Echinobase: leveraging an extant model organism database to build a knowledgebase supporting research on the genomics and biology of echinoderms

Bradley I. Arshinoff[1], Gregory A. Cary[2], Kamran Karimi [1], Saoirse Foley [2], Sergei Agalakov[1], Francisco Delgado[2], Vaneet S. Lotay[1], Carolyn J. Ku[2], Troy J. Pells [1], Thomas R. Beatman[2], Eugene Kim[1], R. Andrew Cameron [3], Peter D. Vize[1], Cheryl A. Telmer [2], Jenifer C. Croce[4], Charles A. Ettensohn[2] and Veronica F. Hinman[2],*

[1]Department of Biological Sciences, University of Calgary, Calgary, AB T2N 1N4, Canada, [2]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA, [3]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA and [4]Laboratoire de Biologie du Développement de Villefranche-sur-Mer (LBDV), Institut de la Mer de Villefranche (IMEV), Sorbonne Université, CNRS, Villefranche-sur-Mer, France

ABSTRACT

Echinobase (www.echinobase.org) is a third generation web resource supporting genomic research on echinoderms. The new version was built by cloning the mature *Xenopus* model organism knowledgebase, Xenbase, refactoring data ingestion pipelines and modifying the user interface to adapt to multispecies echinoderm content. This approach leveraged over 15 years of previous database and web application development to generate a new fully featured informatics resource in a single year. In addition to the software stack, Echinobase uses the private cloud and physical hosts that support Xenbase. Echinobase currently supports six echinoderm species, focused on those used for genomics, developmental biology and gene regulatory network analyses. Over 38 000 gene pages, 18 000 publications, new improved genome assemblies, JBrowse genome browser and BLAST + services are available and supported by the development of a new echinoderm anatomical ontology, uniformly applied formal gene nomenclature, and consistent orthology predictions. A novel feature of Echinobase is integrating support for multiple, disparate species. New genomes from the diverse echinoderm phylum will be added and supported as data becomes available. The common code development design of the integrated knowledgebases ensures parallel improvements as each resource evolves. This approach is widely applicable for developing new model organism informatics resources.

INTRODUCTION

Echinoderms (1) are widely used model organisms that have played a fundamental role in understanding how gene expression is regulated and how developmental processes are controlled and evolve (2–4). The ability to accurately detect gene expression patterns and map gene enhancer elements played an essential role in the uncovering of gene regulatory networks and laid the groundwork for modern systems-level approaches to development (5). Genome data are essential for interpreting such forms of information, and genomic database tools have evolved over the years to support such research efforts. This began with the release of the first complete echinoderm genome- that of the purple sea urchin (*Strongylocentrotus purpuratus*) (6), and the creation of a web resource to host the genome and its associated data called SpBase (7). This resource used Generic Model Organism Database (GMOD) tools (http://gmod.org) to operate a PostgreSQL database backend with a Chado schema and supporting tools such as GBrowse (8) and BLAST (9). An expanded version of this resource, renamed EchinoBase, was then built as additional genomic and transcriptome

information became available for both the purple urchin and seven other echinoderm species (10). This new resource partitioned content by species and added new features such as JBrowse (11) and Textpresso (12). The open-source Content Management System (CMS) Drupal was used to manage the expanded website content and integrate existing pages and features, e.g. the gene search. Data were mostly generated with automated systems, supplemented by community annotation of the purple sea urchin genome. While this resource provided unique and valuable content it lacked an integrated design and core foundational features such as ontologies.

Building a fully functional model organism knowledgebase takes many years of effort and considerable financial resources. The major model organism databases (MODs) for experimentally tractable systems such as mice, amphibians, fruit flies, etc. have all been developed for well over a decade and supported with extensive, novel, computational tools and high quality manual data curation. Although generic database schema and tools are available, specific community requirements and classes of data relevant to particular model organisms vary, and building a new MOD still takes a large team and effort. Most of the features of each resource are driven by their site-specific web applications—a resource for which no fully featured generic option is available. Xenbase is the *Xenopus* MOD that has been funded and developed for 15 years (13,14). Many of the experimental approaches used in amphibians, such as embryo microinjection, microdissection, morpholino knockdowns etc. are the same as those used in echinoderms, and therefore similar types of datasets are generated and interrogated by these communities. Initial Xenbase content was simple community and literature-based material. With time, more sophisticated data interfaces allowed the accumulation of high quality expert curated content and more complex data representations such as gene pages, gene expression data support, and disease support (15). In recent years Xenbase has led the MOD community by developing powerful integrated support for RNA-seq and ChIP-seq content (16) plus phenotype and 'gene expression as a phenotype' features unique to the field. Once again, these also represent techniques widely used in modern echinoderm research. The original Echinobase was built over an ~10-year period, and provided a BLAST interface, JBrowse and gene search functionality with specific content provided on each gene page (7). It was determined that new functionality would be more quickly and effectively provided by replicating the Xenbase software stack and adapting it to the needs of the Echinobase user community while maintaining the core content users have relied upon. This approach allows the new resource to leverage many years of development without replicating the enormous effort involved in building the resource from the ground upwards. In this report we describe the methods and hurdles encountered by this approach, plus the many powerful features of the first public release of the new resource.

The original resource remains available and can be reached at http://legacy.echinobase.org.

MATERIALS AND METHODS

Cloning a MOD code base

The Xenbase application code, along with data loading and exporting scripts, reside in a source control system. Echinobase files are stored within the same code repository and have one of the following relationships to Xenbase code:

i. shared files. The same files are used on both MODs and are species agnostic. These are used for common functionalities, such as BLAST pre- and post-processing,
ii. files specific to only one of the MODs, for example, landing page code,
iii. source files that are shared between resources but have conditional statements to change behaviour at run time depending on the MOD. An example is Gene Pages. Runtime conditional sources are in Java, JavaScript and JSP and drive MOD specific behaviours,
iv. shared files that change before compile time to determine if the code being built will behave as Xenbase or Echinobase when executed. An example is the global configuration files that determine site-wide variables such as the site name, supported species names, etc.

Xenbase and Echinobase source files therefore co-reside in the same development environment, and the development team does not have to deal with two disjointed sets of source files. Two different pre-compile scripts are then used, one to convert the values of the four types of files described above for Xenbase, and the other for Echinobase, to perform the needed changes before building the code. These scripts are highly efficient and take a few seconds to run. Building the code after running one of the scripts results in a Xenbase or Echinobase application.

The built application can then be deployed to a web application server, in our case the IBM WebSphere Application Server. The application server can be configured to support both MOD's specific database connections, meaning either MOD build can easily be deployed to the same application server. The resulting system allows the development team to switch between Xenbase and Echinobase development modes very quickly. It is possible to test bug fixes and new features in both Xenbase and Echinobase modes on a development machine with minimum delay. Going a step further, we have been careful to design the system in such a way that adding support for a third (or more) MOD is possible. In other words, the development solution is efficient, flexible, and scalable. The general features of the two software stacks and the balance between unique, shared and conditional code is illustrated by Supplementary Figure S1.

The details of the various code classes are as follows;

- Xenbase-specific code: SQL, Java, JSP, HTML or JavaScript code, which are meant to run on Xenbase only. Changes are done before compile time. These constitute a small number of files in the code base.
- Echinobase-specific code: SQL, Java, JSP, HTML or JavaScript code, which are meant to run on Echinobase

only. Changes are done before compile time. These constitute a small number of files in the code base.

- Conditional code: SQL, Java, JSP, HTML or JavaScript code, which are common to Xenbase and Echinobase, but use run-time testing to detect the running environment as Xenbase or Echinobase, and behave accordingly. These are files with conditional execution.
- Common code: SQL, Java, JSP, HTML or JavaScript code, which serve both Xenbase and Echinobase with no changes or conditions. These constitute the majority of the code base.

Adding support for Echinobase to the same source code as Xenbase has made it necessary to remove many hard-coded *Xenopus* data dependencies and assumptions. For example, we had hard coded the number of supported organisms (*X. laevis* and *X. tropicalis*), the set of genes per organism (one for *X. tropicalis* and two for the allotetraploid *X. laevis*) (17), and the text displayed on pages. The more flexible and open-ended new code falls in line with our aim of making the resulting solution ready to support MODs beyond Xenbase and Echinobase. The new system also includes TaxonID relationships for data, and as we also store the taxon relationship tree, data can now be analysed in a phylogenetic manner.

Data load

There are two main methods of loading data for Xenbase and Echinobase: one is bulk loads through scripts, that read the data from remote sites or local files (e.g. GFF3) and save them in the database following various quality checks; the other method is adding data through the curation interface. Echinobase, like Xenbase, has been using these scripts to consume data from sites such as NCBI (Entrez), UniProt, Disease Ontology, the GO consortium and others. The various data curation interfaces have also been activated for Echinobase, and manual data entry initiated for features such as publications and experimental reagents (e.g. antisense-morpholino-oligonucleotide sequences, antibodies and guide RNA sequences).

The new expanded JBrowse resource contains a large amount of novel content, including all echinoderm genomes currently annotated by the NCBI data processing systems. This includes genome sequence and gene/CDS and protein sequences from two sea urchin species; *S. purpuratus* and *Lytechinus variegatus*, three sea stars *Acanthaster planci*, *Patiria miniata* and *Asterias rubens* and a feather star, *Anneissia japonica*. Where previous genome builds were available in the legacy resource, we have generated liftover tracks so users can compare the old gene models to the new versions. Various additional datasets have also been mapped to the new genome builds such as various RNA-seq and ATAC-seq data and enhancer RNAs (eRNAs). We also map NCBI gene model names to genomes so users have useful gene identifiers within the browser. Some genomes also have in-house generated transcription factor binding sites, created using Xenbase protocols. Echinobase plans to support most new echinoderm genomes with both genome browser and BLAST support. Full support, where gene models are integrated into the database and linked to genes in other echinoderms on Gene Pages, will be provided for species that are widely used in developmental biology or systems biology- the core focus of the Echinobase resource. Currently these are three species; *S. purpuratus* (both genome version 3 and version 5), *Acanthaster planci* and *Patiria miniata.* We include additional genomes either into BLAST and JBrowse functions or as fully integrated datasets through consultation with the echinoderm researchers.

Hardware and virtualization

The Xenbase private cloud is supported by a pair of Lenovo x3850 × 6 servers and a pair of IBM x3650 M4 servers running VMware vSphere hypervisor. The cluster has a total of 2.6TB RAM and 256 logical CPU cores. The general hardware and software layout is as described in (18), with upgraded hardware. The Echinobase systems were distributed among separate virtual machines (VMs) in a manner similar to (18), to maximise system robustness and manageability. The Echinobase ecosystem presently uses eight VMs: production and test application servers (WebSphere), database (DB2), BLAST+, FTP, JBrowse, Wiki and a general computational resource VM. Other support systems provided by Xenbase include a bulk ftp resource, a source code versioning system, a ticketing and bug tracking tool, and an internal Wiki used for code documentation. All these tools require some level of maintenance and sharing them reduces effort. The use of these established resources once again saved significant development effort.

Loading the clone with Echinoderm literature

The Xenbase literature system uses NCBI's Eutilities (19) to pull in publications with '*Xenopus*' in the title or abstract of papers in PubMed, and these are updated weekly. The PubMed ID is then used to pull additional data from PubMedCentral (https://www.ncbi.nlm.nih.gov/pmc/), for example figures and figure legends when available. This simple approach was not effective for echinoderm literature as there are no terms in common between the diverse species studied, and taxon ID-based searches do not work well in this context. After experimenting with many alternatives, a set of 746 common and Latin names were selected as Eutility search terms. This method identified almost all the papers present in the legacy EchinoBase literature corpus, plus many more. As with Xenbase, the PubMed IDs were then used to both download and parse details on each publication and to collect figures and figure legends from PubMedCentral when available in their Open-i service (20). This method allowed us to generate over 18 000 dynamic pages representing publications, each with links to the various data sources. Xenbase uses a powerful link-matching system to identify terms in titles, abstracts and author lists to build links to gene pages, community member pages, and the anatomical ontology. One drawback of this approach is false positives; papers that have terms that match echinoderm type terms but have no biological relationship, for example 'aster' and 'star'. Such terms are removed using an exclusion list in the link building system. Machine learning is being used to identify false positives in the publication corpus (21). This in-house toolset uses classifier tools

and manually built training sets to identify publications that are loaded, and ones to be rejected. It has a high accuracy up to 98%, when compared to expert curator manual sorting, and outperforms the related NCBI tool LitSuggest (22). The current Echinobase literature dataset has 18 880 echinoderm-based publications.

Manual content curation

In addition to data ingestion and processing, the clone software stack also incorporates the many data curation interfaces available in Xenbase. These sometimes require refactoring when data is attached to an object not yet represented in the Echinobase clone, or when the large number of echinoderm species does not fit with the two species design of Xenbase. The most sophisticated interface currently available is the literature curation tool that allows curators to mark up publications with genes cited, gene expression data, human disease relationships, reagents etc. A second advanced curation interface is available for Gene Pages, where gene names, gene function, symbols and synonyms are commonly edited and tracked. Yet other interfaces are used to process experimental reagents, such as morpholinos, guide RNAs and antibodies. An Echinobase specific addition to these interfaces is the ability to tag papers, reagents etc. with an appropriate echinoderm species attribution selected from a subset of the NCBI echinoderm taxon IDs.

RESULTS

Migrating content to the new resource

The new cloned system imports most of the content found on the Echinobase legacy site directly, either from the NCBI or from genome GFF3 files, so database to database migration was not required. Echinoderm molecular data such as genes, mRNAs and proteins were all loaded using the standard Xenbase data ingestion methodologies. Literature content was loaded directly from PubMed and PubMedCentral as described in Materials and Methods. Static content from the legacy resource was migrated to a new Wiki resource for archiving.

New landing page and site navigation

A new landing page was designed using features available from Xenbase, such as central animated news 'slider', menu styles etc. but adapted to the new resource's content, and with its own unique colour scheme and logo to make the selected portal obvious to the user. As the Xenbase landing page had become content heavy, a minimization was performed for the Echinobase landing page and the content was reduced to the core key data types available for echinoderms. The new interface has simple and direct access to a wide range of tools, such JBrowse, Gene Pages and Publications (Figure 1).

The news slider is updated regularly and points users to new content, conferences and workshops, and interesting new papers. In the upper right of the Landing Page the number of new Gene Pages added, new papers in the literature section, and new employment advertisements in the community section are displayed to allow users to quickly view

new content. Links to tutorials and other community resources are also located on the right hand side of this page (Figure 1).

New Echinobase features

In the first public release of the new Echinobase in mid 2020 multiple core features are supported including Gene Pages, literature, genome browser, a Wiki and community resources. Three genomes and their associated data are fully integrated into the database and available in all site features, such as Gene Pages. These are the purple sea urchin *Strongylocentrotus purpuratus,* the bat star *Patiria miniata* and the Crown of Thorns sea star, *Acanthaster planci.* The green variegated sea urchin *Lytechinus variegatus* is in progress for full integration. Two additional genomes, that of the crinoid *Anneissia japonica* and the sugar sea star *Asterias rubens* are available via the genome browser and BLAST+ services only. JBrowse is the supported Echinobase browser.

Gene Pages

The Xenbase user design philosophy is gene centric, with most types of data, for example publications and gene expression, linked to additional resources via gene identifiers. Each gene in a specific echinoderm genome is linked to both species agnostic content, such as a gene symbol, gene name, protein domains, protein function, GO annotations etc., and species-specific data such as DNA, RNA and protein sequences, chromosome/scaffold maps etc. These data are assembled and displayed on dynamic Gene Pages (Figure 2). Gene Pages also link to publications citing the gene, GO terms (23) from UniProt (24) or on-site curation, gene expression data from cDNA libraries and literature curation, association with human diseases, protein-protein interaction and co-citation associations, and many other types of data relevant to developmental biologists. Gene Pages display this content for all fully supported echinoderm species, *S. purpuratus, P. miniata* and *A. planci,* with content for *L. variegatus* in final testing (see http://test.echinobase.org for the prototype including this species). Priorities for the inclusion of additional species are established through consultation with the community. There are over 38 000 gene pages in the present release. In previous Echinobase generations data from different species were stored and accessed separately; significantly, this is the first tool that allows content from more than one species to be assembled and viewed in parallel. Much of the data enriching these Gene Pages is for the first time available for echinoderm genes and genomes.

As the target user community uses a diverse range of echinoderm species, the Gene Page design was updated to create the flexibility to order page content according to user priorities. Directly under the species agnostic content described above (gene symbol etc.), species-specific data are organized into columns, and the user can select which species content is displayed in each of the three default columns using dropdown menus. The number of columns displayed can also be expanded or contracted in a user defined manner, using the various '+' and '–' icons. Gene
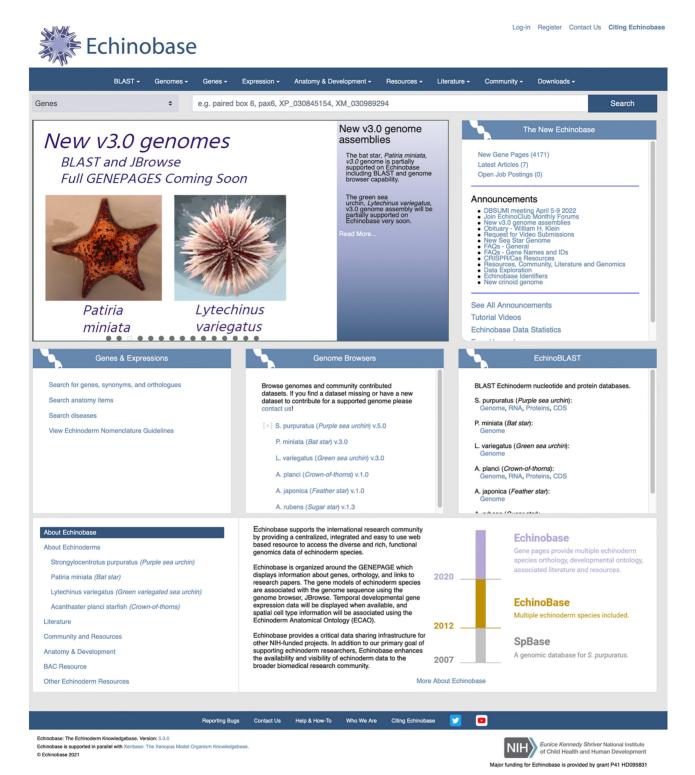
Figure 1. The Echinobase landing page and portal. A central feature of the page is an animated 'news slider' that presents a slide show of community news and new features of the resource. A horizontal navigation menu shared between all Echinobase pages provides consistency, and large 'tiles' of grouped content, e.g. 'Genes and Expression', 'Genome Browsers' and 'BLAST' aggregate commonly used tools and content.
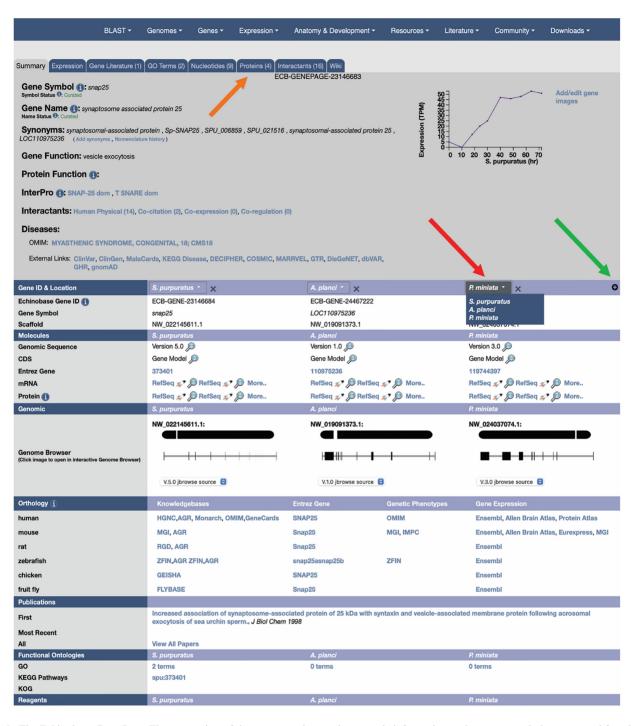
Figure 2. The Echinobase Gene Page. The top section of the page contains species agnostic information such as gene symbols, names and functions. It also contains synonyms, where legacy symbols and short names are stored and make genes locatable via database searches. The synonyms are also used to link Gene Pages to matched terms in the literature section. Due to the large amount of content aggregated on Gene Pages, much of the data is sorted into 'tabs' at the top of the page (orange arrow). Where a tab contains content, this is indicated by a numeral or icon. A transcriptional profile for the gene is also present, currently only for *S. purpuratus*, and will be expanded to additional species as data becomes available. The species specific content is arranged in vertical columns, each topped by a drop down menu that enables the user to define the content of that column (red arrow). In this example the third column header has been selected, displaying the three options for its content. The '+' symbol highlighted by the green arrow allows additional columns to be displayed. As we will be adding both a fourth species and paralogs to the display options in the near future, the ability to customize the data viewed will become even more important. Where we store and integrate multiple genome versions, for example for *S. purpuratus,* the user can also select which genome version is displayed as a gene model via a drop-down menu, and which genome version is linked to the Gene Page in JBrowse. This is only the case when we have integrated multiple versions, and this is only the case in some species. As different genome builds (e.g. *S. purpuratus* v3.1 and v5.0) have different sets of gene models, not all genes will be available in all genome versions. When this happens the number of available genomes will change in the user display. The chromosome maps and gene model exon/intron graphics are generated dynamically from database content by JavaScript, and selecting a new genome version will dynamically display the new gene structure for the selected version. The rocket icon that appears adjacent to many sequence based hyperlinks will load the corresponding sequence into BLAST, while the magnifying glass icon will open the sequence in a display window.

names and orthology between echinoderms are described in more detail in the following sections, but in brief gene names are assigned based on orthology between the purple urchin and human genes, and orthologs in other echinoderms selected via an ortholog table that stores 1:1 orthology relationships between *S. purpuratus* and the selected display species. The field on Gene Pages that displays orthology to genes in other model organisms is generated in a distinct manner. Those data are gathered from NCBI using the same code that generates these links in Xenbase.

While refactoring this code, changes were made to load the species-specific content into an intermediary JSON object, the content of which is pulled and displayed dependent on the user interface selections. This also generates a useful file for users to download the JSON object with all associated gene data. Changing the displayed species does not require rerunning SQL or customized SQL as the data for all species is already present in the JSON object, so is very fast.

Gene Pages display a gene expression profile in the species agnostic section derived from RNA-seq data from purple sea urchin. This is the only species for which we currently have content. Additional species, and multispecies comparisons, will be generated when the content is available. If this graph is clicked, a new view will be opened with a more sophisticated RNA-seq visualization tool that allows users to compare the profiles of different genes, change data transformation method, change graph colour scheme, download the graph as an SVG file etc. Each Gene Page also has multiple tabs where more specific associated data is compiled and available for the selected gene, such as gene expression data, publications citing the gene, associated GO terms, nucleotide sequences available for the gene (genomic, mRNA, EST), protein sequences (genome and cDNA derived), proteins with known interactions and a Wiki page where registered users can add pertinent content (Figure 2). Each tab has an indication that content is present and how much data is in the tab.

Gene Pages can be found by searching in the menu present in the top right corner of every Echinobase page, or the universal search on the landing page. The search method includes gene symbols, gene names and synonyms, so a search using an outdated legacy gene symbol such as Sp-Tbx2/3 or Omb will both load the correct page with the new approved gene symbol, tbx2. This search tool also accepts accession numbers and IDs from major data sources such as NCBI and UniProt.

We are working to also include a fourth species, *L. variegatus* in the database and on Gene Pages, along with the ability to display paralogs- where a gene has been duplicated in one or more echinoderm species. Additional columns can be added via the user interface to accommodate more species than the default three, or columns removed (Figure 2).

JBrowse (11) is the Echinobase supported genome browser. As described above in the Gene Page section, links on each gene page will open the genome browser for the selected species when clicked. Once in JBrowse, the species can be changed using the JBrowse dropdown menu, and the

JBrowse search window used to move to a different gene or genome feature. A video tutorial on how to use JBrowse is provided for users, based on the very similar implementation of JBrowse on Xenbase. If users are unfamiliar with the tool and the provided help files do not resolve the problem, they are encouraged to use the 'Contact Us' link in the top right corner of every page to request help from our team. Each genome we store has supporting BLAST tools. Links from the genome browser to Echinobase Gene Pages, and many other supporting data, are available by clicking on a gene model.

Orthology

Echinobase adopted the methods and best practices developed by the DIOPT consortia to predict orthology, as described in detail by Foley *et al.* (25). These are used in two ways; to assign gene names consistent with human nomenclature according to HUGO Gene Nomenclature Committee (HGNC), and to map relationships between genomes in different echinoderm species. The purple sea urchin gene set from genome build 5 was run against all of the Alliance of Genome Resources (26) maintained genomes by DIOPT (27). This data processing is only partial, as DIOPT relies on external resources to perform some analyses, and these resources do not yet include the purple sea urchin gene set. These results are used as a master orthology set of purple sea urchin to human and other Alliance model organism genomes. We also run the DIOPT tool set internally to build pairwise orthology sets between echinoderm species. To date, Echinobase has integrated seven different methods for inferring such orthologies: InParanoid v4.1 (28,29), PhylomeDB as described in (25) ProteinOrtho (30), SwiftOrtho (31) and FastOrtho (https://github.com/olsonanl/FastOrtho) are run with *e*-values set to 1e–40, with FastOrtho, OMA (32) and OrthoFinder (33) run with default settings. RefSeq identifiers for each hit are converted to Entrez IDs, and outputs are subsequently pooled. Per the 'moderate' criteria outlined by the Alliance of Genome Resources (26), genes from pairs of proteomes may be considered orthologous if three or more prediction tools converge upon a common result. Such hits are therefore treated as orthologous by Echinobase for echinoderm to echinoderm mappings. These mappings are stored in a data table and used to build the relationships used to pull in data and assemble Gene Pages from content derived from multiple species. Genes related by 1:1 orthology mappings can be displayed together and loaded into a single Gene Page.

Nomenclature

Echinobase gene nomenclature guidelines are modelled on those of other major model organisms and MODs, including Xenbase, aiming to provide gene identifiers that are informed by orthology relationships to human genes as established in our orthology pipeline, and apply the HGNC gene names and symbols whenever possible. Full details are available in formal guidelines on Echinobase: https://www.echinobase.org/gene/static/geneNomenclature.jsp and described in detail in (34).
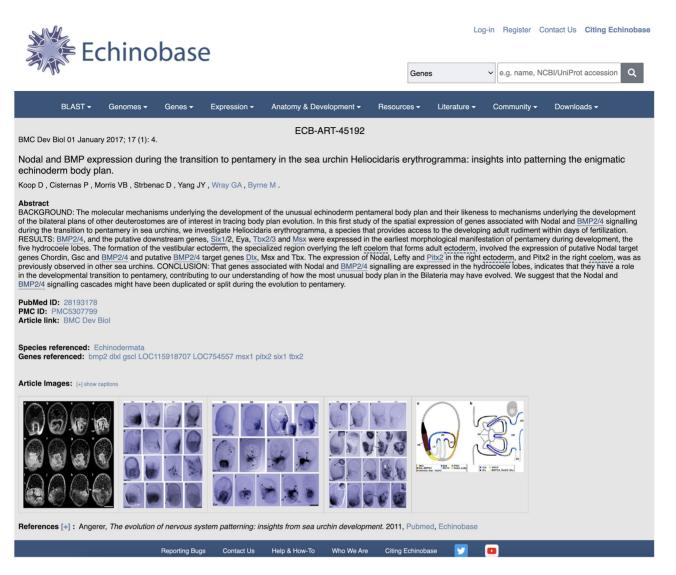
Figure 3. Echinobase literature pages. This screenshot illustrates a paper that has been processed by automated systems only, so all links, identified genes, authors and anatomy terms, figures and legends etc. were added by machine-based methods. The only organism detected by this was 'echinodermata', and a curator must annotate this page to include the species in the title, *Heliocidaris erythrogramma*, add additional genes, experimental reagents (morpholinos etc.), curate the gene expression patterns in the figures etc. When this is done all these data will also be displayed on this page. Authors with Echinobase community pages are hyperlinked, as are terms in the abstract and figure legends matching genes/synonyms, anatomy/synonyms etc. When text matches more than one target, selecting the link will display a disambiguation page.

Anatomy

Echinoderm anatomy and development is now supported by the Echinoderm Anatomical Ontology (ECAO). This tool currently has 443 anatomical and developmental stage terms and 1449 relationships between these terms- details will be published elsewhere. The anatomy and developmental stage terms in the ECAO can be searched via either text based queries or browser using a dynamic JavaScript driven tree viewer: https://www.echinobase.org/anatomy/ecao.do?method=display. The ECAO has been deposited in the public OBO Foundry (http://obofoundry.org/ontology/ecao.html) and can be downloaded there or directly from Echinobase in OWL or OBO formats (https://www.echinobase.org/anatomy/downloadEcao.do?method=display&tabId=1). The ECAO is used to codify a wide variety of content, such as gene expression data, anatomy and literature.

Literature

As described in the methods section, the Echinobase literature section stores over 18 000 publications on echinoderm biology, development, cell biology and genomics. Papers are loaded from PubMed and processed by a custom text classification system that removes irrelevant publications (21). In addition to the standard Title, Authors and Abstract, Echinobase appends many additional types of content, such as figures and figure legends (when available), a list of genes cited in the paper, references cited, links to the source paper and PubMedCentral etc. (Figure 3). Furthermore, this section uses the automated, synonym driven link-matching system pioneered by Xenbase (35). This tool uses gene symbols and synonyms, the anatomical ontology terms and their synonyms, and author names to build reciprocal links to records in other sections. The result is that when viewing an abstract that contains the text for legacy gene symbols,

e.g. 'Sp-Tbx2/3' or 'Omb', this text will be hyperlinked- and if selected, jump straight to the gene page for tbx2. Hyperlinks on author names will take a user directly to the person's Echinobase community page, where additional information such as their other publications are available. Whenever anatomy terms or synonyms are present in an abstract, a link is generated to the corresponding ECAO page, and from this page the tree-viewer tool in that section can be used to explore the relevant anatomy in greater detail. This allows users to flow from papers to Gene Pages, Gene Pages to papers, to authors, to anatomy, etc. in a seamless manner. The use of synonyms in this tool makes it especially powerful given the myriad of terms used to describe gene names and echinoderm anatomy in the literature. Additional data associated with publications, for example reagents such as morpholinos, guide RNAs and antibodies will also be available in the future as curation effort continues. Community members are encouraged to email suggestions and improvements to the curation of their papers.

Data visibility

Applying HGNC consistent gene symbols and names to the extremely diverse echinoderm based literature is a major step in making data from these species visible to the greater biomedical community. Papers using custom, community specific gene names are simply not findable via text based searches, so valuable experimental insights would remain only useful to the community performing echinoderm research without this change. The powerful synonym matching features of the Xenbase code allows users to work with either legacy or HGNC nomenclature while using the universal symbols for generating data reports for consumption by external resources. Major genomics resources such as Ensembl, NCBI and PubMed LinkOut use Xenbase curated gene symbols and names and curated data linked to papers to provide links back to Xenbase. At present manual data improvements, such as gene nomenclature and literature annotation are in early phases and much of the current content is automatically generated, such as gene symbols via orthology, species described and genes studied by text matching, etc. As data curation and processing proceed, Echinobase will work with these external resources to link to our content and make the wealth of data available from research on this phylum available to the wider world.

Help

Various help features are distributed around the resource. These include video tutorials linked from the landing page, information icons (an 'i' with a surrounding circle) are scattered throughout pages at locations where users have previously questioned the display or data, and a 'Contact Us' link at the top of every page where users can request input on an feature of the site. We welcome suggestions from community members for improvements to the user interface, subjects for new tutorials, additions to content, new features that would be useful, or any other aspect of the resource that would aid in our community utilizing the resource.

DISCUSSION

By reusing the database schema, software stack, web application code and curation interfaces of a pre-existing model organism knowledgebase it was possible to build a highly functional novel MOD in approximately one year. The new echinoderm resource has extensive support for molecular data on Gene Pages, a powerful echinoderm literature curation and browsing system and multiple new genomes supported by the JBrowse genome browser. The genome browser displays informative content such as RNA-seq, ATAC-seq, pseudogenes, etc. and in some cases map-over tracks to compare gene models from previous genomes to the latest versions. In the future we will add more Xenbase features to Echinobase, such as more extensive RNA-seq support, work on curating the massive volume of echinoderm based literature, and work to improve the usability and visibility of Echinobase content. Additional features already present in Xenbase can be added and populated with data in a fraction of the time it took to develop each of these systems. This can include extensive reagent support, deep RNA-seq and ChIP-seq data integration, phenotypes, and more. As Xenbase was not designed to be reused in this manner some issues arose, mostly due to hardcoding elements that required more flexibility in the multispecies context of Echinobase. In future development all data will be dealt with via taxon IDs, and it is planned to use the echinoderm phylogenetic tree to add various evolutionary analyses methods to the data corpus. This approach has been effective and produced a powerful knowledgebase with a modest development team, budget, and timeframe. The lessons learned in this effort, and additional flexibility now available through the species attribution methods, will also be applicable to other resources interested in following this approach.

DATA AVAILABILITY

Echinobase; http://www.echinobase.org

Xenbase; http://www.xenbase.org

GMOD, The Generic Model Organism Database project; http://gmod.org

Legacy Echinobase, a VM running the previous generation Echinobase; http://legacy.echinobase.org

FastOrtho, orthology prediction tool; https://github.com/olsonanl/FastOrtho

Echinobase gene nomenclature guidelines; https://www.echinobase.org/gene/static/geneNomenclature.jsp

The ECAO; https://www.echinobase.org/anatomy/ecao.do?method=display

The ECAO at the OBO Foundry; http://obofoundry.org/ontology/ecao.html

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

the 'About' page on each resource's landing page. Training and support from the Xenbase curatorial team at Cincinnati Children's Hospital was essential in building this resource, as was advice from the Echinobase Scientific Advisory Board. The contents of Echinobase are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

REFERENCES

1. Gilpin,D. (2006) Starfish, Urchins, and Other Echinoderms. In: Compass Point Books. Minneapolis.
2. Echinoderms, Part A (2019) In: Foltz,K. and Hamdoun,A. (eds). Methods in Cell Biology. Elsevier, Vol. **150**.
3. Echinoderms, Part B (2019) In: Foltz,K. and Hamdoun,A. (eds). *Methods in Cell Biology*. Elsevier, Vol. **150**.
4. Ettensohn,C.A. (2017) Sea urchins as a model system for studying embryonic development. In: *Reference Module in Biomedical Sciences*. Elsevier.
5. Davidson,E.H. (2009) Network design principles from the sea urchin embryo. *Curr. Opin. Genet. Dev.*, **19**, 535–540.
6. Sea Urchin Genome Sequencing Consortium, Sodergren,E., Weinstock,G.M., Davidson,E.H., Cameron,R.A., Gibbs,R.A., Angerer,R.C., Angerer,L.M., Arnone,M.I., Burgess,D.R. *et al.* (2006) The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science*, **314**, 941–952.
7. Cameron,R.A., Samanta,M., Yuan,A., He,D. and Davidson,E. (2009) SpBase: the sea urchin genome database and web site. *Nucleic Acids Res.*, **37**, D750–D754.
8. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
9. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
10. Cary,G.A., Cameron,R.A. and Hinman,V.F. (2018) Echinobase: tools for echinoderm genome analyses. *Methods Mol. Biol.*, **1757**, 349–369.
11. Skinner,M.E., Uzilov,A.V., Stein,L.D., Mungall,C.J. and Holmes,I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.
12. Müller,H.M., Van Auken,K.M., Li,Y. and Sternberg,P.W. (2018) Textpresso Central: a customizable platform for searching, text mining, viewing, and curating biomedical literature. *BMC Bioinformatics*, **19**, 94.
13. Karimi,K., Fortriede,J.D., Lotay,V.S., Burns,K.A., Wang,D.Z., Fisher,M.E., Pells,T.J., James-Zorn,C., Wang,Y., Ponferrada,V.G. *et al.* (2018) Xenbase: a genomic, epigenomic and transcriptomic model organism database. *Nucleic Acids Res.*, **46**, D861–D868.
14. James-Zorn,C., Ponferrada,V., Fisher,M.E., Burns,K., Fortriede,J., Segerdell,E., Karimi,K., Lotay,V., Wang,D.Z., Chu,S. *et al.* (2018) Navigating xenbase: an integrated xenopus genomics and gene expression database. *Methods Mol. Biol.*, **1757**, 251–305.
15. Nenni,M.J., Fisher,M.E., James-Zorn,C., Pells,T.J., Ponferrada,V., Chu,S., Fortriede,J.D., Burns,K.A., Wang,Y., Lotay,V.S. *et al.* (2019) Xenbase: facilitating the use of xenopus to model human disease. *Front. Physiol.*, **10**, 154.
16. Fortriede,J.D., Pells,T.J., Chu,S., Chaturvedi,P., Wang,D., Fisher,M.E., James-Zorn,C., Wang,Y., Nenni,M.J., Burns,K.A. *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.
17. Vize,P.D., Liu,Y. and Karimi,K. (2015) Database and informatic challenges in representing both diploid and tetraploid xenopus species in xenbase. *Cytogenet. Genome Res.*, **145**, 278–282.
18. Karimi,K. and Vize,P.D. (2014) The Virtual Xenbase: transitioning an online bioinformatics resource to a private cloud. *Database (Oxford)*, **2014**, bau108.
19. Sayers,E. (2010) A General Introduction to the E-utilities. In: *Entrez Programming Utilities Help, National Center for Biotechnology Information*.
20. Demner-Fushman,D., Antani,S., Simpson,M. and Thoma,G.R. (2012) Design and development of a multimodal biomedical information retrieval system. *J. Comput. Sci. Eng.*, **6**, 168–177.
21. Karimi,K., Agalakov,S., Telmer,C.A., Beatman,T.R., Pells,T.J., Arshinoff,B.I., Ku,C.J., Foley,S., Hinman,V.F., Ettensohn,C.A. *et al.* (2021) Classifying domain-specific text documents containing ambiguous keywords. *Database (Oxford)*, **2021**, baab062.
22. Allot,A., Lee,K., Chen,Q., Luo,L. and Lu,Z. (2021) LitSuggest: a web-based system for literature recommendation and curation using machine learning. *Nucleic Acids Res.*, **49**, W352–W358.
23. Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
24. UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
25. Foley,S., Ku,C., Arshinoff,B., Lotay,V., Karimi,K., Vize,P.D. and Hinman,V. (2021) Integration of 1:1 orthology maps and updated datasets into Echinobase. *Database (Oxford)*, **2021**, baab030.
26. Alliance of Genome Resources Consortium (2019) The alliance of genome resources: building a modern data ecosystem for model organism databases. *Genetics*, **213**, 1189–1196.
27. Hu,Y., Flockhart,I., Vinayagam,A., Bergwitz,C., Berger,B., Perrimon,N. and Mohr,S.E. (2011) An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics*, **12**, 357.
28. Remm,M., Storm,C.E. and Sonnhammer,E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.
29. Ostlund,G., Schmitt,T., Forslund,K., Köstler,T., Messina,D.N., Roopra,S., Frings,O. and Sonnhammer,E.L.L. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.*, **38**, D196–D203.
30. Lechner,M., Findeiss,S., Steiner,L., Marz,M., Stadler,P.F. and Prohaska,S.J. (2011) Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics*, **12**, 124.
31. Hu,X. and Friedberg,I. (2019) SwiftOrtho: a fast, memory-efficient, multiple genome orthology classifier. *Gigascience*, **8**, giz118.
32. Altenhoff,A.M., Glover,N.M., Train,C.-M., Kaleb,K., Warwick Vesztrocy,A., Dylus,D., de Farias,T.M., Zile,K., Stevenson,C., Long,J. *et al.* (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res.*, **46**, D477–D485.
33. Emms,D.M. and Kelly,S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.*, **20**, 238.
34. Beatman,T.R., Buckley,K.M., Cary,G.A., Hinman,V.F. and Ettensohn,C.A. (2021) A nomenclature for echinoderm genes. *Database (Oxford)*, **2021**, baab052.
35. Bowes,J.B., Snyder,K.A., Segerdell,E., Jarabek,C.J., Azam,K., Zorn,A.M. and Vize,P.D. (2010) Xenbase: gene expression and improved integration. *Nucleic Acids Res.*, **38**, D607–D612.