FlyCyc: updating the metabolic network for *Drosophila melanogaster*

Steven J Marygold[1*], Phani V Garapati[1], Gil dos Santos[2] and Peter D Karp[3]

1. FlyBase, Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK
2. FlyBase, The Biological Laboratories, Harvard University, Cambridge, MA, USA
3. Bioinformatics Research Group, SRI International, Menlo Park, CA, USA

* email: **sjm41@cam.ac.uk**

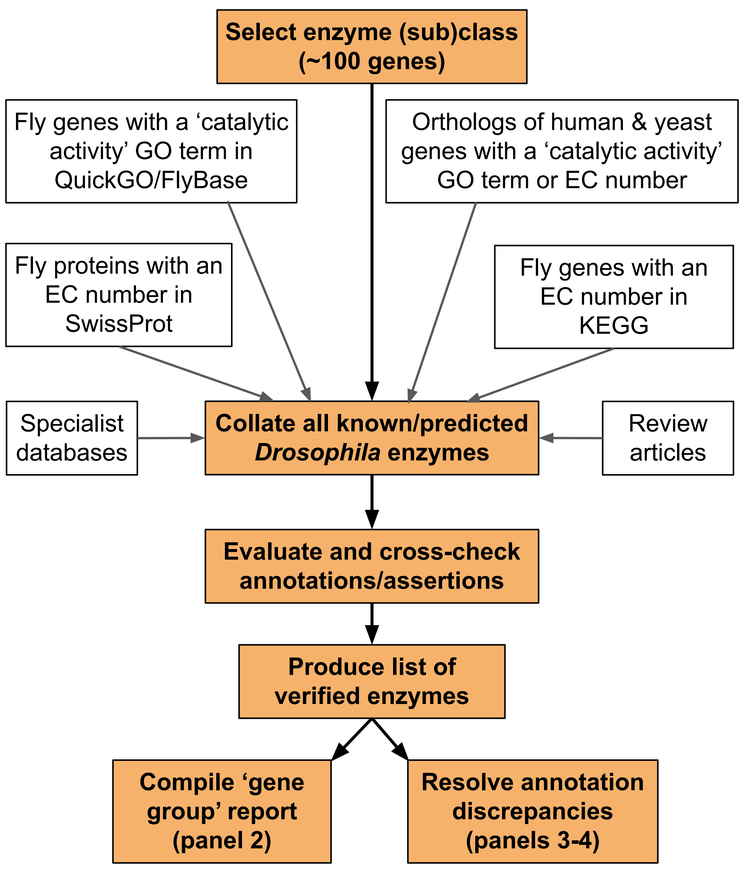
BioCyc is a collection of Pathway/Genome Databases (PGDBs) that represent metabolic networks for over 20,000 species. The BioCyc 'Pathway Tools' software can generate a metabolic reconstruction for a given species, stored as a PGDB, by matching enzymes in the functionally annotated genome of that species with the reactions/pathways in the reference metabolic database (MetaCyc). The quality of the PGDB therefore depends on that of the underlying functional annotations. A PGDB for *Drosophila melanogaster* (FlyCyc) exists but is based on data from a FlyBase release 15 years ago and therefore excludes the changes to genomic or functional annotations made since then.

We have conducted a systematic review of *Drosophila* enzymes, improving the coverage and accuracy of their functional (Gene Ontology molecular function (GO-MF) and Enzyme Commission (EC)) annotations and creating accessible 'gene group' pages for each enzyme class in FlyBase. We verified ~3,750 *Drosophila* enzymes, made ~4,000 changes to manual GO annotations, and organized the enzymes into ~800 hierarchical groups. In so doing, we identified ~400 issues with automated annotation pipelines (InterPro2GO, PAINT, UniRule) and ~300 issues regarding catalytic activity terms within the GO (e.g. incorrect term relationships, missing EC cross-references). Almost all these issues have now been addressed, thereby improving the quality of enzymatic GO annotation for all species.

We have used data from the latest FlyBase release (FB2023_02) and the Pathway Tools software to recompute an updated FlyCyc that incorporates our improved GO/EC annotations as well as the latest genomic and gene nomenclature data. Compared to the previous version, the updated FlyCyc includes >50 additional metabolic pathways and identifies >600 additional enzyme-encoding genes. Ambiguous enzyme mappings and 'pathway holes' are being resolved as far as possible by correcting GO-MF/EC annotations within FlyBase to focus primary curation activities within a single database. The finalized collection of *Drosophila* metabolic pathways will then be made available on the BioCyc website. In addition to providing researchers with improved metabolic pathway diagrams, this update will enhance the functionality of various 'omics data analysis tools available at BioCyc.

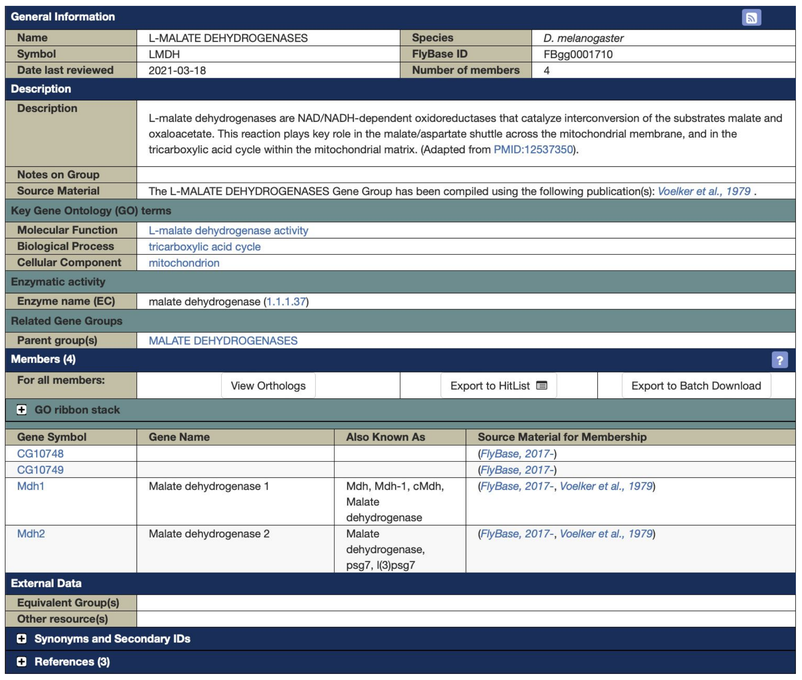
Going forwards, we will review GO cellular component annotations to *Drosophila* enzymes to accurately indicate the subcellular compartment in which they act and, where applicable, the macromolecular complex of which they are a part. We will also perform a systematic annotation review of *Drosophila* transporters, as they also play a critical role in metabolic pathways. These additional enhancements, together with ongoing improvements to enzymatic annotations, will be reflected at FlyCyc by establishing regular 6-monthly synchronizations with the data at FlyBase.

1. Review *Drosophila* enzymes



2. Compile verified enzymes as 'gene groups'

FlyBase 'Gene Groups' are manually-curated collections of functionally-related *D. melanogaster* genes. They are arranged into hierarchies, cross-referenced with applicable GO and EC terms, and provide links to relevant literature, FlyBase tools and equivalent groups of human genes at the HGNC. The organization of enzyme gene groups follows that of the EC/GO, and thereby provides a stable set of verified enzyme lists that can be compared with EC/GO annotation.



General Information

Name	L-MALATE DEHYDROGENASES	**Species**	*D. melanogaster*
Symbol	LMDH	**FlyBase ID**	FBgg0001710
Date last reviewed	2021-03-18	**Number of members**	4

Description

Description	L-malate dehydrogenases are NAD/NADH-dependent oxidoreductases that catalyze interconversion of the substrates malate and oxaloacetate. This reaction plays key role in the malate/aspartate shuttle across the mitochondrial membrane, and in the tricarboxylic acid cycle within the mitochondrial matrix. (Adapted from PMID:12537350).
Notes on Group	
Source Material	The L-MALATE DEHYDROGENASES Gene Group has been compiled using the following publication(s): *Voelker et al., 1979* .

Key Gene Ontology (GO) terms

Molecular Function	L-malate dehydrogenase activity
Biological Process	tricarboxylic acid cycle
Cellular Component	mitochondrion

Enzymatic activity

Enzyme name (EC)	malate dehydrogenase (1.1.1.37)

Related Gene Groups

Parent group(s)	MALATE DEHYDROGENASES

Members (4)

For all members: View Orthologs | Export to HitList | Export to Batch Download

GO ribbon stack

Gene Symbol	Gene Name	Also Known As	Source Material for Membership
CG10748			(*FlyBase, 2017-*)
CG10749			(*FlyBase, 2017-*)
Mdh1	Malate dehydrogenase 1	Mdh, Mdh-1, cMdh, Malate dehydrogenase	(*FlyBase, 2017-*, *Voelker et al., 1979*)
Mdh2	Malate dehydrogenase 2	Malate dehydrogenase, psg7, l(3)psg7	(*FlyBase, 2017-*, *Voelker et al., 1979*)

External Data

Equivalent Group(s)	
Other resource(s)	

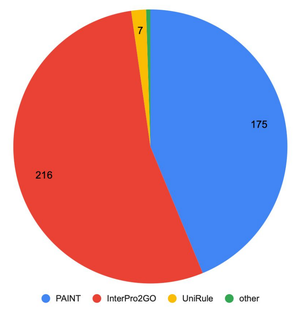
Synonyms and Secondary IDs

References (3)

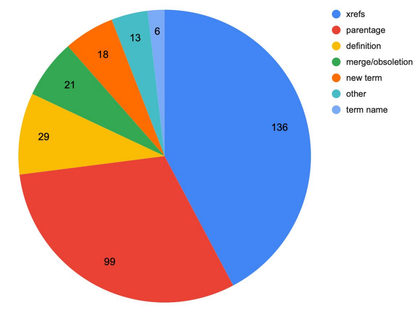
3. Resolve annotation discrepancies

Cause of issue	Action
Erroneous/missing computational GO annotation	GO ticket
Erroneous/missing EC xref in the GO	GO ticket
Erroneous/missing relationship in the GO	GO ticket
Erroneous/missing manual GO annotation	Annotate
Uncurated literature	Curate
Lack of equivalence between GO and EC	n/a
Database asynchrony	n/a
No GO term	GO ticket
Incorrect EC annotations submitted to INSDC	Fix
Erroneous/missing EC/keyword in Swiss-Prot	Helpdesk

~400 go-annotation tickets:



~300 go-ontology tickets:



4. Summary of GO annotation review

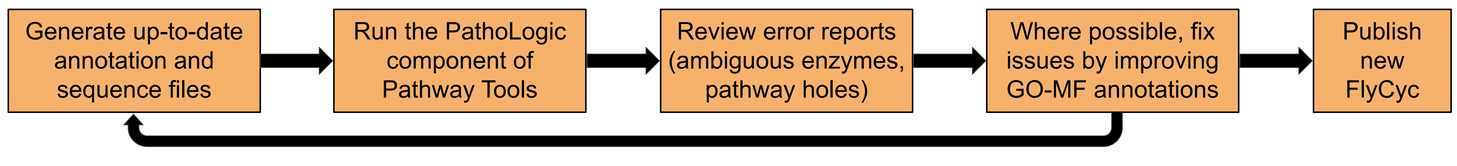
Enzyme class	Number of annotated genes*		
	before review	after review	removed/added
Oxidoreductases	617	621	88 / 92
Transferases	1,317	1,301	222 / 206
Hydrolases	1,781	1,567	440 / 226
Lyases	119	133	13 / 27
Isomerases	96	104	9 / 17
Ligases	111	146	16 / 51
Translocases	133	142	44 / 53
TOTAL	4,174	4,014	832 / 672

* Number of genes annotated to corresponding GO terms in FB2017_05 cf FB2023_02

5. A note on EC annotation & mapping

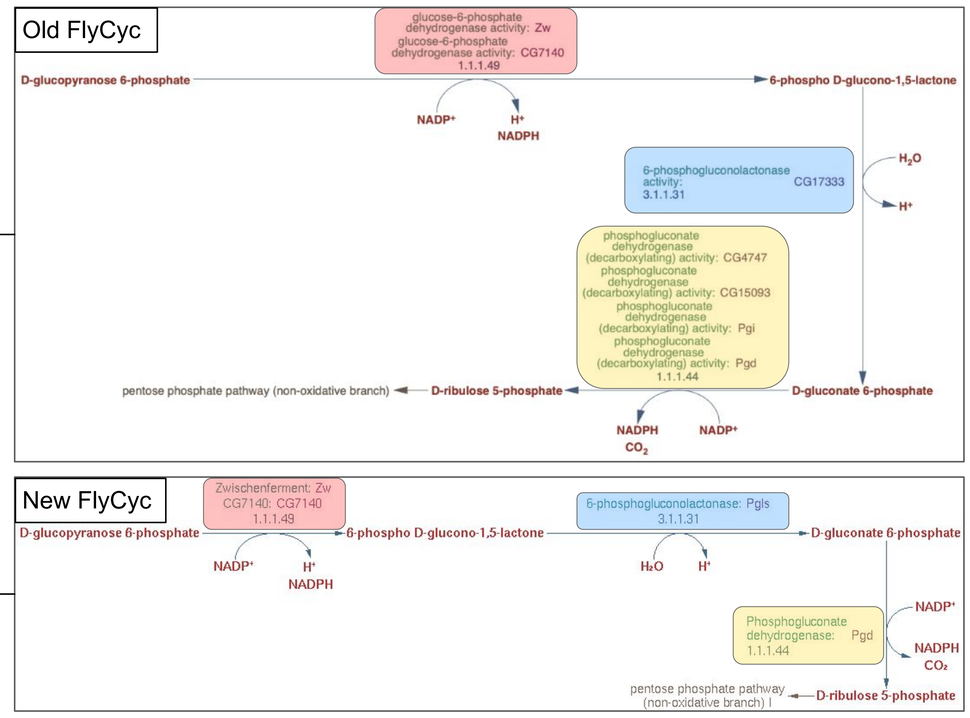
The BioCyc Pathway Tools software requires accurate and comprehensive EC annotation of the genome to correctly assign enzymes to their corresponding metabolic reactions and pathways. EC annotations may either be annotated directly or be inferred via annotations to 'catalytic activity' GO-MF terms and their associated EC cross-references (xrefs). FlyBase employs the latter strategy so that our functional annotation efforts are focussed on the GO and to ensure synchrony between our GO and EC annotations. However, this approach requires accurate GO-to-EC xref mapping (explaining why half of our go-ontology tickets are for xrefs) and does result in missing EC annotations in cases where the GO and EC classification systems diverge.

6. Update FlyCyc



	Old FlyCyc	New FlyCyc
Genome release	5.10	6.51
FlyBase release	FB2008_07	FB2023_02
Total #genes	15,097	17,879
Total #enzymes	3,504	2,467
Total #pathways	230	287
Total #GO terms	170	91,210

Compared to the old pathway diagrams (top), the new diagrams (bottom) use current FlyBase gene symbols (blue highlight), display the FlyBase gene name (red highlight) and incorporate corrected functional annotations (yellow highlight). The example here is the pentose phosphate pathway (oxidative branch).



Acknowledgments: We thank Pascale Gaudet, Harold Drabkin, Marc Feuermann, the InterPro curators and other members of the GO consortium for addressing the hundreds of GO tickets and disputes raised during this work. Thanks also to Helen Attrill for advice on GO annotation, and Kristian Axelson and Ron Caspi for help with EC queries. The core funding for FlyBase is from the National Human Genome Research Institute at the National Institutes of Health (#U41HG000739).