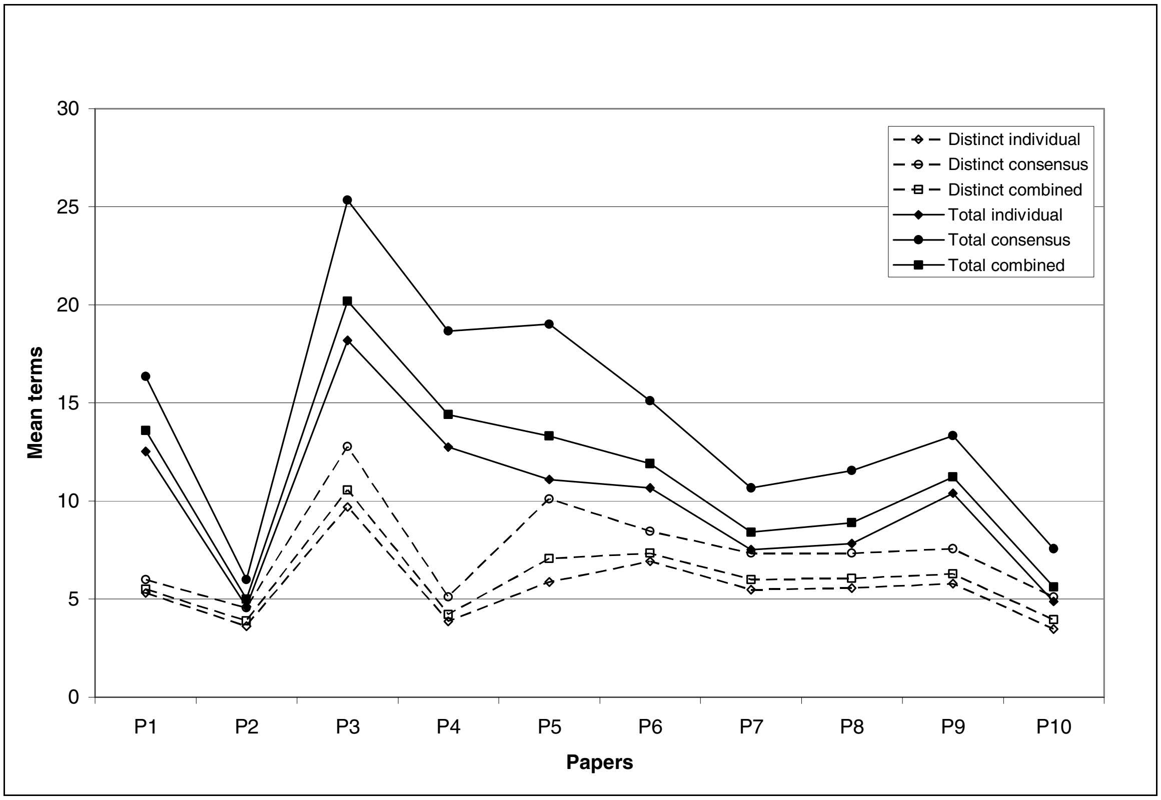


Figure 19. Mean total and distinct terms by paper (MMS)

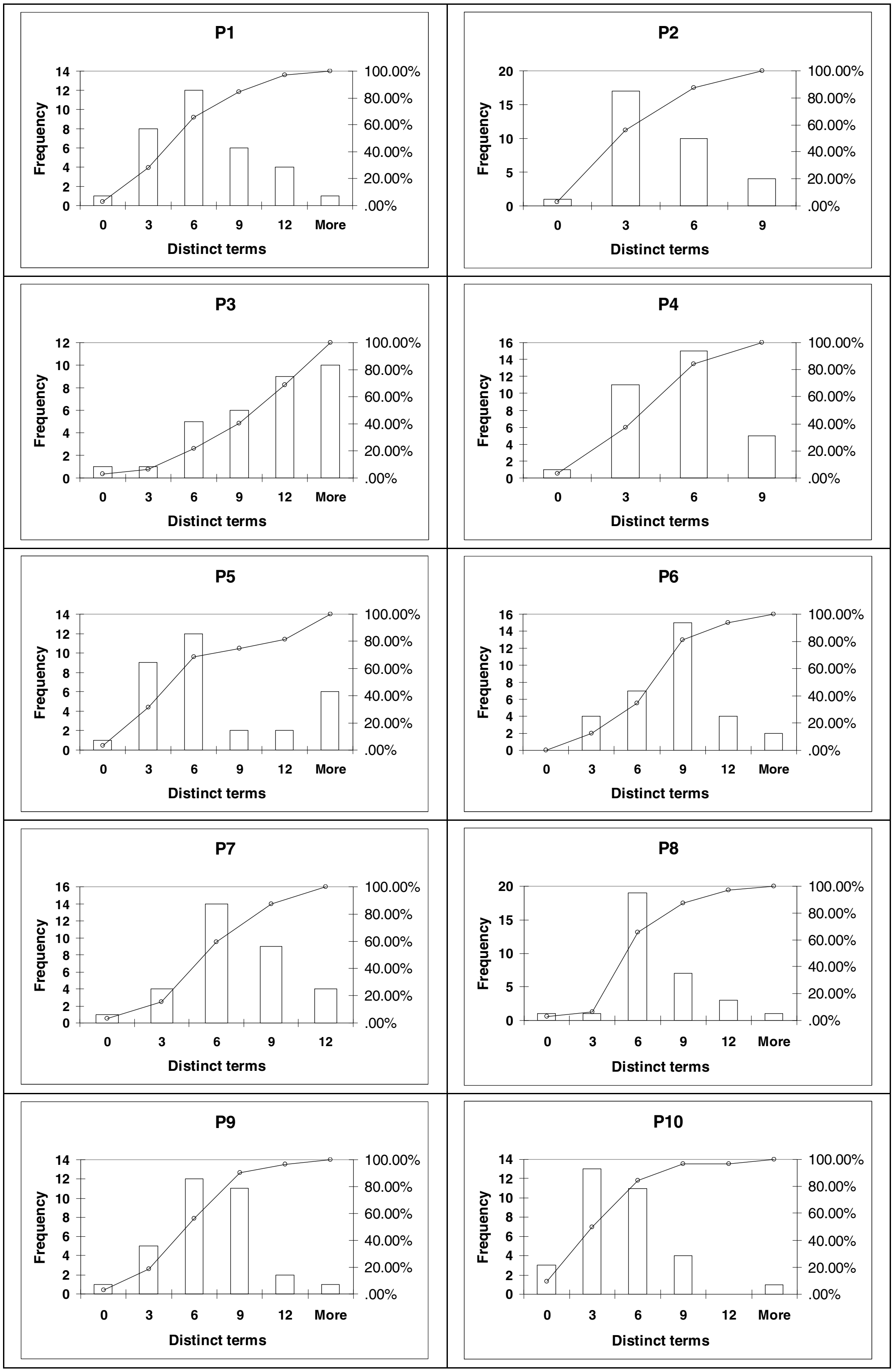


Figure 20. Frequency distributions of distinct terms (MMS papers)

Variation is present in curators' per-annotation GO term selections.

Where Figure 20 showed term variation from the paper perspective, Figure 21 shows the cumulative variation in individual curators' quantities of total and distinct terms across the full paper corpus. The 1,947 terms have a mean of 6.1, median of 5, standard deviation of 4, and a range of 26.

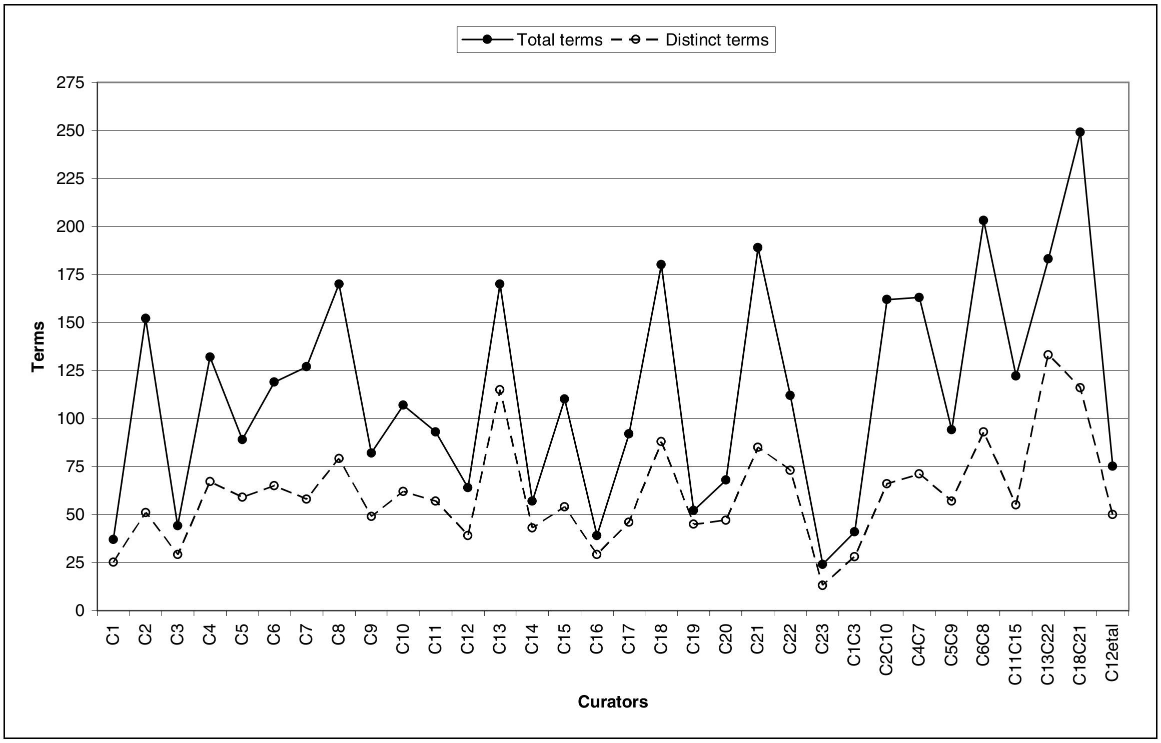


Figure 21. Curators' total and distinct GO terms (MMS)

3.3.4. Facet analysis

This section evaluates a sample of the annotation instances from both the single-MOD and multi-MOD studies against the five GO annotation quality facets defined in section 2.9. As discussed in section 2.1.1, a pilot study of the Reliability facet was conducted with two SMS curators who did not participate in the main study. Their original and repeated annotations are compared below. Their annotations are also compared using the Specificity facet. Two consensus annotations from the SMS and one from the MMS were selected in order to test the viability of the other four facets. The two SMS annotations were chosen by selecting one annotation that had matching instance counts (P12), and one that did not (P3). P12 had 4 total instances, while P3 had 21. The MMS annotation (P2) was chosen because it had a manageable number of total instances (54) for an example, while also having a similar range of instances (13) as half of the papers, and was close to the mean number of instances of the entire corpus (14).

Reliability

For paper RP1, curator R1 originally had four instances, but created only two upon reannotation, as shown in Table 10 (denoted O_n and R_n, respectively). After performing the reannotation task, the curator stated that the evidence for the IMP instances was not as strong as thought during the original reading, and so decided not to make them for the reannnotation.

Table 10. Original and repeated annotation instances for paper RP1 (curator R1)

i	Gene	Aspect	GOID	GO Term	EC	With/from	Qualifier
O_1	JHD1	F	46975	histone-lysine demethylase activity (H3-K36 specific)	IDA		
O_2	JHD1	F	46975	histone-lysine demethylase activity (H3-K36 specific)	IMP		
O_3	JHD1	P	16577	histone demethylation	IDA		
O_4	JHD1	P	16577	histone demethylation	IMP		
R_1	JHD1	F	46975	histone-lysine demethylase activity (H3-K36 specific)	IDA		
R_2	JHD1	P	16577	histone demethylation	IDA		

In this case, the only instance pairs with matching evidence codes are $\{O_1, R_1\}$ and $\{O_3, R_2\}$, so Reliability for RP1 is calculated as:

$P_1A_{O,R}C_{R1\ \mathrm{RELIABILITY}} = 0$

from:

	O_1R_1	O_3R_2	O_2	O_4	Paper level
Gene	1	1	0	0	0
GO aspect	1	1	0	0	0
GO ID	1	1	0	0	0
GO term	1	1	0	0	0
EC	1	1	0	0	0
With/from	-	-	-	-	-
Qualifier	-	-	-	-	-
Overall	1	1	0	0	0

where ‘0’ = mismatch, and ‘1’ = exact match, and ‘-’ = ‘not applicable’.

In this case, we can see that all of the elements in the matched instance pairs were consistent, but the presence of IMP annotations to the same terms in the original but not the reannotation led to inconsistencies. This yielded an overall mismatch at the paper level.

Table 11 shows the initial and subsequent annotations for paper RP2. Similar to RP1, the curator stated that upon a second reading of the paper, the evidence for the SSE1 Function annotation to ‘unfolded protein binding’ was not strong enough to warrant making that annotation. We can also see that while both sets of annotations have three Function instances, O4 is a Component annotation while R4 is a Process annotation. An additional difference is that all four original instances are annotated to the SSE1 gene, while only two of the reannotated instances are; the other two are to the YDJ1 gene. As a result, the only comparable instance pairs are $\{O_1, R_1\}$ and $\{O_2, R_2\}$.

Table 11. Original and repeated annotation instances for paper RP2 (curator R1)

i	Gene	Aspect	GOID	GO Term	EC	With/from	Qualifier
O_1	SSE1	F	5524	ATP binding	IDA		
O_2	SSE1	F	5524	ATP binding	IMP		
O_3	SSE1	F	51082	unfolded protein binding	IMP		
O_4	SSE1	C	5737	cytoplasm	IC	from 51082	
R_1	SSE1	F	5524	ATP binding	IDA		
R_2	SSE1	F	5524	ATP binding	IMP		
R_3	YDJ1	F	30188	chaperone regulator activity	IGI	YDJ1	
R_4	YDJ1	P	6457	protein folding	IC	from 30188	

Reliability for RP2 is calculated as:

$P_2A_{O,R}C_{R1}\ _{\text{RELIABILITY}} = 0$

from:

	O_1R_1	O_2R_2	O_3	O_4	Paper level
Gene	1	1	0	0	0
GO aspect	1	1	1	0	0
GO ID	1	1	0	0	0
GO term	1	1	0	0	0
EC	1	1	0	1	0
With/from	-	-	0	0	0
Qualifier	-	-	-	-	-
Overall	1	1	0	0	0

Table 12 shows the initial and subsequent annotations for paper RP3, annotated by curator R2. In this case, there were six original annotation instances, and seven subsequent. The reannotation set has two instances made to a fourth gene, DRS2, and one fewer made to CDC50 than the original. The comparable instance pairs are {O_1, R_1}, {O_2, R_2}, {O_4, R_5}, and {O_5, R_6}.

Table 12. Original and repeated annotation instances for paper RP3 (curator R2)

i	Gene	Aspect	GOID	GO Term	EC	With/from	Qualifier
O_1	CDC50	P	45332	phospholipid translocation	IGI	LEM3	
O_2	CDC50	C	5802	Glogi trans face	IDA		
O_3	CDC50	F	4012	phospholipid-translocating ATPase activity	IPI	DRS2	contributes to
O_4	DRS2	P	45332	phospholipid translocation	IGI	LEM3	
O_5	DRS2	C	5802	Glogi trans face	IDA		
O_6	LEM3	F	4012	phospholipid-translocating ATPase activity	IPI	DNF1	contributes to
R_1	CDC50	P	30866	actin patch organization	IMP		
R_2	CDC50	C	5802	trans-Golgi network	IDA	DRS2	colocalizes with
R_3	DNF1	C	5933	bud	IDA		
R_4	DNF1	C	5935	bud neck	IDA		
R_5	DRS2	P	30866	actin patch organization	IMP		
R_6	DRS2	C	5802	trans-Golgi network	IDA	CDC50	colocalizes with
R_7	LEM3	C	5933	bud	IDA		

Reliability for RP3 is calculated as:

$P_3A_{O,R}C_{R2\ \text{RELIABILITY}} = 0$

from:

	O_1R_1	O_2R_2	O_4R_5	O_5R_6	O_3	R_3	R_4	Paper level
Gene	1	1	1	1	0	0	0	0
GO aspect	1	1	1	1	0	0	0	0
GO ID	0	1	0	1	0	0	0	0
GO term	0	0	0	0	0	0	0	0
EC	0	1	0	1	0	0	0	0
With/from	0	0	0	0	0	-	-	0
Qualifier	-	0	-	0	0	-	-	0
Overall	0	0	0	0	0	0	0	0

For instance pairs {O_2, R_2} and {O_5, R_6}, the GO IDs matched, but in the six months between annotations, the term name had changed within the GO Component ontology, leading to a mismatch.

The initial and subsequent annotations for paper RP3 are shown in Table 13. The reannotation set has four instances compared with three in the original, with a new Function annotation made to the VTI1 gene.

Table 13. Original and repeated annotation instances for paper RP4 (curator R2)

i	Gene	Aspect	GO ID	GO Term	EC	With/from	Qualifier
O_1	PEP12	F	5486	t-SNARE activity	IDA		
O_2	SNC2	F	5485	v-SNARE activity	IDA		
O_3	TLG1	F	5486	t-SNARE activity	IDA		
R_1	PEP12	F	5486	t-SNARE activity	IDA		
R_2	SNC2	F	5485	v-SNARE activity	IDA		
R_3	TLG1	F	5486	t-SNARE activity	IDA		
R_4	VTI1	F	5486	t-SNARE activity	IDA		

Reliability for RP4 is calculated as:

$P_4A_{O,R}C_{R2\ \mathrm{RELIABILITY}} = 0$

from:

	O_1R_1	O_2R_2	O_3R_3	R_4	Paper level
Gene	1	1	1	0	0
GO aspect	1	1	1	0	0
GO ID	1	1	1	0	0
GO term	1	1	1	0	0
EC	1	1	1	0	0
With/from	-	-	-	-	0
Qualifier	-	-	-	-	-
Overall	1	1	1	0	0

While the three comparable instance pairs were perfectly consistent, the fourth reannotation instance led to a Reliability score of 0.

Consistency

For SMS paper P12, both curator pairs created two annotation instances, as shown in Table 14.

Table 14. SMS SP12 consensus annotation instances; curators SC2C4 and SC6C8

i	Pair	Gene	Aspect	GOID	GO Term	EC	With/from	Qualifier
1	C2C4	GRE2	P	6950	response to stress	IMP	-	-
	C6C8	GRE2	P	6950	resp to stress	IMP	-	-
2	C2C4	GRE2	P	6696	ergosterol biosynthesis	IMP	-	-
	C6C8	GRE2	P	8204	ergosterol metabolism	IMP	-	-

Consistency for SMS SP12 is thus calculated as:

$$\mathrm{P}_{12}\mathrm{A}_{\mathit{1,2}}\mathrm{C}_{\mathit{C2C4,C6C8}\ \mathrm{CONSISTENCY}} = 0$$

from:

	i_1	i_2	Paper level
Gene	1	1	1
GO aspect	1	1	1
GO ID	1	0	0
GO term	1	0	0
EC	1	1	1
With/from	-	-	-
Qualifier	-	-	-
Overall	1	0	0

In this case, we can see that all of the elements in the first pair of instances were exact matches, but the second pair had mismatches for the values of the GO ID and GO term elements. This yields an overall mismatch at the paper level.

For SMS SP3, in which the curator pairs had differing numbers of instances, significant manual manipulation of the annotation data was required prior to evaluation. Since GO annotations are focused on gene products, the instances were ordered by the Gene Name element, and cross-pair matches were searched for. Curator pair SC2C3 identified 6 distinct genes, while SC6C7 identified 4, as shown in Tables 15 and 16. SC2C3 made a total of 12 annotation instances, while SC6C7 made 9.

Table 15. SMS SP3 annotation instances for consensus pair SC2C3

Gene	Aspect	GOID	GO Term	EC	With/from	Qualifier
ARF1	P	6893	Golgi to plasma membrane transport	IPI	FMP50ǀCHS6	
ARF1	P	6893	Golgi to plasma membrane transport	IGI		
BCH1	P	6893	Golgi to plasma membrane transport	IGI		
BUD7	C	30140	trans-Golgi network transport vesicle	IDA		colocalizes_with
BUD7	P	6893	Golgi to plasma membrane transport	IGI		
CHS5	P	282	bud site selection	IMP		
CHS5	P	6893	Golgi to plasma membrane transport	IMP		
CHS5	P	6893	Golgi to plasma membrane transport	IGI		
CHS6	C	30140	trans-Golgi network transport vesicle	IDA		colocalizes_with
CHS6	P	6893	Golgi to plasma membrane transport	IGI		
FMP50	C	30140	trans-Golgi network transport vesicle	IDA		colocalizes_with
FMP50	P	6893	Golgi to plasma membrane transport	IGI		

Table 16. SMS SP3 annotation instances for consensus pair SC6C7

Gene	Aspect	GOID	GO Term	EC	With/from	Qualifier
BCH1	C	5798	Golgi-associated vesicle	IDA		
BCH1	P	6038	cell wall chitin biosynthesis	IGI		
BCH1	P	6038	cell wall chitin biosynthesis	IPI	CHS3	
BUD7	C	5798	Golgi-associated vesicle	IDA		
BUD7	P	6038	cell wall chitin biosynthesis	IGI		
BUD7	P	6038	cell wall chitin biosynthesis	IPI	CHS3	
CHS6	C	5798	Golgi-associated vesicle	IDA		
FMP50	C	5798	Golgi-associated vesicle	IDA		
FMP50	P	6038	cell wall chitin biosynthesis	IPI	CHS3	

The intersection of the two Gene sets is {BCH1, BUD7, CHS6, FMP50}, which leaves SC2C3 with two genes (ARF1 and CHS5) that are not shared. Only FMP50 has the same number of instances across consensus annotations; the other three genes have mismatches. Of the mismatched cases, SC2C3 has only one instance for BCH1, which is made to the GO Biological Process aspect, while SC6C7 has three: two to Process and one to Cellular Component. SC2C3 has 2 instances for BUD7, one Process and one Component, while SC6C7 has two Process and one Component. For the purposes of calculating the Consistency score, instances that were the closest matches were used, and the others were discarded. For example, for BCH1, SC2C3's instance was matched with the second SC6C7 instance, because it was to the same aspect (Process), and had the same Evidence Code (IGI). The full set of instance pairs that the Consistency score was calculated from is shown in Table 17. Instance pairs are labeled 'i_n' for reference purposes.

Table 17. Instance pairs from SMS SP3 for Consistency facet calculation

i	Pair	Gene	Aspect	GOID	GO Term	EC	With/from	Qualifier
1	C2C3	BCH1	P	6893	Golgi to plasma membrane transport	IGI		
	C6C7	BCH1	P	6038	cell wall chitin biosynthesis	IGI		
2	C2C3	BUD7	C	30140	trans-Golgi network transport vesicle	IDA		colocalizes _with
	C6C7	BUD7	C	5798	Golgi-associated vesicle	IDA		
3	C2C3	BUD7	P	6893	Golgi to plasma membrane transport	IGI		
	C6C7	BUD7	P	6038	cell wall chitin biosynthesis	IGI		
4	C2C3	CHS6	C	30140	trans-Golgi network transport vesicle	IDA		colocalizes _with
	C6C7	CHS6	C	5798	Golgi-associated vesicle	IDA		
5	C2C3	FMP50	C	30140	trans-Golgi network transport vesicle	IDA		colocalizes _with
	C6C7	FMP50	C	5798	Golgi-associated vesicle	IDA		
6	C2C3	FMP50	P	6893	Golgi to plasma membrane transport	IGI		
	C6C7	FMP50	P	6038	cell wall chitin biosynthesis	IPI	CHS3	

From these instance matches, Consistency is calculated as:

$P_3A_{1\text{-}6}C_{C2C3,C6C7\ \mathrm{CONSISTENCY}} = 0$

from:

	i_1	i_2	i_3	i_4	i_5	i_6	**Paper level**
Gene	1	1	1	1	1	1	1
GO aspect	1	1	1	1	1	1	1
GO ID	0	0	0	0	0	0	0
GO term	0	0	0	0	0	0	0
EC	1	1	1	1	0	0	0
With/from	-	-	-	-	-	0	0
Qualifier	-	0	-	0	0	-	-
Overall	0	0	0	0	0	0	0

Only the Gene and aspect elements had exact matches across all six instances.

The nine consensus annotations for MMS MP2 are included as Appendix 10. The expert consensus annotation from MC1C3 was used as the reference standard against which to compare the other eight annotations. Applying the same strategy of matching genes as was used for SMS SP3 yielded a set of 16 instance pairs, also attached in Appendix 10. From those instances, Consistency is calculated as:

$P_2A_{1\text{-}16}C_{O1\text{-}O9\ \mathrm{CONSISTENCY}} = 0$

from:

Ref: C1C3	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	i_{11}	i_{12}	i_{13}	i_{14}	i_{15}	i_{16}	**P**
Gene	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GO aspect	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GO ID	1	1	1	0	1	0	0	0	1	0	0	0	0	1	1	1	0
GO term	1	1	1	0	1	0	0	0	1	0	0	0	0	1	1	1	0
EC	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0
With/from	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Qualifier	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Overall	0	1	1	0	1	0	0	0	1	0	0	0	0	1	1	1	0

All of the consensus annotations had instances that matched the reference annotation's Gene and GO Aspect element values. Fourteen of the sixteen matched the Evidence Code, while only half matched the GO ID / terms. While the subset of instances used for this calculation were relatively consistent, it is important to remember that of the original 54 total instances from the nine consensus annotations, less than a third (16) were used in the Consistency calculation.

Specificity

Using the matched instances from Tables 10 - 13, Specificity for papers RP1 - 4 is calculated as:

$P_1A_{O,R}C_{R1 \text{ SPECIFICITY}} = 1$

from:

	O_1R_1	O_3R_2
Exact match	1	1
Mismatch, same lineage	-	-
Mismatch, different lineage	-	-
Broader	-	-
Narrower	-	-
Difference, in nodes	-	-
Nearest common ancestor	-	-

$P_2A_{O,R}C_{R1 \text{ SPECIFICITY}} = 1$

from:

	O_1R_1	O_2R_2
Exact match	1	1
Mismatch, same lineage	-	-
Mismatch, different lineage	-	-
Broader	-	-
Narrower	-	-
Difference, in nodes	-	-
Nearest common ancestor	-	-

$P_3A_{O,R}C_{R2 \text{ SPECIFICITY}} = 0$

from:

	O_1R_1	O_2R_2	O_4R_5	O_5R_6
Exact match	0	1	0	1
Mismatch, same lineage	-	-	-	-
Mismatch, different lineage	1	-	1	-
Broader	-	-	-	-
Narrower	1	-	1	-
Difference, in nodes	6	-	6	-
Nearest common ancestor	9987	-	9987	-

$P_4A_{O,R}C_{R2\ \mathrm{SPECIFICITY}} = 1$

from:

	O_1R_1	O_2R_2	O_3R_3
Exact match	1	1	1
Mismatch, same lineage	-	-	-
Mismatch, different lineage	-	-	-
Broader	-	-	-
Narrower	-	-	-
Difference, in nodes	-	-	-
Nearest common ancestor	-	-	-

In three of the four papers, the terms were exact matches. In RP3, the subsequent annotation for the pairs $\{O_1, R_1\}$ and $\{O_4, R_5\}$ was to a narrower term on a different branch from the original term.

Using the same instance pairs for SMS SP12 as above, we can calculate the Specificity score as:

$P_{12}A_{1,2}C_{C2C4,C6C8\ \mathrm{SPECIFICITY}} = 0$

from:

	i_1	i_2
Exact match	1	0
Mismatch, same lineage	-	1
Mismatch, different lineage	-	-
Broader	-	-
Narrower	-	1
Difference, in nodes	-	1
Nearest common ancestor	-	-

In the case of the first pair of instances, the GO terms were exact matches. The second pair were mismatches, but were adjacent on the GO tree – ‘ergosterol biosynthesis’ is a child term of ‘ergosterol metabolism’.

For SMS SP3, Specificity is calculated as:

$P_3A_{\mathit{1\text{-}6}}C_{\mathit{C2C3,C6C7}\ \text{SPECIFICITY}} = 0$

from:

	i_1	i_2	i_3	i_4	i_5	i_6
Exact match	0	0	0	0	0	0
Mismatch, same lineage	-	1	-	1	1	-
Mismatch, different lineage	1	-	1	-	-	1
Broader	-	1	-	1	1	-
Narrower	1	-	1	-	-	1
Difference, in nodes	11	3	11	3	3	11
Nearest common ancestor	GO: 0016043	GO: 0016023	GO: 0016043	GO: 0016023	GO: 0016023	GO: 0016043

The terms for instance pairs i_1, i_3, and i_6 were on different branches separated by 11 nodes, while those for i_2, i_4, and i_5 were separated by 3 nodes.

For MMS MP2, Specificity is calculated as:

$P_2A_{\mathit{1\text{-}16}}C_{\mathit{O1\text{-}O9}\ \text{SPECIFICITY}} = 0$

from:

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	i_{11}	i_{12}	i_{13}	i_{14}	i_{15}	i_{16}
Exact match	1	1	1	0	1	0	0	0	1	0	0	0	0	1	1	1
Mismatch, same lineage	-	-	-	1	-	1	1	1	-	1	1	1	1	-	-	-
Mismatch, different lineage	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Broader	-	-	-	1	-	1	1	1	-	-	-	-	-	-	-	-
Narrower	-	-	-	-	-	-	-	-	-	1	1	1	1	-	-	-
Difference, in nodes	-	-	-	0	-	0	0	0	-	1	1	1	0	-	-	-
Nearest common ancestor	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

In the case of the first instance, ‘ergosterol biosynthesis’, 4 pairs matched the expert consensus annotation, while the other four chose a broader adjacent term, ‘ergosterol metabolism’. The second instance’s annotations were to three adjacent terms:

```
GO:0006950 : response to stress
      GO:0006970 : response to osmotic stress
            GO:0009651 : response to salt stress
```

with the broadest term assigned by the MOD representatives and the narrower terms by the other participating curators.

Completeness

Given the instances above, Completeness for SMS SP12 is calculated as:

$P_{12}A_{\mathit{1,2}}C_{\mathit{C2C4,C6C8}\ \mathrm{COMPLETENESS}} = 1$

from:

Ref: C2C4	**C6C8**
False positives	0
False negatives	0
Overall	1

Note that the semantic differences (differences in element values) between instances, such as the GO Term differences in i_2, are evaluated by the Consistency measure; Completeness is interested only in instance counts.

Completeness for SMS SP3 is calculated as:

$P_{3}A_{\mathit{1\text{-}9}}C_{\mathit{C2C3,C6C7}\ \mathrm{COMPLETENESS}} = 0$

from:

Ref: C2C3	**C6C7**
False positives	3
False negatives	-6
Overall	0

where the three false positives are two annotations to BCH1 (one Process, one Component) and one to BUD7 (Process), and the six false negatives are the two ARF1 annotations, the three to CHS5, and the one Process annotation to CHS6.

For MMS MP2, Completeness is calculated as:

$P_2A_{\mathit{1\text{-}16}}C_{\mathit{O1\text{-}O9}\ \mathrm{COMPLETENESS}} = 0$

from:

Ref: C1C3	O2	O3	O4	O5	O6	O7	O8	O9
False positives	1	2	1	13	1	7	10	1
False negatives	0	0	0	0	0	0	0	0
Overall	0	0	0	0	0	0	0	0

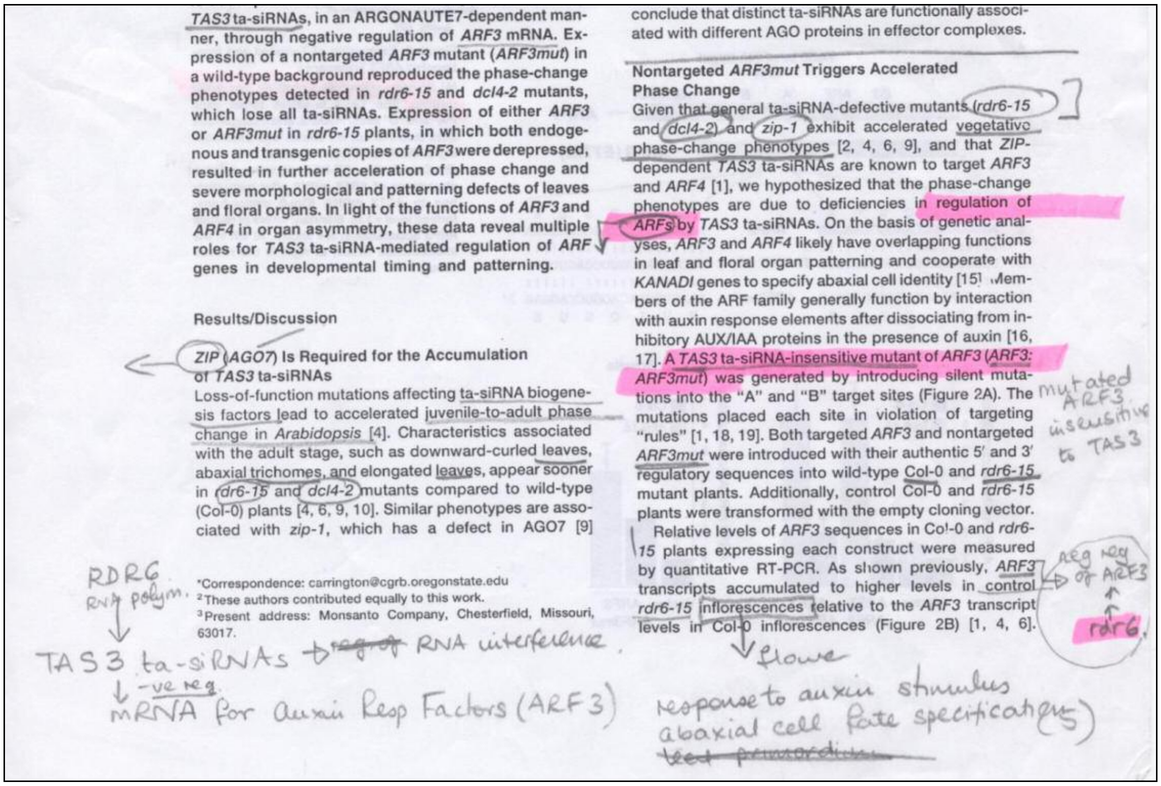
In this case, none of the compared annotations had false negative instances, only false positives.

Validity

To assess Validity, the collections of all consensus annotations from both the SMS and MMS were examined. There was one instance in the SMS and 51 in the MMS that contained a variety of invalid element values, and thus were not used in the analysis of instance counts. These amounted to <1% of the total number of consensus instances obtained in the SMS, and 4% in the MMS. The most common sources of errors were proposed GO terms that lacked direct Gene element referents, and ambiguous relationships to the other instances provided.

3.4. Curators' personal document annotations (MMS)

This section analyzes the personal manual annotations made by a sample of MMS curators (n=6) on the paper copies of the articles used in that study. Figure 22 illustrates an example of the quantity and variety of personal annotations that were made by curators on the document set when curating for GO. (Other types of curation activities could yield different annotations.) This example includes multiple writing media (pen and highlighter), and multiple annotation types, including underlining, highlighting, circling, arrows, and textual notes.



TAS3 ta-siRNAs, in an ARGONAUTE7-dependent manner, through negative regulation of *ARF3* mRNA. Expression of a nontargeted *ARF3* mutant (*ARF3mut*) in a wild-type background reproduced the phase-change phenotypes detected in *rdr6-15* and *dcl4-2* mutants, which lose all ta-siRNAs. Expression of either *ARF3* or *ARF3mut* in *rdr6-15* plants, in which both endogenous and transgenic copies of *ARF3* were derepressed, resulted in further acceleration of phase change and severe morphological and patterning defects of leaves and floral organs. In light of the functions of *ARF3* and *ARF4* in organ asymmetry, these data reveal multiple roles for *TAS3* ta-siRNA-mediated regulation of *ARF* genes in developmental timing and patterning.

Results/Discussion

ZIP (_AGO7_) Is Required for the Accumulation of _TAS3_ ta-siRNAs

Loss-of-function mutations affecting ta-siRNA biogenesis factors lead to accelerated juvenile-to-adult phase change in *Arabidopsis* [4]. Characteristics associated with the adult stage, such as downward-curled leaves, abaxial trichomes, and elongated leaves, appear sooner in *rdr6-15* and *dcl4-2* mutants compared to wild-type (Col-0) plants [4, 6, 9, 10]. Similar phenotypes are associated with *zip-1*, which has a defect in AGO7 [9]

*Correspondence: carrington@cgrb.oregonstate.edu
[2] These authors contributed equally to this work.
[3] Present address: Monsanto Company, Chesterfield, Missouri, 63017.

conclude that distinct ta-siRNAs are functionally associated with different AGO proteins in effector complexes.

Nontargeted _ARF3mut_ Triggers Accelerated Phase Change

Given that general ta-siRNA-defective mutants (*rdr6-15* and *dcl4-2*) and *zip-1* exhibit accelerated vegetative phase-change phenotypes [2, 4, 6, 9], and that *ZIP*-dependent *TAS3* ta-siRNAs are known to target *ARF3* and *ARF4* [1], we hypothesized that the phase-change phenotypes are due to deficiencies in regulation of ARFs by *TAS3* ta-siRNAs. On the basis of genetic analyses, *ARF3* and *ARF4* likely have overlapping functions in leaf and floral organ patterning and cooperate with *KANADI* genes to specify abaxial cell identity [15]. Members of the ARF family generally function by interaction with auxin response elements after dissociating from inhibitory AUX/IAA proteins in the presence of auxin [16, 17]. A *TAS3* ta-siRNA-insensitive mutant of *ARF3* (*ARF3:ARF3mut*) was generated by introducing silent mutations into the "A" and "B" target sites (Figure 2A). The mutations placed each site in violation of targeting "rules" [1, 18, 19]. Both targeted *ARF3* and nontargeted *ARF3mut* were introduced with their authentic 5′ and 3′ regulatory sequences into wild-type Col-0 and *rdr6-15* mutant plants. Additionally, control Col-0 and *rdr6-15* plants were transformed with the empty cloning vector.

Relative levels of *ARF3* sequences in Col-0 and *rdr6-15* plants expressing each construct were measured by quantitative RT-PCR. As shown previously, *ARF3* transcripts accumulated to higher levels in control *rdr6-15* inflorescences relative to the *ARF3* transcript levels in Col-0 inflorescences (Figure 2B) [1, 4, 6].

mutated ARF3 insensitive to TAS3

Neg reg of ARF3
rdr6

flowe

RDR6
RNA polym.

TAS3 ta-siRNAs ~~reg of~~ RNA interference.
-ve reg
mRNA for Auxin Resp Factors (ARF3)

response to auxin stimulus
abaxial cell fate specification(s)
~~leaf primordium~~

Figure 22. Partial manually-annotated document (MMS P5, PMID: 166822356)

1. *Curators' annotations are of a limited number of types, and follow similar styles.*

The six curators studied here used similar types of personal notes when manually annotating the paper copies of the 10 common articles used in the study. The six note types observed were: 1) *highlight* (semi-transparent color marker used to select text), 2) *underline* (single or multiple lines drawn under text or other elements), 3) *circle* (line drawn around text or other elements), 4) *emphasis* (vertical mark drawn next to block of text), 5) *textual note* (words or phrase written in page margins or within text), and 6) *arrow* (drawn to link extant text to notes, or as an indicator of importance). Most curators used all six types at various times, though not always within the same article.

Highlighting, underlining, and textual notes were the most common marks observed. Highlighting and underlining are similar approaches to selecting text, and were often used interchangeably. Some curators clearly preferred one over the other – one curator never used highlights, but underlined extensively – but others used both within the same paper.

2. *Numbers of personal annotation instances varied widely by curator.*

Despite the use of similar annotation types (and the common underlying work task), the curators varied widely in the numbers of annotations instances made per paper. Figure 23 shows the total number of annotations made per paper by each curator to illustrate the amount of variation. Overall, curators 7, 8, and 9 each made nearly twice as many (or more) notes per paper than curators 3, 4, or 6.

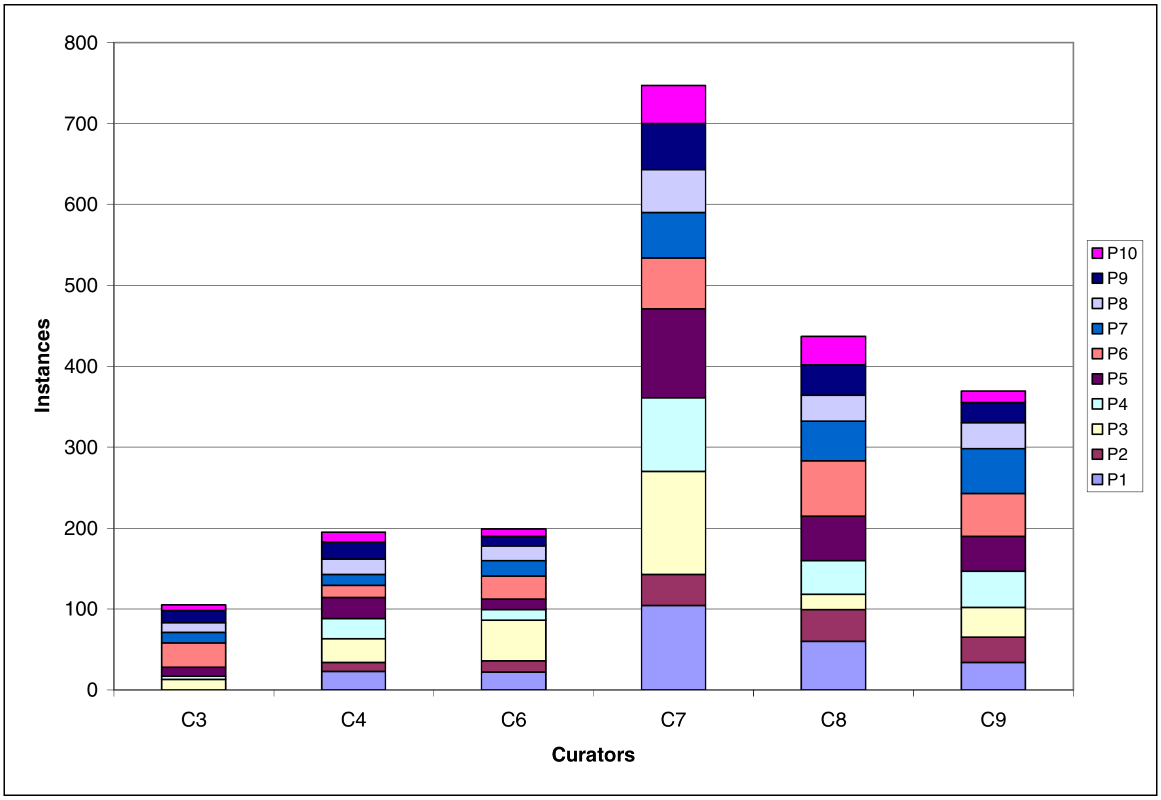


Figure 23. Total annotation instances per curator by paper

Curators also used varying numbers of the six different types of annotations described above. Figure 24 shows boxplots of the curators' annotations by type. The *note* type was the only kind of annotation that all curators employed for one or more papers; each of the other types had at least one curator who did not use it for this set of papers. While Curator 7, who made the most overall annotations, usually has the largest range, Figure 24a shows that the 'highlight' type was the least frequently used by that curator. Underlines were made most frequently by only two of the six curators, C7 and C8, as were circles. Emphasis and arrows were made least frequently, while underlines and highlights were the most often made.

3. *Quantities of curators' annotations varied by paper section.*

Figure 25 shows the cumulative annotations made by all curators to the set of 10 papers, by paper section. For this analysis, the names of paper sections were normalized to 'Abstract', 'Introduction', 'Methods', 'Results', 'Discussion', and 'References', which the majority of the papers followed quite closely. The Results section is clearly where the majority of annotations are made, with nearly 10 times more instances (1,528 total) than the other leading sections, Discussion (169), Introduction (163), and Abstract (152). Methods and References had very few by comparison (23 and 12, respectively). These results partially confirm the curators' self-reported annotation behavior (Section 3.5) when they say they tend to focus on Results and Abstracts when curating papers.

The Acknowledgements section of each paper was examined (when present), but is not shown here as no curator made annotations to it in any of the papers. Paper P10 was the only paper whose title was annotated, by two curators.

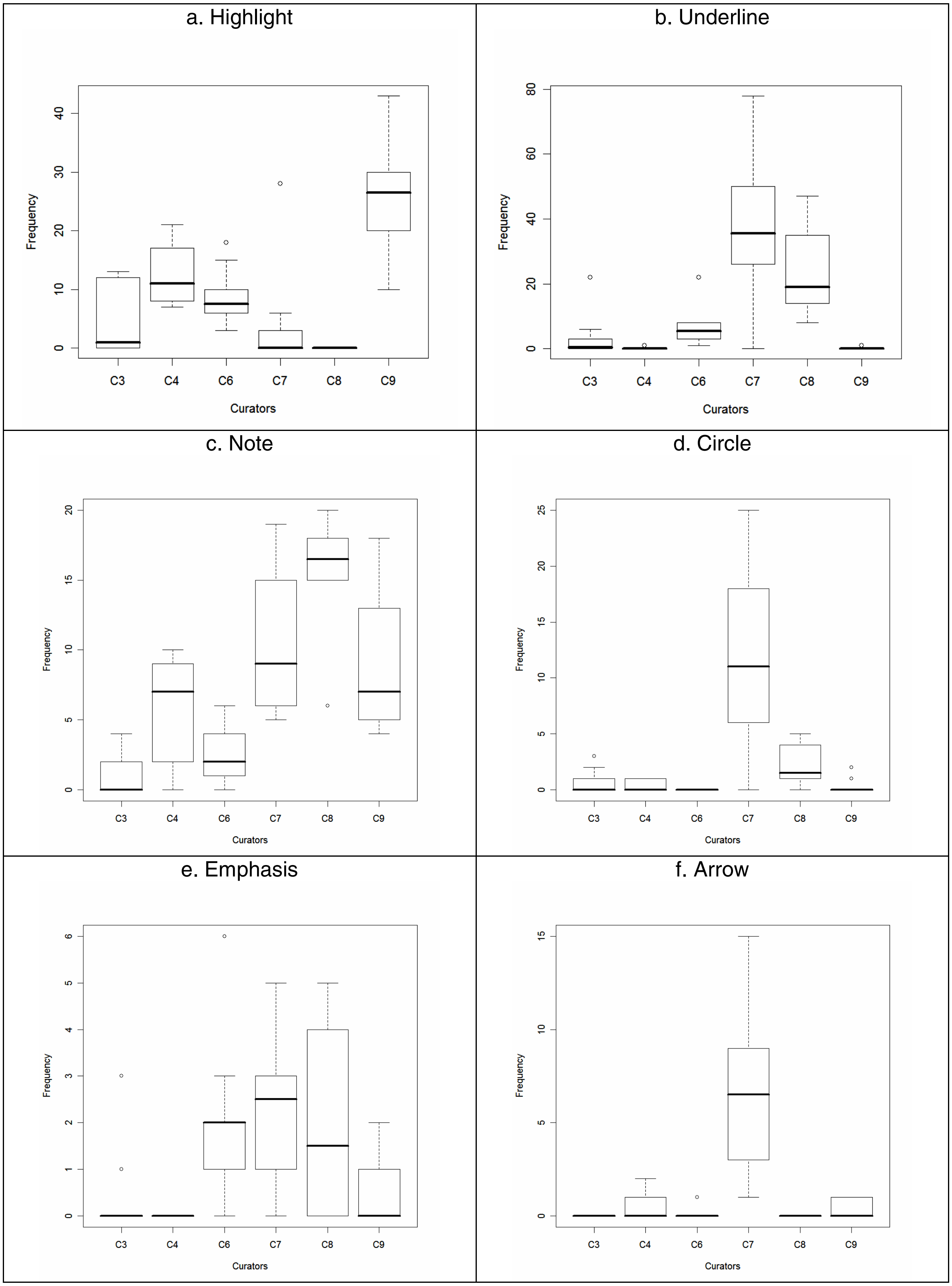


Figure 24a-f. Boxplots of curators' annotations by type (all papers)

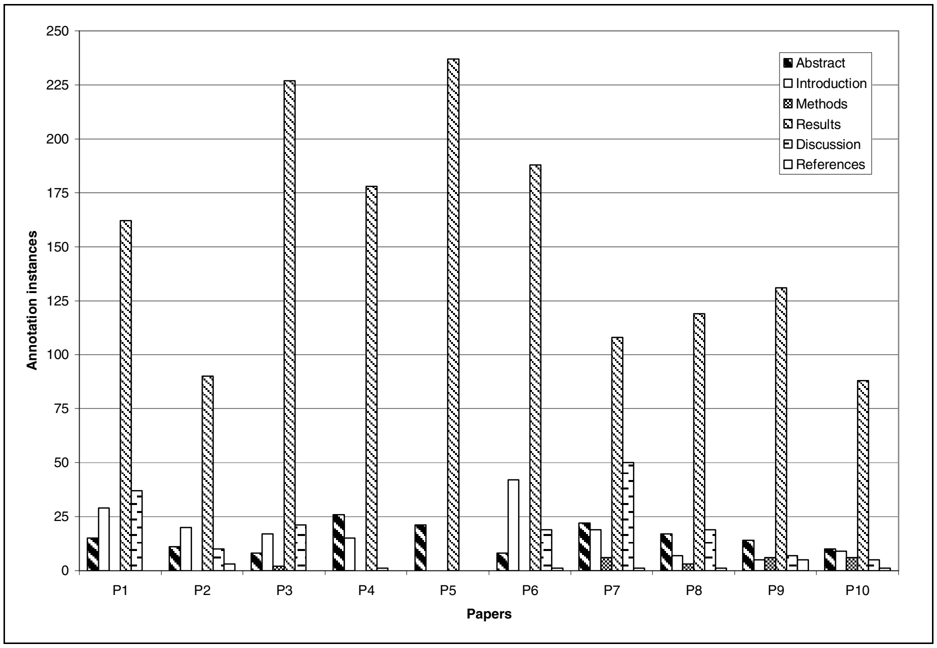


Figure 25. Cumulative annotation instances by paper section (all curators)

4. Personal annotation behavior explains only 9% of instance variation.

To investigate whether personal annotation behavior is related to variation in GO instance counts, the scatterplot in Figure 26 was constructed from the personal annotation counts and GO instance counts of the six curators who participated in this portion of the MMS. An R^2 value of 0.096 means that less than 10% of the variation in GO instances is explained by personal annotation instances (*p*=0.016).

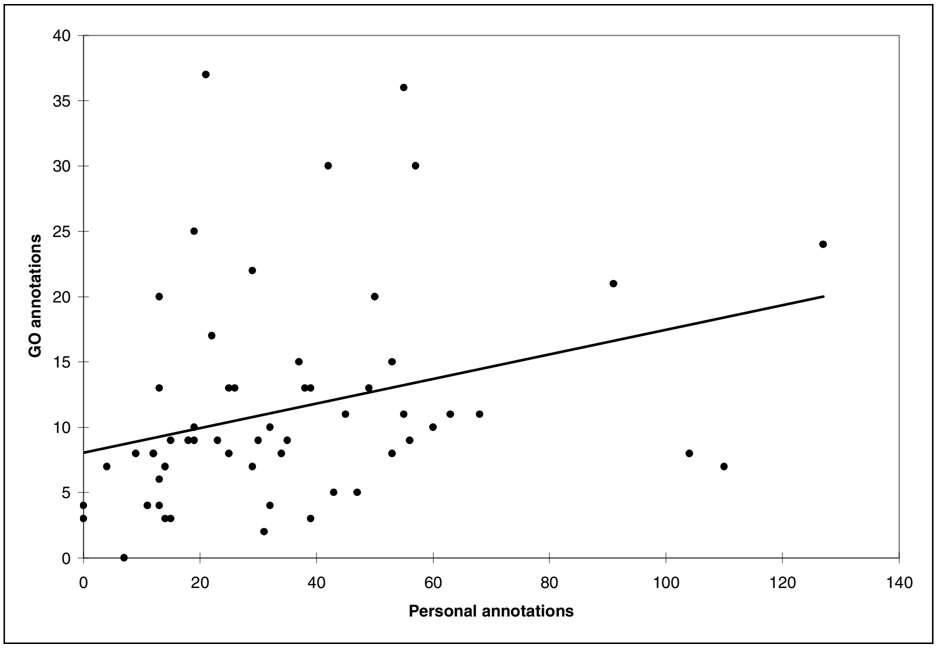


Figure 26. MMS individual GO instances by personal instances

3.5. Workflow analysis

This section describes the analysis of the curator interviews, with regard to differences among the individual work practices curators use when creating GO annotations. A summary of three standardized questions is presented, followed by direct quotes from curators contrasting their specific work practices and experiences.

Article format, ontology browsers, and related resources

Eleven of the fifteen curators interviewed always or almost always curate from electronic versions of articles. Most have large, high-resolution monitors that remove readability issues from the decision to read on the screen versus paper. The main benefits they see to doing so include the ability to search for gene names and other terms in order to find all of the instances within a paper, and the ability to access supplementary information and other resources more easily. Most of the SMS interviewees used the GO browser built into their MOD to access the GO vocabularies while curating, but other frequent choices by both SMS and MMS curators were QuickGO (EBI, 2007), the MGI browser from the Jackson Labs (Jackson Laboratories, 2007), which was used even by non-mouse curators, and AmiGO (GO Browser, 2007). Other reference materials and resources frequently used by curators when annotating included their own MODs, the NCBI Bookshelf (NCBI, 2007) resources, PubMed, Google, Wikipedia, and standard biology and biochemistry textbooks.

Quotes from curators in the following section are cited in the form of '[Study_prefix Curator_ID: transcript_lines]', where 'Study_prefix' is an 'S' for SMS, and an 'M' for

MMS. Quotes are almost always verbatim, but in some cases, words in square brackets are used to remove identifying information, such as the MOD for which a curator works.

Differences in common themes were observed in curator work practices.

Several common themes were present in the interviews, such as the relative utility of abstracts and other information resources in the annotation process; whether organismal or specialty differences are more difficult to overcome than differences in methods used across papers; curation as another layer of peer review; perceptions of the meaning of annotation- and curation quality; and observer effects as a potential source of variation.

1. *If it is important, it will be in the abstract.*

> I always try to read the abstract first, just to try to get a basic idea of what they feel is significant about what they're reporting. [S09: 161-162]
>
> I guess my mindset is, when I first look at a paper, the first thing I really focus on is the abstract, because if they have some groundbreaking new process or function that some protein is involved in, they usually talk about it in the abstract, versus something more peripheral, that might be mentioned in the paper, but not that important in terms of the biology of that gene product. So if I see something in the abstract, that is usually the first key that if they've identified a new function or process for that, that I want to capture that. [S04: 138-143]
>
> I get most of my cues, again, off of the abstract, because I go back to that main question of "What is this paper actually about?" That all should be in the abstract. Some big, important point in the paper, some new function or some new thing that somebody proved, and they didn't even put it in the abstract? It can't be that important. [S06: 479-487]

Some curators said curation experience and organism familiarity are helpful in understanding an abstract and placing the paper in context:

> [I] definitely just read the abstract first. And I'm familiar enough with the genome now that I can often tell if this is something really novel and, you know,

> high priority, or if it's just one more fact to an existing mountain of facts. [S03: 94-98]

One curator takes a different approach, particularly with papers that are on previously uncharacterized proteins:

> Normally I start with the Introduction. I probably would read the abstract last, actually, because the Introduction- especially if it's the first paper on a protein, I like to be led in a little bit slowly, more gently into the subject area - so I'd read the Introduction and then go all the way through Results, Materials and Methods for the species quite often is needed, all the way through the Discussion, mark up the paper, and then I probably would turn to the abstract once I've got fair idea of the GO terms just to double check, actually, that there isn't anything in the abstract [that I have overlooked]. [M07: 130-141]

Use of the abstract as a way to confirm that nothing important was overlooked was a common theme among curators:

> I sometimes look at the abstract first, but I find sometimes I get bogged down in the abstract – it's just too much information too fast. And I sometimes don't want to know the punchline, you know? But every time at the end, I go back to the abstract, because I just want to make sure that I didn't miss anything that the authors [were claiming]. [S01: 180-185]

2. Context influences GO annotation

It seems that the broader work context in which GO annotation is performed heavily influences the curators' specific approaches to selecting, reading, and curating individual papers, beyond the content of the papers themselves. Decisions about which portions of papers to curate, or even to curate the papers at all, are dependent upon factors such as prior annotations to a gene that are extant not only in the database in question, but other resources (e.g., NCBI databases, UniProt) that also cover information about that gene. Database policy questions, such as the strength of one evidence code over another; whether the MOD accumulates evidence versus selectively acquiring and replacing annotations; or the presence and absence of annotations to orthologous genes in other resources, all impact how prospective annotations are made.

Multiple curators said that they use GO annotations in other MODs to provide guidance on creating annotations in their own resources. For example:

> If I find out that some gene is homologous to another gene in mouse or fly, I definitely go to their database and see how they have annotated it. [S02: 213-215]

Another curator said that extant annotations in other resources help drive whether to make them in the curator's own database:

> So one of the things I do is that I go out there and I look first at Entrez Gene, for what GO terms they've got, and also at Uniprot for the human, just so that I can score whether it's got manual annotations, electronic annotations, no annotations at all, or protein incorporated annotations. So I will make a record of that, just so that I know, and that might influence whether I then continue. You know, if it had a lot of manual annotations, I might just go: "that's been done, OK, might not be done brilliantly, but it's got something, and I'm going to put my effort into other things." But sometimes I get very frustrated – you know I might read a paper, or I might see a gene that I know from my previous work, and I know that it's completely underrepresented, and I might just go, "you know, that really isn't helping." So I'll add some terms to it. [M18: 156-172]

One curator who had curated several different organisms was asked whether differences in organisms were more significant that other aspects in affecting annotation performance:

> I actually don't think it was that difficult to go from [my organism] to curating human, mouse and rat. So, you know, this kind of speaks to your question about whether or not you could have a general biology knowledge and curate for any given organism. I guess I tend to think that yes, you probably could. I think the big differences are organisms that use genetic techniques versus those that more typically rely on biochemistry. [M08: 395-399] I think that it's a lot easier to curate different organisms that use similar experimental approaches than it is to even curate the same organism with very different experimental methods or techniques. [M08: 413-415]

Another curator from a different MOD agrees:

> Molecular biology and biochemistry methods are very general. Things like where you're assessing morphology and you're making calls and well, this is mutated and not... a lot of other people are telling me, "man, your [organism is] just, you know, these pictures, they all look normal to me." You know, and they're not – there are subtle differences that you look for. And I think that's the same probably if I were to look at a series of *C. elegans* mutants. You know, they're all the same to me. Yet, the person in that field would be able to point out the subtle differences that make them different. [M09: 248-256]

Lack of knowledge of specialized techniques seems to cause some curators to rely more on an author's statements than on the curator's own expertise:

> I think that when you're not that familiar with the technique, you're much more inclined – or at least I was – to just take the author's word at whatever they say. And you don't really bring that much prejudice to making those kinds of annotations. [M08: 345-348]

One curator in the MMS said that rather than differences in methods, it was specialized terminology within a sub-domain that made the task of curating papers from different organisms difficult:

> Not so much the types of experiments, because – so, the only plant paper was *Arabidopsis*, and that's a genetic system just like [my organism] is a genetic system, so I don't think the types of experiments were different, but the terminology was different, so when they talk about 'pedial' and 'gynoecium' and 'stamen' and 'anther'.... I have done plant science. I have that in my background, but probably not to that degree that they're talking about in the paper, so the terminology was kind of like, "Oh, I don't know what this means," you know; "I have to look it up, or I'm just going to annotate to a very high level term and play it safe." [M06: 253-263]

3. *Familiarity with certain organisms or areas may constrain a curator's personal GO vocabulary.*

Multiple curators in the MMS noted that they realized in curating papers from multiple organisms that their personal GO vocabularies seemed very small, since they tend to focus on certain parts of the ontologies. For example:

> [N]ow that I'm using GO I tend to have a preference for certain nodes of the ontology that I go down more frequently than others. [...] I visit certain GO terms more than others, just simply because I know of their existence, and when I'm reading the paper, if I've got a fair idea if the GO term's right, and I look at the term and it is, I then [would] probably be happy with that one rather than searching all over again, so I think I have a – kind of like a sort of standard vocabulary of GO terms that I visit more often, and those which I have to be a little bit more imaginative where I don't already have an idea of the terms to use. This is where talking [with the other database representative, during the consensus annotation process] was very interesting, because I think just from the type – the randomness of the types of papers that she's been going to in her curation process, I think she's got a different kind of standard vocabulary of her favorite GO terms. And so the clash of those, actually, was really nice, because it just ended up making me appreciate "Oh, I've got to remember to go down that node more often." And I think this is where it was really nice to actually remember that there are 20,000 terms, and that I've really got to remember to go and see those. [M07: 269-288]

4. *Curators are mixed about whether they do (or should) function as an additional layer of peer review.*

One gray area of GO annotation is whether curators should essentially be gatekeepers, functioning as an additional layer of peer review. Some curators may say "all of these papers are already peer reviewed – I'm not going to second-guess an article that was published in *Science*." Others, particularly those with significant lab experience, may say, "you know, I've done this same experiment dozens of times, and never saw data like this. I don't buy it." The following two quotes are representative of the ideas heard in the interviews, and focus on examining and interpreting figures in articles:

> I would say fifty percent of the time I would look at figures. Usually in two cases. Either I can't understand what the experiment was and I try to understand better by looking at figures. You know, where really, they used this mutant or this method, and so on and so on. And another reason is to verify. So, like it's a final check. Ok, I decided, based on the abstract, I decided that this GO term is needed with this evidence code. And as a final kind of double check, I look at the figure to make sure that they, for example, really studied the enzymatic activities. [S08: 198-205]

> I don't normally look at figures. Hardly at all, unless something is confusing or whatever, but – I figure, in the Results, they're going to describe what the

> figures say and, you know, I'm not going to reinterpret the figures, and we're not supposed so be judging the quality of the data. I'm not going to look at the figure and say, you know, they- they're making the wrong conclusion from it. I'm going to just record their conclusions. [S03: 101-105]

5. *Perceptions of annotation and curation quality.*

Curators expressed similar themes about what terms such as 'consistency', 'reliability', 'accuracy', and 'quality' mean to them in relation to curation and GO annotation:

> I'm just very conscious of making sure that anything I've put up there is defensible. You can say where you got it or why. We want people to be able to look at what they have – what we have up there – and trust it. [S06: 527-530]
>
> It is my belief that a user who comes into [our database] should find the current information, and I should strive and make sure that that information, what I put up there is accurate. That's my sense of trying to have good quality. [S02: 347-349]
>
> I think to me, the most important thing is, am I bringing the information from or taking information from that paper that I would expect a user to want to find? And you want to try to represent everybody properly that spent a lot of time working out there. So I try to make sure that I'm not biased based on my background, wanting to curate more from one lab than another, but just properly, trying to represent the information as accurately as I can, so that either if somebody from a small group or somebody that's doing large scale stuff is going to go in and get that information and know that it's correct. [S04: 303-310]
>
> I guess I'm not too paranoid about it, though, because, you know, even if you try to do the best job you can, you can't get it right every time. Not even most of the time, probably, and the whole thing is cumulative. You know, people are going to be coming back to this gene many times. There are going to be new papers, and it's just – it all builds on this. I try, I guess, to compromise speed with accuracy. You know, I'm not going to spend all day on a paper, because that's just not – you just can't do that. [S03: 210-216]

Other curators discussed the idea that it can be frustrating knowing that the volume of new papers that needs to be curated has an impact on factors such as accuracy or quality, and those suffer when enough time is not taken to be certain about annotations, or when papers are complex and challenging and require additional work to understand them. As one curator put it concisely:

> I think that 'quality' in this job is the patience to be complete. [S01: 367-368]

6. *Some variation may be due to observer effects.*

The interviews indicate that some of the variation in instance counts may be due to the fact that the curators knew that their work was being observed and analyzed, and, especially in the MMS, that they would be using their annotations to train novice curators. This theme was noted when curators were asked in interviews what attributes of the study were different from their normal curation practices:

> I think some of us probably over-curated papers, just knowing that we want it to be complete and not miss things. So we maybe curated extra topics that we wouldn't have otherwise. [S04: 246-248]
>
> I think I read the papers much more thoroughly than I would have otherwise. Even the [papers from my organism], because in day-to-day work, I'm trying to get through a lot of papers in a little amount of time. And for this, I was like, "well, you know you're trying to teach people so you really want to extract every little last drop of information that you can out of the paper." And so in some ways I think, you know, we're like overanalyzing our work in this whole workshop. [M09: 162-172]

3.6. Curator background and experience (MMS)

The self-administered questionnaires in each curator's annotation worksheet package in the multi-MOD study were designed to enable comparisons to be made between curators' backgrounds and experiences and their formal GO annotations, as described in section 2.5.2. The original survey instrument is attached as Appendix A2. The responses from each of the four questions are characterized below.

Question 1. Degrees

Question 1 asked participants to list the disciplines of each of their earned degrees. Twenty of the 23 curators held Ph.D degrees at the time of the study. Of the three other participants, two held master's degrees, and one held a recently-earned bachelor's degree. Of the Ph.D holders, only 4 held one or more master's degrees, and one of those was a subsequent degree in bioinformatics. One respondent listed only a Ph.D, and no undergraduate-level degree.

Table 18 shows the distribution of the Ph.D degrees by degree name and area. No attempt was made to combine like degrees, and combinations of subject areas were exact match only. While four curators held genetics degrees, six curators' degrees had 'molecular biology' in their titles.

Table 18. Distribution of curators' Ph.D degrees by degree name and subject area

Degrees	#	%
Genetics	4	20.0
Biochemistry	3	15.0
Biology	2	10.0
Molecular biology	2	10.0
Molecular, Cellular, and Development Biology	2	10.0
Veterinary Sciences	2	10.0
Botany and molecular biology	1	5.0
Plant biochemistry	1	5.0
Plant biochemistry and molecular biology	1	5.0
Plant genetics	1	5.0
Virology	1	5.0
Total	**20**	**100.0**

Subject area	#	%
Molecular biology	6	23.1
Genetics	4	15.4
Biochemistry	3	11.5
Biology	2	7.7
Cellular Biology	2	7.7
Developmental Biology	2	7.7
Plant Biochemistry	2	7.7
Veterinary Sciences	2	7.7
Botany	1	3.8
Plant genetics	1	3.8
Virology	1	3.8
Total	**26**	**100.0**

Question 2. Expertise

Question 2 asked respondents to list their main area or areas of scientific expertise. Multiple responses were permitted. Of participants' first responses, there were 17 distinct areas, as shown in Table 19, with one response of “none”, and one left blank. Only ‘development’ and ‘plant developmental biology’ received more than one response. There were 48 total responses, of which 38 were distinct. Only five were given by more than one curator: ‘development’ (5), ‘cell biology” (2), ‘cell cycle’ (2), ‘immunology’ (2), and ‘plant developmental biology’ (2).

Table 19. Distribution of curators’ areas of expertise (first responses)

Areas of expertise	**#**	**%**
development	4	17.4
plant developmental biology	2	8.7
cell biology	1	4.3
cell cycle	1	4.3
cell wall construction	1	4.3
DNA-protein interactions	1	4.3
DNA methylation	1	4.3
Embryonic development	1	4.3
enzymology	1	4.3
genome evolution	1	4.3
immunology	1	4.3
inflammation	1	4.3
Protein regulation and degradation	1	4.3
proteomics	1	4.3
transposition	1	4.3
regulation of gene expression	1	4.3
virology	1	4.3
"none"	1	4.3
blank	1	4.3
Total	**23**	**100.0**

Question 3. Laboratory Experience

Question 3 asked curators to list those biological organisms with which they have had some amount of laboratory experience. Of participants’ first responses, there were 17 distinct organisms, as shown in Table 20, plus two curators who did not respond. There were 59 total responses reflecting 38 distinct organisms. The number of organisms given by the curators ranged from 0 (2 cases) to 7 (1 case), with a mean of 2.6 and a mode of 2 (7 cases).

Table 20. Curators' laboratory experience by organism (first responses)

Organism	#	%
Arabidopsis	3	13.0
mouse	3	13.0
[blank]	2	8.7
bovine	1	4.3
C. elegans	1	4.3
chicken	1	4.3
Dictyostelium discoideum	1	4.3
Drosophila	1	4.3
E. coli	1	4.3
Erlich ascites tumor cell lines in mice	1	4.3
human tissue	1	4.3
Mammalian cells	1	4.3
Oil palm	1	4.3
sorghum	1	4.3
varied in vitro proteins	1	4.3
wild rodents	1	4.3
Xenopus laevis	1	4.3
yeast	1	4.3
Total	**23**	**100.0**

Question 4. GO Annotation Experience

Question 4 asked respondents to list those biological organisms with which they have had some amount of Gene Ontology annotation experience. Of participants' first responses, there were 11 distinct organisms, as shown in Table 21, plus one curator who responded "none". There were 40 total responses reflecting 17 distinct organisms. The number of organisms given by each curator ranged from 0 (1 case) to 5 (1 case), with a mean of 1.7 and a mode of 1 (14 cases).

Table 21. Curators' GO annotation experience by organism (first responses)

Organism	#	%
Arabidopsis thaliana	3	13.0
chicken (Gallus gallus)	3	13.0
human	3	13.0
mouse	3	13.0
Caenorhabditis elegans	2	8.7
Dictyostelium discoideum	2	8.7
yeast (Saccharomyces cerevisiae)	2	8.7
E. coli	1	4.3
mammals	1	4.3
rat	1	4.3
rice	1	4.3
"none"	1	4.3
Total	**23**	**100.0**

Years of GO Curation Experience

A follow-up question asked of the participating curators via email was how many months or years experience they had in performing GO annotation at the time of the study. Responses were received from 21 of the 23 curators, and ranged from a low of 6 months to a high of 8 years, with a mean of 2.9 years and a standard deviation of 1.7 years. These data were used above in section 3.3 to test whether curators' variation in annotation was influenced by years of GO experience.

3.7. Curator background and experience (SMS)

The single-MOD study did not use a structured questionnaire for collection of information on curators' backgrounds and experience. These data were derived from analysis of the individual semi-structured interviews conducted with the SMS curators, following the five categories of questions in the MMS survey instrument described in section 3.6 above: degrees, expertise, organism experience in the laboratory, organism experience with GO annotation, and duration of GO annotation experience. The protocol used for the interviews is attached as Appendix A1.

Degrees

All 10 of the curators held Ph.D degrees at the time of the study; 6 had also completed one or more post-doctoral research appointments. Three curators had genetics-focused Ph.D degrees, and the other 7 were distributed over 6 other subject specialties, as shown in Table 22.

Table 22. Distribution of curators' Ph.D degrees

Degrees	**#**	**%**
Genetics	3	30
Genetics and molecular biology	2	20
Biophysics	1	10
Bacteriology	1	10
Genetics and biochemistry	1	10
Microbiology, molecular biology and genetics	1	10
Molecular biology	1	10
Total	**10**	**100**

Expertise

Given the small population, the name of the organism with which each of the specialties below are associated was removed to afford the participants more privacy. As Table 23 shows, there were no overlaps in expertise among the 10 curators' verbatim statements of expertise. No normalization of specialties was attempted.

Table 23. Distribution of curators' areas of expertise

Areas of expertise	**#**	**%**
Chloroplast genes	1	10
Cell cycle	1	10
Evolutionary genetics	1	10
Gene therapy	1	10

Areas of expertise	**#**	**%**
Metabolism	1	10
Miotic recombination	1	10
Protein folding	1	10
Splicing	1	10
Transcription	1	10
Transcriptional regulation	1	10
Total	**10**	**100**

Laboratory Experience

The curators predominantly had laboratory experience with the organism that is the subject of their MOD, so those data are omitted here, again to protect their privacy. Given that it is the same across almost all curators, it is unlikely to be a useful discriminant.

GO Annotation Experience

As with Laboratory Experience, curators had most of their GO annotation experience with their MOD’s organism, so this variable was not used for analysis. Several curators had many years of literature curation other than GO, however (exact counts not available).

Years of GO Annotation Experience

At the time of the study, the 10 curators’ experience with GO annotation ranged from a low of 6 months to a high of 8 years, with a mean of 4.1 years and standard deviation of two years.

4. Discussion

4.1. Contributions

This section summarizes the findings and contributions of the two studies.

GO annotation is qualitatively different from indexing in process and intent.

While the GO curation process resembles topical document indexing in some ways (and, in fact, one role curators often have is to index articles to provide access points to specialized topics, such as research methods, that are not captured by standard indexing), the specific process of GO annotation creation, and the subsequent use of annotations by scientists, is quite different. Rather than summarizing what a document is generally 'about', GO annotation focuses on the specific claims authors make, and associates them with specific gene products through the formal GO annotation artifact. End users use GO annotations not to find literature, as users of bibliographic databases do, but to understand what is known about a particular gene product. These differences have implications for not only the study of annotation variation, but also what roles information and library science can reasonably hold in the areas of biomedical curation and database administration. Curators involved in hiring and training novice curators reported in interviews that they believe it to be far easier for subject matter experts (e.g., Ph.D-trained biological scientists) to learn curation and its associated tools (such as ontologies and web-based databases) than it would be for information experts to learn the subject matter necessary to become efficient and effective curators.

GO annotation can be taught but not replicated.

The apprentice model of curator training seems to be the most effective based on curator interviews. By observing experienced curators, curating the same articles, discussing their differing annotations, and then rationalizing the individual annotations into a consensus annotation, novice curators develop the skills necessary to practice GO annotation, including how to read articles differently for curation from the way they are read by students or investigators: which sections are most useful for annotation, which language to focus on, and which to avoid. This approach may not result in perfectly consistent annotation, but it provides a semi-formalized method for teaching the rudiments of curation and transferring tacit knowledge acquired by experienced curators. The annual GO Annotation Camp venue has been extremely important in developing this model and in building both a community of curators and the concept of curation as a career path in science.

Perfect consistency is unlikely given annotation complexity.

If we consider the spectrum of possible outcomes – the potential that there are multiple genes characterized in a paper; that multiple experiment types may be present, yielding different evidence codes and relating to different GO aspects; that differences in workflows, reading behaviors, and experience can affect interpretation; the sheer number of GO terms available from which curators can choose – it may begin to seem unlikely that two or more curators could ever independently create annotations that were exactly the same. We should keep this complexity in mind when reviewing the results above.

A focus on perfect consistency obscures diverse sources of variation.

As shown in 3.2.2, exact match consistency at the paper level was rare: less that 20% in the SMS 4x paper set, 34% in the 2x set (due to zero annotations), and only 50% of MMS papers had one or more pairs of exact matches. Gross numbers on consistency, however, are not useful for summarizing the results, because they aggregate across papers and obscure facet-level variation. Consistency statistics do not provide information about whether the differences between compared annotations are either meaningful or negative. One example is the facet analysis of the annotation instances for MMS paper MP2, where the ten consensus annotations were inconsistent, but the Completeness measure showed that none of the compared annotations had false negative instances, only false positives. One of the goals of developing granular facets of annotation quality was to transcend the limitations of simple consistency measurements.

Several types of variation in GO annotations were observed in the studies. Individual evidence sources had wide ranges of annotation instances. Individual curators made differing amounts of instances across papers, and chose widely different terms, genes, and evidence codes within them. Variation was observed by organism expertise, by group assignment, and between individual and consensus annotations. Years of GO curation experience was found to not be a predictor of annotation instance quantities. The curators themselves had widely different educational, training, and curatorial backgrounds. Interestingly, large numbers of annotation instances seems to counter the idea that curators are conservative about annotating only those claims about which they are very confident. Table 6 showed that while the SMS organism experts' instances for the two papers included in both studies ranged from 1-8, the MMS curators' instances ranged from 0-35.

Perfect consistency is not necessarily optimal.

Perfect consistency among annotators may not be desirable even if it were possible. Slight variation in term selection, for example, may provide broader or deeper coverage of an evidence source, in the same way that multiple indexers of a document may be technically inconsistent in index term selection, but in practice provide different levels of conceptual specificity that lead to multiple useful access points.

Slight differences may also be acceptable to end users and administrators if the annotations are not inaccurate or misleading. Two cases from the Specificity facet calculation in section 3.3.4 are instructive: The three terms in the second instance of MP2 were adjacent, while in SP3, some terms were 11 nodes apart. The former may be an acceptable level of variation, while the latter may not. The setting of thresholds of acceptable variation should, of course, be left to those who create and use the annotations as they deem them suitable. One approach could be the use of GO 'Slims' to smooth variation and provide rubrics for assessing consistency. GO Slims are versions of the GO vocabularies that contain less granular subsets of the entire trees. If a tree is collapsed into smaller numbers of more generalized categories, instances that have term mismatches at the full GO level may be consistent if they fall within the same larger bins of GO Slims. Dolan et al.'s (2005) work employed this method, but, as they note, the variation in how GO Slim bins are constructed can have significant impacts on the calculations of annotation consistency, and the category definitions or variation thresholds for comparison need to be applied uniformly by different evaluators to enable consistent assessments of quality.

There are other reasons why perfect consistency is elusive. While a finite number of GO-annotatable claims exist within a particular evidence source, the concept of a 'consensus' annotation is still only a surrogate for truth, in that even if all annotators of an article agree, they may still be incorrect in terms of one or more of the quality facets. Agreement is no guarantee of accuracy.

The common evaluation practice of summing and averaging judges' scores assumes that they are experts and are qualified to make judgments; even qualified judges vary in the stringency/leniency of their ratings (see Shrout & Fleiss, 1979). Pooling may be useful in certain contexts, but here, since we are more interested in the underlying process of annotation and the individual differences of curators' annotations, pooling or averaging might obscure the information in which we are interested.

One related issue is the relatively large proportion of cases in the SMS that lacked any annotations by one or more curators. The absence of annotation is not necessarily a negative situation. Curators described three possible reasons why they might not make an annotation where it had been expected:

1. There is no annotation present in the paper, in the opinion of that curator.
2. An annotation exists in the paper, but was overlooked by the curator.
3. An annotation exists in the paper, and is recognized by the curator, but is not made because of MOD policy or curator judgment (e.g., it doesn't add anything new that isn't already present in the database (and the MOD's policy is not to

curate every paper); the annotation should be made from a different or earlier paper; etc.).

These are very different reasons, but in the absence of contextual information obtained from observations or other methods, we will be unable to make any claims about why a particular annotation was not made by one curator, but was made by another. Curators seem conservative about making annotations – during interviews several said that they would prefer to not make an annotation at all than to risk making an inaccurate or incorrect one due to uncertainty, given the potential negative impacts on research. The first case seems to have occurred with the SMS 2x coverage article set, where 11 of the 32 papers had zero annotations from both curators, and in the 4x set with papers SP8 and SP9, where the individuals' zero instances were carried over to the consensus annotation. Several curators implicitly addressed the second case in interviews; when asked about differences discovered when discussing their individual annotations in the consensus pairings, they mentioned that the other curator would frequently have identified and annotation that the first curator had missed. The third case is expected to be a small portion of the missing annotations, given the study instructions and the presence of notes from curators in the spreadsheets that explain why they would not make the annotation in practice.

Pairwise matching and comparison of instances is difficult and atypical.

The difficulties associated with matching annotation instances across curators for strict pairwise comparison limits our ability to use formal measures to cases where a large proportion of the curators' annotation instances intersect. Other approaches, such as fuzzy matching or heuristic algorithms may be more suitable. One example could be the use of database approaches to compute the intersection of instance sets and to calculate the Completeness and Specificity metrics to derive relative scores based upon a combination of those three facets.

We can see examples of the drawbacks of the pairwise matching approach in the facet analysis (section 3.3.4). In the MP2 Consistency calculation, only sixteen of the original 54 instances were useful for pairwise comparisons. In the Reliability facet computation for RP1, on the second reading the curator decided that the evidence was not strong enough to warrant IMP annotations in addition to the IDA annotations, which resulted in inconsistent original and subsequent annotations. In this case, if we rely upon the binary measurement of 'consistent/inconsistent' as the method of evaluating performance, the curator is penalized for improving over time.

The annotation quality facets require revision to be more useful in practice.

Section 3.3.4 evaluated the study data against the proposed GO quality facets; here we address the performance of the facets themselves. How useful do they appear to be? What modifications need to be made?

The concept of decomposing 'consistency' into facets such as Specificity and Completeness that are more granular in order to understand different types of variation has face validity. The use of dichotomous attribute values and cumulative binary facet scores of 1 or 0 are not as informative as they could be. A different approach to capturing the granularity of the attributes in a summary measure is needed. The Specificity measure is useful only for determining relative distance. It cannot be used for normative evaluation, because, while the stated goal of GO is to annotate to the most specific appropriate term, we cannot assume that 'more specific' equates to 'better'. In the MMS

MP2 example calculated above, for example, the MOD representatives may have expert knowledge (or other reasons) that caused them to select the broader term ‘response to stress’ rather than the more granular terms chosen by the other participants. Also, since both the depth and the number of nodes per branch can vary by annotation pairs, a normalization method is needed in order to be able to state which of two instances is conceptually closer or more distant when branch depth varies. Different weightings of some facets and attributes may be important for instances where a paper has large or small numbers of annotations; where there are significant discrepancies in curators’ GO term assignments; and in general to capture relative importance of certain attributes of the GO annotation compared to others, which may be context-dependent.

However, more than the shortcomings of the facets, the inability to more easily perform pairwise matching of instances greatly inhibited our ability to evaluate the data using the faceted measures. Effective methods for deriving the intersections of two or more sets of annotation instances would make the use of the measures practical.

Finally, while these measures are designed to evaluate human-created GO annotations, they could also potentially be useful as performance standards for the evaluation of automated systems (such as in a BioCreAtIvE (Hirschman et al., 2005) challenge setting), or for assessing variation between human-curated and automated annotations.

Electronic workflows may inhibit explanations of variation.

The electronic-only workflow model observed in the SMS seems to optimize curator time by eliminating intermediate tasks and artifacts, but further obscures the intellectual linkage between the evidence source and the final GO annotation since it resides only in the curator's head. Significant investments of time were required in the Reliability pilot study to observe curators using the concurrent verbal report method in order to try to apprehend some of the internalized reading- and annotation behaviors and associated decision-making. The Future Work section below describes methods for end-to-end evidence linkage that are being developed that could address this issue.

A useful corpus of high-coverage GO annotations was created.

The prospective nature of the two studies produced a valuable set of nearly 4,000 new GO annotation instances covering five organisms. This data set may be useful for further analysis of variation (as described below in Future Work), for use in training novice curators, as well as for use as training data or reference ('gold standard') data for automated systems for GO annotation and claim extraction. It has also provided the additional benefit of adding to the existing pool of GO annotations. This set of annotations will be made available at a future date for use by other investigators.

4.2. Limitations

These studies have certain limitations, described below, which inhibit generalization of their results.

Participant sample size and composition.

While thirty-one curators participated in the studies (and that number represents a large percentage of the population of GO curators that exist), that is only a sample from a small number of model organism databases, and results might have varied significantly if a different group of curators from different organisms had participated.

Limited facet-level analysis was performed.

Due to the complexity of instance matching described above, only a very small sample of annotations were analyzed using the quality facet metrics in section 3.3.4. We attempted to choose representative annotations, but the small sample size means generalizations about the remainder of the annotations cannot be made, and not enough calculations were made to provide meaningful aggregate statistics at the facet level.

Worktask considerations.

While curators in both studies were instructed in general to perform their annotation tasks as they normally would, differences in quantities of personal annotations (in the paper documents) and GO annotations by organism novices and experts in the MMS might in part be the result of experts annotating differently from their normal approach in order to 'set a good example', or be more thorough to help novices during the discussion portion of the GO meeting understand from where in the documents GO annotations were drawn. Additionally, while the studies were designed to be as naturalistic as possible, certain aspects were different enough from the curators' normal workflow and practices to have affected outcomes. Perhaps the most significant of these occurred in the SMS, where curators were instructed to work alone, and not seek input or advice from other curators when making their annotations. In the individual interviews and focus group that were subsequently conducted, this restriction was the most often repeated negative attribute of the study: most local curators said that their normal work behavior was to discuss annotations and the underlying papers quite frequently with their colleagues, and the inability to do so in this case made the work more difficult and more time-consuming, when they had to search for information online or in reference materials that they may otherwise have been able to turn around and ask a colleague. While the intent of this restriction was to avoid curators converging on the same outcome and obscuring variation, it also forced a situation where the task was somewhat unnatural. However, half of the participating SMS curators work off-site full time, and thus do not have the ability to obtain assistance as easily as those curators who work together. In the individual interviews, the off-site curators reported that they were more likely to return to the pool of uncurated papers an article that contained unfamiliar information than they would be to contact the other curators to seek assistance. Remote curators also tended to have more curation- and work experience as well, though, and may typically have fewer questions about the papers. A related factor is that different MODs have different annotation workflows. In some cases, curators are expected to do several kinds of curation, including GO annotation, while other MODs are more specialized, with some curators focused on GO annotation, and others performing other curation tasks.

5. Conclusions and Future Work

5.1. Conclusions

This project provided the first in-depth investigation of biocurators' Gene Ontology annotation behaviors, both in terms of fine-grained multilevel analysis of annotation instances, as well as contextual factors that may influence curator annotation behavior. Variation in curator annotations was expected prior to conducting this study, but the origins, magnitude, and types of variation were unknown. From a basic science perspective, this project was interested in beginning to understand the origins of differences in human information interaction behavior in the context of Gene Ontology annotation. Given a standard task, with the same evidence sources and controlled vocabularies, why and how do people differ in such aspects as their term selection, identification of genes, reading behavior, and interpretation of experiments? What is the relative difficulty of locating an annotatable gene product, identifying the experiment (for the Evidence Code), and selecting an appropriate GO term? While it may appear that gene products may be easily identified within a paper, to which of them it may be appropriate to make a GO annotation is less clear. While the experiment may be described in detail in the Methods section of a paper, it may not be obvious which Evidence Code should be applied. More work on the cognitive processes underlying these activities of identification and relation is needed, as described below.

In addition to understanding basic human information interaction behavior, the project was interested in: a) outcomes that might assist in the training of novice scientific curators by studying the practices of expert curators; b) testing measures of annotation quality that might be useful to MOD administrators and funding agencies; and c) using the knowledge obtained from the studies to augment the performance of human annotators as well as improving automated approaches to GO annotation. The interviews with experienced curators reinforced the apprentice model of curatorial training, and also provided workflow information that could assist in the development of better training models. The annotation quality facet analysis allowed for a more granular assessment of annotation variation, and could be more useful when refined if methods to address mismatched instances are also developed. Finally, the set of novel annotation instances should be useful for conducting both human- and automated performance improvement activities. The Future Work section below discusses additional projects to help advance these three research objectives.

An example of how endemic variation is in the annotation process is illustrated by curators' differing values in perhaps the most straightforward element of the GO annotation: the aspect, which has only three possible values. Table 24 illustrates eighteen observed variants for this element.

Table 24. Variation in curators' GO aspect values

molecular_function	MF
biological_process	BP
cellular_component	CC
Function	FUNCTION
Process	PROCESS
Component	COMPONENT
function	F
process	P
component	C

And, as the responses to the question about curators' areas of expertise show (Table 15), variation in what terms are used to describe biological concepts extends even to the names of specialties within the domain of biomedical research. While at one level this may indicate the need for ontologies and controlled vocabularies, this work has shown that even when those tools are present, variation in use will persist. Some may eventually be compensated for by automated means, but some will likely remain extremely difficult to evaluate.

5.2. Future work

This section addresses future work to answer additional research questions raised by the results of this project.

Do group dynamics help explain the variation between individual and consensus annotations?

As noted in section 2.7, curators in the SMS were observed while they participated in paired discussions to rationalize differences in their individual annotations to create consensus annotations. While these data were not used in this study, they may help us further understand the origins of individual differences, as well as insights to group dynamics, such as when curators of the same and different levels of expertise rationalize variations in individual annotations.

Do systematic features of the literature lead to variation in annotations?

One unexplored source of variation is the document corpus itself: Do differences in the underlying papers explain part of all of the variation in annotations? Attributes such as document length and complexity, or large or small scale experiments, could be evaluated in relation to instance counts, term specificity, or curator attributes such as organism expertise or years of GO annotation experience. Some data was collected during the individual interviews in both studies on curators' perceived differences in the articles used for annotation, and their behavioral responses.

Does vocabulary size explain variation?

This project did not investigate specific term differences on a large scale outside of the example Specificity calculations. A more comprehensive comparison of the intersection of term sets, particularly from the MMS, could be informative, particularly in conjunction with the experience and expertise contextual data. Also of interest is where divergent terms come from that do not overlap. As with other types of manual classification tasks, we might expect that variation in term assignment would increase as the number of available terms increases, in both breadth and depth. One test of this hypothesis could be the use of the minimal GO ontologies called 'GO Slims' (described in section 4.1) compared with the full GO ontologies, applied to the same evidence sources.

Do curators' indexing outcomes mirror their annotation variation?

During the SMS study, as noted in section 2.1.1, the curators performed a 'full curation' of the papers, so additional data exists that can be used for further analysis, such as evaluation of the consistency of the high-level indexing performed on papers. Since these terms are drawn from a very small controlled vocabulary, it would be interesting to compare consistency of that task with that of the GO annotation component, where curators must choose from a much larger set of terms. Also, rather than capturing the claims of the paper in a formal structure, this type of curation is more similar to a typical article indexing task, where the goal is to identify the ‘main topics’ of a paper, or what the paper is ‘about’. It may be possible to compare the SMS curators’ term selections with the MeSH terms assigned by MEDLINE indexers, as the pilot study in MacMullen (2006a) attempted to do.

Can the information in these studies assist in training novice curators?

Bertrand and Cellier (1995) identified three general indexing strategies based upon indexers’ expertise. Can effective strategies for the creation of GO annotations be identified from the contextual data collected in this study from expert curators? These could be useful for training novice curators in best practices of curation.

How do end users use annotations and perceive GO annotation quality?

The studies reported here focused on GO annotation from the curator perspective, but, as with that area, little investigation has been done of scientists’ use of GO annotations in their everyday work tasks. Work on scientists’ information behaviors, trust, quality perceptions, and roles in creation and revision of GO annotations are needed. What is learned from studies of these kinds could inform not only the further development of measures of GO annotation quality, but may help us to understand curator variation.

At what level of granularity can GO annotations be linked to underlying evidence sources?

The cognitive gap that currently exists between GO annotations and the evidence sources from which they are drawn is another area where work is needed. This could be addressed through possible collaborations with the GOA (Gene Ontology Annotation) group at the European Bioinformatics Institute (EBI), the WormBase MOD, and others who are developing annotation and curation applications that tie the final formal GO annotations directly to the underlying components of the primary source that the curator believes establishes the claims the authors are making (e.g., Couto, et al., 2007). Applications of this kind could be useful at a practical level for training new curators, as well as helping to understand from where curators derive their annotations, and perhaps why they differ.

Research priorities.

This work has shown that a large number of contextual factors exist that may influence the amounts and types of variation in GO annotations. This section has described some approaches to addressing research questions that remain. Two non-contextual areas which should be studied next are a deeper analysis of the term variation observed in the MMS, and studies of GO end-users' information needs and quality perceptions. These studies could be performed concurrently, with GO users assessing the relative importance of selected terms to areas of interest to end users.

APPENDICES

A1. Protocol: Individual interviews

This section provides the semi-structured interview protocols for the individual curator interviews.

A1.1. Single-MOD study

Personal education and experience; work roles

- Describe your educational background [prompts: undergrad major, graduate specialization(s), other training and certification]

- Describe your laboratory experience (if any) [prompts: duration, specialization, organism / disease focus]

- Describe how you came to be involved in biological database curation [prompts: (if applicable) what made you decide to stop working in a lab environment?]

- What did you have to learn early on as a curator to help you better do your work? [prompts: biology/organism knowledge, computer skills, ontology knowledge]

- How much publication experience do you have as a lab scientist? [probe: (if little/significant:) Do you think this makes it more/less difficult for you to make annotations?]

- Do you have lab experience with more than one organism? [probe: How do you think this affects the way you make GO annotations?]

- Do you have curation/annotation experience with more than one organism? [probe: What are some differences you've experienced curating multiple organisms? [prompt: different workflows, different types of organisms' research foci, etc.]

- Describe any specific training you may have had in curation and related topics (e.g,, training in IT, ontologies). [probe: number of GO meetings attended, for example]

- How would you describe your roles as a curator, both to the database, and within science overall? [prompt: compared with lab work, if applicable; what goals are curators trying to achieve through their work and the development of the database?]

Annotation processes and tasks

- Describe the steps you take when making a GO annotation from a scientific article. [draw diagram to confirm]

- Elaborate on the sub-tasks involved in each step; what knowledge, skills, and external tools do you use?

- How much time, on average, does it take to create GO annotations from a typical experimental paper?

- What are some things that you encounter during this process that slow it down or speed it up? [prompt: article type, experiment type, presence/absence of knowledge to understand the paper]

- When examining the source article for assertions from which to annotate, describe what sections of the article you look at, and in what order.

- What specific kinds of information are you looking for when doing this? [prompt: keywords (such as…), methods, tables, figures]

- What are some facets that indicate to you what your annotation may be? [prompt: article type (e.g., experimental vs. review), methodology; do they notice a difference between 'genetics' papers vs 'biochemistry' papers, as one curator does?]

- What do you use as reference material to support your memory and knowledge when building the components of an annotation? [prompt: e.g., when identifying the Evidence Code, is the GO definition list consulted? Laboratory manuals to understand methods?]

- Are there specific types of data you are looking for in each article when you do a full curation? What are they? [supervisor listed them (informally) as "GO, phenotypes, literature topics, new gene names and other general data for a gene."]

- Who do you normally talk with when you have about questions about annotations during curation? [probe: supervisor vs peer. Draw a network diagram of who asks and is asked based on what criteria.]

- Can you give examples of situations where that normal process didn't work, or you had to take a different approach? [prompt: what features of the documents or the assertions inhibited the normal process?]

- What do you think caused that situation? [prompt: type of article, type of experiment, protocol/method, gene, etc. in the article?]

- What was unusual about the exercise compared with your normal work process?

A1.2. Multi-MOD study

(Note: Questions equivalent to those in the single-MOD study about personal education and experience are addressed in the brief written questionnaire attached to the GO annotation template, and thus were not asked during the interviews.)

The goals of this interview are to learn about your personal approach to GO annotation in detail, and to learn how you may have approached this study differently from your normal curation and annotation activities.

Typical annotation process

- Describe your normal or typical personal GO annotation process. [prompts: You might think of a recent experience you've had, or a specific paper that you remember]
- Describe the specific steps you take when reading a paper to make GO annotations. [prompts: Which section do you read first? Next? Do you read the entire paper, or only specific sections? Are there sections you always/never read? In what order do you read the paper?]
- In addition to the paper itself, what other tools or resources do you typically have nearby when annotating? [prompts: MOD-specific curator interface; GO ontology browser; other external reference materials, such as text- or reference books]
- Do you read the article on the screen, or print it out?
- Do you make marks or notes on the article? [prompts: what / how / why]
- Do you take intermediate notes on a separate paper or in a file?

MMS study specifics

- What was different about this study from the way you normally do GO annotation? [prompts: organism diversity; consensus step]
- Have you performed a consensus annotation process at your own database? What was that experience like for you during this study?
- Did you read the papers from organisms you are (less-,un-)familiar with differently than (your own, the ones you're familiar with?) What did you do differently?
- Were there differences in the structure or format of the unfamiliar papers that influenced your annotation? [prompts: paper structure; order of sections; length]
- Were there differences in the underlying experiments or organisms in the unfamiliar papers that influenced your annotation? [prompts: experiment types, terminology, fundamental biological differences (e.g., no leaf morphogenesis in mammals)]
- Do you have any other comments or questions about this study that we haven't discussed?

A2. Protocol: Self-administered curator questionnaire (MMS)

Note: These questions were also used to extract the equivalent data from the individual SMS interviews.

1. In what disciplines are your degrees? (list all) (e.g., BA in biology, BS in chemistry, Ph.D in genetics)

2. What would you say is your main area, or areas, of expertise? (e.g., transcription, DNA repair, development, meiosis)

3. What organisms do you have lab experience with? (pre- and post-doc levels)

4. What organisms do you have GO curation experience with?

Subsequent follow-up question (via email): How many months or years of GO annotation experience did you have at the time of the study?

A3. Protocol: SMS focus group

This section provides the protocol used in the SMS focus group discussions. This is a guide only; the discussion was driven primarily by the curators' responses to each others' statements about curation and annotations.

The practice of annotation

- Given a general model of an annotation process [show diagram], talk about how your individual approaches vary in actual practice. [prompt: different order, iterations, etc.]

- What skills and experience do you think are important for a curator to have to be successful? [prompt: bioscience degree, lab experience, computer experience, specific organism experience]

- What does 'quality' mean in the context of GO annotations? How would you define a 'high-quality' GO annotation? How much variation in each field is acceptable, and are some weighted more heavily than others?

The annotation exercises

- Describe the discussions you had when performing the annotation consensus exercise; what did they consist of? [prompt: arguments from experience, reference to the original article, reference to other sources]

- What were some typical differences between the annotations? [probe: Why do you think they occurred? Do you ever have these types of discussions with other curators in the normal course of your work, prior to finalizing an annotation? Under what conditions does this occur?]

The future of annotation

- What changes do you see in the next few years in biology or informatics that may change the way GO annotation is done?

- Ignoring current technology as much as you can, what would help you as a curator make better annotations? [prompt: where 'better' means 'more', 'faster', 'cheaper', etc. Probe: Include not only technological changes, but changes in the publication process; impact of open access journals, different instantiations of articles, etc.]

A4. Protocol: SMS observations

This section provides the protocols used for the observation of individual curator annotations and consensus annotations. In both cases, the audio recorder was left on even during times when no one was speaking, to enable the calculation of time spent per sub-task and total elapsed time. The researcher recorded field notes and observations about workflows and other events to enhance the recording.

Reliability observations

The curators were instructed 1) to briefly describe their actions and thoughts, but not to make detailed explanations of them; and 2) to try avoid thinking about any memories they may have of the paper, or what they did the first time. [In all four cases, the papers did not look familiar to the curators, and only in certain instances did they remember aspects of them.] Questions from the researcher were typically held until the end, except for brief clarifications or note for the recording about what was taking place.

Consensus observations

The curators were asked to arrive at a consensus by whatever means they felt comfortable. Questions from the researcher were held until the end.

A5. Protocol: Artifact collection and analysis

This section provides the protocol used to collect the individual and consensus group annotations, as well as any manually annotated articles.

SMS study

- Curators who do both individual and consensus annotations remotely should send copies of their annotation spreadsheets to the study coordinator in advance of the meeting.
- Curators who will do individual annotations remotely and consensus annotations at Stanford should bring their spreadsheets to the meeting.
- Curators who do both individual and consensus annotations on site should give their spreadsheets and documents to the local study coordinator.

MMS study

- GO annotation spreadsheets should be emailed prior to the meeting to the local study coordinator to be de-identified and aggregated before being sent to the study PI.
- Participants bringing or sending copies of their manually-annotated articles to the local study coordinator should mark each one with their curator ID, not their name.

A6. Article corpus for Single-MOD study

This appendix lists the articles annotated in the main SMS study, followed by those for the Reliability arm.

P1. Ford AS, Guan Q, Neeno-Eckwall E, Culbertson MR. Ebs1p, a negative regulator of gene expression controlled by the Upf proteins in the yeast Saccharomyces cerevisiae. Eukaryot Cell. 2006 Feb;5(2):301-12. PMID: 16467471

P2. Miao M, Ryan KJ, Wente SR. The integral membrane protein Pom34p functionally links nucleoporin subcomplexes. Genetics. 2006 Mar;172(3):1441-57. Epub 2005 Dec 15. PMID: 16361228

P3. Trautwein M, Schindler C, Gauss R, Dengjel J, Hartmann E, Spang A. Arf1p, Chs5p and the ChAPs are required for export of specialized cargo from the Golgi. EMBO J. 2006 Mar 8;25(5):943-54. Epub 2006 Feb 23. PMID: 16498409

P4. Swanson KA, Hicke L, Radhakrishnan I. Structural basis for monoubiquitin recognition by the Ede1 UBA domain. J Mol Biol. 2006 May 5;358(3):713-24. Epub 2006 Mar 9. PMID: 16563434

P5. Monroe DS Jr, Leitzel AK, Klein HL, Matson SW. Biochemical and genetic characterization of Hmi1p, a yeast DNA helicase involved in the maintenance of mitochondrial DNA. Yeast. 2005 Dec;22(16):1269-86. PMID: 16358299

P6. Mesecke N, Terziyska N, Kozany C, Baumann F, Neupert W, Hell K, Herrmann JM. A disulfide relay system in the intermembrane space of mitochondria that mediates protein import. Cell. 2005 Jul 1;121(7):1059-69. PMID: 15989955

P7. Saerens SM, Verstrepen KJ, Van Laere SD, Voet AR, Van Dijck P, Delvaux FR, Thevelein JM. The Saccharomyces cerevisiae EHT1 and EEB1 genes encode novel enzymes with medium-chain fatty acid ethyl ester synthesis and hydrolysis capacity. J Biol Chem. 2006 Feb 17;281(7):4446-56. Epub 2005 Dec 15. PMID: 16361250

P8. Hoffmann ER, Borts RH. Trans events associated with crossovers are revealed in the absence of mismatch repair genes in Saccharomyces cerevisiae. Genetics. 2005 Mar;169(3):1305-10. Epub 2005 Jan 16. PMID: 15654113

P9. Helmstaedt K, Strittmatter A, Lipscomb WN, Braus GH. Evolution of 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase-encoding genes in the yeast Saccharomyces cerevisiae. Proc Natl Acad Sci U S A. 2005 Jul 12;102(28):9784-9. Epub 2005 Jun 29. PMID: 15987779

P10. McNabb DS, Pinto I. Assembly of the Hap2p/Hap3p/Hap4p/Hap5p-DNA complex in Saccharomyces cerevisiae. Eukaryot Cell. 2005 Nov;4(11):1829-39. PMID: 16278450

P11. Noma A, Kirino Y, Ikeuchi Y, Suzuki T. Biosynthesis of wybutosine, a hyper-modified nucleoside in eukaryotic phenylalanine tRNA. EMBO J. 2006 May 17;25(10):2142-54. Epub 2006 Apr 27. PMID: 16642040

P12. Warringer J, Blomberg A. Involvement of yeast YOL151W/GRE2 in ergosterol metabolism. Yeast. 2006 Apr 15;23(5):389-98. PMID: 16598690

P13. Ishchenko AA, Yang X, Ramotar D, Saparbaev M. The 3'->5' exonuclease of Apn1 provides an alternative pathway to repair 7,8-dihydro-8-oxodeoxyguanosine in Saccharomyces cerevisiae. Mol Cell Biol. 2005 Aug;25(15):6380-90. PMID: 16024777

P14. Wang Z, Jones GM, Prelich G. Genetic analysis connects SLX5 and SLX8 to the SUMO pathway in Saccharomyces cerevisiae. Genetics. 2006 Mar;172(3):1499-509. Epub 2005 Dec 30. PMID: 16387868

P15. Huang LS, Doherty HK, Herskowitz I. The Smk1p MAP kinase negatively regulates Gsc2p, a 1,3-beta-glucan synthase, during spore wall morphogenesis in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A. 2005 Aug 30;102(35):12431-6. Epub 2005 Aug 22. PMID: 16116083

P16. Qiu J, Yoon JH, Shen B. Search for apoptotic nucleases in yeast: role of Tat-D nuclease in apoptotic DNA degradation. J Biol Chem. 2005 Apr 15;280(15):15370-9. Epub 2005 Jan 18. PMID: 15657035

P17. Murakami Y, Siripanyaphinyo U, Hong Y, Tashima Y, Maeda Y, Kinoshita T. The initial enzyme for glycosylphosphatidylinositol biosynthesis requires PIG-Y, a seventh component. Mol Biol Cell. 2005 Nov;16(11):5236-46. Epub 2005 Sep 14. PMID: 16162815

P18. Chuang SM, Madura K. Saccharomyces cerevisiae Ub-conjugating enzyme Ubc4 binds the proteasome in the presence of translationally damaged proteins. Genetics. 2005 Dec;171(4):1477-84. Epub 2005 Aug 22. PMID: 16118187

P19. Duttagupta R, Tian B, Wilusz CJ, Khounh DT, Soteropoulos P, Ouyang M, Dougherty JP, Peltz SW. Global analysis of Pub1p targets reveals a coordinate control of gene expression through modulation of binding and stability. Mol Cell Biol. 2005 Jul;25(13):5499-513. PMID: 15964806

P20. Bebenek K, Garcia-Diaz M, Patishall SR, Kunkel TA. Biochemical properties of Saccharomyces cerevisiae DNA polymerase IV. J Biol Chem. 2005 May 20;280(20):20051-8. Epub 2005 Mar 17. PMID: 15778218

P21. Hoffmann ER, Borts RH. Trans events associated with crossovers are revealed in the absence of mismatch repair genes in Saccharomyces cerevisiae. Genetics. 2005 Mar;169(3):1305-10. Epub 2005 Jan 16. PMID: 15654113

P22. von Janowsky B, Major T, Knapp K, Voos W. The disaggregation activity of the mitochondrial ClpB homolog Hsp78 maintains Hsp70 function during heat stress. J Mol Biol. 2006 Mar 31;357(3):793-807. Epub 2006 Jan 19. PMID: 16460754

P23. Ambrona J, Vinagre A, Ramirez M. Rapid asymmetrical evolution of Saccharomyces cerevisiae wine yeasts. Yeast. 2005 Dec;22(16):1299-306. PMID: 16358308

P24. Rhodin J, Astromskas E, Cohn M. Characterization of the DNA binding features of Saccharomyces castellii Cdc13p. J Mol Biol. 2006 Jan 20;355(3):335-46. Epub 2005 Nov 14. PMID: 16318854

P25. Nolden M, Ehses S, Koppen M, Bernacchia A, Rugarli EI, Langer T. The m-AAA protease defective in hereditary spastic paraplegia controls ribosome assembly in mitochondria. Cell. 2005 Oct 21;123(2):277-89. PMID: 16239145

P26. Caesar R, Warringer J, Blomberg A. Physiological importance and identification of novel targets for the N-terminal acetyltransferase NatB. Eukaryot Cell. 2006 Feb;5(2):368-78. PMID: 16467477

P27. Shaheen HH, Hopper AK. Retrograde movement of tRNAs from the cytoplasm to the nucleus in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A. 2005 Aug 9;102(32):11290-5. Epub 2005 Jul 22. PMID: 16040803

P28. Henry JM, Camahort R, Rice DA, Florens L, Swanson SK, Washburn MP, Gerton JL. Mnd1/Hop2 facilitates Dmc1-dependent interhomolog crossover formation in meiosis of budding yeast. Mol Cell Biol. 2006 Apr;26(8):2913-23. PMID: 16581767

P29. Teo H, Gill DJ, Sun J, Perisic O, Veprintsev DB, Vallis Y, Emr SD, Williams RL. ESCRT-I core and ESCRT-II GLUE domain structures reveal role for GLUE in linking to ESCRT-I and membranes. Cell. 2006 Apr 7;125(1):99-111. PMID: 16615893

P30. Rand JD, Grant CM. The thioredoxin system protects ribosomes against stress-induced aggregation. Mol Biol Cell. 2006 Jan;17(1):387-401. Epub 2005 Oct 26. PMID: 16251355

P31. VerPlank L, Li R. Cell cycle-regulated trafficking of Chs2 controls actomyosin ring stability during cytokinesis. Mol Biol Cell. 2005 May;16(5):2529-43. Epub 2005 Mar 16. PMID: 15772160

P32. Fernandez-Murray JP, McMaster CR. Glycerophosphocholine catabolism as a new route for choline formation for phosphatidylcholine synthesis by the Kennedy pathway. J Biol Chem. 2005 Nov 18;280(46):38290-6. Epub 2005 Sep 19. PMID: 16172116

P33. Mesecke N, Terziyska N, Kozany C, Baumann F, Neupert W, Hell K, Herrmann JM. A disulfide relay system in the intermembrane space of mitochondria that mediates protein import. Cell. 2005 Jul 1;121(7):1059-69. PMID: 15989955

P34. Sterling CH, Sweasy JB. DNA polymerase 4 of Saccharomyces cerevisiae is important for accurate repair of methyl-methanesulfonate-induced DNA damage. Genetics. 2006 Jan;172(1):89-98. Epub 2005 Oct 11. PMID: 16219787

P35. Poulsen P, Lo Leggio L, Kielland-Brandt MC. Mapping of an internal protease cleavage site in the Ssy5p component of the amino acid sensor of Saccharomyces cerevisiae and functional characterization of the resulting pro- and protease domains by gain-of-function genetics. Eukaryot Cell. 2006 Mar;5(3):601-8. PMID: 16524914

P36. Tsai HK, Lu HH, Li WH. Statistical methods for identifying yeast cell cycle transcription factors. Proc Natl Acad Sci U S A. 2005 Sep 20;102(38):13532-7. Epub 2005 Sep 12. PMID: 16157877

P37. Kwan JJ, Warner N, Maini J, Chan Tung KW, Zakaria H, Pawson T, Donaldson LW. Saccharomyces cerevisiae Ste50 binds the MAPKKK Ste11 through a head-to-tail SAM domain interaction. J Mol Biol. 2006 Feb 10;356(1):142-54. Epub 2005 Nov 28. PMID: 16337230

P38. Resnick AC, Snowman AM, Kang BN, Hurt KJ, Snyder SH, Saiardi A. Inositol polyphosphate multikinase is a nuclear PI3-kinase with transcriptional regulatory activity. Proc Natl Acad Sci U S A. 2005 Sep 6;102(36):12783-8. Epub 2005 Aug 25. PMID: 16123124

P39. Castrejon F, Gomez A, Sanz M, Duran A, Roncero C. The RIM101 pathway contributes to yeast cell wall assembly and its function becomes essential in the absence of mitogen-activated protein kinase Slt2p. Eukaryot Cell. 2006 Mar;5(3):507-17. PMID: 16524906

P40. Leger-Silvestre I, Caffrey JM, Dawaliby R, Alvarez-Arias DA, Gas N, Bertolone SJ, Gleizes PE, Ellis SR. Specific Role for Yeast Homologs of the Diamond Blackfan Anemia-associated Rps19 Protein in Ribosome Synthesis. J Biol Chem. 2005 Nov 18;280(46):38177-85. Epub 2005 Sep 12. PMID: 16159874

P41. Su X, Dowhan W. Regulation of cardiolipin synthase levels in Saccharomyces cerevisiae. Yeast. 2006 Mar;23(4):279-91. PMID: 16544270

P42. Kostelansky MS, Sun J, Lee S, Kim J, Ghirlando R, Hierro A, Emr SD, Hurley JH. Structural and functional organization of the ESCRT-I trafficking complex. Cell. 2006 Apr 7;125(1):113-26. PMID: 16615894

P43. Pinsky BA, Kotwaliwale CV, Tatsutani SY, Breed CA, Biggins S. Glc7/protein phosphatase 1 regulatory subunits can oppose the Ipl1/aurora protein kinase by redistributing Glc7. Mol Cell Biol. 2006 Apr;26(7):2648-60. PMID: 16537909

P44. Belanger KD, Gupta A, MacDonald KM, Ott CM, Hodge CA, Cole CM, Davis LI. Nuclear pore complex function in Saccharomyces cerevisiae is influenced by glycosylation of the transmembrane nucleoporin Pom152p. Genetics. 2005 Nov;171(3):935-47. Epub 2005 Aug 22. PMID: 16118201

P45. Lu A, Hirsch JP. Cyclic AMP-independent regulation of protein kinase A substrate phosphorylation by Kelch repeat proteins. Eukaryot Cell. 2005 Nov;4(11):1794-800. PMID: 16278446

P46. Sambade M, Alba M, Smardon AM, West RW, Kane PM. A genomic screen for yeast vacuolar membrane ATPase mutants. Genetics. 2005 Aug;170(4):1539-51. Epub 2005 Jun 3. PMID: 15937126

P47. Young MJ, Theriault SS, Li M, Court DA. The carboxyl-terminal extension on fungal mitochondrial DNA polymerases: identification of a critical region of the enzyme from Saccharomyces cerevisiae. Yeast. 2006 Jan 30;23(2):101-16. PMID: 16491467

P48. Sano T, Kihara A, Kurotsu F, Iwaki S, Igarashi Y. Regulation of the sphingoid long-chain base kinase Lcb4p by ergosterol and heme: studies in phytosphingosine-resistant mutants. J Biol Chem. 2005 Nov 4;280(44):36674-82. Epub 2005 Sep 1. PMID: 16141212

SMS Reliability study papers:

Curator RC1

RP1 (6 month interval):

Tsukada Y, Fang J, Erdjument-Bromage H, Warren ME, Borchers CH, Tempst P, Zhang Y. Histone demethylation by a family of JmjC domain-containing proteins. Nature. 2006 Feb 16;439(7078):811-6. PMID: 16362057

RP2 (12 month interval):

Shaner L, Trott A, Goeckeler JL, Brodsky JL, Morano KA. The function of the yeast molecular chaperone Sse1 is mechanistically distinct from the closely related hsp70 family. J Biol Chem. 2004 May 21;279(21):21992-2001. PMID: 15028727

Curator RC2

RP3 (6 month interval):

Saito K, Fujimura-Kamada K, Furuta N, Kato U, Umeda M, Tanaka K. Cdc50p, a protein required for polarized growth, associates with the Drs2p P-type ATPase implicated in phospholipid translocation in Saccharomyces cerevisiae. Mol Biol Cell. 2004 Jul;15(7):3418-32. PMID: 15090616

RP4 (12 month interval):

Paumet F, Rahimian V, Rothman JE. The specificity of SNARE-dependent fusion is encoded in the SNARE motif. Proc Natl Acad Sci U S A. 2004 Mar 9;101(10):3376-80. PMID: 14981247

A7. Article corpus for Multi-MOD study

This appendix lists the articles annotated in the MMS study by paper number.

P1. Noma A, Kirino Y, Ikeuchi Y, Suzuki T. Biosynthesis of wybutosine, a hyper-modified nucleoside in eukaryotic phenylalanine tRNA. EMBO J. 2006 May 17;25(10):2142-54. Epub 2006 Apr 27. PMID: 16642040

P2. Warringer J, Blomberg A. Involvement of yeast YOL151W/GRE2 in ergosterol metabolism. Yeast. 2006 Apr 15;23(5):389-98. PMID: 16598690

P3. Malone CJ, Misner L, Le Bot N, Tsai MC, Campbell JM, Ahringer J, White JG. The C. elegans hook protein, ZYG-12, mediates the essential attachment between the centrosome and nucleus. Cell. 2003 Dec 26;115(7):825-36. PMID: 14697201

P4. Wu YC, Tsai MC, Cheng LC, Chou CJ, Weng NY. C. elegans CED-12 acts in the conserved crkII/DOCK180/Rac pathway to control cell migration and cell corpse engulfment. Dev Cell. 2001 Oct;1(4):491-502. PMID: 11703940

P5. Fahlgren N, Montgomery TA, Howell MD, Allen E, Dvorak SK, Alexander AL, Carrington JC. Regulation of AUXIN RESPONSE FACTOR3 by TAS3 ta-siRNA affects developmental timing and patterning in Arabidopsis. Curr Biol. 2006 May 9;16(9):939-44. PMID: 16682356

P6. Auldridge ME, Block A, Vogel JT, Dabney-Smith C, Mila I, Bouzayen M, Magallanes-Lundback M, DellaPenna D, McCarty DR, Klee HJ. Characterization of three members of the Arabidopsis carotenoid cleavage dioxygenase family demonstrates the divergent roles of this multifunctional enzyme family. Plant J. 2006 Mar;45(6):982-93. PMID: 16507088

P7. Beigneux AP, Vergnes L, Qiao X, Quatela S, Davis R, Watkins SM, Coleman RA, Walzem RL, Philips M, Reue K, Young SG. Agpat6--a novel lipid biosynthetic gene required for triacylglycerol production in mammary epithelium. J Lipid Res. 2006 Apr;47(4):734-44. Epub 2006 Jan 31. PMID: 16449762

P8. Kim YS, Nakanishi G, Lewandoski M, Jetten AM. GLIS3, a novel member of the GLIS subfamily of Kruppel-like zinc finger proteins with repressor and activation functions. Nucleic Acids Res. 2003 Oct 1;31(19):5513-25. PMID: 14500813

P9. Jagadish N, Rana R, Selvi R, Mishra D, Garg M, Yadav S, Herr JC, Okumura K, Hasegawa A, Koyama K, Suri A. Characterization of a novel human sperm-associated antigen 9 (SPAG9) having structural homology with c-Jun N-terminal kinase-interacting protein. Biochem J. 2005 Jul 1;389(Pt 1):73-82. PMID: 15693750

P10. Abdul KM, Terada K, Yano M, Ryan MT, Streimann I, Hoogenraad NJ, Mori M. Functional analysis of human metaxin in mitochondrial protein import in cultured cells and its relationship with the Tom complex. Biochem Biophys Res Commun. 2000 Oct 5;276(3):1028-34. PMID: 11027586

A8. UNC IRB application and approval

These studies were reviewed and approved by the University of North Carolina's Institutional Review Board on 2006-04-04, and renewed on 2007-03-22, with an expiration date of 2008-03-20. The study number is 06-0044. Portions of the IRB documents that follow are redacted to preserve the anonymity and privacy of study participants.

OFFICE OF HUMAN RESEARCH ETHICS
Institutional Review Board

APPLICATION FOR IRB APPROVAL OF HUMAN SUBJECTS RESEARCH
Version 7-Mar-2006

Part A.1. Contact Information, Agreements, and Signatures

Title of Study: Strategies for information integration using annotation evidence **Date:** 2006-03-27

Name and degrees of Principal Investigator: W. John MacMullen, B.S., M.S.I.S.
Department: School of Information & Library Science Mailing address/CB #: 3360
UNC-CH PID: xxxxxxxxx Pager:
Phone #: 919-xxx-xxxx Fax #: Email Address:

For trainee-led projects: __ undergraduate _x_ graduate __ postdoc __ resident __ other

Name of faculty advisor: Gary Marchionini
Department: School of Information & Library Science Mailing address/CB #: 3360
Phone #: 919-xxx-xxxx Fax #: Email Address:

Name, phone number, email address of project manager or coordinator, if any: none

List **all other project personnel** including co-investigators, and anyone else who has contact with subjects or identifiable data from subjects: none

Name of funding source or sponsor:
__ not funded _x_ Federal __ State __ industry __ foundation __ UNC-CH
__ other (specify): **Sponsor or award number**: NIH/NLM 1 F37 LM009194-01

Include following items with your submission, where applicable.

- Check the relevant items below and include one copy of all checked items 1-11 in the order listed.
- Also include two additional collated sets of copies (sorted in the order listed) for items 1-7.

Check	Item	Total No. of Copies
x	1. This application. One copy must have original PI signatures.	3
x	2. Consent and assent forms, fact or information sheets; include phone and verbal consent scripts.	3
□	3. HIPAA authorization addendum to consent form.	3
□	4. All recruitment materials including scripts, flyers and advertising, letters, emails.	3
x	5. Questionnaires, focus group guides, scripts used to guide phone or in-person interviews, etc.	3
x	6. Protocol, grant application or proposal supporting this submission; (e.g., extramural grant application to NIH or foundation, industry protocol, student proposal).	3
x	7. Documentation of reviews from any other committees (e.g., GCRC, Oncology Protocol Review Committee, or local review committees in Academic Affairs).	3
□	8. Addendum for Multi-Site Studies where UNC-CH is the Lead Coordinating Center.	1
□	9. Data use agreements (may be required for use of existing data from third parties).	1
x	10. Documentation of required training in human research ethics for all study personnel.	1
□	11. Investigator Brochure if a drug study.	1

Principal Investigator: I will personally conduct or supervise this research study. I will ensure that this study is performed in compliance with all applicable laws, regulations and University policies regarding human subjects research. I will obtain IRB approval before making any changes or additions to the project. I will notify the IRB of any other changes in the information provided in this application. I will provide progress reports to the IRB at least annually, or as requested. I will report promptly to the IRB all unanticipated problems or serious adverse events involving risk to human subjects. I will follow the IRB approved consent process for all subjects. I will ensure that all collaborators, students and employees assisting in this research study are informed about these obligations. All information given in this form is accurate and complete.

Signature of Principal Investigator Date

Faculty Advisor if PI is a Student or Trainee Investigator: I accept ultimate responsibility for ensuring that this study complies with all the obligations listed above for the PI.

Signature of Faculty Advisor Date

Department or Division Chair, Center Director (or counterpart) of PI: (or Vice-Chair or Chair's designee if Chair is investigator or otherwise unable to review): I certify that this research is appropriate for this Principal Investigator, that the investigators are qualified to conduct the research, and that there are adequate resources (including financial, support and facilities) available. If my unit has a local review committee for pre-IRB review, this requirement has been satisfied. I support this application, and hereby submit it for further review.

Signature of Department Chair or designee Date

Print Name of Department Chair or designee Department

Part A.2. Summary Checklist

Are the following involved?	Yes	No
A.2.1. Existing data, research records, patient records, and/or human biological specimens?	_x_	__
A.2.2. Surveys, questionnaires, interviews, or focus groups with subjects?	_x_	__
A.2.3. Videotaping, audiotaping, filming of subjects (newly collected or existing)?	_x_	__
A.2.4. Do you plan to enroll subjects from these vulnerable or select populations: a. UNC-CH students or UNC-CH employees? .... b. Non-English-speaking? .... c. Decisionally impaired? .... d. Patients? .... e. Prisoners, others involuntarily detained or incarcerated, or parolees? .... f. Pregnant women? .... g. Minors (less than 18 years)? ***If yes***, give age range: to years	 __ __ __ __ __ __ __	 _x_ _x_ _x_ _x_ _x_ _x_ _x_
A.2.5. a. Is this a multi-site study (sites outside UNC-CH engaged in the research)? b. Is UNC-CH the sponsor or lead coordinating center? ***If yes*****,** include the ***Addendum for Multi-site Studies where UNC-CH is the Lead Coordinating Center***. ***If yes***, will any of these sites be outside the United States? ***If yes***, provide contact information for the foreign IRB.	__ __ __	_x_ __ __
A.2.6. Will there be a data and safety monitoring committee (DSMB or DSMC)?	__	_x_
A.2.7. a. Are you collecting sensitive information such as sexual behavior, HIV status, recreational drug use, illegal behaviors, child/physical abuse, immigration status, etc? b. Do you plan to obtain a federal Certificate of Confidentiality for this study?	__ __	_x_ _x_
A.2.8. a. Investigational drugs? (provide **IND #**) b. Approved drugs for "non-FDA-approved" conditions? *All studies testing substances in humans must provide a letter of acknowledgement from the UNC Health Care Investigational Drug Service (IDS).*	__ __	_x_ _x_
A.2.9. Placebo(s)?	__	_x_
A.2.10. Investigational devices, instruments, machines, software? (provide **IDE #**)	__	_x_
A.2.11. Fetal tissue?	__	_x_
A.2.12. Genetic studies on subjects' specimens?	__	_x_
A.2.13. Storage of subjects' specimens for future research? ***If yes****, see instructions for* ***Consent for Stored Samples****.*	__	_x_
A.2.14. Diagnostic or therapeutic ionizing radiation, or radioactive isotopes, which subjects would not receive otherwise? ***If yes****, approval by the* ***UNC-CH Radiation Safety Committee*** *is required.*	__	_x_
A.2.15. Recombinant DNA or gene transfer to human subjects? ***If yes****, approval by the* ***UNC-CH Institutional Biosafety Committee*** *is required.*	__	_x_
A.2.16. Does this study involve UNC-CH cancer patients? ***If yes****, submit this application directly to the* ***Oncology Protocol Review Committee****.*	__	_x_
A.2.17. Will subjects be studied in the General Clinical Research Center (GCRC)? ***If yes****, obtain the* ***GCRC Addendum*** *from the GCRC and submit complete application (IRB application and Addendum) to the GCRC.*	__	_x_

Part A.3. Conflict of Interest Questions and Certification

The following questions apply to **all investigators and study staff** engaged in the design, conduct, or reporting results of this project **and/or their immediate family members.** For these purposes, "family" includes the individual's spouse and dependent children. "Spouse" includes a person with whom one lives together in the same residence and with whom one shares responsibility for each other's welfare and shares financial obligations.

A.3.1. Currently or during the term of this research study, does any member of the research team or his/her family member have or expect to have:		
(a) A personal financial interest in or personal financial relationship (including gifts of cash or in-kind) with the sponsor of this study?	__ yes	_x_ no
(b) A personal financial interest in or personal financial relationship (including gifts of cash or in-kind) with an entity that owns or has the right to commercialize a product, process or technology studied in this project?	__ yes	_x_ no
(c) A board membership of any kind or an executive position (paid or unpaid) with the sponsor of this study or with an entity that owns or has the right to commercialize a product, process or technology studied in this project?	__ yes	_x_ no
A.3.2. Has the University or has a University-related foundation received a cash or in-kind gift from the Sponsor of this study for the use or benefit of any member of the research team?	__ yes	_x_ no
A.3.3. Has the University or has a University-related foundation received a cash or in-kind gift for the use or benefit of any member of the research team from an entity that owns or has the right to commercialize a product, process or technology studied in this project?	__ yes	_x_ no

If the answer to ANY of the questions above is *yes*, the affected research team member(s) must complete and submit to the Office of the University Counsel the form accessible at http://coi.unc.edu. List name(s) of all research team members for whom any answer to the questions above is *yes*:

____

<u>Certification by Principal Investigator</u>: By submitting this IRB application, I (the PI) certify that the information provided above is true and accurate regarding my own circumstances, that I have inquired of every UNC-Chapel Hill employee or trainee who will be engaged in the design, conduct or reporting of results of this project as to the questions set out above, and that I have instructed any such person who has answered "yes" to any of these questions to complete and submit for approval a Conflict of Interest Evaluation Form. I understand that as Principal Investigator I am obligated to ensure that any potential conflicts of interest that exist in relation to my study are reported as required by University policy.

____ __________________

Signature of Principal Investigator Date

<u>Faculty Advisor if PI is a Student or Trainee Investigator</u>: I accept ultimate responsibility for ensuring that the PI complies with the University's conflict of interest policies and procedures.

____ __________________

Signature of Faculty Advisor Date

Part A.4. Questions Common to All Studies

For all questions, if the study involves only secondary data analysis, focus on your proposed design, methods and procedures, and not those of the original study that produced the data you plan to use.

A.4.1. **Brief Summary**. Provide a *brief* non-technical description of the study, which will be used for internal and external communications regarding this research. Include purpose, methods, and participants. Typical summaries are 50-100 words.

This study investigates how expert human curators extract knowledge from scientific articles and create intellectual linkages ('annotations') between that knowledge and specific gene products in databases of model organism information, with the goal of developing assistive strategies and tools for information integration. The annotation work processes of approximately 15 curators will be studied through observation, interviews, structured tasks, artifact analysis, and group discussions.

A.4.2. **Purpose and Rationale**. Provide a summary of the background information, state the research question(s), and tell why the study is needed. If a complete rationale and literature review are in an accompanying grant application or other type of proposal, only provide a brief summary here. If there is no proposal, provide a more extensive rationale and literature review, including references.

The creation and use of annotations in biological databases is a growing part of biomedical research, due in part to large-scale genome sequencing. At the same time, scientific knowledge in the form of published journal articles continues to grow faster than individuals' abilities to synthesize them, leading to information overload and knowledge fragmentation by specialty. While the linkage of published scientific knowledge with primary data in databases could potentially provide significant integration and defragmentation of scientific knowledge, annotation as a work practice is an understudied area of research. This study will investigate the nature of annotation work processes and artifacts in the context of Gene Ontology annotations in model organism databases, including measures of quality, such as consistency, reliabilty, accuracy, and completeness. [Complete rationale and literature review are in the accompanying grant proposal.]

A.4.3. **Subjects.** *You should describe the subject population even if your study does not involve direct interaction (e.g., existing records).* Specify number, gender, ethnicity, race, and age. Specify whether subjects are healthy volunteers or patients. If patients, specify any relevant disease or condition and indicate how potential subjects will be identified.

The subjects are adult curators of biological databases, typically holding Ph.D degrees in biomedical sciences. The subject population is approximately 15 healthy volunteers of various genders, ethnicities, races, and ages.

A.4.4. **Inclusion/exclusion criteria.** List required characteristics of potential subjects, and those that preclude enrollment or involvement of subjects or their data. Justify exclusion of any group, especially by criteria based on gender, ethnicity, race, or age. If pregnant women are excluded, or if women who become pregnant are withdrawn, specific justification must be provided.

All curators from relevant biological databases are eligible to participate regardless of gender, ethnicity, or race. Children are excluded from this research since none are employed as curators.

A.4.5. **Full description of the study design, methods and procedures.** Describe the research study. Discuss the study design; study procedures; sequential description of what subjects will be asked to do; assignment of subjects to various arms of the study if applicable; doses; frequency and route of administration of medication and other medical treatment if applicable; how data are to be collected (questionnaire, interview, focus group or specific procedure such as physical examination, venipuncture, etc.). Include information on who will collect data, who will conduct procedures or measurements. Indicate the number and duration of contacts with each subject; outcome measurements; and follow-up procedures.

The study addresses three areas related to annotation in one type of biological database (those focused on a particular model organism), covered by the following aims:

Aim 1. *Analysis of annotation processes and artifacts*. This aim studies the artifacts of model organism databases (MODs) in the form of Gene Ontology annotations and the evidence sources (e.g., published scientific articles) from which they are derived, through the use of quantitative and qualitative methods. Specific samples of publicly-available annotations from multiple MODs will be acquired, quantified by measures (such as publication year and language, journal title), and categorized by facets (such as index term overlap). The human and mechanical processes by which annotations in MODs are created and maintained will be studied through individual interviews and observations with each participating subject, and by conducting one or more focus groups with the population of human curators described in A.4.3. (example interview and focus group questions attached). Subjects will be asked to ‘think aloud’ as they perform tasks, in order to understand tacit knowledge and non-obvious work facets. Artifacts relating to the annotation process, such as marked-up source articles, personal notes, and computer screens, will be collected with the subjects’ permission for additional contextual analysis to provide a fuller picture of the process.

Aim 2. *Assessment of annotation quality metrics*. Subsequent to the annotation process analysis in Aim 1, the curators will be asked to perform their normal work processes on a predefined set of scientific articles for the purpose of evaluating the validity of a set of annotation quality metrics. The set of articles will be divided in equal parts among two or more equal groups of curators (quantities depend upon enrollment). Each individual will annotate a subset of his/her group’s documents, with a predefined amount of overlap in shared documents. For example, assume 6 total curators who are randomly assigned to two groups of 3, and a total of 6 unique articles. With a group of 3 curators and a subset of 3 of the 6 total articles, each curator will annotate 2 articles, one of which a second curator in the group will also annotate. Each article will thus be annotated by 2 curators per group. The pairs of curators with shared articles will subsequently rationalize (if different) their individual annotations, leading to a single ‘gold standard’ annotation for that document. Following this step, each group will be assigned the 3 papers from the other group, and will repeat the task without knowledge of the other groups’ annotations. This design provides for 4x coverage of each article, and an ability to compare individual annotations against gold standards devised by independent groups, without an external standard process. The two rounds of annotation can be visualized in the following example (the actual experiment will have randomized group and article assignments and may vary in size):

	Round 1			Curators			
		Group 1			Group 2		
	Papers	C1	C2	C3	C4	C5	C6
	P1	x	x				
Set 1	P2	x		x			
	P3		x	x			
	P4				x	x	
Set 2	P5				x		x
	P6					x	x

Round 2			Curators			
	Group 1			Group 2		
Papers	C1	C2	C3	C4	C5	C6
P1				x	x	
P2				x		x
P3					x	x
P4	x	x				
P5	x		x			
P6		x	x			

The data from these experiments will be analyzed using the quality metrics defined in the attachment "Facets and measures of annotation quality in model organism databases".

Aim 3. *Development of one or more strategies for information integration* from the evidence in Aims 1 and 2. The information collected and analyzed in Aims 1 and 2 (particularly the differences in annotation processes, and in annotations of the same articles made by different curators), will be synthesized using the 'grounded theory' technique and other theory-building approaches, as described in the accompanying research proposal, to attempt to develop one or more strategies for information integration that could be tested in future studies. This work does not involve human subjects.

The work described in these aims (content analysis, interviews, observations, quality experiments, data analysis, and theory development) will be performed by the PI under the supervision of the faculty advisor. The contact with subjects described in aims 1 and 2 will be performed in the same time period in the subjects' normal work setting. The quantitative- and content analyses in Aim 1, and the theory development in Aim 3, will be performed by the PI at different times.

A.4.6. **Benefits to subjects and/or society.** Describe any potential for direct benefit to individual subjects, as well as the benefit to society based on scientific knowledge to be gained; these should be clearly distinguished. Consider the nature, magnitude, and likelihood of any direct benefit to subjects. If there is no direct benefit to the individual subject, say so here and in the consent form (if there is a consent form). Do not list monetary payment or other compensation as a benefit.

The anticipated societal benefits include a greater understanding of an important but little-studied area, as well as the development of one or more strategies for information integration that could reduce information overload and knowledge fragmentation. Research subjects are not expected to derive immediate and specific benefits from the study, but may experience a broader understanding of the nature of the annotation process that could improve their subsequent individual annotation performance.

A.4.7. **Full description of risks and measures to minimize risks.** Include risk of psychosocial harm (e.g., emotional distress, embarrassment, breach of confidentiality), economic harm (e.g., loss of employment or insurability, loss of professional standing or reputation, loss of standing within the community) and legal jeopardy (e.g., disclosure of illegal activity or negligence), as well as known side effects of study medication, if applicable, and risk of pain and physical injury. Describe what will be done to minimize these risks. Describe procedures for follow-up, when necessary, such as when subjects are found to be in need of medical or psychological referral. If there is no direct interaction with subjects, and risk is limited to breach of confidentiality (e.g., for existing data), state this.

No risks to participants are anticipated given the nature of the information they are providing. Interviews and observations will take place in private. Individual performance on the experimental annotation tasks will not be reported by curator name, only by a code as described below.

A.4.8. **Data analysis.** Tell how the qualitative and/or quantitative data will be analyzed. Explain how the sample size is sufficient to achieve the study aims. This might include a formal power calculation or explanation of why a small sample is sufficient (e.g., qualitative research, pilot studies).

The small number of participants in Aims 1 and 2 is not a concern to validity because the nature of the research is focused on the elicitation of process details and the trial of quality measures, not a large-scale assessment of annotation quality in a particular MOD. The experimental data from Aim 2 will be analyzed with standard inter-coder reliability measures such as Cohen's kappa and its variants.

A.4.9. **Will you collect or receive any of the following identifiers?** Does not apply to consent forms.

__ No _x_ Yes *If yes, check all that apply*:

a. _x_ Names
b. __ Telephone numbers
c. __ Any elements of dates (other than year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death. For ages over 89: all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 and older
d. __ Any geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code and their equivalent geocodes, except for the initial three digits of a zip code
e. __ Fax numbers
f. _x_ Electronic mail addresses
g. __ Social security numbers
h. __ Medical record numbers
i. __ Health plan beneficiary numbers
j. __ Account numbers
k. __ Certificate/license numbers
l. __ Vehicle identifiers and serial numbers (VIN), including license plate numbers
m. __ Device identifiers and serial numbers (e.g., implanted medical device)
n. _x_ Web universal resource locators (URLs)
o. __ Internet protocol (IP) address numbers
p. __ Biometric identifiers, including finger and voice prints
q. __ Full face photographic images and any comparable images
r. __ Any other unique identifying number, characteristic or code, other than dummy identifiers that are not derived from actual identifiers and for which the re-identification key is maintained by the health care provider and not disclosed to the researcher

A.4.10. **Confidentiality of the data**. Describe procedures for maintaining confidentiality of the data you will collect or will receive. Describe how you will protect the data from access by those not authorized. How will data be transmitted among research personnel? Where relevant, discuss the potential for deductive disclosure (i.e., directly identifying subjects from a combination of indirect IDs).

For analysis purposes, personal identities are not relevant as data will be aggregated. Each subject will be assigned a unique identifier when transcribing and coding the data. The consent forms and the file linking the identifier with the subject's name will be stored securely as described in A.4.12 and A.4.13 below.

A.4.11. **Data sharing.** With whom will *identifiable* (contains any of the 18 identifiers listed in question 9 above) data be shared outside the immediate research team? For each, explain confidentiality measures. Include data use agreements, if any.

_x_ No one
__ Coordinating Center**:**
__ Statisticians**:**
__ Consultants**:**
__ Other researchers**:**
__ Registries**:**
__ Sponsors:
__ External labs for additional testing:
__ Journals:
__ Publicly available dataset:
__ Other:

A.4.12. **Data security for storage and transmission**. Please check all that apply.

For electronic data:

__ Secure network _x_ Password access __ Encryption

_x_ Other (describe): Audio recordings will not identify participant names.

_x_ Portable storage (e.g., laptop computer, flash drive)

Describe how data will be protected for any portable device: The electronic name/identifier linkage data will be stored in a password-protected file on portable storage media (floppy disk or flash drive) in a locked cabinet for the duration of the study, and then destroyed. Data stored on investigator's laptop and backup disks will be the coded, de-identified data only.

For hardcopy data (including human biological specimens, CDs, tapes, etc.):

_x_ Data de-identified by research team (stripped of the 18 identifiers listed in question 7 above)

__ Locked suite or office

_x_ Locked cabinet

_x_ Data coded by research team with a master list secured and kept separately

_x_ Other (describe): Consent forms will be stored in the same locked cabinet as linkage file.

A.4.13. **Post-study disposition of identifiable data or human biological materials**. Describe your plans for disposition of data or human biological specimens that are identifiable in any way (directly or via indirect codes) once the study has ended. Describe your plan to destroy identifiers, if you will do so.

The name/identifier linkage file described in A.4.12 will be destroyed after the completion of the study by erasure of the electronic media.

Part A.5. The Consent Process and Consent Documentation (including Waivers)

The standard consent process is for all subjects to sign a document containing all the elements of informed consent, as specified in the federal regulations. Some or all of the elements of consent, including signatures, may be altered or waived under certain circumstances.

- If you will obtain consent in any manner, complete **section A.5.1**.
- If you are obtaining consent, but requesting a waiver of the requirement for a signed consent document, complete **section A.5.2**.
- If you are requesting a waiver of any or all of the elements of consent, complete **section A.5.3**.

You may need to complete more than one section. For example, if you are conducting a phone survey with verbal consent, complete sections A.5.1, A.5.2, and possibly A.5.3.

A.5.1. **Describe the process of obtaining informed consent from subjects**. If children will be enrolled as subjects, describe the provisions for obtaining parental permission and assent of the child. If decisionally impaired adults are to be enrolled, describe the provision for obtaining surrogate consent from a legally authorized representative (LAR). If non-English speaking people will be enrolled, explain how consent in the native language will be obtained. Address both written translation of the consent and the availability of oral interpretation. *After you have completed this part A.5.1, if you are not requesting a waiver of any type, you are done with Part A.5.; proceed to Part B.*

Consent will be obtained in writing, in person, using the attached consent form. All potential participants are adult English speakers.

A.5.2. **Justification for a waiver of *written* (i.e., signed) consent**. *The default is for subjects to sign a written document that contains all the elements of informed consent.* Under limited circumstances, the requirement for a signed consent form may be waived by the IRB if either of the following is true:

a. The only record linking the subject and the research would be the consent document and the principal risk would be potential harm resulting from a breach of confidentiality (e.g., study involves sensitive data that could be damaging if disclosed). __ yes __ no
Explain.

b. The research presents no more than minimal risk of harm to subjects and involves no procedures for which written consent is normally required outside of the research context (e.g., phone survey). __ yes __ no
Explain.

If you checked "yes" to either, will consent be oral? Will you give out a fact sheet? Use an online consent form, or include information as part of the survey itself, etc?

→ If you have justified a waiver of written (signed) consent (A.5.2), you should complete A.5.3 *only* if your consent process will not include all the other elements of consent.

A.5.3. **Justification for a full or partial waiver of consent.** *The default is for subjects to give informed consent.* A waiver might be requested for research involving only existing data or human biological specimens (see also Part C). More rarely, it might be requested when the research design requires withholding some study details at the outset (e.g., behavioral research involving deception). In limited circumstances, parental permission may be waived. This section should also be completed for a waiver of HIPAA authorization if research involves Protected Health Information (PHI) subject to HIPAA regulation, such as patient records.

__ Requesting **waiver of some elements** (specify; see SOP 28 on the IRB web site):
__ Requesting **waiver of consent entirely**
If you check either of the boxes above, answer items a-f.. To justify a full waiver of the requirement for informed consent, you must be able to answer "yes" (or "not applicable" for question c) to items a-f. **Insert brief explanations that support your answers.**

a. Will the research involve <u>no greater than minimal risk</u> to subjects or to their privacy? __ yes __ no
Explain.

b. Is it true that the waiver will *not* adversely affect the rights and welfare of subjects? *(Consider the right of privacy and possible risk of breach of confidentiality in light of the information you wish to gather.)* __ yes __ no
Explain.

c. When applicable to your study, do you have plans to provide subjects with pertinent information after their participation is over? *(e.g., Will you provide details withheld during consent, or tell subjects if you found information with direct clinical relevance? This may be an uncommon scenario.)* __ yes __ not applicable
Explain.

d. Would the research be impracticable without the waiver? *(If you checked "yes," explain how the requirement to obtain consent would make the research impracticable, e.g., are most of the subjects lost to follow-up or deceased?).* __ yes __ no
Explain.

e. Is the risk to privacy reasonable in relation to benefits to be gained or the importance of the knowledge to be gained? __ yes __ no
Explain.

If you are accessing patient records for this research, you must also be able to answer "yes" to item f to justify a waiver of HIPAA authorization from the subjects.

f. Would the research be impracticable if you could not record (or use) Protected Health Information (PHI)? *(If you checked "yes," explain how* not *recording or using PHI would make the research impracticable).* __ yes __ no
Explain.

Part B. Questions for Studies that Involve Direct Interaction with Human Subjects

→ *If this does not apply to your study, do not submit this section.*

B.1. **Methods of recruiting.** Describe how and where subjects will be identified and recruited. Indicate who will do the recruiting, and tell how subjects will be contacted. Describe efforts to ensure equal access to participation among women and minorities. Describe how you will protect the privacy of potential subjects during recruitment. *For prospective subjects whose status (e.g., as patient or client), condition, or contact information is not publicly available (e.g., from a phone book or public web site), the initial contact should be made with legitimate knowledge of the subjects' circumstances. Ideally, the individual with such knowledge should seek prospective subjects' permission to release names to the PI for recruitment. Alternatively, the knowledgeable individual could provide information about the study, including contact information for the investigator, so that interested prospective subjects can contact the investigator.* Provide the IRB with a copy of any document or script that will be used to obtain the patients' permission for release of names or to introduce the study. Check with your IRB for further guidance.

The first group of subjects (N=~15) was identified through conversations with the PI of a particular model organism database who is interested in measures of annotation quality. The PI and curatorial staff have committed to participate at a high level. Additional participants from other MODs may be identified and recruited with their help if the first round of the study is useful and informative. No specific additional groups, PIs, or curators have been identified as of the date of this application.

B.2. **Protected Health Information (PHI).** If you need to access Protected Health Information (PHI) to identify potential subjects who will then be contacted, you will need a *limited waiver of HIPAA authorization*. If this applies to your study, please provide the following information.

a. Will the information collected be limited only to that necessary to contact the subjects to ask if they are interested in participating in the study?

b. How will confidentiality/privacy be protected prior to ascertaining desire to participate?

c. When and how will you destroy the contact information if an individual declines participation?

B.3. **Duration of entire study and duration of an individual subject's participation, including follow-up evaluation if applicable.** Include the number of required contacts and approximate duration of each contact.

The portions of the study involving human subjects (Aims 1 and 2) are estimated to over a one-week timeframe in June 2006, with the duration of individual subjects' cumulative participation (including interview, observation, annotation activities, and focus group) estimated at 6-8 hours, distributed over the 1-week period. Minimal follow-up on an individual basis may be required to clarify subjects' answers or descriptions if unclear during data transcription.

B.4. **Where will the subjects be studied?** Describe locations where subjects will be studied, both on and off the UNC-CH campus.

Subjects will be studied in their own work settings (typically academic research environments). There are currently no plans to recruit or study subjects on the UNC-CH campus.

B.5. **Privacy.** Describe procedures that will ensure privacy of the subjects in this study. Examples include the setting for interviews, phone conversations, or physical examinations; communication methods or mailed materials (e.g., mailings should not indicate disease status or focus of study on the envelope).

Interviews and observations will occur in private spaces reserved in subjects' workplace. Since the subjects in the defined group already know each other and work together within the same organization, no attempt will be made to restrict them from discussing the content of the focus group, or their individual experiences, with each other outside of the study. Data resulting from interactions with subjects will be protected as described above in A.4.12 and A.4.13. The consent form notes that the while individual participant names will not be disclosed, the database name may be acknowledged in publications.

B.6. **Inducements for participation.** Describe all inducements to participate, monetary or non-monetary. If monetary, specify the amount and schedule for payments and how this will be prorated if the subject withdraws (or is withdrawn) from the study prior to completing it. For compensation in foreign currency, provide a US$ equivalent. Provide evidence that the amount is not coercive (e.g., describe purchasing power for foreign countries). Include food or refreshments that may be provided.

No monetary compensation is being offered to induce participation in the study. As described in B.1. and B.5., the subjects are part of a single research organization that is interested in the topic of the proposed research for their own benefit.

B.7. **Costs to be borne by subjects.** Include child care, travel, parking, clinic fees, diagnostic and laboratory studies, drugs, devices, all professional fees, etc. If there are no costs to subjects other than their time to participate, indicate this.

No costs other than time are expected to be borne by subjects.

Part C. Questions for Studies using Data, Records or Human Biological Specimens without Direct Contact with Subjects

→ *If this does not apply to your study, do not submit this section.*

C.1. What records, data or human biological specimens will you be using? *(check all that apply)*:

_x_ Data already collected for another research study
__ Data already collected for administrative purposes (e.g., Medicare data, hospital discharge data)
__ Medical records (custodian may also require form, e.g., HD-974 if UNC-Health Care System)
__ Electronic information from clinical database (custodian may also require form)
__ Patient specimens (tissues, blood, serum, surgical discards, etc.)
_x_ Other (specify): Publicly-available documents (e.g., public database entries).

C.2. For each of the boxes checked in 1, how were the original data, records, or human biological specimens collected? Describe the process of data collection including consent, if applicable.

The data are existing annotations from publicly-available databases that may be used in comparison with the data collected from subjects. The public annotation data does not have individual names associated with it.

C.3. For each of the boxes checked in 1, where do these data, records or human biological specimens currently reside?

Principal Investigator's laptop and backups.

C.4. For each of the boxes checked in 1, from whom do you have permission to use the data, records or human biological specimens? Include data use agreements, if required by the custodian of data that are not publicly available.

All data are publicly available from Internet-accessible databases, most federally-funded.

C.5. If the research involves human biological specimens, has the purpose for which they were collected been met before removal of any excess? For example, has the pathologist in charge or the clinical laboratory director certified that the original clinical purpose has been satisfied? Explain if necessary.

__ yes __ no __ not applicable (explain)

C.6. Do all of these data records or specimens exist at the time of this application? If not, explain how prospective data collection will occur.

__ yes _x_ no If no, explain: The data exists today, but is subject to continual revision and updating due to the nature of the information resources.

A9. Participant consent forms

For the single-MOD study, the original consent form as approved by the UNC IRB was used (attached following this page). For the multi-MOD study, the same form was used, but some of the activities listed under the section ‘What will happen if you take part in the study?’ were not performed by the participants, so those portions were marked out. Specifically, for the expert participants, formal observations and concurrent verbal reports were not conducted, and no focus groups were conducted. There were also three GO Camp attendees who submitted their annotations; their forms requested consent for only the item ‘Asking you to perform typical work tasks with specific documents’; all others were marked out, as well as the statement asking for consent to make audio recordings, under the section ‘How will your privacy be protected?’

University of North Carolina-Chapel Hill
Consent to Participate in a Research Study
Adult Participants
Social Behavioral Form

IRB Study # 06-0044
Consent Form Version Date: 9-Mar-2005

Title of Study: Strategies for information integration using annotation evidence

Principal Investigator: W. John MacMullen
UNC-Chapel Hill Department: School of Information and Library Science
UNC-Chapel Hill Phone number: 919-962-8366
Email Address:
Faculty Advisor: Gary J. Marchionini
Funding Source: Microsoft Research; National Library of Medicine

Study Contact telephone number: 919-xxx-xxxx

<u>What are some general things you should know about research studies?</u>
You are being asked to take part in a research study. To join the study is voluntary. You may refuse to join, or you may withdraw your consent to be in the study, for any reason, without penalty.

Research studies are designed to obtain new knowledge. This new information may help people in the future. You may not receive any direct benefit from being in the research study. There also may be risks to being in research studies.

Details about this study are discussed below. It is important that you understand this information so that you can make an informed choice about being in this research study.
You will be given a copy of this consent form. You should ask the researchers named above, or staff members who may assist them, any questions you have about this study at any time.

<u>What is the purpose of this study?</u>
The purpose of this research study is to learn about annotations in biological databases. We are interested in how annotations are made, how they are used, and their features. We are also studying the validity of specific measures of annotation quality, such as consistency, reliability, accuracy, and completeness.

You are being asked to be in the study because as a biological database curator, your expertise is valuable in helping us understand the creation and maintenance of annotations in databases.

<u>How many people will take part in this study?</u>
If you decide to participate, you will be one of approximately 15 people in the study.

How long will your part in this study last?
Your participation in this study will take approximately 6-8 hours, but may be distributed over multiple days at your convenience.

What will happen if you take part in the study?
We are using multiple research methods to understand annotation practices in biological databases. Specifically, we will be:

- *Asking you how you do your work*. We will ask you to describe in detail what your normal work tasks consist of, and how you complete them, including what knowledge and skills are required. This discussion will be audio recorded to ensure accurate transcription.
- *Observing you as you do your work*. By observing how you complete actual work tasks in a natural setting, we may learn more about the annotation process than by discussion alone.
- *Asking you to 'think aloud' while doing your work*. This means that at certain times during the observation, we will ask you to describe what you are thinking and doing as you perform a particular work task. This process will include audio recordings of your spoken thoughts to ensure we have described them accurately.
- *Viewing artifacts related to your work*. Looking at artifacts such as documents, notes, and computer screens you create and use in you work will help us understand what you do.
- *Asking you to perform typical work tasks with specific documents*. For the purposes of evaluating different measures of annotation quality, we will ask you to perform your normal annotation-related tasks using predefined documents to allow for comparisons with other participants. You will be assigned articles at random to annotate. Afterward, you will also collaborate with other curators who have made annotations to the same documents, in order to arrive at a consensus annotation. Your annotations will also be compared to other curators' consensus annotations to see if there are differences.
- *Asking you to participate in a group discussion ('focus group')*. The group will be asked to discuss general questions about annotation in biological databases, and the quality measures being studied. No questions will be directed to you individually, but instead will be posed to the group. You may choose to respond or not respond at any point during the discussion. The focus group discussion will be audio recorded to ensure accurate transcription. Participants will not be identified by their real names in transcripts.

What are the possible benefits from being in this study?
Research is designed to benefit society by gaining new knowledge. You may also expect to benefit by participating in this study by learning more about annotation practices from the research findings and the other curators. This may help you in your future work.

What are the possible risks or discomforts involved from being in this study?
There are no anticipated risks to you as a participant of this study. There may be uncommon or previously unknown risks. You should report any problems to the researcher.

How will your privacy be protected?
You will not be identified by name in any report or publication about this study, but the name of the database may be acknowledged in publications. Although every effort will be made to keep research records private, there may be times when federal or state law requires the disclosure of such records, including personal information. This is very unlikely, but if disclosure is ever

required, UNC-Chapel Hill will take steps allowable by law to protect the privacy of personal information. In some cases, your information in this research study could be reviewed by representatives of the University, research sponsors, or government agencies for purposes such as quality control or safety.

Your participation will be audio recorded in certain cases as described above. Recordings are used to create transcripts to ensure accurate descriptions of your participation. Your real name will not be used in transcripts. The recordings and the files that link your real name to the anonymous identifier used in the transcripts will be stored securely. Only the researcher and faculty advisor will have access to these materials. The original recordings and the file linking your name to the anonymous identifier will be destroyed after the conclusion of the study.

Will you receive anything for being in this study?
You will not receive anything for taking part in this study.

Will it cost you anything to be in this study?
There will be no costs for being in the study other than your participation time.

What if you have questions about this study?
You have the right to ask, and have answered, any questions you have about this research. If you have questions or concerns, you should contact the researcher listed on the first page of this form.

What if you have questions about your rights as a research participant?
All research on human volunteers is reviewed by a committee that works to protect your rights and welfare. If you have questions or concerns about your rights as a research subject you may contact, anonymously if you wish, the Institutional Review Board at 919-966-3113 or by email to IRB_subjects@unc.edu.

- -

Participant's Agreement:

I have read the information provided above. I have asked all the questions I have at this time. I voluntarily agree to participate in this research study.

____ __________________
Signature of Research Participant Date

____
Printed Name of Research Participant

____ __________________
Signature of Person Obtaining Consent Date

____
Printed Name of Person Obtaining Consent

A10. Original consensus annotations for MMS P2 quality facet calculations

Original annotations.

Gene	Aspect	GOID	GO Term	EC	With/from	Qualifier
C1C3						
GRE2	P	6696	ergosterol biosynthesis	IMP		
GRE2	P	6950	response to stress	IMP		
C2C10						
GRE2	P	6950	response to stress	IMP		
GRE2	P	6696	ergosterol biosynthesis	IGI	SGD:S000001049| SGD:S000005224| SGD:S000004815| SGD:S000004540| SGD:S000004442	
GRE2	P	16044	membrane organization and biogenesis	IMP		
C4C7						
GRE2	P	6696	ergosterol biosynthesis	IMP		
GRE2	P	46467	membrane lipid biosynthesis	IMP		
GRE2	P	9651	response to salt stress	IMP		
GRE2	P	9651	response to cation stress	IMP		
C5C9						
GRE2	P	6696	ergosterol biosynthesis	IMP		
GRE2	P	9651	response to salt stress	IMP		
GRE2	P	51592	response to calcium ion	IMP		
C6C8						
GRE2	P	9651	response to salt stress	IMP		
GRE2	P	43157	response to cation stress	IMP		
GRE2	P	6970	ergosterol metabolism	IMP		
GRE2	P	6970	response to osmotic stress	IMP		
ENO1	P	43157	response to cation stress	IEP		
TDH1	P	43157	response to cation stress	IEP		
HXK1	P	43157	response to cation stress	IEP		

Gene	Aspect	GOID	GO Term	EC	With/from	Qualifier
RNR4	P	43157	response to cation stress	IEP		
BGL2	P	43157	response to cation stress	IEP		
ERG10	P	43157	response to cation stress	IEP		
ERG6	P	43157	response to cation stress	IEP		
MVD1	P	43157	response to cation stress	IEP		
GRE2	F	16491	oxidoreductase activity	ISS	NCBI: AK121335	
GRE2	F	47044	3-alpha(or 20-beta)-hydroxysteroid dehydrogenase activity	ISS	EMBL: M38180	
GRE2	F	0252	C-3 sterol dehydrogenase (C-4 sterol decarboxylase) activity	ISS	SGD: S00002969	
C11C15						
GRE2	P	6696	ergosterol biosynthesis	IMP		
GRE2	P	6970	response to osmotic stress	IMP		
GRE2	F	16491	oxidoreductase activity	ISS		
C13C22						
GRE2	P	8204	ergosterol metabolism	IMP		
GRE2	P	43157	response to cation stress	IMP		
GRE2	P	9651	response to salt stress	IMP		
GRE2	P	6950	response to stress	IMP		
GRE2	P	46677	response to antibiotic	IMP		
GRE2	F	16491	oxidoreductase activity	ISS	SGD: S000002969	
ERG10	P	6696	ergosterol biosynthesis	IGI	SGD: S000005511	
ERG6	P	6696	ergosterol biosynthesis	IGI	SGD: S000005511	
ERG19	P	6696	ergosterol biosynthesis	IGI	SGD: S000005511	
C18C21						
GRE2	P	6950	response to stress	IMP		
GRE2	P	9651	response to salt stress	IMP		
GRE2	P	43157	response to cation stress	IMP		

Gene	Aspect	GOID	GO Term	EC	With/from	Qualifier
GRE2	P	8204	ergosterol metabolism	IDA		
GRE2	P	16125	sterol metabolism	IDA		
GRE2	P	16044	membrane organization and biogenesis	IMP		
GRE2	P	19725	cell homeostasis	IMP		
GRE2	P	42391	regulation of membrane potential	IMP		
GRE2	P	6696	ergosterol biosynthesis	IMP		
Ydr541cp	P	16125	sterol metabolism	ISS	Q12068	
Ygl139wp	P	16125	sterol metabolism	ISS	Q12068	
Ygl157wp	P	16125	sterol metabolism	ISS	Q12068	
C12C14C19C20						
GRE2	P	48193	Golgi vesicle transport	IMP		
GRE2	P	6950	response to stress	IMP		
GRE2	P	8204	ergosterol metabolism	IMP		

Evaluation set.

Gene	Aspect	GOID	GO Term	EC	With/from	Qualifier
C1C3						
GRE2	P	6696	ergosterol biosynthesis	IMP		
GRE2	P	6950	response to stress	IMP		
C2C10						
GRE2	P	6950	response to stress	IMP		
GRE2	P	6696	ergosterol biosynthesis	IGI	SGD:S000001049| SGD:S000005224| SGD:S000004815| SGD:S000004540| SGD:S000004442	
C4C7						
GRE2	P	6696	ergosterol biosynthesis	IMP		
GRE2	P	9651	response to salt stress	IMP		
C5C9						
GRE2	P	6696	ergosterol biosynthesis	IMP		
GRE2	P	9651	response to salt stress	IMP		
C6C8						
GRE2	P	9651	response to salt	IMP		

Gene	Aspect	GOID	GO Term	EC	With/from	Qualifier
			stress			
GRE2	P	6970	ergosterol metabolism	IMP		
C11C15						
GRE2	P	6696	ergosterol biosynthesis	IMP		
GRE2	P	6970	response to osmotic stress	IMP		
C13C22						
GRE2	P	8204	ergosterol metabolism	IMP		
GRE2	P	6950	response to stress	IMP		
C18C21						
GRE2	P	6950	response to stress	IMP		
GRE2	P	8204	ergosterol metabolism	IDA		
C12C14C19C20						
GRE2	P	6950	response to stress	IMP		
GRE2	P	8204	ergosterol metabolism	IMP		

Matched instances.

i	Pair	Gene	A	GOID	GO Term	EC	W/F	Q
	C1C3	*GRE2*	*P*	*6696*	*ergosterol biosynthesis*	*IMP*		
1	C2C10	GRE2	P	6696	ergosterol biosynthesis	IGI	[various]	
2	C4C7	GRE2	P	6696	ergosterol biosynthesis	IMP		
3	C5C9	GRE2	P	6696	ergosterol biosynthesis	IMP		
4	C6C8	GRE2	P	6970	ergosterol metabolism	IMP		
5	C11C15	GRE2	P	6696	ergosterol biosynthesis	IMP		
6	C13C22	GRE2	P	8204	ergosterol metabolism	IMP		
7	C18C21	GRE2	P	8204	ergosterol metabolism	IDA		
8	C12C14 C19C20	GRE2	P	8204	ergosterol metabolism	IMP		
	C1C3	*GRE2*	*P*	*6950*	*response to stress*	*IMP*		
9	C2C10	GRE2	P	6950	response to stress	IMP		
10	C4C7	GRE2	P	9651	response to salt stress	IMP		
11	C5C9	GRE2	P	9651	response to salt stress	IMP		
12	C6C8	GRE2	P	9651	response to salt stress	IMP		
13	C11C15	GRE2	P	6970	response to osmotic stress	IMP		
14	C13C22	GRE2	P	6950	response to stress	IMP		
15	C18C21	GRE2	P	6950	response to stress	IMP		
16	C12C14 C19C20	GRE2	P	6950	response to stress	IMP		

REFERENCES

Angier N. Natural Obsessions: Striving to Unlock the Deepest Secrets of the Cancer Cell. Mariner Books, 1999. ISBN: 0395924723.

Ankeny R. The Conqueror Worm: An Historical and Philosophical Examination of the Use of the Nematode *C. elegans* as a Model Organism. Ph.D Dissertation, University of Pittsburgh, 1997.

Bartlett JC, Toms EG. Developing a protocol for bioinformatics analysis: An integrated information behavior and task analysis approach. J American Society for Information Science and Technology 2005; 56(5): 469-482.

Beyer H, Holtzblatt K. Contextual Design: Defining Customer-Centered Systems. San Francisco: Morgan Kaufmann Publishers, 1998. ISBN: 1558604111.

Boles D. The Effect of Subject Matter Familiarity on Inter-Indexer Consistency, Number of Index Terms Supplied and Indexer Use of Author Terminology. Master's Thesis, University of North Carolina, Chapel Hill, School of Information and Library Science, 1989; #1793, 23 pp.

Bowker J. Model Systems in Developmental Biology. Bioessays 1995; 17(5): 451-455. PMID: 7786291.

Brown CM. The changing face of scientific discourse: Analysis of genomic and proteomic database usage and acceptance. J American Society for Information Science and Technology 2003; 54(10): 926-938.

Burian RM. Technique, task definition, and the transition from genetics to molecular genetics: aspects of the work on protein synthesis in the laboratories of J. Monod and P. Zamecnik. J Hist Biol. 1993 Fall;26(3):387-407. PMID: 11613166.

Camon EB, Barrell DG, Dimmer EC, Lee V, Magrane M, Maslen J, Binns D, Apweiler R. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. BMC Bioinformatics 2005; 6 Suppl 1:S17. PMID: 15960829.

de Chadarevian S. Of worms and programmes: Caenorhabditis Elegans and the study of development. Studies in History and Philosophy of Biological and Biomedical Sciences 1998; 29, 81-105.

Chapman WW, Dowling JN. Inductive creation of an annotation schema for manually indexing clinical conditions from emergency department reports. J Biomedical Informatics 2006 Apr;39(2):196-208. PMID: 16230050.

Chen N, Harris TW, Antoshechkin I, Bastiani C, Bieri T, Blasiar D, Bradnam K, Canaran P, Chan J, Chen CK, Chen WJ, Cunningham F, Davis P, Kenny E, Kishore R, Lawson D, Lee R, Muller HM, Nakamura C, Pai S, Ozersky P, Petcherski A, Rogers A, Sabo A, Schwarz EM, Van Auken K, Wang Q, Durbin R, Spieth J, Sternberg PW, Stein LD. WormBase: A comprehensive data resource for Caenorhabditis biology and genomics. Nucleic Acids Research 2005; 33(1): D383-389. PMID: 15608221.

Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Feierbach B, Fisk DG, Hirschman JE, Hong EL, Issel-Tarver L, Nash R, Sethuraman A, Starr B, Theesfeld CL, Andrada R, Binkley G, Dong Q, Lane C, Schroeder M, Botstein D, Cherry JM. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. Nucleic Acids Research 2004; 32(1): D311-314. PMID: 14681421.

Clause, BN. The Wistar Rat as a right choice: establishing mammalian standards and the ideal of a standardized mammal. J Hist Biol. 1993 Summer; 26(2):329-349. PMID: 11623164.

Cohen AM, Hersh WR. The TREC 2004 genomics track categorization task: classifying full text biomedical documents. J Biomedical Discovery and Collaboration, 2006; 1(1):4. PMID: 16722582.

Colosimo ME, Morgan AA, Yeh AS, Colombe JB, Hirschman L. Data preparation and interannotator agreement: BioCreAtIvE Task 1B. BMC Bioinformatics 2005; 6 Suppl 1:S12. PMID: 15960824.

Cooper, WS. Is interindexer consistency a hobgoblin? American Documentation 1969; 20(3): 268-278.

Couto FM, Silva MJ, Lee V, Dimmer E, Camon E, Apweiler R, Kirsch H, Rebholz-Schuhmann D. GOAnnotator: linking protein GO annotations to evidence text. J Biomedical Discovery and Collaboration 2006, 1:19.

Creager ANH. The Life of a Virus: Tobacco Mosaic Virus as an Experimental Model, 1930-1965. The University of Chicago Press, 2002. ISBN: 0226120252.

Detlefsen EG. The information behaviors of life and health scientists and health care providers: characteristics of the research literature. Bull Med Libr Assoc. 1998 Jul;86(3):385-90. PMID: 9681174.

Devos D, Valencia A. Intrinsic errors in genome annotation. Trends in Genetics 2001; 17(8):429-431. PMID: 11485799.

Dolan ME, Ni L, Camon E, Blake JA. A procedure for assessing GO annotation consistency. Bioinformatics 2005 Jun 1;21 Suppl 1:i136-i143. PMID: 15961450.

EBI. European Bioinformatics Institute QuickGO Gene Ontology Browser. http://www.ebi.ac.uk/ego/ (2007-04-10)

Fujimura JH. Constructing 'do-able' problems in cancer research: Articulating alignment. Social Studies of Science 1987; 17, 257–293.

Funk ME, Reid CA. Indexing consistency in MEDLINE. Bull Med Libr Assoc. 1983 April; 71(2): 176–183.

GO. Gene Ontology Consortium. 'Annotation File Format' (GO Annotation Guide). http://geneontology.org/GO.annotation.shtml#file (2007-04-10)

GO Browser. AmiGO Gene Ontology Browser application. http://amigo.geneontology.org (2007-04-10)

GO Camp. Gene Ontology Annotation Camp. http://www.geneontology.org/meeting/AnnotCamp2006info.shtml (2007-04-10)

Green ML, Karp PD. Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. Nucleic Acids Research 2005; 33(13): 4035-4039. PMID: 16034025.

Hilgartner S. Biomolecular databases: New communication regimes for biology? Science Communication, 1995; 17(2), 240-263.

Hine C. Databases as scientific instruments and their role in the ordering of scientific work. Social Studies of Science 2006; 36: 269-298.

Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics. 2005;6 Suppl 1:S1. PMID: 15960821.

Hripcsak G, Wilcox A. Reference standards, judges, comparison subjects: roles for experts in evaluating system performance. J Am Med Inform Assoc. 2001;9:1–15. PMID: 11751799.

Hurwitz FI. A study of indexer consistency. American Documentation 1969; 20(1), 92-94.

Iivonen M. Consistency in the selection of search concepts and search terms. Information Processing & Management 1995; 31(2), 173-190.

Jackson Laboratories. Mouse Genome Informatics GO Browser. http://www.informatics.jax.org/searches/GO_form.shtml (2007-04-10)

Judson HF. The Eighth Day of Creation: Makers of the Revolution in Biology (expanded ed.). Cold Spring Harbor Laboratory Press, 1996. ISBN: 0879694785.

Karp PD. What we do not know about sequence analysis and sequence databases. Bioinformatics 1998; 14(9): 753-754. PMID: 10366280.

Kay LE. Who Wrote the Book of Life? A History of the Genetic Code. Stanford University Press, 2000. ISBN: 0804734178.

Kay LE. The Molecular Vision of Life: Caltech, the Rockefeller Foundation, and the Rise of the New Biology. Oxford University Press, 1996. ISBN: 0195111435.

Kellog EA, Shaffer HB. 1993. Model organisms in evolutionary studies. Syst Biol 42(4):409-414.

Knorr-Cetina K. Epistemic Cultures: How the Sciences Make Knowledge. Harvard University Press, 1999. ISBN: 0674258932.

Kohler RE. Lords of the Fly: Drosophila Genetics and the Experimental Life. University of Chicago Press, 1994. ISBN: 0226450627.

Latour B. Science in Action: How to Follow Scientists and Engineers Through Society. Harvard University Press, 1987. ISBN: 0674792904.

Latour B, Woolgar S. Laboratory Life: The Construction of Scientific Facts. Princeton University Press, 1986. ISBN: 0691094187.

Leonelli S. Understanding in the life sciences - History, Sociology and Philosophy of the Biology and Bioinformatics developed via research on the Model Organism *Arabidopsis thaliana*. Ph.D dissertation (forthcoming), Faculty of Philosophy, Vrije Universiteit, Amsterdam, Netherlands. http://www.ph.vu.nl/~sabina/PhD.htm (2007-04-10)

Lindee MS. Moments of Truth in Genetic Medicine. The Johns Hopkins University Press, 2005. ISBN: 0801881757.

MacMullen WJ. Annotation as Process, Thing, and Knowledge: Multi-domain studies of structured data annotation. SILS Technical Report TR-2005-02. Chapel Hill: University of North Carolina, School of Information and Library Science, Technical Report Series, 2005a.

MacMullen WJ. Inter-database annotation linkages in model organism databases. In Proceedings of the 68th Annual Meeting of the American Society for Information Science & Technology (ASIS&T), Vol. 42, 2005b.

Marshall CC. Annotation: from paper books to the digital library. In Proceedings of the second ACM international conference on Digital libraries (DL ’97), Philadelphia, Pennsylvania, (July 23-26, 1997), pp. 131-140. ISBN: 0897918681.

Marshall CC. Toward an ecology of hypertext annotation. In Proceedings of ACM Hypertext '98, Pittsburgh, PA (June 20-24, 1998) pp. 40-49.

Murphy LS, Reinsch S, Najm WI, Dickerson VM, Seffinger MA, Adams A, Mishra SI. Searching biomedical databases on complementary medicine: the use of controlled vocabulary among authors, indexers and investigators. BMC Complement Altern Med. 2003 Jul 7;3:3. PMID: 12846931.

NCBI. National Center for Biotechnology Information Bookshelf. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books (2007-04-10)

Olson HA, Wolfram D. Indexing consistency and its implications for information architecture: a pilot study. Information Architecture (IA) Summit, March 23-27, 2006, Vancouver, British Columbia, Canada. http://www.iasummit.org/2006/files/175_Presentation_Desc.pdf (2007-04-10)

Ouzounis CA, Karp PD. The past, present and future of genome-wide re-annotation. Genome Biology 2002;3(2):COMMENT2001. PMID: 11864365

Palmer CL. Structures and Strategies of Interdisciplinary Science. Journal of the American Society for Information Science 1999; 50(3): 242-253.

Price DJ de Solla. Little Science, Big Science …And Beyond. Columbia University Press, 1986. ISBN: 0231049579.

Rader KA. Making Mice: Standardizing Animals for American Biomedical Research, 1900-1955. Princeton University Press, 2004. ISBN: 0691016364.

random.org. Random number generator. Randomized sequence function; parameters: {1,16} and {1,32}. http://www.random.org/sequences/ (2007-04-10)

Rolling L. Indexing consistency, quality and efficiency. Information Processing & Management 1981; 17(2), 69-76.

Rose AF, Schnipper JL, Park ER, Poon EG, Li Q, Middleton B. Using qualitative studies to improve the usability of an EMR. J Biomed Inform. 2005 Feb;38(1):51-60. PMID: 15694885.

Rother K, Michalsky E, Leser U. How well are protein structures annotated in secondary databases? Proteins. 2005 Sep 1;60(4):571-576. PMID: 16021624.

Schultz CK, Schultz WL, Orr RH. Comparative indexing: Terms supplied by biomedical authors and by document titles. American Documentation 1965 16(4), 299-312.

Shatkay H, Feldman R. Mining the biomedical literature in the genomic era: an overview. J Computational Biology 2003; 10(6): 821-855. PMID: 14980013.

Sievert MC, Andrews MJ. Indexing consistency in Information Science Abstracts. Journal of the American Society for Information Science, 1991; 42(1), 1-6.

Srinivasan P. Text mining: Generating hypotheses from MEDLINE. J American Society for Information Science and Technology, 2004; 55(5), 396-413.

Stein L. Genome annotation: from sequence to biology. Nature Reviews Genetics 2001; 2(7):493-503. PMID: 11433356.

Takayama Y, Kamimura Y, Okawa M, Muramatsu S, Sugino A, Araki H. GINS, a novel multiprotein complex required for chromosomal DNA replication in budding yeast. Genes Dev. 2003 May 1;17(9):1153-1165. PMID: 12730134.

Ussery DW, Hallin PF. Genome Update: annotation quality in sequenced microbial genomes. Microbiology. 2004 Jul;150(Pt 7):2015-2017. PMID: 15256543.

Whetzel PL, Parkinson H, Causton HC, Fan L, Fostel J, Fragoso G, Game L, Heiskanen M, Morrison N, Rocca-Serra P, Sansone SA, Taylor C, White J, Stoeckert CJ Jr. (2006). The MGED Ontology: a resource for semantics-based description of microarray experiments. Bioinformatics 2006 Apr 1;22(7):866-873. PMID: 16428806.

White HD, McCain KW. Visualization of Literatures. In Williams ME. (Ed.), Annual Review of Information Science and Technology (ARIST), 1997; 32: 99-169.

Wildemuth BM. The effects of domain knowledge on search tactic formulation. J American Society for Information Science and Technology 2004 55(3), 246-258.

Wong JW, Cartwright HM. Deterministic projection by growing cell structure networks for visualization of high-dimensionality datasets. J Biomedical Informatics 2005 Aug;38(4):322-30. PMID: 16084474.

Yeh I, Karp PD, Noy NF, Altman RB. Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). Bioinformatics 2003; 19(2): 241-248. PMID: 12538245.