# **Original article**

# Manual Gene Ontology annotation workflow at the Mouse Genome Informatics Database

### Harold J. Drabkin\* and Judith A. Blake for the Mouse Genome Informatics Database

The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA

\*Corresponding author: Tel: +1 207 288 6650; Fax: +1 207 288 6131; Email: harold.drabkin@jax.org

Submitted 18 June 2012; Revised 26 September 2012; Accepted 3 October 2012

The Mouse Genome Database, the Gene Expression Database and the Mouse Tumor Biology database are integrated components of the Mouse Genome Informatics (MGI) resource (http://www.informatics.jax.org). The MGI system presents both a consensus view and an experimental view of the knowledge concerning the genetics and genomics of the laboratory mouse. From genotype to phenotype, this information resource integrates information about genes, sequences, maps, expression analyses, alleles, strains and mutant phenotypes. Comparative mammalian data are also presented particularly in regards to the use of the mouse as a model for the investigation of molecular and genetic components of human diseases. These data are collected from literature curation as well as downloads of large datasets (SwissProt, LocusLink, etc.). MGI is one of the founding members of the Gene Ontology (GO) and uses the GO for functional annotation of genes. Here, we discuss the workflow associated with manual GO annotation at MGI, from literature collection to display of the annotations. Peer-reviewed literature is collected mostly from a set of journals available electronically. Selected articles are entered into a master bibliography and indexed to one of eight areas of interest such as 'GO' or 'homology' or 'phenotype'. Each article is then either indexed to a gene already contained in the database or funneled through a separate nomenclature database to add genes. The master bibliography and associated indexing provide information for various curator-reports such as 'papers selected for GO that refer to genes with NO GO annotation'. Once indexed, curators who have expertise in appropriate disciplines enter pertinent information. MGI makes use of several controlled vocabularies that ensure uniform data encoding, enable robust analysis and support the construction of complex queries. These vocabularies range from pick-lists to structured vocabularies such as the GO. All data associations are supported with statements of evidence as well as access to source publications.

### Introduction

Mouse Genome Informatics (MGI) is the primary international database resource for the laboratory mouse, providing integrated genetic, genomic and biological data to facilitate the study of human health and disease. The MGI team curates the biomedical literature (11000 publications per year) and normalizes and integrates sequence and functional data about mouse genetics and genomics from almost 50 other external database and informatics resources. MGI organizes curation teams around particular types of data including sequence data, phenotypes, embryonic expression data, comparative and functional information, mouse tumorigenesis and mouse models for human diseases. MGI utilizes multiple bio-ontologies and is the authority for mouse gene and strain nomenclature.

Five projects contribute to this resource. The 'Mouse Genome Database' (1) includes data on gene characterization, nomenclature, mapping, gene homologies among mammals, sequence links, phenotypes, disease models, allelic variants and mutants and strain data. The 'Gene Expression Database' (2) integrates different types of gene expression information from the mouse and provides

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/bas045/439709 by guest on 19 May 2024

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/ licenses/by-nc/3.0/), which permits non-commercial reuse, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com. Page 1 of 10

### Original article

a searchable index of published experiments on endogenous gene expression during development. The 'Mouse Tumor Biology (3) Database' provides data on the frequency, incidence, genetics and pathology of neoplastic disorders, emphasizing data on tumors that develop characteristically in different genetically defined strains of mice. The MGI group is a founding member of the 'Gene Ontology Consortium' (GO, www.geneontology. org, (4)). MGI fully incorporates the GO in the database and provides a GO browser for access to mouse functional annotation. Finally, the 'MouseCyc' database (5) focuses on 'Mus musculus' metabolism and includes cell level processes such as biosynthesis, degradation, energy production and detoxification. It is part of the BioCyc collection of pathway databases created at SRI International (6).

Here, we outline the workflow process for 'one' component of the MGI data acquisition and integration process that associated with the 'Gene Ontology Project at MGI (Figure 1)'. MGI assigns functional annotations (GO terms) to genes and protein products through semi-automated methods and manual curation. Semi-automated annotation strategies include mapping and translating data from the Enzyme Commission, Swiss-Prot, InterPro (see http:// www.geneontology.org/GO.indices.shtml), rat and human



**Figure 1.** GO curation workflow: papers of interest are identified and entered into the database system (triage) and associated with genes (indexed). GO annotations are made using papers selected based on quality control reports and projects. The quality control reports in turn are revised daily based on added annotation.

ranning Parults Browser To OUOSA Express	Current Channel: PubMed
ganzer results proviser to goodsk express	Search: mouse murine mice 55 Results (Search In Results)
Wig My Searches	▽ △ 🥢 🗔 🏘 📓 🖻 🖉 Q. 💋 🖓 🖓 🖉 🛺 🔯 Q. Q. Annotation 🛟
▶ (m) Fri 03/23/12	Pin No Type Arthors Publiched Title Source
▶ []] Fri 03/16/12	1 More realized in the analysis of the source of the sourc
* 100 Mon 03/12/12	a monthack correct to be thank and be accessed of plot chemic to
(0/9) The Journal of biological chemistry	2 🖏 Blumbach K 2012 Feb 24 Dwarfism in Mice Lacking Collagen-bindir J Biol Chem, 201
Prior Dates	
My Alerts	💉 3 🔁 Zhang L, Do 2012 Feb 24 Role of Integrin-beta3 Protein in Macroph J Biol Chem. 201
My Folders	ag
[000] My Citations	4 Luo H, Wu Z 2012 Feb 24 Receptor tyrosine kinase ephb6 regulates J Biol Chem. 201
	Laboratory of Remodeling-related Cardiovascular Diseases, Beijing Anzhen Hospital, Capital Medical University, Ministry of Education, Beijing 100029, China Background: Integrin: #31 is important for the cell migration and proliferation linked to muscle regeneration. Results: In mice with global integrin:#3 KO, an initial macrophage polarization impairs muscle regeneration and stimulates fibratic stir (JEE-G). and exteriors.
	norosis via 16t-β1 production. <b>Conclusion:</b> In bone marrow cells, integrin-β3 expression is necessary for macrophage-dependent processes of muscle repair. <b>Significance:</b> Stimulating integrin-β3 could improve muscle regeneration.

**Figure 2.** QUOSA Information Manager showing part of a paper from an issue of JBC with mouse, murine or mice highlighted. Curator quickly looks at context to select appropriate area of MGI that the paper best fits. The paper shown has been 'tagged' for alleles and GO.

And in case of the local division of the loc	le C	omands	NLF	1													
Туре	e Artic	le 💷	In NLM	? Ye								Searc	ch Cle	ar Modi	fy Add	Delete	
Revi	iew Status		Peer R	eviewed		Is Review Art	icle?	No	-1			# Record	a 17	6005		Y P	
							_				_	1 Necor	us J-	Du		Data	_
Auth	hors Dui	net M; Fra	nklin V;	: Mak E; L	iao X; Tabas	I; Marcel YL						Creek	ted	heb		10/5/2	011
Titl	le Autop	hagy regul	ates chi	lesterol	efflux from n	macrophage foar	a cells v	ia lysoso	mal aci	d lipase.		Hedi	fied	Tiep		2/6/20	12
100000												nodi	fied .	RIAN		2/6/20.	12
Jou	urnal Cel	1 Hetab										J:		<b>1</b> 7	6089		_ ŕ
	_		-					· · · · · ·	_		_	MGI:	-	52	88298		
Date	te 2011 J	iun 8	Volume 113		Issue	•   <u>5</u>	Page	655-67				PubM	led	21	641547		_ 1
21		Selected	Used	Not Used	Never Used			Selected	d Never	r Used		DOI		10	.1016/j.	cmet.2011	L 5
Pr	robes/Seq			x		Tunor							J: -	Add F	Row Dele	ete Row	
Ma	apping			x		SCC		í	1			72.50				_	
Al	llele/Phe	х		x		1			-			Last	]# 118	31675			
Ho	onology			x										1 Searc	ch Result	ts	
Ex	xpression			x								Duiz	et M. C	ell Metab	2011 1	in 8:13(6	H
GO	0	x		x													
No	onen			x													
Sea	arch Data	Sets Using	: 0 PN	D - OR (	default)												
	and stand a					tar											
1					No	tes					- 17						
											l						
1											M						
a.					Abst	tract					_						
c The	e lipid dro	oplet (LD)	is the	major sit	e of cholest	erol storage in	macroph	age foan	cells ar	nd is a	1						
< cho	olesterul e	esters (CE	s), is ]	liberated	from this or	anelle and del	livered t	o cholest	erol ac	ceptors.							
γ Edit I Titl	tors I le I ce I		_	Publisher	Ĭ		Edit	ion I									
γ Edita Mi Titla At Place	tors I le I ce I		_	Publisher	Ĭ		Edit	ion I									X
Y Edit	tors I			Publisher	I	MGD MarkerModu	Edit Ile ei-4-4	ion I	d, PROD	1_MGI, mge	р.х 						X
Y Edit	tors I le I ce I Coemands	Edit	Util	Publisher	I	MGD MarkerModu	Edit	ion I	id, PROD	1_MGI, mgc	d)						N
Y Edit	tors I le I ce I Commands	Edit Gene	Util	Publisher	Ĭ TE ID HEV-0000	MGD MarkerModu	Edit	ion I 2-41 (hj	d, PROD	1_MGi, mge Syebol	4)	Search (	ilear Hoo	dify AM	Delete		X
Y Edit	tors I le I ce I Commands spe official	Edit Gene =] Chronos	Util 	Publisher ities	TIC 10 HCV20000	MGD MarkerModu Festure Type protein codir	Edit ule ei-4-4	ion I i-2-41 (hj	d, PROD Current Atg5	1_MGI, mgr Sysbol		Search C Records	ilear   Hoo	dify Avid	Delete		M
Y Edita II Titla II Place File arker Typ tatus	tors I le I ce I commands pe official - Atgl	Edit Gene 4 Dhromos	Util 	Publisher ities	Ť TEC 1D HEV200000	MGD MarkerModu Feature Type protein codin	Edit ule ei-4-4	ion I I-2-41 (h)	d, PROD Current jitg5 Rdd Row	1_MGI, mgc Symbol Delete Row		Search C Records	ilear Moo 322542 By Feture	dify Air	Delete Deleter D		N
File	tors I le I ce I Comands pp official - AtgE	Edit Gene I Chromos ted 5 (yeast)	Util	Publisher ities	TE ID HCV:00000	MGD MarkerModu Feature Type protein codin	Edit ule ei-4-4	ion I -2-41 (h)	d, PROD Current Jitg5 Rdd Row	1_MGi, mgc Sgebol <u>Delete Row</u>		Search C Records	ilear Moo 322542 By iretire wwh	dify Rid	Delete ▼ ► 9/2/1998 2/6/2012		R
Edit	tors I le I cee I comands pe official - Atg[ utophagy-relations tits Band I	Edit Gene Chronos ted 5 (yeast) ch	Util	Publisher ities 10 - ]	TIC ID HCV20000	MGD MarkerModu	Edit ule ei-4-4 19 gene	ion I	d, PROD Current Rtg5 Rdd Row	1_MGI, mgc Sysbol Belete Row		Search C Records C Created Modified	ilear Moo 322542 By Fettre Math	dify Roll ad_editors	Belete ▼ ► Bate 9/2/1938 2/6/2012		X
File Symbol Symbol Symbol Symbol	tors I le I cee I commands pe	Edit Gene d Chromos ted 5 (yeast) ch	Util	Publisher ities 10 - 1 I 23.24 Date	TIC ID HCV100000 Harker R	MGD MarkerModu	Edit le ei-4-4 ng gene	ion I -2-41 (h)	d, PROD Current Jitg5 Hdd Row	1_MGI, mgc Symbol Belete Row Rodified B		Search C Records C Records Molified Molified	lear Ho B22542 By retire Neh	dify Rid sd_editors 1277186 1277186	Delete Delete		X
Editi Titli t Place File Symbol Symbol Symbol Symbol Symbol Symbol Symbol Symbol Symbol Symbol Symbol	tors I common I	Edit Gene d Chronos ted 5 (yeast) ct Nase sutophagy	Util 	Publisher ities 10 - 1 I 23.24 Bate • 09/02/1990	I           TIC ID           Horker R           Narker R           10           49052	MGD MarkerModu Feature Type protein codin evision Notes Str itation awond EH, FEBS Les	Edit le ei-4-4 19 gene raim-Specif Event t. assigned	ion I 2-41 (h) ic Marker No Reason Not Spec	d, PROD Current Rtg5 Rdd Row tes	1_MGI, mgc Sgebol Belete Row Hodified By		Search C Records C Created Nodified MGI: HGI:	ilear Hoo S22542 By Fettre sain 1 1	dify Avid ad_editors 1277186 1277148 1317267	Delete                                   		X
Edit Titl Titl Place File Symbol Symbol Symbol Symbol Symbol Symbol Symbol Symbol Symbol Symbol Symbol	tors I commons comm	Edst Gene Chronos ted 5 (yesat) ed Nase sutophags RESEN cINN	Util 	Publisher Ities 10 - 1 I 23.24 Bate 09/02/1990 00/14/2000	Image:	MGD MarkerModu Feature Type protein codin evision Notes Str itation amond DH, FEIS Lei awai J., Nature 2000	Edit le ei-4-4 19 gene naim-Specif Event t. assigned L. assigned	ion I 2-41 (h) ic Marker No Reason Not Spec	d, PROD Current Rtg5 Rtd Row tes	1_MGI, mgc Symbol Belete Row Rodified By dbo		Search C Records C Records Midified MGI: MGI: MGI: EC	lear Moo B22542 By Pretro Moh	dify Pod ad_editors 1277196 1277148 1317267 Idd Row Del	Delete Delete Delete 9/2/1998 2/6/2012 1000 100		X
Edit Edit Titl Place File arker Typ Symbol S	tors I comards pe Comards pe official arggi utophagyrelai sbol 100676248;k 100676248;k	Edst Gene Chromos ted 5 (yeast) ed Name sutophags RINEN cDW RDEN DW	Util	Publisher Ities I0 - I I 23.24 Date 00/02/1988 02/14/2002 11/22/2000	TTC 10 HCV200000 January C January C	MGD MarkerModu Feature Type protein codin evision Notes Sta itation awand EH, FEIS Lef awai J, Nature 2000 outo Leftone Informa-	Edit sie ei-4-4 sig gene sim-Specif Event t- assigned assigned assigned assigned	Ion I IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	d, PROD Current Rtg5 Rtd Row cified cified cified	1_MGI, mgr Sambol Belete Row doo doo doo doo		Search C Records Created Hodified HGI: HGI: HGI: Event	lear Moo B22542 By Fetare path 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	dify Rod ad_editors 1277196 1277148 1317267 Idd Rou Del	Delete Delete Date 3/2/1998 2/6/2012 ete Row		X
File File Sumbol	tors I Comands pe Official - Atgl utophagy-relativitic Band I psi 10067tQ4Rik. 10067tQ4Rik.	Edist Cene Chronool ted 5 (yeast) RUSN cDW RUSN cDW RUSN cDW RUSN cDW expressed expressed	Util	Publisher ities 10 - 1 10 -	I           TIC ID           Harker R           J#           C           49032           B5600           65000           R           69001           00055	MGD MarkerModu	Edit	Ion I I-2-41 (hj I-2-41 (hj))) I-2-41 (hj I-2-41 (hj)) I-2-41 (	d, PROD	1_MGI, mgc Symbol Belete Row doo doo doo doo doo		Search () Records   Notified N	lear ho B2542 By Fetire wh B 1 1 1 1 2 4 2 2 3 2 4 2 3 2 5 2 5 4 2 5 4 2 5 4 2 5 4 2 5 4 2 5 4 2 5 4 2 5 4 2 5 4 2 5 5 4 2 5 5 5 5	dify Post sd_editors 1277186 1277148 1317267 Not 3	Pelete Bate 9/2/1998 2/6/2012 ete Row		X
Edit Edit Titl. Titl. File arker Typ Symbol	tors I Comands pe official Atg8 uutophagy-relat tic Band I g51 10067904Rik 10067904Rik 8837	Edst Gene d Chronos ted 5 (yeart) Name autophags RIXEN CINH RIXEN CINH expressed expressed expressed	Util	Publisher ities 10	ТС ID ТСС ID НСУ 200000 НСУ 20000 НСУ 200000 НСУ 20000 НСУ 200000 НСУ 200000 НСУ 200000 НСУ 200000 НСУ 200000 НСУ 20000	MGD MarkerModu	Edit e e-4-4 y pre Event Event Event Event Event Event Event Event Event Event Event Event Event Event Event Edit	ion I -2-41 (hj C Norker No Reason Not Spee Not Spee Not Spee Not Spee	d, PROD	1_MGI, mgo Syebol Belete Row Hodified Bi doo doo doo doo doo		Search C Records C Redified NG1: NG1: EC Event Event	lear Moo BZ2542 By Fetarona B B B B Fetarona I I I I I I I I I I I I I I I I I I I	dify (84) dify (84) diddeditors 1227166 1227148 1227148 Not 5 1 Secret	Neleta Neleta Sr2/1998 Sr2/199		7
Edit Edit Titl: Ti	tors I Comands pe Comands pe official Atg8 utophagy-relation tic Band I g51 10067424Rik 00572 0537 Delete Row pelete Row	Edist Gene Chromos ted 5 (yeast) Nase sutophage RIXEN LINA RIXEN CINA RIXEN CINA RIXEN CINA RIXEN CINA RIXEN CINA	Utili Utili I Position SilioofHC Silioof	Publisher ities 10	ТС ID Начкет Ре Рессоор Инскорон	MGD MarkerModu Feature Type 004 protein codin protein codin evision Notes Sta litation sevind PL, FEBS Les sevind PL, FEBS Les sevind PL, FEBS Les touse Genome Informs touse Genome Informs	Edit	ion II i-2-41 (hj ic Norker No Reason Not Seet Not Seet Not Seet Not Seet	d, PROD Current htp5 Ndd Row cified cified cified cified	1_MGI, mge Syebol Belete Row Abo dbo dbo dbo dbo		Search C Records C Redified NG1: HG1: Event Event Event Event	ilear Moo B22542 By Fetire meh	1 Serrch	Pelete Bate 9/2/1998 9/2/1998 9/2/1998 etc Bow ipecified		
Edit Edit Titl Titl Place File arker Typ Symbol Sym	tors I Comands pe Comands pe official Atg[ utophagy=relation tic Band I g51 10067t24Rik 8837 Delete Row Delete Row 2 = Reference 1 = Referenc	Edist Cene Chromos ted 5 (yeast) RIXEN CINA RIXEN CINA RIXEN RIXEN CINA RIXEN CINA RIXEN	Util Util	Publisher 10	ТС 10 ТС 10 НС/20000 Н	MGD MarkerModu Feature Type Not protein codin evision Notes Sta itation awond DI, FEIS Lei said J. Nature 2000 tous Genome Inform he Jackson Laborato touse Genome Inform he Jackson Laborato	Edit le ei-4-4 9 gene sain-Specif 6, asigned 6, asigned 6, asigned 6, asigned 10(all etc.)	ion II i=2=41 (h) ic Harker No Reason Not Speci Not Speci No	d, PROD) Current Res Res Cified Cified Cified Cified	1_MGI, mge Symbol Belete Row doo doo doo doo doo		Search C Records C Records I NG1: NG1: Event Event Event Event	lear Moo B22542 By Fetire Meh	1 Search	Belete Bate 9/2/1990 9/2/1990 ete Rou iete Rou iete Rou iete Rou		
t Editi Titli t Place t Pla	tors I commands	Edist Cene Chronos ted 5 (yeset) Nase sutophagy RIXEN CINN RIXEN CINN RIXEN RIXEN CINN RIXE	Util Util Util Position S-like (S. Siloshud Silosh	Publisher 10 - 1 1 23.24 Date 03/14/2003 02/14/2003 11/22/2002 03/14/2003 02/14/20	I           The ID           Herker Ro           J           4932           4932           6560           Rother Ro           10           11           65500           8001           6001           10055           10           1015	MGD MarkerModu Feature Type 004 protein codin evision Notes Sta itation Newmond EN, FEIS Lef awai J, Nature 2000 use Genome Inform he Jackson Laboratz buse Genome Inform i.), 5 = Accession	Edit le ei-4-4 9 pre 5 p	Ion II 2-41 (n) 2-41 (n) 	d, PROD Current Pita5 Rdd Rou cified cified cified cified	L_MGI, mge Sysbol Belete Rou Boo doo doo doo doo doo		Search C Records C Records NG1: NG1: Event Event Event Event	lear No B22542 Py Fetire Path 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	arfy Arf	Palata Bata Bata 9/2/1398 9/2/1398 9/2/1398 etc Row etc Row pecified		-
Edit     Edit     Titl	tors I communications commun	Edit Cene Chronos ted 5 (yeast) RISEN cliw RISEN cliw R	Util Util Position S-Tike (S. 311067Pd2 SEQUECE C SEQUECE C	Publisher 10	I           The ID           The ID           Important           Important <td>MGD MarkerModu Feature Type 004 protein codin evision Notes Sti itation lawood CH, FEIS Let isau J, Nature 200 he Jackson Laborat huse Genoe Inform he Jackson Laborat</td> <td>Edit</td> <td>Ion I -2-41 (n) ic Harker No Reason Not Spee Not Spee Not Spee Not Spee Not Spee</td> <td>d, PROD</td> <td>L MGI, mgg Sgabol Belete Row doo doo doo doo doo</td> <td></td> <td>Search C Records C Records C Noti: Noti: Noti: Ecrent Ecrent Ecrent</td> <td>lear No 32542 By Fetire meh 3 1 1 1 1</td> <td>stry out adjusters stry out adjusters stry</td> <td>Belete Bate 9/2/1998 9/2/1997 9/2/1998 9/2/1997 9/2/10 9/2 9/2/10 9/2/10 9/2 9/2/10 9/2 9/2/10 9/2/10 9/2 9/2/10 9/2/10 9/2 9/2/10 9/2/10 9/2 9/2/10 9/2 9/2 9/2 9/2 9/2 9/2 9/2 9/2 9/2 9/2</td> <td></td> <td></td>	MGD MarkerModu Feature Type 004 protein codin evision Notes Sti itation lawood CH, FEIS Let isau J, Nature 200 he Jackson Laborat huse Genoe Inform he Jackson Laborat	Edit	Ion I -2-41 (n) ic Harker No Reason Not Spee Not Spee Not Spee Not Spee Not Spee	d, PROD	L MGI, mgg Sgabol Belete Row doo doo doo doo doo		Search C Records C Records C Noti: Noti: Noti: Ecrent Ecrent Ecrent	lear No 32542 By Fetire meh 3 1 1 1 1	stry out adjusters stry out adjusters stry	Belete Bate 9/2/1998 9/2/1997 9/2/1998 9/2/1997 9/2/10 9/2 9/2/10 9/2/10 9/2 9/2/10 9/2 9/2/10 9/2/10 9/2 9/2/10 9/2/10 9/2 9/2/10 9/2/10 9/2 9/2/10 9/2 9/2 9/2 9/2 9/2 9/2 9/2 9/2 9/2 9/2		
Edit Titl: Titl: File File File File File File File File	tors I Comands pe Comands pe official atophagyrela	Edst Cene Chronos ted 5 (yeast sutophags RIEEN cINN RIEEN cINN CINN RIEEN cINN RIEEN CINN CINN CINN RIEEN CINN RIEEN CINN CINN RIEEN CINN RIEEN CINN CINN RIEEN CINN RIEEN CINN RIEEN CINN RIEEN CINN RIEEN CINN	Util Util Position 5-like (S, 31106/PRG 31106/PRG 31106/PRG 31106/PRG 31106/PRG 31106/PRG 31006/PRG 3106/PRG 3106/PRG 3106/PRG 3106/PRG 3106/PRG 3106/	Publisher 11 10	I           The IB           JB           C           JB           C           JB           C           JB           C           SSG60           Thread           SSG60           C           SSG60           SSG60           SSG60	MGD MarkerModu Feature Type Vision Notes Str itation Second EN, FEBS Le itation He Jackson Laforat house Genome Inform I, J, 5 = Accession Jul;20(13);2275-68	Edit	Ion I -2-41 (n) ic Harker No Reason Not Spec Not Spec Not Spec Not Spec	d, PROD) Current Rdd Row Rdd Row Crified Crified Crified	L MGI, mgg Sysbol Belete Row doo doo doo doo		Search C Records C Records NG1: NG1: NG1: NG1: ECreated NG1: NG1: ECreated NG1: NG1: NG1: NG1: NG1: NG1: NG1: NG1:	lear Moo S22542 Petersen Mon S2542 Petersen Mon S2542 Mon S254 Mon Mon S254 Mon S2 Mon S2 Mon S254 Mon S25 Mon S25 Mon S254 Mon S254 Mon S254 Mon S25 Mon S2 Mon S25 M	stify PAU A states states	Belete Belete 9/2/1998		
Edit Titl Titl File File Subol Subol Subol Subol Subol Subol Subol Subol Tipeg 2 311 3 311 4 C88 5 C88 Add Rou English Contonent English Subol Tipeg 2 311 3 311 4 C88 5 C88 Add Rou Subol Tipeg 2 311 5 C88 Add Rou Subol Sub	tors I Comands pe Comands pe control of the second second of the second of the secon	Edit Gene Chronos ted 5 (yeset RIXEN cDW RIXEN	Util: Util: Position Position Position Position SiliooSH2 S	Publisher ities 10	I The The Re Harter Re I 4932 H 6566 K 10001 H 69300 T 00155 H 10155 H 10155 H 10155 H 10155 H	MGD MarkerModu Feature Type Protein codir protein codir itation amond EH, FEBS Le amin J, Nature 200 iouse Genome Inform i,,, 5 = Accession July20(13):2275-65 Nov 1131(3):527-65 Nov 1131(5) Nov 1131(5):527-65 Nov 1131(5):527-55 Nov 1131(5):527-55 Nov 1131(5):527-55 Nov 1131(5):527 Nov 1131(5):527-55 Nov 1131(5) Nov 110 Nov 1	Edit	Ion I -2-41 (h) is Marker No Reason Not Spee Not Spee Not Spee Not Spee	d, PROD	MGI, mgg Sysbol Belete Row dbo dbo dbo dbo		Search C Records C Created Hodified HGI: HGI: HGI: HGI: HGI: HGI: HGI: HGI:	itar Mo B2542 By Fetrese satisfied	titig elsi diaditors 127746 127746 dd Rou Bell 1 Search	Notes Notes 9/2/1988 2/6/2012 2/6/2012 2/6/2012 2/6/2012 2/6/2012 2/6/2012		
Editi Titl Titl Place File Subol Sub	tors I Comards pe Comards pe controls comards pe controls comards pe controls comards pe controls p	Edit Gene Chronos ted 5 (yeast) RDEN cINA RDEN cINA RDEN cINA RDEN cINA RDEN cINA RDEN cINA RDEN cINA RDEN cINA RDEN cINA	Utili Utili Position Po	Publisher Ities 10 - 1 10 -	I           II           III           IIII           IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	MGD MarkerModu Feature Type evision Notes stitution awanod EN, FEB Le awai J. Nature 200 buse Genome Inform he Jackson Laborat. buse Genome Inform Juli;30(13):8275-65 Nov. 1131(3):857- 11 Jan 10;128(1):17	Edit e-4-4 % gene sam Specif La assigned La assigned L	Ion I -2-41 (h) ic Harker No Reason Not Spee Not Spee Not Spee Not Spee	cified cified cified cified	A MGI, mgs Sapbol Belete Row dbo dbo dbo		Search C Records f Records f Hol: NGI: NGI: Ecent Event Event Event	itear Moo	41fg (24) 41fg (24) 2277166 2277166 2277166 2277166 1 Seerch 1 Seerch	Nelete 9/2/1980 2/6/2012 2/6/20 2/6/2012 2/700 2/70		
Edit Titl Titl Place File Subol Subo	tors I common I	Edit Gene Chromos ted 5 (yeast) RDEN cINA RDEN cINA RDEN cINA RDEN cINA RDEN cINA RDEN cINA RDEN cINA RDEN cINA RDEN cINA RDEN cINA	Util Util Distion 5-like (S.C. 31106/HC 31106/HC 31106/HC 31106/HC 106	Publisher ities 10 - 1 10 -	I           II           III           IIII           IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	MGD MarkerModu Feature Type evision Notes Station aevond EN, FEB Le aevision Notes Station aevond EN, FEB Le aevision Liborat buse Genome Inform - Jackson Laborat Juli 50(13):5275-65 > New 1:191(3):537- I Jan 10:192(3):537- I Jan 10:192(3):537-	Edit	Ion I -2-41 (h) -2-41 (h) 	d, PROD Current Hts5 Hdd Row tes cified cified cified cified	MGI, mge Syebol Belete Row doo doo doo doo		Search C Records F Records F Roll Field NoI: NoI: NoI: Event Event Event Event Created Roll Field Roll Field R	itear Mo B2542 By Petare pah 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	4115y Part   4115y Part   1277106 1277107 10	Nelete Sr2/1988 Sr2/198		
File File	tors I Comands pe Comands pe Comands pe official - Artg4 witchagy-relations artg4 witchagy-relations artg4 witchagy-relations artg4 witchagy-relations artg4 witchagy-relations artg4 witchagy-relations artg4 witchagy-relations artg4 witchagy-relations artg4 witchagy-relations artg4 witchagy-relations artg4 ar	Edit Cene Chronos ted 5 (yeart) Name sutophagy RDSN cDW RDSN cDW C	Util one	Publisher ities 10	I           II           III           IIII           IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	MGD MarkerModu MGD MarkerModu evision Notes Str itation lamond EH, FEB Le and J. Nature 2000 louse Genome Informs he Jackson Laborat louse Genome Informs i,), 5 = Accession Juli20(13):8275-65 0 Nov 11:191(3):827- 11 Jan 10:191(3):827- 11 Jan 10:191(3):827- 11 Jan 10:191(3):827- 11 Jan 10:191(3):827- 11 Jan 10:191(3):827- 15 0 Nov 11:191(3):827- 15 0 N	Edit e Edit g gene ann-Specif asigned a marged asigned asig	ion II i-2-41 (h) ic Harker No Reason Not Seet Not	d, PROD	L_MGI, mge Syebol Belete Row doo doo doo doo		Search C Records C Records I Roll: Mil: Event Event Event Event Event Crosscond Event	al delets	atify Pari d. d. aditors 1277166 1227246 1317267 1 Search 1 Search 1 Search 1 Search 1 Search	Palata Jate 9/2/1990 2/6/2012 2/7012 2/6/2012 2/6/2012 2/6/2012 2/6/2012 2/6/2012 2/		
Edit Titl: Titl: File Sumbol S	tors I Comands Comands pe Comands pe official - Attgl official	Edist Gene d Chromos ted 5 (year) autophage RIXEN CIW RIXEN CIW RIXEN CIW RIXEN CIW RIXEN CIW RIXEN CIW RIXEN CIW RIXEN CIW RESC RIXEN CIW RIXEN CIW RESC RIXEN CIW RESC RIXEN CIW RESC RIXEN CIW RESC RIXEN CIW RESC RIXEN CIW RESC RIXEN CIW RESC RIXEN CIW RESC RIXEN RIXEN CIW RESC RIXEN RIXEN CIW RESC RIXEN CIW RESC RIXEN RI	Utili	Publisher  Ities  I I I I I I I I I I I I I I I I I I	I           II           III           IIII           IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	MGD MarkerModu Feature Type ovision Notes svision Notes Station awond EN, FEBS Le awai J. Nature 200 toue Genose Inform backson Laborat toue Genose Inform Jul;30(13);2275-05 Nov 1:131(3):557-1 Jan 10:132(3):557-1 Jan 10:122(3):557-1 Jan	Edit e e=4-4 g gene =10-Specif 5 Econt 6 enryed 6 enryed 10 (all other 5 2 -27 -27 -27 -27 -2011 Re 18	Ion II IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	d, PROD	1_MGI, mge Sysbol Belete Rou doo doo doo		Search C Records   Created NoI: NGI: Deated NGI: Event Even	al deleti al deleti al transa	stify Pari deletitors deletitors 1927166 19272746 19272746 19272746 19272746 19272746 19272746 19272746 1927246 19274 192746 19274 19274 192746 19274	Notes Sr2/1988 Sr2/19		
Edit Titl: Titl: File Subol Su	tors I common	Edist Gene Chromos ted 5 (yeast) Name sutophagy RIXEN CINN RIXEN CINN RIXEN CINN RIXEN CINN RUSEN CINN RUSEN R	Utili () Position 5-1ike (S., 311067k2 311067k2 311067k2 157067k2 16712 16725 16725 17854 17854 17854	Publisher Ities I I I I I I I I I I I I I I I I I I I	I           Inc 10           Horker R           Image: Inc 10           Horker R           Image: Inc 10           Im	MGD MarkerModu	Edit e ei-4-4 9 gene rain-Specif 5 werged 5 werged 10 (all oth 5 werged 10 (all oth 5 werged 10 (all oth 5 werged 10 (all oth 10 werged 10 w	10n 1 2-41 (n) 2-41 (n) 	d, PROD	L MGI, mge Sysbol Belete Rou doo doo doo doo		Search C Records	ilear Moo B22542 By Peter pah 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4	stify Poil A	Belete Belete S72/1998		

**Figure 3.** (A) MGI Master Bib EI, showing a record for a paper that has been selected for alleles/phenotypes and GO. The paper has not yet been curated for GO (indicated by an X in the selected and X in Not Used). (B) MGI Gene Feature Detail Module. References to be associated with this gene are entered into the lower left hand panel. If a paper is associated with multiple genes, the same paper is entered for each gene. All EIs are TeleUSE graphical user interface applications running under X-windows.

ortholog experimental data and others. Curation of these data sets includes review and resolution of Quality Control Reports generated through the process of data loading and comparisons to existing data in MGI. For the purpose of this paper, we will not discuss these semi-automated integration methods, but rather concentrate on the manual literature curation, which is a vital source of experimental mouse functional data. While we focus on the GO component of MGI in this description, the literature curation process is very similar for other MGI components that curate literature.

The subcomponents of the literature curation process include the following:

(a) Literature triage: identifying and obtaining relevant scientific literature

Each curator is assigned a specific subset of journals from a set of 160 relevant journals out of  ${\sim}650$ 

subscriptions carried by the Jackson Laboratory. Journals are chosen for triage based on the numbers of articles that have been identified and curated for that journal over the previous years. The manuscripts accepted for curation come from a variety of sources. Most are selected from the journals that are regularly triaged. Others are selected on the basis of particular annotation processes such as full curation of the Wnt family of proteins, or in another example, full curation of genes implicated in human diseases. On a yearly basis, the number of papers selected from all journals is tabulated, and the selection of which journals to regularly curate the next year is determined based on the relevancy of the journal publications during the previous year and the number of full time equivalents available for this task. Typically, a few journals are

### PROMINER used for indexing aid

### EXII and EKK immunoblot analysis

Plasma anticoagulated with sodium citrate was fractionated on 4%-12% gradient SDS-polyacrylamide gels (Invitrogen Life Technologies) followed by immunoblotting with human INII (Accurate Chemicals), mouse PKK (R&D Systems) or mouse  $\alpha$ 2-antiplasmin ( $\Delta$ 2 $\Delta$ P; R&D Systems) antibodies. Blots were incubated with secondary fluorophore-labeled antibodies (LI-COR) and imaged on Odyssey Imager (LI-COR). PKK and FXII relative plasma protein levels were determined by densitometry analysis (Imagel 1.43).

### Plasma fXIIa-antithrombin complex ELISA

EXIIa-antithrombin plasma complex levels were measured by sandwich ELISA. Briefly, assay plates were coated with anti-human fXII antibody (Accurate Chemicals) and blocked with 2% BSA before incubation with diluted mouse platelet poor plasma. After extensive washing, fXIIa-antithrombin complex was detected by incubation with HRP-conjugated mitihrombin antibody (Enzygnost human TAT Micro ELISA kit, Siemens). Relative levels of fXIIa-antithrombin complex were calculated using serial dilutions of control mouse plasma as a standard.

#### Ferric chloride-induced inferior vena cava thrombosis

Antithrombotic activity was studied using a well-established ferric chloride (FeCl<sub>3</sub>) induced inferior vena cava (IVC) thrombosis model.<sup>17,18</sup> Total mRNA was purified from vena cava tissue samples and analyzed by RT-PCR for **Pineter Facture (IPM)** mRNA levels. **We** mRNA levels were used to determine the effect of treatment on platelet deposition as a measure of thrombus formation.<sup>19</sup> **IPM** mRNA levels in the IVC tissue exposed to FeCl<sub>3</sub> was normalized to nonexposed vena cava tissue.

#### Stenosis-induced IVC thrombosis

The St Tomas model which uses a combination of reduced blood flow and endothelial damage, was used to study stenosis-induced IVC thrombosis.<sup>18</sup> Briefly, the IVC of male BALB/c mice anaesthetized with 2.5% inhalant isoflurane was exposed via a midline abdominal incision below the left renal vein, and separated from the abdominal aorta. A 6-0 silk tie (Ethicon) was placed behind the vessel and a metal 4-0 suture (Ethicon) was placed Data were analyzed using SPSS software package for Windows Version 14.0 (SPSS Inc). Graphics were constructed using GraphPad Prism Version 5 for Windows (GraphPad Software Inc).

#### Results

#### Systemic delivery of ASO results in suppression of IXII and PKK levels in mice

The role of the intrinsic pathway and contact system in thrombosis in mice has been studied using several approaches, including knockout strategies and specific inhibitors.<sup>9,11,13,20,21</sup> It is clear that fXII is involved in thrombus formation in these models. In vitro,

PKK is required for optimal PKK may have a pro-throm knockout mice are not curren was used to explore the impo effect of PKK and fXII ASO to fXII mRNA in liver, and levels are shown in Figure 1. System twice per week for 3 weeks) of

CLASS: Mouse Found: fXII F12 Q80YC5@SWISSPROT FA12\_MOUSE@SWISSPROT 58992@ENTREZGENE

reduction of target mRNA levels (92% and 95% reduction for PKK and fXII expression, respectively) at the highest tested ASO dose (Figure 1A). ASO treatment produced dose-dependent reduction of the target plasma protein, with maximum reductions of 83% and 85% for PKK and fXII, respectively (Figure 1B-D). Both PKK and fXII ASOs were highly specific for their targets, as demonstrated by unchanged liver mRNA expression of several nontargeted coagulation factor mRNAs. Specifically, fXII ASO did not change mRNA expression of **Examplation factors** II (prothembin), V, VII, VIII, IX, X, IA, as well as **TAFI** and **PKK** ASOs did not change mRNA expression of **Examplation factors** II (prothembin), VII, XI and XII (Figure 1A and data not shown). Both **PKK** and fXII mRNA levels correlated with plasma protein levels of the respective factors (supplemental Figure 1A-C, available on the

Figure 4. Output from PROMINER: PROMINER is used to assist in the gene indexing process. Utilizing official nomenclatures and synonym lists for mouse/rat/human gene names and gene symbols, PROMINER marks up papers for review by a curator, who then associates genes-to-papers in the MGI EI.

dropped and a few are added each year to the formal triage process. The QUOSA application (http://www.quosa.com) is used as an aid to access and identify recent papers as represented in PUBMED containing data about the mouse and determining which component of the database will be curated from the paper (GO, expression, mutant alleles, phenotypes, mapping, tumor). QUOSA is used to retrieve full text PDFs for a journal issue or time frame of interest. Because the experimental organism 'mouse' is often not mentioned in the abstract (3-34% depending on the journal), a curator selects an issue of a journal and then searches the full text of papers containing the keywords 'mouse', 'murine' or 'mice'. The application highlights these terms showing the context of the search keywords, which enables curators to guickly determine whether the experiments described are of a suitable nature to be used for GO annotation (i.e. do the experiments aid in determining the normal function of the gene?) (Figure 2). The number of papers examined and the number selected are somewhat journal dependent. For example, out of  $\sim$ 80–90 papers per each weekly issue of Journal of Biological Chemistry, roughly 60 contain one or more of the three keywords, and of those, the curator may select 10-15 as being relevant for some area of the database. For an issue of Nature, which may have 15-20 research articles per issue, up to 5 may have the keywords and all of them are relevant. Papers selected are uploaded to an in-house server for the next steps.

(b) Adding the publications to the MGI system through the editorial interface

Papers selected as containing data appropriate for GO annotation are entered into a 'master bibliography' module. Figure 3a shows the module's data entry screen. Each record is tagged for the area of use of the database for which it has information. Information about journal, volume, pages and the abstract are automatically obtained nightly from PUBMED using the PMID. The tag for the database area is added manually. At this stage, the paper is not associated with any specific gene. The area tag will automatically change when the paper is actually used for curation. For example, if this paper was used for a GO annotation, the 'used' box would then have an X added automatically.

(c) Indexing the papers to determine the genes being studied

A paper is then associated or indexed to the genes discussed within by adding the paper to each gene's detail module (Figure 3b). PROMINER,

a natural language processing (NLP) application, is used to assist in the gene indexing process (7). Utilizing official nomenclatures and synonym lists for mouse/rat/human gene names and gene symbols, PROMINER marks up papers for review by a curator, who then associates genes-to-papers in the MGI editorial interface (EI) (Figure 4).

Sometimes a paper discusses several genes, but not all of them may be objects for direct GO annotation. For example, a paper describing the effects of a knock out of a particular gene may use analysis of other gene products to analyze the particular processes being affected, but the annotation to involvement in the process would only be made to the gene being knocked out. Currently, the topical areas selected for each paper are not directly tied to the genes associated with the paper. Thus, a paper selected for GO for one of the genes will appear in the GO El interface of unused papers for each of the genes indexed to the paper.

# Curation triage: selecting what genes to annotate

Ideally, all papers would be immediately used for GO annotation, but on average 300 new papers are added to the database each week, only less than half of these are curated each week due to resource limitations. Therefore, various priority selection criteria are used to choose which genes and papers warrant immediate attention. Reports of interest are generated, such as 'genes with no GO annotation but have new indexed literature selected for GO' or 'genes with mutant alleles that have literature selected for GO' (Figure 5). Additionally, participation in various collaborative projects, such as the Reference Genome project (8), or the Protein Ontology (8) defines primary sets of genes to work based on community input.

### Data entry: creating a GO annotation using the EI module for GO

A curator at MGI uses the GO EI module to enter annotations (Figure 6). Curators use the annotation guidelines set forth by the Gene Ontology Consortium (http://www. geneontology.org/GO.annotation.shtml). The MGI interface is gene centric. It is divided into two main data entry sections: the annotation area and the annotation properties area. A list of papers selected for GO that are associated with this gene is shown in the lower right panel. Individual protein isoforms or modified forms can be indicated using annotation properties where an id for a specific isoform can be indicated. After reading a paper, the curator selects the appropriate GO id (1). Next, the reference number

### Partial list of QC reports used to triage GO curation

- Mouse Genes that have Rat/Human Homologs but no GO Annotations
- Mouse Genes with no GO Annotations
- Mouse Genes that have Alleles but no GO Annotations
- Genes with no GO Annotations with references that are selected for GO but have not been used\_
- Genes with references that are selected for GO but have not been used
- <u>Genes with GO Annotations of evidence IEA only and with references that are</u>
   <u>selected for GO but have not been used</u>
- <u>Genes with OMIM Annotations and either GO Annotations of evidence IEA only or</u> <u>no GO Annotations</u>
- <u>All genes with 'root' annotations with new indexed literature</u>
- "Done" Genes with New Literature
- Non-Gene Markers with GO Annotations
- Markers with Annotations to Obsolete GO Terms

Figure 5. GO QC reports used for annotation triage and quality control.



**Figure 6.** MGI GO data entry module: the interface is divided into three main sections: GO annotation, annotation properties and search and reference tracking. Drop-down menus display pick-lists of allowed entries in various fields (evidence property, GO qualifier and evidence codes). Numbered areas: 1, GO ID entry; 2, reference entry; 3, evidence code entry; 4, 'inferred\_from' entry required for certain evidence codes and 5, annotation properties entry.

is added (2), as well as the evidence code (3). If required by the type of evidence code, additional information is added to the inferred from column (4). Once the annotation is saved, information about the cell type that the experiment was done in, or the specific isoform, or tissue, is entered into the annotation properties section (5). Additional ontologies such as the Cell Ontology (9), Mouse Adult (10) and Embryonic Anatomies (11), Protein Ontology (12), and psi-Mod (13) are used in the properties fields. Several of these are used to supply an extension to the annotation which are used in the gene association file (GAF) (5). Curators can use the OBO-EDIT tool (14) to load multiple ontologies to aid in searching for appropriate terms, as well as viewing the chosen term in the context of the rest of the ontology (Figure 7). The data entry module has several built-in features to aid in QC. The GO vocabulary is refreshed daily from the GO site, and only current GO terms can be used, otherwise data entry is prohibited. Only reference identifiers previously entered into 'master bibliography' can be used. Incorrect evidence codes are automatically rejected. There are other mechanisms, such as data loads, that provide GO annotation. In all cases,

provenance is provided. The GO EI also has a built-in report generator that highlights words matching GO terms found in the abstracts of papers selected for GO as an aid to suggesting the type of information and evidence present in a paper (Figure 8).

## Tracking metrics and quality control measures to set priorities for upcoming work

GO annotation metrics in MGI are generated daily. MGI GO curators add on average 200 new annotations per week. Annotations are tracked based on a variety of criteria such as annotation source (MGI curation or data load) and evidence (experimental or predictive, such as through orthology or functional domain). Scripts review changes to the GO structure and provide QC reports for curators noting genes whose annotations may be affected by these changes (Figure 9). Additionally, we use the master bibliography tables and the GO annotations to keep track of various areas that need focus, such as 'genes with



**Figure 7.** OBO-Edit ontology tool used to browse multiple OBO ontologies. The far left panel shows the vocabularies that have been loaded for searching and viewing. The right panel displays the terms in all of the vocabularies that contain the word 'kidney'. The GO term 'kidney mesenchyme morphogenesis' is selected and is visible as a tree view showing its children (middle panel), and as a graphical view showing its parents (lower left).

000	X MGD GOVocAnnot ei-4-4-2-41 (hjd, PROD1_MGI, mgd)
File Commands Report	s Search Clear
Annotation Type	HGI Acc Filter Terence Gene? Yes J
Sort Under Dg IMG, recent Modifi Term Acc ID DAG Vocabulary Te B0:0016021 C integral to m	Cation di Vhome/ng/mg/report/1  Nome/ng/mg/report/1  Nome/ng/mg/report/1  Nome/ng/mg/report/.  Nome/ng/mg/report/.
GD:0016021 C integral to m	enbrane GODrd2,html uniprotloa, 3/25/2012 n
0010043673       C       axon terminus         6010043673       C       axon terminus         6010043673       C       axon terminus         6010043673       C       edmitic spi         6010043871       C       demitic spi         6010043872       C       flagellum         60100430672       C       symptic vesi         6010005625       C       soluble fract         6010005687       C       integral to p         Rdd Row       Delete Row       Edit Term         Stanza       Property       I	Potential New GO References Symbol: Drd2, dopamine receptor D2, Chr 9 Start Date/Time: Tue Mar 27 11:10:49 2012 J:157250, Sahar S, PLoS One 2010;5(1):e8561 Regulation of BMAL1 protein stability and circadian function by GSK3beta-mediated phosphorylation. BACKGROUND: Circadian rhythms govern a large array of physiological and metabolic functions. To achieve plasticity in circadian regulation, proteins constituting the molecular clock machinery undergo various posit-translational modifications (PTMs), which influence their activity and intracellular localization. The core clock protein BMAL1 undergoes several PTMs. Here we report that the Akt-GSK3beta signaling pathway regulates BMAL1 protein stability and activity. PRINCIPAL FINDINGS: GSK3beta phosphorylates BMAL1 specifically on Ser 17 and Thr 21 and primes it for ubiquitylation. In the absence of GSK3beta-mediated phosphorylation, BMAL1 becomes stabilized and BMAL1 dependent circadian gene expression is drinatal neurons. CONCLUSIONS: These findings uncover a previously unknown mechanism of circadian dock control. The GSK3beta kinase phosphorylates BMAL1, an event that controls the stability of the protein and the amplitude of circadian oscillation. BMAL1 phosphorylation appears to be an important regulatory step in maintaining the robustness of the circadian clock.
Evidence Property anatomy	J:161807, Thompson D, PLoS One 2010;5(6):e11038
	BACKGROUND: Drugs of abuse elevate brain dopamine levels, and, in vivo, chronic drug use is accompanied by a selective decrease in dopamine D2 receptor (D2R) availability in the brain. Such a decrease consequently alters the ratio of D1R:D2R signaling towards the D1R. Despite a plethora of behavioral studies dedicated to the understanding of the role of dopamine in addiction, a molecular mechanism responsible for the downregulation of the D2R, in vivo, in response to chronic drug use has yet to be identified. METHODS AND FINDINGS: ETHICS STATEMENT: All animal work was approved by the Gallo Center IACUC committee and was performed in our AAALAC approved facility. In this study, we used wild type (WT) and G protein coupled receptor associated sorting protein-1 (GASP-1) knock out (KO) mice to assess molecular changes that accompany cocaine sensitization. Here, we show that downregulation of D2Rs or upregulation of D1Rs is associated with a sensitized locomotor response to an acute injection of cocaine. Furthermore, we demonstrate that disruption of GASP-1, that targets D2Rs for degradation after endocytosis, prevents cocaine-induced downregulation of D2Rs. As a consequence, mice with a GASP-1 disruption show a reduction in the sensitized locomotor response to cocaine. CONCLUSIONS: Together, our data suggests that changes in the ratio of the D1:D2R could contribute to cocaine-induced behavioral plasticity and demonstrates a role of GASP-1 in regulating both the levels of the D2R and cocaine sensitization.

Figure 8. Report generated using the abstracts of papers selected for GO for the gene being annotated within the GO EI. Text contained in GO terms in each abstract is highlighted.

no GO annotation but have papers that are selected for GO but not used'.

# Annotation presentation and usage

GO data for each gene at MGI are displayed to the public in a 'GO Summary' page. This page displays the GO annotations as a table, summary text or graph. Sample views for the gene *Drd2* are shown in Figure 10. All data assertions in MGI are supported by evidence and citation to the source of the information. For assertions that are associated with controlled vocabularies such as the GO, links are provided to vocabulary browsers that provide the relationships between the assertion and other knowledge in that area of the ontology. Using the table and associated information, MGI provides an automatically generated text description of the GO annotations. MGI also provides a graphical display of GO annotations from the GO detail page for each gene. GO annotations are also shared with the GO Consortium (GOC) through a GAF. This is a tab-delimited file that contains most of the elements of a GO annotation as outlined in the GO EI section above. Presently, only the 'cell type' and 'gene product' annotation properties are included in the GAF. More will be included over time. This file is available on either the GOC web site or along with other data sets, from the MGI FTP site (ftp://ftp. informatics.jax.org/pub/reports/index.html). The GAF and the GO vocabulary file are used as input for many analytical tools such as GO TermFinder (15). Instructions for construction of a GO GAF file are found in GO documentation at http://www.geneontology.org/GO.format.annotation. shtml.

# Information access: NLP and beyond

In general, GO annotation from the mouse experimental literature can be very challenging. Although some groups

### **Curator Report**

### DATE: 12/15/2011 03:03:23

Accession ID	Term	Discrepancy
GO:0008418	protein-N-terminal asparagine amidohydrolase activity	Definition change for Term with annotations. Old Definition: Catalysis of the deamidation of an N-terminal asparagine residue in a peptide or protein. New Definition: Catalysis of the reaction: protein-L-asparagine + H2O = protein-L-asparate + NH3. This reaction is the deamidation of an N-terminal asparagine residue in a peptide or protein. Symbols: Ntan1
GO:0034595	phosphatidylinositol phosphate 5-phosphatase activity	Definition change for Term with annotations. Old Definition: Catalysis of the removal the of the 5-phosphate group of a phosphatidylinositol phosphate. New Definition: Catalysis of the removal of the 5-phosphate group of a phosphatidylinositol phosphate. Symbols: Inpp5k Synj1
GO:0035602	fibroblast growth factor receptor signaling pathway involved in negative regulation of apoptotic process in bone marrow	Definition change for Term with annotations. Old Definition: The series of molecular signals generated as a consequence of a fibroblast growth factor receptor binding to one of its physiological ligands, which stops, prevents, or reduces the frequency, rate or extent of the occurrence or rate of cell death by apoptosis in the bone marrow. New Definition: The series of molecular signals generated as a consequence of a fibroblast growth factor receptor binding to one of its physiological ligands, which stops, prevents, or reduces the frequency, rate or extent of the occurrence or rate of cell death by apoptosis in the bone marrow. Symbols: Fgfr2
GO:0052744	phosphatidylinositol monophosphate phosphatase activity	Definition change for Term with annotations. Old Definition: Catalysis of the reaction: phosphatidylinositol phosphate + H2O = phosphatidylinositol + phosphate. New Definition: Catalysis of the reaction: phosphatidylinositol monophosphate + H2O = phosphatidylinositol + phosphate. Symplex Sympl
GO:0070773	protein-N-terminal glutamine amidohydrolase activity	Definition change for Term with annotations. Old Definition: Catalysis of the deamidation of an N-terminal glutamin residue of a protein. New Definition: Catalysis of the reaction: protein-N-terminal-L-glutamine + H2O = protein-N-terminal-L-glutamate + NH3. This reaction is the deamidation of an N-terminal glutamine residue of a protein. Symbols: Wdyhv1
GO:0071866	negative regulation of apoptotic process in bone marrow	Definition change for Term with annotations. Old Definition: Any process that stops, prevents, or reduces the frequency, rate or extent of the occurrence or rate of cell death by apoptosis in the bone marrow. New Definition: Any process that stops, prevents, or reduces the frequency, rate or extent of the occurrence or rate of cell death by apoptotic process in the bone marrow. Symbols: Left

Figure 9. GO change log report showing changes to the GO and genes with annotations using the term that may need to be looked at.



Figure 10. The GO annotation details for *Drd2* displayed as summary text (A), table (B) or graph (C). Only a portion of each summary is shown. There are 175 annotations total.

have used NLP to expedite the curation of literature (16), this can be especially difficult to do when applied to mouse biology because of the integration of human and mouse studies within the same description of results. The concepts captured by the GO cannot be gleaned just from simple text matching of terms, but must also take into account inferences that reflect a given context. Additionally, an understanding of the nature of an experimental assay is important to correctly use the information as evidence of a particular result. While we continue to work with NLP developers to design a system to automate identification and tagging of papers (17), it is clear that the complexity of understanding the information in a biomedical publication requires the intervention of an experienced biologist-curator in the process.

# **Acknowledgements**

The Mouse Genome Informatics group are Mark Airey, Anna Anagnostopoulos, Randal P. Babiuk, Richard M. Baldarelli, Jonathan S. Beal, Dale A. Begley, Susan M. Bello, Judith A. Blake, Carol J. Bult, Donna L. Burkart, Nancy E. Butler, Jeffrey Campbell, Lori E. Corbani, Howard Dene, Alexander Diehl, Mary E. Dolan, Harold J. Drabkin, Janan T. Eppig, Jacqueline H. Finger, Kim L. Forthofer, Peter Frost, Sharon Giannatto, Jill R. Lewis, Terry F. Hayamizu, David P. Hill, James A. Kadin, Debra M. Krupke, Michelle Knowlton, Monica McAndrews, Susan McClatchy, Ingeborg McCright, David B. Miers, Howie Motenko, Steve Neuhauser, Li Ni, Hiroaki Onda, Janice Ormsby, Jill Recla, Deborah J. Reed, Beverly Richards-Smith, Joel E. Richardson, Martin Ringwald, David Shaw, Robert Sinclair, Dmitry Sitnikov, Constance M. Smith, Cynthia L. Smith, Kevin Stone, John Sundberg, Hamsa Tadepally, Monika Tomczuk, Linda Washburn, Jingjia Xu and Yunxia Zhu.

# Funding

Funding for open access charge: MGI database resources are funded by grants from the National Human Genome Research Institute (HG00330, HG02273), National Institutes of Health/National Institute of Child Health and Human Development (HD062499) and the National Cancer Institute (CA89713).

Conflict of interest. None declared.

# References

- Eppig,J.T., Blake,J.A., Bult,C.J. et al. (2012) The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Res.*, 40, D881–D886.
- Finger, J.H., Smith, C.M., Hayamizu, T.F. *et al.* (2011) The mouse Gene Expression Database (GXD): 2011 update. *Nucleic Acids Res.*, 39, D835–D841.
- 3. Krupke,D.M., Begley,D.A., Sundberg,J.P. *et al.* (2008) The Mouse Tumor Biology database. *Nat. Rev. Cancer*, **8**, 459–465.
- 4. Ashburner, M., Ball, C.A., Blake, J.A. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
- 5. Evsikov, A., Dolan, M., Genrich, M. *et al.* (2009) MouseCyc: a curated biochemical pathways database for the laboratory mouse. *Genome Biol.*, **10**, R84.
- Latendresse, M., Paley, S. and Karp, P.D. (2012) Browsing metabolic and regulatory networks with BioCyc. *Methods Mol. Biol.*, 804, 197–216.
- Hanisch, D., Fundel, K., Mevissen, H.T. et al. (2005) ProMiner: rulebased protein and gene entity recognition. *BMC Bioinformatics*, 6 (Suppl. 1), S14.
- Gaudet, P., Livstone, M.S., Lewis, S.E. *et al.* (2011) Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief. Bioinform.*, **12**, 449–462.
- 9. Bard, J., Rhee, S. and Ashburner, M. (2005) An ontology for cell types. Genome Biol., 6, R21.
- Hayamizu, T., Mangan, M., Corradi, J. et al. (2005) The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. *Genome Biol.*, 6, R29.
- 11. Baldock, R., Bard, J., Brune, R. *et al.* (2001) The Edinburgh Mouse Atlas: using the CD. *Brief. Bioinform.*, **2**, 159–169.
- Natale, D.A., Arighi, C.N., Barker, W.C. et al. (2011) The Protein Ontology: a structured representation of protein forms and complexes. Nucleic Acids Res., 39, D539–D545.
- 13. Montecchi-Palazzi,L., Beavis,R., Binz,P.-A. *et al.* (2008) The PSI-MOD community standard for representation of protein modification data. *Nat. Biotech.*, **26**, 864–866.
- Day-Richter, J., Harris, M.A., Haendel, M. et al. The Gene Ontology OBO-Edit Working Group. (2007) OBO-Edit—an ontology editor for biologists. *Bioinformatics*, 23, 2198–2200.
- Boyle, E.I., Weng, S., Gollub, J. *et al.* (2004) GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20, 3710–3715.
- Van Auken, K., Jaffery, J., Chan, J. et al. (2009) Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. BMC Bioinformatics, 10, 228.
- Dowell,K.G., McAndrews-Hill,M.S., Hill,D.P. et al. (2009) Integrating text mining into the MGI biocuration workflow. *Database*, 2009, bap019.