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

www.informatics.jax.org
I. Gene Ontology (GO@MGI)

MGI is the authoritative source of GO functional annotations to mouse genes & gene products available in MGI.

II. Mammalian Phenotype (MP) Ontology

MGD employs the MP Ontology as a standardized vocabulary that permits robust phenotypic characterization across different domains & species.

III. Adult Mouse Anatomical (MA) Dictionary

GXD uses the Adult MA Dictionary to:

- provide standardized nomenclature for anatomical entities in the postnatal mouse (TS28; adult mouse)
- capture gene expression assay results for all postnatal stages
Designing MGI Bio-Ontologies

All MGI bio-ontologies have been designed with a common generic vocabulary infrastructure

- built as directed acyclic graphs (DAGs) using the OBO-edit Java tool for construction, maintenance & editing
- share a common data model
- use identical annotation tools
- employ equivalent web browsers with browsing & search capabilities to navigate, query, analyze & compare the biological knowledge at hand

The terms constituting the nodes in a DAG represent the kinds of entities that exist within the domain of that ontology.

The edges in a DAG represent specified relations that hold between these entities.
MGI Bio-ontologies/Structured Vocabularies: Using a Generic Annotation Data Model

Vocabulary

DAG

Terms

Definition

Synonyms

MP:0004201

Strain: AEJ

Alleles: bd/bd

Genotype

Strain: C57BL/6

Alleles: Ppp1r3a^{tm1Adpt}/Ppp1r3a^{tm1Adpt}

Annotations

example: MP Annotations

Terms

- Fetal growth retardation
- Dilated renal tubules
- Postnatal lethality
- Respiratory failure
- ...

Evidence Code

<table>
<thead>
<tr>
<th>Evidence Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>EE</td>
</tr>
<tr>
<td>IDA</td>
</tr>
<tr>
<td>TAS</td>
</tr>
</tbody>
</table>

Reference (J:nnnnnn)

- EE J:65322
- IDA J:62648
- TAS J:65371
- ...

Notes

J:65371
J:62648
J:65322

Annotations

example: MP Annotations
I. The GO project at MGI

The GO project (www.geneontology.org/) is a community effort to address the need for consistent descriptions of gene and gene product attributes in different MODs.

- The GO vocabulary system is composed of three orthogonal, canonical, and species-neutral ontologies, including:
  - Molecular function e.g. 'binding', 'adenylate cyclase activity'
  - Biological process e.g. 'growth', 'signal transduction'
  - Cellular component e.g. 'cell envelope', 'nucleolus', 'activin complex'

- As a founding member of the GO Consortium, MGI is an active participant in the development and application of GO.

- MGI fully incorporates the GO in the database and provides:
  - MGI-GO Browser - Access gene information using GO terms as search criteria
  - MGI GO Term Finder - Analyze functional annotations
  - GO Chart Tool - Build GO charts to present GO functional data
Structuring GO terms as a DAG hierarchy

Each orthogonal GO sub-ontology (Molecular Function, Biological Process, Cellular Component) is independently organized as a DAG

- A DAG allows multiple parentage both along is-a and part-of transitive relationships propagated from more specialized (child) terms to less specialized (parent) terms.

- A DAG hierarchy allows both attribution and querying at varying levels of detail.

- Each GO term (node) must conform to the ‘true path’ rule applying to both is-a and part-of relationships. For example:
  - if ‘anion channel activity’ is-a subclass of ‘transporter activity’, it also is-a ‘molecular function’.
  - 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’.

DAGs
MGI-GO Development

All three GO sub-ontologies evolve and expand dynamically to reflect accumulating and changing biological knowledge

MGI-GO curators...

- participate in Curator Interest Groups formed of GOC members and community experts to focus on specific domain areas within the GO ontologies

- attend meetings devoted to ontology content

- create new GO terms & play an active role in augmenting, refining, and reorganizing existing terms and relationships, often in consultation with biological domain specialists

- Recent MGI contributions to the GO include:
  - collaborative implementation of an enhanced representation of immunological content
  - revisions of hierarchical extensions for 'blood pressure regulation' and 'muscle development' in the Biological Process ontology
MGI-GO annotations

- **Current Status (Sept 11, 2008)**

| Genes w/ protein sequence data                  | 27,245 |
| Genes w/ GO annotations                         | 18,078 |
| Genes w/ experimentally-derived GO annotations | 7,274  |
| GO annotations total                            | 133,424|
| Unique refs used for GO annotations             | 11,410 |

- Updated daily
- Available in:
  - MGI-GO Web Browser
  - a variety of files from the MGI FTP site and the official GO web site
- Email suggestions, additions or questions to: mgi-help@informatics.jax.org
Functional Annotation using the Gene Ontology (GO)

MGI's GO project provides functional annotations for mouse gene products using the Gene Ontology.
To browse the GO ontologies:

1. Click an ontology of interest, e.g., Biological Process.
2. Navigate from high-level parent terms to progressively low-level, specific child terms, locate a GO 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'.

GO Term Detail includes:
- Synonyms
- GO:nnnnnnn
- Definition
- Comments (as appropriate)
- Number of paths to term
- Relationship types
To search the GO ontologies:

1. select an ontology of interest (e.g. Biological Process)
2. enter a text string (e.g. 'cell death') or full GO accession number (e.g. GO:0008219) in the Query box
3. click Search

The MGI GO Browser searches all terms containing that string (e.g. 'cell death') plus any synonyms (e.g. 'anoikis'), and returns a list of all matches found per ontology (in this case, Biological Process).

94 Biological Process term(s) matching query "cell death":

- activated T cell apoptosis
- anoikis
- antibody-dependent cellular cytotoxicity
- apoptosis
- autophagic cell death
- B cell apoptosis
- cell death
- compound eye retinal cell programmed cell death
- developmental programmed cell death
- ecdysone-mediated induction of salivary gland cell autophagic cell death
- germ cell programmed cell death
- glial cell apoptosis
- host programmed cell death induced by symbiont
- hydrogen peroxide-mediated programmed cell death
- induction by organism of apoptosis in other organism during symbiotic interaction
- induction by organism of non-apoptotic programmed cell death in other organism during symbiotic interaction
- induction by organism of programmed cell death in other organism during symbiotic interaction
- induction by symbiont of host apoptosis
- induction by symbiont of host programmed cell death
- induction by virus of host apoptosis
- induction of compound eye retinal cell programmed cell death
- induction of non-apoptotic programmed cell death
- induction of non-apoptotic programmed cell death by other organism
Each GO term is linked to all the mouse GENES annotated to that term or its 'descendants'.

**Gene Ontology Browser**

GO term: cytolysis
Synonym: autolysis activity
Synonym: bacteriocin activity
Synonym: bacteriolytic toxin activity
Synonym: holin
Synonym: lysis activity
Synonym: lysis
GO id: GO:0019835
Definition: The rupture of cell membranes and the loss of cytoplasm.
Number of paths to term: 2

- "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**

1. **biological process**
   - **cellular process**
     - **cell death**
       - **cytolysis [GO:0019835]** (24 genes, 26 annotations)
         - autolysis
           - negative regulation of cytolysis +
             - positive regulation of cytolysis +
               - regulation of cytolysis +
                 - programmed cell death +

2. **biological process**
   - **developmental process**
     - **death**
       - **autolysis [GO:0019835]** (24 genes, 26 annotations)
         - autolysis
           - negative regulation of cytolysis +
             - positive regulation of cytolysis +
               - regulation of cytolysis +
                 - programmed cell death +
MGI-GO Query Results - Summary

- Lists all matching genes annotated to desired GO term or its descendants by gene symbol & name
- Chromosome number (or letter) location
- Category
- Evidence Code
- Reference (J:nnnnnn) supporting annotation
- Links to the MGI 'Gene Detail' pages for further gene-centric information

<table>
<thead>
<tr>
<th>GO Term</th>
<th>Gene symbol &amp; name</th>
<th>Chr. Location</th>
<th>Evidence Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0003824</td>
<td>gja3</td>
<td>2</td>
<td>IC</td>
</tr>
<tr>
<td>GO:0003824</td>
<td>gja2</td>
<td>2</td>
<td>IC</td>
</tr>
<tr>
<td>GO:0003824</td>
<td>gja1</td>
<td>2</td>
<td>IC</td>
</tr>
<tr>
<td>GO:0003824</td>
<td>gja4</td>
<td>2</td>
<td>IC</td>
</tr>
</tbody>
</table>

...to 'Gene Detail'
## Prf1 Gene Detail

### Gene symbol/name/ID
Prf1

### Synonym
perform, Psp, Prf-1

### Genetic Map
- Chromosome 10
- Detailed Genetic Map + Link
- Mapping detail

### Sequence Map
- Ctr 10:00756022-00757028 bp, + strand
- (From NCBI annotation of NCBI Build 37)
- Ensembl Coordinates | UCSC Browser | NCBI Map Viewer

### Mammalian homology
- Human, dog, domestic rat (Mammalian Orthologs)
- Comparative Map (Mouse/Human Prf1 + Link)
- TreeView: F226498

### Sequences
- Represented Sequences
  - genomic: 18846
  - transcript: NM_011273, RefSeq, MGI Sequence Detail
  - cDNA: 3102, CTD464
  - polypeptide: 554, not applicable

### Phenotypes
- All alelic variants are lethal in targeted, knock-out strains
- Homozygous null mice exhibit increased susceptibility to viral infection and defective cytotoxic T cell proliferation and mit cell proliferation.
- Associated Human Diseases: 2
- Alleles Associated to Human Diseases: 2

### Polymorphisms
- PRPL1 SNPs within 51686 from dbSNP Build 128

### Gene Ontology classifications
- Process: cytokinesis
- Component: cytokinetic membrane-bounded vesicle, integral to membrane...
- Function: cytokinetic binding

### Expression
- Literature Summary (2 records)
- Allen Brain Atlas

### Other database links
- Ensembl Gene Model: ENSMUSG000000137262
- UniGene: 49024
- DFCI: TC159238, TC159430
- NCBI Mouse Gene Index: LG151957
- Entrez Gene: 18646

### Protein domains
- InterPro ID Description
  - InterPro00010: C2 calcium-dependent membrane targeting
  - InterPro00074: EGF-like Type 3
  - InterPro00184: Membrane attack complex component(C9)complement C9
  - InterPro01372: EGF-like domain, conserved site
- Structural View of InterPro Domain Sequence

### Molecular reagents
- Antibodies (15), Genomic (5)

### References

### Other accession links
- MGI: MGI:12351, MGI:12355
Research by Gots M et al suggests by a direct assay that Prf1 is involved in cytotox. Large reviewed computational analysis by the FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I & II Team suggests that Prf1 is in the extracellular space. Research by Oder-Pikle S et al suggests by a direct assay that Prf1 is in a cytoplasmic membrane-bounded vesicle location.

Prf1 encodes a protein or proteins that contain the following InterPro domains: C2 calcium-binding and Membrane attack complex component/perforin component C9.

**Gene Ontology Evidence Code Abbreviations:**
- IC: Inferred by curator
- IDA: Inferred from direct assay
- IEA: Inferred from electronic annotation
- IGI: Inferred from genetic interaction
- IMP: Inferred from mutant phenotype
- IPI: Inferred from physical interaction
- ISS: Inferred from sequence or structural similarity
- ISO: Inferred from sequence orthology
- TSA: Inferred from sequence alignment
- IAS: Inferred from sequence model
- TAA: Non-traceable author statement
- NA: No biological data available
- RCA: Reviewed computational analysis
- TAS: Traceable author statement
II. The Mammalian Phenotype Ontology (MPO)

MPO permits standardized phenotypic characterization across different domains and species, and supports flexible annotations to background-specified allelic mouse genotypes at varying degrees of granularity.

- Structured as a DAG, currently allowing is-a relationship types
- MPO topmost levels include:
  - major physiological systems
  - development
  - behavior
  - survival
- Over 6,185 MP terms
- Updated daily
- Available in:
  - MGI Web Browser
  - OBO file formats from MGI FTP site
  - various other formats from OBO Download Matrix
- Email suggestions, additions or questions to: pheno@informatics.jax.org
To search the MP Vocabulary:
1. enter a text string or full MP accession number (e.g. MP:0002111) in the Query box
2. click Search
Each MP term is linked to all mouse GENOTYPES annotated to that term or its descendants.
Lists all matching mouse genotypes annotated to selected MP term or its descendants

Reference (J:nnnnnn) supporting annotation

Links to the MGI 'Phenotypic Allele Detail' pages for a full phenotype description...to 'Phenotypic Allele Detail'

In MGI, a genotype is defined as one or more allele pairs describing mutations or QTL plus the genetic background strain(s) where the phenotype is observed.
Phenotypic Allele Detail

Phenotype annotation (in context of genotype)

*Example:* Lrp6<sup>cd</sup> (synonym: "crooked tail")

<table>
<thead>
<tr>
<th>Allotypic Composition</th>
<th>Genotypic Background</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lrp6&lt;sup&gt;cd&lt;/sup&gt;/Lrp6&lt;sup&gt;+&lt;/sup&gt;</td>
<td>A-Lrp6&lt;sup&gt;cd&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

### Limbs/digits/tail
- *abnormal caudal vertebrae morphology (J:13045)*
  - In most cases, the number of abnormal caudal vertebrae is between 3 and 6
  - *small caudal vertebrae (J:13045)*
  - Many caudal vertebrae are shortened

*Kinked tail (J:8433, J:13045)*
- Background Sensitivity: 41% of pups have a crooked tail; crooked tail is more prominent on the A background than others such as DBA, C57BL/6 and 129/Sv (J:10422)
  - The number of affected pups varies from 0-6 (J:13045)

### Skeleton
- *abnormal vertebrae morphology (J:13045)*
  - Abnormalities in the lumbar and sacral regions include lateral displacement of one or more vertebrae, dorsal spine duplication, deletion of a vertebral component and abnormally large parapophyses
  - *abnormal caudal vertebrae morphology (J:13045)*
    - In most cases, the number of abnormal caudal vertebrae is between 3 and 6
    - *small caudal vertebrae (J:13045)*
    - Many caudal vertebrae are shortened
  - *abnormal lumbar vertebrae morphology (J:13045)*
    - 11.6% incidence of malformed lumbar vertebrae
  - *abnormal sacral vertebrae morphology (J:13045)*
    - 14.1% incidence of malformed sacral vertebrae
  - *Vertebral fusion (J:13045)*
    - Sometimes in the lumbar and sacral regions there is fusion of two adjacent vertebrae into a common body with one parapophysis on one side and two parapophyses on the other
## MPO Annotations - Current Status

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MP Terms, Total</td>
<td>6,185</td>
<td>5,712</td>
<td>4,458</td>
</tr>
<tr>
<td>MP Annotations, Total</td>
<td>128,262</td>
<td>101,736</td>
<td>69,972</td>
</tr>
<tr>
<td>MP Annotations, excluding QTL</td>
<td>123,681</td>
<td>97,388</td>
<td>66,110</td>
</tr>
<tr>
<td>Genotypes with MP, Total</td>
<td>29,024</td>
<td>23,666</td>
<td>17,581</td>
</tr>
<tr>
<td>Genotypes with MP, excluding QTL</td>
<td>25,178</td>
<td>19,964</td>
<td>14,249</td>
</tr>
<tr>
<td>Alleles Annotated, excluding QTL / Alleles Total</td>
<td>17,907/29,763</td>
<td>14,807/25,891</td>
<td>11,149/22,775</td>
</tr>
<tr>
<td>Genes Annotated, excluding QTL</td>
<td>8,412</td>
<td>7,334</td>
<td>6,092</td>
</tr>
<tr>
<td>QTL Annotated</td>
<td>2,540</td>
<td>2,427</td>
<td>2,130</td>
</tr>
</tbody>
</table>
MPO Collaborators/Users

MPO continues to expand through collaborative input from other user groups and mutagenesis consortia

- Rat Genome Database (RGD)
- Mouse Mutagenesis Centers
- Human (NCBI/dbSNP)
- OMIA (Online Mendelian Inheritance in Animals)
- Proprietary Databases
- International Mouse Knockout Projects
MPO Development

- **Curation-driven approach**
  New terms are added as required by curators (MGI, RGD, plus others) at annotation time.

- **Systematic expert review**
  MPO sections (e.g. hearing/ear, eye/vision, early CNS development) are periodically subject to content review by biological domain specialists to:
  - refine terms
  - enrich synonyms
  - reorganize existing terms within particular branches

- **Comparisons to other ontologies**
  Proposed new terms are compared to other ontologies to:
  - harmonize definitions
  - collect synonyms
  - keep hierarchical placement logically consistent
  - identify problems with relationship types
**MPO** is an ontology of “pre-coordinated” phenotypes

– The MPO vs the PATO model –

**MPO**: “Pre-composition” of phenotype descriptor at ontology construction time

- **MP:0000708** “thymus hyperplasia”
  - **def**: overdevelopment or increased size, usually due to increased cell number, in the thymus
  - **syn**: hyperplastic thymus; increased thymus cellularity; thymic hyperplasia

**PATO**: “Post-composition” of phenotype descriptor at genotype annotation time

- **PATO:0000645** “hyperplastic”
- **MA:0000142** “thymus”
## Compound clinical/pathological terms used in MPO

<table>
<thead>
<tr>
<th>MP term</th>
<th>MP Definition</th>
<th>Object</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>microphthalmia</td>
<td>“reduced average size of the eyes”</td>
<td>eye</td>
<td>decreased size</td>
</tr>
<tr>
<td>MP:0001297</td>
<td></td>
<td>MA:0000261</td>
<td>PATO:0000587</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FMA:54448</td>
<td></td>
</tr>
<tr>
<td>hydroencephaly</td>
<td>“excessive accumulation of cerebrospinal fluid in the brain, especially the</td>
<td>cerebrospinal fluid</td>
<td>excessive</td>
</tr>
<tr>
<td>MP:0001891</td>
<td>cerebral ventricles, often leading to increased brain size and other brain</td>
<td>MA:0002503</td>
<td>(increased volume)</td>
</tr>
<tr>
<td></td>
<td>trauma”</td>
<td>no FMA term</td>
<td>PATO:0000595</td>
</tr>
</tbody>
</table>

### Single-term approach (e.g. hydroencephaly)

- obviates need for multiple annotations to convey all aspects of a complex phenotype
- minimizes curatorial time
- preserves specificity of commonly cited clinical/pathological descriptors

… but what about *part-of* or other relationship types?
The MPO represents a ‘cross-product’ ontology that can hold, for each term, seamless ‘cross-references’ to other OBO ontologies, including:

- Mouse Anatomy (EMAP, Adult MA)
- Foundational Model of Anatomy (FMA)
- GO Biological Process
- Cell Type (CL) Ontology
- Attribute and Value
- MPATH
- Others...
III. The Adult Mouse Anatomy (MA) Dictionary

GXD has built the Adult MA Dictionary to provide standardized nomenclature for anatomical entities in the postnatal mouse (TS28)

- Initially modeled, as far as possible, on the EMAP dictionary (TS26)
- Structured as a DAG, from body region or organ system to tissue to tissue substructure
- Organized both “spatially” and “functionally” using is-a and part-of relationships
  i.e. adult liver is-an ‘abdomen organ’ AND part-of the ‘hepatobiliary system’
- >2,774 unique Adult MA terms (no cell types except ‘fertilized egg’)
- Updated regularly
- Available in:
  - MGI Web Browser
  - OBO file formats at OBO Foundry site
  - various other formats from the OBO Download Matrix
- Email suggestions, additions or questions to: anatomy@informatics.jax.org
Adult MA Development

- **Anatomical Term Extraction**
  initially constructed based on term extraction from various mouse-specific atlases and anatomy/histology text resources

- **Data-driven approach**
  augmented via extensive evaluation of the published literature and of various anatomically-mapped datasets stored in mouse-specific resources

- **Ongoing refinement by**
  - re-examining hierarchical extensions and term relationships
  - providing definitions, as required
  - including synonyms, as available
  - creating new terms to label microanatomical structures at a level of granularity that is appropriate for querying
FAQs

Stats

Adult MA Browser

Gene Expression Database (GXD)

GXD collects and integrates the gene expression information in MGI. Its primary emphasis is on endogenous gene expression during mouse development.

Adult Mouse Anatomical Dictionary Browser

The Anatomical Dictionary for the Adult Mouse organizes anatomical structures for the postnatal mouse (Thiels stage 30) spatially and functionally, using 'is a' and 'part of' relationships. The ontology will be used to describe expression data for the adult mouse and phenotype data pertinent to anatomy in standardized ways. This browser can be used to view anatomical terms and their relationships in a hierarchical display.

Browse the Anatomical Dictionary for the Adult Mouse

Adult Mouse Anatomical Dictionary

Search the Anatomical Dictionary for the Adult Mouse

Enter any text string or full MGI accession number (include 'MGI' prefix)
Query:  
Search  Reset

Your input is welcome. Please contact us with suggestions, additions, or questions about the Anatomical Dictionary for the Adult Mouse.

The Anatomical Dictionary for the Adult Mouse has been developed by Tony Hayamizu, Mary Nangoy, John Correia and Martin Ringgold as part of the Gene Expression Database (GXD) Project. GXD is funded by NIH grant HD30776. MGI and MGI were supported by postdoctoral fellowships F32 HD00426-01 and F32 HD00426-01.

Contributing Projects
Mouse Genome Database (MGD), Gene Expression Database (GXD), Mouse Tumor Biology (MTB), Gene Ontology (GO), MouseCyc
To browse the Adult MA:

1. click Adult Mouse Anatomical Dictionary to launch the Term Detail for TS28, displaying three top hierarchical levels (superstructures)
2. navigate until you locate the desired anatomical term e.g. ‘pericardium’
Searching the Adult MA Dictionary

To search the Adult MA:
1. enter a text string (e.g. cardium or %cardium) or full MA accession number (e.g. MA:0000080) in the Query box
2. click Search
Adult MA Dictionary - Term Detail

Adult Mouse Anatomy

Term Detail

MA term: pericardium
MA id: MA:00000099
Number of paths to term: 5

- MA:nnnnnnn
- Synonyms (if available)
- Definition (if available)
- Number of paths to term
- Relationship types (is-a and part-of)

No links to gene expression results yet
TS28 is currently represented in an abridged form.

At present, GXD employs an abridged version of TS28 anatomy to annotate & display ‘adult’ gene expression results.
**Adult MA Dictionary – Future Plans**

The Adult MA will be used as a key data aggregator to encode & integrate different types of data pertinent to postnatal mouse anatomy e.g. gene expression patterns & phenotype information curated at MGI

- formally incorporate the expanded Adult MA into MGI and allow display of expression results and phenotype data associated with specific TS28 anatomical structures
- structurally align and fully integrate EMAP with the Adult MA Dictionary to deliver a robust spatial representation of gene activity throughout the lifespan of the mouse
- include *derived-from* relationships that will help query the derivation and destination of any given tissue
- cross-reference to orthogonal ontologies (e.g. GO, MPO, CL, MPATH) to integrate data relevant to expression, biological process, phenotype, cell type and pathology
- 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?*
Acknowledgements

MGI is supported by NIH grants HG000330, HG002273, HD033745, CA089713