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.

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.

For complex gene features or biological noise. The same applies to our annotations. A set of 283 stop-allowed us to annotate many more also provided evidence for gene merges and splits. Additionally, the RNA-seq 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 comparing genomics analysis have been annotated. In two-left panel, transcript cut RNA stop at the first stop codon, while transcript which 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 stop, readthrough, and double-readthrough stop codons. (Figure from Jungreis, et al., 2011)

Transcription start sites

Sharp peak

Two peaks?

Broad peak

modENCODE transcription start site (TSS) mapping produces a frequency distribution of start sites along the genome for each mapped TSS (Cheng et al., 2011). An X-axis type of data is not easily represented in a standard gene annotation, we have found a way to choose a consistent way to represent the transcription start point site for Flybase annotations. Previously, annotations were mapped to the 5'-most supported start site. Annotations for which there is no evidence. The data we now being able to annotate a very high number of new start sites, but raise questions about when to annotate two independent TSS (see 5'-syn example).

New 5' exons

Glycine gene exon

Gene exists

Annotated transcripts

Natural transposon

RNAs and RNase

new spliceosome development

Many new genes (many non-coding, mostly male- specific) have been annotated. Incorporation of new splice sites and expression data and criteria around new genes. The strand-specific splice data has been particularly valuable for annotating new genes (red boxes around RNA-seq data).

New genes – ncRNAs

New genes – antisense RNAs

Permutation problem: how many alternative exons/promoter combinations to capture?

How many 5' UTR variants to capture?

Transcripts that result in truncated polypeptides – alternative splices

Transcripts that result in truncated polypeptides – retained introns

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 that have been annotated (red arrows show annotated 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, 5' UTRs)


Stop-codon readthrough

Jungreis, et al. (2011) Genome Res. 21:2095-2113

Transcription start site data


Impact of High Throughput Data on the Drosophila melanogaster Annotation Set

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

Summary

Improvements to annotations

Annotation Challenges