# Using Bio-ontologies as Data Annotation, Integration & Analytical Tools at the Mouse Genome Informatics Resource

Anna V. Anagnostopoulos, Judith A. Blake, Carol J. Bult, Martin Ringwald, Joel E. Richardson,
James A. Kadin, and Janan T. Eppig

*Abstract*—**The Mouse Genome Informatics (MGI) resource, an in-depth resource for the genetics, genomics and biology of the laboratory mouse, provides free access to integrated data on diverse biological attributes, ranging from sequence to phenotype and disease model representation. MGI advances translational research through an integrated data platform that facilitates acquisition, explicit representation, and semantic querying and interpretation of multi-parametric genome-scale datasets, and fosters interoperability across different model organism systems and disparate data sources. To this end, MGI employs a set of logically rigorous, dynamic, and cross-referenceable ontologies to unambiguously describe current biological knowledge, expedite manual curation, and advance the informatics capacity to execute complex data mining tasks relevant to comparative and functional genomics. Major bio-ontologies developed and implemented at MGI include the Gene Ontology (GO), Mammalian Phenotype (MP) Ontology, and Adult Mouse Anatomical (MA) Dictionary, reviewed in this paper. All these share a common generic vocabulary infrastructure, and utilize identical annotation tools and web-based browsers to reinforce ontology-centric curation and support ontology-driven querying of the vocabularies and the associated knowledgebase.**

## I. Introduction

Ongoing transformation of biological science into an intensely data-driven discipline has led to unprecedented volumes of biological data buried in unstructured text and the attendant accrual of thousands of autonomous and semantically heterogeneous data sources. Spurred on by a growing need for semantic data integration and for database interoperability, biomedical ontologies are increasingly used as shared fundamental knowledge representations that confer semantic standards for annotating and indexing complex biological phenomena, and provide the basis for implementing functional classification and interpretation models. Beyond dictionaries or thesauri, bio-ontologies are controlled, structured vocabularies that formally represent relationships between well-defined bio-specific terms such that this information is human-digestible as well as navigable, parseable and translatable by computer-based systems [1].

This paper summarizes the development and use of bio-ontologies in the semantic annotation, integration, retrieval, and computational analysis of gene function, phenotype, and anatomy-based gene expression information stored in MGI.

## II. Using Bio-ontologies at MGI

The international MGI information system [2], hosted at The Jackson Laboratory, represents a consortium of several bioinformatics research programs working in concert to build a comprehensive, integrated model organism database (MOD) [3]. At the core of the MGI is the Mouse Genome Database (MGD) which provides free integrated access to in-depth genetic, genomic and biological data for the laboratory mouse, and serves as the authoritative source for official mouse gene, allele and strain nomenclature, as well as the international hub for standardized mouse phenotype and disease model representation. Major projects contributing to MGI include the Gene Ontology (GO) project [4], and the Gene Expression Database (GXD) [5], [6], among others. MGI biological data are integrated from multiple sources, ranging from international resource centers to individual investigator laboratories and the biomedical literature, using both automated processes and expert human curation. Data are updated daily, and data access is enabled via dynamically generated web pages, text files available via FTP, and through direct SQL [3].

MGI's primary objectives are to facilitate the use of the mouse as a premier mammalian surrogate for modeling normal development and disease processes in human, and advance the informatics capacity to pose genome-scale and systems-level research questions that accelerate knowledge discovery. To these ends, MGI maintains extensive collaborations with mouse mutagenesis centers, genome sequencing consortia, and other organism-specific or specialized resources. In addition, MGI cooperates with the global bioinformatics community on the development of referential and semantic standards, and applies a variety of standardized nomenclatures and controlled/structured vocabularies to optimize knowledge representation, expedite manual curation, and support complex data mining and analytical or predictive tasks relevant to comparative and

Manuscript received June 15, 2008. MGD and GXD are two components of the MGI database system. MGD is supported by the National Human Genome Research Institute (NHGRI/NIH) under grant HG000330. GXD is supported by the National Institute of Child Health and Human Development (NICHD/NIH) under grant HD033745. The GO project is supported by NHGRI/NIH under grant HG002273.

A. V. Anagnostopoulos is with the MGI resource at The Jackson Laboratory (TJL), Bar Harbor, ME 04609-1500 USA (phone: 207-288-6433; fax: 207-288-6132; e-mail: anna@informatics.jax.org).

J. A. Blake is with MGI at TJL (e-mail: jblake@informatics.jax.org).

C. J. Bult is with MGI at TJL (e-mail: cjb@informatics.jax.org).

M. Ringwald is with MGI at TJL (e-mail: ringwald@informatics.jax.org).

J. E. Richardson is with MGI at TJL (e-mail: jer@informatics.jax.org).

J. A. Kadin is with MGI at TJL (e-mail: jak@informatics.jax.org).

J. T. Eppig is with MGI at TJL (e-mail: jte@informatics.jax.org).

functional genomics. MGI's adherence to semantic standards is a crucial step towards data integration and database interoperability across different model organism systems and disparate data sources or platforms. Moreover, MGI's application of ontology-driven tools fosters new routes to examine gene expression profiles, to map functional features of gene products to complex pathophysiological states, and to establish associations between observed mouse phenotypes and orthologous human gene mutations or distinct nosological entities for which defined mouse genotypes phenomimic the human condition.

Standardized nomenclatures and simple, controlled vocabularies incorporated into the MGI annotation system include strain, gene, and gene product names, allele types, mutation categories, assay types, and developmental stages. Key bio-ontologies developed and used at MGI include the GO, the MP Ontology (MPO), and the Adult MA Dictionary, reviewed here. All three structured vocabularies are accessible from the Vocabularies section of the Search menu at the MGI Home Page [2], and are tightly integrated into relevant data-specific web-based query forms available at MGI [7]–[9].

Importantly, all MGI bio-ontologies have been built as directed acyclic graphs (DAGs) utilizing the versatile OBO-edit Java tool [10], [11] for construction, maintenance, and editing operations. Moreover, MGI has been designed with a common generic vocabulary infrastructure, such that all ontologies implemented in the database system share a common data model, use identical annotation tools, and employ equivalent web browsers to efficiently navigate, query, analyze, and compare the biological knowledge at hand (Fig. 1). The browsing and search capabilities of each ontology web browser will be detailed in the sections below.

### *A. The GO Project at MGI*

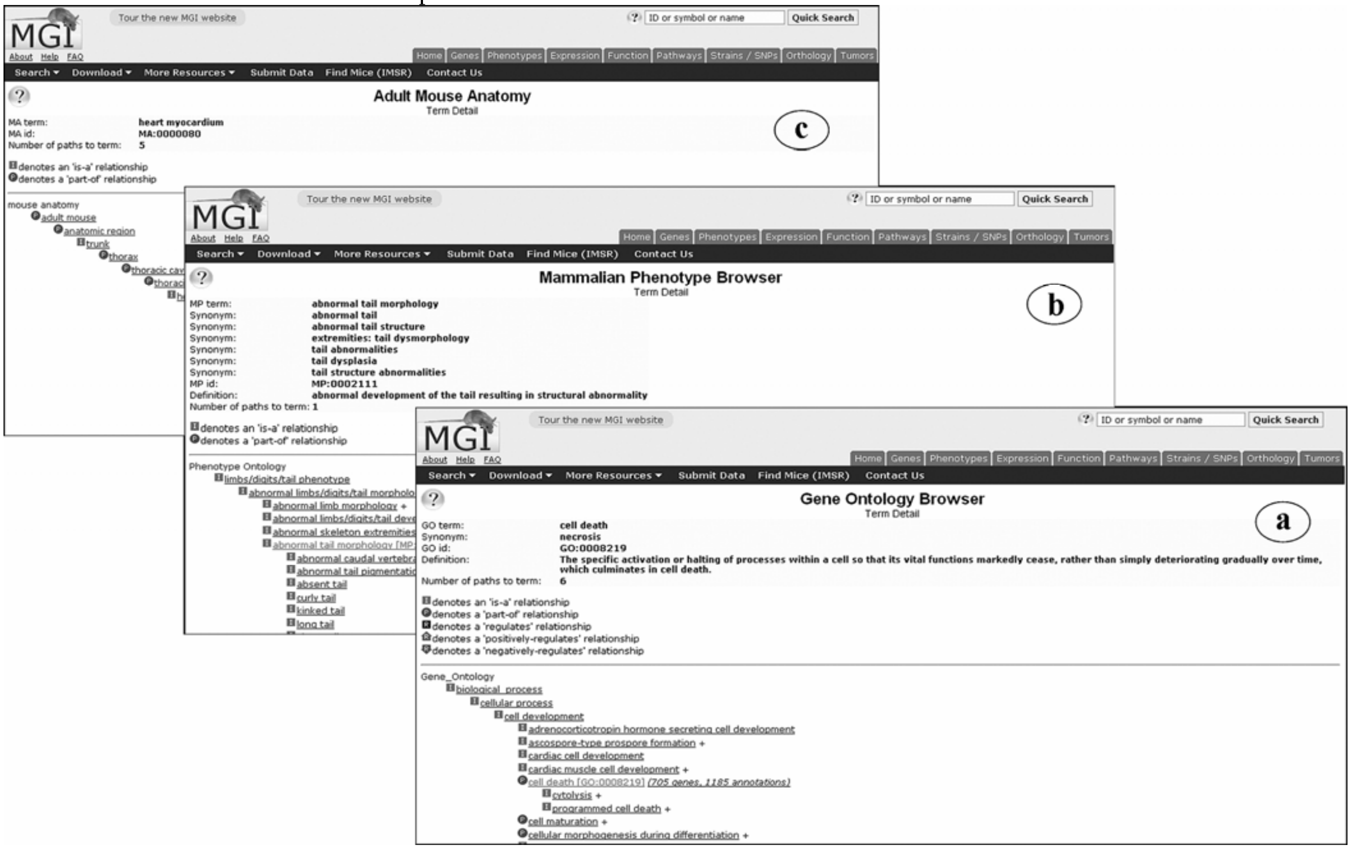
The GO project [12] is a community effort to address the need for consistent descriptions of gene and gene product attributes in different MODs. Specifically, the prototypic GO tripartite vocabulary system represents a widely adopted canonical ontology used to annotate gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. Since its inception in 1998, the GO Consortium (GOC) [13] has grown remarkably to include numerous animal, plant, and microbial genome databases, as well as cross-community resources such as UniProt and InterPro.

As a founding member of the GOC, MGI is an active participant in the development and application of the GO, serving as the authoritative source of precise GO functional annotations to mouse genes and gene products available in MGI [4]. Each MGI-GO functional annotation represents curated or inferred data and includes mandatory information such as the mouse gene identifier (ID), name and symbol, associated GO term, evidence code, and reference(s) supporting this annotation. Gene product annotation records may include qualifiers, such as 'NOT', 'contributes_to', and 'colocalizes_with', that modify the interpretation of an annotation. Briefly, MGI curators review the published scientific literature and associate genes with GO terms and evidence codes, i.e., three-letter designations representing the type of evidence that supports the GO term to gene product association. Experimentally-based GO annotations use a specific set of experimental evidence codes provided at [4]. Inferred annotations are applied by translations of UniProt keywords, InterPro domains, and Enzyme Commission (EC) numbers to GO terms, and primarily use the IEA code (Inferred from Electronic Annotation). Within MGI, GO classifications are accessible in a standard tabular format and as computer generated text paragraphs. Recently added MGI functionality includes graphical displays of GO annotations to individual mouse genes [14], and comparative graphical views of GO annotations to curated mouse-human-rat ortholog sets [15].

Each of the three orthogonal GO sub-ontologies (Biological Process, Cellular Component and Molecular Function) is independently organized as a DAG, a hierarchical tree structure allowing multiple parentage both along *is-a* and *part-of* transitive relationships propagated from more specialized (child) terms to less specialized (parent) terms. In addition, the GOC has recently introduced three new relationship types (*regulates*, *negatively_regulates*, and *positively_regulates*) into the Biological Process ontology [12]. Structuring the GO vocabulary terms as a DAG hierarchy allows both attribution and querying at varying levels of detail. Thus, depending on the experimental evidence for the cellular location of a gene product, a gene may be annotated to the 'nucleus', or to the more specialized term 'nucleolus' in the Cellular Component ontology. Annotations made to either of these terms relate to one another because the 'nucleolus' is *part-of* the 'nucleus'.

Each GO term (node) must conform to the 'true path' rule stating that the pathway from a child term all the way up to its top-level parent(s) must always be true. This rule applies to both *is-a* relationships (i.e., if 'anion channel activity' *is-a* subclass of 'transporter activity', it also *is-a* 'molecular function')**,** and *part-of* relationships (i.e., if 'laminin complex' is *part-of* the 'basal lamina' which, in turn, *is-part* of the 'basement membrane', then 'laminin complex' is itself *part-of* the 'basement membrane'). Moreover, each GO term is assigned a unique numerical identifier (GO:nnnnnnn), a definition, and, as appropriate, one or more synonyms. While unique IDs ensure the referential integrity of the terms, textual definitions support an explicit shared comprehension of the terms for annotation purposes among all organisms. The incorporation of synonyms provides a mechanism for mapping identical concepts with alternate labels, abbreviations and acronyms to a single term within a given ontology.

All three GO vocabularies evolve and expand dynamically to reflect accumulating and changing biological knowledge. In addition to creating new terms, MGI curators play an active role in augmenting, refining, and reorganizing existing terms and relationships, often in consultation with experts in specific domain areas. Recent MGI contributions to the GO include the collaborative implementation of an enhanced representation of immunological content [16], and ongoing revisions of the hierarchical extensions for ‘blood pressure regulation’ and ‘muscle development’ in the Biological Process ontology.


MGI
Tour the new MGI website
ID or symbol or name
Quick Search
About Help FAQ
Home Genes Phenotypes Expression Function Pathways Strains / SNPs Orthology Tumors
Search ▾ Download ▾ More Resources ▾ Submit Data Find Mice (IMSR) Contact Us
?
c
b
a

Adult Mouse Anatomy
Term Detail

MA term: heart myocardium
MA id: MA:0000080
Number of paths to term: 5

denotes an 'is-a' relationship
denotes a 'part-of' relationship

mouse anatomy
adult mouse
anatomic region
trunk
thorax
thoracic cav
thorac
h

Mammalian Phenotype Browser
Term Detail

MP term: abnormal tail morphology
Synonym: abnormal tail
Synonym: abnormal tail structure
Synonym: extremities: tail dysmorphology
Synonym: tail abnormalities
Synonym: tail dysplasia
Synonym: tail structure abnormalities
MP id: MP:0002111
Definition: abnormal development of the tail resulting in structural abnormality
Number of paths to term: 1

denotes an 'is-a' relationship
denotes a 'part-of' relationship

Phenotype Ontology
limbs/digits/tail phenotype
abnormal limbs/digits/tail morpholo
abnormal limb morphology +
abnormal limbs/digits/tail deve
abnormal skeleton extremities
abnormal tail morphology [MP
abnormal caudal vertebra
abnormal tail pigmentatio
absent tail
curly tail
kinked tail
long tail

Gene Ontology Browser
Term Detail

GO term: cell death
Synonym: necrosis
GO id: GO:0008219
Definition: The specific activation or halting of processes within a cell so that its vital functions markedly cease, rather than simply deteriorating gradually over time, which culminates in cell death.
Number of paths to term: 6

denotes an 'is-a' relationship
denotes a 'part-of' relationship
denotes a 'regulates' relationship
denotes a 'positively-regulates' relationship
denotes a 'negatively-regulates' relationship

Gene_Ontology
biological process
cellular process
cell development
adrenocorticotropin hormone secreting cell development
ascospore-type prospore formation +
cardiac cell development
cardiac muscle cell development +
cell death [GO:0008219] *(705 genes, 1185 annotations)*
cytolysis +
programmed cell death +
cell maturation +
cellular morphogenesis during differentiation +

Fig. 1. Partial screenshots of ‘Term Detail’ pages providing individual term information as displayed in the Gene Ontology (a), Mammalian Phenotype (b), and Adult Mouse Anatomy (c) Vocabulary Browsers available at MGI. ‘Term Detail’ pages include the unique numerical identifier for each term, in addition to relevant definitions and/or synonym(s), term-term relationships, and the number of hierarchical paths that lead to the selected term. A plus sign following a term indicates that this term has ‘descendants’, which can be viewed by clicking on the term to expand the relevant portion of the ontology. Only one hierarchical path to the selected term is shown in each case due to space limitations. All three web browsers enable users to view the ontologies in a hierarchical display, and locate a term of interest by searching or browsing. In addition, the GO and MP Browsers provide direct access to other integrated biological knowledge stored in MGI (see text).

As of 21 May 2008, MGI content includes over 28,233 mouse protein-coding genes and 140,626 GO functional annotations to those genes. More than half of the mouse genes (i.e., over 18,671 genes) have at least one GO annotation, and over 8,625 mouse genes have GO annotations derived from experimental data from the laboratory mouse. MGI gene-to-GO annotations are updated daily. In addition to the GO browser described below, a variety of files for MGI gene-to-GO associations are publicly available through the MGI FTP server [17], and the GO web site [18]. Frequently asked questions related to mouse GO annotations are provided at [4]; any other questions or suggestions about this resource can be addressed to mgi-help@informatics.jax.org.

### *1) The MGI GO Browser*

MGI has developed a web browser tool [19] that enables database users to access gene information using the GO functional annotation terms as search criteria. In addition, the gene-centric [7] and gene expression-centric [8] query forms available at MGI provide direct links to the MGI GO Browser to support GO-related queries. Browser users can explore each of the GO sub-ontologies in two ways. *Browsing* enables users to quickly navigate from high-level, broadly descriptive parent terms to progressively low-level, specific child terms, locate a term of interest, and view its semantic relationship to other terms in the hierarchy. A plus sign following a term indicates that this term has ‘descendants’, which can be viewed by clicking on the term to expand the relevant portion of the ontology.

*Searching* requires users to select a desired ontology (the default is to search all three) and enter any text string or GO identifier in the ‘Query’ field. When entering a text string (e.g., ‘death’), the MGI GO Browser searches for all terms containing that string (e.g., ‘cell death’) plus any synonyms (e.g., ‘anoikis’, a synonym of ‘detachment induced cell death’)**,** and returns a list of all matches found *per ontology* in the ‘Query Results’ page. When entering a full GO identifier, the browser searches the GO by the unique ID and returns only an exact match on the ID string. Information on an individual GO term is displayed in the ‘Term Detail’ page, and includes the unique identifier, definition, and synonym(s), all available *is-a*, *part-of*, or *regulates* term relationships, and the number of hierarchical paths that lead to the selected term (Fig. 1a).

Upon querying for a specific GO term, MGI users are able to traverse the DAG structure and retrieve all the mouse genes that are currently annotated to that term or any of its descendants. For instance, querying the Biological Process ontology for ‘cell death’ also retrieves any mouse genes that are annotated to a more specific child term such as ‘cytolysis’, ‘programmed cell death’ or ‘apoptosis’. At time of this publication, the parent term ‘cell death’ and its descendants have been associated with 705 mouse genes and 1185 annotation instances, as indicated in the adjacent hypertext link. Clicking the hyperlink launches a ‘Summary’ page which lists all matching genes by symbol and name, along with their chromosomal location, annotated GO term, evidence code, and supporting reference(s). In addition, each gene symbol is hyperlinked to its corresponding MGI ‘Gene Detail’ page for further gene-centric information, including official gene name and symbol, mapping data, sequences, mammalian orthology, phenotypes, polymorphisms, GO classifications, gene expression data, protein domains, bio-reagents, as well as links to other databases and to the scientific literature.

### *B. The MP Ontology*

MGI curates aberrant mouse phenotypes in the context of mutations (spontaneous, induced or genetically-engineered), strain variations, QTL, and complex traits that serve as plausible models of human biology and disease processes. To this end, MGI employs the MPO as a standardized, DAG-structured vocabulary that permits robust phenotypic characterization across different domains and species, and supports flexible annotations to background-specified allelic mouse genotypes at varying degrees of granularity [20].

Unlike GO, the MPO describes phenomena whose manifestations may result in deviations from an idealized canonical structure, and is used to associate mouse phenotype data with *genotypes* instead of genes. In MGI, genotype is defined as one or more allele pairs describing mutations or QTL and the genetic background strain(s) where the phenotype is observed. Typically, each phenotype annotation associates a given MP term with a genotype, an experimental evidence (EE) code, and the reference or data source supporting this assertion. Where deemed necessary, auxiliary modifying text is annotated to capture phenotypic detail that is either too case-specific to constitute a reusable MP term, or simply not amenable to standardization, such as specifics on the age of onset, incidence or trait penetrance. Background-sensitive notes are also provided to alert researchers to specific strain background effects that modulate the expressivity or pleiotropy of discernible phenotypes.

The topmost levels of the MPO include major physiological systems, behavior, development and survival. Following a general organizing principle, all physiological systems typically bifurcate into morphological and physiological phenotype descriptors at the next node level [20]. Each MP term is assigned a unique numerical identifier (MP:nnnnnnn), definition and synonym(s), and may have multiple parent/child relationships, currently represented by *is-a* relationship types (e.g., a ‘kinked tail’ *is-an* ‘abnormal tail morphology’).

In contrast to the Phenotypic Quality Ontology (PATO) [21], the MPO exemplifies an ontology of ‘pre-coordinated’ phenotypes. Thus, whereas the PATO model deploys entities and qualities as the building blocks of 'post-coordinated' phenotypic descriptions using, for instance, the Adult MA term ‘thymus’ [MA:0000142] and the PATO term ‘hyperplastic’ [PATO:0000644] to compose a ‘hyperplastic thymus’ phenotype representation at annotation time, the MPO endorses direct pre-composition of phenotype descriptors using the term ‘thymus hyperplasia’ [MP:0000708] at ontology construction time. Where necessary, MPO annotators apply widely-adopted compound clinical descriptors to unambiguously represent a single, yet often multi-faceted, pathological concept. A typical example of such a compound term is ‘hydrocephaly’ [MP:0001891], defined as ‘excessive accumulation of cerebrospinal fluid in the brain, especially the cerebral ventricles, often leading to increased brain size and other brain trauma’. In this case, the single-term approach obviates the necessity of multiple annotations to convey all aspects of a complex phenotype, minimizes curatorial time, and helps preserve the specificity of commonly cited clinical and pathological descriptors which may be compromised or lost once the terms are completely deconstructed.

Although highly practical from a curatorial perspective, use of compound MP terms can be challenging in terms of hierarchical assignment as, for instance, *part-of* and other relationship distinctions are particularly arduous to define. Current work focuses on reevaluating the MPO hierarchy to efficiently represent *part-of* and other relationship types. In parallel, efforts are underway to decompose pre-coordinated MP terms into computable logical definitions (cross-products) and render the PATO-style and MPO phenotype descriptions interoperable, using PATO in conjunction with orthogonal ontologies of quality-bearing entities [21]. In fact, the MPO has been designed as a cross-product ontology (22) that can hold, for each term, seamless cross-references to other open biological ontologies (OBO) [23], currently incorporating or associating terms from the GO Biological Process ontology, the Cell-Type (CL) Ontology of the OBO Foundry [24], the Embryonic Mouse Anatomical Dictionary developed by EMAP [25], the Adult MA Dictionary developed by GXD (see below), and the Mouse Pathology ontology (MPATH) developed by Pathbase [26], among others.

Initially built as a high-level classification of cardiovascular and skeletal phenotypes and comprising 105 terms [27], the MPO continues to expand through the dynamic process of data-driven phenotype curation, as well as collaborative input from other user groups, mutagenesis consortia, and biological domain specialists. Active MPO users include the Rat Genome Database [28] and Online Mendelian Inheritance in Animals (OMIA) [29]. As novel and increasingly complex phenotype traits are published, new sets of terms are identified, defined and organized along existing or entirely new hierarchical paths to parallel term usage in the scientific literature. Proposed new terms are additionally compared to other ontologies (see above) in an attempt to harmonize definitions, collect synonyms, and keep hierarchical placement logically consistent. Recently revised portions of the MP vocabulary include the hearing/ear, early CNS development, and vision/eye sections, all of which have been subjected to systematic expert review for term refinement, synonym enrichment, and hierarchical reorganization.

As of 21 May 2008, the MPO contains over 6,094 terms; more than 23,599 mouse genotype records have been annotated to MP terms, totaling over 120,723 MGI phenotypic annotation instances. The MPO is updated daily and is available in browser (see below), in OBO file formats through the MGI FTP site [17], as well as in various other formats from the OBO Download Matrix [30]. Suggestions, additions, or questions about the MPO can be addressed directly to pheno@informatics.jax.org.

#### *1) The MP Browser*

The MP Browser tool is available at [31]. In addition, the gene-centric [7] and phenotype-centric [9] MGI query forms also provide links to the MP Browser to facilitate execution of phenotype-related searches. The browser is designed to reinforce consistent retrieval of mouse phenotypes to the level of known data resolution, be it general or highly specific, offering the ability to query with a high-level phenotype term and retrieve all relevant mutant genotypes annotated to that term or its descendants. Thus, MGI users can query the MP Browser for 'abnormal tail morphology' and retrieve all mouse genotypes annotated to this term and its hierarchical children (e.g., 'abnormal caudal vertebrae morphology', 'abnormal tail pigmentation', 'absent tail', and so on), or specifically request annotations to any of these sub-terms. Following MGI's generic browser paradigm, searching requires users to enter any text string or full MP identifier in the 'Query' field, and returns either a list of all matching items containing that string (plus any synonyms), or an exact ID match in the 'Query Results' page.

Browsing enables MGI users to select a top-level, broadly descriptive phenotype category, such as 'limbs/digits/tail', and quickly navigate to increasingly low-level, granular phenotype terms until they locate a term of interest. As before, a plus sign appearing next to a term indicates the existence of descendants. Information on an individual MP term is displayed in the 'Term Detail' page, and includes the unique MP identifier, definition and synonyms, along with all *is-a* term relationships, and possible hierarchical paths that lead to the selected term (Fig. 1b). Displayed next to the term is a hypertext link indicating the total number of mouse genotypes and annotation instances using this term or any of its descendants, enclosed in parentheses. Clicking the hypertext link launches a 'Summary' page which lists all matching genotypes (i.e., allelic compositions plus genetic background), along with their annotated MP term, and supporting reference. Notably, each constituent allele of an allelic pair is hyperlinked to its corresponding MGI 'Phenotypic Allele Detail' page for a full phenotype description, including all reference-supported MP annotations organized by phenotype category, images of phenotypic genotypes, and established genotype models of human disease.

### *C. The Adult MA Dictionary*

Anatomical ontologies are rapidly emerging as critical data aggregators that enable MODs to encode structural knowledge in ways that can support exploration, mining, and machine-based inference of anatomy-based phenotype and gene expression data within and across species. The GXD component of MGI [5] is a community resource of standardized and image-supported spatiotemporal gene expression information, emphasizing endogenous gene expression patterns during mouse development. GXD collects and integrates primary data from different expression assays, each of which encapsulates gene expression profiles of wild-type and mutant mice at varying degrees of spatial granularity [6]. Currently, both GXD and EMAGE [32] use the Embryonic Mouse Anatomical Dictionary, hereafter referred to as the EMAP ontology [25], to capture detailed gene expression assay results for each successive Theiler stage (TS) during mouse development (TS1 through TS26).

As a logical extension of the EMAP ontology, GXD has built the Adult MA Dictionary [33] to provide standardized nomenclature for anatomical entities in the postnatal mouse (TS28). At present, GXD uses Adult MA terms to annotate expression information pertinent to *all* postnatal stages. While current annotation and display of 'adult' gene expression results employs an abridged version of TS28 anatomy [34], efforts are underway to map expression data directly to the expanded Adult MA Dictionary, reviewed below. Ultimately, the EMAP and adult MA ontologies will be structurally aligned and fully integrated to deliver a robust spatial representation of gene activity spanning the entire lifespan of the laboratory mouse.

Initially modeled, as far as possible, on the EMAP dictionary (TS26) for consistency, the Adult MA ontology is organized hierarchically from body region or organ system to tissue to tissue substructure, using formal naming conventions described in [33]. Each anatomical term is assigned a unique identifier (MA:nnnnnnn), and may have a definition and synonym(s), as available. However, unlike its embryonic counterpart, where each Theiler stage is primarily organized as a straight partonomic (*part-of)* hierarchy allowing only single parent-child relationships, the Adult MA is structured as a DAG, allowing each mouse anatomical structure to be represented as a child of multiple hierarchical parent terms using both *is-a* and *part-of* relationships. Thus, in addition to capturing structural knowledge, the Adult MA attempts to encapsulate some of the functional and spatial relationships between tissues, using the distinction between spatial versus organ system representation as an organizing principle. In this regard, the adult 'liver' concept is both a child of (*is-a*) 'abdomen organ', as well as *part-of* the 'hepatobiliary system'. The planned integration effort will include the representation of *derived-from* types of relationships that will link stage-specific anatomical components at subsequent stages so that it becomes possible to query the derivation and destination of any given tissue.

The root node (TS28) of the Adult MA consists of three top hierarchical levels: 'anatomic region', 'body fluid/substance', and 'organ system'. The initial division of the hierarchy into spatial and organ system components is readily apparent at the first level of substructures below TS28. Thus, anatomical terms encapsulated in the 'anatomic region' superstructure are primarily organized based on spatial localization (e.g., 'body cavity/lining', 'head/neck', 'limb', 'tail', and 'trunk'), while terms represented in the 'organ system' superstructure are organized, as much as possible, according to their respective contribution to a specified functional system (e.g., 'cardiovascular system', 'endocrine system', 'nervous system'). Where appropriate, generic group terms, such as 'blood vessel', 'connective tissue', 'muscle', 'nerve', 'organ', and 'skin', have been applied as sub-terms to represent groups of tissues that are localized in multiple spatial regions. Notably, use of 'pre-coordinated' terms, such as 'thymus epithelium' is reinforced to render anatomical terms unambiguous and readily interpretable.

Early construction of the Adult MA ontology was based on anatomical term extraction from various mouse-specific atlases, and anatomy and histology text resources listed in [33]. Following meticulous term validation, the ontology was augmented via a data-driven approach which entailed extensive evaluation of the published literature and of various anatomically-mapped datasets stored in mouse-specific resources, such as MGI. Currently containing 2,774 terms, the Adult MA continues to be refined by reexamining the hierarchical extensions and term relationships, by adding definitions and synonyms as required, and by creating new terms to label microanatomical structures at a level of granularity that is appropriate for querying.

The Adult MA will be used as key data aggregator to encode and integrate different types of data pertinent to postnatal mouse anatomy, such as gene expression patterns and phenotype information curated at MGI. Eventually, cross-referencing the Adult MA ontology with orthogonal vocabularies, such as the GO, MPO, CL, and Pathbase, will help integrate information relevant to expression, biological process, phenotype, and pathology. This type of integration will in turn enable execution of insightful, multi-parametric queries, such as 'Which mouse growth factors are expressed in the heart and are associated with allelic mutations that result in abnormal cardiac valve morphology?'

The Adult MA ontology is updated regularly and is available in a web browser (see below), in OBO file formats at the OBO Foundry site [23], and in various other formats from the OBO Download Matrix [30]. Suggestions, additions, or questions about the Adult MA can be addressed directly to anatomy@informatics.jax.org.

### *1) The Adult MA Dictionary Browser*

Currently, the Adult MA Dictionary Browser [35] allows users to locate standardized terms for anatomical structures present in the postnatal mouse, and view their relationships in a hierarchical display. Browsing launches a 'Term Detail' page for the root node (TS28) displaying the top level terms in the hierarchy (see above), and provides a starting point for progressively navigating through the ontology to locate specific anatomical structures. Clicking on individual terms results in 'Term Detail' pages displaying parents (super-structures), siblings (structures at the same level), and children (sub-structures). For each selected anatomical term, the 'Term Detail' page displays the unique MA identifier, the definition and synonyms (if available), along with all existing *is-a* or *part-of* term relationships, and possible hierarchical paths that lead to the term (Fig. 1c). As before, a plus sign following a term indicates the existence of child terms (sub-structures). Searching requires users to enter a text string or full MA identifier in the 'Query' field and brings up a 'Query Results' page displaying all structures that match the query. For instance, entering the text string 'cardium' will return a list of matching items, including the terms 'atrium endocardium', 'myocardium', 'dorsal mesocardium', and 'pericardium'. Entering the MA identifier will, of course, return only an exact match on the accession number.

Formal incorporation of the expanded Adult MA ontology into the MGI database system will eventually allow the browser to display expression assay results and phenotype data associated with specific anatomical structures, as is already the case for developmental expression data [34].

## IMPLEMENTATION

MGI is implemented in the Sybase relational database system. A large set of CGI scripts and Java Servlets mediates the user's interaction with the database. For computational users, direct SQL access can be requested through User Support at mgi-help@informatics.jax.org.

## REFERENCES

[1] J. A. Blake and C .J. Bult, "Beyond the data deluge: data integration and bio-ontologies," *Journal of Biomedical Informatics*, vol. 39, pp. 314-320, June 2006.

[2] Mouse Genome Informatics (MGI) Available: http://www.informatics.jax.org

[3] C. J. Bult, J. T. Eppig, J. A. Kadin, J. E. Richardson, J. A. Blake; Mouse Genome Database Group, "The Mouse Genome Database (MGD): mouse biology and model systems," *Nucleic Acids Research*, vol. 36, pp. D724-728, January 2008.

[4] The GO Project at MGI Available: http://www.informatics.jax.org/function.shtml

[5] Gene Expression Database (GXD) Available: http://www.informatics.jax.org/expression.shtml

[6] C. M. Smith, J.H. Finger, T.F. Hayamizu, I.J. McCright, J.T. Eppig, J. A. Kadin, *et al.*, "The mouse Gene Expression Database (GXD): 2007 update," *Nucleic Acids Research*, vol. 35, pp. D618-623, January 2007.

[7] MGI Genes and Markers Query Form Available: http://www.informatics.jax.org/javawi2/servlet/WIFetch?page=markerQF

[8] MGI Gene Expression Data Query Form Available: http://www.informatics.jax.org/javawi2/servlet/WIFetch?page=expresionQF

[9] MGI Phenotypes, Alleles & Disease Models Query Form Available: http://www.informatics.jax.org/searches/allele_form.shtml

[10] J. Day-Richter, M. A. Harris, M. Haendel; Gene Ontology OBO-Edit Working Group, S. Lewis, "OBO-Edit--an ontology editor for biologists," *Bioinformatics*, vol. 23, pp. 2198-2200, August 2007.

[11] OBO-Edit Available: http://www.oboedit.org/

[12] Gene Ontology Home Available: http://www.geneontology.org/

[13] Gene Ontology Consortium, "The Gene Ontology project in 2008," *Nucleic Acids Research*, vol. 36, pp. D440-444, January 2008.

[14] J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, J. E. Richardson; Mouse Genome Database Group, "The mouse genome database (MGD): new features facilitating a model system, " *Nucleic Acids Research*, vol. 35, pp. D630-637, January 2007.

[15] M. E. Dolan and J. A. Blake, "Using ontology visualization to understand annotations and reason about them," in KR-MED 2006 Biomedical Ontology in Action, Baltimore, MD, November 8, 2006, pp. 21–29.

[16] A. D. Diehl, J. A. Lee, R. H. Scheuermann, and J. A. Blake, "Ontology development for biological systems: immunology," *Bioinformatics*, vol. 23, pp. 913-915, April 2007.

[17] MGI Data and Statistical Reports Available: ftp://ftp.informatics.jax.org/pub/reports/index.html

[18] GO Current Annotations Available: http://www.geneontology.org/GO.current.annotations.shtml

[19] MGI GO Browser Available: http://www.informatics.jax.org/searches/GO_form.shtml

[20] C. L. Smith, C. W. Goldsmith, and J. T. Eppig, "The Mammalian Phenotype Ontology as a tool for annotating, analyzing, and comparing phenotypic information," *Genome Biology*, vol. 6, pp. R7, December 2004.

[21] PATO – An ontology of Phenotypic Qualities Available: http://www.bioontology.org/wiki/index.php/PATO:About

[22] D. P. Hill, J. A. Blake, J. E. Richardson, and M. Ringwald, "Extension and integration of the gene ontology (GO): combining GO vocabularies with external vocabularies," *Genome Research*, vol. 12, pp.1982-1991, December 2002.

[23] The Open Biomedical Ontologies (OBO) Available: http://www.obofoundry.org/

[24] J. Bard, S. Y. Rhee, and M. Ashburner, "An ontology of cell types," *Genome Biology*, vol. 6, pp. R21, January 2005.

[25] A. Burger, D. Davidson, and R. Baldock, "Formalization of mouse embryo anatomy," *Bioinformatics*, vol. 20, pp. 259-267, January 2004.

[26] P. N. Schofield, J. B. Bard, C. Booth, J. Boniver, V. Covelli, P. Delvenne, et al., "Pathbase: a database of mutant mouse pathology," *Nucleic Acids Research*, vol. 32, pp. D512-515, January 2004.

[27] J. A. Blake, J. E. Richardson, C. J. Bult, J. A. Kadin, J. T. Eppig; Mouse Genome Database Group, "MGD: the Mouse Genome Database, " *Nucleic Acids Research*, vol. 31, pp. 193-195, January 2003.

[28] Rat Genome Database (RGD) Available: http://rgd.mcw.edu/

[29] Online Mendelian Inheritance in Animals Available: http://www.ncbi.nlm.nih.gov/sites/entrez?db=omia

[30] OBO Download Matrix Available: http://www.berkeleybop.org/ontologies/

[31] Mammalian Phenotype Browser Available: http://www.informatics.jax.org/searches/MP_form.shtml

[32] Edinburgh Mouse Atlas Gene Expression (EMAGE) Available: http://genex.hgu.mrc.ac.uk/Emage/database/emageIntro.html

[33] T.F. Hayamizu, M. Mangan, J. P. Corradi, J. A. Kadin, and M. Ringwald, "The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data," *Genome Biology*, vol. 6, pp. R29, February 2005.

[34] The Mouse Anatomical Dictionary Browser Available: http://www.informatics.jax.org/searches/anatdict_form.shtml

[35] Adult Mouse Anatomical Dictionary Browser Available: http://www.informatics.jax.org/searches/AMA_form.shtml