108–111 *Nucleic Acids Research, 2000, Vol. 28, No. 1*

The Mouse Genome Database (MGD): expanding genetic and genomic resources for the laboratory mouse

Judith A. Blake*, Janan T. Eppig, Joel E. Richardson, Muriel T. Davisson and the Mouse Genome Database Group

The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA

Received October 1, 1999; Accepted October 7, 1999

ABSTRACT

The Mouse Genome Database (MGD) is a comprehensive public database of mouse genomic, genetic and phenotypic information (http://www.informatics.jax.org). This community database provides information about genes, serves as a mapping resource of the mouse genome, details mammalian orthologs, integrates experimental data, represents standardized mouse nomenclature for genes and alleles, incorporates links to other genomic resources such as sequence data, and includes a variety of additional information about the laboratory mouse. MGD scientists and annotators work cooperatively with the research community to provide an integrated, consensus view of the mouse genome while also providing experimental data including data conflicting with the consensus representation. Recent improvements focus on the representation of phenotypic information and the enhancement of gene and allele descriptions.

INTRODUCTION

First released in 1994, MGD has evolved from a mapping and genetics resource to include sequence and genome information and expanded details on the function and role of genes and alleles (1). MGD is one of several Mouse Genome Informatics (MGI) projects underway at The Jackson Laboratory (TJL). Other projects include the Gene Expression Database (GXD) (2) and the Mouse Tumor Biology (MTB) (3) database.

During 1999, two new MGD releases and several incremental software improvements were made. MGD is updated daily through new annotations of mapping, homology, gene information, and from electronically submitted data sets. MGD data coverage was expanded this year with the addition of strain polymorphisms for 129/J versus other inbred strains, the addition of new Recombinent Inbred (RI) strain distribution patterns, improved representation of ESTs and other sequence information, and new information supporting homology assertions. Current status and further enhancements are detailed in this report.

IMPROVEMENTS DURING 1999

In depth coverage of MGD structural organization and overall coverage can be found in previous publications (1,4–6). The numbers below were gathered from database statistics as of September 27, 1999.

Enhanced gene searching and display of gene/marker details

Quick Gene Search and Query Search Forms. A quick gene search is now possible directly from the home page. This search supports wildcards and queries of standard gene symbols/names as well as of synonyms and withdrawn symbols. All web pages now include a search forms selection list just beneath the banner. While the text links in the banner provide access to various menus and other resources, the Search Forms menu provides quick access to the query forms themselves. The Search Forms menu replaces the Query Menu and works with all platforms and browsers.

Enhanced gene/marker detail display. Sequence accession IDs for genes are curated from literature and through MGD annotation efforts. The gene/marker detail display now includes curated sequence links to DDBJ (7), EMBL (8), GSDB (9) and GenBank (10). As an example, see the gene detail for *Tyr* which currently lists nine sequence accession ID links. Collaboration with SWISS-PROT curators has resulted in >4000 gene records being linked to SWISS-PROT (11). The integration of nucleotide and amino-acid sequence links with the gene detail reports results in a highly curated set of data. The user can now click from the gene detail page to extensive information about alleles, sequences, mapping experiments, polymorphisms, references and expression data.

Genes and genetic markers

The number of genes and genetic markers detailed within MGD continues to increase rising to >26 500 this year including >10 560 genes. This includes over 22 800 mapped loci and over 7096 mapped genes. Genetic marker types in MGD include genes, chromosomal aberrations, QTLs, anonymous DNA segments and phenotypically defined Mendelian traits. This year the 10 000th gene with enough biological characterization to be uniquely named and described was annotated in MGD (*Scml2*). Over 3700 genes have been matched with their human ortholog and >1484 with their rat ortholog, each with

*To whom correspondence should be addressed. Tel: +1 207 288 6248; Fax: +1 207 288 6131; Email: jblake@informatics.jax.org

The Mouse Genome Database Group: R. M. Baldarelli, J. S. Beal, R. E. Blackburn, D. Bradt, N. E. Butler, L. E. Corbani, A. P. Davis, J. C. Gilbert, L. H. Glass, P. L. Grant, M. Lennon-Pierce, C. M. Lutz, L. J. Maltais, P. Mani, M. E. May, L. M. McKenzie, J. J. Merriam, J. E. Ormsby, S. Ramachandran, D. J. Reed and C. G. Wray

supporting documentation and evidence assertions. Ortholog relationships are detailed with many other mammalian species as well.

Evidence for homology assertions

Homology reports in MGD have always included the ortholog gene accession IDs where possible, the citation which detail the determination of the homology assertion, and links to specific gene information in other genome databases. In the recent MGD release, the type of evidence used to determine the homology relationship is also provided. For example, a researcher can determine easily which relationships are based solely on sequence similarity, and which on conserved location, sequence similarity and function analysis. The homology criteria used to support orthologous relationship determination conforms to the HUGO comparative mapping guidelines (12).

Homology and comparative genomics

Mammalian gene orthologs are of great interest to researchers. In MGD, most mammalian homology data comes from peer-reviewed publications. Some additional homologies derive from the work of the human and mouse nomenclature committees and from interaction with groups of researchers who are revising and updating gene family relationships. When possible, mouse genes are linked to other databases such as the Genome Database (GDB) (13) so users can obtain further information about the orthologs in other species. The data in MGD can be searched by species, gene symbol, name or map position. A nightly database report includes listings of all mouse/human homologies and mouse/rat homologies sorted by chromosome, gene symbol, or other criteria. Comparative maps can be built using the linkage map building tools.

With the advent of nearly completely sequenced genomes from several of the Human Genome Project organisms, MGD is revising its representation of homologies. From now on, the group of mammalian species for which homology data will be collected will only include selected primates, rodents, experimental and domestic species. MGD will also start to emphasize the relationship of mouse genes to those in other model organisms such as *Drosophila*. These relationships are being represented both in MGD and in the FlyBase database (14).

Mapping status and enhancements

The integration of experimental mapping data continues to be a priority for MGD curators. Over 22 800 genetic markers have been mapped and are represented in MGD. Of the 10 560 genetic markers identified as coding genes, 7096 have associated mapping data and are represented on the mouse genetic map. Queries for mapping data allow the user to request mapping information by the type of mapping assay used (e.g., FISH versus DNA mapping panel).

Among the mapping data sets available are 12 DNA Mapping Panels, composite sets of RI Strain Distribution Patterns, and Recombinant Congenic (RC) Strain Distribution Patterns as well as thousands of smaller sets of mapping data. Graphical map displays are available for linkage, cytogenetic and physical maps. Genetic maps can be constructed with user-defined parameters including choice of data set and type of gene or marker (e.g., only genes involved in neurological disorders). Maps can be built with cross-referencing to homologous genes of mammalian species including human.

Nomenclature

Gene nomenclature continues to play a key role in the data curation process. The integration of genetic and genomic data in MGD depends on the curation of a unique set of gene symbols and names for the laboratory mouse. The MGD Nomenclature Committee works under the guidelines set by the International Committee on Standardized Genetic Nomenclature for Mice (15). Several journals now require review of gene nomenclature as part of the manuscript review process. The MGD nomenclature coordinator works with researchers to clarify and resolve gene nomenclature issues before publication. Scientists can obtain new gene symbols rapidly through the use of the nomenclature electronic submission form (http://www.informatics.jax.org/support/nomen/nomen_submit_form.shtml).

This year, MGD nomenclature experts worked with researchers to revise the nomenclature and orthologous determinations for gene families. Once agreed upon, the gene symbols/names and homologous relationships were updated in MGD. Gene family information is often consolidated and posted on the Web by interested groups of scientists. MGD links to these Web pages and works closely with the scientists to resolve outstanding nomenclature issues. International Nomenclature Workshops, sponsored in part by MGD, are also bringing the community together to resolve issues in gene and gene family nomenclature (16).

Update of function/phenotype information

Enhanced gene descriptions with links to OMIM. MGD supports two representations of gene information. The Gene Detail page reports the gene symbol, name, map position and classification status. Primarily, however, it serves as an integration site from which users can link to more detailed information about the gene including expression data, mapping experiments, homology reports, allele and polymorphism details and links to sequences. In comparison, the Gene Description page (formerly the Mouse Locus Catalog report) provides more detailed information about the gene including a succinct description of the function of the gene product, relationship to other mouse genes and particularly to human genes and information about any phenotypic alleles.

This year the Gene Description page has been re-formatted and compartmentalized into five categories: gene family/protein domains, mouse model/disease, gene/gene product information, expression, and phenotypes/alleles and variants. This new structure allows users to rapidly locate specific information from the text.

MGD welcomes contributions from the research community, especially in the description of phenotypic alleles (contact us through the User Support address below). We encourage users to write short reports on their genes of expertise. Such contributions will be attributed in the Web posting.

Biological ontologies. The standardization of terms and vocabularies within MGD and GXD facilitates data entry and searching. The classification of genes by function, cellular component and biological process using a controlled vocabulary is underway. We are collaborating with the *Drosophila* (14) and *Saccharomyces cerevisiae* (17) database annotators to provide a shared ontology. Each group is using this vocabulary to annotate the genes or gene products. As a collaboration, we

continue to develop the vocabulary and are working toward a shared resource that will allow searches for biological attributes that return cross-species results to users.

Mouse Facts

Mouse Facts pages are a new addition providing information about the laboratory mouse. Currently the pages include the following information: mouse physiology information, mouse genome sequencing progress summaries, genome length estimates, chromosomal length and gene distributions, and links to NIH, molecular resources and other sites.

Electronic data submission

We encourage contributions of electronic data sets from the scientific community. Any type of data that MGI databases maintain can be submitted as an electronic contribution. The most common data submissions this year were mapping data, molecular polymorphism data and mammalian homology information. Each electronic submission receives a permanent database accession ID and is assigned a citation ID with an abstract if appropriate. Information from individual researchers and from community annotation efforts is being incorporated into the MGD resource.

Community outreach and user support

MGD provides extensive user support through online documentation and easy Email or phone access to User Support staff.

User Support WWW access:	http://www.informatics.jax.org/support/support.shtml
Email:	mgi-help@informatics.jax.org
Tel:	+1 207 288 6445
Fax:	+1 207 288 6132

User Support staff develop and maintain online help documentation for MGI database resources, assist users with database questions and provide training and demonstrations for TJL staff, students in courses and workshops conducted at TJL and attendees at scientific meetings. In addition, support staff manage an electronic bulletin board service which includes MGI-LIST, an extremely active list with >1300 researchers subscribed. Users can subscribe directly on the Web. Chromosome Committees rely on User Support for assistance with online submission of their annual Chromosome Committee reports. User Support staff collect feedback and demographic information from researchers (via user registrations and a recent user survey) that can provide input to ongoing development of MGI information resources. Currently, there are >2700 registered users.

All MGD staff are involved in community outreach in various ways. Curatorial staff are frequently called on to investigate a researcher's questions about data in MGD and to assist with data submission. Software staff assist with curation access issues and provide technical support to researchers and organizations having special needs for access beyond the public web interface.

Mirror sites

There are now five mirror sites around the world (UK, Japan, France, Australia and Israel). Mirror sites have the option of downloading FTP update files on a nightly basis and provide users with faster local access to MGD. Most sites update on a regular basis.

IMPLEMENTATION

MGD is implemented in the Sybase relational database system, version 11.5.1. The Web interface comprises a set of static HTML forms and other supporting documents and a large set of CGI scripts, written in Python, mediate the user's interaction with the database. Users may also access MGD directly via SQL. Users requiring an SQL account may contact MGI User Support.

CITING MGD

The following citation format is suggested when referring to specific datasets within MGD: Mouse Genome Database (MGD), Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, Maine (URL: http://www.informatics.jax.org). [Type in date (month, year) when you retrieved the data cited.]

ACKNOWLEDGEMENT

The Mouse Genome Database is supported by NIH grant HG00330.

SUPPLEMENTARY MATERIAL

Links to various MGD web pages are provided at NAR Online.

REFERENCES

1. Blake,J.A., Richardson,J.E., Davisson,M.T., Eppig,J.T. and the *Mouse Genome* Informatics Group (1997) *Nucleic Acids Res.*, **25**, 85–91.
2. Ringwald,M., Mangan,M.E., Eppig,J.T., Kadin,J.A., Richardson,J.E. and the Gene Expression Database Group (1998) *Nucleic Acids Res.*, **27**, 106–112. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 115–119.
3. Bult,C.J., Krupke,D.M. and J.T. Eppig. (1998) *Nucleic Acids Res.*, **27**, 99–105. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 112–114.
4. Blake,J.A., Eppig,J.T., Richardson,J.E., Davisson,M.T. and the Mouse Genome Informatics Group (1998) *Nucleic Acids Res.*, **26**, 130–137.
5. Blake,J.A., Richardson,J.E., Davisson,M.T., Eppig,J.T. and the Mouse Genome Informatics Group (1999) *Nucleic Acids Res.*, **27**, 95–98.
6. Eppig,J.T., Blake,J.A., Davisson,M.T. and Richardson,J.E. (1998) *Methods: Companion Methods Enzymol.*, **14**, 179–190.
7. Sugawara,H., Miyazaki,S., Gojobori,T. and Tateno,Y. (1999) *Nucleic Acids Res.*, **27**, 25–28. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 24–26.
8. Stoesser,G., Tuli,M.A., Lopez,R. and Sterk,P. (1999) *Nucleic Acids Res.*, **27**, 18–24. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 19–23.
9. Skupski,M.P., Booker,M., Farmer,A., Harpold,M., Huang,W., Inman,J., Kiphart,D., Kodira,C., Root,S., Schilkey,F., Schwertfeger,J. *et al.* (1999) *Nucleic Acids Res.*, **27**, 35–38. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 31–32.
10. Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouellette,B.F.F., Rapp,B. and Wheeler,D.L. (1999) *Nucleic Acids Res.*, **27**, 12–17. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 15–18.
11. Bairock,A. and Apweiler,R. (1998) *Nucleic Acids Res.*, **27**, 49–54. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 45–48.
12. Andersson,L.A., Archibald,M., Ashburner,M., Audun,S., Barendse,W., Bitgood,J., Bottema,C., Broad,T., Brown,S., Burt,D. *et al.* (1996) *Mamm. Genome*, **7**, 717–734.
13. Letovsky,S.I., Cottingham,RW., Porter,C.J. and Li,P.W.D. (1998) *Nucleic Acids Res.*, **26**, 94–99.
14. The FlyBase Consortium (1999) *Nucleic Acids Res.*, **27**, 85–88.

15. Maltais,L.J., Blake,J.A., Eppig,J.T. and Davisson,M.T. (1997) *Genomics*, **45**, 471–476.
16. Blake,J.A., Davisson,M.T., Eppig,J.T., Maltais,L.J., Povey,S., White,J.A. and Womack,J.E. (1997) *Genomics*, **45**, 464–468.
17. Ball,C.A., Dolinski,K., Dwight,S.S., Harris,M.A., Issel-Tarver,L., Kasarskis,A., Scafe,C.R., Sherlock,G., Binkley,G., Jin,H. *et al.* (2000) *Nucleic Acids Res.*, **28**, 77–80 (this issue).