Genomics Proteomics Bioinformatics 13 (2015) 64–68

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

RESOURCE REVIEW

Web Resources for Model Organism Studies

Bixia Tang [1,2,a], Yanqing Wang [1,b], Junwei Zhu [1,c], Wenming Zhao [1,*,d]

[1] *Core Genomic Facility, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*
[2] *University of Chinese Academy of Sciences, Beijing 100049, China*

Received 9 January 2015; revised 22 January 2015; accepted 31 January 2015
Available online 20 February 2015

Handled by Zhang Zhang

KEYWORDS

Model organism;
Database;
Genome;
Bioinformatics;
Biology

Abstract An ever-growing number of resources on model organisms have emerged with the continued development of sequencing technologies. In this paper, we review 13 databases of model organisms, most of which are reported by the National Institutes of Health of the United States (NIH; http://www.nih.gov/science/models/). We provide a brief description for each database, as well as detail its data source and types, functions, tools, and availability of access. In addition, we also provide a quality assessment about these databases. Significantly, the organism databases instituted in the early 1990s—such as the Mouse Genome Database (MGD), Saccharomyces Genome Database (SGD), and FlyBase—have developed into what are now comprehensive, core authority resources. Furthermore, all of the databases mentioned here update continually according to user feedback and with advancing technologies.

Introduction

Model organisms were placed at the forefront of biomedical research by the end of the 20th century [1]. Defining the etymology of the term *model organism* is relatively difficult; however, the development of molecular biology technologies during the 1960s and 1970s led to its materialization. The key rationale for the study of model organisms in biomedical research is to examine fundamental mechanisms that may be shared by many or all living entities.

Some model organisms—such as *Drosophila*, mouse, and maize—have long histories of use, whereas others have been developed more recently. Since the common conception of a model organism is changing along with technological advances in genome sequencing and editing, it is difficult to provide a complete list of model organisms; therefore, here we mainly focus on the canonical set of model organisms defined by the National Institutes of Health (NIH) during the 1990s. The available web resources for the 13 covered model organisms are listed in Table S1, and we also provide a quality assessment for each resource (**Table 1**) based on five aspects: (1) webpage esthetics, (2) system performance, (3) data sources, (4) software and tools, and (5) data availability. If all aspects are covered, a full score of 5 points is given.

* Corresponding author.
E-mail: zhaowm@big.ac.cn (Zhao W).
[a] ORCID: 0000-0002-9357-4411.
[b] ORCID: 0000-0002-7985-7941.
[c] ORCID: 0000-0003-4689-3513.
[d] ORCID: 0000-0002-4396-8287.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2015.01.003

Table 1 Major web resources for model organism studies

Name	Link	Main features	Rating	Refs.
MGD	http://www.informatics.jax.org	Well-designed webpage; standard access speed for data; data uploading supported; abundant effective tools; FTP accepted	★★★★★	[2–4]
RGD	http://rgd.mcw.edu/	Well-designed webpage; standard access speed for data; data uploading supported; abundant effective tools; FTP accepted	★★★★★	[5]
SGD	http://www.yeastgenome.org/	Some browser limitation such as Firefox v35 under Win7 system; low access speed for data; data uploading not supported; abundant effective tools; web page downloading accepted	★★★☆☆	[6,7]
dictyBase	http://dictybase.org/	Well-designed webpage; standard access speed for data; data uploading not supported; abundant effective tools; web page downloading accepted	★★★★☆	[8]
WormBase	http://www.wormbase.org	Well-designed webpage; standard access speed for data; data uploading supported; abundant effective tools; FTP accepted	★★★★★	[9]
FlyBase	http://flybase.org/	Well-designed webpage; standard access speed for data; data uploading not supported; abundant effective tools; FTP accepted	★★★★☆	[10]
ZFIN	http://zfin.org/	Well-designed webpage; standard access speed for data; data uploading supported; abundant effective tools; web page downloading accepted	★★★★★	[11]
Xenbase	http://www.xenbase.org	Well-designed webpage; some problems exist such as wild word search 12*; data uploading supported; abundant effective tools; FTP accepted	★★★★☆	[12]
TAIR	http://www.arabidopsis.org/	Well designed webpage; standard access speed for data; data uploading supported; abundant effective tools; FTP accepted	★★★★★	[13]
BeetleBase	http://www.beetlebase.org	Well-designed webpage; standard access speed for data; data uploading supported; abundant effective tools; FTP accepted	★★★★☆	[14]
MyMpn	http://mympn.crg.eu	Well-designed webpage; standard access speed for data; data uploading not supported; abundant effective tools; web page downloading accepted	★★★★☆	[15]
MaizeGDB	http://www.maizegdb.org/	Well-designed webpage; standard access speed for data; data uploading supported; abundant effective tools; FTP accepted	★★★★★	[16]
ASAP	https://asap.genetics.wisc.edu/asap/home.php	Well-designed webpage; standard access speed for data; data uploading supported; abundant effective tools; web page downloading accepted	★★★★★	[17]

Note: The assessment evidence covers 5 aspects: webpage esthetics, system performance, data sources, software and tools, and data availability. The details are as follows: whether webpage is well designed; whether the accessing speed is acceptable; whether data uploading is supported; whether there are abundant effective tools, and whether instant downloading is supported.

Mouse Genome Informatics

The Mouse Genome Informatics (MGI) is a comprehensive bioinformatics resource for the laboratory mouse (*Mus musculus*) that includes the Mouse Genome Database (MGD), Gene Expression Database (GXD), Mouse Tumor Biology (MTB) Database, Gene Ontology (GO), and MouseMine.

Initially designed in 1994, MGD is a key resource of MGI used to curate genome information, as well as to track genome mapping data and to record mouse mutant phenotypes [2]. MGD is mostly suited to inquire about the genetic basis of human disease based on the phenotypic characterization of transgenic mice. Within the database, users can access genetic, genomic, and phenotypic data either curated from public literature by the MGD group or deposited by researchers from all over the world. A ''Quick Search'' function returns a broad range of data from a keyword query, whereas more specific search forms are also available for users, such as genes, markers and SNPs, among others. In addition, MGD supports three vocabulary browsers for users to find all associated annotated information, including GO, the Mammalian Phenotype Ontology (MP), and the Online Mammalian Inheritance in Man (OMIM), as well as the GBrowse and JBrowse graphical browser to view genome interactively. Furthermore, MGD can provide bulk access to certain kinds of information through the Batch Query tool, or the BioMart [3] and InterMine [4] databases. MGD undergoes weekly updates, and supports file transfer protocol (FTP) and direct structure query language (SQL) data retrieval methods. Overall, MGD has become a prime resource for international researchers to obtain biological information on the laboratory mouse.

Rat Genome Database

The Rat Genome Database (RGD) is a database of biological information for studying the laboratory rat (*Rattus norvegicus*) that started more than 10 years ago [5]. RGD allows researchers to investigate mechanisms of human physiology and disease based on experimental research acquired in the rat. The RGD provides several types of disease-focused data, including genetic and genomic data, as well as information on biological pathways, phenotypes, and individual strains. To guarantee quality data, the RGD curates a subset of data manually from literature regarding disease models, molecular pathways, and mammalian phenotypes. The database also utilizes controlled vocabularies and ontologies to present data shared with other data sources, including NCBI and Ensembl, and users can access in-source tools to search and view data on hand. For instance, users can use GBrowse to analyze the genetic or genomic differences among the rat, mouse, and human. Meanwhile, the Virtual Comparative Map function (VCMap: http://www.animalgenome.org/VCmap/) allows users to compare the genomic positions of various genes simultaneously. Similar to the MGD, RGD users can bulk download data from the FTP site.

Saccharomyces Genome Database

The *Saccharomyces* Genome Database (SGD) is an open-source database for querying the genetics and molecular and cellular biology of *Saccharomyces cerevisiae* started in 1993 by David Botstein and colleagues at the Stanford University [6]. The SGD primarily sources manually curated scientific literature to ensure the data quality. Currently, the yeast genome, genes, proteins, and other encoded features are available for user access. SGD employs GO and the Ascomycete Phenotype Ontology (APO) to capture gene product function and mutant phenotypes, respectively. SGD also offers a number of tools for users to access data. For example, users can use YeastMine to retrieve and analyze a variety of data types, including chromosomal features, sequences, and protein features, among others; whereas the GBrowse and Serial Pattern of Expression Levels Locator (SPELL) [7] tools allow users to explore genomic features and analyze microarray gene expression data, respectively.

dictyBase

The recently contrived dictyBase is a genetic database for the social ameba *Dictyostelium discoideum* [8]. Data provided in dictyBase include the organism's complete nuclear genome sequence, mitochondrial genome sequence, the extrachromosomal rRNA genes, expressed sequence tags (ESTs), and annotations information. dictyBase covers genomes of four different species including *D. discoideum*, *Dictyostelium purpureum*, *Dictyostelium fasciculatum*, and *Polysphondylium pallidum*. The resource collects data from major resource centers and projects such as Swiss-Prot and an international consortium of the *Dictyostelium* Genome Project and the Japanese cDNA Project. It also has an integrated central strain repository to facilitate the ordering of *Dictyostelium* strains and other slime molds, known as the Dicty Stock Center. Besides maintaining the comprehensive dataset, dictyBase also provides several tools, including a Genome Browser to display information graphically such as genes, RNA-seq, Dicty alignments, and genome assembly information. Users can utilize BLAST tools to perform comparison searches for nucleotide and amino acid sequences, and the gene page displays integrated information of genes, proteins, gene ontology, and gene orthology. Data can be downloaded directly from the web page and updated on a weekly to monthly basis.

WormBase

WormBase is a curated resource for the model organism *Caenorhabditis elegans* [9]. Available data in WormBase include gene models, allelic variations, mutant phenotypes, anatomy function, expression patterns, gene interactions, and human disease relevance. WormBase hosts the reference genome (WBcel235) that includes 1402 corrections and loads the annotated reference genomic sequence for over 20 species [9]. Besides maintaining comprehensive data, WormBase also provides several user tools. For example, users can view graphical data with the GBrowse and BLAST/BLAT tools. WormMine enables users to query, save, and manipulate these objects, as well as download bulk data. Full text search also provides the ability to do a customized search and users can freely download data from the FTP site. WormBase continues to update and has become a dedicated resource for researchers.

FlyBase

FlyBase is a *Drosophila* genomic database that was founded in 1992 as a resource for collecting and representing related information on the fruit fly (*Drosophila melanogaster*) [10]. FlyBase curates data from the published scientific literature accumulated over the past 20 years, including data detailing animal phenotypes, gene expression, interactions, and GO, among others [10]. Besides providing several basic search tools such as Quick Search and BLAST, FlyBase has also developed several novel tools. These include (1) TermLink, which uses a controlled vocabulary (CV) to classify and annotate queried data; (2) RNA-seq Search tool, which provides a reads per kilobase per million (RPKM) value for searching expression patterns of all annotated genes; (3) FeatureMapper search function for specific genomic features such as functional elements, mutant event locations and reagents in one or more sequence regions, and QueryBuilder for full text searches. Users can download the FlyBase data from its FTP site or use SQL to access the database directly. FlyBase has grown into an integrated and complex database for over 20 years and will continuously update to accommodate data and software tools available.

Zebrafish Information Network

The Zebrafish Information Network (ZFIN) is a database for the model organism zebrafish [11], which includes data on genes, mutations, phenotypes, genotypes, gene expression, orthology, and nucleotide and protein sequences. ZFIN collects data from three primary sources: (1) curated scientific literature; (2) collaborations with major resource centers, such as the Sanger Institute, Ensembl, NCBI and UniProt; and (3) submissions from individual investigators. Besides maintaining the comprehensive data, ZFIN also provides several tools to help users to browse data, such as BLAST and GBrowse for alignment searches. The database also provides search interfaces for various data types such as genes, markers, clones, and gene expression. Users can submit their own data to ZFIN directly and download data from either their search results or the web page. ZFIN updates daily.

Xenbase

Xenbase is a genetic database that maintains the genome builds and gene models for two related species: the allotetraploid *Xenopus laevis* (v7.1) and the diploid *Xenopus tropicalis* (v8.0). Xenbase provides eight module entries, including BLAST, GBrowse, gene expression, genes search, information and resources on *Xenopus* anatomy and development, *Xenopus*-specific reagents and protocols, literature, and community [12]. Users can register an account to submit their own data to Xenbase, as well as download data from the Xenbase FTP site. The database is continually updating. At present, the current version (v3.3.1) has added a substantial number of relevant resources such as microarray data from the Gene Expression Omnibus (GEO), antibody database, and updated the software components over the previous version released in 2008 (v2.01).

The *Arabidopsis* Information Resource

The *Arabidopsis* Information Resource (TAIR) is a genetic and genomic database for *Arabidopsis thaliana* [13]. Data types include genome sequences, gene structure and annotation, metabolic pathways, gene expression, DNA and seed reserve information, genome maps, genetic and physical markers, and information on ecotypes and natural variation, among others. TAIR manually curates public literature, integrates data and resources from other sources—such as GenBank and the *Arabidopsis* Biological Resource Center (ABRC), and allows users to upload own data. TAIR provides several tools for users to query and analyze data, such as Textpresso to extract and process biological literature, N-Browse to view interactive biological networks, and GBrowse that displays multiple genomes to study genome duplication and evolution. Overall, TAIR serves as a community resource for *Arabidopsis* researchers and continues to be an essential resource for plant biologists.

Other resources

BeetleBase

BeetleBase is an integrated resource for the *Tribolium* research community that hosts several data types, including unmapped scaffolds, FGENESH-predicted genes, BAC-end sequences, genetic markers, mutants, GO, sequence ontology, and links to other databases. BeetleBase mainly collects sequence data from public databases and the *Tribolium* research community—including the Human Genome Sequencing Center, NCBI, the *Tribolium* BAC library, and the *Tribolium* Mutant Database, as well as annotation information and mapping results from published literature. BeetleBase also provides several user tools including BLAST for sequence alignments and a BLAT/GMOD-integrated search engine to enhance querying abilities. GBrowse, JBrowse, and Cmap are also provided as map viewers. All the sequences can be downloaded from the BeetleBase FTP site.

MyMpn

The MyMpn database is a web resource for the human pathogen *Mycoplasma pneumoniae*. MyMpn mainly covers the genomics data such as gene, gene essentiality, and operons, transcriptomics data such as microarrays, proteomics data including proteins, Pfam domains, complexes, and peptides, and metabolomics data such as metabolic reactions and growth curves [15]. MyMpn supports GBrowse and MyGBrowser to view genome information and enables users to browse pathway information. Users can download data directly from the MyMpn web page.

Maize Genetics and Genomics Database

First released in 1991, the Maize Genetics and Genomics Database (MaizeGDB) includes a comprehensive dataset encompassing information on genomic regions and loci, allelic variation, markers and probes, gene sequences and products, as well as phenotypic and metabolic information. MaizeGDB

provides several data visualization tools, such as the 'Locus Lookup Tool', 'Bin Viewer', and the MaizeGDB Genome Browser to view relevant data. Users can register with the website and submit data to data center or download data from either the FTP site or webpage. MaizeGDB updates continually.

A Systematic Annotation Package for Community Analysis of Genomes

A Systematic Annotation Package for Community Analysis of Genomes (ASAP) is a database developed to curate sequence and functional characterization data from the *Escherichia coli K-12 MG1655*, the best studied genome of enterobacterial family, as well as 576 other enterobacterial genomes [17]. Data types provided by ASAP include annotations, sequences, expression, experiments, and information on mutant strains. ASAP is an authority-controlled database where guests can view the published genomic data on genomes and experiments and the full data of *K-12 MG1655*. However, users must be registered as an annotator to achieve the complete set of features potentially in a genome and add annotations [17]. ASAP supports data type-specific, full-text, and BLAST searches. The *MG1655* data can be downloaded directly from the webpage. ASAP is currently under development and updates data continuously.

Concluding remarks

The web resources on model organisms have become increasingly enriched in both the amount of data and available analysis tools. The majority of the data surveyed in these databases are manually curated from published scientific literature. All support hyperlinks to connect or share data with other databases. To achieve user flexibility and convenience, these databases also provide several general and dedicated toolsets. Most of the aforementioned databases are open-source and the users can freely download data directly from the resources webpage or FTP site. Furthermore, the databases above continuously update, aiming to become authority resources for model organism study. Although many advantages are listed here, there are shortcomings as well. For instance, different data formats or interface standards are applied in establishing these databases, which bring some limitations, for example data integration, when the users want to access a kind of the data from different databases. We hope that all these databases will be taken care of by well-organized communities and perform the normative data standardization to enhance data sharing and exchange, while also pushing the applications of the model organism data.

Competing interests

The authors declared that there are no competing interests.

Acknowledgements

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences of China (Grant No. XDB13040500).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.gpb.2015.01.003.

References

[1] Dietrich MR, Ankeny RA, Chen PM. Publication trends in model organism research. Genetics 2014;198:787–94.
[2] Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JEMouse Genome Database group. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. Nucleic Acids Res 2014;42:D810–7.
[3] Kasprzyk A. BioMart: driving a paradigm change in biological data management. Database (Oxford) 2011;2011 [bar049].
[4] Smith RN, Aleksic J, Butano D, Carr A, Contrino S, Hu F, et al. InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics 2012;28:3163–5.
[5] Laulederkind SJ, Hayman GT, Wang SJ, Smith JR, Lowry TF, Nigam R, et al. The rat genome database 2013 – data, tools and users. Brief Bioinformatics 2013;14:520–6.
[6] Skrzypek MS, Hirschman J. Using the *Saccharomyces* Genome Database (SGD) for analysis of genomic information. Curr Protoc Bioinformatics 2011:1–3 [Chapter 1: Unit 1.20].
[7] Hibbs MA, Hess DC, Myers CL, Huttenhower C, Li K, Troyanskaya OG. Exploring the functional landscape of gene expression: directed search of large microarray compendia. Bioinformatics 2007;23:2692–9.
[8] Basu S, Fey P, Pandit Y, Dodson R, Kibbe WA, Chisholm RL. DictyBase 2013: integrating multiple Dictyostelid species. Nucleic Acids Res 2013;41:D676–83.
[9] Harris TW, Baran J, Bieri T, Cabunoc A, Chan J, Chen WJ, et al. WormBase 2014: new views of curated biology. Nucleic Acids Res 2014;42:D789–93.
[10] St Pierre SE, Ponting L, Stefancsik R, McQuilton P, FlyBase C. FlyBase 102 – advanced approaches to interrogating FlyBase. Nucleic Acids Res 2014;42:D780–8.
[11] Howe DG, Bradford YM, Conlin T, Eagle AE, Fashena D, Frazer K, et al. ZFIN, the zebrafish model organism database: increased support for mutants and transgenics. Nucleic Acids Res 2013;41:D854–60.
[12] Karpinka JB, Fortriede JD, Burns KA, James-Zorn C, Ponferrada VG, Lee J, et al. Xenbase, the *Xenopus* model organism database; new virtualized system, data types and genomes. Nucleic Acids Res 2014;43:D756–63.
[13] Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 2012;40:D1202–10.
[14] Kim HS, Murphy T, Xia J, Caragea D, Park Y, Beeman RW, et al. BeetleBase in 2010: revisions to provide comprehensive genomic information for *Tribolium castaneum*. Nucleic Acids Res 2010;38:D437–42.
[15] Wodke JA, Alibes A, Cozzuto L, Hermoso A, Yus E, Lluch-Senar M, et al. MyMpn: a database for the systems biology model organism *Mycoplasma pneumoniae*. Nucleic Acids Res 2014;43:D618–23.
[16] Schaeffer ML, Harper LC, Gardiner JM, Andorf CM, Campbell DA, Cannon EK, et al. MaizeGDB: curation and outreach go hand-in-hand. Database (Oxford) 2011;2011 [bar022].
[17] Glasner JD. ASAP, a systematic annotation package for community analysis of genomes. Nucleic Acids Res 2003;31:147–51.