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.

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

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 negatives), while an additional 20 (false positives) 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
- 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

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


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.