FlyBase Gene Model Annotations: The Rule-Benders

767B

Madeline Crosby, Gil dos Santos, Sian Gramates, Beverley Matthews, Susan E. St. Pierre, David Emmert, Pinglei Zhou, Andrew Schroeder, Kathleen Falls, Susan Russo, William Gelbart, and the FlyBase Consortium.

ABSTRACT Biology seems to have exceptions to every rule (if we look hard enough). In the context of the FlyBase annotated gene models in *D. melanogaster*, we summarize aspects of the many exceptional cases we have identified or curated from the literature. These range from non-canonical splices (relatively uncommon), dicistronic and polycistronic transcripts (surprisingly common), and stop-codon read-through (probably not uncommon), to non-canonical translation start codons (not enough data yet) and trans-splicing (uncommon, we hope). Whenever possible, we use Sequence Ontology (SO) terms to flag exceptional genes (Table 1). We also describe how genes affected by mutations in the sequenced strain are treated in FlyBase.

Gene-associated SO term	SO ID
gene_with_dicistronic_mRNA	SO:0000722
gene_with_polycistronic_transcript	SO:0000690
gene_with_stop_codon_read_through	SO:0000697
gene_with_stop_codon_redefined_as_selenocysteine	SO:0000710
gene_with_trans_spliced_transcript	SO:0000459
gene_with_unconventional_translation_start_codon	SO:0001739
gene_with_translation_start_codon_CUG	SO:0001740
gene_with_transcript_with_translational_frameshift	SO:0000712

Table 1. Sequence ontology terms used by FlyBase for exceptional gene models.

Non-canonical splices are relatively rare.

For 99% of the ~58,500 annotated introns, the primary canonical splice donor-acceptor pair GT-AG is used; for most of the remaining 1% the secondary canonical splice donor-acceptor pair GC-AG is used. The frequency of introns for which other splice donor-acceptor pairs are used is very low: only 72 are annotated in current gene models. The distribution is non-random; 10 genes are annotated with more than one non-canonical splice. For many classes (see Table 2), similar alternative splices are common.

Splice donor-acceptor pair	Number in r5.56 (corrections)	Number with RNA-Seq junction support	Number with similar alternative splice	coding/5' UTR
AT-AC (U12)	9	9	1	9/0
AT-AC (U2)	4	4	2	4/0
GT-TG	21 (22)	17	20	12/8 (1 ncRNA)
GT-GG	5	4	4	3/1 (1 ncRNA)
GT-CG	6	6	6	5/1
GT-AT	15 (14)	11	14	11/3
GT-AA	2	2	2	2/0
GA-AG	10	10	4	8/2
GG-AG	0	-	-	-
GT-AC	0	-	-	-
TOTAL	72	63	53	54/15 (2 ncRNA)

Table 2. Tabulation of introns with non-canonical splice sites.

Polycistronic (primarily dicistronic) transcripts are not uncommon.

There are 155 dicistronic gene pairs in the r5.56 annotation release. In addition, there are 8 polycistronic sets with transcripts encoding more than two genes (6 tricistronic and 2 tetracistronic), making a total of 336 genes annotated as sharing one or more transcripts with a neighboring protein-coding gene. Thus, 2.4% of gene models include at least one polycistronic transcript.

Genes that encode small polypeptides are overrepresented among polycistronic genes. There are 11 genes for which all annotated polypeptides are less than 25 amino acids; all are polycistronic, one is also annotated with an alternative monocistronic transcript. Of the 82 genes for which all annotated polypeptides are between 25 and 49 amino acids, 16 (19.5%) are polycistronic. A number in this category correspond to small conserved ORFs found in the UTRs of longer coding gene; these are annotated as separate genes.

Trans-splicing is supported for two genes.

At least two genes in *D. melanogaster* undergo trans-splicing, a process by which a mature mRNA is created by a bimolecular splice between two independently transcribed pre-mRNAs. In both cases, the gene encodes multiple DNA-binding proteins with a common amino BTB/POZ domain and variable carboxy zinc-finger domains. The initial and more dramatic example is *mod(mdg4)* (Labrador, et al, 2001; Dorn, et al., 2001), which encodes over 30 protein isoforms, at least 18 of which appear to be trans-spliced. The second example is *lola*, a gene that encodes at least 20 protein isoforms, one of which shows evidence of being trans-spliced (Horiuchi, et al., 2003). 3' trans-splicing precursors with sufficient support, including transcription start site data, are annotated as separate genes in FlyBase. For *mod(mdg4)* there are 4 polycistronic clusters encoding the 18 different 3' alternative exons. For *lola* a monocistronic trans-splicing precursor that encodes the 3'-most alternative exon has a high level of support.

Non-canonical translation starts have been difficult to identify.

In FlyBase, there are currently 25 genes with transcripts annotated with a non-AUG translation start codon; 11 of these use a CUG start codon. Thus far, no systematic or definitive study of non-AUG translation initiation in Drosophila has been carried out. Individual cases have been discovered more or less by chance; only a handful have been thoroughly characterized (example in Figure 1). This may change in the near future. Recent systematic studies in human and mouse (Ivanov et al., 2011; Ingolia et al., 2011) have allowed identification of many additional non-AUG translation starts; all were near-cognates (differ by one base) of AUG, and CUG was the most common; alternative translation starts were common.

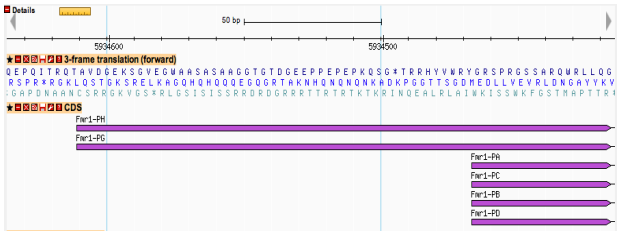

Figure 1. CUG start codon in *Fmr1* results in a 48-aa N-terminal extension. Use of this alternative start codon has been confirmed by Western blot, mutagenesis of reported constructs and rescue constructs (Beerman and Jongens, 2011).

Some exceptional gene models reflect highly conserved phenomena.

Translational frameshift: A programmed translational frameshift occurs in the Ornithine decarboxylase antizyme *(Oda)* from yeast to mammals and serves as a regulatory mechanism controlling the level of polyamines (Ivanov et al., 2000).

HAC1-type intron splice: The X-box binding protein 1 *(Xbp1)* is subject to a specific non-standard intron-processing event that is conserved from yeast to mammals and functions as part of a response to ER stress (Plongthongkum, et al., 2007).

Selenocysteine coding alternative at stop codon: Selenoproteins are produced by incorporation of the amino acid selenocysteine at specific UGA codons and are found in bacteria to mammals. Three selenoproteins have been identified in *D. melanogaster*, in contrast to at least 25 in humans (Roman, et al., 2014).

U12 spliceosome introns: 16 of the 20 introns with sequences recognized by the U12 spliceosome in *D. melanogaster* correspond to a U12-spliced intron in the orthologous human gene (Lin, et al., 2010).

Stop-codon readthroughs appear to be common.

Based primarily on the conservation of protein signatures beyond existing stop codons (Jungreis, et al., 2011), 319 gene are currently annotated with one or more transcripts subject to stop-codon readthrough (Figure 2); in 18 cases a double readthrough is supported. This phenomenon appears to be distinct from the selenocysteine system. A ribosome profiling assay using early embryos and S2 cells confirmed 43 of the previously identified cases and identifies more than 300 new candidates (Dunn et al., 2013).

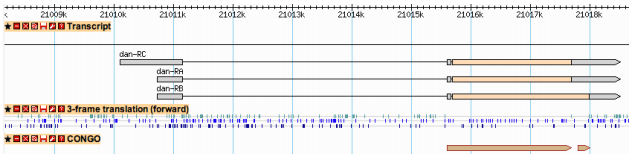

Figure 2. A stop-codon readthrough annotated for dan-RB is supported by CONGO analysis (conservation of protein signatures).

The sequenced genome contains over 50 annotated mutations.

The original *D. melanogaster* sequenced strain (iso-1 in FlyBase) carried the visible mutations *y[1]; cn[1] bw[1] sp[1].* In addition to these known mutations, many other genes in the sequenced genome have been determined to contain gross mutational alterations such as indels, frameshifts or nonsense mutations. These genes are flagged in the Gene Model Comments as "Mutation in sequenced strain." Since one of the goals of genome annotation is to produce a wild-type representation of the proteome, FlyBase replaces the mutant polypeptide sequences with wild-type sequences. This results in a CDS that does not match the corresponding transcript and thus is marked as an exception in RefSeq entries.

References available as handout.