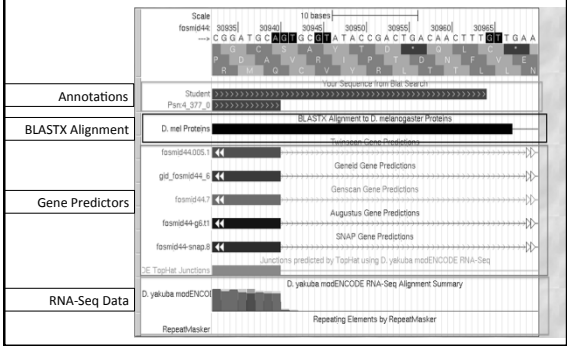
Common Errors in Student Annotation Submissions

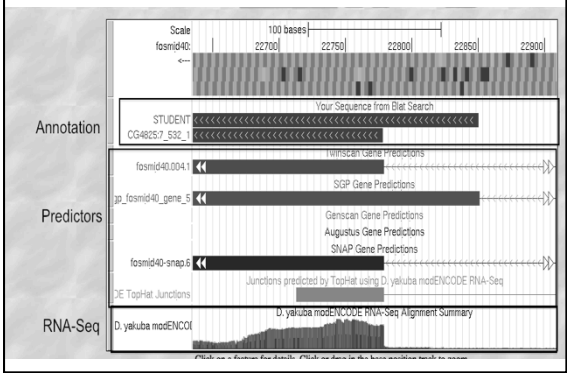
contributions from
Paul Lee, David Xiong, Thomas Quisenberry

- Annotating multiple genes at the same locus based on BLASTX alignments
- Over-reliance on BLAST alignments
- Over-reliance on gene predictors
- Not annotating all genes or all isoforms
- Missing small exons
- Annotating incorrect splice sites

Over-reliance on BLASTX alignment

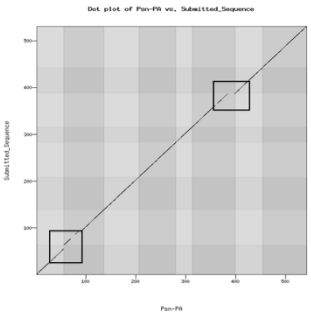


Relying on a single gene predictor



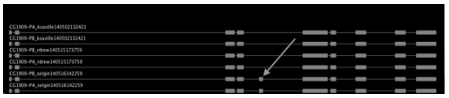
Strategies to resolve common errors

- Dot plot
- TBLASTN / BLASTX with exon by exon strategy
- RNA-Seq
- Identify small coding exons using "Small Exon Finder"
- Use dot plot and peptide sequence alignment to check



An interesting annotation problem: contig34 (Liz Chen's project from Bio 4342), reconciliation by Thomas Quisenberry

Submitted annotations:



Did Liz include an extra exon at 32298-32363? Her model has 10 exons, while the *Drosophila melanogaster* model only has 9.

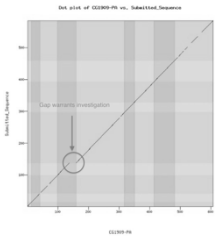
CDS usage map:

Liz's annotation places an extra exon here

Isoform	3_1780_0	4_1780_1	5_1780_2	6_1780_0	7_1780_1	8_1780_0	9_1780_0	10_1780_0	11_1780_1
CG1909-RA	Y	Y	Y	Y	Y	Y	Y	Y	Y
CG1909-RB	Y	Y	Y	Y	Y	Y	Y	Y	Y

Continuing investigation of contig34

Checked other student's submission forms for *CG1909,* the gene in question:



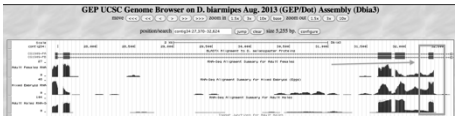
←y-axis= student annotation submission; x-axis=*D. melanogaster* gene model

←Gap (red) indicates residues in *D. melanogaster* gene that are not present in student annotation

←All in all, this dot plot warrants further investigation

contig34 continuing investigation

Check UCSC Genome Browser view for this gene in *D. biarmipes*:



- Above: blue box marks BLASTX alignment and RNA-Seq data in the region of extra exon.
- Right-hand exon (fifth) is supported by RNA-Seq data, conservation.
- Below: TBLASTN results using a.a. sequence of fourth exon in *D. melanogaster* model as the query and nucleotide sequence of contig34 as the subject → two regions of conservation

Range 1: 31870 to 31995 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
88.2 bits(217)	4e-24	Compositional matrix adjust.	40/42(95%)	41/42(97%)	0/42(0%)	+1

```
Query  1      IERGRRLYEQNNQTEAVRTWRSALKGTCQREDCFQLLGYLYQ  42
              IERGRRLYEQNNQTEAVRTWRSALKGTCQ EDCFQLLGYLY+
Sbjct  31870  IERGRRLYEQNNQTEAVRTWRSALKGTCQPEDCFQLLGYLYE  31995
```

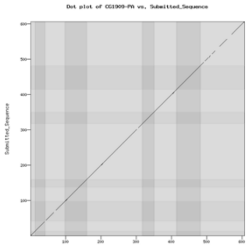
Range 2: 32299 to 32361 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
47.0 bits(110)	5e-10	Compositional matrix adjust.	20/21(95%)	20/21(95%)	0/21(0%)	+1

```
Query  41     YQAHMDWGKFREAIEFSHQQL  61
              YQAHMDWGKFREAIEF HQQL
Sbjct  32299  YQAHMDWGKFREAIEFGHQQL  32361
```

contig34 completed ☺

Gene model checker dot plot output for model including additional exon

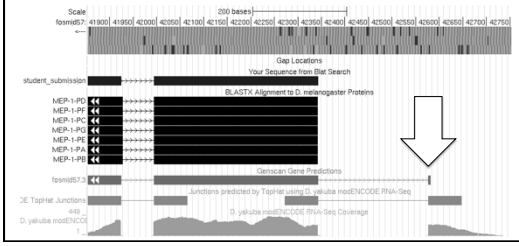


Much better than before!

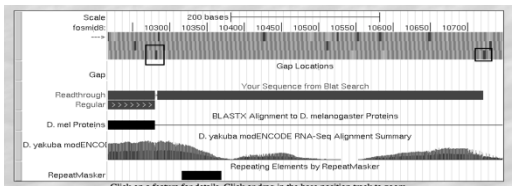
- Amino acid sequence conserved
- Appropriate splice junctions maintaining ORF identified
- Model has 1 more exon

Using RNA-Seq and *TopHat* to identify 5' and 3' UTR, start and stop codons

FlyBase ID	5' Start	3' End	Strand	Phase	Length
9_407_0	2,637,133	2,639,553	+	0	807
10_407_0	2,639,615	2,639,872	+	0	86
11_407_0	2,640,026	2,640,271	+	0	82
12_407_0	2,640,337	2,640,516	+	0	60
13_407_0	2,640,578	2,640,926	+	0	116
14_407_2	2,641,012	2,641,016	+	2	1



Interesting annotation challenges: Read-through stop codons



Comments on Gene Model	
	Gene model reviewed during 5.44
	Gene model reviewed during 5.45
	Stop-codon suppression (UGA) postulated; FBrf0216884.
	gene_with_stop_codon_read_through ; SO:0000697

Jungreis *et al* "Evidence of abundant stop codon read-through in Drosophila and other metazoa." Genome Res. 2011 21: 2096-113.

Interesting annotation challenges Errors in the consensus sequence

```
 Score =  335 bits (858),  Expect(2) = 1e-165, Method: Compositional matrix adjust.
 Identities = 157/173 (91%), Positives = 166/173 (96%), Gaps = 0/173 (0%)
 Frame = +2

Query  1      HLFMEPHPSVSILEADMVDFTGLTTMEVSDLVAILSELFVSFDLAANNNRATRIKFLGD  60
              HLFMEPHPSVSILEADMV+FT LTTM+VS+LVAILSELFVSFDLAAN NRATRIKFLGD
Sbjct  20144  HLFMEPHPSVSILEADMVNFTNLTTMDVSELVAILSELFVSFDLAANQNRATRIKFLGD  20323

Query  61     SYTCVTGIPSYFPTHANACVNQALDMIEISRESVKRRNKKIDLRIGVHSGEILAGIIGLT  120
              SYTCVTGIPSYF  HANACVNQALDMIEISRE S+RRN+KI+LR+GVHSGEILAGIIGLT
Sbjct  20324  SYTCVTGIPSYFSAHANACVNQALDMIEISRESAQRRNQKIELRVGVHSGEILAGIIGLT  20503

Query  121    KWQFDIWSKDVDITNRLESSGLPGMVHISSRTLGLLDNNYVYEEGTDTAKLDP  173
              KWQFDIWSKDVDITNRLE+SGLPGMVHISSRTLGLLDN YVYEEGT+TAK DP
Sbjct  20504  KWQFDIWSKDVDITNRLASGLPGMVHISSRTLGLLDNQYVYEEGTETAKQDP  20662


 Score =  258 bits (658),  Expect(2) = 1e-165, Method: Compositional matrix adjust.
 Identities = 122/138 (88%), Positives = 127/138 (92%), Gaps = 0/138 (0%)
 Frame = +1

Query  165    GTDTAKLDPLLQRSNLSTYLIRSKLPNFEDTDDLEDDNFSLNDYRFSFSFSQDYEDIQVK  224
              G   +  LLQRSNLSTYLIRSKLPNFEDTDDLEDDN SLNDYRFSFSFSQDYEDIQVK
Sbjct  20635  GYRNGQARSLLQRSNLSTYLIRSKLPNFEDTDDLEDDNISLNDYRFSFSFSQDYEDIQVK  20814

Query  225    AQRDMILEVEHMPVNRVQTCKIRRPMHRIAKEDLNEEYRPHLESYYLFTTFRSWRMEWSF  284
              AQRDMILEVEHMPVNRVQ CKIRRP+HRIAK+DINEEY PHLESY+LFTTFRSWRMEWSF
Sbjct  20815  AQRDMILEVEHMPVNRVQACKIRRPLHRIAKKDINEEYTPHLESYFLFTTFRSWRMEWSF  20994

Query  285    NKMRDLLMKYSLGMMVFA  302
              +KMRDLLMKYSLGMM FA
Sbjct  20995  SKMRDLLMKYSLGMMAFA  21048
```

TBLASTN search of exon against contig shows a frame shift in the middle of the exon (problem with 454 sequencing)

To avoid these discrepancies, students should remember to…

- check the dot plot and peptide sequence alignment comparison with *D. melanogaster* (output from Gene Model Checker); be able to explain & defend any differences!
- look for discrepancies by going back to the Gene Record Finder and comparing exon lengths and locations;
- double check all splice sites; check whether any proposed non-canonical splice sites are also observed in the *D. melanogaster* model or nearby species;
- check all final annotation models with BLASTP alignments to the *D. melanogaster* orthologue (higher resolution);
- for 454 sequenced species, check DNA sequence using added Illumina reads or RNA-Seq data if needed.