

1
2 **Updates to the Alliance of Genome Resources Central Infrastructure**

3
4 **Alliance of Genome Resources Consortium**

5
6 **The Alliance of Genome Resources Consortium (alphabetical)**

7 Suzanne A. Aleksander⁹, Anna V. Anagnostopoulos⁵, Giulia Antonazzo¹⁰, Valerio Arnaboldi¹,
8 Helen Attrill¹⁰, Andrés Becerra², Susan M. Bello⁵, Olin Blodgett⁵, Yvonne M. Bradford¹¹, Carol J.
9 Bult⁵, Scott Cain⁸, Brian R. Calvi⁴, Seth Carbon⁶, Juancarlos Chan¹, Wen J. Chen¹, J. Michael
10 Cherry⁹, Jaehyoung Cho¹, Madeline A. Crosby³, Jeffrey L. De Pons⁷, Peter D'Eustachio¹⁵,
11 Stavros Diamantakis², Mary E. Dolan⁵, Gilberto dos Santos³, Sarah Dyer², Dustin Ebert¹², Stacia
12 R. Engel⁹, David Fashena¹¹, Malcolm Fisher¹⁶, Saoirse Foley¹³, Adam C. Gibson⁷, Varun R.
13 Gollapally⁷, L. Sian Gramates³, Christian A. Grove¹, Paul Hale⁵, Todd Harris⁸, G. Thomas
14 Hayman⁷, Yanhui Hu¹⁴, Christina James-Zorn¹⁶, Kamran Karimi¹⁷, Kalpana Karra⁹, Ranjana
15 Kishore¹, Anne E. Kwitek⁷, Stanley J. F. Laulederkind⁷, Raymond Lee¹, Ian Longden³, Manuel
16 Lupaert², Nicholas Markarian¹, Steven J. Marygold¹⁰, Beverley Matthews³, Monica S.
17 McAndrews⁵, Gillian Millburn¹⁰, Stuart Miyasato⁹, Howie Motenko⁵, Sierra Moxon⁶, Hans-
18 Michael Muller¹, Christopher J. Mungall⁶, Anushya Muruganujan¹², Tremayne Mushayahama¹²,
19 Robert S. Nash⁹, Paulo Nuin⁸, Holly Paddock¹¹, Troy Pells¹⁷, Norbert Perrimon¹⁴, Christian
20 Pich¹¹, Mark Quinton-Tulloch², Daniela Raciti¹, Sridhar Ramachandran¹¹, Joel E. Richardson¹¹,
21 Susan Russo Gelbart³, Leyla Ruzicka¹¹, Gary Schindelman¹, David R. Shaw⁵, Gavin Sherlock⁹,
22 Ajay Shrivatsav⁹, Amy Singer¹¹, Constance M. Smith⁵, Cynthia L. Smith⁵, Jennifer R. Smith⁷,
23 Lincoln Stein⁸, Paul W. Sternberg¹, Christopher J. Tabone³, Paul D. Thomas¹², Ketaki Thorat⁷,
24 Jyothi Thota⁷, Monika Tomczuk⁵, Vitor Trovisco¹⁰, Marek A. Tutaj⁷, Jose-Maria Urbano¹⁰,
25 Kimberly Van Auken¹, Ceri E. Van Slyke¹¹, Peter D. Vize¹⁷, Qinghua Wang¹, Shuai Weng⁹,
26 Monte Westerfield¹¹, Laurens G. Wilming⁵, Edith D. Wong⁹, Adam Wright⁸, Karen Yook¹, Pinglei
27 Zhou³, Aaron Zorn¹⁶, Mark Zytkovicz³

28
29 ¹Division of Biology and Biological Engineering 140-18, California Institute of Technology,
30 Pasadena, CA 91125, USA.

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

33 ³The Biological Laboratories, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138,
34 USA.

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

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

37 ⁶Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory,
38 Berkeley, CA

39 ⁷Medical College of Wisconsin - Rat Genome Database, Departments of Physiology and
40 Biomedical Engineering, Medical College of Wisconsin, Milwaukee, WI 53226, USA.

41 ⁸Informatics and Bio-computing Platform, Ontario Institute for Cancer Research, Toronto, ON
42 M5G0A3, Canada.

43 ⁹Department of Genetics, Stanford University, Stanford, CA 94305

44 ¹⁰Department of Physiology, Development and Neuroscience, University of Cambridge,

45 Downing Street, Cambridge CB2 3DY, UK.
46 ¹¹Institute of Neuroscience, University of Oregon, Eugene, OR 97403
47 ¹²Department of Population and Public Health Sciences, University of Southern California, Los
48 Angeles, CA 90033, USA
49 ¹³Department of Biological Sciences, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh,
50 PA 15203
51 ¹⁴Department of Genetics, Howard Hughes Medical Institute, Harvard Medical School, 77
52 Avenue Louis Pasteur, Boston, MA 02115, USA
53 ¹⁵NYU Grossman School of Medicine, New York NY 10016
54 ¹⁶ Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, 3333
55 Burnet Ave, Cincinnati, OH 45229, USA
56 ¹⁷ Department of Biological Sciences, University of Calgary, 507 Campus Dr NW, Calgary, AB
57 T2N 4V8, Canada
58
59

60 **correspondence:** Paul W. Sternberg pws@caltech.edu

61

62 **Abstract**

63 The Alliance of Genome Resources (Alliance) is an extensible coalition of knowledgebases
64 focused on the genetics and genomics of intensively-studied model organisms. The Alliance is
65 organized as individual knowledge centers with strong connections to their research
66 communities and a centralized software infrastructure, discussed here. Model organisms
67 currently represented in the Alliance are budding yeast, *C. elegans*, *Drosophila*, zebrafish, frog,
68 laboratory mouse, laboratory rat, and the Gene Ontology Consortium. The project is in a rapid
69 development phase to harmonize knowledge, store it, analyze it, and present it to the
70 community through a web portal, direct downloads, and APIs. Here we focus on developments
71 over the last two years. Specifically, we added and enhanced tools for browsing the genome
72 (JBrowse), downloading sequences, mining complex data (AllianceMine), visualizing pathways,
73 full-text searching of the literature (Textpresso), and sequence similarity searching
74 (SequenceServer). We enhanced existing interactive data tables and added an interactive table
75 of paralogs to complement our representation of orthology. To support individual model
76 organism communities, we implemented species-specific “landing pages” and will add disease-
77 specific portals soon; in addition, we support a common community forum implemented in
78 Discourse. We describe our progress towards a central persistent database to support curation,
79 the data modeling that underpins harmonization, and progress towards a state-of-the art
80 literature curation system with integrated Artificial Intelligence and Machine Learning (AI/ML).

81

82

83

84

85

86

87

88 **Introduction**

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

95

96 The Alliance of Genome Resources (referred to herein as the Alliance) is a consortium of MODs
97 and the Gene Ontology Consortium (GOC). The mission of the Alliance is to support
98 comparative genomics as a means to investigate the genetic and genomic basis of human
99 biology, health, and disease. To promote sustainability of the core community data resources
100 that make up the Alliance, we implemented an extensible “knowledge commons” platform for
101 comparative genomics built with modular, re-usable infrastructure components that can support
102 informatics resource needs across a wide range of species (Alliance of Genome Resources,
103 2022; Howe et al., 2018; Bult and Sternberg, 2023). In 2022, the Alliance was recognized as a
104 Core Global Biodata Resource by the Global Biodata Coalition (Anderson et al 2017).

105

106 Specifically, the Alliance of Genome Resources is organized as two interdependent units:
107 Alliance Central and the Alliance Knowledge Centers. **Alliance Central** is responsible for
108 developing and maintaining the software for data access and for the coordination of data
109 harmonization and data modeling activities across our members. A primary goal of Alliance
110 Central is to reduce redundancy in systems administration and software development for model
111 organism knowledgebases and to deploy a unified ‘look and feel’ for access to, and display of,
112 common data types and annotations across diverse model organisms and human, following
113 Findability, Accessibility, Interoperability, and Reuse (FAIR) guiding principles. Model organism-
114 specific knowledgebases serve as **Alliance Knowledge Centers**. Knowledge Centers are
115 responsible for expert curation and submission of data to Alliance Central using Alliance Central
116 infrastructure. Knowledge Centers also are responsible for organism-specific user support
117 activities and for providing access to data types not yet supported by Alliance Central. The
118 founding Alliance Knowledge Centers are *Saccharomyces* Genome Database (Engel et al.
119 2022), WormBase (Davis et al. 2022), FlyBase (Gramates et al 2022), Mouse Genome
120 Database (Ringwald et al. 2022), the Zebrafish Information Network (Bradford et al. 2023), Rat
121 Genome Database (Vedi et al 2023), and the Gene Ontology Consortium (Gene Ontology
122 Consortium 2023). The newest member, Xenbase (Fisher et al, 2023), joined the Alliance
123 consortium in 2022.

124

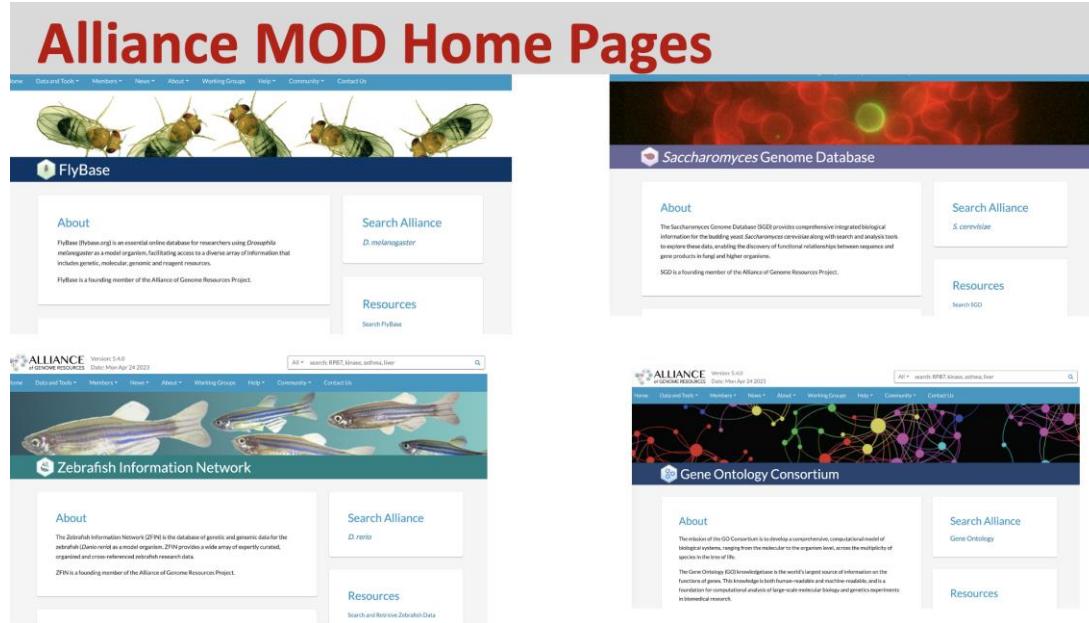
125 Here we describe our progress toward harmonizing information provided by our member
126 resources, our development of a software infrastructure for ingest, curation, storage, analysis,
127 and output of such information, and development of an efficient literature curation system. We
128 also describe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

129

130 **Community Homepages**

131 The Alliance website features landing pages for each model organism in the Alliance

132 consortium. These pages are accessed from the “Members” drop-down menu in the header on
133 every Alliance page. These pages feature MOD-specific-content such as meetings, news, and
134 other MOD-specific resource links. A common template allows users to find the same types of
135 information in each landing page (**Figure 1**). As MODs transition their data and web services to
136 the Alliance, their member pages will evolve into portals hosting additional MOD-specific data,
137 tools, and links to organism-specific resources and will also accommodate the many unique
138 data and tools from individual MODs.
139



140
141 **Figure 1. MOD landing pages at the Alliance Portal.** A common look and feel that allows community-
142 specific content.
143

144 **Paralogy**

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

161
162 The paralogy section is composed of a table (**Figure. 2**), similar to the orthology table, which
163 contains the gene symbol of related paralogs, a calculated rank, alignment length as the
164 number of aligned amino acids, percentage of similarity and identity, and a count of the
165 algorithms or methods which call the paralogous match. The ranking score was developed to
166 sort the paralogs by overall similarity, and was reviewed by curators to display optimally an
167 acceptable rank order for well-studied sets of paralogs. The ranking score considers several
168 factors, including alignment length, percent identity, and the number of paralogy methods that
169 identify the paralog. Additional Information for rank determination and alignment length are
170 available to the users via a clickable help icon located next to those column headers.
171
172 The paralog section was released with Alliance version 6.0.0. Future updates will include the
173 ability to sort and filter the table by column values and the availability of these data via our bulk
174 downloads page. The existing tables on the gene pages for Function, Disease, and Expression
175 all contain checkboxes for "Compare Ortholog Genes" that allow users to search across species
176 for these features. We will add the additional checkbox, "Compare Paralog Genes" to provide
177 similar functionality for paralogous genes in a future Alliance release.
178

Paralogy

Gene symbol	Rank 	Alignment Length (aa) 	Similarity %	Identity %	Method Count	Method								
						Ensembl Compara	HGNC	InParanoid	OMA	OrthoFinder	OrthoInspector	PANTHER	PhylomeDB	SonicParanoid
hlh-27	1	268	99	99	3 of 8	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
hlh-28	2	277	55	39	4 of 8	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
hlh-29	3	279	54	38	4 of 8	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
hlh-26	4	274	48	32	4 of 8	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
ref-1	5	353	38	25	2 of 8	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

179 **Figure 2. Paralog table for *C. elegans* *hlh-25*.** The table presents a ranking of paralogs for the *hlh-25*
180 gene, based on a weighted scoring algorithm that incorporates sequence conservation metrics. It lists the
181 gene symbols, provides the alignment length in amino acids, and quantifies the similarity and identity
182 percentages of genes paralogous to *hlh-25*. The methodology count, indicating the number of algorithms
183 supporting the paralogous relationship, is also included. In this ranking, *hlh-27* is identified as the primary
184 paralog due to its high similarity and identity scores, despite being recognized by fewer methods than *hlh-28*.
185
186

Xenopus in the Alliance

187 Xenbase (Fisher et al 2023), the *Xenopus* knowledgebase, is the first knowledgebase to join the
188 Alliance since the founding members initiated the consortium. *Xenopus* is an amphibian frog

190 species used extensively in biomedical research, and in particular for experimental embryology,
191 cell biology, and disease modeling with genome editing (Carotenuto et al., 2023; Kostiuk and
192 Khokha, 2021). As a non-mammalian air-breathing tetrapod, *Xenopus* represents a valuable
193 evolutionary transition between rodents and zebrafish for comparative genomic studies.
194 Xenbase is a large-scale knowledgebase built on a Chado schema foundation, so has design
195 features related to Alliance foundational members. As a model system, two different *Xenopus*
196 species are used interchangeably; *X. tropicalis* is a diploid that is the preferred system for
197 genome editing and genetics, whereas *X. laevis* is an allotetraploid preferred for use in cell
198 biology studies, microinjection, and microsurgery-style experimentation. *X. tropicalis* has 1:1
199 relationships between most genes and human orthologs (excluding paralogs) (Mitros et al.,
200 2019), whereas *X. laevis* has two copies of most human orthologs. The allotetraploid formed via
201 hybridization of two different frog species (Session et al 2016), and the complexities of genome
202 evolution that subsequently occurred increase the difficulty of identifying orthology of the two *X.*
203 *laevis* genes to their diploid relatives, including humans. Mapping of the diploid *X. tropicalis*
204 genes to their human orthologs was performed with DIOPT, similarly to other model organisms
205 in the Alliance. Because this method does not yet work in the context of an allotetraploid, the
206 Alliance imports the *X. tropicalis* to *X. laevis* paralogy mappings from Xenbase, where they have
207 been established through a combination of synteny analysis and manual curation. Dealing with
208 how to incorporate the two new species with this ploidy complexity was one of the major
209 challenges of adding *Xenopus* to the Alliance.
210

211 On the Xenbase side, as the new member team, Xenbase staff created exporters to upload
212 content, on a regular schedule, formatted in a manner defined by the Alliance data ingest
213 schema and using the Alliance File Management System and API access keys. Currently these
214 data include orthology, the *Xenopus* anatomical ontology, standard gene information, gene
215 expression data, publications, GO term associations, disease associations, anatomical
216 phenotypes, genome details in the Alliance browser, and BLAST capacity. *Xenopus* genes can
217 be found using the Alliance landing page search tool with *Xenopus* genes flagged by Xtr and
218 Xla notations. The two copies of the genes in *X. laevis*, the allotetraploid, are further tagged as
219 '(symbol).L' and '(symbol).S' to denote the genes on the long (L) and short (S) chromosome
220 pairs of this species (e.g., *pax6.L* and *pax6.S*). Alliance release 6.0.0 has Xenbase data for
221 54,000 genes, 19,000 disease associations, over 45,000 gene expression records and more
222 than 7,000 anatomical phenotypes. Expression and phenotype data will be available soon.
223

224 In addition to the rich data made available to the Alliance from *Xenopus* research, this effort also
225 served as a valuable test case for understanding the level of effort and complexities engendered
226 in the addition of new knowledgebases to the Alliance, and the functionality and adaptability of
227 ingest system components.
228

229 **JBrowse sequence detail widget**

230 Delivered in the recent Alliance 6.0.0 release, the "Sequence Detail" section of all gene pages
231 now uses JBrowse and javascript libraries to display an interactive widget that allows users to
232 download DNA and amino acid sequences of genes in several possible configurations: genomic
233 sequence highlighted with UTR, coding and intronic regions, CDS regions, and translated

234 protein for example (**Figure 3**). We will extend the functionality of the widget variant detail
235 pages, where both the wild-type and variant sequences will be provided. When the variant
236 occurs in the context of a protein coding gene, changes to the coding sequence and resulting
237 translated protein will also be displayed and available for download.

238

Sequence Details ?

239

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.

243

244 Model organism BLAST

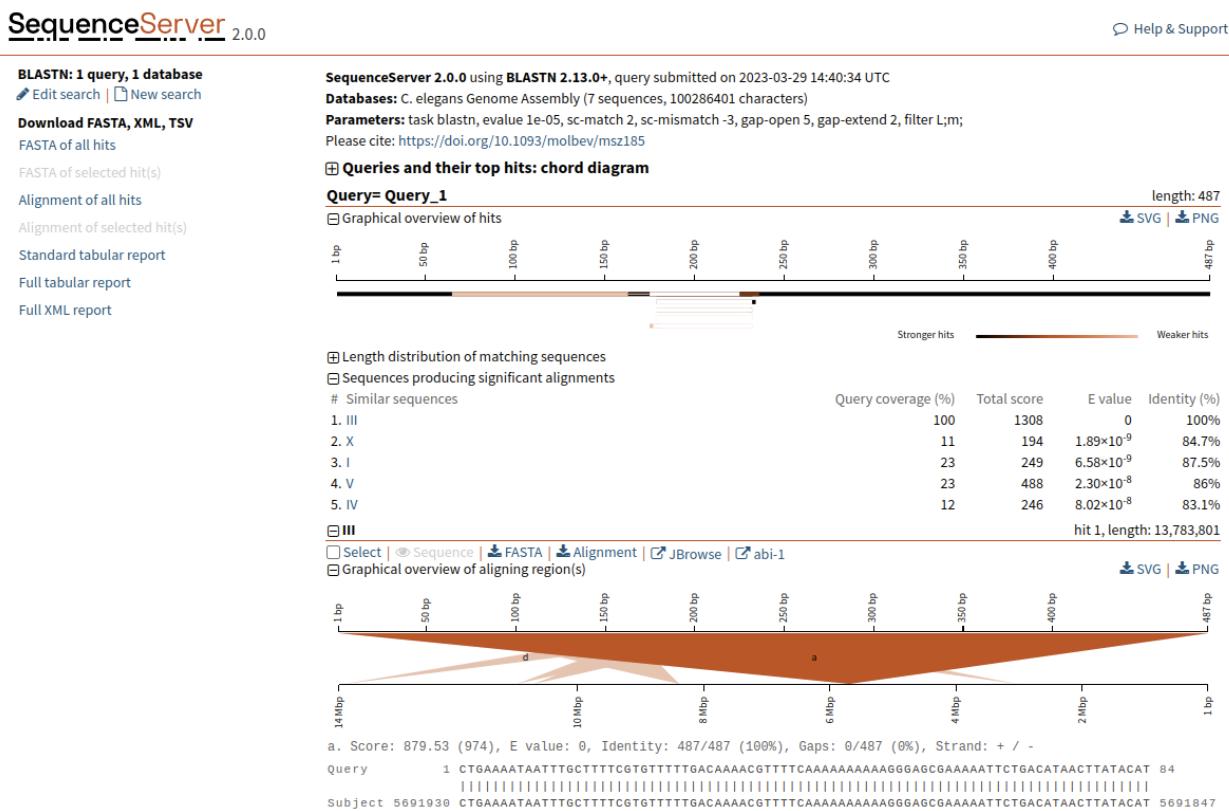
245 For more than two decades, some of the MOD members of the Alliance have hosted their own
246 custom BLAST interfaces (Altschul et al., 1990; e.g., FlyBase Consortium. 1999), which have
247 allowed users to search custom databases related to those model organisms, e.g., subsets of
248 related species or molecular clones and display BLAST hits in Genome Browsers aligned with
249 current gene models. We are now developing an updated and integrated Alliance BLAST that
250 optimizes sequence analysis across model organisms, and we have begun to update BLAST at
251 individual MODs. The new WormBase BLAST is now available online, and simultaneously, the
252 FlyBase BLAST system has been replaced and is currently online, with management facilitated
253 through configuration files on GitHub.

254

255 The Alliance BLAST will significantly improve the user experience. We envision that BLAST
256 systems, currently powered by SequenceServer (Priyam et al. 2019), will deliver an integrated
257 interface by linking results to Genome Browsers and Alliance gene pages (**Figure 4**). This tight
258 connection allows users to navigate seamlessly between their BLAST results and the wealth of
259 information available within the Alliance, enhancing the efficiency and depth of genetic research.
260 For example, users can retrieve BLAST results for a sequence of interest and then easily
261 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.

271

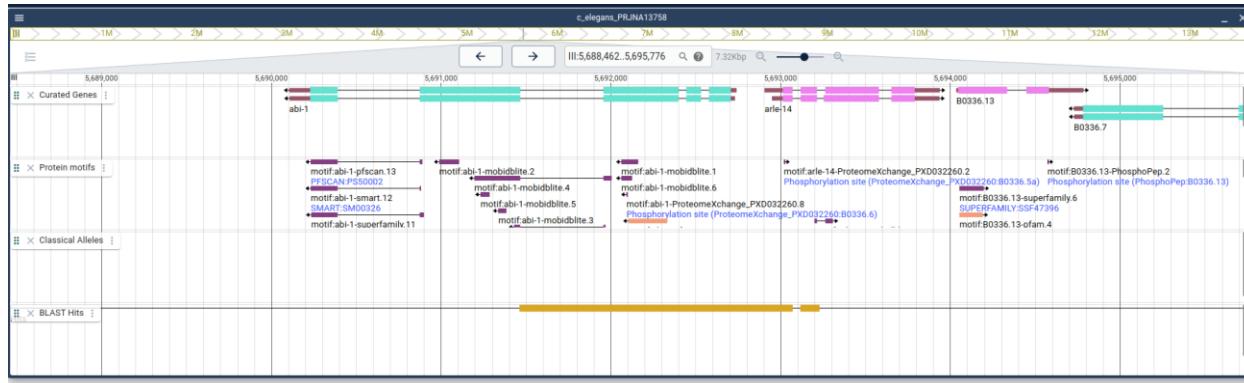


272

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

277

278



279
280
281
282

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.

283

284 **AllianceMine**

285 AllianceMine, a sophisticated, multifaceted search and retrieval tool that utilizes the InterMine
286 software (Smith et al., 2012), offers a unified view of harmonized data, enabling advanced
287 queries across multiple species. For instance, gene lists can be processed as input and
288 simultaneously query different annotations, such as 'Show me genes associated with a (specific
289 disease term)' (**Figure 6**). The results from queries can be combined for further analysis, and
290 saved or downloaded in customizable file formats. Queries themselves can be customized by
291 modifying predefined templates or by creating new templates to access a combination of
292 specific data types. Thus, this powerful tool can be used in multiple ways - for search, discovery,
293 curation, and analysis.

294

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

302

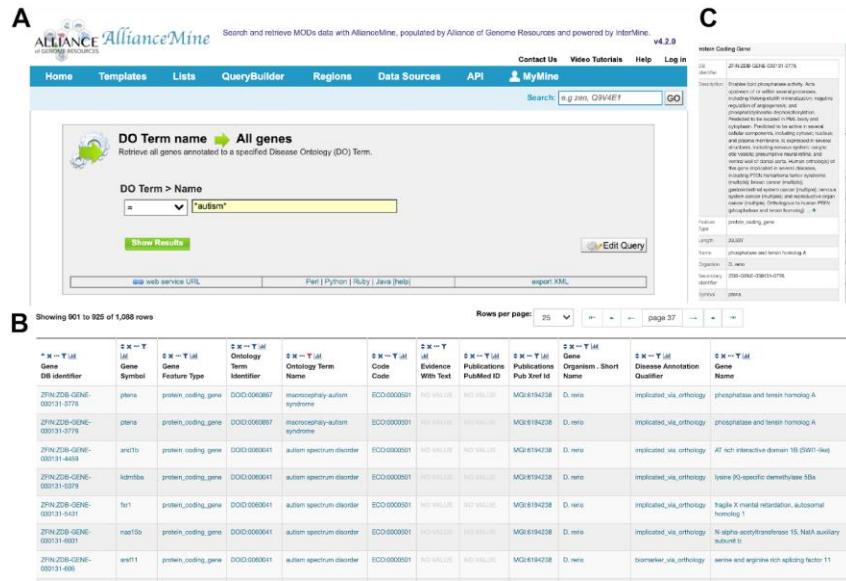


Figure 6. AllianceMine example. Using a simple template, a disease ontology (DO) term is chosen, and all genes associated with this DO term are returned in a downloadable table.

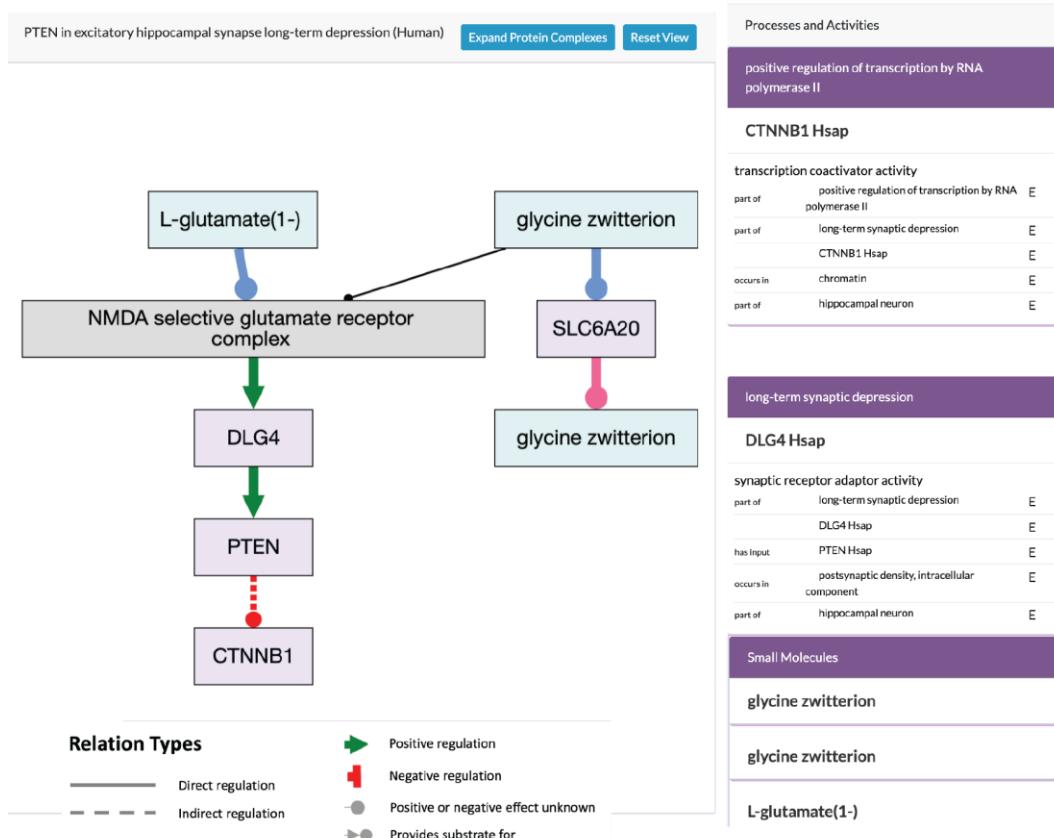
SimpleMine

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, 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 have implemented a pathway display on Alliance gene pages, which presents both GO-CAM (Thomas et al., 2019) and Reactome pathway (Milacic et al. 2024) models. The display queries both the Reactome and GO 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.

333 For signaling pathways, we distinguish between direct and indirect regulation, and between
334 positive, negative, or unknown effects.
335



336
337
338 **Figure 7. Alliance Pathway Viewer.** The pathway widget displays gene products (light purple
339 rectangles), protein complexes (light grey rectangles) and chemicals (light blue rectangles) and the flow
340 of information and material between them (relations). These relations, shown in legend indicate direct or
341 indirect regulation that can be positive, negative or of unknown effect direction.
342

343 **Harmonized Data Models**

344 A key requirement of the data transition from individual MODs to the Alliance infrastructure is
345 collective data harmonization so that existing analogous MOD data classes (types/tables) can
346 be loaded into Alliance databases using a consistent data schema and language for
347 communicating about such data classes. A first step in this process, curators from each
348 Alliance knowledge center communicate about their respective existing data classes with an aim
349 of agreeing on which data classes are analogous and therefore should be treated as a single,
350 consolidated data class in the Alliance infrastructure. Next, curators need to align the properties
351 (table columns) of the consolidated data class to agree on property identity alignment and basic
352 data structure including whether these properties are required and/or defining, what values
353 should be stored for these properties, and whether these entity-property-value
354 associations/triples require their own respective metadata and/or evidence records. We use a
355 data modeling language, the Linked Data Modeling Language (LinkML) for these purposes.

356 Over the last two years, Alliance LinkML modelers have converged on common data modeling
357 patterns that can be reused for each class and property based on the nature of each class
358 property, enabling a standard workflow and implementation to be followed in each case. The
359 LinkML specifications, authored in human-readable YAML files, are used to (programmatically)
360 generate JSON schema specifications that Data Quartermasters (DQMs) can use to generate
361 and validate data files to be submitted to the persistent store. These specifications also inform
362 curation software developers how to generate initial backend (Java models and APIs) and front
363 end infrastructure (curation user interface data tables and detail pages) to be populated once
364 such code is deployed to the production environment of the curation tool and DQMs are ready
365 with their data files. Once DQMs have submitted their data files for a particular data class, the
366 data are loaded into the persistent store with a number of validation and reporting steps (see
367 persistent store architecture description below) and should automatically be populated into the
368 respective data tables and detail pages in the curation interface. The data, having been
369 harmonized, ingested, validated, and displayed to curators in the curation software, can now
370 flow through to the public site according to the data pipeline described (see persistent store
371 architecture description below).

372

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

385

386 **Persistent Store architecture**

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

397

398

399

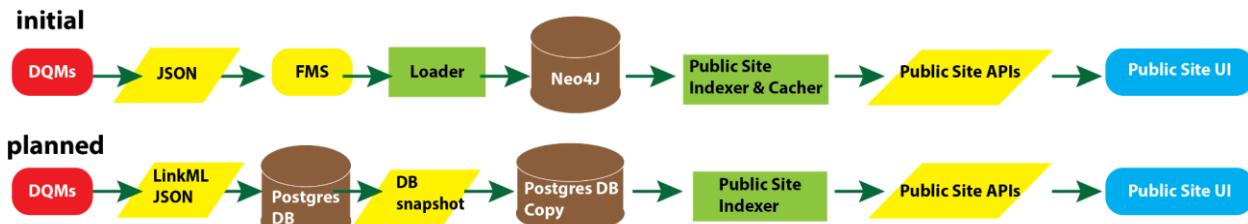


Figure 8. Evolution of Data Flow. Graphical summary showing the design of short term infrastructure initially deployed to support rapid delivery of unified data to the community and the planned production system. Red, data quartermasters at MODs; Yellow, data; Brown, database; Green, transformations;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

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. All of 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 via a cron job to ensure that the latest versions of such data are current. Because 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, particular logistics need to be addressed. These include recognizing that any discrepancies between data previously submitted by a MOD and data newly submitted from the MOD need to be cleaned up programmatically by removing entities in the database not also submitted in the latest file submission.

To enable create-read-update-delete (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

439 transcripts, proteins, other (non-gene) genome features, expression annotations, phenotype
440 annotations, molecular interactions, genetic interactions, gene regulation annotations,
441 antibodies, images, and more. In addition to data tables presenting all entries of a particular
442 data class, the curation tool also has individual entity detail pages (for example, see an allele
443 detail page <https://curation.alliancegenome.org/#/allele/MGI:6446761>) for data entry and editing
444 on a dedicated web page for one particular entity. The curation tool also enables user-specific
445 and MOD-specific custom user settings and preferences to provide a user interface most
446 compatible with individual curators' workflows.
447

Affected Genomic Models Table						
Curie ↑↓	Name ↑↓	Sub Type ↑↓	Taxon ↑↓	Data Provider ↑↓	Updated By ↑↓	Date Updated ↑↓
	Y		Y		Selk	Y
MGI:3720678	Tg(THY1-APP...	genotype	Mus musculu...	MGI		
WB:WBStrain000	FGP29	strain	Caenorhabdit...	WB		
WB:WBStrain000	IE4314	strain	Caenorhabdit...	WB		
MGI:5008182	Akr1a1Gt(OST...	genotype	Mus musculu...	MGI		
MGI:6492714	Atg7 ^{em1} (IMPC...	genotype	Mus musculu...	MGI		
ZFIN:ZDB-FISH-1	crb2a ^{m289/m...}	fish	Danio rerio (...)	ZFIN		

448
449 **Figure 9.** Screenshot of the Alliance curation tool interface showing an example of curated annotations of
450 Affected Genomic Models managed in the persistent store.
451

452 Future development plans for the curation tool include: batch creation of data entities (e.g.,
453 annotations, reagents), batch editing, data history inspection and auditing, undo and review of
454 latest changes, publication constraints (constrain data view and entry to publication currently
455 being curated), customizations and MOD default settings for new entity creation and detail
456 pages, incorporation of data entity and topic tagging information from the ABC literature store,
457 and incorporation of AI/ML into the curation workflow.
458

459 For releases of persistent store data to the Alliance public website, Postgres database
460 snapshots are taken and sent to a separate Postgres instance that feeds the data via the
461 curation APIs (instantiated as a library) into the public site indexer where various data filtering
462 and transformations occur before making those processed data available to our public website
463 APIs via our Elasticsearch index. The Alliance public website user interface, using existing UI
464 infrastructure, is then modified or created to accommodate the data now flowing from the
465 persistent store database.
466

467 **Security, stability and backups**

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

479

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

494

495 **Literature Acquisition**

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

501

502 During the past year, we focused on literature acquisition. Literature acquisition at the Alliance
503 begins with automated, organism-specific PubMed queries to retrieve candidate references for
504 each MOD's corpus. References matching the search criteria are then added to the ABC by
505 assigning an Alliance reference id and importing associated bibliographic information to the
506 database. Subsequently, curators manually sort references as either 'in' or 'out of corpus' based
507 on the curation policies of the MOD and eliminate any false positive results from the initial
508 search. Once references are sorted, they enter MOD-specific curation workflows supported by
509 task-specific ABC curator interfaces to, for example, add reference files, manually tag
510 references with specific entities (e.g., genes, alleles, and data types) and topics (e.g.,

511 phenotypes, anatomic expression) using the Alliance Tags for Papers (ATP) ontology, and
512 merge duplicate references. In addition to adding reference files manually, the full text of 'in
513 corpus' references included in the PubMed Central (PMC) open access set is also automatically
514 downloaded. Curators may also use the ABC to add non-PubMed references. An additional key
515 feature of the ABC is a search interface that allows curators to retrieve references based on
516 various criteria including their in/out of corpus status, bibliographic data, and publication data
517 range, if desired. Reference acquisition functionality can easily be extended to integrate
518 additional MODs into the Alliance infrastructure.

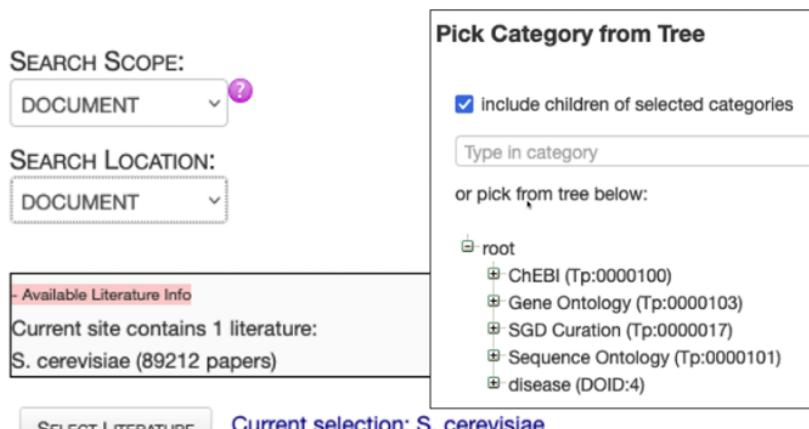
519
520 To facilitate reference data exchange between the Alliance and MOD databases, the MODs
521 provide a mapping file that associates MOD reference CURIEs (Compact Uniform Resource
522 Identifier) with PMIDs, e.g., ZFIN:ZDB-PUB-181026-2 - PMID:30352852. The MODs also
523 provide reference CURIEs and data for references not included in PubMed but used by the
524 MOD, such as internal curation references and those published in a journal not yet indexed at
525 PubMed.

526
527 Over the past 25-30 years, Alliance member databases have independently developed methods
528 to acquire, triage, and curate their respective literatures. Having implemented a common
529 literature curation interface, database, and full text acquisition system, the ABC is now poised to
530 expand its functionality by incorporating ML methods developed by, and in production for, a
531 subset of Alliance members to all groups. For example, automated pipelines that recognize
532 entities (e.g., genes, alleles, strains) as well as data types (e.g., phenotype, genetic interactions)
533 can be developed for new groups with results stored centrally in the Alliance literature database.
534 Incorporating more automated methods will allow faster association of the published literature
535 with relevant biological concepts, information that can be displayed on future Alliance
536 references pages while the papers await detailed full curation. Centralized literature
537 infrastructure will also support other curation pipelines, such as community curation by authors,
538 which can then be more readily implemented for additional Alliance member communities thus
539 providing another avenue by which curated data can be swiftly included in the Alliance. Lastly,
540 the common literature tool will allow Alliance biocurators to coordinate curation of multi-species
541 references that will provide users a facile way to find and view cross-species research exploiting
542 the strengths of each Alliance model organism, a primary goal of the Alliance.

543
544 **Textpresso**
545 Textpresso is a full-text literature search engine that gets power from its single-sentence scope,
546 focus on a specific model organism (or topic), and categories of semantically or biologically
547 related terms (**Figure 10**; Müller et al., 2004; Müller et al. 2018). It has been used extensively by
548 WormBase and SGD curators, as well as *C. elegans* and *S. cerevisiae* researchers in addition
549 to other MODs (Van Auken et al., 2012; Bowes et al., 2013)

550
551 The Alliance is committed to creating Textpresso instances tailored to the unique needs of each
552 member database, all of which will be managed within the Alliance software ecosystem and
553 connected to the ABC as a single reference data source. This will reduce the overhead of
554 managing Textpresso at individual MODs while also simplifying development and deployment of

555 new features. Users will benefit from simplified access to Textpresso from the Alliance website.
556 We also plan to integrate Textpresso searches further into specific Alliance web pages such as
557 gene or allele pages. Users will be able to obtain additional references to biological entities
558 through Textpresso searches, adding information from potentially non-curated literature to the
559 list of curated references currently linked on those pages. Textpresso will be available to
560 Alliance biocurators and to the general public through MOD-customized websites and via APIs
561 for programmatic access.
562



563
564 **Figure 10. Textpresso for SGD literature at the Alliance. ([http://sgd-](http://sgd-textpresso.alliancegenome.org/tpc/search)**
565 **[textpresso.alliancegenome.org/tpc/search](http://sgd-textpresso.alliancegenome.org/tpc/search)**)

566
567 **Artificial Intelligence (AI)**
568 The Alliance member MODs have a track record of implementing ML tools to enhance triage
569 and curation efficiency. Notable examples include RGD's early adoption of UIMA standards and
570 the development of the OntoMate system (Liu et al. 2015) , as well as WormBase's creation of
571 Textpresso (Mueller et al. 2004) and document classifiers for paper triage.
572
573 The rise of Large Language Models (LLMs), like BERT, short for Bidirectional Encoder
574 Representations from Transformers, and ChatGPT, has transformed the NLP landscape, but
575 questions about their accuracy and "hallucinations" remain. The Alliance aims to harness LLMs
576 for tasks such as document classification, Named Entity Recognition (NER), sentence
577 classification, assisted-triage and curation and to build a natural language query system to
578 simplify access to its underlying structured data.
579
580 In the realm of AI/ML, Alliance members have developed classifiers for determining with high
581 accuracy whether papers returned from automated PubMed queries should be kept in their
582 corpus or discarded. The Alliance is developing a central solution by providing this type of in/out
583 corpus classifier to all members.
584
585 Efforts are also underway to improve existing species-specific entity extraction and classification
586 models, with a focus on incorporating human feedback in the loop and continuously training
587 models based on data validated by professional biocurators and community curators. A
588 centralized interface for "topic and entity tag" addition and validation during triage and curation

589 is under development as part of the ABC. The interface allows curators to associate tags with
590 publications and at the same time validate (or invalidate) results extracted from AI/ML methods.
591 This interface will streamline the collection of valuable training and testing sets and will allow a
592 more systematic approach to the creation and comparison of different AI/ML models. Future
593 plans include development of tools for creating training sets and a model manager for tracking
594 ML models' performance. Integration with specialized biocuration tools such as Ontomate and
595 Textpresso is part of the strategy, with a vision of harmonizing AI/ML solutions across member
596 sites.

597

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

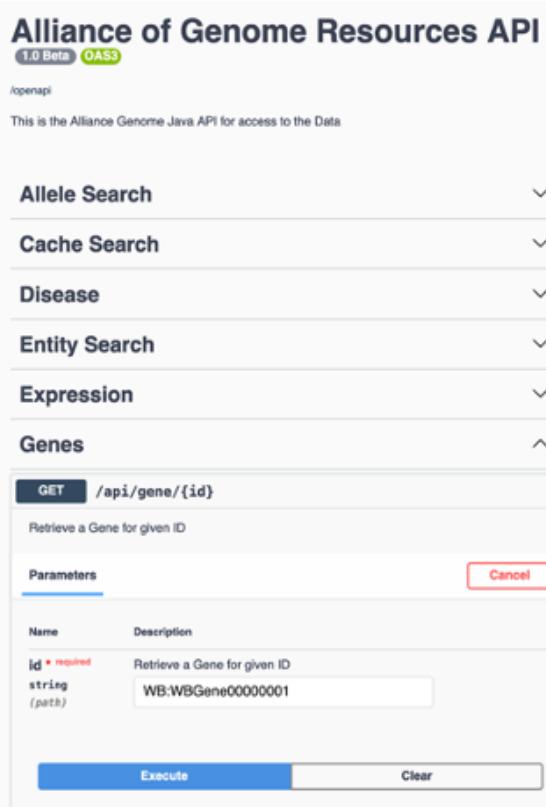
603

604 We will also explore the use of AI/ML in gene function summarization. Included on gene pages
605 at the Alliance are short textual gene summaries based on curated and structured data that
606 provide users a quick overview of gene function. The current automated system for generating
607 gene summaries has produced more than 160,000 summaries (Alliance version 6.0.0) for nine
608 species, including humans (Kishore et al., 2020). However, to increase the coverage of genes
609 further, we will explore the use of LLMs. This is especially relevant for less-studied genes with
610 few curated, structured data, and for scaling and upkeep of the summaries to match the rate of
611 new gene data from publications. We will use prompt engineering and finetuning of LLMs to
612 improve accuracy of the generated summaries. As part of a continual improvement process, we
613 will ask professional biocurators to evaluate summaries, and we will develop a scoring system
614 based on several features such as readability of summaries, inclusion of key gene data, and
615 checking for inaccurate and false data. To improve and keep gene summaries up to date, we
616 plan to retrieve newly published articles that contain gene data that were not available when the
617 LLM was trained and add extracted relevant text from the identified articles to the LLM prompt.
618 To do so, we will use tools such as Textpresso (Muller et al., 2004) and Ontomate (Liu et al.,
619 2015)

620

621 **Application Programming Interfaces (APIs)**

622 Application Programming Interfaces (APIs) are a key component of Alliance Central's data
623 services infrastructure for rapid, modular software development. We currently support a dozen
624 APIs with hundreds of endpoints (**Figures 11, 12**). New APIs will be added as data
625 harmonization and modeling of additional data entities are completed. We will expand public site
626 APIs to generate all data needed for SimpleMine, AllianceMine, etc. from single endpoints.
627 Current APIs include Public site APIs (agr_java_software in the GitHub repo) and APIs available
628 from a public Swagger UI page. Because the public APIs support only GET endpoints, they do
629 not require authentication. All APIs that support both GET and PUT/POST/DELETE endpoints
630 do require authentication. Some of the key API endpoints available at
631 <https://www.alliancegenome.org/swagger-ui/> are: gene-summary, gene-disease, gene-
632 interactions, homologs-species, allele-phenotypes, expression ribbon-summary, etc.
633



The screenshot shows the Swagger UI for the Alliance of Genome Resources API. The top navigation bar includes '1.0 Beta' and 'OAS3' buttons. Below the navigation, a sub-header 'openapi' is visible. The main content area is titled 'Alliance of Genome Resources API'. A sidebar on the left lists categories: 'Allele Search', 'Cache Search', 'Disease', 'Entity Search', 'Expression', and 'Genes'. The 'Genes' section is expanded, showing a 'GET /api/gene/{id}' endpoint. The description for this endpoint is 'Retrieve a Gene for given ID'. Below the description, a 'Parameters' table is shown with one row: 'id * required' (type: string, path) with the value 'WB:WBGene00000001'. At the bottom of the endpoint section are 'Execute' and 'Clear' buttons.

634
635 **Figure 11. Swagger interface for the Alliance APIs.**
636

Responses

Curl

```
curl -X 'GET' \
  'https://www.alliancegenome.org/api/gene/WB:WBGene00000001' \
  -H 'accept: application/json'
```

Request URL

```
https://www.alliancegenome.org/api/gene/WB:WBGene00000001
```

Server response

Code	Details
200	<p>Response body</p> <pre>{ "id": "WB:WBGene00000001", "symbol": "aap-1", "dateProduced": "2023-08-08T21:15:02.000+00:00", "modCrossRefCompleteUrl": "https://www.wormbase.org/db/get?name=WBGene00000001;class=Gene", "species": [{ "name": "Caenorhabditis elegans", "shortName": "Cel", "dataProviderFullName": "WormBase", "dataProviderShortName": "WB", "commonNames": "['worm', 'cel']", "taxonId": "NCBITaxon:6239" }], "synonyms": ["CELE_Y110A7A.18", "Y110A7A.19"], "secondaryIds": [], "geneSynopsis": "aap-1 encodes the <i>C. elegans</i> ortholog of the phosphoinositide 3-kinase (PI3K) p50/p55 adaptor/regulator y subunit; AAP-1 negatively regulates lifespan and dauer development, and likely functions as the sole adaptor subunit for the AGE-1/pi10 PI3K catalytic subunit to which it binds in vitro; although AAP-1 potentiates insulin-like signaling, it is not absolutely required for insulin-like signaling under most conditions.", "automatedGeneSynopsis": "Enables protein kinase binding activity. Involved in dauer larval development, adult lifespan, and insulin receptor signaling pathway. Part of phosphatidylinositol 3-kinase complex. anchor cell; intestine; and neurons. Human ortholog(s) of this gene implicated in several diseases, including SHORT syndrome." }</pre>

637
638

Figure 12. Example of API output.

639

640

Data preservation in external repositories

641
642
643
644
645
646
647
648
649
650
651

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.

652

653
654
655
656
657
658
659

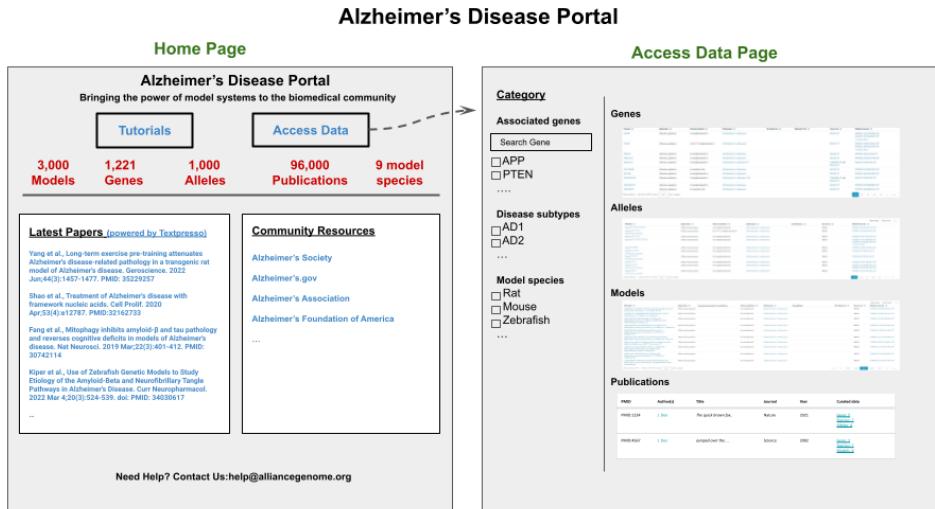
Ontologies that Alliance members maintain are also available from long-term repositories including the OBO Foundry (<https://obofoundry.org/>) and Zenodo (zenodo.org). Annotations related to gene expression, function, phenotype, disease associations, etc. that are generated by Alliance members and are 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**

660 that was established in the early 2000s as a community-based initiative for development and
661 maintenance of biological and biomedical ontologies using standardized practices. The Foundry
662 is the ontology repository of choice for the Alliance because it is widely recognized as an
663 authoritative source of well-maintained ontologies for biology and biomedical research.

664
665 Zenodo is a general purpose repository maintained by CERN (European Council for Nuclear
666 Research) for storing and sharing documents, data, and other digital research materials across
667 many disciplines. Zenodo is a repository of choice for the Alliance, in part, because of the
668 commitment by the European Commission to support Zenodo as long as CERN exists.

669
670 **Disease Portal(s)**

671 Providing users with ready and easy access to curated and harmonized model organism
672 disease data and tools is crucial to accelerate research related to the pathogenesis of human
673 disease. The Alliance has a wealth of disease-relevant data from eight model organism species
674 and human data, such as: genes, alleles and variants implicated in disease, genotypes and
675 strains that serve as disease models, and related data such as modifiers (herbals, chemicals,
676 small molecules, etc.) that ameliorate or exacerbate the disease condition and may serve as
677 candidates for potential drug development. To provide an easy entry point for clinical
678 researchers and human geneticists to access the consolidated data and tools, we are in the
679 process of designing and implementing a topic-specific resource--an Alzheimer's disease (AD)
680 portal that will serve as a paradigm for other disease portals (**Figure 13**). The AD portal will
681 include: orthologous genes in animal model systems, models with a mutation orthologous to one
682 in a patient group, models with a specific set of phenotypes, and/or modifiers that have been
683 shown to alter the disease condition. Building on the experience and pages developed for the
684 AD portal, we will expand this paradigm to other disease portals. Features planned for the
685 disease portal with AD as an example include: a home page with an overview of the data in the
686 portal, an autocomplete search box, links to other AD resources, and a list of the most recent
687 papers from PubMed and/or from the ABC store (see example portal page below). The pages in
688 the portal will be modeled on existing pages at the Alliance and will include gene summaries,
689 alleles and variants, phenotypes, gene interactions, pathways, biological processes (based on
690 GO), gene expression, etc. We also plan to provide visualizations of data analysis, for example,
691 diseases that share genes and protein interactions that may point to common underlying
692 molecular mechanisms. Up-to-date data sets, e.g., genes, strains, modifiers (drugs, chemicals,
693 herbals, etc. shown to either ameliorate or exacerbate phenotypes) will be available as
694 downloadable files. Disease-specific data sets will also be available for query from AllianceMine.
695 We will also provide up-to-date links to disease-specific literature, and search capabilities
696 through literature search engines such as the Textpresso instance dedicated to AD
697 (<http://alzheimer.textpressocentral.org>; corpus size - 96,000 papers).



698

699 **Figure 13. Mockup of the Alzheimer's Disease Portal showing the Home page and the**
700 **Data access page.** These views illustrate the type of information that will be available with a
701 disease-focus.

702

703 Outreach and interactions

704

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

715

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

722

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

727

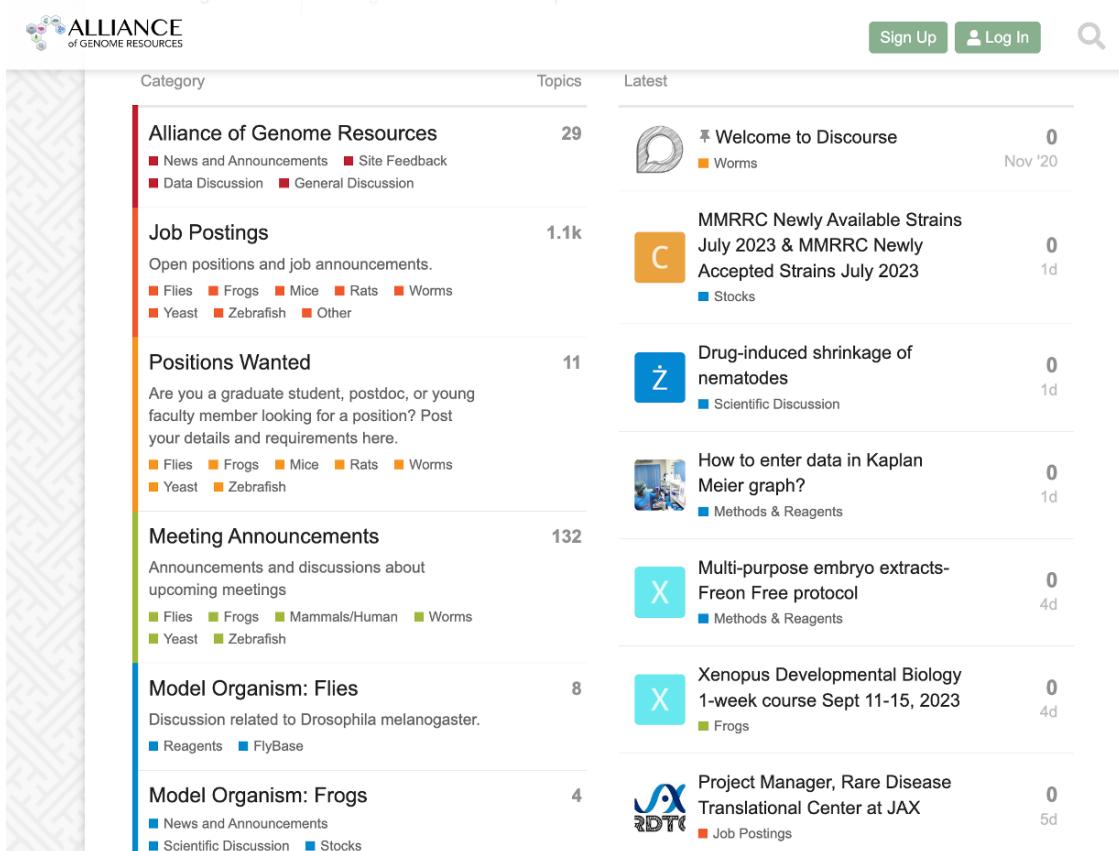
728 **Illustrated tutorials and videos.** We maintain several types of tutorial options that are
729 accessible from the Help menu (<https://www.alliancegenome.org/tutorials>). The most requested
730 types of tutorials are illustrated guides with screenshots on how to use various features of the
731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

734

735 **Alliance User Community Forum.** The Alliance supports a centralized community discussion
736 board implemented in Discourse (<https://community.alliancegenome.org/categories>) (Figure
737 14). Each model organism represented in the Alliance is represented as its own Discourse
738 category with model organism specific threads for news, discussion, and reagent information.
739 The forum also includes categories for job postings, meeting announcements, and general
740 information about the Alliance of Genome Resources. Alliance members with existing on-line
741 community forums are migrating users to the Alliance Central forum.

742

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



The screenshot shows the homepage of the Alliance User Community Forum. At the top, there is a navigation bar with the Alliance logo, a search bar, and buttons for 'Sign Up' and 'Log In'. The main content area is divided into two sections: 'Category' on the left and 'Latest' on the right.

Category:

- Alliance of Genome Resources** (29 topics): Includes News and Announcements, Site Feedback, Data Discussion, and General Discussion.
- Job Postings** (1.1k topics): Open positions and job announcements. Categories include Flies, Frogs, Mice, Rats, Worms, Yeast, Zebrafish, and Other.
- Positions Wanted** (11 topics): For graduate students, postdocs, or young faculty members. Categories include Flies, Frogs, Mice, Rats, Worms, Yeast, and Zebrafish.
- Meeting Announcements** (132 topics): Announcements and discussions about upcoming meetings. Categories include Flies, Frogs, Mammals/Human, Worms, Yeast, and Zebrafish.
- Model Organism: Flies** (8 topics): Discussion related to *Drosophila melanogaster*. Categories include Reagents and FlyBase.
- Model Organism: Frogs** (4 topics): Discussion related to frogs. Categories include News and Announcements, Scientific Discussion, and Stocks.

Latest:

- Worms (Welcome to Discourse, 0 replies, posted Nov '20)
- Stocks (MMRRC Newly Available Strains, 0 replies, posted 1d ago)
- Scientific Discussion (Drug-induced shrinkage of nematodes, 0 replies, posted 1d ago)
- Methods & Reagents (How to enter data in Kaplan Meier graph?, 0 replies, posted 1d ago)
- Methods & Reagents (Multi-purpose embryo extracts-Freon Free protocol, 0 replies, posted 4d ago)
- Frogs (Xenopus Developmental Biology 1-week course Sept 11-15, 2023, 0 replies, posted 4d ago)
- Job Postings (Project Manager, Rare Disease Translational Center at JAX, 0 replies, posted 5d ago)

747
748 **Figure 14. Alliance community forum home page.**
749
750

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

756

757 **Prospects and Challenges**

758

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

767

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

772

773 **AI.** As discussed above, we are actively considering AI/ML applications throughout the project.
774 Our practical approach is driven by us being subject matter experts. Because we have relied on
775 human expert curation, we are in a unique position to evaluate and use the output of various
776 AIs.

777

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

784

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

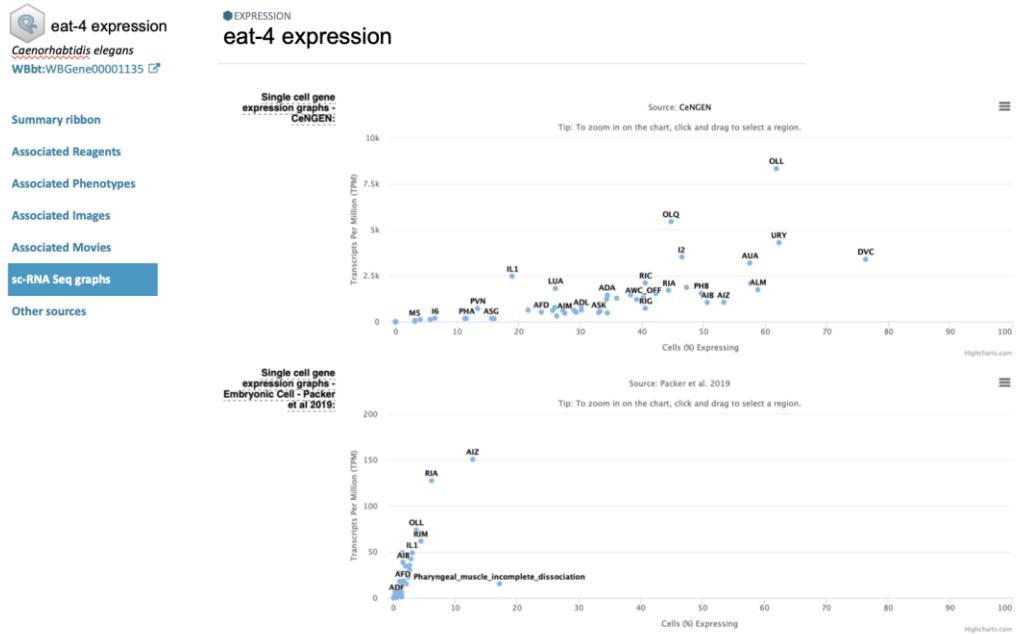
792

793 Several Alliance members also collaborate with publishing groups, such as microPublication
794 Biology (<https://www.micropublication.org/>) or the Genetics Society of America (<https://genetics>-

795 gsa.org/publications/), to streamline pre-publication data integrity verification and curation by
796 curators and authors, enabling MODs to quality-check and work with authors to correct data
797 reporting before publication and promptly incorporate it into Alliance Knowledgebases upon
798 article publication.

799
800 **Dealing with satellite genomes and genetic models.** In addition to the core genomes and
801 associated data, our resources store and present information about the genes and genomes of
802 relatively closely related organisms. For example, WormBase includes some genetically-studied
803 nematodes such as *Caenorhabditis briggsae* that benefit from the rich data models typical of *C.*
804 *elegans*. Genetic screens and positional cloning (Inoue et al., 2007; Sharanya et al. 2012),
805 CRISPR editing (Cohen and Sternberg, 2019; Cohen et al., 2022; Ivanova and Moss 2023), as
806 well as transcriptomic analyses (Jhaveri et al., 2022) are now routinely done in this species. For
807 the Alliance to take on this responsibility of WormBase, we need to support such satellite model
808 organisms. Our plan is to support community gene structure annotation (e.g., for *Drosophila*,
809 Sargent et al, 2020; for *C. elegans*, Moya et al. 2023) using the Apollo curaton system designed
810 specifically for such activity (Dunn et al., 2019).

811
812 **High Throughput expression data and single cell RNA-seq plans**
813 We harmonized high-throughput expression metadata of mouse, rat, yeast, worm, fly, and
814 zebrafish. Users can browse them with species, assay type (microarray, RNA-seq, tiling array,
815 and proteomics), tissue, sex, and curated categories. We plan to add single-cell RNA-seq as a
816 new assay type, making such datasets easily identifiable within our collection, with links to other
817 resources, including Gene Expression Omnibus, EBI single-cell RNAseq Expression Atlas, and
818 CZI CellxGene. To display the information above, we will implement a content-rich expression
819 detail page that will provide a unified way to access all expression data associated with a
820 specific gene, including link outs to other sources and MOD-specific single-cell RNA-seq gene
821 expression graphs (**Figure 15**).



822

823 **Figure 15. Mockup of an Expression Detail page.** This example shows one of the current features of
824 WormBase – single cell data from two studies – displayed on what will be part of an Alliance Gene
825 Expression detail page.

826

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

843

844

845 Acknowledgements

846 We thank our multiple communities for their patience and feedback about the prospect of the
847 Alliance and their love of their own MODs. We also thank the members of our Scientific
848 Advisory Board (Gary Bader, Alex Bateman, Helen Berman, Shawn Burgess, Andrew Chisholm,
849 Phil Hieter, Brian Oliver, Calum Macrae, Titus Brown, Abraham Palmer and Michelle Southard-

850 Smith) for cogent advice, and NHGRI Program Staff (Sandhya Xirasagar, Ajay Pillai, Valentina
851 di Francesco, Sarah Hutchison, and Helen Thompson) for guidance. The core funding for the
852 Alliance is from the National Human Genome Research Institute and the National Heart, Lung
853 and Blood Institute (U24HG010859). The curation of data and their harmonization is supported
854 by National Human Genome Research Institute grants U24HG002659 (ZFIN), U24HG002223
855 (WormBase), U41HG000739 (FlyBase), U24HG001315 (SGD), U24HG000330 (MDG),
856 P41HD064556 (Xenbase), U24HG011851 (Reactome + GO) and U41HG012212 (GO
857 Consortium), as well as grant R01HL064541 from the National Heart, Lung and Blood Institute
858 (RGD), P41HD062499 from the Eunice Kennedy Shriver National Institute of Child Health and
859 Human Development (GXD), and the Medical Research Council-UK grant MR/L001020/1
860 (WormBase). Additional effort was supported by DOE DE-AC02-05CH11231. Curation tools are
861 supported in part by the National Library of Medicine NLM R01LM013871.

862

863

864 References

865

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

868 Altenhoff AM, Train CM, Gilbert KJ, Mediratta I, Mendes de Farias T, Moi D, Nevers Y,
869 Radojkova HS, Rossier V, Warwick Vesztrocy A, Glover NM, Dessimoz C. OMA orthology in
870 2021: website overhaul, conserved isoforms, ancestral gene order and more. Nucleic Acids
871 Res. 2021 Jan 8;49(D1):D373-D379. doi: 10.1093/nar/gkaa1007. PMID: 33174605; PMCID:
872 PMC7779010.

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

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

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

879 Bowes JB, Snyder KA, James-Zorn C, Ponferrada VG, Jarabek CJ, Burns KA, Bhattacharyya B,
880 Zorn AM, Vize PD. The Xenbase literature curation process. Database (Oxford). 2013 Jan
881 9;2013:bas046. doi: 10.1093/database/bas046. PMID: 23303299; PMCID: PMC3540419.

882 Bradford, Y.M., et al., *From multiallele fish to nonstandard environments, how ZFIN assigns*
883 *phenotypes, human disease models, and gene expression annotations to genes*. Genetics,
884 2023. **224**(1).

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

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

891 Carotenuto R, Pallotta MM, Tussellino M, Fogliano C. *Xenopus laevis* (Daudin, 1802) as a
892 Model Organism for Bioscience: A Historic Review and Perspective. *Biology* (Basel). 2023 Jun
893 20;12(6):890. doi: 10.3390/biology12060890. PMID: 37372174; PMCID: PMC10295250.

894 Cohen S, Sternberg P. Genome editing of *Caenorhabditis briggsae* using CRISPR/Cas9 co-
895 conversion marker *dpy-10*. *MicroPubl Biol*. 2019 Oct
896 11:2019:10.17912/micropub.biology.000171. doi: 10.17912/micropub.biology.000171. PMID:
897 32550401; PMCID: PMC7252229.

898 Cohen SM, Wrobel CJ, Prakash SJ, Schroeder FC, Sternberg PW. Formation and function of
899 dauer ascarosides in the nematodes *Caenorhabditis briggsae* and *Caenorhabditis elegans*. *G3*
900 (Bethesda). 2022 Mar 4;12(3):jkac014. doi: 10.1093/g3journal/jkac014. PMID: 35094091;
901 PMCID: PMC8895998.

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

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

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

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

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

917 Fisher M, James-Zorn C, Ponferrada V, Bell AJ, Sundararaj N, Segerdell E, Chaturvedi P,
918 Bayyari N, Chu S, Pells T, Lotay V, Agalakov S, Wang DZ, Arshinoff BI, Foley S, Karimi K, Vize
919 PD, Zorn AM. Xenbase: key features and resources of the *Xenopus* model organism
920 knowledgebase. *Genetics*. 2023 May 4;224(1):iyad018. doi: 10.1093/genetics/iyad018. PMID:
921 36755307; PMCID: PMC10158840.

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

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

928 Gene Ontology, Consortium., *The Gene Ontology knowledgebase in 2023*. *Genetics*, 2023.
929 **224**(1).

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

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

933 Hu Y, Comjean A, Rodiger J, Liu Y, Gao Y, Chung V, Zirin J, Perrimon N, Mohr SE.
934 FlyRNAi.org-the database of the Drosophila RNAi screening center and transgenic RNAi
935 project: 2021 update. *Nucleic Acids Res*. 2021 Jan 8;49(D1):D908-D915. doi:
936 10.1093/nar/gkaa936. PMID: 33104800; PMCID: PMC7778949.

937 Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, Mohr SE. An integrative

938 approach to ortholog prediction for disease-focused and other functional studies. *BMC*
939 *Bioinformatics*. 2011 Aug 31;12:357. doi: 10.1186/1471-2105-12-357. PMID: 21880147; PMCID:
940 PMC3179972.

941 Inoue T, Ailion M, Poon S, Kim HK, Thomas JH, Sternberg PW. Genetic analysis of dauer
942 formation in *Caenorhabditis briggsae*. *Genetics*. 2007 Oct;177(2):809-18. doi:
943 10.1534/genetics.107.078857. Epub 2007 Jul 29. PMID: 17660533; PMCID: PMC2034645.

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

947 Jhaveri N, van den Berg W, Hwang BJ, Muller HM, Sternberg PW, Gupta BP. Genome
948 annotation of *Caenorhabditis briggsae* by TEC-RED identifies new exons, paralogs, and
949 conserved and novel operons. *G3 (Bethesda)*. 2022 Jul 6;12(7):jkac101. doi:
950 10.1093/g3journal/jkac101. PMID: 35485953; PMCID: PMC9258526.

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

954 Larkin A, Marygold SJ, Antonazzo G, Attrill H, Dos Santos G, Garapati PV, Goodman JL,
955 Gramates LS, Millburn G, Strelets VB, Tabone CJ, Thurmond J; FlyBase Consortium. FlyBase:
956 updates to the *Drosophila melanogaster* knowledge base. *Nucleic Acids Res*. 2021 Jan
957 8;49(D1):D899-D907. doi: 10.1093/nar/gkaa1026. PMID: 33219682; PMCID: PMC7779046.
958 (<https://pubmed.ncbi.nlm.nih.gov/33219682/>)

959 Liu W, Laulederkind SJ, Hayman GT, Wang SJ, Nigam R, Smith JR, De Pons J, Dwinell MR,
960 Shimoyama M. OntoMate: a text-mining tool aiding curation at the Rat Genome Database.
961 Database (Oxford). 2015 Jan 25;2015:bau129. doi: 10.1093/database/bau129. PMID:
962 25619558; PMCID: PMC4305386.

963 Milacic M, Rothfels K, Mathews L, Wright A, Jassal B, Shamovsky V, Trinh Q, Gillespie M,
964 Sevilla C, Tiwari K, Ragueneau E, Gong C, Stephan1 R, May B, Haw R, Weiser J, Beavers D,
965 Conley P, Hermjakob H, Stein LD, D'Eustachio P, Wu G (2024) The Reactome Pathway
966 Knowledgebase 2024. *Nucleic Acids Res.*, in press. PMID: [37941124](https://doi.org/10.1093/nar/gkaa1026)

967 Mitros T, Lyons JB, Session AM, Jenkins J, Shu S, Kwon T, Lane M, Ng C, Grammer TC,
968 Khokha MK, Grimwood J, Schmutz J, Harland RM, Rokhsar DS. A chromosome-scale genome
969 assembly and dense genetic map for *Xenopus tropicalis*. *Dev Biol*. 2019 Aug 1;452(1):8-20. doi:
970 10.1016/j.ydbio.2019.03.015. PMID: 30980799.

971 Moya ND, Stevens L, Miller IR, Sokol CE, Galindo JL, Bardas AD, Koh ESH, Rozenich J, Yeo
972 C, Xu M, Andersen EC. Novel and improved *Caenorhabditis briggsae* gene models generated
973 by community curation. *BMC Genomics*. 2023 Aug 25;24(1):486. doi: 10.1186/s12864-023-
974 09582-0. PMID: 37626289; PMCID: PMC10463891.

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

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

984 10.1093/nar/gkac330. PMID: 35552456; PMCID: PMC9252809.

985 Nevers Y, Kress A, Defosset A, Ripp R, Linard B, Thompson JD, Poch O, Lecompte O.

986 OrthoInspector 3.0: open portal for comparative genomics. Nucleic Acids Res. 2019 Jan

987 8;47(D1):D411-D418. doi: 10.1093/nar/gky1068. PMID: 30380106; PMCID: PMC6323921.

988 Oliver, S.G., et al., *Model organism databases: essential resources that need the support of*

989 *both funders and users*. BMC Biol, 2016. **14**: p. 49.

990 Persson E, Sonnhammer ELL. InParanoid-DIAMOND: faster orthology analysis with the

991 InParanoid algorithm. Bioinformatics. 2022 May 13;38(10):2918-2919. doi:

992 10.1093/bioinformatics/btac194. PMID: 35561192; PMCID: PMC9113356.

993 Priyam A, Woodcroft BJ, Rai V, Moghul I, Munagala A, Ter F, Chowdhary H, Pieniak I, Maynard

994 LJ, Gibbins MA et al. 2019. Sequenceserver: A Modern Graphical User Interface for Custom

995 BLAST Databases. *Mol Biol Evol* **36**: 2922-2924.

996 Ringwald, M., et al., *Mouse Genome Informatics (MGI): latest news from MGD and GXD*.

997 Mamm Genome, 2022. **33**(1): p. 4-18.

998 Sargent L, Liu Y, Leung W, Mortimer NT, Lopatto D, Goecks J, Elgin SCR. G-OnRamp:

999 Generating genome browsers to facilitate undergraduate-driven collaborative genome

1000 annotation. PLoS Comput Biol. 2020 Jun 4;16(6):e1007863. doi: 10.1371/journal.pcbi.1007863.

1001 PMID: 32497138; PMCID: PMC7272004.

1002 Session AM, Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S, Fukui A, Hikosaka A,

1003 Suzuki A, Kondo M, van Heeringen SJ, Quigley I, Heinz S, Ogino H, Ochi H, Hellsten U, Lyons

1004 JB, Simakov O, Putnam N, Stites J, Kuroki Y, Tanaka T, Michiue T, Watanabe M, Bogdanovic

1005 O, Lister R, Georgiou G, Paranjpe SS, van Kruisbergen I, Shu S, Carlson J, Kinoshita T, Ohta

1006 Y, Mawaribuchi S, Jenkins J, Grimwood J, Schmutz J, Mitros T, Mozaffari SV, Suzuki Y,

1007 Haramoto Y, Yamamoto TS, Takagi C, Heald R, Miller K, Haudenschild C, Kitzman J,

1008 Nakayama T, Izutsu Y, Robert J, Fortriede J, Burns K, Lotay V, Karimi K, Yasuoka Y, Dichmann

1009 DS, Flajnik MF, Houston DW, Shendure J, DuPasquier L, Vize PD, Zorn AM, Ito M, Marcotte

1010 EM, Wallingford JB, Ito Y, Asashima M, Ueno N, Matsuda Y, Veenstra GJ, Fujiyama A, Harland

1011 RM, Taira M, Rokhsar DS. Genome evolution in the allotetraploid frog *Xenopus laevis*. Nature.

1012 2016 Oct 20;538(7625):336-343. doi: 10.1038/nature19840. PMID: 27762356; PMCID:

1013 PMC5313049.

1014 Sharanya D, Thillainathan B, Marri S, Bojanala N, Taylor J, Flibotte S, Moerman DG, Waterston

1015 RH, Gupta BP. Genetic control of vulval development in *Caenorhabditis briggsae*. G3

1016 (Bethesda). 2012 Dec;2(12):1625-41. doi: 10.1534/g3.112.004598. Epub 2012 Dec 1. PMID:

1017 23275885; PMCID: PMC3516484.

1018 Smith RN, Aleksic J, Butano D, Carr A, Contrino S, Hu F, Lyne M, Lyne R, Kalderimis A,

1019 Rutherford K, Stepan R, Sullivan J, Wakeling M, Watkins X, Micklem G. InterMine: a flexible

1020 data warehouse system for the integration and analysis of heterogeneous biological data.

1021 Bioinformatics. 2012 Dec 1;28(23):3163-5. doi: 10.1093/bioinformatics/bts577. Epub 2012 Sep

1022 27. PMID: 23023984; PMCID: PMC3516146.

1023 Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou LP, Mi H. PANTHER: Making

1024 genome-scale phylogenetics accessible to all. Protein Sci. 2022 Jan;31(1):8-22. doi:

1025 10.1002/pro.4218. Epub 2021 Nov 25. PMID: 34717010; PMCID: PMC8740835.

1026 Thomas, P.D., et al., *Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO*

1027 *annotations to structured descriptions of biological functions and systems*. Nat Genet, 2019.

1028 **51**(10): p. 1429-1433.

1029 Van Auken K, Fey P, Berardini TZ, Dodson R, Cooper L, Li D, Chan J, Li Y, Basu S, Muller HM,
1030 Chisholm R, Huala E, Sternberg PW; WormBase Consortium. Text mining in the biocuration
1031 workflow: applications for literature curation at WormBase, dictyBase and TAIR. Database
1032 (Oxford). 2012 Nov 17;2012:bas040. doi: 10.1093/database/bas040. PMID: 23160413; PMCID:
1033 PMC3500519.

1034 Vedi, M., et al., 2022 *updates to the Rat Genome Database: a Findable, Accessible,*
1035 *Interoperable, and Reusable (FAIR) resource.* Genetics, 2023. **224**(1).

1036 Wood V, Sternberg PW, Lipshitz HD. Making biological knowledge useful for humans and
1037 machines. Genetics. 2022 Apr 4;220(4):iyac001. doi: 10.1093/genetics/iyac001.