Data Integration in the Mouse Genome Informatics (MGI) Database

Donnie Qi, Judith A. Blake, Jim A. Kadin, Joel E. Richardson, Martin Ringwald, Janan T. Eppig, Carol J. Bult and the Mouse Genome Informatics Group
The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA

Abstract

The Mouse Genome Informatics (MGI) Database at The Jackson Laboratory provides a comprehensive public resource about the laboratory mouse. MGI curated data types include gene representation, mapping, expression, gene function, orthology, mutant phenotype, polymorphism data and sequences. Our annotation system combines both computational and manual curation approaches to obtain high quality data representations, all of which are supported by evidence codes and citations. The relational integration of sequence information with existing biological information in MGI enables users to query for sequences using any combination of the biological attributes mentioned above and get results back in a variety of formats. Furthermore, MGI fully incorporates structured vocabularies such as the Gene Ontology (GO), Mammalian Phenotype (MP) Ontology, and the Adult Mouse Anatomy (MA) Ontology, and regularly exchanges data with many major public bioinformatics resources, including NCBI, SWISS-PROT, EBI, the GO, and other model organism databases. The MGI database can be accessed at http://www.informatics.jax.org.

1. Introduction

The Mouse Genome Informatics (MGI) database is the community database for the laboratory mouse [1]. MGI seeks to facilitate the use of the mouse as a model system for understanding human biology and disease by furthering our understanding of the relationship between genotype and phenotype. To achieve this mission MGI focuses on the integration of diverse heterogeneous data types including sequence, polymorphisms, gene expression and function, phenotypes, and mammalian orthology. The development and use of controlled vocabularies and bio-ontologies is key to the data integration efforts of MGI.

The MGI database include several related components, the Mouse Genome Database (MGD), the Gene Expression Database (GXD) [2], the Mouse Genome Sequence (MGS) project, the Mouse Tumor Biology (MTB) database [3], and the Gene Ontology (GO) [4]. The MGI database is updated daily by a staff of professional scientific data curators. All data in MGI are associated with evidence assertions and citations. A summary of MGI data content is listed at: ftp://ftp.informatics.jax.org/pub/reports/MGD_Stats.sql.rpt

2. Integration of sequence data into MGI

The availability of the nearly complete mouse genome sequence and other large sequence datasets presents a challenge of integrating sequence with existing biological information. MGI integrates these emerging public sequences with biology through the curated association of transcript, protein and genomic sequence objects with mouse genes [5, 6]. Our annotation system combines both computational and manual curation approaches to obtain high quality data representations. Sequence attributes represented in MGI include strain name, library name, sequence type, tissue of origin, and sequence length. Also included are the sequence source citations and associations of the sequence with a specific clone collection if appropriate. The relational integration of sequence information with existing biological information in MGI enables users to query for sequences using any combination of the biological attributes mentioned above and get results back in a variety of formats (http://www.informatics.jax.org/searches/sequence_form.shtml). Researchers can download sequence data in FASTA format from MGI or forward sequences directly to a BLAT server for a sequence similarity search.

3. Enhanced gene representations

MGI gene representations are integrated with a wealth of biological information about phenotypic alleles, expression, mammalian orthology, sequences, functional annotations, map position, mutants and strains. Recent enhancements of MGI gene detail pages include listing sequence coordinates (from both NCBI and EnsEMBL annotation of Build 33) and providing links to EnsEMBL ContigView, UCSC Genome Browser and the NCBI MapViewer. To provide a graphical representation of MGI data integrated with the mouse genome sequence we have implemented an interactive genome browser for the mouse using the Generic Genome Browser (GBrowse) [7]. MGI also provides extensive links to other informatics resources, and regularly exchanges data with NCBI, SWISS-PROT, EBI, the GO, and other model organism databases. The following is an example of the MGI gene detail page: http://www.informatics.jax.org/javawi2/servlet/WIFetch?page=markerDetail&key=9936

4. Controlled vocabularies and ontologies

MGI fully incorporates structured vocabularies and ontologies (http://www.informatics.jax.org/menus/vocab_menu.shtml). MGD is the authoritative source for the official names and symbols associated with mouse genes, alleles and strains. Gene ontologies are used extensively for gene annotation in MGI [8]. The Mouse Embryo Anatomy Nomenclature Database [9] and the Anatomical Dictionary for the Adult Mouse [10] have been adopted for annotating data that include anatomical attributes, such as tissue sources for clones and phenotypes. In addition, the Mammalian Phenotype (MP) ontology is used in phenotypic data annotations at MGI [11].

5. Implementation

MGD is implemented in the Sybase relational database system, version 12.5. A large set of CGI scripts and Java Servlets mediate the user's interaction with the database. For computational users, direct SQL access can be requested through User Support. User-requested database reports and a number of widely used data files (generated daily) are available on the ftp site: ftp://ftp.informatics.jax.org/pub/reports/index.html

6. Acknowledgements

The Mouse Genome Database is supported by NIH/NHGRI grant HG00330.

7. References

[1] Bult CJ, Blake JA, Richardson JE, Kadin JA, Eppig JT, and the members of the Mouse Genome Database Group. The Mouse Genome Database (MGD): integrating biology with the genome. *Nucleic Acids Res* 2004, 32: D476-D481.

[2] Hill,D.P., Begley,D.A., Finger,J.H., Hayamizu,T.F., McCright,I.J., Smith,C.M., Beal,J.S., Corbani,L.E., Blake,J.A., Eppig,J.T., Kadin,J.A., Richardson,J.E. and Ringwald,M. The Mouse Gene Expression Database (GXD): updates and enhancements. *Nucleic Acids Res*. 2004, 32: D568–D571

[3] Näf,D., Krupke,D.M., Sundberg,J.P., Eppig,J.T. and Bult,C.J. ((2002)) The mouse tumor biology database: a public resource for cancer genetics and pathology of the mouse. *Cancer Res.* 2002, 62: 1235–1240.

[4] The Gene Ontology Consortium. The Gene Ontology (GO) Database and Informatics Resource. *Nucleic Acids Res*. 2004, 32: D258–D261.

[5] Zhu,Y., King,B.L., Parvizi,B., Brunk,B.P., Stoeckert,C.J.,Jr, Quackenbush,J., Richardson,J., and Bult,C.J. Integrating computationally assembled mouse transcript sequences with the Mouse Genome Informatics (MGI) database. *Genome Biol*., 2003, 4, R16.

[6] Baldarelli,R.M., Hill,D.P., Blake,J.A., Adachi,J., Furuno,M., Bradt,D., Corbani,L.E., Cousins,S., Frazer,K.S., Qi,D. et al. Connecting sequence and biology in the laboratory mouse. *Genome Res.,* 2003, 13, 1505–1519

[7] Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S. The generic genome browser: a building block for a model organism system database. *Genome Res*. 2002, 12(10), 1599-610.

[8] Hill DP, Davis AP, Richardson JE, Corradi JP, Ringwald M, Eppig JT, Blake JA. Strategies for biological annotation of mammalian systems: implementing gene ontologies in mouse genome informatics. *Genomics* 2001, 74:121-128.

[9] Bard JL, Kaufman MH, Dubreuil C, Brune RM, Burger A, Baldock RA, Davidson DR: An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. *Mech Dev* 1998, 74:111-120.

[10] Hayamizu TF, Mangan M, Corradi JP, Kadin JA, Ringwald M. The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. *Genome Biol.* 2005, 6(3): R29.

[11] Smith CL, Goldsmith CA, Eppig JT. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.,* 2005, 6: R7.