Published in final edited form as:

Genesis. 2015 August ; 53(8): 486–497. doi:10.1002/dvg.22873.

Xenbase; core features, data acquisition and data processing

Christina James-Zorn¹, Virgillio G. Ponferrada¹, Kevin A. Burns¹, Joshua D. Fortriede¹, Vaneet S. Lotay², Yu Liu², J. Brad Karpinka², Kamran Karimi², Aaron M. Zorn¹, and Peter D. Vize²

¹Division of Developmental Biology, Cincinnati Children's Hospital, 3333 Burnet Ave, Cincinnati, OH 45229, USA

²Departments of Biological Science and Computer Science, University of Calgary, 2500 University Drive NW, Calgary, AB T2N1N4, Canada

Abstract

Xenbase, the *Xenopus* model organism database (www.xenbase.org), is a cloud-based, web accessible resource that integrates the diverse genomic and biological data from *Xenopus* research. *Xenopus* frogs are one of the major vertebrate animal models used for biomedical research, and Xenbase is the central repository for the enormous amount of data generated using this model tetrapod. The goal of Xenbase is to accelerate discovery by enabling investigators to make novel connections between molecular pathways in *Xenopus* and human disease. Our relational database and user-friendly interface make these data easy to query, and allows investigators to quickly interrogate and link different data types in ways that would otherwise be difficult, time consuming, or impossible. Xenbase also enhances the value of these data through high quality gene expression curation and data integration, by providing bioinformatics tools optimized for *Xenopus* experiments, and by linking *Xenopus* data to other model organisms and to human data. Xenbase draws in data via pipelines that download data, parse the content, and save them into appropriate files and database tables. Furthermore, Xenbase makes these data accessible to the broader biomedical community by continually providing annotated data updates to organizations such as NCBI, UniProtKB and Ensembl. Here we describe our bioinformatics, genome-browsing tools, data acquisition and sharing, our community submitted and literature curation pipelines, text-mining support, gene page features and the curation of gene nomenclature and gene models.

Keywords

Xenopus; model organism database; biocuration; gene expression; genomics; phenotypes; human disease modeling

Introduction

Animal models are essential for biomedical research and frogs of the genus *Xenopus* are among the premier model organisms for vertebrate research. The integration of genomic and biological data has driven a tremendous leap-forward in biomedical research in recent years.

Xenopus research is supported by Xenbase, the *Xenopus* model organism database (MOD). The Xenbase mission is to provide the international research community with a comprehensive, easy to use, web-accessible resource that integrates all of the genomic and biological information from *Xenopus laevis* and *X. tropicalis* research. These two species have many experimental advantages and we incorporate both, despite the complexity caused by dealing with species of different ploidy (Vize et al. 2015). The small diploid species *X. tropicalis* has a short generation time and is well suited for genetics and both species are well suited to embryo microinjection, tissue explants, and genome editing and morpholino based gene knock down. *X. laevis* is also widely used by cell biologists for technologies using egg extracts and by physiologists to express ion channels in oocyte membranes. Due to gene sequence similarity *in situ* hybridization probes designed to *X. laevis* hybridize to and can be used to detect mRNA expression in *X. tropicalis* (e.g. Khokha et al. 2002) and vice versa (Parrain et al. 2012). Injected morpholino oligonucleotides designed against sequences in one species will frequently suppress gene expression when injected into embryos of the other species.

The diverse data types generated by *Xenopus* research require a highly functional database where biological and genomic content can be integrated, compared and indexed. Xenbase was created in 2005 to fulfill this role, establishing itself as an essential resource for this vertebrate model. Xenbase is an object relational database that integrates the vast body of information from *Xenopus* research including genomes, genes, mRNAs, proteins, expression data, function, orthologs and reagents. Xenbase obtains these data from external publicly accessible resources (NCBI, JGI, OMIM and other MODs), from the published literature (PubMed and PubMedCentral) and by direct community submissions. The major goal of Xenbase is to integrate these data, provide high-quality annotation of these data, and to provide analysis tools (e.g., BLAST and Textpresso) optimized for the *Xenopus* experimental system. Our biocuration staff manually annotate incoming information, focusing on gene expression curation and interlinking different data types, which in turn enables powerful computer queries. In addition, Xenbase is the gatekeeper to *Xenopus* gene nomenclature, and thus maintains the official gene names. In doing so, Xenbase links *Xenopus* data to genes in humans and other model organisms, and then provides the enriched annotated data back to NCBI, Ensembl and other MODs through an automated data-sharing infrastructure. By linking *Xenopus* data on each gene page to all of these external resources, *Xenopus* data are accessible to the broader scientific community. Xenbase also provides data sharing infrastructure via our ftp/download resources for unique community generated datasets that are not available elsewhere. Xenbase also acts as a communication hub for the *Xenopus* community. To enhance the utility of the website the Xenbase user interface underwent a major overhaul in 2013. New content is constantly being added through automated and semi-automated input pipelines and through major new data type initiatives. The most recent such additions are to the reagents section, where full database support has been added for antibodies, morpholinos and the *Xenopus* ORFeome (Karpinka et al. 2015), as well as links to genome editing tools (TALENs, CRISPr etc.) that provide *Xenopus* genomic support.

In this paper we describe many of the core Xenbase features in more detail, illustrating how Gene Pages intersect with literature curation, how data mining works, how the Gene Pages are generated and how the combination of manual, automatic and semiautomatic processes, curation and data sharing support research using *Xenopus* and in turn make *Xenopus* data available to other resources.

Xenbase Content

A wide variety of content is stored in Xenbase, and most of it is available through search queries. Some types of data are used for data mining or for building links between other data objects, and these are sometimes not accessible. In Figure 1, many of the types of content are illustrated. Xenbase gathers this content through a variety of processes. Data pipelines update content from remote resources such as the NCBI, UniProtKB, OMIM and other MODs. Once downloaded, various text mining systems automatically process content, build links between identified features and partially annotate the new content. These processes are described in more detail below.

Gene Pages

Many of the different content types illustrated in Figure 1 are assembled and displayed in a gene centric manner via our “Gene Page” feature. The Xenbase Gene Page catalog houses information on over 15,800 dynamically generated pages, which contain information about both the *X. tropicalis* gene and both *X. laevis* homeologs (initially designated as ‘a’ and ‘b’, and now as -L and -S corresponding to the gene loci identified on the long (L) or short (S) chromosomes of the *X. laevis* homeologous pairs). Approximately 30 Structured Query Language (SQL) queries run to build the content of each page. In order to keep the pages uncluttered and easy to navigate, the information is organized in ‘tab’ format. This layout keeps page load time short as content is only loaded when a tab is selected. The ‘Summary’ tab provides the gene name, symbol and synonyms, the genes functional description, whether it is associated with any human diseases at OMIM, interactants calculated via co-citation, and provides summary expression image(s) that depict key features of the expression pattern. The remainder of the page separates content using rows and columns to sort various content types and to organize *X. tropicalis* and *X. laevis* content separately. The ‘Molecules’ section provides links to the genomic, mRNA and protein sequences, as well as a link to load each sequence into our BLAST service. Next is a section showing the genomic organization of the *X. tropicalis*, *X. laevis*-L and *X. laevis*-S gene models via graphical snapshots generated by GBrowse, and links to the different genome builds in our browsers via drop-downs. The ‘Expression’ section in this view provides information on resources used to analyze gene expression data, such as EST profiles, UniGene sets, and a hyperlink to GEO datasets. The vast bulk of expression data are in a different tab and is accessed either by the hyperlink to Expression Images in this view or by selecting the “Expression” tab at the top of the Gene Page. Under ‘Orthology’, we link to the human, mouse (Eppig et al. 2015), zebrafish (Sprague et al. 2001) and chicken (Bell et al. 2004) orthologs and include information on OMIM gene content (Amberger et al. 2009) when available. The Gene Page also highlights the first and most recent literature available that cite this gene under ‘Publications’, we provide a count of how many papers are catalogued in the above

‘Literature’ tab for this gene, and a hyperlink that allows the user to “View All Papers”. ‘Functional Ontologies’ gives GO annotations, and a ‘Reagents’ section catalogs antibodies, morpholinos, cDNA and ORF clones available for this gene. Phenotype curation is currently in development and when available, published phenotypes will also appear on the gene summary page, and more extensively on a separate tab. All of these features can be seen in Figure 2, which illustrates the Xenbase gene page for the gene paired box 6.

Gene Page tabs

The Gene Page ‘Expression’ tab links to image-based, community and text-mined data on gene expression. For about 11,000 Gene Pages, a microarray based transcription profile graph compares *X. laevis* and *X. tropicalis* expression profiles across developmental stages, a community submitted data set from the Kirschner laboratory (Yanai et al. 2011). This is followed by a series of images illustrating gene expression, mostly from *in situ* hybridization data and/or immunolocalization. The images are derived from multiple sources, including direct submissions from laboratories, from large scale screens and from images published in journals. Community submitted gene expression is placed above the images redisplayed from published articles. By clicking on a specific image, a pop-up opens, enlarging the image and showing a table containing manually curated expression data, listing structures in which the gene is expressed at specific developmental stages. These terms are drawn from the *Xenopus* Anatomical Ontology, also known as the XAO (Segerdell et al. 2013), an extensive list of anatomy terms and embryonic stages that form the backbone of gene expression curation. The XAO contains thousands of terms and relationships. Each anatomical system (e.g., central nervous system, alimentary/biliary system), tissue and structure (e.g., neural fold, mesoderm, blastopore), many cell types (e.g., ciliated epidermal cell, migrating neural crest cell) and anatomical regions (e.g., optic field, central retina) are separate terms. XAO cell types are derived directly from the Cell Ontology (Meehan et al., 2011) and cross referenced to this resource by a dbxref. Each XAO term is fully defined, stage restricted (starts_at, ends_at), and related to other terms by ‘is_a’, ‘part_of’, ‘develops_from’ and ‘develops_into’ relationships. Stage restriction means that a term such as ‘tail fin’ is not available for curation at any developmental stage before the tail has formed, nor after the tail has resorbed at metamorphosis. Each term has its own web page where detailed information on the term can be obtained. The Expression tab on the XAO term pages list all genes expressed in that specific tissue or structure, ordering them via images, papers and clone IDs that have been linked to the XAO term.

Bioinformatics and genome browsing

One of our most heavily used resources is our dedicated BLAST VM (virtual machine) server (Westesson et al. 2013). This VM runs BLAST+ (Camacho et al. 2009) and can perform a BLASTN search of the entire *Xenopus* genome in less than 2 seconds. The results of BLAST search can be viewed and hyperlinks are provided to the top of the matched sequence linking the user to the matched region within our genome browser, or to a gene page if the result maps to one (available only when cDNA or proteins are used as the query). The other major software resource is our web browser, GBrowse (Stein et al. 2002). This software was produced by the Generic Model Organism Database (GMOD) consortium and

is an excellent tool for visualizing genome features. Users can zoom in to the nucleotide level using the “GC Content” track and at any time capture the sequence they are viewing in FASTA or other formats by clicking on the floppy disc icon adjacent to the track name. The view of any track can be changed by clicking on the small spanner icon right next to the floppy disc icon. The order of tracks can be changed by clicking and dragging any track header (color highlighted text). When a user clicks on a track feature, for example a gene model, a pop-up window will display links to other content, such as details of the gene model or links to the source of the track or to gene pages. A wide variety of tracks are available on all genomes, and an example of these are illustrated in Figure 3. Not all tracks are available on all genome builds, but we are working towards achieving this goal. Recently we added NCBI gene and transcript models for the *X. tropicalis* v7.1 genome to our GBrowse and will be releasing nested tracks generated by Topoview in the near future (Figure 3). This is necessary as Xenbase already hosts over 100 Next Generation Sequencing tracks displaying data such as ChIP-seq (van Heeringen et al. 2014).

Users can find genomic regions of interest in multiple ways. The most straightforward is directly from a Gene Page. This feature displays a small snapshot depicting the most recent gene model for a particular gene, while a drop down links to other genome builds for which a model for this gene has been annotated. By clicking on the snapshot image, Xenbase will load the genomic region of interest into GBrowse, along with a default set of supporting tracks. Selecting the “Select Tracks” tab on the top left of the browser view can change the set of tracks displayed. A second method is via BLAST. If a user performs a BLAST search against any genome builds, a hyperlink to GBrowse displaying the matched sequence is located at the top of each BLAST hit return. Finally, GBrowse has its own search functionality. In the text field near the upper left corner, users can enter gene symbols and the search will load the corresponding gene model. At the moment this last option is somewhat problematic in *X. laevis* genomes where there are two copies of each gene, however a disambiguation system is in place and users can scroll through search returns and select hits labeled “gene” and find both allo-alleles (formerly referred to the ‘a’ and ‘b’ alleles, and to be updated to ‘L’ and ‘S’ designations).

Curation pipelines for literature and community submitted data

High quality curation of the published *Xenopus* research is the cornerstone of Xenbase content and utility. The diverse body of literature contains many different curatable data types that we want to capture, including; genomes and gene models, DNA, RNA and protein sequences, gene names and symbols, gene function, protein structure, gene expression patterns, new anatomy terms, data on orthology, reagents (morpholinos, antibodies, ORFs), transgenic experiments and transgene structure, phenotypes and disease associations. A major challenge for all MODs is that published literature is not immediately computer readable, and thus comprehensive data annotation requires manual ‘human eyes’ to read and interpret much of the data. To be as effective and efficient as possible, Xenbase utilizes a curation pipeline that integrates automatic, semi-automated and manual processes to index and annotate data using ontologies and standardized procedures that are the accepted best practices as shared across major MODs (Figure 4.). This allows *Xenopus* data to be compared to other animals, and to human disease genes and phenotypes.

The majority of our curatorial effort is to read and annotate the published literature describing *Xenopus* research. Xenbase contains a corpus of ~46,000 papers on *Xenopus*, mirroring PubMed content on this subject. The data we automatically capture includes content such as authors, title, abstract, links to PubMed and the paper at the publisher's website, figures (if we have permission or the paper is open access), figure legends, genes cited, antibodies and morpholinos used in the paper are all automatically captured. Abstracts and figure legends are processed via text-matching for regular expressions to identify XAO terms, gene names and synonyms and authors with Xenbase user accounts, and these matches are used to generate internal hyperlinks. This allows users to seamlessly flow between gene pages, literature, reagents and other features. An interface allows curators to edit the machine generated annotations so incorrect matches and missed content, such as antibodies, morpholinos or genes cited can be improved.

Figure 4 illustrates the data curation pipelines, which include a combination of automated, semi-automated and manual steps. Published literature is added to Xenbase via two methods. The first is automatic weekly scripts that run the NCBI tool E-Utilities (Sayers et al. 2012) to query PubMed for the keyword '*Xenopus*' and parse the content we require from matched papers into Xenbase database tables plus generate a Xenbase ID for the publication. This content is then queried to generate a dynamic article page for each paper illustrating data on; title, authors, full abstract, associated metadata, links to the PubMed ID, PMC ID and the journal website all of which are associated with the Xenbase ID. The second method of data gathering is to perform manual searches of journal eTOCs to identify additional articles lacking the keyword "*Xenopus*" in title or abstract, but relevant to *Xenopus* research. These articles are manually added to Xenbase via a loader that manually runs E-Utilities using the PubMed ID and loads the returned content in a similar way as described above. This system also cross checks the database to prevent duplicate entries using the PubMed ID. In both systems, automated text-matching algorithms identify gene symbols and synonyms, *Xenopus* anatomy ontology (XAO) terms and authors and build links to the appropriate gene pages, XAO pages and registered user (author) pages respectively. The full article PDF is downloaded from PMC or the publisher, and is automatically converted to XML. The text is then indexed using the text-mining software Textpresso, which enables both Xenbase curators and the public to search the full text with complex queries (see Textpresso description below). Due to copyright restrictions, full text content is not directly accessible by users.

For high priority papers where we have publisher permission to redisplay images (currently from approximately 60 journals and open access sources) curators choose to import figures and figure captions from the journal website using custom semi-automatic Java-based image scraping software. This system loads previews of the figures and figure legends, allowing curators to select figures as allowed by our copyright policy (i.e. we have re-display permission from the publisher and/or open access material). Once figures are selected the scraper downloads the appropriate figures and legends and runs them through our image formatting and scaling systems to generate the different jpg images displayed in various contexts. For journals for which we do not have image scrapers, the figures are manually imported using a different custom Java-based system. Again, automated text-mining scripts

use content from various database tables to mark up the figure legends and build links to gene symbols and XAO terms. Our current protocol includes curation of gene names and symbols, gene nomenclature changes, gene expression patterns, morpholinos, antibodies, transgenes and author information. Our team is now developing interfaces to curate several new data types for Xenbase, including mutants, phenotypes and GO annotations, and to update the transgene curation interface to adapt to the latest naming guidelines for transgenic *Xenopus* lines held by laboratories and stock centers.

Gene expression images from large-scale screens are directly submitted to Xenbase. Smaller submissions can be manually added directly to the gene pages through the curatorial interface (Bowes et al. 2010). If the datasets are very large, we build custom loaders to import the data. Large-scale expression screens generate many hundreds to thousands of gene expression images, and the level of redundancy in these very large data sets can be high. For example, we data share with XenMark (Gilchrist and Pollet 2012), and we were both provided with data from a large-scale screen of over 33,300 images. The XenMark team developed image processing software to choose high quality images, e.g. those with strong *in situ* hybridization staining-to background, reducing these data to a more manageable set of ~4,800 images. After initial quality control, we upload images directly to the gene pages, automatically attributing each image to the laboratory/resource of origin, adding probe, accession numbers and associated metadata, and when possible, we capture some curation details such as embryonic stage. All non-literature images are considered “community submitted” and are placed in the curation pipeline and annotated like literature figures. High quality gene expression images from either community and literature classes can be curator-promoted to the summary page of the appropriate gene. Our goal is to provide a developmental series on a gene summary page, with each stage annotated using anatomy terms from the XAO.

Additional expression content is generated by two automated systems. The first uses text mining to identify XAO terms and associated these with papers. This system cannot assign to an individual gene and usually multiple genes are matched per paper. It does however generate a hyperlink so readers can jump to the XAO term and view a full set of genes associated with that term. The second automated system directly associates genes with tissues and stages of expression. This works using the relationships between cDNA clones and their library of origin. For example we have around 10,000 EST sequences from an adult kidney cDNA library. Most of these EST are linked to Xenbase gene IDs via internal BLAST processes. In the example of the adult kidney cDNA library, all gene IDs that have cDNA or EST sequences derived from this library will have the XAO terms for ‘kidney’ stage ‘adult’ added to the expression profiles of these genes. A list of terms, the sum of automated and manual annotations, appears at the top of the Expression tab on Gene P. ages.

Xenbase has also implemented a pipeline to import open access images from the PubMedCentral archive. Information from publicly available *Xenopus* papers from PubMed such as title, author list, and abstract are regularly added to Xenbase as described above. However, PubMed does not provide downloadable images for these papers. Since many *Xenopus* papers are open access, this pipeline regularly runs and uses the FTP protocol to download images and image captions for publications on *Xenopus* from the PubMedCentral

open access subset. These images and captions are added to the Xenbase database if they are not already present. For papers that already have some images, for example those displaying gene expression information, only figures not represented are inserted, while for papers where Xenbase has no figure content, all figures and legends are inserted. Over 10,000 images have already been added to Xenbase by this pipeline. Once loaded the regular text-mining systems process and markup the figure legends as described elsewhere in this paper.

The curated content is available through many different queries. For example, gene expression data are most easily browsable via the Gene Page using the .Expression tab (Figure 2). Alternatively, there is a separate highly customizable expression search system available under the top navigation bar by following the path Expression/Expression Search (see Bowes et al., 2010). Reagents such as antibodies and morpholinos are directly accessible via the top navigation bar under “Reagents”, or via literature pages where annotated reagents are highlighted (and hyperlinked to reagent pages) in between the abstract and the figures (Figure 5).

Text-mining support

Xenbase provides a Textpresso data mining system (Muller et al. 2004) that allows users to perform advanced queries of our corpus of *Xenopus* publications. Full text versions of all papers describing *Xenopus* research are stored in a private database. These articles are then indexed using Textpresso ontologies that identify related terms used widely in queries. For example, a user may search for papers describing “expression” of a gene. The Textpresso ontology treats “expression” as a synonym for “transcription” and other terms, so the query will look for co-usage of the gene name/symbol plus “expression”+ “transcription”+ other ontology terms. The co-usage of terms is restricted to same sentence.

Search returns are presented as marked-up text with matched terms highlighted in different colors and a link is provided to both the Xenbase abstract of the paper and to the source paper at the publisher’s web site. Xenbase curators have taken advantage of the in-context marked up text output to quickly populate new database tables such as our new morpholino feature. By searching for the keywords “morpholinos” and the common abbreviation “MO” we were able to extract >1500 morpholino names and sequences without having to separately download and open each publication. This is just one example of how curators use textpresso to improve the quality and efficiency of their work. Others include researching new terms for the XAO, finding additional references for antibody reagents, and identifying new synonyms for existing gene symbols.

Curation of gene nomenclature and gene models

Virtually all studies of gene expression and function require a complete genome with well-annotated gene models. This is also critical for linking *Xenopus* data to humans and other organisms. To consolidate these tasks, Xenbase administers the official *Xenopus* gene nomenclature, which is based on human gene names approved by the Human Gene Nomenclature Committee (HGNC, Gray et al. 2015) and collaborates with several groups to display and annotate the most recent gene models, as described below.

Xenopus to human gene mapping is automated based on genomic synteny, BLAST and symbol matching from previous assemblies. In the last five years, we assembled a *Xenopus* Gene Nomenclature Committee (XGNC) consisting of 16 experts from around the world. When *Xenopus* genes are incorrectly named (e.g. many of the tubulin, P450 and odorant receptor genes) or when an investigator characterizes a previously unnamed gene, curators identify the correct *Xenopus*-Human orthology and ensure that the proposed name/symbol is consistent with nomenclature guidelines, and not already used by the HGNC. The results are forwarded to the XGNC and HGNC for approval and the new name is registered in the HGCN gene catalog. When the official gene name changes in Xenbase, this cascades to NCBI, Entrez Gene, Ensembl and UniProtKB through automated weekly data exchanges. In a reciprocal process, monthly reports compare Xenbase and human gene names at Entrez Gene. When the human gene names change due to updates by the HGNC, our system updates *Xenopus* gene names, making the outdated name a synonym. We curate and maintain an extensive list of gene synonyms, which are critical to link data from older papers to current gene names. We also manually record nomenclature changes on the gene page Wiki. For example, prior to the official nomenclature the T-box gene *vegT*, was also called *brat*, *xombi* and *antipodean* by different groups. The synonym table allows searches with any of these legacy names to return *vegT* data.

Xenbase also curates gene models. Consistent with other genome projects 5–15% of computer generated gene models in *Xenopus* are predicted to be inaccurate, incomplete or misnamed. Xenbase curators manually annotate *Xenopus* genes, although we do not have the personnel to manually validate all the gene names in the multiple genome builds from *X. tropicalis* and both *X. laevis* sub-genomes and thus focus on 1) corrections as they arise, 2) the 1000 top genes that researchers frequently query, to ensure that these models are accurate; and 3) examine genes that have provisionally machine annotated names. The *Xenopus* Genome Consortium (XGC) is producing new chromosome level assemblies of the *X. tropicalis* and *X. laevis* genomes in the near future. Although the v8.0 assemblies are greatly improved, independent assessments from several XGC laboratories indicate that 5–10% of the gene models are incorrect or incomplete, and another 5–10% are incorrectly named (unpublished data). This is particularly problematic with the tetraploid *X. laevis* where the homeologous chromosome pairs, now to be designated L (long) and S (short) instead of “a” and “b”, can be difficult to resolve (Matsuda et al. 2015).

To begin this process we collaborate with the *Xenopus* Genome Consortium and the JGI, who identified 161 problematic gene models in the new *X. laevis* assembly v7.2 and we implemented the gene annotation software WebApollo (Lee et al. 2013). Curators examine each gene model and the surrounding genomic locus, using synteny with other species, previous assemblies, and newly available RNA-Seq and epigenetic data. Curators correct the gene models using the editing tools in WebApollo and send them back to the JGI for integration into the current assembly. Typical problems involve; 1) models split across scaffolds, 2) models split due to RefSeq mRNAs that are actually not full length, 3) splice acceptor and donor sites or ATGs being disrupted by gaps in the assembly, 4) *X. laevis* homeologs inappropriately labeled with the same -L or -S designation, and 5) small scale translocations between L and S chromosomes (or due to misassembly). Based on the

iterative manual curation needed for the human and mouse genomes, currently in v38, we anticipate that this will be an important ongoing task for new *Xenopus* genome assemblies several of which will be generated in the coming years.

Submitting data to Xenbase

Xenbase enthusiastically welcomes submission of data from the community. This can be in the form of new data produced by laboratories, or as amendments or corrections to existing content. Users can contact Xenbase developers using the “Contact us” link in the top right corner of every page (e.g. Figure 2) for information at any time. If users want to upload data two approaches are available. The first is via a web form. This can be accessed from both the top and bottom navigation bars using the “Submit your data” links. Using this tool users can submit up to 500 MB of data along with details so that curators can store or integrate the submission as appropriate. For larger scale submissions FTP transfer instructions will be provided, and these can be obtained using the “Contact Us” link.

For improvements or corrections to gene models users can use the above methods or they can load edited models directly into Gbrowse using the “Custom Tracks” feature. Once loaded the tracks can be classed as “Public” by the submitter, and when curators are notified of the addition staff will validate the model and modify the appropriate parts of the database. If this option is used, we also request a fasta file of the new model to add to our BLAST database and contact staff so we are aware of the upload.

Data Exports

Every week Xenbase generates a set of reports, or data dumps, containing our most up-to-date content. Almost all of the content described above that is in tables accessible to queries is exported to our ftp site, where it is available for download. Gene IDs are mapped to symbols, synonyms, nucleotide sequences, Affymetrix IDs, XAO terms describing gene expression patterns etc. In order to avoid generating reports whose size makes them difficult to handle, content is broken up into multiple reports. All reports map to Xenbase IDs, so users who wish to use data can always extract content from multiple tables for analysis. This is a rich data source that can be used to analyze RNA- and ChIP-seq data, microarray results, and any project that wants to correlate their results with our annotations or data sets. If users need data that are not in reports, custom reports may be requested and will be added to our weekly exports.

Conclusions

Xenbase continues to be an essential resource for *Xenopus* research and it will be necessary to adapt and grow in order to curate genome scale datasets and the increased use of *Xenopus* for human disease modeling. Xenbase will expand its data content and functionality while continuing to integrating the wealth of genomic and expression data already in Xenbase with new phenotype annotations and data from cutting edge technologies including NGS, genome editing, quantitative proteomics and high resolution imaging, all of which are increasingly being used by researchers utilizing *Xenopus*. In the past decade Xenbase has matured into an essential resource that significantly impacts most *Xenopus* research from experimental

design to data analysis and dissemination: practically all research with *Xenopus* utilizes Xenbase in some way.

Acknowledgments

Grant support: NIH Eunice Kennedy Shriver National Institute of Child Health and Human Development [P41 HD064556]

Xenbase core funding has been provided by the National Institute for Child Health and Human Development through R01HD45776 and P41 HD064556. Additional support has been provided via subcontracts from the European Xenopus Resource center via grants from the Wellcome Trust and BBSRC. We would like to thank all past developers and curators whose enormous contributions still run through our code. We would also like to thank present and past members of our external advisory board for their advice and guidance.

References

- Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.* 2009; 37:D793–796. [PubMed: 18842627]
- Bell GW, Yatskevych TA, Antin PB. GEISHA, a whole-mount *in situ* hybridization gene expression screen in chicken embryos. *Dev Dyn.* 2004; 229:677–687. [PubMed: 14991723]
- Bowes JB, Snyder KA, Segerdell E, Jarabek CJ, Azam K, Zorn AM, Vize PD. Xenbase: Gene Expression and Improved Integration. *Nucleic Acids Res.* 2010; 38:D607–612. [PubMed: 19884130]
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics.* 2009; 10:421–2105-10-421. [PubMed: 20003500]
- Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE. Mouse Genome Database Group. The Mouse Genome Database (MGD): Facilitating mouse as a model for human biology and disease. *Nucleic Acids Res.* 2015; 43:D726–736. [PubMed: 25348401]
- Gilchrist MJ, Pollet N. Databases of gene expression in *Xenopus* development. *Methods Mol Biol.* 2012; 917:319–345. [PubMed: 22956097]
- Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.Org: The HGNC resources in 2015. *Nucleic Acids Res.* 2015; 43:D1079–1085. [PubMed: 25361968]
- Karpinka JB, Fortriede JD, Burns KA, James-Zorn C, Ponferrada VG, Lee J, Karimi K, Zorn AM, Vize PD. Xenbase, the *Xenopus* model organism database; new virtualized system, data types and genomes. *Nucleic Acids Res.* 2015; 43:D756–763. [PubMed: 25313157]
- Khokha MK, Chung C, Bustamante EL, Gaw LW, Trott KA, Yeh J, Lim N, Lin JC, Taverner N, Amaya E, Papalopulu N, Smith JC, Zorn AM, Harland RM, Grammer TC. Techniques and probes for the study of *Xenopus tropicalis* development. *Dev Dyn.* 2002; 225:499–510. [PubMed: 12454926]
- Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH, Elisk CG, Lewis SE. Web Apollo: A web-based genomic annotation editing platform. *Genome Biol.* 2013; 14:R93-2013-14-8-r93. [PubMed: 24000942]
- Matsuda Y, Uno Y, Kondo M, Gilchrist MJ, Zorn AM, Rokhsar DS, Schmid M, Taira M. A new nomenclature of *Xenopus laevis* chromosomes based on the phylogenetic relationship to *Silurana/Xenopus tropicalis*. *Cytogenet Genome Res.* 2015 in press.
- Meehan TF, Masci AM, Abdulla A, Cowell LG, Blake JA, Mungall CJ, Diehl AD. Logical development of the cell ontology. *BMC Bioinformatics.* 2011; 12:6. [PubMed: 21208450]
- Muller HM, Kenny EE, Sternberg PW. Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.* 2004; 2:e309. [PubMed: 15383839]
- Parain K, Mazurier N, Bronchain O, Borday C, Cabochette P, Chesneau A, Colozza G, El Yakoubi W, Hamdache J, Locker M, Gilchrist MJ, Pollet N, Perron M. A large scale screen for neural stem cell markers in *xenopus* retina. *Dev Neurobiol.* 2012; 72:491–506. [PubMed: 22275214]
- Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Krasnov

- S, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Karsch-Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, Wilbur WJ, Yaschenko E, Ye J. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2012; 40:D13–25. [PubMed: 22140104]
- Segerdell E, Ponferrada VG, James-Zorn C, Burns KA, Fortriede JD, Dahdul WM, Vize PD, Zorn AM. Enhanced XAO: The ontology of *Xenopus* anatomy and development underpins more accurate annotation of gene expression and queries on Xenbase. *J Biomed Semantics.* 2013; 4:31-1480-4-31. [PubMed: 24139024]
- Sprague J, Doerry E, Douglas S, Westerfield M. The Zebrafish Information Network (ZFIN): A resource for genetic, genomic and developmental research. *Nucleic Acids Res.* 2001; 29:87–90. [PubMed: 11125057]
- Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S. The Generic Genome Browser: a building block for a model organism system database. *Genome Res.* 2002; 12:1599–1610. [PubMed: 12368253]
- van Heeringen SJ, Akkers RC, van Kruijsbergen I, Arif MA, Hanssen LL, Sharifi N, Veenstra GJ. Principles of nucleation of H3K27 methylation during embryonic development. *Genome Res.* 2014; 24:401–410. [PubMed: 24336765]
- Vize PD, Yu L, Karimi K. Database and informatic challenges in representing both diploid and tetraploid *Xenopus* species in Xenbase. *Cytogenet Genome Res.* 2015 in press.
- Westesson O, Skinner M, Holmes I. Visualizing next-generation sequencing data with JBrowse. *Brief Bioinform.* 2013; 14:172–177. [PubMed: 22411711]
- Yanai I, Peshkin L, Jorgensen P, Kirschner MW. Mapping gene expression in two *Xenopus* species: evolutionary constraints and developmental flexibility. *Dev Cell.* 2011; 20:483–496. [PubMed: 21497761]

Figure 1. Xenbase content. Schematic representation of the major data types stored within Xenbase, and their relationships to each other.

version: [3.5]

A Xenopus laevis and tropicalis resource
©Xenbase 2005-2015Major funding for Xenbase is provided by the National
Institute of Child Health and Human Development,
grant P41 HD064556 **Figure 3.**

Xenbase GBrowse. The Xenbase browser can load three different builds of both *X. tropicalis* (4.1, 7.1 and 8.0) and *X. laevis* (6.0, 7.2 [both J strain] and WT1.0). Multiple gene models, their annotations, and a variety of epigenomic and ancillary data are available for most genome builds. Many tracks launch pop-up windows with further information. In this view a user has clicked on the morpholino feature labelled “cyp26a1 MO1”. This reagent was designed against *X. laevis* cyp26a, so a number of mismatches are present in the alignment to the active view, which is *X. tropicalis* build 7.1. The stacked ChIP-seq tracks

for p300 are displayed using Topoview (http://flybase.org/static_pages/docs/software/topoview.html).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Curation pipeline. The curation pipelines incorporate an integrated set of automatic, semi-automatic and manual processes, where as much data as possible from each publication is curated. A final step is when the paper is entered into the “Xenbase Article Curation Tracker” (XACT) an automatic set of Ruby and Perl scripts that generate a Google fusion table containing all the information on papers imported each week. Curators access this table to monitor the status, content and priority of all papers in the database from the time of its first entry to its complete curation.