Annotation of Contig8

Sakura Oyama
Dr. Elgin, Dr. Shaffer, Dr. Bednarski
Bio 434W
May 2, 2016

Abstract

Contig8, a 45 kb region of the fourth chromosome of *Drosophila ficusphila*, was annotated using the Basic Local Alignment Search Tool (BLAST), FlyBase, Clustal Omega, UCSC Genome Browser, Gene Record Finder, Gene Model Checker, and other tools and databases maintained by the Genomics Education Partnership (GEP). The *Drosophila melanogaster* genome was used as a reference for gene orthology. Three genes were annotated in contig8: *Arl4, CG31997,* and *CG33978*. Conservation relative to *D. melanogaster* orthologs varied across the three genes. The most dramatic changes, including a novel intron and premature stop codons in all isoforms, were observed in *CG33978*. Clustal Omega multiple sequence alignments show that the Calcium-binding EGF-like domain in protein CG33978 is conserved in across multiple *Drosophila* species, despite significant variation in other regions. Due to lack of conservation in 5'UTRs, search regions for transcription start sites (TSSs) of *Arl4* and *CG33978* were proposed. Three likely pseudogenes of *D. melanogaster* gene *mRpL20, Aft6,* and *RpS14a* that were apparently transposed to the dot chromosome by flanking tranpsoons were also annotated in contig8. Conservation data suggests that these transpositions are recent and exist only in *D. ficusphila*. 35.71% of contig8 is composed of repetitious sequences, with nine repeats spanning more than 900 bases. Contig8 is significantly more enriched in repeats relative to the orthologous region in *D. melanogaster*, which has a repeat content of 7.93%. A subset of the repeats identified by RepeatMasker in *D. ficusphila* may be derived from genomic fragments of *Wolbachia* that were found to be integrated into the genome within contig8. Comparison of the relative frames and positions of genes in contig8 with the orthologous region in *D. melanogaster* reveals an inversion involving *Arl4, CG31997*, and *CG33978* in *D. ficusphila*. Additionally, *D. melanogaster* contains non-coding RNA (ncRNA) *CR45198*, which has disappeared in the *D. ficusphila* genome.

Introduction

In eukaryotes, chromatin can be classified as either euchromatin or heterochromatin. Euchromatin is enriched in genes and is associated with active transcription, as its loose packaging allows easy access for RNA polymerase. In contrast, heterochromatin is a tightly packed form of DNA that is generally inaccessible to RNA polymerase and thus associated with silencing. The Muller F element in *Drosophila*, also known as the dot chromosome, appears to be mostly constitutive heterochromatin. Yet, despite being surrounded by the high quantities of repetitious sequences usually associated with heterochromatin formation, most of the approximately eighty genes found on the 1.3 Mb arm of the chromosome are expressed.

With the advent of new computational technologies and next generation sequencing, the genomes of several *Drosophila* species have been sequenced and assembled in order to perform powerful multispecies comparative genomic analyses. These studies can provide valuable insights on how dot chromosome genes can alter chromatin structure in order to recruit the transcription machinery and ultimately become expressed in a heterochromatic environment. This project is part of a larger project to improve the sequence and carefully annotate the F element of *D. ficusphila*, a fly species that diverged from *D. melanogaster* around 10-15 million years ago, an evolutionary distance that is considered the "sweet spot" for comparative genomic analyses for regulator motifs.

Contig8 spans 45,000 bases, and has five features predicted by the *ab initio* gene predictor Genscan, but four genes suggested by conservation (Figure 1). The most upstream prediction, feature 1, was annotated first due to its relatively small size and the congruence of computational and experimental evidence for its structure.

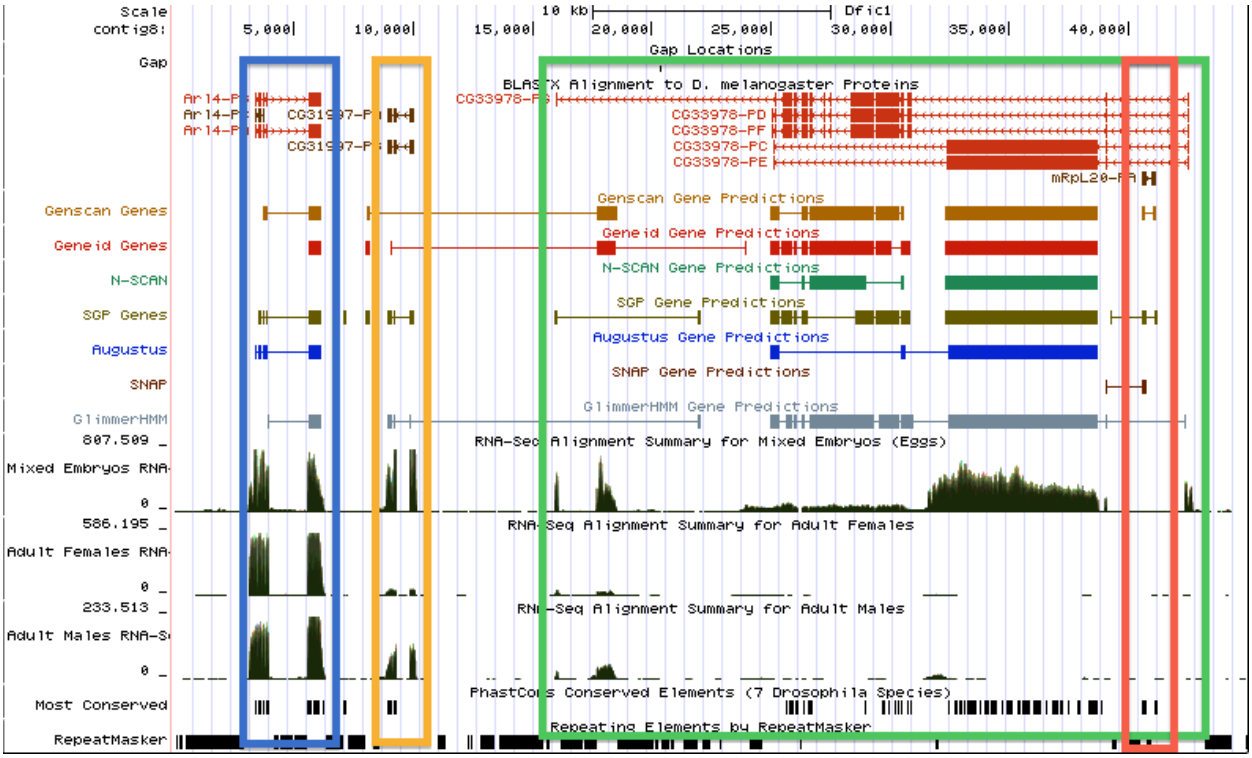


Figure 1: Contig 8 on the GEP UCSC Genome Browser of the *D. ficusphila* Jan. 2016 Dot Assembly. Features 1, 2, 3, and 4 are shown in blue, orange, green, and red respectively.

Feature 1

Using the UCSC Genome Browser, the preliminary properties of feature 1 were assessed (Figure 2). The BLASTX alignment track to *D. melanogaster* shows that the feature is found on the positive strand. The gene prediction tracks display a variable number of exons. However, RNA-Seq data suggests a minimum of five exons, with a large intron between the fourth and fifth exons. In order to determine the *D. melanogaster* ortholog, the predicted peptide from Genscan (query) was used to conduct a BLASTp search of *D. melanogaster* annotated proteins (subject) in FlyBase (Figure 3).

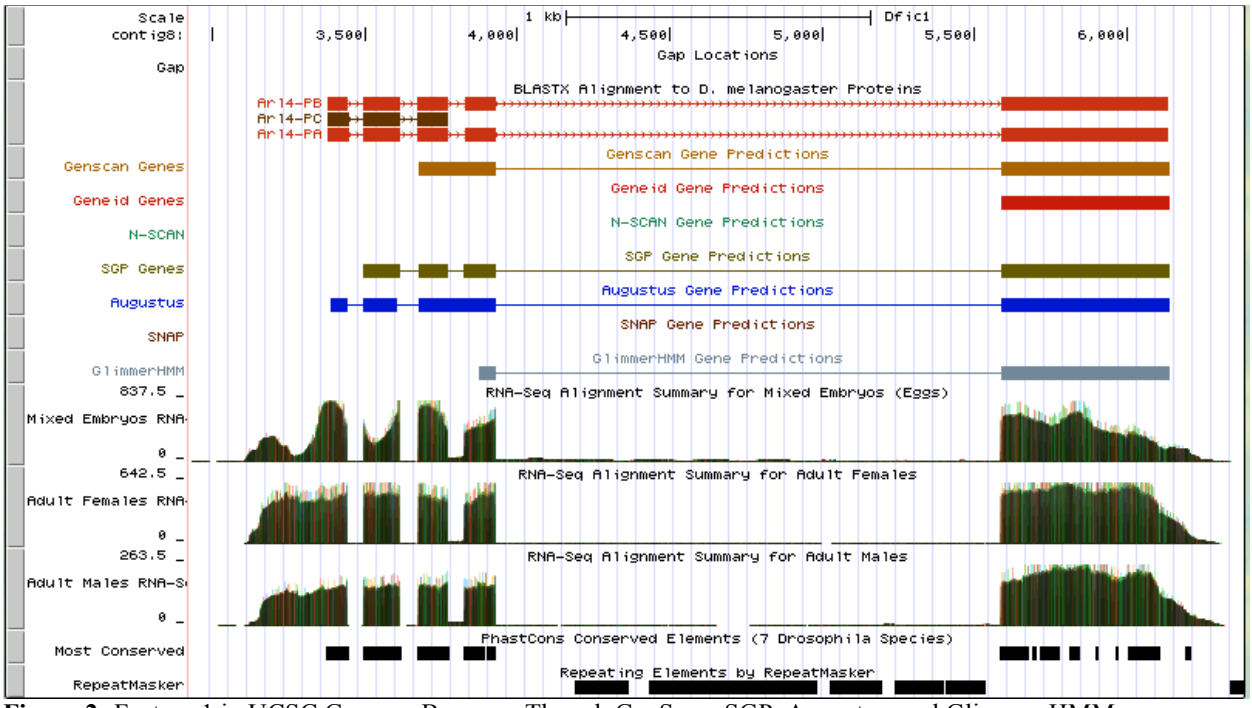


Figure 2: Feature 1 in UCSC Genome Browser. Though GenScan, SGP, Augustus, and Glimmer HMM gene predictions are widely varied, RNA-seq data matches closely with the BLASTx alignment to *D. melanogaster* protein, predicting five exons.

BLAST Hit Summary

Description	Species	Score	E value
Arl4-PA	Dmel	343.199	1.38323e-94
Arl4-PB	Dmel	343.199	1.5036e-94
Arf79F-PJ	Dmel	61.6178	7.09308e-10
Arf79F-PI	Dmel	61.6178	7.09308e-10

>gnl|dmel|FBpp0088220 type=protein; loc=4:complement(join(179264..179329, 179092..179211, 178937..179034, 178778..178880, 177441..177992)); ID=FBpp0088220; name=Arl4-PA; parent=FBgn0039889, FBtr0089153; dbxref=FlyBase:FBpp0088220, GB_protein:AAF59368.2, FlyBase_Annotation_IDs:CG2219-PA, REFSEQ:NP_651905, GB_protein:AAF59368, FlyMine:FBpp0088220, modMine:FBpp0088220; MD5=7d1b1ae8b7a7ed223ab8cfbc6526e200; length=312; release=r6.09; species=Dmel;
Length = 312

HSP # = 1 , **Score =** 343.199 bits (879) , **Expect =** 1.38323e-94
Identities = 169 / 217 (77.9%) , **Positives =** 190 / 217 (87.6%)

Subject FASTA

```
Query:   52          CTDGILFVIDSVDTERMEEAKMELMRTAKCPDNQGVPVLILANKQDLPNACGAIELEKLL   111
                     CTDGILFVIDSVDTERMEEAKMELMRTAKCPDNQGVPVLILANKQDLPNACGA+ELEKLL
Subject: 96          CTDGILFVIDSVDTERMEEAKMELMRTAKCPDNQGVPVLILANKQDLPNACGAMELEKLL   155

Query:   112         GLNELYSPVPNISILTSSNSSSTVNLIGCSMANQSIIESSSTKRTSSHLHSSMIHIKPAL   171
                     GLNELY+PVPNIS+ +SS+SS T+NLIGC ++NQSI + S  ++  SHLHSSMIHIKPAL
Subject: 156         GLNELYNPVPNISMPSSSDSSPTINLIGCRVSNQSITDKSLEEKKESHLHSSMIHIKPAL   215

Query:   172         ETGDQKDTLTEEALPAFIYSHSYNDPADLDQQTQREVKKCFHNKKHNRASSNSMQFRGWY   231
                     E+ D   TL+  AL AFIY  S+N+ A LDQ+  ++VK  FHNKK NR+SSNS+QFRGWY
Subject: 216         ESKDHNSTLSGGALTAFIYPQSHNNSAVLDQKNPQDVKNGFHNKKMNRSSSNSVQFRGWY   275

Query:   232         IQPTCAITGEGLQEGLDALYDMILKRRKINKSQKKKL    268
                     IQPTCAITGEGLQEGLDALYDMILKRRKINKS K+ L
Subject: 276         IQPTCAITGEGLQEGLDALYDMILKRRKINKSNKRNL    312
```

Figure 3: FlyBase BLASTp results for the Genscan predicted polypeptide against the *D. melanogaster* annotated proteins database. There is high sequence similarity to the A and B isoforms of Arl4 protein. However, the first fifty-one amino acids from the Genscan prediction are missing from the alignment.

The BLASTp results show a high sequence similarity to isoforms A and B of *Arl4*. According to the Gene Record Finder, *Arl4* is located on the *D. melanogaster* chromosome four. Since contig8 is known to be part of the fourth chromosome on *D. ficusphila*, it is very likely that *Arl4* is the correct *D. melanogaster* ortholog given the BLASTp e-values and appropriate chromosome. Although high sequence similarity can be observed at the beginning and end of the alignment, the first fifty-one amino acids of the Genscan prediction were completely missing from the alignment.

Gene Record Finder shows that *Arl4* has three isoforms (A, B, and C) in *D. melanogaster* (Figure 4). Isoforms A and B both have five exons and share almost identical coding sequences, but have different 3' untranslated regions (UTR). Isoform C only has four exons. Though its first three exons are identical to isoform A, its fourth exon is very small and nested within the fourth exon region of isoforms A and B. Additionally, the fifth exon in isoforms A and B corresponds to a large 3' UTR in isoform C (Figure 5). By definition, all three isoforms should be located in the same region of contig8.

CDS usage map:

Isoform	2_9580_0	1_9580_0	3_9580_0	4_9580_0	6_9580_1	5_9580_1	7_9580_0
Arl4-PB	1		2	3	4		5
Arl4-PA		1	2	3	4		5
Arl4-PC		1	2	3		4	

Select a row to display the corresponding CDS sequence:

FlyBase ID	5' Start	3' End	Strand	Phase	Size (aa)
2_9580_0	179,329	179,261	-	0	23
3_9580_0	179,211	179,092	-	0	40
4_9580_0	179,034	178,937	-	0	32
6_9580_1	178,880	178,778	-	1	34
7_9580_0	177,992	177,441	-	0	184

Figure 4: Gene Record Finder 6.08 record of *Arl4* in *D. melanogaster*. *Arl4* has three isoforms (A, B, and C) in *D. melanogaster*. The CDS sequence of Arl4-PB is shown.

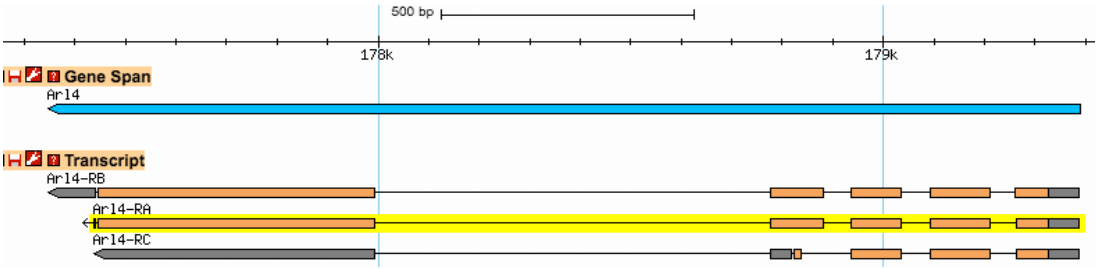


Figure 5: Browser view of *Arl4* in *D. melanogaster* genome. The coding regions are shown in light orange and the untranslated regions are show in gray.

To begin mapping the coordinates of feature 1, a series of pairwise BLASTx searches was conducted using the entire fasta sequence of contig8 as the query and each unique exon (coding sequence, CDS) of *Arl4* as the subject (Figure 6). The BLASTx output was then used to obtain the approximate exon boundaries and the reading frames for each putative exon in Feature 1 (Table 1). CDS7_9580_0, exon five in isoforms A and B, has a size of 184 amino acids and was used to anchor the gene in contig8. The BLASTx output was then used to obtain the approximate exon boundaries and reading frames for each putative exon in Feature 1 (Table 1).

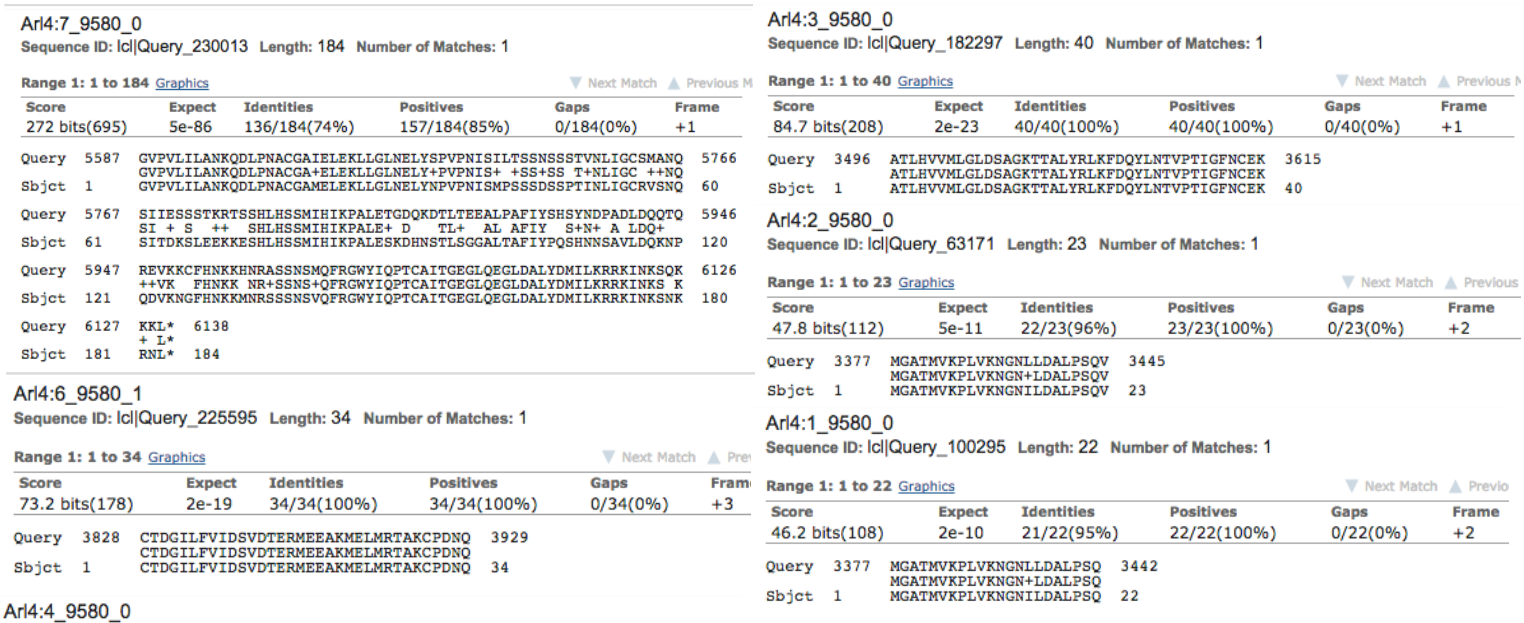


Figure 6: BLASTx alignments of all unique *D. melanogaster Arl4* coding exons (subject) against translated contig 8 DNA sequence (query). All matches covered the entirety of the *Arl4* CDSs except for CDS5_9580_1, which produced no BLASTx alignments. Note that CDS1_9580_0 and CDS2_9580_0 differ by only one amino acid.

FlyBaseID	Coding Exon Size (amino acids)	Query Range	Query Frame	Subject Range
1_9580_0	22	3377-3442	+2	1-22
2_9580_0	23	3377-3445	+2	1-23
3_9580_0	40	3496-3615	+1	1-40
4_9580_0	32	3676-3771	+1	1-32
5_9580_1	6	n/a	n/a	n/a
6_9580_1	34	3828-3929	+3	1-34
7_9580_0	184	5587-6138	+1	1-184

Table 1: Summary table for the approximate exon locations based on BLASTx alignments for Arl4

All coding exons were mapped in their entirety on the forward strand except for CDS5_9580_1, the final exon of isoform C. This lack of alignment was expected because CDS5_9580_1 is reported to be only six amino acids long (Figure 7). Since CDS5_9580_1 is nested within 6_9580_1, the fourth exon of isoforms A and B (Figure 5), the subject was narrowed to the CDS6_9580_1 alignment region within the fasta file. In addition, within general parameters, the "expect threshold" was also raised to ten and "word size" was decreased to two. Within scoring parameters, the "matrix" was changed to PAM30. However, none of these adjustments produced BLASTx matches to CDS5_9580_1 within contig8. Due to the lack of computational data, the UCSC Genome Browser output was closely analyzed for regions of increased expression within the CDS6_9580_1 alignment region. The TopHat junctions, Cufflink transcripts, and Oases transcripts data showed expression from bases 3872-3929 (Figure 8). This region spans approximately 19 amino acids and is thus significantly longer than the fourth exon of *Arl4-PC* in *D. melanogaster*. In order to find the exact mutations that caused the expansion of CDS5_9580_1 in *D. ficusphila,* the original exon in *D. melanogaster* was examined in the UCSC Genome Browser. Where the original amino acid sequence, KNGRS*, was located, the conservation tracks revealed that a mutation from a T to an A at base 170,020 had caused the stop codon to become a lysine in *D. ficusphila* (Figure 9). This mutation was shared by all other

Drosophila species listed in the conservation track except for *D. melanogaster*, D. *yakuba* and *D. erecta* (Figure 10).

CDS usage map:

Isoform	2_9580_0	1_9580_0	3_9580_0	4_9580_0	6_9580_1	5_9580_1	7_9580_0
Arl4-PB	1		2	3	4		5
Arl4-PA		1	2	3	4		5
Arl4-PC		1	2	3		4	

Select a row to display the corresponding CDS sequence:

FlyBase ID	5' Start	3' End	Strand	Phase	Size (aa)
1_9580_0	179,329	179,264	-	0	22
3_9580_0	179,211	179,092	-	0	40
4_9580_0	179,034	178,937	-	0	32
5_9580_1	178,836	178,818	-	1	6

Figure 7: Gene Record Finder 6.08 record of *Arl4* in *D. melanogaster*. CDS details for isoform C shows that the fourth exon is only six amino acids long.

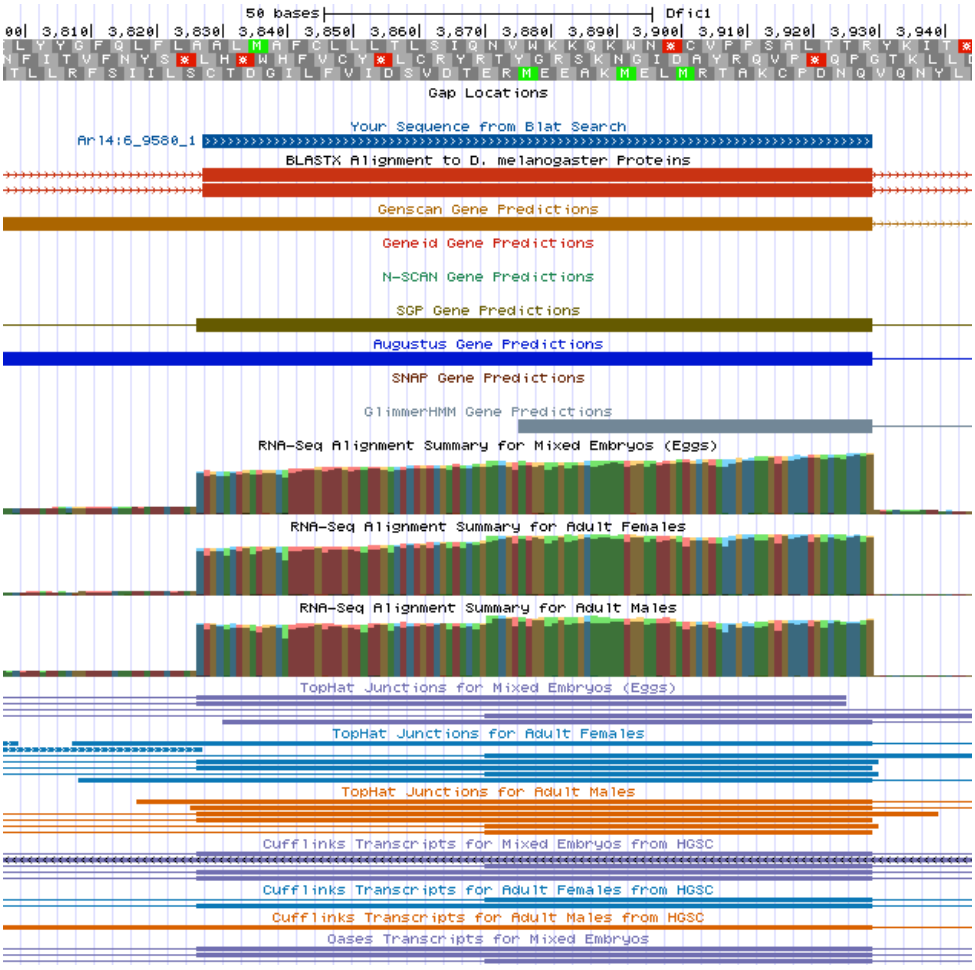


Figure 8: TopHat junctions, Cufflinks transcripts, and Oases transcripts data show a pattern of splice sites and a region of increased expression within the CDS6_9580_1 alignment region that likely originated from CDS5_9580_1 in *D. melanogaster*. Note position of splice sites at base 3870.

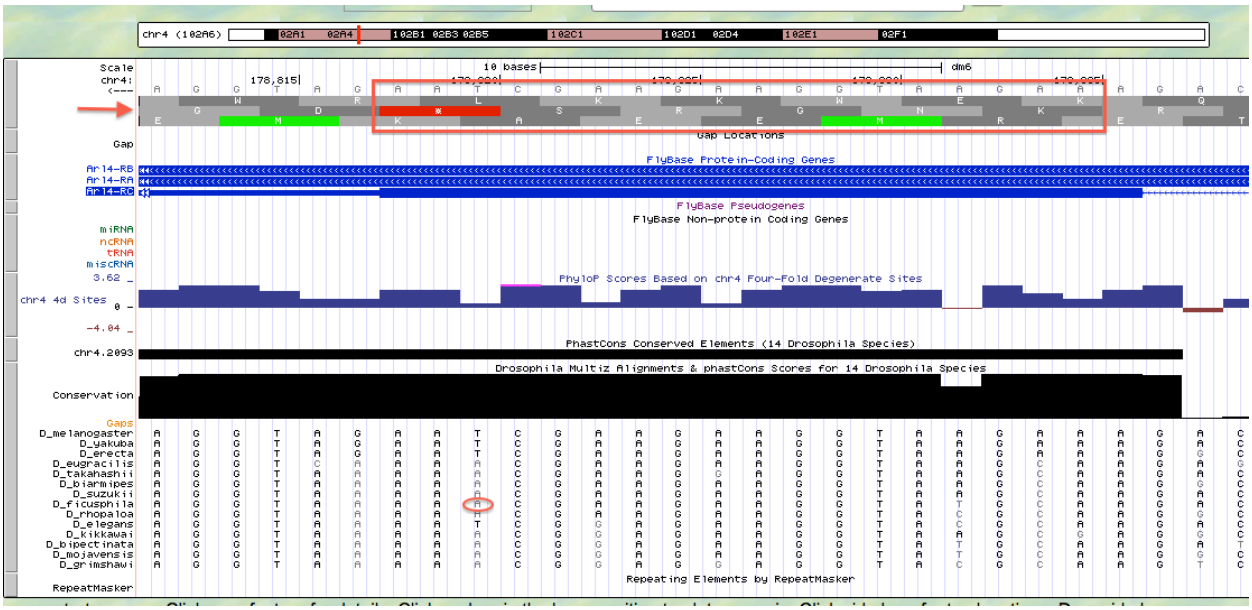


Figure 9: Alignment of original CDS5_9580_1 in *D. melanogaster* in the UCSC Genome Browser. The conservation tracks reveal that the T at base 170,020 has mutated to an A in *D. ficusphila*, changing the stop codon to a lysine. This mutation is shared by all other *Drosophila* species except for *D. yakuba*, *D. erecta*, and *D. elegans.*

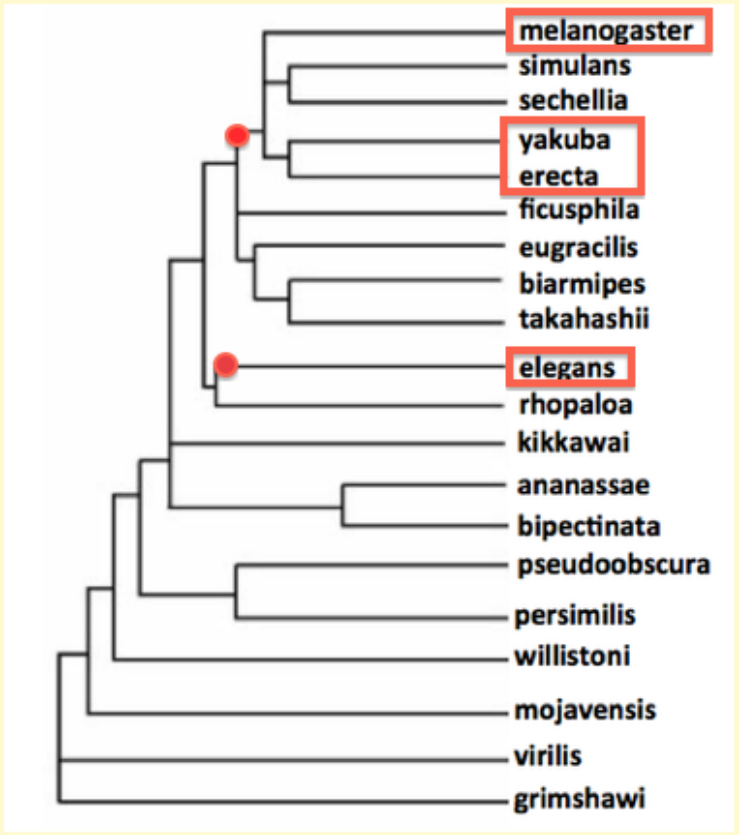


Figure 10: Shared ancestry of *Drosophila* species. The *Drosophila* species enclosed in red contain a stop codon at position 178,818-820. (*D. simulans* and *D. sechellia* are not included in the conservation track in the *D. melanogaster* Genome Browser due to poor quality sequence, but it is likely that the stop codon is conserved in these two species as well.) Thus, the change in CDS5_9580_1 likely occurred following the split of the given species from the rest of the *Drosophila* species, indicated by the red circle. The same mutation likely occurred in *D. elegans*.

To determine exact exon boundaries, the UCSC Genome Browser and the approximate coordinates and reading frames obtained through the BLASTx alignments (see Table 1) were used to locate donor and acceptor

splice sites for each exon. Two start codons were located on the reading frame +2 near the beginning of CDS1_9580_0, the first exon in isoforms A and C, and CDS2_9580_0, the first exon in isoform B (Figure 11). However, if the start codon spanned bases 3389-3391, the first exon of Arl4 in *D. ficusphila* would be approximately three quarters the size of the first exon of Arl4 in *D. melanogaster*. On the other hand, if the start codon spanned bases 3377-3379, the start site coordinates match those obtained from the BLASTx alignment to the *D. melanogaster* sequence. This choice conserves the starting amino acid sequence seen in *D. melanogaster*, MGAT (Figure 6).

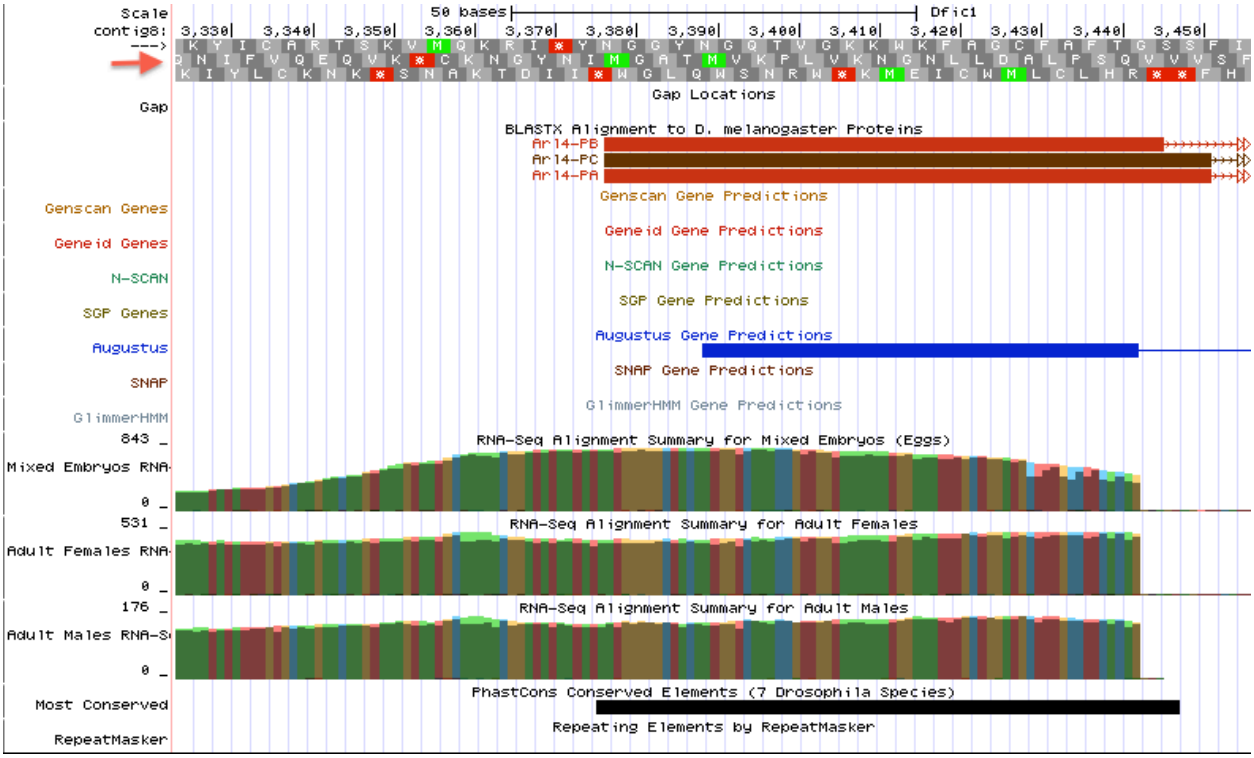


Figure 11: UCSC Genome Browser showing CDS1_9580_0 (exon one in isoform A and C) and CDS2_9580_0 (exon one in isoform B). Two start sites (shown in green) are on reading frame +2 (red arrow). The start codon at base 3377-3379 is supported by amino acid sequence conservation for all three isoforms.

There were three potential GT donor sites close to the 3' end of CDS1_9580_0, the first exon for isoforms A and C, that could potentially be the correct donor site (Figure 12). The GT site at bases 3443-3444 was determined to the most likely donor site, as it matched the BLASTx

results (Table 1), which showed this exon ending in PSQ (Figure 6) and is supported by strong RNA-Seq, TopHat junctions and Cufflinks transcripts data. Since CDS1_9590_0 is on reading frame +2, the most likely donor site at base 3443-3444 is in phase 0 because the donor site is directly adjacent to a complete codon. If the donor site at bases 3443-3444 is indeed the correct donor site, the acceptor site in CDS3_9580_0 should be in phase 0 so that the transcript remains in the correct frame.

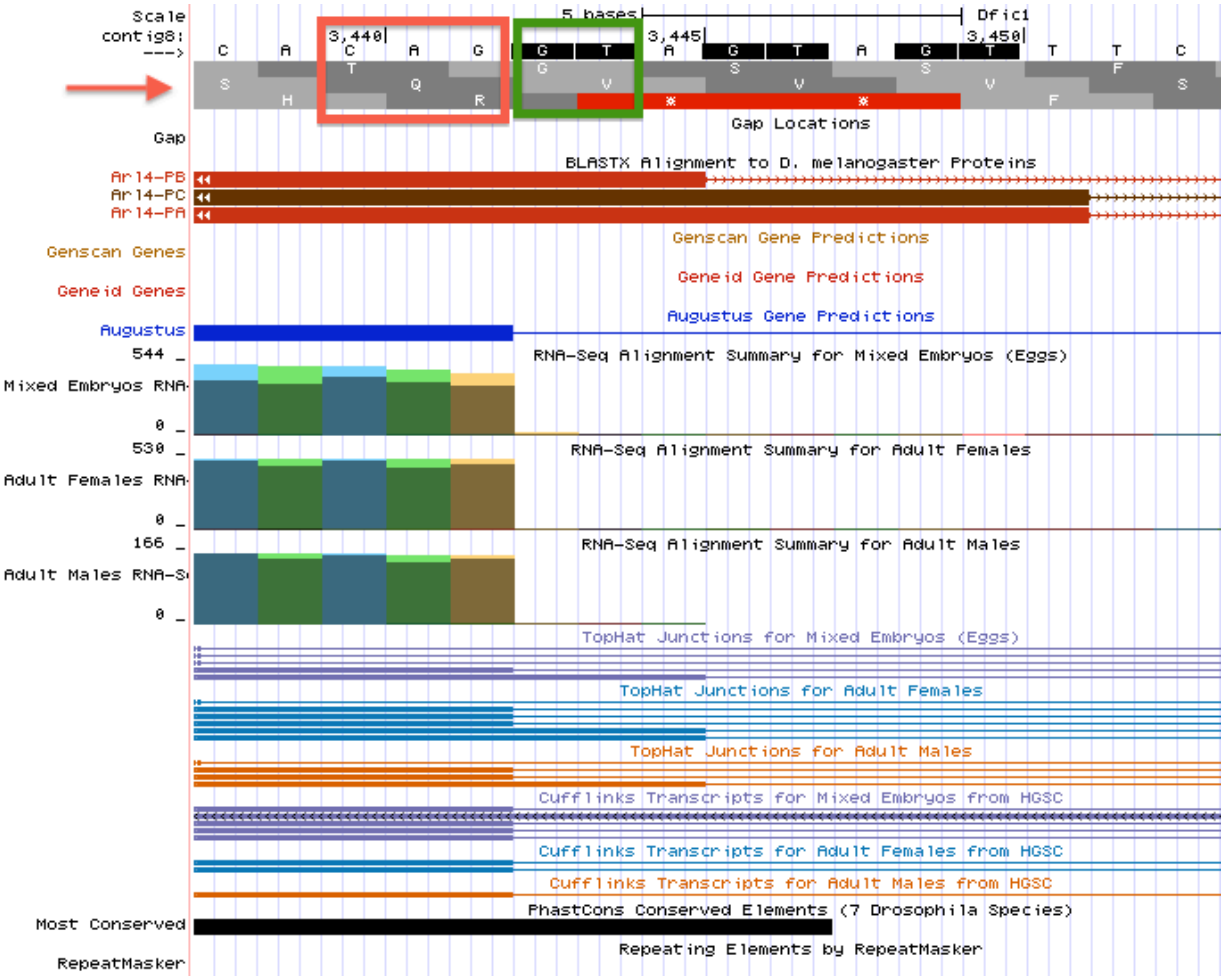


Figure 12: Donor site for CDS1_9580_0 (exon 1 in isoform A and C). The most likely GT donor site is boxed in green. The closest complete codon is boxed in red. The red arrow points to reading frame +2. The most likely donor site at base 3443-3444 is in phase 0.

The same three potential GT donor sites considered for CDS1_9580_0 were also investigated for CDS2_9580_0, the first exon for isoform B. The GT site at bases 3446-3447 was determined to be the most likely donor site, as it matched the BLASTx results (Table 1), which

showed this exon ending in PSQV (Figure 6) and is supported by TopHat junctions data (Figure 13). Since CDS2_9590_0 is on the reading frame +2, the most likely donor site at base 3446-3447 is in phase 0, since the donor site is directly adjacent to a complete codon. This requires a phase 0 acceptor site in CDS3_9580_0, which is in agreement with the conclusions reached after investigation of the donor site for CDS1_9580_0. This is appropriate, as these two exons both splice to exon CDS3_9580_0.

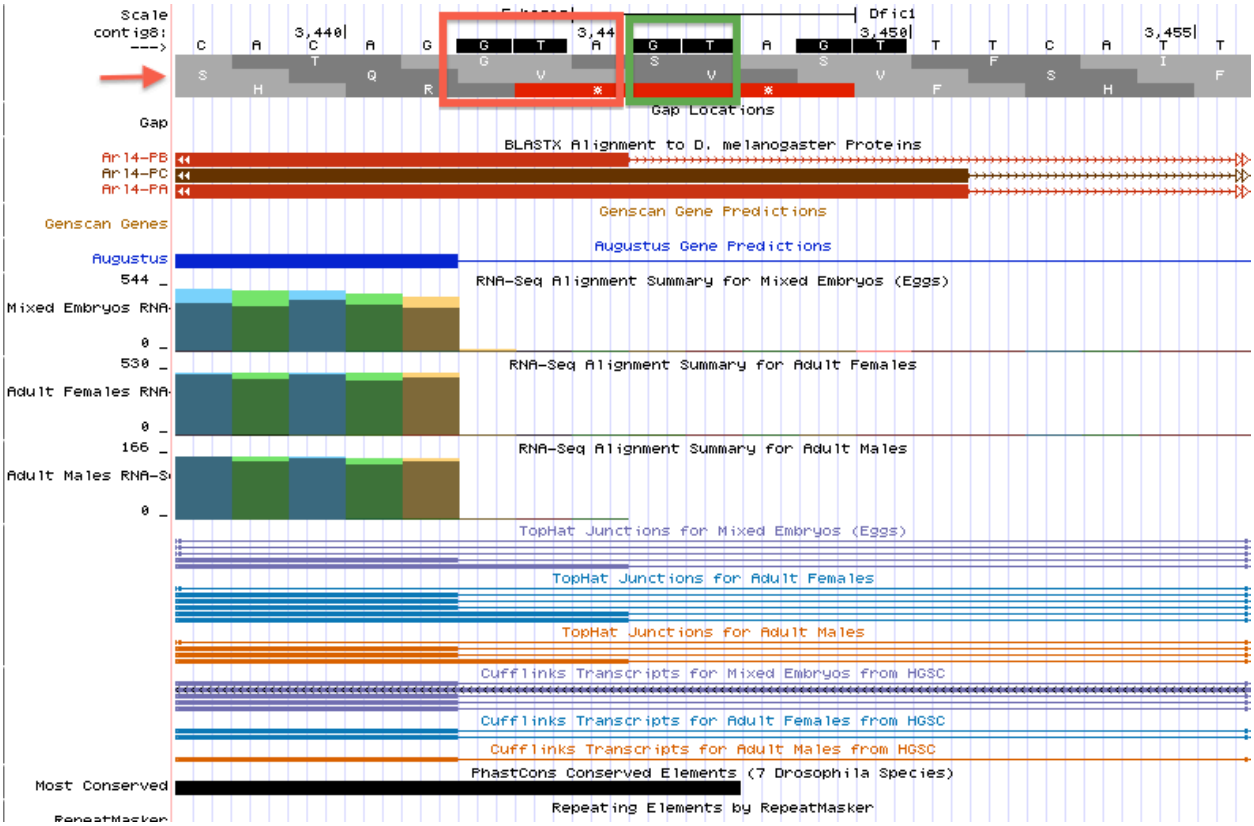


Figure 13: Donor site on CDS2_9580_0 (exon 1 in isoform B). The most likely GT donor site is boxed in green. The closest complete codon is boxed in red. The red arrow points to reading frame +2. The most likely donor site at base 3446-3447 is in phase 0.

The AG acceptor site that corresponds to CDS3_9580_0 spans bases 3494-3495 (Figure 14). This acceptor site matches the site suggested by BLASTx results (Table 1, Figure 6) and is well supported by RNA-Seq data, TopHat junctions, and Cufflinks transcripts. Since CDS3_9580_0 is on the +1 reading frame, the AG acceptor site is in phase 0 and thus is in agreement with the donor sites of both CDS1_9580_0 and CDS2_9580_0, which are in phase 0.

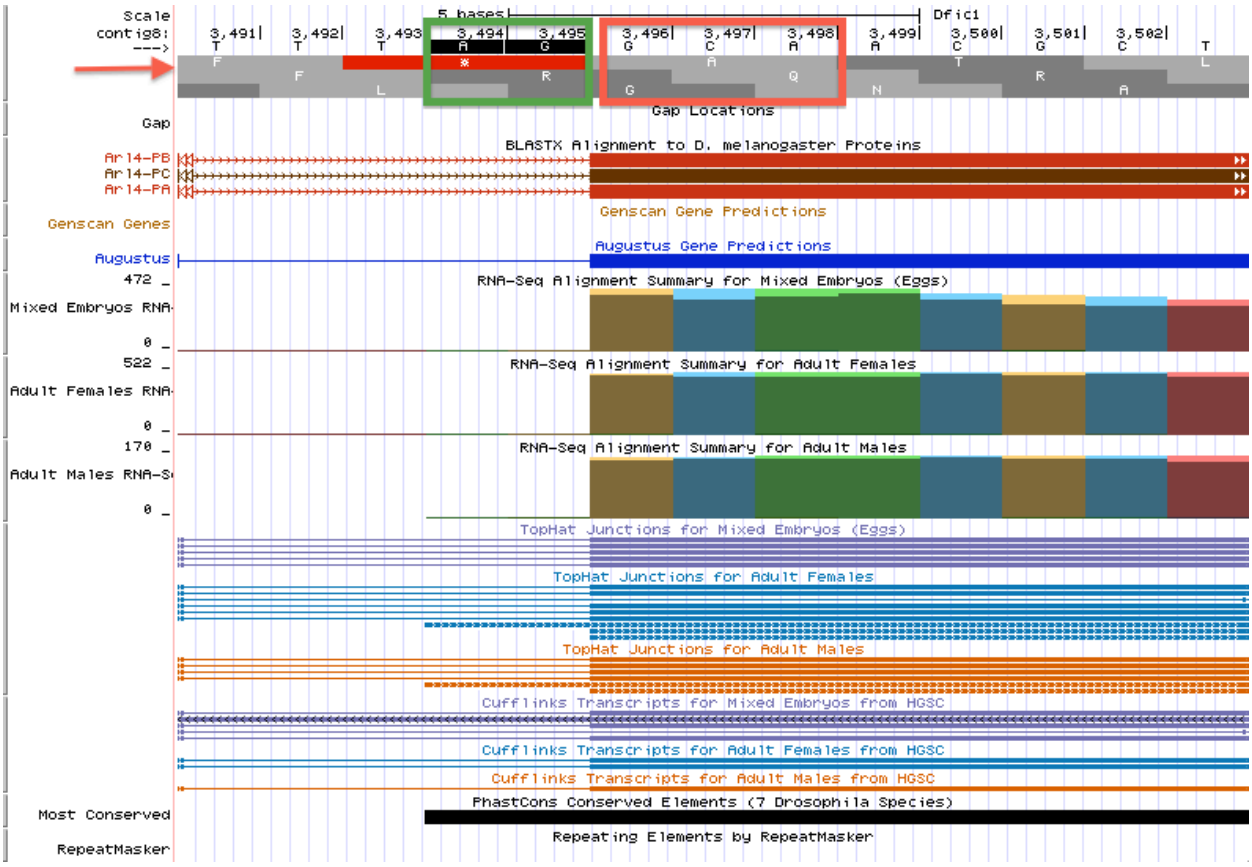


Figure 14: Acceptor site on CDS3_9580_0 (exon 2). The AG acceptor site is boxed in green. The closest complete codon is boxed in red. The red arrow points to reading frame +1.

The GT donor site that corresponds to CDS3_9580_0 spans bases 3616-3617 (Figure 15). This donor site matches the site suggested by BLASTx results (Table 1) and is well supported by RNA-Seq data, TopHat junctions, and Cufflinks transcripts. Since CDS3_9590_0 is on the reading frame +1, the most likely donor site at base 3616-3617 is in phase 0, since the donor site is directly adjacent to a complete codon.

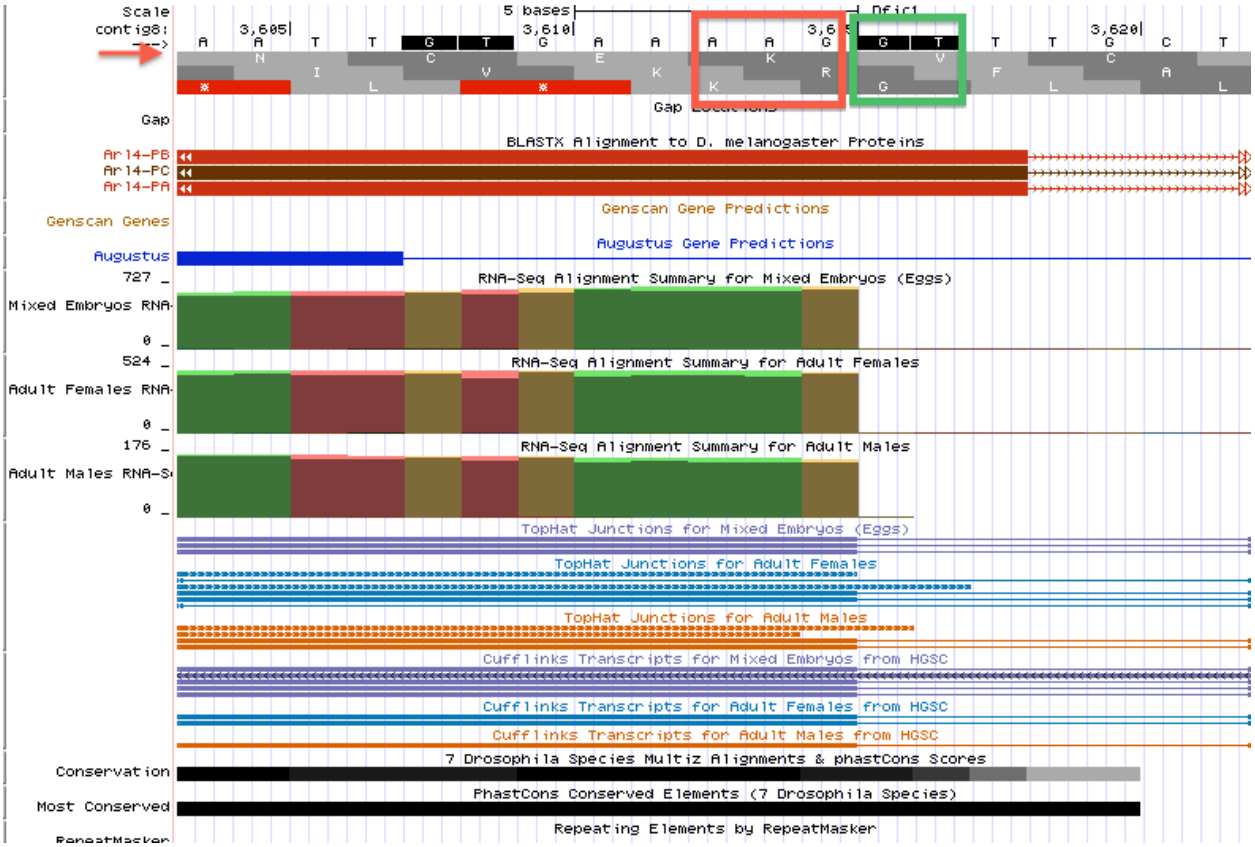


Figure 15: Donor site on CDS3_9580_0 (exon 2). The most likely GT donor site is boxed in green. The closest complete codon is boxed in red. The red arrow points to reading frame +1. The most likely donor site at base 3616-3617 is in phase 0.

The AG acceptor site that corresponds to CDS4_9580_0 spans bases 3675-3676 (Figure 16). This acceptor site matches the site suggested by BLASTx results (Table 1, Figure 6) and is well supported by RNA-Seq data, TopHat junctions, and Cufflinks transcripts. Since CDS4_9580_0 is on the +1 reading frame, the AG acceptor site is in phase 0 and thus is in agreement with the donor site of CDS3_9580_0, which is in phase 0. Another AG can be found further downstream spanning bases 3681-3682. However, this AG is in phase 2 on the +1 reading frame and is thus not in agreement with the phase 0 donor site of CDS3_9580_0. The selected site conserves the starting amino acids of the exon, VQ (Figure 6).

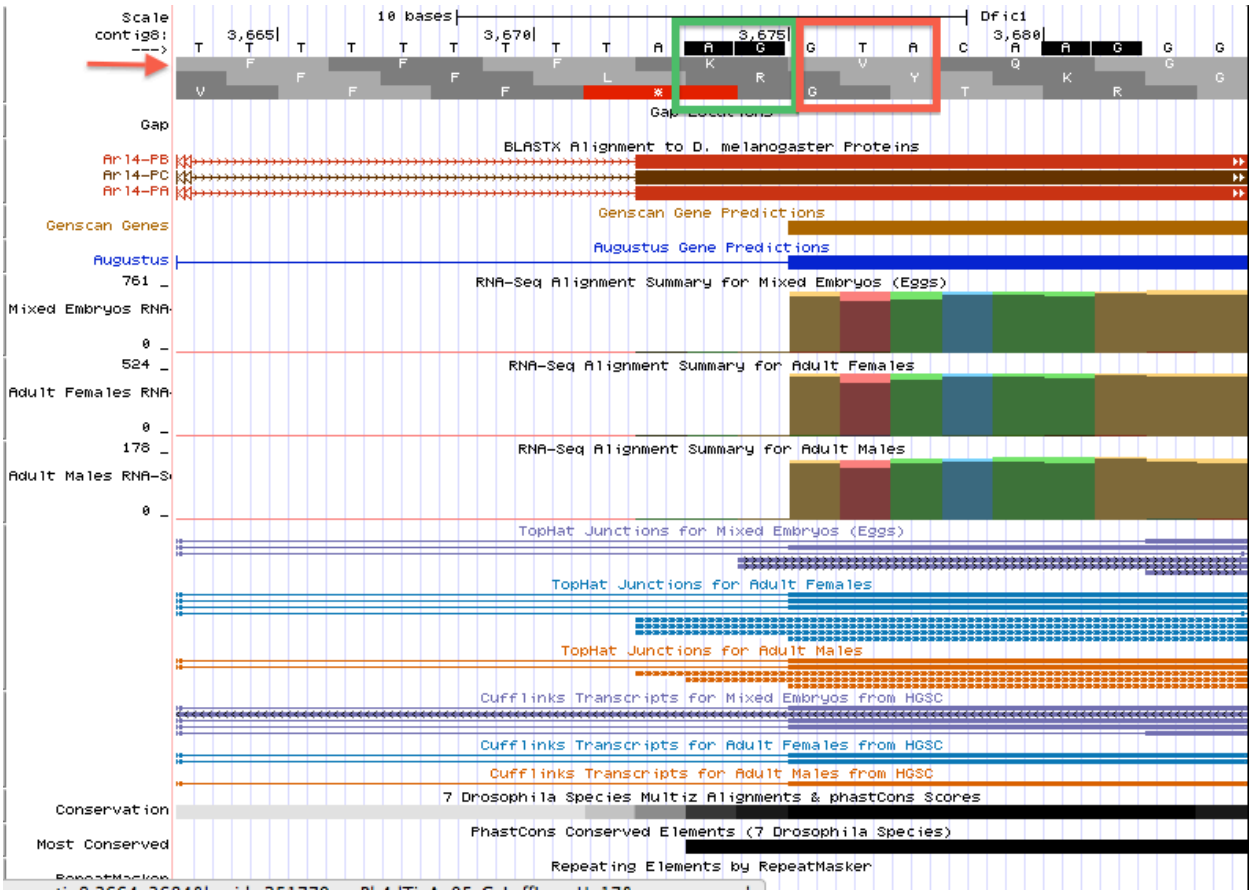


Figure 16: Acceptor site on CDS4_9580_0 (exon 3). The AG acceptor site is boxed in green. The closest complete codon is boxed in red. The red arrow points to reading frame +1.

The GT donor site that corresponds to CDS4_9580_0 spans bases 3774-3775 (Figure 17). This donor site matches the site suggested by BLASTx results (Table 1) and is well supported by RNA-Seq data, TopHat junctions, and Cufflinks transcripts. Since CDS4_9590_0 is on the reading frame +1, the most likely donor site at base 3616-3617 is in phase 2, since there are two nucleotides between the GT and the closest complete codon.

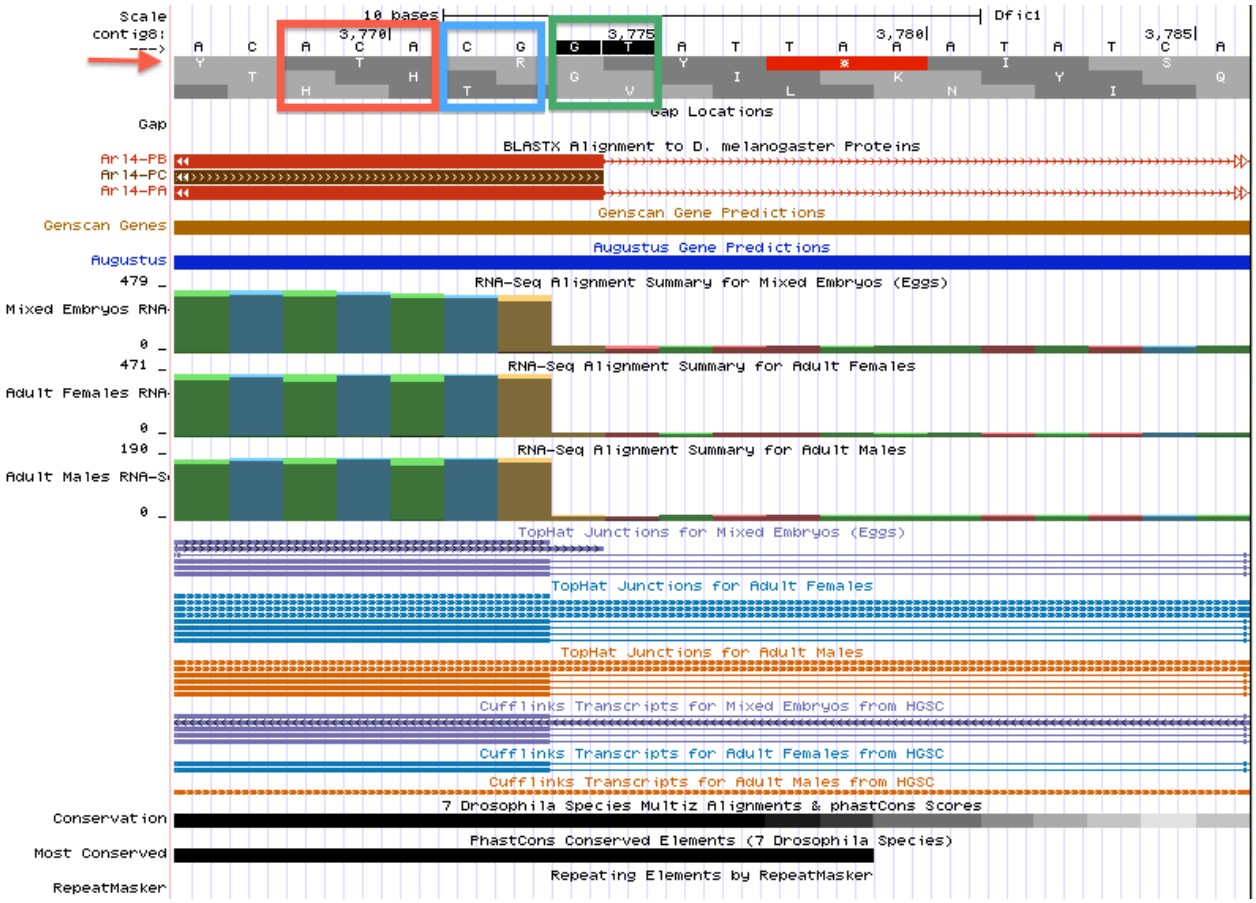


Figure 17: Donor site on CDS4_9580_0 (exon 3). The most likely GT donor site is boxed in green. The closest complete codon is boxed in red. The extra nucleotides in between are boxed in blue. The red arrow points to reading frame +1. The most likely donor site at base 3774-3775 is in phase 2.

In order to find the acceptor site for CDS5_9580_1, the final exon of isoform C, the start of the region within CDS6_9580_1 with increased RNA-Seq data coverage was searched for AGs. The most probably AG acceptor site spans bases 3869-3870 and is supported by TopHat junctions and Cufflinks transcripts data (Figure 18). Since the donor site on CDS4_9580_0 is phase 2, this AG acceptor site must be phase 1. This means that CDS5_9580_1 must be on the +2 frame. Note that this conserves some amino acids, as in *D. melanogaster*, CDS5_9580_0 is KNGRS, while in *D. ficusphila*, the exon begins with TYGRS but then continues (Figure 9). A stop codon is found at bases 3920-3922 on the +2 frame. This is the stop codon for isoform C (Figure 19).

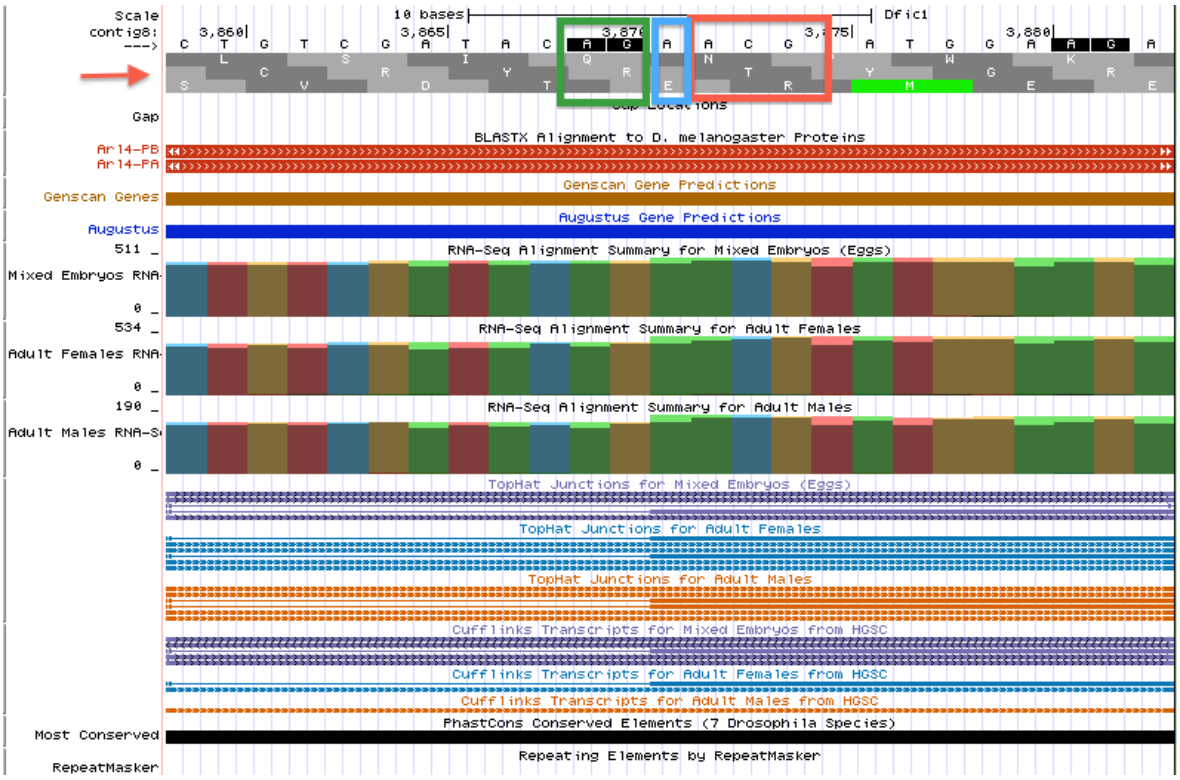


Figure 18: Acceptor site for CDS5_9580_0 (exon 4 in isoform C). The most likely AG acceptor site is boxed in green. The closest complete codon is boxed in red. The extra nucleotides in between are boxed in blue. The red arrow points to reading frame +2. The most likely acceptor site at base 3869-3870 is in phase 1.

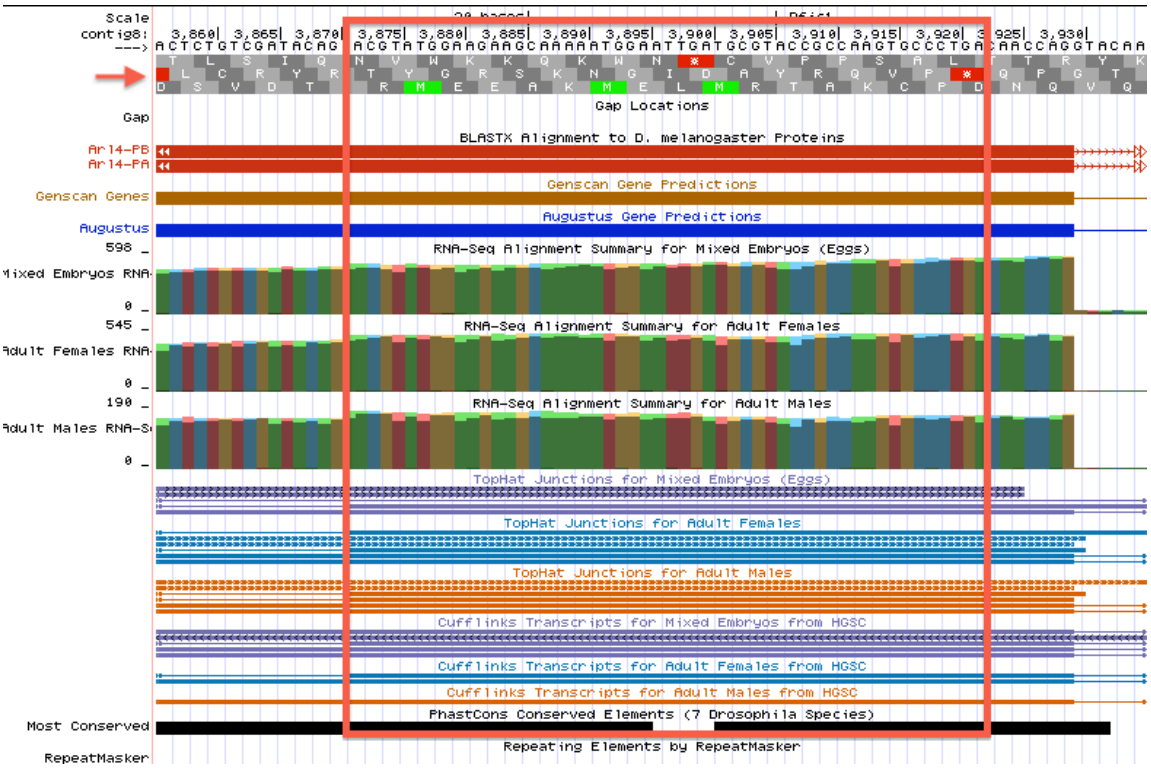


Figure 19: An expanded form of CDS5_9580_1 (exon 4 in isoform C) is found in *D. ficusphila*, spanning bases 3871-3922 on frame +2.

Donor sites and acceptor sites were identified for CDS6_9580_1 and CDS7_9580_0 using similar protocols. The stop codon for isoforms A and B at the 3' end of CDS7_9580_0 on reading frame +1 matches with BLASTx searches (Figure 6, Table 1) and is supported by gene predictors (Figure 20). All donor and acceptor sites matched the approximate BLASTx results (Table 1) and are well supported by RNA-Seq data and TopHat junctions. Thus, all coding exons, as well as the start and stop sites, were annotated for isoforms A, B, and C (Table 2).

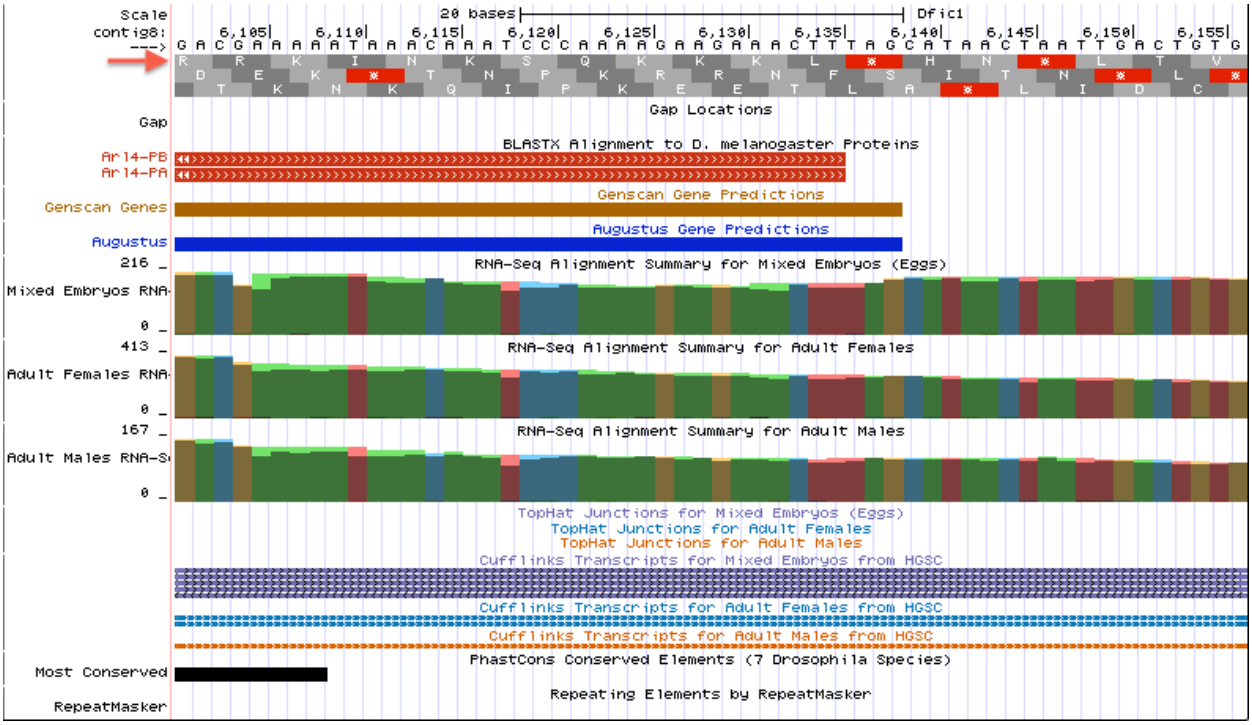


Figure 20: UCSC Genome Browser showing the 3' end of CDS7_9580_0 (exon 5 in isoforms A and C). The red arrow points to reading frame +1. The stop codon is at bases 6136-6138.

FlyBase ID	Coding Exon Size	Begin	End	Isoform A	Isoform B	Isoform C	Frame	Acceptor	Donor	Comments
1_9580_0	22	3377	3442	Exon 1		Exon 1	+2	---	Phase 0	
2_9580_0	23	3377	3445		Exon 1		+2	---	Phase 0	
3_9580_0	40	3496	3615	Exon 2	Exon 2	Exon 2	+1	Phase 0	Phase 0	
4_9580_0	32	3676	3773	Exon 3	Exon 3	Exon 3	+1	Phase 0	Phase 2	
5_9580_1	19	3871	3922			Exon 4	+2	Phase 1	---	Expanded from original exon in *D. melanogaster*
6_9580_1	34	3827	3929	Exon 4	Exon 4		+3	Phase 1	Phase 0	
7_9580_0	184	5587	6138	Exon 5	Exon 5		+1	Phase 0	---	

Table 2: Summary of the proposed gene model components for feature one. Red arrows show corresponding phases between acceptor and donor sites.

The proposed coordinates for feature one were submitted to the Gene Model Checker. The dot plot comparisons between *D. melanogaster* Arl4 and feature one for isoforms A and B are nearly identical since the two isoforms only differ by one base in the first exon. Though large regions of non-homology exist in the fifth exon for isoforms A and B, the exon is anchored on both ends as can be seen by the sequence alignment (Figure 21, 22). The dot plot comparison for isoform C shows very high sequence similarity, with the exception of the final exon (Figure 23). This was expected due to the expansion of the final exon in *D. ficusphila* from 6 to 17 amino acids due to a mutation at the stop codon in *D. melanogaster*.

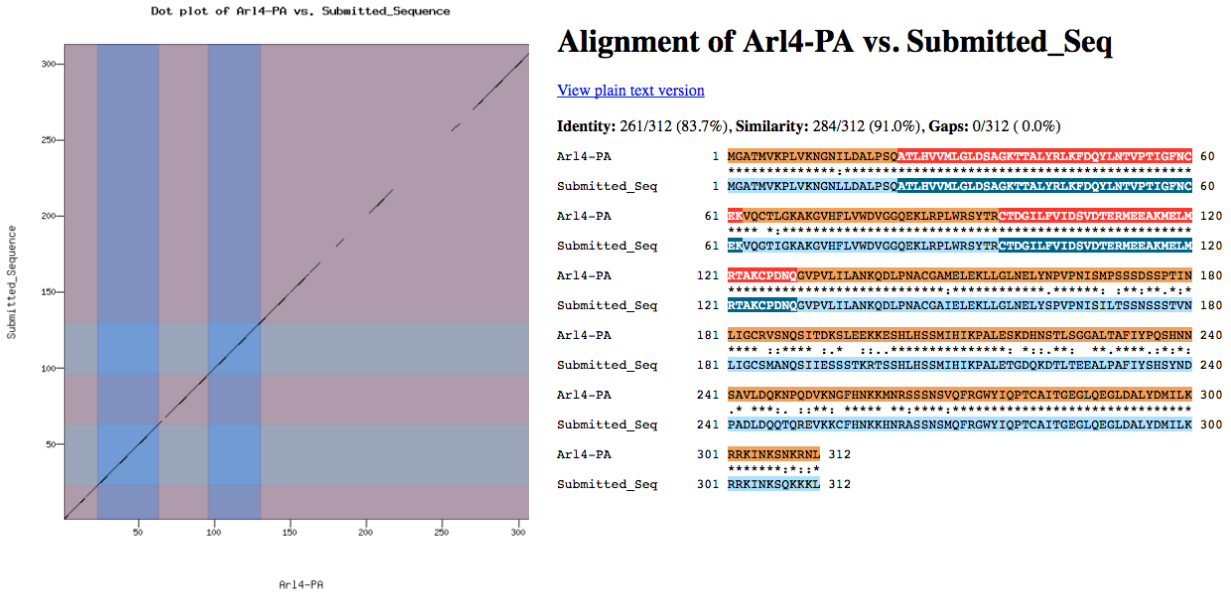


Figure 21: Gene Model Checker output for Isoform A. Left: Dot plot Arl4-PA vs. feature 1 model. Right: Sequence alignment of Arl4-PA and feature 1 model. Though significant regions of low-homology are apparent in the dot plot of the fifth exon, the sequence alignment shows that both ends of the exon are securely anchored.

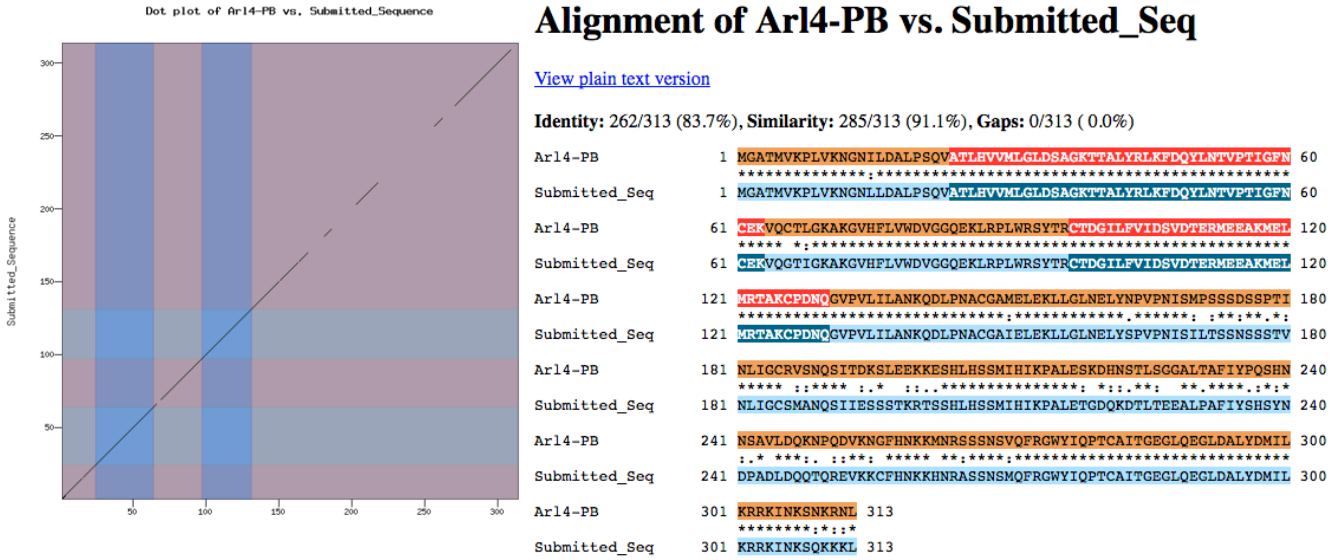


Figure 22: Gene Model checker output for Isoform B. Left: Dot plot of Arl4-PB vs. feature 1 model. Right: Sequence alignment of Arl4-PB and feature 1 model. Though significant gaps are apparent in the dot plot of the fifth exon, the sequence alignment shows that both ends of the exons are securely anchored.

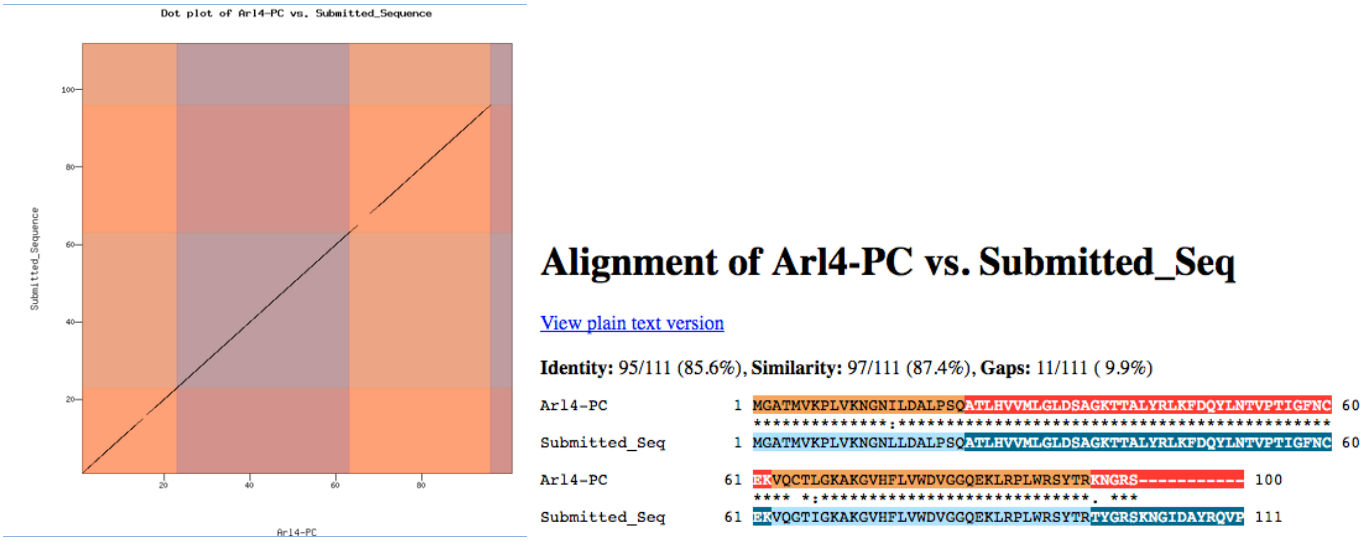


Figure 23: Gene Model Checker output for Isoform C. Left: Dot plot Arl4-PC vs. feature 1 model. Right: Sequence alignment of Arl4_PC and feature 1 model. The poor fourth exon alignment was expected due to its expansion in the *D. ficusphila* genome.

Transcription Start Sites

All isoforms of *Arl4* in *D. melanogaster* have identical 5' untranslated regions (Figure 5). In order to start investigating transcription start sites (TSSs), the 5'UTR of *D. melanogaster* was viewed in the UCSC Genome Browser (Figure 24). The 9-state epigenomic landscape tracks for

BG3 and S2 cells indicate that *Arl4* is actively transcribed in these cell lines despite the final exon in isoforms A and C being located in a heterochromatic region. Furthermore, the DNase I hypersensitive sites (DHS) have significant peaks near the region for all three cell lines. The TSS (Celniker) track only shows one annotated TSS by the modENCODE project (Hoskins, et al. 2011), which suggests that *Arl4* has a peaked promoter, and thus should only have one TSS.

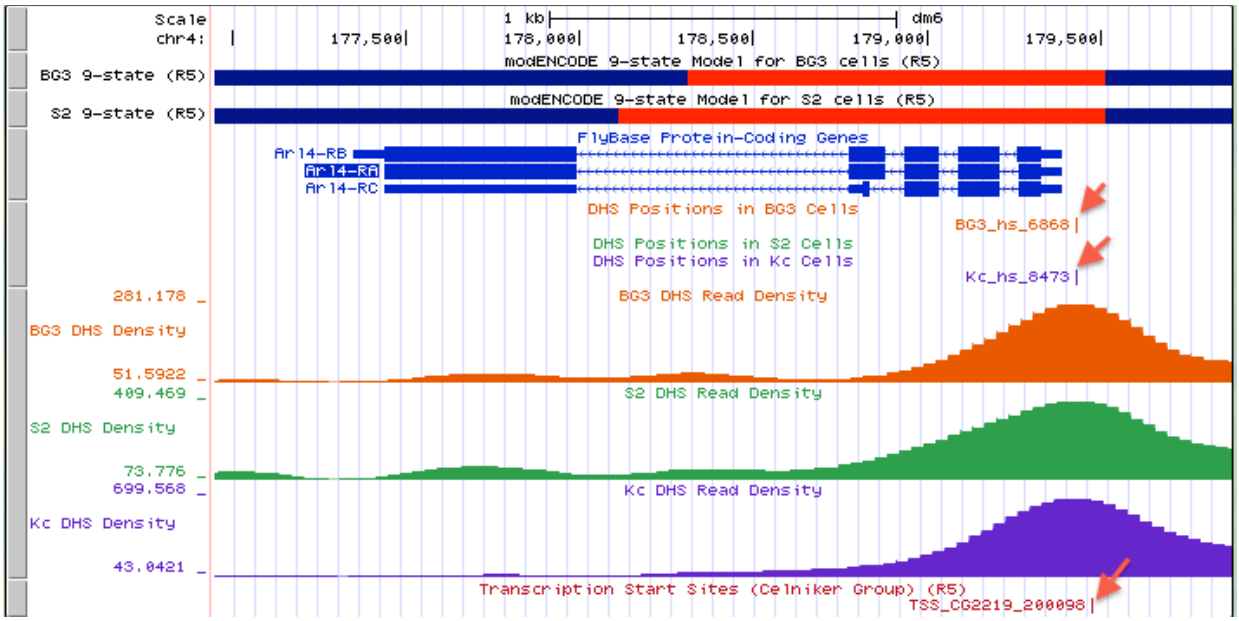


Figure 24: The 5' UTR of *D. melanogaster Arl4*. The 9-state tracks show that the 5' UTR of *Arl4* is within an "Active promoter/TSS region" (red color). A single TSS has been previously annotated, indicated by a red arrow, suggesting that *Arl4* has a peaked promoter.

A pairwise BLASTn alignment with optimized parameters was carried out using the first exon transcript of isoform B in *D. melanogaster* as the query and contig8 as the subject (Figure 25). The modified parameters optimize alignments of sequences with 50-70% identity, which is expected for the less conserved UTRs, and consist of "Word Size = 7," "Match/Mismatch Scores = 1, -1," and "Gap Costs = Existence: 2 Extension: 1." The best BLASTn alignment maps bases 31-129 of isoform B transcript in *D. melanogaster* to bases 3345-3445 in contig8. Bases 61-129 correspond to the coding sequence. The first twenty bases of the 5' UTR are not conserved. If the

size of the 5' UTR in *D. ficusphila* is the same as in *D. melanogaster*, the *D. ficusphila Arl4* TSS is expected to be found at base 3325 on contig8.

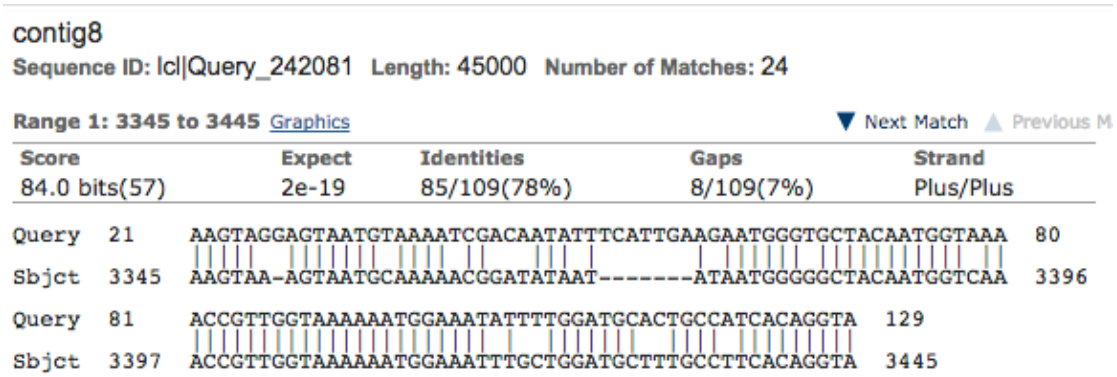


Figure 25: BLASTn alignment of the first exon transcript of isoform B in *D. melanogaster* (query) with contig8 DNA sequence (subject). The first 20 bases of the 5' UTR are not conserved.

However, in the UCSC Genome Browser, RNA-Seq data extend far upstream of base 3325 (Figure 26). Several TopHat reads also extend upstream of base 3325. However, significant TopHat reads do not extend upstream of base 3315. For example, JUNC00003813 and JUNC00001300 have scores of 13 and 1 respectively while JUNC00003817 and JUNC00001174 have scores of 896 and 569 respectively.

All instances of the Initiator (Inr) motif on the plus strand occur upstream of base 3325. One instance occurs at bases 3119-3124, right around the upstream edge of significant RNA-Seq coverage. Since the Inr motif is located at -2 relative to the TSS, the Inr motif at base 3119 maps the TSS to base 3121. The second instance occurs at bases 2979-2984, right around the upstream edge of minimal RNA-Seq coverage. This Inr maps the TSS to base 2981. All of this data suggests that the 5' UTR region is longer in *D. ficsisphila* than in *D. melanogaster*.

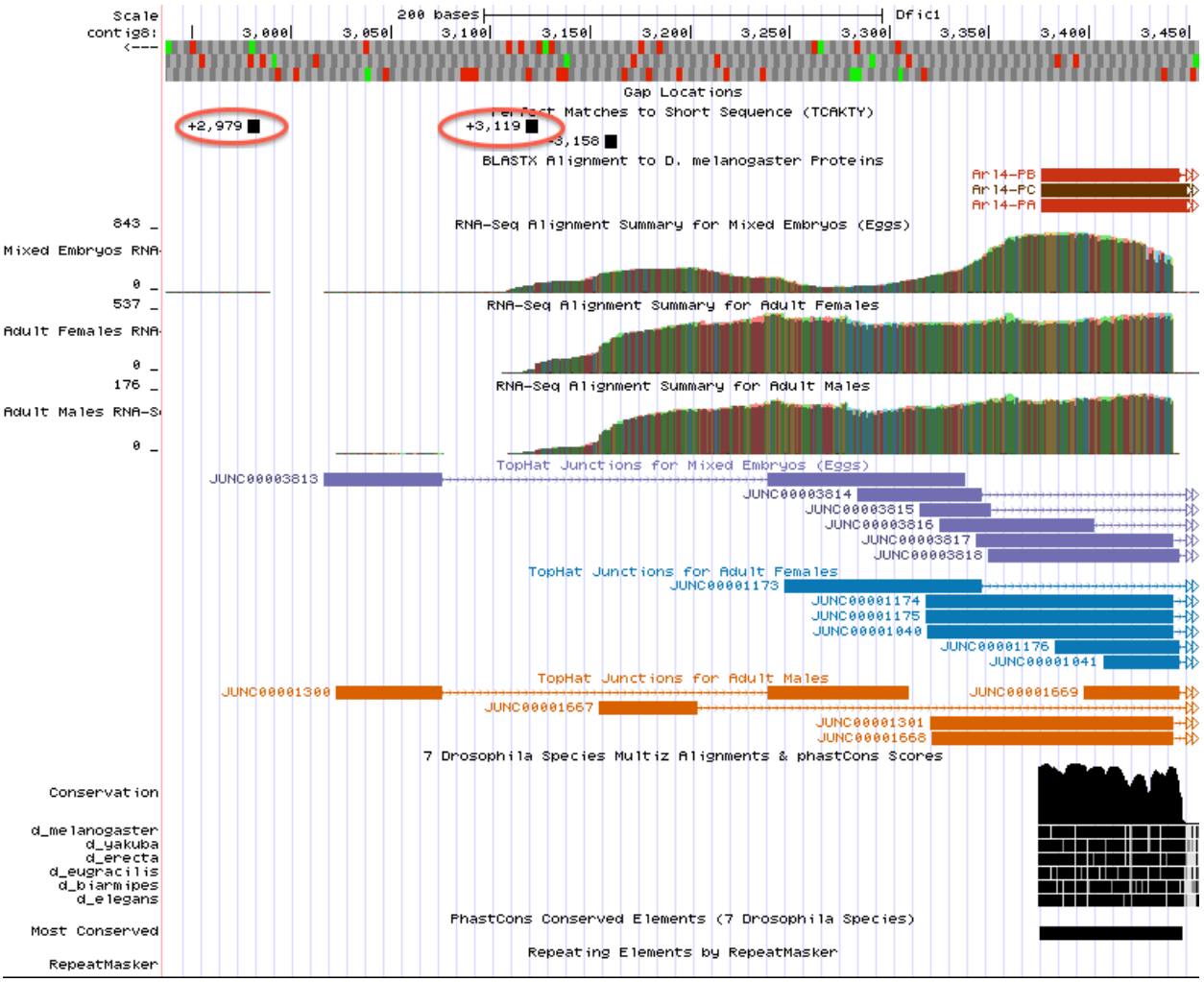


Figure 26: 5'UTR location of *Arl4* on contig8. Two Inr motifs on the plus strand are circled in red. RNA-Seq and TopHat data suggest that the 5'UTR in *D. ficusphila* is longer than in *D. melanogaster*.

In order to further narrow the search region, the positions of other conserved core promoter motifs were investigated in both *D. ficusphila* and *D. melanogaster* (Table 3). A majority of other conserved core promoter motifs do not occur in the expected coordinates on either *D. melanogaster* or *D. ficusphila*. However, the BRE[d] motifs at position +3094 and +3096 give TSSs at position +3117 and +3119 respectively, which closely correspond to the TSS given by the Inr motif at position +3119 (Figure 28)

Motif	Position Relative to TSS	Position in *D. ficusphila*	Position in *D. melanogaster*
BRE^{u}	-38	-	-
TATA Box	-31 or -30	-	-179,981
BRE^{d}	-23	+3094, +3096, +3144, +3203, +3205, +3230, +3249, +3266, +3329, +3331	-179,394 -179,447, -179,629, -179,665, -179,746, -179,916, -179,936, -179,990, -180,060, -180,101
Inr	-2	+2979, +3119	-179,921
MTE	+18	-	-
DPE	+28	+3201	-
Ohler_motif1	NA	-	-
DRE	NA	-	-
Ohler_motif5	NA	-	-
Ohler_motif6	NA	-	-
Ohler_motif7	NA	-	-
Ohler_motif8	NA	-	-

Table 3: Core promoter motifs around bases 3000-3350. The coordinates of the motifs were obtained by using the Short Match tool in the UCSC Genome Browser.

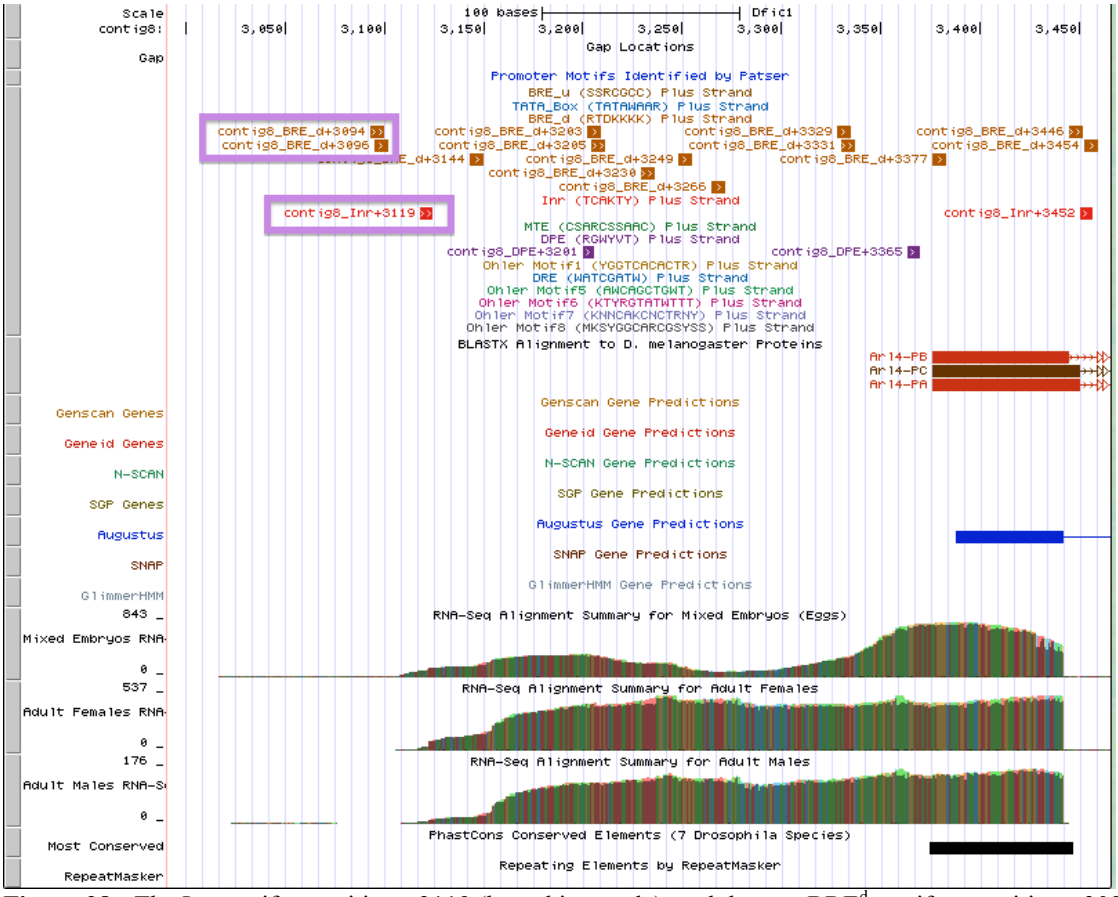


Figure 28: The Inr motif at position +3119 (boxed in purple), and the two BRE^{d} motifs at position +3094 and +3096 (boxed in purple), all converge on a TSS around position +3117-3121, at the upstream edge of the RNA-Seq reads.

A search for regions enriched in RNA Polymerase II (RNA PolII) upstream of *Arl4* in the *D. biarmipes* Aug. 2013 (GEP/Dot) Assembly revealed a region enriched in RNA PolII ChIP-Seq data (Figure 29). However, a BLASTn search using the DNA sequence of this enriched area in *D. biarmipes*, bases 29,300-30,500, (query) against contig8 (subject) did not produce any alignments in the region of interest. Thus, RNA PolII data could not be used to predict the TSS in *D. ficusphila*. The RNA PolII data suggests that the 5' UTR extends over 1000 bases in *D. biarmipes*. This is much longer than the expected length of the 5' UTR in *D. ficusphila*, explaining why the BLASTn search did not produce any significant results.

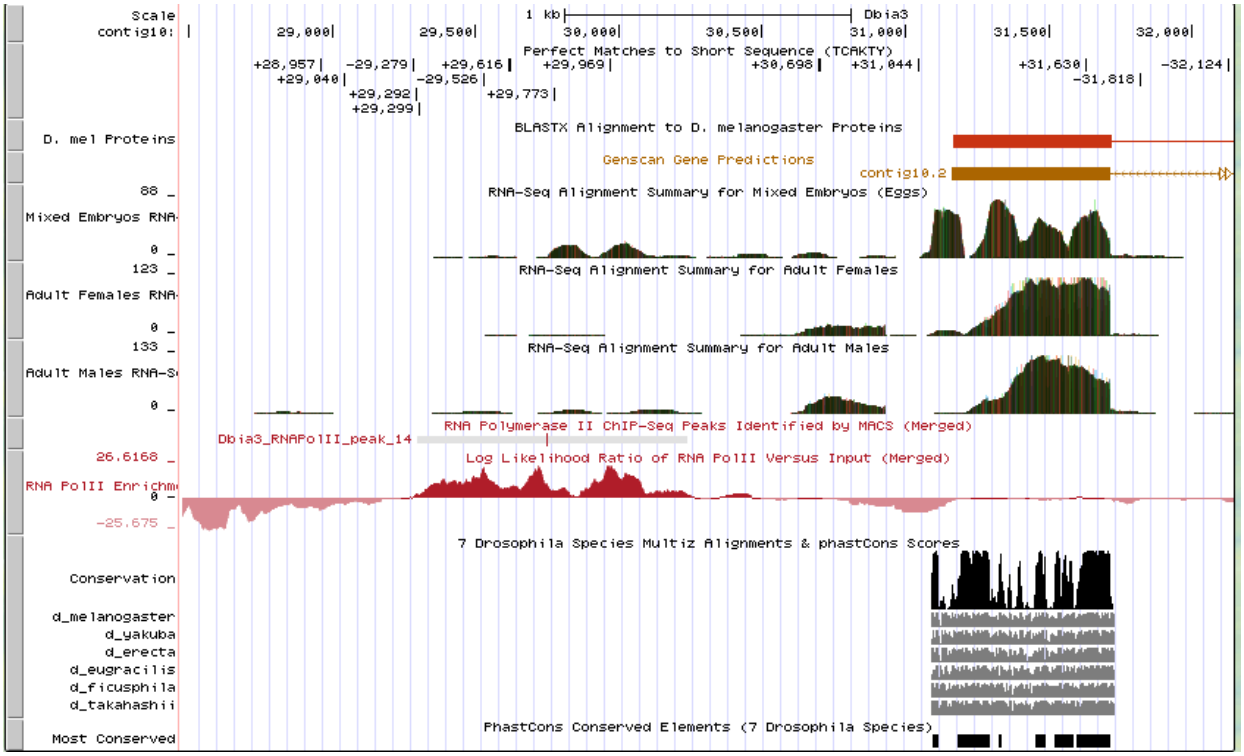


Figure 29: RNA PolII ChIP-Seq data for *D. biarmipes* tissue shows enrichment centered around position +29773. This information cannot be used to inform TSS analysis on *D. ficusphila* because the 5' UTR is not conserved between *D. biarmipes* and *D. ficusphila*.

Based on the information available, the TSS cannot be assigned to one specific site. However, a narrow search can be defined between position +3117-3121 based on data from core

promoter motifs and RNA-Seq data. A wider search region would encompass the minimal RNA-Seq reads up to the Inr motif at position +2979.

Feature 2

Feature 2 was annotated following the same protocol as Feature 1 (Figure 30). The BLASTp search revealed that feature 2 is the *CG31997* ortholog in *D. ficusphila* (Figure 31, 32).

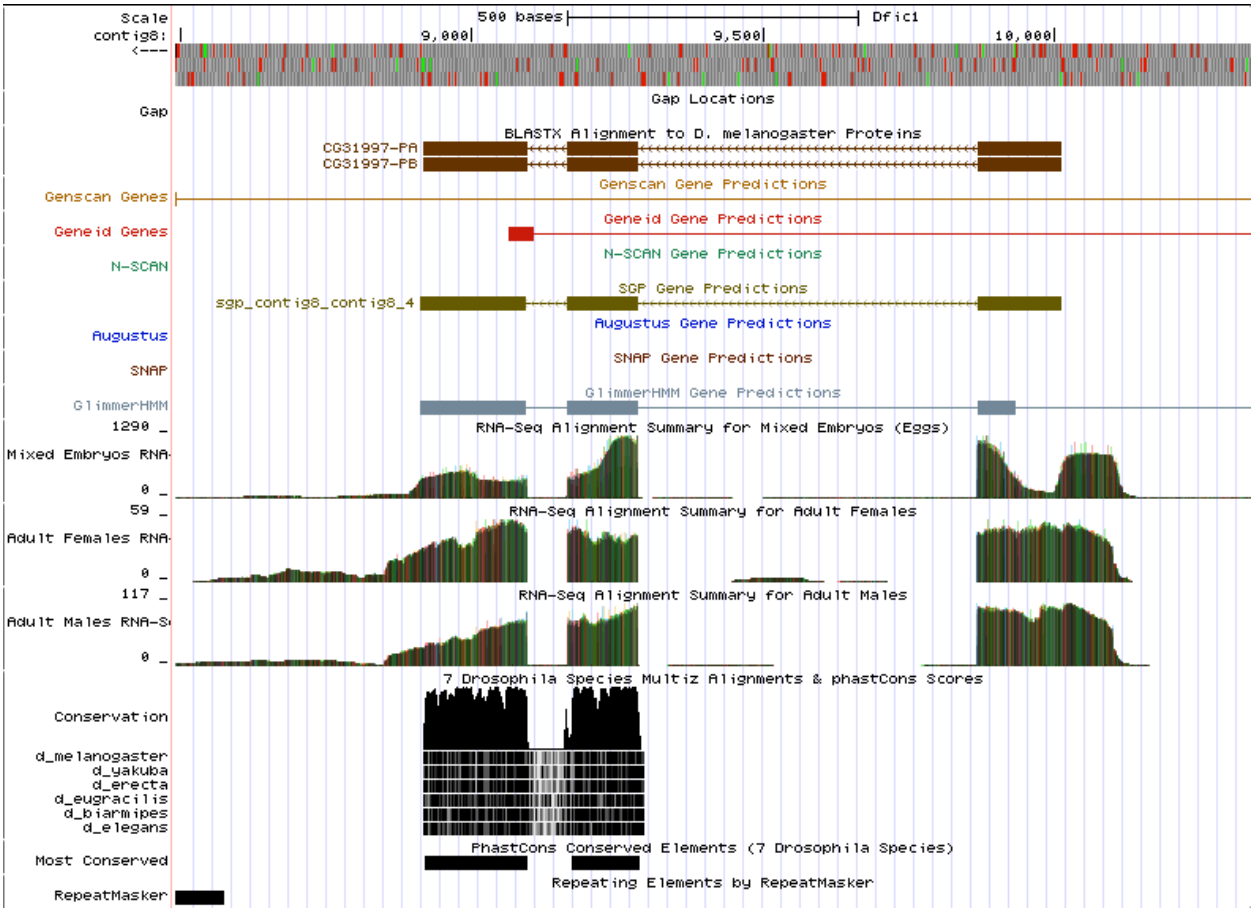


Figure 30: Close up of Feature 2. SGP Gene predictions closely match the RNA-Seq data. Exons two and three are well conserved across many *Drosophila* species.

BLAST Hit Summary				
☑	Description	Species	Score	E value
☑	CG31997-PB	Dmel	239.195	1.06266e-63
☑	CG31997-PA	Dmel	239.195	1.06266e-63
☑	fs(1)Ya-PA	Dmel	28.1054	3.17523
☑	CG18208-PB	Dmel	27.335	5.55341

Figure 31: FlyBase BLASTp results for SGP Gene predicted polypeptide (query) against *D. melanogaster* annotated proteins database (subject). There is high sequence similarity to the A and B isoforms of CG31997 protein.

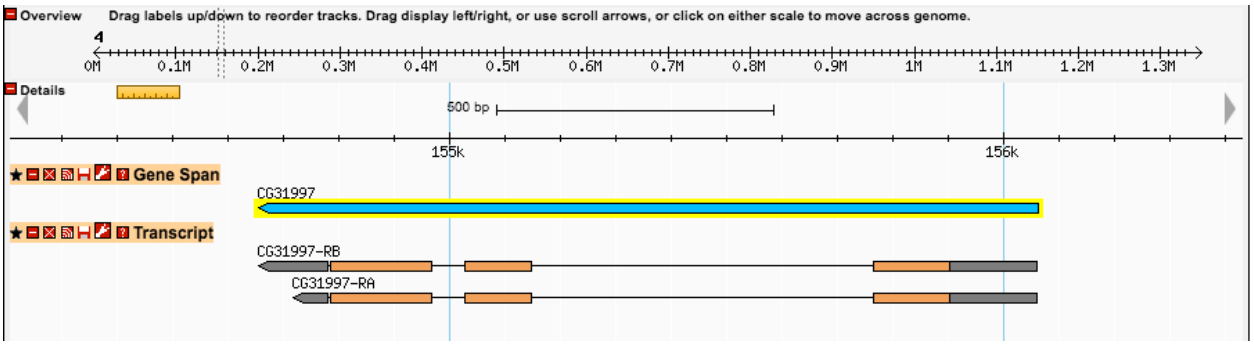


Figure 32: Browser view of *CG31997* (*D. melanogaster* genome). *CG31997* is on the reverse strand and is located on the fourth chromosome. Isoforms A and B have identical coding sequences, but isoform B has a longer 3'UTR.

The approximate BLASTx coordinates were determined by running a pairwise BLASTx search with contig 8 as the query against all of the *D. melanogaster* peptide sequences (Table 3). As for feature 1, the approximate BLASTx coordinates were used to determined exon boundaries.

FlyBase ID	Coding Exon Size (amino acids)	Query Range	Query Frame	Subject Range
1_10861_0	62	10012-9869	-3	1-59
2_10861_2	46	9282-9166	-1	1-46
3_10861_1	39	9093-8917	-1	1-39

Table 4: Summary table for approximate exon locations on BLASTx alignments for CG31997.

Splice sites for all coding exons, as well as the start and stop sites were annotated for isoforms A and B (Table 5). The proposed gene model was submitted to the Gene Model Checker on the GEP website. The dot plot comparisons and protein alignments between the *D. melanogaster CG31997* and feature 2 show that there is significant conservation across all isoforms throughout the entirety of the gene, although as anticipated, there is less conservation in exon 1 (Figure 33).

FlyBase ID	Coding Exon Size	Begin	End	Isoform A	Isoform B	Frame	Acceptor	Donor	Comments
1_9580_0	62	10012	9868	Exon 1	Exon 1	-3	---	Phase +1	
2_9580_0	46	9284	9164	Exon 2	Exon 2	-1	Phase +2	Phase +2	
3_9580_0	39	9094	8911	Exon 3	Exon 3	-1	Phase +1	---	

Table 5: Summary of the proposed gene model for feature 2. Red arrows show corresponding phases between acceptor and donor sites.

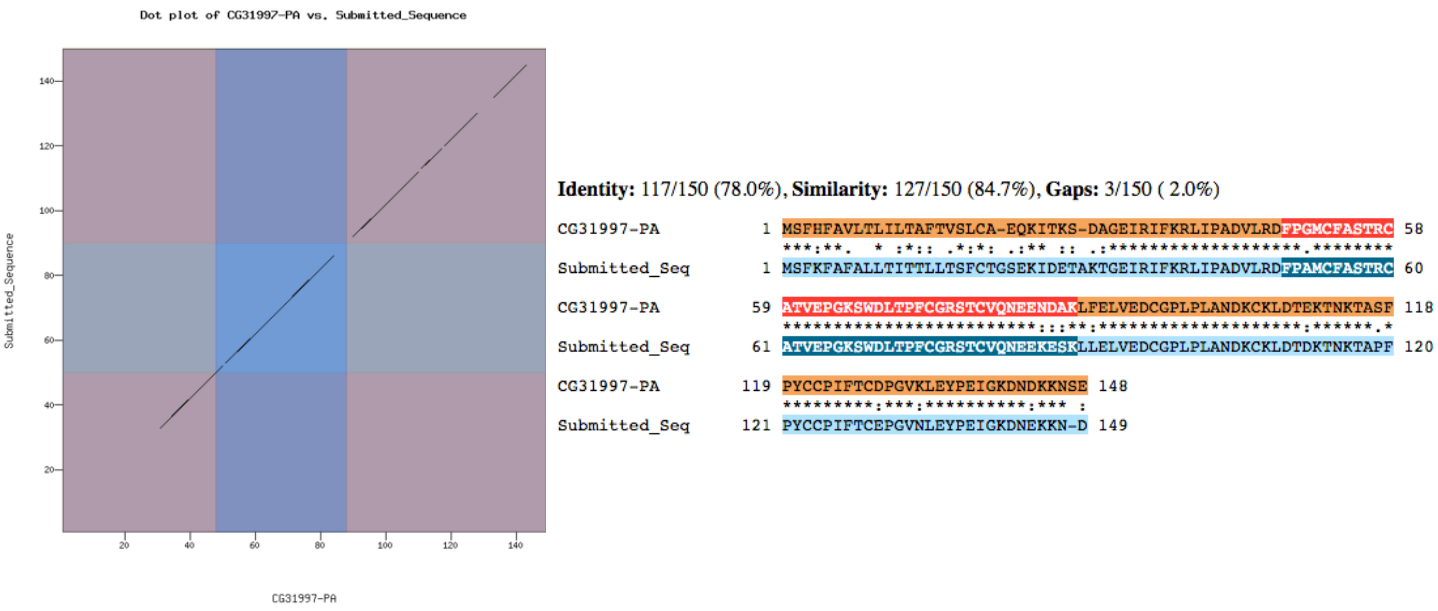


Figure 33: Gene Model Checker output for CG31997 Isoform A. Isoform B has identical outputs because Isoform A and B have identical coding sequences. Left: Dot plot of *D. melanogaster* isoform A vs. feature 2 isoforms. Right: Protein sequence alignment of *D. melanogaster* CG31997 isoform A vs. feature 2 isoforms.

Feature 3

Feature 3 was annotated following the same protocol as Features 1 and 2 (Figure 34). Due to its large size, the protein sequence from GenScan gene prediction contig8.4 was used to conduct a BLASTp search against *D. melanogaster* annotated proteins. The prediction matched to the C and E isoforms of CG33978 (Figure 35). This was expected because the predicted protein sequence used for the BLASTp search is only contained in the C and E isoforms (Figure 36). Based on parsimony, isoforms D, F, and G were also expected to exist in *D. ficusphila*.

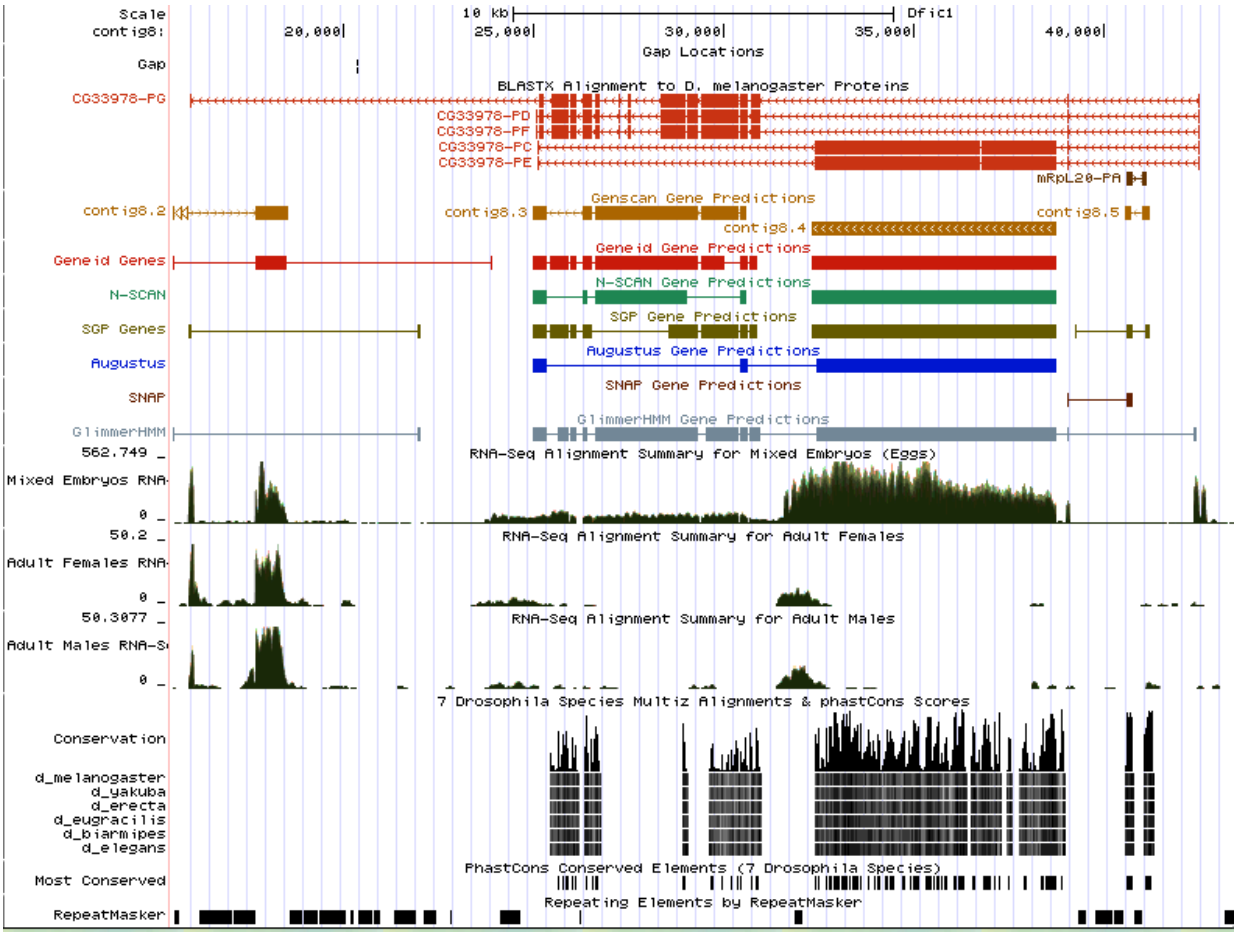


Figure 34: Close-up of Feature 3. All gene predictions broadly match RNA-Seq data from approximately bases 25,000-45,000. This region is also well conserved across many *Drosophila* species.

BLAST Hit Summary				
☑	Description	Species	Score	E value
☑	CG33978-PC	Dmel	1930.99	0
☑	CG33978-PE	Dmel	1930.99	0

Figure 35: FlyBase BLASTp results for GenScan predicted polypeptide contig8.4 (query) against *D. melanogaster* annotated proteins database (subject). There is high sequence similarity to the C and E isoforms of CG33978 protein.

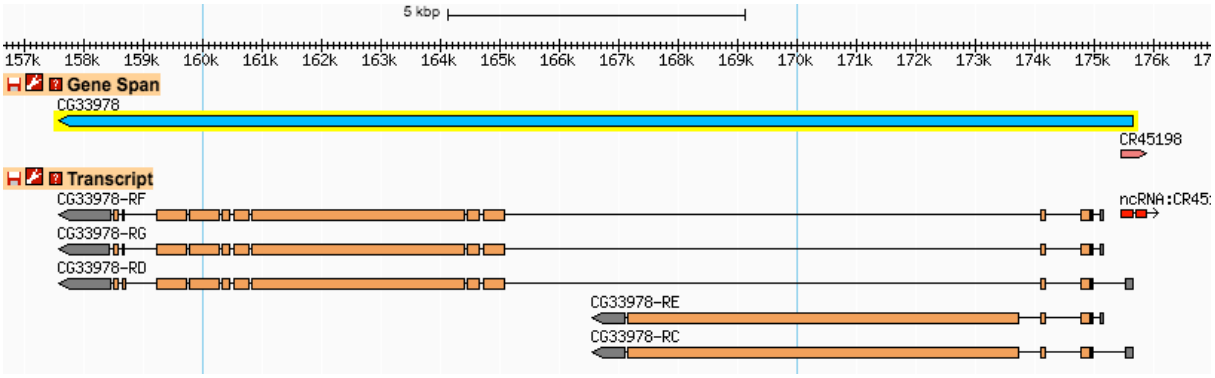


Figure 36: Browser view of *CG33978* in the *D. melanogaster* genome. *CG33978* is on the minus strand and is located on the fourth chromosome. Isoforms D, F, and G span significantly further downstream compared to isoforms C and E.

Exon-by-exon BLASTx searches were then used to find the approximate coordinates for the exon junctions in *D. ficusphila.* Of the fifteen unique coding sequences compiled from the five different isoforms, only eight exons produced alignments (Table 6). In an effort to identify the remaining exons, the subject range was reduced to the corresponding alignment region within the fasta file for each exon. Additionally, within general parameters, the "expect threshold" was raised to ten and "word size" was decreased to two. Within scoring parameters, the "matrix" was changed to PAM30. However, none of these adjustments produced BLASTx matches to these exons within contig8. A majority of these unmapped exons are relatively small and are concentrated in the downstream portion of isoforms D, F, and G. Strangely, there were two BLASTx alignments for CDS6_11602_0, one on the -3 frame that spanned amino acids 1-317 and another on the -1 frame that spanned amino acids 344-1191 (Figure 37). This suggests that a novel intron may have been created within CDS6-11602_0, the fifth and largest exon in isoforms D, F, and G. It was also noticed that the alignments for CDS3_11602_2 and CDS10_11602_0 are significant shorter than expected.

FlyBase ID	**Coding Exon Size (amino acids)**	**Subject Range**	**Query Frame**	**Query Range**
1_11602_0	58			
2_11602_2	17			
3_11602_2	2213	1-2137	-3	38731-32396
4_11602_2	116	17-112	-2	31001-30705
5_11602_0	65	1-61	-2	30635-30459
6_11602_0	1191	1-317, 344-1191	-3, -1	30388-29429, 29325-26614
7_11602_2	79	1-79	-2	26552-26316
8_11602_0	43	1-43	-1	26118-25990
9_11602_2	169	1-169	-2	25922-25434
10_11602_0	163	16-89	-2	25349-25152
11_11602_1	6			
12_11602_0	20			
13_11602_0	26			
14_11602_0	40			
15_11602_0	8			

Table 6: Summary table for approximate exon locations on BLASTx alignments for CG33978 (subject is *D. melanogaster* exons, query is contig 8).

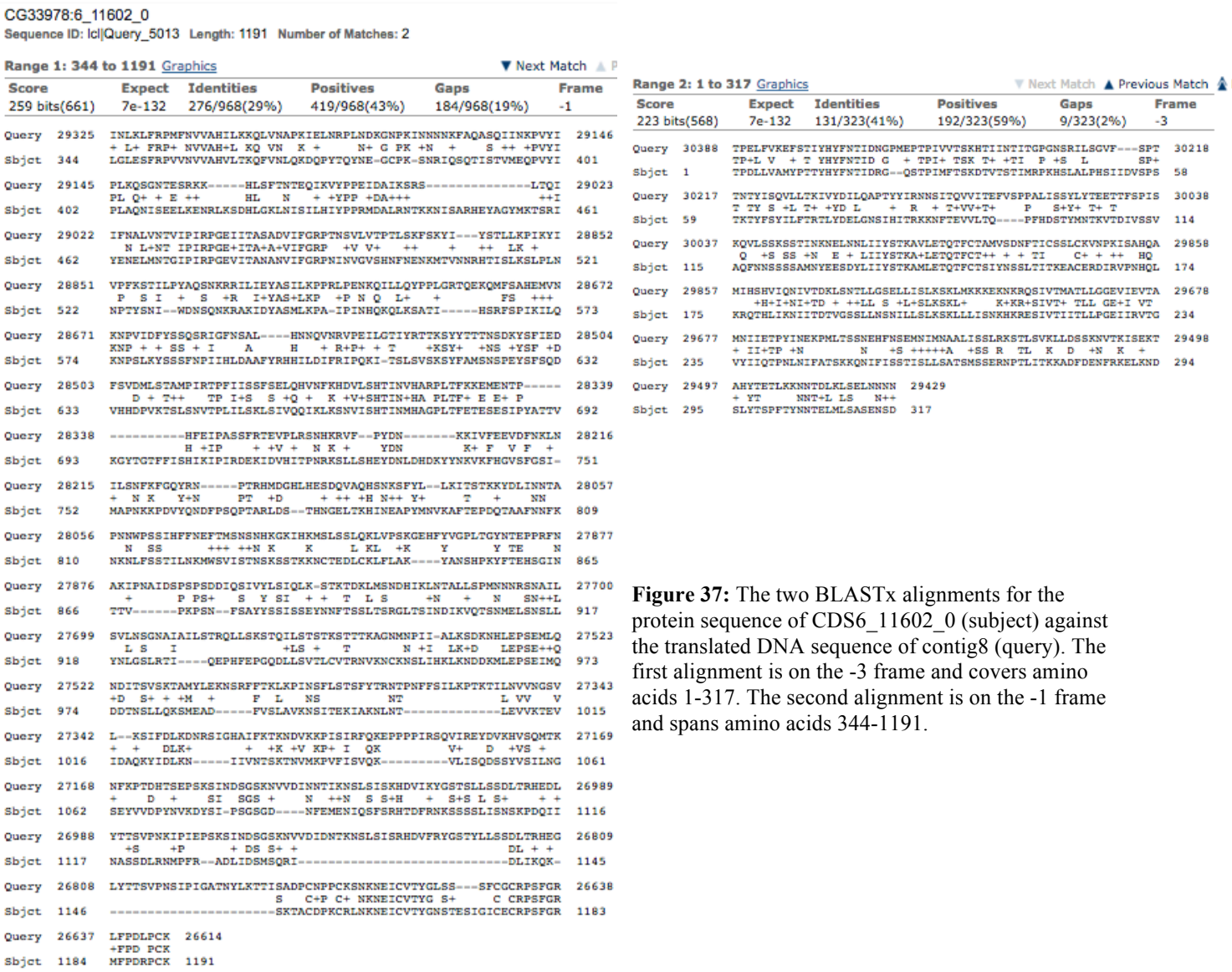


Figure 37: The two BLASTx alignments for the protein sequence of CDS6_11602_0 (subject) against the translated DNA sequence of contig8 (query). The first alignment is on the -3 frame and covers amino acids 1-317. The second alignment is on the -1 frame and spans amino acids 344-1191.

The UCSC Genome Browser was used to closely inspect CDS6_11602_0 for evidence supporting a novel intron. RNA-Seq and gene prediction tracks do, in fact, support this prediction (Figure 38). A TopHat read with a score of 97, JUNC00003850, also supports this prediction (Figure 39). Henceforth, the upstream novel exon will be referred to as CDS6_11602_0A and the downstream novel exon will be referred to as CDS6_11602_0B.

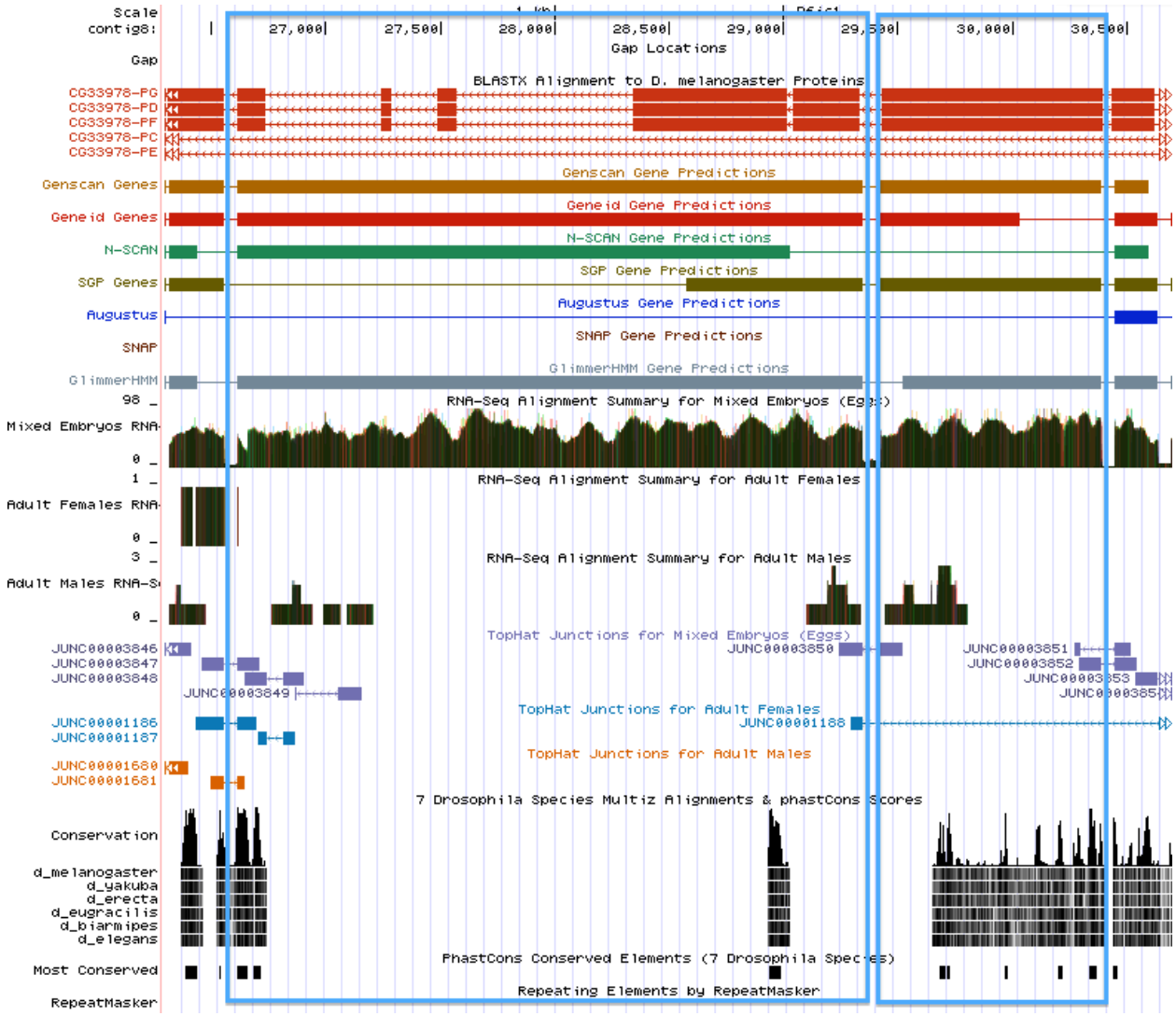


Figure 38: CDS6_11602_0, the fifth exon in isoforms D, F, and G, seems to have acquired a novel intron, splitting this coding region into two parts, boxed in blue. The upstream novel exon will be referred to as CDS6_11602_0A and the downstream novel exon will be referred to as CDS6_11602_0B. Gene prediction and RNA-Seq tracks support this finding.

Figure 39: A significant TopHat read with a score of 97 also supports the prediction that CDS6_11602_0 has acquired a novel intron, dividing the coding region into CDS6_11602_0A and CDS6_11602_0B.

Functional splice sites were identified on both ends of the novel intron. The AG donor site on CDS6_11602_0A was found at bases 20389-20388 on the -3 frame, giving a phase of 0 (Figure 40). The GT acceptor site on CDS6_11602_0B was found at bases 29345-29344 on the -1 frame, giving a phase of 0 (Figure 41). Thus, the AG acceptor site and GT donor site are compatible, further supporting the prediction that CDS6_11602_0 has acquired a novel intron in *D. ficusphila*.

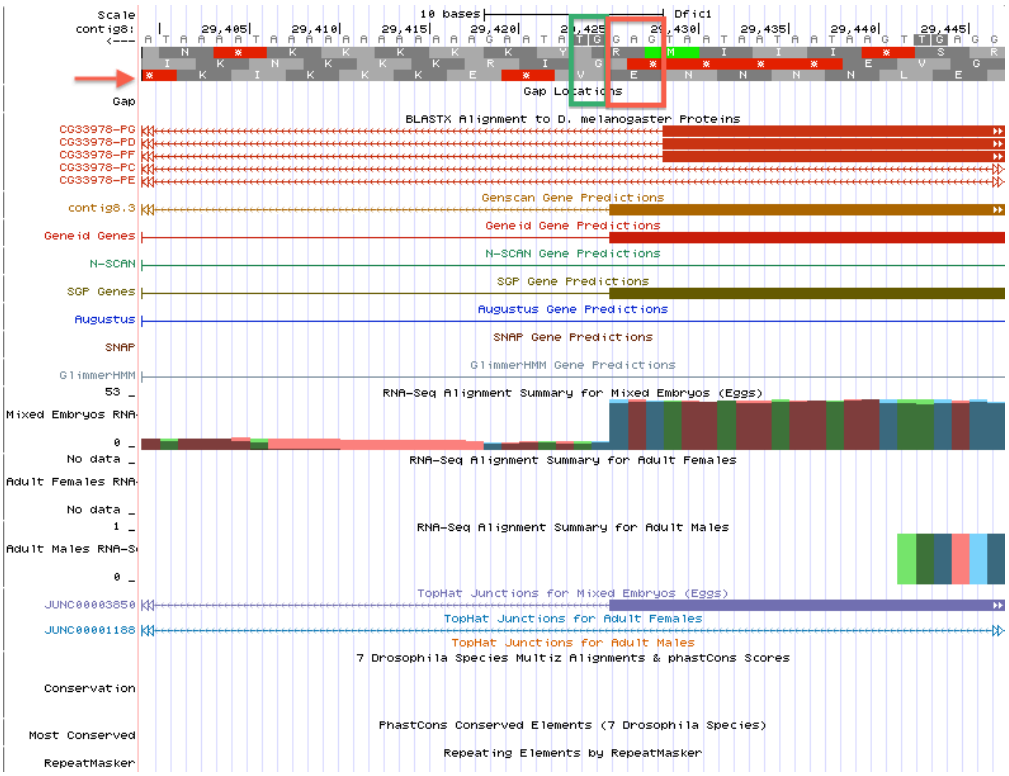


Figure 40: Donor site on upstream portion of CDS6_11602_0. The most likely GT donor site is boxed in green. The closest complete codon is boxed in red. The red arrow points to reading frame -3. The most likely donor site at base 29425-29424 is phase 0.

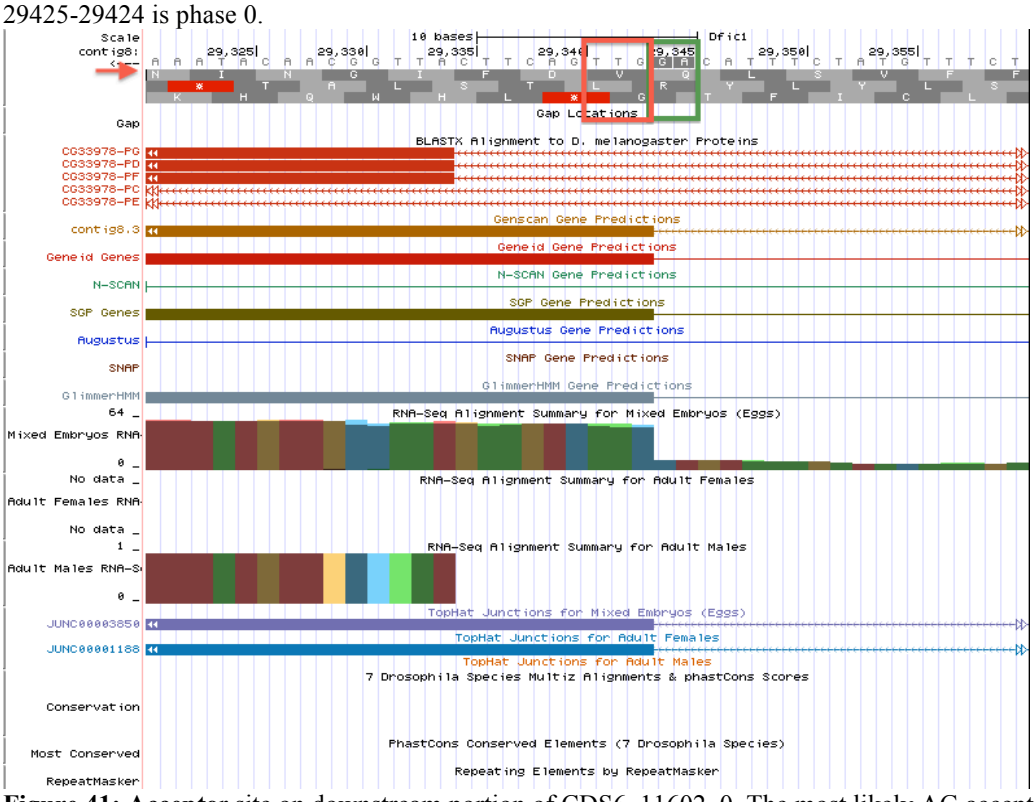


Figure 41: Acceptor site on downstream portion of CDS6_11602_0. The most likely AG acceptor site is boxed in green. The closest complete codon is boxed in red. The red arrow points to reading frame -1. The most likely acceptor site at base 29345-29344 is in phase 0.

By closely examining the UCSC Genome Browser output, CDS1_11602_0 and 2_11602_2, the first and second exons of all isoforms, were annotated despite the lack of BLASTx alignments using RNA-Seq and TopHat data. Close inspection of CDS3_11602_2 revealed a premature stop codon at bases 32,326-32,324 (Figure 42), providing an explanation for the missing seventy-seven amino acids in the BLASTx alignment (Table 6). CDS3_11602_2 is the third and final exon for isoforms C and E. The evolutionary timing of this mutation is unclear due to lack of data available in the conservation tracks. RNA-Seq data continues past the stop codon, suggesting that the rest of the exon still continues to be transcribed as a 5'UTR for isoforms C and E in *D. ficusphila* (Figure 43).

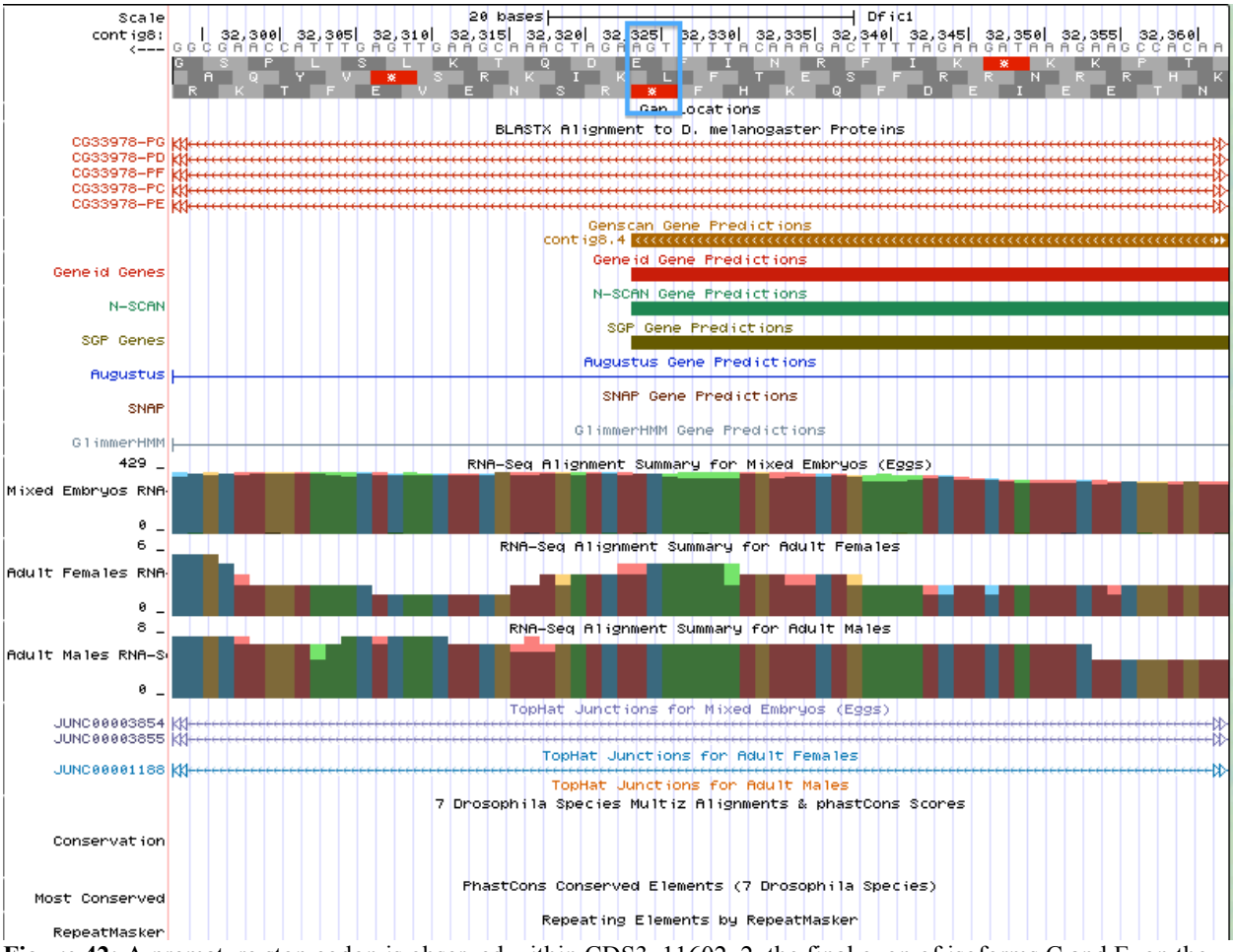


Figure 42: A premature stop codon is observed within CDS3_11602_2, the final exon of isoforms C and E, on the -3 frame at bases 32326-32324.

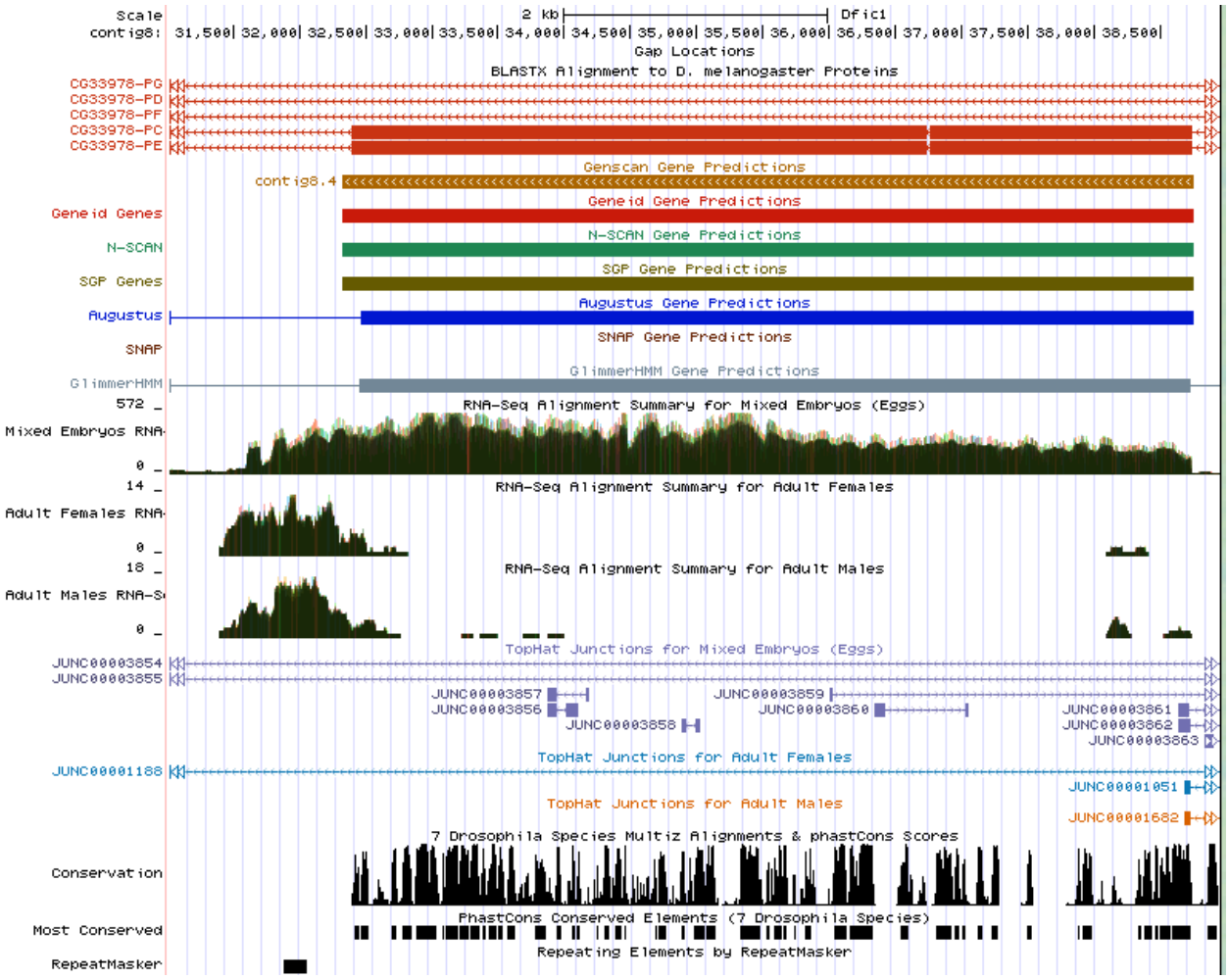


Figure 43: RNA Seq data extends past the premature stop codon at bases 32,324-26, suggesting that the rest of the exon still continues to be transcribed as a 3'UTR for isoforms C and E in *D. ficusphila* mixed embryo RNA.

Similarly, close inspection of CDS10_11602_2 revealed a premature stop codon at bases 25010-25008 (Figure 44), providing an explanation for the missing seventy-five amino acids in the BLASTx alignment (Table 6). CDS10_11602_2 is the ninth exon for isoforms D, F, and G. The evolutionary timing of this mutation is unclear due to lack of data available in the conservation tracks. RNA-Seq data continues past the stop codon, suggesting that the rest of the exon still continues to be transcribed as a 3'UTR for isoforms D, F, and G in *D. ficusphila*. No further splice sites are indicated in the RNA-Seq or TopHat data tracks, indicating no further

downstream exons. This explains why none of the coding sequences downstream of CDS10_11602_2 mapped to contig8 in the BLASTx search.

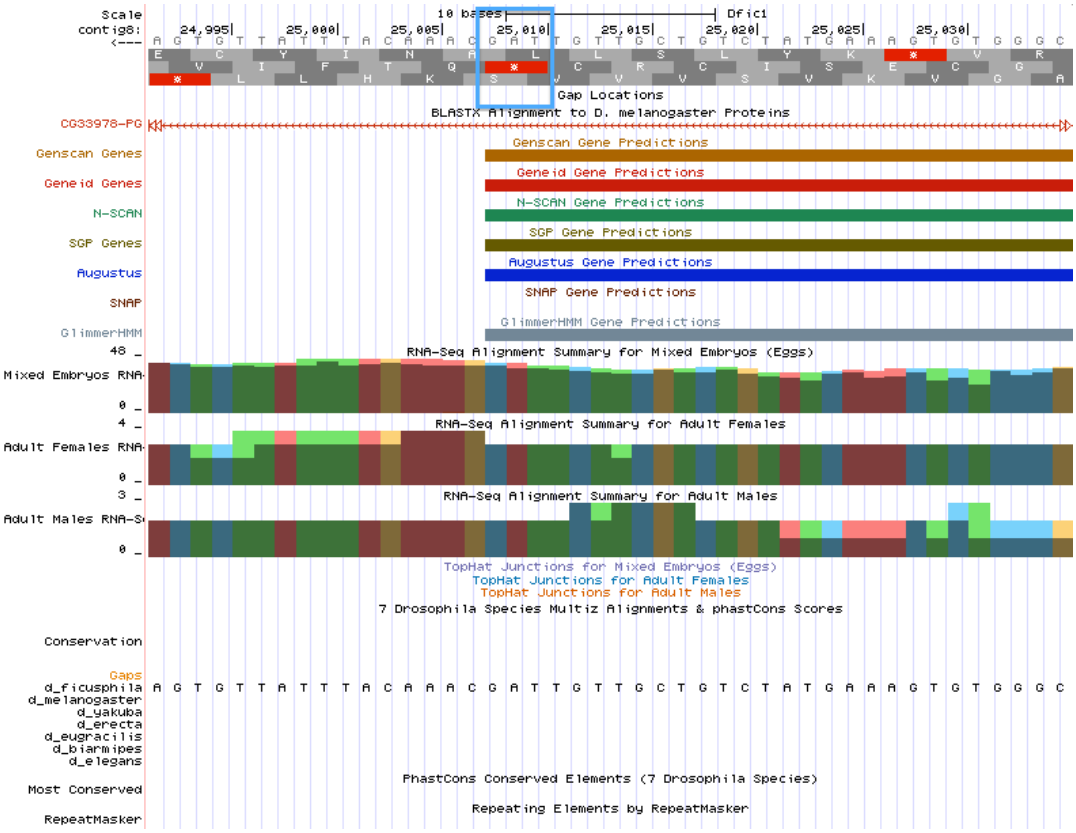
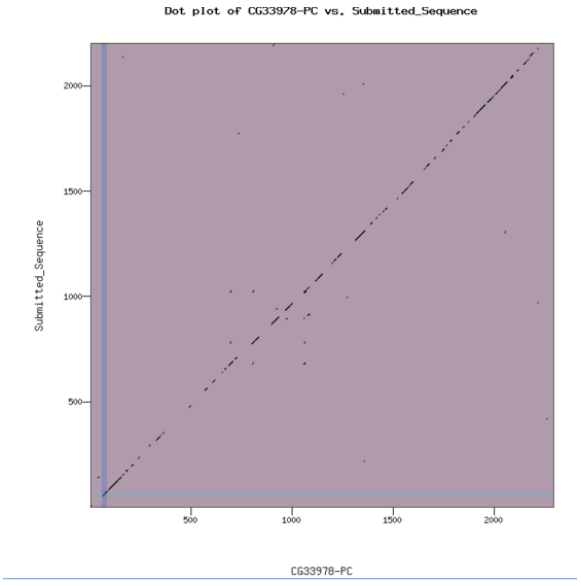


Figure 44: A premature stop codon is observed within CDS10_11602_2, the ninth exon for isoforms D, F, and G, on the -2 frame at bases 25010-25008.

Splice sites for all coding exons, as well as the start and stop were annotated for isoforms C and E and isoforms D, F, and G (Table 7). The proposed gene models were submitted to the Gene Model Checker on the GEP website. The dot plot comparisons and protein alignments between the *D. melanogaster CG33978* and feature 3 show that there is limited conservation throughout the entirety of the gene, particularly for isoforms D, F, and G (Figure 45 and 46).

FlyBase ID	CDS Size	Start	Stop	Isoform C	Isoform D	Isoform E	Isoform F	Isoform G	Frame	Acceptor	Donor
1_11602_0	48	42506	42362	Exon 1	Exon 1	Exon 1	Exon 1	Exon 1	-2	---	+1
2_11602_2	18	39084	39031	Exon 2	Exon 2	Exon 2	Exon 2	Exon 2	-3	+2	+1
3_11602_2	2137	38733	32324	Exon 3		Exon 3			-3	+2	---
4_11602_2	106	31009	30693		Exon 3		Exon 3	Exon 3	-2	+2	+0
5_11602_0	63	30635	30447		Exon 4		Exon 4	Exon 4	-2	+0	+0
6_11602_0A	321	30388	29426		Exon 5		Exon 5	Exon 5	-3	+0	+0
6_11602_0B	910	29343	26613		Exon 6		Exon 6	Exon 6	-1	+0	+1
7_11602_2	80	26554	26316		Exon 7		Exon 7	Exon 7	-2	+2	+0
8_11602_0	43	26118	25989		Exon 8		Exon 8	Exon 8	-1	+0	+1
9_11602_2	160	25924	25446		Exon 9		Exon 9	Exon 9	-2	+2	+0
10_11602_0	121	25370	25008		Exon 10		Exon 10	Exon 10	-2	+0	---
11_11602_1	Does not exist in *D. ficusphila*										
12_11602_0	Does not exist in *D. ficusphila*										
13_11602_0	Does not exist in *D. ficusphila*										
14_11602_0	Does not exist in *D. ficusphila*										
15_11602_0	Does not exist in *D. ficusphila*										

Table 7: Summary of the proposed gene model for feature 3.

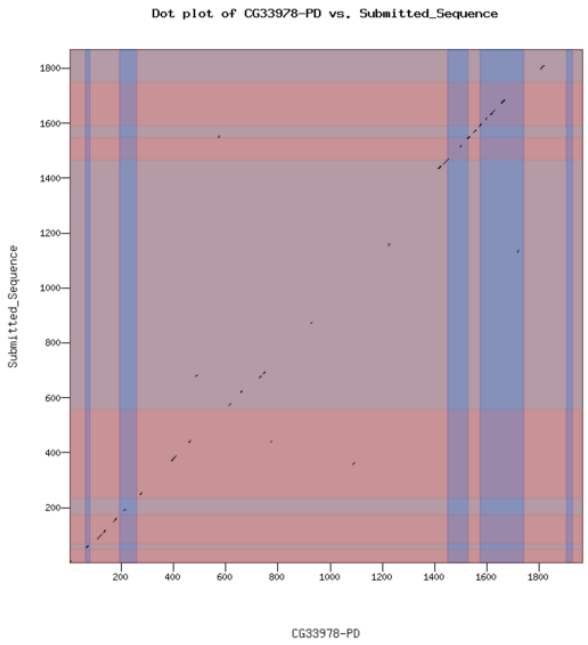


Identity: 1307/2338 (55.9%), **Similarity:** 1632/2338 (69.8%), **Gaps:** 185/2338 (7.9%)

```
CG33978-PC          1 MLQPREDNKISDIFHKKYDTSTSTRRMMYSTYYNFNLSRVLFTTLLFTMMHIETSHAINQ 60
Submitted_Seq       1 MLQPRKTHRIISPCLRNHLSFQFRTRMVIT----------LLNILILITVNTAASQTSNQ 50

CG33978-PC         61 SYHVSPTPTLPAISGDDITTVLLVNNESGHMGDFHGRFLTVRPEIGVLKSTARTFIQDGI 120
Submitted_Seq      51 PYLVSPTPTLPAKVVDEITTVLLLNNESGHVGDIHGRFLTVRPEIGLLTSTARTFIQDGI 110

CG33978-PC        121 TTEFATKIVGTTLNNGRLYAQYLKKSSRVLYENENISPSVVTSWVGEEDSLQTPLVLQSH 180
Submitted_Seq     111 TTEFATKIVGTTLSNGRLYAQYIKKSSRVLFENEHLTPSVVTSWVGERASQQTPQFLQSH 170

CG33978-PC        181 NDLFNIDDSNWQDIDDSIGAHQREFVGNTDFVNEQNTKNKVLEAPALQDKKTISYMGS-- 238
Submitted_Seq     171 NDLFNSDDADWQNVDDNLGVQQVEFVGNTDFITPQTTIVKYSMTPALKENKPITSLESVI 230

CG33978-PC        239 DSNDTLRTDFKKEDEINRLI--EVLSIKHLETYTVKMPFGNYPTSSLNQSHLDH---RKE 293
Submitted_Seq     231 TSNDTLNFNLNGIIEKNKLISDKVFSITNLKTYTIKSNIEDIPNSAFQESATDSYALREE 290

CG33978-PC        294 RKYEHLMQPHTEKKNEDHNQKSRSDHKQVLASVTYYGFADFTTIVGDSVIVFSPSTTQSI 353
Submitted_Seq     291 RKYEMLRQP-------SQGHQSLSTQKPFLATVTYYGFAEFTTIVGDSVIVFSPSTLESS 343

CG33978-PC        354 LHYGHVTSIKGKPTLQHDLIPEEIAQTSKISLSTLKMGTSSGKYEEFVSNISASKMFDKN 413
Submitted_Seq     344 MHFGQVTSIKGNPTLQPDFIHDYHSQNVEITQSSLETVETLEQIELLSKKSSLGSDISAF 403

CG33978-PC        414 ELKNLS--VLKPTNPSPTFAN-NSFNEEQEASSVLSKDIVPFVSKAFNDIISPSFVNGVK 470
Submitted_Seq     404 TIENQSNDILTPQSTASLETNFKSLLVEEAAGNILIPDVLSS-STAKNDLKS-------- 454

CG33978-PC        471 ECTKHNEPSIIKQEGATTVFIDDDPFTSFDDLSGSTSLKLNSANEIKSIKSFEMSKSDEI 530
Submitted_Seq     455 ESTNDNN-NLEKLEGAKTVYIDDDPFSRFVELSDSSSSKSEATDQDSTLNYIHLSNNEDI 513
```

Figure 45: Gene Model Checker output for CG33978 Isoform C. Isoform E has identical outputs because Isoform C and E have identical coding sequences. Top: Dot plot of *D. melanogaster* isoform C vs. feature 3 isoforms. Bottom: Protein sequence alignment of *D. melanogaster* isoforms vs. feature 3 isoforms.

```
CG33978-PC        531 REIVAYQTNINF---TKDSLDINPLFETSN-ENHDFEELCNHTTSQVFLTQMSKANISNE 586
Submitted_Seq     514 KNGTLKTSTVDSDIKNMHSLTEEPIVGTGNIYDTNIKDTCSHTTSQVFLTQMPKP-VQTS 572

CG33978-PC        587 KNSMDTIDIDANPLHAYEIVETTKYYCIQASQSKQTPELDII-VTHETFTKSRGEFVSTY 645
Submitted_Seq     573 NGNIEPIAINDNPLLSYDIVETTKFYCIQASKAEQESGLSTINESHETVTESNGELMSTE 632

CG33978-PC        646 SIDDDIDETTVQDLGDYEGTTENEYESEEEEYESVSDEVDLIYKTLYTTYTYLTTFEGS 705
Submitted_Seq     633 LIDD-VDETTVRGTDDYDVTTENEYEGEDDDYENNADDVDLIYKTLYTTYTYLTTFEGS 691

CG33978-PC        706 RTTISSHTEVLTNIVSSTLLPKAKLKSNTLNYVGENQISDNTIRLIDHKTTEFSQKYSIP 765
Submitted_Seq     692 SSTVASHTEILTNIVSSTLGAEGELKSSNL----EGQFEKASIVL---------PKYTIP 738

CG33978-PC        766 DEIASLLIQESKVNTAESSSQMMTTLR--DDLIYSKTLLTTYTYYTSIFTNNETEVQSRT 823
Submitted_Seq     739 LDIAGLLTKKSTQNDTEVENTISTSTSNFDDLKYSKTLFTTYTYYTSIFKNNDTEVQSRS 798

CG33978-PC        824 EIVTNYITDNVSNNETNFI--RNPQTANLNFLKSQHSLVEEKIDKVVNMAYGMSVNSSML 881
Submitted_Seq     799 EIVTNYITDEMPNSQQNIIISGDSQTSDQSFLSSQ-TLSD------VDMSLGMELKSSSS 851

CG33978-PC        882 RDSSPASFINQNLLDDQVSSESNTEEIIPSATFLLQTSFTTFTFYTTMYVCNDTSIISRH 941
Submitted_Seq     852 NYTASDLSITTNSLEDQVSSESNTEEIIPSATFLLQTSFTTFTFYTTMYVGEDSNIISRL 911

CG33978-PC        942 ETITNVATEALHPTKLISVEDYSLPITYFTTPTYWTKLAKDGEITTLSREETLSNVIQHS 1001
Submitted_Seq     912 ETVTNIATETLKPTKVV-VDDSSFPITYFTTFTYWTKLAKDGEITTLSREETLSNVIQPT 970

CG33978-PC       1002 NGTTLTVNDLKTADKVSSQSTATDSDTENDRNCKSESSLFSTLENQIYQSDITTFYTTYT 1061
Submitted_Seq     971 NVTEIVIDDSSNTEVEASQSTISTKS-------ESNSDYISNSTNADIRSETTTYYTTYT 1023

CG33978-PC       1062 YYTTSYDVNTTIIDSRLETVTNIITSREMSNPMEETDATILMIKPSQPIAYYENPIETIK 1121
Submitted_Seq     1024 YYTTSYEVNNTIIDSRFETVTNLVTAGEVTISASTLD---MMPKSSQSESATENPIETKA 1080

CG33978-PC        1182 GKIPDDSLN---LYKTGLVRLIEGKRIQNSTTTLYQSRVIGTLIENRYAQIIESTSSFFF 1238
Submitted_Seq     1141 SNTPDNSLSSLKMHKTGLVRLIEGKRVQNSTTTLYQSKVIGTVIDNRYAQIIESTSSFFF 1200

CG33978-PC        1239 EKTKSEHILFPTSVEISGMITKNVDLIDVIESTKSEN-DTTVAPYIDQKNTKPESD-NTD 1296
Submitted_Seq     1201 EKYTPTDYLVSSTVQIPVMATANGNFLNSTNNINTESQDATVS------DIKARDDLGTD 1254

CG33978-PC        1297 GLSNHSPQNAKRPFAPVIRPFASRNRPTFAPKQKTLVPSSATIITRSDITPTITATPALK 1356
Submitted_Seq     1255 GLA---PQNTKPLFAPVIRPFASRNRPTFAPKQKTLVPSSATIITRSDITPTITATPALK 1311

CG33978-PC        1357 AVGRYTASRRGIISNVPINLNEPNLSQSSQSSRRLFGRPSKSMYDLIEANSTLQSSNAFS 1416
Submitted_Seq     1312 SAGKYSSSRRGVISNAPINPNEFQSLSQSQQASRRLFGRPTKPLSNSAEANGTLPFNIAFS 1371

CG33978-PC        1417 PSKTRPTSPLRPGVSSRRPNINISYRSSSFPVFRGSGVIANTRLRIKPTSIGIQSSYRPT 1476
Submitted_Seq     1372 PSKNRFVSSSRNVVSSRRQSIN--YRSSNVPGFRGSGIIANSRLRIKPTNSVLVS----T 1425

CG33978-PC        1477 QSTTFATESSSEDSVEEDTSPDEESDVEEGIKNNNNSPLLRFRRPINAPSGFSPPIRQT- 1535
Submitted_Seq     1426 QS--FSKVPSSDSSSEEENSTDEINEEDESPKHTNN-PLTRFRRPLSGPNAFTPAIRQSA 1482

CG33978-PC        1536 GNSPVVSLRRNPLSPRAKISTTTSSTTTTTAKPRARSFQRPIG-PHPRSRPQNTLFPPRG 1594
Submitted_Seq     1483 GNSPAVSLRRNPLSTRSKTSTTTSSTTTTTTKPRTRSFQRPLPFQSARSRPQNSLFPPRG 1542

CG33978-PC        1595 LFQTQIKNENLKVPRTNDSVPNNDFEYE----GGVEDQDLDENIHRRKRSIKYSVRFTTG 1650
Submitted_Seq     1543 LFQSLKKEESSKEVKPNDSATENDPEYVDDDVGNAEESSPQENVYRIKRSVGTLTNPKRN 1602

CG33978-PC        1651 NRVRRQADTVRRNRFRFRRQNQTAT-TKVESKLMDSDNNTEPILTPRDKSSSRFGSRFHS 1709
Submitted_Seq     1603 SRVRRQADALKISRFRFRRQNQTATLTKEDSKLFGPEVRTETTPTSREKISSRFGSRFRS 1662

CG33978-PC        1710 PQVHQLYTQALIAESPTATNHRSIRPSRPTTKRPQFTLREKDANLKG----TASKNFRRQ 1765
Submitted_Seq     1663 PQGQQ----TTLLESTTVANHKSIRPTRPITKRPQFTLREKDVTLKGPSRSSTSNNFRRQ 1718

CG33978-PC        1766 PASGNS---RRPVGNSGNSNSRRLKSYVSYN-KNNLDNGRPSSSPRSRNSNSNTLNRGRS 1821
Submitted_Seq     1719 QATGNASIRRAGTSNTGNSNSRRLKGLVSYNNKNNGDHGRLSNVQRSRSSSSNTLNRGRS 1778

CG33978-PC        1822 RGRNRVDYASDVQSVENDLQTITVTHYIPSEVTVPVVSGHITEYKSIVTAKSSTEIVGPN 1881
Submitted_Seq     1779 RGRNRNEFSPDLQSIEHEVHSITVTHFIPSELTIPVVNGHVTEYKNIVTAKASTEVVGPD 1838

CG33978-PC        1882 DFAVVLGSNGISSTYLNRETSIINIAGATEHTKYLLHESITSSVTFTPTTIRGRKTSFSH 1941
Submitted_Seq     1839 DYMQVLGTNGLSSIYLNREISSINIAGATEHTKYLLHESITSSIITPTTIRGRKTSFSH 1898

CG33978-PC        1942 IIPSTAYSVEYIVTTMQPQISENAPLANILLSQLLLGNLNLPAHPLIGAVGQQNVPSTII 2001
Submitted_Seq     1899 VVPSTAYSVEYLVTTIQPQIPDNAPLANILLSQLLLGNLNLPAQPILGAVGQLT-SATLA 1957

CG33978-PC        2002 PTSVEPITEYRTHTSTYVTTIFDGKSTILPVTFQGKKILTTVYDTTAQTITATEYSVDTI 2061
Submitted_Seq     1958 PTSVEPVTEYRTHTSTYVTTIFDGKSTILPITFQGKKILTTVYDTTAQTITATEYSVDTI 2017

CG33978-PC        2062 INTLPSVQSVHTVGPAAHVNNLLLQQLLIQQQHE--SLVQAISNPTPPQVLLSENLQVLD 2119
Submitted_Seq     2018 VNTVPLPYNTQSIGQAAQVNNLLLQQLLQQQQDGFSLPQAVSKTVSPQIFLSENLQDLE 2077

CG33978-PC        2120 DGTRTSIKSDAIDIEYDSDLVSVSNESQPTKGHRKKSRKSSSGNKRSKQ--------EPS 2171
Submitted_Seq     2078 EGSRLSKQTD-VDFENDSDFMSISSESQPTKSNRKKSRKTSSGHKRSKQRNVAVEPQDPS 2136

CG33978-PC        2172 VITLYVSGRRPGEFSTVLSTVKNEHDHSVSLQKRHAVLEIQPTNSIKYVSSIERSKNIME 2231
Submitted_Seq     2137 VVTLYVSGRRPGEFSTVLSTVGSEYDHSAALQKREAFFEIQPT-----VTSMDHY----- 2186

CG33978-PC        2232 KNDILRDIRSSLTIDIGFNDLPDQTASLESIVGDVDLWFANSNRQSSTVTTFIKSVVV 2289
Submitted_Seq     2187 ----------------------------------IQFPNTEE----IEDFQKHF-- 2202
```



Identity: 715/2094 (34.1%), **Similarity:** 1040/2094 (49.7%), **Gaps:** 354/2094 (16.9%)

```
CG33978-PD          1 MLQPREDNKISDIFHKKYDTSTSTRRMMYSTYYNFNLSRVLFTTLLFTMMHIETSHAINC  60
Submitted_Seq       1 MLQPRKTHRIISPCLRNHLSFQFRTRMVIT----------LLNILLILITVNTAASQTSNQ  50
CG33978-PD         61 SYHVSPTPTLPAISGDERNQLERHSFKEHRETKKFNTSYELKNENNFLHGSISEKVLDEA 120
Submitted_Seq      51 PYLVSPTPTLPAKVVDGHNQINV-SYR----TGRINSRYVL------LHGSISEKIWDEA  99
CG33978-PD        121 RYNLATHLMSNGVELIITNEKEKYIENIGLTTQKVTSNNSQPILITIRSDKPLTLAPSTL 180
Submitted_Seq     100 RYNVATHVMPNGVELIIANEKENSFEKLAVTSKKLVSYNSKPLLITKQFNKPLTLAPSTL 159
CG33978-PD        181 KNQLTLSLPPIRKDPPSIPKNVFVRKICLTTYTYHTTNVKNGSFHLVSKKVVITNSFTED 240
Submitted_Seq     160 EDQMTLTLPLKKQDTVLLPDNSFTTKMCPTTYTYRKTLMENGNMNVVSKDLVISNRLTED 219
CG33978-PD        241 RNYLQASNSMIADVTFSKTPDLLVAMYPTTYHYFNTIDRG--QSTPIMFTSKDTVTSTIM 298
Submitted_Seq     220 R--MTTSDRTISDIISFKTPELFVKEFSTIYHYFNTIDNGPMEPTPIVVTSKHTIINTIT 277
CG33978-PD        299 RPKHSLALPHSIIDVSPSTKTYFSYILFTRTLYDELGNSIHITRKKNFTEVVLTQ----P 354
Submitted_Seq     278 CPGNSRILSGVF---SPTTNTYISQVLLTKIVYDILQAPTYYIRNNSITQVVITEFVSPP 334
CG33978-PD        355 FHDSTYMNTKVTDIVSSVAQFNNSSSSAMNYEESDYLIIYSTKAMLETQTFCTSIYNSSL 414
Submitted_Seq     335 ALISSYLYTEETTFSPISKQVLSSKSSTINKNELNNLIIYSTKAVLETQTFCTAMVSDNF 394
CG33978-PD        415 TITKEACERDIRVPNHQLKRQTHLIKNIITDTVGSSLLNSNILLSLKSKLLLISNKHKRE 474
Submitted_Seq     395 TICSSLCKVNPKISAHQAMIHSHVIQNIVTDKLSNTLLGSELLISLKSKLMKKKEKNKRQ 454
CG33978-PD        475 SIVTIITLLPGEIIRVTGVYIIQTPNLNIFATSKKQNIFISSTISLLSATSMSSERNPTL 534
Submitted_Seq     455 SIVTMATLLGGEVIEVTAMNIIETPYINEKPMLTSSNEHFNSEMNIMNAALISSLRKSTL 514
CG33978-PD        535 ITKKADFDENPRKELKNDSLYTSPFTYNNTELMLSASENSDIYIRTKNKLIHSRLSSATS 594
Submitted_Seq     515 SVKLLDSSKNVTKISEKTAHYTETLKKNNTDLKLSELNNNN------------------- 555
CG33978-PD        595 KTSAIKILGLESFRPVVNVVAHVLTKQFVNLQKDQPYTQYNE-GCPK-SNRIQSQTISTV 652
Submitted_Seq     556 EVDFIGNINLKLFRPMFNVVAHILKKQLVNAPKIELNRPLNDKGNPKINNNNKFAQASQI 615
CG33978-PD        653 MEQPVYIPLAQNISEELKENRLKSDHLGKLNISILHIYPPRMDALRNTKKNISARHEYAG 712
Submitted_Seq     616 INKPVYIPLKQSGNTESRKK-----HLSPTNTEQIKVYPPEIDAIKSRS----------- 659
CG33978-PD        713 YMKTSRIYENELMNTGIPIRPGEVITANANVIFGRPNINVGVSHNFNENKMTVNNRHTIS 772
Submitted_Seq     660 ---LTQIIFNALVNTVIPIRPGEIITASADVIFGRPTNSVLVTPTLSKFSKYI---YSTL 713
CG33978-PD        773 LKSLPLNNPTYSNI--WDNSQNKRAKIDYASMLKPA-IPINHQKQLKSATI-----HSRF 824
Submitted_Seq     714 LKPIKYIVPFKSTILPYAQSNKRRILIEYASILKPPRLPENKQILLQYPPLGRTQEKQMF 773
CG33978-PD        825 SPIKILQKNPSLKYSSSFNPIIHLDAAFYRHHILDIFRIPQKI-TSLSVSKSYFAMSNSP 883
Submitted_Seq     774 SAHEMVNKNPVIDFYSS-QSRIGFNSALHNNQ---VNRVPEILGTIYRTTKSYTTTNSD 829
CG33978-PD        884 EYSFSQDVHHDPVKTSLSNVTPLILSKLSIVQQIKLKSNVISHTINMHAGPLTFETESES 943
Submitted_Seq     830 KYSPIEDFSVDMLSTAMPIRTPFIISSFSELQHVNFKHDVLSHTINVHARPLTFKKEMEN 889
CG33978-PD        944 IPYATTVKGYTGTFFISHIKIPIRDEKIDVHITPNRKSLLSHEYDNLDHDKYYNKVKFHG 1003
Submitted_Seq     890 TP----------------HFEIPASSFRTEVPLRSNHKRVF--PYDN------KKIVFEE 925
CG33978-PD       1004 VSFGSI-MAPNKKPDVYQNDFPSQPTARLDST--HNGELTKHINEAPYMNVKAFTEPDQT 1060
Submitted_Seq     926 VDFNKLNILSNFKFGQYRN-----PTRHMDGHLHESDQVAQHSNKSFYL--LKITSTKKY 978
CG33978-PD       1061 AAFNNFKNKNLFSSTILNKMWSVISTNSKSSTKKNCTEDLCKLFLAK----YANSHPKYF 1116
Submitted_Seq     979 DLINNTAPNNWPSSIHFFNEFTMSNSNEHKGKIHKMSLSSLQKLVPSKGEHFYVGPLTGYN 1038

CG33978-PD       1117 TEHSGINTTVPK------PSNF--SAYYSSISSEYNNFTSSLTSRGLTSINDIKVQTSNM 1168
Submitted_Seq    1039 TEPPRFNAKIPNAIDSPSPSDDIQSIVYLSIQLKSTK-TDKLMSNDHIKLNTALLSPMNN 1097
CG33978-PD       1169 ELSNSLLYNLGSLRTIQ----EPHFEPGQDLLSVTLCVTRNVKNCKNSLIHKLKNDDKML 1224
Submitted_Seq    1098 NRSNAILSVLNSGNAIAILSYRQLLSKSTQILSTSTKSTTTKAGNMNPII-ALKSDKNHL 1156
CG33978-PD       1225 EPSEIMQDDTNSLLQKSMEAD-----FVSLAVKNSITEKIAKNLNT------LEVVKT-- 1271
Submitted_Seq    1157 EPSEMLQNDITSVSKTAMYLEKNSRFFTKLKPINSFLSTSFYTRNTPNFFSILKPTKTIL 1216
CG33978-PD       1272 EVIDA---QKYIDLKNI-----IVNTSKTNVMKPVFISVQK---------VLISQDSSYV 1314
Submitted_Seq    1217 NVVNGSVLKSIFDLKDNRSIGHAIFKTKNDVKKPISIRFQKEPPPPIRSQVIREYDVKHV 1276
CG33978-PD       1315 SILNGSEYVVDPYNVKDYSI-PSGSGD----NFEMENIQSFSRHTDFRNKSSSSLISNSK 1369
Submitted_Seq    1277 SQMTKNFKPTDHTSEPSKSINDSGSKNVVDINNFTKNSLSISKHDVIKYGSTSLLSSDLT 1336
CG33978-PD       1370 PDQIINASSDLRNMPFRAD--LIDSMSQRI-------------------------DL 1399
Submitted_Seq    1337 RHEDLYTTSVPNKIPIEPSKSINDSGSKNVVDIDNTKNSLSISRHDVFRYGSTYLLSSDL 1396
CG33978-PD       1400 IKQK----------------------SKTACDPKCRLNKNEICVTYGNSTESIGICECR 1436
Submitted_Seq    1397 TRHEGLYTTSVPNSIPIGATNYLKTTISADPCNPPCKSNKNEICVTYGLSS---SFCGCR 1453
CG33978-PD       1437 PSFGRMFPDRPCKPTYTYEMRIQTYWTENHLLKFSNNIKSNSLIEYRHISKILLEAADRM 1496
Submitted_Seq    1454 PSFGRLFPDLPCKPTYSYEMKLQTDWVENSLIKLIG-IDMNSTNEYHHIKIIFTDAVDRM 1512
CG33978-PD       1497 VMQSDYRDIFHGVQLQSVFAKT-NDTLMVTYLLQLSENSNEDQLTTVFQKYLRQSNFSIG 1555
Submitted_Seq    1513 IMQSDYRNIFHGAQLQSIFAEEKNNKITATFLLQLSENSDENLLKSTLTKHLRLSNFSIG 1572
CG33978-PD       1556 GTRLYTSREGLQFLTIKDFDECRNEHFYDCSPNAQCFNLIGSYTCSCKKGYVDKSENSLY 1615
Submitted_Seq    1573 GTGLHTSKNGPNILLAFKDFDECRHKDFNDCSNDAHCFNLVGSYTCVCKEGYMDIFPNALY 1632
CG33978-PD       1616 PGRHCLNNIIGCDKCNYNGKCVNVSAEKSHKEIAICKCYAWYAGTKCQVNLKVIILFILT 1675
Submitted_Seq    1633 PGRHCLDNKIGCDKCSYHGKCFQKWADKQQQAITSCKCNSWYAGTKCQVNLKVVIILLLT 1692
CG33978-PD       1676 SGAILSLVLFLFLLIT-KRKKQVDWKTSQLFIASSSLSPSLTTSTKSVREMENVGELCK 1734
Submitted_Seq    1693 MGTLLLALLLSFLLLINIKPKNKGNRIASQFNNASSQLHRSIISSN----------EICE 1742
CG33978-PD       1735 V-MTLISKKEYNEVNNHFISPETHKKTIVIENYTNAIASTKVQDLNIKGAFTQNDILYGS 1793
Submitted_Seq    1743 ISQTMLAKCKSSYI----------KEKSATQNDQTSSAYNTSDDLTVK--------ISSS 1784
CG33978-PD       1794 TDRGSQHEDQINRSLTLRIPRAKFRFPDISQNSCLCHKGKRSSLNI-TGKILVQKDDDYM 1852
Submitted_Seq    1785 SDNNFDEDGHTYRSLTLMIPRAKFRLPFTSKISYINHQSNMELEPCRSGTLTLASRND-- 1842
CG33978-PD       1853 AAYRSFGPSCEKDLKYLTLSDNLFYRGNTVAKADLVPVGY-GVSSLITDQIESPTLDENI 1911
Submitted_Seq    1843 ------------------SDELFYRN------PISPMSYPGISGC------------- 1863
CG33978-PD       1912 SINKAYSNSSSALNEDSNTMTERDLGSTFLLPHTHLYKTDKVCCSSRYYLRNLN 1965
Submitted_Seq    1864 ------------------------------------ESICRC-------- 1869
```

Figure 46: Gene Model Checker output for CG33978 Isoform D. Isoforms F and G have identical outputs because Isoform D, F, and G have identical coding sequences. Top: Dot plot of *D. melanogaster* isoform D vs. feature 3 isoforms. Bottom: Protein sequence alignment of *D. melanogaster* isoforms vs. feature 3 isoforms. Note 3' exons lacking in *D. ficusphila*, as well as the shift from one to two exons for the largest exon.

Transcription Start Sites

Although all isoforms of *CG33978* share an identical first coding exon, isoforms C and D each have unique 5'UTRs, and isoforms E, F, and G all share yet another unique 5'UTR in *D. melanogaster* (Figure 47). In order to start investigating the TSSs, the 5'UTR of *D. melanogaster* was viewed in the UCSC Genome Browser (Figure 48). The 9-state epigenomic landscape tracks for BG3 and S2 cells indicate that *CG33978* is almost entirely located in heterochromatic or heterochromatin-like euchromatin regions and is not actively transcribed in these cell lines. According to FlyBase ModENCODE cell line expression data, the only cell line in which there is moderate expression of *CG33978* is wing disc CME-W2 (Figure 49). The ModENCODE temporal expression data shows that there is moderately high expression in the embryo at 16-18 hours (Figure 50). Furthermore, there are no DHS peaks near the region for all three cell lines. The TSS (Celniker) track shows four TSS previously annotated by the modENCODE project (Hoskins, et al. 2011). TSS_CG33978_196265 likely corresponds to the TSS for isoforms C and D, while the remaining three annotations correspond to isoforms E, F, and G. Thus, in *D. melanogaster*, isoforms C and D have a peaked promoter while isoforms E, F, and G have a broad promoter.

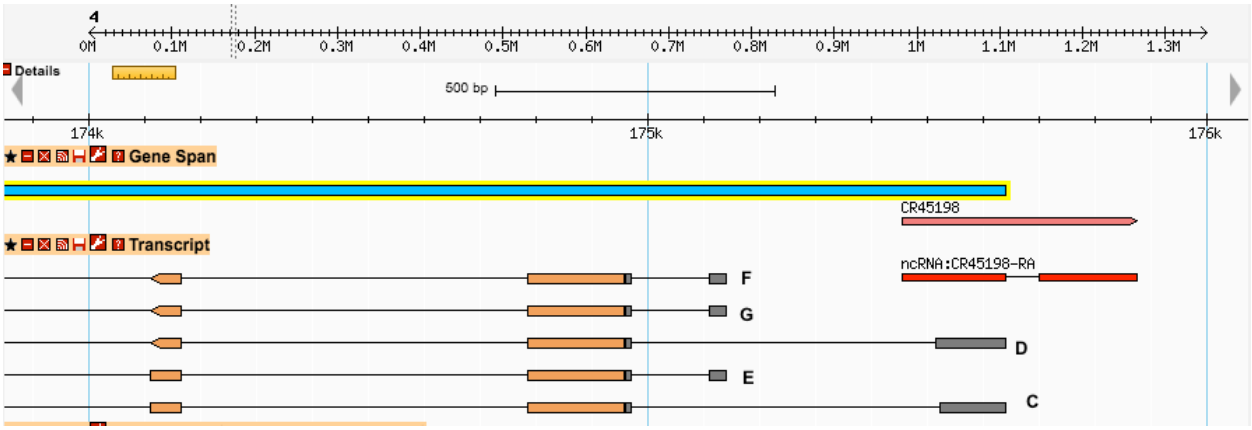


Figure 47: 5' UTR of all *CG33978* isoforms in FlyBase Browser View (*D. melanogaster*). All isoforms are labeled to the right of their respective gene models. Isoforms E, F, and G, have identical 5' UTRs. Isoforms C and D each have unique 5' UTRs.

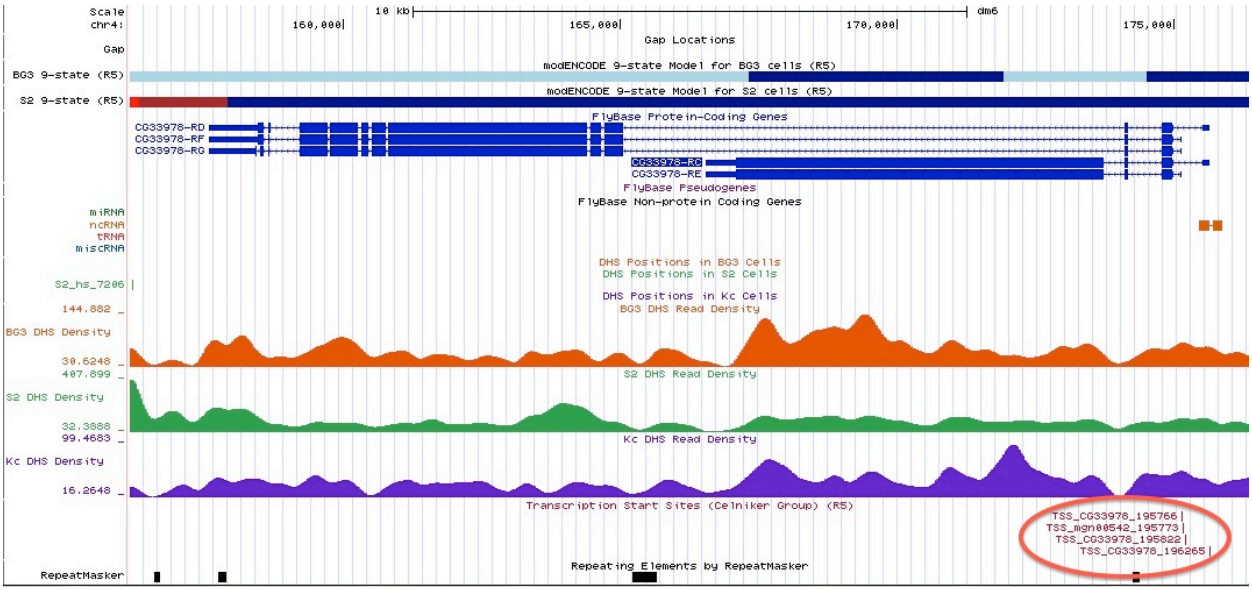


Figure 48: The 5' UTR of *D. melanogaster CG33978*. The 9-state tracks show that the gene, including the 5' UTR, is almost entirely located in heterochromatic (dark blue color) or heterochomatin-like euchromatin (light blue color) regions. Several TSSs has been previously annotated, circled in red.

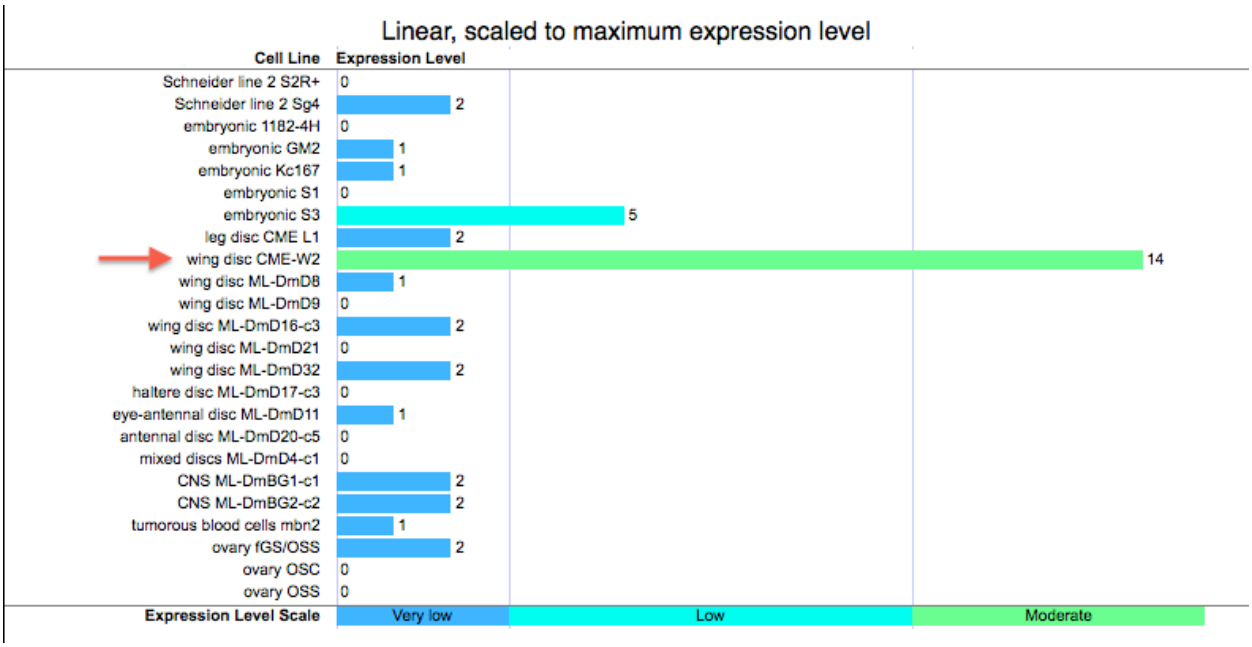


Figure 49: FlyBase ModENCODE cell line expression data shows moderate expression only in the wing disc CME-W2 cell line.

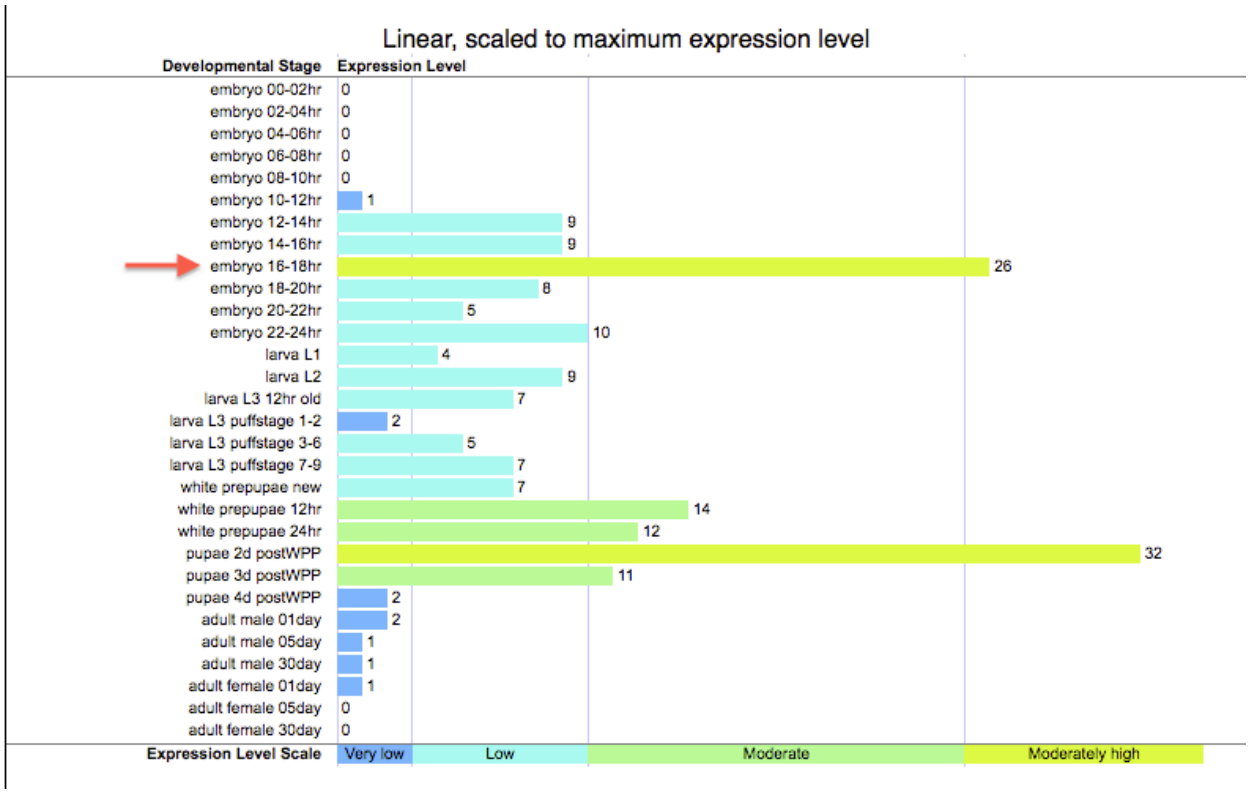


Figure 50: FlyBase ModENCODE temporal expression data shows moderately high expression in the embryo at 16-18 hours.

A pairwise BLASTn alignment with optimized parameters was carried out using the first *D. melanogaster CG33978* transcript as a query and contig8, subrange bases 42350-45000, as the subject. However, regardless of the isoform used, no alignments in the region of interest were found. However, if the size of the 5'UTR has remained relatively constant between *D. melanogaster* and *D. ficusphila*, the TSS for isoform C and D is expected to be found around 700bp upstream of the first coding exon, at approximately position -43,200. The TSS for isoform E, F, and G is expected to be found around 200bp upstream of the first coding exon, at approximately position -42,700.

Upstream of the first *CG33978* coding exon in contig8, there are several significant RNA Seq peaks, indicating possible 5' UTRs (Figure 51). TopHat reads also extend upstream of the first coding exon, though most reads do not have scores above 10, with the exception of

JUNC00003876 which has a score of 37. The RNA Seq data around bases 42570-42700 most likely correspond to the 5'UTR for isoforms E, F, and G. The remaining RNA Seq peaks may correspond to 5'UTRs for isoforms C and D. However, a large repeat that extends from around bases 43,180 to 44,350 prevents their precise identification. Position -43,200, the estimated position of the TSS for isoforms C and D, is contained within this repeat.

A variety of core promoter motifs were found using the Short Match tool on the UCSC Browser (Table 8). Most notably, an Inr motif was found at position -42,702, which corresponds to the upstream edge of the RNA-Seq peak previously predicted to correspond to the 5' UTR for isoforms E, F, and G. This Inr motif gives a TSS coordinate of -42,700. Other motifs that give TSS coordinates in this vicinity include the BRE[d] notif at -42,716 and the DPE motif at -42,673. The TSS for isoforms C and D cannot be narrowed down using core promoter motifs, though an additional Inr motif was found at position -43,875.

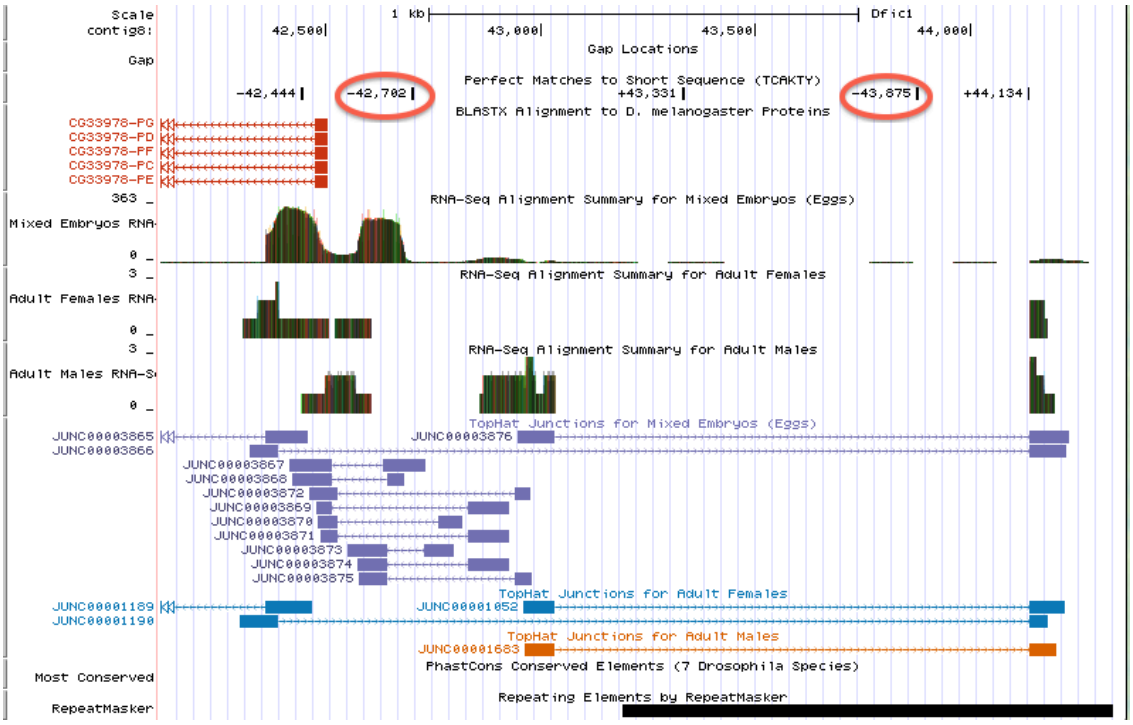


Figure 51: UCSC Genome Browser view of 5' UTR region for *CG33978* in *D. ficusphila*. The RNA-Seq data around bases 42570-42700 most likely correspond to the 5'UTR for isoforms E, F, and G. Inr motifs on the minus strand are circled in red. The Inr motif at position -42,702 likely corresponds to the TSS of isoforms E, F, and G. The large repeat impedes annotation of the TSS for isoforms C and D.

Motif	Position Relative to TSS	Position in *D. ficusphila*	Position in *D. melanogaster*
BRE^u	-38	-	-
TATA Box	-31 or -30	-43379, -43810	
BRE^d	-23	-42537, -42716, -42875, -43106, -43281, -43283, -43301, -43431, -43537, -43557, -43651, -43726, -43731, -43837, -44108, -44256, -44278, -44391, -44400	-175242, -175244, -175281, -175291, -175298, -175539, -175568, -175752, -175820, -175868, -175967, -176014, -176049, -176137, -176461, -176512, -176616, -176792, -176856, -176858, -176911, -177008, -177096
Inr	-2	-42702, -43875	-175137, -175579, -175704, -175755, -176250
MTE	+18	-	-175135, -175220, -175443, -175536, -175770, -175906, -176664, -176732
DPE	+28	-42641, -42673, -42700, -43660, -43831, -42880, -43839, -43982, -44158, -44164, -441143, -44321, -44411	-175135, -175220, -175443, -175536, -175770, -175906, -176664, -176732
Ohler_motif1	NA	-	-
DRE	NA	-	-
Ohler_motif5	NA	-	-
Ohler_motif6	NA	-	-
Ohler_motif7	NA	-	-
Ohler_motif8	NA	-	-

Table 8: Core promoter motifs around bases -44415-42500. The coordinates of the motifs were obtained by using the Short Match tool in the UCSC Genome Browser.

Inspection of the orthologous region in contig9 of the *D. biarmipes* Aug. 2013 (GEP/Dot) Assembly, revealed a region enriched in RNA PolII ChIP Seq data from around bases 35,500-35,850 (Figure 52). Much of this region is conserved across several *Drosophila* species and a small RNA-Seq peak can also be observed. Furthermore, an Inr motif at position +35,722 aligns with the RNA PolII ChIP-Seq peak. Thus, in *D. biarmipes*, +35,744 is a strong candidate for the TSS in isoforms E, F, and G. In *D. biarmipes*, the first coding exon starts at position +36,020. Thus, the TSS for isoforms E, F, and G is located 276 bp upstream of the first coding exon. This is considerably further upstream than the putative TSS for isoforms E, F, and G in *D.*

ficusphila, which is located 196 bp upstream. However, the conservation tracks show that conservation among *Drosophila* species does not extend throughout the entire 5'UTR in *D. biarmipes*. This suggests that the 5' UTR is longer in *D. biarmpes* than *D. ficusphila*. A BLASTn search comparing the DNA sequence of this enriched area in *D. biarmipes* against contig8 produced an alignment that maps the *D. biarmipes* RNA PolII ChIP-Seq data peak to position -42,692 in contig8, near the TSS suggested by the Inr motif (Figure 53, Figure 51).

Based on all of the data available, the TSS for isoforms E, F, and G of *D. ficusphila CG33978* was assigned to bases -42,692-42,700. The search region for the TSS for isoforms C and D was defined as bases 42,800-45,000 in contig 8. However, it is possible that the TSS is located on a neighboring contig. This hypothesis is supported by the fact that non-coding RNA (ncRNA) *CR45198,* which overlaps with the 5'UTR of isoforms C and D in *D. melanogaster*, cannot be found on contig8.

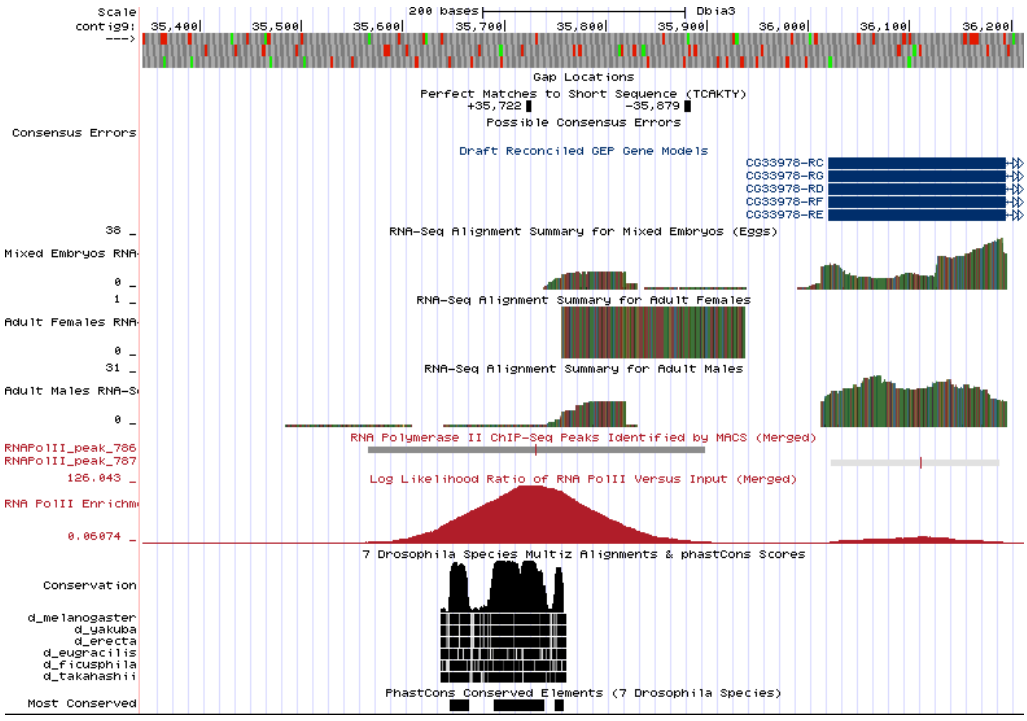


Figure 52: The 5' UTR region of *D. biarmpies CG33978*. The RNA PolII ChIP Seq peak corresponds with the Inr motif at +35,722 and likely represents the TSS for isoforms E, F, and G.

Range 1: 42670 to 42782 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
114 bits(126)	1e-28	93/113(82%)	0/113(0%)	Plus/Minus

```
Query  9      AAACCAGCCAAACCACCTTTTAAAAGCTTGACTACAAAGCAACAACCGGATGAGTCTGTT  68
Sbjct  42782  AAACCAGCCAAACCACATTTGTAAGGTTCGACTGCAGAGCATCAACCGGATGGGTCTGTT  42723

Query  69     ACTGTGCCAGTGAAATCAGTTCTCTGACTTTCGTTTCAACGAACAGACGTGCG  121
Sbjct  42722  ATTGTGTCAGGGAAATCAGTTATCTGACTTTCGTTTCTATAGATAGACCTGCG  42670
```

Figure 53: BLASTn alignment of DNA sequence of RNA PolI ChIP-Seq data enriched region conserved across multiple species (query) against contig 8 (subject).

Feature 4

The BLASTx alignment and several gene prediction tracks showed a putative gene within an intronic region of feature 3 (Figure 54). When a BLASTp search was used to align the predicted peptide from Genscan against the *D. melanogaster* annotated proteins in FlyBase no significant results were found (Figure 55).

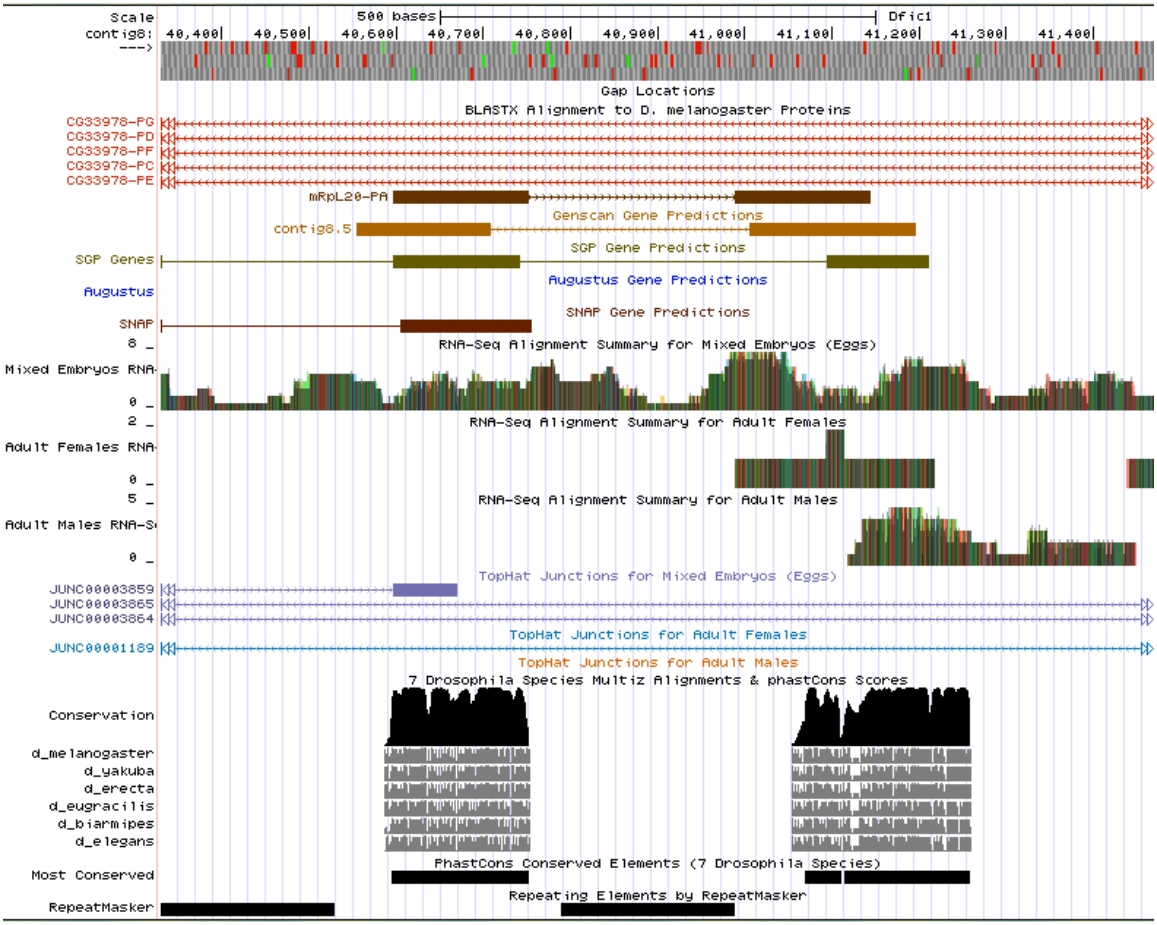


Figure 54: Close up of Feature 4. The region is found within an intronic region of feature 3. RNA-Seq and Top Hat data is inconclusive, but high conservation is observed.

BLAST Hit Summary				
	Description	Species	Score	E value
☑	Dscam1-PBN	Dmel	30.8018	0.258056
☑	Dscam1-PBV	Dmel	30.8018	0.278181
☑	Rpn2-PB	Dmel	28.1054	1.57778
☑	Rpn2-PC	Dmel	28.1054	1.591

Figure 55: FlyBase BLASTp results for GenScan predicted polypeptide contig8.5 (query) against *D. melanogaster* annotated proteins database (subject). No significant results were found.

However, due to the high level of conservation across *Drosophila* species observed in the conservation tracks, in addition to the presence of the RNA-Seq data in the region, annotation efforts were continued. According to the BLASTx alignment to *D. melanogaster* proteins in the UCSC browser, the putative gene is the A isoform of *mRpL20*. The Gene Record Finder shows that only the A isoform of *mRpL20* exists in *D. melanogaster* (Figure 56). Isoform A has three exons. This is notable because the BLASTx alignment in the UCSC browser only has two exons. Additionally, *mRpL20* is located on chromosome 3L in *D. melanogaster* (Figure 57).

Isoform	1_7149_0	2_7149_0	3_7149_1
mRpL20-PA	1	2	3

Select a row to display the corresponding CDS sequence:

FlyBase ID	5' Start	3' End	Strand	Phase	Size (aa)
1_7149_0	13,017,762	13,017,676	-	0	29
2_7149_0	13,017,612	13,017,404	-	0	69
3_7149_1	13,017,349	13,017,193	-	1	52

Figure 56: Gene Record Finder 6.08 record of mRpL20 in *D. melanogaster*. mRpL20 only has one isoform with three exons in *D. melanogaster*.

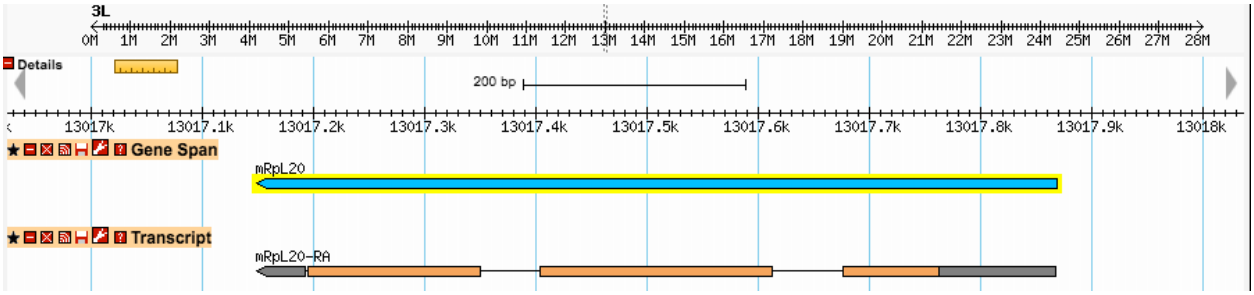
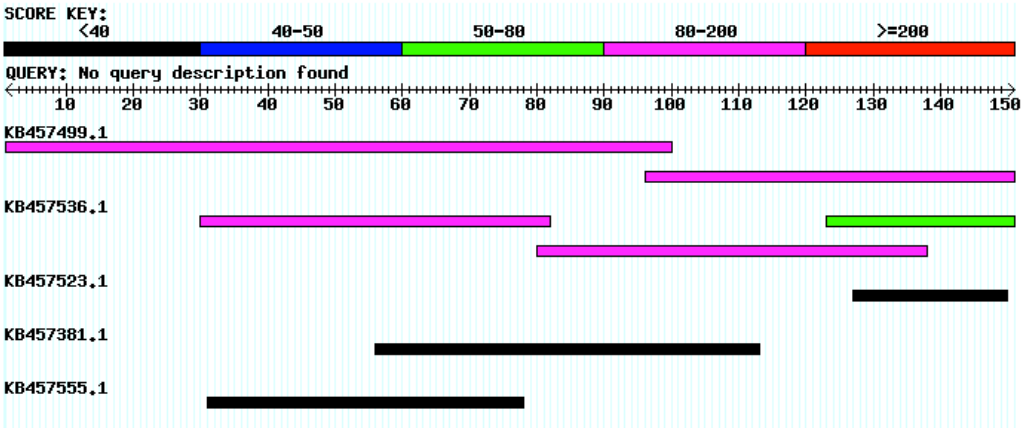


Figure 57: Brower view of *mRpL20* in *D.* melanogaster genome. The coding regions are show in light orange and the translated regions are shown in gray. *mRpL20* is located on the reverse frame, chromosome 3L in *D. melanogaster*.

Due to the overwhelming evidence that feature four may not be the real ortholog of *mRpL20* in *D. ficusphila*, a FlyBase tBLASTn search of the mRpL20 protein sequence in *D. melanogaster* was conducted against the whole genome assembly of *D. ficusphila* (Figure 58). The best alignment mapped to scaffold 7180000454048 and includes all three exons. The second best alignment mapped to scaffold 7180000454086 and does not include the first exon. A BLASTn search of contig8 against the whole genome assembly of *D. ficusphila* mapped to scaffold 7180000454086 (Figure 59). This suggests that feature 4 is a pseudogene derived from *mRpL20*.



BLAST Hit Summary				
	Description	Species	Score	E value
	Drosophila ficusphila unplaced genomic scaffold scf7180000454048, whole genome shotgun sequence	Drosophila ficusphila	191.43	1.87967e-73
	Drosophila ficusphila unplaced genomic scaffold scf7180000454086, whole genome shotgun sequence	Drosophila ficusphila	102.449	3.15032e-22
	Drosophila ficusphila unplaced genomic scaffold scf7180000454073, whole genome shotgun sequence	Drosophila ficusphila	31.187	1.01931
	Drosophila ficusphila unplaced genomic scaffold scf7180000453912, whole genome shotgun sequence	Drosophila ficusphila	30.8018	1.30923

Feature 58: tBLASTn search of *D. melanogaster* mRpL20 amino acid sequence (query) against whole genome assembly of *D. ficusphila* (subject). The best alignment includes all three exons and mapped to scaffold 180000454048. The second best alignment mapped to scaffold 7180000454088 and does not include the first exon.

Query	
Description	Length
Dfic1_dna range=contig8:1-45000 5'pad=0 3'pad=0 strand=+ repeatMasking=none	45000

BLAST Hit Summary				
	Description	Species	Score	E value
	Drosophila ficusphila unplaced genomic scaffold scf7180000454086, whole genome shotgun sequence	Drosophila ficusphila	18521.6	0

Figure 59: BLASTn search of contig8 (query) against the whole genome assembly of *D. ficusphila* (subject) mapped to scaffold 7180000454088.

From the UCSC Genome Browser, the repeat upstream of feature 4 was identified as a helitron (Figure 54). Thus, it is likely that the last two exons of a *mRpL20* duplicate on chromosome three were picked up by the helitron's rolling circle mechanism and transported to chromosome four.

Exon-by-exon BLASTx searches were then used to find the approximate coordinates for the exon junctions in *D. ficusphila* (Table 9). As expected, the first exon could not be found within contig8. The second exon mapped to two frames, amino acids 1-52 and 51-69 on the +2 and +3 frame respectively. This suggests that the *mRpL20* pseudogene has acquired a novel intron. This intron contains an unknown repeat at bases 40790-40988. The third exon also mapped to two frames, amino acids 1-23 and 24-52 on the +3 and +2 frames respectively (Figure 60). Inspection of the subject ranges reveals that the second portion of CDS2_7149_0 and the first portion of 3_7149_1 are essentially contiguous. These observations suggest that this region is derived from a cDNA pseudogene.

FlyBase ID	Coding Exon Size (amino acids)	Subject Range	Query Frame	Query Range
1_7149_0	Not found in contig8			
2_7149_0	69	1-52, 51-69	+2, +3	40598-40753, 40989-41045
3_7149_1	52	1-23, 24-52	+3, +2	41049-41118, 41126-41212

Table 9: Summary table for approximate exon locations on BLASTx alignments for mRpL20 pseudogene.

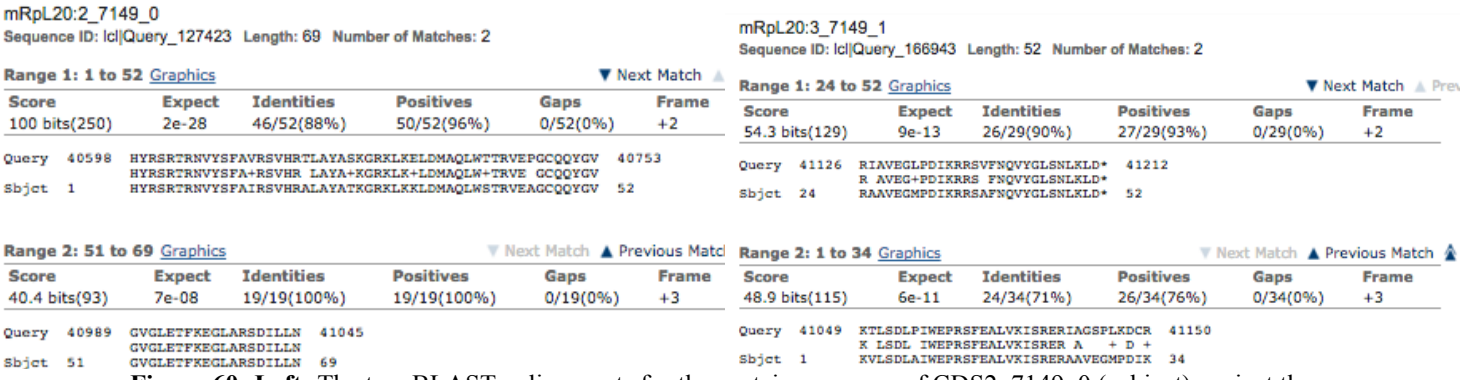


Figure 60: Left: The two BLASTx alignments for the protein sequence of CDS2_7149_0 (subject) against the translated DNA sequence of contig8 (query). The first alignment is on the +2 frame and covers amino acids 1-52. The second alignment is on the +3 frame and spans amino acids 51-69. **Right:** The two BLASTx alignments for the protein sequence of CDS3_7149_0 (subject) against the translated DNA sequence of contig8 (query). The first alignment is on the +3 frame and covers amino acids 1-34. The second alignment is on the +2 frame and spans amino acids 24-52.

Although at first, the conservation tracks seem to suggest that this pseudogene exists in most other *Drosophila* species, the chained alignment tracks for *D. melanogaster*, *D. yakuba*, *D. erecta*, *D. biarmipes*, and *D. elegans* show that the alignments at feature 4 do not belong to the same chain as the surrounding region (Figure 61). This suggests that the alignments at feature 4 belong to the *mRpL20* ortholog and are misaligned to this pseudogene. Thus, it is likely that the *mRpL20* pseudogene on chromosome four only exists in *D. ficusphila*.

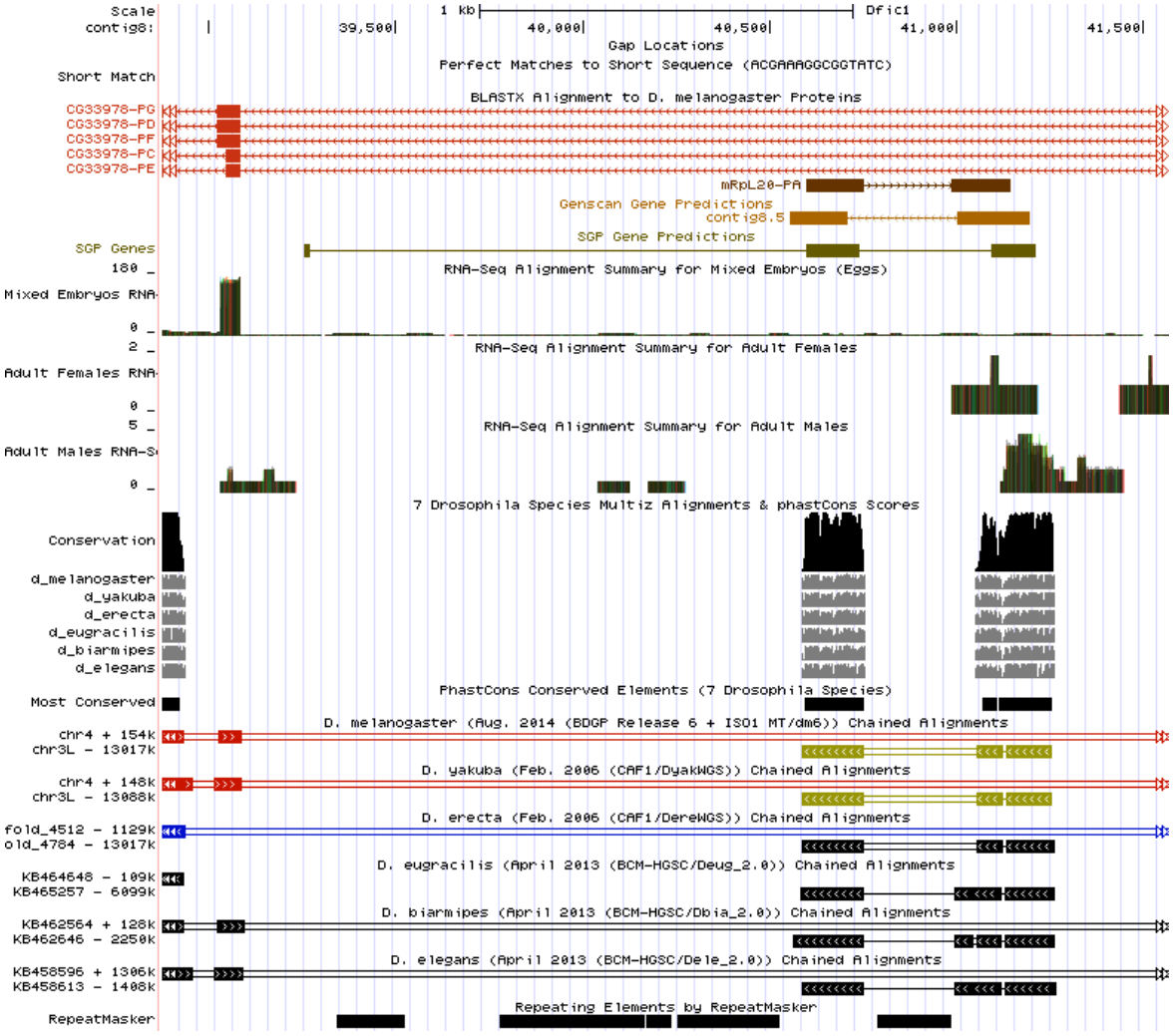


Figure 61: The chained alignment tracks show that alignments at feature 4 belong to a separate chain from the surrounding region. These alignments likely belong to the *mRpL20* ortholog and suggest that the pseudogene only exists in *D. ficusphila*.

Gene Evolution

CG33978

Feature 3, the CG33978 ortholog, was unusual in contig8 in that it had significantly lower conservation relative to the other features. Thus, the predicted polypeptide from the *D. ficusphila*, obtained using Gene Model Checker, was used as a query for a BLASTp search with the reference sequence database as the subject (Figure 62). The BLASTp search identified a Calcium-binding EGF-like domain at position 1573-1620 of the gene. In order to investigate these conservation patterns, a Clustal Omega multiple sequence alignment was carried out with the *Drosophila CG33978* orthologs obtained from FlyBase (Figure 63).

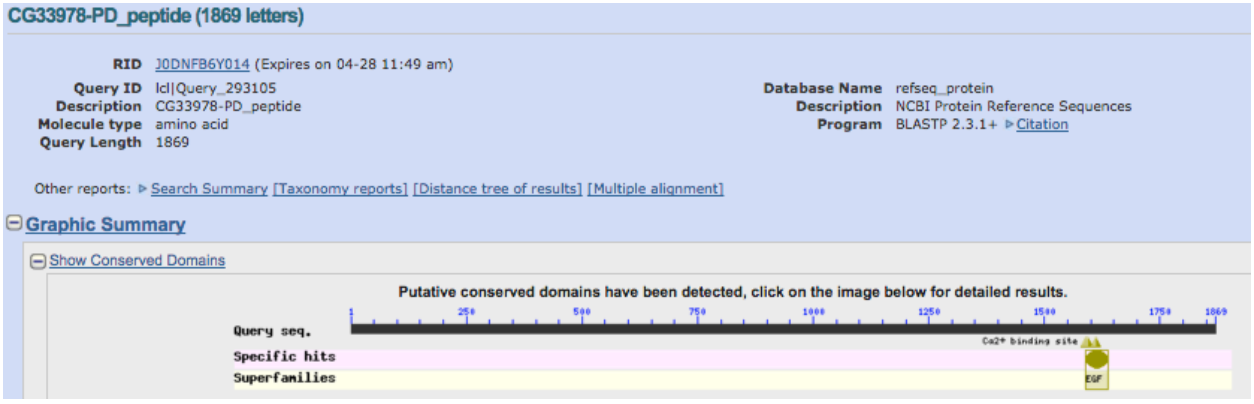


Figure 62: BLASTp alignment of *D. ficusphila CG33978* isoform D peptide sequence (query) against NCBI Refseq database.

Figure 63: Clustal Omega multiple sequence alignment of *Drosophila* CG33978. The calcium-binding EGF-like domain was conserved across species.

The Clustal OMEGA alignment revealed variable conservation among the four *Drosophila* species for most of the alignment. The Calcium-binding EGF-like domain was conserved. This is expected, since the only known molecular function of protein CG33978 is calcium ion binding. Approximately forty amino acids upstream and downstream of the Calcium-binding EGF-like domain were also conserved to essentially the same degree. The lack of conservation in the rest of the *CG33978* gene is interesting, and may be related to the lack of *CG33978* expression in most cell lines across various stages of development. However, further investigation is necessary to determine the precise reasons for variation.

Arl4

The evolution of feature 1, the *Arl4* ortholog, was also investigated. The predicted peptide from the *D. ficusphila* isoform B, obtained using the Gene Model checker, was used as a query for a BLASTp serach with the NCBI reference sequence database as the subject (Figure 64). The BLASTp search identified a P-loop containing Nucleoside Triphosphate Hydrolase (P-loop NTPase) superfamily towards the first half and the last fifty amino acids of the query sequence. This conserved region includes a GTP/Mg^{2+} binding sites, Switch I and II regions, and G1, G2, G3, G4, and G5 boxes. The concentration of highly conserved sites in the first half of the peptide is congruent with the high degree of conservation present in the first three exons in all *Arl4* isoforms. In order to investigate these conservation patterns, a Clustal OMEGA multiple sequence alignment was carried out with *Drosophila Arl4* orthologs obtained from FlyBase (Figure 65).

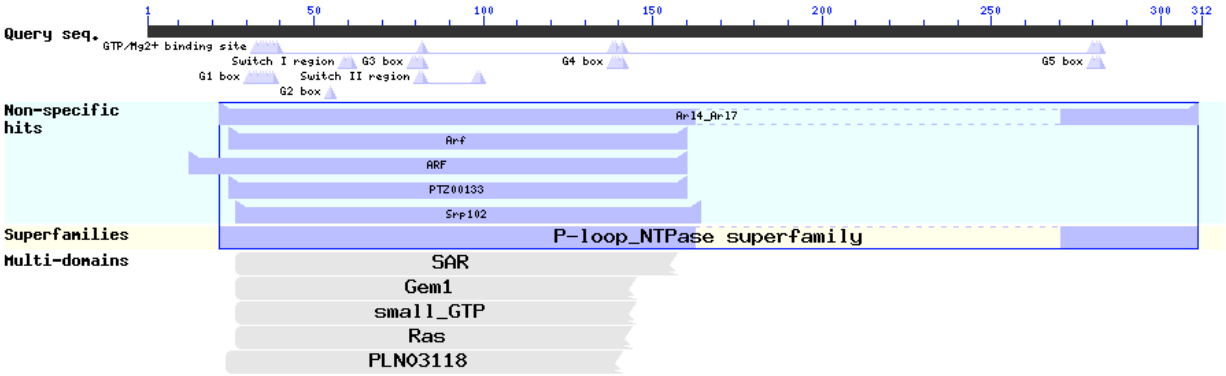


Figure 64: BLASTp alignment of *D. ficusphila Arl4* isoform B peptide sequence (query) against NCBI Refseq database.

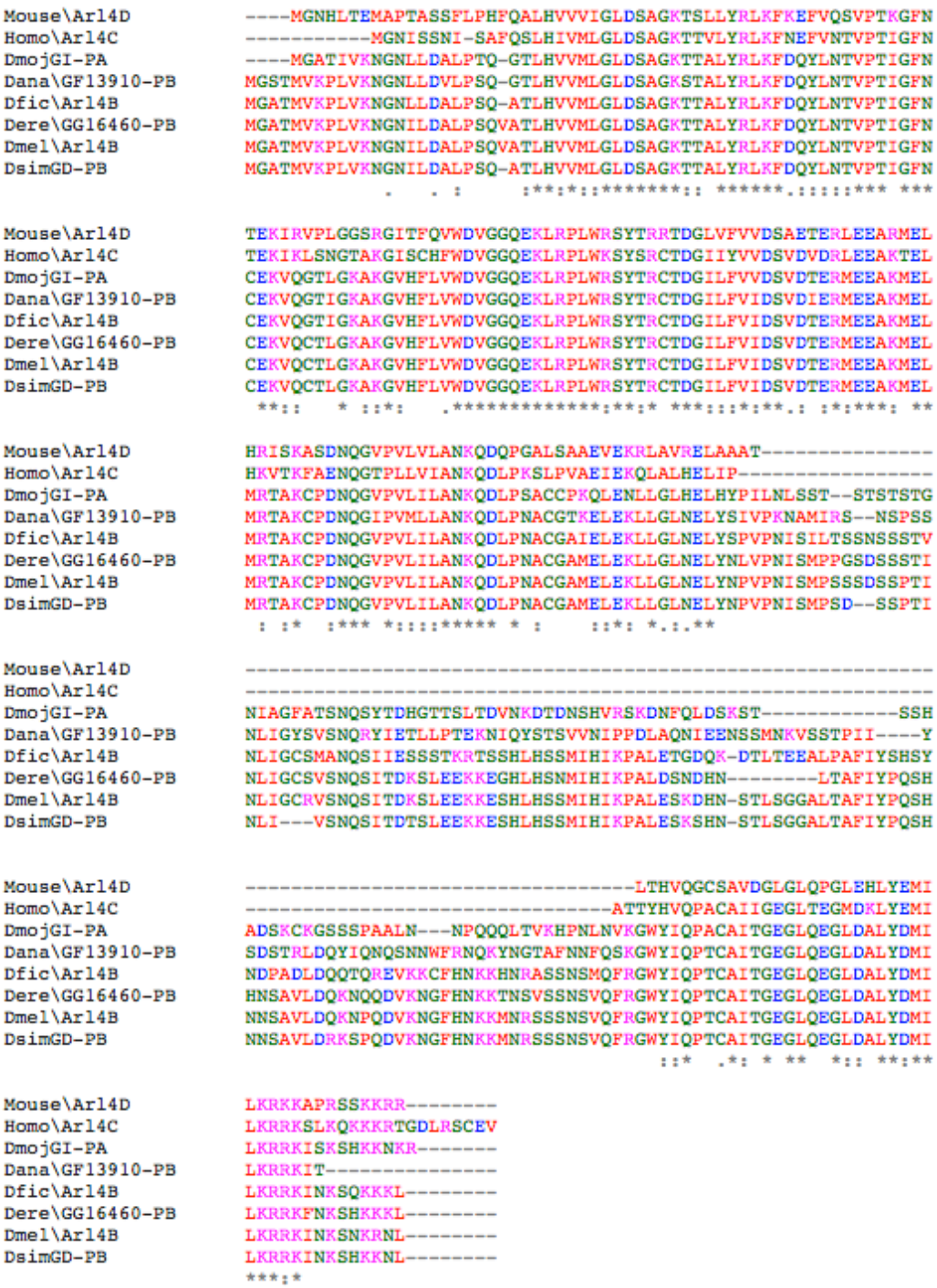


Figure 65: Clustal OMEGA multiple sequence alignment of *Drosohpila Arl4* with human and mouse as outgroups. The P-loop NTPase superfamily conserved domains are highly conserved across all species.

The Clustal OMEGA alignment revealed high degrees of conservation at the P-loop NTPase superfamily conserved domains. There is poor conservation in other regions. Notably,

the human and mouse outgroups do not have a gap in the P-loop NTPase superfamily conserved domain, as is observed in all of the *Drosophila* species.

Repeats

Repetitious sequences were analyzed using the UCSC Table Browser in order to process contig8 using Repeat Masker with the appropriate repeat database (Table 10). The Repeat Masker analysis determined that contig8 is 35.71% repetitive, which corresponds to 16,071/45,000 bp, and contains nine repeats 500 bp or greater (Table 11). The region in *D. melanogaster* that corresponds to contig8 is 7.93% repetitive (from end of *CG33978* to end of *Arl4*), corresponding to 2211/27888 bp, and contains no repeats 500 bp or greater.

```
==================================================
file name: contig8.fasta
sequences:             1
total length:      45000 bp  (44930 bp excl N/X-runs)
GC level:          35.08 %
bases masked:      15988 bp ( 35.53 %)
==================================================
               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:                 5         424 bp    0.94 %
      LINE1            0           0 bp    0.00 %
      LINE2            0           0 bp    0.00 %
      L3/CR1           0           0 bp    0.00 %

LTR elements:          8        2179 bp    4.84 %
      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:         18        4372 bp    9.72 %
     hAT-Charlie       0           0 bp    0.00 %
     TcMar-Tigger      0           0 bp    0.00 %

Unclassified:         46        9096 bp   20.21 %

Total interspersed repeats:    16071 bp   35.71 %


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 %
==================================================
```

```
==================================================
file name: RM2sequpload_1461188981
sequences:             1
total length:      27888 bp  (27888 bp excl N/X-runs)
GC level:          33.93 %
bases masked:       2211 bp ( 7.93 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
Retroelements          1         212 bp    0.76 %
   SINEs:              0           0 bp    0.00 %
   Penelope            0           0 bp    0.00 %
   LINEs:              0           0 bp    0.00 %
    CRE/SLACS          0           0 bp    0.00 %
     L2/CR1/Rex        0           0 bp    0.00 %
     R1/LOA/Jockey     0           0 bp    0.00 %
     R2/R4/NeSL        0           0 bp    0.00 %
     RTE/Bov-B         0           0 bp    0.00 %
     L1/CIN4           0           0 bp    0.00 %
   LTR elements:       1         212 bp    0.76 %
     BEL/Pao           1         212 bp    0.76 %
     Ty1/Copia         0           0 bp    0.00 %
     Gypsy/DIRS1       0           0 bp    0.00 %
       Retroviral      0           0 bp    0.00 %

DNA transposons        6         655 bp    2.35 %
   hobo-Activator      2          97 bp    0.35 %
   Tc1-IS630-Pogo      2         282 bp    1.01 %
   En-Spm              0           0 bp    0.00 %
   MuDR-IS905          0           0 bp    0.00 %
   PiggyBac            0           0 bp    0.00 %
   Tourist/Harbinger   0           0 bp    0.00 %
   Other (Mirage,      1         184 bp    0.66 %
    P-element, Transib)

Rolling-circles        0           0 bp    0.00 %

Unclassified:          5        1344 bp    4.82 %

Total interspersed repeats:      2211 bp    7.93 %


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 %
==================================================
```

Table 10: Left: Summary of all repeats in contig8. 35.53% of all bases are contained in repeats. **Right:** Summary of all repeats in

corresponding region in *D. melanogaster*. 7.93% of all bases are contained in repeats.

Repeat Name	Repeat Class/Family	Strand	Begin	End	Size (bp)
Rnd01_family-108	DNA/MITE	-	4435	4984	549
rnd-1_family-107	DNA/MITE	-	21336	21939	603
rnd-1_family-244	LTR/Pao	-	14599	15235	636
rnd-1_family-89	LTR/Pao	-	13675	14313	638
rnd-6_family-34	RC/Helitron	-	19391	20097	706
rnd-2_family-34	Unknown	+	43192	43902	710
rnd-1_family-14	RC/Helitron	+	2198	2912	714
rnd-4_family-33	RC/Helitron	+	16241	17107	866
rnd-1_family-450	DNA	-	1195	2175	980

Table 11: Details of repeats greater than 500 bp in contig8.

In some species of *Drosophila*, a subset of the repeats identified by RepeatMasker can be derived from genomic fragments of *Wolbachia* that have been integrated into the genome. A BLASTn search comparing the *Wolbachia* endosymbiont of the Riverside strain of *D. simulans* against contig8 revealed eleven alignments under the Expect threshold of 1e-2 (Figure 66).

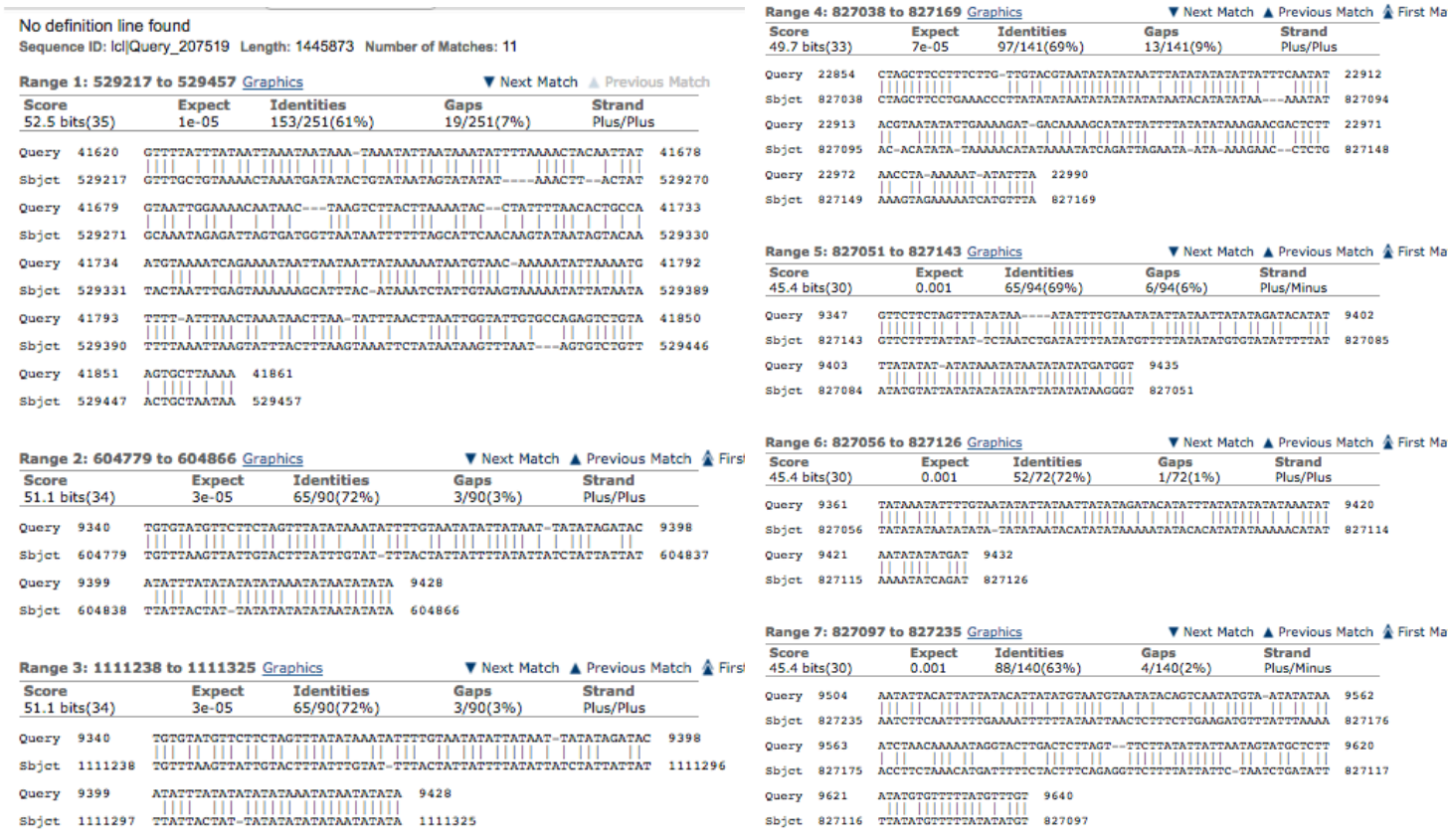


Figure 66: BLASTn results comparing the *Wolbachia* endosymbiont of the Riverside strain of *D. simulans* against contig8.

Synteny

In order to analyze synteny between the annotated genes in contig8 and *D. melanogaster*, the annotated genes were compared with FlyBase *D. melanogaster* Browser View (Figure 67). The *D. melanogaster* region that corresponds to contig8 shows that the three annotated genes are not in the same relative position and orientation. In contig8, *Arl4* is found upstream of *CG31997* on the forward strand. This suggests that an inversion has taken place in *D. ficusphila*. Thus, these genes are not in synteny. In addition, the *mRpL20* paralog found in contig8 is not present in *D. melanogaster*.

The *D. melanogaster* annotations also include a 361 bp non-coding RNA *CR45198* in the same region as the 5' UTR for *CG33978* in *D. melanogaster* that is unaccounted for in *D. ficusphila*. A BLASTn search with the DNA sequence of *CR45198* in *D. melanogaster* (query)

against the contig8 DNA sequence (subject) did not produce any significant results. Since *CR45198* is found on the forward strand and overlaps with the 5'UTR of isoforms C and D of *CG33978* in *D. melanogaster*, *CR45198* should also exist on the forward strand in *D. ficusphila* in the putative 5'UTR region for *CG33978*. The only alignment that fits these parameters had an e-value of 0.87 (Figure 68). It is possible that *CR45198* exists in a neighboring contig as the 5'UTR for *CG33978* could not be identified in contig8. In order to determine whether *CR45198* is present in *D. ficusphila,* a Blastn search was carried out within FlyBase with *CR45198* isoform A gene region as a query and the *D. ficusphila* whole shotgun sequence assembly as the subject. The search did not return any significant results. This gene has therefore been lost in the *D. ficusphila* genome.

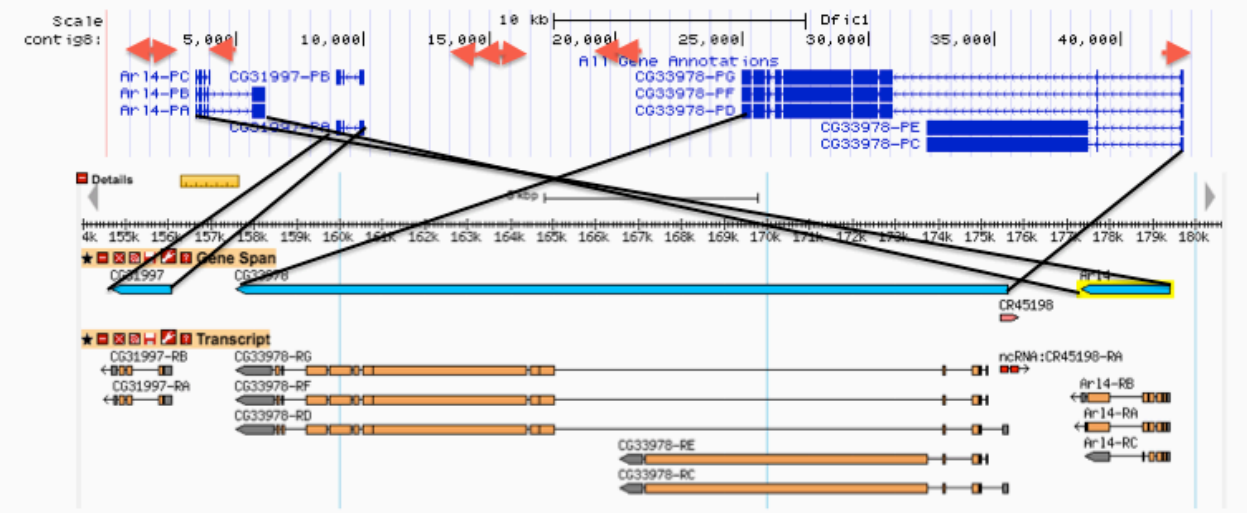


Figure 67: Comparison of gene order and orientation in contig8 and *D. melanogaster.* Repeat elements greater than 500 bp in contig8 are marked with a red arrow. Black alignment lines show inversion of *Arl4* with *CG31997* and *CG33978.*

Range 67: 43053 to 43088 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
24.0 bits(15)	0.87	28/39(72%)	3/39(7%)	Plus/Plus

```
Query  310    ATATCAATAAAATAAGAAATTATAATTTAAAATAAATAA  348
              ||||| | ||| | |||||  ||  ||||||||||||
Sbjct  43053  ATATCGCTTAAACTAAAAAT---AAACTAAAATAAATAA  43088
```

Figure 68: Only alignment in expected region and strand for BLASTn search of DNA sequence of *D. melanogaster CR45198* (query) against contig8 DNA sequence (subject). The alignment is insignificant because it has an e-value of 0.87.

Lastly, any exons predicted by any gene predictor that were unaccounted for, as well as intergenic sequences within contig8 were used as queries for BLASTx searches with non-redundant (nr) database in NCBI to ensure that all genes were annotated. GenScan contig8.2 produced significant alignments to *Atf6* across many *Drosophila* species (Figure 69). However, *Atf6* is located on chromosome 2R in *Drosophila melanogaster*, and is thus unlikely to correspond to a real gene on the dot chromosome. Given its position within a high repeat density region, it is likely that GenScan contig8.2 is a pseudogene of *Dmel\Atf6* that has been transposed from chromosome two to four in the *D. ficusphila* line. Chained alignment tracks show that GenScan contig8.2 aligns to chromosome 2R in *D. melanogaster* and chromosome 2L in *D. yakuba*, further supporting this hypothesis (Figure 70).

Description	Max score	Total score	Query cover	E value	Ident	Accession
uncharacterized protein Dere_GG23144 [Drosophila erecta]	132	132	85%	3e-30	33%	XP_001970863.2
Atf6, isoform B [Drosophila melanogaster]	117	117	84%	3e-25	33%	NP_610159.1
Atf6, isoform C [Drosophila melanogaster]	116	116	85%	7e-25	32%	NP_995745.1
GH21728 [Drosophila grimshawi]	107	107	49%	1e-21	38%	XP_001987094.1
uncharacterized protein Dyak_GE11351, isoform B [Drosophila yakuba]	103	103	84%	7e-21	30%	XP_015045266.1
uncharacterized protein Dyak_GE11351, isoform A [Drosophila yakuba]	103	103	84%	8e-21	30%	XP_002086078.2
uncharacterized protein Dyak_GE11351, isoform D [Drosophila yakuba]	102	102	83%	2e-20	30%	XP_015045268.1
GM26735 [Drosophila sechellia]	102	102	85%	2e-20	34%	XP_002044355.1
uncharacterized protein Dsimw501_GD17476 [Drosophila simulans]	100	100	85%	1e-19	33%	XP_016026006.1
uncharacterized protein Dmoj_GI20675 [Drosophila mojavensis]	97.1	97.1	43%	2e-18	39%	XP_002005817.1
uncharacterized protein Dwil_GK21942 [Drosophila willistoni]	97.1	97.1	49%	2e-18	40%	XP_002063497.2
uncharacterized protein Dvir_GJ20425 [Drosophila virilis]	95.5	95.5	42%	5e-18	38%	XP_002050015.2
uncharacterized protein Dana_GF26908, isoform C [Drosophila ananassae]	91.3	91.3	65%	2e-17	32%	XP_014759184.1
GL22941 [Drosophila persimilis]	89.7	89.7	40%	3e-16	42%	XP_002028233.1
uncharacterized protein Dpse_GA16207, isoform A [Drosophila pseudoobscura pseudoobscura]	89.7	89.7	40%	4e-16	42%	XP_001352340.2
uncharacterized protein Dana_GF26908, isoform A [Drosophila ananassae]	87.4	87.4	65%	2e-15	30%	XP_014759182.1
PREDICTED: cyclic AMP-dependent transcription factor ATF-6 alpha [Ceratitis capitata]	76.3	76.3	40%	9e-12	38%	XP_004526378.1
PREDICTED: cyclic AMP-dependent transcription factor ATF-6 alpha isoform X2 [Bactrocera oleae]	72.4	72.4	50%	2e-10	32%	XP_014091700.1
PREDICTED: cyclic AMP-dependent transcription factor ATF-6 alpha isoform X1 [Bactrocera oleae]	72.4	72.4	50%	2e-10	32%	XP_014091699.1
PREDICTED: cyclic AMP-dependent transcription factor ATF-6 alpha [Bactrocera cucurbitae]	72.4	72.4	40%	2e-10	35%	XP_011191311.1
PREDICTED: uncharacterized protein LOC105225350 [Bactrocera dorsalis]	70.5	70.5	13%	8e-10	77%	XP_011202058.1
PREDICTED: cyclic AMP-dependent transcription factor ATF-6 alpha [Cerapachys biroi]	68.9	68.9	59%	2e-09	30%	XP_011352101.1

Figure 69: BLASTx search of GenScan prediction contig8.2 translated DNA sequence (query) against the nr database (subject) produced significant alignments to *Atf6* across many *Drosophila* species.

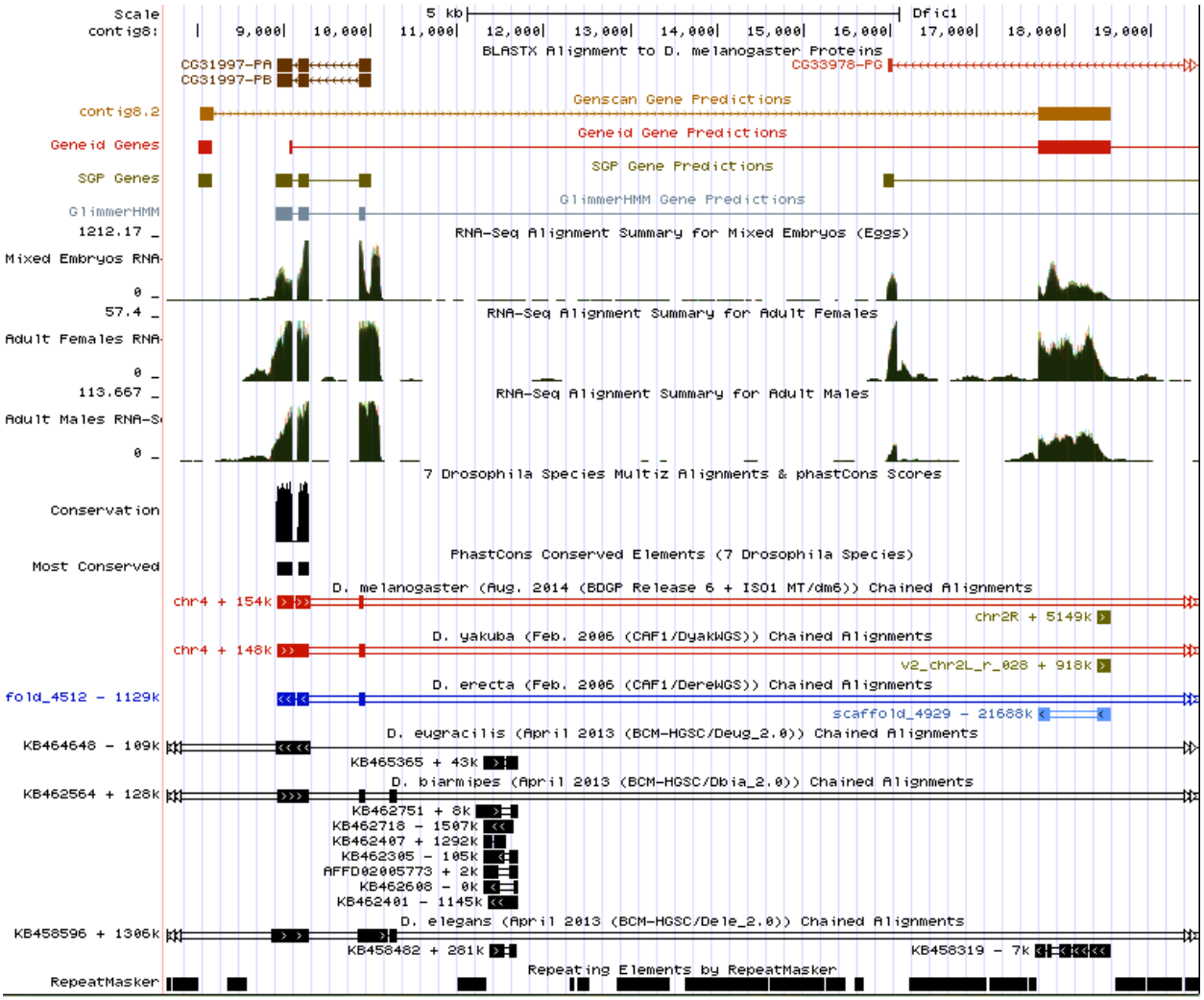


Figure 70: GenScan prediction contig8.2 in the UCSC Genome Browser. Though RNA-Seq coverage is significant on the second exon, chained alignment tracks show that this feature aligns to chromosomes 2R and 2L in *D. melanogaster* and *D. yakuba* respectively, evidence that contig8.2 represents a transposition event, potentially a pseudogene.

Similarly, SGP Genes contig8.2 produced significant alignments to ribosomal protein S14a (RpS14a) across many *Drosophila* species (Figure 71). However, *RpS14a* is located on the X chromosome in all other *Drosophila*, and is thus unlikely to correspond to a real gene on the dot chromosome. Given its position within a high repeat density region, it is likely that SGP Genes contig8.2 is a pseudogene of *RpS14a* that has been transposed from the X chromosome to chromosome four in the *D. ficusphila* line. Chained alignment tracks show that SGP Genes

contig8.2 aligns to chromosome 2R in *D. melanogaster* and chromosome 2L in *D. yakuba*, further supporting this hypothesis (Figure 72).

Description	Max score	Total score	Query cover	E value	Ident	Accession
RH04612p [Drosophila melanogaster]	63.9	63.9	83%	3e-11	54%	AAL48136.1
RpS14a [Drosophila busckii]	63.9	63.9	83%	5e-11	54%	ALC48749.1
ribosomal protein S14a, isoform B [Drosophila melanogaster]	63.9	63.9	83%	5e-11	54%	NP_524884.1
uncharacterized protein Dvir_GJ16536 [Drosophila virilis]	63.9	63.9	83%	5e-11	54%	XP_002057096.1
GM11985 [Drosophila sechellia]	63.9	63.9	83%	5e-11	54%	XP_002044467.1
uncharacterized protein Dmoj_GI21595, isoform A [Drosophila mojavensis]	63.9	63.9	83%	5e-11	54%	XP_002010646.1
GH24437 [Drosophila grimshawi]	63.9	63.9	83%	5e-11	54%	XP_001992030.1
uncharacterized protein Dere_GG19672 [Drosophila erecta]	63.9	63.9	83%	5e-11	54%	XP_001978591.1
uncharacterized protein Dere_GG19671 [Drosophila erecta]	63.9	63.9	83%	5e-11	54%	XP_001978590.1
uncharacterized protein Dana_GF21291, isoform A [Drosophila ananassae]	63.9	63.9	83%	5e-11	54%	XP_001963939.1
GH17341 [Drosophila grimshawi]	62.8	62.8	83%	1e-10	53%	XP_001994623.1
uncharacterized protein Dpse_GA22514 [Drosophila pseudoobscura pseudoobscura]	62.0	62.0	83%	2e-10	53%	XP_002132196.1
uncharacterized protein Dsimw501_GD24764, isoform A [Drosophila simulans]	62.0	62.0	83%	2e-10	53%	XP_016038466.1
uncharacterized protein Dyak_GE15749 [Drosophila yakuba]	62.0	62.0	83%	2e-10	53%	XP_002101212.1
GL18171 [Drosophila persimilis]	62.0	62.0	83%	2e-10	53%	XP_002027050.1
GL15135 [Drosophila persimilis]	62.0	62.0	83%	2e-10	53%	XP_002025532.1

Figure 71: BLASTx search of SGP Genes contig8.2 translated DNA sequence (query) against the nr database (subject) produced significant alignments to RpS14a across many *Drosophila* species.

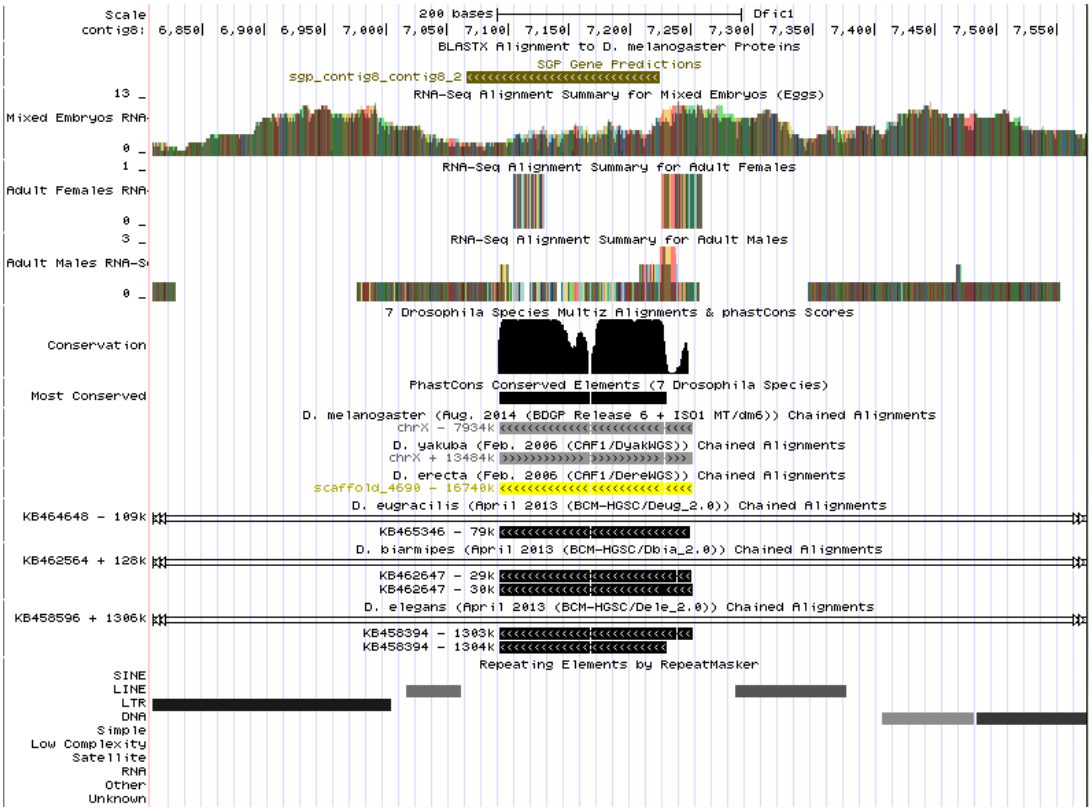


Figure 72: SGP genes contig8.2 in the UCSC Genome Browser. Chained alignment tracks show that this feature aligns to the X chromosome in *D. melanogaster* and *D. yakuba* respectively, evidence that contig8.2 represents a transposition event. Notably, the feature is surrounded by LINEs.

Discussion

Despite being surrounded by high quantities of repetitious sequences usually associated with heterochromatin formation, the approximately eighty genes found on the 1.3 Mb arm of the Muller F element in *Drosophila* are expressed. Contig8 exemplifies this unique nature of the dot chromosome, as evidence of a high level of gene expression was apparent in two of the three annotated genes, *Arl4* and *CG31997,* despite a repeat content of 35.71%. ModENCODE data shows that the remaining annotated gene, *CG33978* is moderately expressed in wing disc CME-W2 cells in the embryo at 16-18 hours and the pupae. Annotation of these genes through integration of information in the UCSC Genome Brower, BLAST, and Clustal Omega multiple sequence alignment revealed varying levels of conservation relative to *D. melanogaster* orthologs. The most notable changes were found in *CG33978*; premature stop codons were found in all isoforms. In addition, TSSs were annotated for *Arl4* and *CG33978*. Further analyses of the TSSs of genes on the fourth chromosome may provide further insights into the mechanism for gene expression in heterochromatic environments.

Contig8 also included three pseudogenes, derived from *mRpL20*, *Atf6*, and *RpS14a,* that were transported by neighboring transposons to the dot chromosome. The presence of these pseudogenes is very interesting because a pseudogene is a rare feature in *Drosophila*. Further characterization of these genes could contribute to future studies that reveal how transposons are able to multiply so rapidly (or be maintained) in the seemingly inaccessible heterochromatic environment of the F element.

Investigation of overall synteny of Contig8 compared to the orthologous region in *D. melanogaster* revealed an inversion involving three genes. This finding is in congruence with previous work by the GEP which showed that the F element has smaller syntenic blocks than

genome averages (3.4-3.6 vs. 8.4-8.8 genes per block), indicating greater rates of inversion despite lower rates of recombination (Leung et al., 2015). In addition, the lack of ncRNA *CR45198* in the *D. ficusphila* genome supports previous research that has found that the relatively low number of ncRNA genes in the *Drosophila* is likely due to the rapid rates of evolution in these types of genes (Drosophila 12 Genomes Consortium, 2007).

The final map of contig8 is shown in Figure 73.

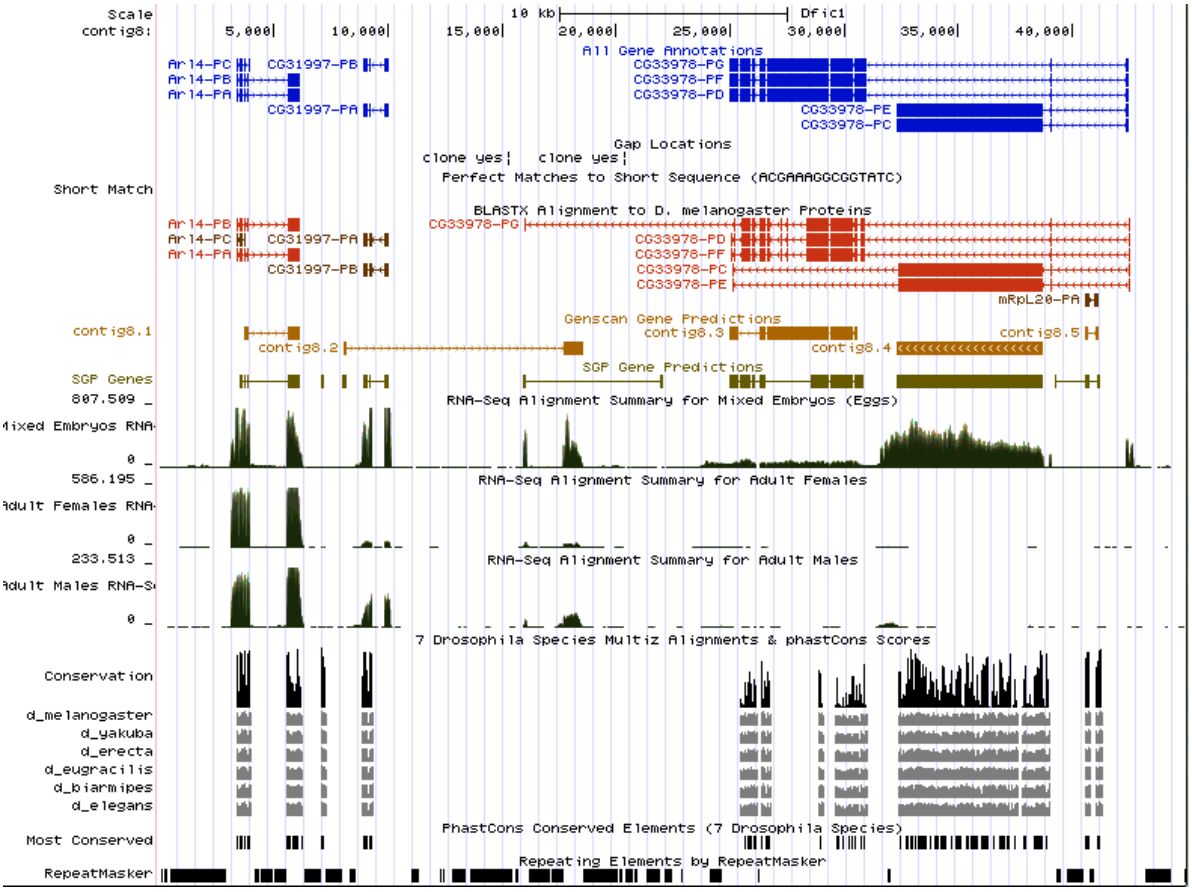


Figure 73: Final map of contig8 with annotated genes in blue custom tracks.

Appendix

Fasta, pep, and gff files are submitted electronically.

Acknowledgements

Thank you to Dr. Elgin, Dr. Shaffer, Dr. Bednarski, Wilson Leung, and Daniel Cui Zhou for their continued support throughout this project.

References

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

Hoskins, R. A., J. M. Landolin, J. B. Brown, J. E. Sandler, H. Takahashi, T. Lassmann, C. Yu, B. W. Booth, D. Zhang, K. H. Wan, L. Yang, N. Boley, J. Andrews, T. C. Kaufman, B. R. Graveley, P. J. Bickel, P. Carninci, J. W. Carlson, and S. E. Celniker. "Genome-wide Analysis of Promoter Architecture in Drosophila Melanogaster." Genome Research 21.2 (2011): 182-92.

Leung, Wilson et al. “Drosophila Muller F Elements Maintain a Distinct Set of Genomic Properties Over 40 Million Years of Evolution.” G3: Genes|Genomes|Genetics 5.5 (2015): 719–740.