Improving enzyme annotation in FlyBase

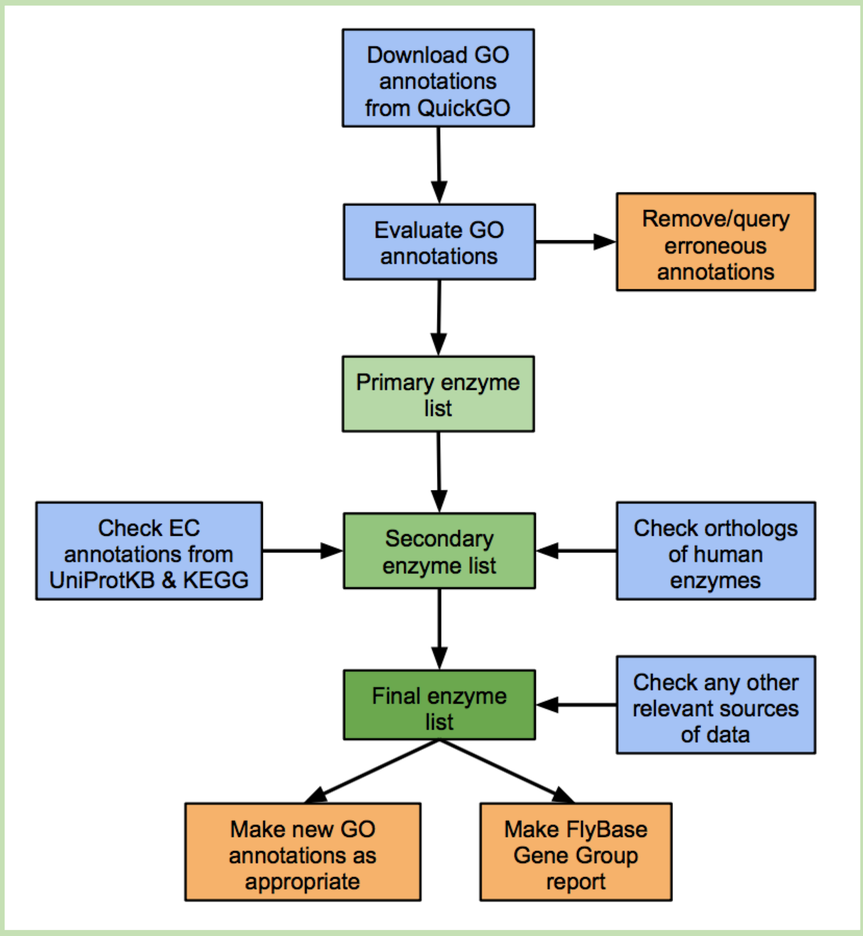
Phani V Garapati, Jingyao Zhang, Alix J Rey and Steven J Marygold
FlyBase, Dept. of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK
Email: sjm41@cam.ac.uk

Drosophila melanogaster **has been used as a model system to study enzyme function for over a century and a substantial proportion (~30%) of its protein-coding genome is known/predicted to encode enzymes. Nonetheless, many *Drosophila* enzymes remain unidentified or poorly/inconsistently classified within biological databases.**

To address these shortcomings, we are systematically reviewing *Drosophila* enzyme data obtained from several key databases, orthology-based searches and the primary literature. After integrating and evaluating these data, we ensure that all verified activities are annotated with the appropriate Gene Ontology (GO) and Enzyme Commission (EC) terms, providing feedback about any discrepancies to the relevant sources.

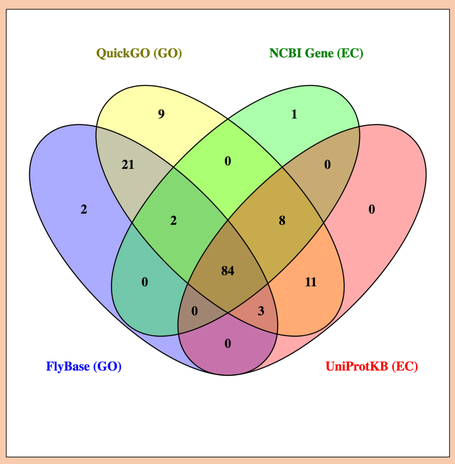
To date, we have reviewed 4 major classes (oxidoreductases, lyases, isomerases and ligases), resulting in new enzyme annotations to >130 genes and the removal of erroneous annotations for >75 genes. These improvements are evident within FlyBase as revised GO data and new EC data fields within Gene Reports. Importantly, these revisions are also exported to key third-party resources, such as UniProtKB, GenBank/NCBI, QuickGO, AmiGO and the Alliance of Genome Resources, thereby improving the accuracy and consistency of enzyme data across sites. Validated enzyme sets are also provided within FlyBase as convenient 'Gene Group' reports.

Reviewing & improving the underlying data



Enzyme class (EC number)	GO term	#Genes before analysis	#Genes after analysis	Genes added / removed	GO annotations added/removed
Oxidoreductases (1.-.-.-)	oxidoreductase activity	616	649	72 / 39	90 / 13
Transferases (2.-.-.-)	transferase activity	1,382	TBD	TBD	TBD
Hydrolases (3.-.-.-)	hydrolase activity	1,877	TBD	TBD	TBD
Lyases (4.-.-.-)	lyase activity	121	130	23 / 14	14 / 8
Isomerases (5.-.-.-)	isomerase activity	97	104	13 / 6	20 / 2
Ligases (6.-.-.-)	ligase activity	112	121	27 / 18	26 / 13

Investigating the discrepancies: focus on *ligases*



Initially, 141 potential *D. melanogaster* ligases were identified by searching GO or EC annotations within 4 different databases: FlyBase, QuickGO, NCBI Gene and UniProtKB. Although most hits (60%) were found in all sources, there were significant differences between them. Reasons for these discrepancies were investigated (see below) and rectified wherever possible. Ultimately, 40 (28%) candidates were discarded (false positives), while an additional 20 (false negatives) were identified via orthology or literature searches, making a total of 121 verified ligases.

Reasons for false *negatives* include:

- Uncurated primary literature
- Incorrect relationships within the GO
- GO annotation pipeline not used in source
- No EC number equivalent to a GO term
- UniProtKB/Swiss-Prot entry lacked EC annotation
- *D. melanogaster* enzyme lacks clear human ortholog
- Human ortholog lacks GO/EC annotation
- Database asynchrony

Reasons for false *positives* include:

- Erroneous manual GO annotations
- Erroneous computational GO annotations
- Incorrect relationships within the GO
- Erroneous EC/keyword annotations in UniProtKB/Swiss-Prot
- Incorrect EC numbers submitted to INSDC by FlyBase
- Incorrect EC numbers submitted to INSDC by researchers
- Database asynchrony

Facilitating access to enzyme data in FlyBase

Enzyme sets are organized into hierarchies:

```
LIGASES [121]
|
|__CARBON-OXYGEN LIGASES [37]
|       |__AMINOACYL-TRNA SYNTHETASES [35]
|       |       |__CYTOPLASMIC AMINO-ACID TRNA SYNTHETASES [16]
|       |       |__MITOCHONDRIAL AMINO-ACID TRNA SYNTHETASES [15]
|       |       |__DUAL-LOCALIZED AMINO-ACID TRNA SYNTHETASES [4]
|       |__AMINOACYL-TRNA SYNTHETASE-LIKE [2]
|
|__CARBON-SULFUR LIGASES [37]
|       |__ACYL-COA SYNTHETASES [25]
|       |__SUCCINATE-COA LIGASES [4]
|       |__UBIQUITIN ACTIVATING ENZYMES (E1) [8]
|
|__CARBON-NITROGEN LIGASES [41]
|       |__ACID-AMMONIA LIGASES [3]
|       |__ACID-AMINO ACID LIGASES [18]
|       |__CYCLO-LIGASES [2]
|       |__BIOTIN CARBOXYLASES [3]
|       |__CARBON-NITROGEN LIGASES, WITH GLUTAMINE AS AMIDO-N-DONOR [9]
|       |__OTHER CARBON-NITROGEN LIGASES [6]
|
|__CARBON-CARBON LIGASES [4]
|       |__PYRUVATE CARBOXYLASES [1]
|       |__COA CARBOXYLASES [3]
|
|__PHOSPHORIC ESTER LIGASES [5]
|       |__DNA LIGASES [3]
|       |__RNA LIGASES [1]
|       |__RNA-3'-PHOSPHATE CYCLASES [1]
|
|__BETA-ALANYL-DOPAMINE SYNTHASES [1]
```

Each set is presented and searchable as a Gene Group Report, which includes links to analysis/download tools, related resources and source references:

General Information			
Name	ACYL-COA SYNTHETASES	Species	*D. melanogaster*
Symbol	ACS	FlyBase ID	FBgg0000835
Date last reviewed	2018-02-01	Number of members	25
Description			
Description	Acyl-coenzyme A (CoA) synthetases catalyze the "activation" of fatty acids by their thioesterification to CoA. This is the initial reaction in fatty acid metabolism, allowing their participation in many fundamental anabolic and catabolic pathways. (Adapted from FBrf0237395).		
Notes on Group			
Source Material	The ACYL-COA SYNTHETASES Gene Group has been compiled by FlyBase curators using the following publication(s): *Watkins et al., 2007*.		
Key Gene Ontology (GO) terms			
Molecular Function	fatty acid ligase activity		
Biological Process			
Cellular Component			
Related Gene Groups			
Parent group(s)	CARBON-SULFUR LIGASES		
Members (25)			
For all members:	View Orthologs | Export to HitList | Export to Batch Download		

Gene Symbol	Gene Name	Also Known As	Source Material for Membership
AcCoAS	Acetyl Coenzyme A synthase	ACS, dACS, ACeCS1	(*FlyBase, 2017-*, *Watkins et al., 2007*)
Acsl	Acyl-CoA synthetase long-chain	l(2)44DEa, l(2)k05304, l(2)05847, dAcsl, FACS	(*FlyBase, 2017-*, *Watkins et al., 2007*)
bgm	bubblegum	BG:DS01514.2	(*FlyBase, 2017-*, *Watkins et al., 2007*)
CG3961		ACSL1	(*FlyBase, 2017-*, *Watkins et al., 2007*)
CG4563			(*FlyBase, 2017-*, *Watkins et al., 2007*)
CG4830			(*FlyBase, 2017-*, *Watkins et al., 2007*)
CG5568			(*FlyBase, 2017-*, *Watkins et al., 2007*)
CG6178			(*FlyBase, 2017-*, *Watkins et al., 2007*)
CG6300			(*FlyBase, 2017-*, *Watkins et al., 2007*)
CG6432			(*FlyBase, 2017-*, *Watkins et al., 2007*)
CG8834			(*FlyBase, 2017-*, *Watkins et al., 2007*)
CG9993			(*FlyBase, 2017-*, *Watkins et al., 2007*)
CG11391			(*FlyBase, 2017-*, *Watkins et al., 2007*)
CG11407			(*FlyBase, 2017-*, *Watkins et al., 2007*)
CG11453			(*FlyBase, 2017-*, *Watkins et al., 2007*)
CG11659			(*FlyBase, 2017-*, *Watkins et al., 2007*)
CG12512			(*FlyBase, 2017-*, *Watkins et al., 2007*)
CG17999			(*FlyBase, 2017-*, *Watkins et al., 2007*)
CG18155			(*FlyBase, 2017-*)
e	ebony	ebony	(*FlyBase, 2017-*, *Watkins et al., 2007*)
Fatp1	Fatty acid transport protein 1	Fatp, l(2)k10307, dFATP	(*FlyBase, 2017-*, *Watkins et al., 2007*)
Fatp2	Fatty acid transport protein 2		(*FlyBase, 2017-*, *Watkins et al., 2007*)
Fatp3	Fatty acid transport protein 3		(*FlyBase, 2017-*, *Watkins et al., 2007*)
hll	heimdall	dbb, BG:DS05899.1	(*FlyBase, 2017-*, *Watkins et al., 2007*)
pdgy	pudgy	BcDNA:GH02901	(*FlyBase, 2017-*, *Watkins et al., 2007*)
External Data			
Equivalent Group(s)	Human Acyl-CoA synthetases (HGNC)		
Other resource(s)			

Synonyms and Secondary IDs

References (3)

Publication Types: All publications 3 | Research paper 1 | FlyBase analysis 2

Filter: 2015, Smith, cell, etc. | Sort by: Year (descending)

FlyBase, 2017-, FlyBase classification of D. melanogaster enzymes.
FlyBase classification of D. melanogaster enzymes. [FBrf0237866]

FlyBase, 2014-, FlyBase Gene Group information.
FlyBase Gene Group information. [FBrf0225556]

Watkins et al., 2007, J. Lipid Res. 48(12): 2736--2750
Evidence for 24 distinct acyl-coenzyme A synthetase genes in the human genome. [FBrf0237395]

EC data (derived from GO annotations) added to Gene Reports:

General Information			
Symbol	Dmel\bgm	Species	*D. melanogaster*
Name	bubblegum	Annotation Symbol	CG4501
Feature Type	protein_coding_gene	FlyBase ID	FBgn0027348
Gene Model Status	Current	Stock Availability	16 publicly available
Enzyme Name (EC)	Long-chain-fatty-acid--CoA ligase (6.2.1.3)		
Gene Snapshot	In progress. Contributions welcome.		
Other Summaries	Auto summary | Gene Group | UniProtKB		
Also Known As	BG:DS01514 .2		
Key Links	ALLIANCE of GENOME RESOURCES | NCBI Gene | Ensembl | UniProtKB		
Families, Domains and Molecular Function			
Gene Group Membership (FlyBase)	ACYL-COA SYNTHETASES		
Protein Family (UniProt, Sequence Similarities)	Belongs to the ATP-dependent AMP-binding enzyme family. Bubblegum subfamily. (Q9V3S9)		
Protein Domains/Motifs	InterPro AMP-dependent synthetase/ligase		
Molecular Function (see GO section for details)	Experimental Evidence long-chain fatty acid-CoA ligase activity Predictions / Assertions fatty acid ligase activity		
Catalytic Activity (EC)	Experimental Evidence ATP + a long-chain fatty acid + CoA = AMP + diphosphate + an acyl-CoA (6.2.1.3) Predictions / Assertions ATP + a long-chain fatty acid + CoA = AMP + diphosphate + an acyl-CoA (6.2.1.3)		

Reference: Garapati et al. (2019) *Towards comprehensive annotation of Drosophila melanogaster enzymes in FlyBase*. Database 2019 bay144 (PMID:30689844)
Funding: FlyBase is supported by a grant from the National Human Genome Research Institute at the U.S. National Institutes of Health (U41HG000739). Support is also provided by the British Medical Research Council (#MR/N030117/1) and FlyBase users all over the world.