Mouse Genome Database

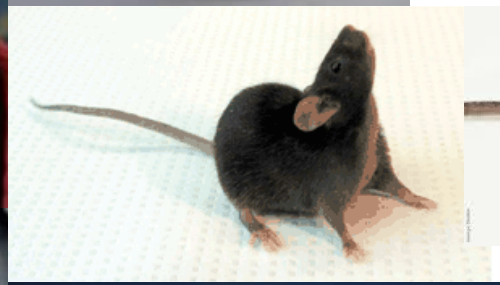
Judith A Blake Ph.D.

Data Resources in Life Sciences Workshop
HFSPO, Strasbourg, France
November 18, 2016

The Jackson Laboratory
Leading the search for tomorrow's cures









gds.nih.gov

Mouse Genome Database (MGD)

Mouse Genome Database program goal

…to facilitate the use of the mouse as a model for heritable human diseases and normal human biology.

Achondroplasia

Homozygous achondroplasia mouse mutant and control

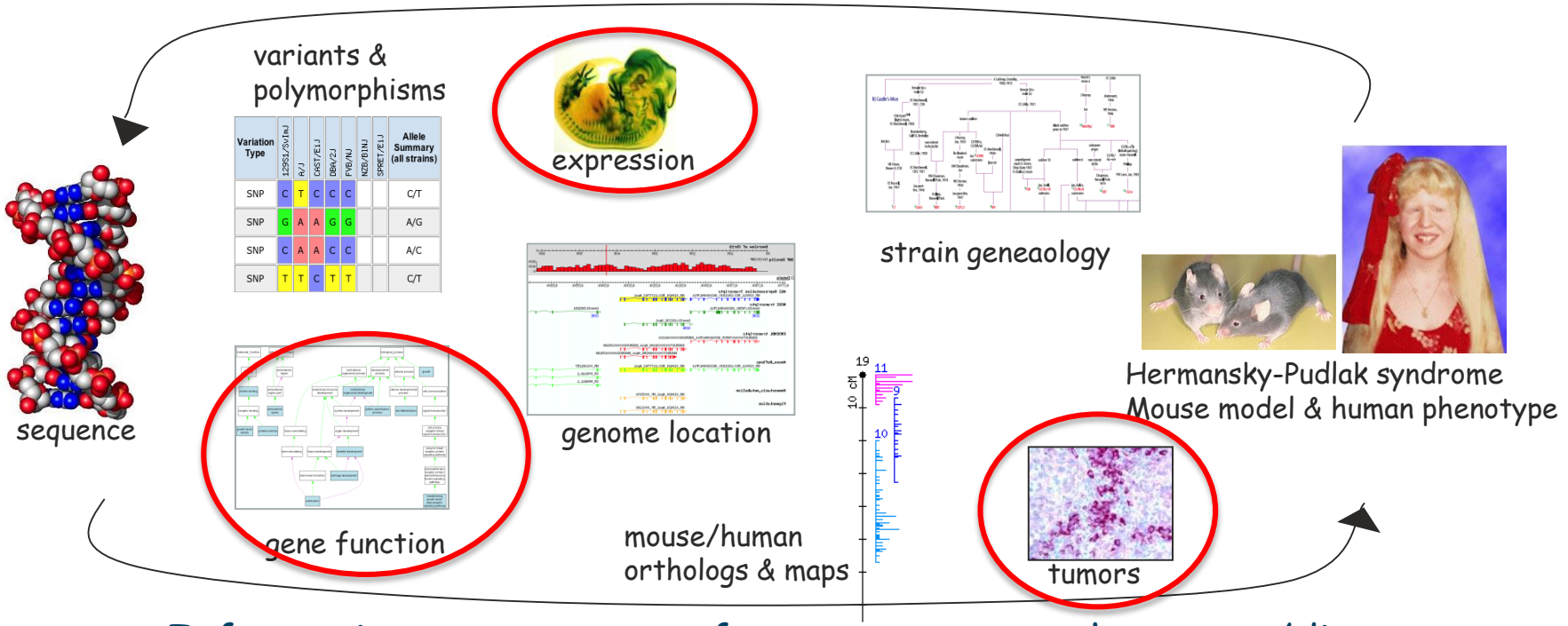
- short domed skull
- short-limbed dwarfism
- malocclusion
- bulging abdomen as adults
- respiratory problems
- shorted lifespan

MGD Mission and User Community

- Mission: To facilitate the use of the laboratory mouse as a model for understanding the genetic and genomic basis of human health and disease.
- Target User Communities:
  - Basic scientists/Mouse geneticists
  - Biomedical/Clinical scientists
  - Computational biologists/Bioinformaticians

Mouse Genome Informatics (MGI)

MGI's primary mission is to facilitate the use of mouse as a model for human biology by providing integrated access to data on the genetics, genomics, and biology of the laboratory mouse.



Information content spans from sequence to phenotype/disease

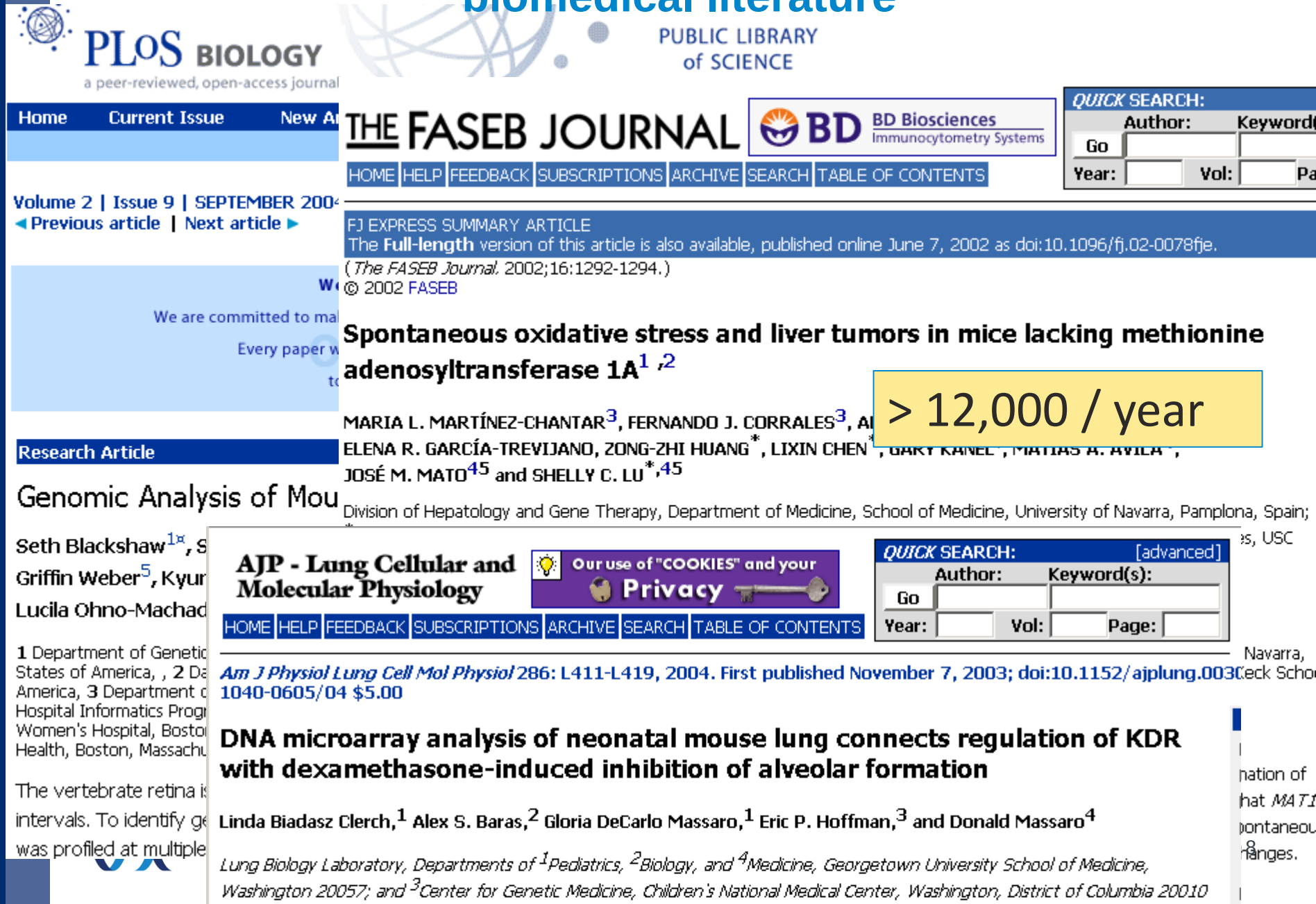
MGI Operating Principles

- ***Data integration*** is key to functional and comparative genomics
  - Allows data to be evaluated in new contexts
- ***Standards*** are key to data integration
  - Terminologies
    - Standardized gene nomenclature, keywords, etc.
  - Ontologies
    - Gene Ontology (GO)
    - Mammalian Phenotype Ontology (MP)
    - Mouse Anatomy (MA)

Automated Data Load Pipelines

Types of Data Integrated	Source	Frequency
Gene Ontology		
Gene Ontology	GO Consortium	weekly
Mouse GO annotations inferred from rat annotations	GO Consortium	weekly
Mouse GO annotations inferred from human annotations	GO Consortium	weekly
GO annotations from GO/PAINT	GO Consortium	weekly
GO annotations from GO/CFP	GO Consortium	weekly
GO annotations from GOC	GO Consortium	weekly
GO annotations from GOA	GOA	weekly
Other Ontologies		
Portions of Sequence Ontology	Sequence Ontology	As needed
Cell Ontology	Cell Ontology	weekly
Protein Information		
Protein Ontology	Protein Ontology	weekly
Protein Ontology annotations for mouse genes	Protein Ontology	weekly
PIR superfamily vocabulary	PIRSF	weekly
PIR superfamily associations for mouse, human, rat genes	PIRSF	weekly
Protein Domains for mouse genes	InterPro	weekly
Protein IDs, EC numbers, PDB IDs, InterPro domain IDs, and GO annotations for mouse genes	UniProt	weekly
Sequence records from SP/TrEMBL	SwissPROT	weekly
SP/TrEMBL sequence associations to mouse genes	SwissPROT	weekly
CCDS IDs for mouse genes	CCDS	weekly
neXtProt IDs for human genes	neXtProt	weekly
Human Disease		
OMIM disease vocabulary	OMIM	weekly
OMIM annotations for human genes	NCBI/OMIM	weekly
Human Phenotype Ontology (HPO)	HPO	As needed
OMIM term mapping to HPO terms	HPO	As needed
Deriving gene level MP/Disease annotations from genotypes		weekly
Gene/Sequence		
Gene records for non-mouse genes including nomenclature, chr location, etc.	NCBI	weekly
NCBI IDs, RefSeq associations for mouse genes	NCBI	weekly
RefSeq mouse sequence records	RefSeq	weekly
GenBank mouse sequence records	GenBank	weekly
Unigene IDs for mouse genes	Unigene	weekly
STS and Broad institute IDs & coordinates	UniSTS	Genome build
Gene Models		
NCBI mouse gene model associations	NCBI	Annot release
Representations of NCBI mouse gene models	NCBI	Annot release
Ensembl mouse gene model associations	Ensembl	Annot release
Representations of Ensembl mouse gene models	Ensembl	Annot release
Ensembl transcript/protein records	Ensembl	Annot release
Havana mouse gene model associations	Havana	Annot release
Representations of Havana mouse gene models	Havana	Annot release
Havana transcript/protein records	Havana	Annot release
Gene Unification: combining Havana, Ensembl, NCBI models into non-redundant set. Creating MGI GFF.		Annot release
Homology		
Homologene clusters	NCBI/Homologene	weekly
HGNC orthologs of mouse genes	HGNC	weekly
Combining HGNC/Homologene into single cluster set		weekly
Alleles and Phenotypes		
Targeted alleles, mutant cell lines from Eucomm and KOMP-CSD	Eucomm/KOMP2	weekly
Derived alleles produced from the KOMP/IMPC pipelines	KOMP2/IMPC	weekly
MP annotations from IMPC	KOMP2/IMPC	weekly
Other Loaded Data		
MicroRNA to gene interactions	mirTarBase, microT-CDS, miRDB, Pictar	On demand
Mouse SNPs	dbSNP	dbSNP build
Publication records	Pubmed	daily
Links to MyGene.info (Wikipedia gene pages)	MyGene.info	weekly
Gene - Metabolic pathway annotations & links	MouseCyc	weekly
Links to QTL Archive	QTL Archive	weekly
Expression (GXD)		
Links to EMAGE	EMAGE	weekly
Links to GEO for mouse genes	GEO	weekly
Links to GENSAT for mouse genes	GENSAT	weekly
Links to ArrayExpress for mouse genes	ArrayExpress	weekly
Links to ABA for mouse genes	Allen Brain Atlas	weekly
Links to expression for Zebrafish curated orthologs	ZFIN	weekly
Links to expression for Chicken curated orthologs	GEISHA	weekly
Links to expression for Xenopus curated orthologs	XenBase	weekly
IMPC lacZ images and annotations	IMPC	2-3 per year

Integration via semi-automated, expert curation of the biomedical literature

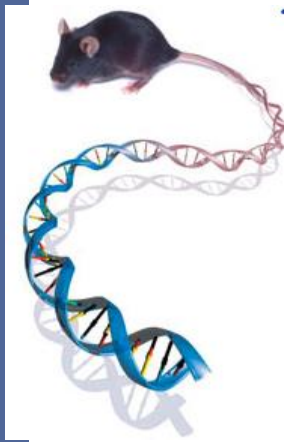


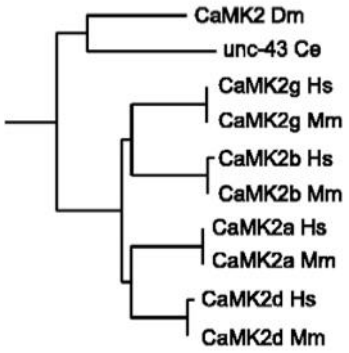
MGD Resource Project focus areas

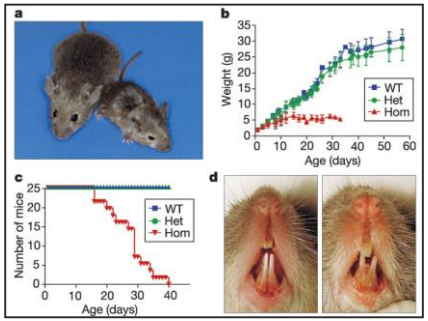
(i) Genome features

(ii) Functional annotations & comparative genomics for mouse

(iii) Mutant alleles, their phenotypic manifestations & associations with the human diseases which they model



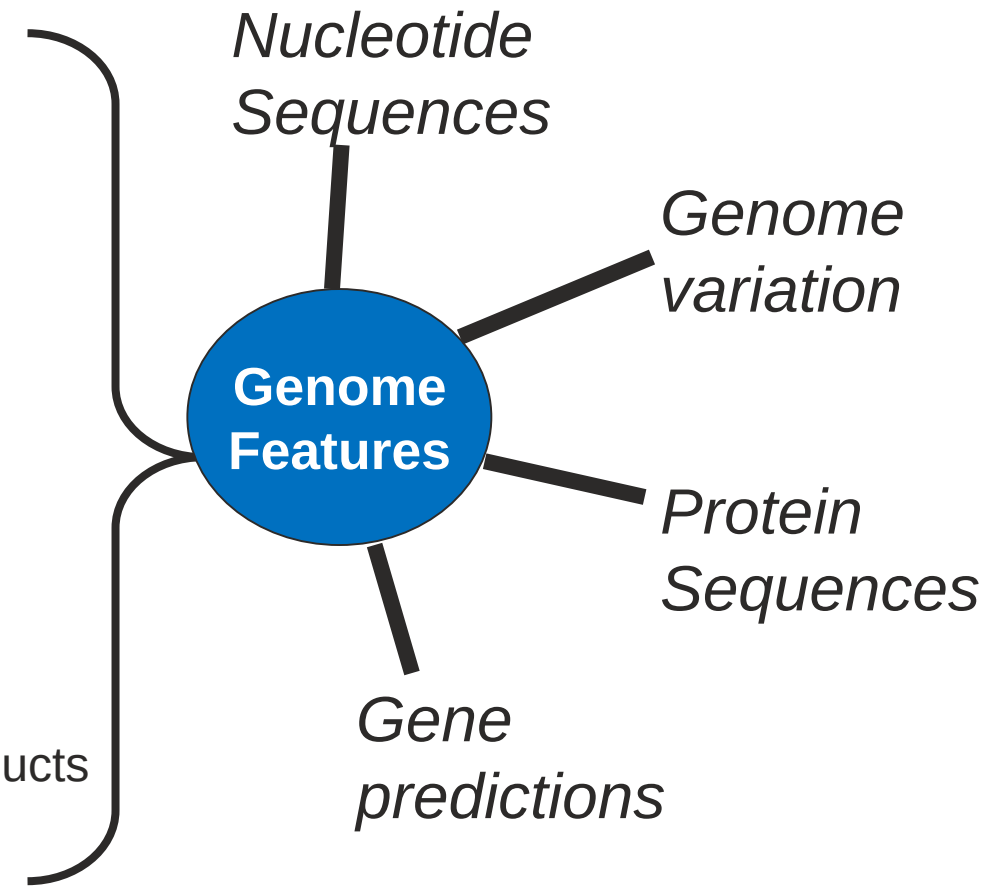




Integrate Sequence with Biology

Biological knowledge and attributes in MGI

- Nomenclature
- Genome location
- Strains
- Polymorphisms
- Orthology
- Expression
- Alleles
- Mutant phenotypes
- Function of gene products
- Literature



Core Sequence Group Functions

- Maintain the unified mouse genome feature catalog
  - Integrate gene predictions from NCBI, Ensembl, VEGA/HAVANA
  - Manage the international mouse genome annotation collaboration
- Maintain associations between MGD/MGI gene records and **external** annotations
  - Ensembl, VEGA, NCBI gene predictions
  - genome variation (SNPs and CCDS)
- Integrate genome features with **internal** annotations
  - alleles and phenotypes
  - gene expression
  - functional annotations
- Support sequence/genome-based access to MGD/MGI via
  - Mouse JBrowse
  - Ensembl, UCSC, and NCBI Map Viewer genome browsers
  - Curated "sequence-to-gene" associations

Search by Gene | Search by Region

Search | Reset | **Search for SNPs by Associated Gene(s)** - from dbSNP Build 142

Associated Genes

Find SNPs Within or Near Specified Genes

Gene Symbol/Name: Search Current symbols

Examples: Fmr1 Pax Dnah5,Dnah6,Dnah7**

For the gene(s) specified above return all SNPs:

- within the gene(s)
- include 2 kb upstream and downstream of the gene(s)
- include 10 kb upstream and downstream of the gene(s)

All SNP function classes will be returned.
You can filter SNPs by function class in the search results.

A.

Strains and Strain Comparisons

Find SNPs Involving these Strains

Select All | Clear All

129S1/SvImJ	BLG2/Ms	FVB	NZW/Lac
129S4/SvJae	BTBR T+ Itpr3tf/J	FVB/NJ	NZW/LacJ
129S6/SvEvTac	BUB	HMI/Ms	PERA
129/Sv	BUB/BnJ	I/LnJ	PERA/EiJ
129X1/Sv	C3H/He	JF1/Ms	PGN2/Ms
129X1/SvJ	C3H/HeJ	KJR/Ms	PL/J
A	C57BL/10J	KK/HlJ	PWD/Ph
A/He	C57BL/6	LG/J	PWD/PhJ
A/HeJ	C57BL/6J	LP/J	RIIIS/J
A/J	C57BLKS/J	MAI/Pas	SAMP1
AKR			
AKR/			
AVZ/			
B10.			
B10.			
BALB			
BALB			
BALB			
BALB			
BALB			
BALB			
BFM/			

Reference Strain for Comparison

No Reference Strain Selected

Return SNPs with alleles in the selected strains:

- Not compared to the Reference Strain
- Different from the Reference Strain
- Same as the Reference Strain

(only applies if a Reference is selected)

Search | Reset

Results

You Searched For...

Gene Symbol/Name: **PAX*** searching current symbols
Include SNPs located: **2 kb upstream/downstream of** specified genes
Selected Strains: **ALL**
Sorted by genome location

Filter SNPs by: dbSNP Function Class

Filtered by:
dbSNP Function Class: Coding-NonSynonymous
Remove All Filters

B.

<< first < prev **1** next > last >> 100
Showing SNP(s) 1 - 8 of 8

Show only selected strains with alleles | **Show all selected strains**

Drag columns to re-order. ⚠ *Column order is reset with Filter or Show option change.*

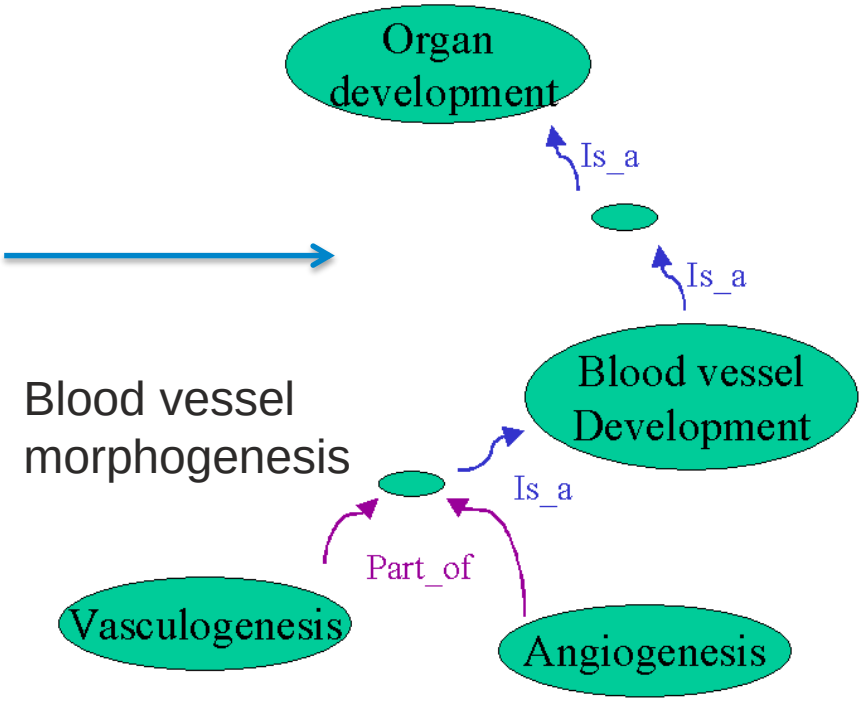
Export: SNPs to Text File

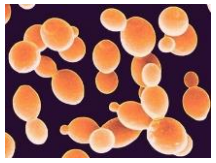
SNP ID (dbSNP Build 142)	Map Position (GRCm38)	Gene & Category	Variation Type	Allele Summary (all strains)	129S1/SvImJ	129X1/SvJ	A/J	AKR/J	BALB/cByJ	BTBR T<+> Itpr3<tf>	C3H/HeJ	C57BL/6J	CAST/EiJ	DBA/2J	FVB/NJ	KK/HlJ	LG/J	MOLF/EiJ	NOD/ShiLtJ	NZW/LacJ	PWD/PhJ	SM/J	WSB/EiJ
rs47736375 MPD | dbSNP | MGI SNP Detail	Chr2:147362490	AI646519 Noncoding-Transcript-Variant Cpgi12386 within coordinates of Pax1 Coding-NonSynonymous Pax1 Locus-Region upstream	SNP	C/T										T			C						C	
rs32116691 MPD | dbSNP | MGI SNP Detail	Chr5:27766191	Del(5D5Mit348-D5Mit389)6Jcs within coordinates of Paxip1 Coding-NonSynonymous	SNP	C/T	C		C	C	C	C	C		C	C	C	C		C	C	C	T		C
rs33824593 MPD | dbSNP | MGI SNP Detail	Chr6:28443529	Pax4 Coding-NonSynonymous	SNP	A/G	G		G	G	G	G	G		A	G	G	G		A	G	G	A		G
rs33826315 MPD | dbSNP | MGI SNP Detail	Chr6:28443910	Pax4 Coding-NonSynonymous Pax4 Intron	SNP	A/G	G		G	G	G	G	G		G	G	G	G		A	G	G	A		G
rs33827147 MPD | dbSNP | MGI SNP Detail	Chr6:28444351	Pax4 Coding-NonSynonymous	SNP	C/T	C		C	C	C	C	C		T	C	C	C		T	C	C	T		T
rs33823858 MPD | dbSNP | MGI SNP Detail	Chr6:28446168	Pax4 Coding-NonSynonymous	SNP	C/T	T		T	T	T	T	T		C	T	T	T		C	T	T	C		
rs4218712	Chr16:91025214	Paxbp1 Coding-NonSynonymous	SNP	C/G	C		C	C	C	C	C	C	C	G	C	C		G	C	C	C		C

We want to place new gene data within context of existing knowledge, but key words aren't enough

Process terms

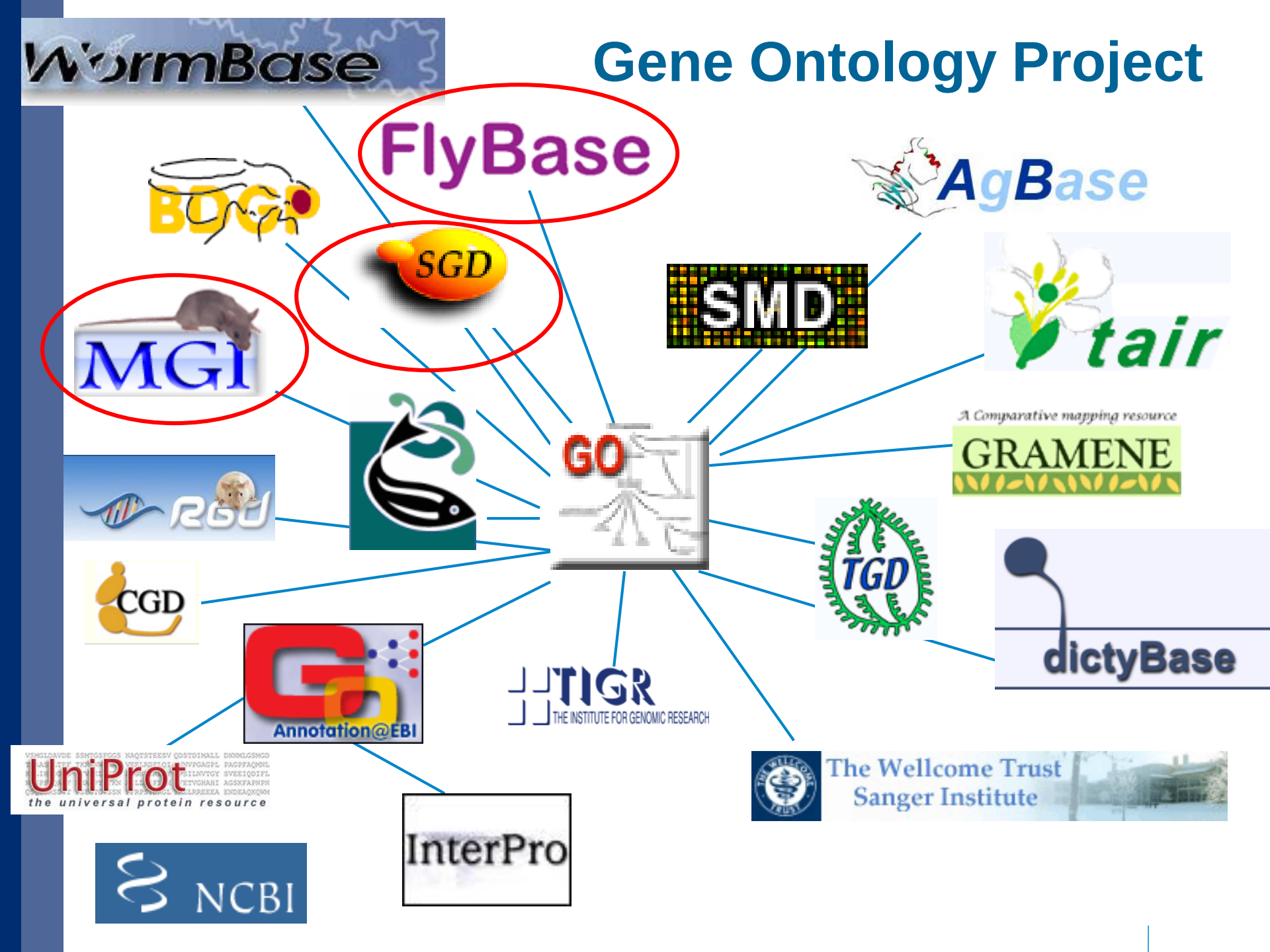
Organogenesis
Blood vessel development
Angiogenesis
Vasculogenesis





All model organism databases and genome annotations streams have the same problem

Gene Ontology Project



Gene Ontology

- Terms are linked by relationships

is_a (is a subtype of)

part_of

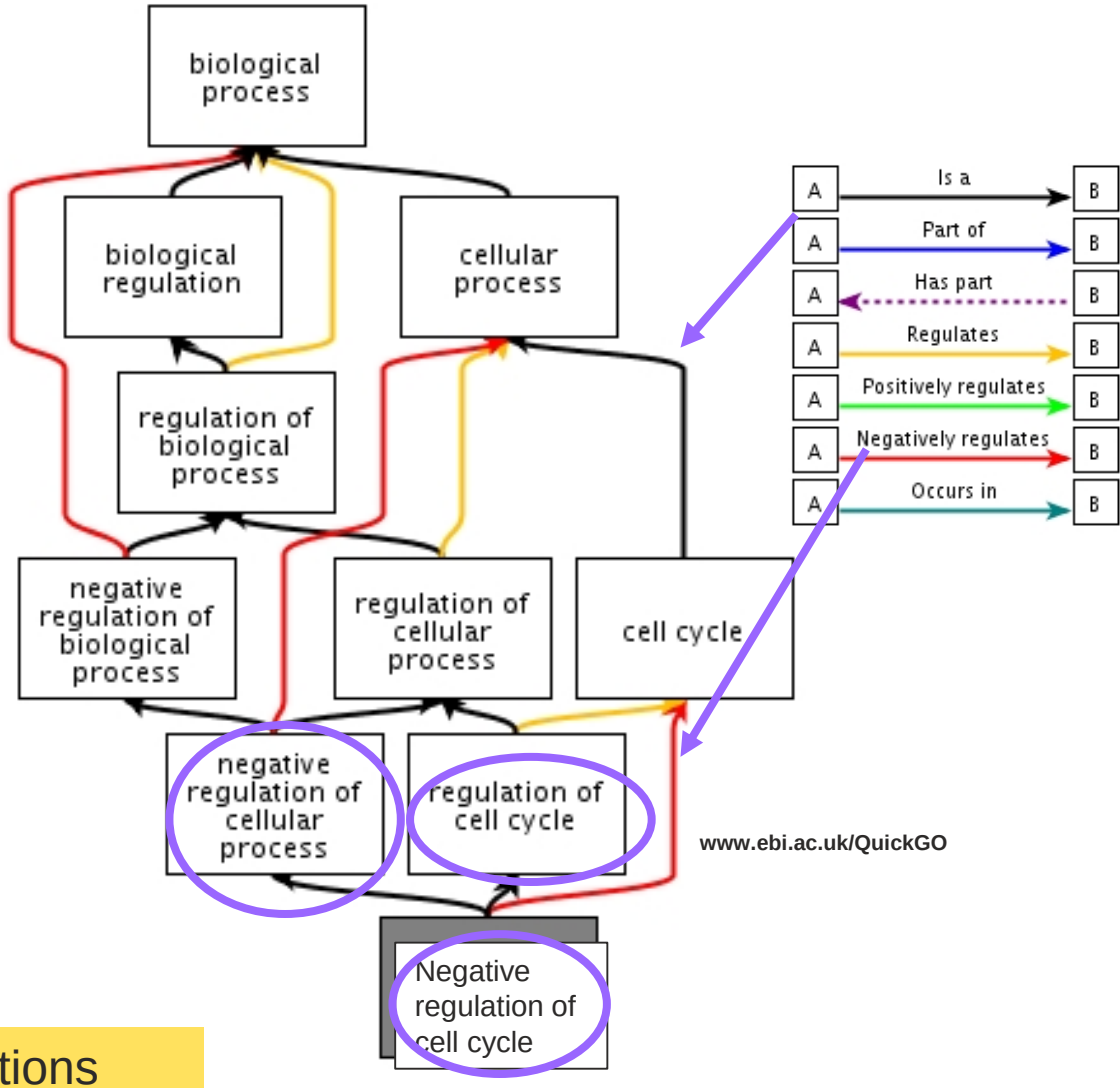
regulates

+ regulates

- regulates

has_part

occurs_in

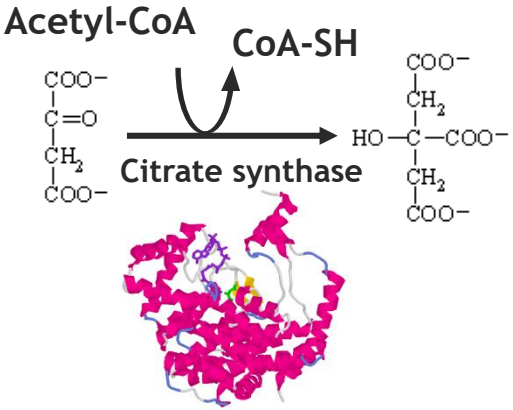


4,280,883 annotations
774,363 gene products
5,633 taxa

Mouse Annotations

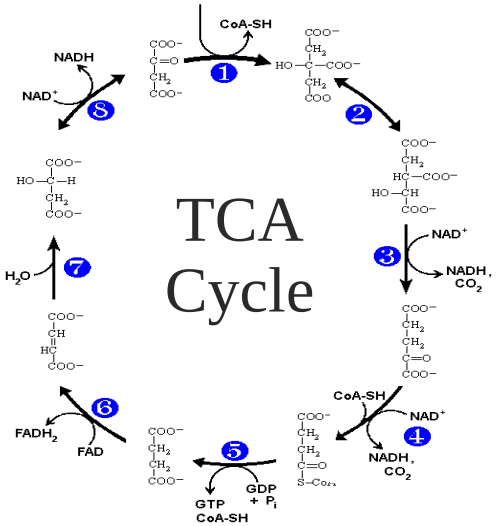
Total Genes:	24,229
Total Annot.:	312,288
Exp. Annot.	86,554
Total Papers:	24,756

Function



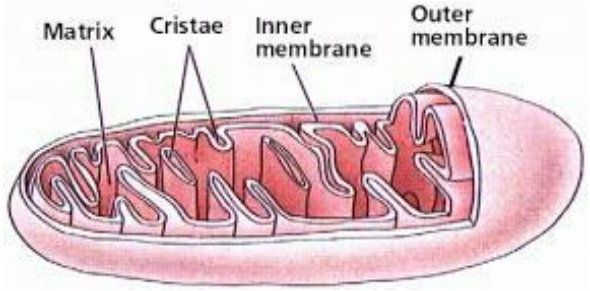
23,398 genes
100,689 annotations

Biological Process



23,883 genes
140,968 annotations

Cellular Component



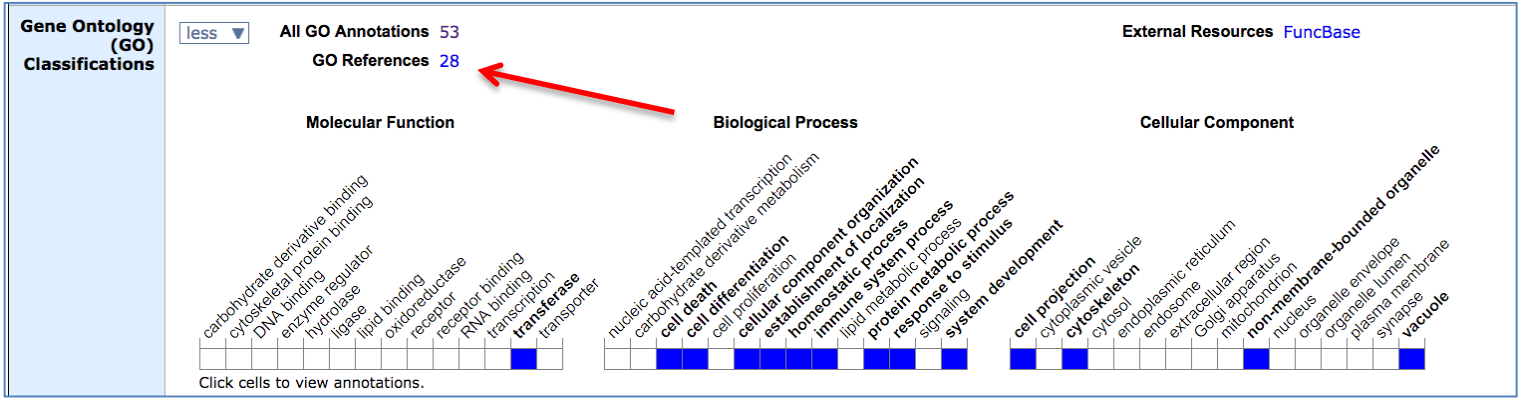
23,828 genes
91,661 annotations

June, 2016

A

Atg5

Autophagy related 5



B

Export: Text File | Excel File

Aspect	Category	Classification Term	Context	Proteoform	Evidence	Inferred From	Reference(s)
Molecular Function	transferase	**contributes_to** Atg8 ligase activity			IBA	PTN000315667	J:161428
Molecular Function		protein binding			IPI	Q9CQY1	J:70539 [PMID:11266458], J:83346 [PMID:12665549], J:164739 [PMID:20723759]
Molecular Function		protein binding			IPI	Q8R1P4	J:80842 [PMID:12482611]
Molecular Function		protein binding	happens in pre-autophagosomal structure membrane. happens in membrane.		IPI	Q8C0J2	J:83346 [PMID:12665549]
Cellular Component		Atg12-Atg5-Atg16 complex			IBA	PTN000315667	J:161428
Cellular Component	vacuole	autophagosome	is a part of germ line stem cell.		IDA		J:70539 [PMID:11266458]
Cellular Component	vacuole	autophagosome			ISA	Q9H1Y0	J:92184 [PMID:15292400]
Cellular Component	cell projection, cytoskeleton, non-membrane-bounded organelle	axoneme			IDA		J:203423 [PMID:24089209]
Cellular Component		cytoplasm	is a part of germ line stem cell.		IDA		J:70539 [PMID:11266458]
Cellular Component		ER-mitochondrion membrane contact site			IDA		J:195131 [PMID:23455425]
Cellular Component		membrane			ISO	Q9H1Y0	J:193725 [PMID:23093945]
Cellular Component		pre-autophagosomal structure membrane			IDA		J:70539 [PMID:11266458], J:195131 [PMID:23455425]
Biological Process		aggrephagy			IMP		J:229367 [PMID:22982048]
Biological Process		aggrephagy			ISO	Q9H1Y0	J:164563
Biological Process	immune system process	antigen processing and presentation of endogenous antigen	happens in medullary thymic epithelial cell.		IMP	MGI:3612279	J:196474 [PMID:23382543]
Biological Process	immune system process	antigen processing and presentation of endogenous antigen	happens in medullary thymic epithelial cell.		IGI	MGI:1338803	J:196474 [PMID:23382543]
Biological Process	cell death	apoptotic process			IEA	KW-0053	J:60000
Biological Process	cellular component organization	autophagosome assembly			IMP		J:70539 [PMID:11266458], J:199115 [PMID:23704209]
Biological Process		autophagy			IMP		J:70539 [PMID:11266458]
Biological Process		autophagy	is a part of cellular response to nitrosative stress.		IMP		J:200700 [PMID:23878245]
Biological Process		blood vessel remodeling	has the participant heart left ventricle.		IMP	MGI:3050453 | MGI:3663625	J:121778 [PMID:17450150]
Biological Process	homeostatic process	cellular homeostasis	happens in liver, regulates the level of gap junction.		IMP	MGI:3612279	J:197783 [PMID:22496425]

Phenotypes and Human Disease Mission

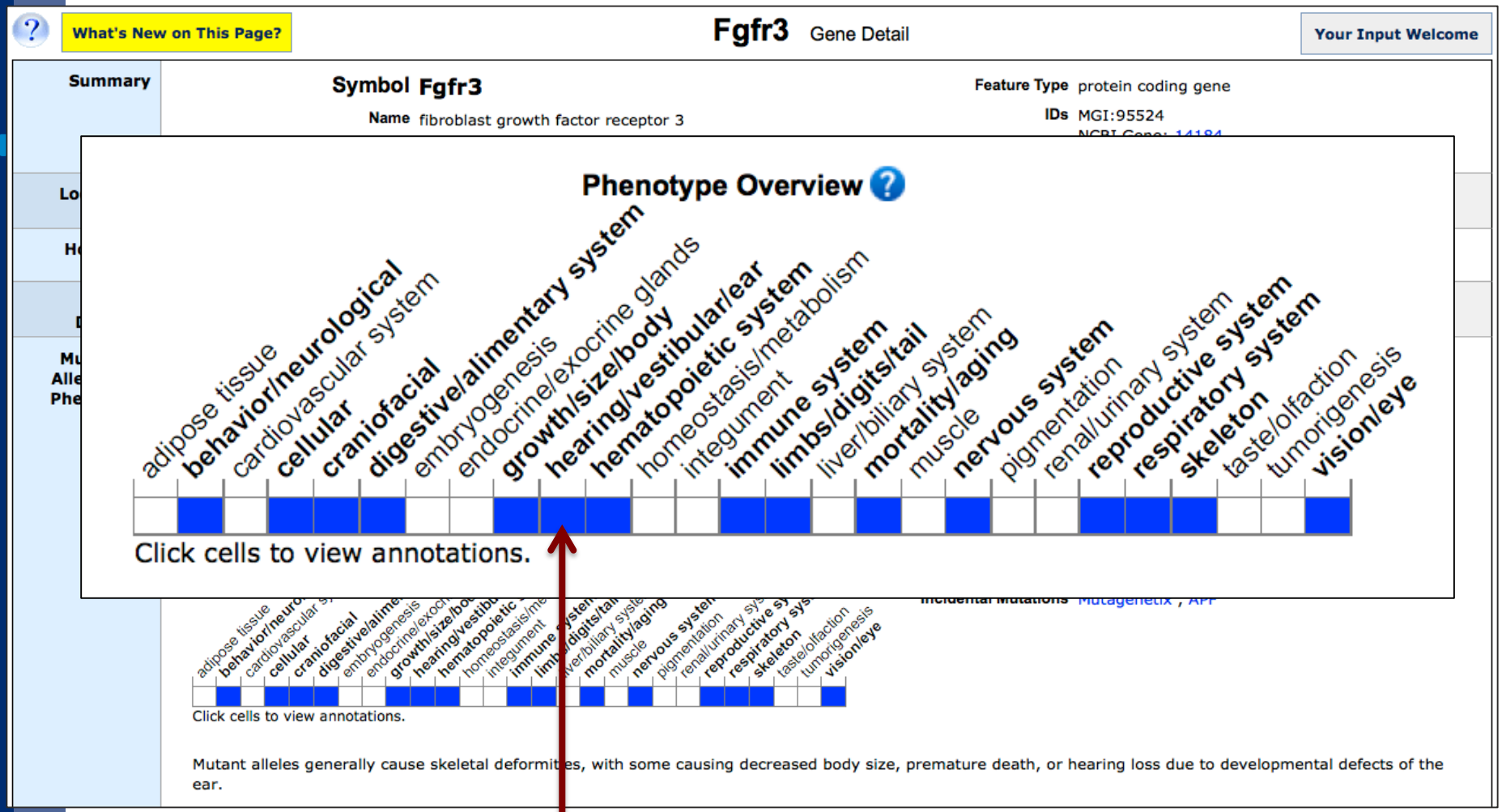
…make comprehensive mouse phenotype and disease model data accessible to researchers, clinicians and computational biologists

- semantic consistency to enable complete data retrieval
- integrated access to all mouse phenotypic variation sources
  (single-gene and genomic mutations, engineered mutations, QTLs, strains)
- data on human disease correlation
- access to mouse phenotype and model data from various approaches
  - Genetic
  - Phenotypic
  - Genomic localization
  - Computational

Status of Phenotype & Disease Data

	May 2015	May 2016	change this yr.
Phenotype terms in MP ontology	10,929	11,366	+437
Mutant alleles cataloged : total	763,035	767,827	+4,792
: in mice	43,602	48,288	+4,686
number of genes represented	23,700	23,742	+42
targeted alleles	56,227	58,702	+2,515
number of genes targeted	16,719	16,822	+103
Alleles w/ phenotype (MP) annotation	37,006	39,464	+2,458
Genotypes with MP annotation	54,996	58,425	+3,429
Total MP annotations	282,370	293,535	+11,165
Mouse genotypes modeling human disease	4,573	4,972	+399
Human Diseases w/1 mouse model(s)	1,359	1,476	+117
QTLs	5,005	10,167	+5162

Gene Level Summary Phenotype Information



Drill down to exp details

Phenotypes Associated with This Genotype

Genotype	Allelic Composition	Find Mice
MGI:4420453 hm2	$Cav1^{tm1Mls}/Cav1^{tm1Mls}$	Using the International Mouse Strain Resource (IMSR)
	Genetic Background: B6.Cg-$Cav1^{tm1Mls}$/J	Mouse lines carrying: $Cav1^{tm1Mls}$ mutation (2 available); any Cav1 mutation (4 available)

♀	phenotype observed in females
♂	phenotype observed in males
N	normal phenotype

cellular

absent caveolae (J:143600)

reproductive system

reduced fertility (J:143600)
- Background Sensitivity: decrease in fertility on a C57BL/6J background

respiratory system

pulmonary edema (J:143600)
- presence of vascular-derived fluid in pulmonary tissues

abnormal lung morphology (J:143600)
- increase in collagen deposition in the lung parenchyma and the periphery of airways
- 60% increase in elastic fiber deposits in the lungs, with thicker layers of elastic fibers primarily around airways and arteries
- however, mutants do not exhibit emphysema

abnormal pulmonary elastic fiber morphology (J:143600)
- 60% increase in elastic fiber deposits in the lungs, with thicker layers of elastic fibers primarily around airways and arteries

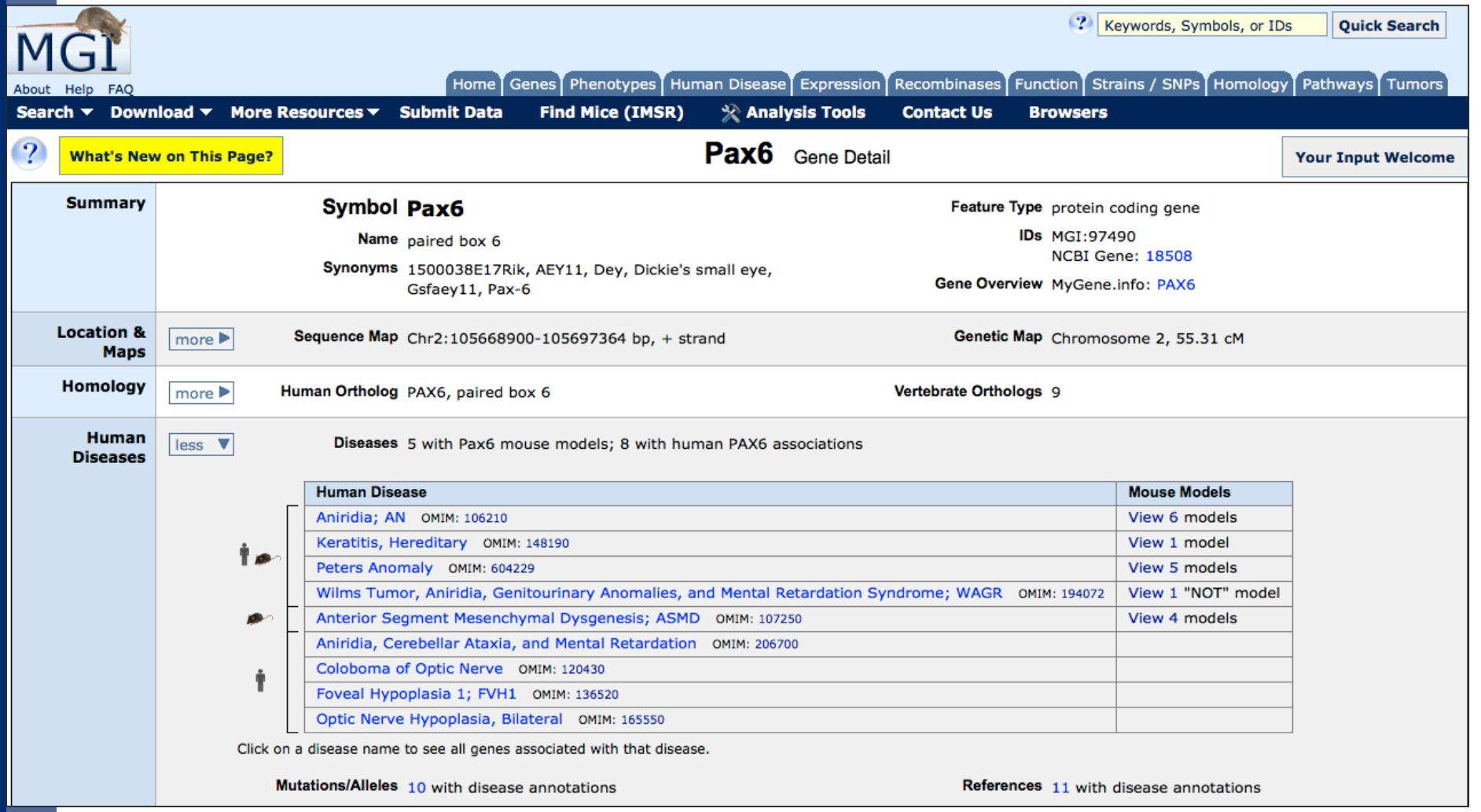
increased lung elastance (J:143600)
- from 3 months on, mutants exhibit a a sustained increase in lung elastance

abnormal respiratory mechanics (J:143600)
- from 3 months on, mutants exhibit altered respiratory mechanics, suggesting stiffening of the lung tissue

increased airway resistance (J:143600)
- from 3 months on, mutants exhibit an increase in airway resistance

All: 154 reference(s)

Human Diseases Section Pages



- table of human diseases OMIM has associated with the human gene as well human diseases modeled in mice with mutations in the gene
- “NOT” models are also shown

Data for which MGD serves as the authoritative source

Data Type	Community Relationship
Unified genome feature catalog	MGD compares/integrates predictions from Ensembl, NCBI, Havana/Vega, produces unified catalog used by NCBI, IMPC…
Gene Ontology (GO) annotations for mouse	MGD does primary curation, integrates data from others, provides definitive mouse GO annotation sets to GO site….
Mouse Phenotype annotations	MGD does primary curation & integrates data from publications & large-scale projects.
Mouse models of human diseases	MGD does primary mouse model curation using disease terms & human gene associations from OMIM and NCBI.
Gene-to-nucleotide sequence association	Co-curation with MGA (Mouse Genome Annotation) group.
Gene-to-protein sequence association	Co-curation with UniProt and Protein Ontology.
Mammalian Phenotype (MP) Ontology	MGD develops & distributes MP. MP is actively used by many groups, e.g., RGD, MRC Harwell, Sanger, IMPC, etc.
Symbols & names for genes & genome features	MGD provides access to International Nomenclature guides, implements policies, coordinates with human and rat groups.
Strain designations	MGD assigns official nomenclature; provides to repositories.

MGD Integrated Data

	September 2016
Number of Genes and Genome Features with nucleotide sequence data	48,285
Number of Genes with protein sequence data	24,682
Number of Mouse genes with Human orthologs	17,102
Number of Mouse genes with Rat orthologs	18,547
Number of Genes with GO annotations	24,237
Total Number of GO annotations	315,086
Number of mutant alleles in mice Genes with targeted mutations	49,038 16,832
Number of QTL	5,493
Number of Genotypes with phenotype annotation (MP)	58,370
Total Number of MP annotations	299,961
Number of Mouse Models (genotypes) associated with Human Diseases	5,021
Number of References in the MGD bibliography	228,740

MGI Software Stats

- 175+ software components (git repos)
- >1 million lines of code
- 25+ Solr indexes on front end
- 75+ types of web pages supported
- 40+ types of data regularly integrated via pipelines
- pipelines

User Statistics

Totals by Year

Calendar Year	PageViews	Sessions	Users
2015	7,641,752	1,483,231	737,998
2014	7,301,998	1,367,272	636,918
2013	7,599,447	1,530,135	675,357

Numbers from Google Analytics

Page view: An instance of an web page being loaded in a browser

Session: A period of time a user is actively engaged with the web site

Users: Defined by unique IP address (will underestimate number of individual users)

User Statistics
FTP site

- 291,903 downloads
- 61,868 distinct files
- 22,424 distinct IP addresses

Commonly accessed reports/data

56,670 MGI_AllGenes.rpt
25,539 index.html (index page for the reports)
5,466 MPheno_OBO.ontology
5,287 HMD_HumanPhenotype.rpt
4,498 mp.owl
4,166 HOM_MouseHumanSequence.rpt
3,842 ALL_CellLine_GeneTrap.rpt
3,442 datasets/incidental_muts/Mutagenix.xlsx
3,417 MGI_PhenotypicAllele.rpt
3,169 MRK_List2.rpt

Overall – Mouse Genome Database

We at MGD…

(i) integrate genetic, genomic, & biological data critical for using the mouse as an experimental model for human biology & disease.

(ii) maintain & enhance MGD as a resource for computational biologists & for translational, clinical, and bench scientists.

(iii) provide training, documentation, & other services to support & reach out to our user community.

(iv) ensure the continued performance, availability and security of MGD's hardware, software, and data.

Thanks

MGI is funded by:

NHGRI grants HG000330, HG002273
NIGMS GM080646
NICHD grant HD062499
NCI grant CA089713

www.informatics.jax.org

The Jackson Laboratory
Leading the search for tomorrow's cures