Mol Genet Genomics (2010) 283:415–425
DOI 10.1007/s00438-010-0525-8

REVIEW

A MOD(ern) perspective on literature curation

Jodi Hirschman · Tanya Z. Berardini · Harold J. Drabkin · Doug Howe

Received: 4 November 2009 / Accepted: 6 February 2010 / Published online: 11 March 2010

Abstract Curation of biological data is a multi-faceted task whose goal is to create a structured, comprehensive, integrated, and accurate resource of current biological knowledge. These structured data facilitate the work of the scientific community by providing knowledge about genes or genomes and by generating validated connections between the data that yield new information and stimulate new research approaches. For the model organism databases (MODs), an important source of data is research publications. Every published paper containing experimental information about a particular model organism is a candidate for curation. All such papers are examined carefully by curators for relevant information. Here, four curators from different MODs describe the literature curation process and highlight approaches taken by the four MODs to address: (1) the decision process by which papers are selected, and (2) the identification and prioritization of the data contained in the paper. We will highlight some of the challenges that MOD biocurators face, and point to ways in which researchers and publishers can support the work of biocurators and the value of such support.

Keywords Annotation · Biocuration · Database · Genome · Literature · Model organism

Communicated by T. Ito.

J. Hirschman, T. Z. Berardini, H. J. Drabkin, D. Howe belong to the Gene Ontology Consortium.

J. Hirschman (✉)
Saccharomyces Genome Database,
Department of Genetics,
Stanford University, Stanford, CA 94305, USA
e-mail: jodi@genome.stanford.edu

T. Z. Berardini
The Arabidopsis Information Resource,
Department of Plant Biology, The Carnegie Institute for Science,
260 Panama St., Stanford, CA 94305, USA

H. J. Drabkin
Mouse Genome Informatics, The Jackson Laboratory,
600 Main St., Bar Harbor, ME 04609, USA

D. Howe
The Zebrafish Information Network, 5291 University of Oregon,
Eugene, OR 97403-5291, USA

Introduction

Biocuration, the practice of collecting and organizing biological data, has roots that go back several thousand years. Formalization began as early as the fourth century BCE when Aristotle grouped organisms by various criteria such as form, mode of reproduction, blood, etc. (Ausdesirk et al. 2004). Linnaeus, Darwin, and others built on this early work through the eighteenth and nineteenth centuries. These types of collections form the core of natural history museums. In the twentieth century, large amounts of sequence data became available. Protein sequences were the first to be collected, followed by RNA sequences, then DNA sequences. With the advent of computer technology came the development of new ways to access and display information. Soon biological databases were created to contain, organize, and make accessible the growing assembly of biological data. Today there are numerous biological databases containing information about a large variety of organisms and covering multiple data types.

Model organism databases (MODs) collect and display information about genes and proteins of individual species. They grew out of the development of genomic sequencing technologies and subsequent release of sequence data for

organisms like *A. thaliana* (Arabidopsis Genome Initiative 2000), *S. cerevisiae* (Goffeau et al. 1997), and *C. elegans* (C. elegans Sequencing Consortium 1998). Associated with the initiation of the human genome sequencing project (Barnhart 1989), MODs are seen as an important tool for guiding investigation of the human genome (Clark 1999; Carroll et al. 2003). Today, MODs have publicly accessible web-based interfaces and specialize in representation of genetic and genomic data generally pertaining to one species or a class of closely related organisms. The authors of this review are each curators for the MODs TAIR (The Arabidopsis Information Resource, Swarbreck et al. 2008), ZFIN (The Zebrafish Information Network, Sprague et al. 2003), MGI (Mouse Genome Informatics, Blake et al. 2006), and SGD (Saccharomyces Genome Database, Dwight et al. 2004), which contain information for the eukaryotic model organisms *Arabidopsis thaliana*, *Danio rerio*, *Mus musculus*, and *Saccharomyces cerevisiae*, respectively. These organisms are representative of the plant, animal, and fungal kingdoms, and their genomes differ greatly from each other. Nevertheless, our work as curators includes many common tasks. In particular, a major component of the biocuration effort at these databases is that of reading and extracting information from the published literature. The goal is to present the disparate data from different experiments in an organized and accessible framework to give biologists a broader perspective than they might get from any one paper. Over the past 10–15 years both the quantity and types of information curated have greatly expanded due to the increased volume of publications, the changing needs of research communities, and advances in research approaches and technology. Here, we focus specifically on how literature curation is done at MODs, with an emphasis on common processes. We will highlight some of the challenges that MOD biocurators face, and point to some ways in which researchers and publishers can support the work of biocurators.

Making molehills out of mountains

Gathering the literature

Curation begins by gathering the relevant body of literature. Typically, papers of interest to a MOD are identified through periodic searches of literature indexed at PubMed. In order to extract information, curators must have access to the full text of the article. Many of the MODs are based at academic institutions that maintain paid subscriptions to many scientific journals. Also, NIH's PubMed Central provides a free digital archive of many journal articles. However, subscriptions are costly and university budgets are strained and not all articles, NIH-funded or not, become available in PubMed Central in a timely fashion. It is often still difficult for curators to obtain the full texts, let alone any supplemental data, of research articles from the journal websites in a consistent and computationally aided manner. Authors who wish to have their data curated into the electronic data stream need to evaluate accessibility of their publication as they choose where to publish.

Identification and prioritization of papers to curate

The number of papers reviewed for curation varies from MOD to MOD and may be anywhere from 100 to 1,000 papers a month. From 2004 to 2009, the average numbers of papers added yearly to TAIR, ZFIN, SGD, and MGI were 2200, 1,000, 3,000, and 11,000, respectively. These numbers are reflective of the size of the research community for each organism. Often, the identification of papers containing data relevant to a particular database involves a paper-by-paper review by curators. Some MODs manually associate each publication with the biological objects of interest, such as genes, while others make associations via electronic methods that match papers with genes and validate these associations manually. The continuous influx of new papers necessitates that each MOD develop a mechanism for prioritizing literature curation. While it is desirable to extract and record every piece of information from the entire literature corpus for each species, the relatively small size of the curation staff at any one MOD dictates that some papers are immediately completely curated, while others take a little longer or are not used at all. At some MODs, highest priority for curation is given to papers describing previously uncharacterized genes or containing functional data that can be used for gene ontology (GO) annotations. At other MODs, priority may be given to papers reporting new mutants and phenotypes. The priority is driven by the needs of the user base for the types of information in the paper, and changes with the changing needs of users.

There are also some publications that will not be curated in the context of the MODs. MODs are gene-centric resources, so certain types of publications are usually not curated by a MOD. For example, toxicology studies describing the lethal effects of chemicals or pharmaceuticals in organisms are generally not curated unless they link the toxicity, resistance to toxicity, or pharmaceutical effects to a gene product in some way. Regardless of the driving force behind the setting of priorities, the end result is identification of a subset of papers that will be carefully read by the curation staff.

Separating the wheat from the chaff

The types of data that are presented in the biological literature are extremely diverse, and choosing which ones to

capture in any given MOD is a challenge. Primary data collected include basic gene-related information such as gene names. There are many more complex data such as specific genotype details, spatial and temporal gene expression patterns, phenotypes, biochemical pathways, and genetic or physical interaction networks. Some MODs provide information about biological reagents, such as mutant and reference strains, DNA materials like transgenic constructs, morpholino oligonucleotides, RNAi constructs and cDNA libraries, and protein reagents like antibodies (Table 1). Many databases also provide structured annotations of gene products using controlled vocabularies, such as those provided by the GO (The Gene Ontology Consortium 2010) that facilitate computational analyses of groups of genes within or between organisms. Which of the more complex data types are tackled depends on the needs of the research community, the priorities set by each MOD's Scientific Advisory Board, and the human and computational resources available. Table 2 presents for each MOD the number of records representing functional gene products, those with literature associated, and those with experimentally supported functional annotation (GO) from literature.

Gene identification and nomenclature

The first step in curating a paper is the identification of the genes and proteins described as well as the organism to which they belong (Fig. 1). Surprisingly, one of the more difficult challenges a biocurator can face is unambiguously linking the gene or genes discussed in a paper to a record in their MOD, particularly when the gene nomenclature is unclear or the research organism is not identified. Inclusion of sequence or database identifiers for each gene in the paper ensures accurate linking between the results presented in that paper and the particular gene or genes described therein, no matter how the name of a gene may change over time. Papers that lack this information may be impossible to curate. Nomenclature standards and collaborative efforts often strive to give orthologous genes the same symbol and name in multiple species. For example, MGI, under the auspices of the International Committee on Standardized Genetic Nomenclature for Mice (Maltais et al. 2002), and the HUGO Gene Nomenclature Committee (HGNC, Bruford et al. 2008), works to co-ordinate nomenclature between mouse, human and rat genes. ZFIN attempts to name zebrafish genes after their human or mouse orthologs. Likewise, ArkDB (Hu et al. 2001) which maintains data on cat, chicken, cow, horse and sheep, adopts the HGNC nomenclature. Consequently, when a paper mentions a gene or protein symbol that is the same in several organisms (such as BRCA1, which is used in more than a dozen species) and it does not state the specific organism of origin or provide a sequence accession number for this particular gene, it is difficult, if not impossible, to determine whether the data in that paper belongs in a particular species-specific database or not. Another nomenclature conundrum for many model organisms is the case where the same symbol is used for more than one gene in a single species. For example, the symbol *PAP1* in *Arabidopsis thaliana* is the primary symbol or alias for four different genes (PURPLE ACID PHOSPHATASE 1, PHOSPHATIDIC ACID PHOSPHATASE 1, PRODUCTION OF ANTHOCYANIN PIGMENT 1, and PHYTOCHROME-ASSOCIATED PROTEIN 1). There are currently 216 similar examples for zebrafish, where a gene has a primary symbol that is the same as an alias for at least one other zebrafish gene. As a result, use of these symbols in the literature is ambiguous, and more information is required to resolve which specific gene is actually described in a particular paper. Additionally, many species have gene duplicates which share a root symbol but are appended with an 'a' or 'b' suffix. For example, when a publication discusses the zebrafish gene 'wnt8', it is unclear which specific gene is meant as zebrafish have both a wnt8a and a wnt8b gene. There is no way to resolve such a case without a sequence accession number, or communication directly with the authors.

Curators of each MOD have tried to formalize the process of naming genes according to the wishes of their respective research communities (Table 3). Not all researchers are aware that such processes exist, and that

Table 1 Data types curated by MGI, SGD, TAIR, and ZFIN

Genetics/genomics	MGI	SGD	TAIR	ZFIN
Genes	X	X	X	X
Alleles	X	X[a]	X	X
Genotypes	X		X	X
Phenotypes	X	X	X	X
Gene expression	X (embryonic)	X	X	X
Sequences	X	X	X	X
Orthology	X	X[a]	X	X
Gene ontology	X	X	X	X
Protein interactions	X (using GO)	X[b]	X[b]	
Gene/allele nomenclature	X	X	X	X
Reagents				
Antibodies	X			X
Transgenic constructs	X			X
Morpholinos				X
Probes	X			X

[a] Limited allele and orthology information are curated as part of phenotype and general literature curation

[b] Interaction data is provided in collaboration with the BioGRID database (http://www.thebiogrid.org/, Breitkreutz et al. 2008)

Table 2 Gene records and literature-based data curated for three data types

	Gene records[a]	Gene records associated with at least one PubMed ID	Gene records with at least one experimental GO annotation	Gene records annotated with expression data[b]	Gene records annotated with at least one mutant allele
MGI	36,123	27,724	8,026	11,115	12,137
SGD	6,412	6,373	5,203	NA	5,555
TAIR	34,770	22,362	6,159	17,940[c]	3,331
ZFIN	30,648	17,991	1,965	10,688	3,374

[a] Includes only those markers that are predicted or have been shown to encode an RNA or protein product

[b] Tissue-specific expression

[c] Based on annotations to plant ontology terms (http://www.plantontology.org)

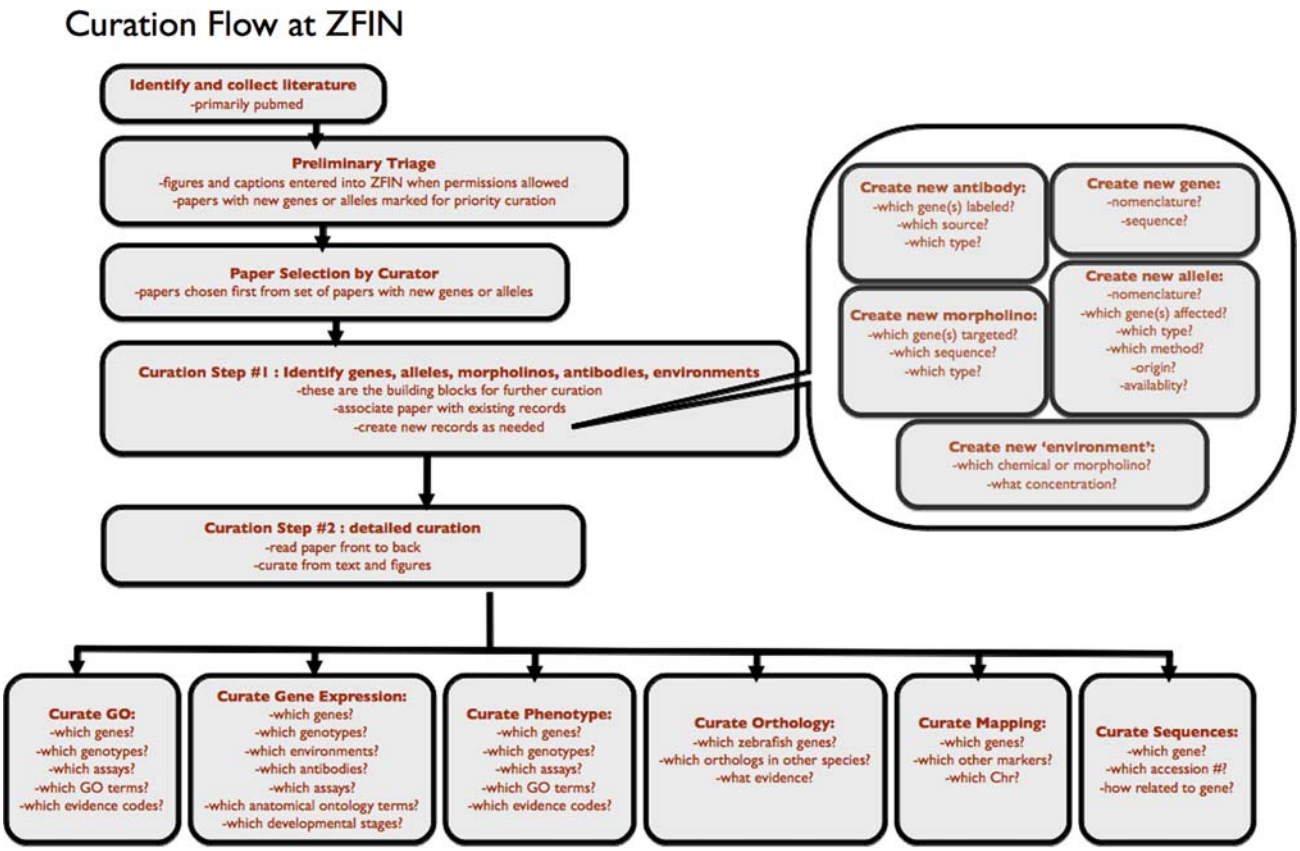

Fig. 1 A typical curation workflow, exemplified by the process at ZFIN. Curation workflows are unique as each MOD strives to best serve its own research community. For example at some MODS, different members of the curation team may enter different types of data, whereas at other MODS a single curator may enter all of the data types from a paper. Additional differences in workflow stem mainly from staffing and other budgetary constraints for each database. However, there are many commonalities in the workflow process, as the questions that must be answered to complete curation of a paper are similar regardless of the MOD. Here, the curation workflow at ZFIN illustrates the order in which certain tasks take place and many of the questions that must be answered at each step. Papers that lack key details can prevent curators from answering questions critical to the curation process, leading to a reduction in the amount or the detail of the curated data

they differ among organisms. In general, validation of gene names is not required as part of the publication process. Consequently, gene names or mutant alleles used in publications sometimes conflict when a gene or mutant already exists with an approved name in the database, or the author-given name is already in use for another gene or mutant. In all such cases, a biocurator must tease out the pertinent information for the appropriate gene by searching through earlier literature cited in the paper or by direct communication with authors. This slows the curation process and unnecessarily increases the risk of data association errors during curation. Journals, reviewers, and authors can

Table 3 Online resources

Home pages	
MGI	http://www.informatics.jax.org
SGD	http://www.yeastgenome.org/
TAIR	http://www.arabidopsis.org/
ZFIN	http://zfin.org/
Nomenclature	
MGI	http://www.informatics.jax.org/mgihome/nomen
SGD	http://www.yeastgenome.org/gene_guidelines.shtml
TAIR	http://www.arabidopsis.org/portals/nomenclature/index.jsp
ZFIN	http://zfin.org/zf_info/nomen.html
Downloads	
MGI	ftp://ftp.informatics.jax.org/pub/reports/index.html
SGD	http://downloads.yeastgenome.org/
TAIR	ftp://ftp.arabidopsis.org/home/tair/
ZFIN	http://zfin.org/zf_info/downloads.html

ensure that published results are associated with the correct genes by using officially approved nomenclature, which can be found at a species-specific authority like a MOD or the HGNC, and by providing identifiers for the genes or proteins discussed in every paper. Researchers unsure of which particular accession identifiers to use are encouraged to contact the appropriate database for guidance.

The information extraction process

Though the exact curation workflow may vary, the basic process of curation is similar at each MOD (Fig. 1). After establishing the identity of the organism, genes and mutants of interest, the biocurator reads the full paper, identifies the data of interest and enters that into a database. As most papers contain multiple types of data (for example, phenotype, orthology, localization, and interaction data), a kind of mental 'multi-tasking' ensues as the biocurator develops an understanding of the experimental concept and the results and then considers whether and how the different data will fit into the database. While this detailed thought process is often straightforward, there are times when deciding which information to extract from the paper and where to put it in the database can entail multiple readings of the paper and involve discussions with other curators or the authors of the paper. The goal of the biocurator is to extract the experimental results published in the paper, to add them to the appropriate sections of the database, and to connect those single pieces of information with the existing data in order to create a larger and more intricate picture of the role of each gene of the organism. There often are data in a paper about genes that are not the main focus but were included in studies. For example, a paper may describe new results about the function of the receptor encoded by the *S. cerevisiae* gene *STE2,* but may also contain ancillary data about one of the subunits of the associated G protein. From the point of view of the curator, all of these data are important, and are therefore added to the appropriate sections of the database.

Distinguishing experimentally supported from inferential assertions

Reading a publication for curation is very different from reading it as a reviewer or bench researcher. It is the curator's job to understand the data and determine how to record them, not to censor them or subject them to another layer of review, since the data have already gone through peer review and thus are accepted as accurate. It is also important to identify the experimental assays behind the data. Because curators are looking for experimentally supported information, they must be able to distinguish experimentally backed assertions from inferential or speculative ones. When authors report the existence of a transmembrane domain in a newly cloned sequence based on computational analysis, the curator must realize that this is not an experiment but merely a prediction and as such cannot be used to assert experimentally verified membrane localization for the protein. In fact, such inferential annotations are most often captured through automated electronic curation pipelines that use the same methods the authors may have used. Curators need to have a strong background in experimental biology and they must possess dedication to scientific detail. Many biocurators have graduate degrees and post-doctoral experience. The eye for detail is especially useful when confirming specifics such as gene names and morpholino sequences and their corresponding target genes. It is not unusual for critical typographical errors to creep into previously reviewed publications. In many cases, it is a MOD biocurator that picks up these mistakes, contacts the corresponding authors and resolves the discrepancy before adding the correct data to their database.

Adapting to new data types

Effective literature curation requires that curators know what kinds of information researchers need to access and how researchers want to access that data. In fact, information needs have changed over time, and efforts to respond to these needs have expanded the scope of curation and the responsibilities of the curator. For example, as full genome assemblies become available, curation of genetic mapping data becomes less important. On the other hand, the technological advances in analyzing thousands of genes or proteins via high-throughput experiments (microarray, GC/MS, etc.) requires new paradigms for data curation and

display. Curators become aware of the latest trends in research by constantly reading new literature, attending scientific conferences and receiving feedback from their scientific advisory committees and user base. When microarray experiments were first published, there was no strategy in place for curating the massive amounts of individual data points and no place to house and display the data. In time, data warehouses and databases like ArrayExpress (Parkinson et al. 2009) and GEO (Barrett et al. 2009) were created to address the long-term storage needs for this type of data. Proteomics advances, be they the isolation of all the proteins in a particular cellular compartment or the identification of all proteins whose expression is triggered by some external stimulus, have also resulted in a large mass of data. This data must be also stored, cross-referenced to the relevant genes and made available to the public long after the original article was published. Individual MODs have made the decision either to mirror such data in their pages or link to these larger multi-organism resources.

Data catch and release

In order to transfer data from the published literature into a database, MODs and other databases have had to (1) develop strategies and tools that enable the curation staff to curate these data; (2) create and modify database schemas to accommodate new data types; and (3) develop the web pages that are accessed by the research community so that they can search for and view the new data. The tools for data curation and the databases that store the interrelated data are designed with input from biocurators, software engineers and database administrators. Careful consideration is given to the design of the database, the various query forms that will be presented to the public, and the web pages that display data to the end user. Does the database format allow the significance of the results to be accurately presented to the user? Can biologically interesting correlations and connections in the data easily be found? Do the controlled vocabularies adequately describe the data, or are new terms needed? How can the volume of data available from large-scale studies be incorporated into the database so that users can access and understand the results? Attention to the experimental details provided in the publication and to data integration allows the users to make complex queries, such as "show me the genes on chromosome 1 that have been experimentally shown to function as a protein kinase" (Fig. 2). The ability to make such detailed queries requires highly detailed curation. Below we discuss how this is done so authors can begin to understand how their data are translated into a MOD during the curation process.

Details of data input

MOD curators add data into their database using software that is designed to make data entry as error-free and accurate as possible. For example, when adding GO annotation to a gene product, the identifier for a GO term must have seven digits and correspond to a currently valid GO term, or input is blocked. The web interfaces incorporate various controlled vocabularies to supply terminology that describes the data in a consistent manner (Fig. 3). The vocabularies can be simple lists, such as those that describe assay types, developmental stages, or chromosome number, or they may be more complex, such as the structured vocabularies of the Cell Ontology (Bard et al. 2005), GO (Ashburner et al. 2000), Mouse Anatomy Ontology (Bard et al. 1998), or plant ontology (Jaiswal et al. 2005). Curators usually develop these vocabularies, and in some cases (such as the GO) they are continuously updated by curators from multiple MODs, with input from experts in relevant fields, as new research findings dictate the need for new terms. These same controlled vocabularies are used in the query forms that users see at the respective MOD web sites, and allow users to more easily analyze large amounts of data for similarities and differences across species. While controlled vocabularies are valuable in helping users group data, they are sometimes not specific enough to describe the subtleties of some results. Therefore, curation of these data often includes an option for curators to add free text. While free text allows for curator freedom in adding experimental details, it is not as amenable to efficient searching as is the use of a controlled vocabulary. A good compromise is the use of both types of data capture. For example, SGD's and MGI's curation of *S. cerevisiae* and *M. musculus* phenotypes includes controlled-vocabulary components as well as free-text fields, where curators often add details about the phenotype or experimental conditions that are critical for interpreting the data (Hancock et al. 2007; Costanzo et al. 2009). In such cases, query interfaces are designed to allow for searching free-text data as well as that using controlled vocabularies.

Importing data from external resources

In addition to data obtained from literature curation, many MODs incorporate a variety of data from outside sources, such as UniProtKB (The UniProt Consortium 2010), Genbank (Benson et al. 2009), and Ensembl (Hubbard et al. 2009). In addition to enhancing the amount of information in the database, imported data provides a quality control function. For example, automated scripts reconcile gene identifiers from imported data with the genes in each database and categorize possible conflicts, such as the

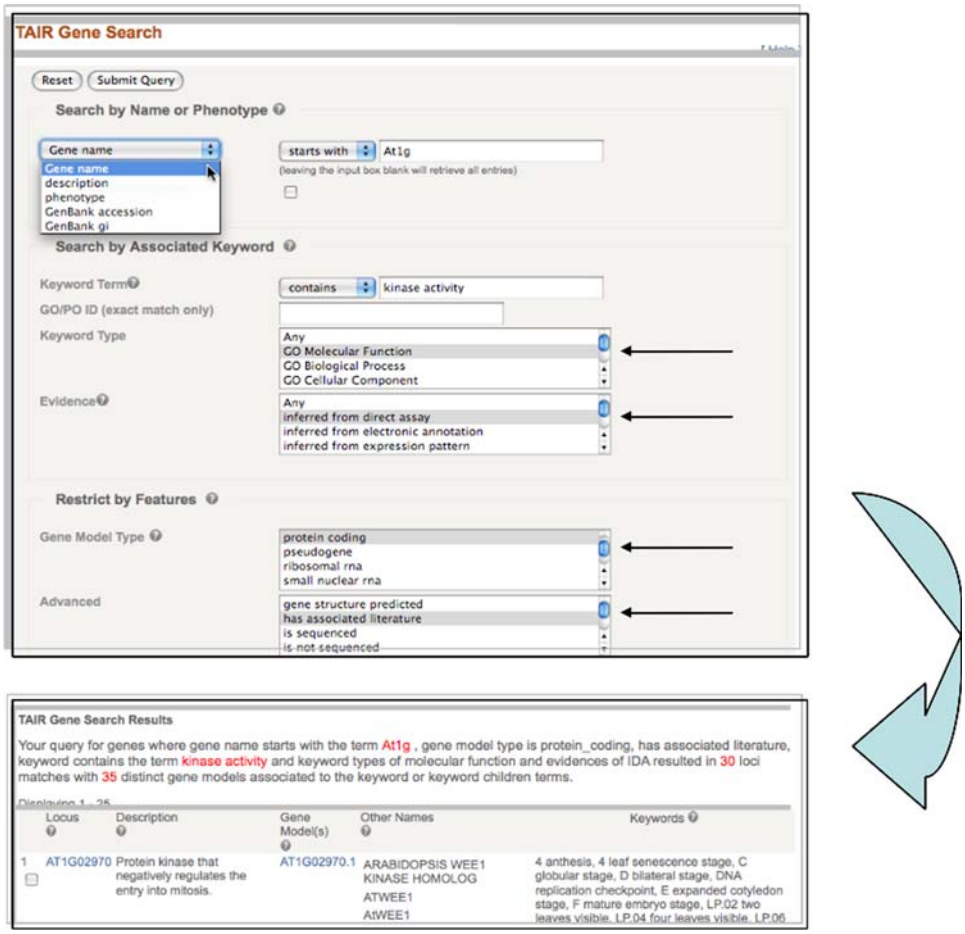

Fig. 2 A TAIR web query form using controlled vocabularies to ask "Find a gene whose symbol begins with At1g and has GO function annotations based on direct assays, and codes for a protein that has literature associated with it."

mapping of one UniProt ID to multiple genes in a MOD, for biocurator review. Manually curated GO annotations can be compared to computer-generated GO annotations provided by UniProtKB to identify missing or contradictory annotations. In cases where information is shared among the databases, or is used for comparison of data across databases, additional effort is required. For example, the Gene Ontology Reference Genomes Project (The Reference Genome Group of the Gene Ontology Consortium 2009) which involves coordinated GO annotation at a dozen MODs, has regular annotation meetings to enhance consistency in the usage of the GO among its members.

Handling data conflicts

Once data are entered into the database, they are available to the public either immediately or following a scheduled database update. Scheduled database updates vary in frequency and may be daily, weekly, monthly or whenever a specified set of curation goals has been accomplished. Since each piece of data is always associated with a reference, the user may examine cases where data appear to be conflicting and decide on the context of the conflict. Users concerned about conflicting (or any other) data are encouraged to contact the originating database. Community input helps the curatorial staff decide how to handle such situations. In general, data that have clearly been shown to be wrong may be removed from the database. For example, in 2006 the *Arabidopsis* gene *FCA* was reported to encode an abscisic acid receptor (Razem et al. 2006). In 2008, the authors of the 2006 paper retracted their claims (Razem et al. 2008) and another group published that FCA does not bind abscisic acid (Risk et al. 2008). Any annotations associated with the first paper were removed from TAIR. On the other hand, data that may appear conflicting as presented in the database are retained with their citation, as long as the paper has not been retracted or the authors have not requested removal. For example, real-time live cell analyses using fluorescently labeled proteins and subcellular markers showed that the *Arabidopsis* protein AUX1 resides at the apical plasma membrane of protophloem cells and at highly dynamic subpopulations of Golgi apparatus and endosomes in all cell types (Kleine-Vehn et al. 2006). In cases like this, both pieces of localization data are valid and are captured in the database.

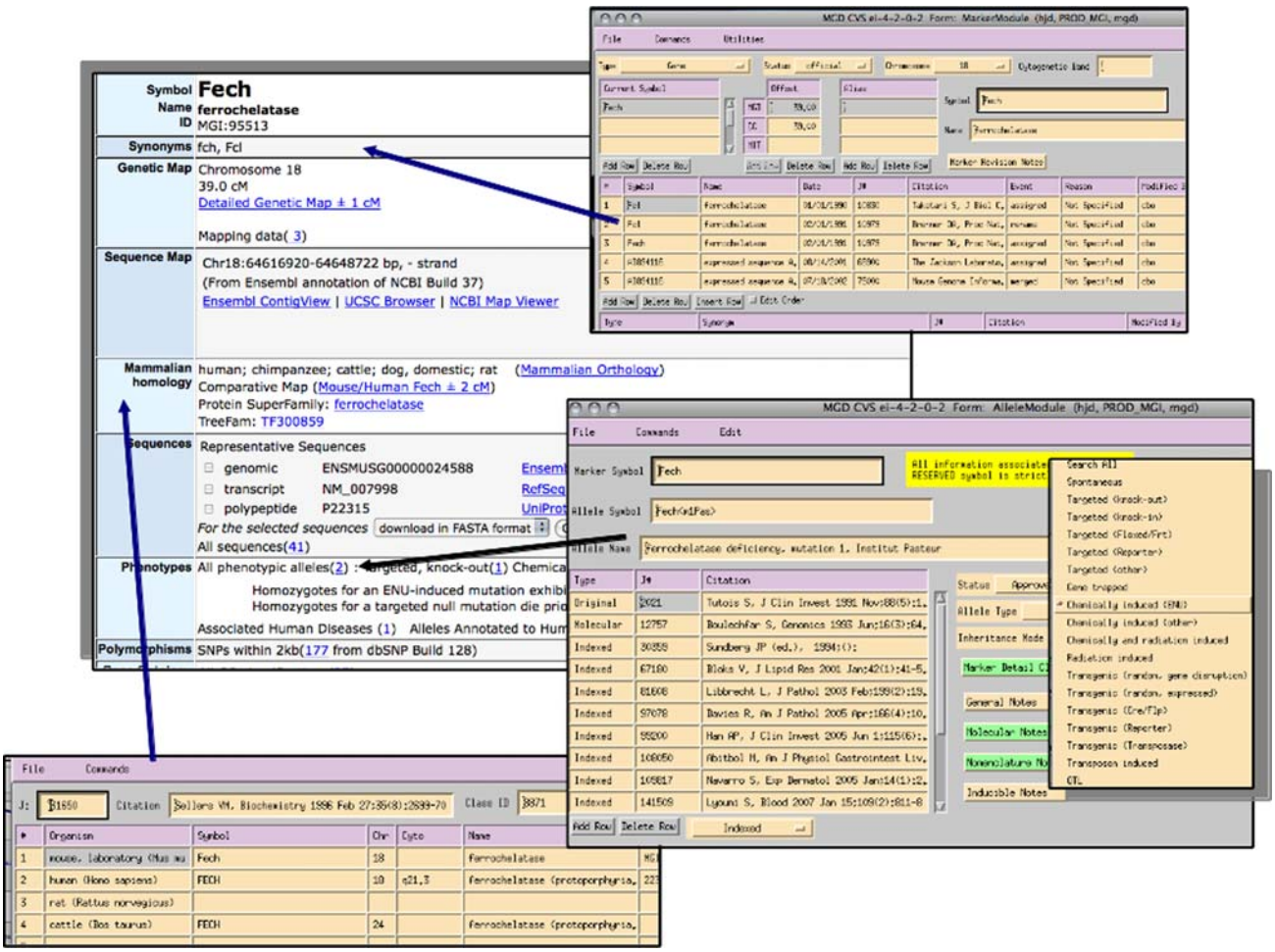

Fig. 3 Snapshot of the Gene Detail page for Fech (*upper left*), and snapshots of MGI editorial interfaces for input of data relating to symbol, name, and synonyms (*upper right*), phenotypic alleles (*bottom right*), including expanded window showing a controlled pick list, and mammalian homology (*bottom left*). *Arrows* point to the relevant section of the gene detail page that the editorial interface addresses

Data 'holes' resulting from incomplete curation of older literature

MODs evolve to incorporate new features in order to support changing research interests and to curate data types that were not tackled in the past. Over time, previously curated literature may become 'incompletely curated' with respect to the current capabilities of a given MOD. For example, a previously curated paper that clearly shows the biological role of a specific gene may not have been used for GO if GO curation was not done at the time that paper was originally curated. Limited curatorial staffing generally makes it impractical to bring curation of all older papers up to date with current curation standards. As a result, 'holes' in the curated data become apparent and can make the curated data seem spotty relative to what a researcher specializing in a field may know from years of reading the literature. Community feedback is essential in pointing out such gaps in the curated data, and helps curators prioritize the curation of older papers to fill such data holes.

Making data available to users

The biological data curated by the MODs are available in several formats not only at their own websites but also at multiple locations on the Internet. The primary location at which a researcher will view information about a single gene is on the MOD's gene detail page (Fig. 3). Using the power of relational databases, data that describes chromosome mapping of a gene can be linked to data discussing the effect of its mutant alleles, or developmental expression of the protein or RNA gene product and displayed on a single page. In addition to the information seen on the individual gene detail pages, MODs also provide data for download purposes in several formats through public ftp sites (Table 3). These sites may house files containing the data that is on the

web pages, as well as more specific data sets and reports. MODs also contribute data to more general resources such as the GO (The Gene Ontology Consortium 2010), UniProtKB (The UniProt Consortium 2010), Ensembl (Hubbard et al. 2009), RefSeq (Pruitt et al. 2007), Entrez Gene (Maglott et al. 2007), and others. In turn the MODS obtain sequence and domain data from the same resources. Thus, the taxon-specific resources compliment those resources that contain information from all taxa.

Displaying the curated data through multiple venues provides greater exposure of the data to users and helps to prevent overlap of curatorial efforts. Most importantly, this data sharing is at the core of our endeavor to connect biological information across organisms, allowing their similarities and their differences to be readily detectable and exploited in new ways by the research community. These inter-connections are made possible when authors provide the necessary details in their publications (Table 4). As authors take these needs into account, the information they have worked so hard to elucidate will be more easily incorporated into the global biological data flow where it can have the biggest impact (Fig. 4).

Biocuration in the future

The primary goal of biocuration is to stitch together a comprehensive, accurate, and up-to-date picture of current biological knowledge. This network of information provides researchers with easy access to detailed and highly cross-linked information that is traceable back to its source. To accomplish this, it is critical that the published data be readily available for curation. Budgetary limitations should not cause limitations in the inclusion of any publication in the curation process. This diminishes the accessibility of the researcher's data.

Recent initiatives to increase access, and institutional support for them, have begun to help change this by providing publication venues where everyone can access publications and their enclosed data freely and in a timely manner. Freely available articles benefit journal publishers by opening new avenues for those journal articles to be found in complex searches conducted against curated data at biological databases. Continued expansion of publication models that provide rapid access to the published literature for the widest possible audience, institutional support for publication in journals with high accessibility, and improved data access collaborations between biological databases and journal publishers will be important if biocuration is to successfully pursue its goal of complete and up-to-date curation of biological knowledge.

Table 4 Recommended guidelines for authors to aid in literature curation

Publish in journals that make full text freely and easily available
Use proper nomenclature for genes and proteins
Supply all relevant sequence identifiers
Clearly indicate the species of each gene and sequence used
Indicate developmental stages where appropriate

As the number of papers increases, several MODs have begun projects to assess the feasibility of utilizing various text-mining tools to aid and streamline the literature curation processes. Aside from technical problems such as obtaining full text in a format suitable for scanning, including figure legends and supplemental data, such tools must be able to accurately recognize key phrases in their proper context, and associate these with various controlled vocabularies (gene lists, phenotype and GO ontologies, etc.). For example, Van Auken et al. have reported a promising test using Textpresso (Muller et al. 2004) to curate GO cellular component annotations at Wormbase (Van Auken et al. 2009). However, it is not clear how such a tool might deal with a much larger literature corpus such as MGI's (at least 1,000 papers a month), or how well this tool might handle curation of complex phenotypes due to conditional knockouts and relate these phenotypes to GO biological process terms, etc. More recently, MGI has evaluated several tools to aid in the bottleneck of associating selected papers to specific genes and reports a possible increase in assignment throughput of 20–40% (Dowell et al. 2009). It is unlikely that such tools will replace the need for expert manual extraction of experimental data within a relevant paper, but they may aid in selecting papers for further scrutiny.

It is hoped that alternative curation models, likely involving more direct participation by the research community, will help address the data gap between the ever-increasing amount of information published in the literature and the amount of data available through curated biological databases. Several community curation models are currently under investigation, including partnerships between journals and databases (Seringhaus and Gerstein 2007; Ceol et al. 2008; Ort and Grennan 2008; Seringhaus and Gerstein 2008), direct editing of database entries by field experts (Menda et al. 2008), and the use of wikis to supplement existing curated databases (see, e.g., http://wiki.yeastgenome.org/index.php/Main_Page). Wikis provide a forum for researchers to discuss issues that may never be published, such as controversies about certain experimental results. Participation of the research community can greatly augment the curator's ability to develop a unified, comprehensive, precise, accurate, and highly cross-referenced view of the current biological knowledge.

Just as biocuration matures, so does our understanding of the complex nature of biology. The concept of a gene,

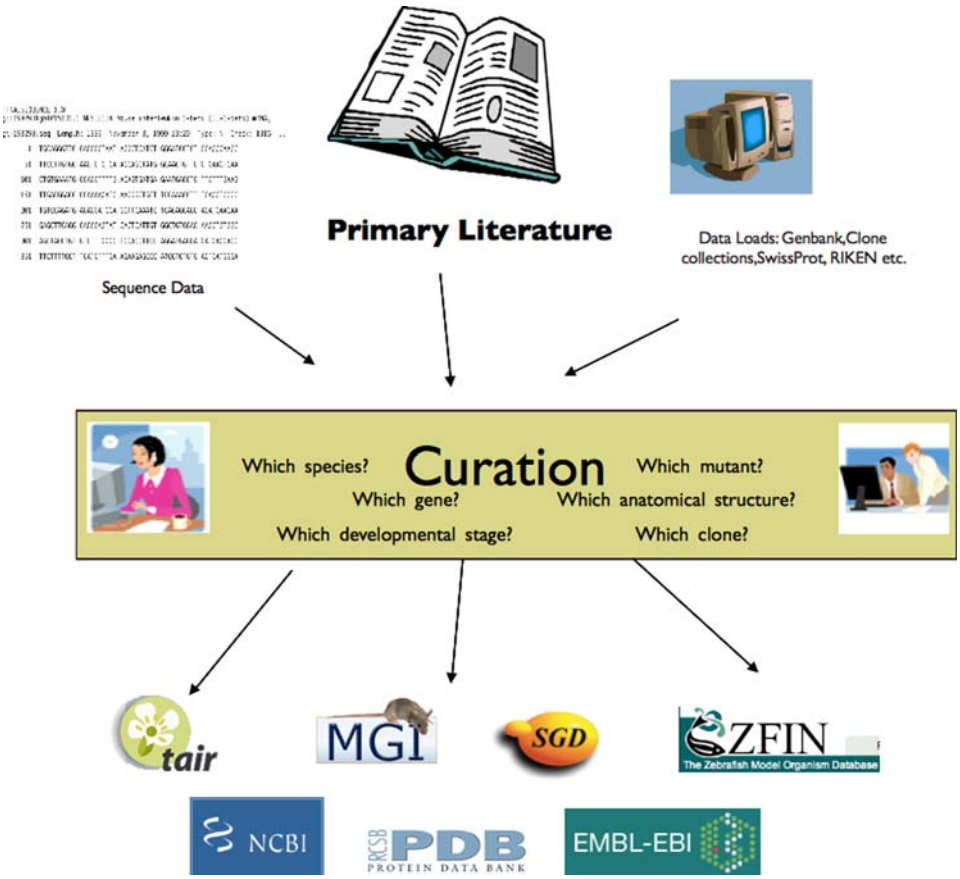

Fig. 4 The global flow of biological data, as presented from a MOD perspective. Curators read the published literature and data that can be extracted for the database is identified and entered. Other sources of data may also be incorporated and in some cases can be used to identify inconsistencies with the literature-derived data. The curation process serves to organize and integrate data into the relational database format for users to easily view what is known and not known about their favorite genes or proteins

one of the most basic tenets of biology, is changing as our understanding of genetics expands. Biocurators will continue to find ways to integrate complex new biological concepts into their existing frameworks. Great challenges lie ahead for bioresearch and biocuration alike. It is becoming apparent that the current gene-based data models used by MODs need to be expanded to allow incorporation, access, and visualization of a growing number of complex entities and processes such as microRNAs, epigenetic influences, gene regulation networks/pathways, and protein complexes. With continued staffing of highly qualified biocurators, sufficient funding, and active collaboration between journals, biological databases and the research community, we can successfully meet these challenges.

Acknowledgments T.Z.B. is supported by grant #P41 HG002273 from the National Human Genome Research Institute (NHGRI) at the United States National Institutes of Health (GO Consortium) and by grant DBI-0417062 from the United States National Science Foundation (TAIR). D.G.H. is supported by grant #P41 HG002659 from the NHGRI at the United States National Institutes of Health. J.H. is supported by grant #P41 HG001315 from the NHGRI at the US National Institutes of Health. H.J.D. is supported by grant #P41 HG000330 from NHGRI and grant # GM080646 from the National Institute of General Medical Sciences. The authors would like to thank Maria Costanzo, David Hill, Eurie Hong, Caroline Koehrer, Mark Berardini, Judith Blake, and J. Michael Cherry for preliminary opinions and commentary on the manuscript.

References

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29

Ausdesirk T, Audesirk G, Byers B (2004) Life on earth, 3rd edn. Pearson Prentice Hall, Saddlebrook

Bard JL, Kaufman MH, Dubreuil C, Brune RM, Burger A, Baldock RA, Davidson DR (1998) An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. Mech Dev 74:111–120

Bard J, Rhee SY, Ashburner M (2005) An ontology for cell types. Genome Biol 6:R21

Barnhart BJ (1989) The Department of Energy (DOE) human genome initiative. Genomics 5:657–660

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R (2009) NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res 37:D885–D890

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2009) GenBank. Nucleic Acids Res 37:D26–D31

Blake JA, Eppig JT, Bult CJ, Kadin JA, Richardson JE, Group MGD (2006) The Mouse Genome Database (MGD): updates and enhancements. Nucleic Acids Res 34:D562–D567

Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bahler J, Wood V, Dolinski K, Tyers M (2008) The BioGRID interaction database: 2008 update. Nucleic Acids Res 36:D637–D640

Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S, Birney E (2008) The HGNC database in 2008: a resource for the human genome. Nucleic Acids Res 36:D445–D448

C. elegans Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282:2012–2018

Carroll PM, Dougherty B, Ross-Macdonald P, Browman K, FitzGerald K (2003) Model systems in drug discovery: chemical genetics meets genomics. Pharmacol Ther 99:183–220

Ceol A, Chatr-Aryamontri A, Licata L, Cesareni G (2008) Linking entries in protein interaction database to structured text: the FEBS Letters experiment. FEBS Lett 582:1171–1177

Clark MS (1999) Comparative genomics: the key to understanding the Human Genome Project. Bioessays 21:121–130

Costanzo MC, Skrzypek MS, Nash R, Wong E, Binkley G, Engel SR, Hitz B, Hong EL, Cherry JM, Saccharomyces Genome Database Project (2009) New mutant phenotype data curation system in the Saccharomyces Genome Database. Database 2009:bap001

Dowell KG, McAndrews-Hill MS, Hill DP, Drabkin HJ, Blake JA (2009) Integrating text mining into the MGI biocuration workflow. Database 2009:bap019

Dwight SS, Balakrishnan R, Christie KR, Costanzo MC, Dolinski K, Engel SR, Feierbach B, Fisk DG, Hirschman J, Hong EL, Issel-Tarver L, Nash RS, Sethuraman A, Starr B, Theesfeld CL, Andrada R, Binkley G, Dong Q, Lane C, Schroeder M, Weng S, Botstein D, Cherry JM (2004) Saccharomyces genome database: underlying principles and organisation. Brief Bioinform 5:9–22

Goffeau A, Aert A, Agostine-Carbone M, Ahmed A, Aigle M et al (1997) The yeast genome directory. Nature 387:5

Hancock JM, Adams NC, Aidinis V, Blake A, Bogue M, Brown SD, Chesler EJ, Davidson D, Duran C, Eppig JT et al (2007) Mouse Phenotype Database Integration Consortium: integration [corrected] of mouse phenome data resources. Mamm Genome 18:157–163

Hu J, Mungall C, Law A, Papworth R, Nelson JP, Brown A, Simpson I, Leckie S, Burt D, Hillyard A, Archibald AL (2001) The ARKdb: genome databases for farmed and other animals. Nucleic Acids Res 29:106–110

Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G, Smith J, Searle S, Flicek P (2009) Ensembl 2009. Nucleic Acids Res 37:D690–D697

Jaiswal P, Avraham S, Ilic K et al (2005) Plant ontology (PO): a controlled vocabulary of plant structures and growth stages. Comp Funct Genomics 6:388–397

Kleine-Vehn J, Dhonukshe P, Swarup R, Bennett M, Friml J (2006) Subcellular trafficking of the Arabidopsis auxin influx carrier AUX1 uses a novel pathway distinct from PIN1. Plant Cell 18:3171–3181

Maglott D, Ostell J, Pruitt KD, Tatusova T (2007) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 35:D26–D31

Maltais LJ, Blake JA, Chu T, Lutz CM, Eppig JT, Jackson I (2002) Rules and guidelines for mouse gene, allele, and mutation nomenclature: a condensed version. Genomics 79:471–474

Menda N, Buels RM, Tecle I, Mueller LA (2008) A community-based annotation framework for linking solanaceae genomes with phenomes. Plant Physiol 147:1788–1799

Muller HM, Kenny EE, Sternberg PW (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol 2:e309

Ort D, Grennan AK (2008) Plant physiology and TAIR partnership. Plant Physiol 146:1022–1023

Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ, Adamusiak T, Brandizi M, Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi SG, Rocca-Serra P, Sansone SA, Sklyar N, Zhao M, Sarkans U, Brazma A (2009) ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res 37:D868–D872

Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35:D61–D65

Razem F, El-Kereamy A, Abrams S, Hill R (2006) The RNA-binding protein FCA is an abscisic acid receptor. Nature 439:290–294

Razem F, El-Kereamy A, Abrams S, Hill R (2008) Retraction. The RNA-binding protein FCA is an abscisic acid receptor. Nature 456:824

Risk J, Macknight R, Day C (2008) FCA does not bind abscisic acid. Nature 456:E5–E6

Seringhaus M, Gerstein M (2007) Publishing perishing? Towards tomorrow's information architecture. BMC Bioinform 8:17

Seringhaus M, Gerstein M (2008) Manually structured digital abstracts: a scaffold for automatic text mining. FEBS Lett 582:1170

Sprague J, Clements D, Conlin T, Edwards P, Frazer K, Schaper K, Segerdell E, Song P, Sprunger B, Westerfield M (2003) The Zebrafish Information Network (ZFIN): the zebrafish model organism database. Nucleic Acids Res 31:241–243

Swarbreck D, Wilks C, Philippe L, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, Radenbaugh A, Singh S, Swing V, Tissier C, Zhang P, Huala E (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res 36:D1009–D1014

The Gene Ontology Consortium (2010) The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res 38:D331–D335

The Reference Genome Group of the Gene Ontology Consortium (2009) The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. PLoS Comput Biol 5:e1000431

The UniProt Consortium (2010) The universal protein resource (UniProt) in 2010. Nucleic Acids Res 38:D142–D148

Van Auken K, Jaffery J, Chan J, Muller HM, Sternberg PW (2009) Semi-automated curation of protein subcellular localization: a text mining-based approach to gene ontology (GO) cellular component curation. BMC Bioinform 10:228