Mouse Genome Informatics (MGI): latest news from MGD and GXD

Martin Ringwald[1] · Joel E. Richardson[1] · Richard M. Baldarelli[1] · Judith A. Blake[1] · James A. Kadin[1] ·
Cynthia Smith[1] · Carol J. Bult[1]

Abstract

The Mouse Genome Informatics (MGI) database system combines multiple expertly curated community data resources into a shared knowledge management ecosystem united by common metadata annotation standards. MGI's mission is to facilitate the use of the mouse as an experimental model for understanding the genetic and genomic basis of human health and disease. MGI is the authoritative source for mouse gene, allele, and strain nomenclature and is the primary source of mouse phenotype annotations, functional annotations, developmental gene expression information, and annotations of mouse models with human diseases. MGI maintains mouse anatomy and phenotype ontologies and contributes to the development of the Gene Ontology and Disease Ontology and uses these ontologies as standard terminologies for annotation. The Mouse Genome Database (MGD) and the Gene Expression Database (GXD) are MGI's two major knowledgebases. Here, we highlight some of the recent changes and enhancements to MGD and GXD that have been implemented in response to changing needs of the biomedical research community and to improve the efficiency of expert curation. MGI can be accessed freely at http://www.informatics.jax.org.

Introduction to Mouse Genome Informatics

As the cost of genome sequencing continues to decrease and as genome editing technologies become widely adopted, the laboratory mouse is more important than ever as a model system for understanding the biological significance of human genetic variation and for research needed to validate and safely advance the emerging practice of genomic medicine.

Mouse Genome Informatics (MGI) is a community genome knowledgebase resource focused on supporting investigations into human biology and disease through the integration of genetic, genomic, and biological data for the laboratory mouse. MGI comprises multiple databases that share common infrastructure and/or common data standards: Mouse Genome Database (MGD) (Blake et al. 2021), Gene Expression Database (GXD) (Baldarelli et al. 2021), Mouse Models of Human Cancer database (MMHCdb) (Krupke et al. 2017), CrePortal (Perry et al. 2021), and the

International Mouse Strain Resource (IMSR) (Eppig et al. 2015b). In this report, we focus on the latest advancements for the two major database components of MGI: MGD and GXD.

The uniqueness of the resources available from MGI is rooted in our emphasis on deep *data integration* which is enabled by the rigorous semantic standards applied to the biological annotations represented in MGI's contributing databases. MGI search capabilities extend far beyond browsing or keyword searches because they allow researchers to combine multiple parameters and concepts in a single search. For example, a query such as "What genes on chromosome 17, studied in mouse models of Spina bifida, are expressed in the neural tube and involved in signaling?" can be answered with a single search at MGI even though the data/information about genome location and feature type, disease/developmental disorder association, and function needed to answer this question with precision comes from many different sources. In contrast, search engines such as PubMed and Google lack the rigor of ontological standardization as well as the underlying information system architecture to support this type of data and knowledge integration. To answer a question like the one above outside of the MGI environment, multiple searches would be required, followed by the daunting and time-consuming tasks of sorting through

✉ Martin Ringwald
 martin.ringwald@jax.org

✉ Carol J. Bult
 carol.bult@jax.org

1 The Jackson Laboratory, Bar Harbor, ME, USA

and collating and quality checking a large resulting data set. The value of MGI as a core community knowledgebase is rooted in our constant extension and integration of new data types, and our support for comparative views of knowledge about human and mouse biology. We continually adapt our operational processes to accommodate new types of experimental data and refine our knowledgebase to reflect evolving understanding of biological systems. Innovation is reflected in our scalable solutions to data acquisition, deep data integration, and execution of FAIR (Findable, Accessible, Interoperable, Reproducible) principles of data management (Wilkinson et al. 2016).

Programmatic access to MGI is supported by MouseMine (Motenko et al. 2015), a data warehouse built with InterMine (Smith et al. 2012). MouseMine provides computational users with powerful features such as customizable queries and reports, lists which can be saved from query results or used to drive subsequent queries, built in enrichment analysis, and interconnections with mines for other model organisms. MouseMine is updated weekly from MGI and includes core data such as the unified mouse genome feature catalog, biological annotations including gene function, phenotype and disease associations, and gene expression data. As well, MouseMine includes the complete genomes (assemblies and gene models) of 19 mouse strains. A comprehensive API provides programmatic access to all features, data, and tools in MouseMine, significantly expanding FAIR access to MGI data.

One of the hallmarks of MGI is the expert curation which underlies robust data integration of knowledge from the peer-reviewed scientific literature and other sources. The mouse literature is expanding at a rate of over 80,000 publications in PubMed per year, but only a fraction of these publications is relevant to the MGI mission. For example, over the years 2017–2020, MGI curators reviewed an average of 31,197 papers per year and selected (aka, triaged) 14,820 of those papers for detailed curation. MGI-relevant publications are published in approximately 900 scientific journals—with most papers coming from approximately 120 journals. Identifying relevant articles requires searching full text, as many of these publications do not mention mouse in the title, abstract, or keywords. A recent enhancement to MGI's literature triage process is the implementation of a machine learning classifier applied to the full text of each article to decide if it is appropriate for at least one of the MGI knowledgebase resources. The performance metrics of precision, recall, and negative predictive value (NPV) for the classifier are 0.85, 0.90, and 0.90, respectively. An NPV of 0.90 means that 90% of the predictions that an article is not relevant to any of the databases that make up MGI are correct and only 10% are false negatives (i.e., incorrectly discarded). The application of machine learning to this critical step in the curation process allows curators to efficiently 'discard' papers that are not relevant, thus saving the hundreds of hours of manuscript review time.

On average, MGI receives 7.5 million page views per year and is accessed by a user base of over 400 thousand individuals that use MGI 1 to 9 times per year. More than 14 thousand users access the site at least 10 times per year. These usage data underestimate the total MGI user base as many researchers access MGI annotations and information from other widely used data resources including, but not limited to, the Alliance of Genome Resources (Alliance of Genome Resources 2020), the Rat Genome Database (RGD) (Smith et al. 2020b), the National Center for Biotechnology Information's (NCBI) Gene resource (Brown et al. 2015), the International Mouse Phenotyping Consortium (IMPC) (Munoz-Fuentes et al. 2018), the Ensembl genome browser (Newman et al. 2018), UCSC Genome Browser's VisiGene (Kuhn et al. 2007), SciCrunch (Bandrowski et al. 2015), the Protein InteraCtion KnowLedgebasE (PICKLE) (Dimitrako-poulos et al. 2020), the Online Gene Essentiality database (OGEE) (Gurumayum et al. 2021), Bio Gene Portal System (BioGPS) (Ringwald et al. 2012), GlyGen (York et al. 2020), JAX Synteny Browser (Kolishovski et al. 2019), Mutant Mouse Resource and Research Center (MMRRC) (Amos-Landgraf et al. 2021), European Mutant Mouse Archive (EMMA) (Hagn et al. 2007), Gene Ontology Consortium (Gene Ontology 2015), and the Online Mendelian Inheritance in Man (OMIM) knowledgebase (Hamosh et al. 2021). Annotations from MGI are also central to the development and application of machine learning and semantic reasoning methods for the prediction of mouse gene function and phenotypes for such projects as GeneWeaver (Baker et al. 2012), Functional Networks of Tissues in Mouse (FNTM) (Goya et al. 2015), Phenodigm (Smedley et al. 2013), and the Monarch Initiative (Shefchek et al. 2020).

Mouse Genome Database (MGD)

MGD is the community model organism database for the laboratory mouse and is a primary resource for biological reference data related to mouse genes and other genome features, functional annotations, phenotypes, and disease models with a strong emphasis on the relationship of these data to human biology and disease. Each of these areas is described below. The resources and annotations for which MGD serves as the authoritative source are shown in Table 1. The primary target user communities for MGD include basic scientists and translational/clinical researchers using mouse as a model organism, computational biologists, and bioinformatics resource development groups. In addition to being a major component of the MGI information system, MGD is a founding member of the Alliance of Genome Resources (Alliance of Genome Resources 2019; Alliance of Genome

Table 1 Resources and annotations for which MGD serves as the authoritative source

Data type or resource	Maintained as	Explanation
Unified mouse genome feature catalog	Catalog of genome features with chromosome location, persistent identifiers, cross-links to other provider identifiers and sequences; Table of the distribution of equivalent genome features across different strains of mice	MGD integrates computational gene predictions from GenCode/Ensembl and NCBI into a single, comprehensive, non-redundant resource. The catalog is the source of mouse genome feature annotations for the Alliance of Genome Resources
Gene Ontology (GO) annotations for mouse	Associations between mouse genes and GO terms incorporating contextual data for function assertions from multiple biomedical ontologies	MGD provides primary curation from the scientific literature and other mouse annotation groups and distributes the definitive set mouse GO annotations
Mammalian Phenotype (MP) Ontology	Ontology of defined phenotype terms and relationships	MGD develops vocabulary terms with community input and distributes MP via multiple ontology resource sites and the MGD FTP site
Mammalian Phenotype Ontology (MP) annotations	Associations between mouse genotypes and MP terms	MGD provides primary curation and integration with alleles, strains, and genotypes
Mouse models of human diseases	Associations between mouse genotypes and human disease terms	MGD provides primary curation with ongoing coordination with Disease Ontology, OMIM, and NCBI for human gene–disease associations
Symbols and names for genes & genome features	Official nomenclature for genome features; synonyms; nomenclature history	MGD is the web host for nomenclature rules set by International Committee on Standardized Genetic Nomenclature for Mice; nomenclature is coordinated with nomenclature groups for human and rat
Symbols and names for mutant alleles, transgenes, and genome rearrangements	Comprehensive catalog of mutations; persistent identifiers; descriptions of mutant construction and inheritance	MGD makes primary assignments; coordinates with many mutagenesis projects
Strain nomenclature	Catalog of mouse strains with persistent identifiers	MGD provides nomenclature and unique, permanent identifiers to multiple resources

Resources Consortium 2019). The Alliance brings together expertly curated annotations from multiple model organisms and human into a single information portal, greatly simplifying the comparison of biological knowledge available for the genomes of diverse model organisms.

Unified gene catalog, variants, and orthology

The foundation for biological annotations (i.e., function, phenotype, expression) in MGI is the MGD unified gene catalog (Zhu et al. 2015). The unified gene catalog is generated by combining the mouse genome annotations generated by GenCode/Ensembl (Frankish et al. 2021) and NCBI (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/) into a canonical, non-redundant catalog of mouse genes and other genome features for the laboratory mouse. We create and maintain the unified gene catalog as a consensus view of mouse genome annotation as many predicted genes are unique to a particular genome annotation pipeline.

The unified gene catalog is generated using a semi-automated analysis pipeline that determines equivalency of two gene predictions (annotations) based on genome coordinate overlap. We use the feature join ("fjoin") algorithm to determine coordinate overlap (Richardson 2006). The inputs for fjoin are files of genome coordinates of predicted features from two different sources that are in General Feature Format (GFF3; http://www.sequenceontology.org/gff3.shtml). The fjoin algorithm is designed to determine overlaps of genome coordinates very efficiently; the comparison of two genome annotation inputs with hundreds of thousands of annotated features takes only minutes to perform. Any genome feature that has genome coordinates can be used as input. fjoin outputs sets of features that have coordinate overlaps and those without overlaps. Genes with coordinate overlap are considered "equivalent." Equivalent genome features are not required to have identical structures and the degree to which features need to overlap to be considered equivalent is an adjustable parameter. We have recently updated all genome coordinates to the latest mouse genome assembly (GRCm39, GenBank accession: GCA_000001635.9) and integrated all genome annotations from GenCode/Ensembl and NCBI into the current MGD unified gene catalog.

Another recent enhancement to MGD's representation of genome features is the addition of annotations available from Ensembl for the genomes of 16 inbred mouse strains (Lilue et al. 2018) and two wild-derived strains (CAROLI/EiJ and PAHARI/EiJ) (Thybert et al. 2018). From MGD, the strain distribution of genome annotations is provided in tabular form allowing researchers to quickly identify genes that are in some strains but not others (Fig. 1A). To support visualization and interaction with multiple annotated genomes at one time, we also developed the Multiple Genome Viewer (MGV) (Richardson et al. 2021). MGV is a highly configurable visualization tool that is integrated with all the biological annotations of mouse genes for phenotype, function, and disease from MGD (Fig. 1B).
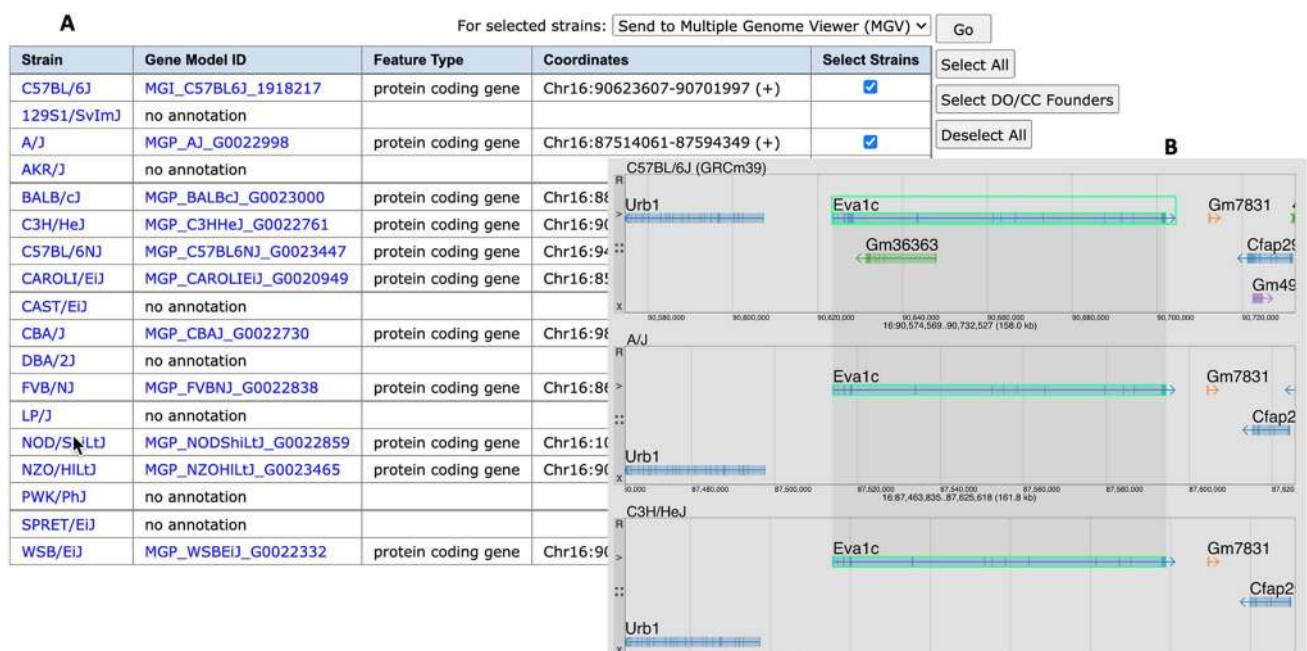
Fig. 1 A Tabular summary of the *Eva1c* gene (MGI:1918217) annotation across multiple strains of mice with three strains selected. **B** Screenshot of the Multiple Genome Viewer showing the *Eva1c* and surrounding annotations for the strains selected in the strain annotation distribution table

Sequence context for phenotypic alleles

The representation of the sequence context of variants in the mouse genome is another major focus for MGD curation efforts. Data for the sequence context of large-scale, high-throughput variants for the laboratory mouse are readily available from the European Variation Archive (EVA; https://www.ebi.ac.uk/eva/). For the set of variants from EVA available in MGD, we provide unique search and data display interfaces that support the identification of nucleotide variants that occur in some strains but not others. In contrast to high-throughput variants, the sequence context for mutant alleles of mouse genes reported in the scientific literature are largely available as free text. The lack of a standard syntax for representing these variants means they are difficult to search for or to use in computational analyses. MGD curators are systematically reviewing text descriptions available for tens of thousands of mutant mouse alleles associated with phenotypes and are converting them from text to Human Genome Variation Society (HGVS) standard notations (den Dunnen 2017). For example, the mutation details in MGD currently for the ENU-induced *Megf8*^{b2b288Clo} allele (MGI:5311364) are as follows: "The causative molecular lesion for the cardiovascular phenotypes is a A to T single point mutation at position 3641 of the cDNA (c.A3641T) (RefSeq NM_001160400). This is predicted to alter an arginine residue to an isoleucine at position 1214 (p.N1241I) in the encoded protein." This allele is now also represented in standard HGVS notation (NC_000073.6:g.25342303A > T) making it far easier to precisely locate the mutation on the mouse reference genome assembly and to use the variation as input into algorithms such a JANNOVAR (Jager et al. 2014) or the Ensembl Variant Effect Predictor (McLaren et al. 2016) to predict the functional consequences of the variation. Curated sequence notations for mouse phenotypic alleles are displayed on gene detail pages for the mouse at the Alliance of Genome Resources website (https://alliancegenome.org) and are also available for download from the Alliance.

Orthology

MGD's representation of orthology for genome features between mouse and other vertebrate species provides the basis for comparing biological knowledge and information between mouse and other vertebrate species. Orthology is central to many of the user interfaces supported by MGD including the Multiple Genome Viewer (see Richardson et al., this issue) and the Human–Mouse Disease Connection (HMDC) (Eppig et al. 2015a). We recently migrated our representation from NCBI's Homologene (Coordinators 2016) to the orthology assertions maintained by the Alliance of Genome Resources which are based on the

DIOPT algorithm (Hu et al. 2011). As the ortholog data from the Alliance are currently limited to human, mouse, rat, zebrafish, fly, worm, and yeast, we use HGNC's HCOP as a source of orthology for vertebrate species not yet in the Alliance set, including chimp, dog, and some agricultural species (Eyre et al. 2007; Yates et al. 2021).

Functional annotation

MGD is the leading source of annotations about the function of mouse protein-coding genes and functional RNAs, providing these data in a variety of formats to computational biologists and bioinformaticians and to general users in human readable formats through the MGI web interface, the Gene Ontology Consortium (GOC) AmiGO portal (Carbon et al. 2009), and the Alliance of Genome Resources (Alliance) (Alliance of Genome Resources 2019; Alliance of Genome Resources Consortium 2019). More than 32,000 mouse genome features (protein-coding and non-coding genes) in MGD have at least one functional annotation with almost 40% of these annotations based on direct experimental evidence. Nearly 500,000 curated GO annotations are available for mouse from MGD.

The annotation standards afforded by the Gene Ontology have been critical to robust comparisons of gene function across many organisms and to gleaning biological insights from the analysis of genome scale data (The Gene Ontology 2019). To make functional annotations using the Gene Ontology even more useful for representing biological systems, MGD curators have been working with the Gene Ontology Consortium to develop Causal Activity Models based on GO (GO-CAM) (Thomas et al. 2019). GO-CAM are models of biological systems and pathways that are generated by linking together multiple GO annotations in a semantically structured manner using the Relation Ontology (Smith et al. 2005). Over 300 GO-CAM models are currently available for mouse currently (https://geneontology.cloud/groups/MGI). Although not widely used by the research community currently, GO-CAMs are an important emerging informatics technology that will contribute to the development of data analysis methodologies that leverage expert curation for computational data integration and hypothesis generation.

Phenotype and disease annotation

The Mammalian Phenotype Ontology (MP) (Smith and Eppig 2012) is essential to the integration of phenotype data in MGD. MP is a structured vocabulary whose development and continued maintenance is centered at MGD with community input. The MP supports general and granular phenotype knowledge and allows robust searching and retrieval for web-based users as well as computational users. Each of the

more than 13,000 vocabulary terms in the MP has an accession ID, name, synonyms, definition, and a reference for the definition. Terms are organized hierarchically from general to specific, so phenotypes can be annotated at the finest level of granularity to reflect current knowledge. Over 360,000 MP annotations have been made in MGD and are associated with 54,177 mutations in 20,009 genes, transgene, and other markers. The MP has been adopted widely including by the Rat Genome Database (Smith et al. 2020b), the International Mouse Phenotyping Consortium (Munoz-Fuentes et al. 2018), and the Monarch Initiative (Mungall et al. 2017).

To enable the comparison of phenotype terms between mouse and human, we collaborate with the developers of the Human Phenotype Ontology (HPO) (Kohler et al. 2019) to generate mappings of phenotype terms between the two ontologies. We are collaborating to ensure that the definitions for the terms in the ontologies are correct and that they are comparable for equivalent terms. As a result, MP and HPO maintain the terminologies in common use among mouse geneticists and human clinicians, respectively, while sharing a common equivalence axiom that allows computational mapping of equivalent terms. This mapping, in turn, enables comparisons between phenotype profiles for mutant mice and for humans with genetic conditions to select the best mouse model or to select the best set of candidate genes.

While many of the nearly 13,000 phenotype terms in the MP have terms in the HPO that parallel each other in name and meaning, other terms share a common meaning but have different names. As well, some concepts are unique to one species or the other, and some can be "matched" or partially matched but are conceptually difficult to establish equivalencies. For example, cataract is a shared term in mouse and human with identical use, whereas absent eyelids (mouse) is equivalent to ablepharon (human), but the specific terms used differ. Complex examples include descriptions of abnormalities of the prostate gland because the anatomical structure is different in human and mouse, even though the organ performs a similar function in the two organisms. A particularly complex area for phenotype comparison is behavior as equivalence of behavioral phenotypes in mouse and human can be difficult to establish. To date, approximately 80% of MP and 50% HPO have been reviewed and agreed on as to Ontology Web Language (OWL) pattern equivalence axioms. MGI, HPO, and the Alliance continue to work collaboratively to develop equivalence axioms across multiple species as members of the POTATO (Phenotype Ontologies Transversing all the Organisms) community (https://zenodo.org/record/2382757).

Disease models

Providing access to published and potential mouse models of human disease is one of the core functions of MGD. The Human–Mouse: Disease Connection (HMDC) (Eppig et al. 2015a) was developed specifically for this purpose. Mouse mutation, and phenotype and disease model data are integrated with human gene-to-disease relationships from the National Center for Biotechnology Information (NCBI) and Online Mendelian Inheritance in Man (OMIM) and with human disease-to-phenotype relationships from the Human Phenotype Ontology (HPO). One of the significant enhancements to HMDC is the use of the Disease Ontology (DO) for standard disease annotations (Bello et al. 2018; Schriml et al. 2019). Previously, most of the disease annotations in MGD and HMDC were based on OMIM phenotype terms. The DO provides a much broader disease terminology for disease annotation but retains cross-links to other terminologies, including OMIM.

One of the major benefits to the HMDC is that researchers can search by human or mouse gene symbols, human or mouse phenotype terms, or human disease terms. The search results are displayed with color coding that indicates how similar the phenotypes and disease annotations are between mouse and human for a given gene (Fig. 2). While gene-to-disease relationship are useful, they can also be misleading as the phenotypes observed for an allelic variant of a gene can differ greatly depending on the genetic background. Thus, HMDC provides a genotype level overview of mouse models that can help researchers find the most appropriate model for their disease of interest.

Gene Expression Database (GXD)

The mouse Gene Expression Database (GXD) collects and integrates different types of gene expression information, with a focus on endogenous gene expression in wild-type and mutant mice. Because GXD is an integral component of the larger MGI system, its expression data are combined with the genetic, functional, and phenotypic data collected by MGD. This enables many biologically and biomedically relevant search capabilities and facilitates insights into the molecular mechanism of health and disease. Here, we provide a brief update on GXD data content and then highlight some of the new interface features made possible by the high level of data integration in MGI.

Classical types of expression data

For many years, GXD has collected mouse developmental expression data from RNA in situ hybridization, immunohistochemistry, in situ reporter (knock-in) RT-PCR, northern blot, and western blot experiments (Finger et al. 2017; Ringwald et al. 1999; Smith et al. 2015). These data are acquired through systematic curation of the scientific literature and by collaborations with large-sale expression projects

Fig. 2 Screenshot from MGD's Human–Mouse Disease Connection (HMDC) **A** Results of search for the gene *Fgfr3* (MGI:95524) showing the ribbon format overview of phenotype term annotation similarities of mouse and human. Cells with orange indicate annotations for the human gene for a phenotype term. Cells with blue shading indicate annotations for a mouse gene for a phenotype term. Cells with both orange and blue indicate a phenotype term shared by mouse and human. Depth of color indicates the number of published annotations.

Mouse genes starting with "Tg" represent transgenic mice carrying either the mouse or human *Fgfr3* gene. **B** Arrow indicates the results of clicking on a cell for a disease term showing specific disease annotations at the gene level for human and at the genotype level for the mouse. The "Find Mice" button is a link to the International Mouse Strain Resource (IMSR) (Eppig et al. 2015b) which lists the availability of mouse models from repositories around the world

(Diez-Roux et al. 2011; Harding et al. 2011; Koscielny et al. 2014; Magdaleno et al. 2006; Visel et al. 2004). Curators annotate these data in detail using official gene, mouse strain and allele nomenclature, controlled vocabularies, and an extensive anatomy ontology (Hayamizu et al. 2015), thus enabling data integration and search capabilities. Expression images are included in expression records when available. See http://www.informatics.jax.org/assay/MGI:6147780 for an immunohistochemistry record. As of June 15, 2021, GXD contains detailed expression data for over 15,400 genes, including data from numerous strains of wild-type mice and from > 6600 mutants. GXD now holds > 418,000 images and > 1.83 million expression results from classical types of expression data.

RNA-Seq and microarray expression experiments

Searchable index of RNA-Seq and microarray expression experiments

We recently extended GXD to capture information about RNA-Seq and microarray expression experiments. It remains

difficult to find specific expression experiments in the Gene Expression Omnibus (GEO) (Clough and Barrett 2016) and ArrayExpress (Athar et al. 2019) data archives based on biological parameters because sample annotations use free-form metadata provided by data submitters. Therefore, we generated a standardized, searchable index of RNA-seq expression and microarray experiments in ArrayExpress and GEO to help researchers quickly and reliably find data sets of interest (Smith et al. 2020a). We evaluated all mouse microarray and RNA-seq expression data sets in ArrayExpress (including thousands of experiments imported from GEO); identified those that reported endogenous gene expression in wild-type and mutant mice (in keeping with GXD's scope but including data from postnatal mice as well); and used ontologies, controlled vocabularies, and standard genetic nomenclature to annotate sample metadata (i.e., tissue, age, sex, strain, mutation carried) for each experiment. We also indicated experimental categories (baseline tissue study, wild-type versus mutant comparison) and experimental variables for these experiments. As of June 18, 2021, we have evaluated 16,983 mouse microarray or RNA-seq data sets (including bulk and single-cell RNA-seq data), identified 3199 data sets

◀**Fig. 3** Anatomical comparison of expression and phenotype data. **A** Comparing expression and phenotype data for a specific anatomical structure. The Mouse Developmental Anatomy Browser allows users to search for anatomical structures and to look up the expression and phenotype data associated with these structures. 'Endocrine pancreas' is shown as example. Using the Tree View (lower right), one can explore the anatomy by expanding and collapsing the hierarchy. The links indicated by arrows in the Tree View section lead to the expression and phenotype data associated with the selected anatomical structure and its substructures. The link (arrow) in the Term Detail section (upper right) leads to the Mammalian Phenotype (MP) browser, listing the MP terms mapped to the selected anatomical structure, and phenotype data associated with these MP terms. **B** Comparing expression and phenotype pattern for a specific gene. The Gene Expression + Phenotype Matrix displays the expression and phenotype data for a selected gene in the same anatomical matrix view. The gene *Ins2* is shown as an example, with the 'endocrine system' expanded along the anatomy axis. The wild-type expression pattern of *Ins2* is displayed in the first column (gold header), the following columns show the anatomical structures phenotypically affected in different *Ins2* mutant mice (different *Ins2* alleles). The coloring of the matrix cells gets progressively darker as the number of expression and phenotype annotations increases; the conventions are defined in the matrix legend (inset)

as relevant for GXD, and completed the metadata annotation for 3188 data sets.

The searchable index is available via the RNA-Seq and Microarray Experiment Search (http://www.informatics.jax.org/gxd/htexp_index). It is current with regard to experiments in ArrayExpress. As ArrayExpress no longer imports experiments from GEO, we are implementing additional procedures to obtain experiments directly from GEO. Our goal is to have a complete and non-redundant representation of RNA-Seq and microarray expression experiments from both repositories.

Curation and integration of RNA-seq expression data

The EBI Expression Atlas project selects high-quality RNA-seq data sets from ArrayExpress and GEO and uses a standardized processing pipeline, starting from the primary data, to generate consistently processed transcript per millions (TPM) values (Papatheodorou et al. 2020). We have imported these TPM level data from the Expression Atlas for those RNA-seq experiments that are within GXD's scope (as determined by our indexing work described above). We have processed these data further to enable their full integration into GXD. Taking advantage of our sample metadata annotations (described above), we identify the unique biological replicate sets for each experiment; determine the averaged quantile-normalized TPM value for each gene per biological replicate set; and assign each of these TPM values a Present/Absent call, using the TPM range bins employed by the Expression Atlas (Baldarelli et al. 2021). Because GXD uses Present/Absent annotations for classical expression assays, this last step permits the full integration of RNA-Seq data

with classical types of expression data in database searches, filters, and displays. To date, we have loaded expression data for 70 RNA-Seq experiments in the GXD metadata index. These include 1846 distinct samples that were condensed to 631 biological replicate sets, representing 88 distinct anatomical structures and 85 distinct mouse strains. Mouse genes represented in Expression Atlas RNA-Seq TPM files cover the entire Ensembl transcriptome (nearly 55,000 genome features), including protein-coding and non-coding RNA genes. Total RNA-Seq assay results amount to nearly 35 million, with comprehensive genome coverage for each experiment. While our literature curation effort focuses on mouse development, we have always accepted postnatal data from electronic submission. However, it is worth noting that the incorporation of RNA-Seq data has led to a large increase in postnatal expression data in GXD. As the RNA-Seq data are fully integrated with classical types of expression data, they are accessible through all the search, filtering, and display tools that GXD provides.
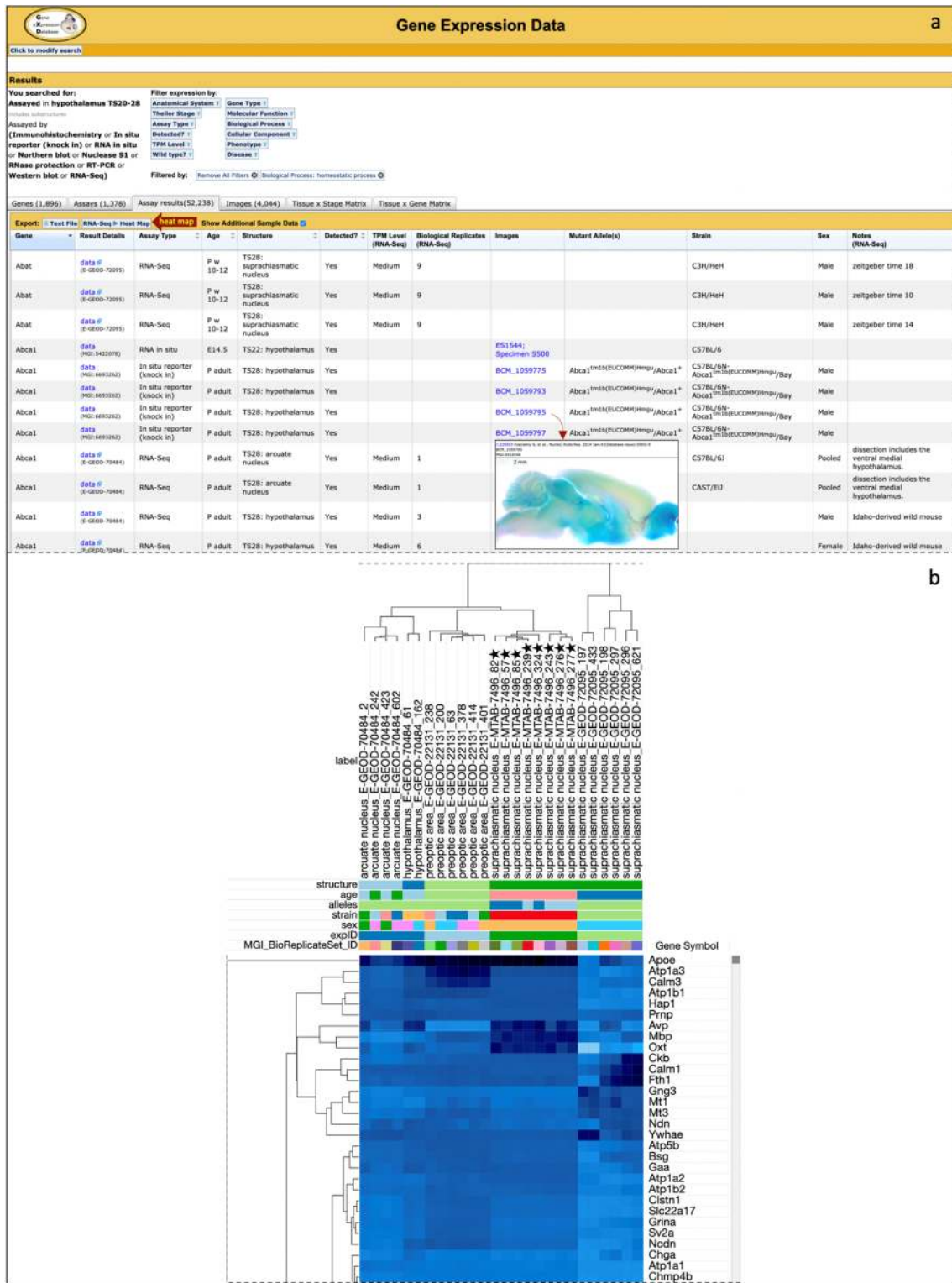
New interface features: searching and analyzing expression data in the larger context

The GXD Home Page (http://www.informatics.jax.org/expression.shtml) provides the best entry point to all the features and resources provided by GXD. Graphical tiles provide a quick overview of, and access to, GXD's search functions. For first time users, a one-page flow chart describes the GXD interface. A Highlights section alerts users of newly added features and data. Additional graphical links are provided to more information (About GXD), to guidelines for electronic data submissions (Fast Track Your Data), to current data content (GXD Statistics), and to Contact information. Tabbed fields at the bottom provide information about new features, curation policies, access to Help documentation, and Links to other resources.

GXD's interface utilities have been described previously (Finger et al. 2015, 2017). Here, we focus on search and display features that take advantage of the close integration of expression data with genetic, functional, and phenotypic information in MGD/MGI.

Expression data and image search

GXD's most versatile search form not only supports basic expression queries such as 'Where and when is a given gene expressed?' or 'What genes are expressed in a specific tissue?' but it also enables much more complex searches such as 'What genes located in a specified genomic interval are expressed in a specific tissue?,' thus supporting candidate gene queries, or facilitating insights into cis-regulation of gene expression. Further, one can search for expression data of sets of genes

Gene Expression Data a

Click to modify search

Results

You searched for:
Assayed in hypothalamus TS20-28
Includes substructures
Assayed by
(Immunohistochemistry or In situ reporter (knock in) or RNA in situ or Northern blot or Nuclease S1 or RNase protection or RT-PCR or Western blot or RNA-Seq)

Filter expression by:
Anatomical System ?   Gene Type ?
Theiler Stage ?   Molecular Function ?
Assay Type ?   Biological Process ?
Detected? ?   Cellular Component ?
TPM Level ?   Phenotype ?
Wild type? ?   Disease ?

Filtered by: Remove All Filters ⊗ Biological Process: homeostatic process ⊗

Genes (1,896)   Assays (1,378)   Assay results(52,238)   Images (4,044)   Tissue x Stage Matrix   Tissue x Gene Matrix

Export: Text File RNA-Seq Heat Map heat map Show Additional Sample Data

Gene	Result Details	Assay Type	Age	Structure	Detected?	TPM Level (RNA-Seq)	Biological Replicates (RNA-Seq)	Images	Mutant Allele(s)	Strain	Sex	Notes (RNA-Seq)
Abat	data (E-GEOD-72095)	RNA-Seq	P w 10-12	TS28: suprachiasmatic nucleus	Yes	Medium	9			C3H/HeH	Male	zeitgeber time 18
Abat	data (E-GEOD-72095)	RNA-Seq	P w 10-12	TS28: suprachiasmatic nucleus	Yes	Medium	9			C3H/HeH	Male	zeitgeber time 10
Abat	data (E-GEOD-72095)	RNA-Seq	P w 10-12	TS28: suprachiasmatic nucleus	Yes	Medium	9			C3H/HeH	Male	zeitgeber time 14
Abca1	data (MGI:5422078)	RNA in situ	E14.5	TS22: hypothalamus	Yes			ES1544; Specimen S500		C57BL/6		
Abca1	data (MGI:6693262)	In situ reporter (knock in)	P adult	TS28: hypothalamus	Yes			BCM_1059775	Abca1tm1b(EUCOMM)Hmgu/Abca1+	C57BL/6N-Abca1tm1b(EUCOMM)Hmgu/Bay	Male	
Abca1	data (MGI:6693262)	In situ reporter (knock in)	P adult	TS28: hypothalamus	Yes			BCM_1059793	Abca1tm1b(EUCOMM)Hmgu/Abca1+	C57BL/6N-Abca1tm1b(EUCOMM)Hmgu/Bay	Male	
Abca1	data (MGI:6693262)	In situ reporter (knock in)	P adult	TS28: hypothalamus	Yes			BCM_1059795	Abca1tm1b(EUCOMM)Hmgu/Abca1+	C57BL/6N-Abca1tm1b(EUCOMM)Hmgu/Bay	Male	
Abca1	data (MGI:6693262)	In situ reporter (knock in)	P adult	TS28: hypothalamus	Yes			BCM_1059797	Abca1tm1b(EUCOMM)Hmgu/Abca1+	C57BL/6N-Abca1tm1b(EUCOMM)Hmgu/Bay	Male	
Abca1	data (E-GEOD-70484)	RNA-Seq	P adult	TS28: arcuate nucleus	Yes	Medium	1			C57BL/6J	Pooled	dissection includes the ventral medial hypothalamus.
Abca1	data (E-GEOD-70484)	RNA-Seq	P adult	TS28: arcuate nucleus	Yes	Medium	1			CAST/EiJ	Pooled	dissection includes the ventral medial hypothalamus.
Abca1	data (E-GEOD-70484)	RNA-Seq	P adult	TS28: hypothalamus	Yes	Medium	3				Male	Idaho-derived wild mouse
Abca1	data (E-GEOD-70484)	RNA-Seq	P adult	TS28: hypothalamus	Yes	Medium	6				Female	Idaho-derived wild mouse

b

label:
arcuate nucleus_E-GEOD-70484_2
arcuate nucleus_E-GEOD-70484_242
arcuate nucleus_E-GEOD-70484_423
arcuate nucleus_E-GEOD-70484_602
hypothalamus_E-GEOD-70484_61
hypothalamus_E-GEOD-70484_162
preoptic area_E-GEOD-22131_238
preoptic area_E-GEOD-22131_200
preoptic area_E-GEOD-22131_63
preoptic area_E-GEOD-22131_378
preoptic area_E-GEOD-22131_414
preoptic area_E-GEOD-22131_401
suprachiasmatic nucleus_E-MTAB-7496_82★★
suprachiasmatic nucleus_E-MTAB-7496_57★
suprachiasmatic nucleus_E-MTAB-7496_85★★
suprachiasmatic nucleus_E-MTAB-7496_239★
suprachiasmatic nucleus_E-MTAB-7496_324★★
suprachiasmatic nucleus_E-MTAB-7496_243★
suprachiasmatic nucleus_E-MTAB-7496_276★
suprachiasmatic nucleus_E-MTAB-7496_277★
suprachiasmatic nucleus_E-GEOD-72095_197
suprachiasmatic nucleus_E-GEOD-72095_433
suprachiasmatic nucleus_E-GEOD-72095_198
suprachiasmatic nucleus_E-GEOD-72095_297
suprachiasmatic nucleus_E-GEOD-72095_296
suprachiasmatic nucleus_E-GEOD-72095_621

structure
age
alleles
strain
sex
expID
MGI_BioReplicateSet_ID

Gene Symbol:
Apoe
Atp1a3
Calm3
Atp1b1
Hap1
Prnp
Avp
Mbp
Oxt
Ckb
Calm1
Fth1
Gng3
Mt1
Mt3
Ndn
Ywhae
Atp5b
Bsg
Gaa
Atp1a2
Atp1b2
Clstn1
Slc22a17
Grina
Sv2a
Ncdn
Chga
Atp1a1
Chmp4b

defined by Function, Phenotype, or Disease association. For example, one can search for genes that are associated with 'spina bifida' and expressed in the 'neural tube'; or for genes that are involved in 'signal transduction' and expressed in the 'eye' of 'Pax6-mutant' mice. These

◄**Fig. 4** Hypothalamus expression results in GXD for genes involved in homeostasis. **A** Partial search result set shown from a search for expression data assayed in mouse hypothalamus and then filtered for genes annotated to GO Biological Process term: homeostatic process. Classical GXD expression data are integrated with RNA-Seq data via curated source metadata. Assay-level filters allow results to be narrowed by selected sample metadata fields (Anatomical System, Theiler Stage, Assay Type and whether mutant or wild-type mice were assayed). Gene-level filters allow results to be refined by qualitative (Detected?) and quantitative (TPM Level) expression. Gene-level filters also refine results to sets of genes annotated to biological systems of interest (Gene Ontology, Mammalian Phenotype Ontology, Disease Ontology). Expression results are organized into separate views/tabs. The default Assay Results view (shown truncated at the dotted line) displays sample-level metadata for each assayed gene (assay reference column not shown). For classical expression assay types, links in the images column lead to complete assay details including thumbnails of corresponding images, from which details for each image can be accessed (insert shows image detail from (Koscielny et al. 2014)). For RNA-Seq data, a row represents the consolidation of biological replicates, and TPM Level bins are shown for each gene, based on the average, quantile-normalized TPM values from corresponding biological replicate samples with the number of biological replicate samples provided. Clicking the RNA-Seq Heat Map button (red heat map arrow) renders a Morpheus RNA-Seq heat map of the query results. Additional views/tabs for the expression result set (not shown) feature search results organized by Genes, Assays, Images, or two matrix views of the data (Tissue x Stage matrix and Tissue x Gene matrix). Selected filters affect results on all display tabs. **B** RNA-Seq heat map of GXD results from the filtered query rendered with Morpheus (partial heat map shown). Columns represent distinct sample bioreplicate sets and are labeled by a combination of anatomical structure, experiment ID, and bioreplicate set ID (stars represent mutant samples). Metadata shared between samples are indicated by distinct colors in the metadata rows above the gene rows. Colored cells in gene rows reflect average, quantile-normalized TPM values for corresponding columns using a TPM value color scheme that accounts for an expansive dynamic TPM range (dark blue (high) to light blue (low), gray = below threshold, not present in the section of the heat map shown). Genes and samples were clustered using the Hierarchical Clustering feature of Morpheus (Euclidean distance, complete linking). Although the heat map is cropped to show only a few of the 1,893 genes in the result set, both gene-level and sample-level clustering are evident. Notably, the mutant samples (stars) cluster from the wild-type samples and are derived from mice with a mutation in the *Sox2* gene (*Sox2*^*tm1.1Lan*^, ArrayExpress: E-MTAB-7496, (Cheng et al. 2019)

diverse search parameters can be used to quickly find genes of specific research interest.

Developmental anatomy browser

This browser, used to search for anatomical structures and to look up the associated expression data, has been enhanced to display phenotype data associated with corresponding structures as well (Smith et al. 2019). Thus, it is now easy to compare the expression and phenotype data for specific anatomical structures (Fig. 3A).

Gene expression + phenotype comparison matrix

This interactive matrix views, accessible from the phenotype and expression ribbons on MGI gene detail pages, allows users to compare gene expression and phenotype data for a given gene (Fig. 3B). Based on our previously developed Tissue-by-Gene expression matrix views, the Gene Expression + Phenotype Comparison Matrix displays both expression and phenotype data in the mouse developmental anatomy framework, visually juxtaposing the tissues where a gene is normally expressed against tissues where mutations in that gene cause abnormalities. The anatomy axis of the view can be expanded and collapsed, allowing users to explore correlations between gene expression and phenotype at different levels of detail.

Expanded search summaries with enhanced filtering capabilities

As illustrated at www.informatics.jax.org/mgihome/GXD/FirstTimeUsers.shtml, all of GXD's expression data search forms lead to the same multi-tabbed displays that summarize data at different levels of detail: Genes, Assays, Assay Results, Images, and in the form of two different Matrix Views (Tissue × Stage and Tissue × Gene). With the incorporation of RNA-Seq data, search summaries have become much more voluminous. We have, therefore, added new data columns to these summaries, supporting new sorting options, and we have added many new filtering capabilities. New sortable data columns include information about TPM-Level (for RNA-Seq data), Strain, and Sex. New filters have been added to filter RNA-Seq expression results by TPM-Level and to filter gene sets based on their gene type (protein-coding, non-coding RNA, etc.); based on high-level Mammalian Phenotype (MP) and Disease Ontology (DO) categories (Bello et al. 2018; Smith and Eppig 2012); and according to their molecular functions, the biological processes they are involved in, and the cellular components in which they are found (high-level Gene Ontology (GO) categories (The Gene Ontology 2019). As illustrated in Fig. 4A, these new filters can be combined with each other and with the expression-related filters that we developed previously (anatomical system, developmental stage, assay type, detected (yes/no), etc.). This provides researchers with new and powerful means to quickly and efficiently narrow down and analyze expression data according to their specific interests, or as a more general exploratory tool.

Further analysis of RNA-Seq expression results using Morpheus

One can now employ GXD's search tools and filters to create customized RNA-Seq data sets (as described above) and, by

merely clicking a button, export these data into Morpheus (Fig. 4B). Morpheus, a versatile heat map visualization and analysis tool developed at the Broad Institute (https://software.broadinstitute.org/morpheus/), offers myriad utilities for further display and analysis, including sorting, filtering, hierarchical clustering, nearest neighbor analysis, and visual enrichment. The export function transmits GXD's sample annotations to Morpheus, where they can be readily used for sorting and filtering.

Summary

Virtually all advances in medicine rely on the use of animal models to generate, test, and evaluate hypotheses relevant to human biology. As the cost of genome-scale sequencing continues to decrease and new technologies for genome editing become widely adopted, the laboratory mouse is more important than ever as a model system for understanding the biological significance of human genetic variation and for advancing the emergence of genomic medicine. In this era of genome-scale, data-driven biomedical research, the knowledgebases that are part of the MGI environment play a pivotal role in standardizing, integrating, and disseminating information about the laboratory mouse as a model system for understanding human biology and disease processes. MGI's focus on annotation and nomenclature standards ensures the broadest possible impact of the resource in contributing to the development of new tools and hypotheses for advancing our understanding of the human genome and how it functions and for advancing genomic approaches to improving human health.

Data availability MGI is freely available and accessible from http://www.informatics.jax.org/. Proposed new terms, definitions, synonyms, and changes to the Mammalian Phenotype (MP) Ontology are tracked in GitHub (https://github.com/obophenotype/mammalian-phenotype-ontology/issues). MP ontology files, Adult Mouse Anatomy (MA), and Mouse Developmental Anatomy (EMAPA) ontology files are available in multiple formats from the OBO Foundry (www.obofoundry.org), Ontobee (www.ontobee.org), Ontology Lookup Service (OLS, www.obofoundry.org), and Bioportal (bioportal.bioontology.org). MP and MA ontology files are also available on the MGI reports download site (http://www.informatics.jax.org/downloads/reports/MPheno_OBO.ontology and http://www.informatics.jax.org/downloads/reports/adult_mouse_anatomy.obo).

Code availability All source code for the MGI system is available at https://github.com/mgijax.

Declarations

References

Alliance of Genome Resources C (2019) The Alliance of Genome Resources: building a modern data ecosystem for model organism databases. Genetics 213:1189–1196

Alliance of Genome Resources C (2020) Alliance of Genome Resources Portal: unified model organism research platform. Nucleic Acids Res 48:D650–D658

Alliance of Genome Resources Consortium T (2019) Alliance of Genome Resources Portal: unified model organism research platform. Nucleic Acids Res gkz813

Amos-Landgraf J, Franklin C, Godfrey V, Grieder F, Grimsrud K, Korf I, Lutz C, Magnuson T, Mirochnitchenko O, Patel S, Reinholdt L, Lloyd KCK (2021) The Mutant Mouse Resource and Research Center (MMRRC): the NIH-supported National Public Repository and Distribution Archive of Mutant Mouse Models in the USA. Mamm Genome

Athar A, Fullgrabe A, George N, Iqbal H, Huerta L, Ali A, Snow C, Fonseca NA, Petryszak R, Papatheodorou I, Sarkans U, Brazma A (2019) ArrayExpress update—from bulk to single-cell expression data. Nucleic Acids Res 47:D711–D715

Baker EJ, Jay JJ, Bubier JA, Langston MA, Chesler EJ (2012) GeneWeaver: a web-based system for integrative functional genomics. Nucleic Acids Res 40:D1067-1076

Baldarelli RM, Smith CM, Finger JH, Hayamizu TF, McCright IJ, Xu J, Shaw DR, Beal JS, Blodgett O, Campbell J, Corbani LE, Frost PJ, Giannatto SC, Miers DB, Kadin JA, Richardson JE, Ringwald M (2021) The mouse Gene Expression Database (GXD): 2021 update. Nucleic Acids Res 49:D924–D931

Bandrowski A, Brush M, Grethe JS, Haendel MA, Kennedy DN, Hill S, Hof PR, Martone ME, Pols M, Tan S, Washington N, Zudilova-Seinstra E, Vasilevsky N, Resource Identification Initiative Members are listed here hwfonm (2015) The Resource Identification Initiative: a cultural shift in publishing. F1000Res 4:134

Bello SM, Shimoyama M, Mitraka E, Laulederkind SJF, Smith CL, Eppig JT, Schriml LM (2018) Disease Ontology: improving and unifying disease annotations across species. Dis Model Mech 11:dmm032839

Blake JA, Baldarelli R, Kadin JA, Richardson JE, Smith CL, Bult CJ, Mouse Genome Database G (2021) Mouse Genome Database (MGD): knowledgebase for mouse-human comparative biology. Nucleic Acids Res 49:D981–D987

Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, Tolstoy I, Tatusova T, Pruitt KD, Maglott DR, Murphy TD (2015) Gene: a gene-centered information resource at NCBI. Nucleic Acids Res 43:D36-42

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, Ami GOH, Web Presence Working G (2009) AmiGO: online access to ontology and annotation data. Bioinformatics 25:288–289

Cheng AH, Bouchard-Cannon P, Hegazi S, Lowden C, Fung SW, Chiang CK, Ness RW, Cheng HM (2019) SOX2-dependent transcription in clock neurons promotes the robustness of the central circadian pacemaker. Cell Rep 26:3191–3202

Clough E, Barrett T (2016) The gene expression omnibus database. Methods Mol Biol 1418:93–110

Coordinators NR (2016) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 44:D7-19

den Dunnen JT (2017) Describing sequence variants using HGVS nomenclature. Methods Mol Biol 1492:243–251

Diez-Roux G, Banfi S, Sultan M, Geffers L, Anand S, Rozado D, Magen A, Canidio E, Pagani M, Peluso I, Lin-Marq N, Koch M, Bilio M, Cantiello I, Verde R, De Masi C, Bianchi SA, Cicchini J, Perroud E, Mehmeti S, Dagand E, Schrinner S, Nurnberger A, Schmidt K, Metz K, Zwingmann C, Brieske N, Springer C, Hernandez AM, Herzog S, Grabbe F, Sieverding C, Fischer B, Schrader K, Brockmeyer M, Dettmer S, Helbig C, Alunni V, Battaini MA, Mura C, Henrichsen CN, Garcia-Lopez R, Echevarria D, Puelles E, Garcia-Calero E, Kruse S, Uhr M, Kauck C, Feng G, Milyaev N, Ong CK, Kumar L, Lam M, Semple CA, Gyenesei A, Mundlos S, Radelof U, Lehrach H, Sarmientos P, Reymond A, Davidson DR, Dolle P, Antonarakis SE, Yaspo ML, Martinez S, Baldock RA, Eichele G, Ballabio A (2011) A high-resolution anatomical atlas of the transcriptome in the mouse embryo. PLoS Biol 9:e1000582

Dimitrakopoulos GN, Klapa MI, Moschonas NK (2020) PICKLE 3.0: Enriching the human Meta-database with the mouse protein interactome extended via mouse-human orthology. Bioinformatics

Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, Mouse Genome Database G (2015a) The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. Nucleic Acids Res 43:D726-736

Eppig JT, Motenko H, Richardson JE, Richards-Smith B, Smith CL (2015b) The International Mouse Strain Resource (IMSR): cataloging worldwide mouse and ES cell line resources. Mamm Genome 26:448–455

Eyre TA, Wright MW, Lush MJ, Bruford EA (2007) HCOP: a searchable database of human orthology predictions. Brief Bioinform 8:2–5

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 53:510–522

Finger JH, Smith CM, Hayamizu TF, McCright IJ, Xu J, Law M, Shaw DR, Baldarelli RM, Beal JS, Blodgett O, Campbell JW, Corbani LE, Lewis JR, Forthofer KL, Frost PJ, Giannatto SC, Hutchins LN, Miers DB, Motenko H, Stone KR, Eppig JT, Kadin JA, Richardson JE, Ringwald M (2017) The mouse Gene Expression Database (GXD): 2017 update. Nucleic Acids Res 45:D730–D736

Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, Sisu C, Wright JC, Armstrong J, Barnes I, Berry A, Bignell A, Boix C, Carbonell Sala S, Cunningham F, Di Domenico T, Donaldson S, Fiddes IT, Garcia Giron C, Gonzalez JM, Grego T, Hardy M, Hourlier T, Howe KL, Hunt T, Izuogu OG, Johnson R, Martin FJ, Martinez L, Mohanan S, Muir P, Navarro FCP, Parker A, Pei B, Pozo F, Riera FC, Ruffier M, Schmitt BM, Stapleton E, Suner MM, Sycheva I, Uszczynska-Ratajczak B, Wolf MY, Xu J, Yang YT, Yates A, Zerbino D, Zhang Y, Choudhary JS, Gerstein M, Guigo R, Hubbard TJP, Kellis M, Paten B, Tress ML, Flicek P (2021) Gencode 2021. Nucleic Acids Res 49:D916–D923

Gene Ontology C (2015) Gene Ontology Consortium: going forward. Nucleic Acids Res 43:D1049-1056

Goya J, Wong AK, Yao V, Krishnan A, Homilius M, Troyanskaya OG (2015) FNTM: a server for predicting functional networks of tissues in mouse. Nucleic Acids Res 43:W182-187

Gurumayum S, Jiang P, Hao X, Campos TL, Young ND, Korhonen PK, Gasser RB, Bork P, Zhao XM, He LJ, Chen WH (2021) OGEE v3: Online GEne Essentiality database with increased coverage of organisms and human cell lines. Nucleic Acids Res 49:D998–D1003

Hagn M, Marschall S, Hrabe de Angelis M (2007) EMMA—the European mouse mutant archive. Brief Funct Genomic Proteomic 6:186–192

Hamosh A, Amberger JS, Bocchini C, Scott AF, Rasmussen SA (2021) Online Mendelian Inheritance in Man (OMIM(R)): Victor McKusick's magnum opus. Am J Med Genet A

Harding SD, Armit C, Armstrong J, Brennan J, Cheng Y, Haggarty B, Houghton D, Lloyd-MacGilp S, Pi X, Roochun Y, Sharghi M, Tindal C, McMahon AP, Gottesman B, Little MH, Georgas K, Aronow BJ, Potter SS, Brunskill EW, Southard-Smith EM, Mendelsohn C, Baldock RA, Davies JA, Davidson D (2011) The GUDMAP database—an online resource for genitourinary research. Development 138:2845–2853

Hayamizu TF, Baldock RA, Ringwald M (2015) Mouse anatomy ontologies: enhancements and tools for exploring and integrating biomedical data. Mamm Genome 26:422–430

Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, Mohr SE (2011) An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinform 12:357

Jager M, Wang K, Bauer S, Smedley D, Krawitz P, Robinson PN (2014) Jannovar: a java library for exome annotation. Hum Mutat 35:548–555

Kohler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine JP, Gargano M, Harris NL, Matentzoglu N, McMurry JA, Osumi-Sutherland D, Cipriani V, Balhoff JP, Conlin T, Blau H, Baynam G, Palmer R, Gratian D, Dawkins H, Segal M, Jansen AC, Muaz A, Chang WH, Bergerson J, Laulederkind SJF, Yuksel Z, Beltran S, Freeman AF, Sergouniotis PI, Durkin D, Storm AL, Hanauer M, Brudno M, Bello SM, Sincan M, Rageth K, Wheeler MT, Oegema R, Lourghi H, Della Rocca MG, Thompson R, Castellanos F, Priest J, Cunningham-Rundles C, Hegde A, Lovering RC, Hajek C, Olry A, Notarangelo L, Similuk M, Zhang XA, Gomez-Andres D, Lochmuller H, Dollfus H, Rosenzweig S, Marwaha S, Rath A, Sullivan K, Smith C, Milner JD, Leroux D, Boerkoel CF, Klion A, Carter MC, Groza T, Smedley D, Haendel MA, Mungall C, Robinson PN (2019) Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res 47:D1018–D1027

Kolishovski G, Lamoureux A, Hale P, Richardson JE, Recla JM, Adesanya O, Simons A, Kunde-Ramamoorthy G, Bult CJ (2019) The JAX Synteny Browser for mouse-human comparative genomics. Mamm Genome 30:353–361

Koscielny G, Yaikhom G, Iyer V, Meehan TF, Morgan H, Atienza-Herrero J, Blake A, Chen CK, Easty R, Di Fenza A, Fiegel T, Griffiths 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, consortium I (2014) The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. Nucleic Acids Res 42:D802-809

Krupke DM, Begley DA, Sundberg JP, Richardson JE, Neuhauser SB, Bult CJ (2017) The Mouse Tumor Biology Database: a comprehensive resource for mouse models of human cancer. Cancer Res 77:e67–e70

Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pedersen JS, Hsu F, Hinrichs AS, Harte RA, Diekhans M, Clawson H, Bejerano G, Barber GP, Baertsch R, Haussler D, Kent WJ (2007) The UCSC genome browser database: update 2007. Nucleic Acids Res 35:D668-673

Lilue J, Doran AG, Fiddes IT, Abrudan M, Armstrong J, Bennett R, Chow W, Collins J, Collins S, Czechanski A, Danecek P, Diekhans M, Dolle DD, Dunn M, Durbin R, Earl D, Ferguson-Smith A, Flicek P, Flint J, Frankish A, Fu B, Gerstein M, Gilbert J, Goodstadt L, Harrow J, Howe K, Ibarra-Soria X, Kolmogorov M, Lelliott CJ, Logan DW, Loveland J, Mathews CE, Mott R, Muir P, Nachtweide S, Navarro FCP, Odom DT, Park N, Pelan S, Pham SK, Quail M, Reinholdt L, Romoth L, Shirley L, Sisu C, Sjoberg-Herrera M, Stanke M, Steward C, Thomas M, Threadgold G, Thybert D, Torrance J, Wong K, Wood J, Yalcin B, Yang F, Adams DJ, Paten B, Keane TM (2018) Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. Nat Genet 50:1574–1583

Magdaleno S, Jensen P, Brumwell CL, Seal A, Lehman K, Asbury A, Cheung T, Cornelius T, Batten DM, Eden C, Norland SM, Rice DS, Dosooye N, Shakya S, Mehta P, Curran T (2006) BGEM: an in situ hybridization database of gene expression in the embryonic and adult mouse nervous system. PLoS Biol 4:e86

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F (2016) The ensembl variant effect predictor. Genome Biol 17:122

Motenko H, Neuhauser SB, O'Keefe M, Richardson JE (2015) MouseMine: a new data warehouse for MGI. Mamm Genome 26(7-8):325–330. https://doi.org/10.1007/s00335-015-9573-z

Mungall CJ, McMurry JA, Kohler S, Balhoff JP, Borromeo C, Brush M, Carbon S, Conlin T, Dunn N, Engelstad M, Foster E, Gourdine JP, Jacobsen JO, Keith D, Laraway B, Lewis SE, NguyenXuan J, Shefchek K, Vasilevsky N, Yuan Z, Washington N, Hochheiser H, Groza T, Smedley D, Robinson PN, Haendel MA (2017) The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res 45:D712–D722

Munoz-Fuentes V, Cacheiro P, Meehan TF, Aguilar-Pimentel JA, Brown SDM, Flenniken AM, Flicek P, Galli A, Mashhadi HH, Hrabe de Angelis M, Kim JK, Lloyd KCK, McKerlie C, Morgan H, Murray SA, Nutter LMJ, Reilly PT, Seavitt JR, Seong JK, Simon M, Wardle-Jones H, Mallon AM, Smedley D, Parkinson HE (2018) The International Mouse Phenotyping Consortium (IMPC): a functional catalogue of the mammalian genome that informs conservation. Conserv Genet 19:995–1005

Newman V, Moore B, Sparrow H, Perry E (2018) The Ensembl Genome Browser: strategies for accessing eukaryotic genome data. Methods Mol Biol 1757:115–139

Papatheodorou I, Moreno P, Manning J, Fuentes AM, George N, Fexova S, Fonseca NA, Fullgrabe A, Green M, Huang N, Huerta L, Iqbal H, Jianu M, Mohammed S, Zhao L, Jarnuczak AF, Jupp S, Marioni J, Meyer K, Petryszak R, Prada Medina CA, Talavera-Lopez C, Teichmann S, Vizcaino JA, Brazma A (2020) Expression Atlas update: from tissues to single cells. Nucleic Acids Res 48:D77–D83

Perry MN, Smith CM, Onda H, Ringwald M, Murray SA, Smith CL (2021) Annotated expression and activity data for murine recombinase alleles and transgenes: the CrePortal resource. Mamm Genome

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

Richardson JE, Baldarelli RM, Bult CJ (2021) Multiple genome viewer (MGV): a new tool for visualization and comparison of multiple annotated genomes. Mamm Genome

Ringwald M, Mangan ME, Eppig JT, Kadin JA, Richardson JE (1999) GXD: a gene expression database for the laboratory mouse. The Gene Expression Database Group. Nucleic Acids Res 27:106–112

Ringwald M, Wu C, Su AI (2012) BioGPS and GXD: mouse gene expression data-the benefits and challenges of data integration. Mamm Genome 23:550–558

Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, Felix V, Jeng L, Bearer C, Lichenstein R, Bisordi K, Campion N, Hyman B, Kurland D, Oates CP, Kibbey S, Sreekumar P, Le C, Giglio M, Greene C (2019) Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res 47:D955–D962

Shefchek KA, Harris NL, Gargano M, Matentzoglu N, Unni D, Brush M, Keith D, Conlin T, Vasilevsky N, Zhang XA, Balhoff JP, Babb L, Bello SM, Blau H, Bradford Y, Carbon S, Carmody L, Chan LE, Cipriani V, Cuzick A, Della Rocca M, Dunn N, Essaid S, Fey P, Grove C, Gourdine JP, Hamosh A, Harris M, Helbig I, Hoatlin M, Joachimiak M, Jupp S, Lett KB, Lewis SE, McNamara C, Pendlington ZM, Pilgrim C, Putman T, Ravanmehr V, Reese J, Riggs E, Robb S, Roncaglia P, Seager J, Segerdell E, Similuk M, Storm AL, Thaxon C, Thessen A, Jacobsen JOB, McMurry JA, Groza T, Kohler S, Smedley D, Robinson PN, Mungall CJ, Haendel MA, Munoz-Torres MC, Osumi-Sutherland D (2020) The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res 48:D704–D715

Smedley D, Oellrich A, Kohler S, Ruef B, Sanger Mouse Genetics P, Westerfield M, Robinson P, Lewis S, Mungall C (2013) PhenoDigm: analyzing curated annotations to associate animal models with human diseases. Database 2013:bat25

Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C (2005) Relations in biomedical ontologies. Genome Biol 6:R46

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 RN, Aleksic J, Butano D, Carr A, Contrino S, Hu F, Lyne M, Lyne R, Kalderimis A, Rutherford K, Stepan R, Sullivan J, Wakeling M, Watkins X, Micklem G (2012) InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics 28:3163–3165

Smith CM, Finger JH, Hayamizu TF, McCright IJ, Xu J, Eppig JT, Kadin JA, Richardson JE, Ringwald M (2015) GXD: a community resource of mouse Gene Expression Data. Mamm Genome 26:314–324

Smith CM, Hayamizu TF, Finger JH, Bello SM, McCright IJ, Xu J, Baldarelli RM, Beal JS, Campbell J, Corbani LE, Frost PJ, Lewis JR, Giannatto SC, Miers D, Shaw DR, Kadin JA, Richardson JE, Smith CL, Ringwald M (2019) The mouse Gene Expression Database (GXD): 2019 update. Nucleic Acids Res 47:D774–D779

Smith CM, Kadin JA, Baldarelli RM, Beal JS, Blodgett O, Giannatto SC, Richardson JE, Ringwald M (2020a) GXD's RNA-Seq and Microarray Experiment Search: using curated metadata to reliably find mouse expression studies of interest. Database 2020

Smith JR, Hayman GT, Wang SJ, Laulederkind SJF, Hoffman MJ, Kaldunski ML, Tutaj M, Thota J, Nalabolu HS, Ellanki SLR, Tutaj MA, De Pons JL, Kwitek AE, Dwinell MR, Shimoyama ME (2020b) The Year of the Rat: The Rat Genome Database at 20: a multi-species knowledgebase and analysis platform. Nucleic Acids Res 48:D731–D742

The Gene Ontology C (2019) The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res 47:D330–D338

Thomas PD, Hill DP, Mi H, Osumi-Sutherland D, Van Auken K, Carbon S, Balhoff JP, Albou LP, Good B, Gaudet P, Lewis SE, Mungall CJ (2019) Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. Nat Genet 51:1429–1433

Thybert D, Roller M, Navarro FCP, Fiddes I, Streeter I, Feig C, Martin-Galvez D, Kolmogorov M, Janousek V, Akanni W, Aken B, Aldridge S, Chakrapani V, Chow W, Clarke L, Cummins C, Doran A, Dunn M, Goodstadt L, Howe K, Howell M, Josselin AA, Karn RC, Laukaitis CM, Jingtao L, Martin F, Muffato M, Nachtweide S, Quail MA, Sisu C, Stanke M, Stefflova K, Van Oosterhout C, Veyrunes F, Ward B, Yang F, Yazdanifar G, Zadissa A, Adams DJ, Brazma A, Gerstein M, Paten B, Pham S, Keane TM, Odom DT, Flicek P (2018) Repeat associated mechanisms of genome evolution and function revealed by the Mus caroli and Mus pahari genomes. Genome Res 28:448–459

Visel A, Thaller C, Eichele G (2004) GenePaint.org: an atlas of gene expression patterns in the mouse embryo. Nucleic Acids Res 32:D552-556

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, t Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3:160018

Yates B, Gray KA, Jones TEM, Bruford EA (2021) Updates to HCOP: the HGNC comparison of orthology predictions tool. Brief Bioinform

York WS, Mazumder R, Ranzinger R, Edwards N, Kahsay R, Aoki-Kinoshita KF, Campbell MP, Cummings RD, Feizi T, Martin M, Natale DA, Packer NH, Woods RJ, Agarwal G, Arpinar S, Bhat S, Blake J, Castro LJG, Fochtman B, Gildersleeve J, Goldman R, Holmes X, Jain V, Kulkarni S, Mahadik R, Mehta A, Mousavi R, Nakarakommula S, Navelkar R, Pattabiraman N, Pierce MJ, Ross K, Vasudev P, Vora J, Williamson T, Zhang W (2020) GlyGen: computational and informatics resources for glycoscience. Glycobiology 30:72–73

Zhu Y, Richardson JE, Hale P, Baldarelli RM, Reed DJ, Recla JM, Sinclair R, Reddy TB, Bult CJ (2015) A unified gene catalog for the laboratory mouse reference genome. Mamm Genome 26:295–304