Sarah Jacobs
Bio 434W
April 18, 2014

Annotation of Contig9 from *Drosophila biarmipes* Dot Chromosome

ABSTRACT:

The goal of this project is to annotate all genes and other important features on Contig9 from the *Drosophila biarmipes* dot chromosome. The Genscan predictions for Contig9 were used as a starting point for the annotation, and a variety of strategies including BLAST alignments with known *D. melanogaster* genes, RNA-Seq data, TopHat data, and other gene predictors were used to refine the annotation. Contig9 from the *D. biarmipes* dot chromosome is expected to be somewhat syntenic with the orthologous region in *D. melanogaster*, and synteny is an important tool for the annotation process. Contig9 was determined to have three complete genes, which are orthologs of the *D. melanogaster* genes *CG32000, CG32006* and *CG33978*. While *CG32000* and *CG32006* are syntenic in *D. biarmipes* and *D. melanogaster*, *CG33978* appears to have undergone an inversion in *D. biarmipes* relative to *D. melanogaster*. This inversion appears to be present in *D. eugracilis* as well, but it remains unclear when this inversion occurred in evolutionary history. Furthermore, Contig9 has a putative paralog of *Arl4* that is not present in *D. melanogaster*. The mechanism by which this paralog arose is uncertain. The *Arl4* paralog does not seem to exist in the closely related *Drosophila* species available for comparison, and it is potentially the result of a recent duplication event. Overall, improving the annotation of *D. biarmipes*, as well as other *Drosophila* species, allows us to use a comparative genomics approach to study both evolutionary processes and gene structure and function.

INTRODUCTION:

Comparative genomics is an important tool for understanding sequence divergence and elucidating evolutionary processes across species. Previous studies have sequenced the genomes of multiple *Drosophila* species, and detailed annotation data exists for *D. melanogaster*. Improving the annotation of additional *Drosophila* species, such as *D. biarmipes*, allows us to use the comparative genomics approach to study evolutionary change across multiple *Drosophila* genomes. By comparing the sequence, synteny and repeat content of the genomes of various *Drosophila* species, we can better understand gene structure and function. For example, peptide coding sequences that are highly conserved across multiple species are likely important for protein function.

The goal of my project is to annotate Contig9 from the *Drosophila biarmipes* dot chromosome (Figure 1). I will attempt to annotate all genes and pseudogenes, as well as analyze the repeat content of the region. Furthermore, I will compare Contig9 to the orthologous region of *D. melanogaster* in order to determine if synteny has been preserved.

GENES:

Contig9 from the *Drosophila biarmipes* dot chromosome was viewed using the Genome Education Partnership (GEP) UCSC Genome Browser Mirror (Figure 1). Genscan predicted that Contig9 contains six features: three single exon features and three multi-exon features (Figure 2). Based on a preliminary assessment of the BLASTX alignment to *D. melanogaster* proteins, I hypothesized that the Genscan predictions were likely highly inaccurate, especially from bases 30 to 60 kb. Several of the single exon features predicted by Genscan appear to align to multi-exon genes predicted by the BLASTX alignment. Annotation of Contig9 may be complicated by the presence of genes with multiple isoforms.

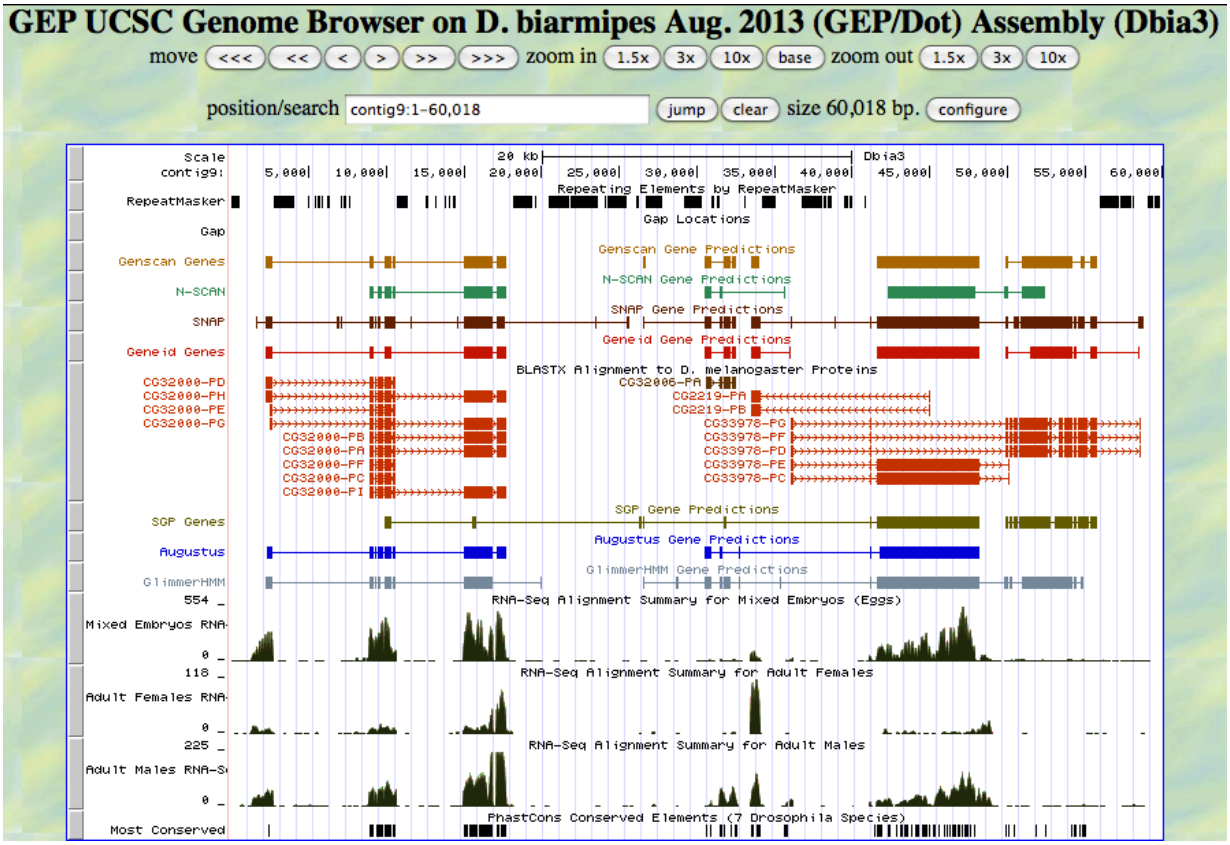


Figure 1: GEP UCSC Genome Browser view of Contig9 on *D. biarmipes* August 2013 (GEP/Dot) Assembly (Dbia3).

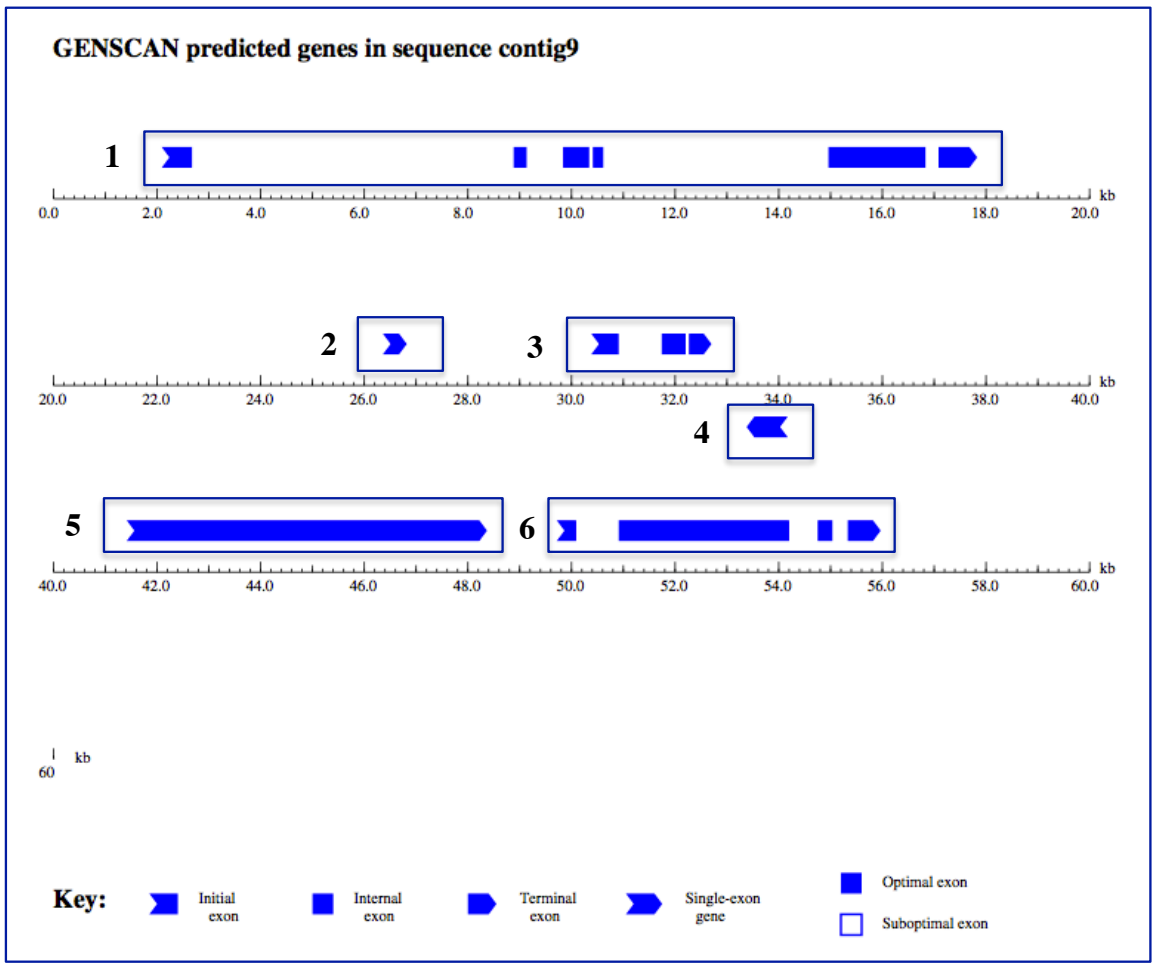


Figure 2: Genscan predicted six features for Contig9.

Gene *CG32000*:

I first investigated Genscan Feature 1, which appears to align to various isoforms of CG32000 in *D. melanogaster* based on the BLASTX alignment shown in Figure 1. The first step was to use the predicted protein sequence for Genscan Exon 5 of Feature 1 (the largest open reading frame) to perform a FlyBase BLASTp search against all annotated proteins for *D. melanogaster* (Genscan Exon 5 is boxed in red in Figure 3). Figure 4 shows the top alignments generated by the BLASTp search. The best match was to CG32000-PG, and the next best matches are to other isoforms of CG32000. Figure 5 shows the alignment of the predicted protein sequence for Genscan Exon 5 to CG32000-PG. The alignment is as expected for a comparison of *D. biarmipes* with *D. melanogaster*: there are regions with high similarity interspersed with a few regions of little or no similarity. Based on this initial comparison, Feature 1 is likely orthologous to CG32000 in *D. melanogaster*.

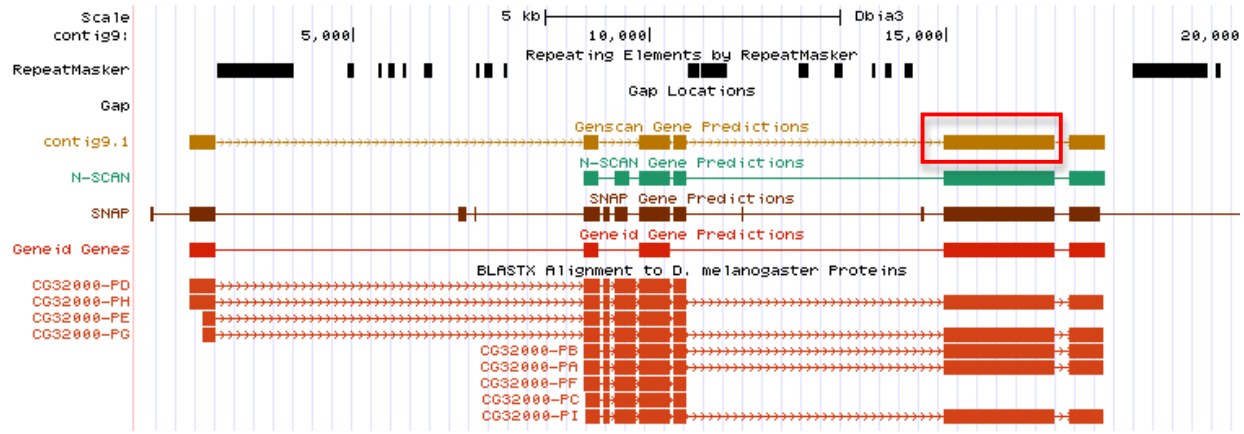


Figure 3: Genscan Feature 1 on GEP UCSC Genome Browser view of Contig9 of *D. biarmipes* (August 2013). Genscan Exon 5 of Feature 1 is boxed in red.

BLAST Hit Summary

Description	Species	Score	E value
CG32000-PG	Dmel	1875.14	0
CG32000-PI	Dmel	1873.98	0
CG32000-PH	Dmel	1873.21	0
CG32000-PB	Dmel	1871.67	0
CG32000-PA	Dmel	1871.67	0
CG32000-PE	Dmel	473.396	6.70822e-133
CG32000-PC	Dmel	473.011	7.23136e-133
CG32000-PD	Dmel	473.011	8.33339e-133
CG32000-PF	Dmel	472.241	1.26475e-132
CG6230-PA	Dmel	206.838	1.08638e-52
CG40625-PB	Dmel	77.0258	1.33117e-13
CG4301-PA	Dmel	70.0922	1.48454e-11

Figure 4: BLAST Hit Summary from BLASTp search aligning the predicted protein sequence for Genscan Exon 5 of Feature 1 against all annotated proteins for *D. melanogaster*.

>gnl|dmel|FBpp0089013 type=protein; loc=4:join(155344..155590, 157505..157793, 157849..157958, 158017..158378, 158433..158936, 159000..159200, 160896..162753, 162826..163421); ID=FBpp0089013; name=CG32000-PG; parent=FBgn0052000, FBtr0089451; dbxref=GB_protein:AAS64610.1, FlyBase:FBpp0089013, FlyBase_Annotation_IDs:CG32000-PG, REFSEQ:NP_995587, GB_protein:AAS64610, FlyMine:FBpp0089013, modMine:FBpp0089013; MD5=236d6f96e1a9759e999c37819e2863c1; length=1388; release=r5.56; species=Dmel; Length = 1388

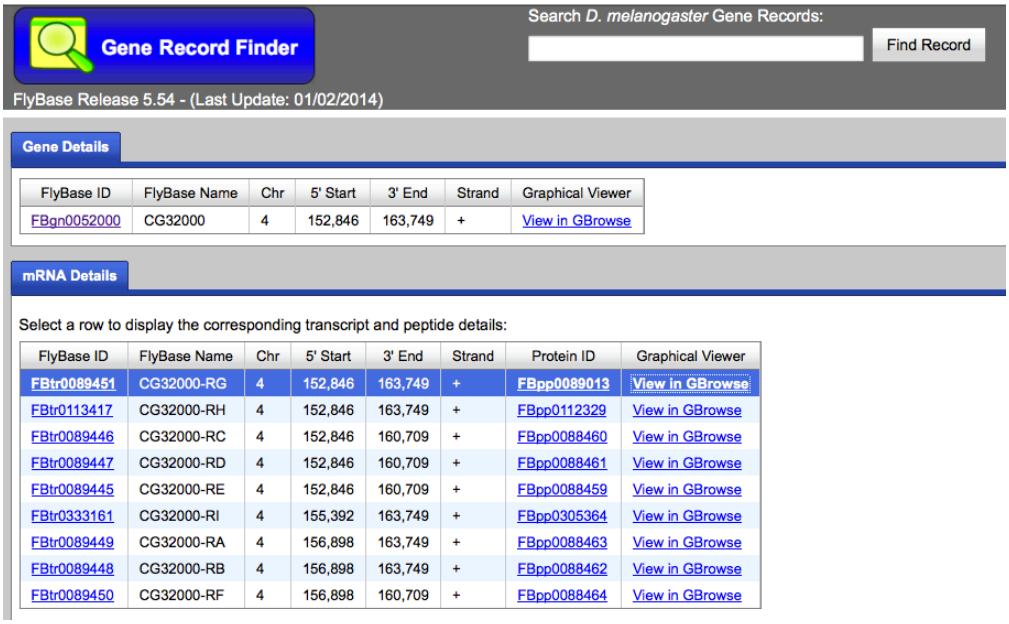
HSP # = 1 , **Score =** 1875.14 bits (4856) , **Expect =** 0
Identities = 905 / 1058 (85.5%) , **Positives =** 967 / 1058 (91.4%) , **Gaps =** 9 / 1058 (0.9%)

Subject FASTA

```
Query:   230         NQDVLQKTVYNTGNAWVVDPRGLSKEFPTRALVPGDIIEIPSSGCTMQCDAVLLSGNCIL   289
                    NQDVLQKTVYNTGNAWVVD +GLSKE PTRA+VPGDIIEIPSSGCT+ CDA+L+SGNCIL
Subject: 337         NQDVLQKTVYNTGNAWVVDHKGLSKELPTRAIVPGDIIEIPSSGCTLHCDAILISGNCIL   396

Query:   290         DESMLTGESVPVTKTPLPSKRDMIFDKTEHARHTLFCGTKVIQTRYIGSKKVLAFVINTG   349
                    DESMLTGESVPVTKTPLPSKRDMIFDKTEHARHTLFCGTKVIQTRYIGSKKVLAFVINTG
Subject: 397         DESMLTGESVPVTKTPLPSKRDMIFDKTEHARHTLFCGTKVIQTRYIGSKKVLAFVINTG   456

Query:   350         NITAKGELIRSILYPPPVDYKFEQDSYKFIQFLAVIACVGFIYTLVTKIMRGTDPVKIAV   409
                    NITAKGELIRSILYPPPVDYKFEQDSYKFIQFLA+IACVGFIYTLVTKI+RGTDPVKIAV
Subject: 457         NITAKGELIRSILYPPPVDYKFEQDSYKFIQFLAIIACVGFIYTLVTKILRGTDPVKIAV   516

Query:   410         ESLDLITIVVPPALPAAMTVGRFYAQKRLKASEIFCISPRSINVAGGINCCCFDKTGTLT   469
                    ESLDLITIVVPPALPAAMTVGRFYAQKRLK SEIFCISPRSINVAG INCCCFDKTGTLT
Subject: 517         ESLDLITIVVPPALPAAMTVGRFYAQKRLKTSEIFCISPRSINVAGSINCCCFDKTGTLT   576

Query:   470         EDGLDMWGVVPKSSTNQFQIPLKTVDRLPYDHFLFGMVTCHSITIMNGILMGDPLDLKMF   529
                    EDGLDMWGVVPKSSTNQFQIPLK+VDRLP+DHFLFGMVTCHSITI+NG +MGDPLDLKMF
Subject: 577         EDGLDMWGVVPKSSTNQFQIPLKSVDRLPFDHFLFGMVTCHSITILNGRMMGDPLDLKMF   636

Query:   530         NSTGWTIEDSNNIPDNEKYGILYPTILRQPRICSSDLRSSDSKKKSQISRQSSVDDLLAT   589
                     STGW +EDSNNIPD EKYGILYPTILRQPR   S +  ++S  K++I RQSSVDDLLAT
Subject: 637         ESTGWELEDSNNIPDTEKYGILYPTILRQPRGGLSGMAETESGSKNEIKRQSSVDDLLAT   696
```

Figure 5: Alignment of predicted protein sequence for Genscan Exon 5 (query) with CG32000-PG (subject), which was the best match identified by BLAST.

The next step was to use the GEP Gene Record Finder to further investigate the *CG32000* gene (Figure 6). The FlyBase accession number for *CG32000* is FBgn0052000. There are nine isoforms in *D. melanogaster* labeled A-I.



FlyBase ID	FlyBase Name	Chr	5' Start	3' End	Strand	Graphical Viewer
FBgn0052000	CG32000	4	152,846	163,749	+	View in GBrowse

FlyBase ID	FlyBase Name	Chr	5' Start	3' End	Strand	Protein ID	Graphical Viewer
FBtr0089451	CG32000-RG	4	152,846	163,749	+	FBpp0089013	View in GBrowse
FBtr0113417	CG32000-RH	4	152,846	163,749	+	FBpp0112329	View in GBrowse
FBtr0089446	CG32000-RC	4	152,846	160,709	+	FBpp0088460	View in GBrowse
FBtr0089447	CG32000-RD	4	152,846	160,709	+	FBpp0088461	View in GBrowse
FBtr0089445	CG32000-RE	4	152,846	160,709	+	FBpp0088459	View in GBrowse
FBtr0333161	CG32000-RI	4	155,392	163,749	+	FBpp0305364	View in GBrowse
FBtr0089449	CG32000-RA	4	156,898	163,749	+	FBpp0088463	View in GBrowse
FBtr0089448	CG32000-RB	4	156,898	163,749	+	FBpp0088462	View in GBrowse
FBtr0089450	CG32000-RF	4	156,898	160,709	+	FBpp0088464	View in GBrowse

Figure 6: Gene Record Finder entry for *D. melanogaster* gene *CG32000*. The FlyBase ID number for *CG32000* is FBgn0052000. *CG32000* appears to have 9 isoforms labeled A-I.

Figure 7 shows the FlyBase GBrowse2 view of the *CG32000* transcript in *D. melanogaster*. The various isoforms differ in terms of both translated and untranslated coding sequences.

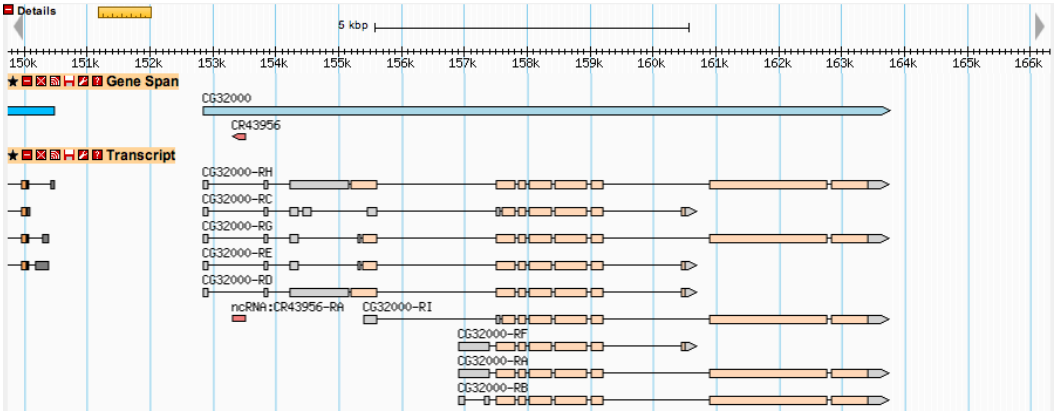


Figure 7: FlyBase GBrowse2 view showing coding regions of *CG32000* in *D. melanogaster*. Gray boxes represent untranslated exons and orange boxes represent translated exons.

The next step was to confirm that Contig9 of *D. biarmipes* contains the entire CG32000 protein sequence. Peptide sequences for each translated exon were obtained through the GEP Gene Record Finder. Table 1a shows the exons for all nine isoforms of CG32000 in *D.*

melanogaster and Table 1b summarizes the location of each exon in *D. melanogaster*. For each exon, a BLASTX search was performed comparing the *D. melanogaster* CG32000 peptide sequence to the unmasked Contig9 sequence. The GEP UCSC Genome Browser Mirror was then used to locate the exact boundaries of each exon.

Isoform	5_1767_0	7_1766_0	11_1759_0	13_1766_2	13_1767_0	14_1766_1	15_1766_2	16_1766_0	17_1766_0	18_1767_0	19_1766_0	20_1766_2
CG32000-RG		Y		Y		Y	Y	Y	Y		Y	Y
CG32000-RH	Y			Y		Y	Y	Y	Y		Y	Y
CG32000-RC					Y	Y	Y	Y	Y	Y		
CG32000-RD	Y			Y		Y	Y	Y	Y	Y		
CG32000-RE		Y		Y		Y	Y	Y	Y	Y		
CG32000-RI					Y	Y	Y	Y	Y		Y	Y
CG32000-RA			Y	Y		Y	Y	Y	Y		Y	Y
CG32000-RB			Y	Y		Y	Y	Y	Y		Y	Y
CG32000-RF			Y	Y		Y	Y	Y	Y	Y		

Table 1a: Summary of all nine isoforms of CG32000 and their respective exons in *D. melanogaster*.

FlyBase_ID	Dmel_chrom	Dmel_start	Dmel_end	Dmel_strand
CDS_FBgn0052000:5_1767_0	4	155155	155590	+
CDS_FBgn0052000:7_1766_0	4	155344	155590	+
CDS_FBgn0052000:11_1759_0	4	157379	157403	+
CDS_FBgn0052000:13_1766_2	4	157505	157793	+
CDS_FBgn0052000:13_1767_0	4	157552	157793	+
CDS_FBgn0052000:14_1766_1	4	157849	157958	+
CDS_FBgn0052000:15_1766_2	4	158017	158378	+
CDS_FBgn0052000:16_1766_0	4	158433	158936	+
CDS_FBgn0052000:17_1766_0	4	159000	159200	+
CDS_FBgn0052000:18_1767_0	4	160448	160522	+
CDS_FBgn0052000:19_1766_0	4	160896	162753	+
CDS_FBgn0052000:20_1766_2	4	162826	163421	+

Table 1b: FlyBase ID number and location for the twelve translated exons for CG32000 in *D. melanogaster*.

Exon 5_1767_0:

BLASTX was used to compare the unmasked Contig9 sequence (query) to the 5_1767_0 peptide sequence (subject). The low complexity filter was turned off in order to ensure the alignment would cover low complexity regions. Figure 8 shows the alignment. All 145 residues of the 5_1767_0 peptide sequence are covered by the alignment. There is a 54% identity and regions of high similarity are interspersed with regions of little similarity, which is expected.

CG32000:5_1767_0
Sequence ID: lcl|236061 Length: 145 Number of Matches: 5

Range 1: 1 to 145 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
156 bits(394)	2e-46	Compositional matrix adjust.	79/145(54%)	105/145(72%)	0/145(0%)	+3

```
Query  2238  MFTSQYKSTDSRNETLQLNSLPNLDKNSKFDVLEFETEEDLQFPQIAAKNKLLKLSWWNS  2417
             MF SQ K+ DS +ET+ L SLPN  ++ KF   + ET++DL FP IAAKNK+LKLSWW+S
Sbjct  1     MFASQSKACDSPSETVHLQSLPNHIEDLKFRNSDVETDDDLHFPGIAAKNKILKLSWWHS  60

Query  2418  PAEVHEIEANLKHHRKNDNGRNNKTEKLENNNSLVNGCVSTPARSVSLLQYIRPDQERVS  2597
             P ++H +        +K+D    +NK +K+ENNN+LVNGC  T ARSV LL+Y RPDQ+  S
Sbjct  61    PVDMHYVATESVQPKKSDKVPSNKIKKVENNNTLVNGCSKTSARSVPLLKYNRPDQDGDS  120

Query  2598  EAQETNILEPSVDELYPRDSHRLVD  2672
             E   T++LEP+VDE+Y +DS RLVD
Sbjct  121   EENITSVLEPNVDEIYSKDSERLVD  145
```

Figure 8: Alignment resulting from BLASTX search using the unmasked Contig9 sequence as the query and the 5_1767_0 peptide sequence as the subject.

The GEP UCSC Genome Browser Mirror was used to confirm the exon boundaries. Figure 9 shows that Exon 5_1767_0 starts at base 2238 with a methionine start codon in frame +3 as predicted by the BLASTX alignment. The start site is supported by multiple gene predictors including Genscan, Geneid, SNAP and GlimmerHMM, as well as the BLASTX alignment. The mRNA data extends past the 5' end, but that is expected due to the 5' untranslated region.

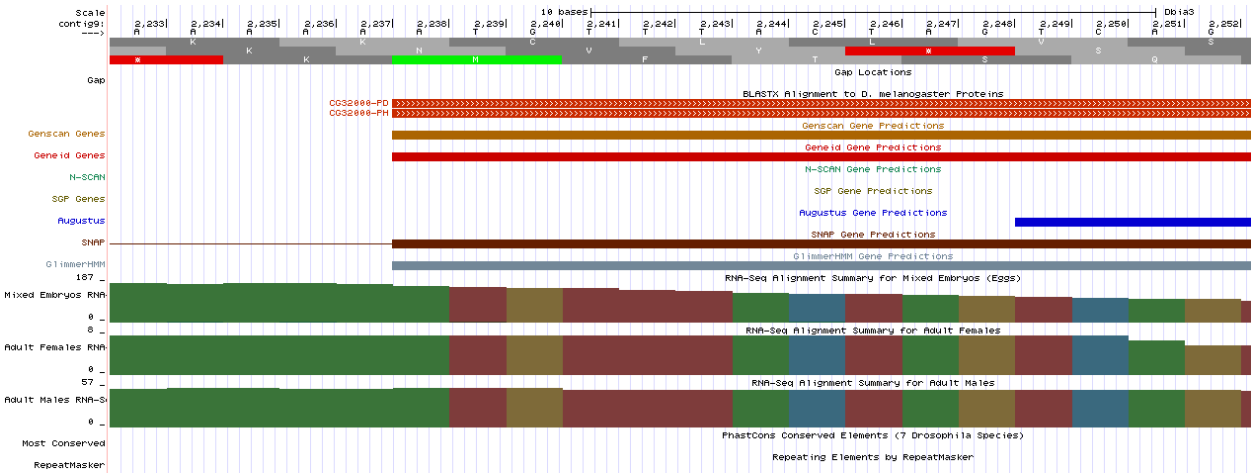


Figure 9: Beginning of Exon 5_1767_0 at base 2238.

Figure 10 shows that Exon 5_1767_0 ends at base 2673. While the BLASTX alignment extends only to base 2672, the 2673 3' splice site is supported by multiple gene predictors, mRNA data, TopHat data and the presence of a "GT" donor site at bases 2674-2675.

Figure 10: End of Exon 5_1767_0 at base 2673.

Exon 7_1766_0:

BLASTX was used to compare the unmasked Contig9 sequence (query) to the 7_1766_0 peptide sequence (subject). The low complexity filter was turned off. Figure 11 shows the alignment. All 82 residues of the 7_1766_0 peptide sequence are covered by the alignment. There are regions of high similarity interspersed with regions of little similarity, which is expected.

CG32000:7_1766_0

Sequence ID: lcl|17305 Length: 82 Number of Matches: 3

Range 1: 1 to 82 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
85.1 bits(209)	1e-22	Compositional matrix adjust.	43/82(52%)	58/82(70%)	0/82(0%)	+3

```
Query  2427  VHEIEANLKHHRKNDNGRNNKTEKLENNNSLVNGCVSTPARSVSLLQYIRPDQERVSEAQ  2606
             +H +        +K+D    +NK +K+ENNN+LVNGC   T ARSV LL+Y RPDQ+   SE
Sbjct  1     MHYVATESVQPKKSDKVPSNKIKKVENNNTLVNGCSKTSARSVPLLKYNRPDQDGDSEEN  60

Query  2607  ETNILEPSVDELYPRDSHRLVD  2672
              T++LEP+VDE+Y +DS RLVD
Sbjct  61    ITSVLEPNVDEIYSKDSERLVD  82
```

Figure 11: Alignment resulting from BLASTX search using the unmasked Contig9 sequence as the query and the 7_1766_0 peptide sequence as the subject.

Figure 12 shows the region of Contig9 where Exon 7_1766_0 is predicted to start based on the BLASTX alignment. Exon 7_1766_0 is an alternative initial exon that ends at the same location as Exon 5_1767_0. However, there is no methionine in the +3 reading frame after the methionine that initiates 5_1767_0. I hypothesize that this start codon does not exist in *D. biarmipes*.

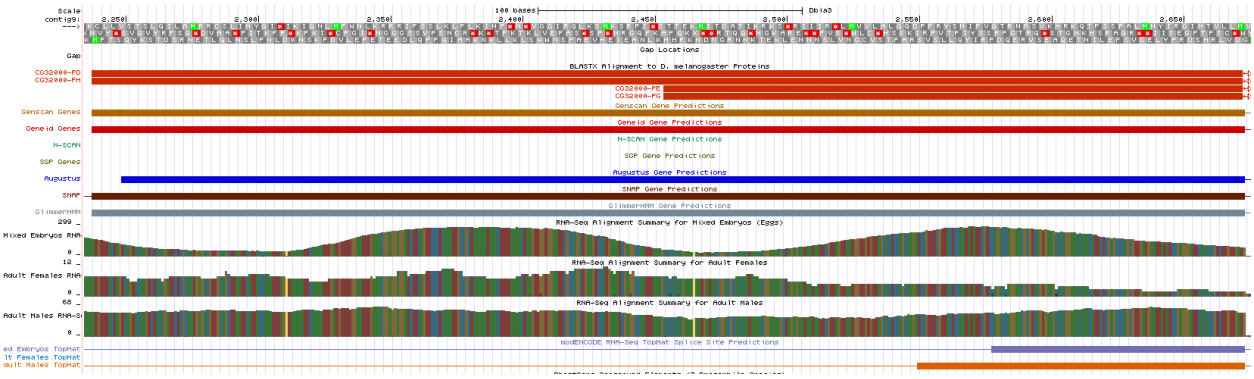


Figure 12: Region where Exon 7_1766_0 was predicted to start. Image shows the entire alignment to Exon 5_1767_0 and 7_1766_0: no methionine is present in frame +3 after the methionine at the beginning of Exon 5_1767_0.

In order to further investigate if Exon 7_1766_0 no longer exists as an initial exon in *D. biarmipes*, I performed a FlyBase BLASTp search aligning the *D. melanogaster* Exon 7_1766_0 peptide sequence to the annotated protein database of several *Drosophila* species (Figure 13). While the methionine is conserved in *D. erecta*, both *D. eugracilis* and *D. biarmipes* have a valine (V) at the equivalent location. I hypothesize that Exon 7_1766_0 acquired mutations at some point after *D. erecta* and *D. eugracillis* diverged such that the start codon no longer exists. This is a reasonable hypothesis because the methionine start codon is encoded by "AUG" and valine is encoded by "GUG," so a single point mutation could transform methionine to valine. I conclude that exon 7_1766_0 does not exist as a start codon in *D. biarmipes*.

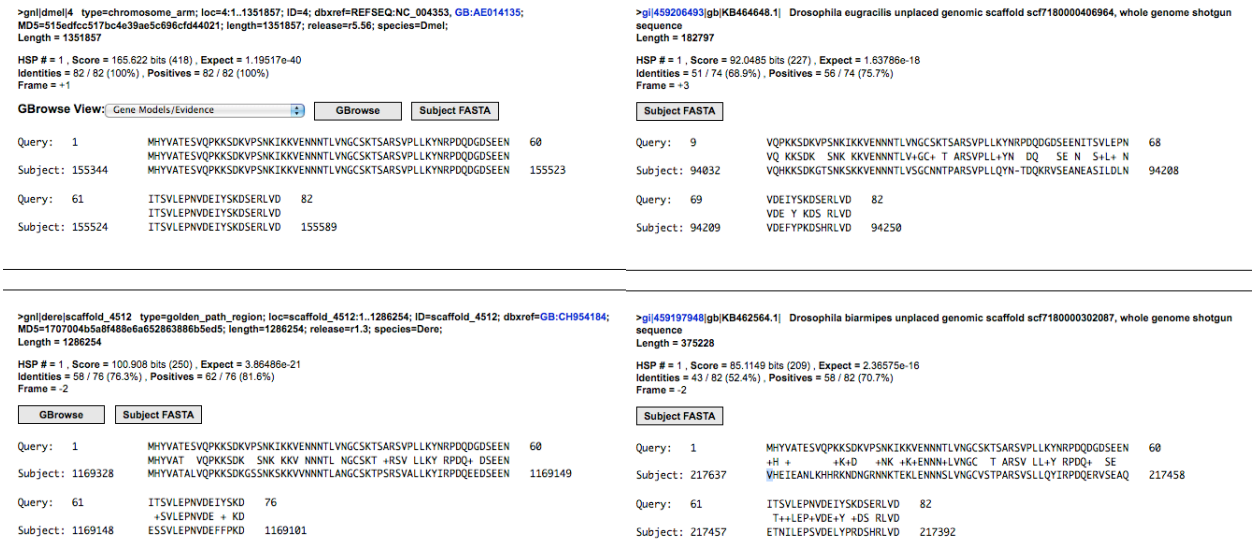


Figure 13: Alignment of *D. melanogaster* Exon 7_1766_0 peptide sequence to the annotated protein database for *D. melanogaster* (upper-left), *D. erecta* (lower-left), *D. eugracilis* (upper-right) and *D. biarmipes* (lower-right). Methionine start codon is conserved in *D. erecta*, but both *D. eugracilis* and *D. biarmipes* now have a valine (V) at the equivalent location.

Exon 11_1759_0:

Exon 11_1759_0 is an eight-residue long alternative initial exon in *D. melanogaster*. BLASTX did not produce any significant matches, so the Small Exons Finder was used to identifiy possible locations. The range of bases to search was selected based on the location of the adjacent exons. The Small Exons Finder identified potential exons at bases 8023 to 8047 and 8758 to 8782 (Figure 14).

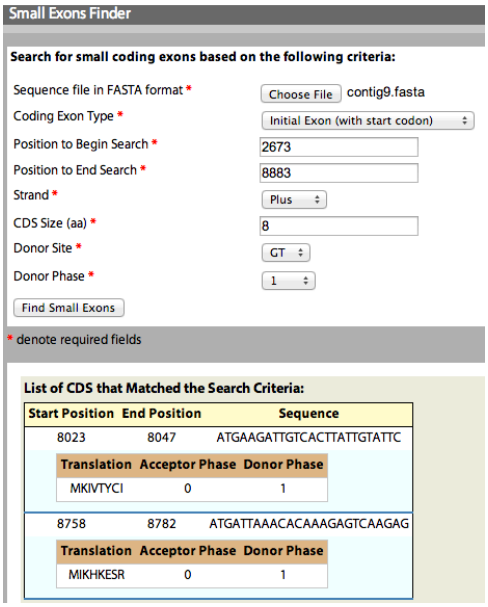


Figure 14: Results of Small Exon Finder. Potential exons located at bases 8023 to 8047 and 8758 to 8782.

I hypothesize that Exon 11_1759_0 is located at bases 8758 to 8782 in frame +1 (Figure 15). This location is supported by mRNA data and TopHat Junction data. The donor phase is 1 as expected (based on later exons) and there is a GT splice donor site in the correct location.

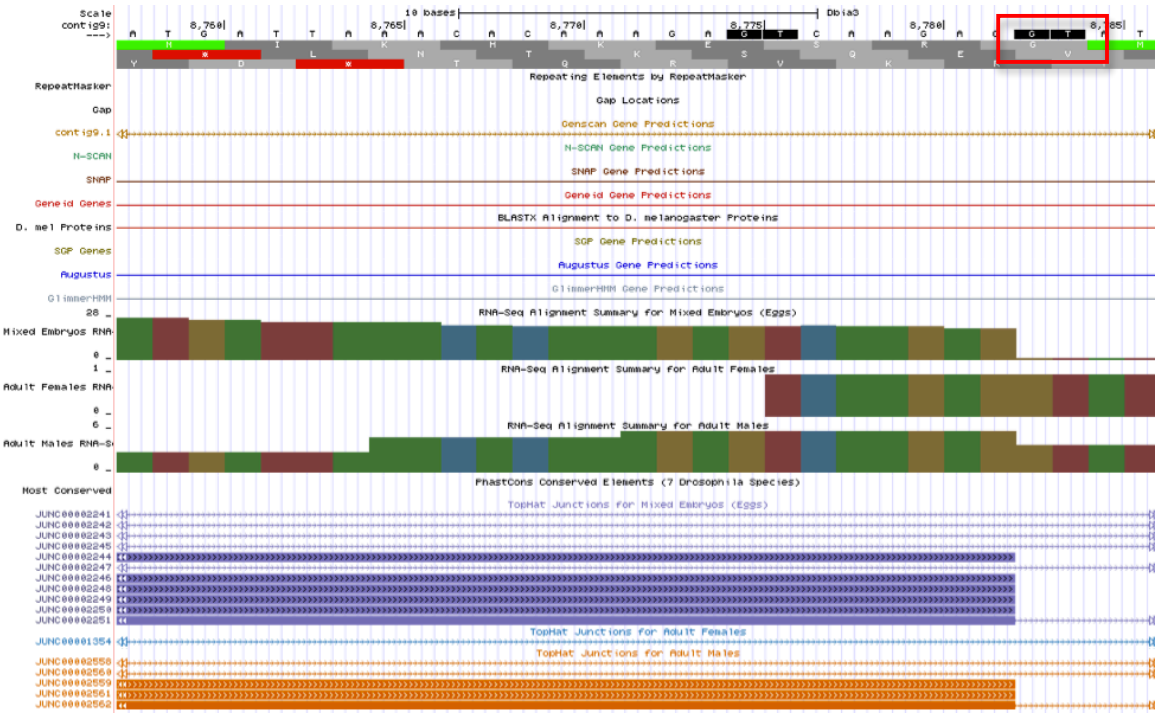


Figure 15: Location of Exon 11_1759_0 in reading frame +1. Preferred donor site is boxed in red.

Exon 13_1766_2:

BLASTX was used to compare the unmasked Contig9 sequence (query) to the 13_1766_2 peptide sequence (subject). The low complexity filter was turned off. Figure 16 shows the alignment. All 95 residues of the 13_1766_2 peptide sequence are covered by the alignment, and the alignment shows 82% identity.

CG32000:13_1766_2
Sequence ID: lcl|14239 Length: 95 Number of Matches: 5

Range 1: 1 to 95 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
167 bits(423)	5e-51	Compositional matrix adjust.	78/95(82%)	85/95(89%)	0/95(0%)	+2

```
Query  8885  SKPCTGLLNQDQEDQMNVSGYRRSHIRTAICWACIILTGGLLRLVLHWWRHLYLYATCSR  9064
             SKPC GLLN DQEDQM V GYRRS +RT  CWACI LTGGLLRLVLHWWRHLYLYATCS+
Sbjct  1     SKPCIGLLNPDQEDQMKVCGYRRSLMRTGFCWACIFLTGGLLRLVLHWWRHLYLYATCSQ  60

Query  9065  CPLEEAEQVLVTEDYQGKHKVYHVKQIRVLTANHF  9169
             C LEEAEQVLVTEDYQGKHK+YHVKQI+VLT+++
Sbjct  61    CSLEEAEQVLVTEDYQGKHKMYHVKQIQVLTSSNL  95
```

Figure 16: Alignment resulting from BLASTX search using the unmasked Contig9 sequence as the query and the 13_1766_2 peptide sequence as the subject.

Figure 17 shows the beginning and end of Exon 13_1766_2. Exon 13_1766_2 starts at base 8883 as supported by gene predictors, mRNA data and TopHat data. There is an "AG" splice acceptor site at bases 8881 to 8882. The BLASTX track (in Figure 17) suggests the exon should start earlier, but closer examination reveals that BLAST incorrectly over-extended. Exon

13_1766_2 ends at base 9171 as supported by gene predictors, mRNA data and TopHat data. There is a "GT" splice donor site at bases 9172 to 9173. The BLASTX alignment ends earlier, but that is expected because Figure 16 shows that last few residues of the *D. melanogaster* sequence are not well conserved in *D. biarmipes*.

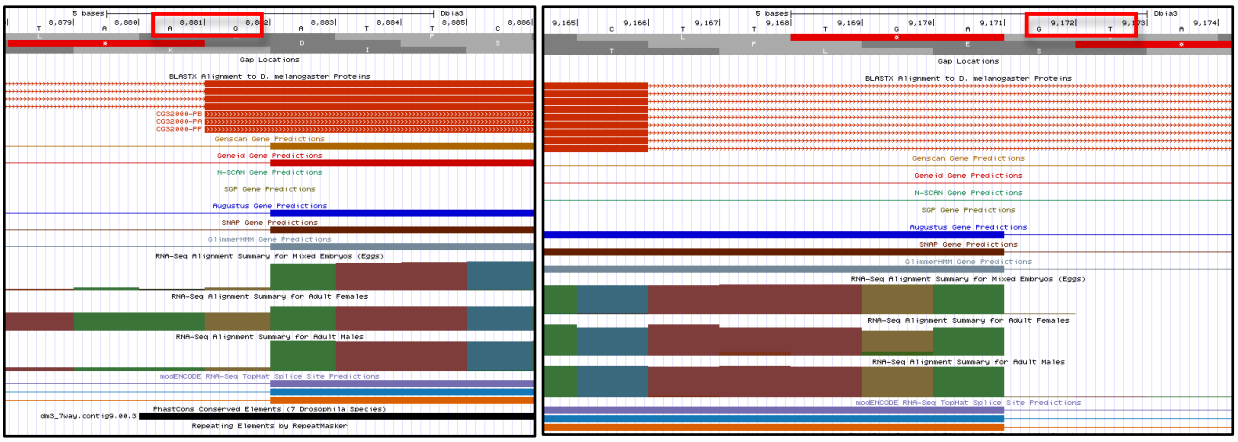


Figure 17: Beginning (left) and end (right) of Exon 13_1766_2 in reading frame +2. Donor/acceptor sites are boxed in red.

Exon 13_1767_0:

BLASTX was used to compare the unmasked Contig9 sequence (query) to the 13_1767_0 peptide sequence (subject). The low complexity filter was turned off. Figure 18 shows the alignment. All 80 residues of the 13_1767_0 peptide sequence are covered by the alignment, and the alignment shows 81% identity.

CG32000:13_1767_0

Sequence ID: lcl|261677 **Length:** 80 **Number of Matches:** 3

Range 1: 1 to 80 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
139 bits(351)	1e-41	Compositional matrix adjust.	65/80(81%)	72/80(90%)	0/80(0%)	+2

```
Query  8930  MNVSGYRRSHIRTAICWACIILTGGLLRLVLHWWRHLYLYATCSRCPLEEAEQVLVTEDY  9109
             M V GYRRS +RT  CWACI LTGGLLRLVLHWWRHLYLYATCS+C LEEAEQVLVTEDY
Sbjct  1     MKVCGYRRSLMRTGFCWACIFLTGGLLRLVLHWWRHLYLYATCSQCSLEEAEQVLVTEDY  60

Query  9110  QGKHKVYHVKQIRVLTANHF  9169
             QGKHK+YHVKQI+VLT+++
Sbjct  61    QGKHKMYHVKQIQVLTSSNL  80
```

Figure 18: Alignment resulting from BLASTX search using the unmasked Contig9 sequence as the query and the 13_1767_0 peptide sequence as the subject.

Exon 13_1767_0 begins at base 8930 (Figure 19). The start site is supported by the BLASTX alignment and the presence of a methionine start codon in the expected frame (+2).

Exon 13_1767_0 and Exon 13_1766_2 both end at base 9171, so see Figure 17 for further details.

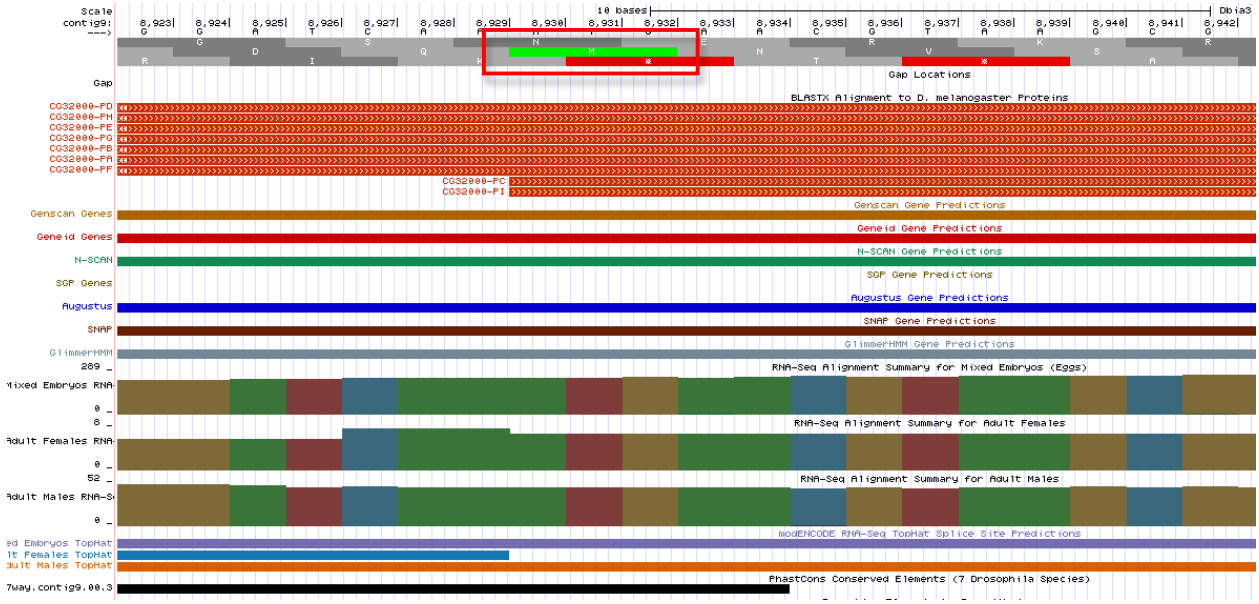


Figure 19: Start of Exon 13_1767_0 in frame +2.

Exon 14_1766_1:

BLASTX was used to compare the unmasked Contig9 sequence (query) to the 14_1766_1 peptide sequence (subject). The low complexity filter was turned off. Figure 20 shows the alignment. All 36 residues of the 14_1766_1 peptide sequence are covered by the alignment. The ends of the alignment show high similarity while the middle shows a weaker similarity. The alignment shows 53% identity.

CG32000:14_1766_1

Sequence ID: lcl|207837 **Length:** 36 **Number of Matches:** 4

Range 1: 1 to 36 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
36.2 bits(82)	8e-07	Compositional matrix adjust.	19/36(53%)	25/36(69%)	1/36(2%)	+1

```
Query  9238  ALLKKERPDAELMNVGSD-AEHAVQLSVHFTSAQFK  9342
             LL+KE+    E  ++  D  E+ +QLSVHFTSAQFK
Sbjct  1     TLLEKEQQSIERTHIECDHVENVLQLSVHFTSAQFK  36
```

Figure 20: Alignment resulting from BLASTX search using the unmasked Contig9 sequence as the query and the 14_1766_1 peptide sequence as the subject.

Exon 14_1766_1 starts at base 9237 and ends at base 9343 (Figure 21). The start site is supported by gene predictors, mRNA data, TopHat data and an "AG" splice acceptor site at bases 9235 to 9236. The BLASTX alignment is over-extended and does not predict the correct

start site. The end site is supported by gene predictors, mRNA data, TopHat data and a "GT" splice donor site at bases 9344 to 9345.

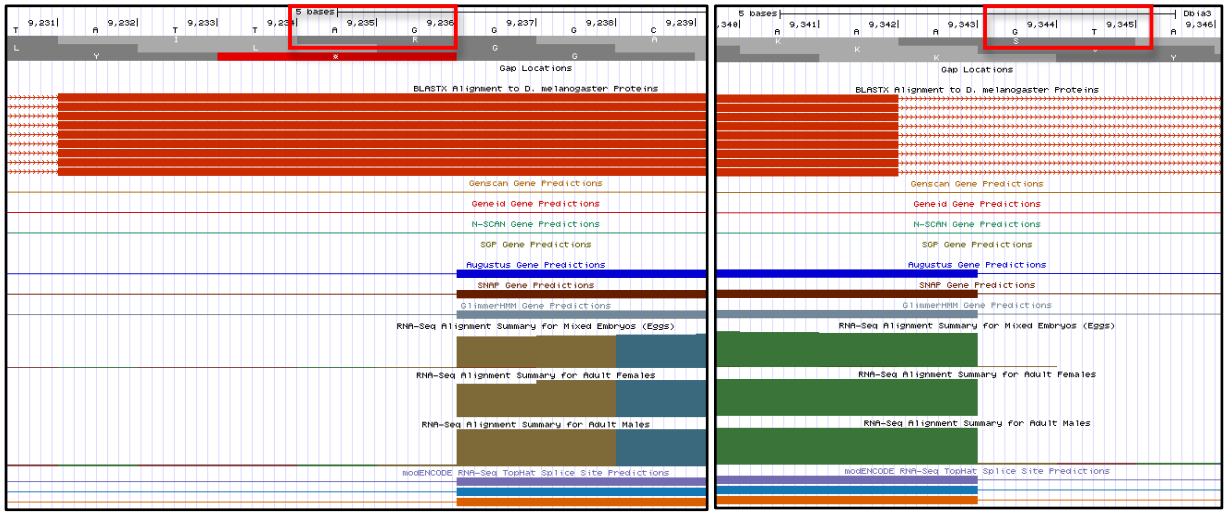


Figure 21: Start (left) and end (right) of Exon 14_1766_1 in reading frame +1. Splice acceptor/donor sites are boxed in red.

Exon 15_1766_2:

BLASTX was used to compare the unmasked Contig9 sequence (query) to the 15_1766_2 peptide sequence (subject). The low complexity filter was turned off. Figure 22 shows the alignment. All 120 residues of the 15_1766_2 peptide sequence are covered by the alignment, and the alignment shows 91% identity.

CG32000:15_1766_2
Sequence ID: lcl|44861 Length: 120 Number of Matches: 11

Range 1: 1 to 120 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
228 bits(582)	8e-72	Compositional matrix adjust.	109/120(91%)	114/120(95%)	0/120(0%)	+3

```
Query  9420  CSAIRIFRCKQLVYAWNNNLNCFQRINGLDLNIPCSYYHQQRGLTAREQISRRIVFGDNE  9599
             CS+IRIFRCKQLVYAWNNN N FQRINGLDLNIPCSYYHQQRGL   EQISRRIVFGDNE
Sbjct  1     CSSIRIFRCKQLVYAWNNNTNRFQRINGLDLNIPCSYYHQQRGLPVHEQISRRIVFGDNE  60

Query  9600  ITVPLRDIKTLLFLEVLNPFYVFQIFSVILWFTYDYYYYACVILLMSIFGITMSILQTKK  9779
             ITVPLRD KTLLFLEVLNPFYVFQ+FSVILWFTYDYYYYACVILLMS+FGIT+S+LQTKK
Sbjct  61    ITVPLRDFKTLLFLEVLNPFYVFQLFSVILWFTYDYYYYACVILLMSVFGITVSVLQTKK  120
```

Figure 22: Alignment resulting from BLASTX search using the unmasked Contig9 sequence as the query and the 15_1766_2 peptide sequence as the subject.

Exon 15_1766_2 starts at base 9418 (Figure 23). The start site is supported by multiple gene predictors, mRNA data, TopHat data and the presence of an "AG" splice acceptor site at bases 9416 to 9417. Exon 15_1766_2 ends at base 9779 (Figure 24). The end site is supported by multiple gene predictors, the BLASTX alignment, mRNA data and TopHat data. I hypothesize

that there is a non-canonical “GC” splice donor site at bases 9780 to 9781. This model is further supported by the presence of a non-canonical “GC” splice sites at the equivalent location in other *Drosophila* species, such as *D. yakuba* and *D. erecta* (Figure 25).

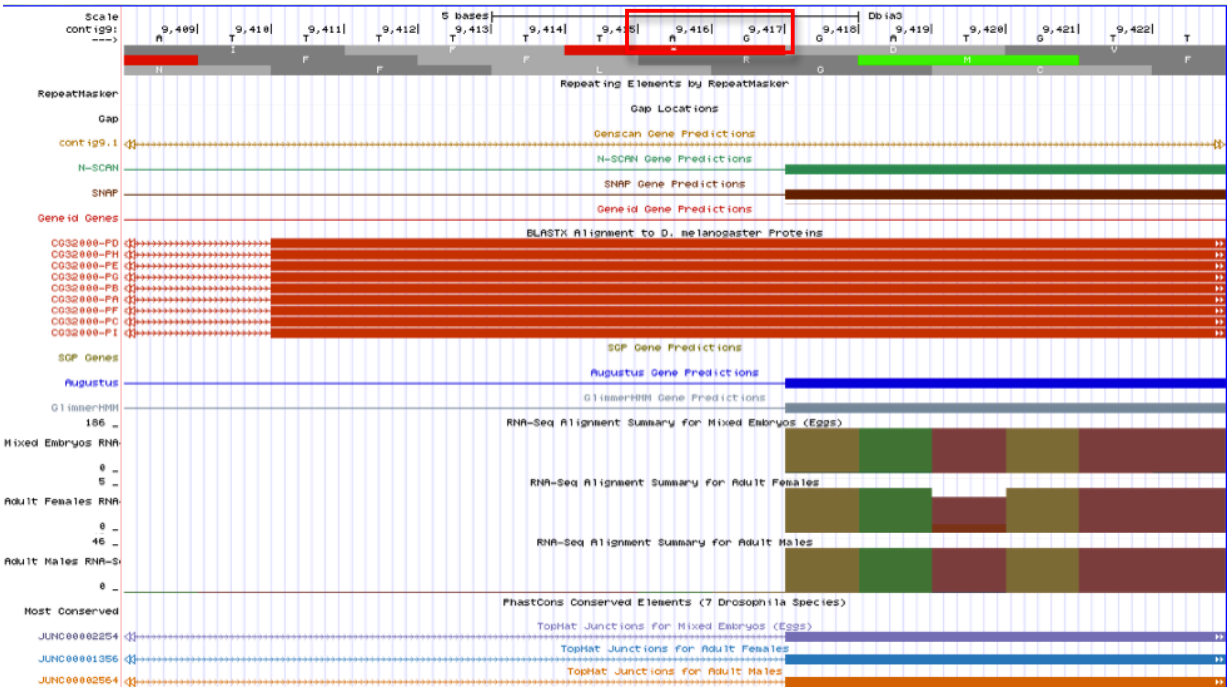


Figure 23: Start of Exon 15_1766_2 in reading frame +3. The splice acceptor site is boxed in red.

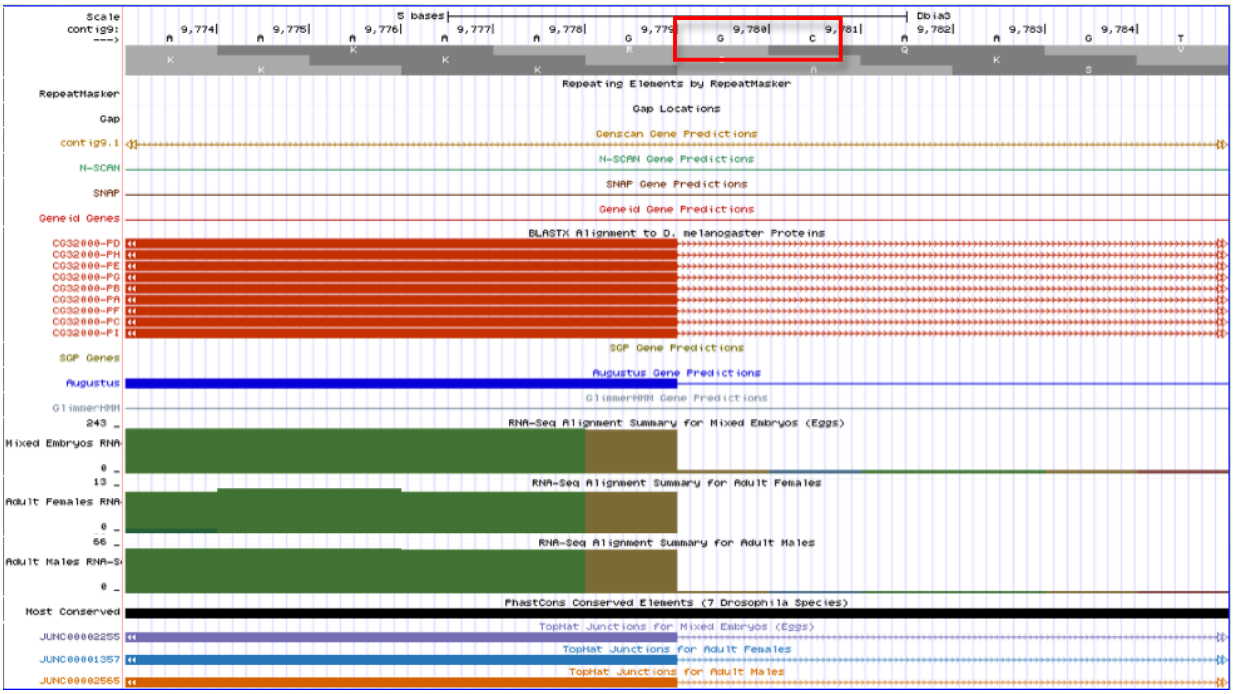


Figure 24: End of Exon 15_1766_2 in reading frame +3. The non-canonical “GC” splice donor site is boxed in red.

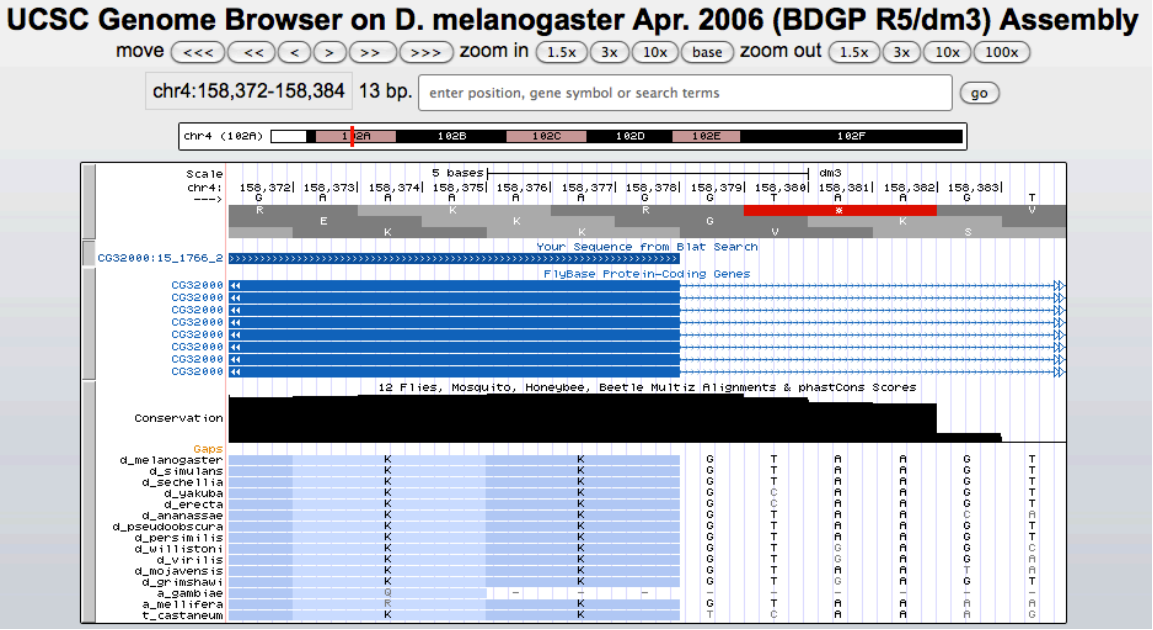


Figure 25: Equivalent splice site in other *Drosophila* species. *D. yakuba* and *D. erecta* also have a non-canonical "GC" splice donor site.

Exon 16_1766_0:

BLASTX was used to compare the unmasked Contig9 sequence (query) to the 16_1766_0 peptide sequence (subject). The low complexity filter was turned off. Figure 26 shows the alignment. All 168 residues of the 16_1766_0 peptide sequence are covered by the alignment, and the alignment shows 95% identity.

CG32000:16_1766_0

Sequence ID: lcl|48049 Length: 168 Number of Matches: 4

Range 1: 1 to 168 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
335 bits(860)	3e-108	Compositional matrix adjust.	159/168(95%)	165/168(98%)	0/168(0%)	+3

```
Query  9837   NQDVLQKTVYNTGNAWVVDPRGLSKEFPTRALVPGDIIEIPSSGCTMQCDAVLLSGNCIL  10016
              NQDVLQKTVYNTGNAWVVD +GLSKE PTRA+VPGDIIEIPSSGCT+ CDA+L+SGNCIL
Sbjct  1      NQDVLQKTVYNTGNAWVVDHKGLSKELPTRAIVPGDIIEIPSSGCTLHCDAILISGNCIL  60

Query  10017  DESMLTGESVPVTKTPLPSKRDMIFDKTEHARHTLFCGTKVIQTRYIGSKKVLAFVINTG  10196
              DESMLTGESVPVTKTPLPSKRDMIFDKTEHARHTLFCGTKVIQTRYIGSKKVLAFVINTG
Sbjct  61     DESMLTGESVPVTKTPLPSKRDMIFDKTEHARHTLFCGTKVIQTRYIGSKKVLAFVINTG  120

Query  10197  NITAKGELIRSILYPPPVDYKFEQDSYKFIQFLAVIACVGFIYTLVTK  10340
              NITAKGELIRSILYPPPVDYKFEQDSYKFIQFLA+IACVGFIYTLVTK
Sbjct  121    NITAKGELIRSILYPPPVDYKFEQDSYKFIQFLAIIACVGFIYTLVTK  168
```

Figure 26: Alignment resulting from BLASTX search using the unmasked Contig9 sequence as the query and the 16_1766_0 peptide sequence as the subject.

Exon 16_1766_0 starts at base 9837 (Figure 27). The start site is supported by multiple gene predictors, BLASTX, mRNA data, TopHat data, and an "AG" splice acceptor site at bases

9835 to 9836. Exon 16_1766_0 ends at base 10340 (Figure 28). The end site is supported by multiple gene predictors, mRNA data, TopHat data and a “GT” splice donor site at bases 10341-10342. BLASTX incorrectly over-extends in the genome browser view shown in Figure 28—Figure 26 correctly shows that the alignment ends at base 10340.

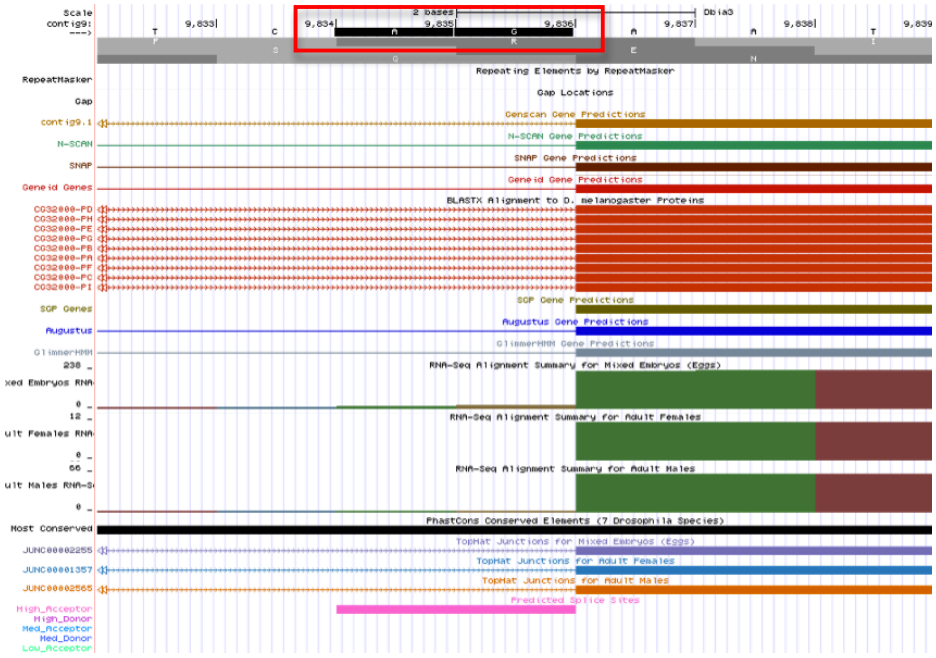


Figure 27: Start of Exon 16_1766_0 in reading frame +3. The splice acceptor site is boxed in red.

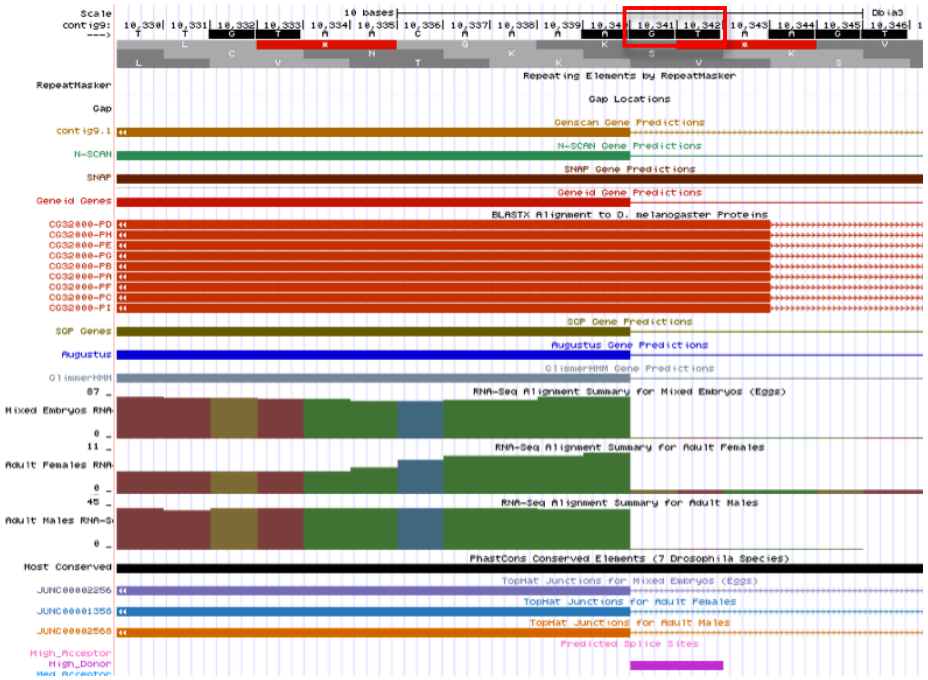


Figure 28: End of Exon 16_1766_0 in reading frame +3. The splice donor site is boxed in red.

Exon 17_1766_0:

BLASTX was used to compare the unmasked Contig9 sequence (query) to the 17_1766_0 peptide sequence (subject). The low complexity filter was turned off. Figure 29 shows the alignment. All 67 residues of the 17_1766_0 peptide sequence are covered by the alignment, and the alignment shows 96% identity.

CG32000:17_1766_0
Sequence ID: lcl|27951 Length: 67 Number of Matches: 8

Range 1: 1 to 67 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
132 bits(332)	2e-39	Compositional matrix adjust.	64/67(96%)	65/67(97%)	0/67(0%)	+2

```
Query  10412  IMRGTDPVKIAVESLDLITIVVPPALPAAMTVGRFYAQKRLKASEIFCISPRSINVAGGI  10591
              I+RGTDPVKIAVESLDLITIVVPPALPAAMTVGRFYAQKRLK SEIFCISPRSINVAG I
Sbjct  1      ILRGTDPVKIAVESLDLITIVVPPALPAAMTVGRFYAQKRLKTSEIFCISPRSINVAGSI  60

Query  10592  NCCCFDK  10612
              NCCCFDK
Sbjct  61     NCCCFDK  67
```

Figure 29: Alignment resulting from BLASTX search using the unmasked Contig9 sequence as the query and the 17_1766_0 peptide sequence as the subject.

Exon 17_1766_0 starts at base 10412 as supported by multiple gene predictors, mRNA data, TopHat data and an "AG" splice acceptor site at 10410-10411 (Figure 30). The BLASTX alignment also starts at base 10412 as shown in Figure 29, and is incorrectly over-extended in Figure 30. Exon 17_1766_0 ends at base 10612 as supported by multiple gene predictors, mRNA data, TopHat data, BLASTX and a "GT" splice donor site at 10613-10614 (Figure 31).

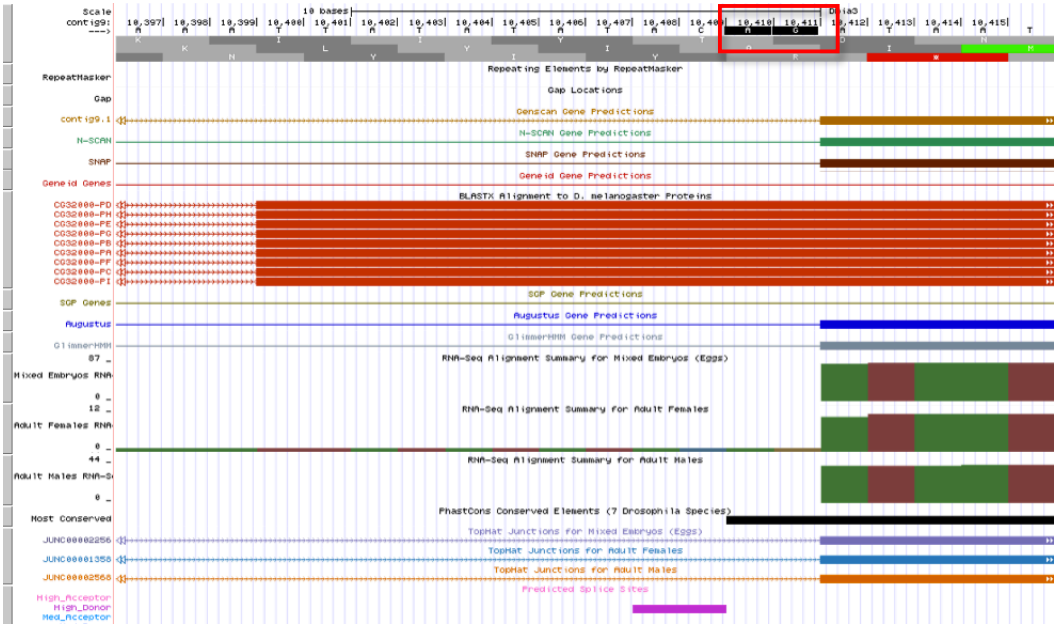


Figure 30: Start of Exon 17_1766_0 in reading frame +2. The spliced acceptor site is boxed in red.

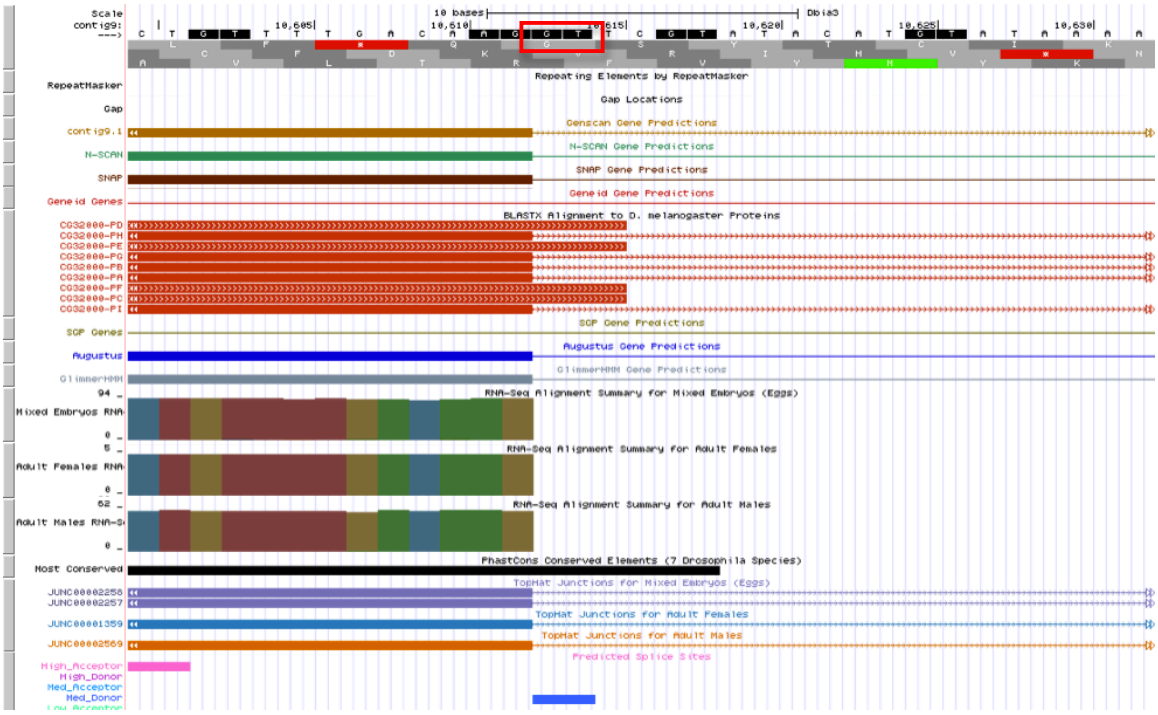


Figure 31: End of Exon 17_1766_0 in reading frame +2. The splice donor site is boxed in red.

Exon 18_1767_0:

Exon 18_1767_0 is a 25 amino acid long terminal exon in *D. melanogaster*. BLASTX did not produce any significant matches, so the Small Exons Finder was used to identifiy possible locations. The Small Exons Finder identified a potential exon at bases 14526 to 14600 (Figure 32).

Search for small coding exons based on the following criteria:

Sequence file in FASTA format * Browse... contig9.fasta

Coding Exon Type * Terminal Exon (with stop codon)

Position to Begin Search * 10600

Position to End Search * 15000

Strand * Plus

CDS Size (aa) * 25

Acceptor Phase * 0

Find Small Exons

* denote required fields

List of CDS that Matched the Search Criteria:

Start Position	End Position	Sequence
14526	14600	CCTTTTGTAGTGTCCGATGATCCAGTTCTTCAGTACATTTCTGACATGGTCCTTAATGGCGACTTTAATCTTTAG

Translation	Acceptor Phase	Donor Phase
PFVVSDDPVLQYISDMVLNGDFNL*	0	0

Figure 32: Results of Small Exons Finder.

Exon 18_1767_0 starts at base 14526 and ends at base 14600 (Figure 33). The exon location is supported by an "AG" splice acceptor site at bases 14524-14525, mRNA coverage data in the region, and a stop codon at bases 14598-14600. This is likely to be the correct location for Exon 18_1767_0.

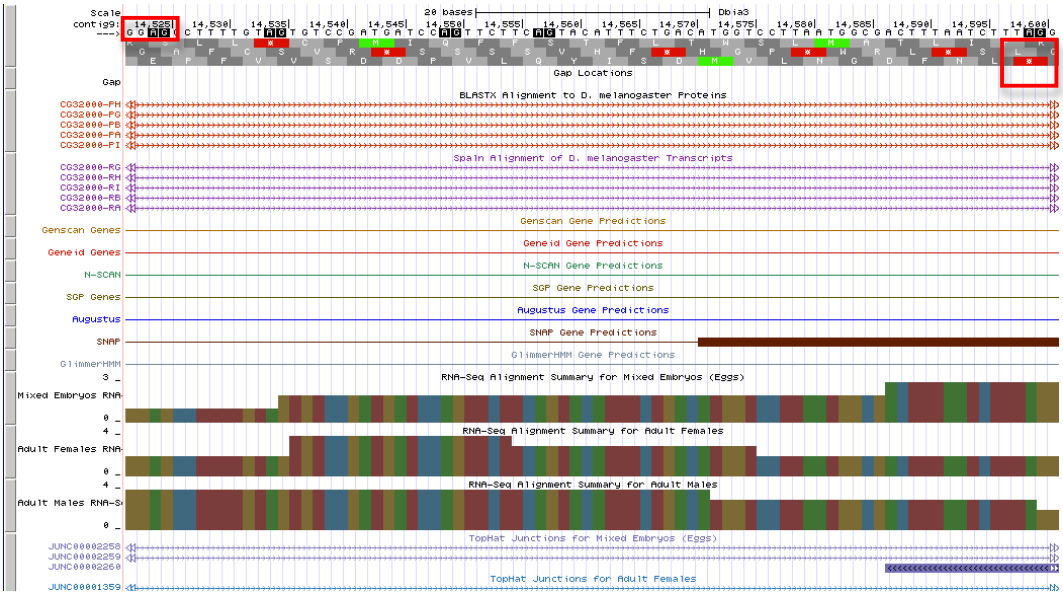


Figure 33: Location of Exon 18_1767_0 in reading frame +3. The splice acceptor site and stop codon are boxed in red.

Exon 19_1766_0:

BLASTX was used to compare the unmasked Contig9 sequence (query) to the 19_1766_0 peptide sequence (subject). The low complexity filter was turned off. Figure 34 shows the alignment. 618 out of 619 residues of the 19_1766_0 peptide sequence are covered by the alignment, and the alignment has regions of high similarity interspersed with regions of lower similarity with an overall 82% identity.

CG32000:19_1766_0
Sequence ID: lcl|28563 Length: 619 Number of Matches: 6

Range 1: 1 to 618 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
1040 bits(2688)	0.0	Compositional matrix adjust.	512/625(82%)	552/625(88%)	9/625(1%)	+2

```
Query  14960  TGTLTEDGLDMWGVVPKSSTNQFQIPLKTVDRLPYDHFLPGMVTCHSITIMNGILMGDPL  15139
              TGTLTEDGLDMWGVVPKSSTNQFQIPLK+VDRLP+DHFLPGMVTCHSITI+NG +MGDPL
Sbjct  1      TGTLTEDGLDMWGVVPKSSTNQFQIPLKSVDRLPFDHFLPGMVTCHSITILNGRMMGDPL  60

Query  15140  DLKMFNSTGWTIEDSNNIPDNEKYGILYPTILRQPRICSSDLRSSDSKKKSQISRQSSVD  15319
              DLKMF STGW +EDSNNIPD EKYGILYPTILRQPR   S +  ++S  K++I RQSSVD
Sbjct  61     DLKMFESTGWELEDSNNIPDTEKYGILYPTILRQPRGGLSGMAETESGSKNEIKRQSSVD  120

Query  15320  DLLATVGISQNQKNFDHGIVREFPFTSSLQRMSVITRCLSDQVFNVYCKGSPEMLQTLCT  15499
              DLLATVGIS +QKNFDHGIVREFPFTS+LQRMSV+TRCLSDQVFNVYCKGSPEML+ LC
Sbjct  121    DLLATVGISPSQKNFDHGIVREFPFTSALQRMSVVTRCLSDQVFNVYCKGSPEMLKKLCK  180

Query  15500  PQSLPDNYSEQLSTFAKRGYRIIAIAFKPLVQKMNYSKIQRLSREEVESNLEFLGFVVLE  15679
              PQSLPDNYS+QLS FAK+GYRIIAIAFK L  KMNY+K+QRLSREEVE+N+EFLGFV+LE
Sbjct  181    PQSLPDNYSQQLSEFAKKGYRIIAIAFKALSHKMNYTKVQRLSREEVENNMEFLGFVILE  240

Query  15680  NRLKPDTAKVIKALNSAKIRTIMITGDNILTAMSVARDCGIVSASQAVITVHADPLGNSG  15859
              NRLKPDT KVI ALN+AKIRTIMITGDNILTA+SVARDCGIVS SQ+VITVHADP+G+S
Sbjct  241    NRLKPDTTKVINALNAAKIRTIMITGDNILTAISVARDCGIVSPSQSVITVHADPIGDSA  300

Query  15860  NIQSN-NTDSIVDKKTARLYKLQYTLDLGSKISSPSLLTKCNSNHIPKENREDKFSTVG  16036
              NIQ+N  T+   D  + + YKL YTLDLGSK S   L  +  NSN    E  E  F+
Sbjct  301    NIQTNTGTECNFDNSSDKHYKLHYTLDLGSKTSRAYLFKSCFNSNLFDPETPE--FTAQV  358

Query  16037  DQTLFRLESTNSLVNESSSSYAESILPTSDSLASVKTTDTWTHNEPTDVDAELGINRTQD  16216
               +T+F +ESTNSLVNES+SSYAES LPTSDSLASVKT DTWTHN      DAELGI  T D
Sbjct  359    GKTIFHMESTNSLVNESTSSYAESGLPTSDSLASVKTIDTWTHN-----DAELGIKHTPD  413

Query  16217  ESW-RQHYIFAMDGGTWQIIKDHFPDEMAILLTRGSIYARMSPDQKQALVIELQNLDYCV  16393
              ESW RQ  IFAMDG TWQI+KD+FP+EM ILLTRGSIYARMSPDQKQALVIELQNLDYCV
Sbjct  414    ESWRRQECIFAMDGKTWQIVKDYFPEEMEILLTRGSIYARMSPDQKQALVIELQNLDYCV  473

Query  16394  AMCGDGANDCGALKVAHAGISLSETEASIASPFTSRNPTISAVLKVIKEGRAALVTSFGI  16573
              AMCGDGANDCGALKVAHAGISLSETEASIASPFTSRNPTISAVLKVIKEGRAALVTSFGI
Sbjct  474    AMCGDGANDCGALKVAHAGISLSETEASIASPFTSRNPTISAVLKVIKEGRAALVTSFGI  533

Query  16574  FKYMAAYSLVQFISVMVLYSIDSNLTDKQYLYVDLGLISIFAFFFGKTGSFDGNLVEQVP  16753
              FKYMAAYSLVQFISVM+LYSIDSNLTDKQYLYVDLGLISIFAFFFGKT SFDG LVEQVP
Sbjct  534    FKYMAAYSLVQFISVMILYSIDSNLTDKQYLYVDLGLISIFAFFFGKTESFDGNLVEQVP  593

Query  16754  LSSLISSTPLASLLLHLTVVTPGFQV  16828
              LSSLISSTPLASLLLHLTVVT FQV
Sbjct  594    LSSLISSTPLASLLLHLTVVTAFQV  618
```

Figure 34: Alignment resulting from BLASTX search using the unmasked Contig9 sequence as the query and the 19_1766_0 peptide sequence as the subject.

Exon 19_1766_0 starts at base 14960 and ends at base 16832 (Figure 35). The start site is supported by multiple gene predictors, BLASTX, mRNA data, TopHat data and an "AG" splice acceptor site at bases 14958 to 14959. The end side is supported by multiple gene predictors, mRNA data, TopHat data and a "GT" splice donor site at bases 16833 to 16834. The BLASTX alignment ends 1 codon before the predicted end site, which makes sense because the last codon (I) of the peptide sequence did not align in Figure 34.

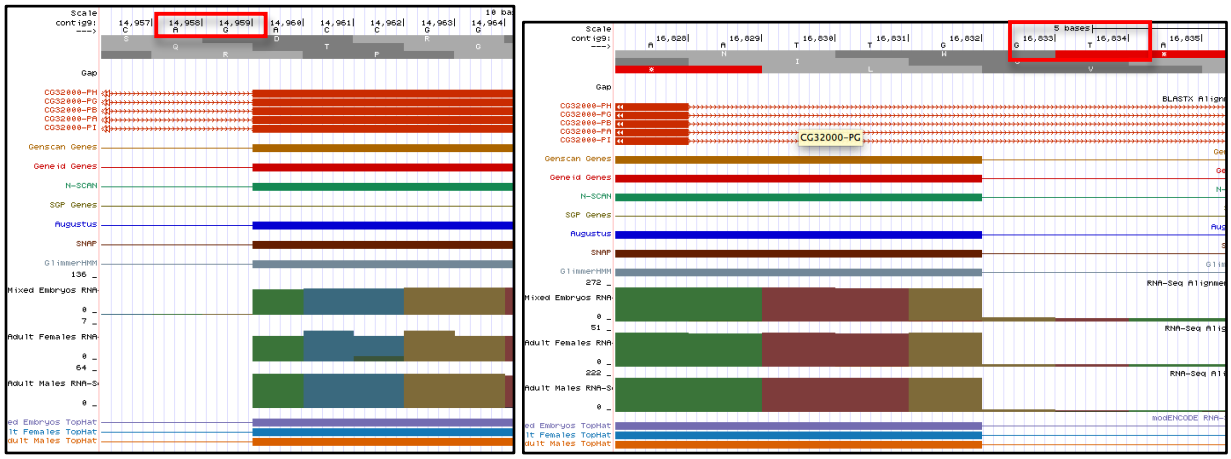


Figure 35: Start (left) and end (right) of Exon 19_1766_0 in reading frame +2.

Exon 20_1766_2:

BLASTX was used to compare the unmasked Contig9 sequence (query) to the 20_1766_2 peptide sequence (subject). The low complexity filter was turned off. Figure 36

shows the alignment. 196 out of 198 residues of the 20_1766_2 peptide sequence are covered by the alignment, and the alignment has an 87% identity.

CG32000:20_1766_2
Sequence ID: lcl|7431 Length: 198 Number of Matches: 7

Range 1: 1 to 196 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
363 bits(933)	2e-117	Compositional matrix adjust.	170/196(87%)	185/196(94%)	0/196(0%)	+2

```
Query  17087  WIHLHQEPWFKPFQPADEDHLGCYENYTMFCISSFQYIILAFVFSKGAPYRKPLWSNFPL  17266
              W+HLHQ+PWFK F+PADEDHLGC+ENYTMFCISSFQYIILAFVFSKGAPYRKPLWSN+PL
Sbjct  1      WVHLHQQPWFKAFEPADEDHLGCFENYTMFCISSFQYIILAFVFSKGAPYRKPLWSNWPL  60

Query  17267  CLAFIVNLCIIVYLVIYPSDWVANFFQLIVPPVMSFRYMMLVYGAAAFSCHIFVESFLVE  17446
              CLAFIVNLCIIVYLV+YPSDWVA+FFQLIVPP M FRY+ML YGAA+F CHIFVESFLVE
Sbjct  61     CLAFIVNLCIIVYLVLYPSDWVASFFQLIVPPTMRFRYVMLAYGAASFICHIFVESFLVE  120

Query  17447  YIVFKKYQVKRDKNWVTAKQKYLRLEYDISTIKNWPPITEVYEPNNCSDWEVDQPTYVSL  17626
              Y+VFKKYQVKR+KNWVT+KQKY+RLE+DIS IKNWPPITEVYEPNN  D E +QPTYVSL
Sbjct  121    YLVFKKYQVKREKNWVTSKQKYMRLEHDISNIKNWPPITEVYEPNNLIDCETEQPTYVSL  180

Query  17627  HAEQNHDTQFGKFPGF  17674
              HAEQNHDTQ GKFPGF
Sbjct  181    HAEQNHDTQLGKFPGF  196
```

Figure 36: Alignment resulting from BLASTX search using the unmasked Contig9 sequence as the query and the 20_1766_2 peptide sequence as the subject.

Exon 20_1766_2 starts at base 17085 and ends at base 17680 (Figure 37). The start site is supported by multiple gene predictors, mRNA data, TopHat data and an "AG" splice acceptor site at bases 17083 to 17084. The end site is supported by multiple gene predictors, mRNA data, and a stop codon at bases 17678 to 17680.

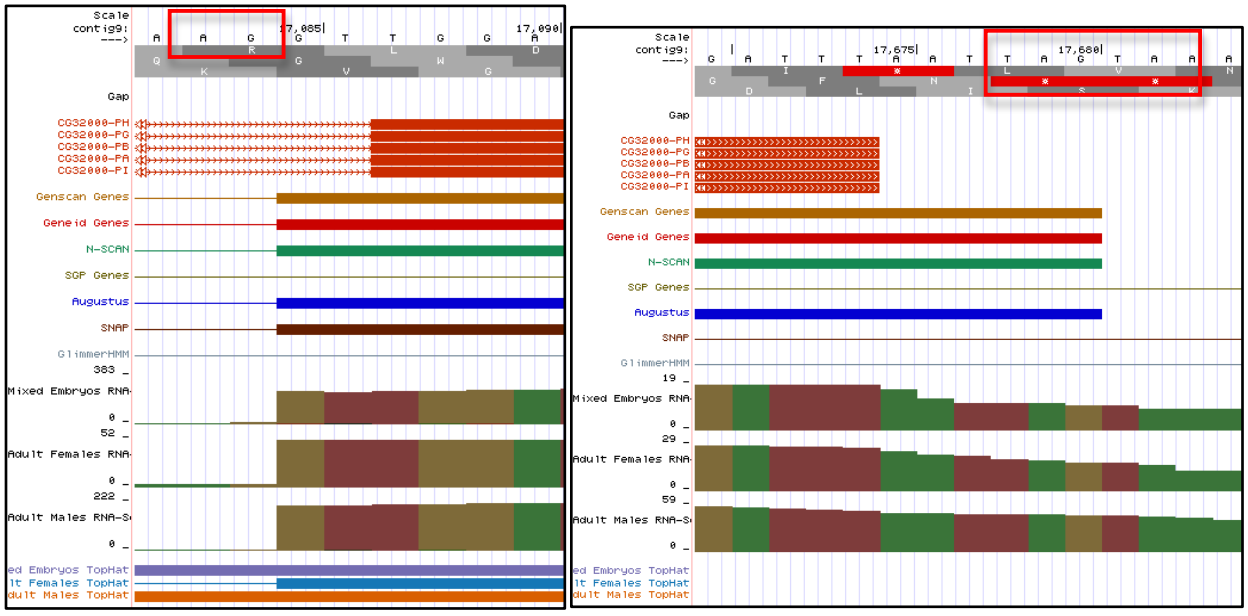


Figure 37: Start (left) and end (right) of Exon 20_1766_2 in reading frame +2. Splice acceptor site and stop codon are boxed in red.

TopHat Data:

TopHat data supports the proposed exons and suggests at least two upstream untranslated exons (see Figure 38).

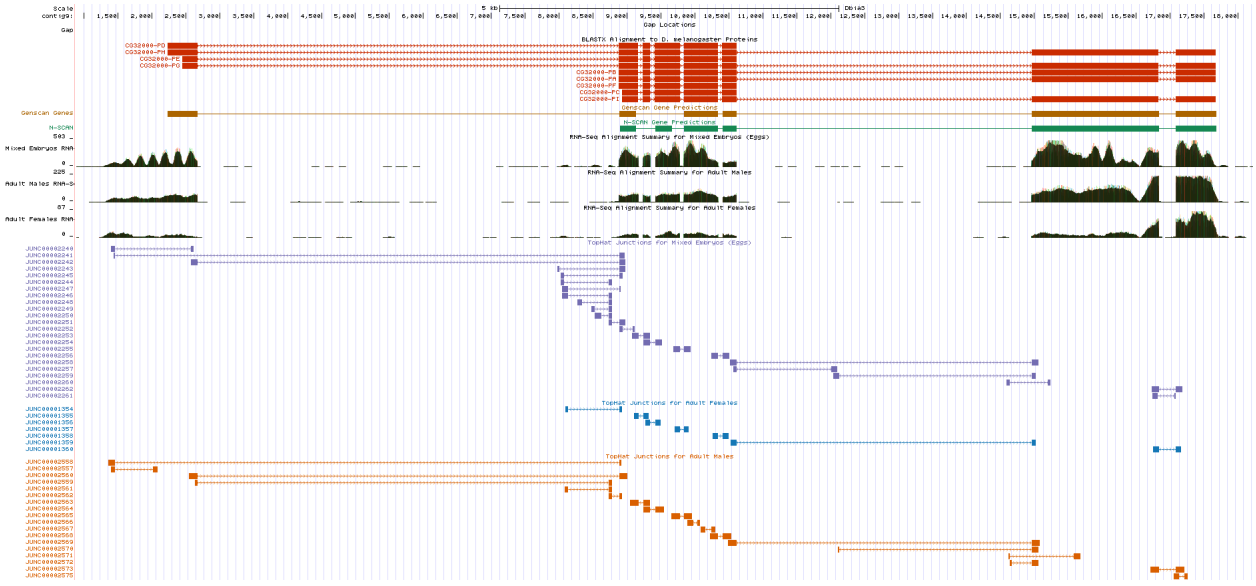


Figure 38: Proposed CG32000 ortholog with TopHat Data at bottom of Figure.

Table of Exons:

Table 2 summarizes the location of all translated exons for CG32000 in both *D. melanogaster* and *D. biarmipes*. As explained earlier, Exon 7_1766_0 is an alternative initial exon in *D. melanogaster* that I hypothesize does not exist in *D. biarmipes*. There are two terminal exons: 18_1767_0 (stop codon at bases 14598 to 14600) and 20_1766_2 (stop codon at bases 17678 to 17680).

D. melanogaster				**Contig9 of *D. biarmipes* Dot chromosome**					
Exon FlyBase ID	Start	End	Length (bp)	Frame	Start	Phase	End	Phase	Length (bp)
CDS_FBgn0052000:5_1767_0	155155	155590	436	+3	2238	0	2673	1	436
CDS_FBgn0052000:7_1766_0	155344	155590	247	Alt. initial exon, does not exist in *D. biarmipes*.					
CDS_FBgn0052000:11_1759_0	157379	157403	25	+1	8758	0	8782	1	25
CDS_FBgn0052000:13_1766_2	157505	157793	289	+2	8883	2	9171	2	289
CDS_FBgn0052000:13_1767_0	157552	157793	242	+2	8930	0	9171	2	242
CDS_FBgn0052000:14_1766_1	157849	157958	110	+1	9237	1	9343	1	107
CDS_FBgn0052000:15_1766_2	158017	158378	362	+3	9418	2	9779	0	362
CDS_FBgn0052000:16_1766_0	158433	158936	504	+3	9837	0	10340	0	504
CDS_FBgn0052000:17_1766_0	159000	159200	201	+2	10412	0	10612	0	201
CDS_FBgn0052000:18_1767_0	160448	160522	75	+3	14526	0	14600	0	75
CDS_FBgn0052000:19_1766_0	160896	162753	1858	+2	14960	0	16832	1	1873
CDS_FBgn0052000:20_1766_2	162826	163421	596	+2	17085	2	17680	0	596

Table 2: Table of exons for CG32000. The start and end points of each exon are shown for both *D. melanogaster* and *D. biarmipes*.

Gene Model Checker:

Gene Model Checker was used to confirm the proposed gene models for the following isoforms: CG32000-PA, CG32000-PB, CG32000-PC, CG32000-PD, CG32000-PF, CG32000-PH, and CG32000-PI. Figure 39 shows the dot plot comparisons of the *D. biarmipes* gene models to *D. melanogaster*. Isoforms CG32000-PE and CG32000-PG could not be confirmed using the Gene Model Checker because both use Exon 7_1766_0 as their initial exon, which does not appear to exist in *D. biarmipes*. Apart from the initial exon, CG32000-PE has the same polypeptide sequence as CG32000-PD and CG32000-PG has the same polypeptide sequence as CG32000-PH. However, the isoforms have different untranslated regions. Determining whether isoforms CG32000-PE and CG32000-PG exist in *D. biarmipes* will require further investigation.

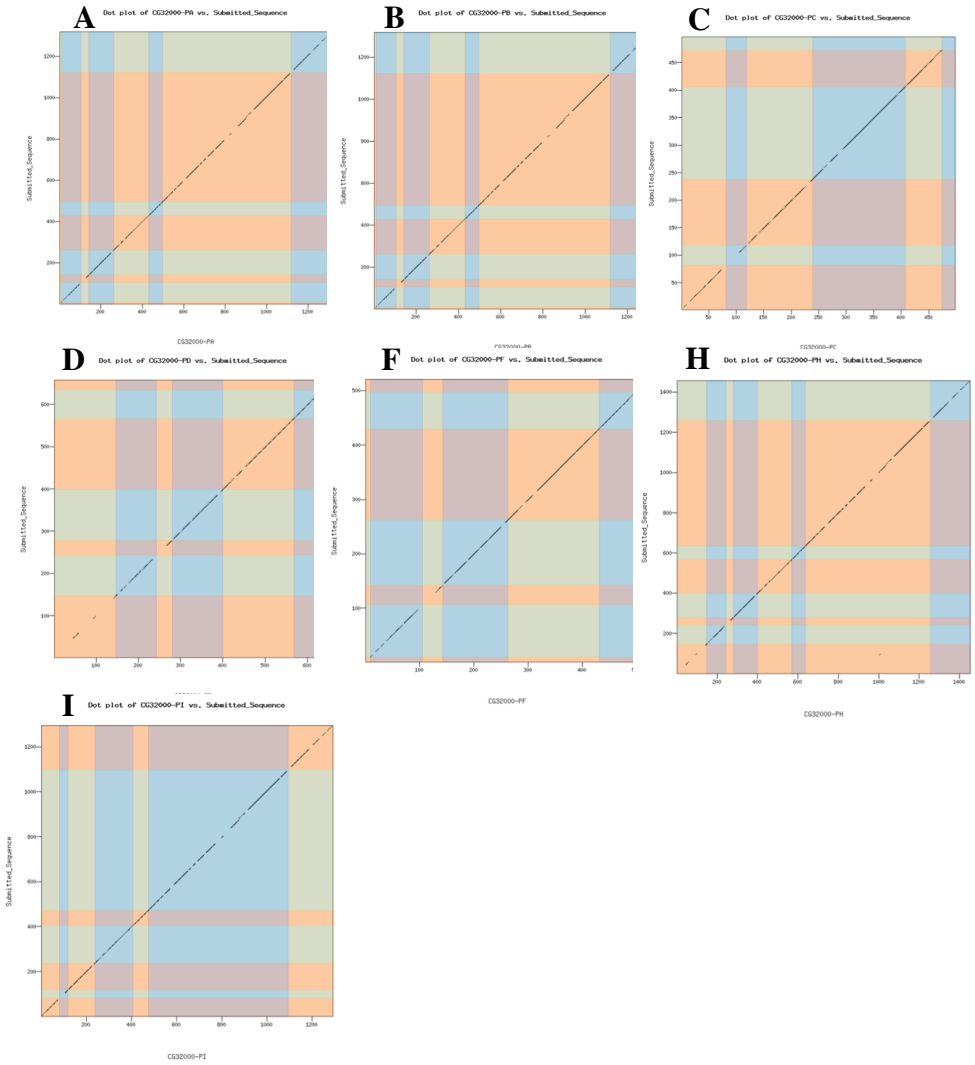


Figure 39: Dot plot results from Gene Model Checker for isoforms CG32000-PA, CG32000-PB, CG32000-PC, CG32000-PD, CG32000-PF, CG32000-PH, and CG32000-PI

Untranslated Regions:

The 5' untranslated regions (UTRs) for the various *CG32000* isoforms are complex, and likely expand beyond the 5' boundary of Contig9 for isoforms E, H, C, D and G. Figure 40 shows the FlyBase GBrowse2 schematic of the 5' UTRs of the *D. melanogaster CG32000* gene. The exons are labeled with their FlyBase ID number.

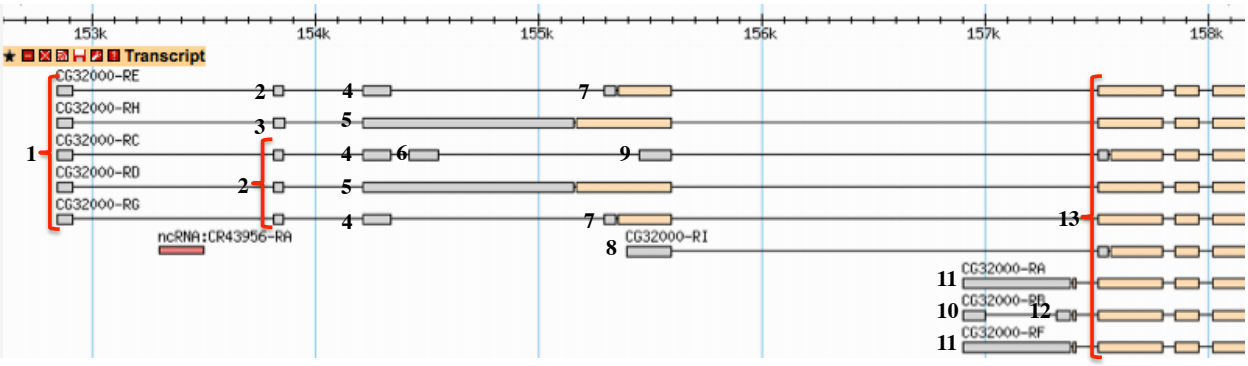


Figure 40: FlyBase GBrowse2 view of transcript details for *CG32000* in *D. melanogaster*. Gray boxes represent untranslated exons and yellow boxes represent translated exons. Figure is zoomed in on the 5' end of *CG32000*. Exons are labeled with their FlyBase ID number.

In order to investigate the 5' UTRs, the transcripts for the untranslated exons in *D. melanogaster* were obtained through the GEP Gene Record Finder. For each exon, a BLASTn search was performed to compare the unmasked Contig9 sequence (query) to the *D. melanogaster CG32000* nucleotide sequence (subject). The BLASTn parameters were adjusted to provide optimal results for the expected low percent identity between UTRs in *D. melanogaster* and *D. biarmipes* (no low complexity filter, Match/Mismatch Scores set to +1/-1, Gap Costs set to Existence: 2, Extension: 1, and Word size set to 7). RNA-Seq and TopHat data from the GEP UCSC Genome Browser Mirror was used in conjunction with the BLASTn results to determine a tentative model for the 5' UTRs. Table 3 summarizes the results of the attempt to annotate the 5' UTRs. Figure 41 shows all of Contig9 that is upstream of the first coding exon of *CG32000* (exon 5_1767_0). I expect to find the 5' UTRs for isoforms C, D, E, G, H and I in this region. Figure 42 shows region of Contig9 where I expect to find to 5' UTRs for isoforms A, B and F of *CG32000*.

Exon	Size (bp)	BLASTn Results	GEP UCSC Genome Browser	Annotation
1	68	Inconclusive. No significant alignments in the appropriate region.	No TopHat data or RNA-Seq, probably not included in Contig9.	Probably located beyond the 5' boundary of Contig9.
2 & 3	46 & 50	No significant alignments on the plus strand in the appropriate region	No significant data.	Might not be on Contig9, need wet lab data to confirm location.
4	126	Bases 63 to 115 of Exon 4 align to bases 1320 to 1372 of Contig9 with 72% identity.	Supported by RNA-Seq data, potentially TopHat junctions	Probably located in the region of bases 1275 to 1400. Need wet lab data to confirm.
5	1377	This exon includes coding exon 5_1767_0. Bases 115 to 1377 of Exon 5 align to bases 1454 to 2673 of Contig9 with 68% identity. Bases 63 to 115 of Exon 5 align to bases 1320 to 1372 of Contig9 with 72% identity.	Significant RNA-Seq coverage from bases 1275 to 2673.	Need wet lab data to confirm, but untranslated region probably spans from around base 1275 to base 2238 (beginning of coding exon 5_1767_0).
6	130	Bases 29 to 130 of Exon 6 align to bases 1561 to 1662 of Contig9 with 81% identity.	RNA-Seq data in region. No TopHat data to support junction with Exons 4 and 9.	Potentially located in region identified by BLASTn, need wet lab data to confirm if exon structure is the same as in *D. melanogaster*.
7	294	Bases 1 to 294 of Exon 7 align to bases 2380 to 2673 of Contig9 with 73% identity.	RNA-Seq data in region. Includes *D. melanogaster* coding exon 7_1766_0, which does not exist in *D. biarmipes*.	Need wet lab data to confirm if this exon exists in *D. biarmipes*.
8	199	Bases 9 to 199 of Exon 8 align to bases 2483 to 2673 of Contig9 with 76% identity.	RNA-Seq data in region.	Potentially located in region identified by BLASTn, need wet lab data to confirm if exon exists in *D. biarmipes*.
9	140	Bases 5 to 140 of Exon 9 align to bases 2538 to 2673 of Contig9 with 76% identity.	RNA-Seq data in region.	Potentially located in region identified by BLASTn, need wet lab data to confirm if exon exists in *D. biarmipes*.
10	100	Bases 6 to 95 of Exon 10 align to bases 8229 to 8320 of Contig9 with 66% identity.	Figure 42 shows scattered RNA-Seq data in this region; exon location is supported by JUNC00002248. Bases 7900 to 8200 also have RNA-Seq data; this region is supported by TopHat data. There is a match to initiator sequence TCAKTY at base 8,137.	There is likely an untranslated exon in the region of bases 7900 to 8320, but wet lab data is needed to confirm the exact location and exon structure.
11	506	Bases 6 to 313 of Exon 11 align to bases 8229 to 8526 of Contig9 with 66% identity.	Some RNA-Seq data in the region. There is a match to initiator sequence TCAKTY at base 8,137.	Potentially located in region identified by BLASTn, need wet lab data to confirm if exon exists in *D. biarmipes*.
12	87	Inconclusive.	Some RNA-Seq data in expected region. For example, coverage from bases 8734 to 8758, directly before predicted coding exon 11_1759_0. Also top hat data to support a splice junction at base 8734.	Need wet lab data to confirm exon existence and location in *D. biarmipes*.
13	289	Bases 1 to 287 of Exon 13 align to bases 8883 to 9169 of Contig9 with 85% identity.	Significant RNA-Seq coverage and TopHat data.	Exon 13 corresponds to coding exon 13_1766_2. In isoforms C and I bases 8883 to 8929 are untranslated.

Table 3: Summary of attempt to annotate 5' UTRs for *CG32000*.

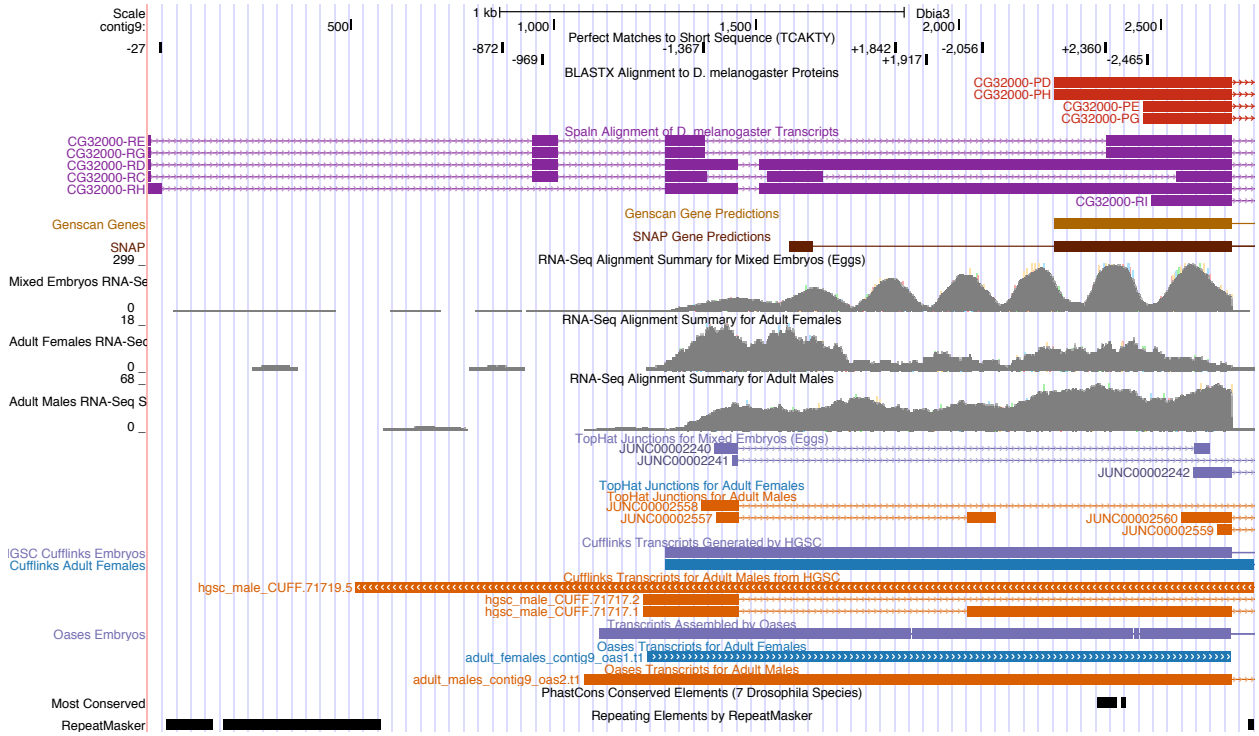


Figure 41: Region of Contig9 upstream of 5_1767_0, the first coding exon of *CG32000*.

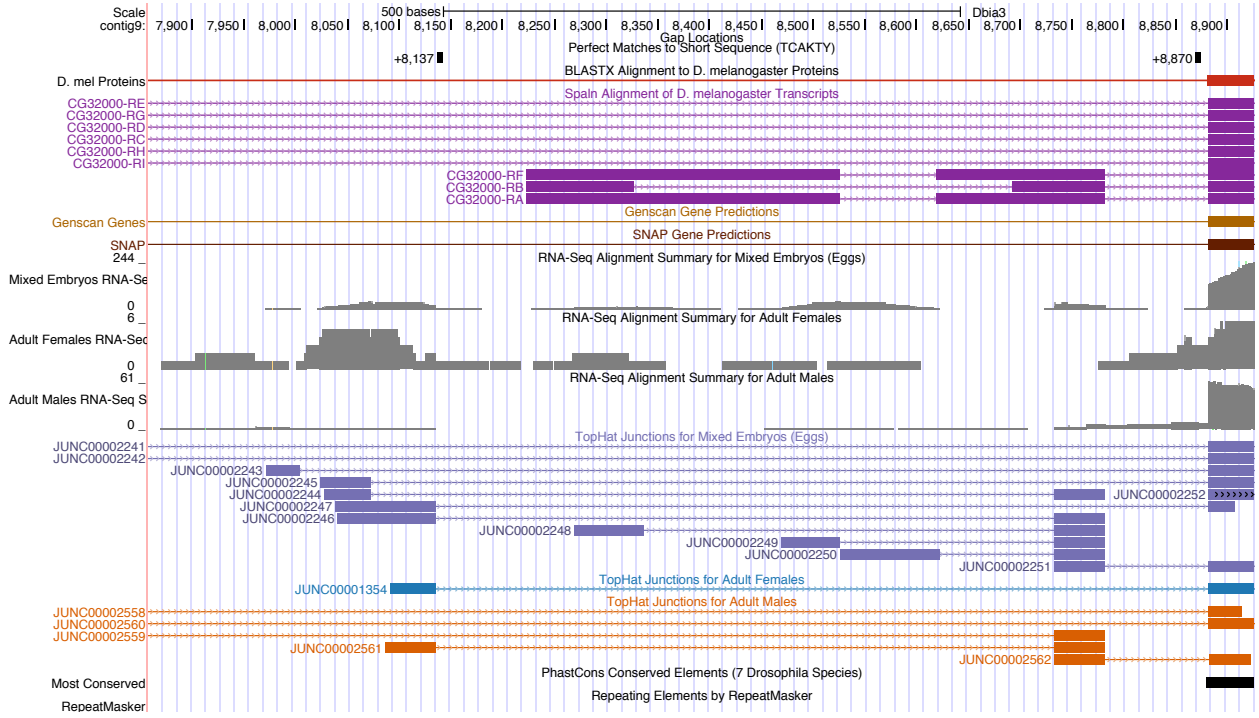


Figure 42: Region of Contig9 where I expect to find the 5' UTRs for isoforms A, B and F of *CG32000*.

Overall, confirming the structure of the 5' UTRs for the various isoforms of *CG32000* will require further evidence through wet lab experiments.

There are only two different 3' UTRs for *CG32000* in *D. melanogaster*. As shown in Figure 43, isoforms C, D, E and F have a terminal exon (18_1767_0) and 3' UTR encoded by Exon 18 and isoforms A, B, G, H and I have a terminal exon (20_1766_2) and 3'UTR encoded by Exon 20.

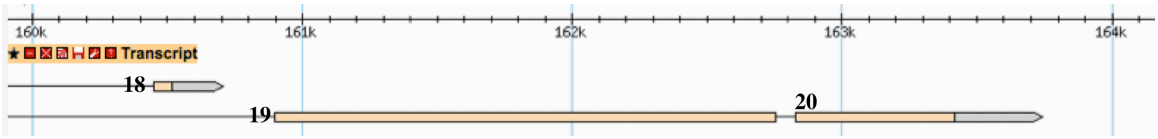


Figure 43: FlyBase GBrowse2 view of transcript details for *CG32000* in *D. melanogaster*. Gray boxes represent untranslated exons and yellow boxes represent translated exons. Figure is zoomed in on the 3' end of *CG32000*. Exons are labeled with their FlyBase ID number. Isoforms C, D, E and F have a terminal exon (18_1767_0) and 3' UTR encoded by Exon 18. Isoforms A, B, G, H and I have a terminal exon (20_1766_2) and 3'UTR encoded by Exon 20.

The same procedure that was used to annotate the 5' UTRs was again used to annotate the 3' UTRs. The BLASTn search using Exon 18 as the subject and the unmasked Contig9 sequence as the query was inconclusive, which was expected because coding exon 18_1767_0 was not very well conserved and only tentatively annotated. Figure 44 shows the predicted location of coding exon 18_1767_0, which is from bases 14526 to 14600. There is some RNA-Seq data from bases 14600 to 14725, which could be the 3' UTR. The 3' UTR could end at base 14642, which is the boundary predicted by SNAP and supported by a drop-off in RNA-Seq data. Further evidence and wet lab experiments are necessary to confirm the location of Exon 18 in *D. biarmipes*.

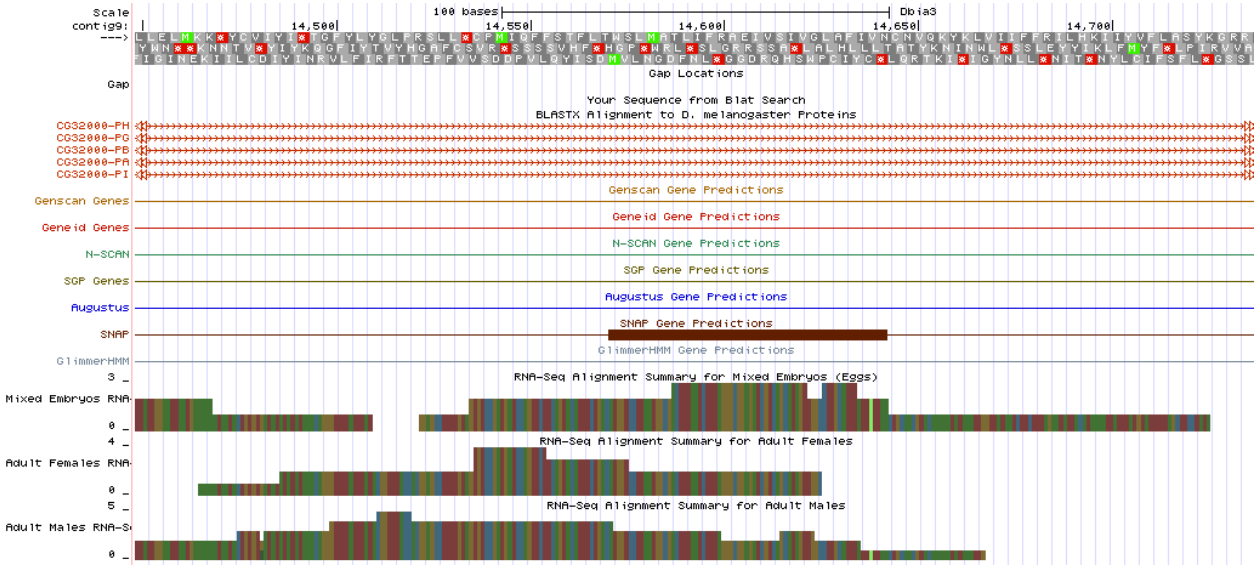


Figure 44: Location of coding exon 18_1767_0, which is predicted to be at bases 14526 to 14600, on Contig9.

Exon 20 is 924 bases long, and the BLASTn search using Exon 20 as the subject and the unmasked Contig9 sequence as the query predicted that bases 1 to 845 of Exon 20 align to bases 17085 to 17973 of Contig9 with 75% identity. Figure 45 shows the end of terminal exon 20_1766_2, which is located at bases 17085 to 17680. The alignment with the *D. melanogaster* transcript extends to base 18036 and weak RNA-Seq coverage extends to approximately base 18150. Further evidence and wet lab experiments are necessary to confirm the extent of the 3'UTR in *D. biarmipes*.

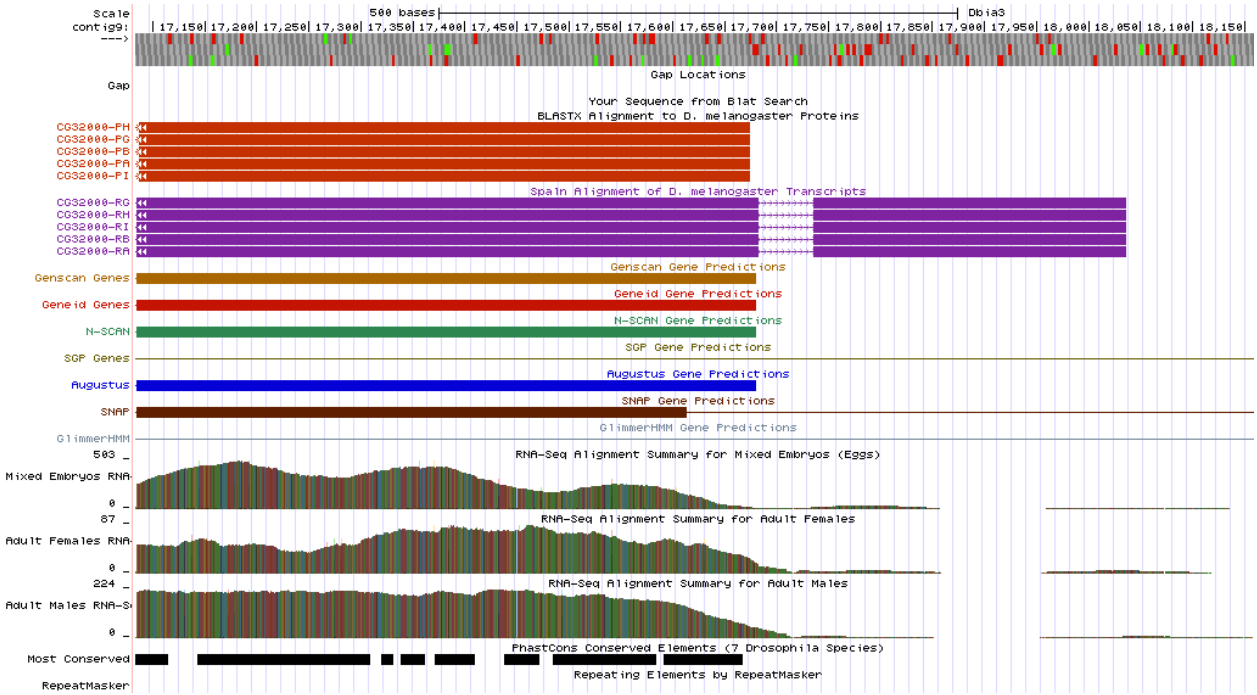


Figure 45: The end of terminal exon 20_1766_2, which is located at bases 17085 to 17680.

***CR43956*:**

In *D. melanogaster*, *CR43956* is a non-coding RNA on the negative strand that is located in the 5' UTR of *CG32000* (Figure 46). A BLASTn search comparing the unmasked Contig9 sequence (query) to the FlyBase sequence for *CR43956* in *D. melanogaster* (subject) was inconclusive (Figure 47). Based on the annotation of the 5' UTR of *CG32000*, it is likely that *CR43956* is located beyond the 5' boundary of Contig9. Further evidence is necessary to annotate this feature.

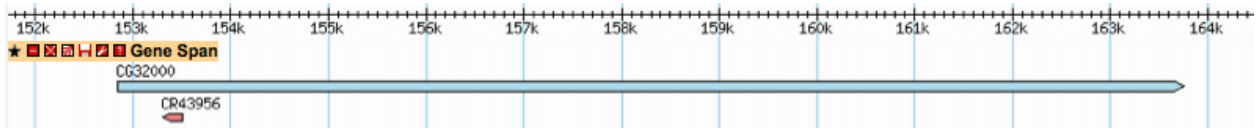


Figure 46: FlyBase GBrowse2 view of the location of *CR43956* in *D. melanogaster*.

Sequence ID: lcl|180823 Length: 201 Number of Matches: 9

Range 1: 79 to 92 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
16.8 bits(10)	7.3	12/14(86%)	0/14(0%)	Plus/Minus

```
Query  180  TTCCTTCTTGCAAC  193
            || |||||||| |||
Sbjct  92   TTGCTTCTTGTAAC  79
```

Range 2: 19 to 43 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Identities	Gaps	Strand
21.1 bits(13)	0.37	21/27(78%)	2/27(7%)	Plus/Plus

```
Query  390  ATTAAATTAGCTGCCATAAAAACTGGT  416
            ||| |||  ||| | ||||||||| ||
Sbjct  19   ATTGAAT--GCTACAATAAAAACTAGT  43
```

Range 3: 126 to 196 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Identities	Gaps	Strand
18.3 bits(11)	2.7	45/73(62%)	4/73(5%)	Plus/Plus

```
Query  1362  CAATGAAATGAGTAATACCGTAAAATTTTATGACAAAAACATAACATACATTCGAA-ATA  1420
             || ||| ||  | |||||||     | |  | | | || |  |||| | ||  ||| | |
Sbjct  126   CACTGACATTGGCAATACCGGTGCAGTCAAAAAAATAACCCGAACAAATAT--GAATAGA  183

Query  1421  TAG-AAATGTTTA  1432
              |||  ||| ||||
Sbjct  184   TAGCTAATCTTTA  196
```

Range 4: 115 to 130 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Identities	Gaps	Strand
16.8 bits(10)	7.3	13/16(81%)	0/16(0%)	Plus/Minus

```
Query  1362  CAATGAAATGAGTAAT  1377
             || |||||||  | |||
Sbjct  130   CAGTGAAATGGGAAAT  115
```

Figure 47: Results in expected region of Contig9 for BLASTn search comparing the unmasked Contig9 sequence (query) to the FlyBase sequence for *CR43956* in *D. melanogaster* (subject).

Gene *CG32006*:

Genscan Feature 3 appears to align to *CG32006* in *D. melanogaster* based on the BLASTX alignment shown in Figure 48. The presence of a gene in the region of Genscan Feature 3 is further supported by RNA-Seq, TopHat and other gene predictors.

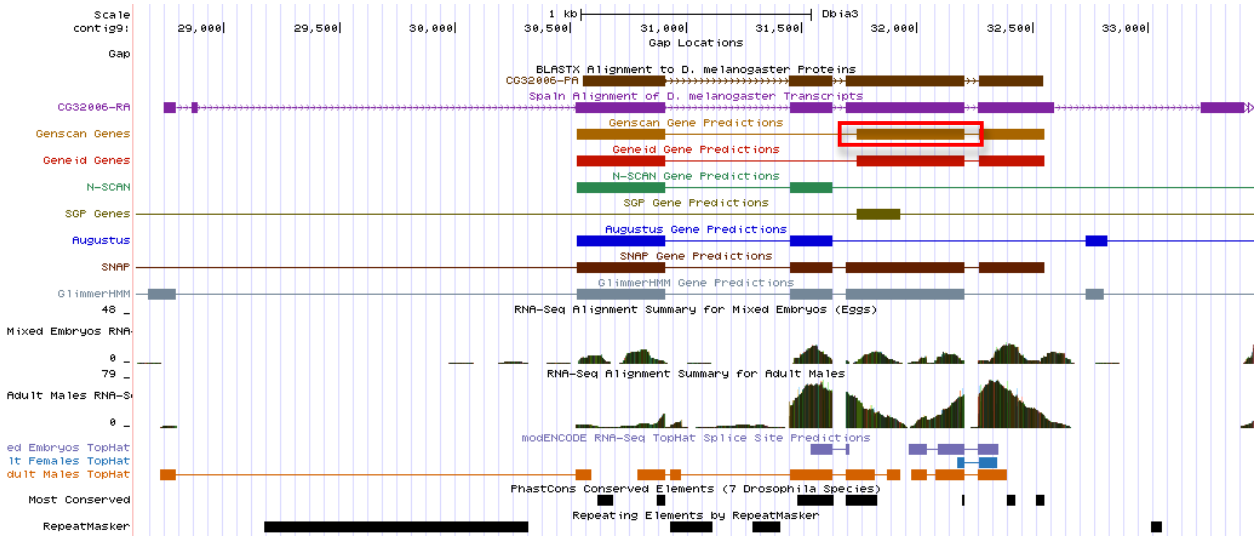


Figure 48: Genscan Feature 3 on GEP UCSC Genome Browser view of Contig9 of *D. biarmipes* (August 2013). Genscan Exon 2 of Feature 3 is boxed in red.

The first step was to use the predicted protein sequence for Genscan Exon 2 of Feature 3 (boxed in red in Figure 48) to perform a FlyBase BLASTp search against all annotated proteins for *D. melanogaster*. Figure 49 shows the alignments generated by the BLASTp search; the best match is clearly to CG32006-PA. Feature 3 is likely orthologous to *CG32006* in *D. melanogaster*.

BLAST Hit Summary

	Description	Species	Score	E value
☑	CG32006-PA	Dmel	246.899	2.35814e-65
☑	Cep135-PC	Dmel	30.0314	4.23801
☑	Cep135-PB	Dmel	30.0314	4.60678
☑	Cep135-PA	Dmel	30.0314	4.68431

Figure 49: BLAST Hit Summary from BLASTp search aligning the predicted protein sequence for Genscan Exon 2 of Feature 3 against all annotated proteins for *D. melanogaster*.

According to the GEP Gene Record Finder, *CG32006* has one isoform (A), which has four translated exons (Figure 50). Figure 51 shows the FlyBase GBrowse2 view of the *CG32006* transcript in *D. melanogaster*.

Gene Details

FlyBase ID	FlyBase Name	Chr	5' Start	3' End	Strand	Graphical Viewer
FBgn0052006	CG32006	4	171,390	174,164	+	View in GBrowse

mRNA Details

Select a row to display the corresponding transcript and peptide details:

FlyBase ID	FlyBase Name	Chr	5' Start	3' End	Strand	Protein ID	Graphical Viewer
FBtr0089170	CG32006-RA	4	171,390	174,164	+	FBpp0088237	View in GBrowse

Transcript Details | **Polypeptide Details**

Options: Export All Unique CDS's to FASTA | Export All CDS's for Selected Isoform to FASTA | Download CDS Workbook

CDS usage map:

Isoform	2_1769_0	3_1769_1	4_1769_0	5_1769_0
CG32006-RA	Y	Y	Y	Y

Select a row to display the corresponding CDS sequence:

FlyBase ID	5' Start	3' End	Strand	Phase	Length
2_1769_0	172,377	172,756	+	0	126
3_1769_1	172,821	173,004	+	1	61
4_1769_0	173,065	173,547	+	0	161
5_1769_0	173,610	173,894	+	0	95

Figure 50: Gene Record Finder entry for *D. melanogaster* gene *CG32006*. The FlyBase ID number for *CG32006* is FBgn0052006.

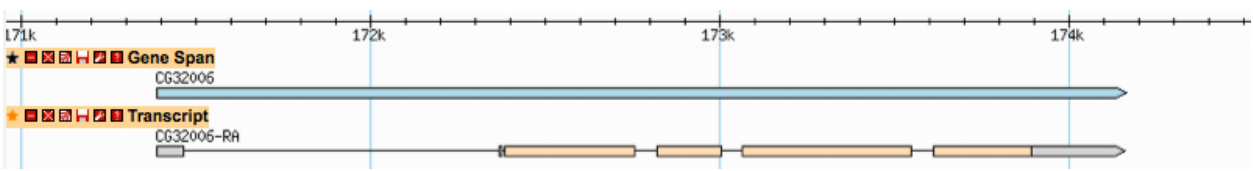


Figure 51: FlyBase GBrowse2 view showing the transcript details of *CG32006* in *D. melanogaster*. Gray boxes represent untranslated regions and orange boxes represent translated exons.

The next step was to confirm that Contig9 of *D. biarmipes* contains the entire CG32006 protein sequence. Peptide sequences for each translated exon were obtained through the GEP Gene Record Finder. For each exon, a BLASTX search was performed comparing the *D. melanogaster* CG32006 peptide sequence to the unmasked Contig9 sequence. Table 4 summarizes the results of the BLASTX searches.

Subject	Subject Start	Subject End	Query	Query Start	Query End	E-value	% Identity	Frame
2_1769_0	1	126	Contig9	30531	30908	8e-22	51%	+ 3
3_1769_1	1	61	Contig9	31451	31633	6e-37	93%	+ 2
4_1769_0	1	161	Contig9	31694	32203	2e-46	52%	+ 2
5_1769_0	2	95	Contig9	32267	32551	2e-36	67%	+ 2

Table 4: Results of BLASTX searches comparing the translated exon sequences for CG32006 (subject) to the unmasked Contig9 sequence (query).

The results of the BLASTX searches provide strong evidence that this feature is orthologous to *CG32006* in *D. melanogaster*. In order to confirm the location of each translated exon, the GEP UCSC Genome Browser was used to examine the available RNA-Seq, TopHat, Cufflinks, Oases and gene predictor data. Table 5 summarizes the location of all translated exons for *CG32006* in *D. biarmipes*. 5_1769_0 is the terminal exon, for which the stop codon is located at bases 32549 to 32551. The annotations were fairly straightforward, with the exception of picking the appropriate splice acceptor site for exon 5_1769_0. As shown in Figure 52, there are two possible splice acceptor sites that result in the exon starting in phase 0 in the +2 frame. The start site at base 32264 is supported by TopHat data for mixed embryos and adult males, as well as a single RNA-Seq read. The start site at base 32267 is supported by TopHat data for adult females, as well as most of the RNA-Seq data. I selected base 32264 as the start site to ensure that no amino acids are missing from the annotation, but this exon requires further investigation.

D. melanogaster				**Contig9 of *D. biarmipes* Dot chromosome**					
FlyBase_ID	Start	End	Length (bp)	Frame	Start	Phase	End	Phase	Length (bp)
CDS_FBgn0052006:2_1769_0	172377	172756	380	+3	30531	0	30910	2	380
CDS_FBgn0052006:3_1769_1	172821	173004	184	+2	31450	1	31633	0	184
CDS_FBgn0052006:4_1769_0	173065	173547	483	+2	31694	0	32203	0	510
CDS_FBgn0052006:5_1769_0	173610	173894	285	+2	32264	0	32551	0	288

Table 5: Table of exons for *CG32006*. The start and end points of each exon are shown for both *D. melanogaster* and *D. biarmipes*.

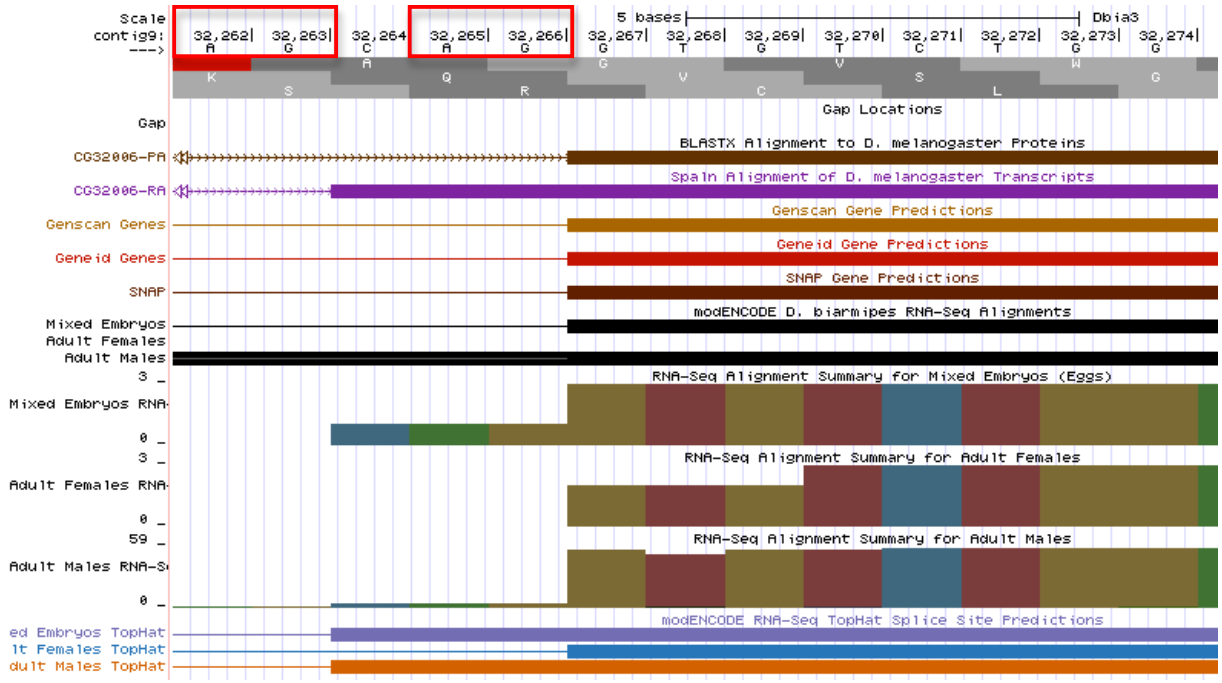


Figure 52: The two possible splice acceptor sites for exon 5_1769_0 are boxed in red.

Gene model checker was used to confirm the proposed gene model for CG32006-PA. Figure 53 shows the dot blot comparison of the *D. biarmipes* gene model (vertical axis) to *D. melanogaster* (horizontal axis). While there are apparent gaps in the dot plot, the overall alignment to the *D. melanogaster* gene has a 60% identity and 72.7% similarity.

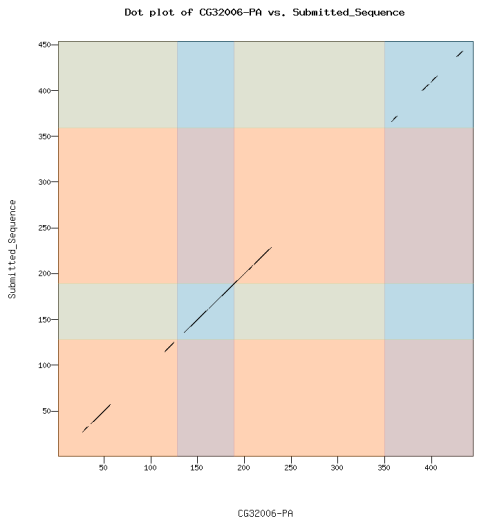


Figure 53: Dot plot result from Gene Model Checker for CG32006-PA.

In order to investigate the UTRs for *CG32006* in *D. biarmipes*, the transcript details were examined in Gene Record Finder (Figure 54). Exon 1 is part of the 5' UTR, and exon 2 includes

5’ UTR as well as the coding exon 2_1769_0. Exon 5 includes coding terminal exon 5_1769_0, as well as the 3’ UTR (see Fig. 51).

Exon usage map:

Isoform	1	2	3	4	5
CG32006-RA	Y	Y	Y	Y	Y

Select a row to display the corresponding exon sequence:

FlyBase ID	5' Start	3' End	Strand	Length
1	171,390	171,462	+	73
2	172,368	172,756	+	389
3	172,821	173,004	+	184
4	173,065	173,547	+	483
5	173,610	174,164	+	555

Figure 54: Transcript details for *CG32006-RA* in Gene Record Finder. *CG32006-RA* has five exons.

Figure 55 shows the region of Contig9 where I expect to find the 5’ UTRs for *CG32006*. A BLASTn search comparing Exon 1 (subject) to the unmasked Contig9 sequence (query) was inconclusive. However, examining the genome browser reveals an initiator sequence at base 28729 and RNA-Seq coverage in the surrounding region. This potential UTR extends to about base 28793 as supported by RNA-Seq, TopHat, and a “GT” splice donor site at bases 28794 to 28795. TopHat JUNC00002577 supports a junction between this region and the initial coding exon. Hence, there is potentially a 5’ untranslated exon at bases 28729 to 28793.

In *D. melanogaster* there is also a small 5’ UTR immediately adjacent to the initial coding exon. A BLASTn search comparing Exon 2 (subject) to the unmasked Contig9 sequence (query) showed that bases 5 to 388 of Exon 2 align to bases 30526 to 30909 of Contig9 with 67% identity. Examining the genome browser reveals RNA-Seq, TopHat and an “AG” splice acceptor site at bases 30523 to 30524 support that the 5’ UTR is located at bases 30525 to 30530.

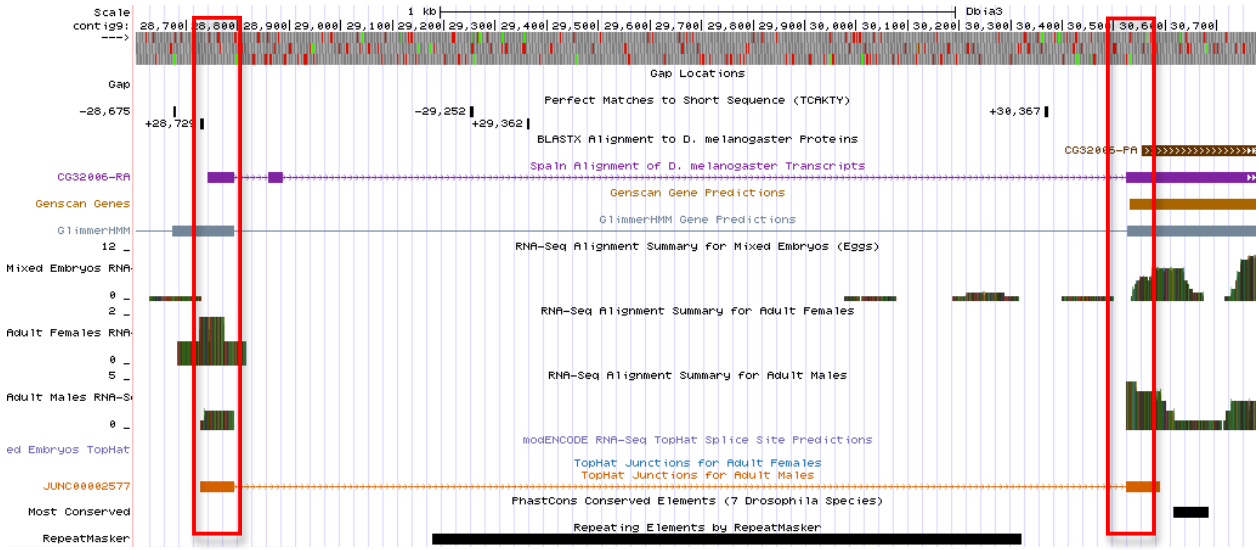


Figure 55: Region of Contig9 containing the 5' UTRs for *CG32006*. The proposed 5' UTRs are boxed in red.

In order to investigate the 3' UTR, I conducted a BLASTn search comparing Exon 5 (subject) to the unmasked Contig9 sequence (query). Exon 5 is 555 bases long, and bases 7 to 301 of Exon 5 align to bases 32270 to 32564 of Contig9 with 75% identity. As expected, the translated portion of Exon 5, which corresponds to terminal exon 5_1769_0, is well conserved while the UTR is less conserved. It is not possible to annotate the 3' UTR precisely, but based on RNA-Seq data it likely extends from base 32552 (immediately after the stop codon, reading frame +2) to around base 32679 (Figure 56).

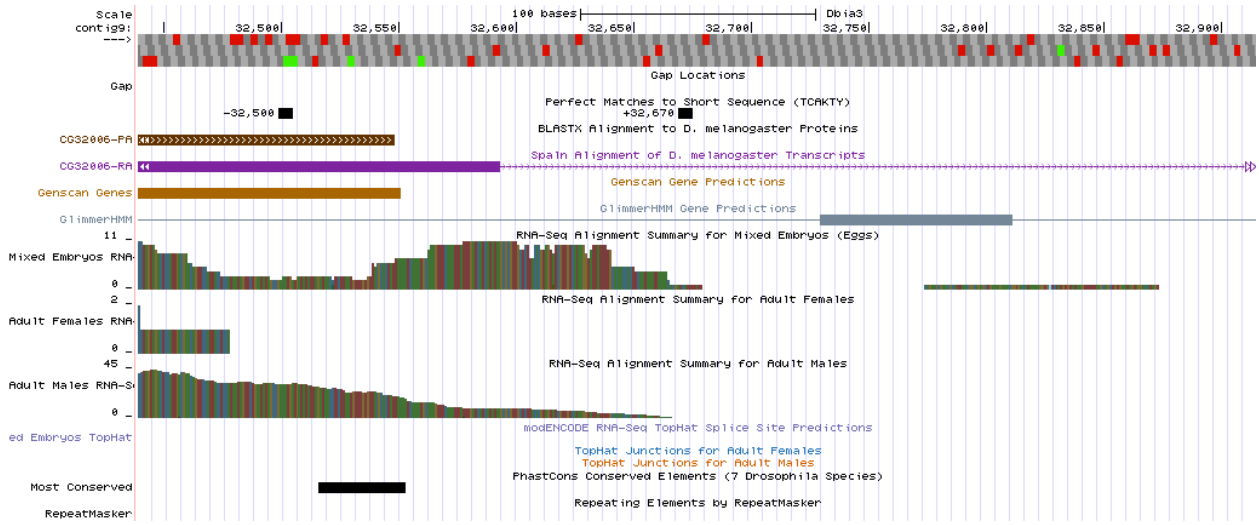


Figure 56: 3' UTR of the *CG32006* gene in *D. biarmipes*. The 3' UTR is located at around bases 32552 to 32679.

***CG2219* (*Arl4*) Paralog:**

Genscan Feature 4 appears to align to isoforms A and B of *CG2219* in *D. melanogaster*

based on the BLASTX alignment shown in Figure 57. *CG2219* is named *ADP ribosylation factor-like 4* (*Arl4*). According to Gene Record Finder, there are three isoforms (A, B and C) of *Arl4* in *D. melanogaster* (Figure 58). In *D. melanogaster*, *Arl4* has a total of seven coding exons, while there appears to be only one region of significant alignment on Contig9 that does not overlap with other features. The FlyBase GBrowse2 diagram shows that in *D. melanogaster*, *Arl4* is not located at the 5' end of *CG32006*, but is further downstream (Figure 59). Based on this initial analysis, I hypothesized that Genscan Feature 4 may be a duplication of *Arl4*, and the complete *Arl4* gene may be located further downstream.

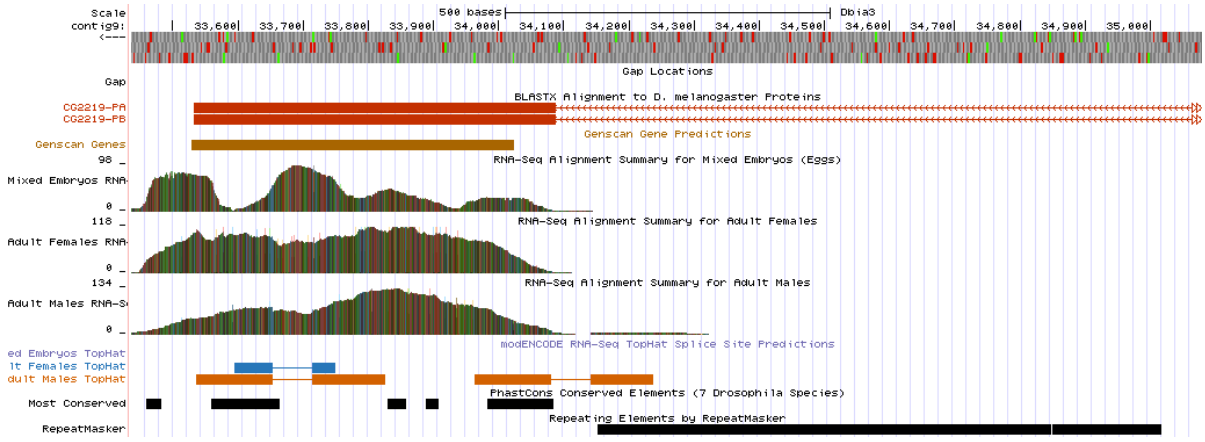
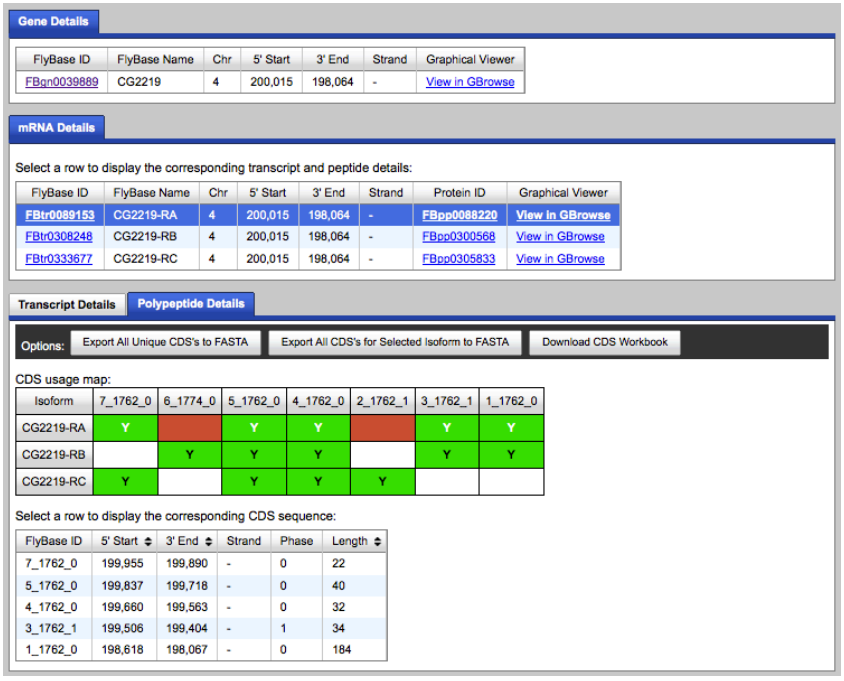


Figure 57: Genscan Feature 4 aligns to *CG2219-PA* and *CG2219-PB*. There is a large "unknown" repeat to the 3' end of the alignment. The TopHat data in the region is likely spurious.



Gene Details

FlyBase ID	FlyBase Name	Chr	5' Start	3' End	Strand	Graphical Viewer
FBgn0039889	CG2219	4	200,015	198,064	-	View in GBrowse

mRNA Details

Select a row to display the corresponding transcript and peptide details:

FlyBase ID	FlyBase Name	Chr	5' Start	3' End	Strand	Protein ID	Graphical Viewer
FBtr0089153	CG2219-RA	4	200,015	198,064	-	FBpp0088220	View in GBrowse
FBtr0308248	CG2219-RB	4	200,015	198,064	-	FBpp0300568	View in GBrowse
FBtr0333677	CG2219-RC	4	200,015	198,064	-	FBpp0305833	View in GBrowse

Transcript Details | Polypeptide Details

Options: Export All Unique CDS's to FASTA | Export All CDS's for Selected Isoform to FASTA | Download CDS Workbook

CDS usage map:

Isoform	7_1762_0	6_1774_0	5_1762_0	4_1762_0	2_1762_1	3_1762_1	1_1762_0
CG2219-RA	Y		Y	Y		Y	Y
CG2219-RB		Y	Y	Y		Y	Y
CG2219-RC	Y		Y	Y	Y		

Select a row to display the corresponding CDS sequence:

FlyBase ID	5' Start	3' End	Strand	Phase	Length
7_1762_0	199,955	199,890	-	0	22
5_1762_0	199,837	199,718	-	0	40
4_1762_0	199,660	199,563	-	0	32
3_1762_1	199,506	199,404	-	1	34
1_1762_0	198,618	198,067	-	0	184

Figure 58: Gene Record Finder entry for *D. melanogaster* gene *CG2219* (*Arl4*). The FlyBase ID number for *CG2219* is FBgn0039889.

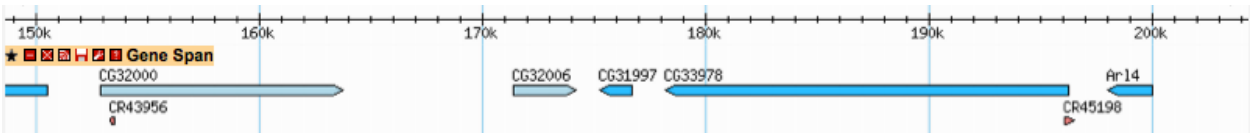


Figure 59: FlyBase GBrowse2 view of *Arl4*, showing location and orientation relative to other genes in *D. melanogaster*.

In order to confirm that Genscan Feature 4 best aligns to *Arl4* and not a different ADP ribosylation factor gene, the predicted protein sequence for Genscan Feature 4 was used to perform a FlyBase BLASTp search against all annotated proteins for *D. melanogaster*. Figure 60 shows the top alignments generated by the BLASTp search. The best matches are clearly *Arl4-PB* and *Arl4-PA*.

BLAST Hit Summary

	Description	Species	Score	E value
☑	Arl4-PB	Dmel	235.343	1.69935e-62
☑	Arl4-PA	Dmel	235.343	1.8627e-62
☑	Arf79F-PJ	Dmel	39.6614	0.00125896
☑	Arf79F-PI	Dmel	39.6614	0.00125896
☑	Arf79F-PH	Dmel	39.6614	0.00125896

Figure 60: BLAST Hit Summary from BLASTp search aligned the predicted protein sequence for Genscan Feature 4 against all annotated proteins for *D. melanogaster*.

The next step was to determine which exon(s) of *Arl4* are encoded by Genscan Feature 4. BLASTX was used to compare the unmasked Contig9 sequence (query) to the 3' terminal coding exon (1_1762_0) for isoforms A and B (subject) (Figure 61). The alignment spans all 184 residues of the 1_1762_0 peptide sequence with a 77% identity. The alignment is located on negative strand of Contig9 at bases 34081 to 33530, which encompasses all of Genscan Feature 4. Genscan Feature 4 thus likely corresponds to exon 1_1762_0 of *Arl4*. Examining the Genome Browser reveals that the feature is located on the -3 frame, and there is a start codon beginning at base 34109, followed by an open reading frame, and a terminal exon at bases 33532 to 33530.

CG2219:1_1762_0

Sequence ID: lcl|69835 Length: 184 Number of Matches: 3

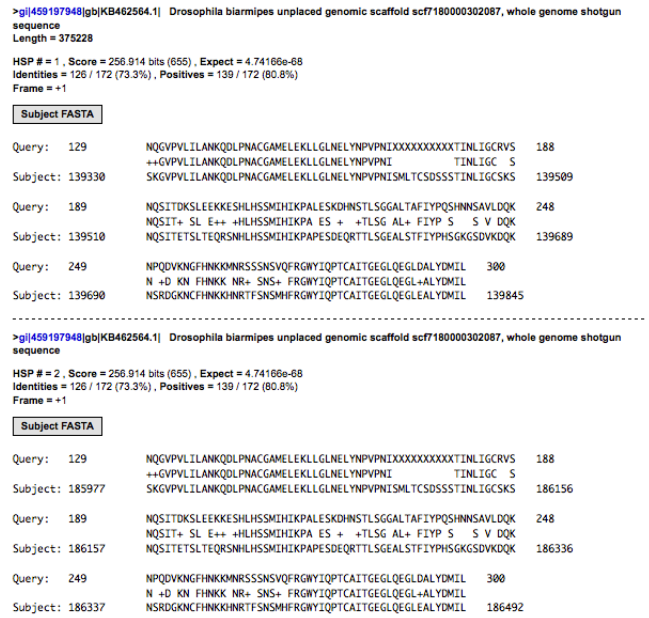
Range 1: 1 to 184 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
275 bits(703)	6e-87	Compositional matrix adjust.	142/184(77%)	157/184(85%)	0/184(0%)	-3

```
Query  34081  GVPVLILANKQDLPNACGAMELEKLLGLNELYNPVPNISMLTCSDSSSTINLIGCSKSNQ  33902
              GVPVLILANKQDLPNACGAMELEKLLGLNELYNPVPNISM + SDSS TINLIGC  SNQ
Sbjct  1      GVPVLILANKQDLPNACGAMELEKLLGLNELYNPVPNISMPSSSDSSPTINLIGCRVSNQ  60

Query  33901  SITETSLTEQRSNHLHSSMIHIKPAPESDEQRTTLSGEALSTFIYPHSGKGSDVKDQKNS  33722
              SIT+ SL E++ +HLHSSMIHIKPA ES +  +TLSG AL+ FIYP S    S V DQKN
Sbjct  61     SITDKSLEEKKESHLHSSMIHIKPALESKDHNSTLSGGALTAFIYPQSHNNSAVLDQKNP  120

Query  33721  RDGKNCFHNKKHNRTFSNSMHFRGWYIQPTCAITGEGLQEGLEALYDMILKRRKLNKSHK  33542
              +D KN FHNKK NR+ SNS+ FRGWYIQPTCAITGEGLQEGL+ALYDMILKRRK+NKS+K
Sbjct  121    QDVKNGFHNKKMNRSSSNSVQFRGWYIQPTCAITGEGLQEGLDALYDMILKRRKINKSNK  180

Query  33541  KKL*  33530
              + L*
Sbjct  181    RNL*  184
```

Figure 61: Alignment resulting from BLASTX search using the unmasked Contig9 sequence as the query and the 1_1762_0 peptide sequence as the subject.

In order to determine if the complete *Arl4* gene is located elsewhere in the *D. biarmipes* dot chromosome, the complete peptide sequence for *D. melanogaster Arl4-PB* was used to perform a FlyBase tBLASTn search against the genome assembly for *D. biarmipes*. Figure 62 shows that residues 129 to 300 of *Arl4-PB* align to two places in *D. biarmipes* with an e-value of 4.74166e-68 and a 73.3% identity. The alignments are exactly identical, which supports the hypothesis that the duplication of exon 1_1762_0 is a recent event.



>gi|459197948|gb|KB462564.1| Drosophila biarmipes unplaced genomic scaffold scf7180000302087, whole genome shotgun sequence
Length = 375228

HSP # = 1 , Score = 256.914 bits (655) , Expect = 4.74166e-68
Identities = 126 / 172 (73.3%) , Positives = 139 / 172 (80.8%)
Frame = +1

Subject FASTA

```
Query:   129          NQGVPVLILANKQDLPNACGAMELEKLLGLNELYNPVPNIXXXXXXXXXXTINLIGCRVS   188
                      ++GVPVLILANKQDLPNACGAMELEKLLGLNELYNPVPNI          TINLIGC  S
Subject: 139330       SKGVPVLILANKQDLPNACGAMELEKLLGLNELYNPVPNISMLTCSDSSSTINLIGCSKS   139509

Query:   189          NQSITDKSLEEKKESHLHSSMIHIKPALESKDHNSTLSGGALTAFIYPQSHNNSAVLDQK   248
                      NQSIT+ SL E++ +HLHSSMIHIKPA ES +  +TLSG AL+ FIYP S   S V DQK
Subject: 139510       NQSITETSLTEQRSNHLHSSMIHIKPAPESDEQRTTLSGEALSTFIYPHSGKGSDVKDQK   139689

Query:   249          NPQDVKNGFHNKKMNRSSSNSVQFRGWYIQPTCAITGEGLQEGLDALYDMIL   300
                      N +D KN FHNKK NR+ SNS+ FRGWYIQPTCAITGEGLQEGL+ALYDMIL
Subject: 139690       NSRDGKNCFHNKKHNRTFSNSMHFRGWYIQPTCAITGEGLQEGLEALYDMIL   139845
```

>gi|459197948|gb|KB462564.1| Drosophila biarmipes unplaced genomic scaffold scf7180000302087, whole genome shotgun sequence

HSP # = 2 , Score = 256.914 bits (655) , Expect = 4.74166e-68
Identities = 126 / 172 (73.3%) , Positives = 139 / 172 (80.8%)
Frame = +1

Subject FASTA

```
Query:   129          NQGVPVLILANKQDLPNACGAMELEKLLGLNELYNPVPNIXXXXXXXXXXTINLIGCRVS   188
                      ++GVPVLILANKQDLPNACGAMELEKLLGLNELYNPVPNI          TINLIGC  S
Subject: 185977       SKGVPVLILANKQDLPNACGAMELEKLLGLNELYNPVPNISMLTCSDSSSTINLIGCSKS   186156

Query:   189          NQSITDKSLEEKKESHLHSSMIHIKPALESKDHNSTLSGGALTAFIYPQSHNNSAVLDQK   248
                      NQSIT+ SL E++ +HLHSSMIHIKPA ES +  +TLSG AL+ FIYP S   S V DQK
Subject: 186157       NQSITETSLTEQRSNHLHSSMIHIKPAPESDEQRTTLSGEALSTFIYPHSGKGSDVKDQK   186336

Query:   249          NPQDVKNGFHNKKMNRSSSNSVQFRGWYIQPTCAITGEGLQEGLDALYDMIL   300
                      N +D KN FHNKK NR+ SNS+ FRGWYIQPTCAITGEGLQEGL+ALYDMIL
Subject: 186337       NSRDGKNCFHNKKHNRTFSNSMHFRGWYIQPTCAITGEGLQEGLEALYDMIL   186492
```

Figure 62: tBLASTn alignment of peptide sequence for *Arl4-PB* (query) with *D. biarmipes* nucleotide assembly (subject).

To further investigate the location of *Arl4* in *D. biarmipes* and the hypothesis that part of *Arl4* is duplicated, BLAT was used to search the *Arl4-PB* protein sequence against the *D. biarmipes* April 2013 (BCM-HGSC) Assembly (DbioWGS2) (Figure 63). As expected, *Arl4-PB* aligns in two distinct locations. Part of *Arl4-PB* aligns immediately to the right of *CG32006*, while the full *Arl4-PB* gene appears to be located further downstream. In order to investigate if this duplication is present in other *Drosophila* species, the alignment nets for *D. melanogaster*, *D. erecta*, *D. ficusphila*, *D. eugracilis*, *D. takahashii* and *D. ananassae* were selected. Based on the alignment nets and tBlastn searches against *Drosophila* species genomes, the duplication does not seem to exist in nearby species. Due to the low quality of the *D. takahashii* alignment net, it is difficult to say if the duplication exists in *D. takahashii*.

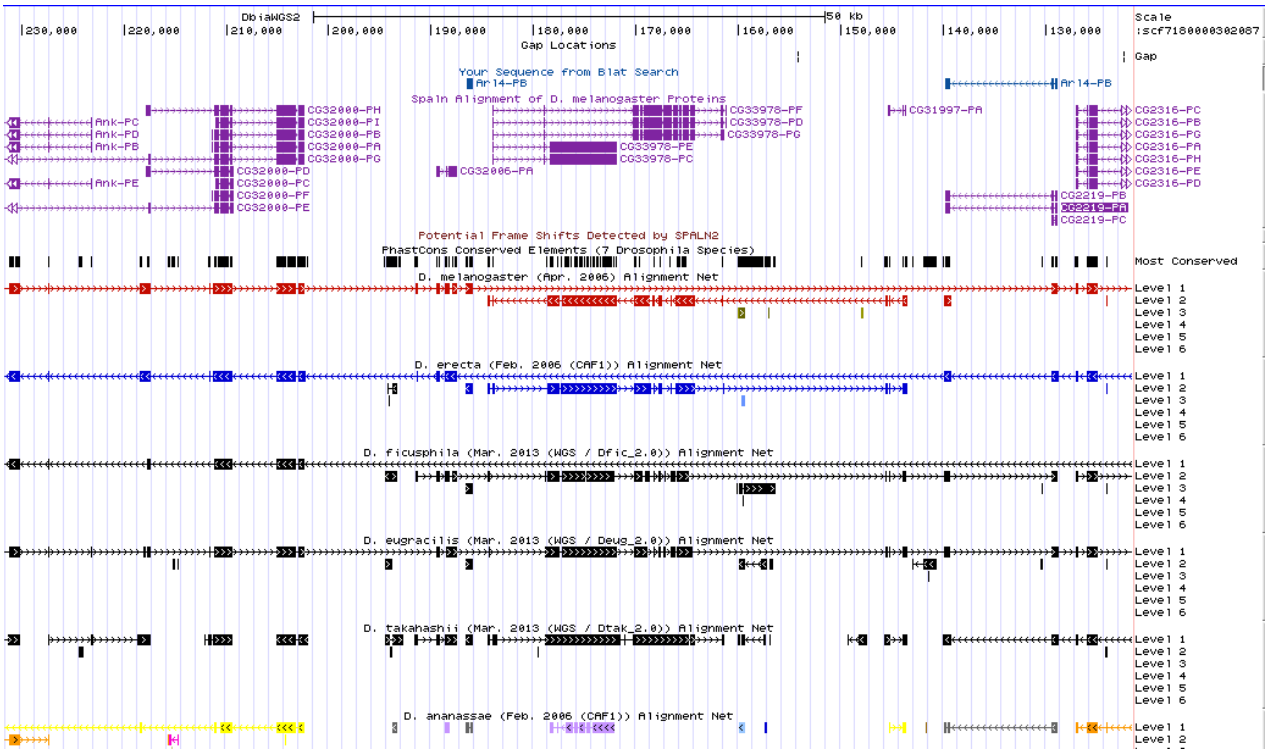


Figure 63: Results of BLAT search of the *Arl4-PB* protein sequence against the *D. biarmipes* April 2013 (BCM-HGSC) Assembly. Alignment nets for other *Drosophila* species are shown.

Overall, Genscan Feature 4 likely corresponds to a paralog of *Arl4*. Based on a comparison with nearby *Drosophila* species and the complete conservation between the paralog and the actual *Arl4* gene (see Fig. 62), the paralog likely arose from a recent duplication event. The mechanism by which the duplication arose is unknown; however there is a large "unknown" repeat adjacent to the paralog (see Figure 57) that could be involved. Attempts to classify this inverted repeat were unsuccessful—searching the DNA sequence against the Genetic Information Research Institute (GIRI) database suggested that the repeat is conserved in *D. takahashii*, but it is not well characterized (Figure 64). At this point, there is not enough data to reject the hypothesis that this paralog is a functional gene, as there is an ORF and RNA-Seq data in the region (which could be misplaced and could actually belong to the complete *Arl4* gene). Further evidence is necessary to classify this putative *Arl4* paralog.

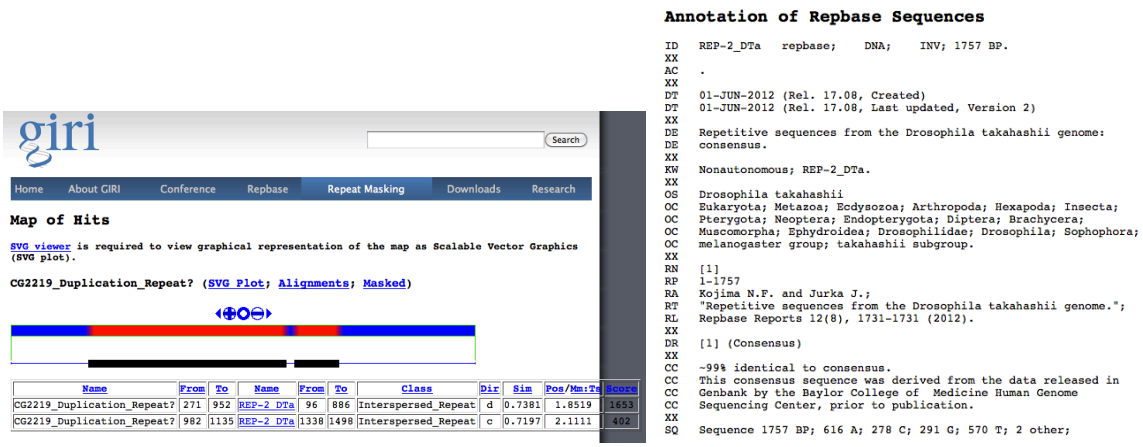


Figure 64: GIRI annotation information for repeat adjacent to the proposed *Arl4* paralog.

Gene *CG33978:*

Genscan Features 5 and 6 appear to align to various isoforms of *CG33978* in *D. melanogaster* based on the BLASTX alignment shown in Figure 65. The first step was to use the predicted protein sequence for Genscan Feature 5 to perform a FlyBase BLASTp search against all annotated proteins for *D. melanogaster*. Figure 66 shows the top alignments generated by the BLASTp search; the best match is clearly to *CG33978*. Genscan Features 5 and 6 are likely orthologous to *D. melanogaster* gene *CG33978*.

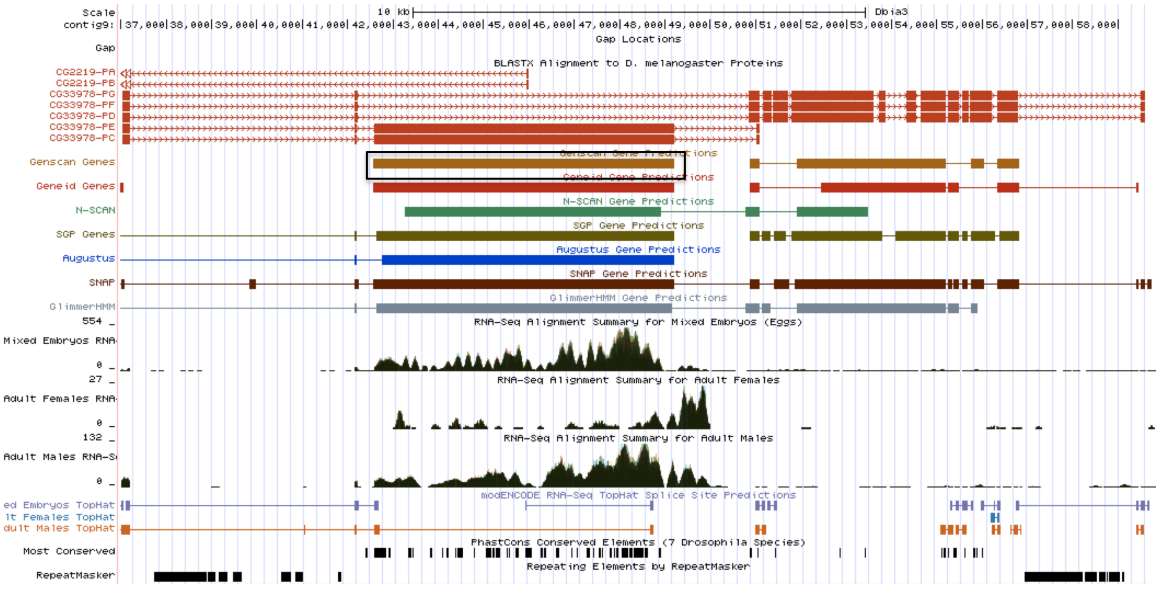


Figure 65: Genscan Feature 5 (boxed in black) and Genscan Feature 6 on GEP UCSC Genome Browser view of Contig9 of *D. biarmipes* (August 2013).

BLAST Hit Summary				
	Description	Species	Score	E value
☑	CG33978-PC	Dmel	2168.27	0
☑	CG33978-PE	Dmel	2168.27	0
☑	dp-PR	Dmel	59.3066	4.7017e-08
☑	dp-PT	Dmel	59.3066	4.78081e-08

Figure 66: BLAST Hit Summary from BLASTp search aligning the predicted protein sequence for Genscan Feature 5 against all annotated proteins for *D. melanogaster*.

According to the GEP Gene Record Finder, *CG33978* (FlyBase ID FBgn0053978) has five isoforms in *D. melanogaster* (C, D, E, F and G) and fifteen translated exons. Table 6a shows the exons for all five isoforms of *CG33978* in *D. melanogaster* and Table 6b summarizes the location of each exon in *D. melanogaster*.

Isoform	18_13859_0	16_13859_2	13_1759_2	14_13859_2	12_13859_0	10_13859_0	8_13859_2	6_13859_0	4_13859_2	2_13859_0	5_1759_1	0_13859_1	3_1759_0	2_1759_0	1_1759_0
CG33978-RD	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y			Y	
CG33978-RF	Y	Y		Y	Y	Y	Y	Y	Y	Y		Y		Y	
CG33978-RG	Y	Y		Y	Y	Y	Y	Y	Y	Y		Y	Y		Y
CG33978-RC	Y	Y	Y												
CG33978-RE	Y	Y	Y												

Table 6a: Summary of all five isoforms of *CG33978* and their respective exons in *D. melanogaster*.

FlyBase_ID	Dmel_chrom	Dmel_start	Dmel_end	Dmel_strand
CDS_CG33978:18_13859_0	4	195411	195585	-
CDS_CG33978:16_13859_2	4	194737	194790	-
CDS_FBgn0053978:13_1759_2	4	187722	194362	-
CDS_CG33978:14_13859_2	4	185341	185690	-
CDS_CG33978:12_13859_0	4	185088	185282	-
CDS_CG33978:10_13859_0	4	181448	185021	-
CDS_CG33978:8_13859_2	4	181151	181389	-
CDS_CG33978:6_13859_0	4	180958	181087	-
CDS_CG33978:4_13859_2	4	180393	180901	-
CDS_CG33978:2_13859_0	4	179848	180338	-
CDS_FBgn0053978:5_1759_1	4	179270	179330	-
CDS_CG33978:0_13859_1	4	179270	179288	-
CDS_FBgn0053978:3_1759_0	4	179126	179203	-
CDS_FBgn0053978:2_1759_0	4	179084	179203	-
CDS_FBgn0053978:1_1759_0	4	179045	179068	-

Table 6b: FlyBase ID number and location for the fifteen translated exons for *CG33978* in *D. melanogaster*.

Figure 67 shows the FlyBase GBrowse2 view of the *CG33978* transcript in *D. melanogaster*. *CG33978* is inverted in *D. biarmipes* relative to its orientation in *D. melanogaster*. This potential inversion will be discussed in further detail in the "**SYNTENY**" section.

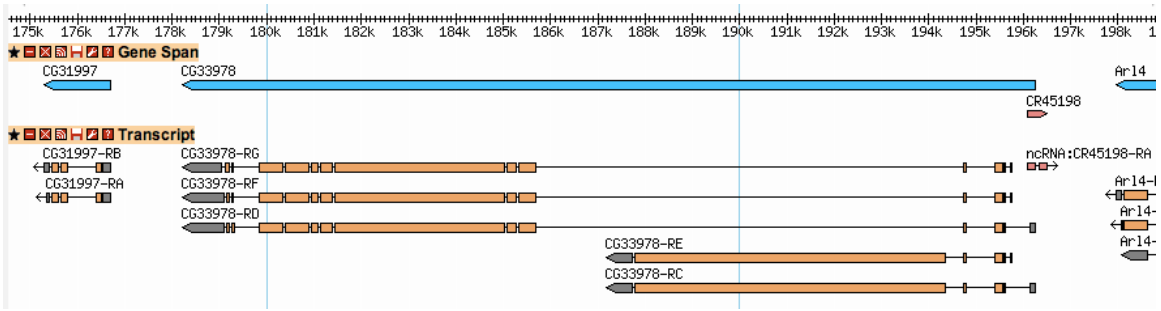


Figure 67: FlyBase GBrowse2 view showing the transcript details of *CG33978* in *D. melanogaster*. Gray boxes represent untranslated regions and orange boxes represent translated exons.

The next step was to confirm that Contig9 of *D. biarmipes* contains the entire CG33978 protein sequence. Peptide sequences for each translated exon were obtained through the GEP Gene Record Finder. For each exon, a BLASTX search was performed comparing the *D. melanogaster* CG33978 peptide sequence to the unmasked Contig9 sequence. Table 7 summarizes the results of the BLASTX searches.

Subject	Size (aa)	Subject Start	Subject End	Query	Query Start	Query End	E-value	% Identity	Frame
18_13859_0	58	1	58	Contig9	36021	36194	2e-06	36%	+ 3
16_13859_2	17	1	17	Contig9	41156	41206	0.023	59%	+ 2
13_1759_2	2213	1	2213	Contig9	41592	48218	0	62%	+ 3
14_13859_2	116	6	116	Contig9	49786	50088	3e -25	50%	+ 1
12_13859_0	65	5	65	Contig9	50158	50340	5e-17	56%	+ 1
10_13859_0	1191	1	100	Contig9	50405	50710	5e-164	47%	+ 2
		128	1191		50791	54198	5e-164	35%	+ 1
8_13859_2	79	1	79	Contig9	54258	54494	2e-28	63%	+ 3
6_13859_0	43	1	43	Contig9	54562	54690	4e-16	72%	+ 1
4_13859_2	169	1	169	Contig9	54754	55275	5e-50	57%	+ 1
2_13859_0	163	1	159	Contig9	55332	55793	1e-23	36%	+ 3
5_1759_1	20	No significant alignments.		Contig9					
0_13859_1	6	No significant alignments.		Contig9					
3_1759_0	26	1	26	Contig9	58509	58586	1e-08	69%	+ 3
2_1759_0	40	1	39	Contig9	58509	58625	4e-10	56%	+ 3
1_1759_0	8	No significant alignments.		Contig9					

Table 7: Results of BLASTX searches comparing the translated exon peptide sequences for CG33978 (subject) to the unmasked Contig9 sequence (query).

The results of the BLASTX searches provide strong evidence that this feature is orthologous to *CG33978* in *D. melanogaster*. The next step was to use the Genome Browser to refine the annotation of each exon using RNA-Seq, TopHat, Oases, Cufflinks and other gene predictors. Several exons proved somewhat difficult to annotate. As shown in Table 7, 10_13859_0 aligned to Contig9 in two separate sections, one in the +2 frame and one in the +1

frame. Moreover, three small exons, 5_1759_1, 0_13859_1 and 1_1759_0, could not be located using BLASTX. The annotation process for these four more complicated exons will be described in detail in the following sections.

Exon 10_13859_0:

BLASTX predicts that 10_13859_0 aligns to Contig9 in frame +2 from bases 50405 to 50710 (Figure 68, boxed in black on the left) and in frame +1 from bases 50791 to 54198 (Figure 68, boxed in black on the right). For the alignment from bases 50791 to 54198, the BLASTX alignment on the genome browser incorrectly breaks the alignment into four separate exons. There are no stop codons in frame +1 in this region, and gene predictors such as Genscan, SNAP and GlimmerHMM correctly predict that it should be a single exon.

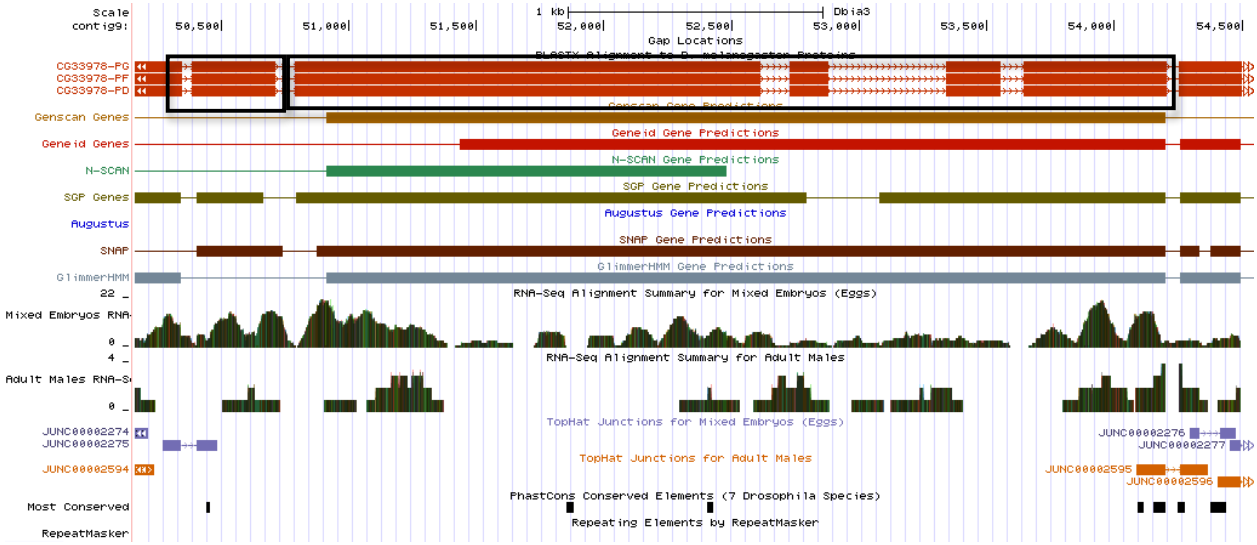


Figure 68: BLASTX predicts that 10_13859_0 aligns to Contig9 in frame +2 from bases 50405 to 50710 (boxed in black on the left) and in frame +1 from bases 50791 to 54198 (boxed in black on the right).

The next question is whether 10_13859_0 exists as two separate exons in different frames in *D. biarmipes* as predicted by the initial BLASTX search. There are no TopHat junctions in the region from bases 50405 to 54198, and there is RNA-Seq coverage over most of the region. Also, residues 101 to 127 of 10_13859_0 are missing from the BLASTX alignment and should be located in this region. Examining the gap from bases 50711 to 50790 between the two alignments reveals a mononucleotide run of A's (Figure 69, boxed in red). Mononucleotide runs frequently result in sequencing errors, so I decided to examine the two mRNA reads that cover

this mononucleotide run (Figure 70). Both reads have one more "A" base in the mononucleotide run than the Contig9 consensus sequence.

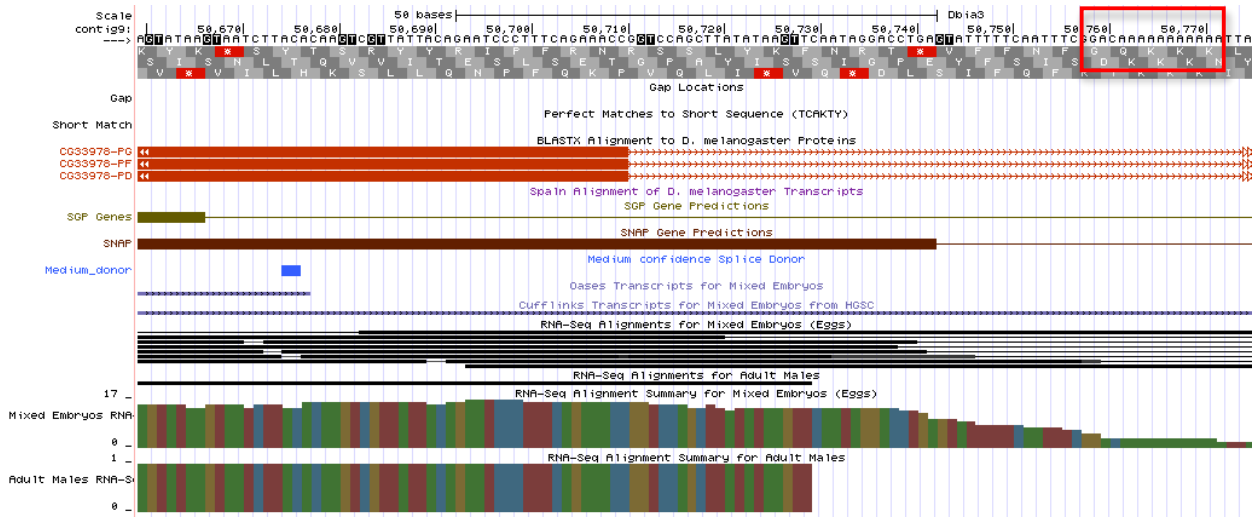
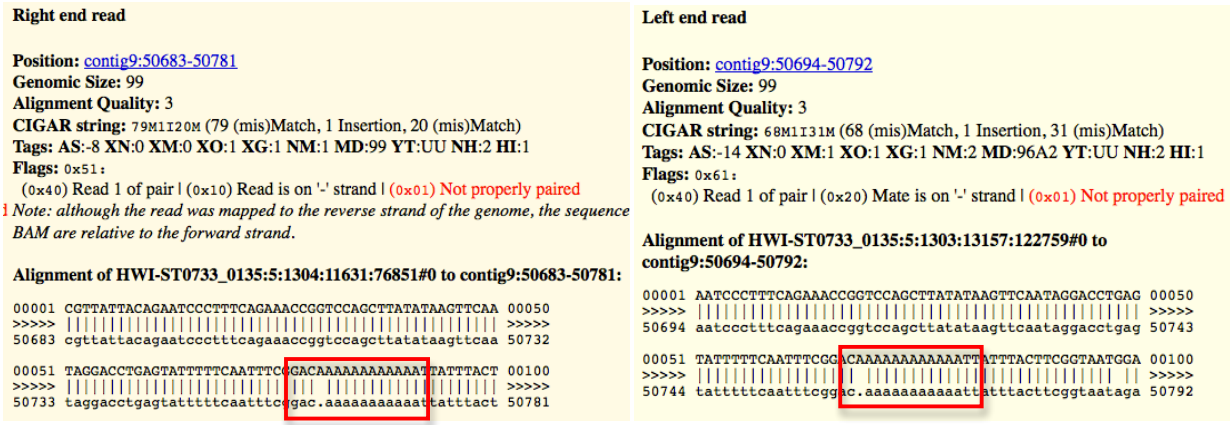


Figure 69: End of alignment of 10_13859_0 to contig9 in frame +2 at base 50710. To the right of the alignment, there is a mononucleotide run of A's that is boxed in red.



Right end read

Position: contig9:50683-50781
Genomic Size: 99
Alignment Quality: 3
CIGAR string: 79M1I20M (79 (mis)Match, 1 Insertion, 20 (mis)Match)
Tags: **AS**:-8 **XN**:0 **XM**:0 **XO**:1 **XG**:1 **NM**:1 **MD**:99 **YT**:UU **NH**:2 **HI**:1
Flags: 0x51:
(0x40) Read 1 of pair | (0x10) Read is on '-' strand | (0x01) Not properly paired
Note: although the read was mapped to the reverse strand of the genome, the sequence BAM are relative to the forward strand.

Alignment of HWI-ST0733_0135:5:1304:11631:76851#0 to contig9:50683-50781:

```
00001 CGTTATTACAGAATCCCTTTCAGAAACCGGTCCAGCTTATATAAGTTCAA 00050
>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>
50683 cgttattacagaatccctttcagaaaccggtccagcttatataagttcaa 50732

00051 TAGGACCTGAGTATTTTTCAATTTCGGACAAAAAAAAAAAATTATTTACT 00100
>>>>> ||||||||||||||||||||||||| ||| |||||||||||||||||| >>>>>
50733 taggacctgagtatttttcaatttcggac.aaaaaaaaaaattatttact 50781
```

Left end read

Position: contig9:50694-50792
Genomic Size: 99
Alignment Quality: 3
CIGAR string: 68M1I31M (68 (mis)Match, 1 Insertion, 31 (mis)Match)
Tags: **AS**:-14 **XN**:0 **XM**:1 **XO**:1 **XG**:1 **NM**:2 **MD**:96A2 **YT**:UU **NH**:2 **HI**:1
Flags: 0x61:
(0x40) Read 1 of pair | (0x20) Mate is on '-' strand | (0x01) Not properly paired

Alignment of HWI-ST0733_0135:5:1303:13157:122759#0 to contig9:50694-50792:

```
00001 AATCCCTTTCAGAAACCGGTCCAGCTTATATAAGTTCAATAGGACCTGAG 00050
>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>
50694 aatcctttcagaaaccggtccagcttatataagttcaataggacctgag 50743

00051 TATTTTTCAATTTCGGACAAAAAAAAAAAATTATTTACTTCGGTAATGGA 00100
>>>>> |||||||||||||||| |||||||||||||||||||||||||| >>>>>
50744 tatttttcaatttcggac.aaaaaaaaaaattatttacttcggtaataga 50792
```

Figure 70: Alignment of two mRNA reads to Contig9 in the region spanning the mononucleotide run of A's. For both alignments, the mRNA sequence is on the top and the Contig9 consensus sequence is on the bottom. The additional A in the mRNA sequences is boxed in red.

Based on this evidence, the Sequence Updater was used to create a VCF file that adds an additional A to the contig9 consensus sequence at base 50762. A BLASTX search comparing 10_13859_0 (subject) to the updated unmasked Contig9 sequence (query) showed that residues 1 to 1191 of 10_13859_0 align to bases 50405 to 54199 of Contig9 in frame +2 with a 36% identity. Exon 10_13859_0 is thus present as a single exon in Contig9.

Exon 5_1759_1 and 0_13859_1:

Exons 5_1759_1 and 0_13859_1 overlap in *D. melanogaster*; 5_1759_1 is 20 amino acids long while 0_13859_1 is only 6 amino acids long. Both end at the same base. Small Exons

Finder located several candidate exons for both, but none were plausible. Figure 71 shows the region of Contig9 where I expect to find 5_1759_1 and 0_13859_1,which is between 2_13859_0 (ends at base 55813) and 3_1759_0/2_1759_0 (start at base 58506). There is likely spurious RNA-Seq coverage in the region due to the presence of repeating elements. SNAP gene predictor and TopHat JUNC00002281 and JUNC00002282 both suggest the presence of an exon at around base 58400.

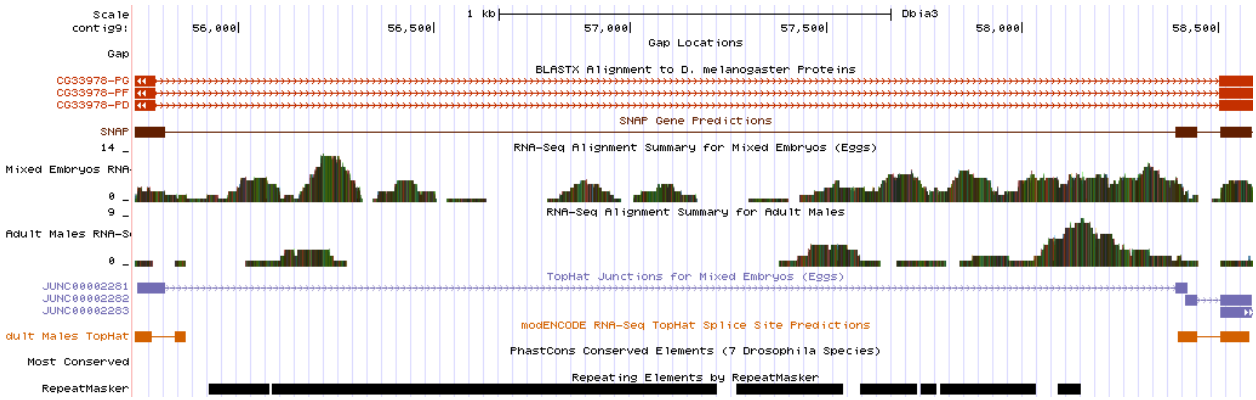


Figure 71: Expected location of 5_1759_1 and 0_13859_1, which is between 2_13859_0 (ends at base 55813) and 3_1759_0/2_1759_0 (start at base 58506).

Figure 72 shows a zoomed in view of the proposed location for 5_1759_1 and 0_13859_1 on Contig9. Both exons are in the +1 frame, and "AG" splice acceptor sites were picked to give phase 1 (based on the annotation of the connecting exon). The "AG" splice acceptor site for 5_1759_1 is located at bases 58390 to 58391, and the actual exon starts at base 58392. The acceptor site for 0_13859_1 is located at bases 58432 to 58433, and the actual exon starts at base 58434. Both exons end at base 58446, and there is a "GT" splice donor site at bases 58447 to 58448.

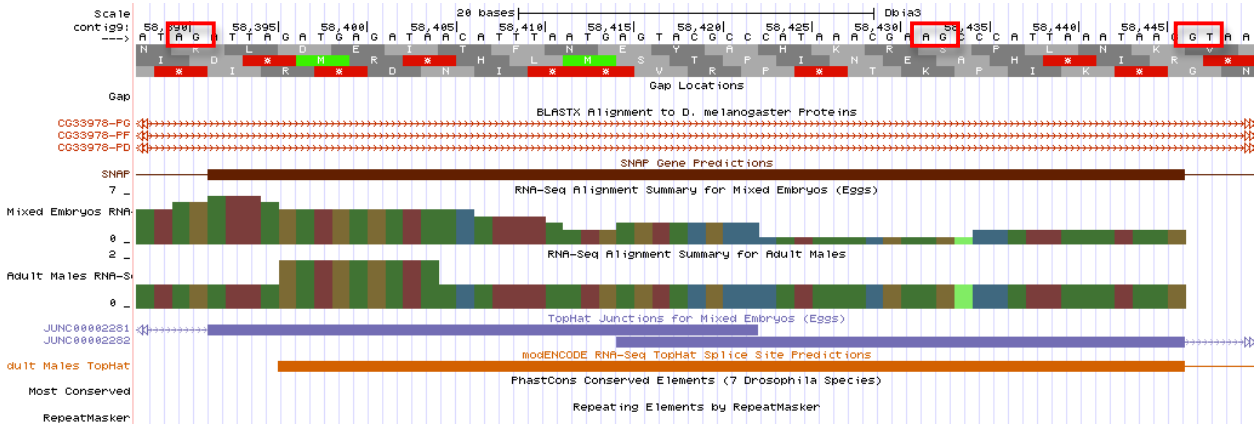


Figure 72: Proposed location of 5_1759_1 and 0_13859_1 on Contig9. The two splice acceptor sites and the splice donor site are boxed in red.

Exon 1_1759_0:

In *D. melanogaster*, exon 1_1759_0 is the last possible terminal exon. I expected to find 1_1759_0 downstream of 3_1759_0, which ends at base 58586. Examining the genome browser suggests that 1_1759_0 is located at bases 58644 to 58679 in the +3 frame (Figure 73). The exon is supported by SNAP gene predictor, RNA-Seq data, and TopHat JUNC00002283. There is an "AG" splice acceptor site at bases 58642 to 58643, and a terminal exon at bases 58677 to 58679.

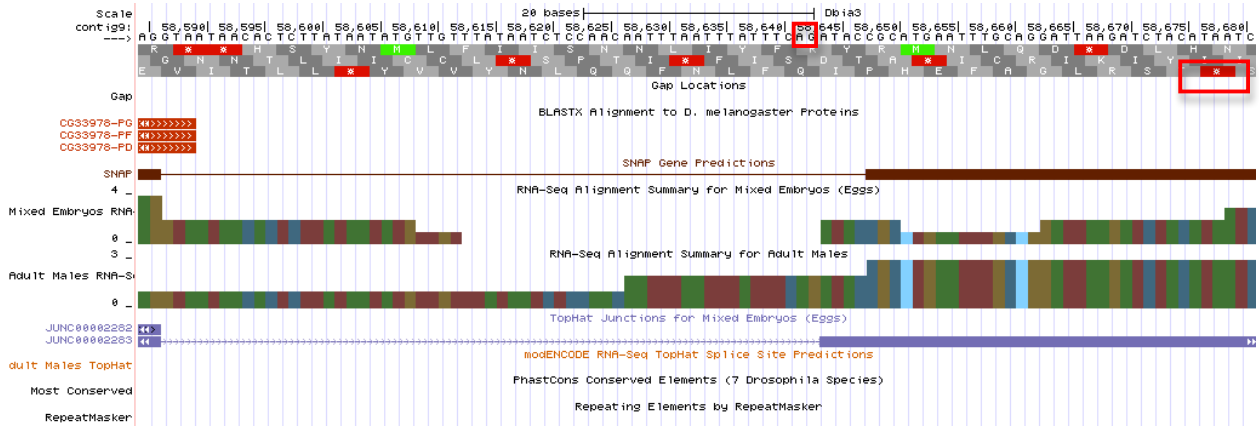


Figure 73: Location of terminal exon 1_1759_0. "AG" splice acceptor site and terminal exon are boxed in red.

Table of Exons:

Table 8 summarizes the location of all exons for *CG33978* in both *D. melanogaster* and *D. biarmipes*. 18_13859_0 is the initial exon for all five isoforms. There are three terminal exons: 13_1759_2 (stop codon at bases 48216 to 48218), 2_1759_0 (stop codon at bases 58602 to 58604) and 1_1759_0 (stop codon at bases 58677-58679).

D. melanogaster				**Contig9 of *D. biarmipes* Dot chromosome**					
FlyBase_ID	Start	End	Length (bp)	Frame	Start	Phase	End	Phase	Length (bp)
CDS_CG33978:18_13859_0	195411	195585	175	3	36021	0	36195	1	175
CDS_CG33978:16_13859_2	194737	194790	54	2	41154	2	41207	1	54
CDS_FBgn0053978:13_1759_2	187722	194362	6641	3	41590	1	48218	0	6629
CDS_CG33978:14_13859_2	185341	185690	350	1	49775	2	50088	0	314
CDS_CG33978:12_13859_0	185088	185282	195	1	50146	0	50340	0	195
CDS_CG33978:10_13859_0	181448	185021	3574	2	50405	0	54199	Frame 1, 1	3795
CDS_CG33978:8_13859_2	181151	181389	239	3	54256	2	54494	0	239
CDS_CG33978:6_13859_0	180958	181087	130	1	54562	0	54691	1	130
CDS_CG33978:4_13859_2	180393	180901	509	1	54752	2	55275	0	524
CDS_CG33978:2_13859_0	179848	180338	491	3	55332	0	55813	2	482
CDS_FBgn0053978:5_1759_1	179270	179330	61	1	58392	1	58446	0	55
CDS_CG33978:0_13859_1	179270	179288	19	1	58434	1	58446	0	13
CDS_FBgn0053978:3_1759_0	179126	179203	78	3	58506	0	58586	0	81
CDS_FBgn0053978:2_1759_0	179084	179203	120	3	58506	0	58604	0	99

CDS_FBgn0053978:1_1759_0	179045	179068	24	3	58644	0	58679	0	36

Table 8: Table of exons for *CG33978*. The start and end points of each exon are shown for both *D. melanogaster* and *D. biarmipes*.

Gene Model Checker:

Gene Model Checker was used to confirm the proposed gene models for the following isoforms: CG33978-PC, CG33978-PD, CG33978-PE, CG33978-PF and CG33978-PG. Figure 74 shows the dot plot comparisons of the *D. biarmipes* gene models to *D. melanogaster*. The slope of the dot plots all looks correct, but there are gaps as is to be expected due to the low level of conservation of some of the exons. Gene Model Checker flagged several potential issues. First of all, exon 4_13859_2 has a non-canonical "GC" splice donor site. This site is supported by RNA-Seq and TopHat data (Figure 75). Secondly, the boundaries of some of the smaller exons that did not have BLASTX alignments and were identified using RNA-Seq, TopHat and gene predictors need to be investigated further.

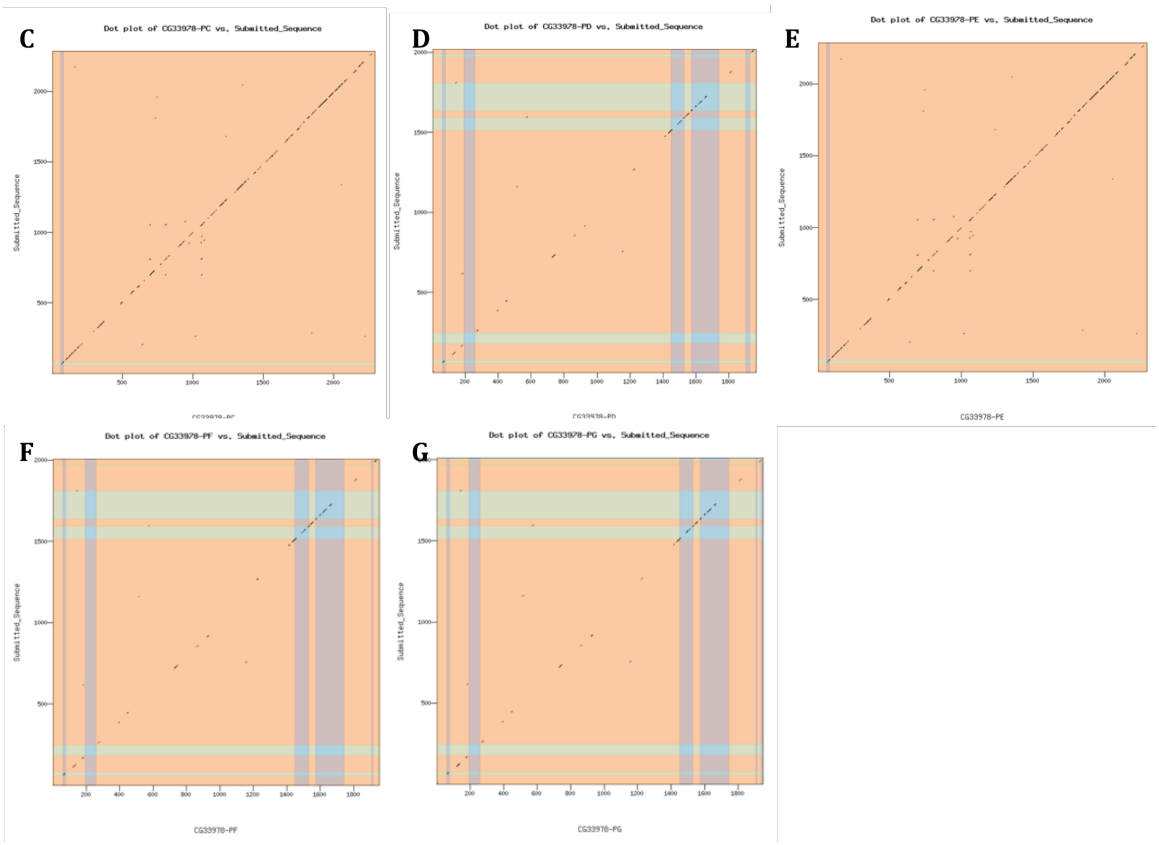


Figure 74: Dot plot results from Gene Model Checker for *CG33978*. Each dot plot is labeled with the isoform letter. The *D. biarmipes* sequence is on the vertical axis and the *D. melanogaster* sequence is on the horizontal axis.

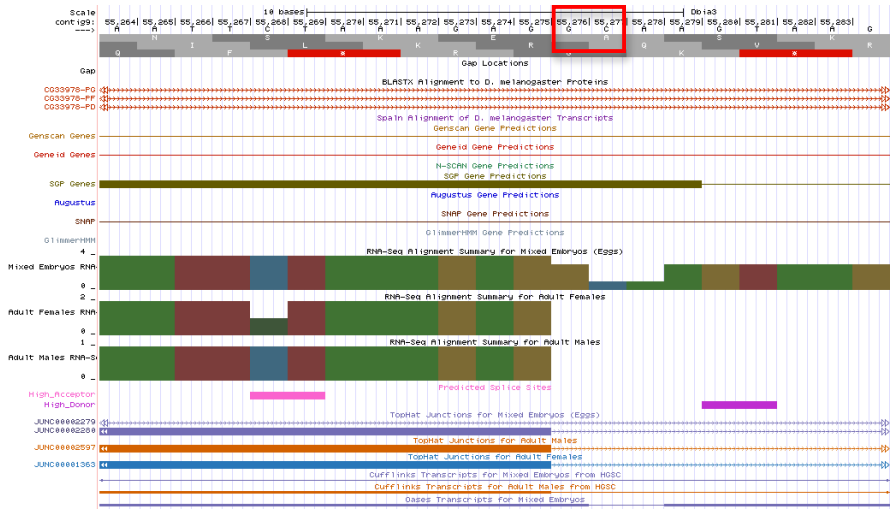


Figure 75: Non-canonical "GC" splice donor site for exon 4_13859_2 is boxed in red.

Untranslated Regions:

The 5' UTR was approximately annotated, but further investigation is required to determine the exact 5' UTR for each isoform. All five isoforms are predicted to have one of two possible untranslated exons upstream of the initial coding exon. Figure 76 shows the location of a potential 5' untranslated exon.

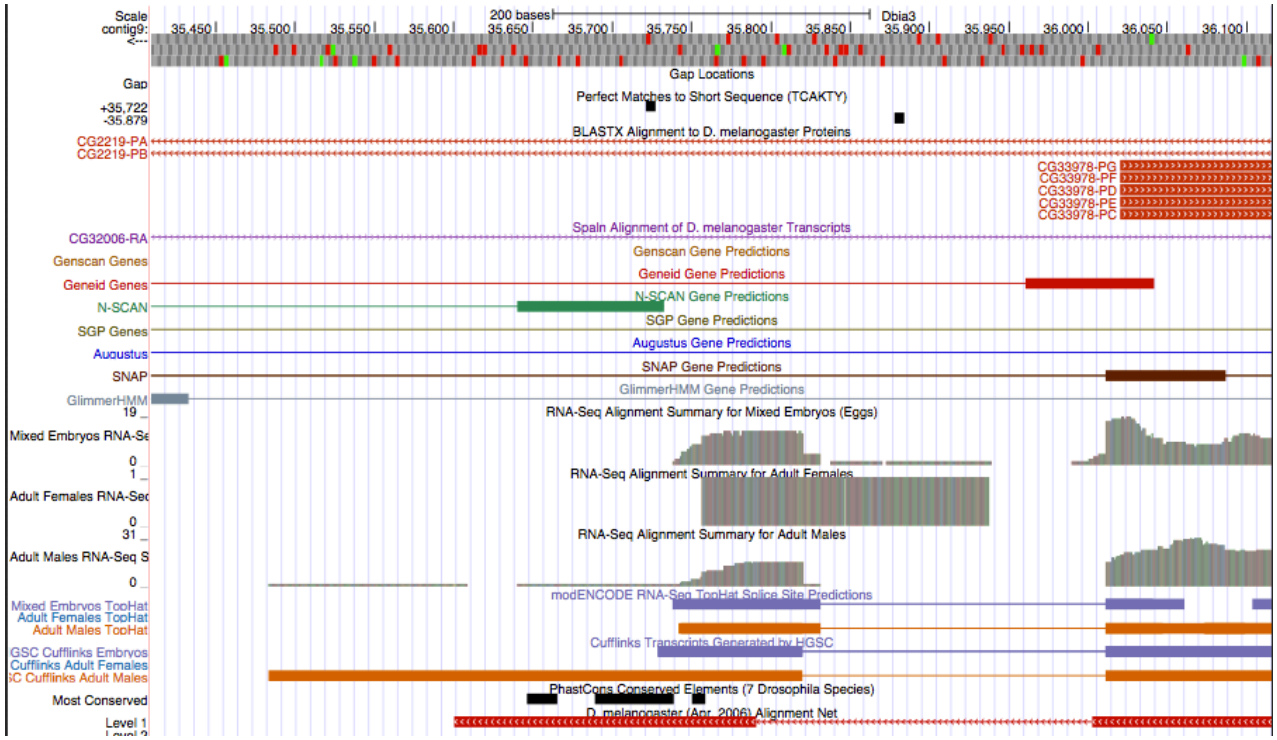


Figure 76: Possible 5' UTR of *CG33978* in *D. biarmipes*.

There is predicted initiator sequence at base 35722. RNA-Seq and TopHat suggest that the exon extends to approximately base 35831, and there is a “GT” potential splice donor site at bases 35832 to 35833. Moreover, there is a small 5’ UTR immediately adjacent to the initial coding exon (18_13859_0, starts at base 36021). There is an “AG” splice acceptor site at bases 36010 to 36011, and the 5’ UTR spans from base 36012 to 36020 as supported by RNA-Seq and TopHat.

The 3’ UTR for isoforms C and E begins at base 48219, immediately following terminal exon 13_1759_2. The extent of the 3’ UTR is unclear. A BLASTn search aligning the transcript for 13_1759_2 and the associated 3’ UTR (exon 13 on Gene Record Finder) to Contig9 produced an alignment that extended to base 48268 of Contig9, but the last 500 bases of exon 13 were missing from the alignment. Based on RNA-Seq data and Cufflinks Transcripts, the 3’ UTR probably extends to somewhere between bases 49000 to 49550 (Figure 77).

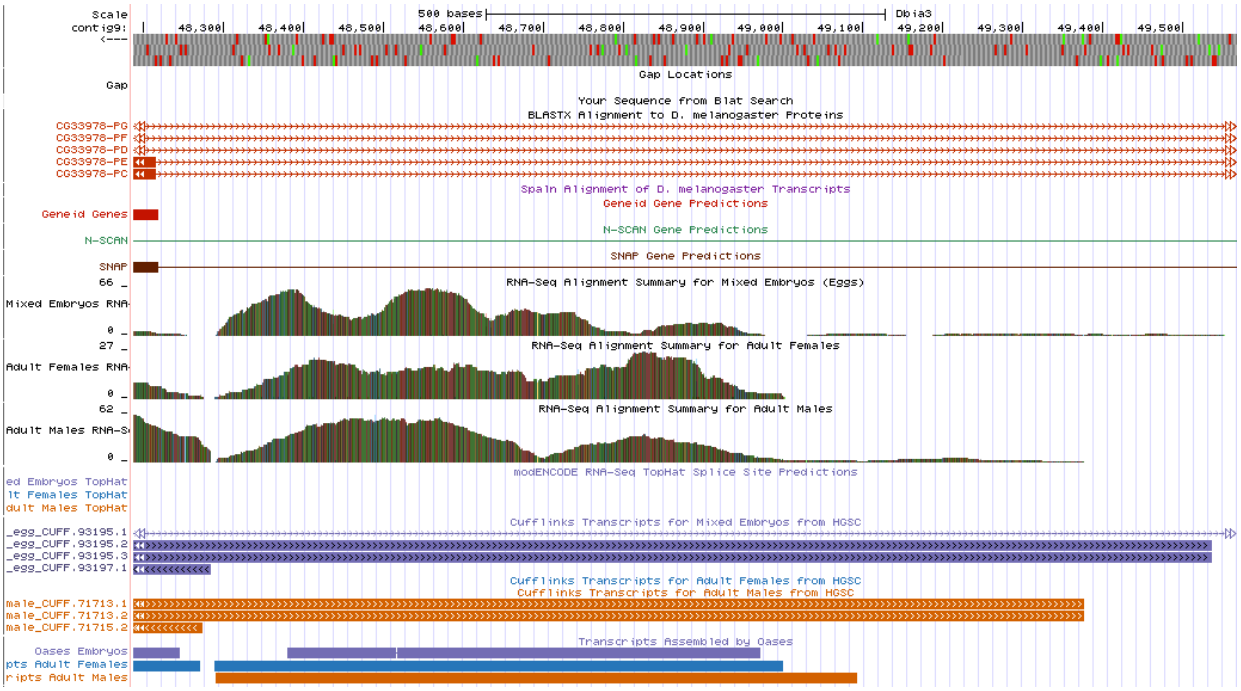


Figure 77: 3’ UTR for *CG33978-PC* and *CG33978-PE*.

In *D. melanogaster*, the 3’ UTR for isoforms D and F starts immediately after the terminal coding exon, 2_1759_0. This 3’ UTR includes 1_1759_0, the terminal coding exon for isoform G, as well as the 3’ UTR for isoform G. A BLASTn search aligning the transcript for the 3’ UTR to Contig9 was inconclusive, but an approximate annotation can be made based on RNA-Seq, Cufflinks and Oases (Figure 78). The 3’ UTR for isoforms D and F starts at base

58605 and extends to approximately base 58992. The 3' UTR for isoform G is located at bases 58680 to 58992. These annotations are tentative and require further evidence to confirm.

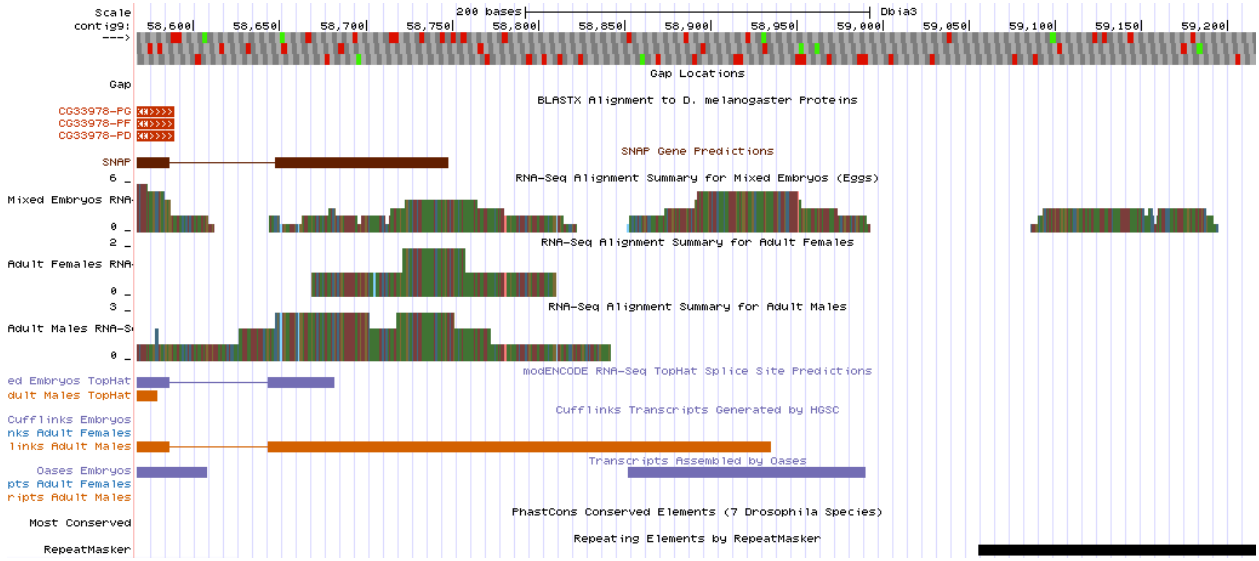


Figure 78: 3' UTR for *CG33978-PD*, *CG33978-PF* and *CG33978-PG*.

General Comments:

In *D. melanogaster*, *CR45198-RA* is a non-coding RNA that overlaps with *CG33978*. A BLASTn search aligning the nucleotide sequence for *CR45198-RA* (query) against Contig9 (subject) was inconclusive. Further investigations are necessary to confirm the existence of CR45198-RA on Contig9.

The region between *CG3200* and *CG32006* on Contig9 (approximately bases 17500 to 31000) contains many repeating elements and does not have any BLASTX alignments (see Fig. 1). Genscan Feature 2 is in this region. To confirm that this region does not contain any genes, I performed a FlyBase BLASTX search comparing the nucleotide sequence for this region (query) against *D. melanogaster* annotated proteins (subject). There were no significant alignments. I also performed a FlyBase BLASTp search comparing the Genscan Feature 2 peptide sequence (query) against *D. melanogaster* annotated proteins (subject), which did not produce any significant alignments. Based on these searches as well as synteny with *D. melanogaster*, I am fairly confident that this region does not contain any genes.

CLUSTAL:

I performed a Clustal analysis on *CG32000* in order to investigate its conservation across multiple *Drosophila* species. For the comparison, I used the peptide sequence for CG32000-PH

in *D. biarmipes* and *D. melanogaster* because it is the longest isoform. For *D. erecta, D. yakuba, D. pseudoobscura* and *D. mojavensis,* I used the peptide sequence for the CG32000 ortholog available through FlyBase. These annotations are the result of gene predictor models, and are not hand-curated or isoform-specific. However, they are adequate for the purpose of a simple comparison to identify conserved domains.

The Clustal analysis revealed that there are regions of high similarity interspersed with regions of weak similarity (see Figure 79 for an example). These regions of highly similarity are likely conserved domains that are essential for the protein function, whereas the regions of weak similarity probably encoded nonessential peptides. Highly conserved regions include bases 170-235, 290-300, 309-407, 436-734, 755-935, 968-978, 1017-1052 and 1080-1238 of the consensus sequence (ranges are approximate).

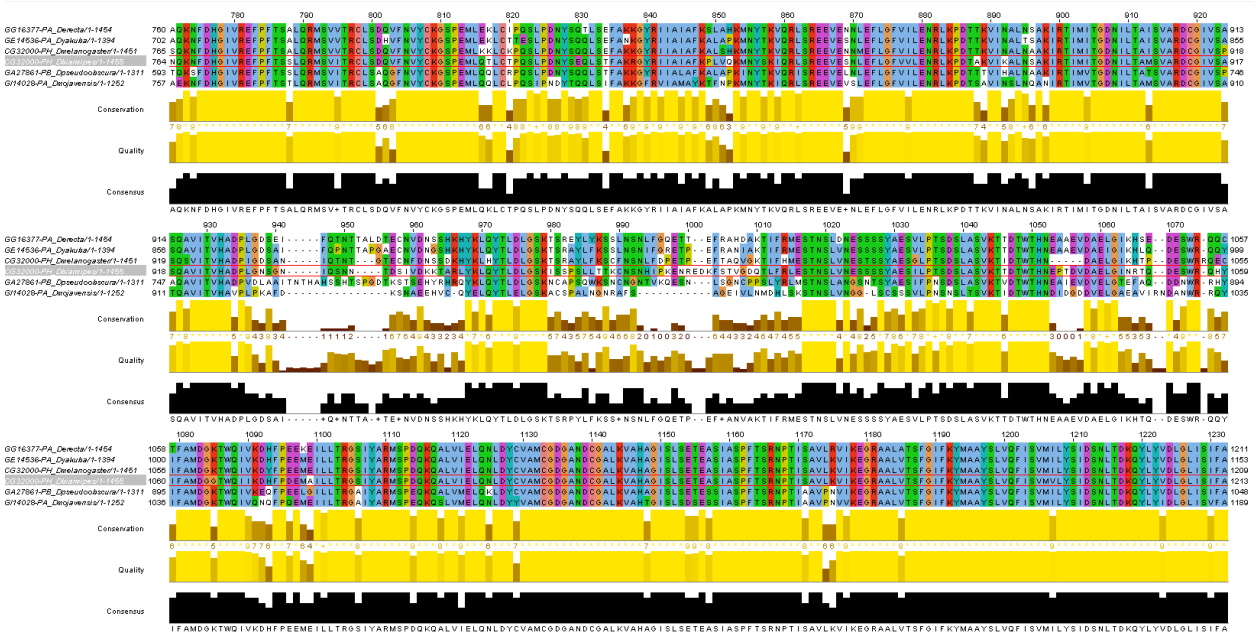


Figure 79: Jalview View of part of the Clustal analysis comparing *CG32000* orthologs from several *Drosophila* species.

EVOPRINTER:

EvoPrinterHD was used to analyze the putative 5' UTR of *CG33978* in order to gather further evidence for the annotation of the transcription start site. Bases 35474 to 36209, which include the 5' upstream region and initial coding exon, were used to perform the EvoPrinterHD analysis. Figure 80 shows the alignment of the sequence to the orthologous regions of the *D. erecta, D. yakuba* and *D. melanogaster* genomes. Black capital letters indicate conservation in all four species. A BLAT search was used to align the conserved region from bases 146 to 315 of

the alignment consensus (query) against the *D. biarmipes* dot chromosome (subject) (Figure 81). The alignment spans the initiator sequence identified by Short Match at base 35722. This provides further evidence that the transcription start site is located at around base 35722, and includes the sequence "TCAGTT."

Figure 80: EvoPrinterHD alignment of bases 35474 to 36209 of Contig9 to the orthologous regions of the *D. erecta*, *D. yakuba* and *D. melanogaster* genomes

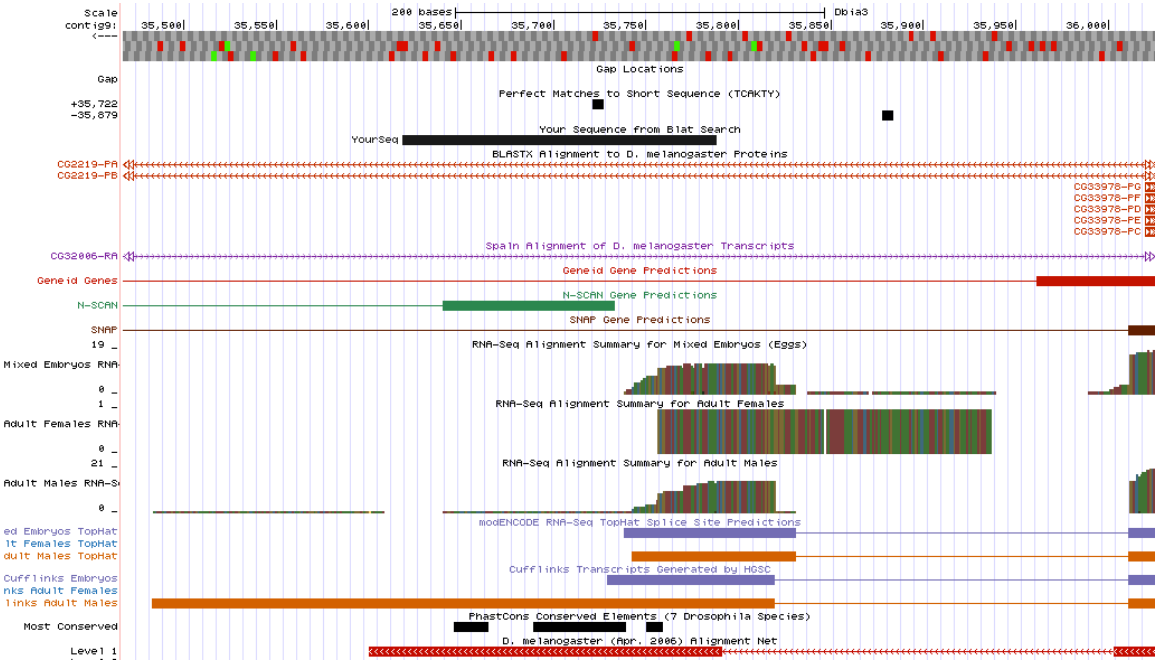


Figure 81: Alignment on Contig9 resulting from a BLAT search comparing the conserved region from bases 146 to 315 of the alignment (query) against the *D. biarmipes* dot chromosome (subject).

REPEATS:

Figure 82 shows the RepeatMasker analysis of the repeat content of Contig9. 29.39% of Contig9 is repetitive DNA. 3.33% is LINEs, 0.75% is LTR elements, 0.33% is DNA elements and 25.98% is unclassified. Most of the unclassified repeats belong to the "RC/Helitron" repeat class/family. Table 9 show the repeat class/family and location on Contig9 for all repeats longer than 500 bases.

```
==================================================
file name: contig9.fasta
sequences:             1
total length:      60018 bp  (60018 bp excl N/X-runs)
GC level:         33.82 %
bases masked:      17640 bp ( 29.39 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:                0            0 bp    0.00 %
      ALUs            0            0 bp    0.00 %
      MIRs            0            0 bp    0.00 %

LINEs:                3         2000 bp    3.33 %
      LINE1           0            0 bp    0.00 %
      LINE2           0            0 bp    0.00 %
      L3/CR1          1         1247 bp    2.08 %

LTR elements:         3          453 bp    0.75 %
      ERVL            0            0 bp    0.00 %
      ERVL-MaLRs      0            0 bp    0.00 %
      ERV_classI      0            0 bp    0.00 %
      ERV_classII     0            0 bp    0.00 %

DNA elements:         1          199 bp    0.33 %
     hAT-Charlie      0            0 bp    0.00 %
     TcMar-Tigger     0            0 bp    0.00 %

Unclassified:       101        15593 bp   25.98 %

Total interspersed repeats:    18245 bp   30.40 %


Small RNA:            0            0 bp    0.00 %

Satellites:           0            0 bp    0.00 %
Simple repeats:       0            0 bp    0.00 %
Low complexity:       0            0 bp    0.00 %
==================================================

* most repeats fragmented by insertions or deletions
  have been counted as one element
```

Figure 82: Repeat content of Contig9 as identified by RepeatMasker.

Matching repeat	Repeat class/family	Strand	Contig9 Start	Contig9 End	Size (bp)
rnd-1_family-9	LINE/CR1	+	18162	19408	1247
rnd-2_family-32	RC/Helitron	C	56218	57221	1004
rnd-4_family-146	RC/Helitron	C	29603	30320	718
rnd-4_family-146	RC/Helitron	C	37221	37880	660
rnd-5_family-8014	LINE/Jockey	+	26721	27347	627
rnd-4_family-146	RC/Helitron	+	56458	57062	605

Table 9: Repeat class/family and location on Contig9 for all repeats longer than 500 bases.

SYNTENY:

Figure 83 compares gene order and orientation between Contig9 of the *D. biarmipes* dot chromosome and the orthologous region in *D. melanogaster*. Due to potential inversions and rearrangements, it was necessary to understand the gene structure of *D. biarmipes* past the 3' end of Contig9. I also included a gene map for the region of the *D. biarmipes* April 2013 Assembly that corresponds to Contig9 and extends about 30,000 bases past the 3' end of Contig9. Synteny is only partially preserved between *D. biarmipes* and *D. melanogaster*. *CG32000* and *CG32006* are in the same order and orientation in both species. The putative *D. biarmipes Arl4* paralog does not exist in *D. melanogaster*, and it is in the opposite orientation compared to the actual *Arl4* gene. As discussed earlier, the presence of a large repeat element near the *Arl4* paralog suggest it may have arisen from a retro-transposition event. *CG33978* and *CG31997* (not on Contig9) are inverted in *D. biarmipes* relative to *D. melanogaster*. Based on the alignment nets with other *Drosophila* species in Figure 63, the inversion of *CG33978* may exist in *D. eugracilis* as well. The actual *Arl4* gene is located in the same order and orientation in both species.

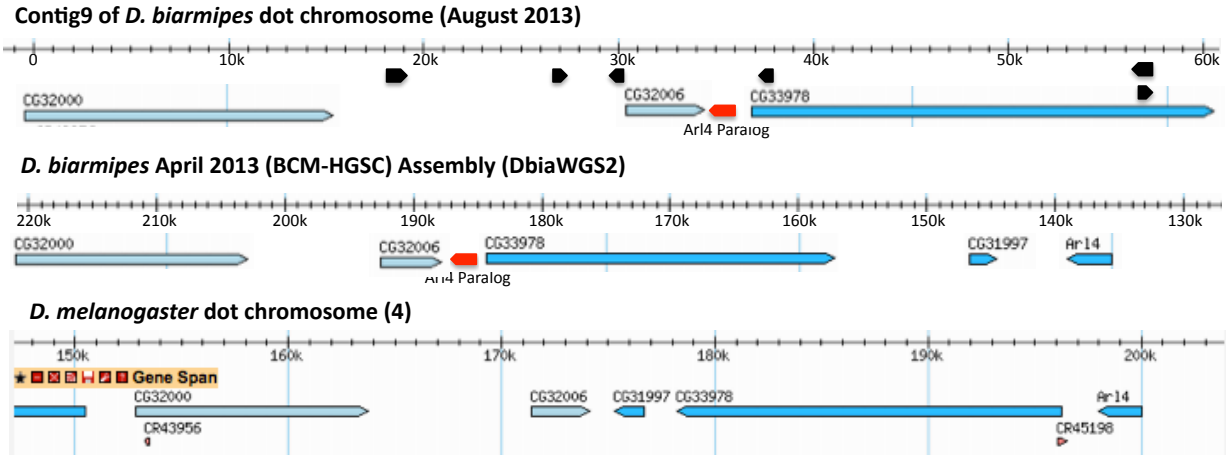


Figure 83: Comparison of gene order and orientation in *D. biarmipes* and *D. melanogaster*. Repeat elements >500 bp long are mapped in black to the gene map for Contig9.

DISCUSSION:

While synteny between *D. biarmipes* and *D. melanogaster* is not entirely conserved, all genes annotated on Contig9 of the *D. biarmipes* dot chromosome are located on the dot chromosome of *D. melanogaster* (also known as chromosome 4 or the Muller F Element). This is consistent with prior analysis, which suggests that genes tend to stay on the same chromosome across *Drosophila* species. Moreover, prior research has suggested that inversions are not uncommon. It is entirely plausible that *CG33978* underwent an inversion at some point in the

evolutionary history between *D. melanogaster* and *D. biarmipes*. The putative *Arl4* paralog requires further investigation. Pseudogenes are uncommon in *Drosophila,* hence I am hesitant to call this paralog a pseudogene. I hypothesize that the paralog is potentially functional, and it is possible that it has been co-opted by a neighboring gene. Overall, the annotation of Contig9 is supported by BLAST alignments, RNA-Seq, TopHat and gene predictors. Confirming the annotation, especially the *Arl4* paralog and the UTRs for all genes, may require further experiments.

APPENDIX:

All fasta, pep and GFF sequence files will be submitted electronically.

ACKNOWLEDGMENTS:

Thank you to Dr. Elgin, Dr. Shaffer, Wilson Leung, Nicholas Spies and Kevin Ko for their help with annotating this project.

REFERENCES:

Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203-218.