REVIEW

# Mouse Genome Database: From Sequence to Phenotypes and Disease Models

**Janan T. Eppig,* Joel E. Richardson, James A. Kadin, Cynthia L. Smith, Judith A. Blake, and Carol J. Bult, the MGD Team**

The Jackson Laboratory, Bar Harbor, Maine

Received 6 May 2015; Revised 30 June 2015; Accepted 2 July 2015

**Summary: The Mouse Genome Database (MGD, www.informatics.jax.org) is the international scientific database for genetic, genomic, and biological data on the laboratory mouse to support the research requirements of the biomedical community. To accomplish this goal, MGD provides broad data coverage, serves as the authoritative standard for mouse nomenclature for genes, mutants, and strains, and curates and integrates many types of data from literature and electronic sources. Among the key data sets MGD supports are: the complete catalog of mouse genes and genome features, comparative homology data for mouse and vertebrate genes, the authoritative set of Gene Ontology (GO) annotations for mouse gene functions, a comprehensive catalog of mouse mutations and their phenotypes, and a curated compendium of mouse models of human diseases. Here, we describe the data acquisition process, specifics about MGD's key data areas, methods to access and query MGD data, and outreach and user help facilities. genesis 53:458–473, 2015. © 2015 The Authors. Genesis Published by Wiley Periodicals, Inc.**

**Key words:** mouse; database; genomics; gene function; human disease models

## INTRODUCTION

The laboratory mouse is widely recognized as an important animal model for investigating genetic and cellular systems relevant to human biology and disease. It has the advantages of being a small mammal in which all life stages can be accessed, for which a completely sequenced, well annotated reference genome is publicly available, and for which there are myriad genomic tools available for comparative and experimental manipulation. These include thousands of unique inbred strains listed in the International Mouse Strain Resource (www.findmice.org), which catalogs resources available in pubic repositories (currently over 32,000 strains and more than 200,000 embyronic stem (ES) cell lines).

In addition to the mouse reference genome sequence from the C57BL/6J inbred strain, 17 additional strains have been deeply sequenced (Keane *et al.*, 2011), with sequence for an additional 11 inbred strains available, but not yet published (K. Hunter and T.M. Keane, see http://www.sanger.ac.uk/resources/mouse/genomes/). For many strains, extensive single nucleotide polymorphism (SNP) data exist (e.g., dbSNP, Sayers *et al.*, 2012; Center for Genome Dynamics SNP data http://cgd.jax.org/cgdsnpdb/; Sanger Mouse Genomes Project, http://www.sanger.ac.uk/resources/mouse/genomes/), as well as data on genome structural variation (Yalcin *et al.*, 2012).

A number of large-scale mouse mutagenesis programs have created ready access to pre-generated, preserved mutations on defined genetic backgrounds. These

Members of the MGD Team: Anna Anagnostopoulos, Richard M. Baldarelli, Jonathan S. Beal, Susan M. Bello, Joanne Berghout, Olin Blodgett, Nancy Butler, Jeffrey W. Campbell, Lori E. Corbani, Howard Dene, Harold J. Drabkin, Kim L. Forthofer, Peter Frost, Sharon C. Giannatto, Paul Hale, Lucie Hutchins, Michelle Knowlton, MeiYee Law, Jill R. Lewis, Monica McAndrews, David B. Miers, Howie Motenko, Li Ni, Hiroaki Onda, Janice E. Ormsby, Wendy Pitman, Jill M. Recla, Beverly Richards-Smith, David R. Shaw, Dmitry Sitnikov, Kevin R. Stone, Monika Tomcuk, Linda L. Washburn, Yunxia Zhu

*Correspondence to: Janan T. Eppig, The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA E-mail: janan.eppig@jax.org

Contract grant sponsor: National Human Genome Institute (NHGRI) of the NIH, Contract grant number: HG000330

Published online 27 July 2015 in
Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/dvg.22874

include the International Gene Trap Consortium, producing ≈190,000 gene-trapped ES cell lines (Nord *et al.*,2006[1]), a number of *N*-ethyl-*N*-nitrosourea (ENU)-mutagenesis projects producing over 3,300 defined mutations (e.g., see Andrews *et al.*, 2012; Boles *et al.*, 2009; Clark *et al.*, 2004), and the International Knockout Mouse Consortium (IKMC) generating targeted mutations in all mouse protein-coding genes in ES cell lines, with over 14,000 genes successfully targeted to date (Bradley *et al.*, 2012). Genetic engineering methodologies including CRISPR/Cas (Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR associated sequences) engineering (Wang *et al.*, 2013) and elegant, technologies that allow spatiotemporal control of gene ablation and regulation are providing new insight into gene function through the use of conditional mutagenesis (Murray *et al.*, 2012). Other new resources for complex traits and QTL (Quantitative Trait Loci), including the Collaborative Cross (Collaborative Cross Consortium, 2012; Threadgill and Churchill, 2012) and Diversity Outbred mice (Churchill *et al.*, 2012; Svenson *et al.*, 2012) are beginning to bear fruit in analysis of complex and multigenic diseases (Chesler, 2014; Logan *et al.*, 2013; Philip *et al.*, 2011).

These exquisite genetic resources for mice help unravel the functional components of the mouse (and mammalian) genomes, at the level of the individual gene, complex interactions, and systemic effects (Bucan *et al.*, 2012). The large-scale efforts of the International Phenotyping Consortium (Brown and Moore, 2012), with a goal to systematically phenotype 20,000 mouse lines developed from the IKMC knockout ES cell line resource, are currently producing a global sweep of high throughput phenotyping data that add new breadth to the careful, detailed phenotyping done by hundreds of researchers worldwide and captured in the MGD.

The Mouse Genome Database (MGD) has, over the course of its 25+ year history, moved from linkage maps to sequence maps, from visual to molecular phenotypes, and from small data (single genes, biochemical variation) to genome size data (gene sets, comparative phenotypes, genome-wide variants), adapting to the changing scientific landscape and the improvements in computation and information technology. This article provides an overview of MGD. First, we describe the underlying requirements for data standards and integration that facilitate robust data retrieval and the process of data acquisition in MGD. This is followed by specifics about the breadth and depth of data content found in MGD and the multiple ways users can access and query these data and work with data sets, including through the web interface and customized searches. Next, a section is devoted to MGD's community interactions that supply documentation, online help, workshops, and ongoing communication to users; as well as receiving from users input on MGD's data, access and tools, and new research data submissions. Finally, we comment on the ongoing changes and challenges of maintaining and evolving MGD as an international community data resource for mouse genomics and biology.

## DATA INTEGRATION AND THE ROLE OF SEMANTIC AND DATA STANDARDS

Data integration is an underlying principle for MGD. To truly utilize the power of the mouse as a model system, disparate data must be brought into a common framework. The integration of data from many sources requires applying common semantic standards across those data. This process adds value to the data and creates a true knowledgebase, where new biological insights can emerge from novel data connections based on common, shared, identified objects (Fig. 1).

### Nomenclatures and Biological Ontologies

Human language is rich with descriptors and alternate ways of expressing the same or similar concepts. For scientific discourse, however, use of common terminology is essential to ensure that identical entities are clearly delineated. Unique identifiers (accession IDs) are used successfully in databases for this purpose, but generally human discussion relies upon words and symbols. This fact (along with vagaries of manuscript formats) has made precision in attempts at automating data extraction from the scientific literature via Natural Language Processing (NLP) difficult and further emphasizes the necessity for standard nomenclatures, ontologies, and the addition of consistent metadata to biological data sets and publications.

A simple example illustrates the value of using standard nomenclatures. Suppose an investigator is interested in finding information about the mouse gene *Tap*. The problem is that five different mouse genes have been published using this gene symbol. Fortunately for this investigator, the non-standard uses of this symbol have been captured by MGD as synonyms to the correct gene nomenclature and all of these genes will be returned from doing a search[2] for *Tap*. The user will still need to examine the gene alternatives to determine which is his/her actual gene of interest among those called *Tap* in the literature (in this case, each is on a unique chromosome, which might be enough of a clue). Two key points are: (1) that the investigator

[1] The Gene Trap Consortium is no longer generating new mutations, but ES cell lines from this effort are available to researchers. See the International Strain Resource (www.findmice.org).

[2] To perform this search, enter "Tap" in the Quick Search tool or the Genes and Marker Query form. Search results indicate Tap as a synonym in the right-hand column. See additional detail in later sections of this review on the Quick Search and Advanced Search Forms.

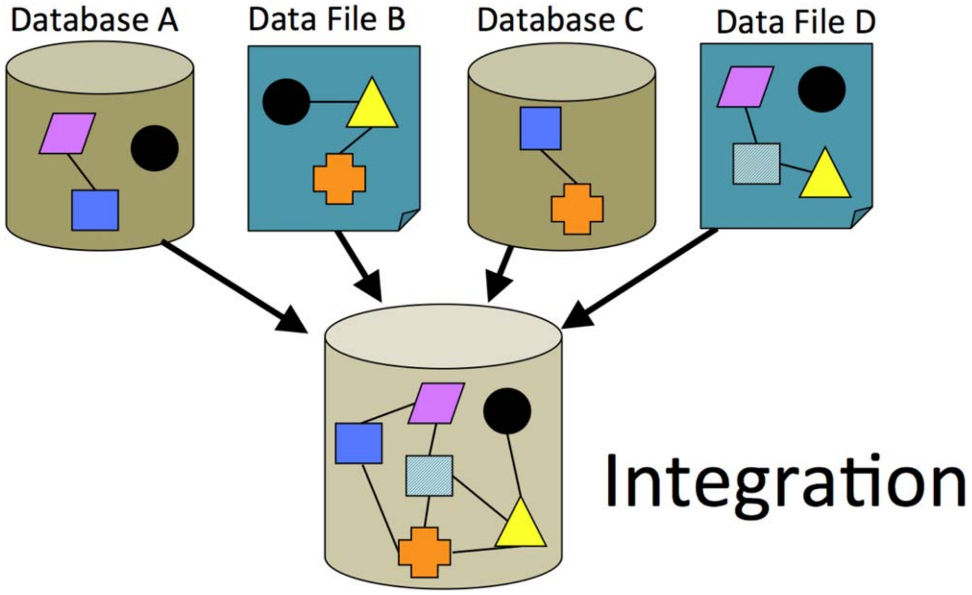


**FIG. 1.** Data integration. The process of data integration includes: (**a**) gathering data from various sources (other database resources, electronic files from online submissions or literature curation, etc.), (**b**) identifying common objects among the input, (**c**) resolving conflicts or inconsistent information or discovering missing data, and (**d**) assembling the relationships among the resolved data. Integration is key to knowledge discovery. For example, in this diagram the integrated data of the pink object is now seen to have a relationship to the orange object via its association with the dark blue square and light blue square, although the pink and orange objects were not coincident in the incoming data.

might have pursued a disastrous path by basing new experiments on data that was not about the gene he/she thought it was, or developed an erroneous view of *Tap* function that combined information about multiple genes had he/she not taken advantage of MGD's nomenclature curation; and (2) conversely, using non-standard nomenclature in publications may result in one's work not being recognized or misinterpreted. MGD is the authoritative source for nomenclatures for genes, genome features, mutations, alleles, structural variants, and mouse strains. Nomenclature guidelines are available at: www.informatics.jax.org/nomen and assistance with new nomenclature is available by emailing nomen@jax.org.

Biological ontologies go beyond naming objects. Ontologies are structured vocabularies and the placement of terms within that structure relay information about the relationships of terms to one another and the type of relationship involved. Within MGD, the Gene Ontology (Gene Ontology Consortium, 2000) and the Mammalian Phenotype Ontology (Smith *et al*., 2005) are used extensively for function and phenotype annotations, respectively. In addition, the EMAP/EMAPA (Edinburgh Mouse Atlas Project/Edinburgh Mouse Atlas Project Abstract version) ontology of mouse developmental anatomy (Finger *et al*., 2015; Hayamizu *et al*., 2013) is used to coordinate anatomical term use among the ontologies.

The Gene Ontology (GO) was initiated over 15 years ago by MGD, the *Saccharomyces* Genome Database (SGD), and *Drosophila* Database (Flybase) (GO Consortium, 2000) and is composed of three distinct ontologies, for molecular function, biological process, and cellular component. GO has been widely adopted and applied to gene and gene product annotations in hosts of organism databases and is commonly used for enrichment analyses of large data sets (Gene Ontology Consortium, 2015). In MGD, GO is used to annotate mouse gene and gene product functions.

MGD began developing the Mammalian Phenotype Ontology (MP) in 2003 as a replacement for textual descriptions of mouse mutants, which were becoming too numerous to properly maintain and effectively search. The MP Ontology has also been adopted by the Rat Genome Database (RGD, www.rgd.mcw.edu), the Medical Research Council, Harwell, UK, the Sanger Institute, UK, and a number of large-scale mutagenesis centers such as the ENU mutagenesis groups (e.g., see Ayadi *et al*., 2012; Goldowitz *et al*., 2004; Morgan *et al*., 2010), and the International Mouse Phenotyping Consortium (IMPC) that is phenotyping the targeted knockouts developed by the IKMC (Koscielny *et al*., 2014).

Use of standard semantics via nomenclatures and ontologies enables wide dissemination of scientific results and ready integration with other data applying these same standards. In the end, this standardization greatly benefits the community, enabling comprehensive searches in the data space and producing better returns with more complete results.

## DATA ACQUISITION: COMBINING AUTOMATION AND CURATION

The size and complexity of MGD data has grown tremendously since its inception, as has the ability of the scientific community to generate larger and larger data sets at a rapid pace. MGD has responded by continually improving its processes, automating data acquisition where feasible, developing semi-automated curation and data verification tools, and streamlining curation procedures. Data submissions directly from large-scale projects and individual laboratories are encouraged (see Community: Outreach and User Support section below).

Today, the bulk of MGD data are loaded and integrated automatically from over 50 external sources. Additional data are loaded from files prepared by MGD data analysts or are data relationships computed internally by MGD. In some cases, MGD loads are required to create new database objects and relationships. These loads, which occur weekly, include quality control checks that generate reports with any anomalies found. These reports are reviewed and anomalies resolved (either with the data provider or via curator analysis) as part of the curation and standardization process for MGD.

Critically important data continue to be published in peer-reviewed journals, with more detail and specificity than those produced in large-scale projects and of particular significance in validating large-scale data inferences. The key for MGD is to identify, select, and curate the set of publications providing the highest quality data for new biological understanding and providing the best new information to enable hypothesis generation and relevance for human health.

Scientific curators and analysts identify and prioritize newly published articles for inclusion in MGD based on novelty (e.g., new mutant allele created, new gene function described) and whether a model for human disease is included. Articles are tagged with the genes and mutant alleles described, ensuring that nomenclature is correct. In the last 5-year period, the average number of new publications indexed and/or curated in MGD is ∼12,000 each year.

Prioritization and full curation of articles follow different strategies, depending on article content. For example, for GO annotations, the highest priority is to assign gene or gene-product function to those genes that have no previously known function or function has only been inferred by sequence similarity to another species. In contrast, new disease models and new or extended phenotyping data are high priority for capturing phenotypic manifestations.

## KEY DATA CLASSES IN MGD

### The Unified Gene Catalog

MGD produces and maintains a current unified, non-redundant catalog of mouse genes and genome features, including, but not limited to, protein-coding genes, pseudogenes, non-coding RNA genes, quantitative trait loci (QTL), and heritable phenotypic markers. The gene catalog's foundation is the unified set of genome annotations resulting from the most recent assembly of the mouse reference sequence (C57BL/6J) as provided by NCBI (Sayers *et al.*, 2012), Ensembl (Flicek *et al.*, 2013 ), and Havana/Vega (Wilming *et al.*, 2008). We use the fjoin (feature join) algorithm developed by Richardson (2006) at MGD to determine the equivalency of predicted genes by examining genome coordinate overlap of exons. Where gene predictions are non-conflicting (i.e., either a genome feature in one data set overlaps only one feature in another data set or a gene prediction is unique to a data set) those features are loaded into MGD. Where complex overlap relationships are detected (i.e., a genome feature in one data set overlaps more than one feature in another data set (1:*n* or *n*:1); or multiple such overlaps are detected (*n*:*m*), quality control (QC) reports are generated and inspection of the data by a curator and communication with the data providers follows to reconcile the data. Updates are processed periodically as released by the annotation providers. Ongoing analysis of conflicting data is done in collaboration with the Mouse Genome Annotation group consisting of curators and analysts from MGD, Ensembl, NCBI, Havana/Vega, and the UCSC (University of California, Santa Cruz) browser team.

### Functional Annotations Using GO

MGD is one of the founding groups of the Gene Ontology Consortium (Gene Ontology Consortium, 2000). MGD is a major contributor to the ongoing development and refinement of the Ontology itself, in both its terminology base and structure. MGD is the authoritative source for GO functional annotations of mouse protein-coding and functional RNA genes. Mouse GO annotations are derived by curation of the scientific literature, through inferences based on orthology from experimental data in other organisms, notably human and rat (Reference Genome Group of the Gene Ontology Consortium, 2009), and through inferences based on protein structural domains (Mitchell *et al.*, 2015). MGD's mouse GO annotations are contributed to the GO Consortium Gene Ontology database, resulting in their wide distribution to other resources and use in large-scale functional analyses.

### Genetic Mutations and Variants

MGD maintains a comprehensive catalog of mouse mutations that have arisen spontaneously or have been induced by mutagens (e.g., chemical, radiation) or created through genetic engineering. For large mutagenesis projects, MGD develops project-specific loads which define the nomenclature that needs to be created, as well as a standard mutation creation statement utilizing appropriate parameter for each new mutant allele (e.g., genomic insertion site, transgene construct characteristics), strain on which the mutation is developed and/or bred, etc. Load processes include running scripts that generate QC reports from load log files to aid curators in identifying, communicating, and resolving data issues with the providers.

Despite significant large-scale mutation-generating projects, individual investigators continue to create many of their own new mouse mutations. This is no surprise, as large-scale projects tend to create new mutations in many genes, but all are of the same type and for in-depth studies for the discovery of gene function, genetic interactions, and strain-specific modifier or epigenetic effects, very specific mutations (or sets of mutations) are required. Thus, although MGD has added thousands of mutations from large-scale mutagenesis programs, 150–200 new mutant alleles are incorporated from individual laboratories each month through a combination of literature curation and direct online submission from investigators. (See User Input and Data Submission section below.)

## Phenotypes

MGD captures phenotypes from defined genotypes carried on various strain backgrounds using the Mammalian Phenotype (MP) Ontology (Smith and Eppig, 2012). Each genotype, described by a set of one or more mutant alleles in one or more genes, and carried in a specific strains background, can display an array of different phenotypes. For example, mutations in the gene *Kit* (kit oncogene), which has had considerable study, illustrate the large combinatory explosion that can occur with phenotypic data. *Kit* has 153 mutant alleles documented in MGD, 112 of which have been reported with one or more distinct “genome-type” (different allelic combinations and/or different strain backgrounds and/or a compound genotype with multiple mutant genes). An additional 41 alleles of *Kit* are gene trapped or from IKMC and only exist in ES cell lines. As of early May 2015, for all genes, MGD includes 54,909 distinct “genome-types” with 276,712 MP annotations, and these phenotype annotations include mutant alleles for 10,955 genes.

The Mammalian Phenotype (MP) Ontology is essential for the curation of phenotype data. It is a structured vocabulary whose initial development and continued maintenance is centered at MGD. Building and maintaining the MP Ontology is a dynamic process. New terms are added to the MP Ontology as curators require new terms (usually more granular terms) to annotate phenotypic studies. Suggestions for new terms, definitions, synonyms, and changes in ontology organization are tracked in SourceForge, (http://sourceforge.net/p/obo/mammalian-phenotype-requests/). In addition, work on particular branches of the MP Ontology has been spurred by the need for outside projects to use the MP Ontology, (e.g., we are working with the Cardiovascular Development Consortium (CvDC) funded by NHLBI’s Bench to Bassinet program which is, in part, creating new mouse ENU mutations for congenital heart defects, Li *et al*., 2015). New terms are being added to the MP Ontology to complete and harmonize congenital heart defect terms with those used in pediatric clinics. As of early May 2015, the MP Ontology contained more than 10,900 phenotypic terms.

The MP Ontology can support general and granular phenotype knowledge and allows robust searching and retrieval for web-based users as well as computational users. Each vocabulary term has an accession identifier, term name, synonyms, definition, and a reference for the definition. Terms are placed within a DAG (directed acyclic graph) vocabulary structure with relationships between parent and child nodes. Terms are organized from general to specific, allowing phenotypes to be annotated at the finest level of granularity possible to reflect current knowledge.

## Mouse Models for Human Disease

MGD curates the association between mutant mouse genotype/strain combinations (genome-type) and human diseases, where those mouse genotypes are experimentally determined models of human diseases. This provides an accurate representation of the disease model, as just as in humans, the phenotypic manifestation of a disease-causing mutation may differ significantly depending on the composition of the genetic background that may include modifying loci.

Annotations to human disease are based on published experimental assertions using human disease terms from OMIM (Online Mendelian Inheritance in Man, http://omim.org). There are five types of mouse models to human disease relationships used by MGD: models with phenotypic similarity to human disease where the causative gene in both the human disease and the mouse model involves mutations in orthologous genes; models with phenotypic similarity to human disease, but where the etiologies are distinct (mutations are not in orthologous genes); models with phenotypic similarity to human disease, but where the etiology is unknown; models with phenotypic similarity to human disease, and the mouse carries a transgene insertion (sometimes of a mutant human gene); and “not” models, where phenotypic similarity to the human disease was expected (usually based on creating a knockout of the human ortholog), but the resulting mice did not display symptoms of the human disease. As of May 2015, more than 4,530 mouse genome-types have been associated with at least one OMIM disease, covering over 1,350 human diseases. See

**Table 1**
*Common Ways to Search MGD and the Data That Can Be Retrieved**

| Search form | Access from | What is searched for | What is retrieved |
|---|---|---|---|
| Quick Search (most used access method) | Located prominently at the top of the topics index boxes on the homepage (www.informatics.jax.org) and in the upper right corner of other web pages | Broad searches through all MGD data, including nomenclature, vocabulary/ ontology terms, and annotations; provides less search specificity, but greater breadth in a simple search | Genome Features (protein coding genes, non-coding RNA genes, QTL, symbol synonyms, etc.) and Vocabulary Terms (GO terms, Phenotype terms, Diseases, etc.) with links to data details |
| Genes and Markers Query | Use Search pull-down on the navigation bar (Genes) or select Genes from topic index boxes on the homepage | Genes and genome features using search parameters: nomenclature, feature type (*e.g*., protein coding gene), genome location, InterPro Domains, GO, Phenotypes, or Diseases annotations | Genes and genome features with location, maps, homologs, mutants and alleles, GO annotations, embryonic expression, sequence and protein links, references |
| Phenotypes, Alleles & Disease Model Search | Use Search pull-down on the navigation bar (Phenotypes) or select Phenotypes & Mutant Alleles from topic index boxes on the homepage | Mutant or genetically engineered alleles, transgenes, QTL, etc. using search parameters: phenotype or disease terms, gene or allele nomenclature, genome location, allele generation method and/or allele attributes, or allele project collections | Summary of alleles for specified parameters and displaying allele attributes, system level phenotypes, and human diseases modeled. Links to genotypes, phenotypes, and disease annotations |
| Human-Mouse Disease Connection | Use Search pull-down on the navigation bar (Human Disease) or select Human-Mouse: Disease Connection from the topic index boxes on the homepage | Mouse or human genes and orthologs, genome locations, phenotypes and disease terms. Searches accept mouse or human values as input; VCF files and text files of genes or IDs are also accepted input | Grid and table views showing mouse/human orthologs, phenotype classes, and human diseases/disease models fitting the parameters entered |
| Mouse Genome Browser | Use Search pull-down on the navigation bar (Mouse Genome Browser) or select the Browser under the Genes topic from the homepage | Search by chromosome and genome coordinates | Optionally turn on tracks for mutant alleles, SNPs, QTL, phenotypes, or switch to viewing human GRCh38 build or pseudo genomes for strains other than the C57BL/6J reference genome |
| Vocabulary Browsers | Use Search pull-down on the navigation bar and, hovering over the Vocabularies section, select GO, Mammalian Phenotype Ontology, OMIM | Search or browse vocabulary terms and select term of interest; GO and Phenotype Ontologies terms are displayed hierarchically; OMIM terms are displayed alphabetically | GO and MP terms provide definitions, synonyms and links to all annotations for the selected term; OMIM terms link to MGD Disease Model pages and to OMIM entries |

*This is not an exhaustive list of search methods or data that can be retrieved from MGD. Users are encouraged to explore MGD, visit the "Getting Started" section on the homepage (www.informatics.jax.org), the Help, and FAQ sections in the upper left of each web page, and to contact User Support (email mgi-help@jax.org) for additional questions and assistance.

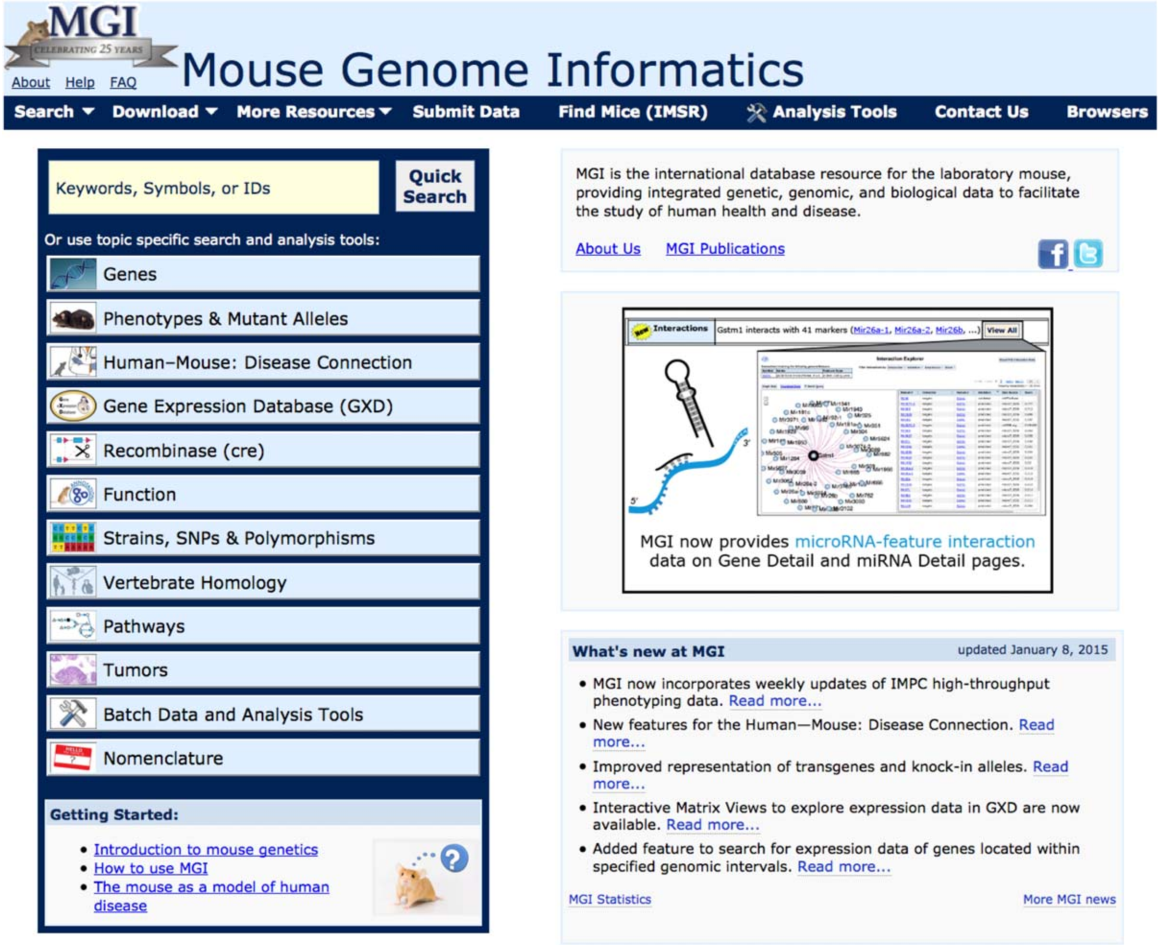
the section below on the Human-Mouse: Disease Connection interface that can be used to examine both experimentally supported mouse models, as well as suggest candidates for mouse models, and potential human gene candidates based on mouse models.

## WEB INTERFACE ACCESS TO MGD

MGD provides an extensive set of search forms, browsing tools, pre-generated downloadable data files, analysis tools, developer tools, various informational pages, and numerous links to other resources. A guide to several common ways to search MGD and the data that can be retrieved using those methods can be found in Table 1.

### MGD Homepage

Users are invited to explore MGD at www.informatics.jax.org (Fig. 2). The homepage features access to the


MGI
CELEBRATING 25 YEARS
About Help FAQ
Mouse Genome Informatics
Search ▾ Download ▾ More Resources ▾ Submit Data Find Mice (IMSR) Analysis Tools Contact Us Browsers
Keywords, Symbols, or IDs
Quick Search
Or use topic specific search and analysis tools:
Genes
Phenotypes & Mutant Alleles
Human–Mouse: Disease Connection
Gene Expression Database (GXD)
Recombinase (cre)
Function
Strains, SNPs & Polymorphisms
Vertebrate Homology
Pathways
Tumors
Batch Data and Analysis Tools
Nomenclature
Getting Started:
Introduction to mouse genetics
How to use MGI
The mouse as a model of human disease
About Us
MGI Publications
Interactions Gstm1 interacts with 41 markers (Mir26a-1, Mir26a-2, Mir26b, ...) View All
3'
5'
MGI Statistics
More MGI news

MGI is the international database resource for the laboratory mouse, providing integrated genetic, genomic, and biological data to facilitate the study of human health and disease.

MGI now provides microRNA-feature interaction data on Gene Detail and miRNA Detail pages.

What's new at MGI updated January 8, 2015

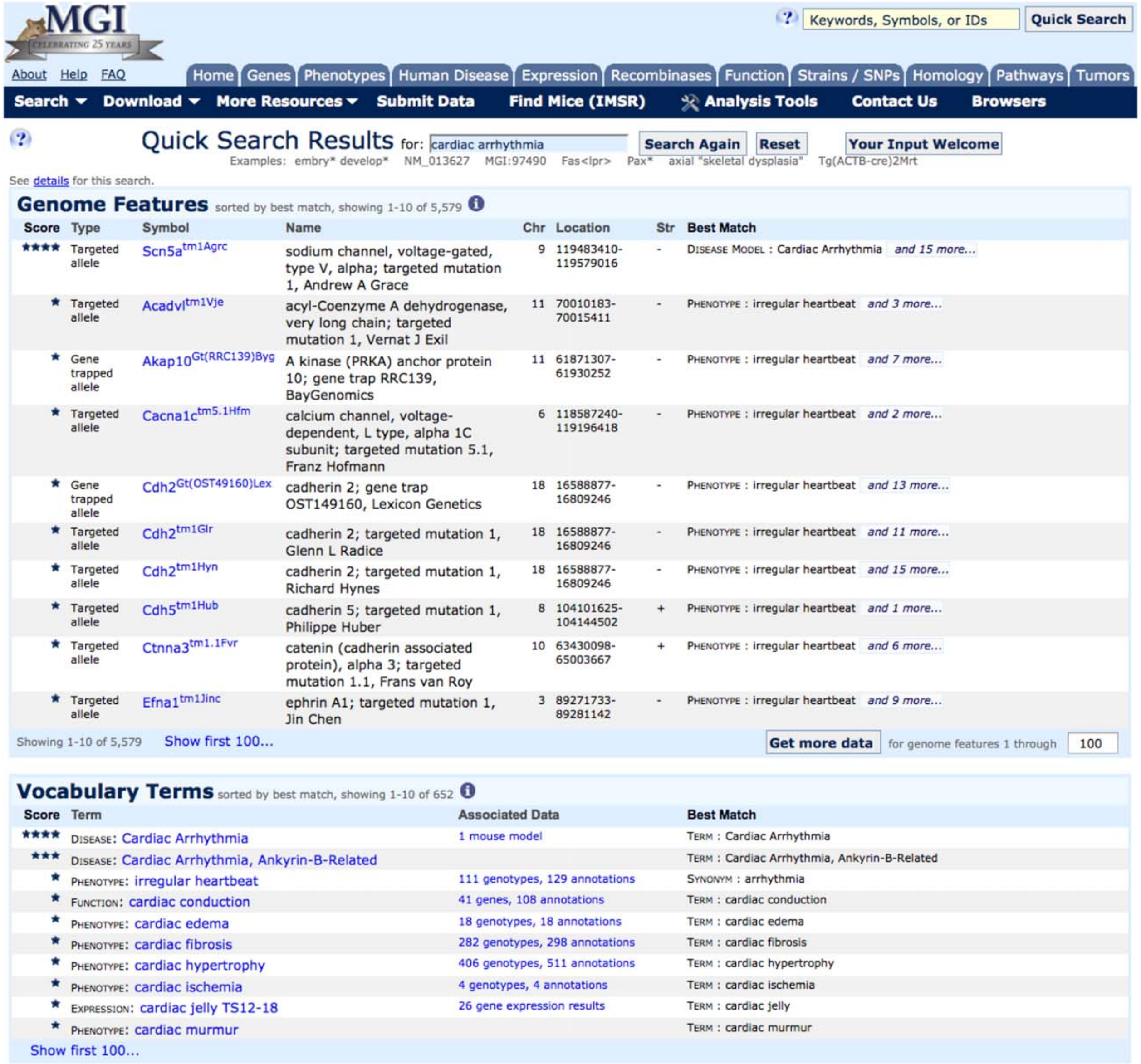
- MGI now incorporates weekly updates of IMPC high-throughput phenotyping data. Read more...
- New features for the Human—Mouse: Disease Connection. Read more...
- Improved representation of transgenes and knock-in alleles. Read more...
- Interactive Matrix Views to explore expression data in GXD are now available. Read more...
- Added feature to search for expression data of genes located within specified genomic intervals. Read more...

**FIG. 2.** MGD Homepage. This figure shows the top portion of the homepage including the table of contents for the database. In the topic-specific boxes (left), the major areas of data content are presented. Each topic-specific box is a link to a sub-page describing contents of that topic, links to specific search pages for that topic, FAQs, and links to documentation and collaborators. Beneath the topic index on the homepage is a "Getting Started" section of particular interest to new users. The "Quick Search" at the top of the table of contents is described further in Figure 3. The right side of the homepage includes links to and "About" page and to "MGI Publications," a rotating informational image, a "What's New" section, providing information on the latest software and web changes, and MGI statistics on data content. The bottom section (not shown) holds items of "Community Interest." The navy blue navigation bar at the top appears on each web page and allows users to quickly jump to the area of interest within MGD. The "Search," "Download," and "More Resources" items are each pull-down menus leading to additional choices.

Quick Search tool, a popular starting point for many purposes. There is a menu of topic areas found at the Mouse Genome Informatics site, where one can explore specific interests, and a "Getting Started" section for new users. The navy blue navigation bar at the top is present on all web pages, making it easy to jump from one data area to another. In the navigation bar, the "Search," "Download," and "More Resources" choices have pull-down menus as an additional navigation aid. Also available on the homepage is information about MGI publications, new features, and links to database statistics and former news. In the following sections of this review, we highlight key MGD features to provide a general flavor of the web interface. Users are encouraged to take advantage of our outreach and user help (see below) for assistance and questions about specific elements of the MGD web site.

## Quick Search Tool

A primary access point into MGD is through use of the *Quick Search* tool, available from nearly every web page on the site (Fig. 3). The Quick Search tool is useful for those already familiar with MGD's web pages, who are looking to access a very specific piece of information quickly (e.g., find gene "X"). The Quick Search tool also provides a broad sweep across MGD data, providing an easy entry point for users unfamiliar with MGD content (e.g., what kinds of


MGI
CELEBRATING 25 YEARS
Keywords, Symbols, or IDs
Quick Search
About Help FAQ
Home Genes Phenotypes Human Disease Expression Recombinases Function Strains / SNPs Homology Pathways Tumors
Search Download More Resources Submit Data Find Mice (IMSR) Analysis Tools Contact Us Browsers
Search Again
Reset
Your Input Welcome
Get more data

Quick Search Results for: cardiac arrhythmia

Examples: embry* develop* NM_013627 MGI:97490 Fas<lpr> Pax* axial "skeletal dysplasia" Tg(ACTB-cre)2Mrt

See details for this search.

**Genome Features** sorted by best match, showing 1-10 of 5,579

| Score | Type | Symbol | Name | Chr | Location | Str | Best Match |
|---|---|---|---|---|---|---|---|
| ★★★★ | Targeted allele | Scn5a$^{tm1Agrc}$ | sodium channel, voltage-gated, type V, alpha; targeted mutation 1, Andrew A Grace | 9 | 119483410-119579016 | - | DISEASE MODEL : Cardiac Arrhythmia *and 15 more...* |
| ★ | Targeted allele | Acadvl$^{tm1Vje}$ | acyl-Coenzyme A dehydrogenase, very long chain; targeted mutation 1, Vernat J Exil | 11 | 70010183-70015411 | - | PHENOTYPE : irregular heartbeat *and 3 more...* |
| ★ | Gene trapped allele | Akap10$^{Gt(RRC139)Byg}$ | A kinase (PRKA) anchor protein 10; gene trap RRC139, BayGenomics | 11 | 61871307-61930252 | - | PHENOTYPE : irregular heartbeat *and 7 more...* |
| ★ | Targeted allele | Cacna1c$^{tm5.1Hfm}$ | calcium channel, voltage-dependent, L type, alpha 1C subunit; targeted mutation 5.1, Franz Hofmann | 6 | 118587240-119196418 | - | PHENOTYPE : irregular heartbeat *and 2 more...* |
| ★ | Gene trapped allele | Cdh2$^{Gt(OST49160)Lex}$ | cadherin 2; gene trap OST149160, Lexicon Genetics | 18 | 16588877-16809246 | - | PHENOTYPE : irregular heartbeat *and 13 more...* |
| ★ | Targeted allele | Cdh2$^{tm1Glr}$ | cadherin 2; targeted mutation 1, Glenn L Radice | 18 | 16588877-16809246 | - | PHENOTYPE : irregular heartbeat *and 11 more...* |
| ★ | Targeted allele | Cdh2$^{tm1Hyn}$ | cadherin 2; targeted mutation 1, Richard Hynes | 18 | 16588877-16809246 | - | PHENOTYPE : irregular heartbeat *and 15 more...* |
| ★ | Targeted allele | Cdh5$^{tm1Hub}$ | cadherin 5; targeted mutation 1, Philippe Huber | 8 | 104101625-104144502 | + | PHENOTYPE : irregular heartbeat *and 1 more...* |
| ★ | Targeted allele | Ctnna3$^{tm1.1Fvr}$ | catenin (cadherin associated protein), alpha 3; targeted mutation 1.1, Frans van Roy | 10 | 63430098-65003667 | + | PHENOTYPE : irregular heartbeat *and 6 more...* |
| ★ | Targeted allele | Efna1$^{tm1Jinc}$ | ephrin A1; targeted mutation 1, Jin Chen | 3 | 89271733-89281142 | - | PHENOTYPE : irregular heartbeat *and 9 more...* |

Showing 1-10 of 5,579 Show first 100... Get more data for genome features 1 through 100

**Vocabulary Terms** sorted by best match, showing 1-10 of 652

| Score | Term | Associated Data | Best Match |
|---|---|---|---|
| ★★★★ | DISEASE: Cardiac Arrhythmia | 1 mouse model | TERM : Cardiac Arrhythmia |
| ★★★ | DISEASE: Cardiac Arrhythmia, Ankyrin-B-Related | | TERM : Cardiac Arrhythmia, Ankyrin-B-Related |
| ★ | PHENOTYPE: irregular heartbeat | 111 genotypes, 129 annotations | SYNONYM : arrhythmia |
| ★ | FUNCTION: cardiac conduction | 41 genes, 108 annotations | TERM : cardiac conduction |
| ★ | PHENOTYPE: cardiac edema | 18 genotypes, 18 annotations | TERM : cardiac edema |
| ★ | PHENOTYPE: cardiac fibrosis | 282 genotypes, 298 annotations | TERM : cardiac fibrosis |
| ★ | PHENOTYPE: cardiac hypertrophy | 406 genotypes, 511 annotations | TERM : cardiac hypertrophy |
| ★ | PHENOTYPE: cardiac ischemia | 4 genotypes, 4 annotations | TERM : cardiac ischemia |
| ★ | EXPRESSION: cardiac jelly TS12-18 | 26 gene expression results | TERM : cardiac jelly |
| ★ | PHENOTYPE: cardiac murmur | | TERM : cardiac murmur |

Show first 100...

**FIG. 3.** Results of a Quick Search query. In this example, the search term *cardiac arrhythmia* was used. The results page displays two sections: Genome Features, delineating genes and genome features matched in the search (10 displayed of 5,579 results that can be viewed sequentially) and Vocabulary Terms, showing various annotations in MGD to phenotype, disease, expression, function, and protein domain terms matched in the search (10 of 652 results shown). Results are returned by weighting, so that the "best matched" results appear first. Rarely does one need to scroll through many pages to discover their term of interest.

information are available on "apoptosis" or related to "cardiomyopathy"). The Quick Search tool provides a variety of links to follow for more details.

The Quick Search tool is located prominently on the homepage at the top of the topic index and on all other pages at the upper right corner. The Quick Search tool allows users to enter keywords, symbols, or IDs from multiple databases (e.g., RefSeq, Ensembl, PubMed, MGI). Results are returned as a "weighted list by best match" in two sections: (1) by Genome Feature, where the search targets nomenclature or annotations to genes or genome feature symbols, names, former symbols, allele symbols, synonyms, etc. and (2) by Vocabulary Terms, where the search targets data annotations made against specific vocabulary and ontology terms employed in MGD. In this case, the results returned will be one or more terms from the Mouse Developmental or Adult Anatomies, the Mammalian Phenotype Ontology, Gene Ontology, OMIM disease terms, and protein domains and families with links to the term and links to the annotated data.

## Advanced Search Forms

Depending on the search terms or IDs entered, the Quick Search tool can return thousands of results to sort through. In such cases, using more specific terms with the Quick Search tool or using one of the more targeted topic-specific search forms that allows multiparameter searching can produce better results. For example, consider the results obtained with the Quick Search when entering the term *neuron*, which yields 9,218 Genome Feature matches and 1,122 Vocabulary matches (as of May 2015). The Genome Feature search results include both nomenclature matches to gene names and synonyms, as well as phenotype, disease, and functional annotations to the genome features. In contrast, entering the more specific term *"motor neuron degeneration"* (with quotes for exact match) yields 103 Genome Feature matches and 2 Vocabulary matches. Note that motor neuron degeneration *without* quotes will return a large number of results, as this quick search is performed on the entire term, motor neuron degeneration, as well as each word and each combination of words in the 3-word string (motor, neuron, degeneration, motor neuron, motor degeneration, neuron degeneration).

Using MGD's Advanced Search forms allows users to more precisely specify the parameters of a search and immediately narrow the scope of the search. The Advanced Search forms are accessed from any web page using the pull-down menu under "Search" in the navy blue navigation bar at the top of the page and choosing a search form under the listed topics. For example, one can use the more targeted *Genes and Markers Query* form (Fig. 4), accessed under the "Search" pull-down, submenu Genes. Once on the *Genes and Markers Query* form, entering the term *motor neuron degeneration* in the Gene/Marker box searches *only* MGD nomenclature and returns two genes where motor neuron degeneration is part of the name of mutant alleles of these genes.

## Mouse Genome Browser

MGD's Mouse Genome Browser is implemented using the open-source genome browser tool, JBrowse (Skinner *et al.*, 2009). JBrowse provides interactive web graphical displays of genome data and allows for frequent data updates and incorporation of large data files from MGD and external data sources. Currently, MGD's JBrowse implementation includes the following data types from MGD: genome features, alleles, phenotypes, QTL, and gene traps; from Ensembl: genes and regulatory features; from NCBI: genes and consensus CDS (cDNAs); and from Wellcome Trust Sanger Institute: Havana/Vega genes and Sanger SNPs. Also available for viewing is the human GRCh38 genome build and a pseudo genome for the CAST/EiJ mouse strain based on the analysis of sequence alignments (Munger *et al.*, 2014). The integration of biological annotations with large-scale sequence data and the mouse reference genome allow users to explore a genome view and biological context simultaneously. There are several ways to access the Mouse Genome Browser, including from the navigation bar of each web page under the "Search" pull-down menu, where it is listed directly as the last item in the pull-down. In addition, the Mouse Genome Browser link appears on "Search" submenus of the Genes or Sequences pull-down menu items, and under the Genes topic box on the homepage (Fig. 5).

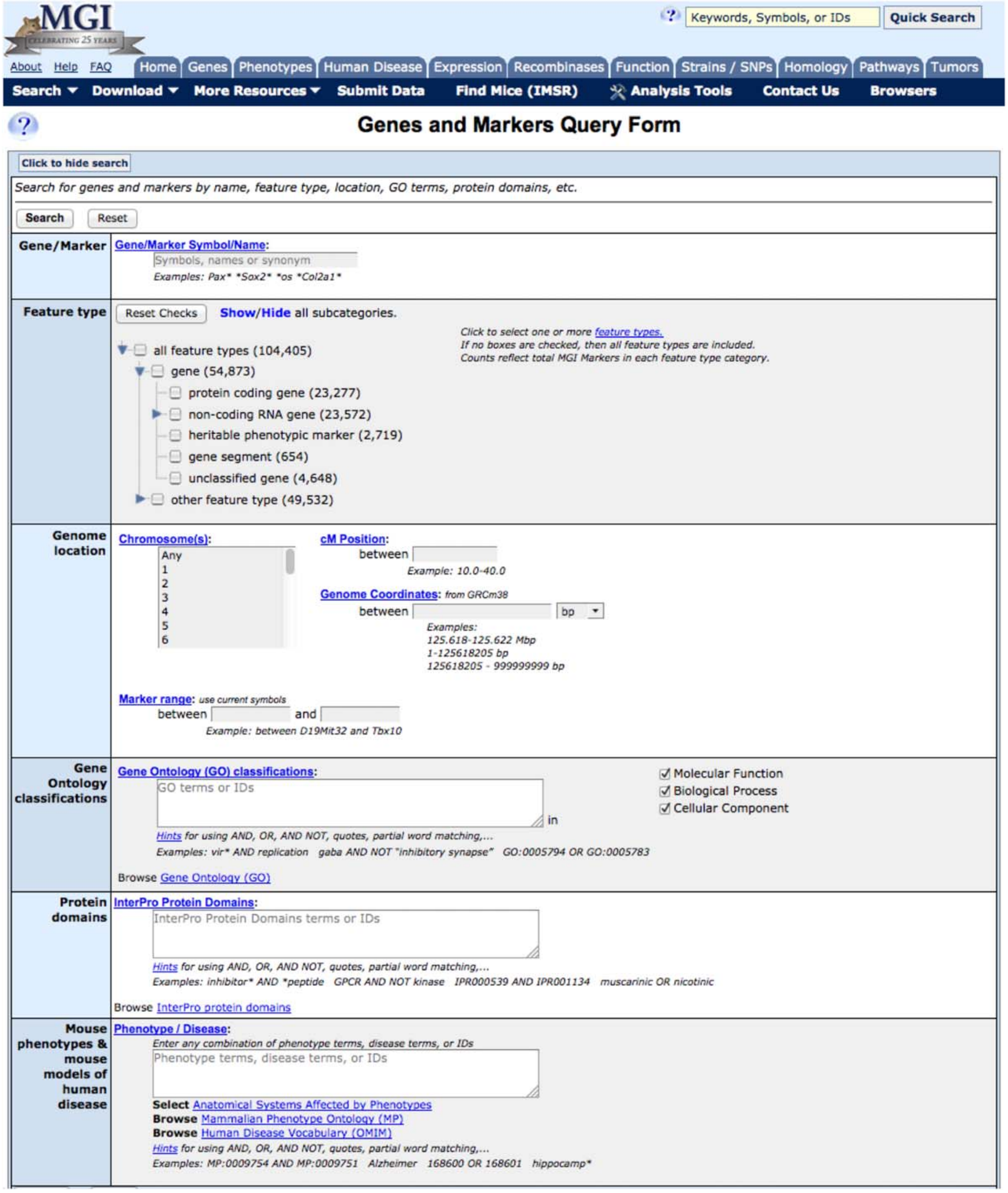
## Vocabulary/Ontology Browsers

As vocabularies and ontologies are key in mouse data annotation, MGD provides users with the ability to browse the terminologies. For simple vocabularies, such as OMIM terms, an alphabetical listing is provided with links from the disease term to the MGD Human Disease and Mouse Model Page, and from the OMIM ID to the disease entry at www.omim.org. For ontologies, including GO, Mammalian Phenotype, and Developmental and Adult Anatomies, the ontologies are presented hierarchically, reflecting the ontology structure and links are provided to all annotations associated with each term. These browsers are accessed via the navigation bar, under the Search pull-down menu and the sub-menu "Vocabularies."

## Human-Mouse: Disease Connection Interface

The Human-Mouse: Disease Connection (HMDC, Eppig *et al.*, 2015) is a translational tool developed to aid researchers in finding mouse models of human diseases and in identifying novel candidate disease genes. HMDC is accessed at www.diseasemodel.org, or by selecting "Human-Mouse: Disease Connection" from the topic boxes on the MGD homepage, or using the "Search" pull-down menu in the navigation bar and selecting "Human Disease."

Taking advantage of MGD's integrated mouse and human genomic, phenotypic, and disease data, users can approach the HMDC from a human or mouse perspective, searching for symbols, names, or IDs for genes, genome locations, or disease or phenotype terms. VCF (variant call format) files and text files of genes or gene IDs may be used as input as well. The initial search returns a grid (Fig. 6) with rows of human and mouse orthologs, and columns of phenotypes and human disease terms that are color-coded (blue for mouse data; orange for human data). Important in this example of Ehlers–Danlos syndrome, Type I, is that the ortholog pair human *COL5A1*and mouse *Col5a1* genes are both known to be associated with this disease as denoted by the intersecting bi-colored grid cell. Ehlers–Danlos syndrome, Type I is also associated with human genes *COL1A1* and *COL5A2* as denoted by orange grid


MGI
CELEBRATING 25 YEARS
Keywords, Symbols, or IDs
Quick Search
About Help FAQ
Home Genes Phenotypes Human Disease Expression Recombinases Function Strains / SNPs Homology Pathways Tumors
Search ▾ Download ▾ More Resources ▾ Submit Data Find Mice (IMSR) Analysis Tools Contact Us Browsers
Click to hide search
Search
Reset
Reset Checks

Genes and Markers Query Form

Search for genes and markers by name, feature type, location, GO terms, protein domains, etc.

**Gene/Marker** Gene/Marker Symbol/Name: Symbols, names or synonym
*Examples: Pax* *Sox2* *os *Col2a1**

**Feature type** **Show/Hide** all subcategories.

*Click to select one or more feature types.*
*If no boxes are checked, then all feature types are included.*
*Counts reflect total MGI Markers in each feature type category.*

- □ all feature types (104,405)
  - □ gene (54,873)
    - □ protein coding gene (23,277)
    - □ non-coding RNA gene (23,572)
    - □ heritable phenotypic marker (2,719)
    - □ gene segment (654)
    - □ unclassified gene (4,648)
  - □ other feature type (49,532)

**Genome location** Chromosome(s): Any 1 2 3 4 5 6

cM Position: between
*Example: 10.0-40.0*

Genome Coordinates: *from GRCm38*
between bp
*Examples:*
*125.618-125.622 Mbp*
*1-125618205 bp*
*125618205 - 999999999 bp*

Marker range: *use current symbols*
between and
*Example: between D19Mit32 and Tbx10*

**Gene Ontology classifications** Gene Ontology (GO) classifications: GO terms or IDs in
☑ Molecular Function
☑ Biological Process
☑ Cellular Component
*Hints for using AND, OR, AND NOT, quotes, partial word matching,...*
*Examples: vir* AND replication gaba AND NOT "inhibitory synapse" GO:0005794 OR GO:0005783*
Browse Gene Ontology (GO)

**Protein domains** InterPro Protein Domains: InterPro Protein Domains terms or IDs
*Hints for using AND, OR, AND NOT, quotes, partial word matching,...*
*Examples: inhibitor* AND *peptide GPCR AND NOT kinase IPR000539 AND IPR001134 muscarinic OR nicotinic*
Browse InterPro protein domains

**Mouse phenotypes & mouse models of human disease** Phenotype / Disease:
*Enter any combination of phenotype terms, disease terms, or IDs*
Phenotype terms, disease terms, or IDs
**Select** Anatomical Systems Affected by Phenotypes
**Browse** Mammalian Phenotype Ontology (MP)
**Browse** Human Disease Vocabulary (OMIM)
*Hints for using AND, OR, AND NOT, quotes, partial word matching,...*
*Examples: MP:0009754 AND MP:0009751 Alzheimer 168600 OR 168601 hippocamp**

**FIG. 4.** Genes and Marker Query Form. This Advanced Query Form illustrates the further precision in searching that one can obtain using specific Query Forms versus the Quick Search method. Here one can specify gene(s) by name or symbol, as one can in the Quick Search. But, in addition, data can be further specified by one or more characteristics—feature type, genome location, GO terms, Protein domains, and mouse phenotypes or human disease terms. To access the *Genes and Markers Query* form use the "Search" pull-down menu in the navigation bar and follow the Genes section to select *Genes and Markers Query* or click the Gene topic box in the content boxes on the homepage and follow the appropriate link.

cells. The mouse orthologs of both of these genes (*Col1a1* and *Col5a2*) have known phenotypic mutations in MGD, and thus are ready-made candidate genes for experimental validation as new Ehlers–Danlos syndrome, Type I models. Conversely, there is a mouse model for Ehlers–Danlos syndrome, Type I with mutations in mouse gene *Lum* (lumican), but the human ortholog has not yet been associated with this disease. Thus human *LUM* may

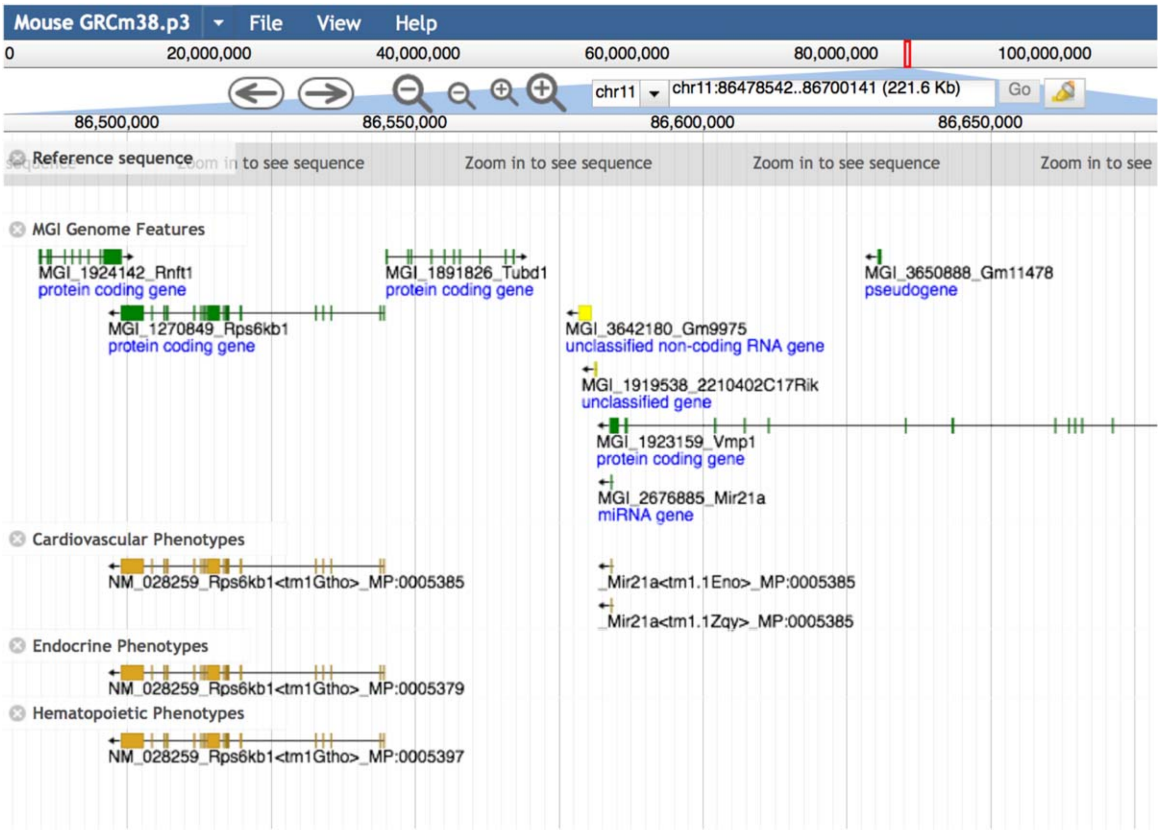


**FIG. 5.** Mouse Genome Browser results. Here the Mouse Genome Browser is providing a view of 221.6 kb of chromosome 11 from nucleotide 86478542 to 86700141. Genome features are displayed at the top (green) gene structures and the associated phenotypes shown in gold. There are several ways to access the Mouse Genome Browser, including from the navigation bar of each web page under the "Search" pull-down menu, where it is listed directly as the last item in the pull-down. In addition, the Mouse Genome Browser link appears on "Search" submenus of the Genes or Sequences pull-down menu items, and under the Genes topic box on the homepage.

be a new candidate to consider as causative in human patients or may be in a critical pathway for Ehlers–Danlos syndrome. Each colored grid cell is an active link to additional data detail, leading to specific genotype information, MP ontology annotations, links to obtaining mouse models, references, and more. Other data views are seen using the "Genes" and "Diseases" tabs located above the grid.

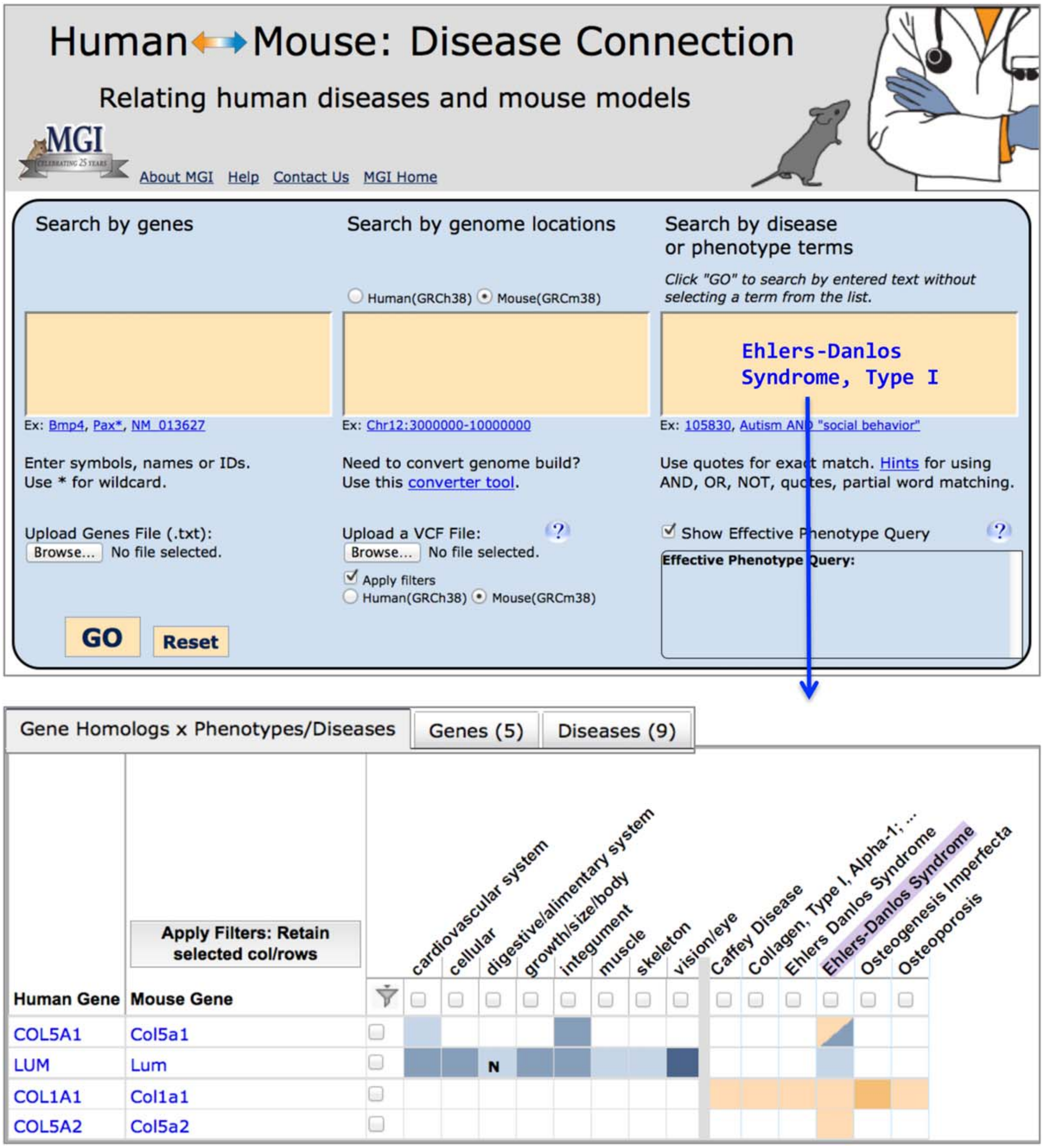
## WORKING WITH LARGE DATA SETS IN MGD

### Downloading MGD Data

Many web pages throughout the MGD website provide the ability to download the current data being viewed in text or Excel format. In addition, MGD maintains a large set of text files with specific subsets of data on the FTP site. These are frequently downloaded by collaborators and other biomedical resources that regularly incorporate MGD data, as well as by computational users. The files are updated weekly. Further, complete database dumps are accessible, along with a schema browser. A link to an index of commonly used FTP files (All MGI Reports) arranged by topic area with a description of their file contents and download link can be found in the "Download" pull-down menu on the navigation bar at the top of all MGD web pages. Users also can select a specific topic area to view subsets of MGD data reports or select Batch Query to create their own unique reports. Additional resources can be found using the "More Resources" pull-down on the navigation bar and selecting More Resources Index.

### Batch Query

Users often come to MGD with a set of genes of interest and the desire to retrieve additional information about that set of genes. The MGD Batch Query tool (Bult *et al.*, 2008) allows the user to type in, paste, or upload thousands of gene symbols or IDs from various data sources (MGI, Ensembl, RefSeq, etc.) and specify the additional information desired (GO annotations, mutant alleles, MP annotations, RefSNP IDs, Human Disease annotations, etc.). The Batch Query tool also serves as a convenient "ID converter," so that, for example a set of MGI IDs can bring back IDs for the same



Human ⟷ Mouse: Disease Connection

Relating human diseases and mouse models

Search by genes

Ex: Bmp4, Pax*, NM_013627

Enter symbols, names or IDs.
Use * for wildcard.

Upload Genes File (.txt):
No file selected.

Search by genome locations

○ Human(GRCh38) ● Mouse(GRCm38)

Ex: Chr12:3000000-10000000

Need to convert genome build?
Use this converter tool.

Upload a VCF File:
No file selected.

☑ Apply filters
○ Human(GRCh38) ● Mouse(GRCm38)

Search by disease
or phenotype terms

*Click "GO" to search by entered text without selecting a term from the list.*

Ehlers-Danlos
Syndrome, Type I

Ex: 105830, Autism AND "social behavior"

Use quotes for exact match. Hints for using AND, OR, NOT, quotes, partial word matching.

☑ Show Effective Phenotype Query

**Effective Phenotype Query:**

| Human Gene | Mouse Gene |  | cardiovascular system | cellular | digestive/alimentary system | growth/size/body | integument | muscle | skeleton | vision/eye | Caffey Disease | Collagen, Type I, Alpha-1; ... | Ehlers Danlos Syndrome | Ehlers-Danlos Syndrome | Osteogenesis Imperfecta | Osteoporosis |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  |  | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ |
| COL5A1 | Col5a1 | □ | 🟦 |  |  |  | 🟦 |  |  |  |  |  |  | 🟧🟦 |  |  |
| LUM | Lum | □ | 🟦 | 🟦 | N | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 |  |  |  | 🟦 |  |  |
| COL1A1 | Col1a1 | □ |  |  |  |  |  |  |  |  | 🟧 | 🟧 | 🟧 | 🟧 | 🟧 | 🟧 |
| COL5A2 | Col5a2 | □ |  |  |  |  |  |  |  |  |  |  |  | 🟧 |  |  |

**FIG. 6.** Human-Mouse: Disease Connection (HMDC), www.diseasemodel.org. The top panel shows the upper portion of the HMDC homepage with three distinct search boxes to allow searching by either mouse or human genes, genome locations, or disease or phenotype terms. Note that options are provided to upload a gene file or a VCF file to use as search parameters as well. In this example Ehlers–Danlos syndrome, Type I was entered in the disease/phenotype term box. The lower panel shows the resulting grid display where human and mouse orthologs are shown in rows and phenotypes and diseases are shown in columns. Blue indicates mouse data; orange indicates human data. The highlighted Ehlers–Danlos syndrome column shows both human *COL5A1* and mouse *Col5a1*, respectively, are associated to the disease. Mouse gene *Lum* and human genes *COL1A1* and *COL5A2* are associated to this human disease as well, but not coincidentally. These data suggest that mice with mutations in *Col1a1* and *Col5a2* should be examined for phenotypes correlated to human Ehlers–Danlos syndrome and that human patients with Ehlers–Danlos phenotypes might be checked for mutations in the *LUM* gene.

genes in Ensembl, NCBI Gene, and VEGA. The results are displayed as a paginated table with links to underlying data in MGD and the results table can be downloaded. The Batch Query tool can be found from the homepage by (1) selecting “Batch Data and Analysis Tools” in the topic selection boxes or, (2) using the navigation bar at the top of the page and selecting Batch Query from the “Search” pull-down menu or from the “Download” pull-down menu or following the “Analysis Tools” link in the bar.

### MouseMine

MouseMine (Sullivan *et al.*, 2013) is the mouse-specific instance of InterMine (Kalderimis *et al.*, 2014), an open source software framework supporting data warehousing with an interface of powerful tools for data querying, retrieval, and analysis. Users can access Mines of other Model Organisms (MODs) using common interfaces and tools, thus greatly lowering the barrier to exploring data in different MODs (see Lyne *et al.*, 2015 for an extensive review of InterMine). MouseMine is accessed via www.mousemine.org, by selecting MouseMine from the "Search" pull-down menu on the MGD web page navigation bar or by selecting "Batch Data and Analysis Tools" from the topic boxes on the homepage.

MouseMine includes most of the core data present in MGD (e.g., genes, genome features, genome coordinates, alleles and mutations, mouse strains and genotypes, ontologies, functional annotations, phenotype annotations, human disease annotations, vertebrate homologies, references, etc.), Data are updated weekly from MGD. In addition, other data sets not currently included in MGD are integrated in MouseMine through common object identifiers, such as homology data from PANTHER (Protein Analysis THrough Evolutionary Relationships, Mi *et al.*, 2010).

The MouseMine interface is not as user friendly as the MGD web interface, but provides a powerful tool for developing specific database queries that are not addressed using the MGD web query forms. Thus, MouseMine is a highly valuable tool for knowledgeable users and computational biologists. With MouseMine complex queries can be specified and display content can be individualized. Users can also save or download results for future analysis, export results to other tools such as Galaxy (Goecks *et al.*, 2010) or Genomespace (http://www.genomespace.org), or apply MouseMine's enrichment analysis tools. As well, MouseMine also provides a RESTful web services API (Applications Program Interface).

## COMMUNITY: OUTREACH AND USER SUPPORT

### User Input and Data Submission

MGD encourages user contributions in many ways. On each Gene Detail page and Allele Detail page, a "Your Input Welcome" button appears where users can comment on the page they are viewing by contributing new data or references or commenting if an error is noticed.

A suite of data submission tools is provided by clicking the "Submit Data" option in the MGD navigation bar at the top of each web page. These tools allow investigators to submit a new locus for establishing official nomenclature, submit a new variant or mutant allele, register a new mouse strain or submit phenotype data or expression data. Investigators with large or specialized data sets can contact MGD for assistance in formatting and contributing their data.

### User Support

MGD provides a variety of User Support services. Each MGD web page includes a "Contact Us" option on the navigation bar that leads to a web form for comments and questions. An active email address, mgi-help@jax.org is monitored 5-days per week for questions regarding data content or nomenclature, data access, bug reports, and data submission. User Support staff maintain an active community Listserve with nearly 2,000 active users participating in forum discussions.

### Online Help

MGD maintains extensive online documentation and tutorials covering current data content and interfaces. These include information about what kinds of data are integrated into MGD, how those data are annotated, help with searching, and definitions and explanations of the fields displayed in search results. In addition, Frequently Asked Questions (FAQ) pages provide illustrated examples that walk the user through common searches (e.g., How do I search for a gene in a specific genomic interval?). Links to "Help" and "FAQ" pages are located on all web pages in the upper left corner beneath the MGI logo. The homepage also includes a "Getting Started" section located beneath the topic index in the left column (Fig. 2). Newly added features and significant data content changes are highlighted in the "News and Release Announcements" section of the homepage.

### Digital Media

MGD is on Facebook (www.facebook.com/mgi.informatics) and maintains a Twitter account (MouseGenomeInfo@mgi_mouse). A newly released iPhone app, MGI GenomeCompass, includes the ability to save "favorite" genes, phenotypes, or diseases, allowing the user to receive automated updates whenever they log in. An Android app is in the planning stage.

### MGD Workshops

MGD staff frequently provide individual or group tutorials in conjunction with courses and conferences at the Jackson Laboratory and provide off-site workshops at various institutions as feasible. When MGD staff attend off-site conferences, these can be coupled with local workshops or MGD staff can schedule customized workshops on request (see http://www.informatics.jax.org/mgihome/support/Roadshow_description.pdf).

## EVOLUTION AND ONGOING CHALLENGES FOR MGD

The issues of growing and maintaining a viable bioinformatics resource for the research community are complex. MGD plays a critical role in the ecosystem of major biological databases. To maintain its necessary rapid and continued growth in data size and changing data types means continually searching for new efficiencies on every level, evaluating trends (what data types are no longer relevant? what are the next challenges from new biological data types), and reaching for new emerging collaborations.

Some clear examples of MGD's evolution include:

1. At the beginning, genes were placed on linkage maps; there were no sequences, only a few RFLPs (restriction fragment length polymorphisms); we did not know how many genes there might be; we sought to provide better maps, and identify new mutations that might help (indirectly) find new genes. As the Human Genome Project gathered momentum and the mouse began to be sequenced in earnest, attention turned to sequencing (still painfully slow and difficult to analyze); but grew in capability and analytical tool collections. MGD moved first to support physical maps and radiation hybrid maps, and finally to contig and sequence maps.
2. Mutations were initially detected by observation of spontaneous visible phenotypes. Radiation and certain mutagens began to be used for mutation induction, though their mechanisms were poorly understood. Later studies using ENU mutagenesis, where dosage was studied and mutagenesis maximized, became a standard and generated thousands of new mutations for study. Genetic engineering in mice became possible and propelled mice into a wunderkind position with this unique and powerful tool that remains a mainstay of mutagenesis and enables highly innovative and specifically designed genetic changes.
3. Sequencing technology continues to get cheaper, and with it the ability to sequence multiple inbred strains for variant analysis and to quickly identify the precise DNA change(s) responsible for a mutation, whether spontaneous or induced. We soon will be sequencing all our experimental animals and have individual profiles for each mouse.

The above are punctuated paradigm shifts brought about by adoption of new biological technology at a community scale, which continues to flourish and produce innovation for tomorrow's research. The database resources in this issue of *Genesis* represent those that have been able to adapt to the kinds of major research shifts mentioned above, as well as being adaptive to other disruptive advances in computer technologies necessary to provide the infrastructure for continued database advances.

## LITERATURE CITED

Andrews TD, Whittle B, Field MA, Balakishnan B, Zhang Y, Shao Y, Cho V, Kirk M, Singh M, Xia Y, Hager J, Winslade S, Sjollema G, Beutler B, Enders A, Goodnow CC. 2012. Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: An immediate source for thousands of new mouse models. Open Biol 2:120061.

Ayadi A, Birling MC, Bottomley J, Bussell J, Fuchs H, Fray M, Gailus-Durner V, Greenaway S, Houghton R, Karp N, Leblanc S, Lengger C, Maier H, Mallon AM, Marschall S, Melvin D, Morgan H, Pavlovic G, Ryder E, Skarnes WC, Selloum M, Ramirez-Solis R, Sorg T, Teboul L, Vasseur L, Walling A, Weaver T, Wells S, White JK, Bradley A, Adams DJ, Steel KP, Hrabě de Angelis M, Brown SD, Herault Y. 2012. Mouse large-scale phenotyping initiatives: Overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. Mamm Genome 23:600–610.

Boles MK, Wilkinson BM, Maxwell A, Lai L, Mills AA, Nishijima I, Salinger AP, Moskowitz I, Hirschi KK, Liu B, Bradley A, Justice MJ. 2009. A mouse chromosome 4 balancer ENU-mutagenesis screen isolates eleven lethal lines. BMC Genetics 10:12.

Bradley A, Anastassiadis K, Ayadi A, Battey JF, Bell C, Birling MC, Bottomley J, Brown SD, Burger A, Bult CJ, Bushell W, Collins FS, Desaintes C, Doe B, Economides A, Eppig JT, Finnell RH, Fletcher C, Fray M, Frendewey D, Friedel RH, Grosveld FG, Hansen J, Herault Y, Hicks G, Horlein A, Houghton R, Hrabe de Angelis M, Huylebroeck D, Iyer V, de Jong PJ, Kadin JA, Kaloff C, Kennedy K, Koutsourakis M, Lloyd KC, Marschall S, Mason J, McKerlie C, McLeod MP, von Melchner H, Moore M, Mujica AO, Nagy A, Nefedov M, Nutter LM, Pavlovic G, Peterson JL, Pollock J, Ramirez-Solis R, Rancourt DE, Raspa M, Remacle JE, Ringwald M, Rosen B, Rosenthal N, Rossant J, Ruiz Noppinger P, Ryder E, Schick JZ, Schnutgen F, Schofield P, Seisenberger C, Selloum M, Simpson EM, Skarnes WC, Smedley D, Stanford WL, Stewart AF, Stone K, Swan K, Tadepally H, Teboul L, Tocchini-Valentini GP, Valenzuela D, West AP, Yamamura K, Yoshinaga Y, Wurst W. 2012. The mammalian gene function resource: The International Knockout Mouse Consortium. Mamm Genome 23:580–586.

Brown SD, Moore MW. 2012. The International Mouse Phenotyping Consortium: Past and future perspectives on mouse phenotyping. Mamm Genome 23:632–640.

Bucan M, Eppig JT, Brown S. 2012. Mouse genomics programs and resources. The mouse: Pushing the boundaries. Mamm Genome 23:479–489.

Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA; Mouse Genome Database Group. 2008. The Mouse Genome Database (MGD): Mouse biology and model systems. Nucleic Acids Res 36:D724–D728.

Chesler EJ. 2014. Out of the bottleneck: The Diversity Outcross and Collaborative Cross mouse populations in behavioral genetics research. Mamm Genome 25:3–11.

Churchill GA, Gatti DM, Munger SC, Svenson KL. 2012. The Diversity Outbred mouse population. Mamm Genome 23:713–718.

Clark AT, Goldowitz D, Takahashi JS, Vitaterna MH, Siepka SM, Peters LL, Frankel WN, Carlson GA, Rossant J, Nadeau JH, Justice MJ. 2004. Implementing large-scale ENU mutagenesis screens in North America. Genetica 122:51–64.

Collaborative Cross Consortium. 2012. The genome architecture of the Collaborative Cross mouse genetic reference population. Genetics 190:389–401.

Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE; The Mouse Genome Database Group. 2015. The Mouse Genome Database (MGD): Facilitating mouse as a model for human biology and disease. Nucleic Acids Res 43:D726–D736.

Finger JH, Smith CM, Hayamizu TF, McCright IJ, Xu J, Eppig JT, Kadin JA, Richardson JE, Ringwald M. 2015. The mouse gene expression database: New features and how to use them effectively. Genesis. doi:10.1002/dvg.22864.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Garcia-Giron C, Gordon L, Hourlier T, Hunt S, Juettemann T, Kahari AK, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T, McLaren WM, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sheppard D, Sobral D, Taylor K, Thormann A, Trevanion S, White S, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Harrow J, Herrero J, Hubbard TJ, Johnson N, Kinsella R, Parker A, Spudich G, Yates A, Zadissa A, Searle SM. 2013. Ensembl 2013. Nucleic Acids Res 41:D48–D55.

Gene Ontology Consortium. 2000. Gene ontology: Tool for the unification of biology. Nature Genet 25:25–29.

Gene Ontology Consortium. 2015. Gene Ontology Consortium: Going forward. Nucleic Acids Res 43: D1049–D1056.

Goecks J, Nekrutenko A, Taylor J, Galaxy Team. 2010. Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11:R86.

Goldowitz D, Frankel WN, Takahashi JS, Holtz-Vitaterna M, Bult C, Kibbe WA, Snoddy J, Li Y, Pretel S, Yates J, Douglas J, Swanson DJ. 2004. Large-scale mutagenesis of the mouse to understand the genetic bases of nervous system structure and function. Brain Res Mol Brain Res 132:105–115.

Hayamizu TF, Wicks MN, Davidson DR, Burger A, Ringwald M, Baldock RA. 2013. EMAP/EMAPA ontology of mouse developmental anatomy: 2013 update. J Biomed Semantics 4:15.

Kalderimis A, Lyne R, Butano D, Contrino S, Lyne M, Heimbach J, Hu F, Smith R, Stepan R, Sullivan J, Micklem G. 2014. InterMine: Extensive web services for modern biology. Nucleic Acids Res 42: W468–W472.

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, Furlotte NA, Eskin E, Nellaker C, Whitley H, Cleak J, Janowitz D, Hernandez-Pliego P, Edwards A, Belgard TG, Oliver PL, McIntyre RE, Bhomra A, Nicod J, Gan X, Yuan W, van der Weyden L, Steward CA, Bala S, Stalker J, Mott R, Durbin R, Jackson IJ, Czechanski A, Guerra-Assuncao JA, Donahue LR, Reinholdt LG, Payseur BA, Ponting CP, Birney E, Flint J, Adams DJ. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477:289–294.

Koscielny G, Yaikhom G, Iyer V, Meehan TF, Morgan H, Atienza-Herrero J, Blake A, Chen CK, Easty R, Di Fenza A, Fiegel T, Grifiths M, Horne A, Karp NA, Kurbatova N, Mason JC, Matthews P, Oakley DJ, Qazi A, Regnart J, Retha A, Santos LA, Sneddon DJ, Warren J, Westerberg H, Wilson RJ, Melvin DG, Smedley D, Brown SD, Flicek P, Skarnes WC, Mallon AM, Parkinson H. 2014. The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. Nucleic Acids Res 42:D802–D809.

Li Y, Klena NT, Gabriel GC, Liu X, Kim AJ, Lemke K, Chen Y, Chatterjee B, Devine W, Damerla RR, Chang C, Yagi H, San Agustin JT, Thahir M, Anderton S, Lawhead C, Vescovi A, Pratt H, Morgan J, Haynes L, Smith CL, Eppig JT, Reinholdt L, Francis R, Leatherbury L, Ganapathiraju MK, Tobita K, Pazour GJ, Lo CW. 2015. Global genetic analysis in mice unveils central role for cilia in congenital heart disease. Nature 521:520–524.

Logan RW, Robledo RF, Recla JM, Philip VM, Bubier JA, Jay JJ, Harwood C, Wilcox T, Gatti DM, Bult CJ, Churchill GA, Chesler EJ. 2013. High-precision genetic mapping of behavioral traits in the diversity outbred mouse population. Genes Brain Behav 12: 424–437.

Lyne R, Sullivan J, Butano D, Contrino S, Heimbach J, Hu F, Kalderimis A, Lyne M, Smith RN, Stepan R, Balakrishnan R, Binkley G, Harris T, Karra K, Moxon SA, Motenko H, Neuhauser S, Ruzicka L, Cherry M, Richardson J, Stein L, Westerfield M, Worthey E, Micklem G. 2015. Cross-organism analysis using InterMine. Genesis. doi:10.1002/dvg.22869.

Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. 2010. PANTHER version 7: Improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Res 38:D204–D210.

Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S, Sangrador-Vegas A, Scheremetjew M, Rato C, Yong SY, Bateman A, Punta M, Attwood TK, Sigrist CJ, Redaschi N, Rivoire C, Xenarios I, Kahn D, Guyot D, Bork P, Letunic I, Gough J, Oates M, Haft D, Huang H, Natale DA, Wu CH, Orengo C, Sillitoe I, Mi H, Thomas PD, Finn RD. 2015. The InterPro protein families database: The classification resource after 15 years. Nucleic Acids Res 43:D213–D221.

Morgan H, Beck T, Blake A, Gates H, Adams N, Debouzy G, Leblanc S, Lengger C, Maier H, Melvin D, Meziane H, Richardson D, Wells S, White J, Wood J, EUMODIC Consortium de Angelis MH, Brown SD, Hancock JM Mallon AM. 2010. EuroPhenome: A repository for high-throughput mouse phenotyping data. Nucleic Acids Res 38:D577–D585.

Munger SC, Raghupathy N, Choi K, Simons AK, Gatti DM, Hinerfeld DA, Svenson KL, Keller MP, Attie AD, Hibbs MA, Graber JH, Chesler EJ, Churchill GA. 2014. RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. Genetics 198:59–73.

Murray SA, Eppig JT, Smedley D, Simpson EM, Rosenthal N. 2012. Beyond knockouts: cre resources for conditional mutagenesis. Mamm Genome 23:587–599.

Nord AS, Chang PJ, Conklin BR, Cox AV, Harper CA, Hicks GG, Huang CC, Johns SJ, Kawamoto M, Liu S, Meng EC, Morris JH, Rossant J, Ruiz P, Skarnes WC, Soriano P, Stanford WL, Stryke D, von Melchner H, Wurst W, Yamamura K, Young SG, Babbitt PC, Ferrin TE. 2006. The International Gene Trap Consortium Website: A portal to all publicly available gene trap cell lines in mouse. Nucleic Acids Res 34: D642–D648.

Philip VM, Sokoloff G, Ackert-Bicknell CL, Striz M, Branstetter L, Beckmann MA, Spence JS, Jackson BL, Galloway LD, Barker P, Wymore AM, Hunsicker PR, Durtschi DC, Shaw GS, Shinpock S, Manly KF, Miller DR, Donohue KD, Culiat CT, Churchill GA, Lariviere WR, Palmer AA, O'Hara BF, Voy BH, Chesler EJ. 2011. Genetic analysis in the Collaborative Cross breeding population. Genome Res 21:1223–1238.

Reference Genome Group of the Gene Ontology Consortium. 2009. The Gene Ontology's Reference Genome Project: A unified framework for functional annotation across species. PLoS Comput Biol 5: e1000431.

Richardson JE. 2006. fjoin: Simple and efficient computation of feature overlaps. J Comput Biol 13:1457–1464.

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Krasnov S, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Karsch-Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, Wilbur WJ, Yaschenko E, Ye J. 2012. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 40: D13–D25.

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. 2009. JBrowse: A next-generation genome browser. Genome Res 19:630–638.

Smith CL, Eppig JT. 2012. The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. Mamm Genome 23:653–668.

Smith CL, Goldsmith C-AW, Eppig JT. 2005. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol 6:R7.

Sullivan J, Karra K, Moxon SA, Vallejos A, Motenko H, Wong JD, Aleksic J, Balakrishnan R, Binkley G, Harris T, Hitz B, Jayaraman P, Lyne R, Neuhauser S, Pich C, Smith RN, Trinh Q, Cherry JM, Richardson J, Stein L, Twigger S, Westerfield M, Worthey E, Micklem G. 2013. InterMOD: Integrated data and tools for the unification of model organism research. Sci Rep 3:1802.

Svenson KL, Gatti DM, Valdar W, Welsh CE, Cheng R, Chesler EJ, Palmer AA, McMillan L, Churchill GA. 2012. High-resolution genetic mapping using the Mouse Diversity Outbred population. Genetics 190: 437–447.

Threadgill DW, Churchill GA. 2012. Ten years of the collaborative cross. G3 2:153–156.

Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, Jaenisch R. 2013. One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell 153:910–918.

Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, Harrow JL. 2008. The vertebrate genome annotation (Vega) database. Nucleic Acids Res 36:D753–D760.

Yalcin B, Adams DJ, Flint J, Keane TM. 2012. Next-generation sequencing of experimental mouse strains. Mammalian Genome 23:490–498.