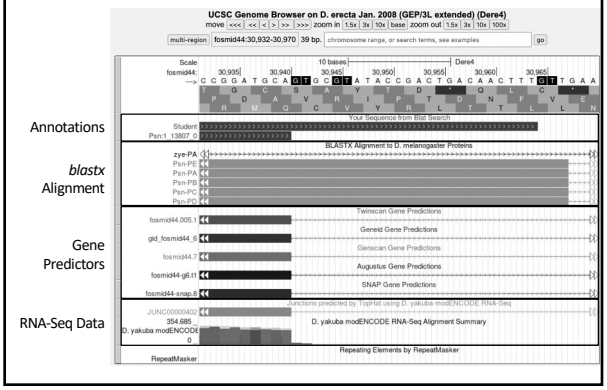
Common Errors in Student Annotation Submissions

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

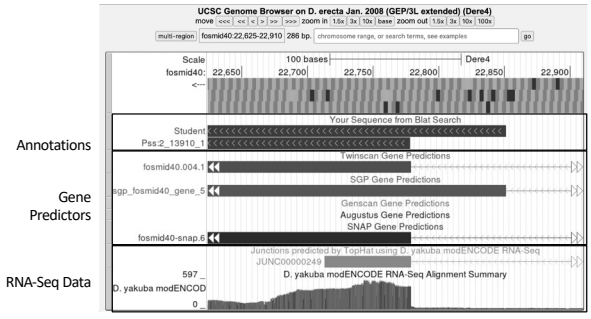
1

Over-reliance on *blastx* alignment



2

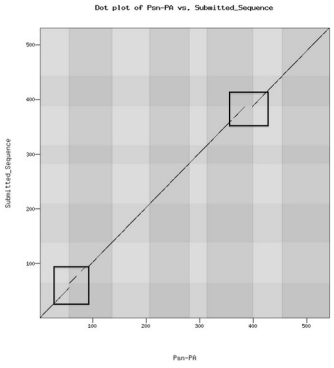
Relying on a single gene predictor



3

Strategies to resolve common errors

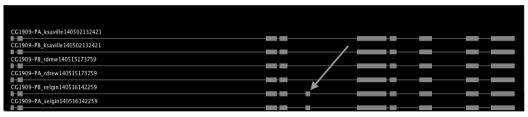
- Dot plot
- *tblastn* / *blastx* with exon-by-exon strategy
- RNA-Seq
- Identify small coding exons using "Small Exons Finder"
- Use dot plot and peptide sequence alignment to check



4

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.

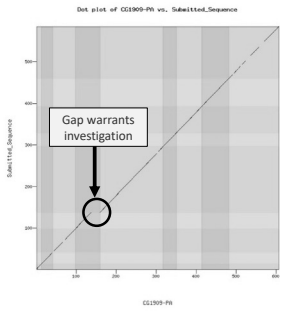
Isoform	1_9477_0	2_9477_1	3_9477_2	4_9477_0	5_9477_1	6_9477_0	7_9477_0	8_9477_0	9_9477_1
CG1909-PA	1	2	3	4	5	6	7	8	9
CG1909-PB	1	2	3	4	5	6	7	8	9

Liz suggested that there is an extra exon between 4_9477_0 and 5_9477_1

5

Continuing investigation of contig34

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

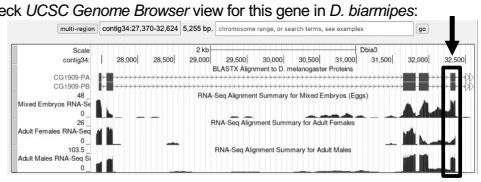


- y-axis = student annotation submission; x-axis = *D. melanogaster* gene model
- Gap (red) indicates residues in *D. melanogaster* gene that are not present in student annotation
- All in all, this dot plot warrants further investigation

6

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	Identities	Positives	Gaps	Frame
88.2 bits(217)	2e-25	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

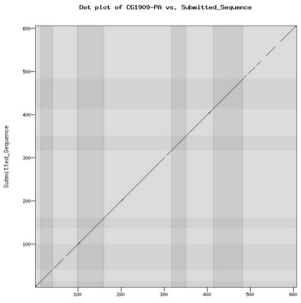
Score	Expect	Identities	Positives	Gaps	Frame
48.5 bits(114)	2e-11	20/21(95%)	20/21(95%)	0/21(0%)	+1

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

7

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

8

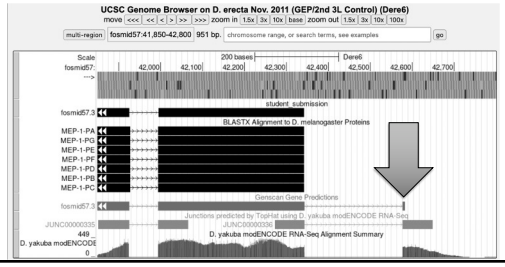
Strategies to identify small exons, particularly those with start and stop codons: Use RNA-Seq and TopHat to identify the 5' and 3' UTRs.

Select a row to display the corresponding CDS sequence:

FlyBase ID	5' Start	3' End	Strand	Phase	Size (aa)
1_6436_0	2,637,133	2,639,553	+	0	807
2_6436_0	2,639,615	2,639,872	+	0	86
3_6436_0	2,640,026	2,640,271	+	0	82
4_6436_0	2,640,337	2,640,516	+	0	60
5_6436_0	2,640,578	2,640,926	+	0	116
6_6436_2	2,641,012	2,641,016	+	2	1

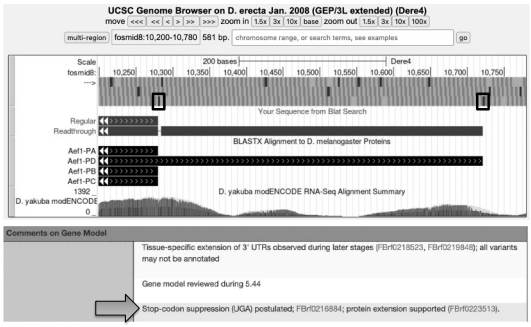
Sequence viewer for MEP-1: MEP-1:6_6436_2

```
>MEP-1:6_6436_2
*
```



9

Interesting annotation challenges: Read-through stop codons



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

10

Interesting annotation challenges Errors in the consensus sequence

Query:
CG32301:7_10861_1

Subject:
fosmid40 from the *D. erecta* Nov. 2011 (GEP/2nd 3L Control) (Dere6) assembly

Range 1: 20144 to 20662 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
321 bits(823)	5e-160	157/173(91%)	166/173(95%)	0/173(0%)	+2 ← Frame +2

```
Query  1      HLFMEPHPEVSILEADMVDFTGLTTTMEVSDLVAILHELFVSFDLAANHNRATRIKFLGD  60
              HLFMEPHPEVSILEADMV+FT LTTTM+VS+LVAILHELFVSFDLAAN NRATRIKFLGD
Sbjct  20144  HLFMEPHPEVSILEADMVNFTNLTTTMDVSELVAILHELFVSFDLAANQNRATRIKFLGD  20323

Query  61     SYTCVTGIPSYFPTHANACVNQALDMIEISREVSKRRNKKIDLRIGVHSGEILAGIIGLT  120
              SYTCVTGIPSYF  HANACVNQALDMIEISRE S+RRN+KI+LR+GVHSGEILAGIIGLT
Sbjct  20324  SYTCVTGIPSYFSAHANACVNQALDMIEISREASQRRNQKIELRVGVHSGEILAGIIGLT  20503

Query  121    KWQFDIWSKDVDITNRLESSGLPGMVHISSRTLGLLDNHYVYVEEGTDTAKLDP  173
              KWQFDIWSKDVDITNRLE+SGLPGMVHISSRTLGLLDN YVYEEGT+TAK DP
Sbjct  20504  KWQFDIWSKDVDITNRLEASGLPGMVHISSRTLGLLDNQYVYEEGTETAKQDP  20662
```

Range 2: 20662 to 21048 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Identities	Positives	Gaps	Frame
252 bits(644)	5e-160	121/129(94%)	125/129(96%)	0/129(0%)	+1 ← Frame +1

```
Query  174    LLQRSNLSTYLIRSRLPNFEDTDDLEDDNFSLNDYRFSFSFSQDYEDIQVKAQRDMILEV  233
              LLQRSNLSTYLIRSRLPNFEDTDDLEDDN SLNDYRFSFSFSQDYEDIQVKAQRDMILEV
Sbjct  20662  LLQRSNLSTYLIRSRLPNFEDTDDLEDDNISLNDYRFSFSFSQDYEDIQVKAQRDMILEV  20841

Query  234    EHMPVNRVQTCKIRRPMHRIAKEDINEEYRFHLESYYLFTTFRSWRMEWSFNKMRDLLMK  293
              EHMPVNRVQ CKIRRP+HRIAK+DINEEY FHLESY+LFTTFRSWRMEWSF+KMRDLLMK
Sbjct  20842  EHMPVNRVQACKIRRPLHRIAKKDINEEYTFHLESYFLFTTFRSWRMEWSFSKMRDLLMK  21021

Query  294    YSLGMMVFA  302
              YSLGMM FA
Sbjct  21022  YSLGMMAFA  21048
```

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

11

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* ortholog (higher resolution);
- for 454 sequenced species, check DNA sequence using added Illumina reads or RNA-Seq data if needed.

12