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.

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. barmipes:

• 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

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

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

Interesting annotation challenges:

Read-through stop codons

Interesting annotation challenges:
Errors in the consensus sequence

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.