OXFORD GENETICS

GENETICS, 2025, iyaf237
https://doi.org/10.1093/genetics/iyaf237
Advance Access Publication Date: 3 November 2025
Knowledgebase and Database Resources

Xenbase: 25 years of integrating molecular and biomedical data from *Xenopus*

Stanley Chu,[1] Andrew J. Bell ,[2] Vaneet Lotay,[1] Ngoc Ly,[2] Troy J. Pells,[1] Taejoon Kwon ,[3] Sergei Agalakov,[1] Virgilio Ponferrada,[2] Courtney Lenz,[1] Christina James-Zorn ,[2] Brad Arshinoff,[1] Erik Segerdell,[2] DongZhuo Wang,[1] Konrad Thorner,[2] James D. Wasmuth,[4] Malcolm Fisher ,[2] Kamran Karimi ,[1] Aaron M. Zorn,[2,5] Peter D. Vize [1,*]

[1]Department of Biological Sciences, University of Calgary, Calgary, Alberta T2N1N4, Canada
[2]Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, United States
[3]Department of Biomedical Engineering, Ulsan National Institute of Science and Technology, Ulsan 44919, Republic of Korea
[4]Faculty of Veterinary Medicine, University of Calgary, Calgary, Alberta T2N1N4, Canada
[5]College of Medicine, University of Cincinnati, Cincinnati, OH 45229, United States

*Corresponding author: Department of Biological Sciences, University of Calgary, Calgary, Alberta T2N1N4, Canada. Email: pvize@ucalgary.ca

The *Xenopus* model organism knowledgebase, Xenbase (www.xenbase.org), bridges a wide variety of data types including genomes, anatomy, phenotypes, proteins, diseases and more. The goal of Xenbase is to support *Xenopus* molecular, cell and developmental biology research, to make these data available to the broader biomedical ecosystem, and accelerate the translation of *Xenopus* research into knowledge that will improve human health. Connections are made between data through relationships in our core data model and via a series of ontologies that serve as graph-based maps that can be traversed in various dimensions to find connections within our vast corpus of data. Data is input by a team of expert curators applying FAIR data management principles and also via automated pipelines and data processing routines. While our main focus is embryonic development and cell biology, these are often the underlying causes of compromised human health and are therefore invaluable for exploring the medical impacts of DNA sequence variants identified through patient exome or whole genome sequencing. One of the foundational elements in Xenbase with our gene-centric data structure is genomes, and we have recently vastly improved the quality of these core resources for both *Xenopus laevis* and *Xenopus tropicalis*. These and an extensive suite of other improvements are described, including updates and upgrades in content types, software and systems.

Keywords: knowledgebase; database; animal models; *Xenopus*; genomics; FAIR data

Introduction

Model organism knowledgebases (MOKs) such as Xenbase, RRID: SRC_003280 (Fisher et al. 2023), provide a pivotal step in all modern laboratory research. As most experiments are exploring the function of either genes, or cellular processes driven by gene products, researchers need high quality genomes combined with the available published data on the genomic regions of interest to design their experiments and to understand the outputs generated. Examples of data attached to a gene and its protein include when and where it is expressed, its molecular function, how it is regulated in various scenarios, publications on experiments involving the gene, and what phenotypes and diseases it is associated with. As different model organisms have distinct experimental approaches to which they are best suited, results from other systems must also be easy to trace, for example from genetic screens in the fruit fly, or a double-mutant phenotype in yeast compared to an mRNA microinjection in a frog embryo. Xenbase achieves this in many ways, but in part by integrating our *Xenopus* derived content into the Alliance of Genome Resources and by reciprocal links with other organism specific MOKs. An MOK serves as the bridge between these data and also serves as a community hub for researchers using the system it serves— in our example the system is the amphibian models *Xenopus laevis* and *Xenopus tropicalis*. These two frogs have different but overlapping experimental advantages; *X. laevis* is optimal for microinjection, tissue recombination, cell biology and any other method easier to perform in a physically large embryo, and *X. tropicalis* optimal for genetics, transgenics and any experiment requiring a short generation time or diploid genome (Harland and Grainger 2011; Philpott 2021).

Xenbase began as a simple HTML based website focused on images showing gene expression patterns from *in situ* hybridization, then developed more complex features such as JavaScript driven visualization tools and local BLAST search capabilities, rather than linking to external resources at NCBI. At the time moving to our own BLAST service cut search times from hours (or even overnight) to seconds. We then moved to a basic relational database backend accessed via a Java web application. From the original small database of community member names, addresses

Received on 05 September 2025; accepted on 20 October 2025

and images to the modern suite of virtual machines, cloud resources and the current vast 200 GB database of many genomes, hundreds of thousands of images and complex data processing pipelines and systems took decades of development by software developers and data curators. As we built new features and added new data types we leveraged the latest technologies, approaches and practices developed by other MOKs and the newest data standards, including implementing FAIR management principles to make data findable, accessible, interoperable, and reusable (Wilkinson et al. 2016). A major evolution in our resource was the addition of an expert curation team that enhances the value of content through manual annotation and data integration (James-Zorn et al. 2013). With complex data sets, such as a publication describing experiments using genome editing to generate a developmental phenotype, or the jumble of files that make up an RNA-seq dataset in GEO, automated methods cannot yet accurately process and sort the rich and invaluable content into a database. These data require sophisticated curation interfaces and a highly trained team to work effectively. Our present system is a combination of automatic pipelines and data processing routines including curation via term matching and regular expressions, semi-automatic systems that perform some steps that are supplemented with manual finishing, and purely manual processes.

Achievements at Xenbase over the past two years include a major software migration including database and web application server, large scale genome annotation improvements, increased use of machine learning for article filtering (Karimi et al. 2021), reworked community pages (persons, labs, organizations, jobs), synteny visualization systems, greatly improved integration into the Alliance, enriched GO annotation (including ortholog GO) and remapping core data sets to the new v10 genomes.

Genome improvement

One of the major Xenbase achievements of the past two years in vastly improved gene annotation in our two core genomes, specifically improving gene names, symbols and characterizing previously uncharacterized genes/proteins. 93% of *X. tropicalis* gene models are now annotated with at least an informative symbol or name, with 76% being fully annotated for both (here we use the term "gene annotation" as a process to assign gene name and/or gene symbol to the entity on the genome to infer function, not as a process to define the gene structure, such as exon-intron boundaries, on the genome). Details on the programmatic and manual methods used to achieve this will be published elsewhere (Kwon *et al.*, in prep). As our core genome resources function as the anchor points of so many data types in Xenbase, the improvements radiate throughout our resource, so a brief description is included here. The *X. tropicalis* v10 (Bredeson et al. 2024) annotation gave 14,877 models gene names and symbols, with 1,168 partially annotated and 5,887 with no useful information. The 2025 version generated by Xenbase improved this to 16,517 fully annotated (76%) and 5,315 partially annotated (24%) protein coding genes. The improvements to the *X. laevis* v10 annotation (GCA_017654675.1) are ongoing and more significant, increasing from 21,661 to 25,753 fully annotated models, though this does leave 22.6% presently unannotated. We will continue to further improve the naming and characterization of *Xenopus* genes in the next year. Updated gene model annotations are available live on gene pages as they are entered into the database, and are also used to update our genome resource GFF3 and GTF files periodically so can also be viewed in JBrowse and are present in the resource files downloaded and used for local bioinformatics analyses. Changes to the GFF3 and GTF resource files are tracked and versioned on our downloads server and updates are performed quarterly, so there is minor lag in new annotations appearing on JBrowse and JBrowse2.

As we administer gene nomenclature on behalf of the *Xenopus* research community, gene symbols and names applied in the annotation process are the official versions. Our naming methods are performed in collaboration with the HGNC (McCarthy et al. 2023) and NCBI RefSeq (Goldfarb et al. 2025). Systems are in place to keep Xenbase and NCBI symbols and names synchronized and to avoid feedback errors being introduced. Changes to *Xenopus* gene names are recorded on the Xenbase Gene Page wiki. Nomenclature guidelines are provided to aid community members in recommending new gene names.

Xenbase uses Gene Pages starting with genome-based gene models (Karimi et al. 2018) as a foundation to bridge many different types of data, from DNA sequences and gene expression patterns to mutant phenotypes to human disease (Nenni et al. 2019). The improved genome annotation has resulted in the generation of over 6,400 new Gene Pages, in addition to updating and improving orthology annotations on thousands more. Each of these new pages (and page updates) serves as a branch point pulling in a vast set of additional collected data, including publications, protein interactants, protein domain structures and functional predictions, RNA-seq experiments with data on the gene's differential expression, GO terms and more, so adding 6,400 new Gene Pages radiates into a massive level of improved data connectivity. Many of these connections, for example between a gene and a cell type in which it is expressed, can further radiate outwards via our core ontologies that link (in this example) a cell type to a detailed graph-based map of all cells, tissues and systems within our study organism.

RNA-Seq and ChIP-seq data

The Xenbase RNA-seq and ChIP-seq visualization system is one of the most advanced available. Every *Xenopus* RNA-seq/ChIP-seq data set in GEO (that passes our QC) is both machine and manually curated, then aligned against the corresponding genome and processed through a series of bioinformatic pipelines. Details on the methods have been published elsewhere (Fortriede et al. 2020). The results can be viewed using a variety of custom built tools, from JBrowse tracks to heatmaps, hive plots and data tables. The normalized read-count tables are stored directly in our database and their content can be processed and displayed using SQL based methods within Xenbase, or downloaded for custom local analyses. For ChIP-seq both wiggle tracks and peak-calls are displayed. We recently reprocessed this massive corpus of raw sequence data, realigning the original reads from each GEO set against the new v10 genomes and reanalyzing the output, see Table 1 for counts of major datasets processed for the most common methods. Smaller numbers of tracks generated by less common methods, for example CLIP-seq and MNase-seq, are also available.

Table 1. Examples of tracks generated from GEO data sets that support the v10 genomes in Xenbase.

Technology	X. laevis	X. tropicalis
ATAC-seq	2	25
Bisulfite-seq		2
RIP-seq	7	
RNA-seq	513	605
WTTS-seq		13

Users can now view all these data against v10 genomes in JBrowse. While we do maintain the results generated against v9 as both JBrowse tracks and the original data files (e.g. normalized read-count tables, peak calls etc.) on our download server, results that rely on database table querying will now all show v10 results. In addition to tracks in JBrowse, we provide other tools to interrogate our processed GEO data, including table views and various JavaScript driven data visualization systems, such as heatmaps and hive plots (Fortriede et al. 2020). Additional visualization systems will be added to maximize the new sequencing method data, in addition to JBrowse tracks, as development time allows.

Genome browser resources

Our core goal is to support research using *Xenopus laevis* and *Xenopus tropicalis*, and to this end tmany new tracks and features have been added to the browsers supporting genomes of these two species in addition to the GEO processed data described in Table 1. Our core browser remains JBrowse (Skinner et al. 2009), but we have also released a beta version of JBrowse2 (Diesh et al. 2023) with a limited set of supporting tracks and will soon have all v10 genomes and their supporting data available on this system. While the new features of JBrowse2 are powerful adjuncts to *Xenopus* research, it takes a considerable time to get used to mining data in a new browser, consequently we will also maintain the legacy JBrowse tools and tracks for as long as this is technically possible.

One feature we work on constantly, but not always on a resource user's radar, is improving the richness of the "About this track" information available on every track. This is accessed once a track has been loaded using the drop-down symbol at the end of the track name. It opens to display as many resources relevant to the track as possible, including links to its source, links to publications, and paths to access other pertinent resources at Xenbase. The same dropdown that brings up this invaluable information also includes other useful options users may be unaware of, such as the ability to change display characteristics, which is particularly useful when configuring a browser display for use in a publication figure.

Many of the tracks we support in JBrowse are also made available to load into a UCSC browser track hub, including the v10 genomes. Instructions for how to view the Xenbase genome tracks in UCSC are provided under the main horizontal navigation bar on every Xenbase page under "Genomes". The tracks converted to work in UCSC vary with genome versions, and if specific tracks are not present, users are encouraged to directly request them via the "Contact Us" link in the top right corner of every page.

The move to v10 genomes as our prime data sources also required reprocessing of reagent tracks that use alignment as a visualization method, for example morpholinos (>1,000) and guide RNAs (>500). Our web application generates and integrates alignment results into the database as each reagent is added, so reprocessing these as a batch required novel methods. Updated reagent mappings can be viewed both in the reagent resource page, and in a GFF3 track on each genome.

Xenbase computational resources are also made available to other amphibian systems that do not have an MOK including basic browser support for their genomes. We host such genomes on JBrowse and also provide a BLAST database as a conduit to genomic regions, as especially in less well annotated genomes alignment is often the most reliable method to identify sought genes. Support is currently available for the urodele *Ambystoma mexicanum* v6.0 and anurans *Nanorana parkeri* v1.0, *Rana catesbeiana* v2.0, and *Hymenochirus boettgeri* 1.0.

Synteny

While synteny between *Xenopus* chromosomes and those of other model systems is important for establishing orthology, our target *Xenopus* species also have a unique synteny situation with *X. tropicalis* being a diploid and *X. laevis* an allotetraploid. Within the *Xenopus laevis* allotetraploid genome two distinct sets of chromosomes reside, each from a distinct ancestral species, and each subgenome evolves through interactions with the other (Session et al. 2016). Often this results in downregulating, and sometimes inactivating or deleting, one of the two alloalleles. One ancestral chromosome set (and the genes within it) is labeled with an "L" suffix, and the other with an "S", indicative of the longer chromosome lengths within the L ancestral species relative to the shorter chromosome set from the S ancestor (Session et al. 2016). Interestingly, *X. tropicalis* is predicted to have evolved from a common ancestor at around the same time as the genome fusion event that led to the allotetraploid genome of *X. laevis* so the L and S alloalleles and *X. tropicalis* genes have similar levels of conservation to each other. This complex genomic history means being able to view syntenic relationships between the two *X. laevis* subgenomes and the diploid *X. tropicalis* genome is a unique and important feature for our user community and experiment planning. Other synteny resources do not (to our knowledge) deal with the subgenome issues.

Two methods have been developed to display synteny data, both between the subgenomes and the diploid genome, and in one example between these and other model systems (human, mouse and chick). The first is a graphical tool that displays genes as colored arrows and compares their order and orientation on chromosomes, based on the SimpleSynteny tool (Veltri et al. 2016). This pipeline runs on a backend server on demand using the GFF files to identify the gene of interest and its adjacent cohort, and generates the graphical output that is returned to the web application and the user. Orthology calls for this tool are based on existing orthology assertions captured in Xenbase or direct symbol matching. In the example shown in Fig. 1, illustrating the gene layout surrounding the *shh* gene, where the L subgenome in *X. laevis* and the *X. tropicalis* chromosome have the same organization as human, while the *X. laevis* S subgenome has undergone a rearrangement. For more detailed comparisons between conserved blocks across species including non-coding alignments, we also provide deep aligned genome comparisons from the NCBI Comparative Genome Viewer (CGV) (Rangwala et al. 2024) displayed in JBrowse2 (Diesh et al. 2023) on Xenbase. To display this data we obtained pairwise genome alignments from the NCBI CGV in GFF format. Those data organize local alignments into longer syntenic blocks to produce the best-scoring alignment. The GFF alignments from CGV were converted by Xenbase into PAF format to enable synteny visualization in our local JBrowse2 (Fig. 2). In order to use this method for a specific gene users need to launch JBrowse2 under "Genomes" in horizontal navigation bar, then select "Synteny" from the JBrowse2 blue menu bar, followed by the option "Open orthologs in multi-way synteny viewer" then input a gene symbol.

UniProt

A large-scale harmonization effort addressed long-standing issues with the UniProt reference resource for *Xenopus* proteins (UniProt Consortium 2023). These issues arose through a number of factors, such as the lack of, or problems with, ENSEMBL annotations of the *Xenopus* genomes and also the accumulation of

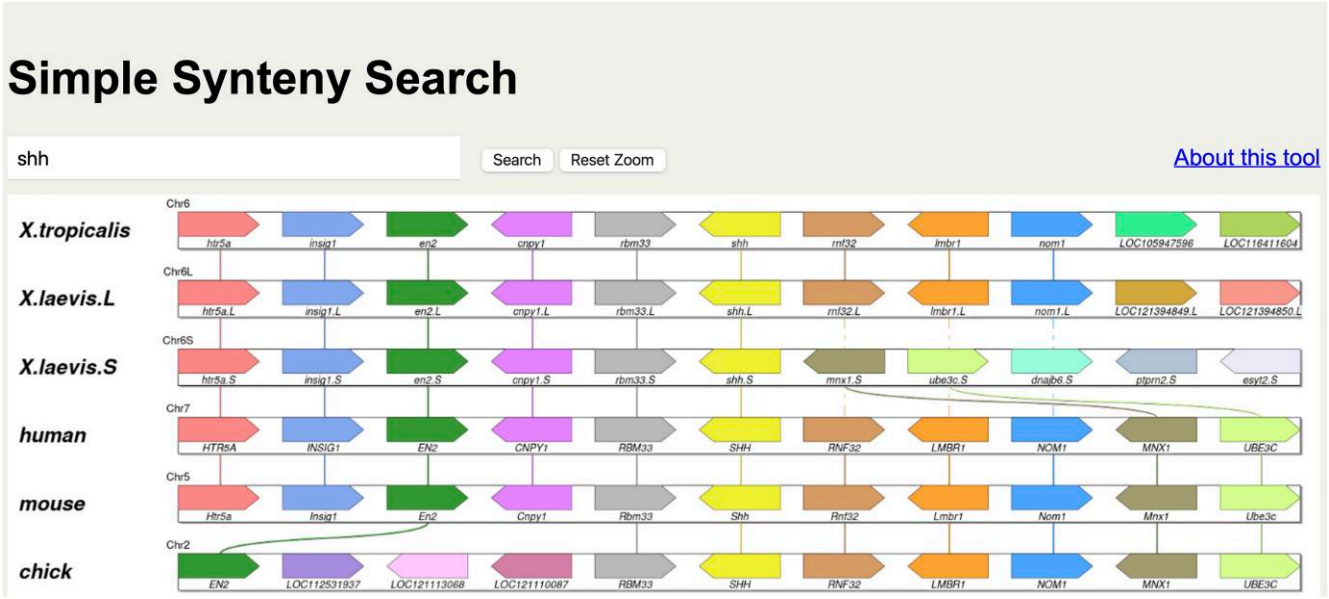

Fig. 1. Simple synteny screenshot, based on the *shh* gene. This visualization is available under the Genomes dropdown menu, and also on every individual Gene Page.

layers of data within UniProt over time. This work was carried out as a collaboration between Xenbase and UniProt. It was initiated by negotiating with UniProt to rerun their annotation systems against NCBI reference sets, rather than ENSEMBL (their standard approach) for *Xenopus* proteins. Xenbase then performed systematic analyses and cleanups of the output, removing redundancies where multiple identical sequences were represented with separate IDs, and updating IDs to the latest RefSeq and newer genome assemblies. The results for the new *X. tropicalis* reference protein set were then shared with, and deployed by, UniProt, and will be kept synchronized in an ongoing manner. A related project is underway to similarly address the *X. laevis* reference protein set in parallel with our genome improvement efforts. As UniProt reference sets are used to seed many other external systems, for example the Quest for Orthologs project (Altenhoff et al. 2024), these results will filter out throughout the broader bioinformatics and biomedical communities.

GO

Xenbase performs manual GO annotation on published papers based on experiments using *Xenopus*, and submits the results to the GO Consortium through the NOCTUA curation tools at GO Central (Gene Ontology Consortium 2023). As *Xenopus* genes and proteins have relatively sparse curated GO coverage compared to other systems, this work plays an important role in improving the value of *Xenopus* molecular and experimental data and leveraging its importance to understanding foundational elements of human health and disease. There are around 5,000 experimentally supported GO annotations for *Xenopus* genes, compared to 494,000 for human. While we support the available *Xenopus* GO terms via a tab on every Gene Page in Xenbase and in our data exports, due to this sparseness, and the essential role GO annotations play in many forms of data analysis, we recently added the GO ribbon widget (Fig. 3).

This tool from the GO consortium pulls the richer content available for orthologous genes to enhance our GO term representation while we work on improving them manually. As different model systems use different experimental techniques, the results between species will always differ, and be enriched by inclusion of orthologous data. The GO consortium (2023) ribbon is similar to that deployed on Alliance Gene Pages, and is once again available on Xenbase Gene Page GO tabs. We now display GO terms from a suite of vertebrate (human, mouse, rat and zebrafish) and one invertebrate (*Drosophila*) model organisms. The ribbon retrieves annotations via the GO Application Program Interface (API) so provides the latest available information. Annotations are grouped into high level GO terms (GO subsets/slims), and then further divided into three categories: molecular function, biological process, and cellular component. A color gradient from dark green to white is used to indicate the number of annotations available for a given subset in a heatmap display. When a cell in the ribbon is selected, more detailed information about the annotations in that subset are also retrieved from the GO API, and then displayed in a table below the ribbon. This now typically includes many hundreds of GO terms associated with each gene, a vast improvement. In the example shown in Fig. 3 for the *shh* gene the number of ortholog GO terms is increased from 34 to 360. We have also enhanced our GO term GAF files provided for user downloads by incorporating ortholog derived data for human, mouse, rat, chicken, zebrafish and *Drosophila* as ISO annotations. To generate these annotations, ortholog UniProt accession numbers were mapped to Xenbase gene IDs. Only annotations supported by experimental evidence codes (EXP, IDA, IPI, IMP, IGI, IEP), or selected non-experimental codes—ISS (sequence similarity), TAS (traceable author statement), and IC (inferred by curator)—were included in the mapping process. The corresponding ortholog UniProt accessions were recorded in the with/from field, and these annotations were collapsed to a single ISO entry for each unique gene + GO term pair, with all identified ortholog UniProts listed in a single string.

Ontologies

A number of in-house generated and external ontologies bridge data within Xenbase, in addition to the GO ontology just discussed.

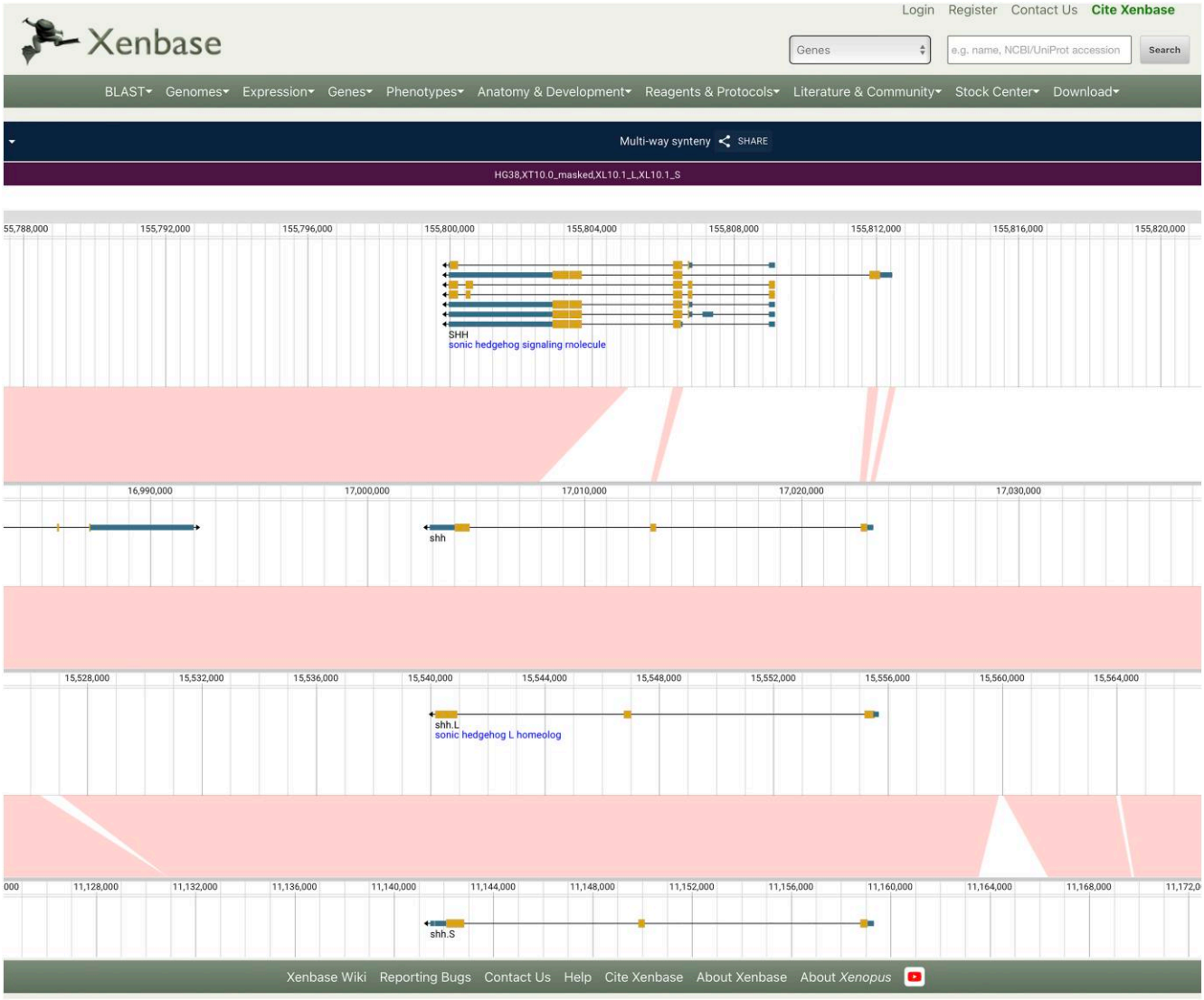

Fig. 2. Deep synteny generated by pairwise alignments in CGV, once again based on the *shh* gene. These data are displayed within JBrowse2.

Xenbase-generated ontologies include the *Xenopus* Anatomy Ontology, the XAO (Segerdell et al. 2013), the *Xenopus* Phenotype Ontology, the XPO (Fisher et al. 2022) and a number of smaller ontologies used to index data such as experimental methods and reagent types (XBED) and small molecules (XSMO). The most recent XAO release (v11) had 55 new terms and 1,830 total terms. A new release (v3) of the XPO has 873 new terms, with 22,276 phenotype classes in total. The XPO update includes 96 new terms for changes in the level of a chemical compound in an anatomical location, a class of phenotypes that has not appeared in the XPO previously. In addition to anatomical phenotypes Xenbase also curates changes in gene expression as a separate class of impacts on the embryo. For information on visualization of gene expression and expression as a phenotype see (Fisher et al. 2022). The XAO and XPO are dynamic and have Github term request methods for creating and modifying terms. As the XPO is built programmatically using design patterns, addition of a single term can result in a large number of changes to the system (Fisher et al. 2022). We also use several external ontologies, such as the Disease Ontology (Baron et al. 2024), Gene Ontology (Gene Ontology Consortium 2023), Sequence Ontology (Mungall et al. 2011) and Uberon (Mungall et al. 2012). Each ontology is applied to Xenbase data using a suite of tools, some script based that use ID or term matching and others via software interfaces that allow curators to apply the ontology to content within the database. Each ontology has overhead and requires regular updating and testing, maintenance at repositories etc.

The graph-based structure of the core ontologies allows queries to walk between data types that are not directly tagged with a term. For example the different anatomical elements of the eye are all linked by an XAO relationship term, so data associated with the lens of the eye can be included (or excluded) from data sets on the retina of the eye (Vize 2024). Another example would be genes in the same signaling pathway (via GO) expressed in the same developmental time window. Cross-references to other ontologies are also stored within each, so searches can work across many dimensions as they use relationships coded within each ontology to link data together.

Xenbase at the alliance of genome resources

Xenbase became a member of the Alliance of Genome Resources in 2022 and data on *Xenopus* became available as one of the core model systems represented in the Alliance portal late that year (Alliance of Genome Resources Consortium 2024). Both *X. tropicalis*

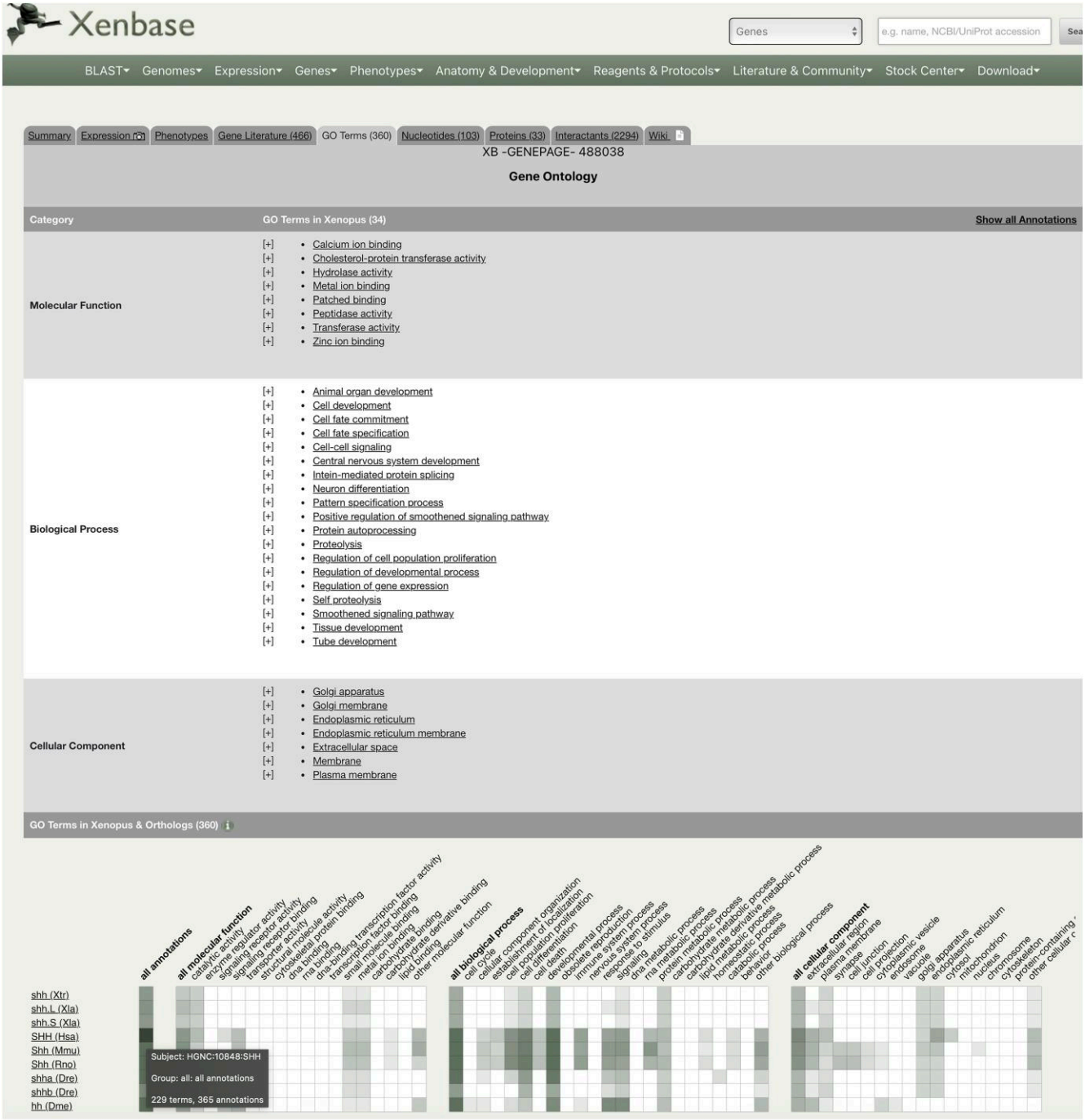

Fig. 3. Orthologous GO terms generated by the GO API (Gene Ontology Consortium 2023) for the *shh* gene displayed within Xenbase. The 34 *Xenopus* derived annotations are displayed in the top section, and the 360 total annotations in the lower heatmap.

and X. *laevis* data are included. While Xenbase has its own internal curated and stable orthology relationships, we had to get our system to work with that of the Alliance. The Alliance uses DIOPT (Hu et al. 2011) to establish orthology, which works effectively for the diploid *X. tropicalis* but not for the allotetraploid *X. laevis* due to the complexities just discussed in the above section on synteny. To resolve this situation the Alliance uses DIOPT to predict *X. tropicalis* orthology to genes in other represented organisms, and then pulls Xenbase derived *X. tropicalis* to *X. laevis* L and S gene relationships. The original Alliance data loads used the *X. tropicalis* v9 genome based DIOPT predictions, which only matched 9,220 Xenbase *X. tropicalis* genes and around 1.1× this many *X. laevis* genes. A new data load was performed in early 2025 using DIOPT predictions based on the v10 *X. tropicalis* genome, which more than doubled the number of matches to 20,569, of which ~16,000 mapped into the Alliance, and around 23,000 mapped to *X. laevis* genes via Xenbase relationships. This vastly improved integration means that not only is *Xenopus* data covering most of the genome deeply integrated into Alliance resources, the suite of third party resources that pull data from the Alliance also have access to our content, for example the Monarch Initiative (Putman et al. 2024).

The Alliance portal includes an InterMine instance (Kalderimis et al. 2014), AllianceMine, so sets of data can be analyzed against

the Alliance data corpus, custom queries or queries based on InterMine templates can be performed, and the InterMine API used to run queries directly from scripts in a variety of languages on *Xenopus* data (Alliance of Genome Resources Consortium 2024). Updates to *Xenopus* content at the Alliance is via a Xenbase private Java based API, the code for which is available upon request.

New gene page features

We constantly improve and update features on our Gene Pages, which serve both as a summary of the most commonly sought data and as a gateway to the vast amount of information associated with every gene. Most of the data is within Xenbase itself, but we also provide links to external resources and maintenance includes fixing links to these– as they change constantly, and also improving the quality of links as we can. New features include:

- improved visual cues on available tab content, such as GO term depth, wiki data entry etc.
- a link to the AlphaFold (Jumper et al. 2021) prediction of the protein 3D structure (uses the UniProt reference proteome as source)
- nucleotides tab reorganized and improved, added RefSeq sequences and updated sorting and labeling of tab contents
- updated suite of links to sequence variant and human phenotype databases
- links to clone and plasmid suppliers updated as the resources change (also applies to animal suppliers)
- Gene Page search refactored for speed, new search fields added, e.g. InterPro domains
- Incorporation of the GO orthologs widget
- link to synteny visualization tool
- protein-protein interactants tool updated to allow more levels of interaction and more complex networks

Data sharing

Xenbase both exports and imports data from a wide variety of resources. Our exports are either directly to our user base via a series of data reports that are updated weekly, or to informatics resources who collect our content, usually on a monthly basis. These external resources, for example NCBI and UniProt, typically import a Xenbase flat file and parse it into their system. These files are stored on the Xenbase download server and can be accessed at download.xenbase.org or via the "Download" link on the right-hand side of the navigation bar on every Xenbase page.

When data is exchanged bidirectionally with external resources, it is essential to establish formal agreements to mitigate the risk of feedback loops. Such loops may arise when data retrieved from an external source is curated at both ends—whether through manual annotation or *in silico* processing—particularly in cases where mappings, symbols, or identifiers are subject to revision. One strategy we employ to address this challenge is the designation of a single authoritative source for specific subsets of the data. While effective under certain conditions, this approach is often inadequate for complex datasets that integrate contributions from multiple stakeholders. A complementary strategy is the use of controlled, non-overlapping update cycles, which compartmentalize modifications and thereby prevent the emergence of feedback loops.

In addition to static files Xenbase also provides API access to some content. This is the method used by the Alliance of Genome Resources to pull *Xenopus* data from Xenbase for incorporation into their systems via a private API. Another set of private RESTful APIs serve to connect the main Java application to BLAST servers. This replaces a SOAP-driven mechanism, and has resulted in noticeable improvements in BLAST reliability and performance.

To support external resources wishing to link to Xenbase Gene Pages, an API call allows programmatic verification of the existence of a given Gene Page. Additionally, while data from our Gene Pages is exported to a series of weekly data reports (see download.xenbase.org) we also support programmatic download of individual gene page contents via an API that delivers such contents as JSON objects. More detailed information about these calls are available at https://xenbase.org/xenbase/static/apis-links.jsp

Infrastructure and systems

Both software and hardware support for Xenbase has been extensively upgraded. On the software side we have undergone our first large scale technology move from our decades long academic partnership with IBM and their DB2 database and WebSphere Application Server to open source alternatives PostgreSQL and Apache Tomcat. This was done as open source technologies have developed to the point where the required features essential to our ecosystem are now fully supported, and we can avoid both the resource intensity and licensing complexities of proprietary tools. The open source options are vastly easier to maintain and will reduce systems overhead in the long term. Both DB2 and PostgreSQL databases are relational and mostly use standard SQL, so the basic technology was similar. Likewise, WebSphere and Tomcat are both Java application servers, so once again no major shift in technology or language was required. Despite the fundamentally common technologies, each migration was a significant software engineering task, as many features built into Enterprise systems are not available in the open source options. There were other large scale tasks such as conversion from DB2 SQL dialect to PostgreSQL dialect, and rewriting database triggers and functions to the different embedded language. A significant percentage of the SQL queries had to be rewritten to work better with PostgreSQL optimizer. As a benefit this effort allowed us to utilize the trigram index available in PostgreSQL but not in DB2, and improved performance of several search queries by orders of magnitude.

With the more widespread use of bots to harvest data from the Internet, we noticed major "bot attacks" that would affect site performance for weeks at a time, and bloated download volumes and wasted computing resources. In response, we modified existing code to detect spurious data access and handle these without adding unnecessary load to the database. Realizing that harvesting web content is needed for many AI applications, we followed a balanced approach that does not reject bot access outright, but at the same time does not slow down the site for human users.

On the hardware side we acquired two Lenovo ThinkSystem SR635 V3 servers, each with 48 cores and 768 GB of RAM. Having two servers provides load balancing and fault tolerance, as explained in Karimi and Vize (2014). We have kept the older servers (installed in 2017) operational and use them to run lower priority tasks. Our storage system has been replaced with an all-SSD Lenovo DE4000F system with 90 TB of usable storage space, set up as RAID 6. All our virtual machines have been upgraded to the latest version of Ubuntu Linux.

In addition to the user interfaces we build to provide gateways for our community to query and browse content, we also write curator interfaces that provide our professional curatorial staff the tools they require to add value to content and build links between data objects. Such interfaces are not visible to visitors;

curators must login to the system using alternative secure systems. With each type of new content such tools must be generated, secure management established, and the entire suite of such code must be maintained. The first versions of such systems were quite simple, for example allowing curators to link a gene to a publication, but these rapidly progressed to tools loading ontologies and more complex options. In recent years we have produced a number of advanced curation interfaces allowing staff to efficiently search for, add and edit Xenbase data elements. This includes support for multi-tab editing e.g. curators can edit different transgenic/mutant lines in separate tabs at the same time. Interfaces for adding/editing users, researchers or contributors, laboratories, organizations, books, job postings have all been revamped so that they are more user friendly and reliable.

A new copyright statement system was created to enhance the accuracy of copyright statements for papers in the modern multi-step publishing cycle, which is now more complex and has changed with time as papers move through from early release to final and final to print versions. Each publisher's negotiated copyright statement is stored in the database and assigned to papers as part of the curation process. Some papers may have unique copyright statements and these are also accepted by the system. Historically loaded papers have been assigned copyright statements based on the declared publisher and copyright owner. New papers that have not had a copyright statement assigned are evaluated by curators and display an interim default statement until curation is complete.

Testing systems

Every time we commit new code to such a complex software ecosystem the possibility of conflicts with other elements of Xenbase arises. For this reason, we operate a tightly controlled software release and testing regimen. New releases of Xenbase are tested for defects prior to deployment on a dedicated test environment that is closely aligned with the production environment, including VMs running test-specific copies of the production machines, such as test-JBrowse, test-PostgreSQL and test-Tomcat virtual machines. Testing of new code is carried out using an open-source Python software tool called Behave that uses WebDrivers for popular browsers to run a suite of regression and integration tests across all Xenbase web application modules; this has replaced our older Ruby based Cucumber-Watir system. The tests emulate user and curator behavior in a browser to uncover new software bugs that may have been inadvertently introduced as a result of new development work. Tests are written and maintained by a quality assurance analyst in order to avoid developer bias. They are written to ensure specific data availability, correct logical processing, and proper output display across the site and utilize the Gherkin behavior-driven development language that allows new tests to be defined rapidly in a plain language syntax (Wynne 2017) and maintains continuity with our older Cucumber system. Regression tests uncover software bugs that have been inadvertently introduced as a result of new development work. Integration tests are similar but are designed to uncover problems occurring between software modules. Pre-deployment testing using Behave has been an effective way of avoiding defects from entering production. The automated testing suite is also supplemented by manual testing for changes outside the scope of the test suite.

Modernizing legacy modules in Xenbase

Some Xenbase code is 25 years old. Updating legacy modules and tools in Xenbase involves overhaul of both user interfaces and underlying code to address technical aspects and meet evolving user needs. Over the past two years, updates to modules like the community portal, transgenic lines interface, and expression search functionality have been driven by community feedback and insights from partner resources like the Alliance of Genome Resources and other MOKs, and technical necessity. The process typically begins with redesigning the user interface, creating mockups to visualize improved layouts and interactions that enhance usability and accessibility. On the backend, we update the database and data model by adding new database table rows or modifying existing ones to support new features or data types. The Java Server Pages (JSPs) and their associated SQL queries are rewritten to optimize performance and ensure compatibility with modern standards, often replacing outdated logic with streamlined implementations using tools like MyBatis for more efficient data access and mapping. Additionally, we incorporate modern frameworks and libraries to replace deprecated dependencies, ensuring scalability and maintainability while aligning with current best practices.

When cleaning up old code, we focus on removing or replacing inefficient, redundant, or obsolete components to improve reliability and maintainability. This includes eliminating hardcoded values and replacing them with configurable parameters, refactoring complex JSP logic into modular Java classes or services, and updating SQL queries to leverage indexed fields for faster performance. Outdated libraries or unsupported dependencies, such as those tied to 20-year-old frameworks, are replaced with modern alternatives, alongside MyBatis for streamlined database interactions. Code comments and documentation are updated to reflect changes, and automated tests are updated to include the new features to ensure stability. Security vulnerabilities, such as those from outdated input validation or session management, are addressed by implementing modern security practices. These updates reduce bugs, enhance performance, and create a codebase that is easier for future developers to maintain and extend, ultimately delivering a more robust and user-friendly experience for Xenbase users.

Future prospects

Xenbase has been running as a suite of virtual machines in a private cloud on our own servers since 2013 (Karimi and Vize 2014). Moving from a private resource to the public cloud has been one of our recent technology goals. Engineering a solution that maintains all of our features and performance in a cost effective manner has been challenging, but is a major goal in the immediate future.

New content goals include completing annotation of the *X. laevis* genome and full harmonization of NCBI, ENSEMBL and *Xenopus* Genome Consortium annotations, improving gene models using RNA-seq data, support for paralogs and non-coding RNA genes, deploying single-cell support, human variant support and testing AI tools for improving both code development and for data processing. Our partner resources are also developing enhanced interfaces to leverage data from model organisms and backend tools to improve data processing, and we will adopt such methods when they will support research using *Xenopus*.

Data availability

All Xenbase content is available freely at www.xenbase.org. Wherever possible no login is required. Whenever data is to be

entered by a user or curator login is of course essential. Our download site makes our core datasets and weekly exports of database content simple and free to access, and if users have difficulty locating content they are encouraged to request assistance using the "Contact Us" link near the top right corner of every page. Scripts, application code and APIs are available upon request to any non-commercial resource.

Acknowledgments

Many skilled developers and curators have driven development of Xenbase over the past 25 years, including Boluwatife Osifaluj, Sedat Demiriz, Vy Ngo, Nadia Bayyari, Nivitha Sundararaj, Eugene Kim, Praneet Chaturvedi, Mardi Nenni, Ying Wang, Elizabeth Wilke, Joshua Fortriede, Kevin A. Burns, Vicente Pader, Shraddha Kadam, Yu Liu, Brad Karpinka, Jacqueline Lee, Jeff Bowes, Chris Jarabek, Kevin Snyder, Lea Randall, Bishnu Bhattacharyya, Kenan Azam, Bianca Maters, Angela Fan, Ross Gibb and Etienne Noumen.

Funding

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under Award Number P41HD064556. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of interest

The authors declare no conflicts of interest.

Literature cited

Alliance of Genome Resources Consortium. 2024. Updates to the alliance of genome resources central infrastructure. Genetics. 227: iyae049. https://doi.org/10.1093/genetics/iyae049.

Altenhoff A *et al.* 2024. New developments for the quest for orthologs benchmark service. NAR Genom Bioinform. 6:lqae167. https://doi.org/10.1093/nargab/lqae167.

Baron JA *et al.* 2024. The DO-KB knowledgebase: a 20-year journey developing the disease open science ecosystem. Nucleic Acids Res. 52:D1305–D1314. https://doi.org/10.1093/nar/gkad1051.

Bredeson JV *et al.* 2024. Conserved chromatin and repetitive patterns reveal slow genome evolution in frogs. Nat Commun. 15:579. https://doi.org/10.1038/s41467-023-43012-9.

Diesh C *et al.* 2023. JBrowse 2: a modular genome browser with views of synteny and structural variation. Genome Biol. 24:74. https://doi.org/10.1186/s13059-023-02914-z.

Fisher M *et al.* 2023. Xenbase: key features and resources of the Xenopus model organism knowledgebase. Genetics. 224:iyad018. https://doi.org/10.1093/genetics/iyad018.

Fisher ME *et al.* 2022. The Xenopus phenotype ontology: bridging model organism phenotype data to human health and development. BMC Bioinformatics. 23:99. https://doi.org/10.1186/s12859-022-04636-8.

Fortriede JD *et al.* 2020. Xenbase: deep integration of GEO & SRA RNA-Seq and ChIP-seq data in a model organism database. Nucleic Acids Res. 48:D776–D782. https://doi.org/10.1093/nar/gkz933.

Gene Ontology Consortium. 2023. The gene ontology knowledgebase in 2023. Genetics. 224:iyad031. https://doi.org/10.1093/genetics/iyad031.

Goldfarb T *et al.* 2025. NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. Nucleic Acids Res. 53:D243–D257. https://doi.org/10.1093/nar/gkae1038.

Harland RM, Grainger RM. 2011. Xenopus research: metamorphosed by genetics and genomics. Trends Genet. 27:507–515. https://doi.org/10.1016/j.tig.2011.08.003.

Hu Y *et al.* 2011. An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics. 12:357. https://doi.org/10.1186/1471-2105-12-357.

James-Zorn C *et al.* 2013. Xenbase: expansion and updates of the Xenopus model organism database. Nucleic Acids Res. 41: D865–D870. https://doi.org/10.1093/nar/gks1025.

Jumper J *et al.* 2021. Highly accurate protein structure prediction with AlphaFold. Nature. 596:583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kalderimis A *et al.* 2014. InterMine: extensive web services for modern biology. Nucleic Acids Res. 42:W468–W472. https://doi.org/10.1093/nar/gku301.

Karimi K *et al.* 2021. Classifying domain-specific text documents containing ambiguous keywords. Database. 2021:baab062. https://doi.org/10.1093/database/baab062.

Karimi K *et al.* 2018. Xenbase: a genomic, epigenomic and transcriptomic model organism database. Nucleic Acids Res. 46: D861–D868. https://doi.org/10.1093/nar/gkx936.

Karimi K, Vize PD. 2014. The virtual Xenbase: transitioning an online bioinformatics resource to a private cloud. Database. 2014: bau108. https://doi.org/10.1093/database/bau108.

McCarthy FM *et al.* 2023. The case for standardizing gene nomenclature in vertebrates. Nature. 614:E31–E32. https://doi.org/10.1038/s41586-022-05633-w.

Mungall CJ, Batchelor C, Eilbeck K. 2011. Evolution of the sequence ontology terms and relationships. J Biomed Inform. 44:87–93. https://doi.org/10.1016/j.jbi.2010.03.002.

Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. 2012. Uberon, an integrative multi-species anatomy ontology. Genome Biol. 13:R5. https://doi.org/10.1186/gb-2012-13-1-r5.

Nenni MJ *et al.* 2019. Xenbase: facilitating the use of Xenopus to model human disease. Front Physiol. 10:154. https://doi.org/10.3389/fphys.2019.00154.

Philpott A. 2021. The use of Xenopus for cell biology applications. Cold Spring Harb Protoc. 6:221–225. https://doi.org/10.1101/pdb.top105528.

Putman TE *et al.* 2024. The monarch initiative in 2024: an analytic platform integrating phenotypes, genes and diseases across species. Nucleic Acids Res. 52:D938–D949. https://doi.org/10.1093/nar/gkad1082.

Rangwala SH *et al.* 2024. The NCBI comparative genome viewer (CGV) is an interactive visualization tool for the analysis of whole-genome eukaryotic alignments. PLoS Biol. 22:e3002405. https://doi.org/10.1371/journal.pbio.3002405.

Segerdell E *et al.* 2013. Enhanced XAO: the ontology of Xenopus anatomy and development underpins more accurate annotation of gene expression and queries on Xenbase. J Biomed Semantics. 4:31. https://doi.org/10.1186/2041-1480-4-31.

Session AM *et al.* 2016. Genome evolution in the allotetraploid frog Xenopus laevis. Nature. 538:336–343. https://doi.org/10.1038/nature19840.

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. 2009. JBrowse: a next-generation genome browser. Genome Res. 19: 1630–1638. https://doi.org/10.1101/gr.094607.109.

UniProt Consortium. 2023. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res. 51:D523–D531. https://doi.org/10.1093/nar/gkac1052.

Veltri D, Wight MM, Crouch JA. 2016. SimpleSynteny: a web-based tool for visualization of microsynteny across multiple species. Nucleic Acids Res. 44:W41–W45. https://doi.org/10.1093/nar/gkw330.

Vize PD. 2024. Anatomical ontologies. In: Vize PD, Zahn N, Vize SF, editors. Xenopus normal table redux. Ozymandias. p. 246–251.

Wilkinson MD *et al.* 2016. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 3:160018. https://doi.org/10.1038/sdata.2016.18.

Wynne M. 2017. The cucumber book. In: Hellesoy A, Tooke S, editors. Behaviour-driven development for testers and developers. 2nd ed Pragmatic Programmers, LLC.

Editor: J. Blake