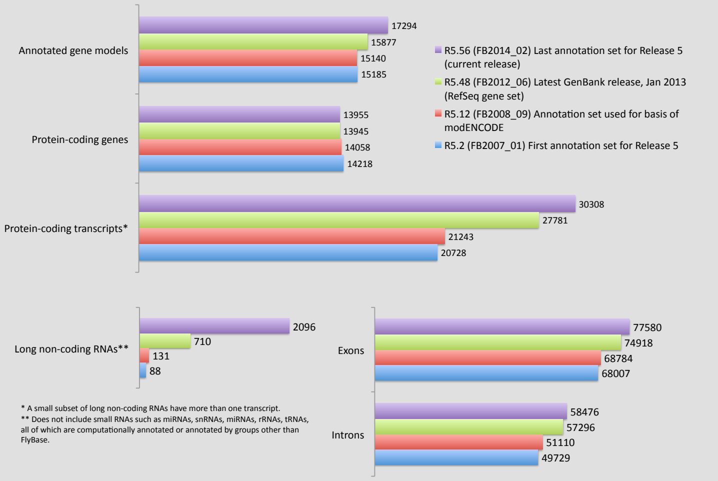
FlyBase Gene Model Annotations: Impact of High Throughput Data 774C

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

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 new 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 767B 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 also 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. Interesting aspects of new annotations include new genes (coding, non-coding, and antisense), many genes with alternative transcripts with very long 3' UTRs (up to 15-18kb), and a stunning mismatch in the number of male-specific genes (roughly 10 percent of all annotated gene models) vs. female-specific genes (fewer than 1 percent). Challenges reamain 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



* A small subset of long non-coding RNAs have more than one transcript.
** Does not include small RNAs such as miRNAs, snRNAs, miRNAs, rRNAs, tRNAs, all of which are computationally annotated or annotated by groups other than FlyBase.

RNA-Seq Coverage Data

New Genes

Long non-coding RNAs (lncRNAs)

- Strand-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 lncRNAs has increased 16X since release 5.12.

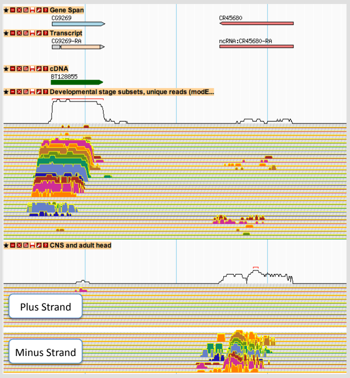
Coding vs. non-coding

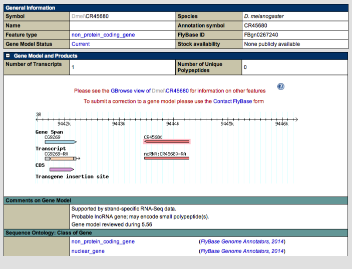
- In absence of other proteomic support, conservation across closely-related species is considered, especially conservation of ATG start site.
- Without evidence of conservation, gene is categorized as non-coding and a comment added indicating that it may encode a polypeptide.

Extended UTRs

Annotating 3' Extents:

- If a polyadenylated cDNA is available, most transcripts are extended 3' to the last non-A nucleotide of the longest cDNA.
- If RNA-Seq coverage data support 3' UTR sequences beyond those present in a cDNA, at least one transcript is extended 3' to the approximate terminus supported by the RNA-Seq data (see red bracket in panel below).
- Many extended 3' UTRs have been annotated. There are 2772 transcripts with the "extended 3' UTR" comment found on the transcript report.
- See panel in upper right (*corto* gene) for additional example





General Information			
Symbol	Dmel\CR45680	Species	*D. melanogaster*
Name		Annotation symbol	CR45680
Feature type	non_protein_coding_gene	FlyBase ID	FBgn0267240
Gene Model Status	Current	Stock availability	None publicly available

⊟ Gene Model and Products

Number of Transcripts	1	Number of Unique Polypeptides	0

Please see the GBrowse view of Dmel\CR45680 for information on other features

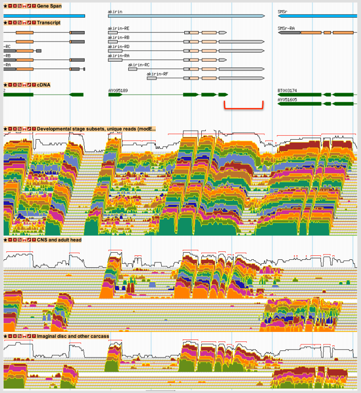
To submit a correction to a gene model please use the Contact FlyBase form

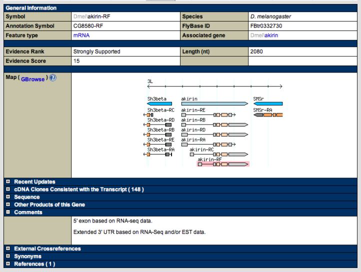
Comments on Gene Model

Supported by strand-specific RNA-Seq data.
Probable lncRNA gene; may encode small polypeptide(s).
Gene model reviewed during 5.56

Sequence Ontology: Class of Gene

non_protein_coding_gene	(*FlyBase Genome Annotators, 2014*)
nuclear_gene	(*FlyBase Genome Annotators, 2014*)





General Information			
Symbol	Dmel\akirin-RF	Species	*D. melanogaster*
Annotation Symbol	CG8580-RF	FlyBase ID	FBtr0332730
Feature type	mRNA	Associated gene	Dmel\akirin
Evidence Rank	Strongly Supported	Length (nt)	2080
Evidence Score	15		

Map (GBrowse)

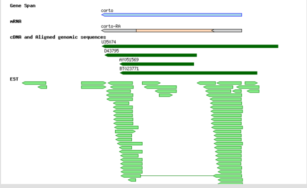
⊞ Recent Updates
⊞ cDNA Clones Consistent with the Transcript (148)
⊞ Sequence
⊞ Other Products of this Gene
⊟ Comments

5' exon based on RNA-seq data.
Extended 3' UTR based on RNA-Seq and/or EST data.

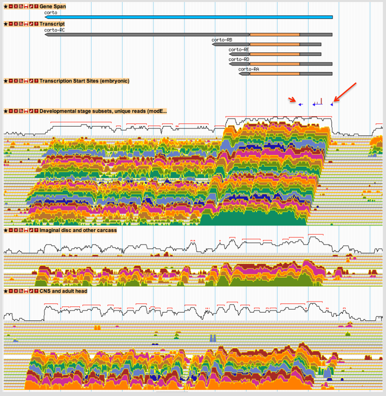
⊞ External Crossreferences
⊞ Synonyms
⊞ References (1)

Transcription Start Sites

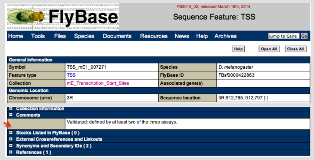
Release 5.12



Release 5.56


Gene Span
corto
Transcript
corto-RC
corto-RB
corto-RE
corto-RD
corto-RA
Transcription Start Sites (embryonic)
Developmental stage subsets, unique reads (modE...
Imaginal disc and other carcass
CNS and adult head

Long arrow (in figure to right): Links to Sequence Feature report for TSS (below), includes "Comment" describing validity of TSS. "Validated TSS" are annotated at the 90% cummulative frequency point. "Supported" TSS are annotated only if they align with other evidence. Short arrow (in figure to right) indicates unannotated TSS.



FB2014_02, released March 18th, 2014

Sequence Feature: TSS

General Information			
Symbol	TSS_mE1_007271	Species	*D. melanogaster*
Feature type	TSS	FlyBase ID	FBsf0000422863
Collection	mE_Transcription_Start_Sites	Associated gene(s)	
Genomic Location			
Chromosome (arm)	3R	Sequence location	3R:912,785..912,797 [-]

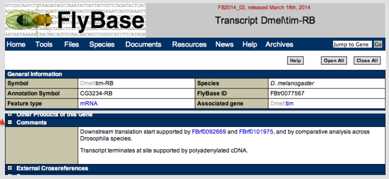
⊞ Collection Information
⊟ Comments

Validated: defined by at least two of the three assays.

⊞ Stocks Listed in FlyBase (0)
⊞ External Crossreferences and Linkouts
⊞ Synonyms and Secondary IDs (2)
⊞ References (1)

Translation 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 these cases comments are added and appear in the "Comment" section of the relevant transcript report.



FB2014_02, released March 18th, 2014

Transcript Dmel\tim-RB

General Information			
Symbol	Dmel\tim-RB	Species	*D. melanogaster*
Annotation Symbol	CG3234-RB	FlyBase ID	FBtr0077567
Feature type	mRNA	Associated gene	Dmel\tim

⊞ Other Products of this Gene
⊟ Comments

Downstream translation start supported by FBrf0092669 and FBrf0101975, and by comparative analysis across Drosophila species.

Transcript terminates at site supported by polyadenylated cDNA.

⊞ External Crossreferences

RNA-Seq Exon Junctions

	Release 5.45 (May 2012)	Release 5.56 (March 2014)
Total RNA-Seq Junctions (modENCODE)	71082	71382
Annotated Introns	57986	58476
Annotated Junctions (Junctions corresponding to annotated introns)	53734 (92.7%)	57363 (98.1%)
Analysis of Annotated Junctions	Average Read Counts: 4724 (modENCODE) 289 (Baylor) Average Entropy Score*: 4.987	Average Read Count: 4452 (modENCODE) 272 (Baylor) Average Entropy Score: 4.993
Unannotated Junctions	17348	14019
Analysis of Unannotated Junctions	Average Read Counts: 110 (modENCODE) 3 (Baylor) Average Entropy Score: 3.641	Average Read Counts: 79 (modENCODE) 1.8 (Baylor) Average Entropy Score: 3.523

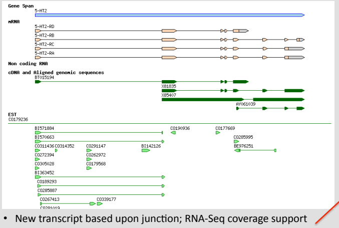
Alternative Transcripts: Permutations and combinations (2012 guidelines)

- Alternative transcripts are annotated based on cDNA/ EST data, RNA-Seq data, and community data.
- Almost all alternative transcripts are now supported by RNA-based data.
- Frequently, RNA-Seq junction data support many alternative splices within the 5' UTR of a gene. For a given TSS, all such splices may not be annotated.
- RNA-Seq junctions that are of much lower frequency than alternative junctions may not be annotated
- Excluding low-frequency junctions, all alternative splices within the CDS and all promoters are represented, but not necessarily all possible combinations.

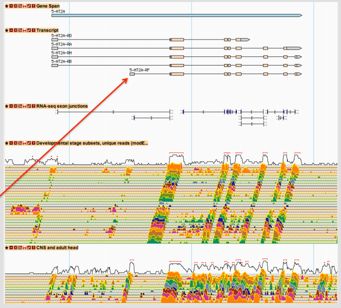
*Entropy: The entropy score is a function of both the total number of reads that map to a given junction and the number of different offsets to which those reads map and the number that map at each offset. Thus, junctions with multiple reads mapping at each of the possible windows across the junction will be assigned a higher entropy score, than junctions where many reads map to only one or two positions. (Graveley, BR et al. 2011).

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

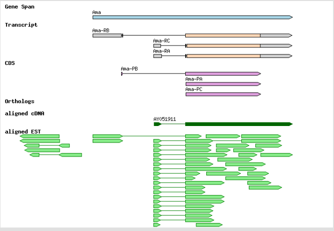
Release 5.12

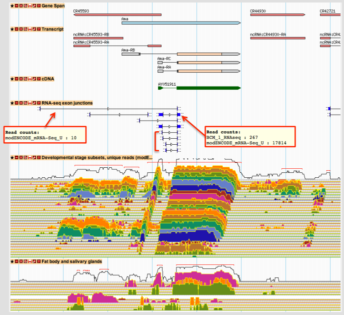


Release 5.56



- New transcript based upon junction; RNA-Seq coverage support especially strong in CNS.
- Evidence as of 5.12 had no support for alternative 5' end.
- Read count for junction supporting long 5' intron is 136. Read count for junction supporting short 5' intron is 38.





- Low frequency junctions are not annotated. Note 5' unannotated junction (with readcount box) and junctions within 5' UTR (red bracket)
- Gene Model comments indicate when junctions that fall within the gene area are not annotated.
- Identification of lncRNA on opposite strand based on RNA-Seq coverage and junction data