

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

Annotations

BLASTX Alignment

Gene Predictors

RNA-Seq Data

Relying on a single gene predictor

Annotation

Predictors

RNA-Seq

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	Y	Y
CG1909-RB	Y	Y	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: 31876 to 31999
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     IERGRRLYFQNGQVAVVWVNSALGTCQREDCFLGLYLY 42
            IERGRRLYFQNGQVAVVWVNSALGTCQREDCFLGLYLY+
Sbjct 31876 IERGRRLYFQNGQVAVVWVNSALGTCQREDCFLGLYLYE 31995

Range 2: 32299 to 32361
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    YQAMGKGRFREALIFRQQL 61
            YQAMGKGRFREALIFRQQL
Sbjct 32299 YQAMGKGRFREALIFRQQL 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_D	2,637,133	2,639,553	+	0	807
10_407_D	2,639,815	2,639,872	+	0	86
11_407_D	2,640,028	2,640,271	+	0	82
12_407_D	2,640,337	2,640,516	+	0	60
13_407_D	2,640,578	2,640,926	+	0	118
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: FBR0216884
- gene_with_stop_codon_read_through : SO:000697

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 (898), Expect(2) = 1e-165, Method: Compositional matrix adjust.
157/173 (91%), Positives = 166/173 (96%), Gaps = 0/173 (0%)
Frame = +2
Query 1     ILHSEHFFVLEADNVPGLTTFVEVDLALHLELVEFLAANNRATRIKFLD 60
            ILHSEHFFVLEADNVPFT LTTWVSIWALHLELVEFLAANNRATRIKFLD
Sbjct 20144 ILHSEHFFVLEADNVPGLTTFVEVDLALHLELVEFLAANNRATRIKFLD 20323

Query 61     STYCVGIFSYF HANACVQALWTEISRE SRRRKLIRIVSSEILAGIQLT 120
            STYCVGIFSYF HANACVQALWTEISRE SRRRKLIRIVSSEILAGIQLT
Sbjct 20324 STYCVGIFSYF HANACVQALWTEISRE SRRRKLIRIVSSEILAGIQLT 20503

Query 121    KNQFDNSKVDITNLSLSGLPQVHISSTLGLLQHWVYVETQVADLQ 173
            KNQFDNSKVDITNLSLSGLPQVHISSTLGLLQHWVYVETQVADLQ
Sbjct 20504 KNQFDNSKVDITNLSLSGLPQVHISSTLGLLQHWVYVETQVADLQ 20662

Score = 258 bits (658), Expect(2) = 1e-165, Method: Compositional matrix adjust.
122/138 (88%), Positives = 127/138 (92%), Gaps = 0/138 (0%)
Frame = +1
Query 165    QVTAFALDPLQNRSLSTLIRSLRPFVETDGLDQFSLNDFRFSFQGVYQVK 224
            QVTAFALDPLQNRSLSTLIRSLRPFVETDGLDQFSLNDFRFSFQGVYQVK
Sbjct 20635 QVTAFALDPLQNRSLSTLIRSLRPFVETDGLDQFSLNDFRFSFQGVYQVK 20814

Query 225    AQRMGLVEIDPQVWVQCKIRAPRIRLAKDINEIYFSELETLFTFRNRKNSF 284
            AQRMGLVEIDPQVWVQCKIRAPRIRLAKDINEIYFSELETLFTFRNRKNSF
Sbjct 20815 AQRMGLVEIDPQVWVQCKIRAPRIRLAKDINEIYFSELETLFTFRNRKNSF 20994

Query 285    NKRDLMLKYSLSQWFA 302
            NKRDLMLKYSLSQWFA
Sbjct 20995 NKRDLMLKYSLSQWFA 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.