# Impact of High Throughput Data on the *Drosophila melanogaster* Annotation Set

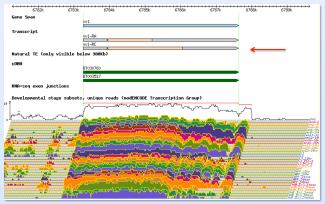
Beverley Matthews, Madeline Crosby, Gil dos Santos, Sian Gramates, Susan St. Pierre, William Gelbart, and the FlyBase Consortium, Harvard University, Cambridge, MA, 02138

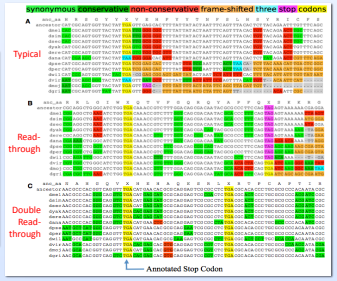
## Summary

Over the past two years, high throughput data from the modENCODE project has been incorporated into FlyBase and used to manually refine the Drosophila melanogaster annotation set. The RNA-seq data, consisting of exon junctions and coverage data from 30 developmental stages as well as strand-specific data from tissues and cell lines, has led to the creation of numerous new exons, UTR extensions, and alternative splice forms. It has also provided evidence for gene merges and splits. Additionally, the RNA-seq data has allowed us to annotate many more ncRNAs and some anti-sense RNAs. The modENCODE transcription start site data is being incorporated using the 90% point as the starting point for our annotations. A set of 283 stop-codon readthrough predictions has been annotated.

With the abundance of data come some challenges. Many of the low frequency RNA-seq exon junctions lead to truncated polypeptides, which could be interesting regulatory features or biological noise. The same applies to cDNAs with retained introns. We have established criteria for when to create annotations from these events. For complex gene models, the possible combinations of alternative exons and/or alternative promoters may be more than we can practically annotate. We have established standardized comments to indicate when we have not created annotations from the low frequency exon junctions or from all of the possible exon combinations. Our revised gene model annotation guidelines are available at http://flybase.org/static_pages/docs/refman/refman-G.html#G8.

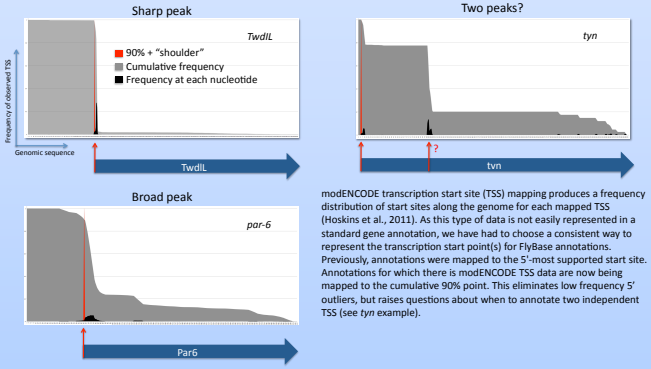
### Stop codon readthrough






The 283 cases of stop codon readthrough identified by Jungreis et al., 2011 on the basis of comparative genomic analysis have been annotated. In the left panel, transcript vvl-RA stops at the first stop codon, while transcript vvl-RC has been annotated to extend through the first stop codon and stop at the second one. The right panel shows alignments of 12 Drosophila species in the region of the annotated stop codons for three genes which illustrate the protein-coding evolutionary signatures for typical, readthrough, and double-readthrough stop codons. (Figure from Jungreis, et al., 2011)

### Transcription start sites



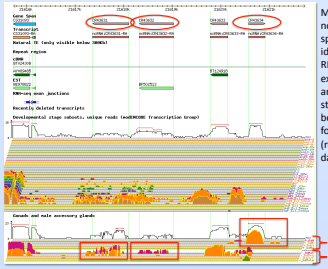

modENCODE transcription start site (TSS) mapping produces a frequency distribution of start sites along the genome for each mapped TSS (Hoskins et al., 2011). As this type of data is not easily represented in a standard gene annotation, we have had to choose a consistent way to represent the transcription start point(s) for FlyBase annotations. Previously, annotations were mapped to the 5'-most supported start site. Annotations for which there is modENCODE TSS data are now being mapped to the cumulative 90% point. This eliminates low frequency 5' outliers, but raises questions about when to annotate two independent TSS (see *tyn* example).

## Improvements to annotations

Incorporation of cDNA and RNA-seq junction and expression data has led to many annotation updates including new 5' exons/promoters, new internal exons, additional alternative splice forms, gene merges, gene splits, and extended 3' UTRs. It has also resulted in the creation of many new genes, both coding and non-coding.

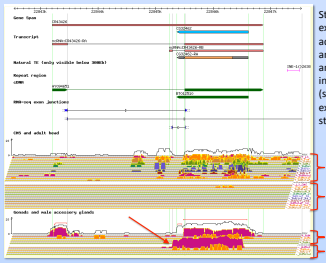
### New 5' exons


### New genes – ncRNAs



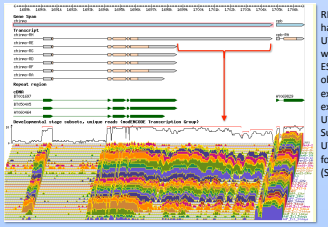

Many new genes (many non-coding, mostly male-specific) have been identified on the basis of RNA-seq junction and expression data (red circles around new genes). The strand-specific data has been particularly valuable for annotating ncRNA genes (red boxes around RNA-seq data).

### New genes – antisense RNAs




Strand-specific RNA-seq expression data provides additional evidence that antisense cDNAs should be annotated. This is often seen in a male-specific pattern (see male-specific expression on the minus strand, red arrow)

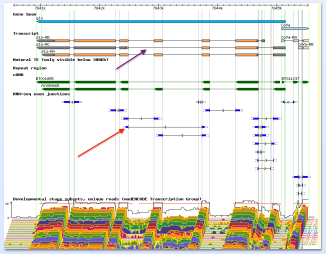
### Extended 3' UTRs



RNA-seq expression data has been used to extend 3' UTRs. Previously, 3' UTRs were based on the longest ESTs/cDNAs. Where observed with RNA-seq expression profile expression data, extended 3' UTRs are now annotated. Surprisingly, very long 3' UTRs have been observed, for some genes, up to 18kb (Smibert, et al., 2011).

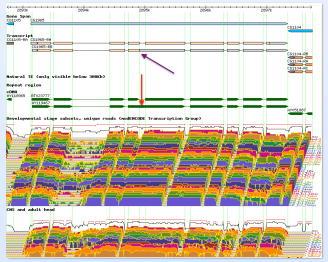
## Annotation Challenges

### Transcripts that result in truncated polypeptides – alternative splices



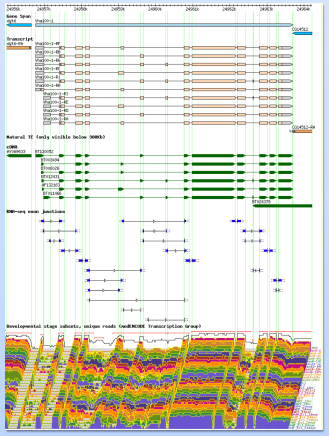
Low frequency exon junctions (red arrow) often lead to frameshifts, causing truncated polypeptides (purple arrow). Alternative exon junctions within coding regions are annotated as separate transcripts if in frame and at least 1% of the highest junction count for the gene. If out of frame, they are only annotated if the level is at least 10%.

### Transcripts that result in truncated polypeptides – retained introns



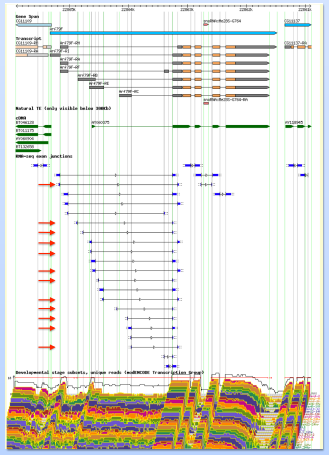
cDNAs with retained introns (red arrow) often lead to transcripts with truncated polypeptides (purple arrow). Annotations are created from these if supported by RNA-seq expression data. We have plans to reclassify these as non-coding alternative transcripts of protein-coding genes.

### Permutation problem- how many alternative exon/promoter combinations to capture?



For genes with multiple promoters and/or multiple sites of alternative splicing or mixing and matching of entire exons, the number of possible permutations quickly becomes very large. It is impractical to represent all possible combinations and it is often unknown which combinations are actually produced in vivo. We do not attempt to annotate all possible combinations but rather, represent each alternative exon and promoter in at least one annotation.

### How many 5' UTR variants to capture?



Priority is placed on creating annotations that encode new CDS variants. 5' UTR variants are captured if they represent separate promoters or constitute a significant portion of the total transcripts. In this example, many more 5' exon variants are supported than have been annotated (red arrows show unannotated junctions).

**Link to FlyBase Gene Model Annotation Guidelines**
http://flybase.org/static_pages/docs/refman/refman-G.html#G8

**modENCODE papers:**

RNA-Seq (coverage, junctions, 3' UTRs)
Graveley, et al. (2011) Nature 471:473-479
Cherbas, et al. (2011) Genome Res. 21:301-314
Smibert, et al. (2012) Cell Reports 23 Feb. 2012
Stop-codon readthrough
Jungreis, et al. (2011) Genome Res. 21:2096-2113
Transcription start site data
Hoskins, et al. (2011) Genome Res. 21:182-192