D724–D728 *Nucleic Acids Research, 2008, Vol. 36, Database issue*
doi:10.1093/nar/gkm961

The Mouse Genome Database (MGD): mouse biology and model systems

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

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

Received September 19, 2007; Revised October 14, 2007; Accepted October 15, 2007

ABSTRACT

The Mouse Genome Database, (MGD, http://www.informatics.jax.org/), integrates genetic, genomic and phenotypic information about the laboratory mouse, a primary animal model for studying human biology and disease. MGD data content includes comprehensive characterization of genes and their functions, standardized descriptions of mouse phenotypes, extensive integration of DNA and protein sequence data, normalized representation of genome and genome variant information including comparative data on mammalian genes. Data within MGD are obtained from diverse sources including manual curation of the biomedical literature, direct contributions from individual investigator's laboratories and major informatics resource centers such as Ensembl, UniProt and NCBI. MGD collaborates with the bioinformatics community on the development of data and semantic standards such as the Gene Ontology (GO) and the Mammalian Phenotype (MP) Ontology. MGD provides a data-mining platform that enables the development of translational research hypotheses based on comparative genotype, phenotype and functional analyses. Both web-based querying and computational access to data are provided. Recent improvements in MGD described here include the association of gene trap data with mouse genes and a new batch query capability for customized data access and retrieval.

INTRODUCTION

The Mouse Genome Database (MGD) is an integrated database of genetic, genomic and phenotypic data for the laboratory mouse (1–3). MGD is a core database component of the Mouse Genome Informatics (MGI) database resource (http://www.informatics.jax.org), the community model organism database for the laboratory mouse. Other resources that are integrated with MGD as part of the MGI resource include the Gene Expression Database (GXD) (4), the Mouse Tumor Biology Database (MTB) (5) and the Gene Ontology project (6).

MGD facilitates translational biomedical research by integrating data that enhances the use of the laboratory mouse as a model animal system for studying human biology and disease processes. MGD supports genome-scale electronic data mining through its integration of diverse data and use of semantic standards. Primary data types maintained in MGD include sequences, genetic and physical maps, genes and their functions, gene families, strains, mutant phenotypes, SNPs and other polymorphisms, animal models of human disease and mammalian homology (Table 1). The diverse data in MGD are integrated through a combination of expert human curation and automated processes that evaluate when different data refer to the same gene. MGD employs controlled and structured vocabularies (i.e. ontologies) to facilitate knowledge representation and data retrieval. Examples of vocabularies and ontologies that are used for annotation include the Gene Ontology (GO), and the Mammalian Phenotype (MP) Ontology (7). Mouse genes and gene products in MGD are also associated with Online Mendelian Inheritance in Man (OMIM) human phenotype terms, InterPro protein domain descriptions

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

†The Mouse Genome Database Group: M.T. Airey, A. Anagnostopoulos, R. Babiuk, R.M. Baldarelli, M. Baya, J.S. Beal, S.M. Bello, D.W. Bradt, D.L. Burkart, N.E. Butler, J. Campbell, L.E. Corbani, S.L. Cousins, D.J. Dahmen, H. Dene, M.E. Dolan, K.L. Forthofer, K.S. Frazer, P. Frost, M. Hall, M. Knowlton, J.R. Lewis, I. Lu, L.J. Maltais, M. McAndrews-Hill, S. McClatchy, M.J. McCrossin, J. Mason, D.B. Miers, L.A. Miller, L. Ni, H. Onda, J.E. Ormsby, T.B.K. Reddy, D.J. Reed, B. Richards-Smith, D.R. Shaw, R. Sinclair, D. Sitnikov, C.L. Smith, P. Szauter, M. Tomczuk, M.A. Updegraff, L.L. Washburn, I.T. Witham and Y. Zhu.

Table 1. Summary of MGD data content (5 September 2007)

MGD data statistics	5 September 2007
Number of genes with sequence data	29 031
Number of genes (including uncloned mutations)	36 473
Number of markers (including genes)	72 020
Markers mapped	67 536
Genes with protein sequence information	24 961
Genes with GO annotations	18 049
Mouse/human orthologs	16 927
Mouse/rat orthologs	15 801
Genes with one or more phenotypic alleles	7205
Genes with targeted alleles	4751
Phenotypic alleles	18 491
Phenotype alleles that are targeted mutations	10 536
Human diseases with one or more mouse models	790
QTLs	3601
Number of references	119 121
Mouse RefSNPs	6 348 628
Mouse nucleotide sequences integrated into the MGI system (includes ESTs)	>8 400 000

and PIR protein super family classifications. The staff of MGD collaborates with members of other large genome informatics resources to maintain a comprehensive catalog of mouse genes and genome features, and to resolve inconsistencies in the representation of mouse genome features as needed. MGD is the authoritative source for mouse gene, allele and strain nomenclature and GO annotations for mouse gene function. MGD contains the most comprehensive source of mouse phenotype information and associations between human diseases and mouse models.

Researchers can query MGD using simple keywords, vocabulary browsers and web-based query forms. Keywords can include any free text including gene symbols, anatomical terms, strain names, phenotypes and disease terms, etc. MGD also provides several vocabulary browsers to support browsing of the database content using controlled vocabulary terms. For example, MGD's Human Disease Vocabulary Browser supports access to mouse genotype information that has been cross-referenced to human disease terms in OMIM. Finally, the MGD web-based query forms allow users to formulate queries of differing degrees of specificity. For example, using the Genes and Markers Query form in MGD, one can query for a list of all genes on mouse Chromosome 1. The Genes and Markers query form can also support more complex, biologically relevant queries that leverage the data integration in MGD. The query above, for example, could be refined to return a list of genes on mouse Chromosome 1 where the genes are associated with eye dysmorphology and have been annotated as transcription factors.

Data in MGD are updated daily. Data access is accomplished via dynamically generated web pages, text files available via FTP (updated nightly) and through direct SQL (Structured Query Language; user account is required). In general, there are 4–6 major software releases per year to support access and display of new data types. A recent summary of MGD content is shown in Table 1.

IMPROVEMENTS DURING 2007

Inclusion of gene trap data

MGD now includes the association of gene trap mutant cell IDs with mouse genes. The data for the gene trap are obtained on a regular basis from the dbGSS division of GenBank. The data in dbGSS include sequences associated with both exon and gene traps. The mouse data in dbGSS are expected to grow markedly as a consequence of several initiatives that have begun to generate a knockout allele for every mouse gene (8). Gene trap associations for a given gene or marker are included in the 'Other Database Links' section of the gene detail report. Currently in MGD, 10 708 mouse genes are associated with at least one gene trap. There are over 124 000 genomic survey sequences from NCBI's dbGSS database that cannot be unambiguously associated with mouse genes.

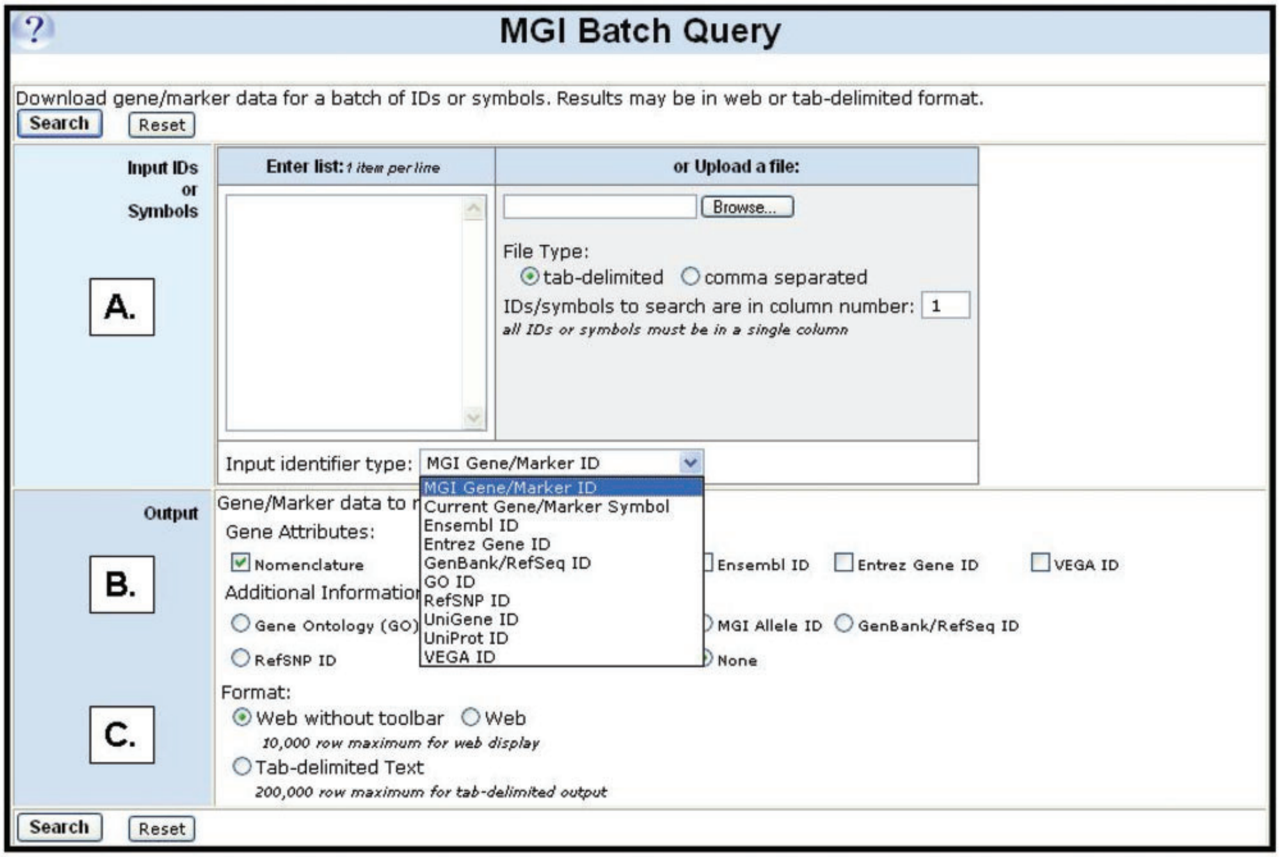
Queries for specific gene traps can be accomplished by querying MGD using the dbGSS sequence accession identifiers. In addition, tab-delimited reports of gene trap in MGD can be viewed or downloaded from the MGI FTP site. The FTP reports include 'gene traps associated with markers' and 'gene traps not associated with markers'. The report of gene trap sequences that cannot be associated with a mouse gene includes a brief notation describing why the sequence cannot be associated to a gene. (e.g. no match to the reference genome sequence, multiple good matches to the reference genome sequence, etc.)

Batch query tool

The new MGI Batch Query Tool (http://www.informatics.jax.org/javawi2/servlet/WIFetch?page=batchQF) provides the ability to access information about nomenclature, genome location, function or phenotype associations for many genes/markers in a single query (Figure 1). The allowable input into the Batch Query Tool includes current gene symbols, Ensembl gene ids, EntrezGene ids, VEGA gene ids, MGI ids, RefSeq ids and GenBank sequence accession ids. These data can be uploaded as a file or pasted into a text box on the query form. Users can specify the desired output and output format (web or tab-delimited text). The Batch Query Tool is particularly useful for researchers who use non-MGI gene accession identifiers in their analyses but who want to connect those identifiers to the rich functional and phenotypic annotations for mouse genes contained in MGD.

Links to TreeFam and other comparative resources

The comparative data resources in MGD now includes links to the TreeFam (9) resource and the availability of graphical displays of mammalian genes organized by the OrthoDisease (10) sets (Figure 2). TreeFam provides curated information about ortholog and paralog assignments and the evolutionary history of various gene families. Hypertext links to TreeFam are from the Genes and Markers or the Mammalian Orthology detail pages in MGD.



MGI Batch Query

Download gene/marker data for a batch of IDs or symbols. Results may be in web or tab-delimited format.

Input IDs or Symbols

Enter list: *1 item per line*

or Upload a file:

File Type:
● tab-delimited ○ comma separated

IDs/symbols to search are in column number: 1

all IDs or symbols must be in a single column

Input identifier type: MGI Gene/Marker ID

Output

Gene/Marker data to r

Gene Attributes:
☑ Nomenclature ☐ Ensembl ID ☐ Entrez Gene ID ☐ VEGA ID

Additional Informatio
○ Gene Ontology (GO) ○ MGI Allele ID ○ GenBank/RefSeq ID
○ RefSNP ID ○ None

Format:
● Web without toolbar ○ Web
10,000 row maximum for web display
○ Tab-delimited Text
200,000 row maximum for tab-delimited output

Figure 1. Screenshot showing the new MGI Batch Query Tool. Inputs into the query form **(A)** can be lists of sequence or gene identifiers. The output of the query **(B)** can be gene identifiers from other resources, genome location, gene nomenclature, functional annotations and phenotype annotations. Output from the batch query form can be displayed as a web form or as a tab-delimited file **(C)**.

The OrthoDisease orthology resource (http://orthodisease.cgb.ki.se/) provides eukaryotic orthology sets based on InParanoid analysis for genes in 26 organisms (11). These orthology sets are organized in relationship to 3409 diseases as represented in OMIM (12).

OTHER INFORMATION

Mouse Gene, allele and strain nomenclature

MGD is responsible for assigning unique symbols and names to mouse genes, alleles and strains following the guidelines set by the 'International Committee on Standardized Genetic Nomenclature for Mice' (http://www.informatics.jax.org/nomen). This official nomenclature is widely disseminated through regular data exchange and curation of shared links between MGI and other bioinformatics resources. MGD staff works with editors of journal publications to promote adherence to mouse nomenclature standards in publications.

The MGD nomenclature group works closely with nomenclature specialists for human (http://www.genenames.org/) and rat (http://rgd.mcw.edu) to provide consistent nomenclature for mammalian species. The mouse and human nomenclature committees collaborate with scientific experts in specific domain areas to represent the latest knowledge about gene families such as the alpha-tubulin family (13) or the RCAN gene family (14). The MGD nomenclature coordinator can be contacted by email (nomen@informatics.jax.org).

Electronic data submission

MGD accepts contributed data sets for any type of data maintained by the database. The most frequent types of contributed data are mutant and phenotypic allele information originating with the large mouse mutagenesis centers and repositories that contribute to the International Mouse Strain Resource (IMSR, http://www.imsr.org) (9). Each electronic submission receives a permanent database accession ID. All data sets are associated with their source, either a publication or an electronic submission reference. Details about data submission procedures can be found at: http://www.informatics.jax.org/mgihome/submissions/submissions_menu.shtml.

Community outreach and user support

MGD User Support can be accessed through online documentation and easy email or phone access to User Support Staff.

- World wide web: http://www.informatics.jax.org/mgihome/support/support.shtml

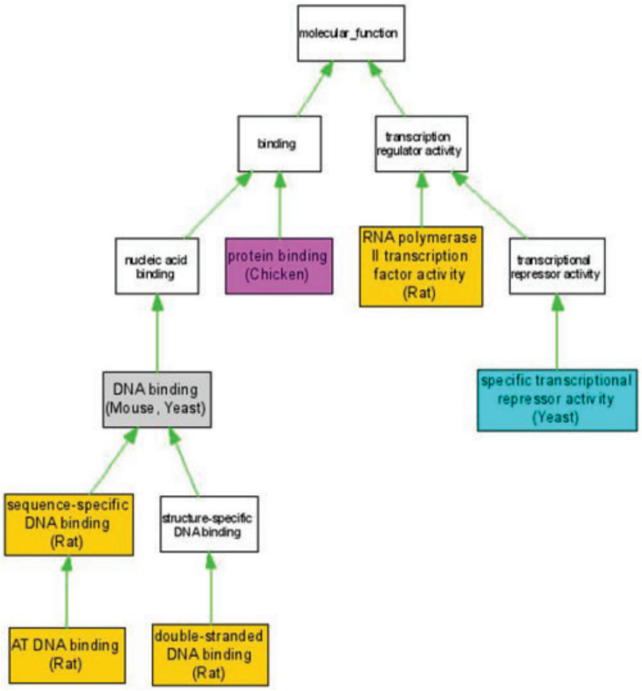

molecular_function
binding
transcription regulator activity
nucleic acid binding
protein binding (Chicken)
RNA polymerase II transcription factor activity (Rat)
transcriptional repressor activity
DNA binding (Mouse, Yeast)
specific transcriptional repressor activity (Yeast)
sequence-specific DNA binding (Rat)
structure-specific DNA binding
AT DNA binding (Rat)
double-stranded DNA binding (Rat)

Figure 2. Screenshot showing the Gene Ontology Molecular Function annotation graph for a gene associated with OMIM disorder 'Aniridia, type II' (OMIM id 106210). The graph displays experimental GO annotations for the human gene (PAX6) associated with this disorder as well as annotations for orthologous genes in other organisms (mouse, rat, nematode, chicken and yeast) based on the OrthoDisease set. The graph nodes are color-coded to indicate the organism that is the source of the annotation. The full graph and table of annotations can be viewed at: http://proto.informatics.jax.org/prototypes/GOgraphEX/OrthoDisease_Graphs/OMIM_DisorderGraphs/106210.html

- Email access: mgi-help@informatics.jax.org
- Telephone access: 1 207 288 6445
- FAX access: 1 207 288 6132

MGD User Support staff are also available for on-site training on the use of MGD and other MGI data resources. The traveling tutorial program includes lectures, demos and hands-on tutorials that can be customized according to the research interests of the audience.

Other Outreach. MGI-LIST (http://www.informatics.jax.org/mgihome/lists/lists.shtml) is a moderated and active email bulletin board supported by the MGD User Support group.

HIGH-LEVEL OVERVIEW OF THE MAIN COMPONENTS AND IMPLEMENTATION

MGD is implemented in the Sybase relational database management system with ~180 tables within which the biological information is stored. BLAST-able databases and genome assembly files for sequence data are stored outside the relational database. An editing interface and automated load programs are used to input data into the MGD system. The editing interface (EI) is an interactive, graphical application used by curators. Automated load programs that integrate larger data sets from many sources into the database include quality control (QC) checks and processing algorithms that integrate the bulk of the data automatically and identify issues to be resolved by curators or the data provider. Thus, through EI and automated loads, we acquire and integrate large amounts of data into a high-quality, knowledgebase.

Public data access is provided through the web interface (WI) where users can interactively query and download our data through a web browser. MouseBLAST allows users to do sequence similarity searches against a variety of rodent sequence databases that are updated weekly from selected sequence databases from NCBI, UniProt and other providers. Mouse GBrowse allows users to visualize mouse data sets against the genome as a series of linear tracks. FTP reports are a major source for other data providers who link to or use MGD data in their products, and for computational biologists who use MGD data in their analyses. Programmatic access to MGD via web services is under development. All MGD files and programs are openly and freely available.

CITING MGD

For a general citation of the Mouse Genome Informatics (MGI) resource please cite this article. In addition, the following citation format is suggested when referring to data sets specific to the MGD component of MGI: 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/NHGRI grant HG000330. Funding to pay the Open Access publication charges for this article was provided by HG000330.

Conflict of interest statement. None declared.

REFERENCES

1. Eppig,J.T., Blake,J.A., Bult,C.J., Kadin,J.A., Richardson,J.E. and the Mouse Genome Informatics group. (2007) The Mouse Genome Database (MGD): new features facilitating a model system. *Nucleic Acids Res.*, **35**, 637.
2. Blake,J.A., Eppig,J.T., Bult,C.J., Kadin,J.A., Richardson,J.E. and Mouse Genome Database Group. (2006) The Mouse Genome Database (MGD): updates and enhancements. *Nucleic Acids Res.*, **34**, D562–D567.
3. Eppig,J.T., Bult,C.J., Kadin,J.A., Richardson,J.E., Blake,J.A. and The Mouse Genome Database Group. (2005) The Mouse Genome Database (MGD): from genes to mice – a community resource for mouse biology. *Nucleic Acids Res.*, **33**, D471–D475.
4. 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. *et al*. (2004) The Mouse Gene Expression Database (GXD): updates and enhancements. *Nucleic Acids Res.*, **32**, D568–D571.Available at http://www.informatics.jax.org/mgihome/GXD/aboutGXD.shtml

5. Krupke,D.M., Naf,D., Vincent,M.J., Allio,T., Mikaelian,I., Sundberg,J.P., Bult,C.J. and Eppig,J.T. (2005) The Mouse Tumor Biology Database: integrated access to mouse cancer biology data. *Exp. Lung Res.*, **31**, 259–270. Available at http://tumor.informatics.jax.org
6. The Gene Ontology Consortium. (2008) The Gene Ontology (GO) project in 2008. *Nucleic Acids Res.*, **36**, in press. http://www.geneontology.org
7. Smith,C.L., Goldsmith,C.A. and Eppit,J.T. (2005) The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.*, **6**, R7.
8. The International Mouse Knockout Consortium. (2007) A mouse for all reasons. *Cell*, **128**, 9–13. Available at http://www.nih.gov/science/models/mouse/knockout/
9. Li,H., Coghlan,A., Ruan,J., Coin,L.J., Heriche,J.K., Osmotherly,L., Li,R., Liu,T., Zhang,Z. *et al.* (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.*, 34, D572–D580. Available at http://www.treefam.org/
10. Dolan,M.E. and Blake,J.A. (2006). Using ontology visualization to coordinate cross-species functional annotation for human disease genes. Computer-based medical systems. *Proceedings of the Nineteenth IEEE Symposium on Computer-Based Medical Systems (CBMS'06): Ontologies for Biomedical Systems*. Salt Lake City, Utah, pp. 583–587.
11. O'Brien,K.P, Westerlund,I. and Sonnhammer,E.L. (2004) OrthoDisease: a database of human disease orthologs. *Hum. Mutat.*, **24**, 112–119.
12. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**(Database issue), D514–D517.
13. Khodiyar,V.K., Maltais,L.J., Sneddon,K.M.B., Smith,J.R., Shimoyama,M., Cabral,F., Dumontet,C., Dutcher,S.K., Harvey,R.J. *et al.* (2007) A revised nomenclature for the human and rodent alpha-tubulin gene family. *Genomics*, **90**, 285–289.
14. Davies,K.J.A., Ermak,G., Rothermel,B.A., Pritchard,M., Heitman,J., Ahnn,J., Henrique-Silva,F., Crawford,D., Canaider,S. *et al.* (2007) Renaming the *DSCR1/Adapt78* gene family as *RCAN*:regulators of calcineurin. *FASEB J.*, **21**, doi:10.1096/fj.06-7246com.