

The Mouse Genome Database (MGD): integration nexus for the laboratory mouse

Judith A. Blake*, Janan T. Eppig, Joel E. Richardson, Carol J. Bult and J. A. Kadin and the Mouse Genome Database Group

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

Received October 2, 2000; Accepted October 4, 2000

ABSTRACT

The Mouse Genome Database (MGD) is the community database resource for the laboratory mouse, a key model organism for interpreting the human genome and for understanding human biology and disease (<http://www.informatics.jax.org>). MGD provides standard nomenclature and consensus map positions for mouse genes and genetic markers; it provides a curated set of mammalian homology records, user-defined chromosomal maps, experimental data sets and the definitive mouse 'gene to sequence' reference set for the research community. The integration and standardization of these data sets facilitates the transition between mouse DNA sequence, gene and phenotype annotations. A recent focus on allele and phenotype representations enhances the ability of MGD to organize and present data for exploring the relationship between genotype and phenotype. This link between the genome and the biology of the mouse is especially important as phenotype information grows from large mutagenesis projects and genotype information grows from large-scale sequencing projects.

INTRODUCTION

The Mouse Genome Database (MGD) has been available to the scientific community via the Web since 1994 (1). MGD represents mouse genetic, genomic and biological data of many kinds, including gene identification, allelic variants, sequence links, maps and mapping data, mutant and strain phenotype descriptions, orthologous gene relationships, functional classifications and molecular segment data. Data are gathered and integrated from many sources including genome centers, individual laboratory groups and the scientific literature. MGD provides a rich set of links to other key information resources and engages in active and ongoing collaborations for annotating mouse data with SWISS-PROT, RIKEN, NCBI and others.

The MGD is the major database component of the Mouse Genome Informatics (MGI) database system ([http://](http://www.informatics.jax.org)

www.informatics.jax.org) at The Jackson Laboratory. The MGI system consists of an integrated community database with associated resources dedicated to providing genetic, genomic and biological information about the laboratory mouse. Other components of the integrated MGI system are the Gene Expression Database (GXD) (2) and the Mouse Genome Sequence (MGS) project. Scientific curators from the entire MGI system, as well as curators for the Mouse Tumor Biology (MTB) (3) prototype database, work together to provide the core information of MGD including standardized gene descriptions, mammalian ortholog assertions and sequence information.

During 2000, two new MGD software releases and several incremental software improvements were made. MGD is updated daily through additional annotations of new genes, location data, mammalian homologies and new alleles. MGD data coverage was expanded this year with extensive annotation of mouse genes to controlled vocabularies describing function and cellular location and with major enhancements to the representation of alleles. Recent enhancements are detailed in this report. MGD structural organization and previous enhancements are described in previous publications (4–6).

IMPROVEMENTS DURING 2000

New representation of alleles

Until recently, MGD represented alleles in textual reports and as attributes of genes. With the expansion of gene targeting experiments and the initiation of large-scale chemical mutagenesis projects, this representation of alleles was inadequate. Alleles are now represented as unique objects in the database and each allele is fully annotated (Fig. 1). Standardization of annotation was partially accomplished through the introduction of several controlled vocabularies such as 'strain of origin', 'molecular mutation' and 'mode of inheritance'. These changes are the beginning of an overall extension in the way MGD is managing increased allele and phenotype information.

Genealogies of mouse inbred strains

The differences between inbred strains are ever more important as we expand the use of mouse systems as models for human genetic diseases and seek to understand differences in disease susceptibility

*To whom correspondence should be addressed. Tel: +1 207 288 6248; Fax: +1 207 288 6132; Email: jblake@informatics.jax.org

Mouse Genome Informatics
Main Menu | MGI Home | User Support | Help Documents | Submissions | Citations
Markers | Molecular | Homology | Mapping | Expression | Strain/Polymorphism | Ref | Acrid

Query Form

Genes, Markers and Phenotypes
Query Results -- Details

Type: Gene
Symbol: *Vdr*
Name: vitamin D receptor
Chromosome: 5
cM Position: 56.0
MGI Accession ID: MGI:103076

Synonyms:

- Nr1h1

Additional Information:

- Mammalian Homology
- Marker Mapping Data (2)
- Phenotypic Alleles (1)
- RFLP/PCR Polymorphism (1)
- GXD Index Data (3)
- Molecular Probes and Segments (5)
- References (14)

Gene Classifications: (You can browse the Gene Ontology (GO) Classification)

Category	Classification Term	Evidence	Reference
Biological Process	axial development	inferred from mutant phenotype	J42054
Biological Process	calcium ion homeostasis	inferred from mutant phenotype	J42054
Biological Process	developmental processes	inferred from mutant phenotype	J42054
Biological Process	embryogenesis and morphogenesis	inferred from mutant phenotype	J42054
Biological Process	transcription regulation	electronic annotation	J60000
Cellular Component	nucleus	electronic annotation	J60000
Molecular Function	DNA binding	electronic annotation	J60000

Other Database Links for this Marker:

Acc ID	Links	Reference
AF017775	(DBJ, EMBL, GSD, GenBank)	J42054
D31969	(DBJ, EMBL, GSD, GenBank)	J22664
F48281	(SWISS-PROT (EBI), SWISS-PROT (SIB))	J53168
44170	(UniGene)	J57747

History (symbol/name/reference/event/date):

- Vdr*/vitamin D receptor/J.22664/assigned/08/21/1995

Other MGI Accession IDs: MGD-MRK-23940

Phenotypic Alleles
Query Results -- Details

Allele Symbol: *Vdr*^{tm1.1ka}
Allele Name: targeted mutation 1, Shigeki Kato
Gene Symbol: *Vdr*
Gene Name: vitamin D receptor
MGI Accession ID: MGI:1859929
Original Reference: J42054
Type: Transgene induced (gene targeted)
Strain of Origin: CBA and C57BL/6J
Mode of Inheritance: Not Specified

Molecular Description
Mutation(s): Deletion caused by insertion of targeting vector

Synonyms:

Symbol	Reference
VDR-/-	J42054

Additional Notes
ES cell line = TT2

Associated Phenotype Controlled Terms

- postnatal lethal
- infertile
- alopecia
- growth retardation
- reduced vibrissae
- flat face
- short nose
- hypocalcaemia
- abnormal bone formation
- impaired folliculogenesis
- uterine hypoplasia

Relationship to Human Genes and Disease

- Orthologous Human Gene: VITAMIN D RECEPTOR, VDR OMIM
- Phenotypic Similarity to Human Syndrome in Orthologous Human Gene: VITAMIN D-RESISTANT RICKETS WITH END-ORGAN UNRESPONSIVENESS TO 1,25-DIHYDROXYCHOLECALCIFEROL OMIM
- Phenotypic Similarity to Human Syndrome in Orthologous Human Gene: OSTEOPOROSIS, INVOLUTIONAL OMIM

Figure 1. MGD gene and allele detail pages for *Vdr*. Gene detail report for *Vdr*, the gene for the vitamin D receptor, and the new allele detail report for an allele of the *Vdr* gene. The gene detail report includes annotation to the GO as well as extensive links to sequence representations for this gene. The allele detail report includes information about the source of the allele as well as details of the phenotype presentation and information about its utility as a model for human phenotypes.

between strains. MGD staff collaborated in the publication of detailed genealogies of inbred strains of mice (7). A poster illustrating these origins and relationships is freely available at the MGI web site (<http://www.informatics.jax.org/mgihome/genealogy/>) and, with journal subscription, at the *Nature Genetics* web site (<http://genetics.nature.com/mouse>). This chart will be updated and the annotations extended in an ongoing curation effort for these data. Researchers are invited to submit additional information on genealogical and genetic data of inbred mouse strains to MGD at submissions@informatics.jax.org.

Expanded representation of sequence information

Recently, much attention has been focused on an expanded association of mouse genes with sequence information. With the surge in new gene discovery as a result of genome sequencing projects, correctly associating genes with sequence data (and thereby associating sequences with biological information about genes, alleles and phenotypes) is an important focus of genomic integration. MGD and GXD collaboratively provide the gene to sequence associations for the genomics

community. We are collaborating with SWISS-PROT, UniGene and LocusLink in our 'gene to sequence' integration efforts. While the SWISS-PROT effort has been ongoing for several years, the coordinated links to sequence data through collaborations with LocusLink and UniGene groups are new this year.

We provide gene map position and official nomenclature for mouse genes. Based upon our 'gene to sequence' associations, LocusLink and MGD curators share the annotation of new gene records to sequence reports. MGD provides the official nomenclature for all genes. Links from LocusLink are made back to MGD reports both as an attribute of the mapping data and with the recognition that the LocusLink record contains official gene symbols and names as provided by MGD.

In the UniGene collaboration, the associations between sequence and gene made by MGD are the basis for the association of a UniGene record and a mouse gene. The UniGene association then provides the basis for the putative associations between the EST sequences grouped by the UniGene clustering algorithm and MGI records for I.M.A.G.E. clones. The integrated nature

of the MGI system results in the capability to search the data by various sequence IDs, gene names or other parameters to recover specific subsets of information.

MouseBLAST

A new feature provided via the MGI web site is the MouseBLAST server (B.L.King *et al.*, manuscript in preparation). This utility provides an entry point into MGD using the WU-BLAST 2.0 set of sequence similarity searching programs (WU-BLAST 2.0 software package; <http://blast.wustl.edu>). Users can search sequences against several locally maintained nucleotide and protein sequence databases that are grouped by species (e.g., all rodent sequences, mouse sequences, rat sequences, etc). If a returned sequence is associated with a gene in MGD, a link to that gene is embedded in the WU-BLAST 2.0 output. MouseBLAST was established through a collaboration between MGD and MGS.

Update of function/phenotype information

Biological ontologies. MGD is putting extensive effort into the development of controlled structured vocabularies. Development of these vocabularies permits both complex structured annotations as well as more sophisticated query capabilities. As founding members of the Gene Ontology (GO) consortium (8), MGI curators annotate the function and cellular location of gene products using the GO vocabularies. These annotations are presented on detailed gene reports, are available via our ftp site and are regularly submitted to the GO web site where they can be queried as part of a shared model organism resource (<http://www.geneontology.org/>). Details of the implementation of the functional annotation of mouse genes using the GO are available on the MGI site (<http://www.informatics.jax.org/mgihome/GO/ontology.shtml>).

Disease models. The laboratory mouse is the premier animal model used for the study of human diseases. A recent enhancement of MGD is the association of mouse mutants with a structured, controlled set of disease terms. The association is of three types. A mouse model may (i) exhibit a disease phenotype, (ii) be studied as a model for a human disease or (iii) be the ortholog of a human gene that is associated with a disease. A new phenotype query form allows users to query the database by a disease term and recover a set of objects, genes, alleles and/or strains that have been associated with the disease.

Nomenclature. The curation of a unique set of symbols and names for mouse genes facilitates the integration of genetic and genomic data in MGD. The MGD Nomenclature Committee assigns unique symbols and names to mouse genes under the guidelines set by the International Committee on Standardized Genetic Nomenclature for Mice and in conjunction with researchers and collaborators such as the HUGO Nomenclature committee (9). Several journals such as *Nature Genetics* and *Genomics* 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 rapidly reserve symbols for newly identified genes through the use of the nomenclature electronic submission form (http://www.informatics.jax.org/mgihome/nomen/nomen_submit_form.shtml).

Additional enhancements

Enhanced gene/marker representations. In response to user requests, we have replaced the two fields, marker symbol and marker name, with a single symbol/name field on most query forms. Any search term in this field queries official symbols and names as well as synonyms.

Enhancements to the gene/marker display include providing better representation of map positions for QTL markers and revising the semantics regarding synonyms. Linkage positions for QTLs are typically vague. So, the mapped position of QTLs in MGD is now reported as the 'cM position of the peak correlated region/marker'. In the case of synonyms, these unofficial symbols and names are now included in the combined symbol/name searches.

This year, MGD nomenclature experts worked with researchers to revise the nomenclature and orthologous determinations for many gene families. Once agreed upon, the gene symbols/names and homologous relationships were updated in MGD. MGD also links to specialized gene family web pages and recent updates of mouse gene families are posted from the MGI homepage.

OTHER INFORMATION

Electronic data submission

We encourage contribution 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. For information about submitting data, see http://www.informatics.jax.org/mgihome/submissions/submissions_menu.shtml.

Community outreach and user support

MGD provides extensive user support through online documentation and easy email or telephone access to User Support Staff.

User Support WWW access:

<http://www.informatics.jax.org/mgihome/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 participants at courses and conferences conducted at The Jackson Laboratory and at other venues for scientific meetings. In addition, support staff manage an electronic bulletin board service which includes MGI-LIST, an extremely active list with over 1600 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 over 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 with special needs for access beyond the public web interface. In addition, MGD staff conduct demonstrations and present posters at numerous scientific meetings each year.

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. A large set of CGI scripts, written in Python, mediate the user's interaction with the database. For computational users, direct SQL access can be requested through User Support. User-requested special SQL reports and a number of widely used data files (generated daily) are available on the ftp site.

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.]

ACKNOWLEDGEMENTS

The Mouse Genome Database is supported by NIH grant HG00330. The MouseBLAST is supported by NIH:HG01559 and DOE:DE-FG02-99ER62850.

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) The Mouse Genome Database (MGD). A comprehensive public resource of genetic, phenotypic and genomic data. *Nucleic Acids Res.*, **25**, 85–91.
2. Ringwald, M., Eppig, J.T., Kadin, J.A., Richardson, J.E. and the Gene Expression Database Group (2000) GXD: a Gene Expression Database for the laboratory mouse: current status and recent enhancements. *Nucleic Acids Res.*, **28**, 115–119. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 98–101.
3. Bult, C.J., Krupke, D.M. and J.T. Eppig. (1998) Electronic access to mouse tumor data: the Mouse Tumor Biology Database (MTB) project. *Nucleic Acids Res.*, **27**, 99–105.
4. Blake, J.A., Eppig, J.T., Richardson, J.E., Davisson, M.T. and the Mouse Genome Database Group (2000) The Mouse Genome Database (MGD): expanding genetic and genomic resources for the laboratory mouse. *Nucleic Acids Res.*, **28**, 108–111.
5. Blake, J.A., Richardson, J.E., Davisson, M.T., Eppig, J.T. and the Mouse Genome Database Group (1999) The Mouse Genome Database (MGD): genetic and genomic information about the laboratory mouse. *Nucleic Acids Res.*, **27**, 95–98.
6. Blake, J.A., Eppig, J.T., Richardson, J.E., Davisson, M.T. and the Mouse Genome Informatics Group (1998) The Mouse Genome Database (MGD): a community resource. Status and enhancements. *Nucleic Acids Res.*, **26**, 130–137.
7. Beck, J.A., Lloyd, S., Hafezparast, M., Lennon-Pierce, M., Eppig, J.T., Festing, M.F.W. and Fisher, M.C. (2000) Genealogies of mouse inbred strains. *Nature Genet.*, **24**, 23–25.
8. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.E., Harris, M.A., Hill, D.P. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
9. Maltais, L.J., Blake, J.A., Eppig, J.T. and Davisson, M.T. (1997) Rules and Guidelines for Mouse Gene Nomenclature: A condensed version. *Genomics*, **45**, 471–476.