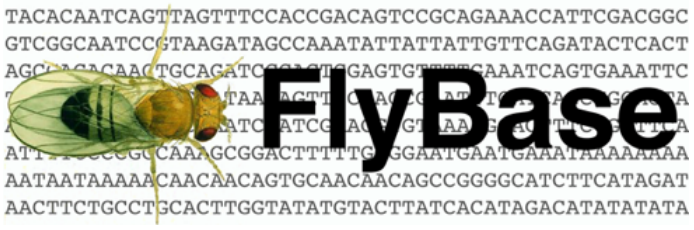
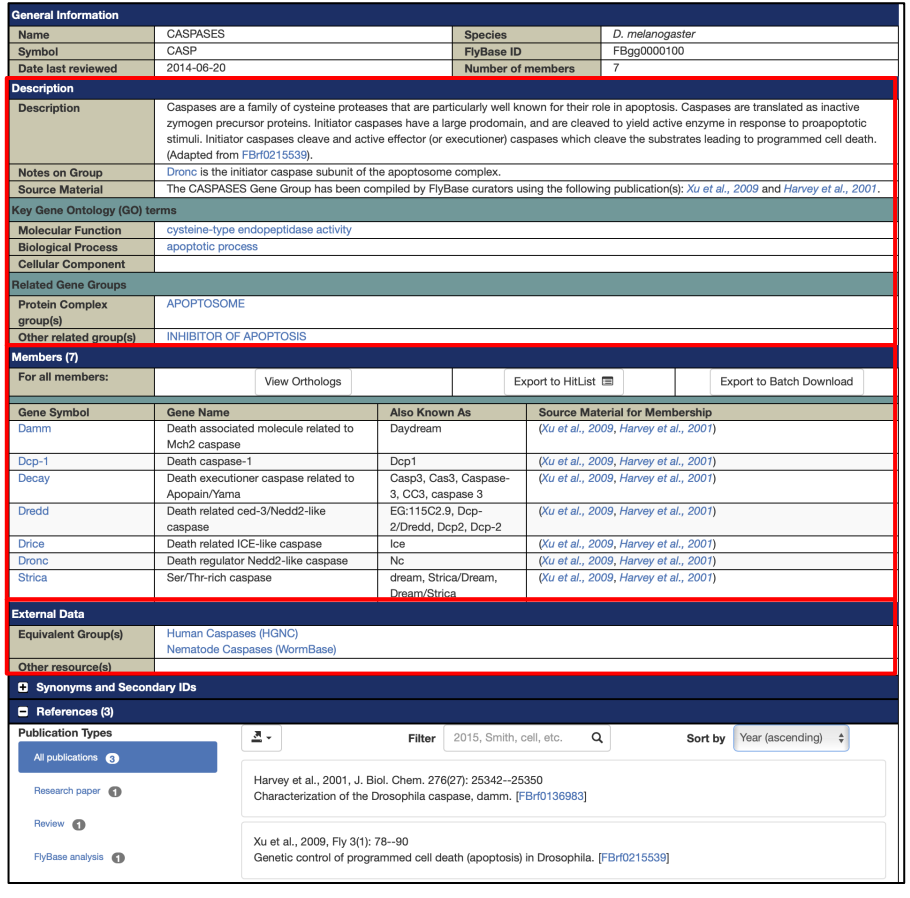


Towards comprehensive annotation of *Drosophila melanogaster* enzymes in FlyBase

Steven Marygold
FlyBase, University of Cambridge, UK
sjm41@cam.ac.uk

‘Gene Groups’ in FlyBase



General Information

Name	CASPASES	Species	*D. melanogaster*
Symbol	CASP	FlyBase ID	FBgg0000100
Date last reviewed	2014-06-20	Number of members	7

Description

Description	Caspases are a family of cysteine proteases that are particularly well known for their role in apoptosis. Caspases are translated as inactive zymogen precursor proteins. Initiator caspases have a large prodomain, and are cleaved to yield active enzyme in response to proapoptotic stimuli. Initiator caspases cleave and active effector (or executioner) caspases which cleave the substrates leading to programmed cell death. (Adapted from FBrf0215539).
Notes on Group	Dronc is the initiator caspase subunit of the apoptosome complex.
Source Material	The CASPASES Gene Group has been compiled by FlyBase curators using the following publication(s): *Xu et al., 2009* and *Harvey et al., 2001*.

Key Gene Ontology (GO) terms

Molecular Function	cysteine-type endopeptidase activity
Biological Process	apoptotic process
Cellular Component	

Related Gene Groups

Protein Complex group(s)	APOPTOSOME
Other related group(s)	INHIBITOR OF APOPTOSIS

Members (7)

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

Gene Symbol	Gene Name	Also Known As	Source Material for Membership
Damm	Death associated molecule related to Mch2 caspase	Daydream	(*Xu et al., 2009*, *Harvey et al., 2001*)
Dcp-1	Death caspase-1	Dcp1	(*Xu et al., 2009*, *Harvey et al., 2001*)
Decay	Death executioner caspase related to Apopain/Yama	Casp3, Cas3, Caspase-3, CC3, caspase 3	(*Xu et al., 2009*, *Harvey et al., 2001*)
Dredd	Death related ced-3/Nedd2-like caspase	EG:115C2.9, Dcp-2/Dredd, Dcp2, Dcp-2	(*Xu et al., 2009*, *Harvey et al., 2001*)
Drice	Death related ICE-like caspase	Ice	(*Xu et al., 2009*, *Harvey et al., 2001*)
Dronc	Death regulator Nedd2-like caspase	Nc	(*Xu et al., 2009*, *Harvey et al., 2001*)
Strica	Ser/Thr-rich caspase	dream, Strica/Dream, Dream/Strica	(*Xu et al., 2009*, *Harvey et al., 2001*)

External Data

Equivalent Group(s)	Human Caspases (HGNC) Nematode Caspases (WormBase)
Other resource(s)	

Synonyms and Secondary IDs

References (3)

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

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

Harvey et al., 2001, J. Biol. Chem. 276(27): 25342--25350
Characterization of the Drosophila caspase, damm. [FBrf0136983]

Xu et al., 2009, Fly 3(1): 78--90
Genetic control of programmed cell death (apoptosis) in Drosophila. [FBrf0215539]

Summary of Gene Group data (FB2019_02):	
Total number of groups	1,031
Number of genes in groups	6,279
- as % of all genes	35%
- as % of protein-coding genes	45%

D786–D792 *Nucleic Acids Research, 2016, Vol. 44, Database issue* *Published online 13 October 2015*
doi: 10.1093/nar/gkv1046

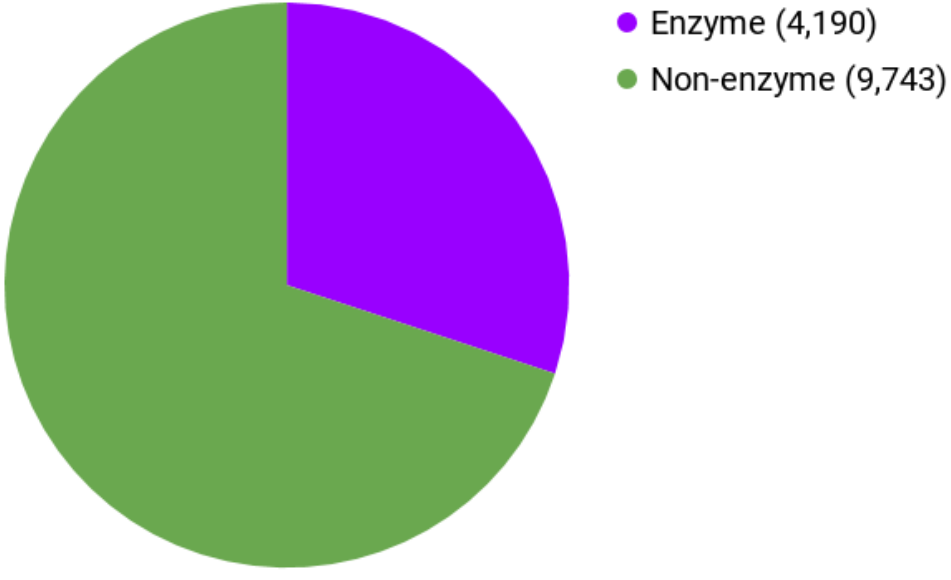
FlyBase: establishing a Gene Group resource for *Drosophila melanogaster*

Helen Attrill[1], Kathleen Falls[2], Joshua L. Goodman[3], Gillian H. Millburn[1], Giulia Antonazzo[1], Alix J. Rey[1], Steven J. Marygold[1,*] and the FlyBase consortium[†]

[1]Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH, UK, [2]The Biological Laboratories, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA and [3]Department of Biology, Indiana University, Bloomington, IN 47405, USA

Received September 14, 2015; Accepted October 01, 2015

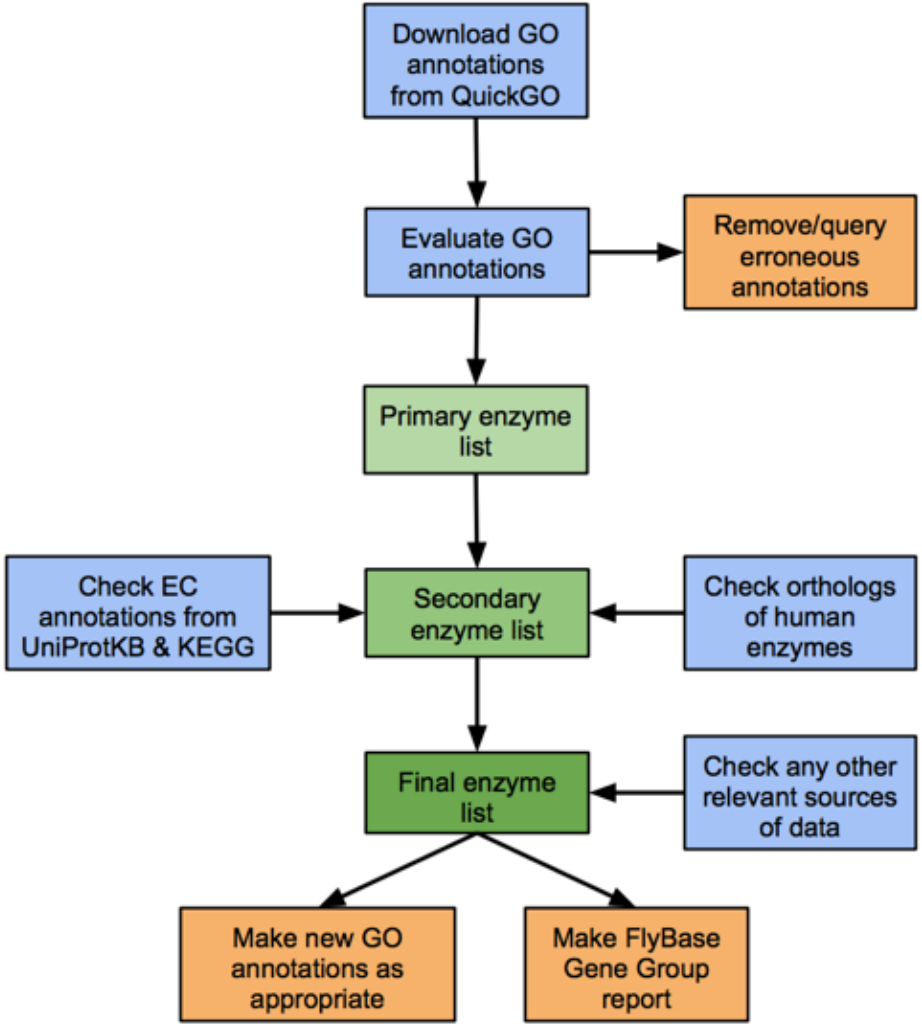
Drosophila melanogaster enzymes



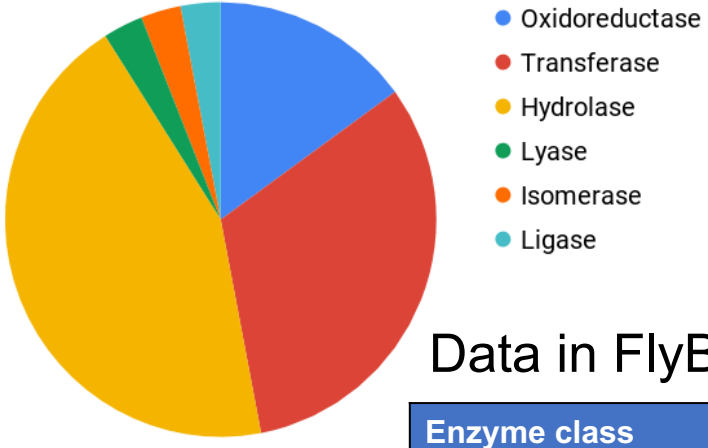
Sources of enzyme data:

- Gene Ontology annotations
- Enzyme Commission annotations
- Protein domains
- Primary literature
- Specialist databases
- Orthologs

Method


Download GO annotations from QuickGO
Evaluate GO annotations
Remove/query erroneous annotations
Primary enzyme list
Check EC annotations from UniProtKB & KEGG
Secondary enzyme list
Check orthologs of human enzymes
Final enzyme list
Check any other relevant sources of data
Make new GO annotations as appropriate
Make FlyBase Gene Group report

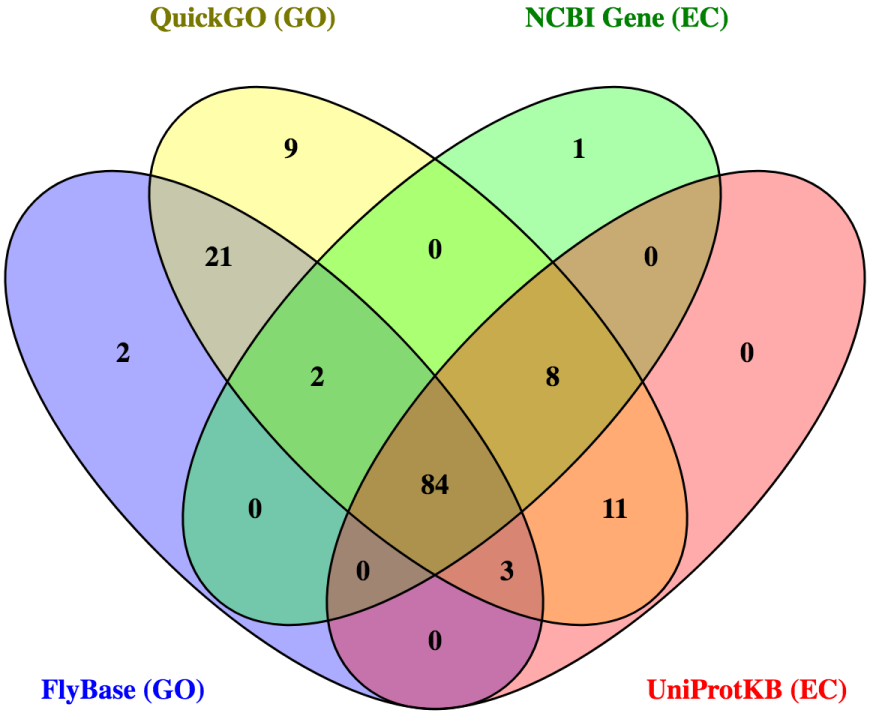
Summary of improvements to date



Data in FlyBase:

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

Case study: ligases



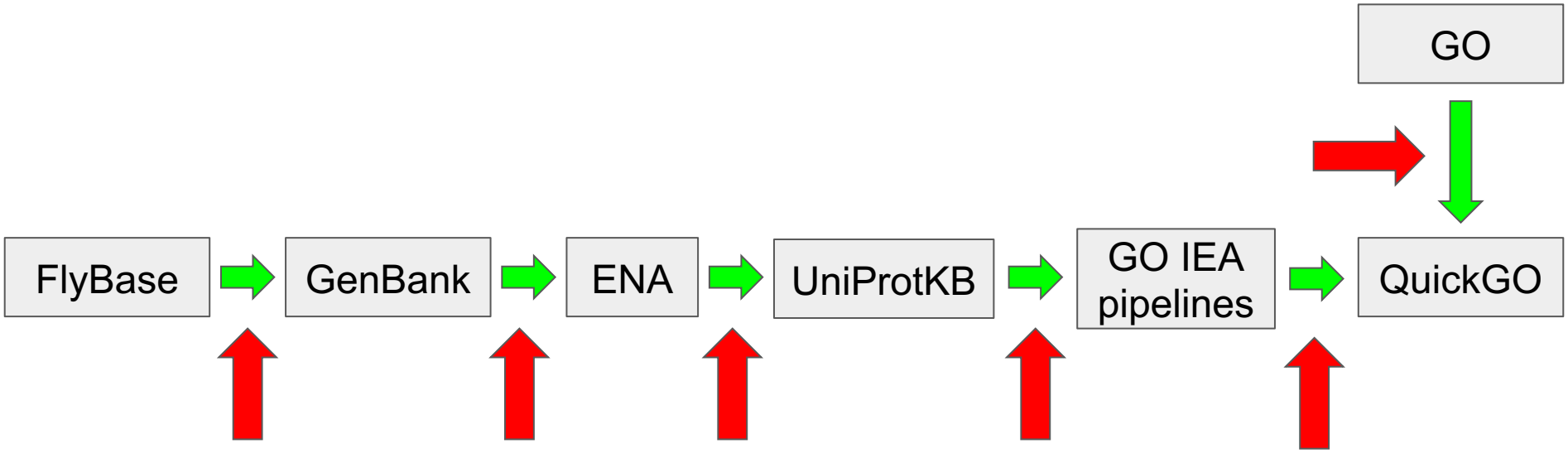
Data in FlyBase:

	Before (FB2017_05)	After (FB2018_05)
# Genes from GO search	112	121
• # false positives	18	0
• # false negatives	27	0
# Genes in Gene Group	n/a	121

Ligases - false positives/negatives

Cause	#false positives	#false negatives
Incorrect relationships within the GO	1	38
Uncurated primary literature		23
Erroneous computational GO annotations (UnitProtKB-Keyword2GO, PAINT, InterPro2GO)	16	
Erroneous manual GO annotations	15	
Database asynchrony - expected	6	3
Erroneous/missing EC/keyword annotations in UniProtKB/Swiss-Prot	7	1
No EC number equivalent to a GO term		8
GO annotation pipeline not used in source		4
Incorrect EC numbers submitted to INSDC	3	
Database asynchrony - unexpected	2	

Example of data flow



Take-home messages

1. Described an effective (low throughput) method for reviewing & improving enzyme annotations

2. No single database/approach gives accurate/comprehensive answer
    a. Same query, different results

3. Some discrepancies are expected, but others are avoidable
    a. Databases should better indicate their data sources/versions/policies
    b. New/additional checks could help to reduce discrepancies

4. Primary sources and third-parties share responsibility for accuracy

5. Essential that biocurators give feedback on core resources (e.g. GO, UniProt)

Acknowledgments

Phani Garapati

Lana Zhang

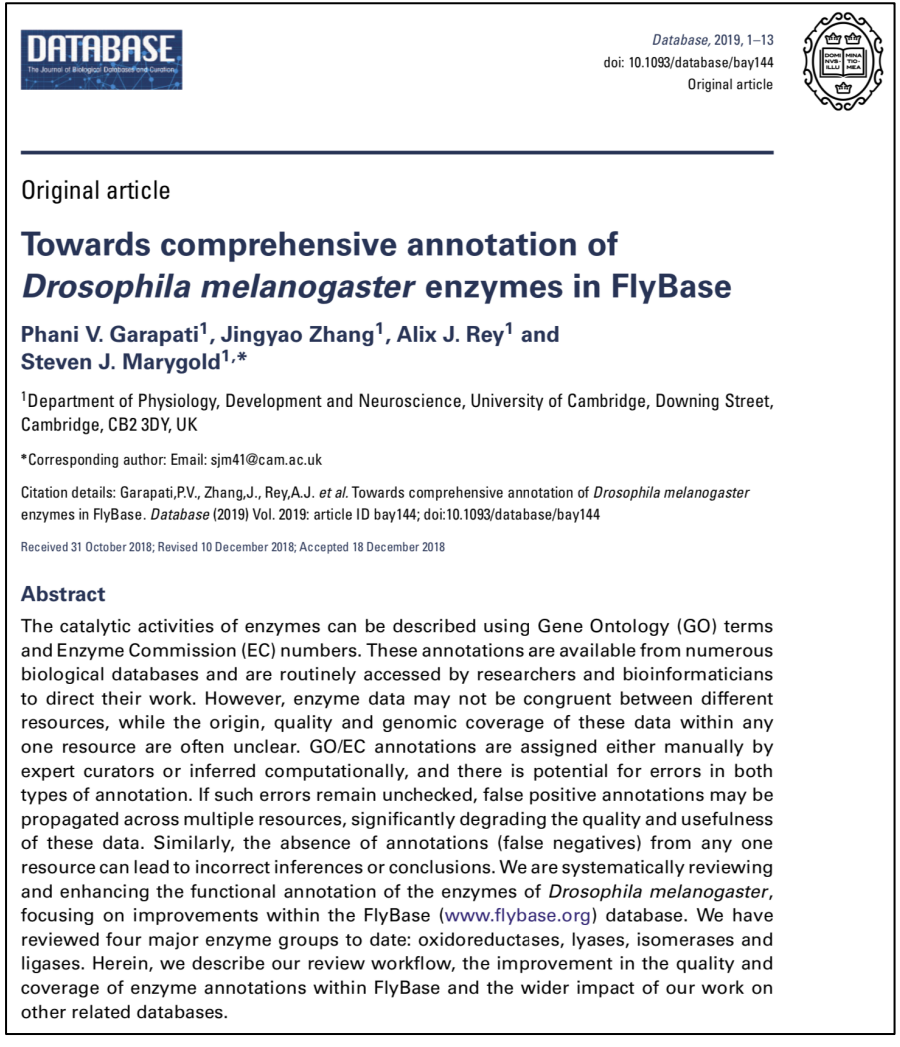
Alix Rey

Helen Attrill

GO Consortium, especially Harold Drabkin

FlyBase Consortium

Funding: NHGRI at the US NIH (U41HG000739)



DATABASE
The Journal of Biological Databases and Curation

Database, 2019, 1–13
doi: 10.1093/database/bay144
Original article

Original article

Towards comprehensive annotation of *Drosophila melanogaster* enzymes in FlyBase

Phani V. Garapati[1], Jingyao Zhang[1], Alix J. Rey[1] and Steven J. Marygold[1,*]

[1]Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge, CB2 3DY, UK

*Corresponding author: Email: sjm41@cam.ac.uk

Citation details: Garapati,P.V., Zhang,J., Rey,A.J. *et al.* Towards comprehensive annotation of *Drosophila melanogaster* enzymes in FlyBase. *Database* (2019) Vol. 2019: article ID bay144; doi:10.1093/database/bay144

Received 31 October 2018; Revised 10 December 2018; Accepted 18 December 2018

Abstract

The catalytic activities of enzymes can be described using Gene Ontology (GO) terms and Enzyme Commission (EC) numbers. These annotations are available from numerous biological databases and are routinely accessed by researchers and bioinformaticians to direct their work. However, enzyme data may not be congruent between different resources, while the origin, quality and genomic coverage of these data within any one resource are often unclear. GO/EC annotations are assigned either manually by expert curators or inferred computationally, and there is potential for errors in both types of annotation. If such errors remain unchecked, false positive annotations may be propagated across multiple resources, significantly degrading the quality and usefulness of these data. Similarly, the absence of annotations (false negatives) from any one resource can lead to incorrect inferences or conclusions. We are systematically reviewing and enhancing the functional annotation of the enzymes of *Drosophila melanogaster*, focusing on improvements within the FlyBase (www.flybase.org) database. We have reviewed four major enzyme groups to date: oxidoreductases, lyases, isomerases and ligases. Herein, we describe our review workflow, the improvement in the quality and coverage of enzyme annotations within FlyBase and the wider impact of our work on other related databases.