Updates to the Alliance of Genome Resources Central Infrastructure

Alliance of Genome Resources Consortium

Revised for Genetics 2024

The Alliance of Genome Resources Consortium (alphabetical)

Suzanne A. Aleksander[9], Anna V. Anagnostopoulos[5], Giulia Antonazzo[10], Valerio Arnaboldi[1], Helen Attrill[10], Andrés Becerra[2], Susan M. Bello[5], Olin Blodgett[5], Yvonne M. Bradford[11], Carol J. Bult[5], Scott Cain[8], Brian R. Calvi[4], Seth Carbon[6], Juancarlos Chan[1], Wen J. Chen[1], J. Michael Cherry[9], Jaehyoung Cho[1], Madeline A. Crosby[3], Jeffrey L. De Pons[7], Peter D'Eustachio[15], Stavros Diamantakis[2], Mary E. Dolan[5], Gilberto dos Santos[3], Sarah Dyer[2], Dustin Ebert[12], Stacia R. Engel[9], David Fashena[11], Malcolm Fisher[16], Saoirse Foley[13], Adam C. Gibson[7], Varun R. Gollapally[7], L. Sian Gramates[3], Christian A. Grove[1], Paul Hale[5], Todd Harris[8], G. Thomas Hayman[7], Yanhui Hu[14], Christina James-Zorn[16], Kamran Karimi[17], Kalpana Karra[9], Ranjana Kishore[1], Anne E. Kwitek[7], Stanley J. F. Laulederkind[7], Raymond Lee[1], Ian Longden[3], Manuel Luypaert[2], Nicholas Markarian[1], Steven J. Marygold[10], Beverley Matthews[3], Monica S. McAndrews[5], Gillian Millburn[10], Stuart Miyasato[9], Howie Motenko[5], Sierra Moxon[6], Hans-Michael Muller[1], Christopher J. Mungall[6], Anushya Muruganujan[12], Tremayne Mushayahama[12], Robert S. Nash[9], Paulo Nuin[8], Holly Paddock[11], Troy Pells[17], Norbert Perrimon[14], Christian Pich[11], Mark Quinton-Tulloch[2], Daniela Raciti[1], Sridhar Ramachandran[11], Joel E. Richardson[11], Susan Russo Gelbart[3], Leyla Ruzicka[11], Gary Schindelman[1], David R. Shaw[5], Gavin Sherlock[9], Ajay Shrivatsav[9], Amy Singer[11], Constance M. Smith[5], Cynthia L. Smith[5], Jennifer R. Smith[7], Lincoln Stein[8], Paul W. Sternberg[1], Christopher J. Tabone[3], Paul D. Thomas[12], Ketaki Thorat[7], Jyothi Thota[7], Monika Tomczuk[5], Vitor Trovisco[10], Marek A. Tutaj[7], Jose-Maria Urbano[10], Kimberly Van Auken[1], Ceri E. Van Slyke[11], Peter D. Vize[17], Qinghua Wang[1], Shuai Weng[9], Monte Westerfield[11], Laurens G. Wilming[5], Edith D. Wong[9], Adam Wright[8], Karen Yook[1], Pinglei Zhou[3], Aaron Zorn[16], Mark Zytkovicz[3]

[1]Division of Biology and Biological Engineering 140-18, California Institute of Technology, Pasadena, CA 91125, USA.

[2]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.

[3]The Biological Laboratories, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA.

[4]Department of Biology, Indiana University, Bloomington, IN 47408, USA.

[5]The Jackson Laboratory for Mammalian Genomics, Bar Harbor, ME, 04609, USA.

[6]Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA

[7]Medical College of Wisconsin - Rat Genome Database, Departments of Physiology and Biomedical Engineering, Medical College of Wisconsin, Milwaukee, WI 53226, USA.
[8]Informatics and Bio-computing Platform, Ontario Institute for Cancer Research, Toronto, ON M5G0A3, Canada.
[9]Department of Genetics, Stanford University, Stanford, CA 94305
[10]Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3DY, UK.
[11]Institute of Neuroscience, University of Oregon, Eugene, OR 97403
[12]Department of Population and Public Health Sciences, University of Southern California, Los Angeles, CA 90033, USA
[13]Department of Biological Sciences, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15203
[14]Department of Genetics, Howard Hughes Medical Institute, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA
[15]NYU Grossman School of Medicine, New York NY 10016
[16] Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave, Cincinnati, OH 45229, USA
[17] Department of Biological Sciences, University of Calgary, 507 Campus Dr NW, Calgary, AB T2N 4V8, Canada

correspondence: Paul W. Sternberg pws@caltech.edu

Running Title: Alliance of Genome Resources

Keywords: database; knowledgebase; software; text mining; data integration; *Drosophila*; yeast; *C. elegans*; zebrafish; mouse

Abstract

The Alliance of Genome Resources (Alliance) is an extensible coalition of knowledgebases focused on the genetics and genomics of intensively-studied model organisms. The Alliance is organized as individual knowledge centers with strong connections to their research communities and a centralized software infrastructure, discussed here. Model organisms currently represented in the Alliance are budding yeast, *C. elegans*, *Drosophila*, zebrafish, frog, laboratory mouse, laboratory rat, and the Gene Ontology Consortium. The project is in a rapid development phase to harmonize knowledge, store it, analyze it, and present it to the community through a web portal, direct downloads, and Application Programming Interfaces

(APIs). Here we focus on developments over the last two years. Specifically, we added and enhanced tools for browsing the genome (JBrowse), downloading sequences, mining complex data (AllianceMine), visualizing pathways, full-text searching of the literature (Textpresso), and sequence similarity searching (SequenceServer). We enhanced existing interactive data tables and added an interactive table of paralogs to complement our representation of orthology. To support individual model organism communities, we implemented species-specific "landing pages" and will add disease-specific portals soon; in addition, we support a common community forum implemented in Discourse software. We describe our progress towards a central persistent database to support curation, the data modeling that underpins harmonization, and progress towards a state-of-the art literature curation system with integrated Artificial Intelligence and Machine Learning (AI/ML).

Introduction

As has been discussed at length elsewhere (e.g., Oliver et al. 2016; Wood et al. 2022), model organism knowledgebases (a.k.a. model organism databases; MODs) provide daily utility to researchers for the design and interpretation of experiments, to computational biologists for curated datasets, and to genomic researchers for annotated genomes. Some of the major uses of the MODs have been one-stop shopping for all information about a particular gene or obtaining cleansed datasets with standard metadata for computational analyses.

The Alliance of Genome Resources (referred to herein as the Alliance) is a consortium of MODs and the Gene Ontology Consortium (GOC). The mission of the Alliance is to support comparative genomics to investigate the genetic and genomic basis of human biology, health, and disease. To promote sustainability of the core community data resources that make up the Alliance, we implemented an extensible "knowledge commons" platform for comparative genomics built with modular, re-usable infrastructure components that can support informatics resource needs across a wide range of species (Alliance of Genome Resources, 2022; Howe et al. 2018; Bult and Sternberg, 2023). In 2022, the Alliance was recognized as a Core Global Biodata Resource by the Global Biodata Coalition (Anderson et al 2017).

Specifically, the Alliance of Genome Resources is organized as two interdependent units: Alliance Central and the Alliance Knowledge Centers. ***Alliance Central*** is responsible for developing and maintaining the software for data access and for the coordination of data harmonization and data modeling activities across our members. A primary goal of Alliance Central is to reduce redundancy in systems administration and software development for model organism knowledgebases and to deploy a unified 'look and feel' for access to, and display of, common data types and annotations across diverse model organisms and human, following Findability, Accessibility, Interoperability, and Reuse (FAIR) guiding principles. Model organism-specific knowledgebases serve as ***Alliance Knowledge Centers.*** Knowledge Centers are responsible for expert curation and submission of data to Alliance Central using Alliance Central infrastructure. Knowledge Centers also are responsible for organism-specific user support activities and for providing access to data types not yet supported by Alliance Central. The founding Alliance Knowledge Centers are *Saccharomyces* Genome Database (Engel et al. 2022), WormBase (Davis et al. 2022; Sternberg et al., 2024), FlyBase (Gramates et al. 2022),

Mouse Genome Database (Ringwald et al. 2022), the Zebrafish Information Network (Bradford et al. 2023), Rat Genome Database (Vedi et al. 2023), and the Gene Ontology Consortium (Gene Ontology Consortium 2023). The newest member, Xenbase (Fisher et al, 2023), joined the Alliance consortium in 2022.

Here we describe our progress toward harmonizing information provided by our member resources, our development of a software infrastructure for ingest, curation, storage, analysis, and output of such information, and development of an efficient literature curation system. We start by describing new features in our web portal at AllianceGenome.org.

The Web Portal

Community Homepages. The Alliance website features landing pages for each model organism in the Alliance consortium. These pages are accessed from the “Members” drop-down menu in the header on every Alliance page. These pages feature MOD-specific-content such as meetings, news, and other MOD-specific resource links. A common template allows users to find the same types of information in each landing page (Figure 1). As MODs transition their data and web services to the Alliance, their member pages will evolve into portals hosting additional MOD-specific data, tools, and links to organism-specific resources.

Xenopus in the Alliance. Xenbase, the *Xenopus* knowledgebase (Fisher et al 2023), is the first knowledgebase to join the Alliance since the founding members initiated the consortium. *Xenopus* is an amphibian frog species used extensively in biomedical research, and in particular for experimental embryology, cell biology, and disease modeling with genome editing (Carotenuto et al. 2023; Kostiuk and Khokha 2021). As a non-mammalian air-breathing tetrapod, *Xenopus* represents a valuable evolutionary transition between rodents and zebrafish for comparative genomic studies. Xenbase is built on the same underlying data schema (structure) as FlyBase (Chado). Two different *Xenopus* species are used interchangeably as a model system: *X. tropicalis* is a diploid that is the preferred system for genome editing and genetics, whereas *X. laevis* is an allotetraploid preferred for use in cell biology studies, microinjection, and microsurgery-style experimentation. *X. tropicalis* has 1:1 relationships between most genes and human orthologs (excluding paralogs) (Mitros et al. 2019), whereas *X. laevis* has two copies of most human orthologs. The allotetraploid formed via hybridization of two different frog species (Session et al 2016), and the complexities of genome evolution that subsequently occurred increase the difficulty of identifying orthology of the two *X. laevis* genes to their diploid relatives, including humans. Mapping of the diploid *X. tropicalis* genes to their human orthologs was performed as with the other organisms in the Alliance (see below). Because this method does not yet work in the context of an allotetraploid, the Alliance imports the *X. tropicalis* to *X. laevis* paralogy mappings from Xenbase, where they have been established through a combination of synteny analysis and manual curation; this was one major challenges in adding *Xenopus* to the Alliance.

Xenbase created software to upload content on a regular schedule formatted for the current Alliance data ingest schema. Currently these data include orthology, the *Xenopus* anatomical ontology, standard gene information, gene expression data, publications, GO term associations,

disease associations, anatomical phenotypes, and genome details. *Xenopus* genes can be found using the Alliance landing page search tool with *Xenopus* genes flagged by *Xtr* and *Xla* notations. The two copies of the genes in *X. laevis*, the allotetraploid, are further tagged as '*(symbol).L*' and '*(symbol).S*' to denote the genes on the long (L) and short (S) chromosome pairs of this species (e.g., *pax6.L* and *pax6.S*). Alliance release 6.0.0 has Xenbase data for 54,000 genes, 19,000 disease associations, over 45,000 gene expression records and more than 7,000 anatomical phenotypes. Expression and phenotype data will be available in about a year.

In addition to the rich data made available to the Alliance from *Xenopus* research, this effort also served as a valuable test case for understanding the level of effort and complexities engendered in the addition of new knowledgebases to the Alliance, and the functionality and adaptability of ingest system components.

New gene page section: paralogy. Gene pages now include a Paralogy section populated with data from the Drosophila Research & Screening Center (DRSC) Integrative Ortholog Prediction Tool (DIOPT) version 9.1 developed by the DRSC (Hu et al. 2011, 2020). The assembly of protein sets and algorithmic inferences of their orthology from various sources was first centralized by the DRSC and then exported to the Alliance Central. We include the same data sources used for orthology, when these resources also provide paralogy information. Specifically, these resources have performed well on the standardized benchmarking from the Quest for Orthologs (QfO) Consortium (Nevers et al. 2022). Orthologous Matrix (OMA) (Altenhoff et al. 2021) and PANTHER (Thomas et al. 2022) datasets were retrieved through the QfO benchmark portal (https://orthology.benchmarkservice.org), and Compara data were acquired directly from the EBI Compara FTP site. In addition, the DRSC conducted local analyses using Inparanoid (Persson and Sonnhammer. 2022), OrthoFinder (Emms and Kelly 2019), OrthoInspector (Nevers et al. 2019), and sonicParanoid (Cosentino and Iwasaki 2019) using the UniProt 2020 reference proteome set (UniProt Consortium 2023), the same set used in the downloaded datasets, to ensure consistency. Direct data submissions from PhylomeDB (Fuentes et al. 2022) and the *Saccharomyces* Genome Database (SGD; Engel et al. 2022) were also integrated into the dataset.

The new paralogy section comprises a table (**Figure 2**), similar to the orthology table, that contains the gene symbol of related paralogs, a calculated rank, alignment length as the number of aligned amino acids, percentage of similarity and identity, as well as a count of the algorithms or methods that call the paralogous match. The ranking score was developed to sort the paralogs by overall similarity, and was reviewed by curators to display optimally an acceptable rank order for well-studied sets of paralogs. The ranking score considers several factors, including alignment length, percent identity, and the number of paralogy methods that identify the paralog. Additional Information for rank determination and alignment length are available to the users via a clickable help icon located next to those column headers.

The paralog section was released with Alliance version 6.0.0. Forthcoming updates will include the ability to sort and filter the table by column values and the availability of these data via our

bulk downloads page. The existing tables on the gene pages for Function, Disease, and Expression all contain checkboxes for "Compare Ortholog Genes" that allow users to search across species for these features. We will add the additional checkbox, "Compare Paralog Genes" to provide similar functionality for paralogous genes in a future Alliance release.

JBrowse sequence detail widget. A recent Alliance 6.0.0 release includes a new “Sequence Detail’" section of all gene pages that uses JBrowse and javascript libraries to display an interactive widget that allows users to download DNA and amino acid sequences of genes in several possible configurations: genomic sequence highlighted with UTRs, coding and intronic regions, CDS regions, and translated protein for example (Figure 3). In the next few releases, we will extend the functionality of the widget variant detail pages, where both the wild-type and variant sequences will be provided. When the variant occurs in the context of a protein coding gene, changes to the coding sequence and resulting translated protein will also be displayed and available for download.

Model organism BLAST. For more than two decades, some of the MOD members of the Alliance have hosted their own custom BLAST interfaces (Altschul et al. 1990; e.g., FlyBase Consortium. 1999), that have allowed users to search custom databases related to those model organisms, e.g., subsets of related species or molecular clones and display BLAST hits in Genome Browsers aligned with current gene models. We are now developing an updated and integrated Alliance BLAST, powered by SequenceServer (Priyam et al. 2019), that optimizes sequence analysis across model organisms. We have begun to update BLAST for the individual MODs. The new WormBase BLAST is now available online, and can currently be accessed via the tools menu on wormbase.org. The results are linked to Genome Browsers and Alliance gene pages (**Figure 4**). This tight connection allows users to navigate seamlessly between their BLAST results and the wealth of information available within the Alliance, enhancing the efficiency and depth of genetic research. For example, users can retrieve BLAST results for a sequence of interest and then easily navigate across Genome Browsers for different organisms, with a comparison to different tracks revealing how that sequence aligns with gene models, variants, and experimental tools (**Figure 5**). From a project perspective, developing Alliance BLAST with a common cloud-optimized infrastructure will increase efficiency by reducing the cost of compute overhead and eliminating the need to manage separate MOD systems, which will then allow more focus on developing new functionality to support researchers. Our focus in the upcoming year is directed toward enhancing the user interface, reflecting our commitment to providing an intuitive platform for researchers in model organism genetics. We plan to produce more analysis tools as part of the evolving Alliance portal, thereby broadening the range of resources available for genetic research within the community.

AllianceMine. AllianceMine, a sophisticated, multifaceted search and retrieval tool that utilizes the InterMine software (Smith et al. 2012), offers a unified view of harmonized data, enabling advanced queries across multiple species. For instance, gene lists can be processed as input and simultaneously query different annotations, such as 'Show me genes associated with a (specific disease term)' (**Figure 6**). The results from queries can be combined for further analysis, and saved or downloaded in customizable file formats. Queries themselves can be

customized by modifying predefined templates or by creating new templates to access a combination of specific data types. Thus, this powerful tool can be used in multiple ways, namely, for search, discovery, curation, and analysis.

AllianceMine currently showcases harmonized data encompassing genes, diseases, Gene Ontology (GO), orthology, expression, alleles, variants, and FASTA formatted genome sequences. The tool also offers predefined queries or "templates" for cross-species searching. Continual optimization will ensure timely data synchronization with the main Alliance site, as well as integration of newly harmonized data types. Another aspect of improvement will be the addition of more templates, widgets, and pre-compiled lists, which can serve as logical input for templated queries.

SimpleMine. We designed SimpleMine for biologists to get essential information for a list of genes without any command-line or programming skill, or patience to learn the awesome power of AllianceMine discussed above. Users can submit a list of gene names or IDs to access more than 20 types of essential data with which they are associated. The results are one line per gene with detailed information separated by four types of separators: tab, comma, bar, and semicolon. Users can choose to display the output as HTML or to download a tab-delimited file. Alliance SimpleMine contains ten species curated by the Alliance MODs. It provides easy gene name/ID conversion among MOD ID, public name (the commonly used name for public consumption), NCBI, PANTHER, Ensembl, and UniProtKB. Users can find summarized anatomic and temporal expression patterns, variants, genetic and physical interactions. Other essential gene information includes disease association and orthologs among all ten species. The infrastructure of SimpleMine allows users to perform species-specific searches for lists of genes that take about two seconds to return results, or mixed-species searches that take about 10 seconds to complete.

Pathway displays with metabolites (GO Causal Activity Models; GO-CAMs). We implemented a pathway display on Alliance gene pages that presents both GO-CAM (Thomas et al. 2019) and Reactome pathway (Milacic et al. 2024) models. The display queries both the Reactome and GO Application Programming Interfaces (APIs) and shows the number of pathways from each resource that contain the gene of interest. If a gene appears in multiple pathways, users can select which pathway to display. For the GO-CAM models, the viewer has been improved relative to previous releases of the Alliance website (**Figure 7**). First, the layout has been improved to show clearly the overall causal flow through a pathway, from top to bottom and branching as necessary. Second, the viewer displays not only the activities of genes/proteins in a pathway, but also metabolites, which is particularly useful for visualizing metabolic pathways. These metabolites may be either intermediates in a pathway, or regulators of a protein activity. For signaling pathways, we distinguish between direct and indirect regulation, and between positive, negative, or unknown effects.

Harmonized Data Models

The transition of data from individual MODs to the Alliance infrastructure requires data harmonization so that existing analogous MOD data classes (types/tables) can be loaded into

Alliance databases using a consistent schema and language. The first step is for biocurators from each Alliance knowledge center to agree on which data classes are analogous and can be treated as a single, consolidated data class. The biocurators then align the properties (table columns) of the consolidated data class, including identifiers, types of values, and whether entity-property-value associations/triples require their own respective metadata and/or evidence records. To enable this process the Linked Data Modeling Language (LinkML). We now have a standard workflow and common data modeling patterns that have streamlined the process, which we expect to complete in the next year. The LinkML specifications, authored in human-readable files, are used to programmatically generate JavaScript Object Notation (JSON) schema specifications, which allow Data Quartermasters (DQMs) to move data to the persistent store. These specifications also inform curation software developers how to generate initial backend (Java models and APIs) and front-end infrastructure (curation user interface data tables and detail pages). Once DQMs have submitted their data files for a particular data class, the data are loaded into the persistent store and validated (see persistent store architecture description below) and thus automatically populated into data tables and the curation interface. The data, having been harmonized, ingested, validated, and displayed to curators in the curation software, can now flow through to the public site according to the data pipeline described (see persistent store architecture description below).

Many Alliance data classes have completely (or nearly completely) harmonized data models in LinkML (see https://github.com/alliance-genome/agr_curation_schema) including: disease annotations, alleles, variants, expression annotations, and references. Although many other data classes have partially harmonized models, ongoing and future harmonization efforts will focus on completing harmonized models for the remaining curated data classes: genes, transcripts, proteins, non-transcribed genome features, affected genomic models (AGMs; strains, genotypes, fish), phenotype annotations, molecular and genetic interactions, gene regulation annotations, high-throughput expression dataset metadata (including for RNA-seq, single-cell RNA-seq, and proteomics datasets), species, reagents such as DNA clones and antibodies, images, persons, laboratories, companies, and various entity set classes like gene sets, which can be used for storing assay results and performing downstream analyses like ontology term enrichment, alignments, and other entity set processing calculations.

Persistent Store architecture

We have designed a powerful database system that can handle most of the demands of our project including curation, analysis, and display of the data (**Figure 8**). Specifically, we created a database using Postgres for long-term and persistent storage of Alliance curated data contributed by Alliance member MODs. In parallel to the existing (drop-and-reload) data pipeeline (Alliance 2022), DQMs from each MOD now submit data according to our new LinkML schema in JSON format directly to the persistent store for ingestion, validation, and curation via create-read-update-delete (CRUD) operations enabled by a curation API library and Prime React user interface (UI). A data pipeline has been established to provide data from the persistent store Postgres database to our Alliance public website APIs and front-end web user interfaces and to other tools and services.

LinkML-based JSON files are ingested into Postgres with validation to ensure: (1) recognition of submitted entities such as genes, alleles, affected genomic models (AGMs; e.g., strains, genotypes), publications, experimental conditions, and ontology terms, (2) recognition of references to such entities in annotations and associations, (3) no entry of duplicate entities, and (4) proper handling of obsolete entities. Every file load is accompanied by a report (in Postgres and the curation UI) indicating (1) the recognized MD5 sum and size of the (uncompressed) file submitted, (2) the success or failure of the load, (3) the number of entities recognized in the submitted file, (4) the number of distinct entities loaded into Postgres, (5) the number and identity of entities (if any) that failed to load and the reason for the failure, (6) a link to download the submitted file, (7) the corresponding compatible LinkML model/schema version, and (8) the MOD data release version corresponding to the data in the file submitted. This information can be used by DQMs, curators, and developers to keep track of the fidelity of the data transfer and troubleshoot any issues that arise. Ontology (and other external resource) loads are updated nightly to ensure that the latest versions of such data are current. The source of truth for MOD data will be transitioned over to the Alliance infrastructure in phases, beginning with a few data types from a few MODs and expanding over time to eventually include all (relevant) data types from all participating MODs; as part of this process, legacy issues with data are cleaned up.

To enable CRUD operations on persistent store data, curation APIs and a curation user interface accessible with Okta authentication have been implemented (**Figure 9**). Curators can now access data tables for the following data types: genes, alleles, variants, affected genomic models (AGMs; e.g. strains, genotypes), publications (accessed via Alliance Bibliographic Central (ABC) APIs), experimental conditions, constructs, disease annotations, molecules (not already managed by Chemical Entities of Biological Interest (ChEBI)), ontology terms, and controlled vocabularies and their terms. CRUD operations have been fully enabled for disease annotations, experimental conditions, and controlled vocabularies, read-update operations have been enabled for alleles and variants, and read operations are enabled for the remaining data types. Work is underway to fully enable CRUD operations on all remaining data classes and their attributes including new data tables for transcripts, proteins, other (non-gene) genome features, expression annotations, phenotype annotations, molecular interactions, genetic interactions, gene regulation annotations, antibodies, images, and more. In addition to data tables presenting all entries of a particular data class, the curation tool also has individual entity detail pages (for example, see an allele detail page https://curation.alliancegenome.org/#/allele/MGI:6446761) for data entry and editing on a dedicated web page for one particular entity. The curation tool also enables user-specific and MOD-specific custom user settings and preferences to provide a user interface most compatible with individual curators' workflows.

In the next year, the curation tool will include batch creation of data entities (e.g., annotations, reagents), batch editing, data history inspection and auditing, undo and review of latest changes, publication constraints (constrain data view and entry to publication currently being curated), customizations and MOD default settings for new entity creation and detail pages, incorporation of data entity and topic tagging information from the ABC literature store (see

below), and incorporation of AI/ML into the curation workflow.

For releases of persistent store data to the Alliance public website, Postgres database snapshots are taken and sent to a separate Postgres instance that feeds the data via the curation APIs (instantiated as a library) into the public site indexer where various data filtering and transformations occur before making those processed data available to our public website APIs via our Elasticsearch index. The Alliance public website user interface, using existing UI infrastructure, is then modified or created to accommodate the data now flowing from the persistent store database.

Security, stability and backups. All services and data provided by the Alliance to its community are hosted on Amazon web services (AWS). This provides us with industry leading availability of up to 99.99% on services like EC2, which we use to host our virtual servers. We use additional AWS-managed services such as Elastic Beanstalk for application deployment, AWS Relational Database Service for hosting our relational (Postgres) databases, and Amazon OpenSearch Service for hosting our search indexes, which all provide automatic updates and maintenance for increased reliability. All files hosted at the Alliance of Genome Resources are stored in S3 buckets, which ensures industry leading durability and availability. Furthermore, we make daily backups of our relational databases and have processes in place that enable easy restore of those backups in case of failure or data corruption. All Search indexes are derived from the persistent relational database and can be regenerated at any moment when required.

We make use of separated subnets between public-facing and private systems, and only services requiring public access are given public IP addresses, ensuring that public-facing services such as our curation interface can be accessed by our curators world-wide (through Okta Authentication), although the supporting back-end services such as the supporting databases can be kept private. Access to all services is furthermore restricted to allow access only to the required ports and services through the use of AWS Security Groups to control the allowed network traffic. AWS IAM users, groups, and roles are used to control the allowed AWS operations and access among Alliance developers. In all cases, the principle of least privilege is applied, so that the potential attack surface is reduced to a minimum (for example by not granting blanket AWS admin permissions to developers who do not have an AWS admin function). Access keys to any system can be revoked when misuse of those access keys is detected. We also configured our github repositories to be scanned automatically for accidental secret credential leakages through the use of GitGuardian software.

Literature Acquisition

We designed and are implementing a literature system, Alliance Bibliographic Central (ABC), that will support curation, and in the future, end users. The ABC supports the tasks of reference acquisition, triage, and curation workflow management. Specifically, the ABC is an ecosystem of online tools and supporting Alliance databases that manage all references and related metadata that are ‘in corpus’ for the member MODs.

Literature acquisition at the Alliance begins with automated, organism-specific PubMed queries

to retrieve candidate references for each MOD's corpus. References matching the search criteria are then added to the ABC by assigning an Alliance reference identifier and importing associated bibliographic information to the database. Subsequently, curators manually sort references as either 'in' or 'out of corpus' based on the curation policies of the MOD and eliminate any false positive results from the initial search. While many thousands of papers are published each year, only some have information that is currently curated. For example, in 2022, the curatable literature size after triage was: 3181 for ZFIN, 3221 for SGD,; 2130 for FlyBase, 1419 for WormBase, and 437 for Xenbase. Once references are sorted, they enter MOD-specific curation workflows supported by task-specific ABC curator interfaces to, for example, add reference files, manually tag references with specific entities (e.g., genes, alleles, and data types) and topics (e.g., phenotypes, anatomic expression) using the Alliance Tags for Papers (ATP) ontology, and merge duplicate references. In addition to adding reference files manually, the full text of 'in corpus' references included in the PubMed Central (PMC) open access set is also automatically downloaded. Curators may also use the ABC to add non-PubMed references. An additional key feature of the ABC is a search interface that allows curators to retrieve references based on various criteria including their in/out of corpus status, bibliographic data, and publication data range, if desired. Reference acquisition functionality can easily be extended to integrate additional MODs into the Alliance infrastructure.

To facilitate reference data exchange between the Alliance and MOD databases, the MODs provide a mapping file that associates MOD reference CURIEs (Compact Uniform Resource Identifier) with PMIDs, e.g., ZFIN:ZDB-PUB-181026-2 - PMID:30352852. The MODs also provide reference CURIEs and data for references not included in PubMed but used by the MOD, such as internal curation references and those published in a journal not yet indexed at PubMed.

Over the past 25-30 years, Alliance member databases have independently developed methods to acquire, triage, and curate their respective literatures. Having implemented a common literature curation interface, database, and full text acquisition system, the ABC is now poised to expand its functionality by incorporating ML methods developed by, and in production for, a subset of Alliance members to all groups. For example, automated pipelines that recognize entities (e.g., genes, alleles, strains) as well as data types (e.g., phenotype, genetic interactions) can be developed for new groups with results stored centrally in the Alliance literature database. Incorporating more automated methods will allow faster association of the published literature with relevant biological concepts, information that can be displayed on future Alliance references pages while the papers await detailed full curation. Centralized literature infrastructure will also support other curation pipelines, such as community curation by authors, which can then be more readily implemented for additional Alliance member communities thus providing another avenue by which curated data can be swiftly included in the Alliance. Lastly, the common literature tool will allow Alliance biocurators to coordinate curation of multi-species references that will provide users a facile way to find and view cross-species research exploiting the strengths of each Alliance model organism, a primary goal of the Alliance.

Textpresso. Textpresso is a full-text literature search engine that gets power from its single-sentence scope, focus on a specific model organism (or topic), and categories of semantically or biologically related terms (**Figure 10**; Müller et al. 2004; Müller et al. 2018). It has been used extensively by WormBase and SGD curators, as well as *C. elegans* and *S. cerevisiae* researchers in addition to other MODs (Van Auken et al. 2012; Bowes et al. 2013).

The Alliance is committed to creating Textpresso instances tailored to the unique needs of each member database, all of which will be managed within the Alliance software ecosystem and connected to the ABC as a single reference data source. This will reduce the overhead of managing Textpresso at individual MODs while also simplifying development and deployment of new features. Users will benefit from simplified access to Textpresso from the Alliance website. We also plan to integrate Textpresso searches further into specific Alliance web pages such as gene or allele pages. Users will be able to obtain additional references to biological entities through Textpresso searches, adding information from potentially non-curated literature to the list of curated references currently linked on those pages. Textpresso will be available to Alliance biocurators and to the general public through MOD-customized websites and via APIs for programmatic access.

Artificial Intelligence (AI). The Alliance member MODs have a track record of implementing ML tools to enhance literature triage and curation efficiency. Notable examples include RGD's early adoption of standard software architectures such as UIMA (Unstructured Information Management Architecture, an Apache.org project) and the development of the OntoMate system (Liu et al. 2015) an ontology-driven literature search engine, as well as WormBase's creation of Textpresso (Mueller et al. 2004) and document classifiers for paper triage.

The rise of Large Language Models (LLMs), such as BERT (Bidirectional Encoder Representations from Transformers), and ChatGPT, has transformed the natural language processing (NLP) landscape, but questions about their accuracy and "hallucinations" remain. The Alliance is developing LLMs for tasks such as document classification, Named Entity Recognition (NER), sentence classification, computationally assisted triage and curation and to build a natural language query system to simplify access to its underlying structured data.

Alliance members have developed AI/ML classifiers for determining with high accuracy whether papers returned from automated PubMed queries should be kept in their corpus or discarded (Jiang et al. 2020) and classifiers that can determine whether specific data types relevant for curation are present in a document (Fang et al. 2012). The Alliance is developing a central solution by providing these types of classifiers to all members.

Efforts are also underway to improve existing species-specific entity extraction and classification models, with a focus on incorporating human feedback in the loop and continuously training models based on data validated by professional biocurators and community curators. A centralized interface for “topic and entity tag” addition and validation during literature triage and curation is under development as part of the ABC. The interface allows curators to associate tags with publications and at the same time validate (or invalidate) results extracted from AI/ML

methods. This interface will streamline the collection of valuable training and testing sets and will allow a more systematic approach to the creation and comparison of different AI/ML models.

Furthermore, the Alliance is adopting Evidence and Conclusion Ontology (ECO) terms to record systematically the type of evidence, e.g. neural network method evidence, and assertion method, e.g. automatic assertion, used for reference flagging and triage. This is especially relevant for topic and entity tags. Using ECO terms aligns with FAIR data principles and offers transparency in curation workflows.

Application Programming Interfaces (APIs)

Application Programming Interfaces (APIs) are a key component of Alliance Central's data services infrastructure for rapid, modular software development. We currently support a dozen APIs with hundreds of endpoints (**Figure 11**). New APIs will be added as data harmonization and modeling of additional data entities are completed. We will expand public site APIs to generate all data needed for SimpleMine, AllianceMine, etc. from single endpoints. Current APIs include Public site APIs (agr_java_software in the GitHub repo) and APIs available from a public Swagger UI page. Because the public APIs support only GET endpoints, they do not require authentication. All APIs that support both GET and PUT/POST/DELETE endpoints do require authentication. Some of the key API endpoints available at https://www.alliancegenome.org/swagger-ui/ are: gene-summary, gene-disease, gene-interactions, homologs-species, allele-phenotypes, expression ribbon-summary, etc.

Data preservation in external repositories

The Alliance of Genome Resources is committed to the long-term preservation of digital objects (annotations) and resources (e.g., ontologies and software) that are central to the management and integration of functional knowledge about the genomes of diverse model organisms. As part of this commitment, the annotations and resources generated by Alliance members are integrated into many long-standing external public bioinformatic resources (e.g., Ensembl, UniProt, NCBI). Distribution of Alliance annotations from multiple sources provides a degree of redundancy that contributes to data stability and preservation. Alliance maintained ontologies and annotations and are also deposited into third party repositories that fulfill Open Science principles (see below). Leveraging community repositories ensures the data products and resources remain accessible to the research community even if the Alliance and/or its members cease operations.

Ontologies that Alliance members maintain are also available from long-term repositories including the OBO Foundry (https://obofoundry.org/) and Zenodo (zenodo.org). Annotations related to gene expression, function, phenotype, disease associations, etc. that are generated by Alliance members and are available on the Alliance Data Downloads page are archived in Zenodo. Software developed as part of the Alliance of Genome Resources knowledge commons platform is available from GitHub (https://github.com/alliance-genome). The external repositories used by the Alliance of Genome Resources include the *OBO Foundry* that was established in the early 2000s as a community-based initiative for development and maintenance of biological and biomedical ontologies using standardized practices. The Foundry

is the ontology repository of choice for the Alliance because it is widely recognized as an authoritative source of well-maintained ontologies for biology and biomedical research. *Zenodo* is a general purpose repository maintained by CERN (European Council for Nuclear Research) for storing and sharing documents, data, and other digital research materials across many disciplines. Zenodo is a repository of choice for the Alliance, in part, because of the commitment by the European Commission to support Zenodo as long as CERN exists.

Outreach and interactions

The Alliance Helpdesk. We established a common help desk email address (help@alliancegenome.org) that is featured prominently on the Alliance website header and footer under "Contact Us". All inquiries submitted using this email are logged as tickets in the Alliance Jira software system. Members of the User Support Working Group respond to user questions and inquiries in a timely manner, typically within 48 hours. Time to resolve user inquiries depends on the nature of the question or request. The Jira system tracks open tickets, forward tickets, tracks their active/resolved status, and classifies them by subject. We use the information, in part, to evaluate the design and utility of our user interfaces. For example, if particular questions or subjects arise frequently, we re-evaluate the design and wording of the search form and/or results display that caused user confusion.

Online documentation. We provide extensive user documentation about using the Alliance data resources under the Help menu on the homepage (https://www.alliancegenome.org/help). The online documentation provides guidance on such topics as how to use the search functions, defines acceptable field parameters, and provides explanations of the displayed results. The User Support Working Group also works closely with the User Interface Working Group and the Developers to craft text for tooltips displayed on user interfaces.

Frequently Asked Question (FAQ) pages. The FAQ/Known Issues page provides answers to commonly asked questions about the Alliance and also describes any known issues associated with a particular software release. The link to the FAQ page is featured prominently on the Alliance home page under the Help menu.

Illustrated tutorials and videos. We maintain several types of tutorial options that are accessible from the Help menu (https://www.alliancegenome.org/tutorials). The most requested types of tutorials are illustrated guides with screenshots on how to use various features of the Alliance web portal. When new functionality is released, we post to social media channels and issue "Tweetorials." Short video tutorials are disseminated through the Alliance YouTube channel.

Alliance User Community Forum. The Alliance supports a centralized community discussion board implemented in Discourse (https://community.alliancegenome.org/categories) (**Figure 12**). Each model organism represented in the Alliance is represented as its own Discourse category with model organism specific threads for news, discussion, and reagent information. The forum also includes categories for job postings, meeting announcements, and general

information about the Alliance of Genome Resources. Alliance members with existing on-line community forums are migrating users to the Alliance Central forum.

Users are not required to register to access the forum but must register to post messages, questions, and announcements. On average, ~1,000 users a day access the forum. Posts include jobs open and sought, news, meeting announcements and discussion of research approaches, reagents and interpretation.

Social Media. In addition to a News and Events header that links to software release notes and other Alliance Central updates, the Alliance uses standard social media venues to engage with the user community, including FaceBook (www.facebook.com/alliancegenome/), Twitter (now, X) (twitter.com/alliancegenome), Mastodon (https://genomic.social/@AllianceGenome), and Bluesky (https://bsky.app/profile/alliancegenome.bsky.social).

Prospects and Challenges

The tail of not-yet harmonized data. One challenge in the central Alliance infrastructure providing support for the union of MOD and GO features is the many unique dataset displays and tools that have evolved in the individual MODs over two decades. Among the 8 resources this comprises 150 years of branch length! Although horizontal tool transfer has occurred, it is not complete. We are taking a few approaches to this problem. In some cases, where the data are stand-alone, we will simply move the data and code to the Alliance. In the short term we will likely run tools off their existing servers. As tools age out, we will evaluate whether there is a broader mandate for that feature, and if so, implement it in the context of the Alliance.

There are types or aspects of our data that can be harmonized but have not yet been so. We adopted LinkML to help with harmonization because it provides a common language to represent structured data. The use of this language has spread to the point where our progress on harmonization is much more rapid.

AI. As discussed above, we are actively considering AI/ML applications throughout the project. Our practical approach is driven by us being subject matter experts. Because we have relied on human expert curation, we are in a unique position to evaluate and use the output of various AIs. Future plans include development of tools for creating training sets and a model manager for tracking ML models' performance. Integration with specialized biocuration tools such as Ontomate and Textpresso is part of the strategy, with a vision of harmonizing AI/ML solutions across member sites.

We will also explore the use of AI/ML in gene function summarization. Included on gene pages at the Alliance are short textual gene summaries based on curated and structured data annotations that provide users a quick overview of gene function. The current automated system for generating gene summaries has produced more than 160,000 summaries (Alliance version 6.0.0) for nine species, including humans (Kishore et al. 2020). However, to increase the coverage of genes further, we will explore the use of LLMs. This is especially relevant for

less-studied genes with few curated, structured data, and for scaling and upkeep of the summaries to match the rate of new gene data from publications. Leveraging LLMs to generate gene summaries for less studied genes, particularly those with limited curated data, offers the advantage of automatically uncovering relevant publications that may not have been previously curated. In principle, AI might be able to enhance or replace the automatically-generated textual gene summaries for both well studied and less studied genes.

We will use prompt engineering and finetuning of LLMs to improve accuracy of the generated summaries. As part of a continual improvement process, we will ask professional biocurators to evaluate summaries, and we will develop a scoring system based on several features such as readability of summaries, inclusion of key gene data, and checking for inaccurate and false data. To improve and keep gene summaries up to date, we plan to retrieve newly published articles that contain gene data that were not available when the LLM was trained and add extracted relevant text from the identified articles to the LLM prompt. To do so, we will use tools such as Textpresso (Muller et al. 2004) and Ontomate (Liu et al. 2015).

Community curation. Some Alliance MODs employ community curation pipelines to engage authors in curation of their papers. For example, FlyBase utilizes the Fast Track Your Paper (FTYP) (Bunt et al 2012; Larkin et al. 2021) tool that allows users to curate scientific papers, identify data types, and associate relevant genes with the reference. Authors using FTYP to ensure their papers appear quickly on the FlyBase website, help highlight data needing manual curation, and prioritize their publication for further curation.

Similarly, WormBase developed ACKnowledge (Author Curation to Knowledgebase; Arnaboldi et al. 2020), a semi-automated curation tool that lets authors curate their publications with the help of ML. Authors receive an email with a link to a form pre-populated by document-level classifiers that identify data types and several NER pipelines that extract lists of entities. Authors can correct and validate the extracted data using the form and submit curated information to WormBase. We will continue to provide these services to our community and develop a unified infrastructure which will help expand the service to other member communities.

Several Alliance members also collaborate with publishing groups, such as microPublication Biology (https://www.micropublication.org/) or the Genetics Society of America (https://genetics-gsa.org/publications/), to streamline pre-publication data integrity verification and curation by curators and authors, enabling MODs to quality-check and work with authors to correct data reporting before publication and promptly incorporate it into Alliance Knowledgebases upon article publication.

Dealing with satellite genomes and genetic models. In addition to the core genomes and associated data, our resources store and present information about the genes and genomes of relatively closely related organisms. For example, WormBase includes some genetically-studied nematodes such as *Caenorhabditis briggsae* that benefit from the rich data models typical of *C. elegans*. Genetic screens and positional cloning (Inoue et al. 2007; Sharanya et al. 2012), CRISPR editing (Cohen and Sternberg, 2019; Cohen et al. 2022; Ivanova and Moss 2023), as

well as transcriptomic analyses (Jhaveri et al. 2022) are now routinely done in this species. For the Alliance to take on this responsibility of WormBase, we need to support such satellite model organisms. Our plan is to support community gene structure annotation (e.g., for *Drosophila*, Sargent et al, 2020; for *C. elegans*, Moya et al. 2023) using the Apollo curaton system designed specifically for such activity (Dunn et al. 2019).

High Throughput expression data and single cell RNA-seq plans. We harmonized high-throughput expression metadata of mouse, rat, yeast, worm, fly, and zebrafish. Users can browse them with species, assay type (microarray, RNA-seq, tiling array, and proteomics), tissue, sex, and curated categories. We plan to add single-cell RNA-seq as a new assay type, making such datasets easily identifiable within our collection, with links to other resources, including Gene Expression Omnibus, EBI single-cell RNA-seq Expression Atlas, and CZI CellxGene, to display the information above, we will implement a content-rich expression detail page that will provide a unified way to access all expression data associated with a specific gene, including link outs to other sources and MOD-specific single-cell RNA-seq gene expression graphs (**Figure 13**).

Disease Portal(s). Providing users with ready and easy access to curated and harmonized model organism disease data and tools is crucial to accelerate research related to the pathogenesis of human disease. The Alliance has a wealth of disease-relevant data from eight model organism species and human data, such as: genes, alleles and variants implicated in disease, genotypes and strains that serve as disease models, and related data such as modifiers (herbals, chemicals, small molecules, etc.) that ameliorate or exacerbate the disease condition and may serve as candidates for potential drug development. To provide an easy entry point for clinical researchers and human geneticists to access the consolidated data and tools, we are in the process of designing and implementing a topic-specific resource--an Alzheimer's disease (AD) portal that will serve as a paradigm for other disease portals (**Figure 14).** The AD portal will include: orthologous genes in animal model systems, models with a mutation orthologous to one in a patient group, models with a specific set of phenotypes, and/or modifiers that have been shown to alter the disease condition. Building on the experience and pages developed for the AD portal, we will expand this paradigm to other disease portals. In addition to the specific disease portals we also plan to provide "compare" functionalities across diseases. Features planned for the disease portal with AD as an example include: a home page with an overview of the data in the portal, an autocomplete search box, links to other AD resources, and a list of the most recent papers from PubMed and/or from the ABC store (see example portal page below). The pages in the portal will be modeled on existing pages at the Alliance and will include gene summaries, alleles and variants, phenotypes, gene interactions, pathways, biological processes (based on GO), gene expression, etc. We also plan to provide visualizations of data analysis, for example, diseases that share genes and protein interactions that may point to common underlying molecular mechanisms. Up-to-date data sets, e.g., genes, strains, modifiers (drugs, chemicals, herbals, etc. shown to either ameliorate or exacerbate phenotypes) will be available as downloadable files. Disease-specific data sets will also be available for query from AllianceMine. We will also provide up-to-date links to disease-specific literature, and search capabilities through literature search engines such as the Textpresso

instance dedicated to AD (http://alzheimer.textpressocentral.org; corpus size - 96,000 papers). Not all papers are curatable by the MODs given their extensive but not comprehensive data models, and thus literature search will remain important.

The Alliance in the ecosystem of knowledgebases. The Alliance has a unique and complementary role relative to other informatics resources that support comparative biology. For example, NCBI's new Comparative Genomics Resource (CGR; Bornstein et al 2023) focuses on developing analysis tools and resources for *sequence-based* genome comparisons across a large number of species, the Alliance focuses on standardized annotations, harmonized biological concepts, and comparison of *biological knowledge*. The CGR supports comparative sequence analysis for all eukaryotes whereas the Alliance is primarily focused on model organisms used widely in biomedical research. These model organisms have a tremendous amount of highly valuable genetic, transgenic, and phenotypic data generated with multiple types of assays and are uniquely represented by the Alliance Knowledge Centers. The CGR uses the standardized gene summaries from the Alliance and follows nomenclature and ontology standards developed and maintained by Alliance members. For sequence analysis, the Alliance leverages sequence-based analysis tools developed and maintained by the CGR. Resource developers by and large appreciate the magnitude of the tasks we face in order to provide researchers with the information they need and strive to fill in the many gaps in services.

Acknowledgements

We thank our multiple communities for their patience and feedback about the prospect of the Alliance and their love of their own MODs. We also thank the members of our Scientific Advisory Board (Gary Bader, Alex Bateman, Helen Berman, Shawn Burgess, Andrew Chisholm, Phil Hieter, Brian Oliver, Calum Macrae, Titus Brown, Abraham Palmer and Michelle Southard-Smith) for cogent advice, and NHGRI Program Staff (Sandhya Xirasagar, Ajay Pillai, Valentina di Francesco, Sarah Hutchison, and Helen Thompson) for guidance.

Funding

The core funding for the Alliance is from the National Human Genome Research Institute and the National Heart, Lung and Blood Institute (U24HG010859). The curation of data and their harmonization is supported by National Human Genome Research Institute grants U24HG002659 (ZFIN), U24HG002223 (WormBase), U41HG000739 (FlyBase), U24HG001315 (SGD), U24HG000330 (MGD), P41HD064556 (Xenbase), U24HG011851 (Reactome + GO) and U41HG012212 (GO Consortium), as well as grant R01HL064541 from the National Heart, Lung and Blood Institute (RGD), P41HD062499 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (GXD), and the Medical Research Council-UK grant MR/L001020/1 (WormBase). Additional effort was supported by DOE DE-AC02-05CH11231. Curation tools are supported in part by the National Library of Medicine NLM R01LM013871.

Competing Interests

The authors declare no competing interests.

References

Alliance of Genome Resources, C., *Harmonizing model organism data in the Alliance of Genome Resources.* Genetics, 2022. **220**(4).

Altenhoff AM, Train CM, Gilbert KJ, Mediratta I, Mendes de Farias T, Moi D, Nevers Y, Radoykova HS, Rossier V, Warwick Vesztrocy A, Glover NM, Dessimoz C. OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. Nucleic Acids Res. 2021 Jan 8;49(D1):D373-D379. doi: 10.1093/nar/gkaa1007. PMID: 33174605; PMCID: PMC7779010.

Altschul SF, Gish W, Miller W, Eugene W. Myers, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

Anderson, W.P. and G. Global Life Science Data Resources Working, *Data management: A global coalition to sustain core data.* Nature, 2017. **543**(7644): p. 179.

Bornstein, K., et al., *The NIH Comparative Genomics Resource: addressing the promises and challenges of comparative genomics on human health.* BMC Genomics, 2023. **24**(1): p. 575.

Bowes JB, Snyder KA, James-Zorn C, Ponferrada VG, Jarabek CJ, Burns KA, Bhattacharyya B, Zorn AM, Vize PD. The Xenbase literature curation process. Database (Oxford). 2013 Jan 9;2013:bas046. doi: 10.1093/database/bas046. PMID: 23303299; PMCID: PMC3540419.

Bradford, Y.M., et al., *From multiallele fish to nonstandard environments, how ZFIN assigns phenotypes, human disease models, and gene expression annotations to genes.* Genetics, 2023. **224**(1).

Bult, C.J. and P.W. Sternberg, *The alliance of genome resources: transforming comparative genomics.* Mamm Genome, 2023.

Bunt SM, Grumbling GB, Field HI, Marygold SJ, Brown NH, Millburn GH; FlyBase Consortium. Directly e-mailing authors of newly published papers encourages community curation. Database (Oxford). 2012 May 2;2012:bas024. doi: 10.1093/database/bas024. PMID: 22554788; PMCID: PMC3342516.

Carotenuto R, Pallotta MM, Tussellino M, Fogliano C. Xenopus laevis (Daudin, 1802) as a Model Organism for Bioscience: A Historic Review and Perspective. Biology (Basel). 2023 Jun 20;12(6):890. doi: 10.3390/biology12060890. PMID: 37372174; PMCID: PMC10295250.

Cohen S, Sternberg P. Genome editing of *Caenorhabditis briggsae* using CRISPR/Cas9 co-conversion marker *dpy-10*. MicroPubl Biol. 2019 Oct 11;2019:10.17912/micropub.biology.000171. doi: 10.17912/micropub.biology.000171. PMID: 32550401; PMCID: PMC7252229.

Cohen SM, Wrobel CJJ, Prakash SJ, Schroeder FC, Sternberg PW. Formation and function of dauer ascarosides in the nematodes Caenorhabditis briggsae and Caenorhabditis elegans. G3 (Bethesda). 2022 Mar 4;12(3):jkac014. doi: 10.1093/g3journal/jkac014. PMID: 35094091;

PMCID: PMC8895998.

Cosentino S, Iwasaki W. SonicParanoid: fast, accurate and easy orthology inference. Bioinformatics. 2019 Jan 1;35(1):149-151. doi: 10.1093/bioinformatics/bty631. PMID: 30032301; PMCID: PMC6298048.

Davis, P., et al., *WormBase in 2022-data, processes, and tools for analyzing Caenorhabditis elegans.* Genetics, 2022. **220**(4).

Dunn NA, Unni DR, Diesh C, Munoz-Torres M, Harris NL, Yao E, et al. Apollo: democratizing genome annotation. PLoS Comput Biol. 2019;15:e1006790.

Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019 Nov 14;20(1):238. doi: 10.1186/s13059-019-1832-y. PMID: 31727128; PMCID: PMC6857279.

Engel SR, Wong ED, Nash RS, Aleksander S, Alexander M, Douglass E, Karra K, Miyasato SR, Simison M, Skrzypek MS, Weng S, Cherry JM. New data and collaborations at the Saccharomyces Genome Database: updated reference genome, alleles, and the Alliance of Genome Resources. Genetics. 2022 Apr 4;220(4):iyab224. doi: 10.1093/genetics/iyab224. PMID: 34897464; PMCID: PMC9209811.

Fang R, Schindelman G, Van Auken K, Fernandes J, Chen W, Wang X, Davis P, Tuli MA, Marygold SJ, Millburn G, Matthews B, Zhang H, Brown N, Gelbart WM, Sternberg PW. Automatic categorization of diverse experimental information in the bioscience literature. BMC Bioinformatics. 2012 Jan 26;13:16. doi: 10.1186/1471-2105-13-16. PMID: 22280404; PMCID: PMC3305665.

Fisher M, James-Zorn C, Ponferrada V, Bell AJ, Sundararaj N, Segerdell E, Chaturvedi P, Bayyari N, Chu S, Pells T, Lotay V, Agalakov S, Wang DZ, Arshinoff BI, Foley S, Karimi K, Vize PD, Zorn AM. Xenbase: key features and resources of the Xenopus model organism knowledgebase. Genetics. 2023 May 4;224(1):iyad018. doi: 10.1093/genetics/iyad018. PMID: 36755307; PMCID: PMC10158840.

FlyBase C. 1999. The FlyBase database of the Drosophila Genome Projects and community literature. *Nucleic Acids Res* **27**: 85-88.

Fuentes D, Molina M, Chorostecki U, Capella-Gutiérrez S, Marcet-Houben M, Gabaldón T. PhylomeDB V5: an expanding repository for genome-wide catalogues of annotated gene phylogenies. Nucleic Acids Res. 2022 Jan 7;50(D1):D1062-D1068. doi: 10.1093/nar/gkab966. PMID: 34718760; PMCID: PMC8728271.

Gene Ontology, Consortium., *The Gene Ontology knowledgebase in 2023.* Genetics, 2023. **224**(1).

Gramates, L.S., et al., *FlyBase: a guided tour of highlighted features.* Genetics, 2022. **220**(4).

Howe, D.G., et al., *Model organism data evolving in support of translational medicine.* Lab Anim (NY), 2018. **47**(10): p. 277-289.

Hu Y, Comjean A, Rodiger J, Liu Y, Gao Y, Chung V, Zirin J, Perrimon N, Mohr SE. FlyRNAi.org-the database of the Drosophila RNAi screening center and transgenic RNAi

project: 2021 update. Nucleic Acids Res. 2021 Jan 8;49(D1):D908-D915. doi: 10.1093/nar/gkaa936. PMID: 33104800; PMCID: PMC7778949.

Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, Mohr SE. An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics. 2011 Aug 31;12:357. doi: 10.1186/1471-2105-12-357. PMID: 21880147; PMCID: PMC3179972.

Inoue T, Ailion M, Poon S, Kim HK, Thomas JH, Sternberg PW. Genetic analysis of dauer formation in Caenorhabditis briggsae. Genetics. 2007 Oct;177(2):809-18. doi: 10.1534/genetics.107.078857. Epub 2007 Jul 29. PMID: 17660533; PMCID: PMC2034645.

Ivanova M, Moss EG. Orthologs of the C. elegans heterochronic genes have divergent functions in *C. briggsae*. Genetics. 2023 Oct 3:iyad177. doi: 10.1093/genetics/iyad177. Epub ahead of print. PMID: 37788363.

Jhaveri N, van den Berg W, Hwang BJ, Muller HM, Sternberg PW, Gupta BP. Genome annotation of Caenorhabditis briggsae by TEC-RED identifies new exons, paralogs, and conserved and novel operons. G3 (Bethesda). 2022 Jul 6;12(7):jkac101. doi: 10.1093/g3journal/jkac101. PMID: 35485953; PMCID: PMC9258526.

Jiang X, Li P, Kadin J, Blake JA, Ringwald M, Shatkay H. Integrating image caption information into biomedical document classification in support of biocuration. Database (Oxford). 2020 Jan 1;2020:baaa024. doi: 10.1093/database/baaa024. PMID: 32294192; PMCID: PMC7159034.

Kostiuk V, Khokha MK. Xenopus as a platform for discovery of genes relevant to human disease. Curr Top Dev Biol. 2021;145:277-312. doi: 10.1016/bs.ctdb.2021.03.005. Epub 2021 Apr 23. PMID: 34074532; PMCID: PMC8734201.

Larkin A, Marygold SJ, Antonazzo G, Attrill H, Dos Santos G, Garapati PV, Goodman JL, Gramates LS, Millburn G, Strelets VB, Tabone CJ, Thurmond J; FlyBase Consortium. FlyBase: updates to the Drosophila melanogaster knowledge base. Nucleic Acids Res. 2021 Jan 8;49(D1):D899-D907. doi: 10.1093/nar/gkaa1026. PMID: 33219682; PMCID: PMC7779046. (https://pubmed.ncbi.nlm.nih.gov/33219682/)

Liu W, Laulederkind SJ, Hayman GT, Wang SJ, Nigam R, Smith JR, De Pons J, Dwinell MR, Shimoyama M. OntoMate: a text-mining tool aiding curation at the Rat Genome Database. Database (Oxford). 2015 Jan 25;2015:bau129. doi: 10.1093/database/bau129. PMID: 25619558; PMCID: PMC4305386.

Milacic M, Rothfels K, Mathews L, Wright A, Jassal B, Shamovsky V, Trinh Q, Gillespie M, Sevilla C, Tiwari K, Ragueneau E, Gong C, Stephan1 R, May B, Haw R, Weiser J, Beavers D, Conley P, Hermjakob H, Stein LD, D'Eustachio P, Wu G (2024) The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res*., in press. PMID: 37941124

Mitros T, Lyons JB, Session AM, Jenkins J, Shu S, Kwon T, Lane M, Ng C, Grammer TC, Khokha MK, Grimwood J, Schmutz J, Harland RM, Rokhsar DS. A chromosome-scale genome assembly and dense genetic map for Xenopus tropicalis. Dev Biol. 2019 Aug 1;452(1):8-20. doi: 10.1016/j.ydbio.2019.03.015. PMID: 30980799.

Moya ND, Stevens L, Miller IR, Sokol CE, Galindo JL, Bardas AD, Koh ESH, Rozenich J, Yeo

C, Xu M, Andersen EC. Novel and improved Caenorhabditis briggsae gene models generated by community curation. BMC Genomics. 2023 Aug 25;24(1):486. doi: 10.1186/s12864-023-09582-0. PMID: 37626289; PMCID: PMC10463891.

Müller HM, Van Auken KM, Li Y, Sternberg PW. Textpresso Central: a customizable platform for searching, text mining, viewing, and curating biomedical literature. BMC Bioinformatics. 2018 Mar 9;19(1):94. doi: 10.1186/s12859-018-2103-8. PMID: 29523070; PMCID: PMC5845379.

Nevers Y, Jones TEM, Jyothi D, Yates B, Ferret M, Portell-Silva L, Codo L, Cosentino S, Marcet-Houben M, Vlasova A, Poidevin L, Kress A, Hickman M, Persson E, Piližota I, Guijarro-Clarke C; OpenEBench team the Quest for Orthologs Consortium; Iwasaki W, Lecompte O, Sonnhammer E, Roos DS, Gabaldón T, Thybert D, Thomas PD, Hu Y, Emms DM, Bruford E, Capella-Gutierrez S, Martin MJ, Dessimoz C, Altenhoff A. The Quest for Orthologs orthology benchmark service in 2022. Nucleic Acids Res. 2022 Jul 5;50(W1):W623-W632. doi: 10.1093/nar/gkac330. PMID: 35552456; PMCID: PMC9252809.

Nevers Y, Kress A, Defosset A, Ripp R, Linard B, Thompson JD, Poch O, Lecompte O. OrthoInspector 3.0: open portal for comparative genomics. Nucleic Acids Res. 2019 Jan 8;47(D1):D411-D418. doi: 10.1093/nar/gky1068. PMID: 30380106; PMCID: PMC6323921.

Oliver, S.G., et al., *Model organism databases: essential resources that need the support of both funders and users.* BMC Biol, 2016. **14**: p. 49.

Persson E, Sonnhammer ELL. InParanoid-DIAMOND: faster orthology analysis with the InParanoid algorithm. Bioinformatics. 2022 May 13;38(10):2918-2919. doi: 10.1093/bioinformatics/btac194. PMID: 35561192; PMCID: PMC9113356.

Priyam A, Woodcroft BJ, Rai V, Moghul I, Munagala A, Ter F, Chowdhary H, Pieniak I, Maynard LJ, Gibbins MA et al. 2019. Sequenceserver: A Modern Graphical User Interface for Custom BLAST Databases. *Mol Biol Evol* **36**: 2922-2924.

Ringwald, M., et al., *Mouse Genome Informatics (MGI): latest news from MGD and GXD.* Mamm Genome, 2022. **33**(1): p. 4-18.

Sargent L, Liu Y, Leung W, Mortimer NT, Lopatto D, Goecks J, Elgin SCR. G-OnRamp: Generating genome browsers to facilitate undergraduate-driven collaborative genome annotation. PLoS Comput Biol. 2020 Jun 4;16(6):e1007863. doi: 10.1371/journal.pcbi.1007863. PMID: 32497138; PMCID: PMC7272004.

Session AM, Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S, Fukui A, Hikosaka A, Suzuki A, Kondo M, van Heeringen SJ, Quigley I, Heinz S, Ogino H, Ochi H, Hellsten U, Lyons JB, Simakov O, Putnam N, Stites J, Kuroki Y, Tanaka T, Michiue T, Watanabe M, Bogdanovic O, Lister R, Georgiou G, Paranjpe SS, van Kruijsbergen I, Shu S, Carlson J, Kinoshita T, Ohta Y, Mawaribuchi S, Jenkins J, Grimwood J, Schmutz J, Mitros T, Mozaffari SV, Suzuki Y, Haramoto Y, Yamamoto TS, Takagi C, Heald R, Miller K, Haudenschild C, Kitzman J, Nakayama T, Izutsu Y, Robert J, Fortriede J, Burns K, Lotay V, Karimi K, Yasuoka Y, Dichmann DS, Flajnik MF, Houston DW, Shendure J, DuPasquier L, Vize PD, Zorn AM, Ito M, Marcotte EM, Wallingford JB, Ito Y, Asashima M, Ueno N, Matsuda Y, Veenstra GJ, Fujiyama A, Harland RM, Taira M, Rokhsar DS. Genome evolution in the allotetraploid frog Xenopus laevis. Nature.

2016 Oct 20;538(7625):336-343. doi: 10.1038/nature19840. PMID: 27762356; PMCID: PMC5313049.

Sharanya D, Thillainathan B, Marri S, Bojanala N, Taylor J, Flibotte S, Moerman DG, Waterston RH, Gupta BP. Genetic control of vulval development in Caenorhabditis briggsae. G3 (Bethesda). 2012 Dec;2(12):1625-41. doi: 10.1534/g3.112.004598. Epub 2012 Dec 1. PMID: 23275885; PMCID: PMC3516484.

Smith RN, Aleksic J, Butano D, Carr A, Contrino S, Hu F, Lyne M, Lyne R, Kalderimis A, Rutherford K, Stepan R, Sullivan J, Wakeling M, Watkins X, Micklem G. InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics. 2012 Dec 1;28(23):3163-5. doi: 10.1093/bioinformatics/bts577. Epub 2012 Sep 27. PMID: 23023984; PMCID: PMC3516146.

Sternberg et al., 2024. WormBase article, Genetics

Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou LP, Mi H. PANTHER: Making genome-scale phylogenetics accessible to all. Protein Sci. 2022 Jan;31(1):8-22. doi: 10.1002/pro.4218. Epub 2021 Nov 25. PMID: 34717010; PMCID: PMC8740835.

Thomas, P.D., et al., *Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems.* Nat Genet, 2019. **51**(10): p. 1429-1433.

UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023 Jan 6;51(D1):D523-D531. doi: 10.1093/nar/gkac1052. PMID: 36408920; PMCID: PMC9825514.

Van Auken K, Fey P, Berardini TZ, Dodson R, Cooper L, Li D, Chan J, Li Y, Basu S, Muller HM, Chisholm R, Huala E, Sternberg PW; WormBase Consortium. Text mining in the biocuration workflow: applications for literature curation at WormBase, dictyBase and TAIR. Database (Oxford). 2012 Nov 17;2012:bas040. doi: 10.1093/database/bas040. PMID: 23160413; PMCID: PMC3500519.

Vedi, M., et al., *2022 updates to the Rat Genome Database: a Findable, Accessible, Interoperable, and Reusable (FAIR) resource.* Genetics, 2023. **224**(1).

Wood V, Sternberg PW, Lipshitz HD. Making biological knowledge useful for humans and machines. Genetics. 2022 Apr 4;220(4):iyac001. doi: 10.1093/genetics/iyac001.

Legends

Figure 1. MOD landing pages at the Alliance Portal. A common look and feel that allows community-specific content.

Figure 2. Paralog table for *C. elegans hlh-25*. The table presents a ranking of paralogs for the *hlh-25* gene, based on a weighted scoring algorithm that incorporates sequence conservation metrics. It lists the gene symbols, provides the alignment length in amino acids, and quantifies

the similarity and identity percentages of genes paralogous to *hlh-25*. The methodology count, indicating the number of algorithms supporting the paralogous relationship, is also included. In this ranking, *hlh-27* is identified as the primary paralog due to its high similarity and identity scores, despite being recognized by fewer methods than *hlh-28*.

Figure 3. Sequence detail widget. Chosen views of a specific gene are readily available for copying as plain text or with highlights. 5' region of the human PLAA gene.

Figure 4. Screenshot of results from the Alliance SequenceServer BLAST tool. The results have been enhanced relative to the default Sequence Server results page by the addition of links to Alliance JBrowse and to the corresponding gene page (in this case, *C. elegans* abi-1) at the Alliance website for each BLAST hit.

Figure 5. Output of a BLAST search After a user clicks on the JBrowse link for a BLAST hit they are directed to the web service where they will see a track for the BLAST hit and how the hit aligns with other tracks.

Figure 6. AllianceMine example. Using a simple template, a disease ontology (DO) term, in this case "autism," is chosen, and all genes associated with this DO term are returned in a downloadable table.

Figure 7. Alliance Pathway Viewer. The pathway widget displays gene products (light purple rectangles), and chemicals (light blue rectangles) and the flow of information and material between them (relations). These relations, shown in legend indicate direct or indirect regulation that can be positive, negative or of unknown effect direction. For metabolites, grey-blue shows that a metabolite mediates the information flow between gene products. In addition, blue lines with circles indicate input to a reaction; pink indicates output of a reaction.

Figure 8. Evolution of Data Flow. Graphical summary showing the design of short term infrastructure initially deployed to support rapid delivery of unified data to the community and the planned production system. Red, data quartermasters at MODs; Yellow, data; Brown, database; Green, transformations; Blue, user interface.

Figure 9. Alliance Curation tool. Screenshot of the Alliance curation tool interface showing an example of curated annotations of Affected Genomic Models managed in the persistent store.

Figure 10. Textpresso for **SGD literature at the Alliance**. (http://sgd-textpresso.alliancegenome.org/tpc/search)

Figure 11. Swagger interface for the Alliance APIs.

Figure 12. Alliance community forum home page.

Figure 13. Mockup of an Expression Detail page. This example shows one of the current features of WormBase – single cell data from two studies – displayed on what will be part of an Alliance Gene Expression detail page.

Figure 14. Mockup of the Alzheimer's Disease Portal showing the Home page and the Data access page. These views illustrate the type of information that will be available with a disease-focus.

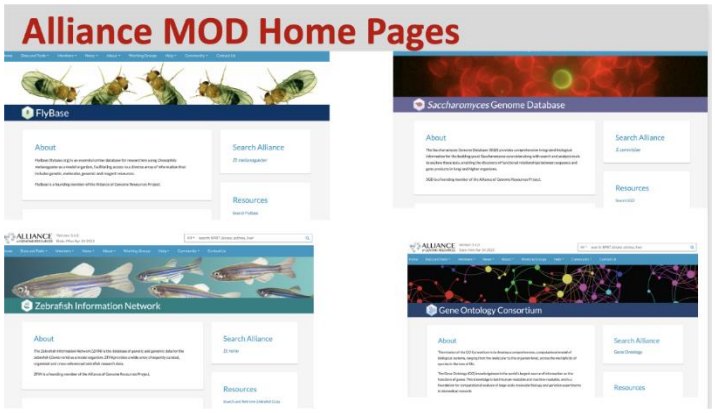

Figure 1
94x53 mm (x DPI)

Paralogy

Gene symbol	Rank	Alignment Length (aa)	Similarity %	Identity %	Method Count	Method: Ensembl Compara	HCNC	Treefam-9	OMA	OrthoFinder	OrthoInspector	PANTHER	PhylomeDB	SonicParanoid	SQD
hlh-27	1	268	99	99	3 of 8	☑	☐	☐	☑	☐	☑	☐	☐	☐	
hlh-28	2	277	55	39	4 of 8	☑	☐	☐	☑	☐	☑	☑	☐		
hlh-29	3	279	54	38	4 of 8	☑	☐	☐	☑	☐	☑	☑	☐		
hlh-26	4	274	48	32	4 of 8	☑	☐	☐	☑	☐	☑	☑	☐		
ref-1	5	353	38	25	2 of 8	☐	☐	☐	☑	☐	☑	☐	☐		

Figure 2
82x33 mm (x DPI)

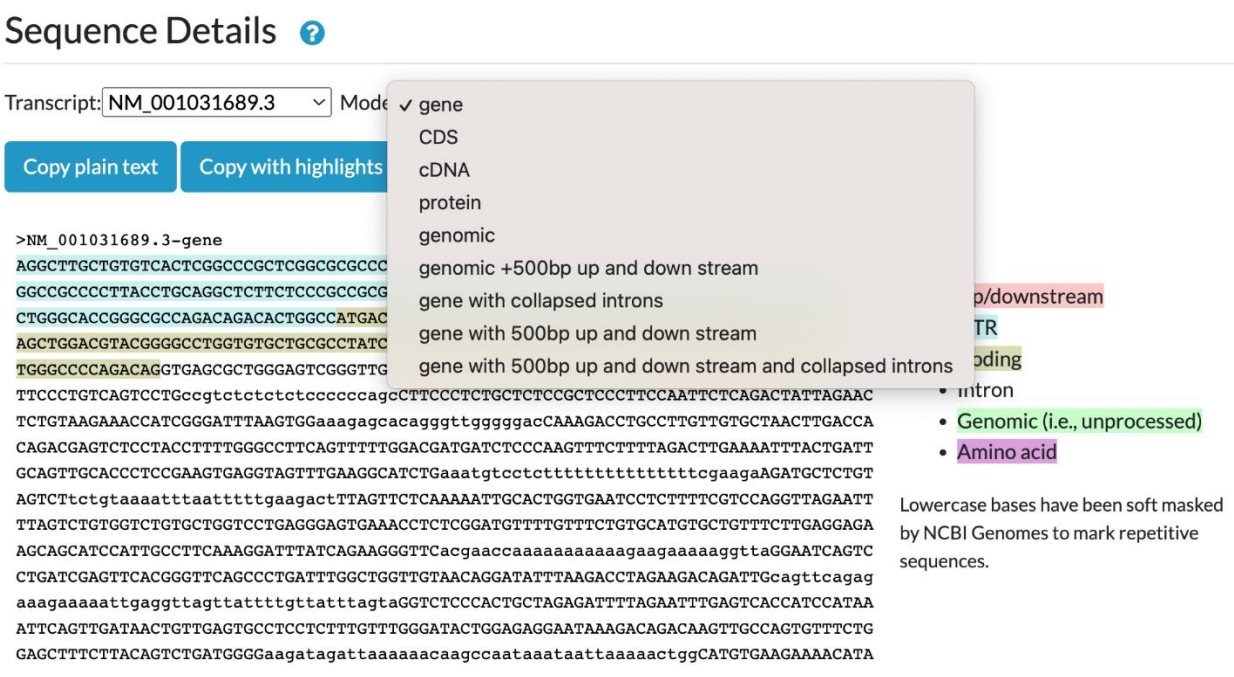

Figure 3
165x87 mm (x DPI)

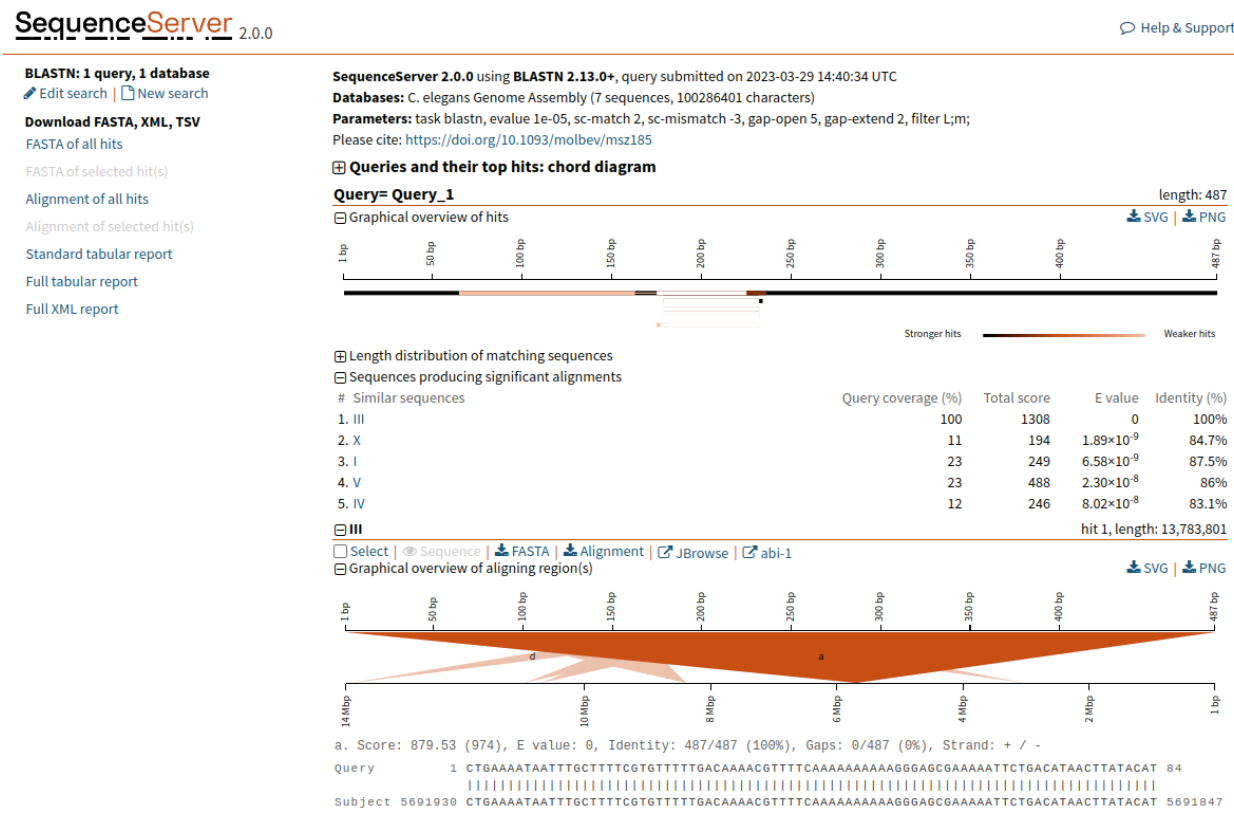

Figure 4
165x107 mm (x DPI)

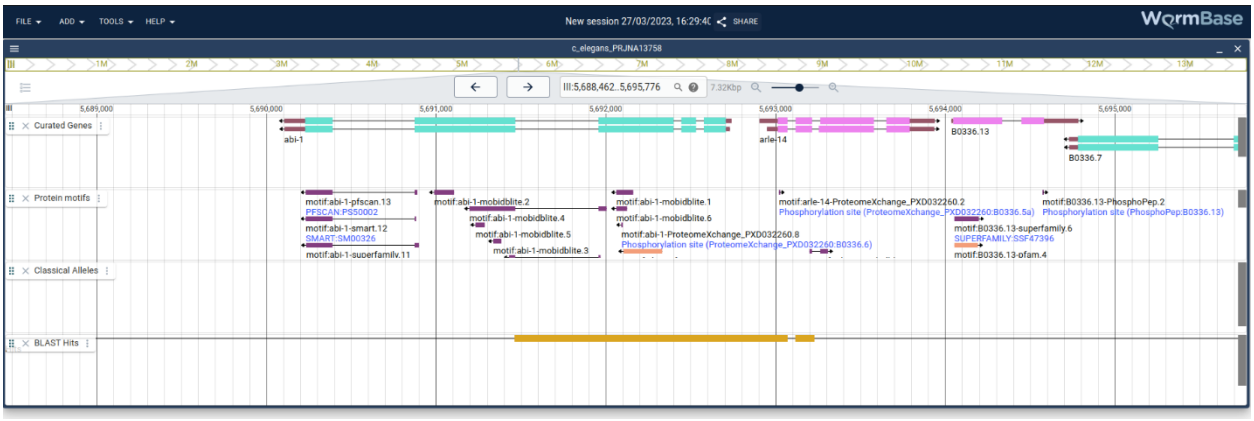

Figure 5
165x55 mm (x DPI)

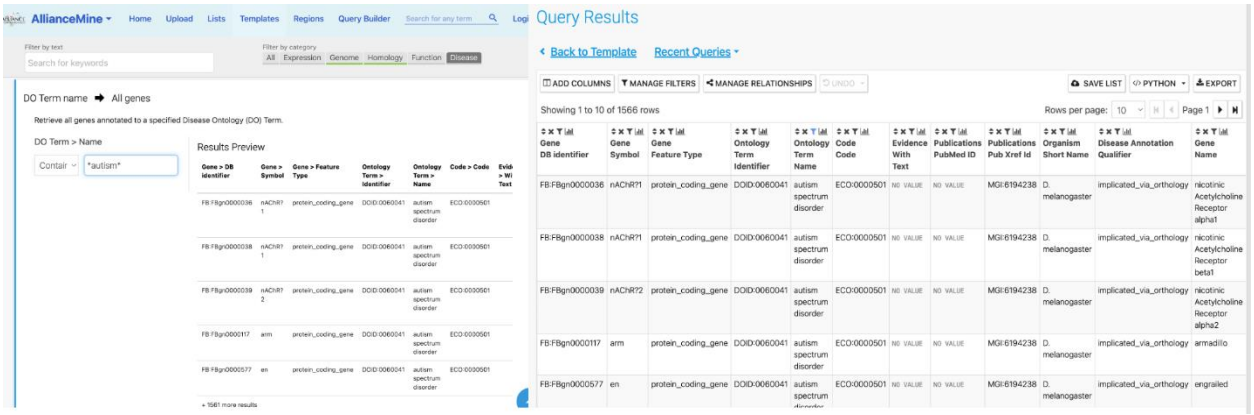

Figure 6
165x54 mm (x DPI)

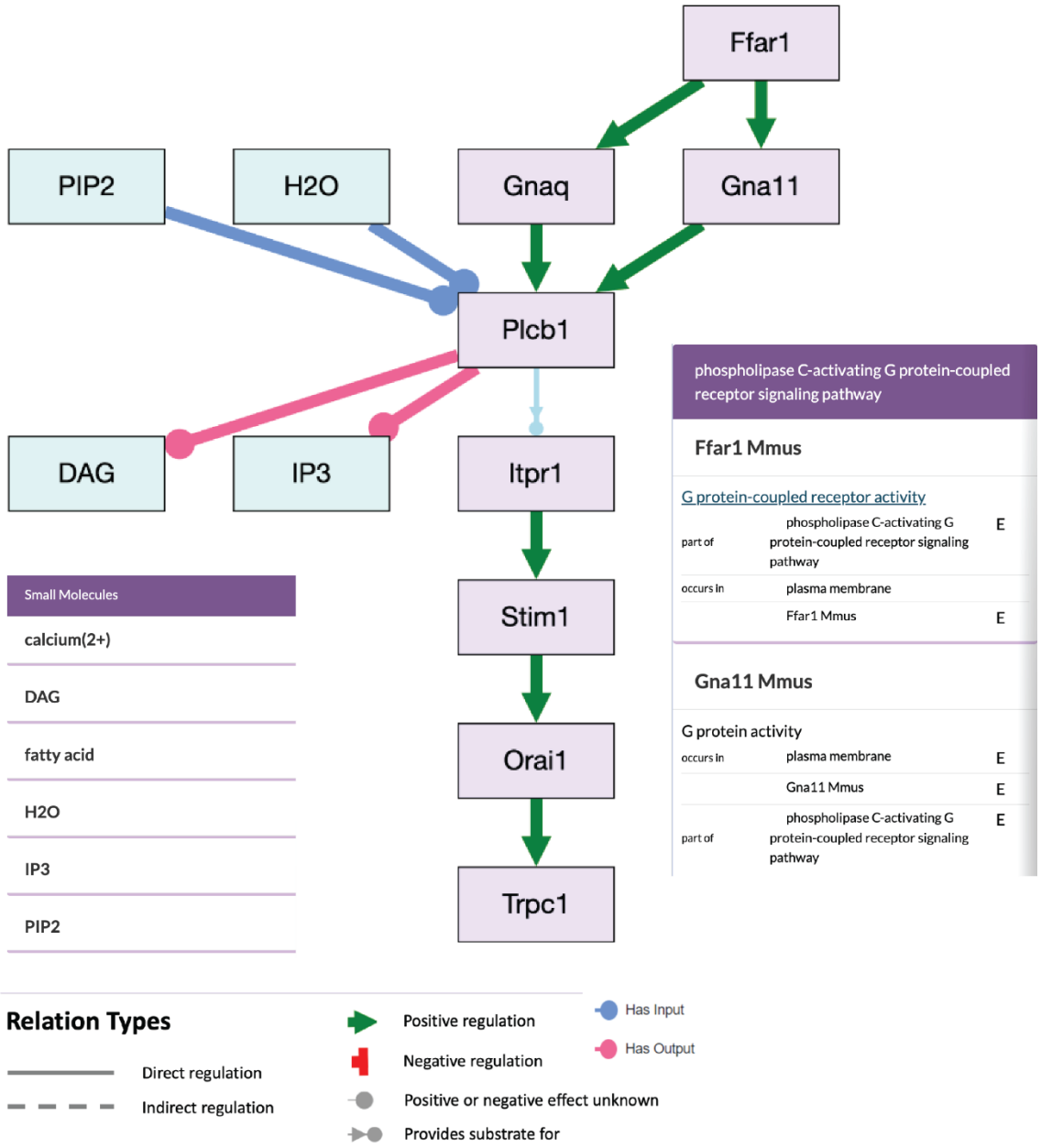

Figure 7
165x173 mm (x DPI)

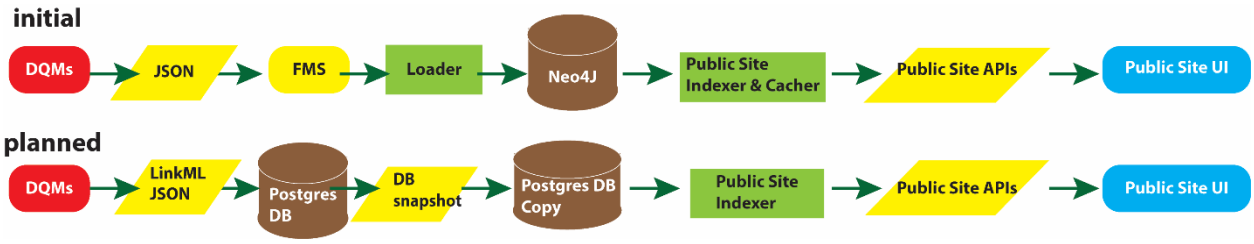

Figure 8
165x31 mm (x DPI)

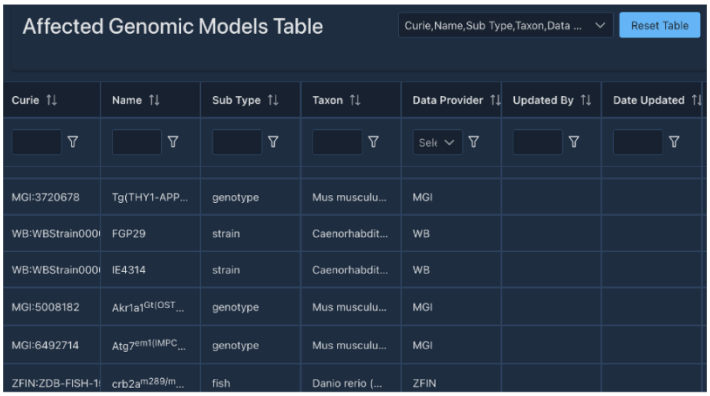

Figure 9
93x51 mm (x DPI)

Figure 10
61x34 mm (x DPI)

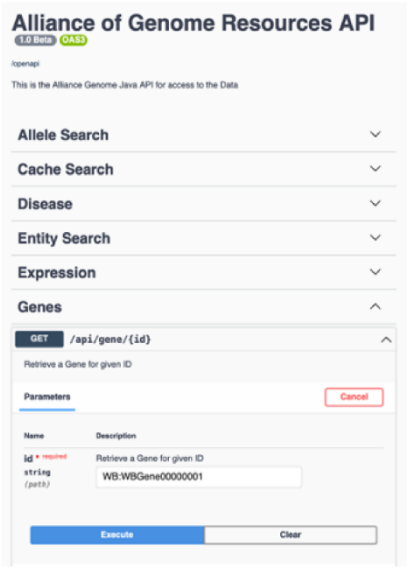

Figure 11
54x75 mm (x DPI)

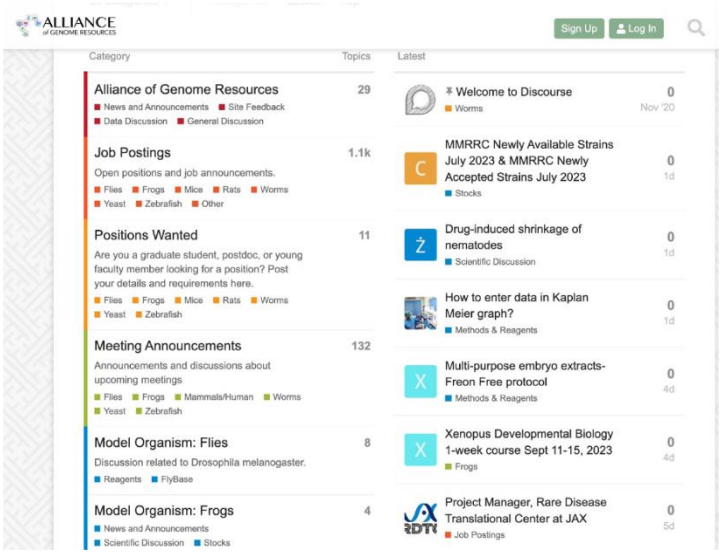

Figure 12
94x72 mm (x DPI)

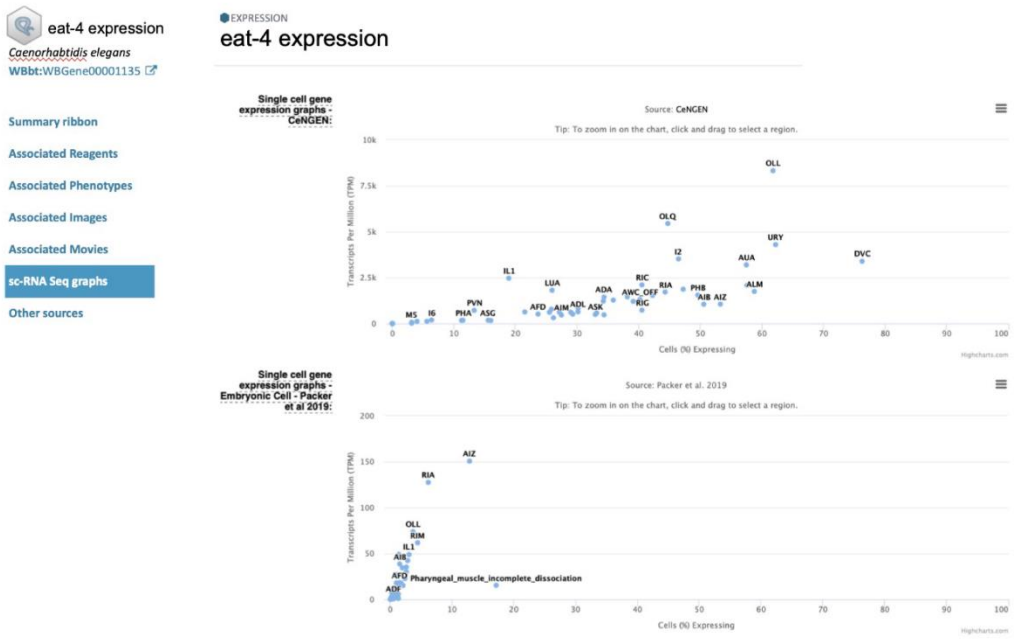

Figure 13
142x91 mm (x DPI)

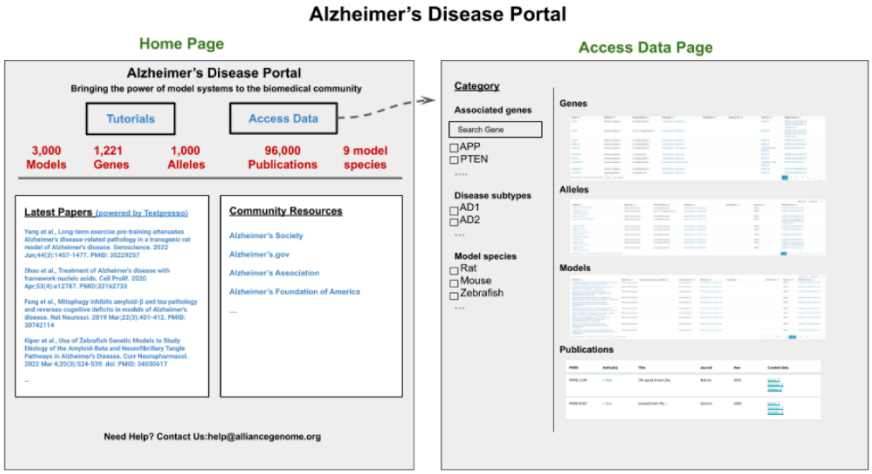

Figure 14
124x70 mm (x DPI)