FlyBase Gene Model Annotations: Impact of High Throughput Data


Abstract

We report the current status of the FlyBase annotated gene set for D. melanogaster and highlight improvements based on high-throughput data. The FlyBase annotated gene set consists entirely of manually annotated gene models (with the exception of some classes of small non-coding RNAs). All gene models have now been reviewed using evidence from next-high-throughput datasets, primarily from the modENCODE project. These datasets include RNA-Seq coverage data, RNA-Seq junction data, transcription start site profiles, and translation stop-codon read-through predictions (see poster 7948 for discussion of stop-codon read-through data). We describe how this flood of new data was incorporated into new annotation guidelines. FlyBase has adopted a philosophy of excluding low-confidence and low-frequency data from gene model annotations; we do not attempt to represent all possible permutations in the case of complex and modularly organized genes. This has allowed us to produce a high-confidence, manageable gene annotation dataset that is available as bulk download files, in gene reports, and on GBrowse views. Interestingly aspects of new annotations include new genes (coding, non-coding, and unannotated), many genes with alternative transcripts with very long 5' UTRs (up to 15,189bp), and a stunning increase in the number of male-specific genes (roughly 10 percent of all annotated gene models) vs. female-specific genes (far fewer than 1 percent).

Challenges remain for gene model annotation, for instance, identification of functional small polypeptides and detection of alternative translation starts.

Gene model annotation statistics: counts at significant timepoints

RNA-Seq Coverage Data

New Genes

Long non-coding RNAs (lncRNAs)

• Sequence-specific coverage data is required to reliably annotate lncRNAs.

• Tissue-specific lncRNAs are common, especially male-specific and CNS-specific. Very few female-specific lncRNAs are annotated.

• Number of lncRNA has increased 16x since release 5.12.

Coding vs. non-coding

• In absence of other intrinsic support, conservation across closely-related species is considered, especially conservation of ATG start sites.

• Without evidence of conservation, genes is categorized as non-coding and a comment added indicating that it may encode a polypeptide.

Transcription Start Sites

Release 5.12

Release 5.16

Alternative Transcripts: Permutations and combinations (2012 guidelines)

• Alternative transcripts are annotated based on cDNA/EST data, RNA-Seq data, and community data.

• Almost all alternative transcripts are now supported by RNA-Seq data.

• Frequently, RNA-Seq junction data supports many alternative splices within the 5' UTR of a gene; for a given TSS, all such splices may not be annotated.

• Transcripts that are of lower frequency than alternative junctions may not be annotated.

• Scaffolding low frequency junctions, all alternative splices within the CDS and all promoters are represented, but not necessarily all possible splices.

Transcription Start Sites

The Apollo annotation tool sets the translation start site to the 5'-most in-frame ATG. But, in cases supported by the literature (including conservation patterns across Drosophila species), a non-ATG translation start site, or a downstream ATG may be used. In those cases comments are added and appear in the "Comment" section of the relevant transcript report.

RNA-Seq Exon Junctions

New 5' end based on junction (and coverage data)

• New transcript based upon junction; RNA-Seq coverage supports spanning CDS ON.

• Evidence as of 5.12 that no support for alternative 5' end

• Read count for junction supporting 5' end is 136. Read count for junction supporting short 5' intron is 36.

Extended UTRs

Annotated 3' UTRs

• IF polyadenylated RNA is available, most transcripts are extended 3' to the last non-A nucleotide of the longest CDS.

• RNA-Seq coverage data supports 3' UTR sequences beyond those present in cDNA. At least one transcript in extended 3' to the approximate terminus supported by the RNA-Seq data (i.e red bracket is in a panel below).

• Many extended 3' UTRs have been annotated. There are 2727 transcripts with the extended 3' UTR annotated based on the transcript context.

• See panel in upper right (corta gene) for additional example

FlyBase

Sequence Feature View

RNA-Seq Coverage Data

Extended UTRs

Annotated 3' UTRs