HHS Public Access

Author manuscript

Methods Cell Biol. Author manuscript; available in PMC 2019 January 07.

Published in final edited form as:

Methods Cell Biol. 2011 ; 104: 311–325. doi:10.1016/B978-0-12-374814-0.00017-3.

Data Extraction, Transformation, and Dissemination through ZFIN

Douglas G. Howe, **Ken Frazer**, **David Fashena**, **Leyla Ruzicka**, **Yvonne Bradford**, **Sridhar Ramachandran**, **Barbara J. Ruef**, **Ceri Van Slyke**, **Amy Singer**, and **Monte Westerfield**

ZFIN, the Zebrafish Model Organism Database, University of Oregon, Eugene, Oregon, USA

Abstract

The publication of a research article is the beginning of the digital life of its associated data. In this article, we will present an overview of how data are incorporated into ZFIN, with a particular emphasis on helping researchers make their work accessible to online databases.

I. Introduction

Genetics research flows through an online ecosystem of highly interconnected databases that acquire, aggregate, and re-distribute various kinds of information. In this article, we provide a brief introduction to the role of ZFIN (http://zfin.org) in the scientific data ecosystem, with an emphasis on helping researchers understand, contribute to, and benefit from the expanding world of online science data.

As the zebrafish model organism database, ZFIN aims to be the central information resource for the zebrafish researcher. ZFIN acquires, organizes, and shares a wide variety of zebrafish genomic, genetic, and phenotypic data. The data in ZFIN are continuously updated and comprehensive.

ZFIN provides powerful search interfaces, designed to address typical research questions such as: Where is my gene of interest expressed in the embryo? What are its mutant alleles and phenotypes? Where is the gene located in the genome? Are there mammalian orthologs of this gene? Are there reagents that label a specific gene product or downregulate expression of a gene? How should a new gene or allele be named? In addition, many different data files are available for download, giving researchers and bioinformaticians the ability to formulate queries geared to specific interests.

A major portion of data in ZFIN is curated from literature. Section II will provide insight into literature curation and the flow of data after entry into ZFIN. It will also present details on data exchange and integration with other major genome resources. Section III includes guidelines and tools to help authors make their manuscripts more accessible for rapid and accurate data curation and provides examples of ambiguity in published literature that could delay curation and the online accessibility of research data.

Section IV provides information for users interested in submitting research data or protocols. ZFIN hosts the Community Wiki to facilitate rapid online communication of tips and

resources, with sections for antibodies and experimental protocols. Researchers working on large-scale screens may contribute datasets directly to ZFIN.

II. Extracting Data and Data Flow

A. The Digital Life of Information Extracted from Publications

Modern hypothesis generation for genetic and genomic research depends heavily upon computational approaches to sift through large volumes of rapidly changing biological knowledge. A vast network of digital data sources and interconnecting data streams is being tapped into when a researcher conducts a search for genes with specific functions, gene expression patterns, phenotypes, or nucleotide sequences. Many of the data populating these digital data sources are manually curated from the literature. ZFIN is one such data source, and the digital life of published genecentric zebrafish research data originates at ZFIN. Once published data are digitized and integrated into the ZFIN database, they may be downloaded for use in other research efforts and analyses. In addition to individual research, data from the ZFIN database are also downloaded and presented by a number of major independent data repositories including the National Center for Biotechnology Information (NCBI), the Ensembl Project (Ensembl), and the Universal Protein Resource (UniProt). Giving published data a new digital life involves the work of ZFIN's scientific curators who cross check details such as gene identification and morpholino sequences and target genes, and disambiguate other details such as labeled anatomical structures and gene identification before these data are entered into the ZFIN database. Data corrections may occur at this step. Unlike the hard copies of published work, curated digital data can be easily updated in the future when new information comes to light or when errors are identified. Data that are viewed at ZFIN are captured into the database from four main sources: manual curation of published data, direct data submissions, data imported from external sources, and user feedback. In this section we focus on the digital life your published data acquires after being captured into ZFIN.

B. Data on the Move

Once published data have been successfully entered into the ZFIN database they immediately enter an expansive network of biological knowledge which is mobile, widely disseminated, searched, downloaded, and mined by thousands of researchers the world over every day. Curated data sets originating at ZFIN are downloaded hundreds of times each year by researchers, tool developers, and a number of major online resources including NCBI Gene, Ensembl, and UniProt Knowledgebase (UniProtKB) for presentation on the corresponding pages in those resources (Table I).

For example, Gene Ontology (GO) annotations made at ZFIN from published data can be found on NCBI Gene pages, embedded in UniProtKB and Ensembl records, and displayed on the European Bioinformatics Institute (EBI) QuickGO site, and the Gene Ontology Consortium's AmiGO site. These GO annotations are also used for gene enrichment analyses in various commercial and free software packages including AmiGO (Carbon *et al.*, 2009), DAVID (Huang *et al.*, 2009), eGOn (Beisvag *et al.*, 2006), GOEAST (Zheng and

Wang, 2008), and Agilent Technologies GeneSpring package. A summary of the major data exchanges is presented in Table II.

C. Data Exchange with Other Genome Resources

ZFIN participates in collaborations with a number of other genome resources to exchange, integrate, and display datasets. These collaborations provide opportunities to increase functionality and robustness of data in ZFIN. Periodic large-scale information exchange between databases reduces duplication of effort and expertise needed to acquire, update, and maintain the integrity of these data. Integration and display of data from disparate sources promotes increased functionality by eliminating the need for users to do this integration across databases on their own. Finally, integration of various datasets increases the robustness of ZFIN not only by adding data to the existing data types we capture, but also by providing new types of data that drive the development of ZFIN content and features.

The data exchanged and frequency of exchange differs for particular genome resource collaborations and is outlined below, but the process for each is similar. Files with data content requested by ZFIN are made available for pickup at the FTP sites of the collaborating databases. The data files are then run through a series of quality control scripts and/or data analysis tools to ensure incoming data are correctly associated with data in ZFIN. Data failing the quality control and consistency checks are analyzed by a ZFIN curator.

Incoming data that satisfy the quality control checks or are resolved manually are then integrated with the corresponding ZFIN content, displayed on ZFIN webpages, and attributed using a “data load” reference unique to each specific database collaboration. Subsequent data loads follow the same process, and the incoming data replaces the data from the previous load as identified by the data load reference. Any significant updates or corrections identified during the load process are resolved through direct communication between database staff, and corrections are made manually. Through these processes, the databases are kept updated and in sync.

1. Wellcome Trust Sanger Institute—The Wellcome Trust Sanger Institute is doing sequencing, and gene and transcript annotation of the zebrafish genome. Sequencing, annotating, and identifying the full complement of genes and their products are worthy endeavors in their own right; however, these data are maximally useful to the research community when they are integrated with information about expression patterns, phenotypes, homology, and function. ZFIN provides manual curation of these diverse data types and, in conjunction with the Sanger Institute, plays a pivotal role in fully integrating genome sequence, annotation, and experimentally derived data from the primary literature.

Gene, transcript, and clone annotation is done at Sanger by the Human and Vertebrate Analysis and Annotation team (HAVANA). Their annotations are stored in an internal database and become available to the public when a new version of the zebrafish Vertebrate Genome Annotation Database (Vega; Jekosch, 2004) is released. Just prior to a Vega release, the most current gene, transcript, and clone annotation are made available to ZFIN. Analysis and integration of transcript sequences are facilitated by use of our “redundancy pipeline”, a

series of scripts utilizing ZFIN's internal Basic Local Alignment Search Tool (BLAST) server, to compare incoming sequences with those already found in ZFIN in an effort to keep redundancy out of the database.

The results of the redundancy pipeline are analyzed by ZFIN curators who then make an expedited effort to merge, split, rename, and add genes to ZFIN to accommodate the integration of newly annotated genes, transcripts, and clones. Sanger and ZFIN then exchange database identifiers and establish external links to each other's database resources. Following this consolidation step, ZFIN ID mappings to the Sanger genes and transcripts are made publicly available for Sanger to include in the Vega release.

ZFIN curators provide nomenclature support by reviewing the novel genes annotated by Sanger and using sequence and orthology evidence to rename them with standardized, informative nomenclature. Changes to database records are synchronized, through automatic updates, to maintain consistency. Personal correspondence between databases is used to update annotation between releases.

2. **UniProtKB—**UniProtKB is a premier protein database that provides high-quality annotation, a wide variety of content, and a large number of links to other protein resources fully integrated into each database record. These characteristics make UniProtKB a valuable data resource with which to collaborate.

ZFIN obtains a flat file from the UniProtKB FTP site approximately every 4 weeks to update UniProtKB-derived data in ZFIN. This file is processed through a series of quality control scripts that attempt to match ZFIN genes and UniProtKB records based on common protein and nucleotide sequences. UniProtKB records that fail to match any ZFIN gene record or match records but show inconsistencies are set aside for manual curation.

The ZFIN gene and UniProtKB record associations that are unambiguous are used to load protein data derived from UniProtKB content into ZFIN. A load script populates ZFIN with links to UniProtKB, ENZYME, Pfam, PROSITE, and the Integrated Resource of Protein Domains And Functional Sites InterPro. UniProtKB comments are added to the "Gene Products" section of ZFIN gene records. UniProtKB keywords, in addition to InterPro domains and Enzyme Commission (EC) numbers, are used to drive automated electronic annotation of genes with GO terms.

ZFIN makes a file available for UniProtKB to facilitate linking UniProtKB records to ZFIN gene records. This file is updated weekly, and made available by FTP. UniProtKB establishes reciprocal links to ZFIN through the file exchange.

3. **NCBI—**The widespread use of genome resources at NCBI and the high quality of annotation and integration of genomic data necessitates links to and collaboration with NCBI. Linking to NCBI provides ZFIN users a direct gateway to a wide variety of genome data and analysis tools.

A weekly file exchange and series of quality control scripts populates ZFIN with links to NCBI Gene, Reference Sequence (RefSeq) sequences, GenPept sequences, and UniGene

clusters. Further, database links and nomenclature for ZFIN gene orthologs are checked for consistency. A report is generated to identify missing or incorrect NCBI Gene, Online Mendelian Inheritance in Man (OMIM), and Mouse Genome Informatics (MGI) identifiers as well as instances where nomenclature does not match.

ZFIN makes available to NCBI weekly updates of all ZFIN genes and their associated data. NCBI uses these files to supplement their own records and to establish reciprocal links back to ZFIN gene records thus providing their users with additional gene information.

It is important to understand these different data sources and analysis tools for what, at first glance, may appear to be the same data. Each data source may filter the data they load into their system, or they may present only a subset of the data available from the original source, or they may update the data at different frequencies. All of these differences affect the results of an analysis. Understanding the data flow ensures that the most current and complete data set is used to conduct analyses. For example, the most current and complete GO annotation set for zebrafish is provided by ZFIN. It can be obtained from the Gene Ontology Consortium CVS repository or by FTP (ftp://ftp.geneontology.org/pub/go/gene-associations/gene_association.zfin.gz). When publishing, users should be sure to record the version number and download date of any downloaded data and the source from which those data were obtained. This is critically important to ensure reproducibility of results.

III. Ensuring Accurate and Rapid Data Dissemination

A. The Importance of Accuracy and Precision in Publications

To understand the information in a piece of scientific literature fully and correctly, the objects being reported on must be unambiguously identifiable by the reader. ZFIN curators must unambiguously identify genes, morpholinos, alleles, transgenic lines, genotypes, antibodies, anatomical structures, developmental stages, genetic backgrounds, and experimental conditions. Ambiguity in any of these data is likely to prompt an email to authors from a ZFIN curator or a colleague looking for clarification of the details. Precision and accuracy is critically important in executing experiments. Using the same precision and accuracy to describe data in publications will maximize their value by allowing readers to interpret published results accurately with the least effort, while accelerating the curation process and maximizing the probability of curated data being associated with the correct genes. See Table III for examples of ambiguity and ways to avoid them.

B. Acquiring and Using Proper Nomenclature

Using correct and current nomenclature for zebrafish genes, mutants, and transgenic lines will ensure that published data can be correctly integrated into the electronic stream. When preparing research articles specific, correct, approved names of genes or mutants being described are important for accurate identification. The use of “hedgehog gene” in a research publication might lead the reader to reasonably infer that the author is discussing the *shha* gene; however, the data associated with this statement cannot be immediately curated into ZFIN, because there are several members of the hedgehog gene family in zebrafish. In order to annotate these data, ZFIN curators must contact the authors to confirm which gene was

described, thus increasing the time it takes for the data to enter the digital data stream. The use of specific and current nomenclature in research manuscripts results in the unambiguous identification and integration of data into the ZFIN database, thus allowing this information to be accessed rapidly by the broader scientific community.

ZFIN is the governing body for zebrafish nomenclature and is here to assist with nomenclature needs (nomenclature@zfin.org). When naming a new gene, it is important first to conduct a sequence similarity search using BLAST in ZFIN and/or Ensembl to determine whether the gene in question has already been identified and named. There is always the possibility that the gene has a name and symbol, even if there are no publications describing it, due to collaborative efforts by ZFIN and Sanger to name unidentified genes. If there are no identical matches for the gene sequence, closely related genes, such as genes created by genome duplication, should be checked. If the analysis of sequence identity and conserved synteny suggests that the gene of interest is a duplicate of a named zebrafish gene, the proposed gene name and symbol should be comprised of the existing gene name with an “a” or “b” appended to the end. The gene already listed in ZFIN will be renamed by adding an “a” or “b” to the original symbol to indicate the paralogous relationship of the genes.

If the gene of interest has not been named, the gene sequence should be analyzed using BLAST at NCBI to identify human and mouse orthologs. In general, zebrafish genes are named after their human and mouse orthologs using the mammalian symbol for the zebrafish gene following the zebrafish nomenclature guidelines. Zebrafish gene symbols are in lower case italics, with no “z” or “zf” prefixes. If the gene in question is a member of a gene family, the “root” symbol of the gene family in zebrafish, should be used, with gene members numbered sequentially. If there is no gene family in zebrafish, but there is in human or mouse, the established “root” symbol from human or mouse should be used for the zebrafish gene following zebrafish nomenclature guidelines for gene names and symbols (http://zfin.org/zf_info/nomen.html#1).

For those cases where the gene of interest is not a duplicate of a known zebrafish gene, does not have a known human or mouse ortholog, and does not appear to be a member of a gene family, the gene name and symbol that is being proposed should be checked against ZFIN to ensure that it is not already in use. The proposed gene name and symbol should also be checked against the Human Gene Nomenclature Database (http://www.genenames.org/) and Mouse Genome Informatics (http://www.informatics.jax.org/) to ensure that it is not already in use for a different gene or mutant in human and/or mouse. After a suitable gene name and symbol has been determined, the zebrafish nomenclature committee should be contacted for confirmation and approval (nomenclature@zfin.org). The inclusion of sequence or sequence accession numbers, mapping information, and additional data on orthology, including phylogenetic trees, is recommended when submitting a name and symbol to the nomenclature committee to ensure immediate processing of the request. Once approved, ZFIN will assign and reserve the symbol for the novel gene.

Unambiguous identification of genes, sequences, reagents, fish, stages, and anatomical terms in articles or direct data submissions will greatly increase the speed with which these objects

can be incorporated into the ZFIN database and subsequently disseminated to other databases.

C. Accession IDs and Catalog Numbers

Accessions, sequences, and catalog numbers provide unambiguous means of identifying the data objects in your paper. A GenBank/EMBL/DDBJ accession number is the most precise means of matching genes in a publication to genes in the ZFIN database. Transient identifiers such as gene prediction identifiers should be avoided. A ZFIN database ZDB, NCBI Gene or Ensembl identifier allows similar identification of genes, transcripts, and other objects.

Reporting the exact sequences of the morpholinos used in your paper prevents ambiguity and supports correction of errors. If several different morpholinos targeting the same gene are used in a study, it is helpful to indicate which one was used in each figure, and at which concentration. The morpholino sequence and notes regarding the morpholino target (e.g., splice blocking) are reported on ZFIN morpholino pages.

Reagents listed in the materials and methods should include unique identifiers. This information is particularly important for antibodies, where multiple products are often available that match the description in the methods section.

D. Utilizing Ontologies

ZFIN uses multiple ontologies to curate and display information (see Table IV). Ontologies consist of a controlled vocabulary of terms and their relationships to one another, and thus provide a framework for standardizing data collection and searching.

Ontologies are created and maintained by the bioinformatics community, and are continually improved and updated. ZFIN maintains the Zebrafish Anatomical Ontology (ZFA), which describes the anatomical structures and cells found in zebrafish. The anatomy terms are related to one another by structural (is type of, is part of) and developmental (develops from) relationships.

Zebrafish gene expression patterns and phenotypes are recorded using the ZFA, GO (cellular component), CL, BSPO, and ZFS ontologies. Phenotype data are captured using the ZFA, GO, CL, BSPO, ZFS, and PATO ontologies. Because ontologies are an integral part of ZFIN, expression patterns, and phenotypes in publications are translated by curators into the highly structured vocabularies of ontologies. For example, built-in constraints prevent curation of a specific anatomical structure outside the stage range specified in the Zebrafish Anatomical Ontology: expression of a gene in a 24-h post-fertilization (hpf) fish cannot be annotated to retina, because the retina first appears at the Prim-15 stage (30–36 h). The expression can be annotated to the structure the retina develops from (optic cup) or is part of (eye), provided that structure exists at 24 h.

Researchers can improve the speed of data entry into the electronic data stream by getting to know the major ontologies, helping to update them, and by using the most precise terms possible in their publications and ZFIN direct-data submissions.

IV. Submitting Data Directly to ZFIN

Submitting data directly to ZFIN provides another avenue for disseminating and integrating your data. This approach offers many benefits including:

- obviating the need for individual laboratories to maintain publicly accessible web resources
- providing data that are fully integrated and searchable in the context of other zebrafish data
- small data sets, not substantial enough for an entire paper and data from high throughput projects can be integrated into the expansive network of zebrafish data

ZFIN provides two avenues for disseminating your data: a Community Wiki and integration of your data into the ZFIN database.

The research community can share protocols and information about antibodies using the ZFIN Community Wiki (http://wiki.zfin.org). Everyone is encouraged to participate.

ZFIN also accepts direct submission of expression, phenotype, and transgenic insertion site mapping. The data can be either an extension of previously published work or part of a large-scale project. There are a few requirements to ensure successful integration of directly submitted data. Contacting ZFIN as early as possible with submission details will help to facilitate integration of data.

A. Submitting Expression and Phenotype Data

The use of well-defined, controlled vocabulary terms to annotate gene expression and phenotype data is essential to facilitate inter- and intra-species comparisons and searches against the database. As described earlier, annotations in ZFIN derive from several biological ontologies (see Section III.D). To integrate a directly submitted data set, it must be annotated using terms from these same ontologies. For projects already storing their data in a database and annotating with ontology terms, it is relatively simple to submit data. For others, there is a free program, Phenote (http://www.phenote.org/) that supports annotation of gene expression and phenotype data using the same ontologies utilized by ZFIN. Because Phenote manages the ontologies, all that is needed is to choose the correct configuration and start annotating. Phenote is a program under active development, so it is recommended to check the release notes before installing and report any bugs so they can be fixed in the next release.

B. Submitting Transgenic Insertion Sites to ZFIN

GBrowse, a generic genome browser (http://gmod.org), is used by ZFIN to host the Sanger Institute's genome assembly based on fully sequenced BAC and PAC clones from the Tübingen strain with Vega gene and transcript annotation (Ashurst *et al.*, 2005; Jekosch, 2004). GBrowse allows users to upload and display custom tracks for their own use. To set up a collaboration to display transgenic integration sites on a public GBrowse track at ZFIN, specific types of information are needed for each insertion as shown in Table V.

C. Using the Community Wiki to Share Protocols and Antibody Data

The Community Wiki provides an easy way for the zebrafish research community to share protocols and antibody data. All that is necessary is to create an account, login to the wiki, and post your information. Only you will be able to edit your postings; others are encouraged to submit comments. In addition to community submitted protocols, the wiki contains protocols from *The Zebrafish Book*, 5th edition. Protocols are organized into sections corresponding to the chapters of *The Zebrafish Book*, 5th edition. The Antibody Wiki differs slightly in that a specific form is provided for antibody submission. The model antibody (https://wiki.zfin.org/display/AB/example-antibody), and existing antibody submissions can assist in navigating the form. The Antibody Wiki is also populated with antibodies curated at ZFIN from zebrafish publications. Community members are encouraged to share their tips for best results or other comments about these antibodies.

V. Conclusions

ZFIN is a manually curated and highly integrated resource for zebrafish research data. Data curated at ZFIN are shared and integrated with many major bioinformatics resources and are available for use by the larger scientific community. Accurate and rapid dissemination of these research data relies on the adoption and use of proper nomenclature to identify objects and specific ontologies to describe anatomical terms, growth stages, phenotypes, and functions.

Acknowledgments

The authors would like to thank the zebrafish community and our external advisory board for their continued support and guidance. They also thank the entire ZFIN staff for their dedication and commitment to ZFIN's success. ZFIN is supported by the National Institutes of Health (P41 HG002659, R01 HG004838, and R01 HG004834).

References

Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, Meidl P, Searle SM, Stalker J, Storey R, Trevanion S, Wilming L, and Hubbard T (2005). The Vertebrate Genome Annotation (Vega) database. Nucleic Acids Res. 33 (Database issue), D459–D465. [PubMed: 15608237]

Bard J, Rhee SY, and Ashburner M (2005). An ontology for cell types. Genome Biol. 6, R21. [PubMed: 15693950]

Beisvag V, Jünge FK, Bergum H, Jølsum L, Lydersen S, Günther CC, Ramampiaro H, Langaas M, Sandvik AK, and Laegreid A (2006). GeneTools – application for functional annotation and statistical hypothesis testing. BMC Bioinformatics 7, 470. [PubMed: 17062145]

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, and AmiGO, Hub Web Presence Working Group (2009). AmiGO: online access to ontology and annotation data. Bioinformatics 25(2), 288–289. [PubMed: 19033274]

Huang DW, Sherman BT, and Lempicki RA (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nat. Protoc 4(1), 44–57. [PubMed: 19131956]

Eilbeck K, Lewis S, Mungall CJ, Yandell M, Stein L, Durbin R, and Ashburner M (2005). The sequence ontology: a tool for the unification of genome annotations. Genome Biol. 6, R44. [PubMed: 15892872]

Jekosch K (2004). The zebrafish genome project: sequence analysis and annotation. Methods Cell Biol 77, 225–239. [PubMed: 15602914]

Kawakami K, Takeda H, Kawakami N, Kobayashi M, Matsuda N, and Mishina M (2004). A transposon-mediated gene trap approach identifies developmentally regulated genes in zebrafish. Dev. Cell 7(1), 133–144. [PubMed: 15239961]

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A,Mungall CJ, Consortium OBI, Leontis N, Rocca-Serra P, Ruttenberg A,Sansone S-AA, Scheuermann RH, Shah N, Whetzel Whetzel. P., and Lewis S (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol 25, 1251–1255. [PubMed: 17989687]

Zheng Q, and Wang XJ (2008). GOEAST: aweb-based software toolkit for Gene Ontology enrichment analysis. Nucleic Acids Res. 36 (Web Server issue), W358–W363. [PubMed: 18487275]

Table I

Number of unique downloaders of specific curated data types from January 1–December 3, 2010

Curated data type	Unique downloaders (Jan 1–Dec 3, 2010)
Orthology	588 (Human)
	256 (Mouse)
Gene ontology	418[a]
Markers (alleles, genes, etc.)	367
Morpholinos	264
Phenotypes	244
Wild-type expression	190
Antibodies	186

[a]Estimate extrapolated from actual counts (38) for the 4-week period preceding December 14, 2010.

Table II

A summary of data exchanges between ZFIN and other sites

Sites presenting curated data originating at ZFIN	Data types
NCBI Gene	Nomenclature, gene ontology annotations
Sanger Institute (Ensembl, Vega)	Nomenclature, gene ontology annotations
UniProtKB	Nomenclature, gene ontology annotations
AmiGO	Gene ontology annotations
LAMHDI	Phenotype
Neuroscience Information Framework	Phenotype, expression
Bgee	Expression
Phenoscape	Phenotype
4DXpress	Expression

Table III

Commonly encountered examples of ambiguity in published literature and proposals to maintain specificity in describing data

Ambiguity	Solution
Citing a prior publication for a morpholino, transgene, mutant line, reagent, etc., where the cited article contains ambiguity in the data (multiple morpholinos, alleles, etc.)	Cite the earlier paper, but also include the morpholino sequence, allele/Tg line designation, catalog number, etc., directly in the paper
Referring to a gene by the root symbol rather than the full official symbol	Duplicated zebrafish genes typically have a numeric or alphabetic suffix relative to their mammalian orthologs. Include the full current zebrafish gene symbol and a nucleotide accession number to avoid ambiguity. The nucleotide accession number ensures continued correct gene identification in the event the gene symbol changes in the future
Reporting the use of a commercially available antibody by citing only the company	It is common for commercial vendors of antibodies to have multiple antibodies against a target protein. Include the company and the catalog number to ensure the antibody can be identified unambiguously in the future
Referring to the use of a mutant by the locus or gene name	Many genes/loci have more than one allele. Avoid this ambiguity by including an allele designation for every mutant used

Table IV

Ontologies used to curate and display data in ZFIN

Ontologies used in ZFIN
Zebrafish Anatomical Ontology (ZFA)
Zebrafish Stage Ontology (ZFS)
Gene Ontology (GO) (http://www.geneontology.org/)
Phenotypic Quality Ontology (PATO; http://obofoundry.org/wiki/index.php/PATO:Main Page)
Cell Ontology (CL; Bard *et al.*, 2005)
Sequence Ontology (SO; Eilbeck *et al.*, 2005)
Spatial Ontology (BSPO)

Many of these ontologies can be found at the Open Biological and Biomedical Ontologies site (OBO Foundry; Smith *et al.*, 2007).

Table V

An example of the specific information needed to display integration sites on a public GBrowse track (from Kawakami *et al.*, 2004)

Information type	Data
Construct name	Gt(T2KSAG)
ZFIN construct ID	ZDB-GTCONSTRCT-070430–1
Line number	nksag014
ZFIN feature ID	ZDB-ALT-080827–8
GenBank Accession ID (for the sequence at the integration site)	AB175057