The Comprehensive Functional Annotation of Mouse Genes and Gene Products using The Gene Ontology (GO)

Li Ni, Mary E. Dolan, Alex D. Diehl, Harold J. Drabkin, David P. Hill, Dmitry Sitnikov and Judith A. Blake
The Jackson Laboratory, Bar Harbor, ME 04609

The Mouse Genome Informatics Database (MGI), integrates genetic, genomic, biological, and phenotypic data about the laboratory mouse to support the discovery of the genetic basis of functional mechanisms underlying heritable diseases. The Gene Ontology (GO), is a set of three structured vocabularies used by model organism databases such as MGI to identify the roles gene products play in the life of an organism. GO provides an ontology-driven functional annotation system that facilitates high-quality gene annotation for all species.

The GO Reference Genome Project is a shared annotation effort among the GO Consortium members who provide annotations for the nine primary model organisms such as mouse, fly and yeast. Starting with the set of genes implicated in human disease processes, the GO curators with the Model Organism Databases are coordinating their efforts in providing comprehensive annotations for the human disease genes and their orthologs. As the curators are simultaneously working on the same set of genes, they are also updating the ontologies, providing an orthology set for these organisms, and improving documentation of the GO annotation processes. The comprehensive annotations of the well-studied model organisms provide broad and deep annotation of the reference genomes and serve as a basis for the annotation of emerging genomes via sequence similarity matrices. Here we report progress for the MGI component of the GO Reference Genome Project.

The Gene Ontology project is supported by NHGRI grant HG0002273. The MGI project is supported by NIH grants HG000330.

Reference genomes

Model organism databases coordinate efforts to provide comprehensive literature-based annotations for selected genes and homologs.

Mouse, Human, Arabidopsis, Dictyostelium, S. pombe, S. cerevisiae, C.elegans, E. coli, Drosophila, Chicken, Zebrafish, Rat

Target genes

Each month a set of target genes is chosen for annotation. The criteria for selection are:

1. **Implicated in human disease;**
2. **Part of a biochemical or signaling pathway;**
3. **Current "hot" genes;**
4. **Conserved homology across many species.**

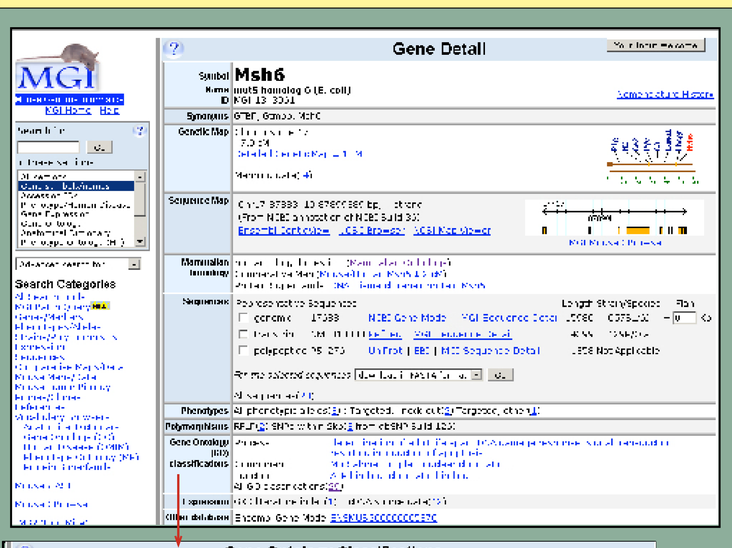
As of October 2007, 273 genes have been selected. One selected gene, human *MSH6*, has homologs in nine other species.

Organism	Gene
Mus musculus	*Msh6*
Rattus musculus	*Msh6-predicted*
Caenorhabditis elegans	*msh-6*
Drosophila melanogaster	*Dmel/CG7003*
Dictostelium discoideum	*MSH6*
Danio rerio	*msh6*
Arabidopsis thaliana	*MSH6, MSH6-1*
Saccharaomyces cerevisiae	*MSH6*
Schizosaccharomyces pombe	*msh6*
Escherichia coli	*mutS*

ACHE	ATXN10	DPYS	MSH2	PIP5K3
ACR	ATXN2	DRD2	MSH3	PML
ACTC1	ATXN3	DRD3	MSH6	POLA1
ACVR1	ATXN7	DRD4	MYBPC3	PPT1
ACVR1B	B4GALN2	DRD5	MYH11	PPT1
ACVR2A	B4GALT1	EIFAK3	MYH6	PRKAG2
ACVRB2	BAX	EIF2B1	MYH7	PSEN1
ACVRL1	BCL10	EIF2B2	MYH8	PSEN2
ADSL	BCL2	EIF2B3	MYH9	RECQL4
AGGF1	BCL3	EIF2B4	MYL2	RMRP
AGXT	BCL6	EIF2B5	MYL3	RPS19
ALAS2	BCL7A	ERCC6	MYLK2	SBDS
ALB	BCL8	ERCC8	MYO1A	SETX
ALDH5A1	BLM	FOXC1	MYO3A	SHROOM3
ALG12	BRCA2	FOXC2	MYO5A	SHROOM4

GO annotations at MGI

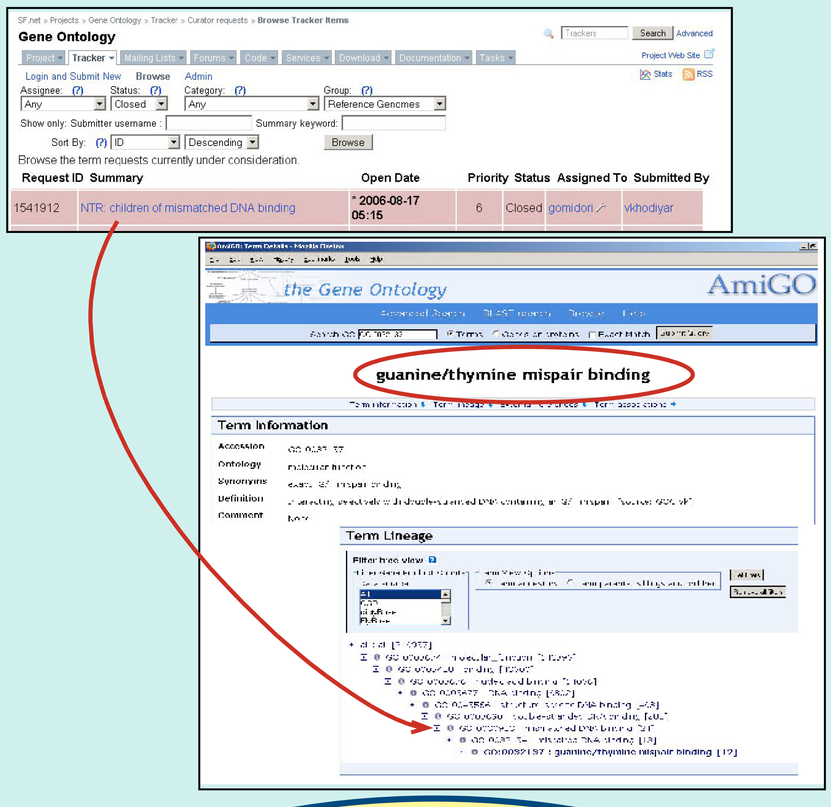
The Mouse Genome Informatics (MGI) Database provides integrated access to data on the genetics, genomics and biology of the laboratory mouse. The gene detail page is the starting point to access data for a particular gene, including GO annotation.



MGI GO annotations are coordinated with other homolog annotations for the selected gene to provide an overview of the gene's function.

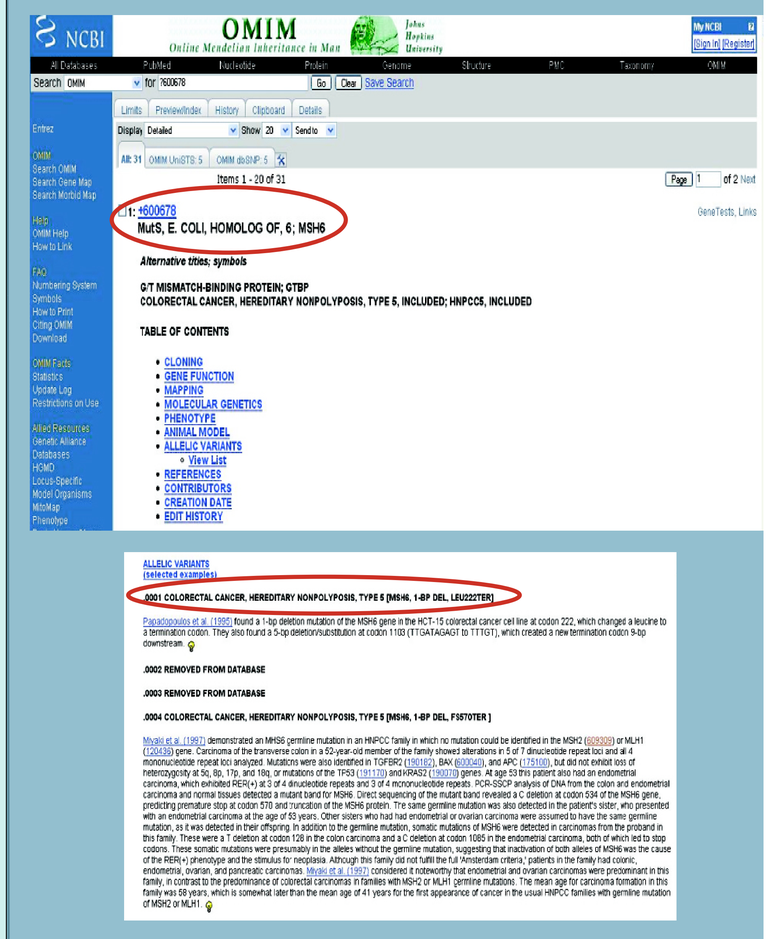
Curation-driven Ontology Improvement

Coordinated curation leads to modification and improvement of the ontology. For example, the new term "guanine/thymine mispair binding" was created for reference genome annotation because experimental data showed that the human *MSH6* gene product, as well as the orthologous genes in mouse, zebrafish and yeast, are involved in this function.



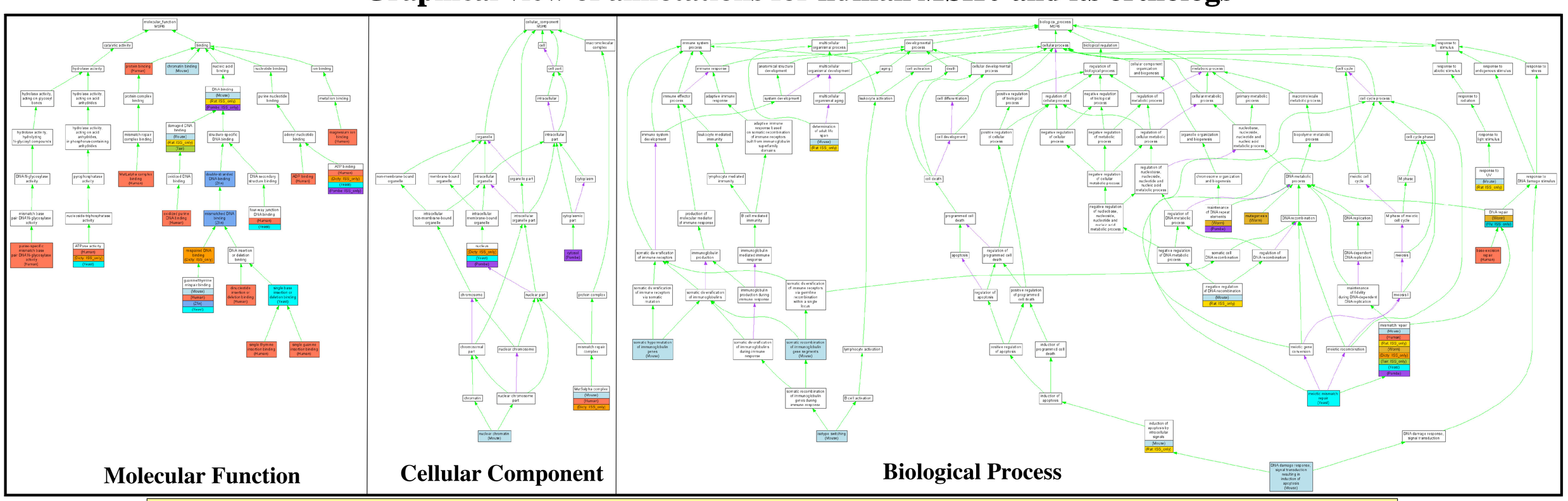
Our goal to achieve comprehensive annotation is an effort to capture all the unique features of a gene product that can be characterized in the Gene Ontology (GO). Detailed comprehensive annotation coordinated across species presents an overview of a gene that a biologist would recognize as describing that particular gene.

Many genes are chosen that are associated with human disease



OMIM records can provide information about the human disease.

Graphical view of annotations for human *MSH6* and its orthologs



Graphical views provide the "big picture" of how genes function in all model organisms. In this view, different colors in the boxes represent annotations from different model organisms.

http://www.informatics.jax.org **http://www.geneontology.org**