Kevin Lin
Dr. Elgin
Bio 434W
2 May 2016

Annotating contig43

Abstract

Contig43, a 38500bp region on the dot chromosome (Muller F element) of *Drosophila ficusphila,* was annotated to create gene models including coding exons and transcription start sites (TSS) based on orthology to *D. melanogaster*. A variety of tools were used in creating annotation gene models including Genscan, BLAST, Gene Record Finder, Flybase, and the various data tracks of the UCSC genome browser. Additionally, RepeatMasker was used to identify and classify repeats in the contig. Contig43 was found to contain 27% repeats, with four repeats longer than 500bp. All isoforms of the three genes present in the contig – *eIF4G*, *mGluR*, and *4E-T* – were annotated for coding exons and transcription start sites. Synteny is not preserved in terms of relative gene order or orientation as there appears to be an inversion involving *mGluR* and *eIF4G* in *D. ficusphila*. The data from this project will be used to improve our understanding of the role genomic features have on gene expression.

Introduction

The eukaryotic genome can be broadly considered to be in one of two categories based on interphase chromatin packaging: euchromatin and heterochromatin. Euchromatin is loosely packaged chromatin generally located on chromosome arms; these regions are actively transcribed and gene rich. Heterochromatin is tightly packaged chromatin generally located at telomeres and centromeres; such packaging is associated with gene silencing. Unlike the generally well understood mechanisms regulating euchromatic gene expression, much remains to

be learned about heterochromatic gene expression. For example, the *D. melanogaster* Muller F element is almost entirely heterochromatic by most measures, yet contains many actively transcribed genes. This unique mixture of characteristics makes the Muller F element an ideal candidate for comparative genomics analysis to study gene expression in heterochromatin. Comparative genomics is a powerful research tool used to compare genomic features across multiple organisms. Through comparative analysis of euchromatic and heterochromatic genes across multiple species of *Drosophila*, we hope to better understand the roles sequence characteristics and organizational features play in controlling gene expression. *D. ficusphila* was chosen specifically due to its ideal evolutionary distance from *D. melanogaster* for searching for conserved regulatory motifs.

The goal of this annotation project is to create gene models for all of the genes in contig43 of the *D. ficusphila* dot chromosome (F element). A complete gene model for this project can be considered to contain exact beginning and ending sites for each exon as well as identification of the putative transcription start and stop sites. Each gene model must be consistent with known biological constraints and be supported by the available experimental evidence. Annotation was completed by first identifying the probable *D. melanogaster* ortholog and then using the *D. melanogaster* gene model to search for the coding regions and transcription start sites in the *D. ficusphila* contig.

Genes Overview

Genscan predicted three genes on contig43 (Figure 1), which will be referred to as features 1, 2, and 3 in the order that they appear on the contig. Features 1 and 2 are complete gene predictions whereas feature 3 is missing the initial exon. Viewing this contig on the *D. ficusphila* dot chromosome UCSC genome browser (Figure 2) reveals that the three features

predicted by Genscan correspond to high scoring blastx alignments of *D.melanogaster* genes *mGluR*, *eIF4G*, and *4E-T*, indicative of high conservation.

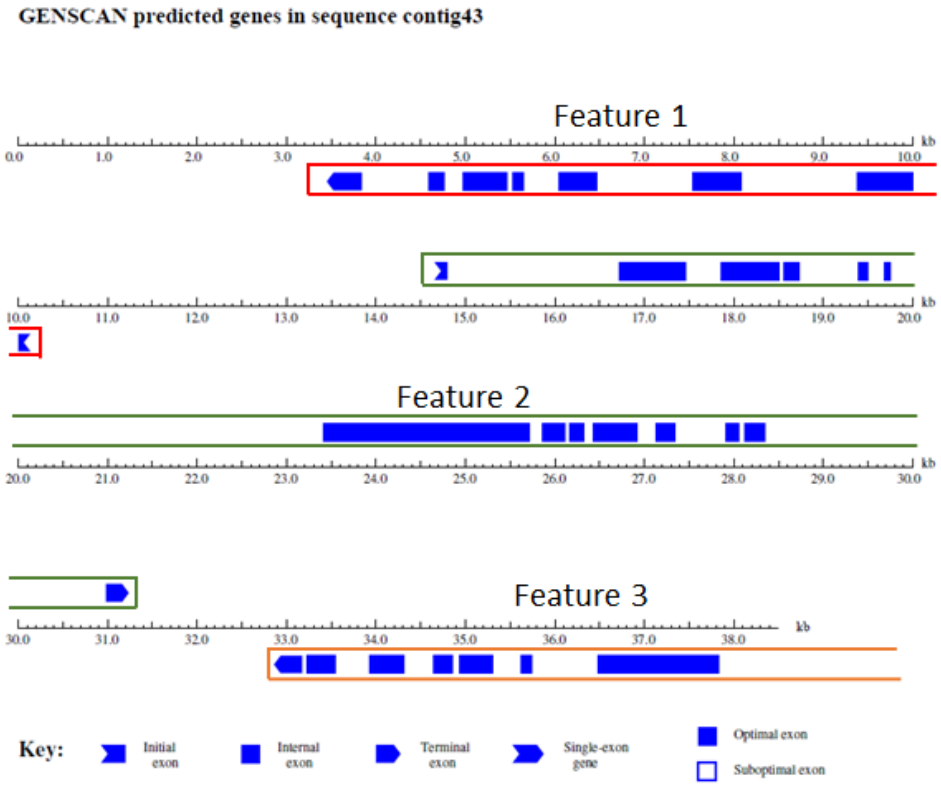


Figure 1: Genscan prediction for contig43. Genscan predicted three putative genes on the contig, labeled feature 1 (red box), feature 2 (green box), and feature 3 (orange box).

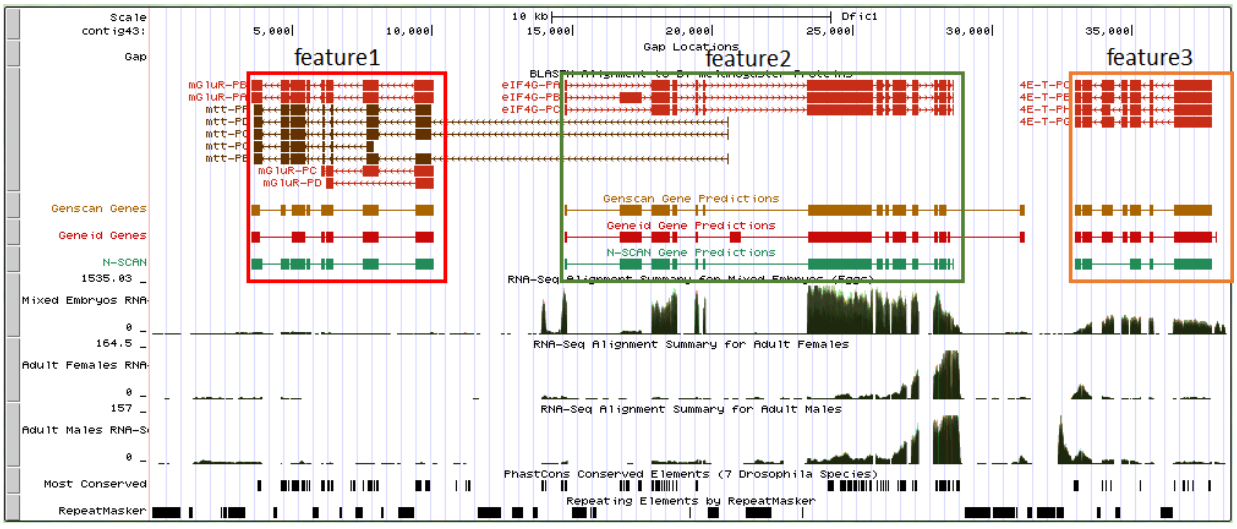
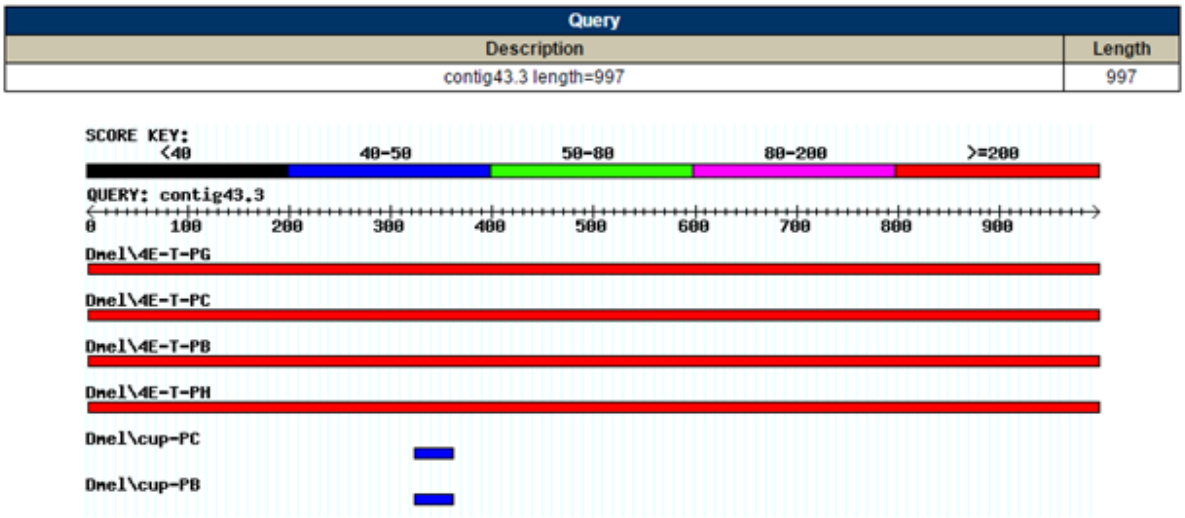


Figure 2: UCSC genome browser of D. ficusphila dot chromosome contig43. Genscan features 1 (red box), feature 2 (green box), and feature 3 (orange box) are marked with their corresponding *D. melanogaster* protein blastx alignments.

Annotation of Feature 3

Feature 3 was selected as the first putative gene to be annotated due to its high conservation with the *D. melanogaster 4E-T* gene and the relatively low number of exons. The amino acid sequence of the Genscan prediction for feature 3 (query) was used to search a database of *D. melanogaster* protein amino acid sequences (subject) using Flybase blastp (Figure 3). The results showed alignments to *4E-T* isoforms G, C, B, and H with an E value of zero, and alignments to cup C and B isoforms with an E value 1.95e-05. Since the blastp alignments to *4E-T* were much stronger than the alignments to cup, the probable *D. melanogaster* ortholog of feature 3 is the *4E-T* gene.



BLAST Hit Summary

Description	Species	Score	E value
4E-T-PG	Dmel	1132.09	0
4E-T-PC	Dmel	1132.09	0
4E-T-PB	Dmel	1132.09	0
4E-T-PH	Dmel	1071.61	0
cup-PC	Dmel	49.2914	1.94991e-05
cup-PB	Dmel	49.2914	1.94991e-05

Figure 3: Flybase blastp of feature 3 (query) searched against the Flybase D. melanogaster proteins amino acid sequences (subject). There are four matches with E-value of zero to isoforms of *4E-T*, and two matches with much worse scores and E-values to cup isoforms. Thus, the ortholog of features 3 was determined to be the *4E-T* gene.

Gene Record Finder was used to find the overall gene model and amino acid sequence of each exon for all four isoforms of *4E-T* in *D. melanogaster* (Figure 4). All four isoforms have identical coding exons except for exon 6, which is shortened in the H isoform relative to the B, C, and G isoforms. A view of the isoforms of *4E-T* in *D. melanogaster* can be seen in GBrowse, which shows that exon 6 of the H isoform is shorter due to alternative splicing at the 5' end (Figure 5).

CDS usage map:

Isoform	1_10866_0	2_10866_1	3_10866_0	4_10866_2	5_10866_2	7_10866_2	6_10866_2	8_10866_2	9_10866_2
4E-T-PB	1	2	3	4	5	6		7	8
4E-T-PC	1	2	3	4	5	6		7	8
4E-T-PG	1	2	3	4	5	6		7	8
4E-T-PH	1	2	3	4	5		6	7	8

Figure 4: Gene Record Finder CDS usage map of D. melanogaster 4E-T isoforms B,C,G, and H. Isoforms B, C, and G share a common sixth exon while isoform H has an alternatively spliced sixth exon. All other coding exons are identical between the 4 isoforms.

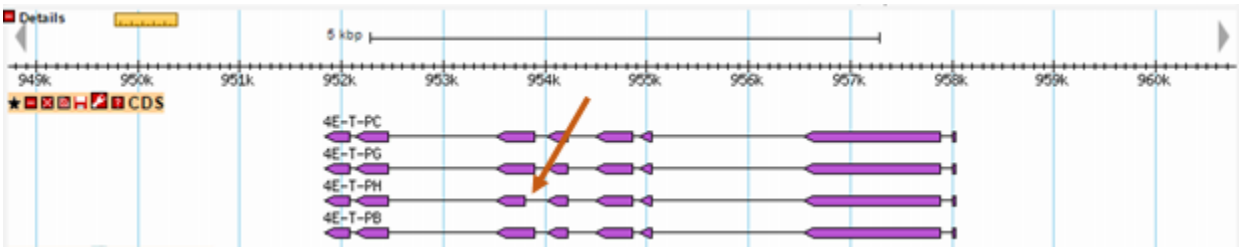


Figure 5: GBrowse view of the D. melanogaster 4E-T isoforms. Exon 6 of the H isoform (orange arrow) is shortened relative to exon 6 of the B, C, and G isoforms.

Coding exons of the *4E-T* gene were identified on the contig by performing a pairwise blastx alignment of the *D. melanogaster* amino acid sequence of each coding exon (subject) to the DNA sequence of contig43 (query). The process was initiated using exon 2 due to its large size, which made it more likely that a meaningful alignment could be found to contig43. The blastx results of searching for the *D. melanogaster 4E-T* exon 2 (subject) with contig43 (Figure

6) showed an alignment containing all 448 bases of the *D. melanogaster* exon with an E-value 1e-161 corresponding to bases 37829 to 36483 reading in the negative frame. This result gives the approximate location for the remainder of the coding exons and restricts other coding exons to also be in the negative frame. Note that blastx searches should usually been performed with compositional matrix adjust turned off to more easily find weaker alignments, but this error was not noticed until after the analysis was performed for all genes. However in this project, this error did not hinder finding the coding exons because the genes in contig43 are well conserved.

Download Graphics

4E-T:2_10866_1

Sequence ID: lcl|Query_224719 Length: 448 Number of Matches: 1

Range 1: 1 to 448 Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
502 bits(1292)	1e-161	Compositional matrix adjust.	277/455(61%)	342/455(75%)	13/455(2%)	-3

```
Query  37829  YTRADLLALRYESKSRQRPQCTNRTELHTLGFWKINFNAASLNVVNNFLNQNKHRLSPEA  37650
              Y++ DLLALRYE KSRQRPQC+ R EL TLGFWKIN N A+L V + + NQNK+RLSPEA
Sbjct  1      YSKVDLLALRYEGKSRQRPQCSTRLELQTLGFWKINLNTAALTVSSAYSNQNKNRLSPEA  60

Query  37649  DNSTLNCSNNGSISSRRALRNRERANNYYQRFIPADSLQAGGGEEKDKDSQADRQSFKLP  37470
              DNS+L CSN+ SISSRRA+RNRERANNYYQRF+P DSL   G E+KDKD+ +  Q +KL
Sbjct  61     DNSSLICSNSSSISSRRAMRNRERANNYYQRFVPTDSLLISG-EDKDKDALSHGQPYKLN  119

Query  37469  VIDHRSISSSHLMPAFAKKRFAAVTGGNGVENSEATVDTSASYRRESKGTPVS-PSRKTI  37293
              +IDHRSISSSHLMPAFAK+RF    G N  E++E  ++T AS     KG   S PSRK
Sbjct  120    IIDHRSISSSHLMPAFAKRRFVISKGSNSEESNEG-INTCAS-----KGKAASSPSRKGS  173

Query  37292  ELDSFETRLNYT-PDHDVGSSSSPTFSTTRQERRIGSGRLLPRNDNWEYKGQKSKEITLD  37116
              ELD+ ET LN+  PDHD   SSSPTFST+RQERRIGSGRLLPR+DNW+YK +K+ E +++
Sbjct  174    ELDTAETCLNFVQPDHDQCMSSSPTFSTSRQERRIGSGRLLPRSDNWDYKNEKTVEASIE  233

Query  37115  AENDSAPIESGGGSGVNQTNQSQHRHRTFSGRLVDRVPEPTDRRFQYDTKRTLDRQGVSG  36936
               E +++P  SG  S +NQ NQSQHR RTFSGRLV+RVPE TDRRFQYD+K++ DRQG++
Sbjct  234    NEKETSPNGSGSTSSLNQHNQSQHRSRTFSGRLVERVPEVTDRRFQYDSKKSFDRQGINN  293

Query  36935  RRLSNKESNNNQSRGKRGNSYQILEEPEWFSGGPTSQLETIDLHGFDELENTEECSGEKD  36756
              RR+S KE  + QSR KRGNSY I EEPEWFS GP SQLETIDLHGF++LE  EE S  +D
Sbjct  294    RRISGKEPFSTQSRSKRGNSYLIHEEPEWFSAGPKSQLETIDLHGFEDLEKNEERSVTED  353

Query  36755  Y-DQFSHKDKKLNVQATNDKSSRRSSNASLN--DPSPLDDMKHIGENKLTFTQNLSEATN  36585
                +Q    DK L+ QA+ D++S R+SN SLN  +  P D+ KH  EN +T  QN ++  +
Sbjct  354    KNNQIQQLDKNLDAQASKDEASMRNSNDSLNFREVIPSDEKKHTDENVVTSIQNSTDLGH  413

Query  36584  QN-NHSSQLHYNQSSESEFNFDAFLNIHPLDHSLM  36483
               N N   Q+  +Q+ ESEFNFDAFLN+HPLD+S++
Sbjct  414    PNKNKPIQMQPSQNPESEFNFDAFLNMHPLDNSVL  448
```

Figure 6: Blastx alignment of D. melanogaster 4E-T exon 2 amino acid sequence (subject) to contig43 DNA sequence (query).

Coding exons 3-7, including both variations of exon 6, were found using the same method of searching with the *D. melanogaster* amino acid sequence of the exon, looking at contig43 DNA sequence using blastx. Example blastx results can be seen in Figures 7, 8, and 9 for exons 3, 4, and 5 respectively. The shorter length of exons 3 and 5 allowed for multiple blastx alignments to meet the E-value threshold. Under these circumstances the correct exon was determined by checking the location, the frame, and the quality of the alignment. For example, the blastx results for the exon 3 search (Figure 7) showed three matches that were all in the negative frame. However, only one alignment was in the approximately correct location of the *4E-T* gene on contig43 (33000-38000 based on the blastx alignment of the contig as a whole), and this alignment also had the highest score, lowest E value, highest percent identity, and most complete coverage. Therefore, the approximate location of exon 3 on contig43 could be determined to be at 35742-35620. Similar techniques were used to find exons 4, 5, 6, 7, and 8. A summary of the approximate location of each exon found using this blastx technique can be seen in Table 1.

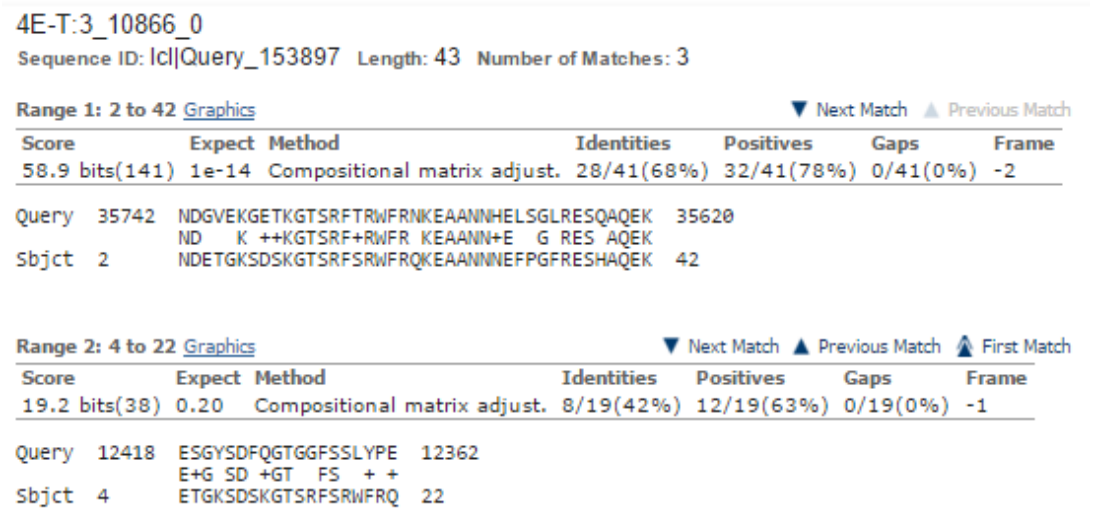
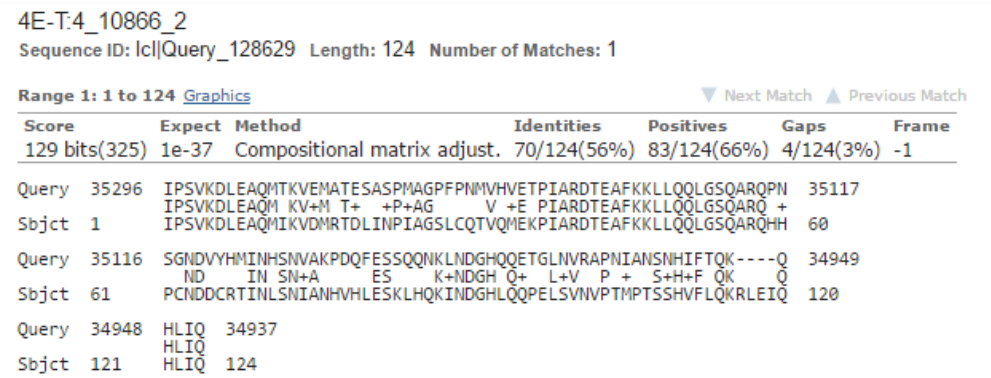


Figure 7: Blastx alignment of D. melanogaster 4E-T exon 3 amino acid sequence (subject) to contig43 DNA sequence (query). The first alignment has the most complete coverage, lowest E-value, is in a negative frame, and falls within the region covered by feature 3. Thus, the first alignment was chosen as the alignment to exon 3. The third match is not shown.



4E-T:4_10866_2

Sequence ID: lcl|Query_128629 Length: 124 Number of Matches: 1

Range 1: 1 to 124 Graphics ▼ Next Match ▲ Previous Match

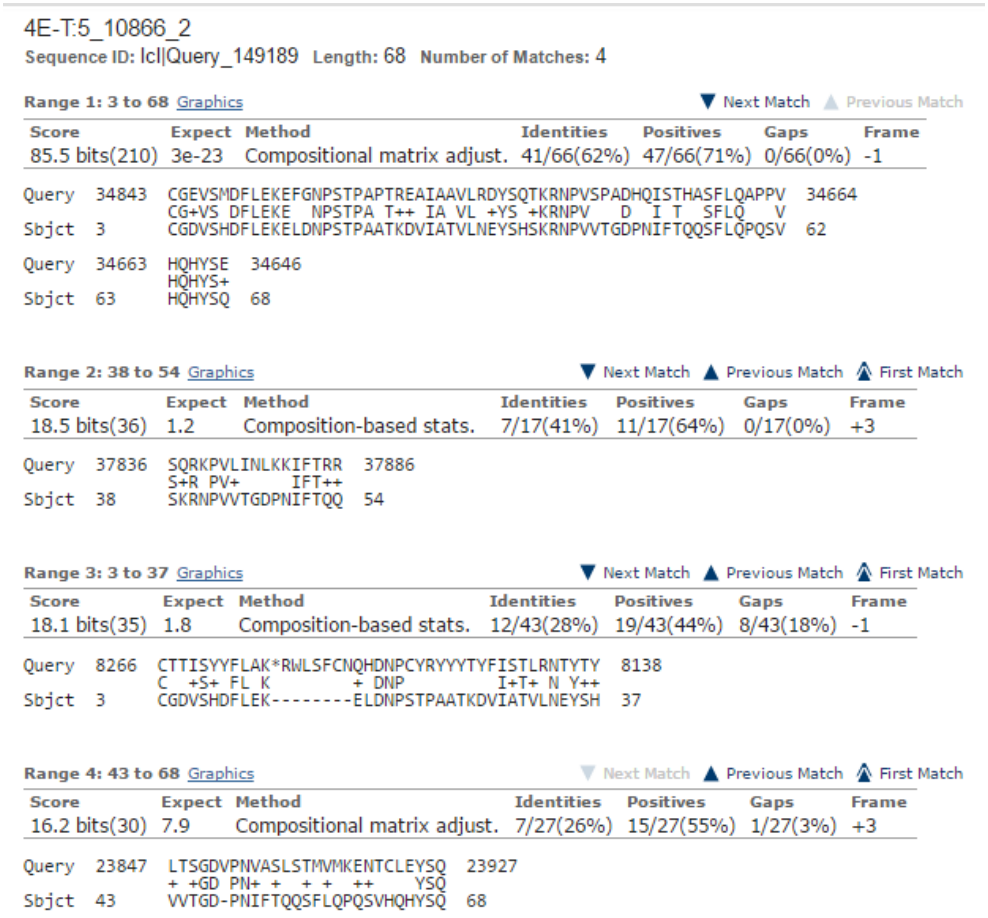
Score	Expect	Method	Identities	Positives	Gaps	Frame
129 bits(325)	1e-37	Compositional matrix adjust.	70/124(56%)	83/124(66%)	4/124(3%)	-1

```
Query  35296  IPSVKDLEAQMTKVEMATESASPMAGPFPNMVHVETPIARDTEAFKKLLQQLGSQARQPN  35117
              IPSVKDLEAQM KV+M T+  +P+AG     V +E PIARDTEAFKKLLQQLGSQARQ +
Sbjct  1      IPSVKDLEAQMIKVDMRTDLINPIAGSLCQTVQMEKPIARDTEAFKKLLQQLGSQARQHH  60

Query  35116  SGNDVYHMINHSNVAKPDQFESSQQNKLNDGHQQETGLNVRAPNIANSNHIFTQK----Q  34949
                ND    IN SN+A     ES    K+NDGH Q+  L+V  P +  S+H+F QK    Q
Sbjct  61     PCNDDCRTINLSNIANHVHLESKLHQKINDGHLQQPELSVNVPTMPTSSHVFLQKRLEIQ  120

Query  34948  HLIQ  34937
              HLIQ
Sbjct  121    HLIQ  124
```

Figure 8: Blastx alignment of D. melanogaster 4E-T exon 4 amino acid sequence (subject) to contig43 DNA sequence (query). There was only one alignment, which has E-value 1e-37, in a negative frame, and falls within the region covered by feature 3. Thus this alignment was chosen as the alignment to exon 4.



4E-T:5_10866_2

Sequence ID: lcl|Query_149189 Length: 68 Number of Matches: 4

Range 1: 3 to 68 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
85.5 bits(210)	3e-23	Compositional matrix adjust.	41/66(62%)	47/66(71%)	0/66(0%)	-1

```
Query  34843  CGEVSMDFLEKEFGNPSTPAPTREAIAAVLRDYSQTKRNPVSPADHQISTHASFLQAPPV  34664
              CG+VS DFLEKE  NPSTPA T++ IA VL +YS +KRNPV   D  I T  SFLQ   V
Sbjct  3      CGDVSHDFLEKELDNPSTPAATKDVIATVLNEYSHSKRNPVVTGDPNIFTQQSFLQPQSV  62

Query  34663  HQHYSE  34646
              HQHYS+
Sbjct  63     HQHYSQ  68
```

Range 2: 38 to 54 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
18.5 bits(36)	1.2	Composition-based stats.	7/17(41%)	11/17(64%)	0/17(0%)	+3

```
Query  37836  SQRKPVLINLKKIFTRR  37886
              S+R PV+     IFT++
Sbjct  38     SKRNPVVTGDPNIFTQQ  54
```

Range 3: 3 to 37 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
18.1 bits(35)	1.8	Composition-based stats.	12/43(28%)	19/43(44%)	8/43(18%)	-1

```
Query  8266  CTTISYYFLAK*RWLSFCNQHDNPCYRYYYTYFISTLRNTYTY  8138
             C  +S+ FL K          + DNP        I+T+ N Y++
Sbjct  3     CGDVSHDFLEK--------ELDNPSTPAATKDVIATVLNEYSH  37
```

Range 4: 43 to 68 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
16.2 bits(30)	7.9	Compositional matrix adjust.	7/27(26%)	15/27(55%)	1/27(3%)	+3

```
Query  23847  LTSGDVPNVASLSTMVMKENTCLEYSQ  23927
              + +GD PN+ +  + +  ++     YSQ
Sbjct  43     VVTGD-PNIFTQQSFLQPQSVHQHYSQ  68
```

Figure 9: Blastx alignment of D. melanogaster 4E-T exon 5 amino acid sequence (subject) to contig43 DNA sequence (query). There were four alignments, but the first alignment has the best E-value, is in the negative frame, and falls within the region covered by feature 3, and covers the entire *D. melanogaster* sequence. Thus, the first alignment was chosen as the alignment to exon 5.

Table 1: Approximate exon locations of D. ficusphila 4E-T isoforms based on blastx searches. Note that exon 1 could not be found through the blastx searches.

Exon (all isoforms unless otherwise noted)	Approximate location	Frame
1	n/a	n/a
2	37829-36483	-3
3	35742-35620	-2
4	35296-34937	-1
5	34849-34646	-1
6 (B, C, G)	34316-33933	-3
6 (H)	34223-33933	-3
7	33544-33233	-1
8	33173-32934	-3

The first exon could not be located using the blastx technique previously described, most likely due to its small size of eight amino acids. Raising the expect threshold to 1e+50 still yielded no results from the blastx search so the Small Exon Finder was used to search for the first exon. Searching the entire minus strand from the end of contig43 (base 38500) to the beginning of the second exon of *4E-T* (base 37829) for an eight amino acid initial exon yielded a single match to an eight amino acid initial exon from bases 37990 to 37965 (Figure 10). The sequence of these eight amino acids is MDVTKSKS, which only aligns to the corresponding *D. melanogaster* sequence of MDTSKISA at the first two amino acids. However, the two sequences have similar composition, which would be expected for orthologous sequences. Support for this exon by other data tracks was needed given the relatively poor conservation at the amino acid level. Thus, this region in *D. ficusphila* was more closely analyzed using the UCSC Genome browser, specifically looking at the translated nucleotide, RNA expression, and TopHat junction tracks (Figure 11). All of these lines of evidence support the hypothesis that this is the correct identification of the first exon by the Small Exon Finder. (For all relevant figures of the UCSC genome browser, the reading frame will be indicated by a green box, canonical splice donor/acceptor sites by a blue box, and the selected splice donor/acceptor site by a red box.)

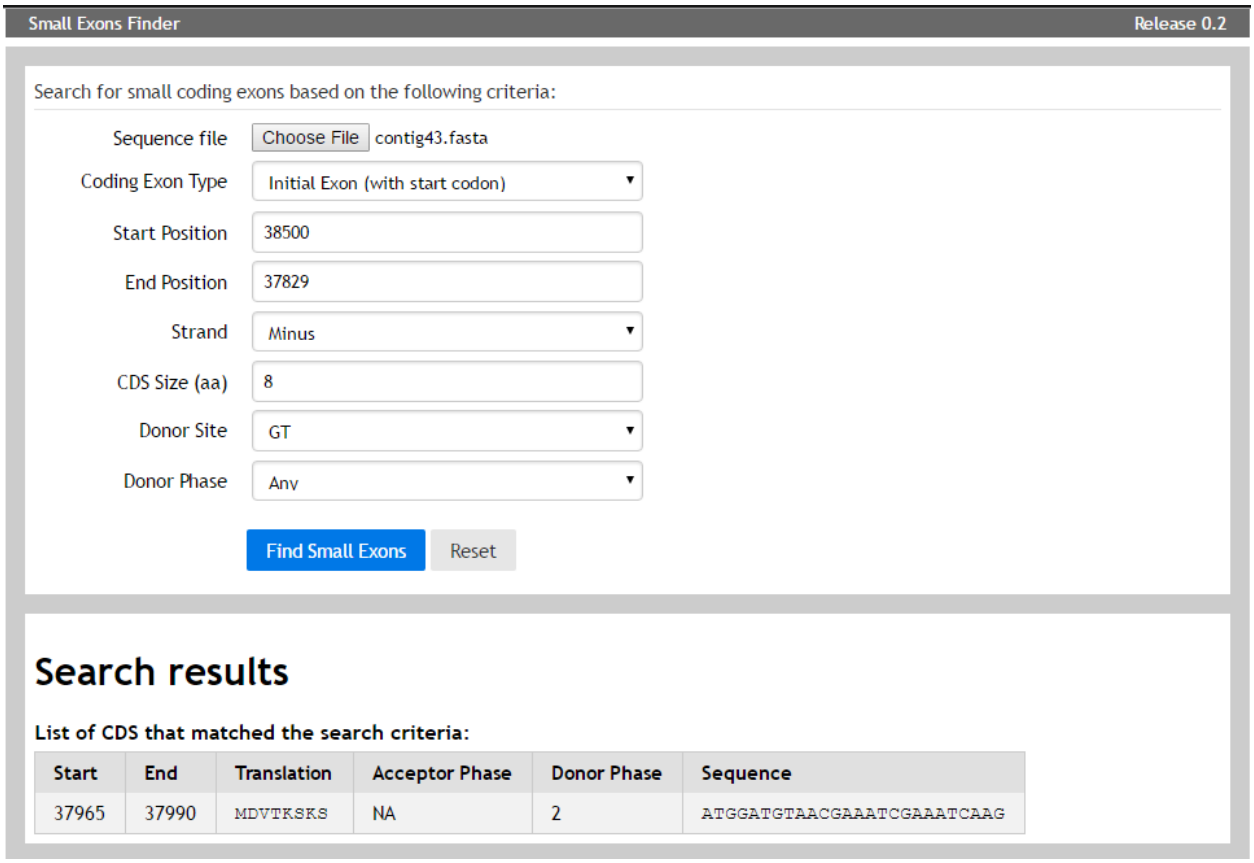


Figure 10: Small Exon Finder search for the first exon of 4E-T. Note that the Donor Phase could have been specified here upon annotation of exon 2.

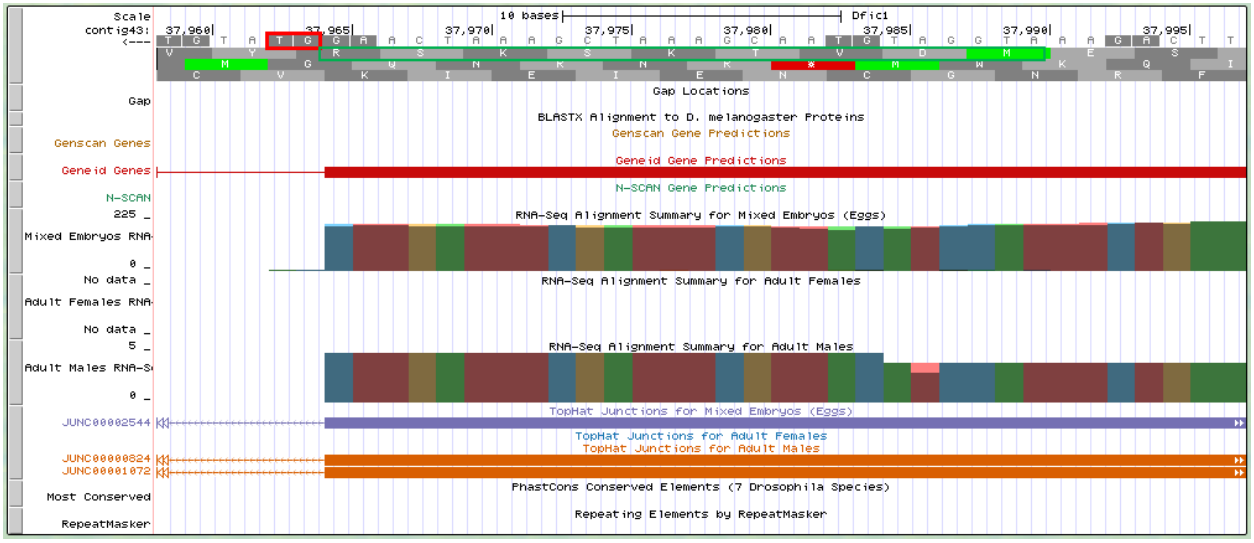


Figure 11: UCSC Genome browser view of the first exon found by Small Exon Finder. The exon begins at base 37990 in frame -1 and ends at base 37965. The donor site is supported by RNA expression data and TopHat junctions.

Splice sites were determined by using the UCSC genome browser to search for canonical splice sites, GT for donor sites and AG for acceptor sites, near the edges of the blastx alignment for each exon. Splice sites should be consistent with available RNA-Seq expression and TopHat junction data. In addition, the phase of the donor site of one exon and the phase of the acceptor site of the following exon must add up to either zero or three to preserve the reading frame. The vast majority of splice sites for the *4E-T* gene were unambiguous and supported by all tracks of data. Figures 12 through 18 illustrate determining splice sites for the first four coding exons and the stop codon of the *4E-T* gene.

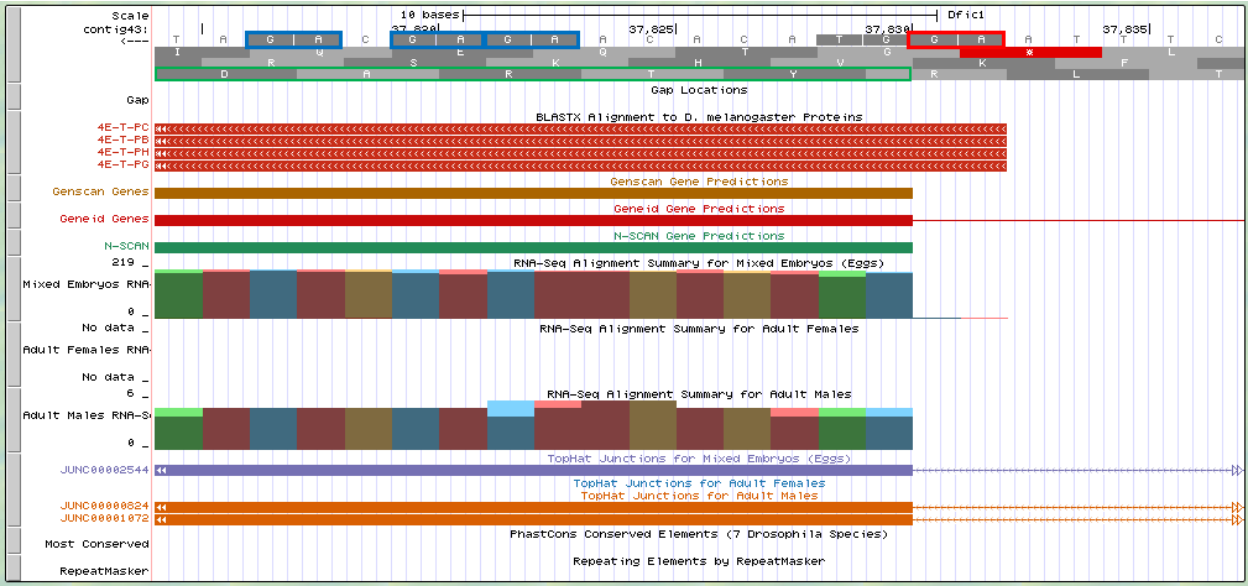


Figure 12: UCSC genome browser view of splice acceptor site of exon 2. Exon 2 is in frame -3. The chosen acceptor site is consistent with RNA expression and TopHat junction data and is in phase 1, which is consistent with the previous splice donor site being in phase 2.

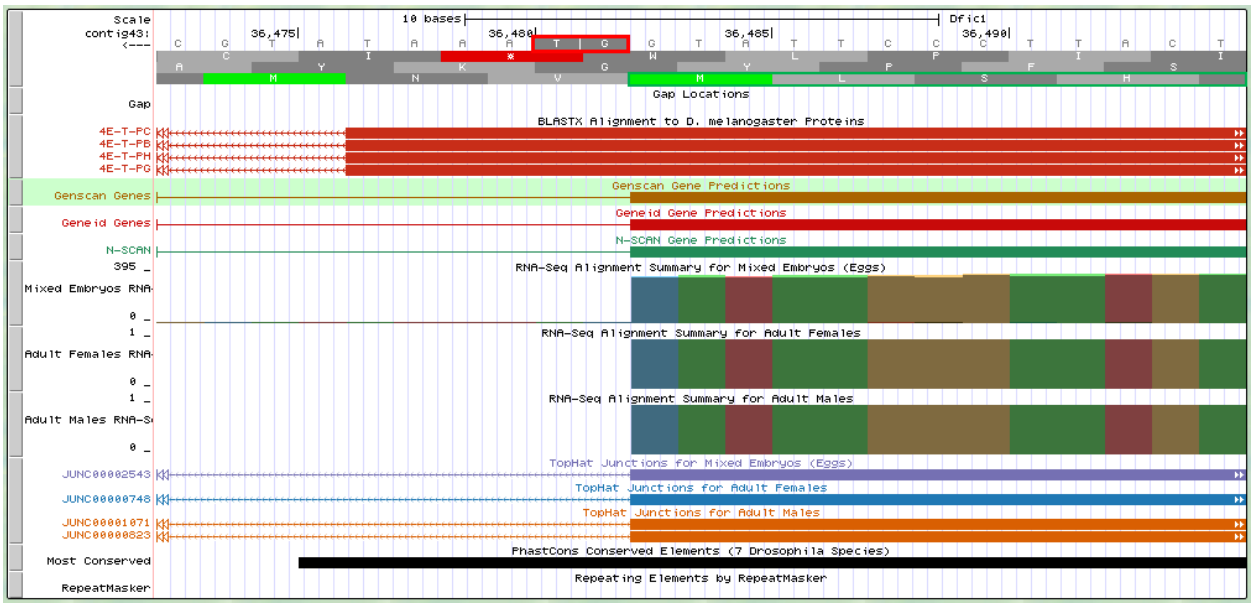


Figure 13: UCSC genome browser view of splice donor site of exon 2. Exon 2 is in frame -3. The chosen donor site is consistent with RNA expression and TopHat junction data and is in phase 0, which is consistent with the following splice acceptor site being in phase 0.

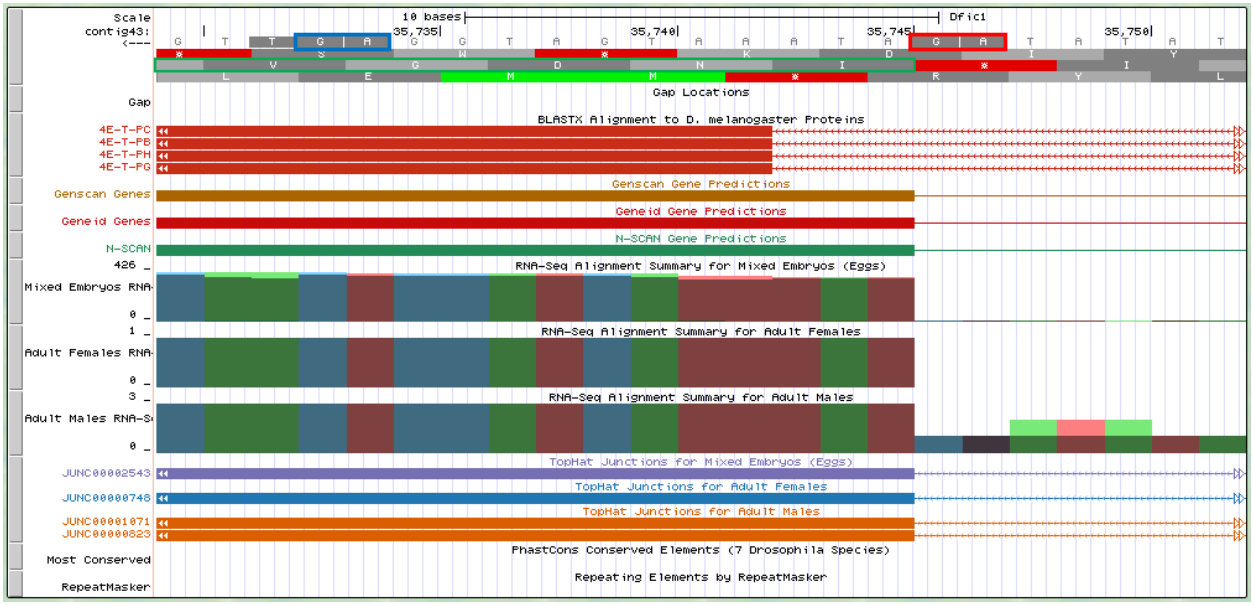


Figure 14: UCSC genome browser view of splice acceptor site of exon 3. Exon 3 is in frame -2. The chosen acceptor site is consistent with RNA expression and TopHat junction data and is in phase 0, which is consistent with the previous splice donor site being in phase 0.

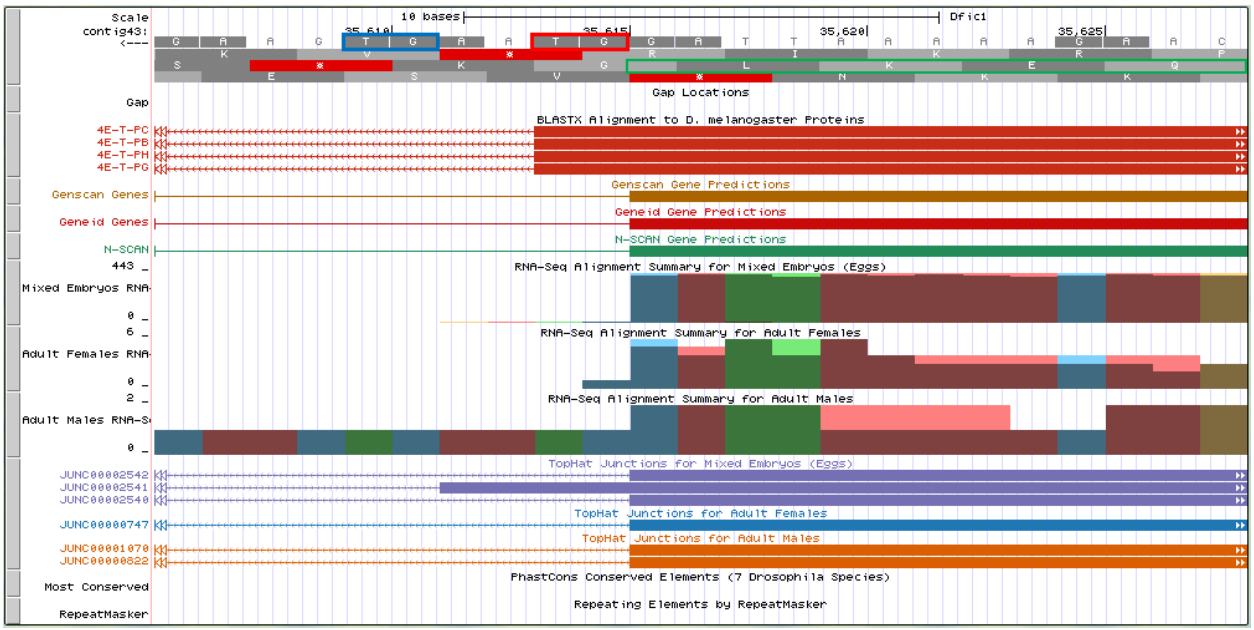


Figure 15: UCSC genome browser view of splice donor site of exon 3. Exon 3 is in frame -2. The chosen donor site is consistent with RNA expression and TopHat junction data and is in phase 1, which is consistent with the following splice acceptor site being in phase 2.

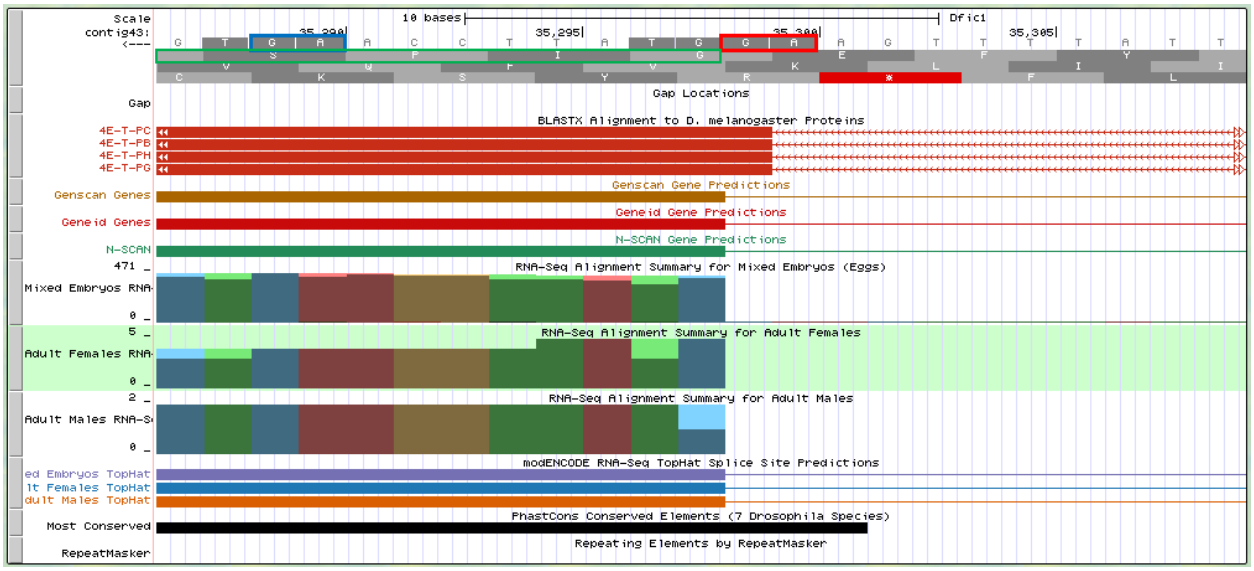


Figure 16: UCSC genome browser view of splice acceptor site of exon 4. Exon 4 is in frame -2. The chosen acceptor site is consistent with RNA expression and TopHat junction data and is in phase 2, which is consistent with the previous splice donor site being in phase 1.

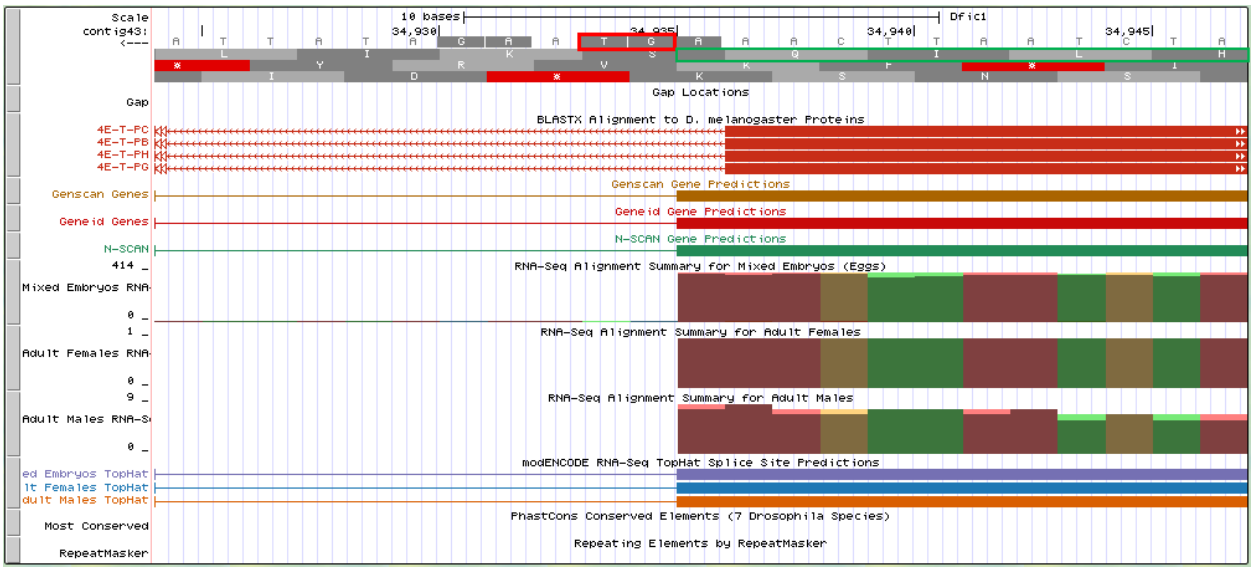


Figure 17: UCSC genome browser view of splice donor site of exon 4. Exon 4 is in frame -1. The chosen acceptor site is consistent with RNA expression and TopHat junction data and is in phase 1, which is consistent with the following splice acceptor site being in phase 2.

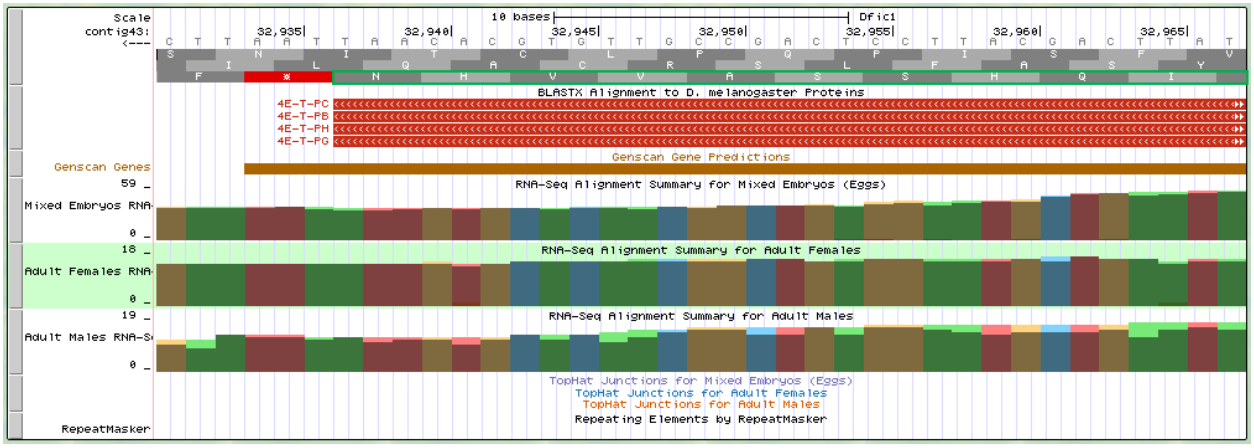


Figure 18: UCSC genome browser view of the stop codon of 4E-T. The terminal exon is in frame -3 and reaches a canonical stop codon.

Exon 6 of the *4E-T* H isoform needed additional attention because the splice acceptor site is contained within the coding region of exon 6, found in isoforms B, C, and G. Any RNA expression from the H isoform is covered by the data from the other isoforms. Therefore, the other data tracks of the UCSC Genome Browser (Figure 19) were used to determine which splice

acceptor site belonged to exon 6 of isoform H. The acceptor site needed to be in phase 2 as the previous donor site was in phase 1. Only one of the two potential splice acceptor sites near the end of the blastx alignment fit this criteria, which was the same site in agreement with the blastx alignment to the *D. melanogaster 4E-T* isoform H.

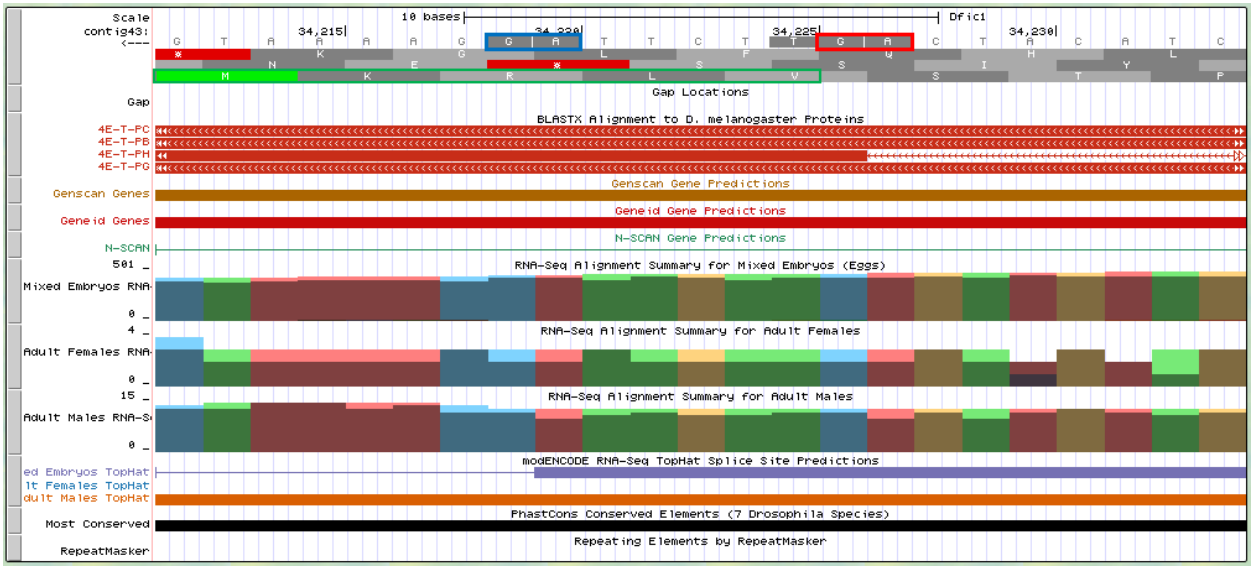


Figure 19: UCSC genome browser near the beginning of the blastx alignment to exon 6 of 4E-T isoform H. Only the first acceptor site is in the correct phase (2) and thus it was chosen as the acceptor site.

Table 2 shows the coding regions for all exons and the location and phase of all splice sites. The results are consistent with the knowledge that donor and acceptor sites of the same intron must add to zero or three in order to preserve the reading frame. The gene model described in Table 2 was entered into the Gene Model Checker, which confirmed viability of the gene model. Figure 20 shows a dot plot of the alignment of the amino acid sequence of the *D. ficusphila 4E-T* gene model (isoform B) and the amino acid sequence of the *D. melanogaster 4E-T* gene. The overall alignment shows good conservation and no gaps between the two species, further supporting the proposed model. The protein alignment shown in Figure 21 shows the same information in more detail, further confirming the conservation and lack of large gaps.

Figure 22 and 23 show the dot plot and protein alignment for the *4E-T* H isoform (which differs in exon 6), which also shows conservation to *D. melanogaster* and no large gaps.

Table 2: Gene model for D. ficusphila 4E-T, isoforms B, C, G, and H.

Exon (all isoforms unless otherwise noted)	Location	Exon size	Frame	Acceptor Phase	Donor Phase
1	37990-37965	26	-1	n/a	2
2	37830-36483	348	-3	1	0
3	35745-35616	130	-2	0	1
4	35298-34936	363	-1	2	1
5	34851-34645	207	-1	2	1
6 (B, C, G)	34318-33932	387	-3	2	1
6 (H)	34225-33932	294	-3	2	1
7	33546-33232	315	-1	2	1
8	33175-32937	239	-3	2	n/a

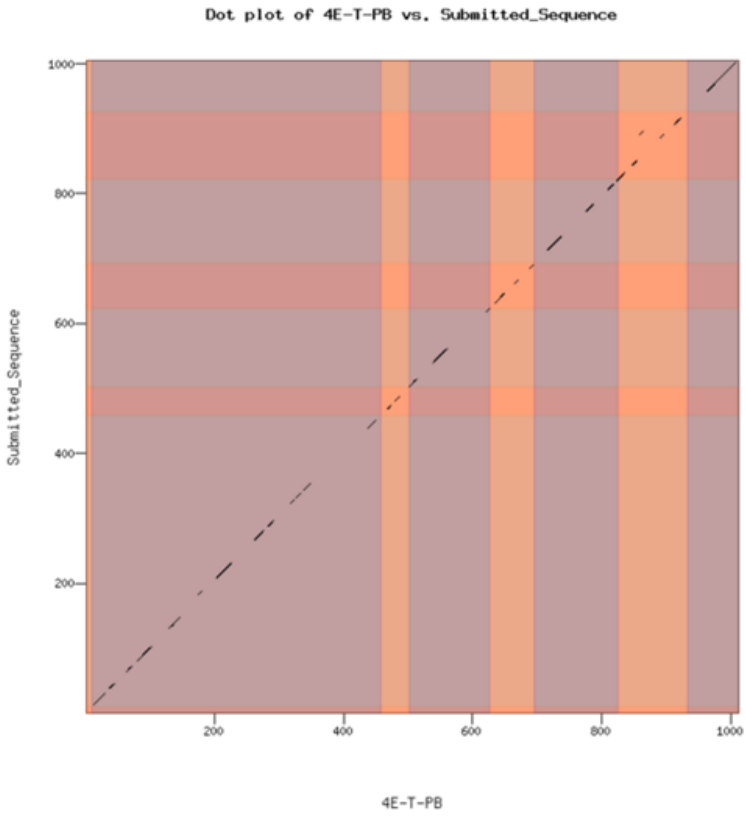


Figure 20: *Dot plot of the submitted gene model for D. ficusphila 4E-T isoforms B, C, and G (y-axis) against the D. melanogaster 4E-T-PB amino acid sequence.* The overall alignment shows a continuous line of slope 1, indicating conservation of all exons with no gaps between the two species.

Figure 21: Protein alignment of the submitted gene model for D. ficusphila 4E-T isoforms B, C, and G. The alignment shows conservation throughout all exons and no large gaps, confirming the plausibility of the submitted gene model.

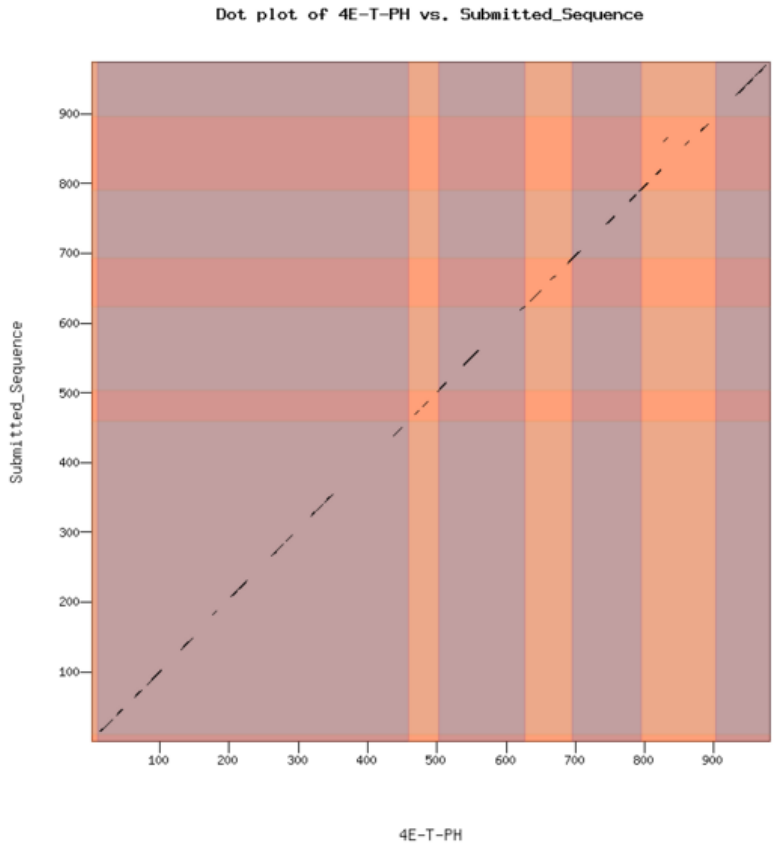


Figure 22: Dot plot of the submitted gene model for D. ficusphila 4E-T isoform H (y-axis) against the D. melanogaster 4E-T-PH amino acid sequence. The dotted line with constant slope 1 indicates conservation with no major gaps in alignment to *D. melanogaster*.

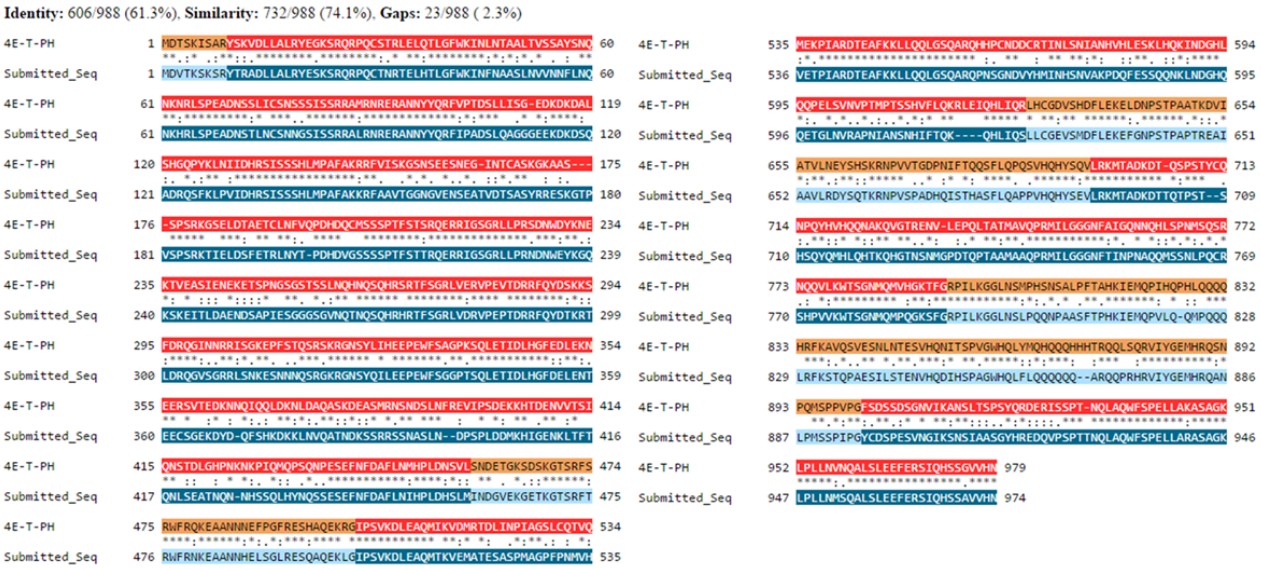


Figure 23: Protein alignment of the submitted gene model for D. ficusphila 4E-T isoform H. Similar to the alignment for isoforms B, C, and G, this alignment shows conservation throughout all exons and no large gaps, confirming the plausibility of the submitted gene model.

In addition to the coding exons, the putative TSS of the *4E-T* gene in *D. ficusphila* was annotated as part of the annotation process. The TSS in *D. melanogaster* was found using Gene Record Finder (Figure 24), which shows that isoforms B, C, and H share a common TSS. The TSS of isoform G is located 222bp downstream of the shared TSS for isoforms B, C, and H. A blastn search using modified parameters (word size 7, +1/-1 match/mismatch, 2 existence 1 extension gap cost, no masking low complexity) was used to align the 5' noncoding exon of isoform C (query) to contig43 (subject). Isoform C was chosen because it has the largest 5' noncoding exon and also overlaps the noncoding exons of isoform B, G, and H. The best alignment (Figure 25) shows a putative TSS at base 38355 for the isoforms B, C, and G, which also suggests that the putative TSS for isoform H is 222 bp away at base 38137. The putative TSS for isoform H was quickly checked by comparing the nucleotide sequence starting at base 38137 to the *D. melanogaster* sequence. Both started with bases "TGTGTA", supporting the putative TSS of isoform H at base 38137.

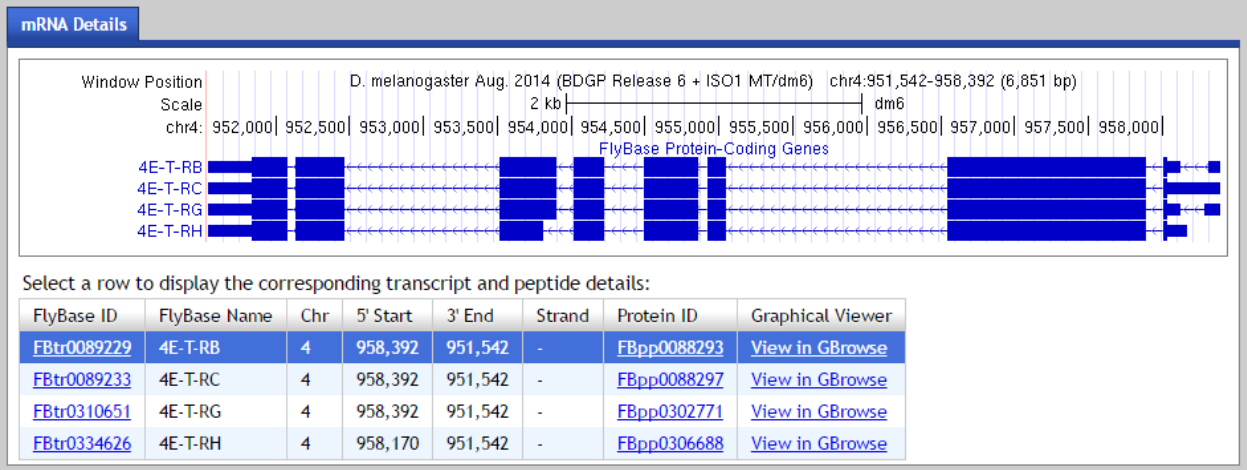


Figure 24: Gene Record Finder mRNA details of 4E-T isoforms in D. melanogaster. Isoforms B, C, and G share a TSS at 958392 while isoform H has a TSS 222bp downstream at 958170.

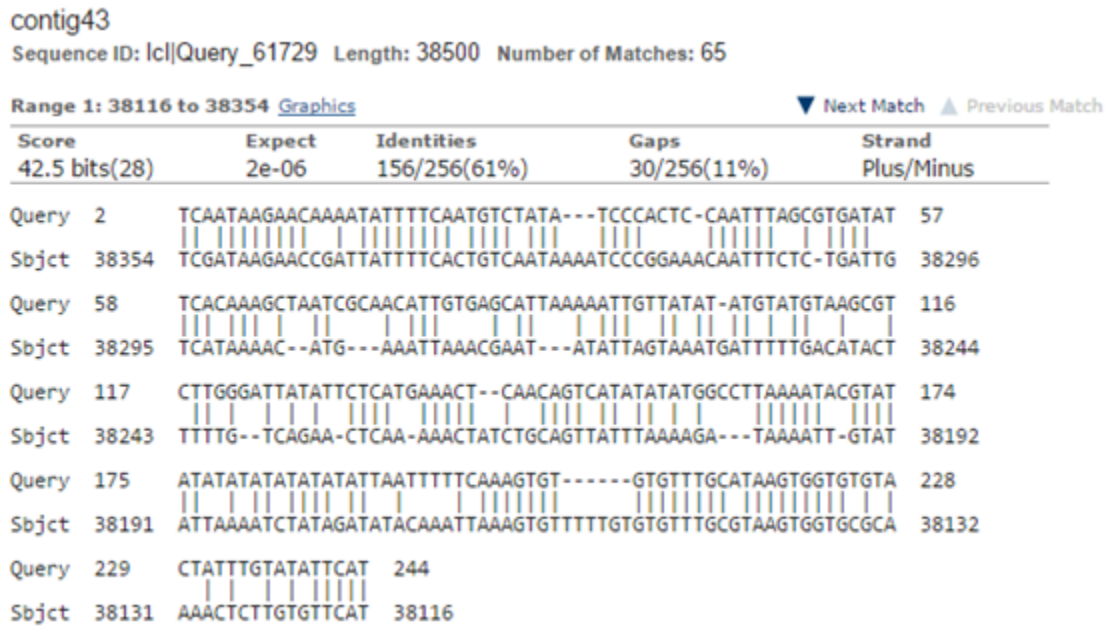


Figure 25: Highest scoring blastn alignment of 5' noncoding exon of 4E-T isoform C (query) to contig43 (subject). Extending the alignment to reach the first base of the query, the orthologous TSS in contig43 for *4E-T* isoforms B, C, and G would be 38355.

Using the UCSC Genome Browser's DNase I hypersensitivity (DHS) tracks and Celinker TSS prediction track for *D. melanogaster*, the TSS can be classified as either peaked (single TSS, single DHS position), intermediate (single TSS, multiple DHS positions within 300bp window), broad (multiple TSS, multiple DHS positions within 300bp window), or insufficient evidence (no TSS or DHS positions). The browser for the *4E-T* TSS in *D. melanogaster* (Figure 26) shows a single TSS and a single DHS position, so it can be classified as a peaked promoter. There is no evidence in the browser for a TSS for isoform H. However, a lack of evidence in these data tracks does not prove the lack of a TSS being present. RNA-Seq expression data seen in the *D. ficusphila* genome browser supports the putative TSS (Figure 27).

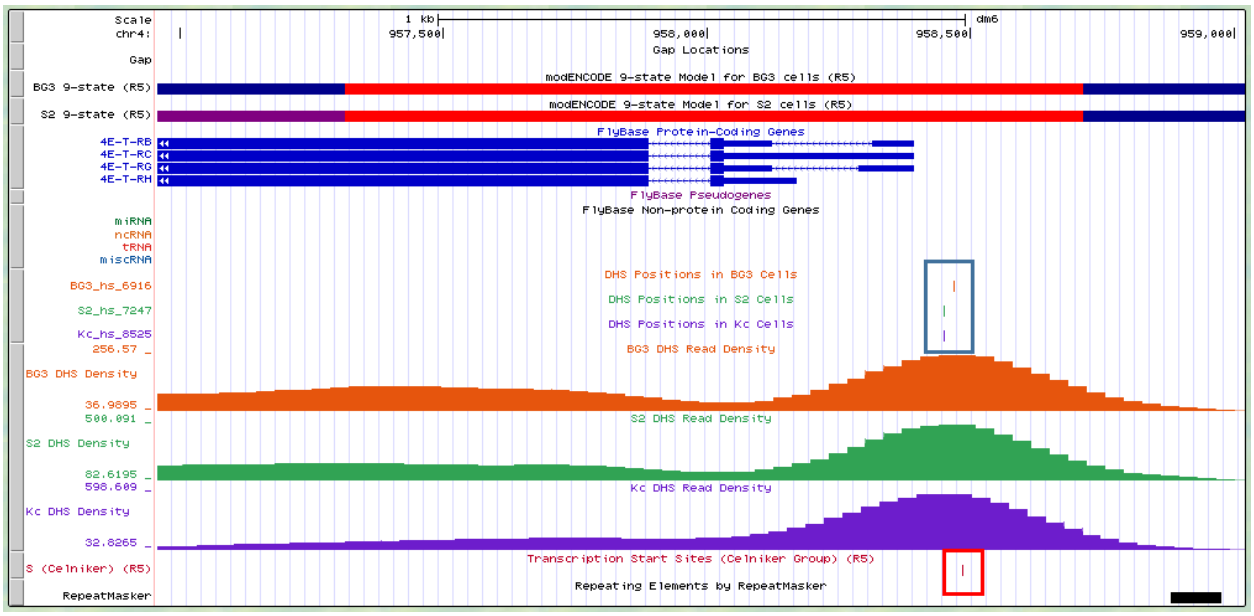


Figure 26: UCSC Genome Browser view of D. melanogaster 4E-T TSS region. There is a single annotated Celniker TSS (red box) and a single DNase I hypersensitivity position in a given cell type (blue box), which would classify this promoter as a peaked promoter.

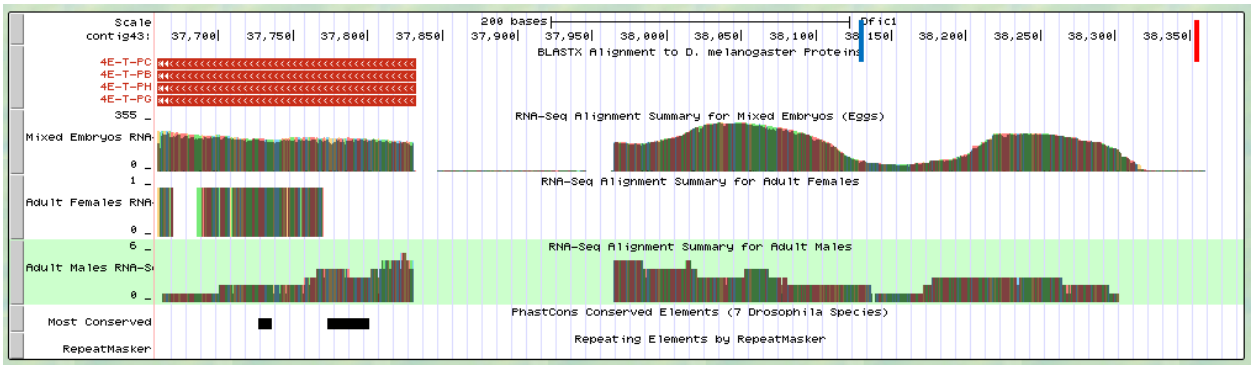


Figure 27: UCSC Genome Browser view of D. ficusphila 4E-T TSS region. The putative TSS for isoforms B, C, and G is marked by the red line and the putative TSS for the H isoform is marked by the blue line. There is supporting RNA-Seq expression data in the regions downstream of the putative TSS.

Sites near the previously identified TSS for *4E-T* in contig43 of base 38355 for isoforms B, C, and G and 38137 for isoform H were searched for core promoter motifs. A region from bases 37950 to 38500 was searched for motifs corresponding to Bre^{u}, TATA, Bre^{d}, Inr, MTE, DPE, DRE, Ohler 1, Ohler 5, Ohler 6, Ohler 7, and Ohler 8 notifs. The motifs found in the search region are listed in Table 3. All of the motifs found in the search region corresponded to Bre^{u}

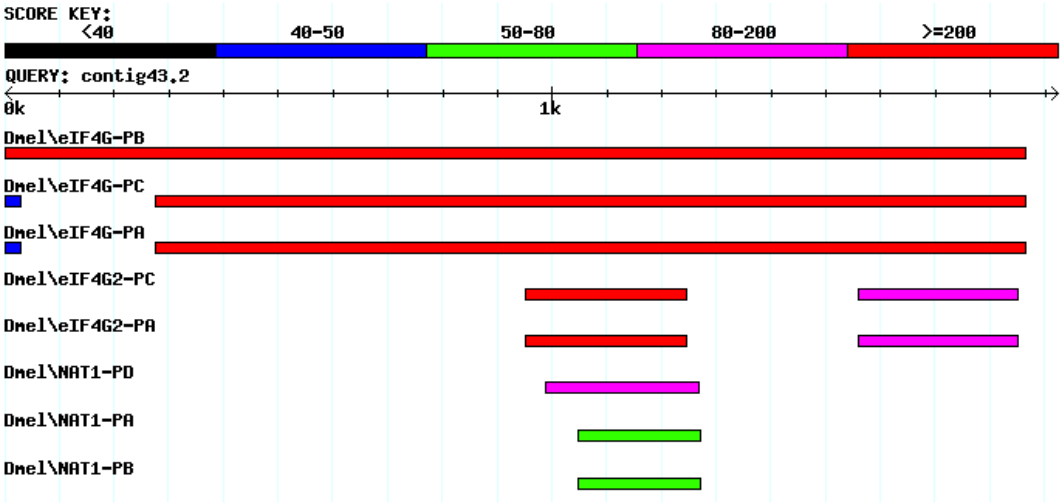
and DPE, which occur with relatively high frequency by chance. There is a single Bre^{d} motif 3 bases from its expected position relative to the TSS of isoform H. However, given the imperfect location and frequency of the motif occurring by chance, the presence of this motif is not very informative. Thus, there is no supporting evidence for either of the putative TSS found in *D. ficusphila* based on conserved core promoter motifs. The majority of genes do not have conserved core promoter motifs, so this finding is not contradictory to previous results, but rather does not offer any additional supporting evidence.

Table 3: Motifs found near 4E-T putative TSS (contig43 bases 37950-38500). The Bre^{u} motif highlighted in yellow is 3 bases away from its expected position and may potentially be informative.

Motif	Position relative to TSS	Expected position for TSS (38355, 38137)	Found Motifs
Bre u	-38	38393, 38175	
TATA	-31 or -30	38385 or 38386. 38137 or 38138	
Bre d	-23	38378, 38150	37954-37960, 37956-37962, 37958-39964, 38045-38051, 38060-38066, 38068-38074, 38147-38153. 38149-38155, 38155-38161, 38157-38163, 38251-38257, 38439-38445, 38458-38464, 38460-38466
Inr	-2	38357, 38139	
MTE	18	38337, 38119	
DPE	28	38337, 38109	37983-37988, 38174-38179, 38211-38216
Ohler 1	N/A	N/A	
DRE	N/A	N/A	
Ohler 5	N/A	N/A	
Ohler 6	N/A	N/A	
Ohler 7	N/A	N/A	
Ohler 8	N/A	N/A	

Annotation of Features 2

The procedure used to annotate feature 3 was used to annotate feature 2. The amino acid sequence of feature 2 (query) was used to search a database of *D. melanogaster* proteins (subject) using Flybase blastp (Figure 28). The results showed matches to isoforms of *eIF4G*, *eIF4G*2, and NAT1. The alignments to *eIF4G* contained the longest matches to the query and had the lowest E-value. Thus, the most probable *D. melanogaster* ortholog to feature 2 is the *eIF4G* gene.



Send BLAST Hits to HitList

BLAST Hit Summary				
	Description	Species	Score	E value
☑	eIF4G-PB	Dmel	2676.35	0
☑	eIF4G-PC	Dmel	2363.57	0
☑	eIF4G-PA	Dmel	2363.57	0
☑	eIF4G2-PC	Dmel	248.054	6.83763e-65
☑	eIF4G2-PA	Dmel	248.054	6.83763e-65
☑	NAT1-PD	Dmel	137.117	1.69986e-31
☑	NAT1-PA	Dmel	68.9366	4.77134e-11
☑	NAT1-PB	Dmel	68.9366	4.77134e-11

Figure 28: Flybase blastp of feature 2 (query) searched against Flybase D. melanogaster proteins amino acid sequences (subject). There are 3 matches with E-value of zero to *eIF4G* isoforms with no other proteins matching with equally strong E-values or scores. Therefore, *eIF4G* was determined to be the orthologous gene to feature 2.

The overall gene model and amino acid sequence of all isoforms the *eIF4G* gene were found using Gene Record Finder (Figure 29). All three isoforms have the same exons with the exception that isoform B contains a single additional exon after the first exon. The gene models for *eIF4G* isoforms can be seen in GBrowse (Figure 30), which shows the additional exon found in the B isoform.

CDS usage map:

Isoform	1_1905_0	2_1905_0	3_1905_0	4_1905_0	5_1905_2	6_1905_1	7_1905_2	8_1905_0	9_1905_2	10_1905_0	11_1905_2	12_1905_0	13_1905_0	14_1905_2	15_1905_2
eIF4G-PA	1		2	3	4	5	6	7	8	9	10	11	12	13	14
eIF4G-PB	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
eIF4G-PC	1		2	3	4	5	6	7	8	9	10	11	12	13	14

Figure 29: Gene Record Finder CDS usage map of eIF4G isoforms A, B, and C. All isoforms are identical with the exception of B, which contains an additional coding exon after the initial exon.

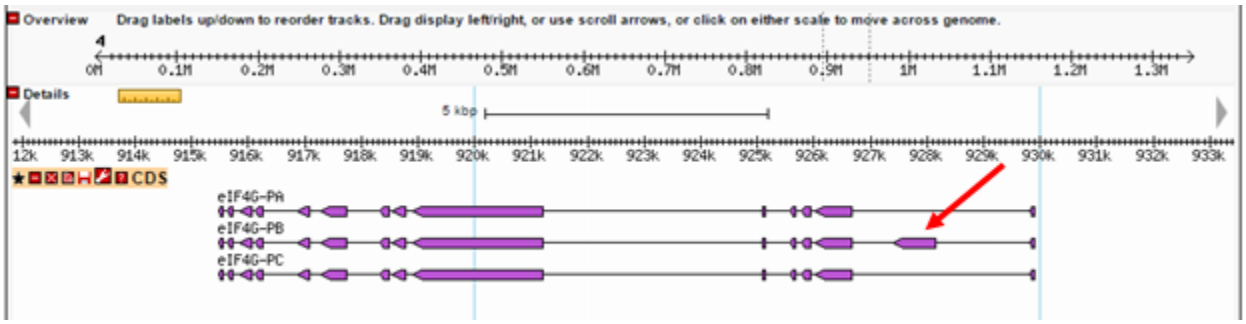


Figure 30: Gbrowse view of the eIF4G gene. The large additional second exon in the B isoform highlighted with the red arrow.

Pairwise blastx alignments were used to find the coding exons of *eIF4G* by searching with the amino acid sequence of each coding exon (subject) to the DNA sequence of contig43 (query). All exons were unambiguously found and the search results were used to locate the approximate splice junctions for each exon. Exact coordinates of the splice sites were located by searching for canonical splice donor and acceptor sites in the *D. ficusphila* UCSC Genome Browser near the approximate splice junction sites found by the blastx searches. All splice sites for the coding exons of *eIF4G* isoforms were supported by unambiguous RNA-Seq expression and TopHat junction data and were in the correct phase to maintain an open reading frame.

Coding regions for all exons and the phase of all splice sites for the isoforms of *eIF4G* are shown in Table 4. The gene models described by Table 4 were put into the Gene Model Checker, which confirmed viability of the gene model for all isoforms. Figure 31 and Figure 32 show the dot plot and protein alignment respectively for isoform B of *eIF4G*, while Figure 33 and Figure 34 show the dot plot and protein alignment respectively for equivalent isoforms A and C of *eIF4G*. Both dot plots and protein alignments show no gaps and high conservation between *D. ficusphila* and *D. melanogaster eIF4G* genes.

Table 4: Gene model for D. ficusphila eIF4G, isoforms A, B, and C.

Exon (all isoforms unless otherwise noted)	Location	Exon size	Frame	Acceptor Phase	Donor Phase
1	14727-14801	75	+3	n/a	0
2 (B only)	16714-17460	747	+1	0	0
3	17852-18499	648	+2	0	0
4	18562-18736	175	+1	0	1
5	19392-19503	112	+2	2	2
6	19688-19743	56	+3	1	1
7	23413-25709	297	+3	2	0
8	25864-26107	244	+1	0	1
9	26167-26318	152	+3	2	0
10	26424-26913	490	+3	0	1
11	27124-27347	224	+3	2	0
12	27914-28060	147	+2	0	0
13	28121-28343	223	+2	0	1
14	28402-28503	102	+3	2	1
15	28570-28628	59	+3	2	n/a

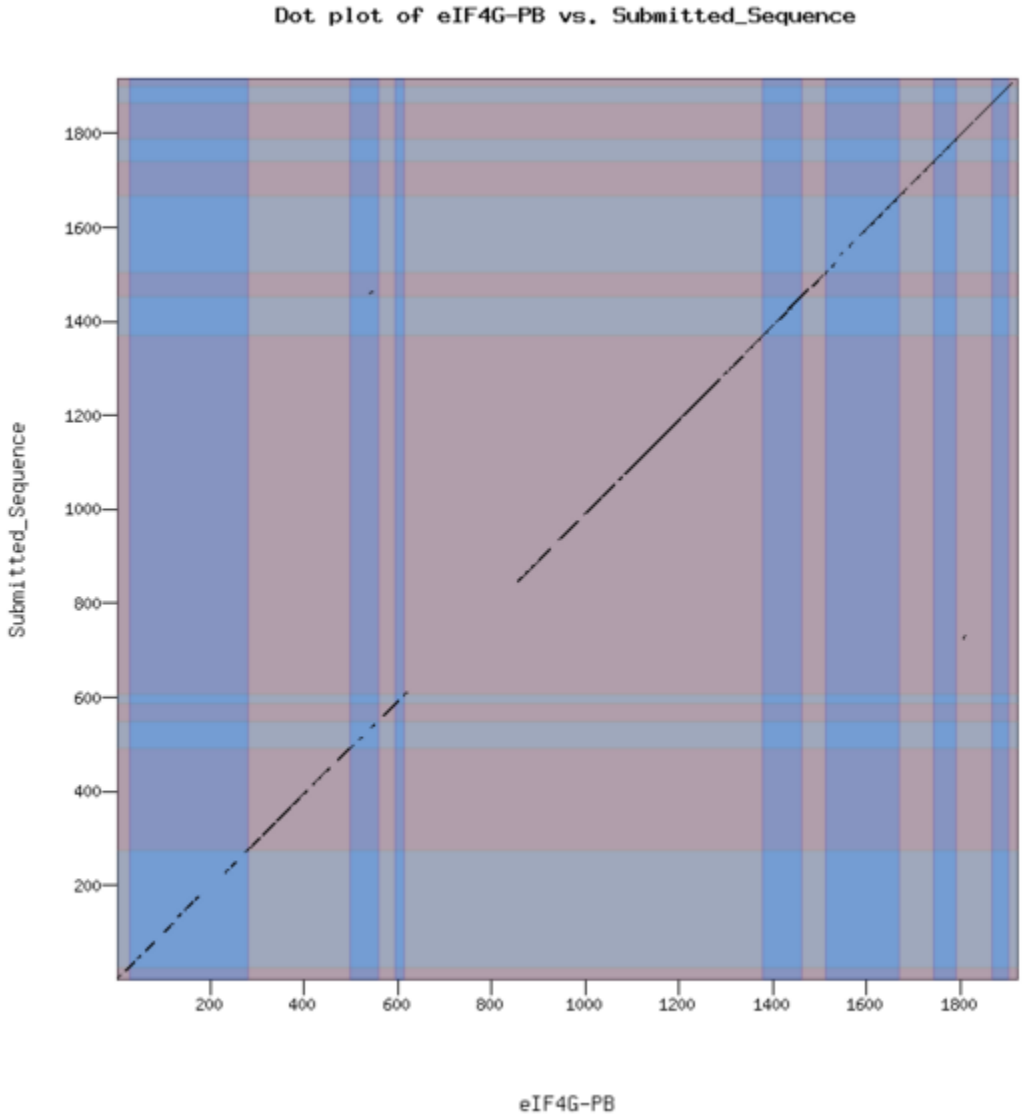


Figure 31: Dot plot of the submitted gene model for D. ficusphila eIF4G isoform B (y-axis) against the D. melanogaster eIF4G-PB amino acid sequence. The constant slope of the plot indicates no major gaps and the mostly solid line indicates conservation to *D. melanogaster* in all but the beginning of the seventh exon.

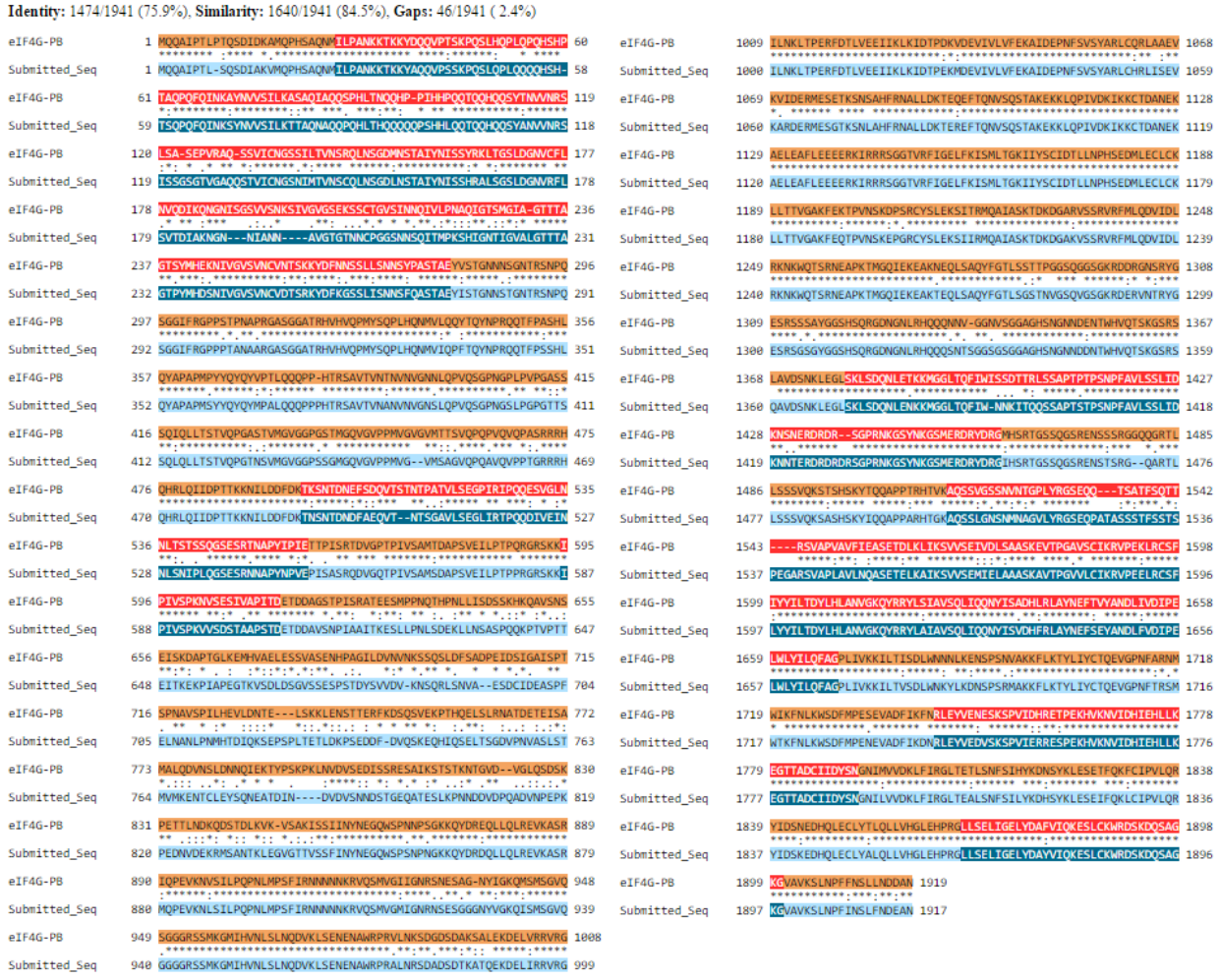


Figure 32: Protein alignment of the submitted gene model for D. ficusphila eIF4G isoform B. The alignment shows no major gaps and conservation to *D. melanogaster*.

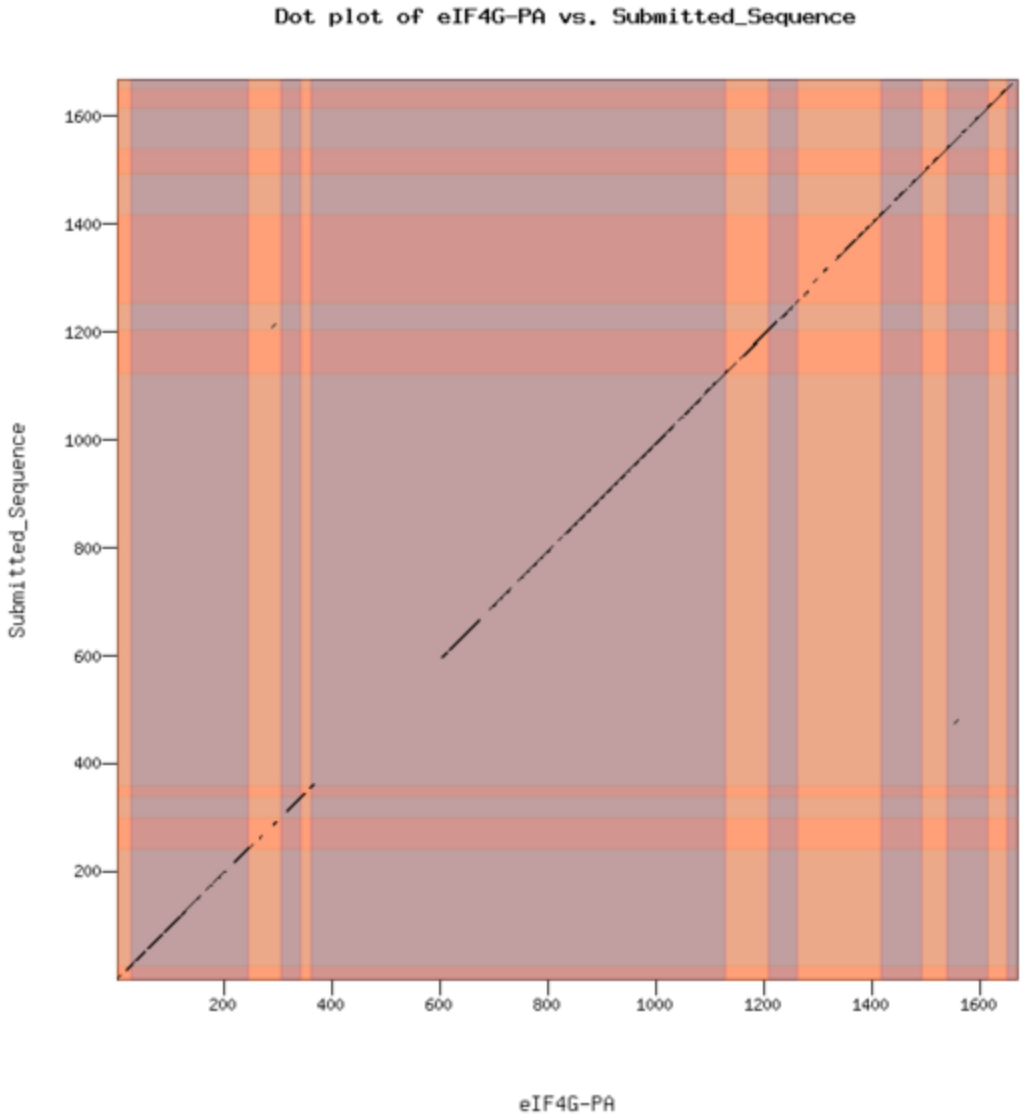


Figure 33: Dot plot of the submitted gene model for D. ficusphila eIF4G isoforms A and C (y-axis) against the D. melanogaster eIF4G-PA amino acid sequence. This dot plot is identical to that of isoform B except it does not contain an alignment corresponding to the second exon of isoform B.

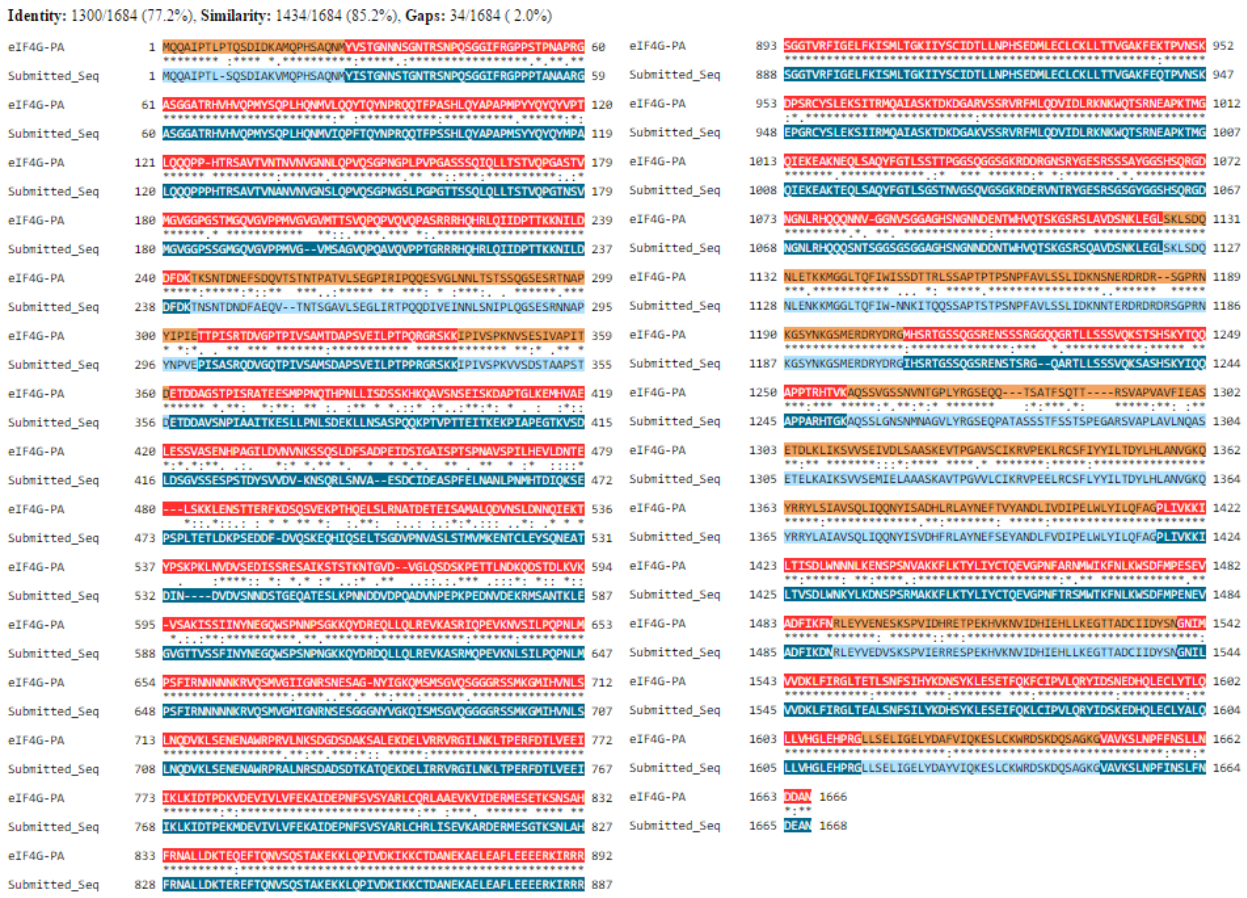


Figure 34: Protein alignment of the submitted gene model for D. ficusphila eIF4G isoforms A and C. Similar to the alignment for isoform B, the alignment shows no major gaps and conservation to *D. melanogaster*.

Putative TSS for *eIF4G* isoforms were found using the same method for finding the putative TSS for *4E-T* isoforms. Gene Record Finder of the mRNA transcript (Figure 35) shows that isoforms A, B, and C share a common TSS in *D. melanogaster*. Using the first untranslated exon of the C isoform, a blastn search using previously described modified parameters identified the orthologous TSS in *D. ficusphila* at base 13893 in contig43 (Figure 36). Viewing the region around the *eIF4G* TSS in the *D. melanogaster* UCSC Genome Browser (Figure 37) showed two annotated Celniker TSS and a single DHS position, which would most closely fall into the category of a broad promoter. The first Celniker position corresponded to the position from the blastn alignment and the second Celniker position was added as an additional putative TSS in *D.*

ficusphila at base 13844 in contig43. RNA-Seq expression data seen in the *D. ficusphila* genome browser supports the putative TSS at base 13887 but does not rule out the putative TSS at base 13844 (Figure 38).Bases 13700-14000 of contig43 were searched for core promoter motifs; the results can be found in Table 5. Bre[d] and DPE motifs were found but not at expected positions, and a DRE motif was found, which does not have a defined position relative to the transcription start site.

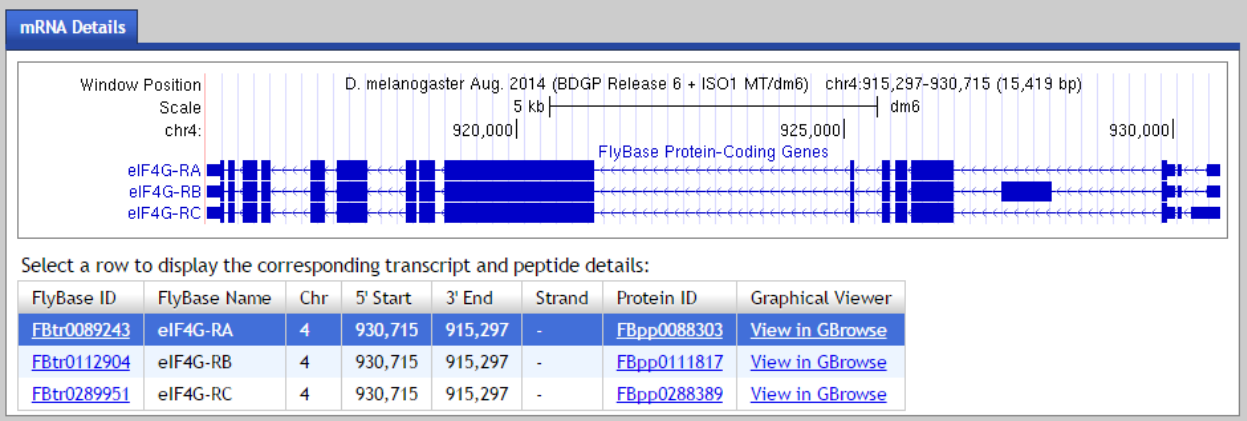


Figure 35: Gene Record Finder mRNA details of eIF4G isoforms in D. melanogaster. All three isoforms share a TSS at base 930715 on the fourth chromosome.

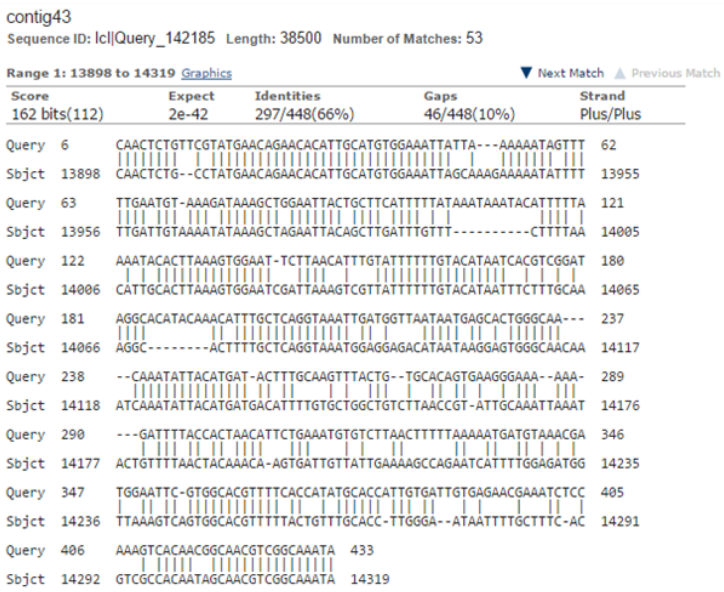


Figure 36: Blastp alignment of first noncoding exon of eIF4G isoform C (query) to contig43 (subject). From the 434bp long query sequence, bases 6-433 aligned. Extending the alignment would have base 13893 in contig43 correspond to base 1 of the query, so contig43 base 13893 was determined to be the putative TSS of *eIF4G* in *D. ficusphila*.

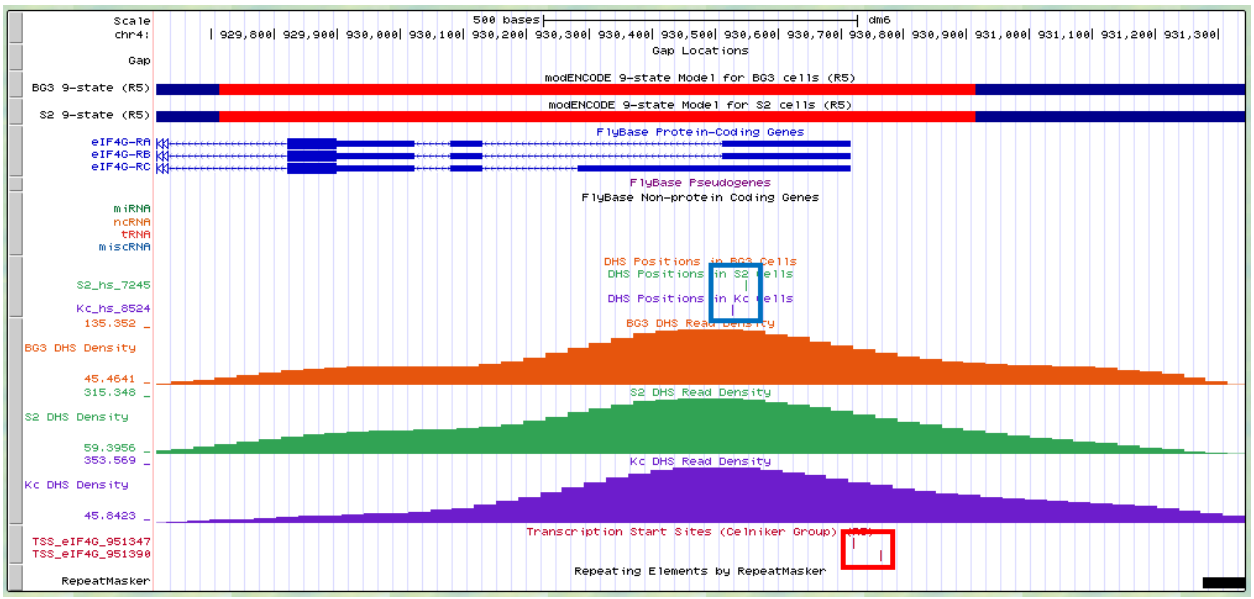


Figure 37: UCSC Genome Browser view of D. melanogaster eIF4G TSS region. There are two annotated Celniker TSS positions (red box) at bases 730721 (roughly corresponding to the blastn alignment at 730715) and 730764 (49 bases upstream of the blastn alignment). These Celniker positions correspond to bases 13887 and 13844 in contig43 respectively. There are also DHS positions corresponding to the peak boxed in blue.

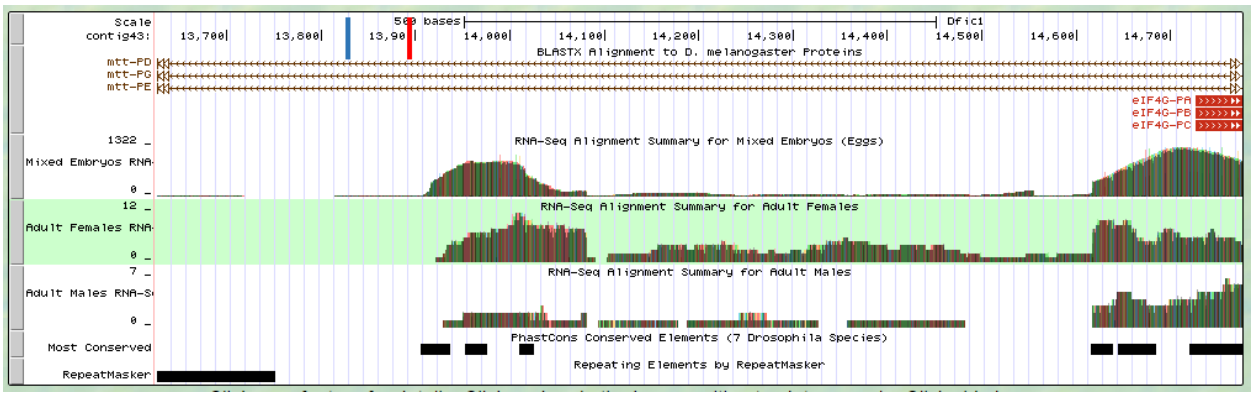


Figure 38: UCSC Genome Browser view of D. ficusphila eIF4G TSS region. The putative TSS based on blastn and the first Celniker annotated TSS is indicated by the red line and the putative TSS based on the second Celniker annotated TSS is indicated by the blue line. The RNA-Seq expression evidence better supports the putative TSS from the blastn alignment and the Celniker annotation at base 13893, but this data cannot rule out the putative TSS at base 13844.

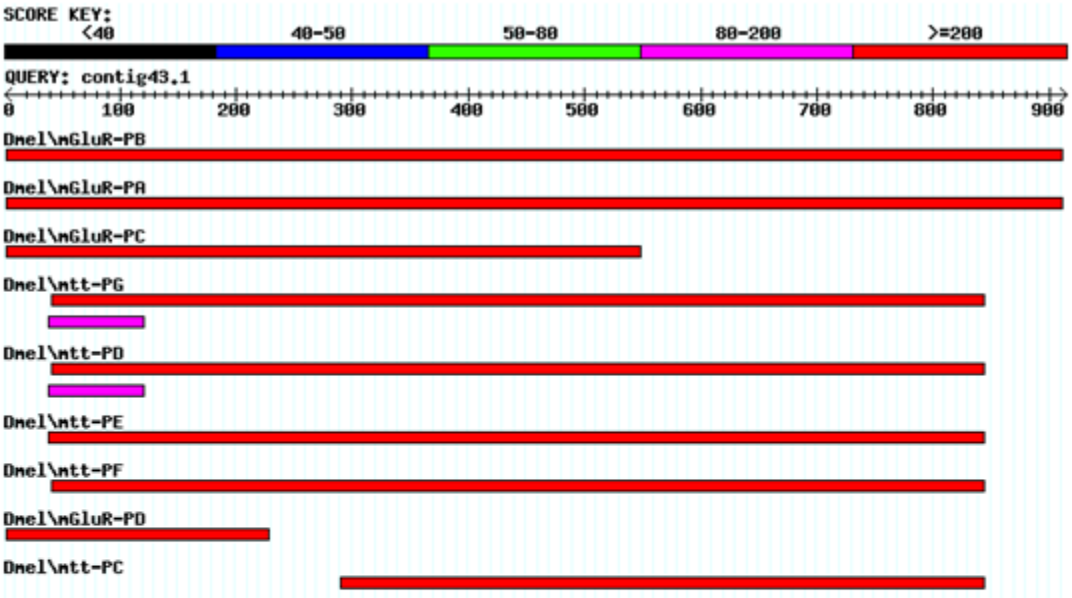
Table 5: Motifs found near eIF4G putative TSS (contig43 bases 13700-14000). The single DRE motif highlighted in green, which does have an expected prediction, may offer slight evidence to support this transcription start site. Expected positions for both putative TSS are listed in the third column.

Motif	Position relative to TSS	Expected position for TSS (13893, 13844)	Found Motifs
Bre u	-38	13855, 13006	
TATA	-31 or -30	13862 or 13863, 13813 or 13814	
Bre d	-23	13870, 13821	13726-13732, 13799-13805, 13840-13846, 13949-13955, 13951-13957, 13991-13997
Inr	-2	13891, 13842	
MTE	18	13914, 13862	
DPE	28	14021, 13872	13759-13764
Ohler 1	N/A	N/A	
DRE	N/A	N/A	13835-13842
Ohler 5	N/A	N/A	
Ohler 6	N/A	N/A	
Ohler 7	N/A	N/A	
Ohler 8	N/A	N/A	

Feature 1

Feature 1 was annotated by using the same procedure used to annotate features 2 and 3. Flybase blastp was used to align the amino acid sequence of feature 1 (query) to a database of *D. melanogaster* proteins (subject). The search results showed alignments to various isoforms of *mGluR* and *mtt* genes (Figure 39). While the *mGluR* isoforms generally show alignments with higher score and lower E-values, this information is not sufficient to determine the orthologous gene given the E-value of *mtt* isoforms is on the order of e-168. To confirm the ortholog of the feature, Gbrowse was used to determine the gene model and location of *mGluR* and *mtt* in *D.*

melanogaster (Figures 40 and 41). The browser shows that in *D. melanogaster*, the coding region of *mtt* spans 46kb, is located next to the *Mal-A8* gene, and is on chromosome 2R. In contrast, in *D. melanogaster*, the coding region of *mGluR* spans 8kb, is located between *eIF4G* and *4E-T* genes, and is on chromosome 4. In addition, the gene model for *mGluR* is consistent with the model for feature 1. Based on the evidence, the orthologous gene for feature 1 is *mGluR*.



BLAST Hit Summary				
✔	Description	Species	Score	E value
✔	mGluR-PB	Dmel	1575.84	0
✔	mGluR-PA	Dmel	1575.84	0
✔	mGluR-PC	Dmel	966.452	0
✔	mtt-PG	Dmel	589.341	4.5868e-168
✔	mtt-PD	Dmel	589.341	5.60374e-168
✔	mtt-PE	Dmel	588.571	1.01334e-167
✔	mtt-PF	Dmel	587.8	1.43863e-167
✔	mGluR-PD	Dmel	426.017	8.00233e-119
✔	mtt-PC	Dmel	375.555	1.32594e-103
✔	CG10738-PF	Dmel	53.1434	1.49485e-06

Figure 39: Flybase blastp of feature 1 (query) searched against Flybase D. melanogaster proteins amino acid sequences (subject). There are three alignments to *mGluR* isoforms A, B, and C that have E-value of zero as well as alignments to *mGluR* isoform D and various isoforms of *mtt* that have E-values stronger than e-100 that align to long regions of feature 1.

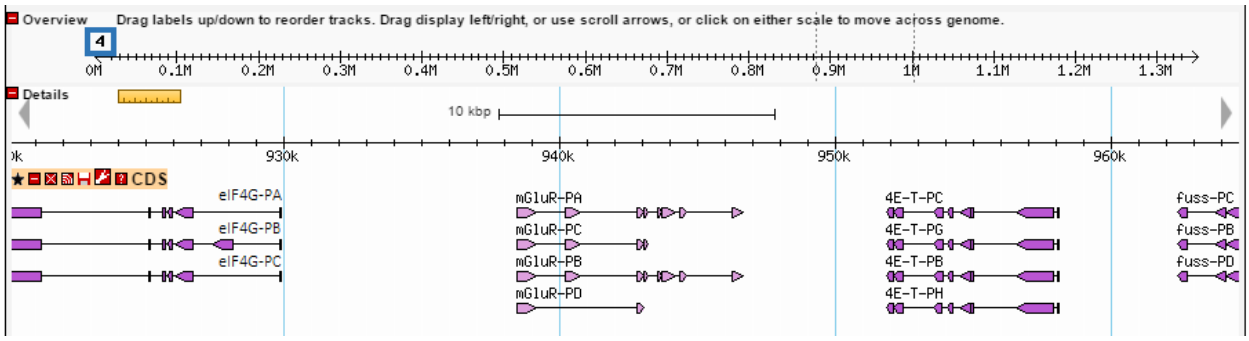


Figure 40: Gbrowse view of the mGluR gene and surrounding genes in D. melanogaster. In *D. melanogaster*, the coding region of *mGluR* spans approximately 8kb, is located on chromosome 4 (blue box), and is between *eIF4G* and *4E-T* genes.

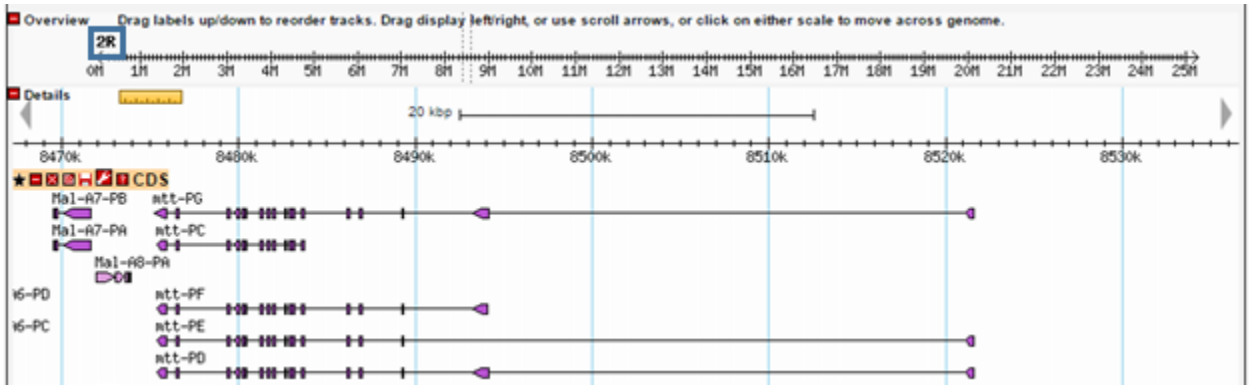


Figure 41: Gbrowse view of the mtt gene and surrounding genes in D. melanogaster. In *D. melanogaster*, the coding region of *mtt* spans approximately 46kb, is located on chromosome 2R (blue box), and flanked on one end by *Mal-A8* and *Mal-A7* genes.

Gene Record Finder was used to find the overall gene model and amino acid sequences for each exon of *mGluR* in *D. melanogaster* (Figure 42). Isoforms A and B have the same coding sequences while isoforms C and D have alternatively spliced exons and do not contain many exons found in isoforms A and B. A detailed view of the gene models for the different isoforms can be seen using the Gbrowse view of the gene (Figure 43).

CDS usage map:

Isoform	1_1702_0	2_1702_2	3_1702_1	4_1702_2	5_1702_2	6_1702_2	7_1702_0	8_1702_0	9_1702_2	10_1702_0
mGluR-PB	1	2	3			4	5	6	7	8
mGluR-PD	1			2						
mGluR-PA	1	2	3			4	5	6	7	8
mGluR-PC	1	2	3		4					

Figure 42: Gene Record Finder CDS usage map of mGluR isoforms A, B, C, and D. Isoforms A and B share the same coding exons while isoforms C and D have unique coding sequences.

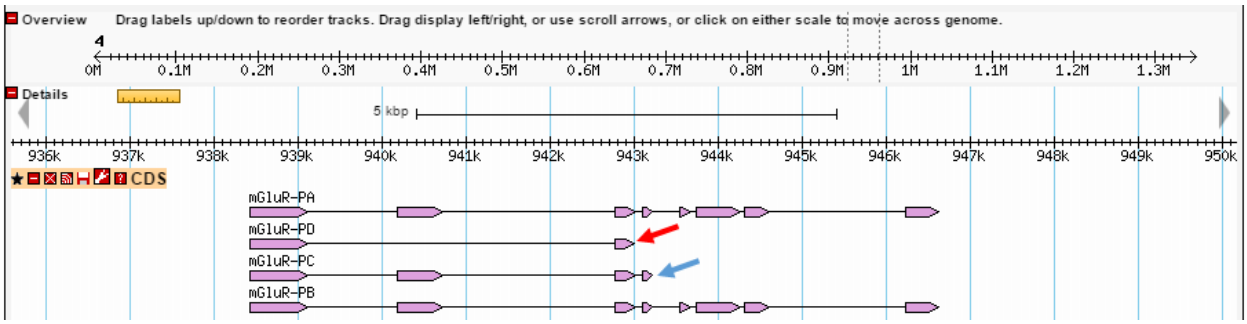


Figure 43: Gbrowse view of the mGluR gene coding exons. The D isoform shares its first exon with the other three isoforms but has an alternatively spliced second exon (red arrow) in the same region as the third exon of the other isoforms. The C isoform shares its first, second, and third exons with isoforms A and B, but has an alternatively spliced fourth exon (blue arrow) when compared to isoforms A and B.

Coding exons of *mGluR* were found by aligning coding exon amino acid sequences from *D. melanogaster* (subject) to contig43 (query) using pairwise blastx searches. Approximate locations of all coding exons were found using the blastx searches. The precise exon boundaries were determined by searching for splice donor and acceptor sites in the *D. ficusphila* UCSC Genome Browser. The majority of splice sites could be unambiguously determined by comparing phases of donor and acceptor sites, RNA-Seq expression data, and TopHat junction.

There were, however, two stop codons in *mGluR* that needed additional. For the A and B isoforms, the blastx alignment was five amino acids short of the actual stop codon due to a lack of homology, so the stop codon needed to be adjusted (Figure 44). For the D isoform, there is an

in-frame stop codon in the terminal exon (Figure 45). The conservation track in the orthologous region of the *D. melanogaster* genome browser shows that only *D. ficusphila* has this in-frame stop codon (Figure 46). Because the analysis was performed based on conservation to *D. melanogaster*, the D isoform was called to still be present in *D. ficusphila* but terminating prematurely due to lack of evidence to indicate it is no longer transcribed at all.

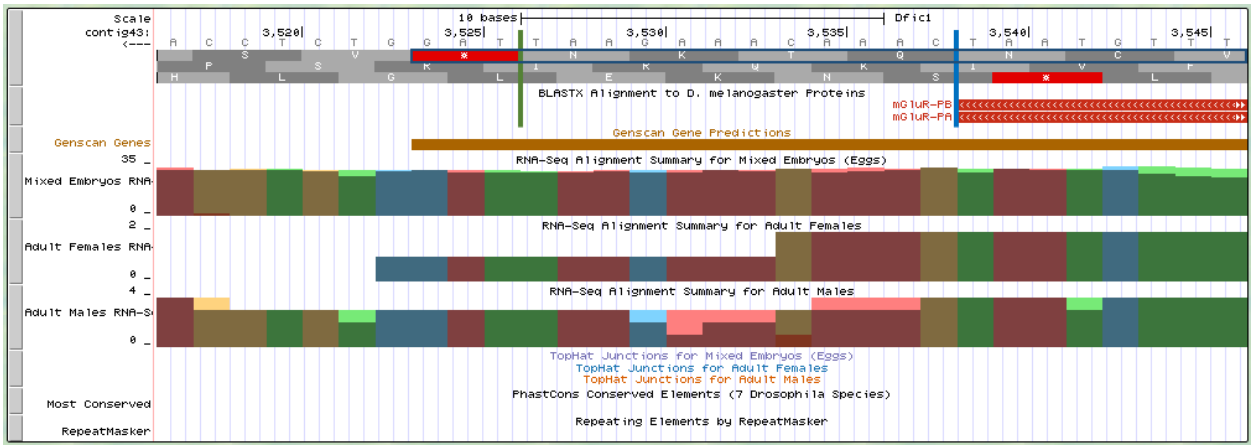


Figure 44: UCSC Genome Browser view of the terminal end of the last exon for isoforms A and B. The pairwise blastx alignment for the last exon ends at base 3539 (blue line) but the browser indicates that the final exon continues to base 3527 (green line).

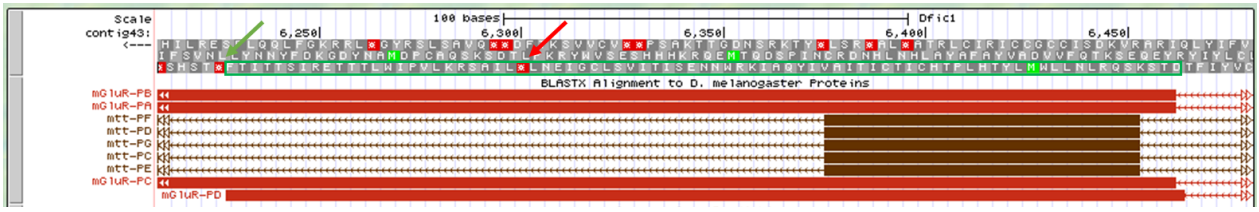


Figure 45: UCSC Genome Browser view of the terminal exon of mGluR isoform D. The red arrow indicates the in-frame stop codon and the green arrow indicates the stop codon orthologous to the one found in *D. melanogaster* (reading frame -3). Note that isoforms A, B, and C are read in frame -2.

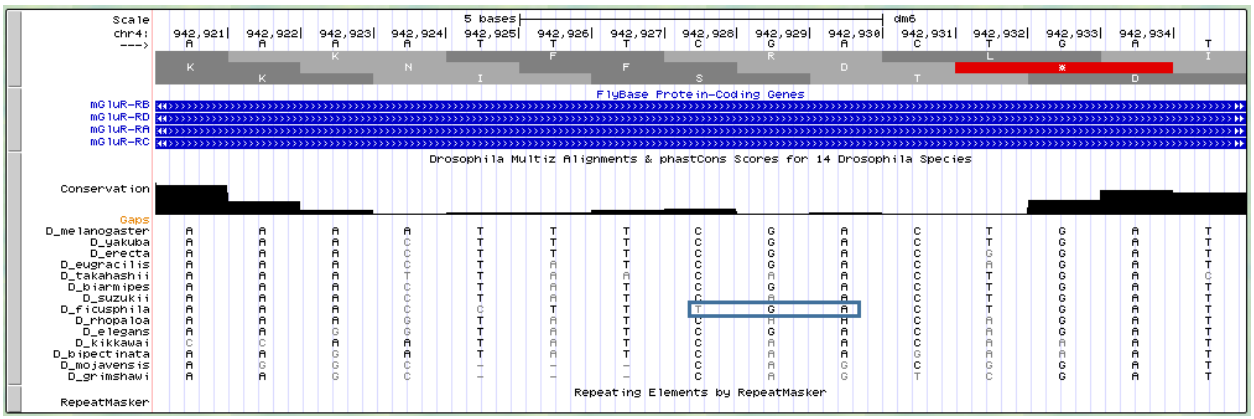


Figure 46: D. melanogaster UCSC Genome Browser view of orthologous region to in-frame stop codon of mGluR isoform D in D. ficusphila. The conservation track indicates that there is a mutation corresponding to base 942928 in *D. melanogaster* in only *D. ficusphila*, changing from a C to a T, resulting in a TGA stop codon.

The final gene models for all isoforms are shown in Tables 6-8. All gene models were put into Gene Model Checker, which confirmed viability of all isoforms. Figure 47 and Figure 48 show the dot plot and protein alignment for isoforms A and B, while Figure 49 and Figure 50 show the dot plot and protein alignment for isoform C. These graphs show no gaps and high conservation between *D. ficusphila* and *D. melanogaster*. The dot plot and protein alignment for isoform D can be seen in Figure 51 and Figure 52. These graphs both show that the *D. ficusphila* gene model is missing amino acids compared to *D. melanogaster* for the second exon due to the premature stop codon in the second exon, as well as poor conservation to *D. melanogaster* for the second exon.

Table 6: Gene model for D. ficusphila mGluR, isoforms A and B

Exon	Location	Exon size	Frame	Acceptor Phase	Donor Phase
1	10062-9381	682	-2	n/a	1
2	8081-7538	544	-2	2	2
3	6463-6303	161	-3	1	1
4	6145-6039	107	-3	2	0
5	5650-5525	126	-1	0	0
6	5464-4945	520	-1	0	1
7	4886-4591	296	-2	2	0
8	3916-3527	390	-1	0	n/a

Table 7: Gene model for D. ficusphila eIF4G, isoform C

Exon	Location	Exon size	Frame	Acceptor Phase	Donor Phase
1	10062-9381	682	-2	n/a	1
2	8081-7538	544	-2	2	2
3	6463-6303	161	-3	1	1
4	6145-6030	116	-3	2	n/a

Table 8: Gene model for D. ficusphila eIF4G, isoform D.

Exon	Location	Exon size	Frame	Acceptor Phase	Donor Phase
1	10062-9381	682	-2	n/a	1
2	6463-6303	161	-3	2	n/a

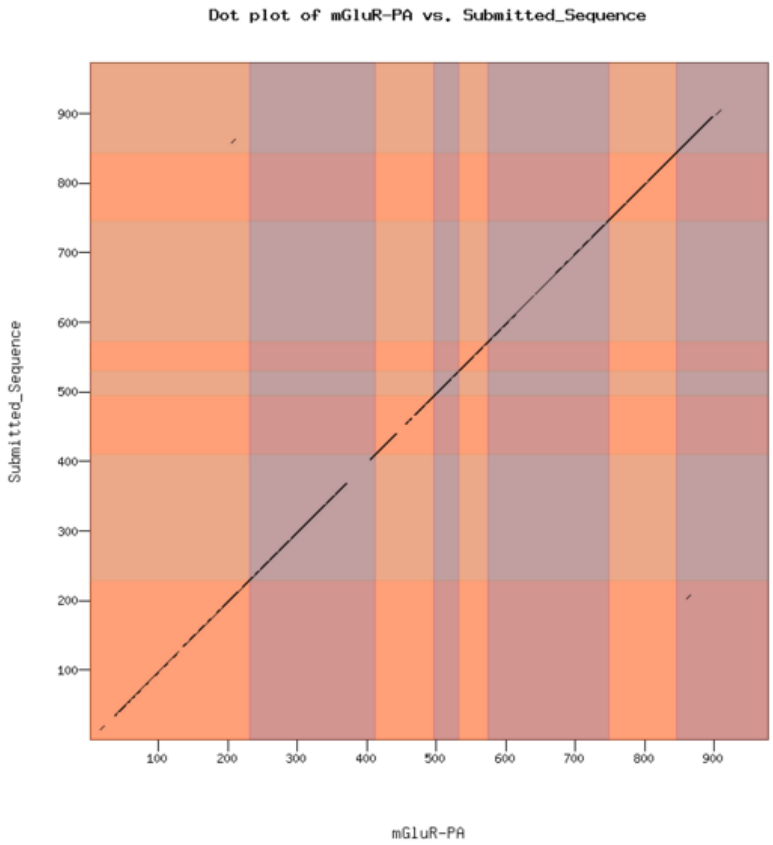


Figure 47: *Dot plot of the submitted gene model for D. ficusphila mGluR isoforms A and B (y-axis) against the D. melanogaster mGluR-PA amino acid sequence.* The dot plot shows no large gaps in the alignment and conservation across most exons, with the exception of the second half of the terminal exon.

Figure 48: Protein alignment of the submitted gene model for D. ficusphila mGluR isoforms A and B. The alignment shows conservation throughout all exons and no large gaps, confirming the plausibility of the submitted gene model.

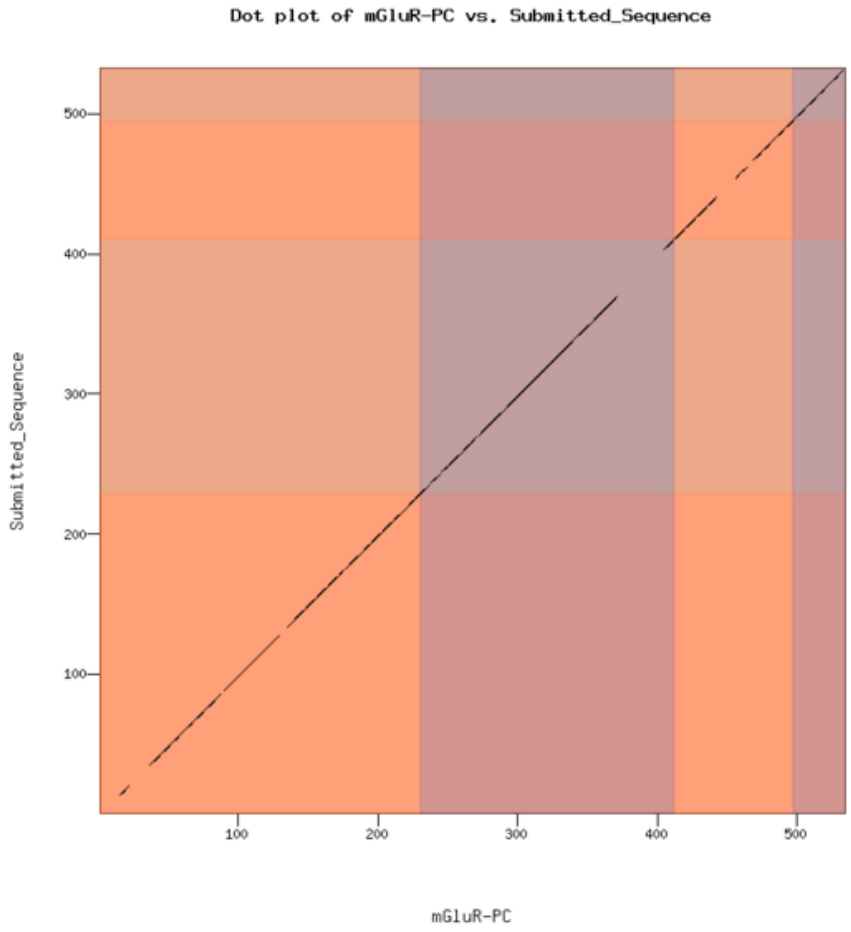


Figure 49: *Dot plot of the submitted gene model for D. ficusphila mGluR isoform C (y-axis) against the D. melanogaster mGluR-PC amino acid sequence.* The plot shows no large gaps and conservation for the majority of the proposed gene.

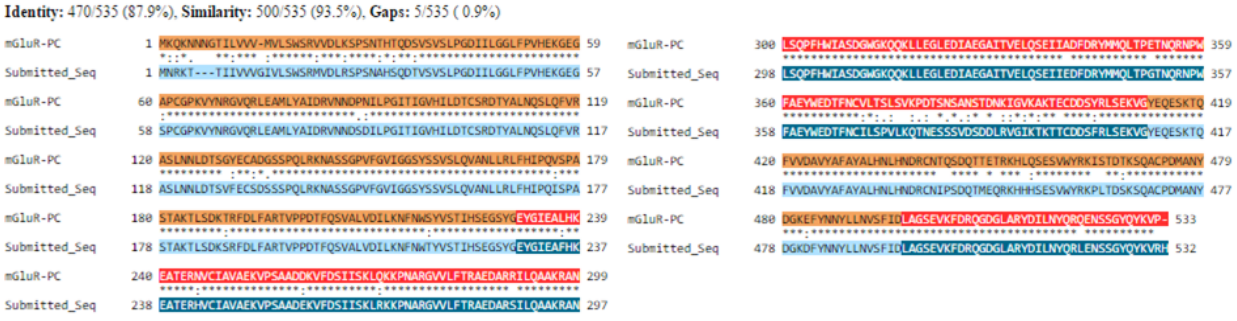


Figure 50: Protein alignment of the submitted gene model for D. ficusphila mGluR isoform C. The alignment shows conservation throughout all exons and no large gaps, confirming the plausibility of the submitted gene model.

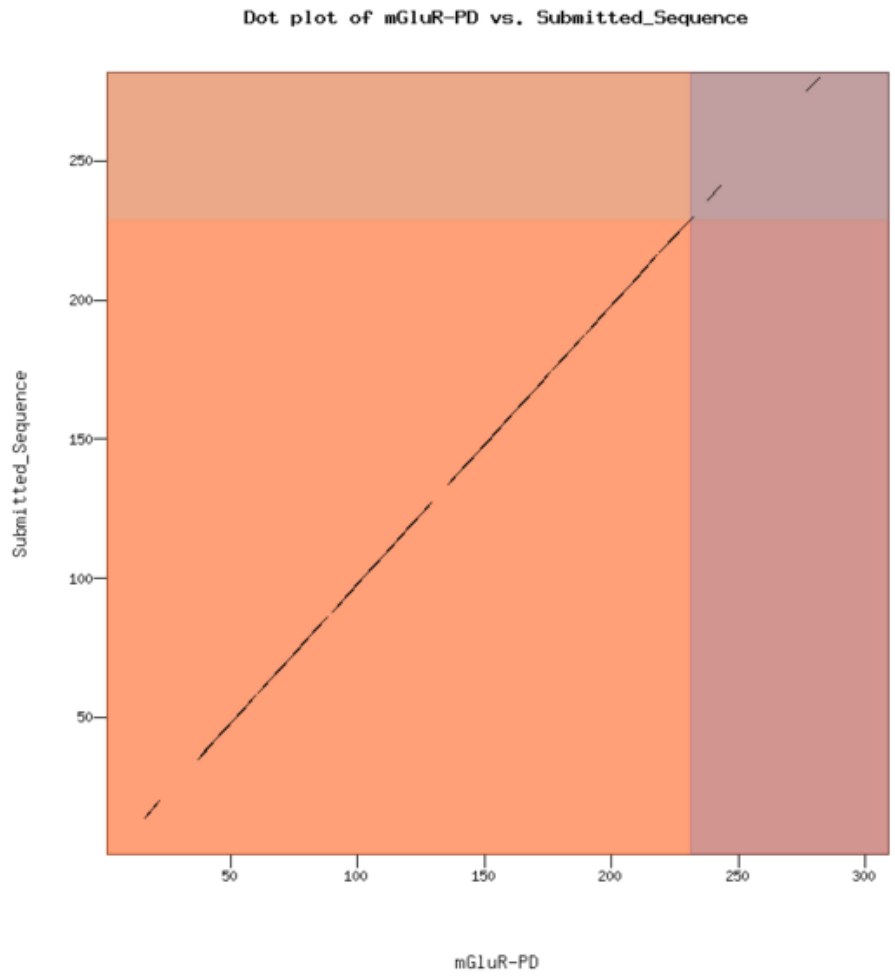


Figure 51: *Dot plot of the submitted gene model for D. ficusphila mGluR isoform D (y-axis) against the D. melanogaster mGluR-PD amino acid sequence.* The plot shows generally poor conservation within the second exon and shows the missing sequences when compared to the *D. melanogaster* second exon of the *mGluR* D isoform due to the premature stop codon in *D. ficusphila* (Figure 45).

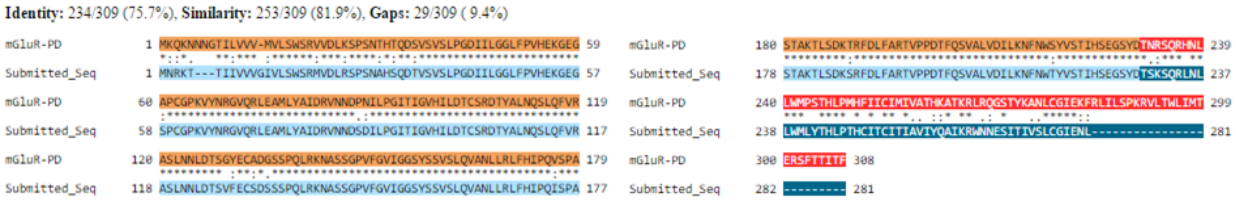


Figure 52: Protein alignment of the submitted gene model for D. ficusphila mGluR isoform D. The alignment shows the highly conserved first exon and the poorly conserved second exon with the in-frame stop codon, prematurely terminating this isoform in *D. ficusphila*.

Using the technique previously described for *4E-T* and *eIF4G* genes, the putative TSS for *mGluR* was annotated for *D. ficusphila*. All four isoforms of *mGluR* share a common TSS in *D. melanogaster* as seen in the mRNA transcript data from Gene Record Finder (Figure 53). A blastn search using the untranslated exon, which is the same in all isoforms, as the query and contig43 as the subject shows base 11224 in contig43 aligning to the TSS in *D. melanogaster* (Figure 54). Annotated Celniker TSS positions and DNase I hypersensitivity positions were visualized in the *D. melanogaster* UCSC Genome Browser (Figure 55). There is a single Celniker TSS position corresponding to base 11299 in contig43, 75 bases upstream of the TSS at base 11224 aligned by blastn. There are no DHS positions, thus this promoter can be classified as peaked. A look at the *D. ficusphila* genome browser shows the putative TSS at base 11224 to be better supported by RNA-Seq expression data, but does not rule out the possibility of the putative TSS at base 11299 (Figure 56). Table 9 shows core promoter motifs between bases 11100 and 11400 in contig43. An Inr motif was found at its expected position relative to the blastn determined TSS and a DPE motif was found 1 base away from its expected position relative to the blastn determined TSS. Thus, there is strong evidence from the core promoter motifs to annotate the TSS in *D. ficusphila* at base 11224, the TSS determined by the blastn alignment, over base 11299, which corresponds to the Celniker TSS annotation in *D. melanogaster*.

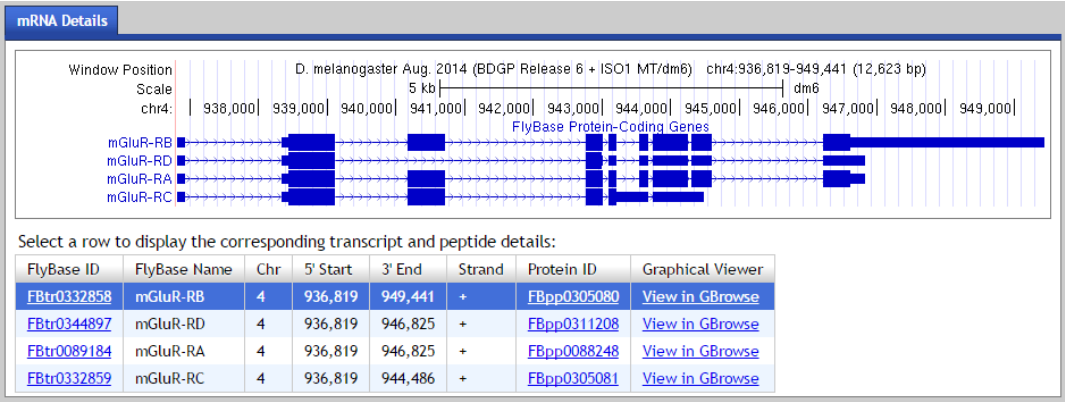


Figure 53: Gene Record Finder mRNA details of mGluR isoforms in D. melanogaster. All four isoforms share a TSS at base 936819 on the fourth chromosome.

contig43

Sequence ID: lcl|Query_70261 Length: 38500 Number of Matches: 62

Range 1: 11142 to 11224 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
44.0 bits(29)	2e-07	66/97(68%)	14/97(14%)	Plus/Minus

```
Query  1      AGTCAACCAACTGGTAATGGTAGGACAAGACGTGCGCGTATTAGTTAAATATAAAAAAGG  60
              |||||||  ||||  |||| || |||||||||| |||||||  |    |||| || ||
Sbjct  11224  AGTCAACATACTGAGAATGCTAAGACAAGACGTACGCGTATCA--CAAAT-TAGAA----  11172

Query  61     TTTTAAGAAATGTTGCGATAAAGTTGTATAAGAATTT  97
               ||| ||   |  ||||||||||||| |||||
Sbjct  11171  -------CAATTTTCGGTCAAAGTTGTATAATAATTT  11142
```

Figure 54: Blastn alignment of first noncoding exon of mGluR (query) to contig43 (subject). All 97 bases of the query sequence were aligned, and the putative TSS of *D. ficusphila* can be determined to be base 11224 in contig43.

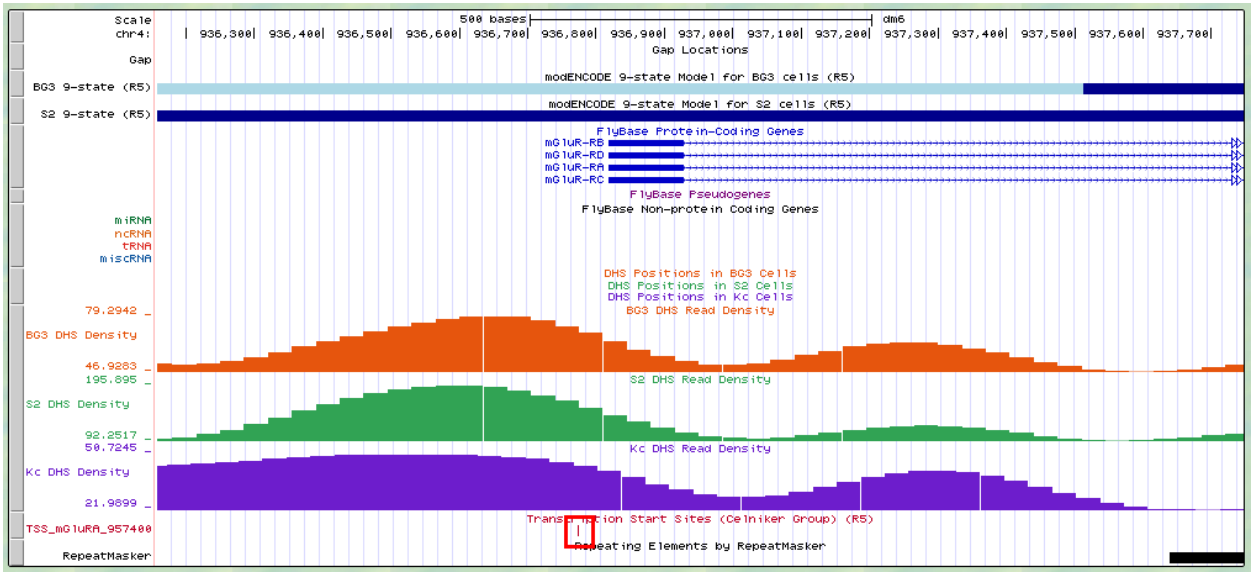


Figure 55: UCSC Genome Browser view of D. melanogaster mGluR TSS region. There is a single Celniker annotated TSS position (red box) and no DHS positions. Note that this TSS region is classified as heterochromatin (blue) or heterochromatin-like (light blue) in the two modENCODE tracks on the top of the browser, indicating that this gene is unlikely to be highly transcribed in the cell lines the data was obtained from.

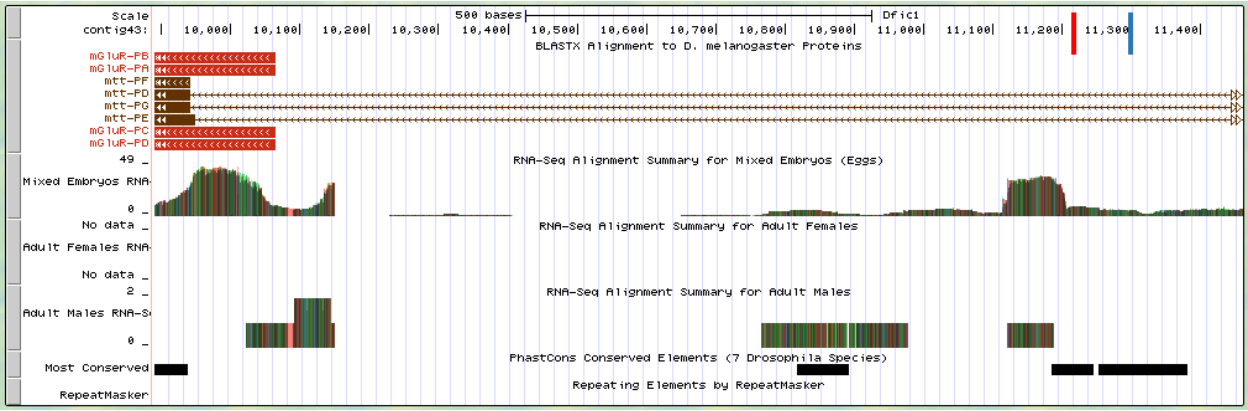


Figure 56: UCSC Genome Browser view of D. ficusphila mGluR TSS region. The putative TSS based on blastn is shown by the red line (11224) and the putative TSS based on the Celniker annotated TSS is shown by the blue line (11299). The RNA-Seq expression evidence and occurrence of transcription factor motifs better supports the putative TSS from the blastn alignment; the site is close to a small peak of RNA-Seq expression data.

Table 9: Motifs found near mGluR putative TSS (contig43 bases 11100-11400). Two motifs highlighted in green, an Inr found at its expected position from the TSS and a DPE motif found one base away from its expected position from the TSS, both support the annotation of 11224 as the TSS in *D. ficusphila*.

Motif	Position relative to TSS	Expected position for TSS (11224, 11299)	Found Motifs
Bre u	-38	11262, 11337	
TATA	-31 or -30	11254 or 11255, 11329 or 11330	
Bre d	-23	11247, 11322	11280-11286, 11283-11289. 11329-11335, 11351-11357
Inr	-2	11226, 11301	11221-11226
MTE	18	11206, 11281	
DPE	28	11196, 11271	11152-11157, 11192-11197, 11228-11233
Ohler 1	N/A	N/A	
DRE	N/A	N/A	
Ohler 5	N/A	N/A	
Ohler 6	N/A	N/A	
Ohler 7	N/A	N/A	
Ohler 8	N/A	N/A	

Clustal Omega

A Clustal Omega analysis was performed on *mGluR* isoform A for species *D. ficusphila, D. melanogaster, D. simulans, D. yakuba, D. erecta, D. virilis,* and *D. grimshawi* (Figure 57). In terms of evolutionary distance from *D. melanogaster*, *D. ficusphila* is more distant than *D. simulans, D. yakuba,* and *D. erecta*, but closer than *D. virilis* and *D. grimshawi*. The Clustal analysis shows the evolution of the gene across multiple *Drosophila* species and shows that the *mGluR* gene is generally highly conserved between the species analyzed, except for the beginning and end of the peptide and a small region in the middle of the peptide.

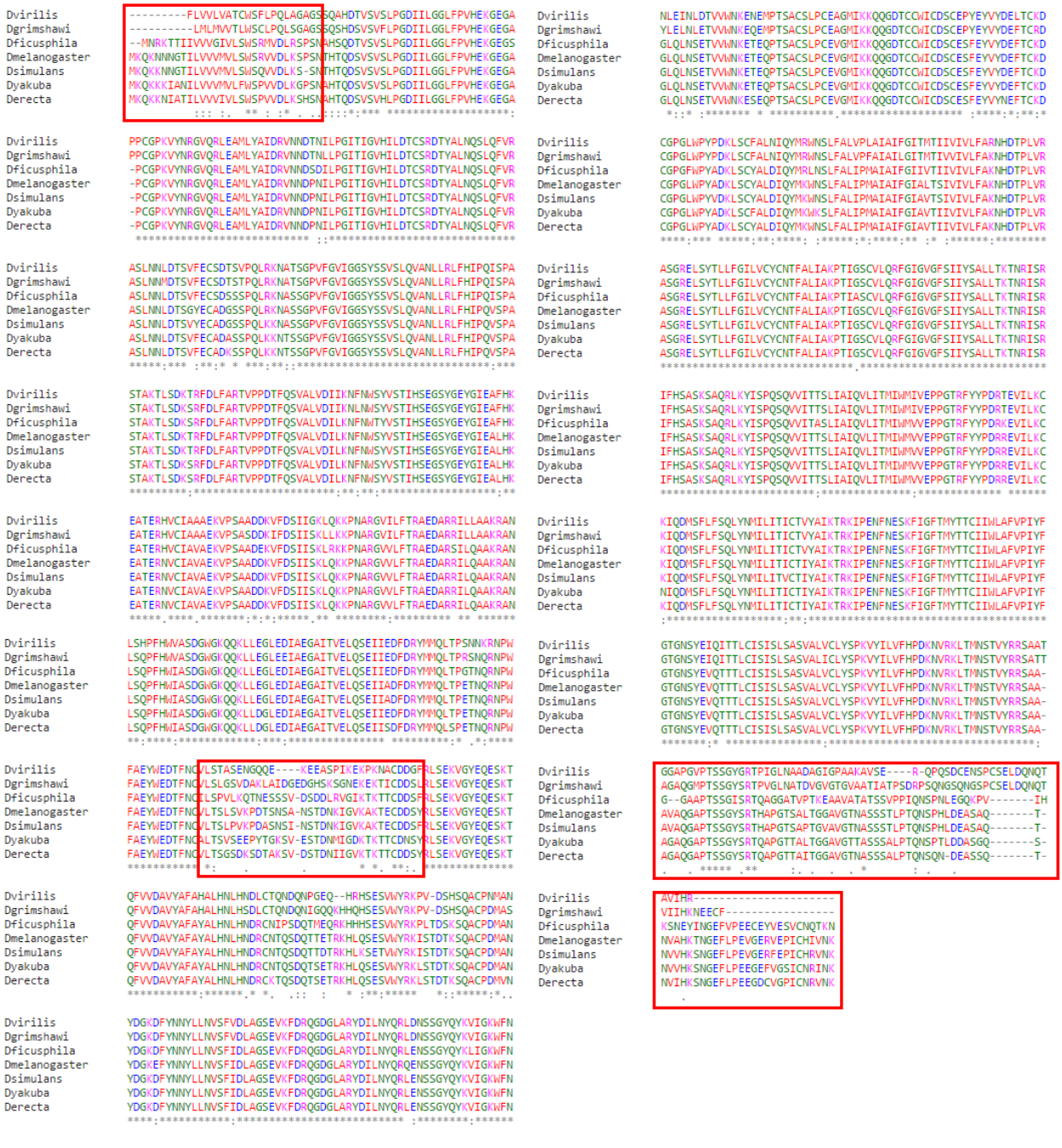


Figure 57: Clustal Omega Alignment of mGluR isoform A across seven species of Drosophila. The gene is generally highly conserved but there are some regions of lower conservation boxed in red, primarily located at the beginning and end of the peptide sequence but also including a region in the middle of the protein.

Repeats

The UCSC Table Browser was used to annotate repetitious sequences in contig43 and the orthologous region in *D. melanogaster* using RepeatMasker (Table 10). Due to a rearrangement of the region in *D. ficusphila* relative to *D. melanogaster*, the orthologous *D. melanogaster* region was defined as the region containing the *4E-T*, *eIF4G*, and *mGluR* genes. Both regions show a similar percentage of repeats, with 27.02% repeats in *D. ficusphila* contig43 and 27.08% repeats in the orthologous region on chromosome 4 in *D. melanogaster*. Contig43 had 4 repeats longer than 500bp, which are shown in Table 11. Of these four repeats longer than 500bp, two are DNA TcMar-Mariner DNA transposons, one is a DNA transposon, and one is unknown. Contig43 was also searched for fragments of *Wolbachia* by performing a blastn search, aligning CP001391.1 (*Wolbachia* endosymbiont of the Riverside strain of *D. simulans*) (query) to contig43 (subject) at an expect threshold of 1e-2. No results were found at that expect threshold. Increasing the expect threshold resulted in some short matches (<50bp) with e-values between 1e-2 and 10 (Figure 58), but these relatively weak alignments are unlikely to indicate that fragments of *Wolbachia* have been integrated into the genome.

Table 10: D. ficusphila contig43 and orthologous region of D melanogaster (chr4:915,000-960,000) RepeatMasker summary statistics.

	D. ficusphila	*D. melanogaster*
item count	52	49
item bases	10,402 (27.02%)	12,188 (27.08%)
item total	10,629 (27.61%)	12,202 (27.11%)
smallest item	29	19
average item	204	249
biggest item	1,007	1,102
smallest score	225	12
average score	1,332	1,389
biggest score	8,663	8,991

Table 11: Large (>500bp) repeats in contig43.

swScore	prec. div	perc. del	perc. ins	Query sequence	pos. begin	pos. end	size	repName	repClass	Rep Start	Rep End	Rep Left	id
5778	3.1	0.1	0	contig43	21178	21845	667	rnd-1_family-23	DNA/TcMar-Mariner	-69	1221	553	28
5975	6.2	0	2	contig43	11641	12425	784	rnd-1_family-12	Unknown	4	772	0	16
7330	3.4	0.3	0.1	contig43	29027	29932	905	rnd-1_family-46	DNA	-26	908	1	31
8663	2.6	1.6	0.2	contig43	19	1025	1006	rnd-1_family-23	DNA/TcMar-Mariner	270	1290	0	1

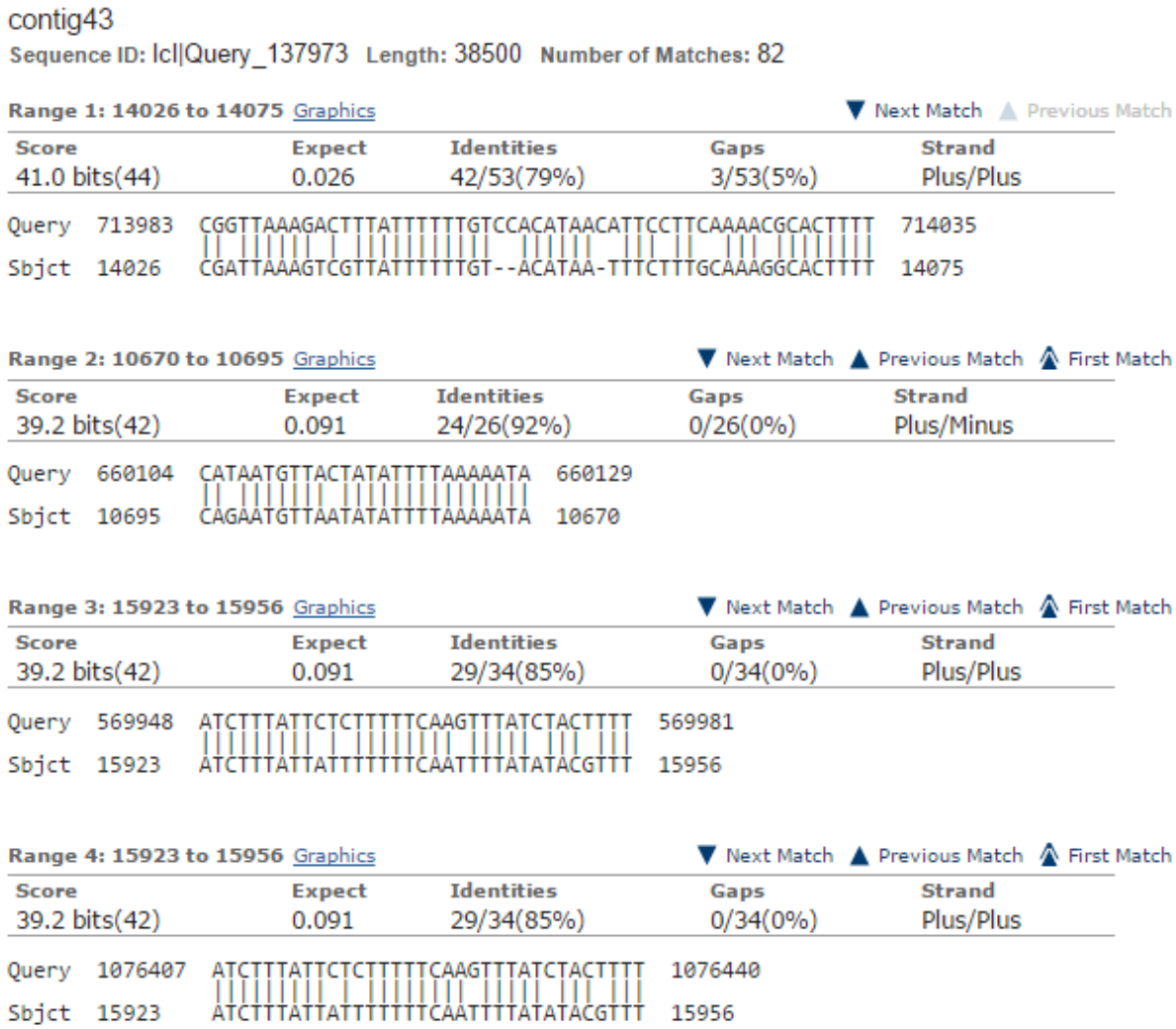


Figure 58: Blastn aligning CP001391.1 (query) to contig43 (subject). The search was performed with expect threshold of 10. Only the top four results are shown. These short matches with E-values greater than 1e-2 are unlikely to represent genomic fragments of *Wolbachia* integrated into the *D. ficusphila* genome.

Synteny

The previously annotated orthologous region on the fourth chromosome of *D. melanogaster* was compared to the annotated contig43 to analyze synteny between the annotated genes in contig43 and the corresponding genes in *D. melanogaster* (Figure 59). It is apparent that synteny has not been preserved as the genes are in different positions and orientations when comparing *D. melanogaster* to *D. ficusphila*. However, here is no way to determine with certainty which gene moved based on this data. Zooming out to view a larger portion of the *D. melanogaster* genome browser (Figure 60) shows that the gene order and orientation with increasing base position in *D. melanogaster* is unc-13 (minus), *eIF4G* (minus), *mGluR* (plus), *4E-T* (minus), *fuss* (minus). The adjacent regions to contig43 can be viewed by analyzing contig42 and contig44 (Figure 61). The gene order and orientation with increasing base position in *D. ficusphila* is unc-13 (minus), *mGluR* (minus), *eIF4G* (plus), *4E-T* (minus), *fuss* (minus). Based on this information, the most parsimonious explanation for the loss of synteny is a single inversion containing *eIF4G* and *mGluR* in the clade containing *D. ficusphila* but not *D. melanogaster*, which would explain the change in relative gene order and orientation with a single event.

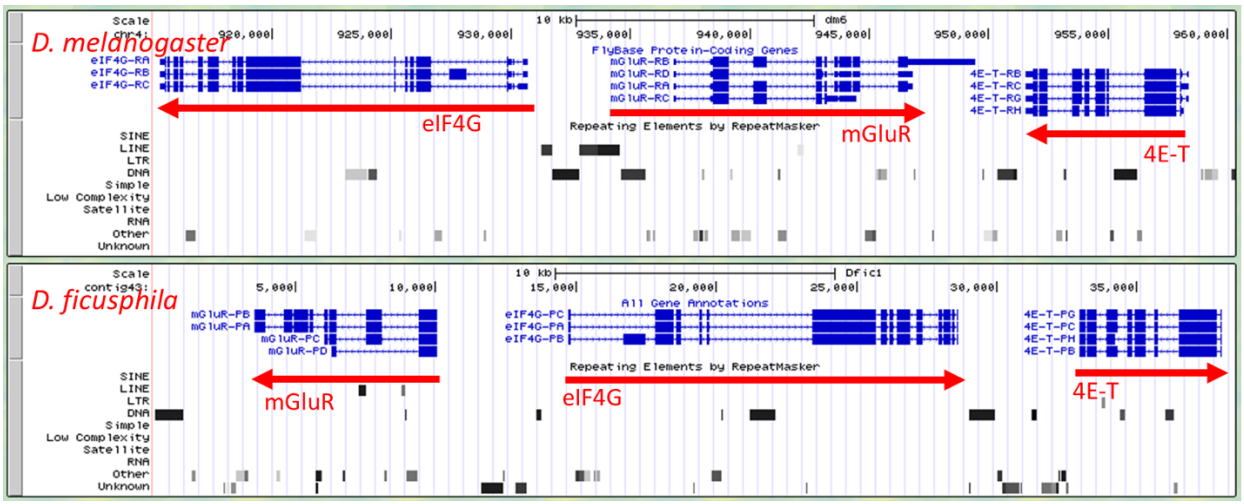


Figure 59: Comparison of genome browser of contig43 to orthologous region of D. melanogaster. Red arrows labeled with the corresponding gene names have been added to more easily visualize the orientation of each gene. Synteny has clearly not been preserved as both the relative gene order and orientation have changed going from *D. melanogaster* to *D. ficusphila*.

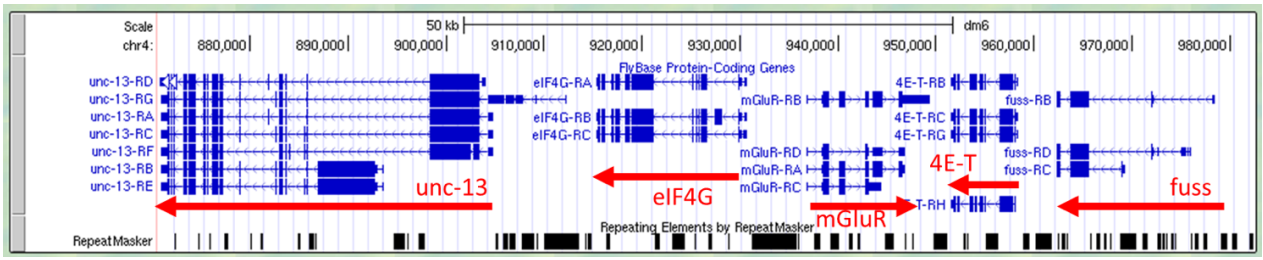


Figure 60: D. melanogaster Genome browser view of regions surrounding the region orthologous to contig43. Red arrows labeled with the corresponding gene names have been added to more easily visualize the orientation of each gene.

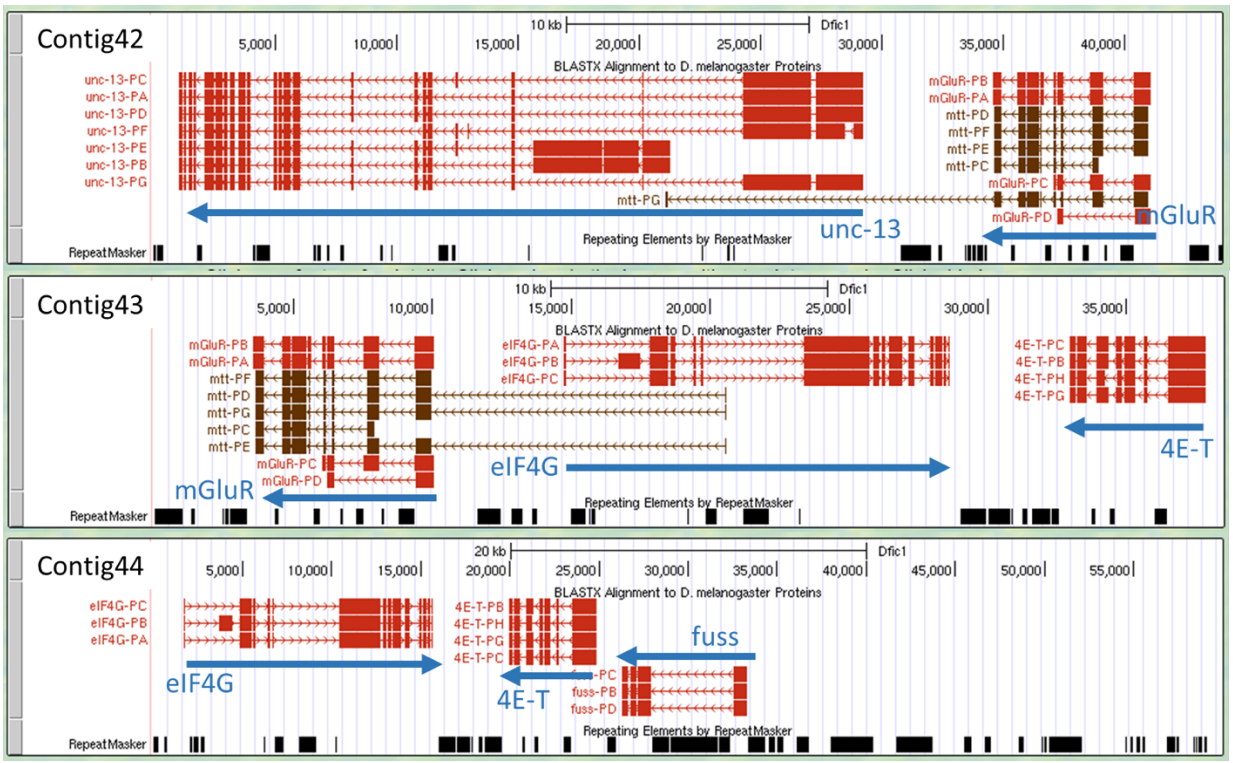


Figure 61: Genome browser views for contig42, contig43, and contig44. Blue arrows labeled with the corresponding gene names have been added to more easily visualize orientation of each gene. Given the overlap between contigs, it is possible to determine relative gene order and orientation of genes surrounding contig43.

Discussion

This annotation project annotated contig43 for coding exons and transcription start sites for all genes present. There are three genes on contig43 -- *mGluR*, *eIF4G*, and *4E-T*. All isoforms present in *D. melanogaster* also appear to be present in contig43, with the possible exception of *mGluR*-PD, given the in-frame stop codon resulting from a single base change in the terminal exon. TSS for all three genes were annotated based on conserved untranslated exons, DNase I hypersensitivity sites, Celniker TSS annotations, RNA-Seq expression data, and presence of core promoter motifs. Contig43 contains 27% repetitious sequences and this region also shows an inversion containing two genes relative to *D. melanogaster*. These data are consistent with previous findings by the GEP that found F elements had greater transposon density (25%-50%) when compared to euchromatic reference regions and greater rates of inversion (Leung et al.,

2015). In conclusion, this annotation project completed the annotation of contig43 and will contribute to the eventual annotation of *D. ficusphila*. A final map of the contig can be seen in Figure 62.

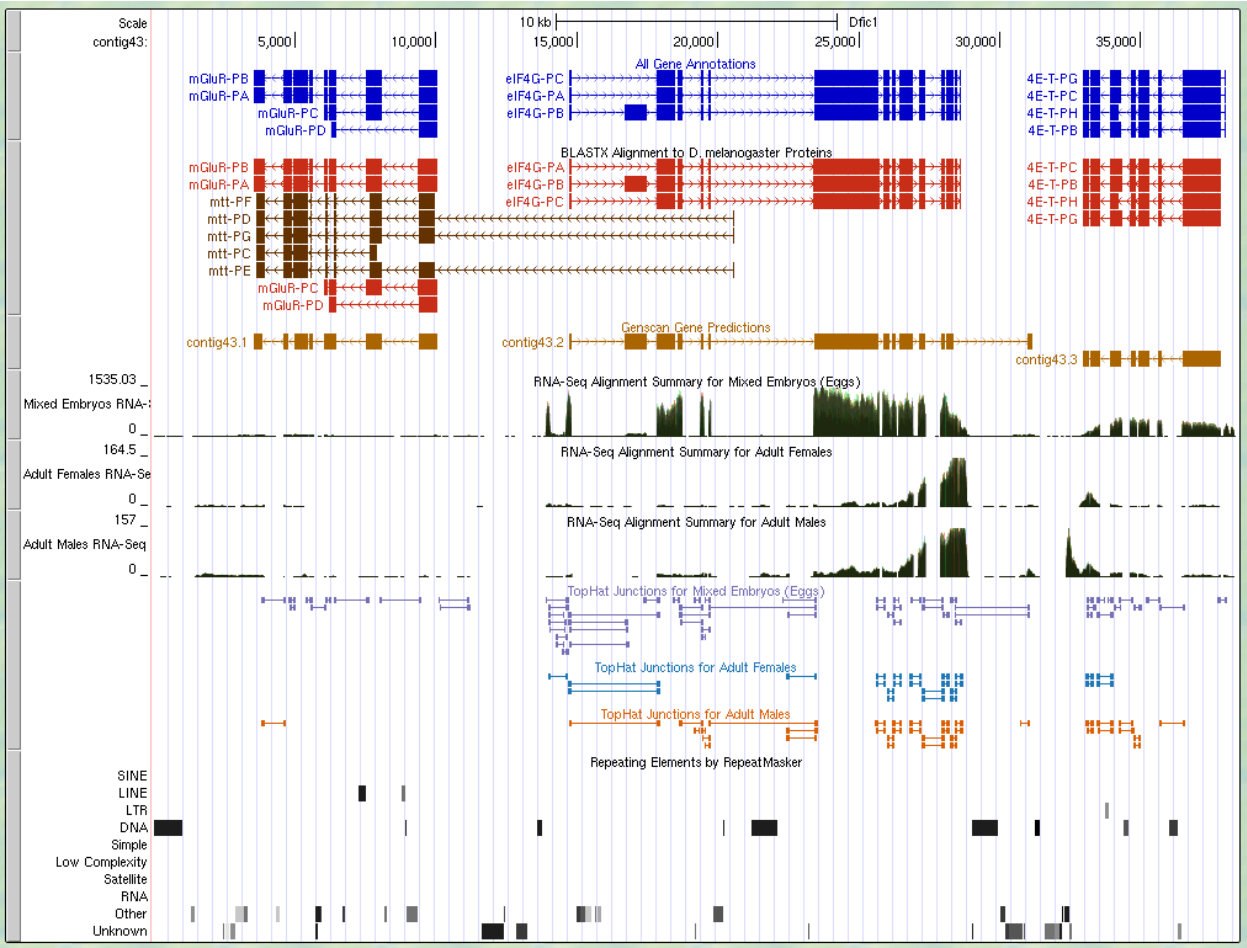


Figure 62: Final map of contig 43. The top tracks in blue are the annotated coding exons for the genes present in contig43.

Acknowledgements

I would like to thank Dr. Elgin, Dr. Shaffer, Dr. Bednarski, Wilson Leung, Yu He, and Daniel Zhou for their help with this project and throughout the entire semester of Bio4342.

References

Leung, W. et al (2015). Drosophila Muller F Elements Maintain a Distinct Set of Genomic Properties Over 40 Million years of Evolution. G3: Genes Genomes Genetics 5(5): 719-740.