D586–D592 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 27 October 2009*
doi:10.1093/nar/gkp880

The Mouse Genome Database: enhancements and updates

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

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

Received September 15, 2009; Accepted October 1, 2009

ABSTRACT

The Mouse Genome Database (MGD) is a major component of the Mouse Genome Informatics (MGI, http://www.informatics.jax.org/) database resource and serves as the primary community model organism database for the laboratory mouse. MGD is the authoritative source for mouse gene, allele and strain nomenclature and for phenotype and functional annotations of mouse genes. MGD contains comprehensive data and information related to mouse 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 for MGD are obtained from diverse sources including manual curation of the biomedical literature and 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 and use of biomedical ontologies such as the Gene Ontology and the Mammalian Phenotype Ontology. Recent improvements in MGD described here includes integration of mouse gene trap allele and sequence data, integration of gene targeting information from the International Knockout Mouse Consortium, deployment of an MGI Biomart, and enhancements to our 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 central component of the Mouse Genome Informatics (MGI) database resource (http://www.informatics.jax.org), the community model organism database for the laboratory mouse. Other MGI data resources integrated with MGD includes the Gene Expression Database (GXD) (4), the Mouse Tumor Biology Database (MTB) (5), the Gene Ontology (GO) project (6) and the MouseCyc database of biochemical pathways (7). Data in MGD are updated daily. There are typically four to six major software releases per year to support access and display of new data types.

The primary data types maintained in MGD include mouse genes and other genome features along with their function and phenotype annotations, associations of genome features with nucleotide and protein sequences, genetic and physical maps, gene families, mutant phenotypes, SNPs and other polymorphisms animal models of human disease, and mammalian homology. A recent summary of MGD content is shown in Table 1.

MGD is the authoritative source for mouse gene, allele and strain nomenclature, Gene Ontology annotations for mouse gene function, and Mammalian Phenotype (MP) Ontology (8) annotations for phenotype associations. MGD contains the most comprehensive source of mouse phenotype information and associations between human diseases and mouse models. MGI curatorial staff acquire data by direct data loads from other databases, from direct submission from researchers and from published literature. To facilitate data integration, MGI employs recognized standards for genetic nomenclature and functional annotation to describe mouse sequence data, genes,

*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, A. D. Diehl, M. E. Dolan, K. L. Forthofer, K. S. Frazer, P. Frost, D. E. Geel, M. Hall, M. Knowlton, J. R. Lewis, L. J. Maltais, M. McAndrews-Hill, S. McClatchy, M. J. McCrossin, J. Mason, T. F. Meehan, D. B. Miers, L. A. Miller, L. Ni, H. Onda, J. E. Ormsby, D. J. Reed, B. Richards-Smith, D. R. Shaw, R. Sinclair, D. Sitnikov, C. L. Smith, P. Szauter, M. Tomczuk, L. L. Washburn, I. T. Witham, Y. Zhu.

© The Author(s) 2009. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Table 1. Summary of MGD data content (10 September 2009)

MGD data statistics	10 September 2009
Genes with nucleotide sequence data	28 891
Genes with protein sequence data	26 255
Genes (including uncloned mutations)	36 323
Genes with GO annotations	18 167
Mouse/human orthologs	17 787
Mouse/rat orthologs	16 768
Genes with one or more mutant alleles[a]	17 227
Genes with one or more phenotypic alleles[b]	8363
Total mutant alleles[a]	524 527
Phenotypic alleles[b]	22 666
Targeted alleles	13 721
Gene trapped alleles	501 232
Human diseases with one or more mouse models	964
QTLs	4248
Number of references	146 597
Mouse RefSNPs	10 089 692

[a]Mutant alleles include those occurring in mice and/or in ES cell lines.
[b]Phenotypic alleles include only those mutant alleles present in mice.

strains, expression data, alleles and phenotypes. All data associations in MGD are supported with evidence and citations.

Researchers can query MGD using keyword searches, vocabulary browsers and advanced web-based query forms. Keyword search supports the use of the wildcard characters (i.e.*) for broad searches and the use of quotation marks for specific phrases search. MGD also provides vocabulary browsers for GO annotations, MP annotations and Human Disease Term annotations to support browsing of the database content. The web-based query forms in MGD allow, users to construct queries of differing degrees of specificity. For example, using the Genes and Markers Query form in MGD, a researcher query broadly for all genes on mouse Chromosome 3 or specifically for genes on Chromosome 3 that are associated with specific phenotypes and/or functions (i.e. show me all genes on mouse Chromosome 3 that are associated with respiratory distress and that have been annotated functionally as being enzymes). The MGI MouseBLAST server allows users to interrogate the MGI database using nucleotide and/or protein sequences. Access to data in MGD is also facilitated by summary data files that are updated nightlyand available for download via FTP, and through direct SQL (Structured Query Language; user account is required).

The staff of MGD collaborates with members of other large genome informatics resources including NCBI (http://www.informatics.jax.org), Ensembl (http://www.ensembl.org), UCSC Genome Browser (http://genome.ucsc.edu) and the Vertebrate Genome Annotation (Vega) group (http://vega.sanger.ac.uk/index.html), to maintain a comprehensive catalog of mouse genes and other genome features, and also to resolve inconsistencies in the representation of mouse genome features as needed. Biological annotations for mouse genes based on MGD curation are incorporated into scores of external informatics resources and software products.

NEW IN 2009

Completing the representation of Mouse Gene Traps

Release of 4.3 of MGD added over 500 000 mouse ES cell lines and sequences for gene traps from the NCBI Genome Survey Sequences Database (dbGSS), including those from the International Gene Trap Consortium (IGTC), and from Lexicon Genetics. Database records for gene-trap alleles in MGD now include the following information:

(i) sequence tag information, including genome coordinates,
(ii) cell line IDs from organizations supplying gene trap sequence data to dbGSS,
(iii) information about the parent stem cell line and the gene trap vector used to produce each mutant,
(iv) information about whether a mouse has been produced from the gene trap mutant cell line,
(v) phenotypic data for specific mouse genotypes carrying this gene trap allele,
(vi) identification of any mouse models with phenotypic similarity to human diseases associated with the allele,
(vii) links to the International Mouse Strain Resource (IMSR) for access to available mouse strains and cell lines, and
(viii) official allele nomenclature.

In addition to the rich annotation details for gene trap alleles (Figure 1), the location and structure of the gene traps in a genomic context are available from mouse GBrowse (Figure 2). GBrowse contains separate tracks for DNA and RNA-based gene traps. In addition, there is a summary track which displays the number of traps per gene. Since GBrowse includes the gene predictions from NCBI, Vega and Ensembl in individual tracks, it is straightforward to compare the location of the gene traps relative to multiple gene predictions.

Gene trap data are easily accessed from gene detail pages via hypertext links in the Phenotypes section of the report. Direct queries for gene traps in MGD can be accomplished using the dbGSS sequence accession identifiers or by searching for specific parameters on the Phenotypes, Alleles and Disease Query form. Tab-delimited reports of gene traps in MGD can be viewed or downloaded from the MGI FTP site.

Incorporation of International Knockout Mouse Consortium data

The International Knockout Mouse Consortium (IKMC) is a broad based international effort to generate knockout alleles for every mouse gene (9,10). As IKMC generates ES cell lines carrying new targeted mutant alleles, these are incorporated into MGD and provide official nomenclature and MGI identifiers. Thus, IKMC alleles are accessible with all other mouse mutant alleles. As IKMC mutant ES cell lines are used to produce mice and those mice are phenotyped, data will be available in MGD for comparative phenotyping with all other extant mouse mutant data.

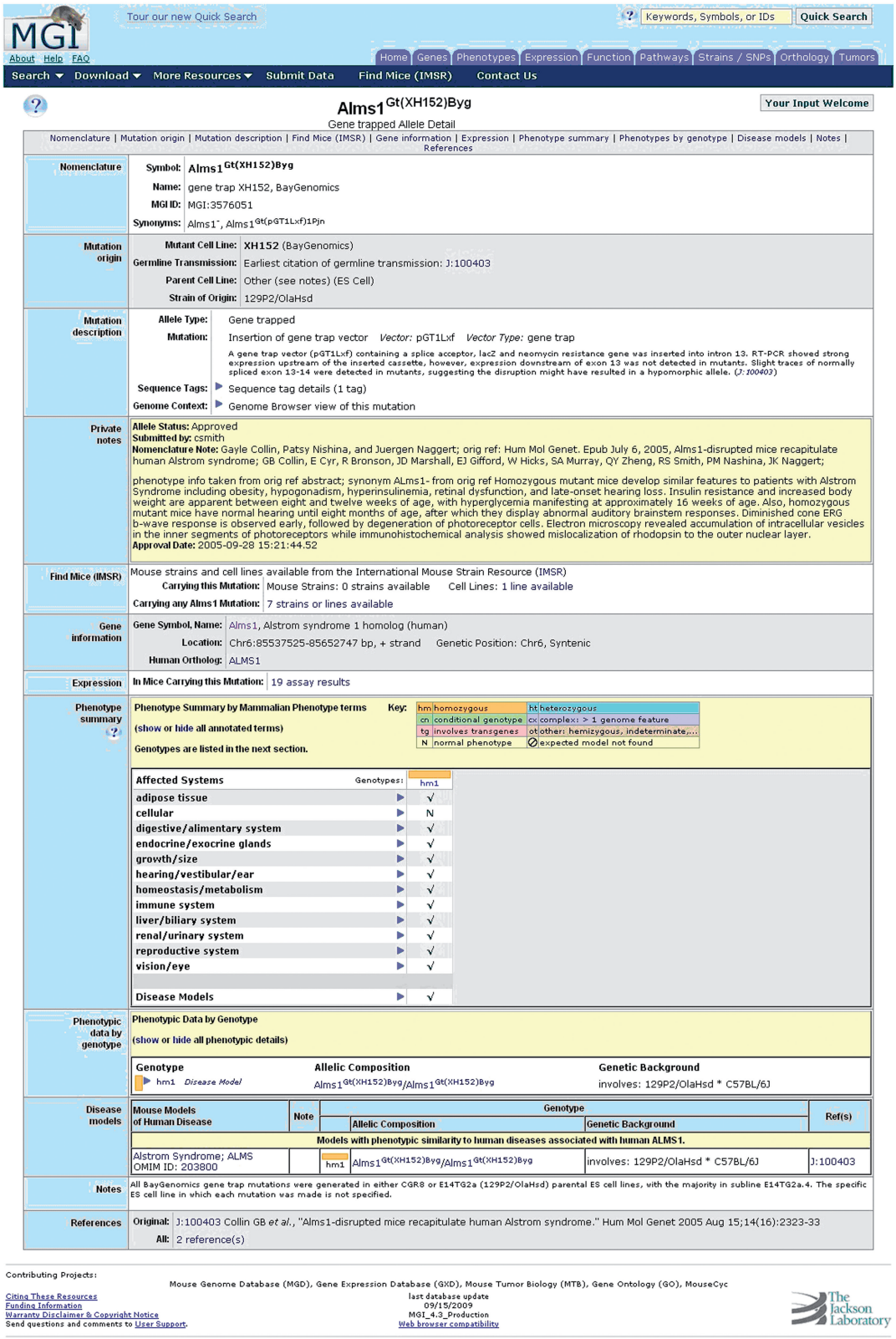

Figure 1. Screen shot demonstrating the new gene trap allele detail page for a BayGenomics gene trap in the *Alms1* gene.

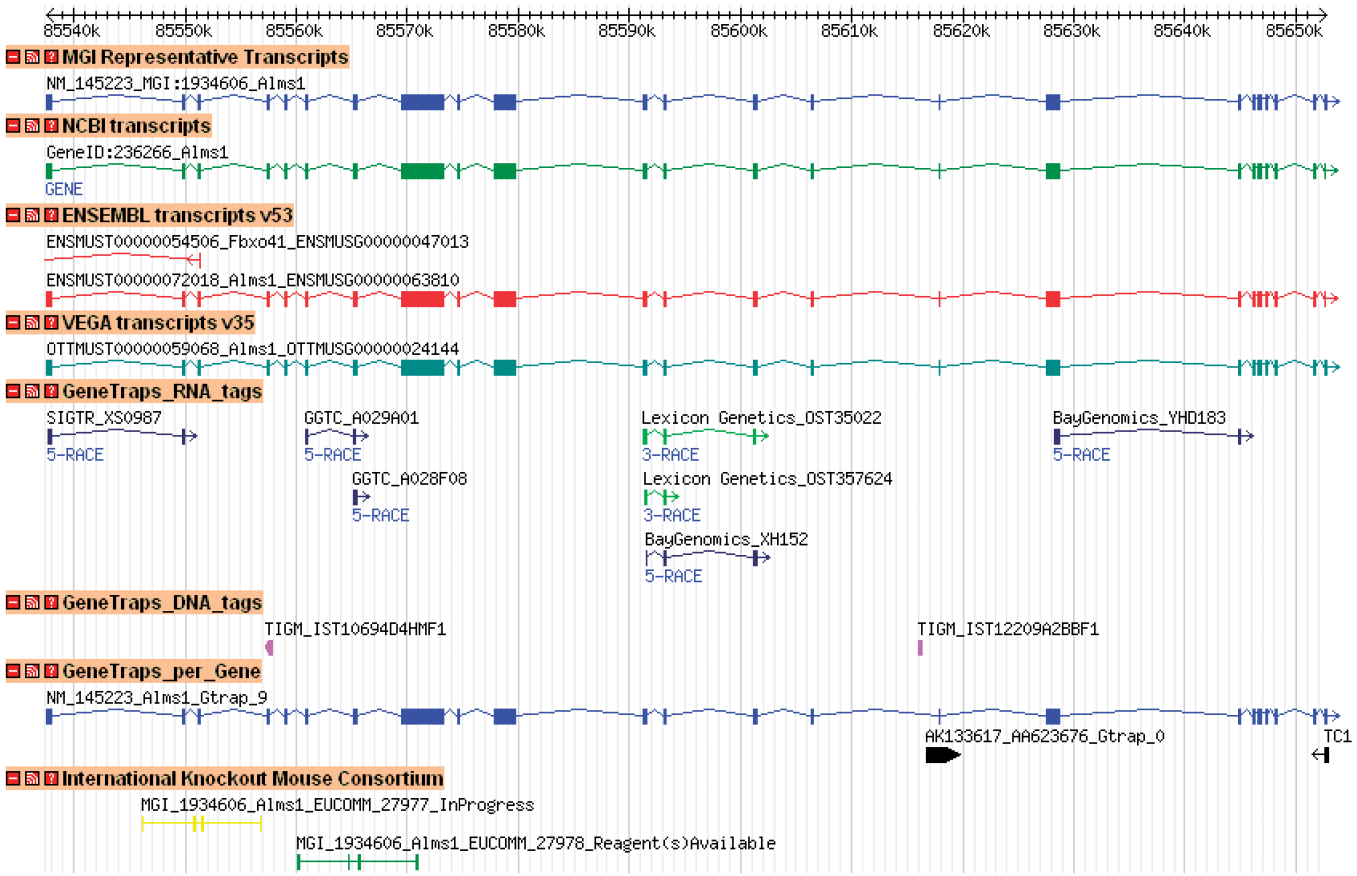

Figure 2. Screen shot of MGI GBrowse showing gene traps and gene targeting projects from the IKMC. Figure shows mouse chromosome 6 region 85577061-85692283 (NCBI Build 37).

Primary access to IKMC progress and resources is available through a common web portal (http://www.knockoutmouse.org). To facilitate access to IKMC information and resources from within MGI, curated links to the IKMC web site are now available from MGI gene detail pages and also from tracks in mouse GBrowse (Figure 2).

Enhanced Batch Query Tool capabilities

The Batch Query Tool is particularly useful for researchers who use non-MGI mouse 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. The initial release of the MGI Batch Query Tool (http://www.informatics.jax.org/javawi2/servlet/WIFetch?page=batchQF) provided the ability to access information about nomenclature, genome location, function, or phenotype associations for many genes/markers in a single query (2). Allowable input into the Batch Query Tool included 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 specified the desired output and output format (web or tab-delimited text). Recent additions to the Batch Query Tool include the ability to use Affymetrix microarray probe identifiers as inputs into the query tool and the ability to download phenotype and functional annotation terms, gene expression data from the Gene Expression Database (GXD), and human disease terms that are associated with the user supplied id lists. In addition, the Batch Query Tool will now accept mixed lists of identifiers as input, for example. MGI:96677, Pax6, 16590, OTTMUSG00000015949, Q3UFR6, which are, respectively, an MGI accession identifier, a gene symbol, an Entrez Gene identifier, a VEGA mouse gene identifier, and a UniProt protein record accession identifier.

MGI BioMart

To support cross database integration and data mining, MGI now supports a BioMart application (Figure 3). BioMart is a 'query-oriented data management system' that is designed to support a federated approach to data integration (11). The unique aspect of the MGI BioMart query tool relative to existing data access mechanisms for MGI is that the resources supports the ability of users to combine data and annotations from MGI with data from external databases, such as the gene annotation

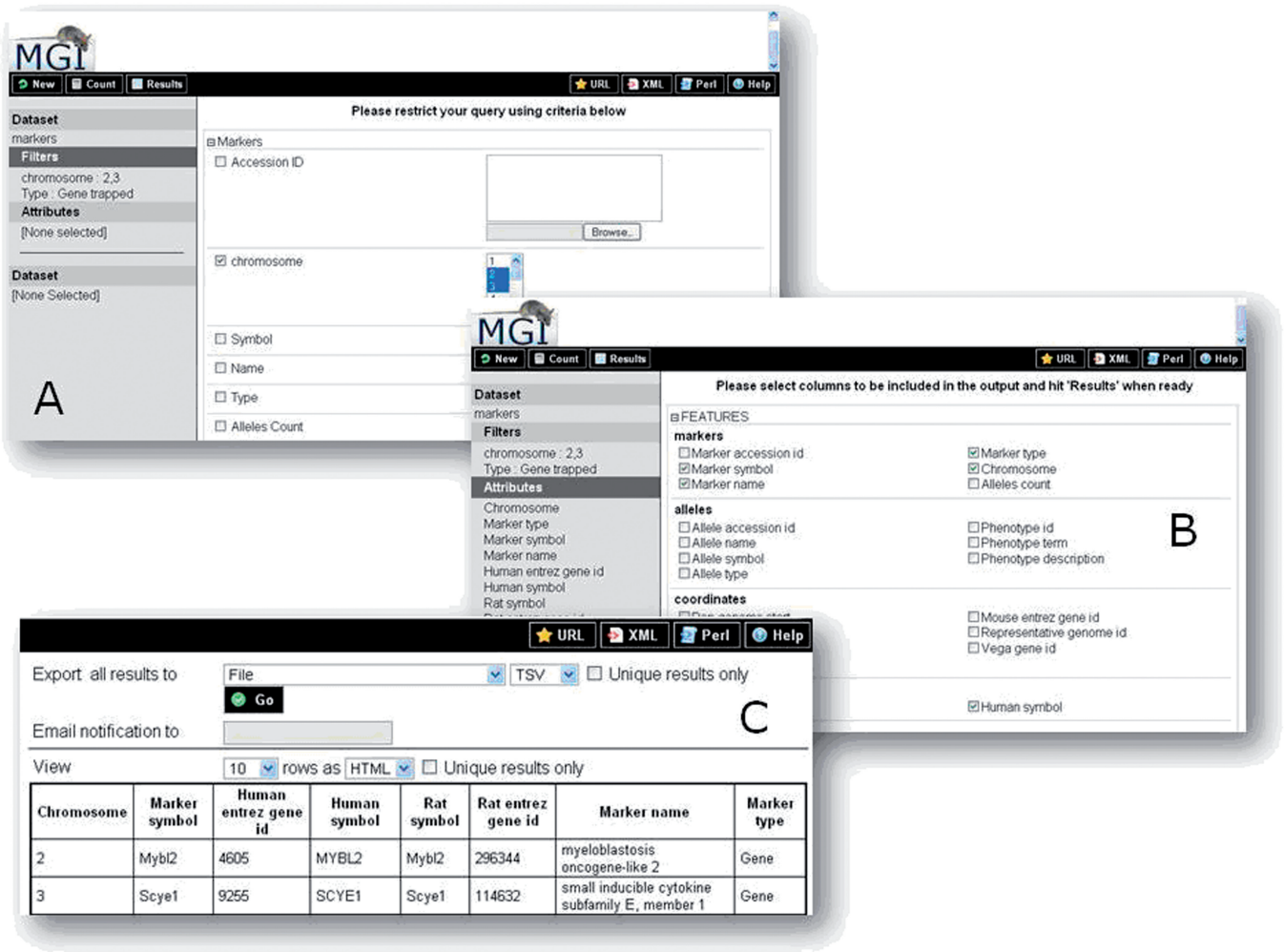

Figure 3. Screen shots of the MGI BioMart. To create a data set in BioMart users select the database of interest (**A**), select the attributes they wish to include in their results (**B**), and save the results in one of several possible format (**C**). Sets of results can be refined iteratively and can be combined with data from external BioMarts.

data from Ensembl (http://www.ensembl.org), 'on the fly'. The BioMart also supports iterative query refinement and allows users to save query results in a variety of output formats.

OTHER INFORMATION

Mouse gene, allele and strain nomenclature

MGD is the authoritative source of symbols and names for mouse genes, alleles and strains. The nomenclature in MGD follows 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 members work with editors of journal publications to promote adherence to mouse nomenclature standards in publications.

To support consistency of nomenclature across multiple mammalian species, members of the MGD nomenclature group coordinate gene names and symbols with nomenclature specialists from the Human Gene Nomenclature Committee (HGNC) (http://www.genenames.org/) and the rat genome database (RGD; http://rgd.mcw.edu). 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 NLR gene family (12). The MGD nomenclature coordinator can be contacted by email (nomen@informatics.jax.org).

Electronic data submission

MGD accepts contributed data sets from individuals and organizations 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 (13)]. 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.

Suggestions and corrections to the representation of data and information in MGD can be submitted using the 'Your Input Welcome' link which appears in the upper right hand corner of gene and allele detail pages.

Community outreach and user support

The MGD resource has full-time staff members who are dedicated to user support and training. Members of the User Support team can be contacted via email, web requests, phone or FAX.

(i) World wide web: http://www.informatics.jax.org/mgihome/support/support.shtml
(ii) Email access: mgi-help@informatics.jax.org
(iii) Telephone access: +1-207-288-6445
(iv) FAX access: +1-207-288-6132

MGD User Support staff are 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.

Online training materials for MGD and other MGI data resources are available as FAQs and on-demand help documents. In addition, a freely available Mouse Genome Informatics tutorial is available via Open Helix (http://www.openhelix.com/mgi).

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. The MGI list serve has over 2100 subscribers. On an average there are three posts per day.

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 to MGD is provided primarily 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 (SOAP) is also supported (http://www.informatics.jax.org/mgihome/other/web_service.shtml). All MGD files and programs are openly and freely available.

CITING MGD

For a general citation of the MGI resource please cite this article. In addition, the following citation format is suggested when referring to datasets 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 retrieve the data cited.]

FUNDING

National Institutes of Health National Human Genome Research Institute grant HG000330. Funding for open access charge: National Institutes of Health grant HG000330.

Conflict of interest statement. None declared.

REFERENCES

1. Blake,J.A., Bult,C.J., Eppig,J.T., Kadin,J.A., Richardson,J.E. and the Mouse Genome Database Group. (2009) The Mouse Genome Database genotypes::phenotypes. *Nucleic Acids Res.*, **37**, D712–D719.
2. Bult,C.J., Eppig,J.T., Kadin,J.A., Richardson,J.E., Blake,J.A. and the Mouse Genome Database Group. (2008) The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res.*, **36**, D724–D728.
3. Eppig,J.T., Blake,J.A., Bult,C.J., Kadin,J.A., Richardson,J.E. and the Mouse Genome Database Group. (2007) The Mouse Genome Database (MGD): new features facilitating a model system. *Nucleic Acids Res.*, **35**, D630–D637.
4. Smith,C.M., Finger,J.H., Hayamizu,T.F., McCright,I.J., Eppig,J.T., Kadin,J.A., Richardson,J.E. and Ringwald,M. (2007) The mouse Gene Expression Database (GXD): 2007 update. *Nucleic Acids Res.*, **35**, D618–D623.
5. Krupke,D.M., Begley,D.A., Sundberg,J.P., Bult,C.J. and Eppig,J.T. (2008) The Mouse Tumor Biology database. *Nat. Rev. Cancer*, **8**, 459–465.
6. The Gene Ontology Consortium. (2008) The Gene Ontology (GO) project in 2008. *Nucleic Acids Res.*, **36**, D440–D444.
7. Evsikov,A., Dolan,M., Genrich,M.J., Patek,E. and Bult,C.J. (2009) MouseCyc: a curated biochemical pathways database for the laboratory mouse. *Genome Biol.*, **10**, R84.

8. Smith,C.L., Goldsmith,C.A. and Eppig,J.T. (2005) The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.*, **6**, R7.
9. The International Mouse Knockout Consortium. (2007) A mouse for all reasons. *Cell*, **128**, 9–13.
10. Collins,F.S., Finnell,H., Rossant,J. and Wurst,W. (2007) A new partner for the international knockout mouse consortium. *Cell*, **129**, 235.
11. Smedley,D., Haider,S., Ballester,B., Holland,R., London,D., Thorisson,G. and Kasprzyk,A. (2009) BioMart – biological queries made easy. *BMC Genomics*, **10**, 22.
12. Ting,J.P., Lovering,R.C., Alnemri,E.S., Bertin,J., Boss,J.M., Davis,B.K., Flavell,R.A., Girardin,S.E., Godzik,A., Harton,J.A. *et al*. (2008) The NLR gene family: a standard nomenclature. *Immunity*, **28**, 285–287.
13. Strivens,M. and Eppig,J.T. (2004) Visualizing the laboratory mouse: capturing phenotype information. *Genetica*, **122**, 89–97.