Chiraag Kapadia

BIOL 434W

Dr. Elgin

16 May 2017

Annotation of *Drosophila eugracilis* Contig22

Abstract

Contig22 is a 45,000 bp region on the *Drosophila eugracilis* F element. Genes identified within contig22 were annotated using the Genomics Education Partnership UCSC Genome Browser mirror, conservation to *D. melanogaster*, *ab initio* gene predictors, and RNA-seq data. The orthologous region in the *Drosophila melanogaster* reference genome contained three genes, *4E-T*, *mGluR*, and *fuss*. All three genes, and their respective isoforms, were identified in *D. eugracilis*. The orthologous region in *D. melanogaster* contained three non-coding RNAs (ncRNAs), *CR45201*, *CR44030*, and *sphinx*. Orthologous ncRNAs could not be identified in contig22 by sequence similarity. The predicted transcription start sites for *D. eugracilis 4E-T, mGluR,* and *fuss* are within peaked promoters. Gene evolution analysis of the 4E-T-PB coding region revealed high conservation among insect species and putative functional domains. Despite this conservation, no meaningful *4E-T* three-dimensional structure could be predicted. *RepeatMasker* classified 27.23% of contig22 as repetitious according to a species-specific repeat library. Of these repeats, eight were greater than 500 nucleotides. The arrangement of genes on contig22 demonstrates complete synteny with *D. melanogaster*. All coding regions demonstrate high sequence conservation to their respective *D. melanogaster* orthologs. Annotation insights from contig22 will be combined with those from additional regions along the chromosome and

will guide future comparative genomics and phylogenetic footprinting approaches to better understand *Drosophila* F element gene regulation.

Introduction

Eukaryotic DNA is packaged into chromatin, initially defined as euchromatin and heterochromatin. Euchromatic DNA is loosely packaged, replicated throughout in S phase, contains a low prevalence of repetitious sequences, and is associated with domains populated with active genes. In contrast, heterochromatic DNA is densely packaged, replicated late in S phase, repeat-rich, and associated with transcriptionally silent domains that are generally gene-poor. The *Drosophila melanogaster* fourth chromosome is referred to as the Müller F element, or dot chromosome. The *D. melanogaster* 1.3 Mb distal arm contains a normal gene density (about 80 genes), but contains a three-fold increase in the prevalence of repetitious sequences compared to euchromatic chromosome arms. Repetitious sequences are frequently targets for heterochromatin formation. As expected, the *D. melanogaster* F element is nearly entirely heterochromatic and displays H3K9me2/3 and HP1 silencing marks. Despite these silencing signals, genes on the F element are actively transcribed. This anomaly makes the F element a candidate of interest for comparative analysis across *Drosophila* species. Comparison of conserved sequence features across *Drosophila* species should provide insights into mechanisms of gene regulation for transcriptionally active heterochromatic genes.

The *Drosophila eugracilis* F element was recently sequenced and remains to be annotated to guide subsequent comparative genomics analysis. The *D. melanogaster* genome is well-characterized and was used for reference during annotation. The *D. eugracilis* genome is evolutionarily separated from *D. melanogaster* by 10-15 million years and is thus an ideal

evolutionary distance for the identification of conserved regulatory motifs. Annotation of putative coding exons in *D. melanogaster* provides the foundation for predicting accompanying gene transcription start sites (TSSs). Using a comparative genomics approach, TSSs conserved across several *Drosophila* species can be identified and used in phylogenetic footprinting analysis with the aim of identifying motifs that help determine how F element genes are expressed within their heterochromatic context.

Contig22 is a 45,000 bp region on the *D. eugracilis* F element. To yield a complete annotation of the *D. eugracilis* F element, annotation insights from contig22 will be combined with those from additional regions along the chromosome. *Ab initio* gene predictors identified three features of interest within contig22 (Figure 1). Feature 2 was associated with the most RNA-seq data and showed general agreement between BLAST alignments and *ab initio* gene predictors. Thus, Feature 2 was selected for initial annotation.

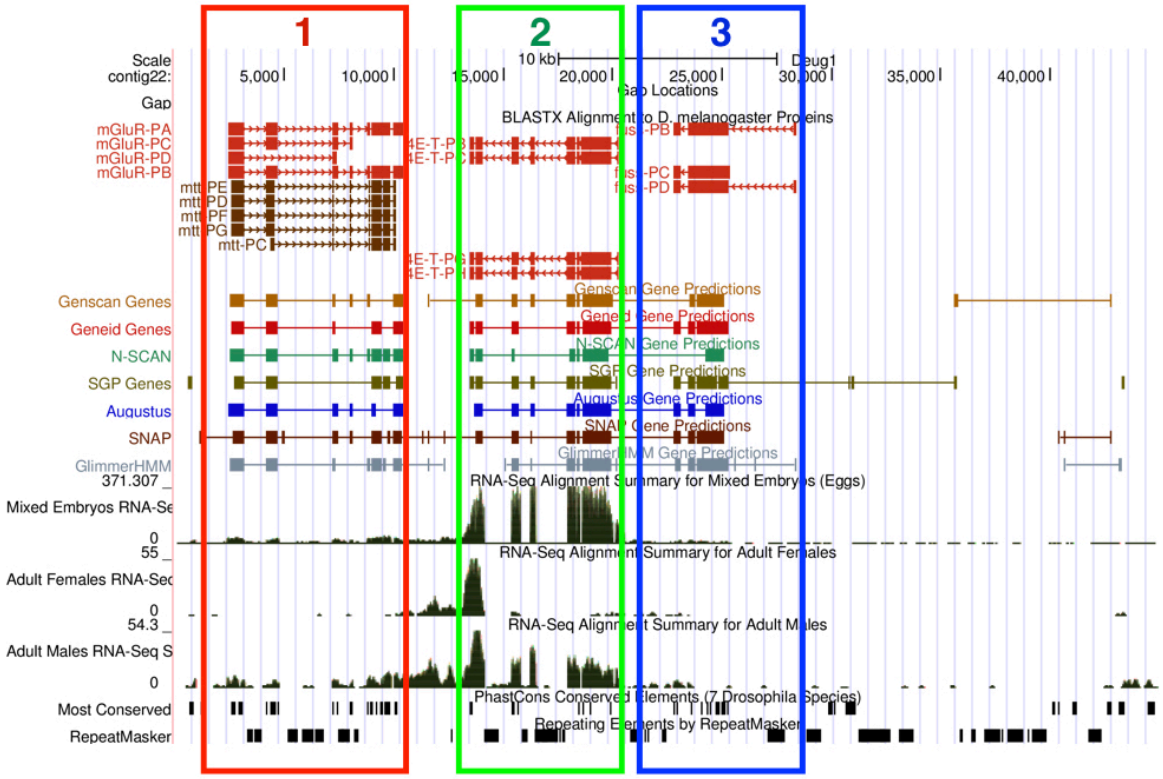


Figure 1: Overview of *D. eugracilis* contig22 as displayed on the GEP UCSC Genome Browser mirror. The *ab initio* gene predictors and initial blastx alignment identified three regions that appear to contain distinct genes. Features 1, 2, and 3 are displayed in red, green, and blue boxes, respectively. Note that Feature 2 is associated with greater RNA-seq reads than the flanking features.

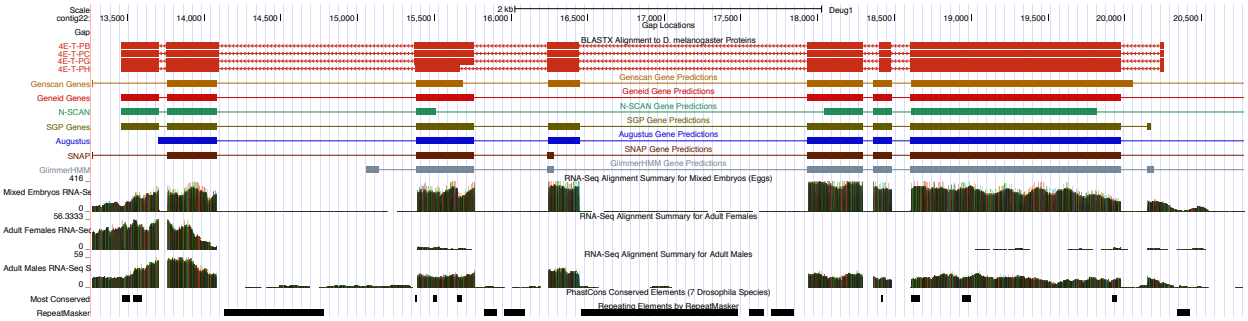


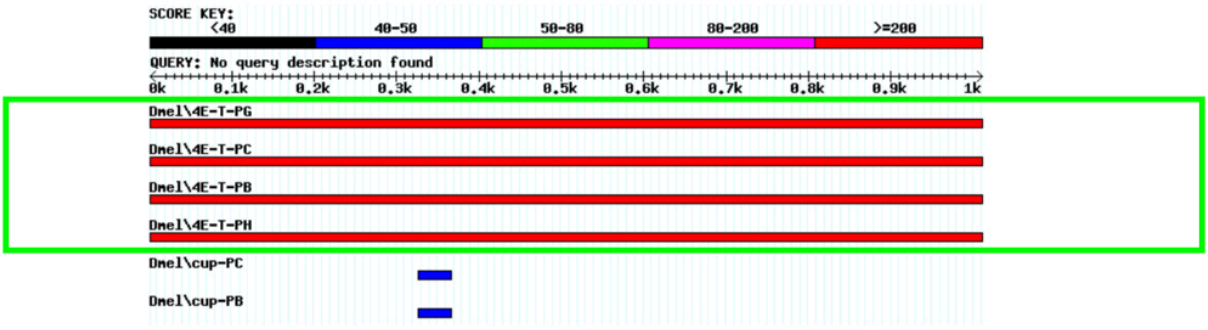
Figure 2: Feature 2 on *D. eugracilis* contig22. Initial blastx alignment suggests four isoforms with eight exons each. Gene predictors agree roughly on the placement of coding regions and RNA-seq data overlaps roughly with predicted exons. Repetitious elements are identified in putative intronic regions adjacent to transcribed exons, recapitulating a unique feature of F element genes.

Feature 2 Annotation

A detailed view of Feature 2 is shown in Figure 2. The SGP gene predictor suggests eight coding exons that overlap roughly with the blastx alignment to *D. melanogaster* proteins. Other *ab initio* gene predictors agree in coding exon placement, but suggest this coding region extends into that of Feature 3. RNA-seq data demonstrates that Feature 2 is highly transcribed, especially in the embryonic stage. *RepeatMasker* identifies repetitive elements in putative intronic sequences within Feature 2. Thus, despite the presence of a high repeat density, there is transcriptional output at Feature 2, a representative of F element gene behavior.

Feature 2 contains the putative *D. eugracilis* ortholog of *4E-T*

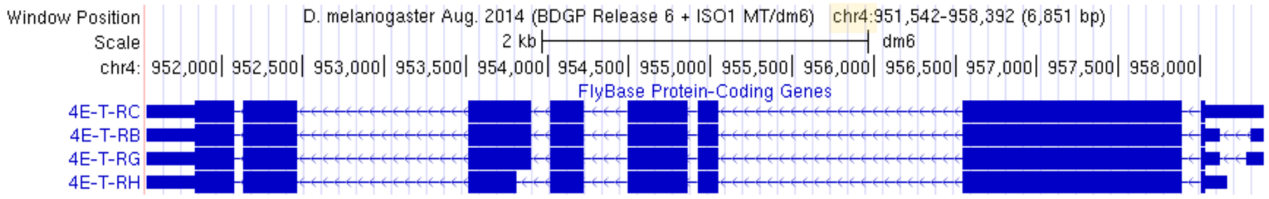
To identify the *D. melanogaster* ortholog for Feature 2, the protein sequence predicted by the SGP gene predictor was used as query in a blastp search against the FlyBase *D. melanogaster* annotated proteins (AA) database (subject). The B, C, G, and H isoforms of *4E-T* shared the lowest (most significant) E-value. The next greatest E-values stemmed from matches to the B and C isoform of *cup*, but were several orders of magnitude greater than *4E-T* matches. Thus, the gene at Feature 2 is likely the *4E-T* ortholog (Figure 3). Preliminary blastx alignment identified Feature 2 with four *4E-T* isoforms as well, supporting identification of this ortholog.



BLAST Hit Summary			
Description	**Species**	**Score**	**E value**
4E-T-PG	Dmel	2100.48	0
4E-T-PC	Dmel	2100.48	0
4E-T-PB	Dmel	2100.48	0
4E-T-PH	Dmel	1921.36	0
cup-PC	Dmel	45.0542	0.000399182
cup-PB	Dmel	45.0542	0.000399182

Figure 3: blastp search using SGP predicted protein sequence in Feature 2 (query) against the *D. melanogaster* annotated proteins (AA) database (subject). Four isoforms of the same protein, *4E-T*, match with lower E-value than the next best match. Thus, Feature 2 likely contains an ortholog of *4E-T*.

A



B CDS usage map:

Isoform	1_10825_0	2_10825_1	3_10825_0	4_10825_2	5_10825_2	7_10825_2	6_10825_2	8_10825_2	9_10825_2
4E-T-PC	1	2	3	4	5	6		7	8
4E-T-PB	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

Isoforms with unique coding exons:

Unique isoform(s) based on coding sequence	Other isoforms with identical coding sequences
4E-T-PC	4E-T-PB, 4E-T-PG
4E-T-PH	

Figure 4: Gene structure for *D. melanogaster 4E-T* provided by FlyBase. A) *4E-T* in *D. melanogaster* is located on the fourth chromosome (F element) (highlighted). There are four isoforms with two distinct coding regions. **B)** The B, C, and G isoforms of *4E-T* produce identical coding sequences. 4E-T-H utilizes a distinct sixth exon.

FlyBase Release 6.13 was used to analyze the *4E-T* gene structure. *D. melanogaster 4E-T* is located on the F element (chromosome 4) (Figure 4A), providing further evidence that Feature 2 contains the ortholog of *4E-T*. The *D. melanogaster* ortholog contains four isoforms that translate to two distinct protein products (Figure 4B). Isoforms B, C, and G yield identical protein coding regions. 4E-T-PH contains a shorter sixth exon and is otherwise identical in coding sequence to the other isoforms (Figure 4A). Alternative splicing occurs primarily at the 5'-UTR.

The isoform 4E-T-PB was selected for initial annotation. The coding DNA sequence (CDS) table for *D. melanogaster* 4E-T-PB is shown in Figure 5. To identify the approximate CDS locations on contig22, pairwise blastx alignment was performed using the amino acid sequence of each *D. melanogaster 4E-T* exon as the subject and the contig22 DNA sequence as the query. The exons were mapped in order of longest exon size to shortest. Exon 1 was unable to be identified using this approach due to its short (25 nucleotide) sequence not producing any significant alignments. Approximate locations of each exon based on the conservation found by blastx are described in Table 1. Blastx search results for each exon are shown in Figure 6.

FlyBase ID	5' Start	3' End	Strand	Phase	Size (aa)
1_10825_0	958,031	958,006	-	0	8
2_10825_1	957,889	956,545	-	1	448
3_10825_0	955,050	954,921	-	0	43
4_10825_2	954,863	954,489	-	2	124
5_10825_2	954,226	954,020	-	2	68
7_10825_2	953,899	953,513	-	2	128
8_10825_2	952,461	952,138	-	2	107
9_10825_2	952,077	951,839	-	2	79

Figure 5: *D. melanogaster* 4E-T-PB exon usage as annotated on FlyBase. Notice all exons lie on the minus strand. Each orthologous exon will be subject to blastx pairwise alignment against contig22 to map putative coding exon boundaries.

		D. eugracilis ortholog					
FlyBase_ID	Isoform Usage	contig22 start	contig22 end	Frame	*D. mel*. exon size	D. eug. exon start	D. eug. exon end
4E-T:1_10825_0	B, C, G, H	-	-	-	-	-	-
4E-T:2_10825_1	B, C, G, H	18607	19974	-1	448	1	448
4E-T:3_10825_0	B, C, G, H	18364	18477	-1	43	1	42
4E-T:4_10825_2	B, C, G, H	17935	18291	-1	124	1	124
4E-T:5_10825_2	B, C, G, H	16243	16443	-1	68	1	68
4E-T:7_10825_2	B, C, G	15385	15747	-1	128	4	128
4E-T:6_10825_2	H	15385	15663	-1	97	1	97
4E-T:8_10825_2	B, C, G, H	13761	14081	-2	107	1	107
4E-T:9_10825_2	B, C, G, H	13459	13698	-1	79	1	79

Table 1: blastx alignment maps *D. melanogaster 4E-T* coding exons (subject) to contig22 (query). Exon 1 was too small to produce a significant blastx alignment. Exons 4E-T:3_10825_0 and 4E-T:7_10825_2 showed incomplete alignment at exon ends. All *D. melanogaster* exons occupy the minus strand of chromosome four.

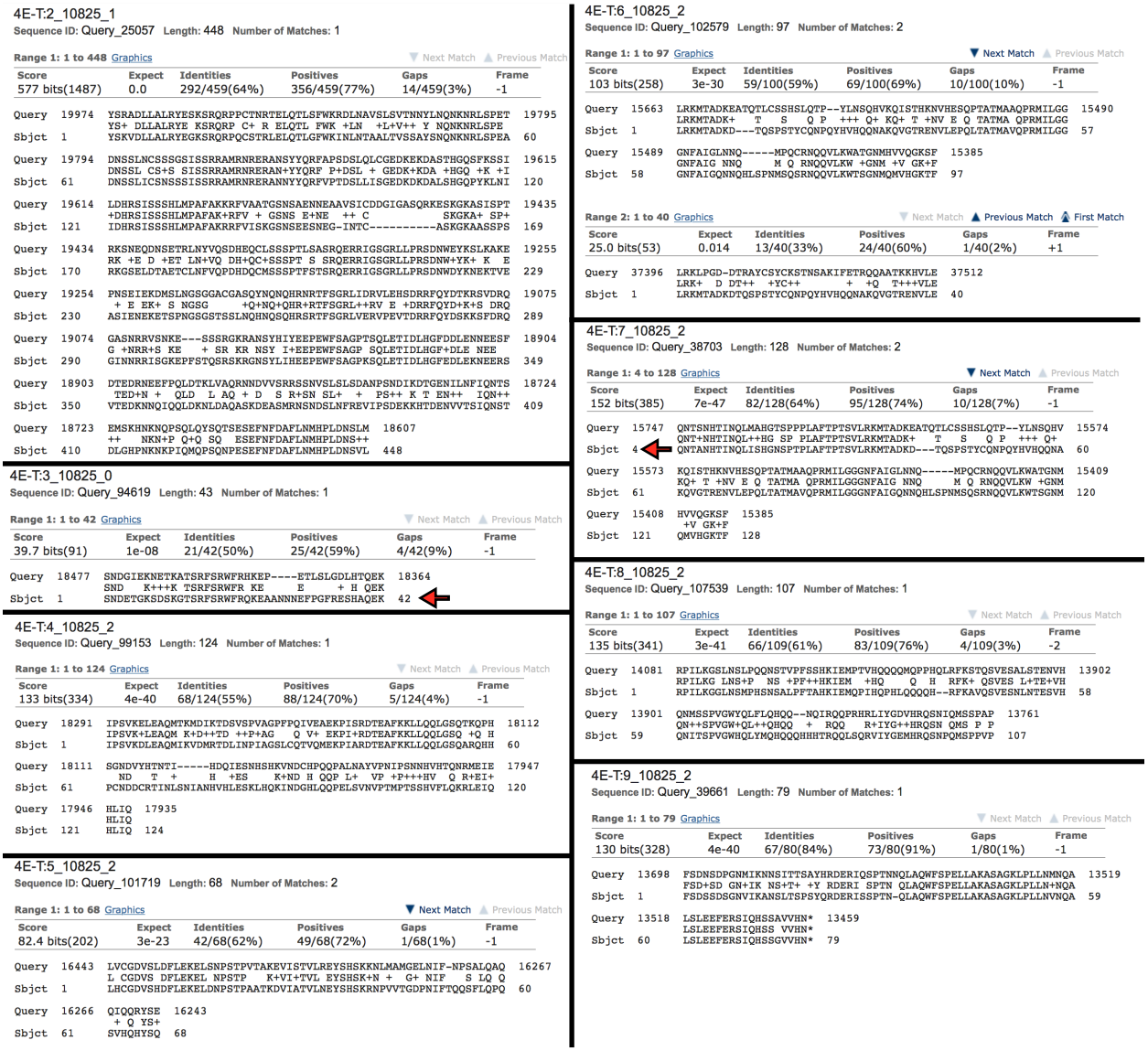


Figure 6: Pairwise blastx alignment for each *D. melanogaster 4E-T* coding exon (subject) against the contig22 nucleotide sequence (query). Each *D. melanogaster* exon is identified by FlyBase ID. Red arrows indicate missing amino acids at exon ends. Exon alignments are separated by a black line and each alignment displays the frame and approximate contig22 coordinates for the exon.

Alignments of exons 4E-T:3_10825_0 and 4E-T:7_10825_2 showed missing residues compared to their *D. melanogaster* ortholog. Splice site identification will reveal whether the lack of aligned residues results from deletions in *D. eugracilis* or alignment mismatches. These discrepancies will be further investigated during subsequent splice site identification and CDS coordinate refinement. All coding exons alignments were collinear compared to the *D. melanogaster 4E-T* reference.

The coordinates for exon 1 were identified by inspection of RNA-seq reads. To identify the eight amino acids comprising the exon 1 CDS, the closest instance of RNA-seq data prior to exon 2 was examined. Sequence inspection revealed frame -3 to contain an eight amino acid coding region that initiates with a methionine, suggesting this region is the exon 1 CDS. This locus yields transcriptional output that decreases sharply almost immediately after the final putative exon 1 amino acid (Figure 7), providing further evidence that this region is spliced to assemble mature mRNA. Additionally, the SGP *ab initio* gene predictor identified the same region to be a coding exon. Since the CDS size, RNA-seq output, and a gene predictor support this decision, the eight amino acids identified tentatively comprise the *D. eugracilis* candidate ortholog to 4E-T:1_10825_0. Confirming the predicted exon 1 splice site is compatible with that of exon 2 remains necessary to conclude this exon is the *D. eugracilis* ortholog.

Splice Site Identification and CDS coordinate refinement

CDS boundaries were next refined using RNA-seq data and TopHat junction predictions. RNA-seq data allows single nucleotide resolution of intron-exon boundaries and was used to identify splice sites and exact CDS boundaries. Canonical splice sites in *D. eugracilis 4E-T* were

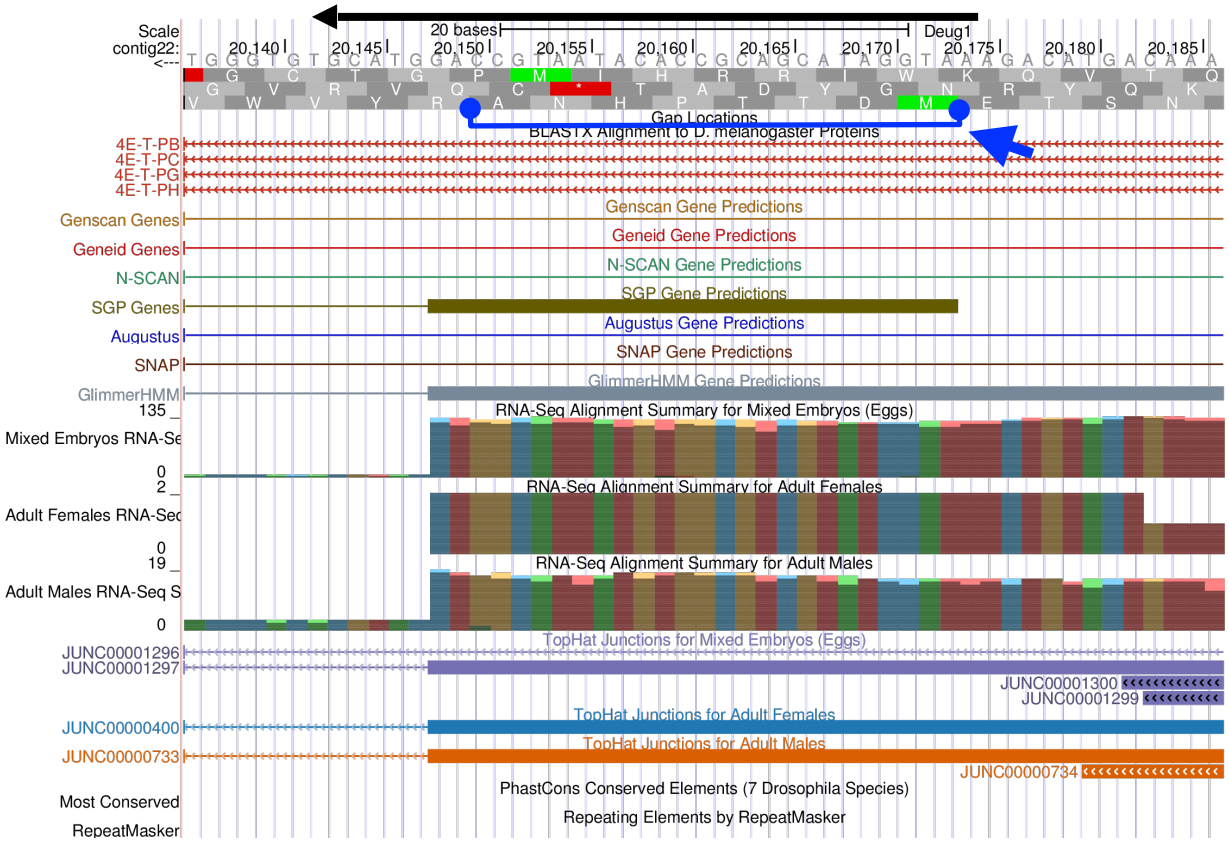


Figure 7: Sequence inspection upstream of exon 2 reveals a transcribed region that translates to eight amino acids and initiates with methionine. The amino acids indicated in blue correspond to the expected size and initial amino acid for exon 1. RNA-seq and TopHat data suggests that this region is spliced to exon 2, a known coding exon. This region is likely the orthologous *4E-T* exon 1 in *D. eugracilis*.

used at both donor (GT) and acceptor (AG) sites. Knowledge of the adjacent exon frame was combined with inspection of RNA-seq data and splice signals to define CDS boundaries. The sum count of phase nucleotides between all splice donor and acceptor pairs was confirmed to be a multiple of three; no annotated CDS boundaries resulted in a frameshift. Any ambiguities in splice site nucleotide usage were resolved using RNA-seq data and TopHat junctions.

To confirm the identified exon 1 can be functional, the exon 1 splice donor site and exon 2 splice acceptor site were examined. RNA-seq data, Top-Hat predictions, and knowledge of exon phase was applied to determine the exact CDS boundaries for exon 1 and the exact initial coding region nucleotide for exon 2. The initiating methionine for *D. eugracilis 4E-T* is at

nucleotide 20173 in reading frame -3 (Figure 7). The exon 1 splice donor is in phase 2 and the exon 2 splice acceptor is in phase 1 (Figure 8), confirming exon 1 is compatible with exon 2. The end of exon 1 was set to include phase nucleotide and extended to nucleotide 20148. The first nucleotide in exon 2 was set to include the phase nucleotide as extended to nucleotide 19975.

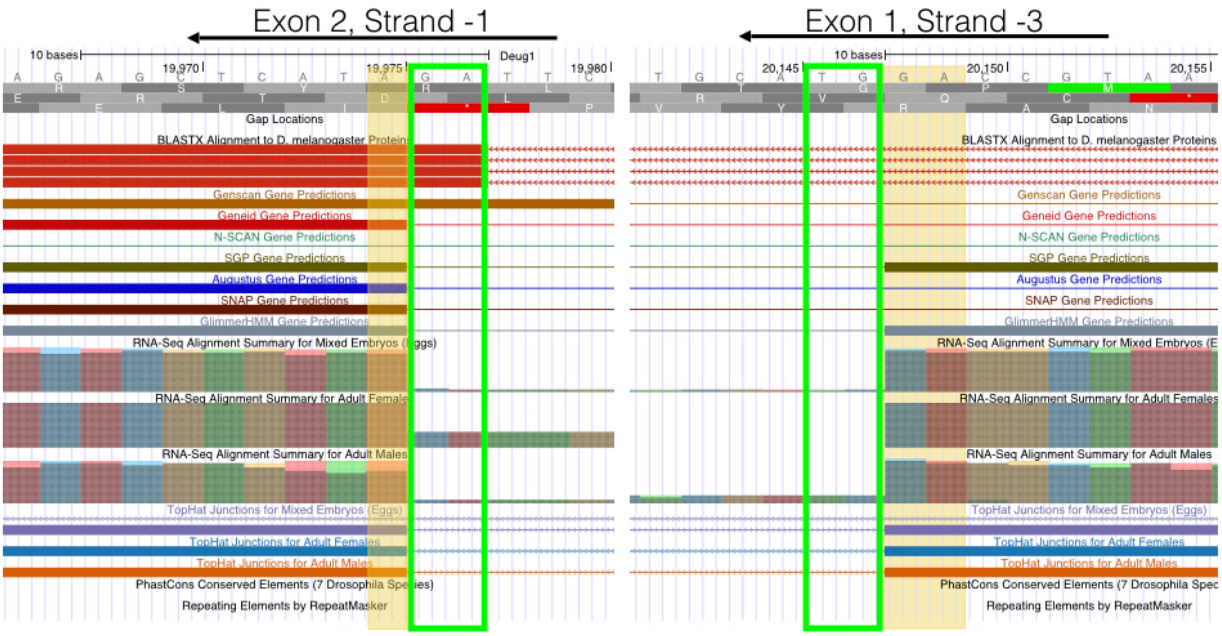


Figure 8: Splice junction connecting exon 1 and exon 2. The exon 1 splice donor is phase 2 and exon 2 splice acceptor is phase 1. Exon 1 is on strand -3, and exon 2 is on strand -1.

The next splice site boundary was examined to refine CDS coordinates. The best exon 2 splice donor candidate is in phase 0 and the exon 3 splice acceptor candidate is in phase 0. RNA-seq data and Top-Hat predictions supports these assignments (Figure 9).

The following splice sites between exon 3 and exon 4 were inspected. The candidate exon 3 splice donor is in phase 1 and the candidate exon 4 splice acceptor is in phase 2; both exons are on the -1 strand (Figure 10). RNA-seq data and Top-Hat predictions supported these phase assignments and CDS boundary assignments.

The splice site between exon 4 and exon 5 was examined next. The candidate exon 4

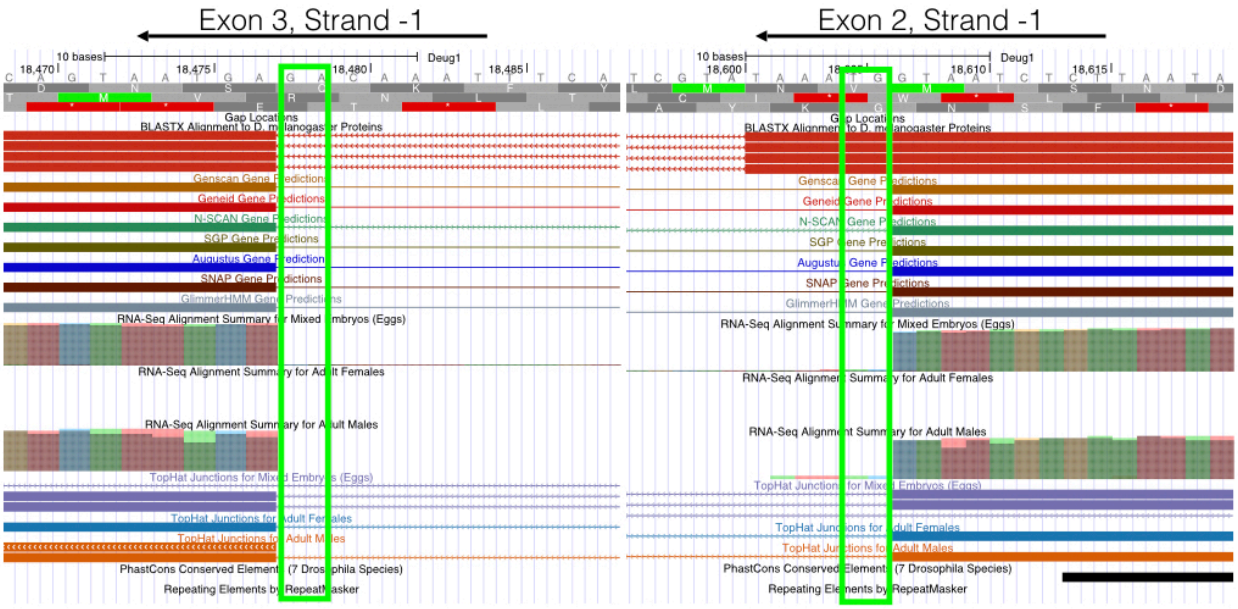


Figure 9: Splice junction connecting exon 2 and exon 3. The exon 2 splice donor is phase 0 and exon 3 splice acceptor is phase 0. Exon 2 is on strand -1, and exon 3 is on strand -1.

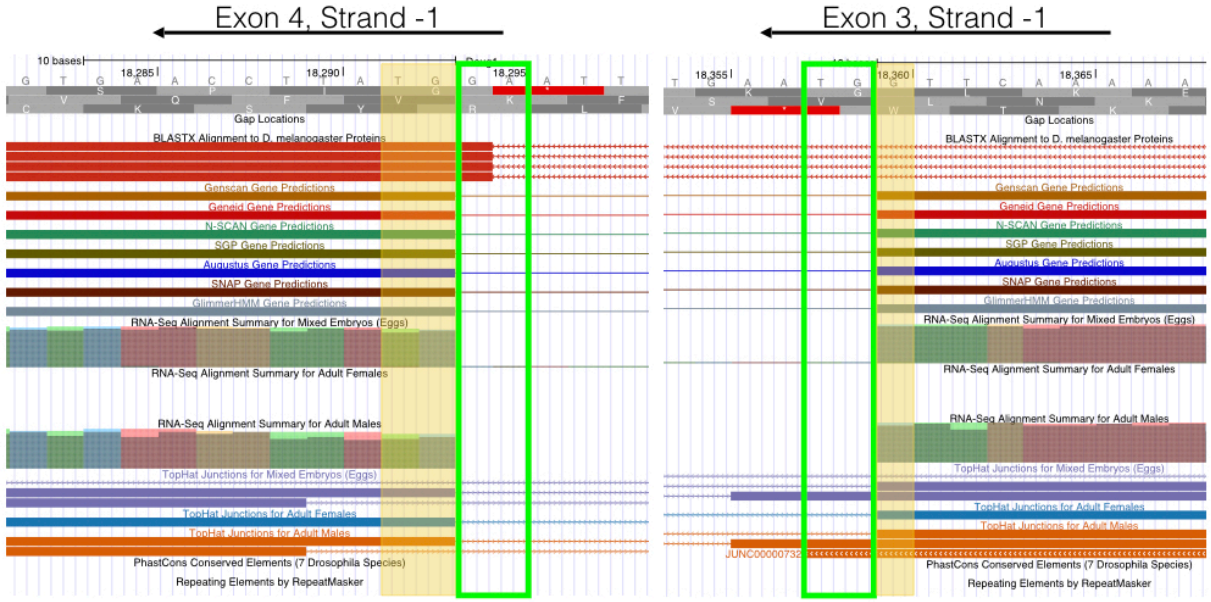


Figure 10: Splice junction connecting exon 3 and exon 4. The exon 3 splice donor is phase 1 and exon 4 splice acceptor is phase 2. Exon 3 is on strand -1, and exon 4 is on strand -1.

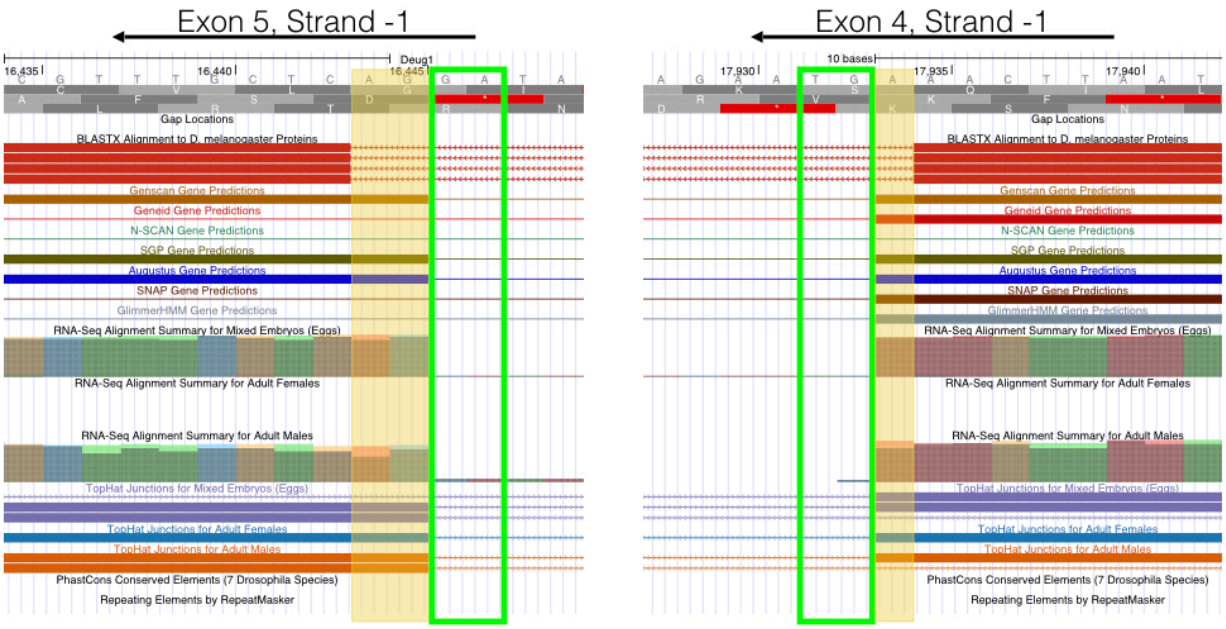


Figure 11: Splice junction connecting exon 4 and exon 5. The exon 4 splice donor is phase 1 and exon 5 splice acceptor is phase 2. Exon 4 is on strand -1, and exon 5 is on strand -1.

splice donor is in phase 1 and the candidate exon 5 splice acceptor is in phase 2; both exons are on the -1 strand (Figure 11). RNA-seq data and Top-Hat predictions corroborated CDS boundary assignment.

Isoform 4E-T-PH in *D. melanogaster* uses an alternative splice site that results in a shorter exon 6 (4E-T:6_10825_2) compared to isoforms 4E-T-PB/C/G. The orthologous shorter exon was identified in *D. eugracilis* by blastx alignment and splice site agreement. The candidate exon 5 splice donor is in phase 1 and the candidate shorter exon 6 splice acceptor is in phase 2; both exons are on the -1 strand (Figure 12).

This process was applied to identify the best candidate splice donor and acceptor sites for all *4E-T* exons and to refine the CDS coordinates. Amino acids missing at exon boundaries from blastx alignment were reincorporated in the predicted *D. eugracilis* CDS coordinates based on RNA expression data. Approximate CDS coordinates were corrected based on splice nucleotides,

RNA-seq support, and Top-Hat predictions. The stop codon for this gene can be seen in Figure 13. The *4E-T* gene model with exact CDS boundaries is described in Table 2.

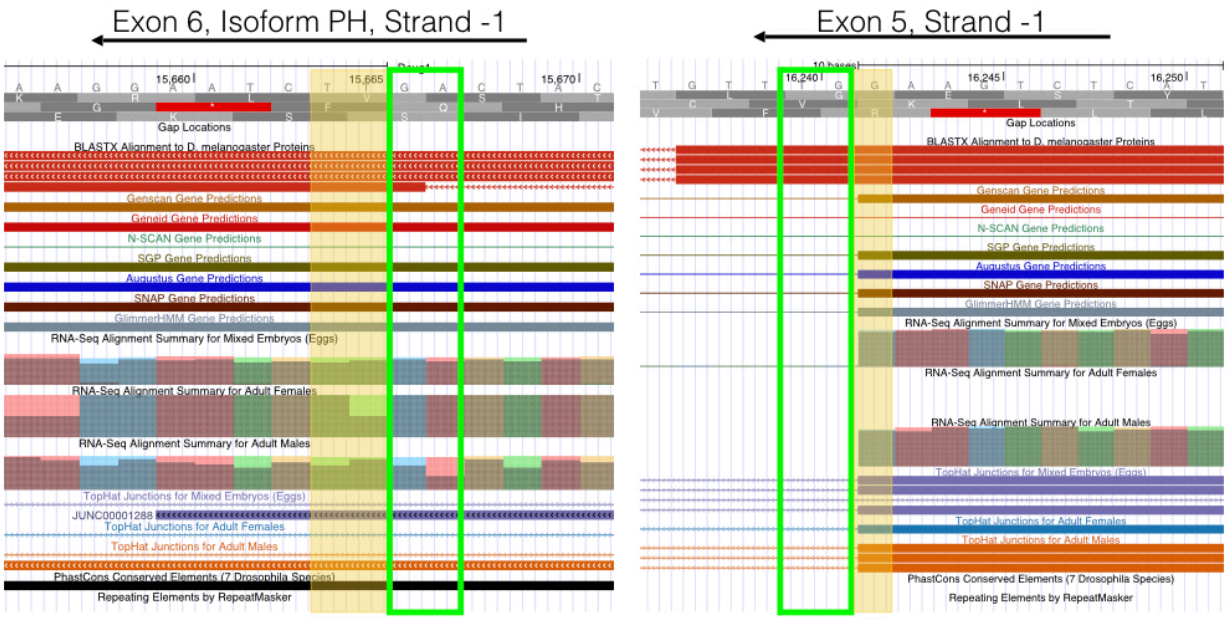


Figure 12: Splice junction connecting exon 5 and the alternatively spliced exon 6. The exon 5 splice donor is phase 1 and exon 6 splice acceptor is phase 2. Exon 5 is on strand -1, and exon 6 is on strand -1.

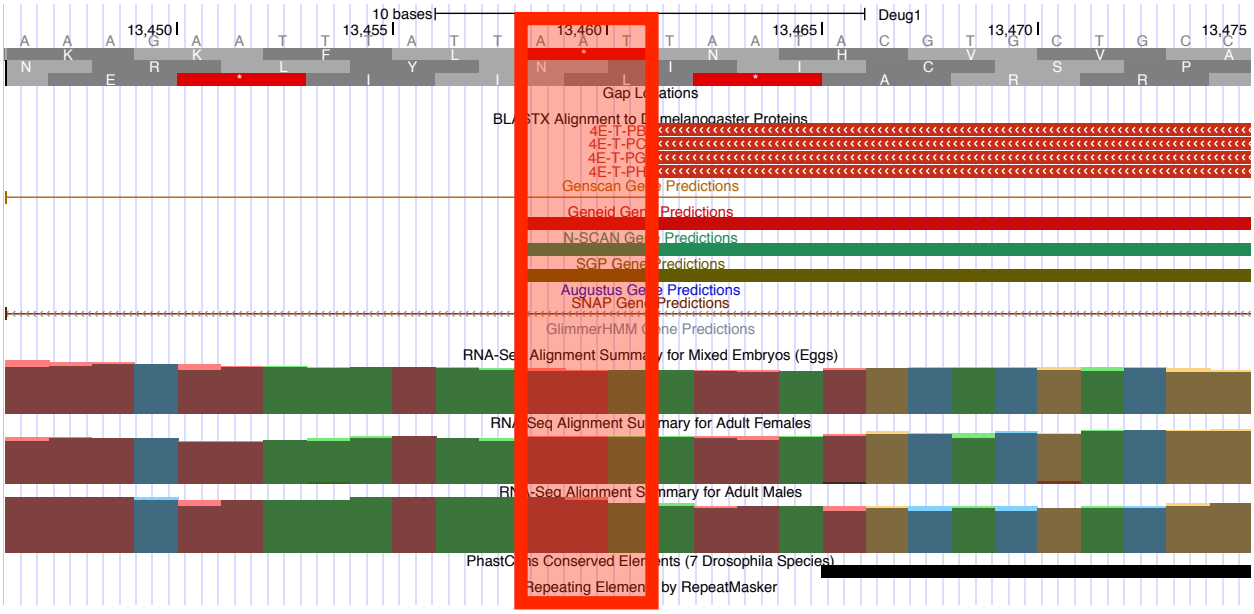


Figure 13: Stop codon observed in exon 8. Exon 8 is in frame -1. The stop codon displayed in red is corroborated by blastx homology to the *D. melanogaster* homolog and matches the coding sequence end as suggested by gene predictors. In addition, the final exon coding sequence demonstrated conservation to the *D. melanogaster* ortholog (Figure 6).

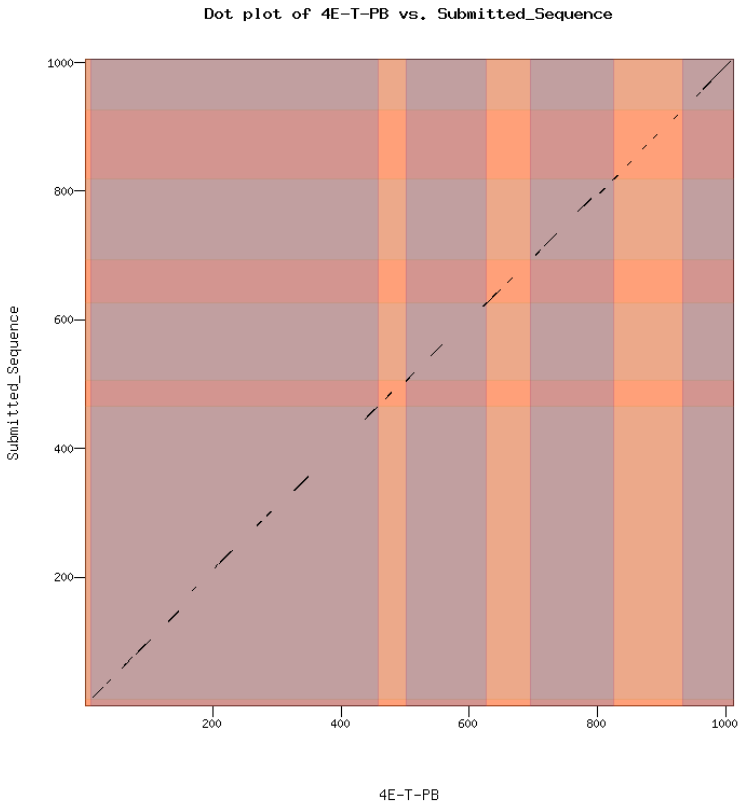
	D. eugracilis ortholog								
FlyBase_ID	Exon Number	Isoform Usage	End	Start	Frame	Donor phase	Acceptor phase	Exon Size (nt)	D. mel Exon Size (nt)
4E-T:1_10825_0	1	B, C, G, H	20148	20173	-3	2	-	25	25
4E-T:2_10825_1	2	B, C, G, H	18607	19975	-1	0	1	1368	1344
4E-T:3_10825_0	3	B, C, G, H	18360	18477	-1	1	0	117	129
4E-T:4_10825_2	4	B, C, G, H	17934	18293	-1	1	2	359	374
4E-T:5_10825_2	5	B, C, G, H	16242	16445	-1	1	2	203	206
4E-T:7_10825_2	7	B, C, G	15384	15758	-1	1	2	374	386
4E-T:6_10825_2	6	H	15384	15665	-1	1	2	281	293
4E-T:8_10825_2	8	B, C, G, H	13760	14083	-2	0	2	323	323
4E-T:9_10825_2	9	B, C, G, H	13462	13700	-1	-	0	238	238

Table 2: Gene model for *D. eugracilis* ortholog of *4E-T*. CDS boundaries were refined by splice site inspection and corroborated by RNA-seq data and Top-Hat predictions. Exon splice site phases are summarized. All putative *D. eugracilis* exons roughly match the size of their respective orthologs.

Gene Model Verification

The gene model for *D. eugracilis 4E-T* isoform B was verified using the GEP Gene Model Checker. All four isoforms passed validation. The B, C, and G isoforms produce identical proteins so only the 4E-T-PB similarity dot-plot and alignment are shown (Figure 14). *D. eugracilis* 4E-T-PB demonstrates high similarity to its *D. melanogaster* ortholog overall. While regions of mismatch occur in exons, all exon boundaries are conserved upon inspection of the pairwise exon alignment (Figure 14B). A similar pattern is seen for 4E-T-PH (Figure 15). *4E-T* is conserved with reference to the *D. melanogaster* ortholog and the most parsimonious model for comparative annotation passes validation.

A



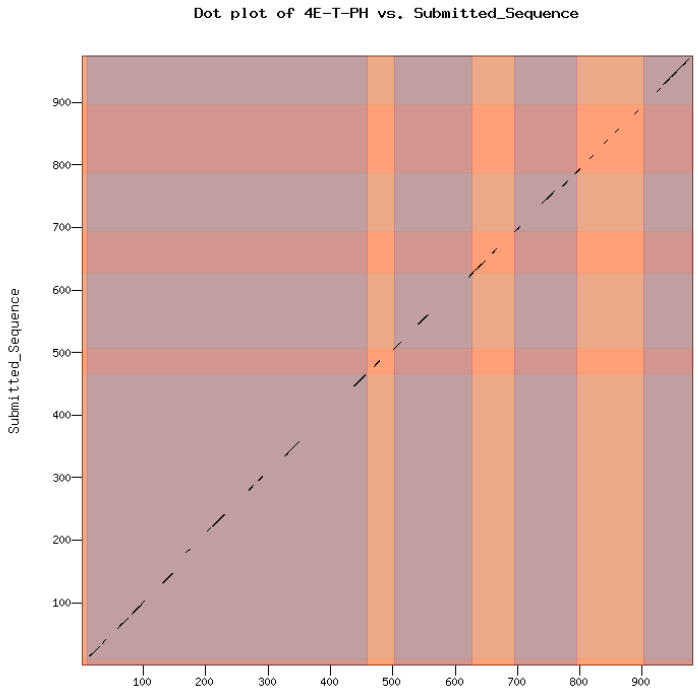
B

Identity: 644/1025 (62.8%), **Similarity:** 780/1025 (76.1%), **Gaps:** 35/1025 (3.4%)

```
4E-T-PB          1 MDTSKISARYSKVDLLALRYEGKSRQRPQCSTRLELQTLGFWKINLNTAALTVSSAYSNQ  60
                   ***:  .****:.********.****** *:.* ******.*** :**:.:*:*:. * **
Submitted_Seq    1 MDTTPHNARYSRADLLALRYESKSRQRPPCTNRTELQTLSFWKRDLNAVSLSVTNNYLNQ  60

4E-T-PB         61 NKNRLSPEADNSSLICSNSSSISSRRAMRNRERANNYYQRFVPTDSLLISGEDKDKDALS 120
                   ********:***** **.*.*****************.******.*:*** :.****:*** :
Submitted_Seq   61 NKNRLSPETDNSSLNCSSSGSISSRRAMRNRERANSYYQRFAPSDSLQLCGEDKEKDAST 120

4E-T-PB        121 HGQPYKLNIIDHRSISSSHLMPAFAKRRFVISKGSNSEESNEG-INTC---------AS 169
                   ***.:* .*:*****************;*** :.**** *.**. :. *          *
Submitted_Seq  121 HGQSFKSSILDHRSISSSHLMPAFAKKRFVAATGSNSAENNEAAVSICDDGIASQRKES 180

4E-T-PB        170 KGKAASSPSRKGSELDTAETCLNFVQPDHDQCMSSSPTFSTSRQERRIGSGRLLPRSDNW 229
                   ****: **:**..* *.:** **:**.**:**:*****:*:*********************
Submitted_Seq  181 KGKASISPTRKSNEQDNSETRLNYVQSDHEQCLSSSPTLSASRQERRIGSGRLLPRSDNW 240

4E-T-PB        230 DYKNEKTVEASIENEKETSPNGSGSTSSLNQHNQSQHRSRTTFSGRLVERVPEVTDRRFQY 289
                   :**. *: *.. * **: * ****.:.  .*:**.***.*******:;** * :******
Submitted_Seq  241 EYKSLKAKEPNSEIEKDMSLNGSGGACGASQYNQNQHRNRTTFSGRLIDRVLEHSDRRFQY 300

4E-T-PB        290 DSKKSFDRQGINNRRISGKEPFSTQSRSKRGNSYLIHEEPEWFSAGPKSQLETIDLHGFE 349
                   *:*:*.**** .***:*.**   :.**.**.*** *:**********.***********:
Submitted_Seq  301 DTKRSVDRQGASNRRVSNKE---SSSRGKRANSYHIYEEPEWFSAGPTSQLETIDLHGFD 357

4E-T-PB        350 DLEKNEERSVTEDKNNQIQQLDKNLDAQASKDEASMRNSNDSLNFREVIPSDEKKHTDEN 409
                   ***:***   ***:*::: ***.:* ** .:* .* *.** **.: :. **:: *.*.**
Submitted_Seq  358 DLENNEESFDTEDRNEEFPQLDTKLVAQRNNDVVSRRSSNVSLSLSDANPSNDIKDTGEN 417

4E-T-PB        410 VVTSIQNSTDLGHPNKNKPIQMQPSQNPESEFNFDAFLNMHPLDNSVTSNDETGKSDSKG 469
                   ::. ***::::.: ***:* *:* **..******************:;***   *.::*.
Submitted_Seq  418 ILNFIQNTSEMSKHNKNQPSQLQYSQTSESEFNFDAFLNMHPLDNSLKSNDGIEKNETKA 477

4E-T-PB        470 TSRFSRWFRQKEAANNNEFPGFRESHAQEKRGIPSVKDLEAQMIKVDMRTDLINPIAGSL 529
                   *********:**.    *  .: : *:*** ******:***** *:*::** :.*:**.:
Submitted_Seq  478 TSRFSRWFRHKEP----ETLSLGDLHTQEKLGIPSVKELEAQMTKMDIKTDSVSPVAGPF 533

4E-T-PB        530 CQTVQMEKPIARDTEAFKKLLQQLGSQARQHHPCNDDCRTINLSNIANHVHLESKLHQKI 589
                   * *: ****:******************::* *. **  :* .:     * ::**:  :*:

Submitted_Seq  534 PQIVEAEKPISRDTEAFKKLLQQLGSQTKQPHSGNDVYHTNTI-----HDQIESNHSHKV 588

4E-T-PB        590 NDGHLQQPELSVNVPTMPTSSHVFLQKRLEIQHLIQRLHCGDVSHDFLEKELDNPSTPAA 649
                   ** * *** *.. **.:*:..**. *:*:**:****** ***** *******.*****.:
Submitted_Seq  589 NDCHPQQPALNAYVPNIPSNNHVHTQNRMEIEHLIQRLVCGDVSLDFLEKELSNPSTPVT 648

4E-T-PB        650 TKDVIATVLNEYSHSKRNPVVTGDPNIFTQQSFLQPQSVHQHYSQEEHSQNTANHTINQL 709
                   ;*:**:***.******;* :. *: ***.  * **.*.::*:**::*  .***:*******
Submitted_Seq  649 AKEVISTVLREYSHSKKNLMAMGELNIFNP-SALQAQQIQQRYSEDLLPQNTSNHTINQL 707

4E-T-PB        710 ISHGNSPTPLAFTPTSVLRKMTADKD-TQSPSTYCQNPQYHVHQQNAKQVGTRENVLEPQ 768
                   ::**.**.*****************: **: .:  .   :::.*:.**:.*::** *.*
Submitted_Seq  708 MAHGTSPPPLAFTPTSVLRKMTADKEATQTLCSSHSLQTPYLNSQHVKQISTHKNVHESQ 767

4E-T-PB        769 LTATMAVQPRMILGGGNFAIGQNNQHLSPNMSQSRNQQVLKWTSGNMQMVHGKTFGRPIL 828
                    ******.*************** ***      *.*.********:;***:;*:**:******
Submitted_Seq  768 PTATMAAQPRMILGGGNFAIGLNNQ-----MPQCRNQQVLKWATGNMHVVQGKSFGRPIL 822

4E-T-PB        829 KGGLNSMPHSNSALPFTAHKIEMQPIHQPHLQQQQH--RFKAVQSVESNLNTESVHQNIT 886
                   **.***:*:.**:::**:::***** .:** :  * *  ***:.***** *.**.****::
Submitted_Seq  823 KGSLNSLPQQNSTVPFSSHKIEMPTVHQQQQMQPPHQLRFKSTQSVESALSTENVHQNMS 882

4E-T-PB        887 SPVGWHQLYMQHQQQHHHTRQQLSQRVIYGEMHRQSNPQMSPPVPGFSDSSDSGNVIKAN 946
                   *****:**::****  :: ***  :*:***::***** ***.*.*.***.**.**:** *
Submitted_Seq  883 SPVGWYQLFLQHQQ--NQIRQQPRHRLIYGDVHRQSNIQMSSPAPNFSDNSDPGNMIKNN 940

4E-T-PB        947 SLTSPSYQRDERISSPTN-QLAQWFSPELLAKASAGKLPLLNVNQALSLEEFERSIQHSS 1005
                   *:*:.:*:*****.**** ************************;*****************
Submitted_Seq  941 SITTSAYHRDERIQSPTNNQLAQWFSPELLAKASAGKLPLLNMNQALSLEEFERSIQHSS 1000

4E-T-PB       1006 GVVHN 1010
                   .****
Submitted_Seq 1001 AVVHN 1005
```

Figure 14: Dot-plot and pairwise exon alignment produced by Gene Model Checker for 4E-T-PB/ PC/PG. Dot plot shows overall conservation to homologous exons in *D. melanogaster*. Exon alignment shows regions of mismatch within exons, but conserved residues at exon ends.

A



B

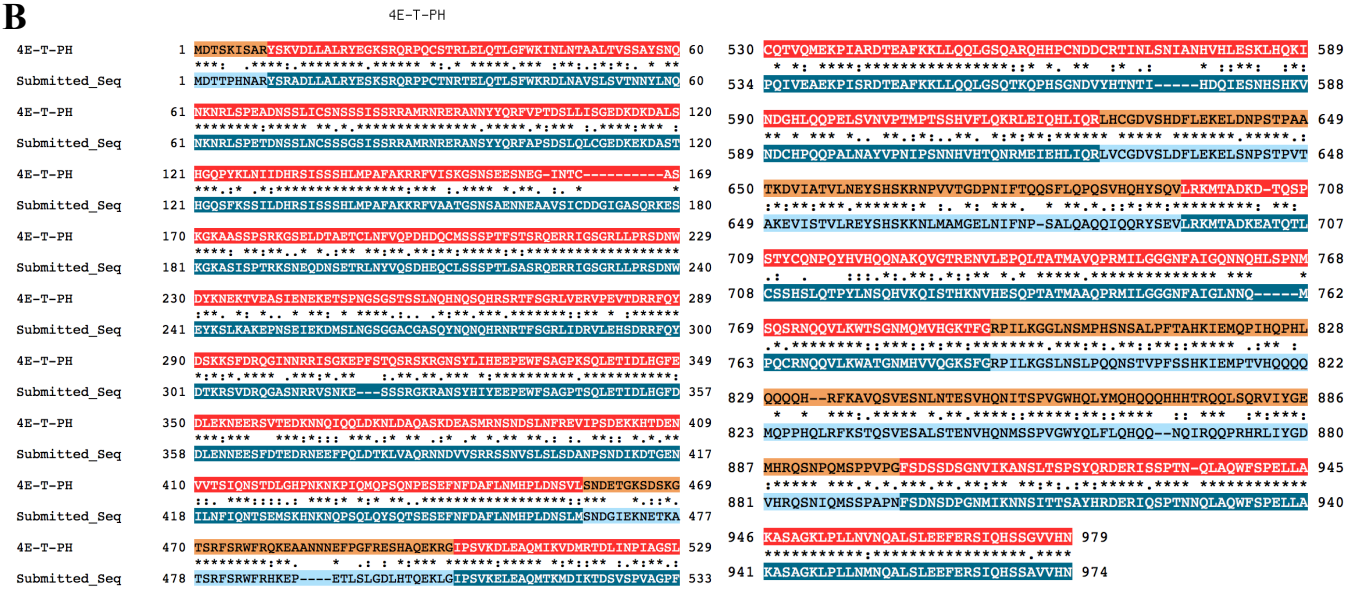


Figure 15: Dot-plot and pairwise exon alignment produced by Gene Model Checker for 4E-T-PH. Dot plot shows overall conservation to homologous exons in *D. melanogaster*. Exon alignment shows regions of mismatch within exons, but conserved residues at exon ends.

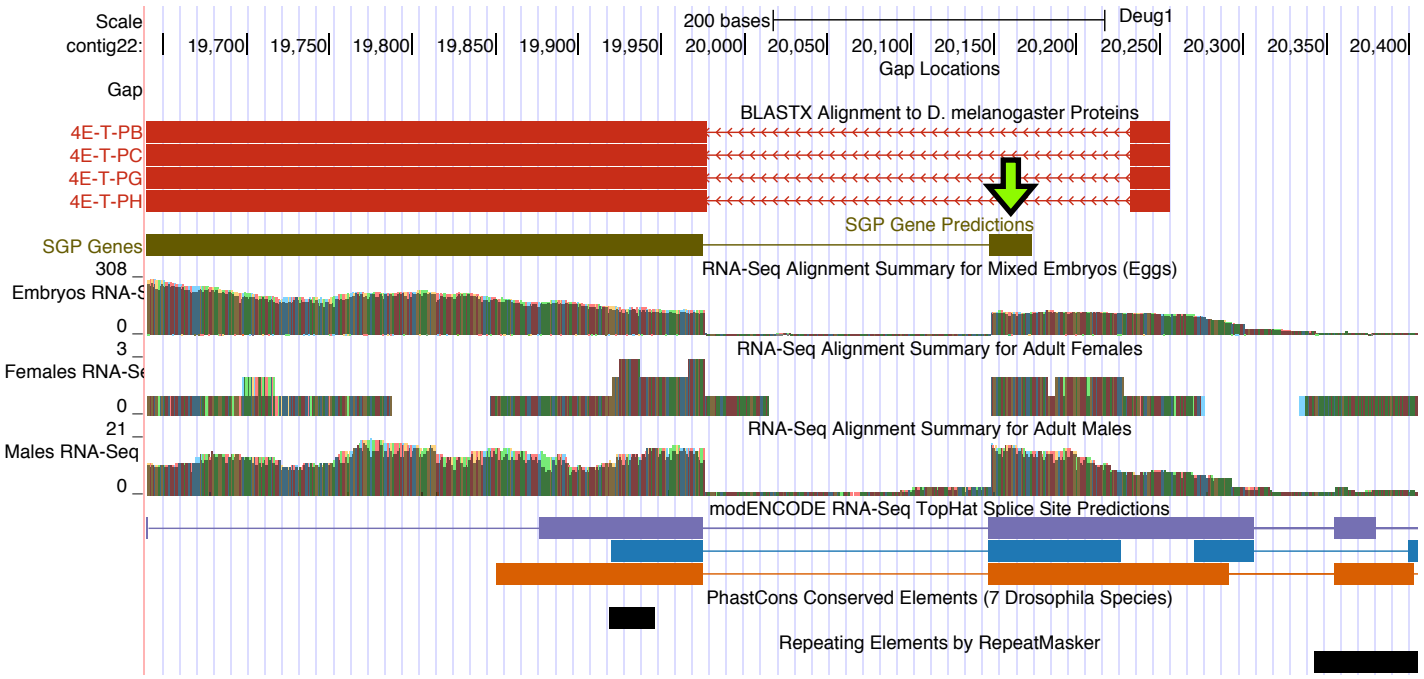
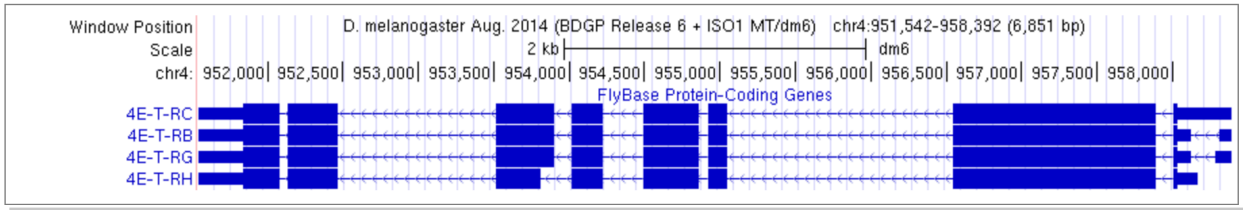


Figure 16: Putative *D. eugracilis 4E-T* 5'-UTR. RNA-seq reads extend upstream of the initial coding exon (green arrow), suggesting that the TSS occupies this region. The SGP gene predictor identified the conserved first coding exon as confirmed during coding exon annotation (green arrow).

Transcription Start Site Identification

An overview of the *D. eugracilis 4E-T* 5'-untranslated region (5'-UTR) is shown in Figure 16. RNA-seq reads extend upstream of the first coding exon, suggesting that the TSS may be present in this upstream region. In *D. melanogaster*, there are four *4E-T* isoforms. Isoforms B, C, and G share a transcription start site, and isoform H initiates transcription downstream from the other isoforms (Figure 17). Examination of the transcription start region in *D. melanogaster* demonstrates that this locus is transcriptionally accessible in BG3 and S2 cells according to the 9-state epigenetic model (Figure 18). Clustered immediately upstream of the B, C, and G isoforms, single DNaseI hypersensitive sites (DHSs) are observed in BG3, S2, and Kc cells and a single TSS is annotated by the modENCODE project. RAMPAGE and CAGE experiments corroborate these TSS placements, although both peak about 100 nucleotides downstream (Figure 18). There exist three classifications for promoters. Peak promoters are defined by



FlyBase ID	FlyBase Name	Chr	5' Start	3' End	Strand
FBtr0089233	4E-T-RC	4	958,392	951,542	-
FBtr0089229	4E-T-RB	4	958,392	951,542	-
FBtr0310651	4E-T-RG	4	958,392	951,542	-
FBtr0334626	4E-T-RH	4	958,170	951,542	-

Figure 17: *D. melanogaster 4E-T* FlyBase record. *4E-T* consists of four different isoforms utilizing two distinct transcription start sites (green arrows). The distinct TSSs are 218 nucleotides apart.

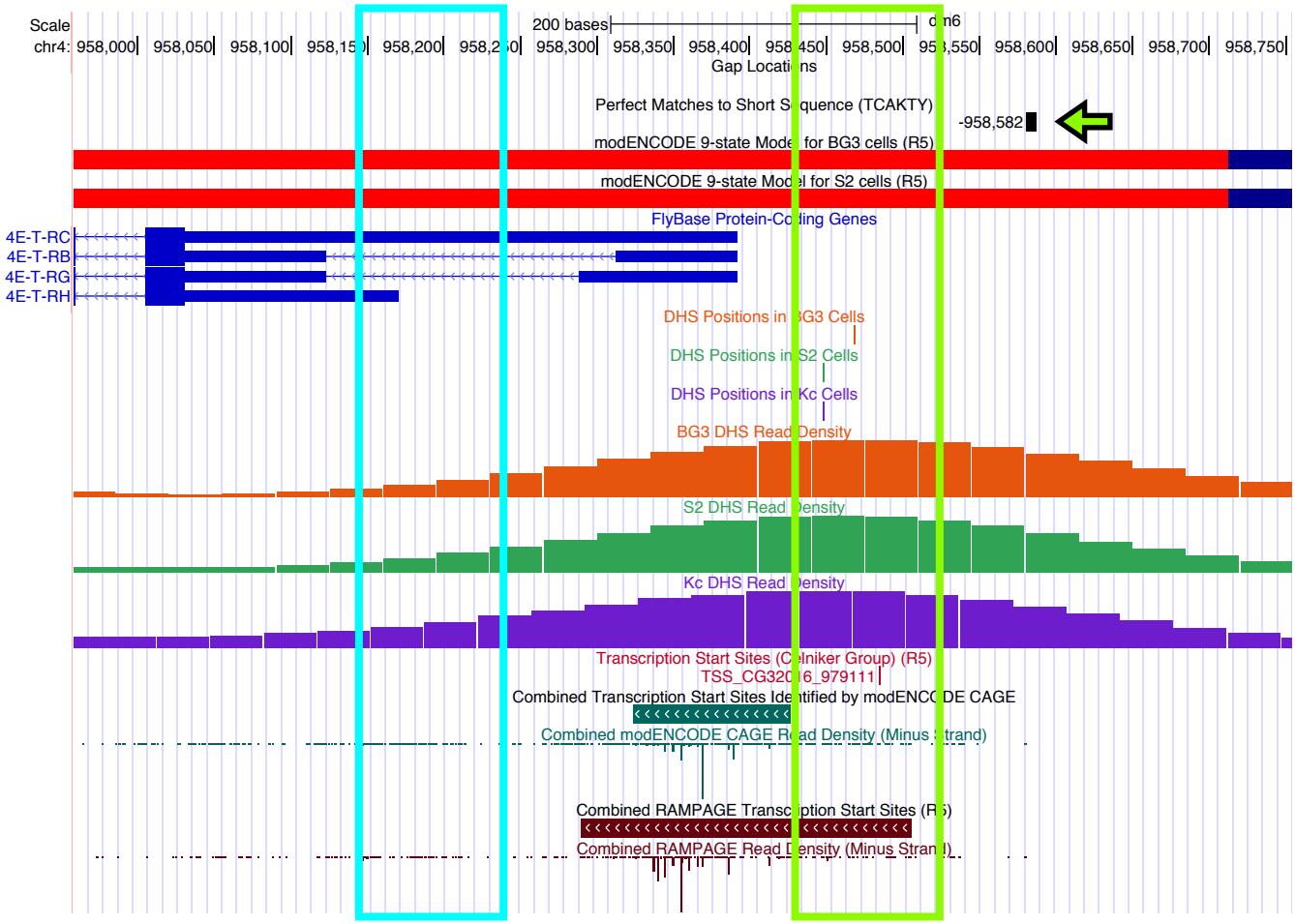


Figure 18: ***D. melanogaster 4E-T* 5'-UTR and promoter region.** The four *4E-T* isoforms are shown in blue. The nine-state epigenetic model defines the promoter region as transcriptionally accessible (red track). DHS positions, TSSs, CAGE peaks and RAMPAGE peaks all occur around the same region, suggesting a peak promoter (green box). There is an Inr motif present, but about 150 nucleotides upstream of the predicted TSS and DHS positions (green arrow), suggesting it is not relevant to transcription initiation at the predicted TSS. There is no evidence to support a distinct TSS for 4E-T-RH (cyan box).

possessing a single DHS and a single annotated TSS; intermediate promoters have a single DHS and multiple TSSs, or vice versa; broad promoters have multiple DHS positions and multiple annotated TSSs. Using this classification, the B, C, and G isoforms of *D. melanogaster 4E-T* utilize a peak promoter.

To identify the first exon ortholog in *D. eugracilis*, pairwise blastn alignment was performed using the first exon sequence of 4E-T-RC (query) and the contig22 sequence (subject). The first 4E-T-RC exon was used because it shared a TSS with the B and G isoforms and has the longest exon 1 sequence region, thus providing more nucleotides to guide alignment. To account for a greater incidence of mutations in non-coding regions, blastn parameters were made more sensitive. Word size was set to 7; scores for matches and mismatches were set to 1 and -1, respectively; the gap introduction and extension costs were set to 2 and 1, respectively, and low complexity regions were not filtered. The highest-scoring alignment placed the *D. eugracilis* TSS at position 20580 on contig22. Exactly 298 nucleotides of 387 nucleotide query sequence (86.6%) aligned to contig 22, and 61% of the aligned region matched exactly (Figure 19). This position is upstream of the predicted first *4E-T* coding exon and upstream of most RNA-seq reads.

To define a putative, orthologous, TSS in *D. eugracilis*, the sequence surrounding position 20580 on contig22 was inspected for core promoter motifs. An Inr motif was observed at position 20579, corroborating the evidence of an orthologous TSS in the region (Figure 20). Using this motif, the 4E-T-RB/C/G TSS is predicted at position 20582. There are a few RNA-seq reads extending upstream of this TSS, suggesting transcription begins prior to position 20582 in some cases. To account for these reads, a narrow search region was defined upstream of the

Sequence ID: Query_19957 **Length:** 45000 **Number of Matches:** 109

Range 1: 20250 to 20580 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
66.8 bits(45)	1e-13	206/335(61%)	41/335(12%)	Plus/Minus

```
Query  1      TTCAATAAGAACAAAATATTTTCAATGTCTAT---ATCCCACTCC-AATTTAGCGTGATA  56
              ||| |||||| |   | |  ||| | |||| ||   |||||||||| |||||  ||||||
Sbjct  20580  TTCGATAAGACCCGATTCATTTTACTGTCAATCATATCCCACTCCTAATTTGACGTGATT  20521

Query  57     TTCACAAAGCTAA-----TCGCAACATTGTGAGCATTAAAAATTGT-TA--TATATGTAT  108
              |||| |||   |||     | | || |  | | || ||||  |||| ||  |||| ||||
Sbjct  20520  TTCAGAAAAATAAAATTATTGGAAAAC-GCCAACAATAAAT-TTGTCTACTTATAAGTAT  20463

Query  109    GTAAGCGTCTTGGGATTATATTCTCATGAAACTCAACAGTCATAT--ATATGGCCTTAAA  166
                ||    |       |  |  ||  ||  || |  |  |    ||||  ||||   | |||
Sbjct  20462  TGAAAT-TAAAAAAAACAACTTAACAAAAACCGAATGATATATATGTATATACACACAAA  20404

Query  167    ----------ATACGTATATATATATATATATAT-TAATTTTTCAAA-GTGTGTGTTTGC  214
                       |||  |||||||||||||||||| ||| | |||  | || ||||||||||
Sbjct  20403  TTTGTAATATATATCTATATATATATATATATATATAAATATTCGTATGTATGTGTTTGC  20344

Query  215    ATAAGTGGTGTGTAC----TATTTGTATATTCATTGTA-----CAATTT--CTCCCGTTC  263
              |||||||||  | |||      ||||  |||||| | |   |      |||||   | |  |||
Sbjct  20343  ATAAGTGGTATATACAAAAAATTTTTATATTAAATAGAAACAGGAATTTAGTTTCTTTTC  20284

Query  264    ATATAGGAGAGTGTTCTGTTTTTAATAATTTTGGA  298
               |  |||  |||  | | |  ||||  |||||||
Sbjct  20283  CTTCAGGGGAGAATT-TTTATATAATTGTTTTGGA  20250
```

Figure 19: blastn alignment of the *D. melanogaster* 4E-T-PC first exon (query) against *D. eugracilis* contig22 (subject). Alignment predicts the transcription start site is at position 20580 on contig 22 (green). The query length was 387 nucleotides; of the 298 nucleotides aligned, 61% matched contig22 exactly.

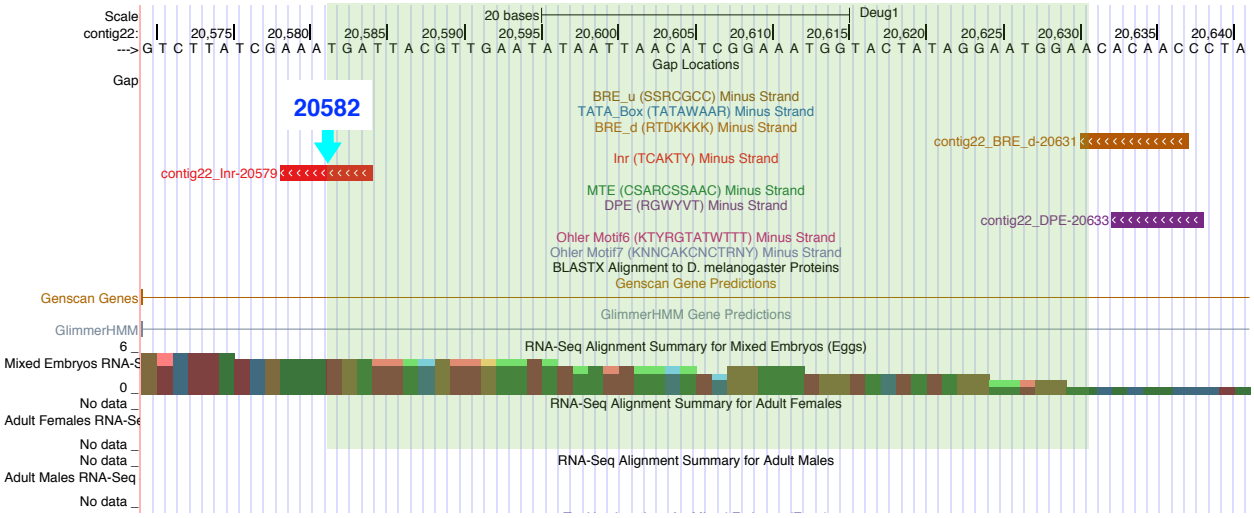


Figure 20: Putative *D. eugracilis 4E-T* TSS and narrow search region. An Inr motif is immediately upstream of the TSS defined by blastn alignment (cyan arrow). Thus, the TSS is predicted at position 20582. RNA-seq reads extend upstream of the putative TSS. The region from position 20582 to 20631, where the count of RNA-seq reads drops to one, is defined as the narrow search region (green shading).

Motif	*D. mel.* position	*D. eug.* position
BRE^{d}	958180, 958182, 958184, 958371, 958445, 958616, 958729, 958757	20345, 20347, 20349, 20631, 20679, 20731, 20733, 20749, 20774
TATA Box	958716	20365
Inr	958582	20559, 20579
DPE	958242, 958532	20633, 20708

Table 3: Core promoter motifs observed in the 300 nucleotides flanking the *D. melanogaster* annotated TSS and the putative *D. eugracilis* TSS. The highlighted motif supports the TSS position in *D. eugracilis*. All motifs lie in the same direction as 4E-T. No instances of the BRE^{u}, MTE, DRE, or Ohler core promoter motifs were observed in the region considered.

putative TSS until the number of RNA-seq reads reduced to one (nucleotides 20582 to 20631) (Figure 20). Interestingly, no orthologous Inr motif is observed at the *D. melanogaster* TSS (Figure 18), suggesting *D. melanogaster* evolved distinct measures for ensuring *4E-T* transcription. All core promoters motifs within 300 nucleotides of the predicted *4E-T* TSSs in *D. melanogaster* and *D. eugracilis* are described in Table 3.

To define the TSS for *D. eugracilis* 4E-T-RH, the orthologous region in *D. melanogaster* was examined. There is no evidence for a 4E-T-RH TSS (Figure 18, blue box). The FlyBase transcript report for *D. melanogaster* 4E-T-RH described the evidence for the isoform as "weakly supported" with a 3/10 evidence score (Figure 21). In addition, there are no cDNA clones that support the transcription initiation site of this isoform. This evidence suggests 4E-T-PH may not be expressed from its own TSS in *D. melanogaster*. To search for a *D. eugracilis* 4E-T-PH TSS, the predicted *D. melanogaster* exon 1 of isoform H was used as query for blastn alignment as described previously. The predicted 4E-T-RB/C/G can be identified in the previous blastn alignment (Figure 19). However, when using the isoform H exon 1 sequence as query, no resulting alignments overlapped with the *4E-T* 5'-UTR region in contig22, suggesting that the *4E-T* H isoform 5'-UTR is weakly conserved (data not shown). In addition, RNA Pol II ChIP-

seq reads in *Drosophila biarmipes* do not demonstrate a secondary peak that could correspond to a unique TSS for the orthologous H isoform. Taken together, no evidence accumulates to support the existence of an orthologous 4E-T-RH isoform in *D. eugracilis*.

General Information			
Symbol	Dmel\4E-T-RH	**Species**	*D. melanogaster*
Annotation Symbol	CG32016-RH	**FlyBase ID**	FBtr0334626
Feature type	mRNA	**Associated gene**	Dmel\4E-T
Evidence Rank	Weakly Supported	**Length (nt)**	3376
Evidence Score	3		

Figure 21: *D. melanogaster* 4E-T-PH FlyBase transcript report. Isoform H is characterized as weakly supported with an evidence score of 3/10 (green arrows).

Gene Evolution

4E-T displayed high RNA expression, especially at the embryonic stage, suggesting a conserved role for the protein. To determine if *4E-T* possesses conserved domains, the peptide sequence for *D. eugracilis* 4E-T-PB was used as query for blastp search against non-redundant (nr) sequences. Of the 1005 amino acid protein, residues 10-733 were identified by the Conserved Domain Database (CDD) as part of the EIF4E-T superfamily (Figure 22). Insects not in the *Drosophila* genus were selected for comparative alignment. The predicted orthologous 4E-T-PB protein sequences for *Rhagoletis zephyria* (Snowberry fruit fly), *Dendroctonus ponderosae* (Mountain pine beetle), *Cephus cinctus* (Wheat stem sawfly), and *Dinoponera quadriceps* (South American ant) were used as input for *Clustal Omega* multiple sequence alignment. All selected non-*Drosophila* isoforms were predicted sequences; thus, the predicted *4E-T* isoform with

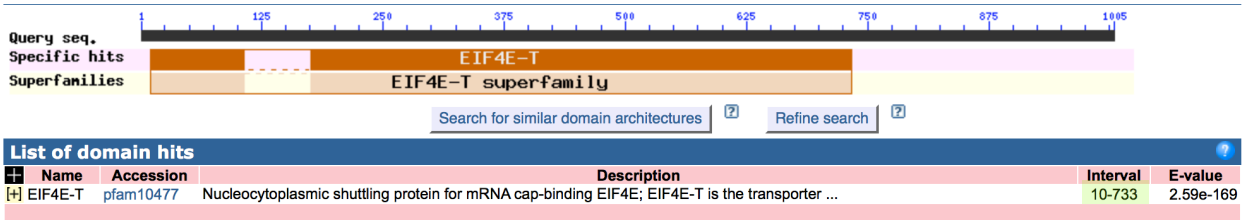


Figure 22: Putative *D. eugracilis 4E-T* contains a conserved domain. The Conserved Domains Database identifies residues 10 to 733 (green shading) as part of the EIF4E-T superfamily.

highest similarity to *D. eugracilis* 4E-T-PB was selected for each species. Alignment displays residue regions of high conservation across insect species within the superfamily domain (Figure 23). Pairwise alignment with human *4E-T* (not shown) directed identification of the *4E-T* nuclear localization signal (Figure 23, green arrow). Overall, sequence alignment illustrates divergence among insect species but retention of possibly functional domains.

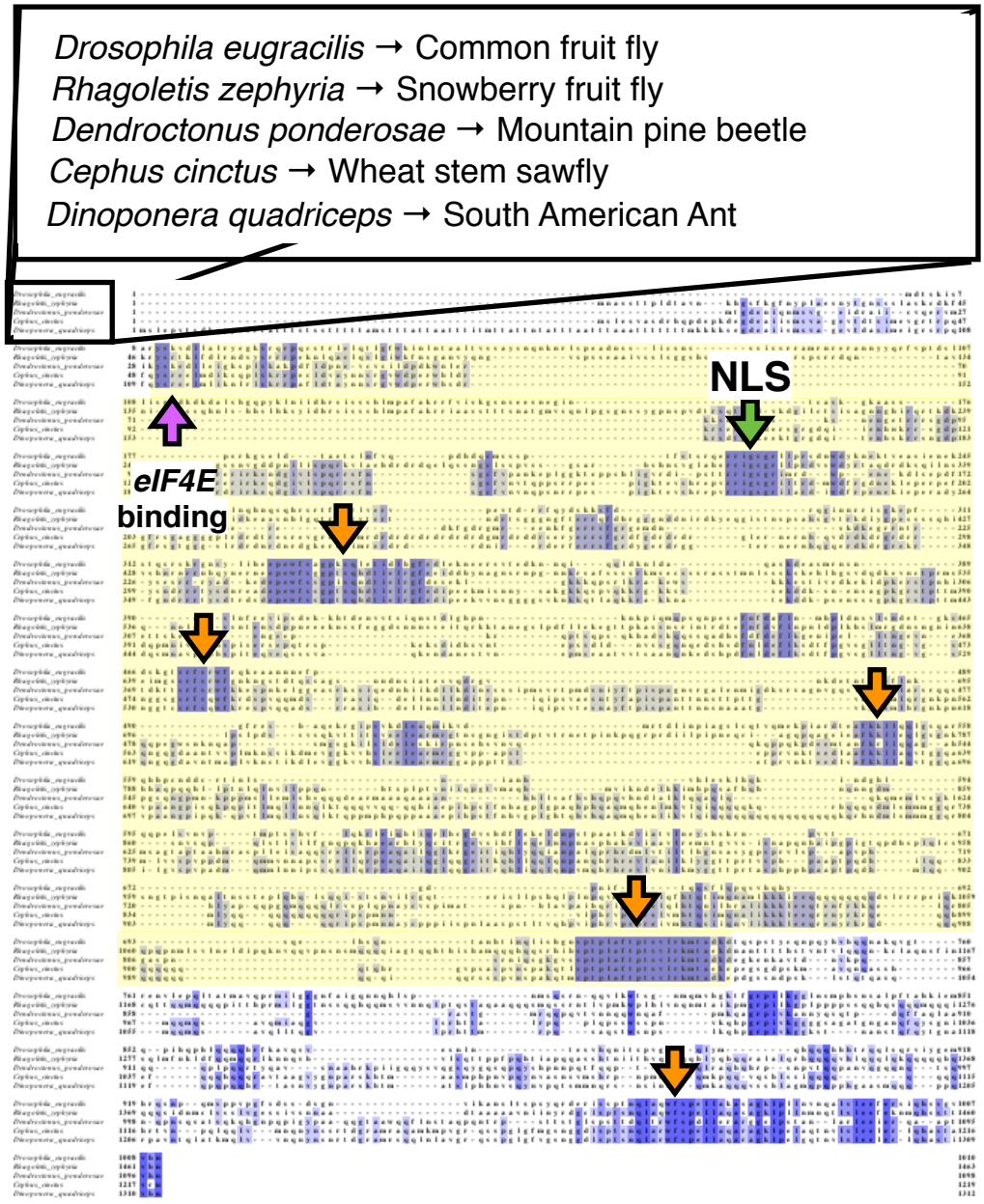


Figure 23: *Clustal Omega* multiple sequence alignment with insect *4E-T* homologs reveals conserved domains. The alignment of five insect species is shown. The EIF4E-T superfamily identified by CDD is shown in yellow shading. The conserved NLS (green arrow) was identified by comparison to the annotated human *4E-T*. An *eIF4E* binding site (purple arrow) was identified from structural predictions. Other possible conserved domains are indicated with an orange arrow and were selected upon qualitative assessment of the alignment, without any guiding criteria.

Structure Analysis

To illustrate functional domains at the three-dimensional structure level, the tertiary structure of *4E-T* was next predicted. The peptide sequence for *D. eugracilis* 4E-T-PB was used as input to PHYRE2 fold recognition server using default settings. The most significant result aligned to a known structure of *D. melanogaster 4E-T*. The result had 99.7% confidence and the aligned sequence displayed 76% percent identity to the *D. melanogaster* ortholog. However, 83% of the protein structure is disordered (Figure 24A), indicating a predicted structure that is

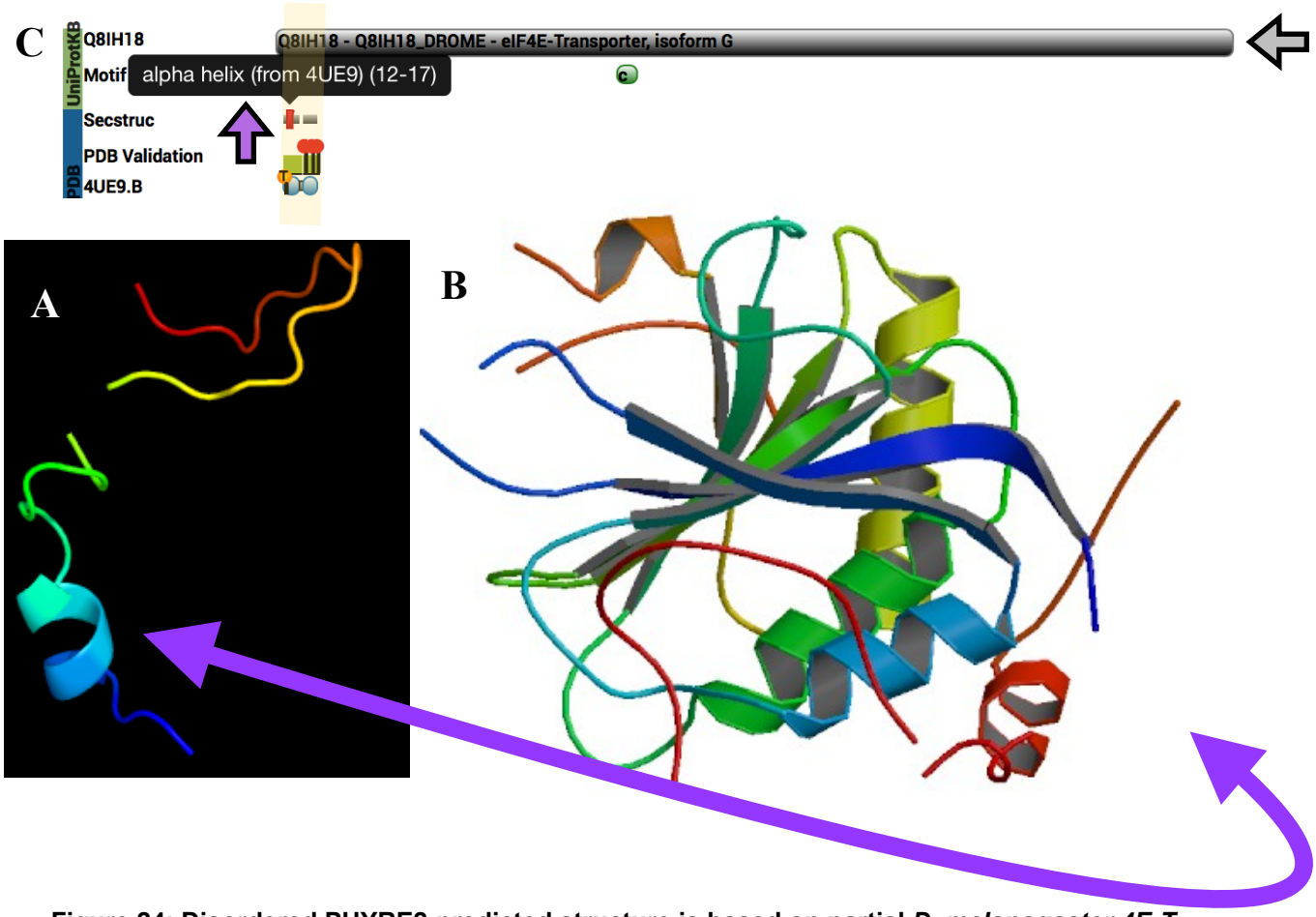


Figure 24: Disordered PHYRE2-predicted structure is based on partial *D. melanogaster 4E-T* structure and reveals *eIF4E* binding interface. A) The PHYRE2-predicted structure for *D. eugracilis 4E-T* is shown. The protein is highly disordered. **B)** The prediction is based on the partial, disordered, *D. melanogaster 4E-T* structure (red helix) that is part of the *eIF4E* crystal structure (PDB ID: 4UE9). The purple arrow shows *4E-T* in the PHYRE2 prediction and PDB structure. **C)** Only a small fraction (yellow shading) of the entire 4E-T coding sequence (grey arrow) is accounted for in the crystal structure. The domain that binds *eIF4E* is an alpha helix composed of residues 12-17 (purple arrow).

biologically and thermodynamically unfavorable. The subject structure utilized for three-dimensional alignment was for *eIF4E* and only a portion of *4E-T* was crystalized. Due to an incomplete source *4E-T* crystal structure, Phyre2 predicted the structure of residues with sequence homology and failed to predict meaningful tertiary structure for the remaining protein (Figure 24B). However, the PDB crystal structure used by Phyre2 (PDB ID: 4UE9) identified residues 12-17 as an *eIF4E* binding interface (Figure 24C). These residues show weak conservation among insect species at the sequence level (Figure 23, purple arrow).

Feature 1 contains the putative *D. eugracilis* ortholog of *mGluR*

A detailed view of Feature 1 is shown in Figure 25. To identify the *D. melanogaster* ortholog for Feature 1, the protein sequence predicted by the Genscan gene predictor was used as query in a blastp search against the FlyBase *D. melanogaster* annotated proteins (AA) database (subject). The blastp search suggests Feature 1 contains the *D. eugracilis* ortholog of *mGluR* (Figure 26). In addition, the blastp search aligned with significance to *mtt*. *mtt* is located on the second chromosome in *D. melanogaster* and *mGluR* is located on the fourth chromosome in *D. melanogaster* (Figure 27A); thus, parsimony suggests the orthologous region in *D. eugracilis* chromosome four contains the *mGluR* ortholog.

To guide subsequent annotation, the *mGluR* gene structure was analyzed. The *D. melanogaster* ortholog is located on the fourth chromosome and contains four isoforms that translate to three distinct protein products (Figure 27B). Isoforms A and B yield identical protein coding regions. mGluR-PC and mGluR-PD contain truncated coding regions relative to the other isoforms (Figure 27A).

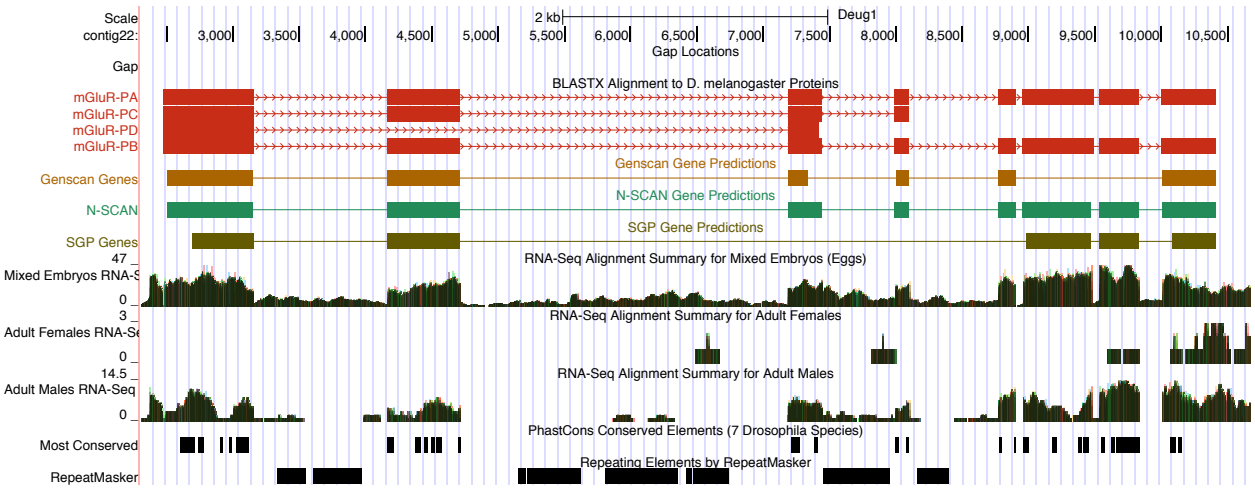
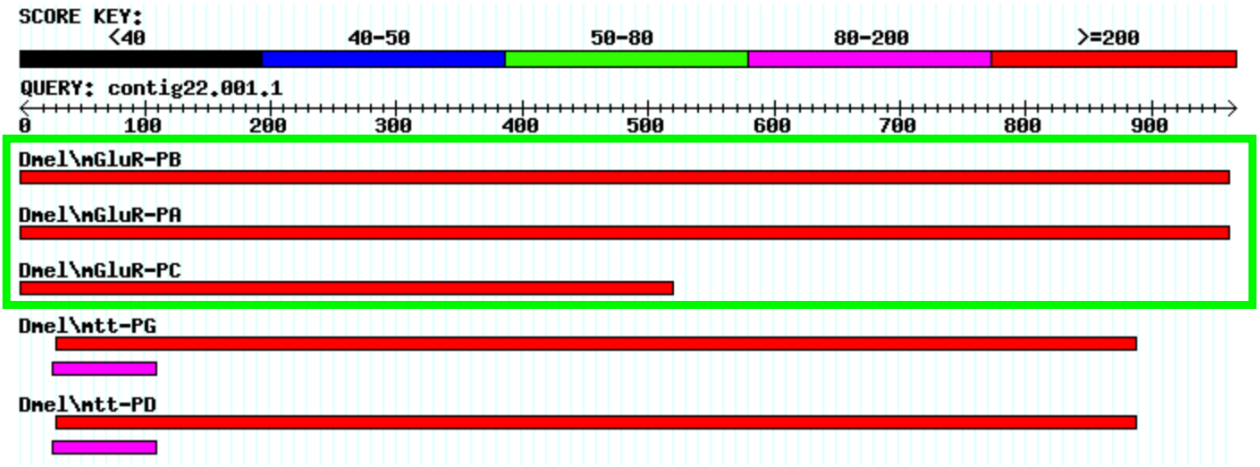


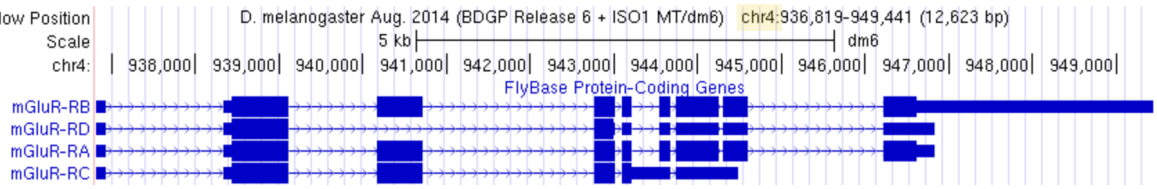
Figure 25: Feature 1 on *D. eugracilis* contig22. Initial blastx alignment suggests four isoforms with between two and eight exons. Gene predictors agree roughly on the placement of coding regions and RNA-seq data overlaps roughly with predicted exons.



BLAST Hit Summary				
☑	Description	Species	Score	E value
☑	mGluR-PB	Dmel	1786.54	0
☑	mGluR-PA	Dmel	1786.54	0
☑	mGluR-PC	Dmel	984.941	0
☑	mtt-PG	Dmel	697.967	0
☑	mtt-PD	Dmel	697.197	0
☑	mtt-PF	Dmel	696.041	0

Figure 26: blastp search using Genscan predicted protein sequence in Feature 1 (query) against the *D. melanogaster* annotated proteins (AA) database (subject). Three isoforms of the same protein, *mGluR*, match with greater score than the next best match. Thus, Feature 1 likely contains an ortholog of *mGluR*.

A



B CDS usage map:

Isoform	1_1692_0	2_1692_2	3_1692_1	4_1692_2	5_1692_2	6_1692_2	7_1692_0	8_1692_0	9_1692_2	10_1692_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					

Isoforms with unique coding exons:

Unique isoform(s) based on coding sequence	Other isoforms with identical coding sequences
mGluR-PB	mGluR-PA
mGluR-PD	
mGluR-PC	

C

FlyBase ID	FlyBase Name	Chr	5' Start	3' End	Strand	Protein ID	Graphical Viewer
FBtr0332858	mGluR-RB	4	936,819	949,441	+	FBpp0305080	View in GBrowse
FBtr0344897	mGluR-RD	4	936,819	946,825	+	FBpp0311208	View in GBrowse
FBtr0089184	mGluR-RA	4	936,819	946,825	+	FBpp0088248	View in GBrowse
FBtr0332859	mGluR-RC	4	936,819	944,486	+	FBpp0305081	View in GBrowse

Figure 27: Gene structure for *D. melanogaster mGluR* provided by FlyBase. A) *mGluR* in *D. melanogaster* is located on the fourth chromosome (F element) (highlighted). There are four isoforms with three distinct coding regions. **B)** The A and B isoforms of *mGluR* produce identical coding sequences. mGluR-H and mGluR-D utilizes a distinct sixth exon. **C)** *D. melanogaster mGluR* utilizes one transcription start site (green arrow).

Approximate CDS locations on contig22 were predicted using the pairwise blastx exon alignment approach as described earlier. The common *D. eugracilis mGluR* initiation methionine is shown in Figure 28. Approximate locations of each exon based on the conservation found by blastx are described in Table 4. CDS boundaries were next refined using RNA-seq data and TopHat junction predictions as described previously. The *mGluR* gene model with exact CDS boundaries is described in Table 5.

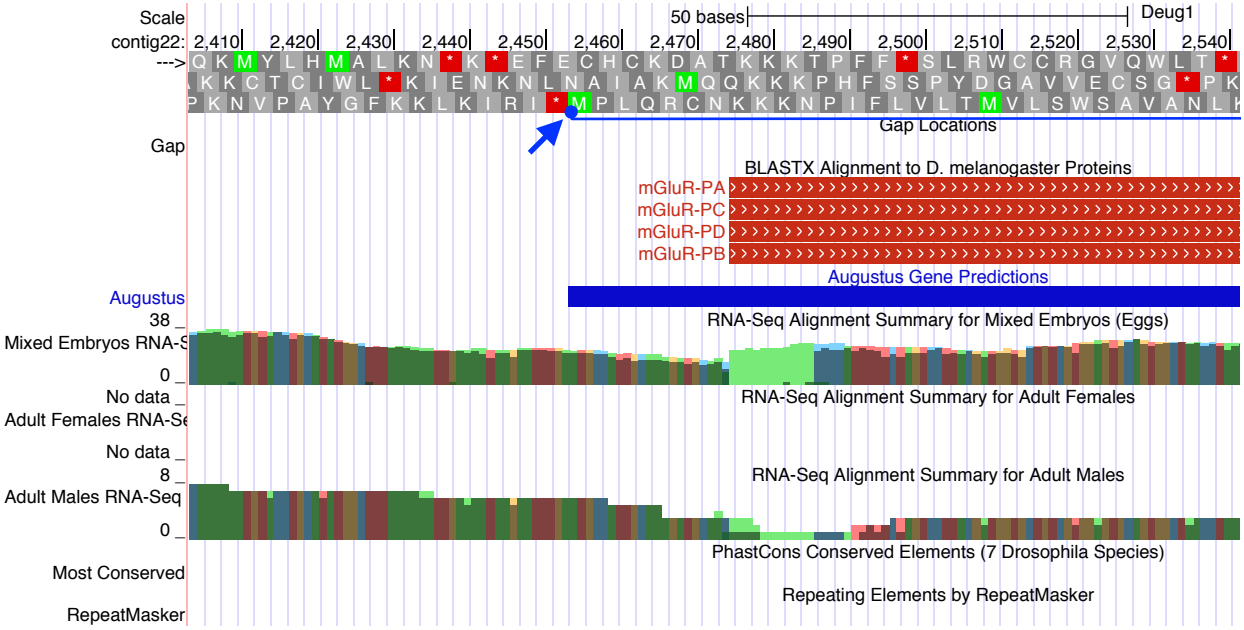


Figure 28: *mGluR* exon 1 initiation methionine. The common first amino acid for putative *mGluR* isoforms is shown in frame +3 with a blue arrow. The Augustus gene predictor corroborates the initial methionine assignment.

FlyBase_ID	Isoform Usage	contig22 start	contig22 end	Frame
mGluR:1_1692_0	A, B, C, D	2454	3153	+3
mGluR:2_1692_2	A, B, C	4169	4712	+1
mGluR:3_1692_1	A, B, C	7189	7442	+2
mGluR:4_1692_2	D	7189	7424	+3
mGluR:5_1692_2	C	7994	8103	+1
mGluR:6_1692_2	A, B	7994	8100	+1
mGluR:7_1692_0	A, B	8775	8900	+3
mGluR:8_1692_0	A, B	8957	9476	+2
mGluR:9_1692_2	A, B	9539	9834	+1
mGluR:10_1692_0	A, B	10011	10415	+3

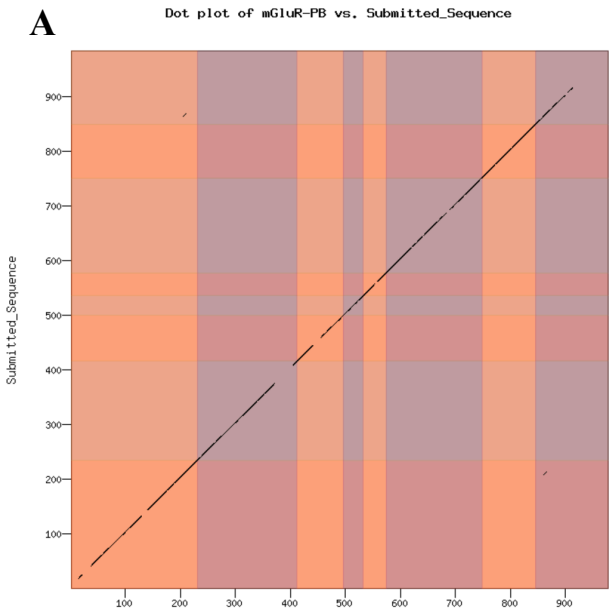
Table 4: blastx alignment maps *D. melanogaster mGluR* coding exons (subject) to contig22 (query). All *D. melanogaster* exons aligned to contig22. These CDS boundary coordinates will be refined by splice site inspection. All *D. melanogaster* exons occupy the positive strand of chromosome four.

	D. eugracilis ortholog								
FlyBase_ID	Exon Number	Isoform Usage	Start	End	Frame	Donor Phase	Acceptor Phase	Exon Size (nt)	*D. mel* Exon Size (nt)
mGluR:1_1692_0	1	A, B, C, D	2454	3153	+3	1	-	699	687
mGluR:2_1692_2	2	A, B, C	4169	4712	+1	2	2	543	543
mGluR:3_1692_1	3	A, B, C	7189	7442	+2	1	1	253	253
mGluR:4_1692_2	4	D	7189	7424	+3	-	2	235	238
mGluR:5_1692_2	5	C	7994	8103	+1	-	2	109	115
mGluR:6_1692_2	6	A, B	7994	8100	+1	0	2	106	106
mGluR:7_1692_0	7	A, B	8775	8900	+3	0	0	125	125
mGluR:8_1692_0	8	A, B	8957	9476	+2	1	0	519	519
mGluR:9_1692_2	9	A, B	9539	9834	+1	0	2	295	295
mGluR:10_1692_0	10	A, B	10011	10415	+3	-	0	404	395

Table 5: Gene model for *D. eugracilis* ortholog of *mGluR*. CDS boundaries were refined by splice site inspection and corroborated by RNA-seq data and Top-Hat predictions. Exon splice site phases are summarized. All putative *D. eugracilis* exons roughly match the size of their respective orthologs.

Gene Model Verification

The gene model for *D. eugracilis mGluR* was verified using the GEP Gene Model Checker. All four isoforms passed validation. The A and B isoforms produce identical proteins so only the mGluR-PB similarity dot-plot and alignment are shown (Figure 29). The mGluR-PC similarity dot-plot and alignment are shown in Figure 30. The mGluR-PD similarity dot-plot and alignment are shown in Figure 31. Regions of mismatch occur in several exons, however, all exon boundaries in all isoforms display conservation upon inspection of the pairwise exon alignment (Figures 29B, 30B and 31B). *mGluR* is conserved with reference to the *D. melanogaster* ortholog and the most parsimonious model for comparative annotation passes validation.



B

mGluR-PB

Identity: 864/984 (87.8%), **Similarity:** 905/984 (92.0%), **Gaps:** 8/984 (0.8%)

```
mGluR-PB          1 MK----QKNNNGTILVVVMVLSWSRVVDLKSPSNTHTQDSVSVSLPGDIILGGLFPVHEK 56
                    *     :*::*  :**:.****** *.:****.*** **:********************
Submitted_Seq     1 MPLQRCNKKKNPIFLVLTMVLSWSAVANLKSPGNTHIQDTVSVSLPGDIILGGLFPVHEK 60
mGluR-PB         57 GEGAPCGPKVYNRGVQRLEAMLYAIDRVNNDPNILPGITIGVHILDTCSRDTYALNQSLQ 116
                    **.*******************************.********************************
Submitted_Seq     61 GESAPCGPKVYNRGVQRLEAMLYAIDRVNNDSNILPGITIGVHILDTCSRDTYALNQSLQ 120
mGluR-PB        117 FVRASLNNLDTSGYECADGSSPQLRKNASSGPVFGVIGGSYSSVSLQVANLLRLFHIPQV 176
                    *********** :**:*.***:**************************************:
Submitted_Seq    121 FVRASLNNLDTSVFECSDASSPKLRKNASSGPVFGVIGGSYSSVSLQVANLLRLFHIPQI 180
mGluR-PB        177 SPASTAKTLSDKTRFDLFARTVPPDTFQSVALVDILKNFNWSYVSTIHSEGSYGEYGIEA 236
                    ***********:**********************************************
Submitted_Seq    181 SPASTAKTLSDKSRFDLFARTVPPDTFQSVALVDILKNFNWSYVSTIHSEGSYGEYGIEA 240
mGluR-PB        237 LHKEATERNVCIAVAEKVPSAADDKVFDSIISKLQKKPNARGVVLFTRAEDARRILQAAK 296
                    :************************** ****************************
Submitted_Seq    241 FHKEATERNVCIAVAEKVPSAADDKVFDLIISKLQKKPNARGVVLFTRAEDARRILQAAK 300
mGluR-PB        297 RANLSQPFHWIASDGWGKQQKLLEGLEDIAEGAITVELQSEIIADFDRYMMQLTPETNQR 356
                    ********************************************** ************ ****
Submitted_Seq    301 RANLSQPFHWIASDGWGKQQKLLEGLEDIAEGAITVELQSEIIEDFDRYMMQLTPGTNQR 360
mGluR-PB        357 NPWFAEYWEDTFNCVLTSLSVKPDTSNSANSTDNKIGVKAKTECDDSYRLSEKVGYEQES 416
                    **************:*:..:* * .:* * .:*:* * *:**** * :*::********:***
Submitted_Seq    361 NPWFAEYWEDTFNCILAPVSEKNNTLNIVDSSDVKTGMKAKTTCIESFRLSEKVGYQQES 420
mGluR-PB        417 KTQFVVDAVYAFAYALHNLHNDRCNTQSDQTTETRKHLQSESVWYRKISTDTKSQACPDM 476
                    ************************ **.***:* *** :********:************
Submitted_Seq    421 KTQFVVDAVYAFAYALHNLHNDRCITQNDQTSEHRKHQHSESVWYRKLSTDTKSQACPDM 480

mGluR-PB        477 ANYDGKEFYNNYLLNVSFIDLAGSEVKFDRQGDGLARYDILNYQRQENSSGYQYKVIGKW 536
                    *******:**************:********************** **************
Submitted_Seq    481 ANYDGKDFYNNYLLNVSFIDLTGSEVKFDRQGDGLARYDILNYQRLENSSGYQYKVIGKW 540
mGluR-PB        537 FNGLQLNSETVVWNKETEQPTSACSLPCEVGMIKKQQGDTCCWICDSCESFEYVYDEFTC 596
                    ****************:*** **************************************
Submitted_Seq    541 FNGLQLNSETVVWNKEAEQPISACSLPCEVGMIKKQQGDTCCWICDSCESFEYVYDEFTC 600
mGluR-PB        597 KDCGPGLWPYADKLSCYALDIQYMKWNSLFALIPMAIAIFGIALTSIVIVLFAKNHDTPI 656
                    ************************:::******* *********:* **************
Submitted_Seq    601 KDCGPGLWPYADKLSCYALDIQYIRWNSLFALTPMAIAIFGIAVTIIVIVLFAKNHDTPI 660
mGluR-PB        657 VRASGRELSYTLLFGILVCYCNTFALIAKPTIGSCVLQRFGIGVGFSIIYSALLTKTNRI 716
                    ******************:******:****: ************************
Submitted_Seq    661 VRASGRELSYTLLFGILVCFCNTFALVAKPTVVSCVLQRFGIGVGFSIIYSALLTKTNRI 720
mGluR-PB        717 SRIFHSASKSAQRLKYISPQSQVVITTSLIATQVLITMIWMVVEPPGTRFYYPDRREVIL 776
                    ************************************:************************ ****
Submitted_Seq    721 SRIFHSASKSAQRLKYISPQSQVVITASLITAQVLITMIWMVVEPPGTRFYYPDRTEVIL 780
mGluR-PB        777 KCKIQDMSFLFSQLYNMILITICTIYAIKTRKIPENFNESKFIGFTMYTTCIIWLAFVPI 836
                    ************************:**************************************
Submitted_Seq    781 KCKIQDMSFLFSQLYNMILITICTVYAIKTRKIPENFNESKFIGFTMYTTCIIWLAFVPI 840
mGluR-PB        837 YFGTGNSYEVQTTTLCISISLSASVALVCLYSPKVYILVFHPDKNVRKLTMNSTVYRRSA 896
                    ************************************************************
Submitted_Seq    841 YFGTGNSYEVQTTTLCISISLSASVALVCLYSPKVYILVFHPDKNVRKLTMNSTVYRRSA 900
mGluR-PB        897 AAVAQGAPTSSGYSRTHAPGTSALTGGAVGTNASSSTLPTQNSPHLDEASAQTNVAHKT- 955
                    ** :* * ********:*.*::. ****:*:.*.:*::***.* :*:* . *:.*.:** 
Submitted_Seq    901 AAGTQVAQTSSGYSRTQAAGSTVPTGGAIGSSACASAIPTQSSHNLNEDPNQASVVNKTS 960

mGluR-PB        956 ---NGEFLPEVGERVEPICHIVNK 976
                        **** .*  * . ** : *:.
Submitted_Seq    961 ESKNGEFPSEERECAGPIYNRVKD 984
```

Figure 29: Dot-plot and pairwise exon alignment produced by Gene Model Checker for mGluR-PB. A) Dot plot shows overall conservation to homologous exons in *D. melanogaster*. **B)** Exon alignment shows regions of mismatch within exons, but conserved residues at exon ends.

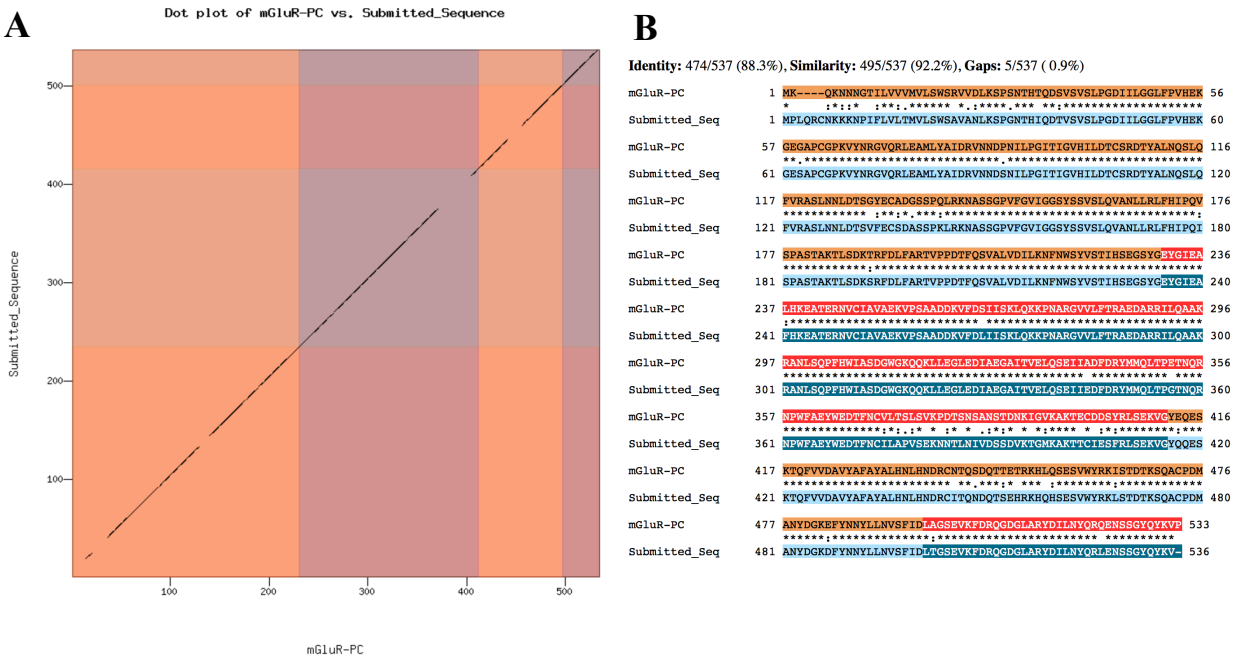


Figure 30: Dot-plot and pairwise exon alignment produced by Gene Model Checker for mGluR-PC. A) Dot plot shows overall conservation to homologous exons in *D. melanogaster*. **B)** Exon alignment shows regions of mismatch within exons, but conserved residues at exon ends.

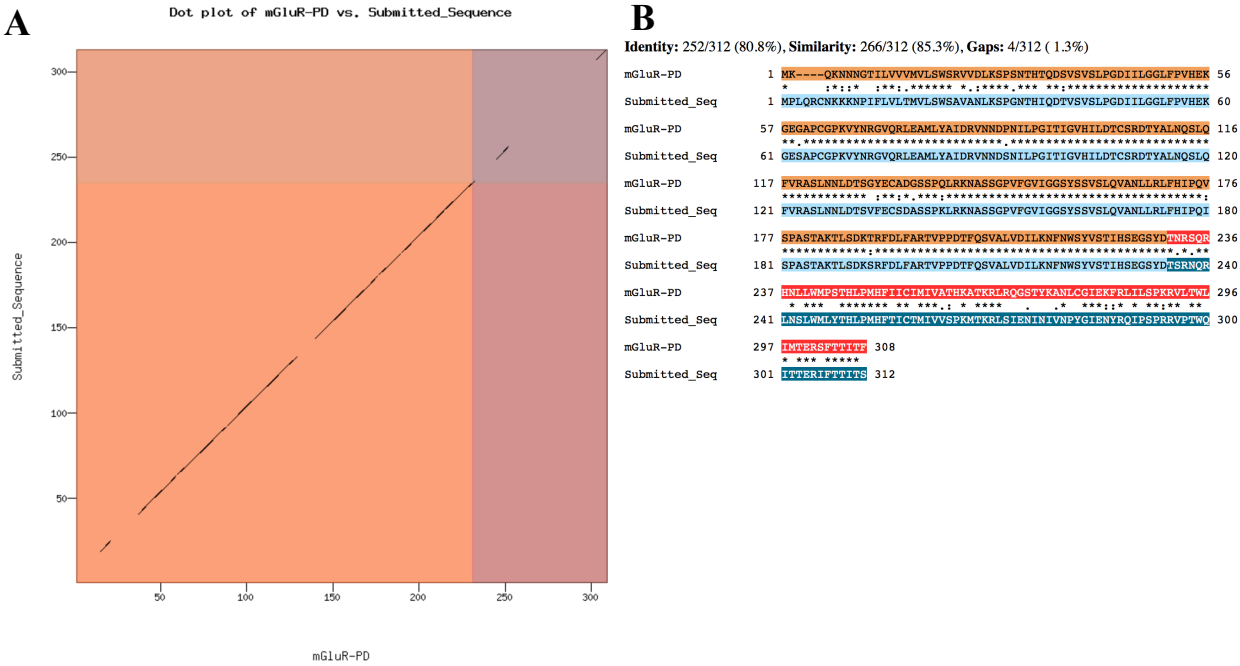


Figure 31: Dot-plot and pairwise exon alignment produced by Gene Model Checker for mGluR-PD. A) Dot plot shows overall conservation to homologous exons in *D. melanogaster*. **B)** Exon alignment shows regions of mismatch within exons, but conserved residues at exon ends.

Transcription Start Site Identification

The TSS for *mGluR* was identified as described previously. In *D. melanogaster*, all four *mGluR* isoforms share a single transcription start site (Figure 27C). Examination of the transcription start region in *D. melanogaster* demonstrates that this locus is heterochromatin-like or heterochromatic in BG3 and S2 cells, respectively, according to the 9-state epigenetic model (Figure 32), indicating a lack of transcription in these cell types. Thus, there are no annotated DHSs, but the DHS read density displays a single broad peak. There is an annotated TSS, and CAGE and RAMPAGE data corroborates a single prominent peak. Thus, the *D. melanogaster* *mGluR* promoter is classified as peaked.

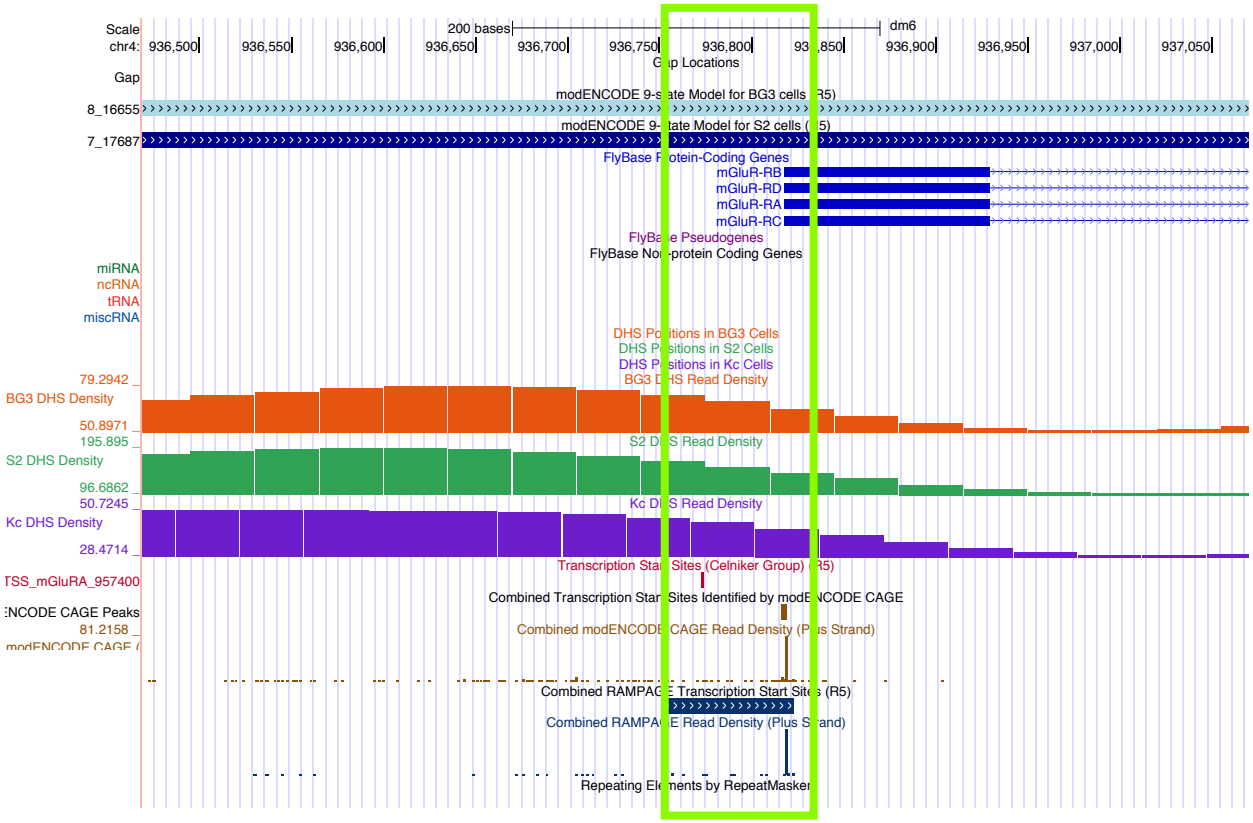


Figure 32: *D. melanogaster mGluR*-UTR and promoter region. The four *mGluR* isoforms are shown in blue. The nine-state epigenetic model defines the promoter region as heterochromatin-like or heterochromatic (light blue and blue tracks). DHS positions, TSSs, CAGE peaks and RAMPAGE peaks all occur around the same region, suggesting a peaked promoter (green box).

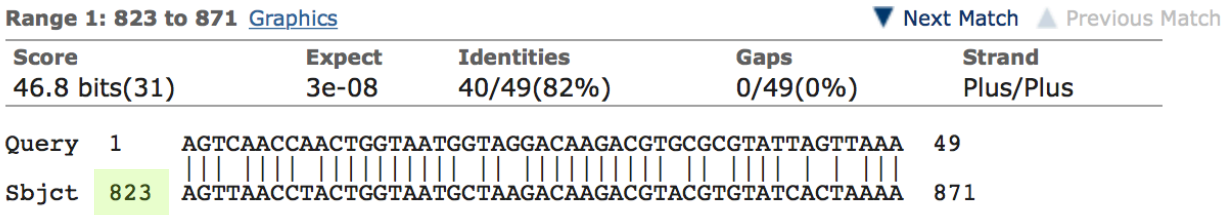


Figure 33: blastn alignment of the *D. melanogaster* mGluR-RB first exon (query) against *D. eugracilis* contig22 (subject). Alignment predicts the transcription start site is at position 823 on contig 22 (green). The query length was 111 nucleotides; of the 49 nucleotides aligned, 82% matched contig22 exactly.

To identify the first exon ortholog in *D. eugracilis*, pairwise blastn alignment was performed for the first mGluR-RB exon as described earlier. The highest-scoring alignment placed the *D. eugracilis* TSS at position 823 on contig22. Exactly 49 nucleotides of 111 nucleotide query sequence (44.14%) aligned to contig 22, and 82% of the aligned region matched exactly (Figure 33).

This TSS placement was refined upon identification of corroborating core promoter motifs. The *D. eugracilis mGluR* core promoter region contains an Inr motif at 821 and an DPE motif at position 850 (Figure 34). In addition, the orthologous region in *D. melanogaster* contains a DPE motif that corroborates the *D. melanogaster mGluR* TSS. Taken together, the evidence suggests *D. eugracilis mGluR* utilizes a TSS at position 823. However, RNA-seq reads extend further upstream from position 823. To account for these reads, a narrow search region was defined from the TSS position to nucleotide 465 on contig 22 (Figure 34). All core promoters motifs within 300 nucleotides of the predicted *4E-T* TSSs in *D. melanogaster* and *D. eugracilis* are described in Table 6.

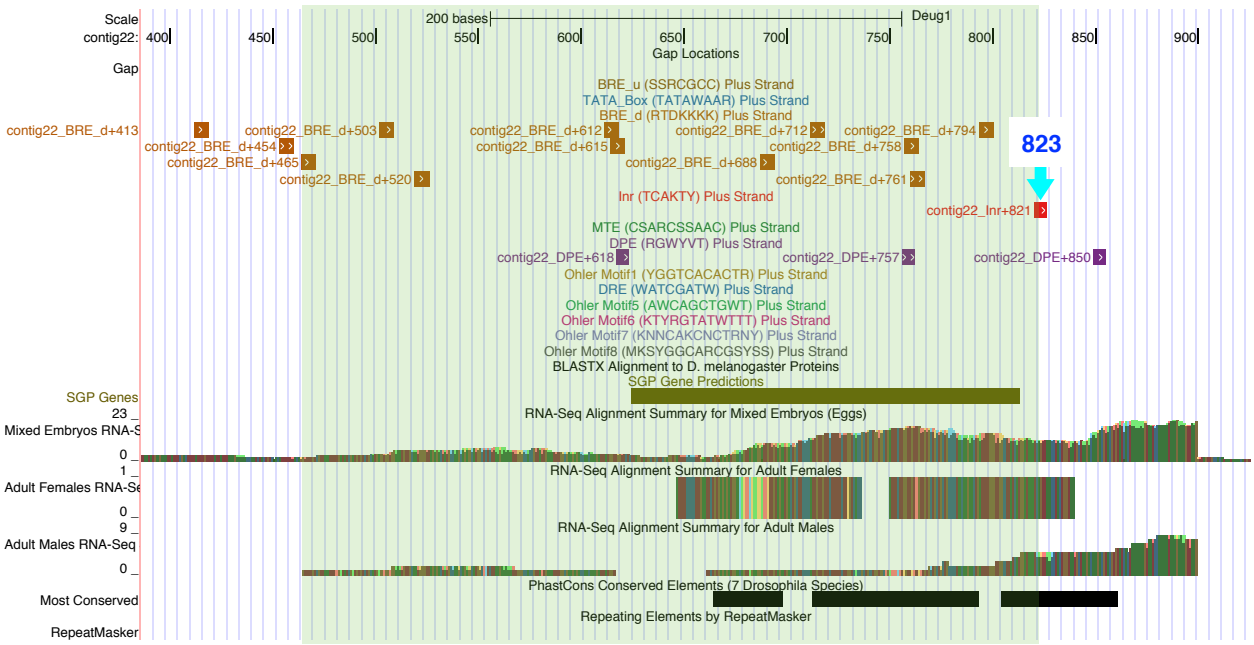


Figure 34: Putative *D. eugracilis mGluR* TSS and narrow search region. An Inr motif is immediately flanking the TSS defined by blastn alignment (cyan arrow). Thus, the TSS is predicted at position 823. A DPE motif at nucleotide 850 is an appropriate distance from the predicted TSS to be function and supports the TSS position. RNA-seq reads extend upstream of the putative TSS. The region from position 823 to 465, where the count of RNA-seq reads drops to one, is defined as the narrow search region (green shading).

Motif	*D. mel.* position	*D. eug.* position
BRE^{d}	936686, 936696, 936706, 936708, 936,754, 936757	794
TATA Box	936868	
Inr	936817	821
DPE	936753, 936846, 936900, 936974, 937018	850

Table 6: Core promoter motifs observed in the 300 nucleotides flanking the *D. melanogaster* annotated *mGluR* TSS and the putative *D. eugracilis mGluR* TSS. The highlighted motifs support the TSS position in the respective species. All motifs lie in the same direction as 4E-T. No instances of the BRE^{u}, MTE, DRE, or Ohler core promoter motifs were observed in the region considered.

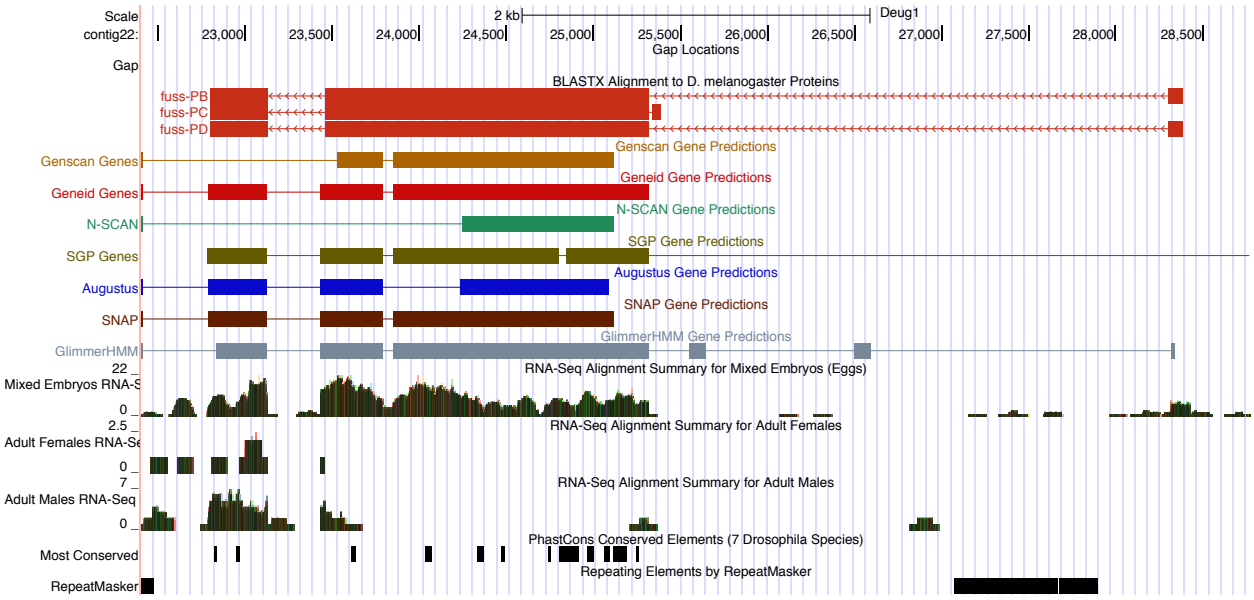
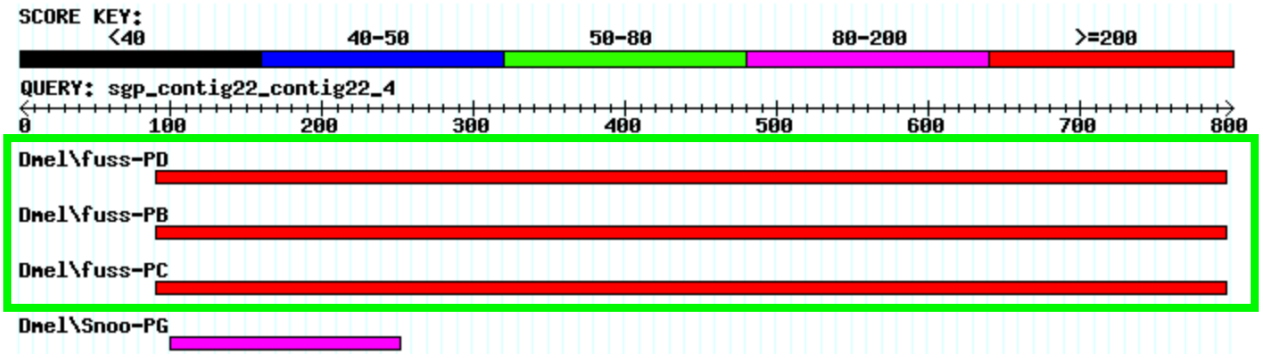


Figure 35: Feature 3 on *D. eugracilis* contig22. Initial blastx alignment suggests three isoforms with three exons each. Gene predictors agree roughly on the placement of coding regions and RNA-seq data overlaps roughly with predicted exons.

Feature 3 contains the putative *D. eugracilis* ortholog of *fuss*

A detailed view of Feature 3 is shown in Figure 35. To identify the *D. melanogaster* ortholog for Feature 3, the protein sequence predicted by the SGP gene predictor was used as query in a blastp search against the FlyBase *D. melanogaster* annotated proteins (AA) database (subject). The blastp search suggests Figure 3 contains the *D. eugracilis* ortholog of *fuss* (Figure 36). To guide subsequent annotation, the *fuss* gene structure was analyzed. The *D. melanogaster* ortholog contains three isoforms that translate to two distinct protein products (Figure 37B). Isoforms B and D yield identical protein coding regions. fuss-PC contains a distinct initial coding exon and is otherwise identical in coding sequence to the other isoforms (Figure 37A).

Approximate CDS locations on contig22 were predicted using the pairwise blastx exon alignment approach described earlier. The common *D. eugracilis* fuss-RB and fuss-RC initiation methionine is shown in Figure 38. The first exon for fuss-PC did not produce any significant blastx alignments; thus the GEP small exons finder tool was used to predict the exon location

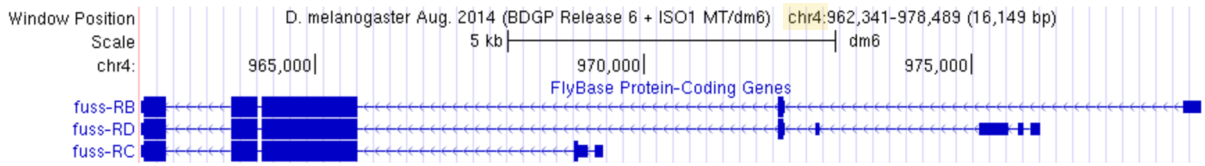


BLAST Hit Summary

	Description	Species	Score	E value
☑	fuss-PD	Dmel	915.605	0
☑	fuss-PB	Dmel	915.605	0
☑	fuss-PC	Dmel	914.064	0
☑	Snoo-PG	Dmel	102.834	1.38157e-21

Figure 36: blastp search using SGP predicted protein sequence in Feature 3 (query) against the *D. melanogaster* annotated proteins (AA) database (subject). Three isoforms of the same protein, *fuss*, match with greater score than the next best match. Thus, Feature 3 likely contains an ortholog of *fuss*.

A



B CDS usage map:

Isoform	1_9572_0	2_9572_0	3_9572_0	4_9572_2	5_9572_0
fuss-PB	1		2	3	4
fuss-PD	1		2	3	4
fuss-PC		1	2	3	4

Isoforms with unique coding exons:

Unique isoform(s) based on coding sequence	Other isoforms with identical coding sequences
fuss-PB	fuss-PD
fuss-PC	

C

FlyBase ID	FlyBase Name	Chr	5' Start	3' End	Strand	Protein ID	Graphical Viewer
FBtr0301335	fuss-RB	4	978,489	962,341	-	FBpp0290550	View in GBrowse
FBtr0334625	fuss-RD	4	976,035	962,341	-	FBpp0306687	View in GBrowse
FBtr0301336	fuss-RC	4	969,374	962,341	-	FBpp0290551	View in GBrowse

Figure 37: Gene structure for *D. melanogaster fuss* provided by FlyBase. A) *fuss* in *D. melanogaster* is located on the fourth chromosome (F element) (highlighted). There are three isoforms with four distinct coding regions each. **B)** The B and D isoforms of *fuss* produce identical coding sequences. fuss-PC utilizes a distinct initial coding exon. **C)** *D. melanogaster fuss* utilizes three transcription start sites (green arrows).

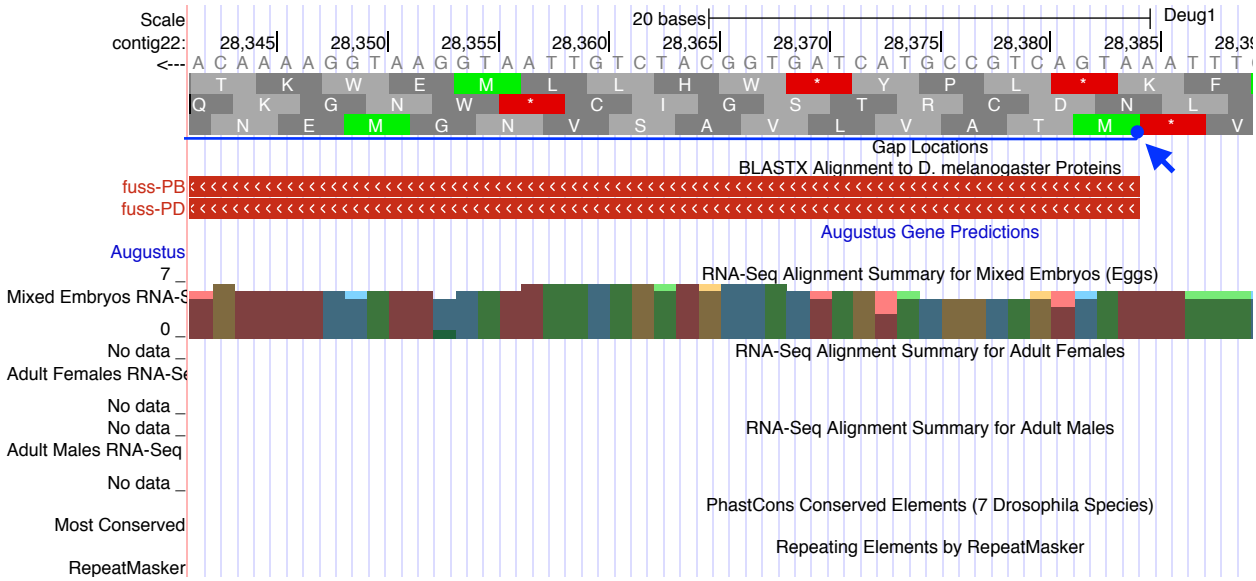


Figure 38: fuss-PB and fuss-PD exon 1 initiation methionine. The common first amino acid for putative *fuss* B and D isoforms is shown in frame -3 with a blue arrow.

within the orthologous *D. eugracilis* region (Figure 39). Small exons finder found a possible fuss-PC exon 1 from 27459 to 27376 on contig22 (Figure 39). There is no substantial RNA-seq or TopHat data to support this assignment and the region lies within a region marked as a repeat by *RepeatMasker* (Figure 40); however, this region remains the most parsimonious assignment for fuss-PC exon 1. In addition, it is possible fuss-PC is not expressed in *D. eugracilis* or only expressed in a narrow, tissue-specific context such that transcript detection was not favorable.

Approximate locations of each exon based on the conservation found by blastx and (in the above case) the small exons finder are described in Table 7. CDS boundaries were next refined using available RNA-seq data and TopHat junction predictions as described previously. Interestingly, *D. melanogaster* contained a non-canonical splice donor site between exons 4 and 5; however, a canonical splice site is observed in *D. eugracilis* at the orthologous splice site. The *fuss* gene model with exact CDS boundaries is described in Table 8.

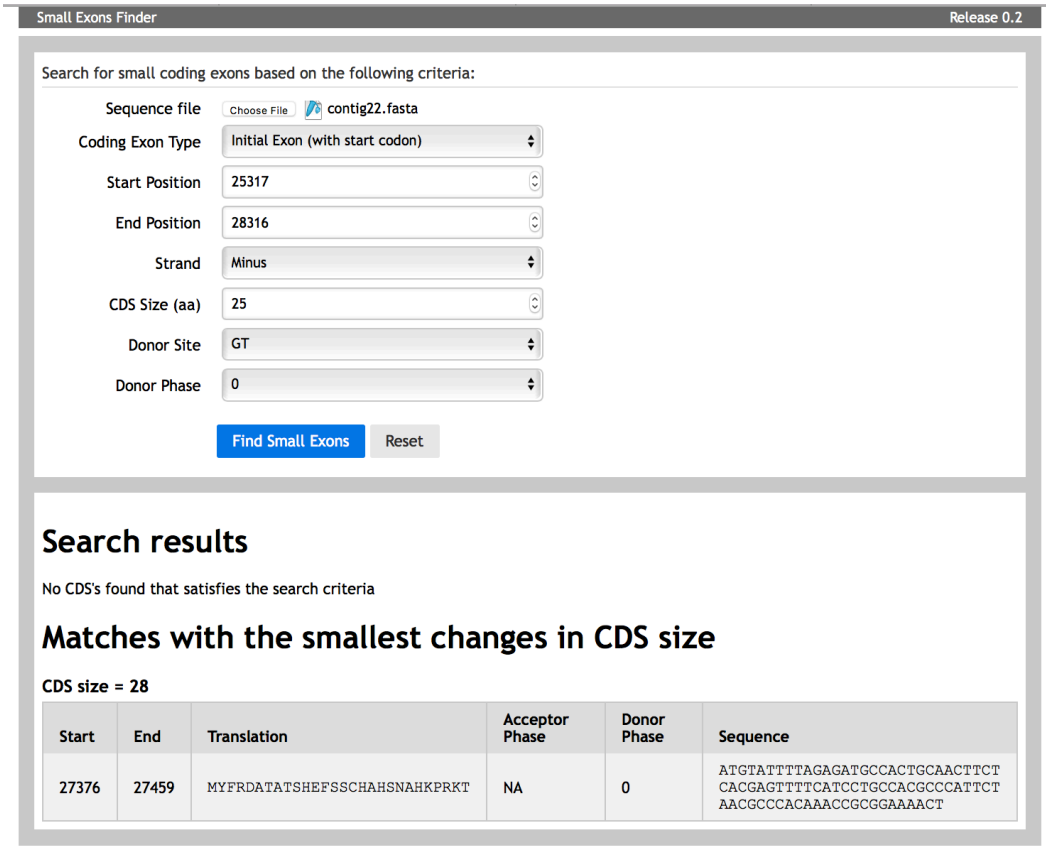


Figure 39: GEP Small Exons Finder input and output for identifying fuss-PC exon 1. The region between the common *fuss* exon 2 and fuss-PB exon 1 was searched for a small exon. One exon was predicted that matched splice phase and predicted size for fuss-PC exon 1. Thus, fuss-PC exon 1 was assigned these coordinates.

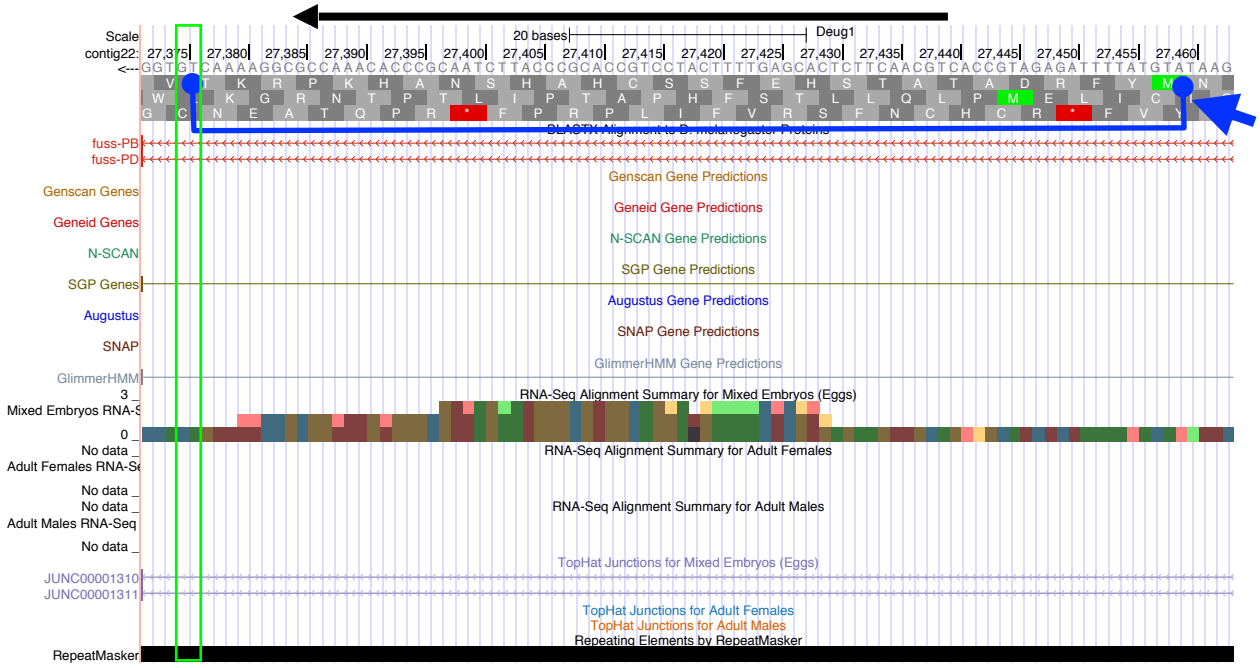


Figure 40: fuss-PC exon 1. The amino acids indicated in blue correspond to the expected size and initial amino acid for fuss-PC exon 1. No RNA-seq reads or TopHat data supports this assignment. The predicted splice donor site is in phase (phase 0) with the fuss-PC-exon 2 splice acceptor site. Splice donor nucleotides are boxed in green. Parsimony suggests this region is likely the orthologous fuss-PC exon 1 in *D. eugracilis*.

FlyBase_ID	Isoform Usage	contig22 start	contig22 end	Frame
fuss:1_9572_0	B, D	28316	28396	-3
fuss:2_9572_0	C	27376	27459	-1
fuss:3_9572_0	B, C, D	23851	25317	-1
fuss:4_9572_2	B, C, D	23464	23775	-1
fuss:5_9572_0	B, C, D	22807	23127	-1

Table 7: blastx alignment and GEP small exon finder maps *D. melanogaster fuss* coding exons (subject) to contig22 (query). All *D. melanogaster* exons aligned to contig22. Isoform 2, the first coding exon for fuss-PC, was identified using the small exon finder tool. These CDS boundary coordinates will be refined by splice site inspection. All *D. melanogaster* exons occupy the positive strand of chromosome four.

	D. eugracilis ortholog								
FlyBase_ID	Exon Number	Isoform Usage	End	Start	Frame	Donor Phase	Acceptor Phase	Exon Size (nt)	*D. mel* Exon Size (nt)
fuss:1_9572_0	1	B, D	28316	28384	-3	0	-	68	92
fuss:2_9572_0	2	C	27376	27459	-1	0	-	83	74
fuss:3_9572_0	3	B, C, D	23850	25317	-1	1	0	1467	1467
fuss:4_9572_2	4	B, C, D	23434	23789	-1	0	2	355	397
fuss:5_9572_0	5	B, C, D	22792	23127	-1	-	0	320	353

Table 8: Gene model for *D. eugracilis* ortholog of *fuss*. CDS boundaries were refined by splice site inspection and corroborated by RNA-seq data and Top-Hat predictions. Exon splice site phases are summarized. *D. melanogaster* contained a non-canonical splice site between exons 4 and 5; however, a canonical splice site is observed in *D. eugracilis* at the orthologous splice site (orange). All putative *D. eugracilis* exons roughly match the size of their respective orthologs.

Gene Model Verification

The gene model for *D. eugracilis fuss* was verified using the GEP Gene Model Checker. All three isoforms passed validation. The B, and D isoforms produce identical proteins so only the fuss-PB similarity dot-plot and alignment are shown (Figure 41). The fuss-PC similarity dot-plot and alignment are shown in Figure 42. Regions of mismatch occur in exons, however, all exon boundaries in fuss-PB display conservation upon inspection of the pairwise exon alignment (Figure 41B). The fuss-PC first exon does not show any identity to the *D. melanogaster* ortholog (Figure 42B). In addition, the *D. melanogaster* fuss-PC first exon shows no sequence homology to any other sequenced *Drosophila* species (not shown). Given the lack of sequence homology, it

is possible that fuss-PC is not expressed in D. eugracilis. Overall, *fuss* is conserved with reference to the *D. melanogaster* ortholog and the most parsimonious model for comparative annotation passes validation.

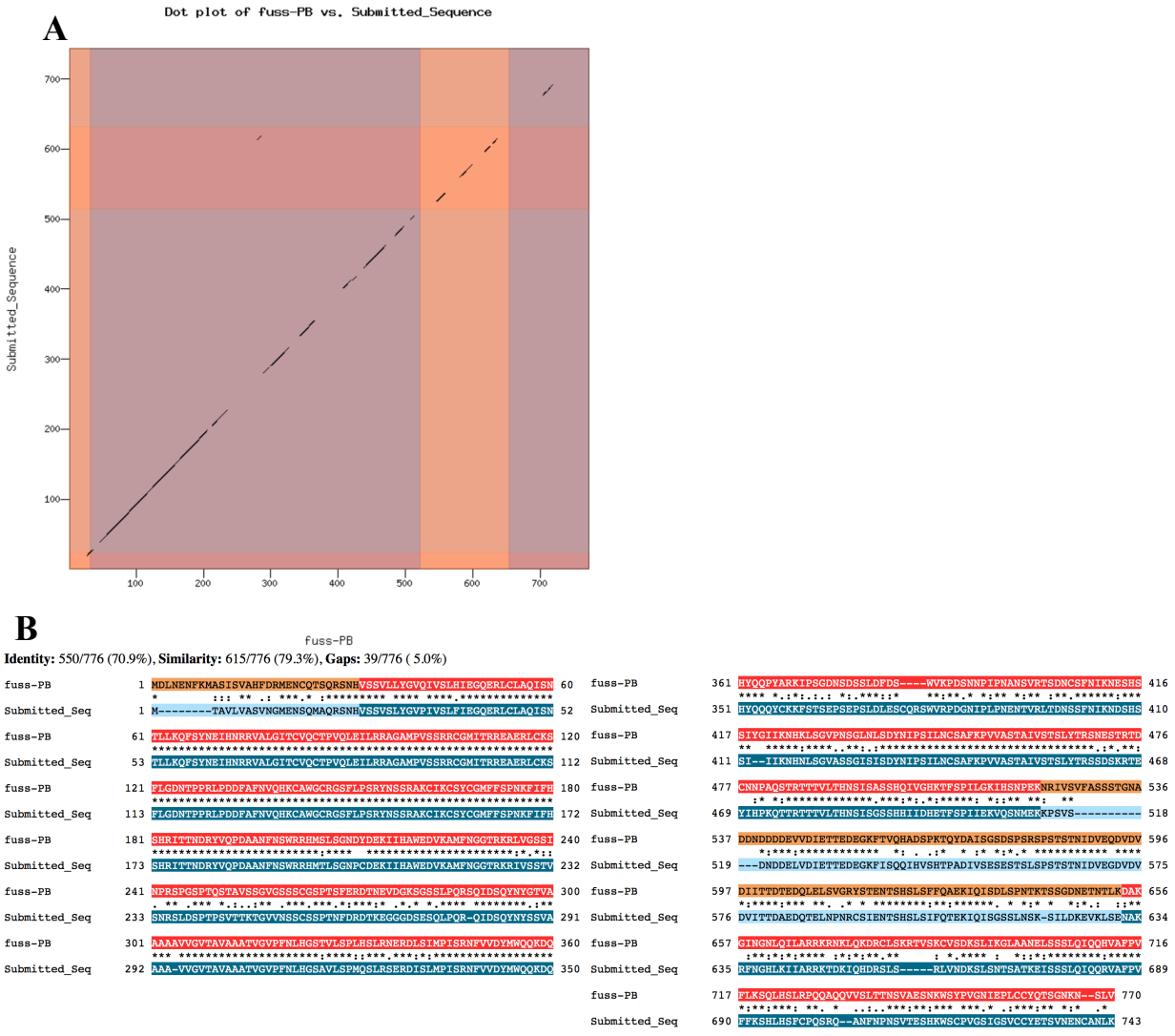


Figure 41: Dot-plot and pairwise exon alignment produced by Gene Model Checker for fuss-PB. A) Dot plot shows overall conservation to homologous exons in *D. melanogaster*. **B)** Exon alignment shows regions of mismatch within exons, but conserved residues at exon ends.

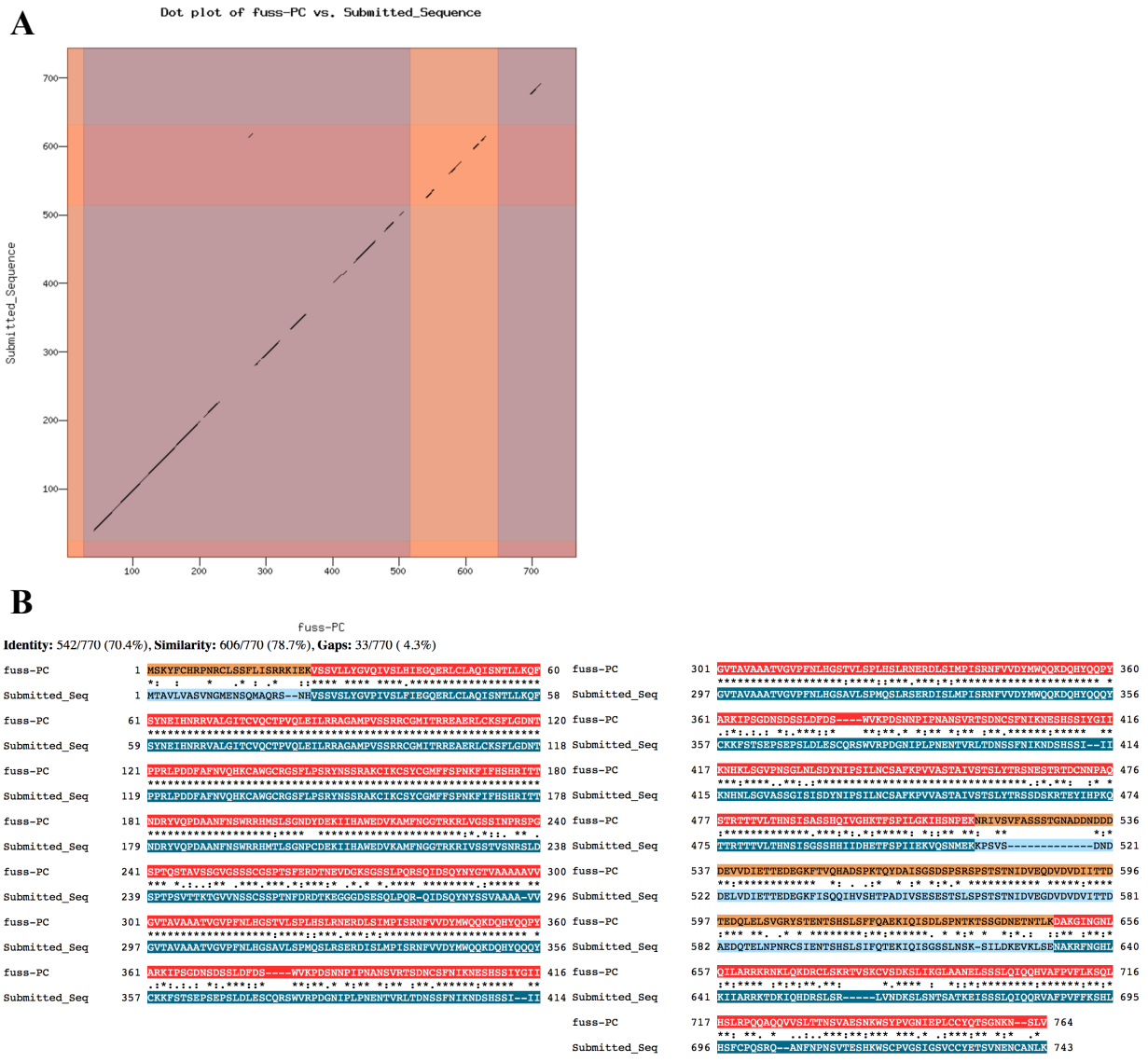


Figure 42: Dot-plot and pairwise exon alignment produced by Gene Model Checker for fuss-PC. **A)** Dot plot shows overall conservation to homologous exons in *D. melanogaster*. **B)** Exon alignment shows regions of mismatch within exons. fuss-RC exon 1 does not show high conservation to the *D. melanogaster* ortholog.

Transcription Start Site Identification

The TSSs for *fuss* was identified as described previously. In *D. melanogaster*, there are three *fuss* isoforms. All isoforms utilize distinct TSSs (Figure 37C).

Examination of the fuss-RB transcription start region in *D. melanogaster* demonstrates that this locus is heterochromatin-like or heterochromatic in BG3 and S2 cells, respectively, according to the 9-state epigenetic model (Figure 43). Thus, there are no annotated DHSs, but the DHS read density displays a single broad peak. There is an annotated TSS and CAGE and RAMPAGE data corroborates a single prominent peak, thus, the *D. melanogaster* promoter is classified as peaked.

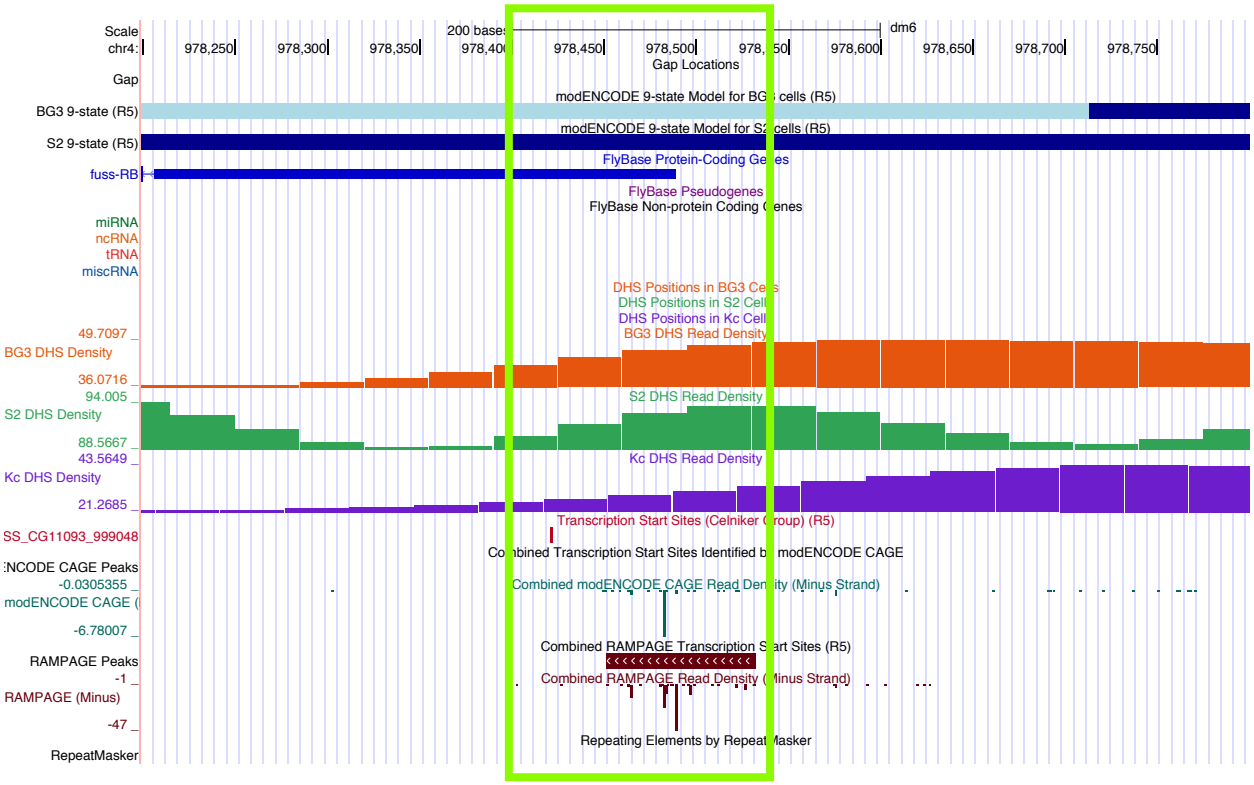


Figure 43: *D. melanogaster* fuss-RB-UTR and promoter region. The fuss-RB isoform is shown in blue. The nine-state epigenetic model defines the promoter region as heterochromatin-like or heterochromatic (light blue and blue tracks) in BG3 and S2 cells, respectively. TSSs, CAGE peaks and RAMPAGE peaks all occur around the same region, suggesting a peak promoter (green box).

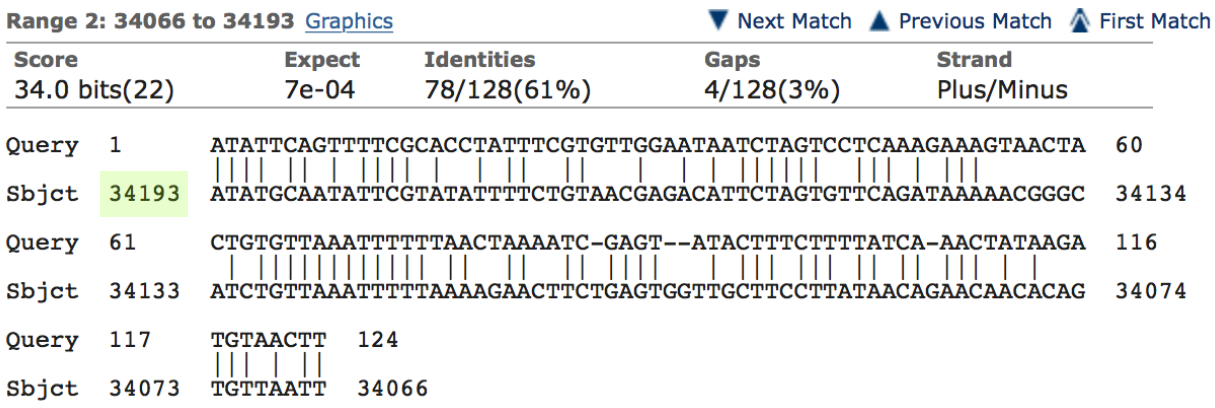


Figure 44: blastn alignment of the *D. melanogaster* fuss-RB first exon (query) against *D. eugracilis* contig22 (subject). Alignment predicts the transcription start site is at position 34193 on contig 22 (green). The query length was 283 nucleotides; of the 128 nucleotides aligned, 61% matched contig22 exactly.

To identify the first exon ortholog for *D. eugracilis* fuss-RB, pairwise blastn alignment was performed as described earlier. The highest-scoring alignment in the expected region on contig22 placed the *D. eugracilis* fuss-RB TSS at position 34193 on contig22. Exactly 128 nucleotides of 283 nucleotide query sequence (45.22%) aligned to contig 22, and 61% of the aligned region matched exactly (Figure 44). This TSS placement was refined upon identification of corroborating core promoter motifs. The *D. eugracilis* fuss-RB core promoter region contains a DPE motif at position 34158 (Figure 45). In addition, the orthologous region in *D. melanogaster* contains a DPE motif that corroborates the *D. melanogaster* fuss-RB TSS. Taken together, the evidence suggests *D. eugracilis* fuss-RB utilizes a TSS at position 34191. However, RNA-seq reads extend further upstream from position 34191. To account for these reads, a narrow search region was defined from the TSS position to nucleotide 34229 on contig 22 (Figure 45). All core promoters motifs within 300 nucleotides of the predicted fuss-RB TSSs in *D. melanogaster* and *D. eugracilis* are described in Table 9.

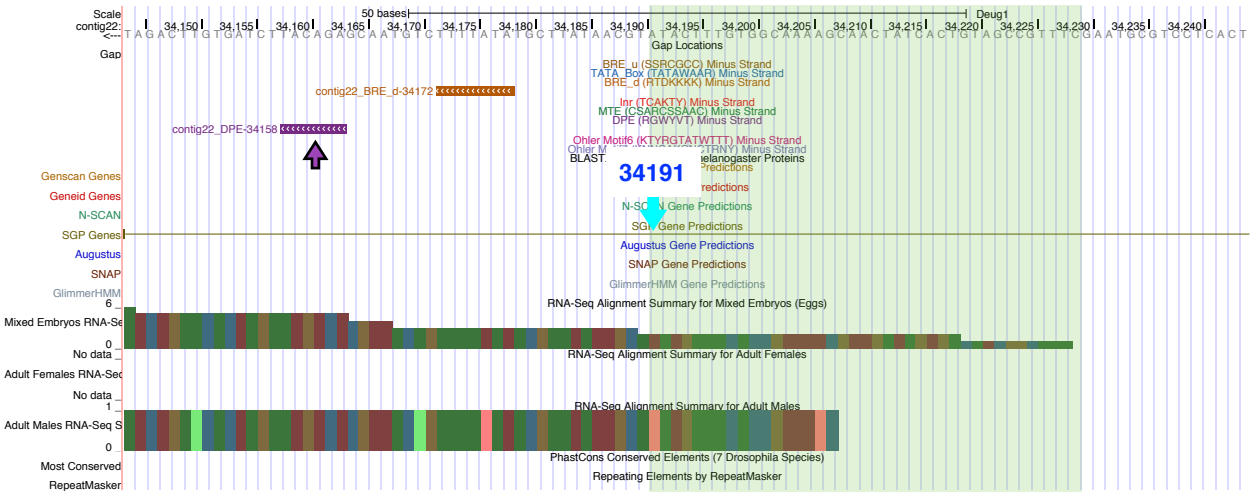


Figure 45: Putative *D. eugracilis fuss-PB* TSS and narrow search region. A DPE motif at nucleotide 34158 (purple arrow) is an appropriate distance from the predicted blastn TSS to be functional and refined the TSS position to position 34191 (cyan arrow). RNA-seq reads extend upstream of the putative TSS. The region from position 34191 to 34229, where the count of RNA-seq reads drops to zero, is defined as the narrow search region (green shading).

Motif	*D. mel.* position	*D. eug.* position
BRE^{d}	978414, 978459, 978493, 978495, 978581, 978628, 978664, 978885	34023, 34053, 34098, 34172, 34306, 34334, 34362, 34384, 34388, 34428
TATA Box	978311, 978338	33929
Inr	978480, 978621, 978667	
DPE	978371, 978445	33975, 34158, 34260

Table 9: Core promoter motifs observed in the 300 nucleotides flanking the *D. melanogaster* annotated fuss-RB TSS and the putative *D. eugracilis* fuss-RB TSS. The highlighted motifs support the TSS position in the respective species. All motifs lie in the same direction as 4E-T. No instances of the BRE^{u}, MTE, DRE, or Ohler core promoter motifs were observed in the region considered.

Turning to fuss-RC, examination of the transcription start region in *D. melanogaster* demonstrates that this locus is heterochromatic in BG3 and S2 cells according to the 9-state epigenetic model (Figure 46). There are no DHSs peaks and there is no peak in DHS read density in this region. An annotated TSS, CAGE data, and RAMPAGE data suggest a single TSS, thus, the *D. melanogaster* fuss-RC promoter is classified as peaked.

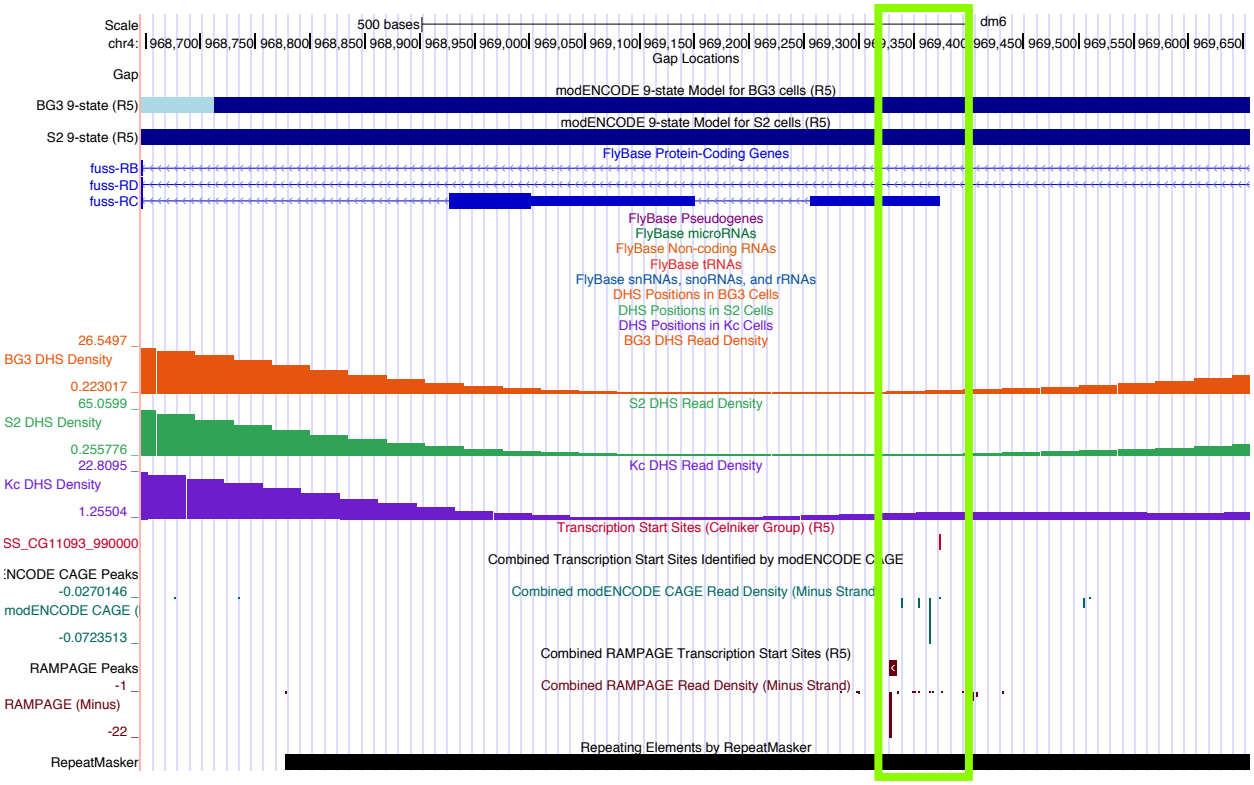


Figure 46: *D. melanogaster fuss-RC*-UTR and promoter region. The fuss-RC isoform is shown in blue. The nine-state epigenetic model defines the promoter region as heterochromatic (blue tracks). TSS annotations, CAGE peaks and RAMPAGE peaks all occur around the same region, suggesting a peaked promoter (green box).

To identify the first exon ortholog for *D. eugracilis* fuss-RC, a pairwise blastn alignment was performed as described earlier. The highest-scoring alignment in the expected region on contig22 placed the *D. eugracilis* fuss-RC TSS at position 28458 on contig22. Exactly 19 nucleotides of 119 nucleotide query sequence (15.97%) aligned to contig 22, and 84% of the aligned region matched exactly (Figure 47). However, the aligned region is a low-complexity region and likely a spurious alignment given its high E-value. No core promoter motifs corroborated the TSS placement and there is a single RNA-seq read initiating from this putative TSS (Figure 48); however, this region is the most parsimonious ortholog of fuss-RC exon 1. No search region is defined for the fuss-RC TSS because no RNA-seq reads extend upstream from the TSS position. All core promoters motifs within 300 nucleotides of the predicted fuss-RC

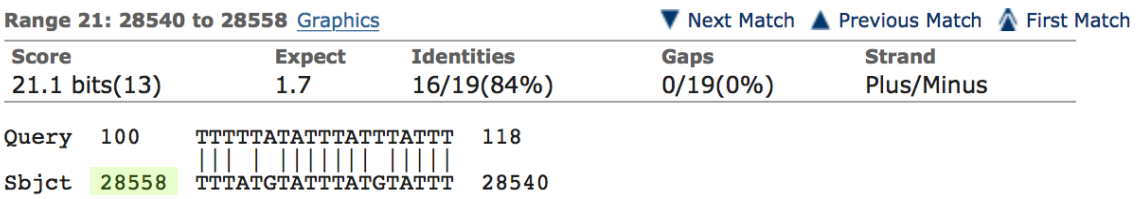


Figure 47: blastn alignment of the *D. melanogaster* fuss-RC first exon (query) against *D. eugracilis* contig22 (subject). Alignment predicts the transcription start site is at position 28458 on contig 22 (green), by extrapolation of the alignment. The query length was 119 nucleotides; of the 19 nucleotides aligned, 84% matched contig22 exactly. While this alignment provides the most parsimonious assignment for the fuss-RC TSS, the aligned sequence is of low-complexity and has a high E-value of 1.7. Therefore, it is probable this alignment is spurious.

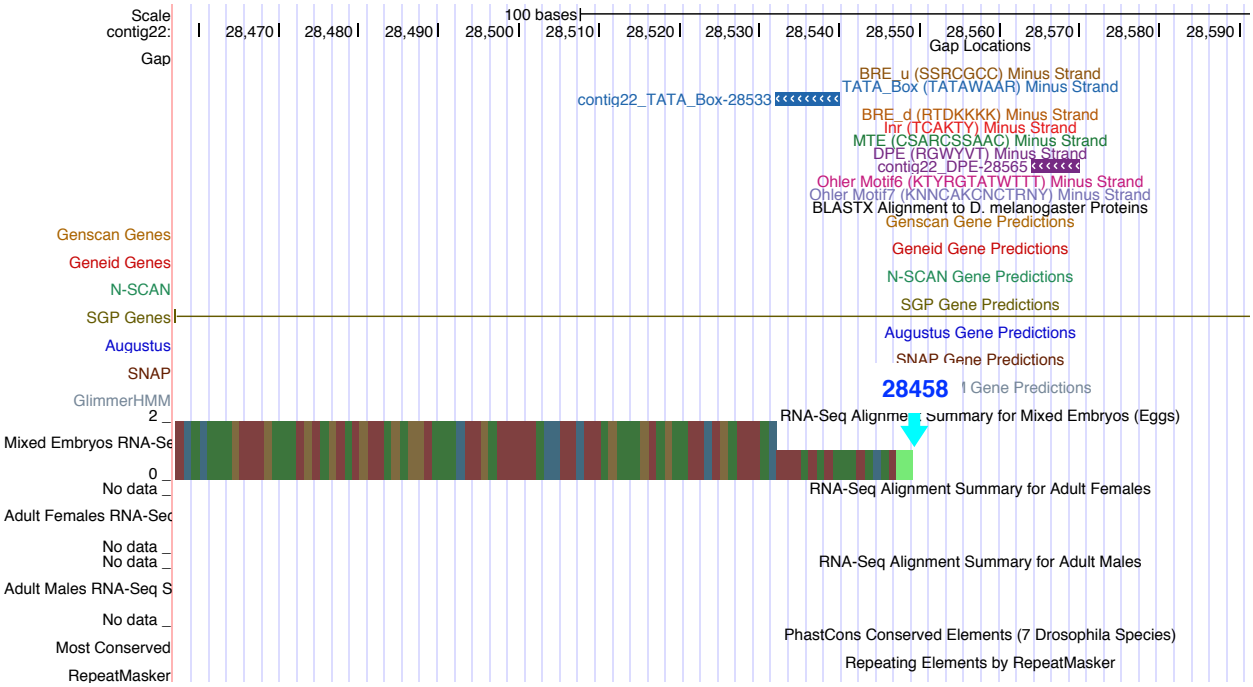


Figure 48: Putative *D. eugracilis fuss-PC* TSS and narrow search region. No core promoter motifs support the predicted TSS at 28458 (cyan arrow). A single RNA-read initiates from the predicted TSS.

Motif	*D. mel.* position	*D. eug.* position
BRE^{d}	969165, 969207, 969209, 969227, 969254, 969272, 969274, 969276, 969278, 969281, 969289, 969321, 969473, 969540, 969542, 969621, 969648, 969650	28161, 28202, 28290, 28295, 28305, 28403, 28425, 28692
TATA Box	969149, 969168	28431, 28533
DPE	969283, 969323, 969398, 969449	28262, 28565, 29680

Table 10: Core promoter motifs observed in the 300 nucleotides flanking the *D. melanogaster* annotated fuss-RC TSS and the putative *D. eugracilis* fuss-RC TSS. No motifs support the TSS position in the respective species. All motifs lie in the same direction as *fuss*. No instances of the Inr, BRE^{u}, MTE, DRE, or Ohler core promoter motifs were observed in the region considered.

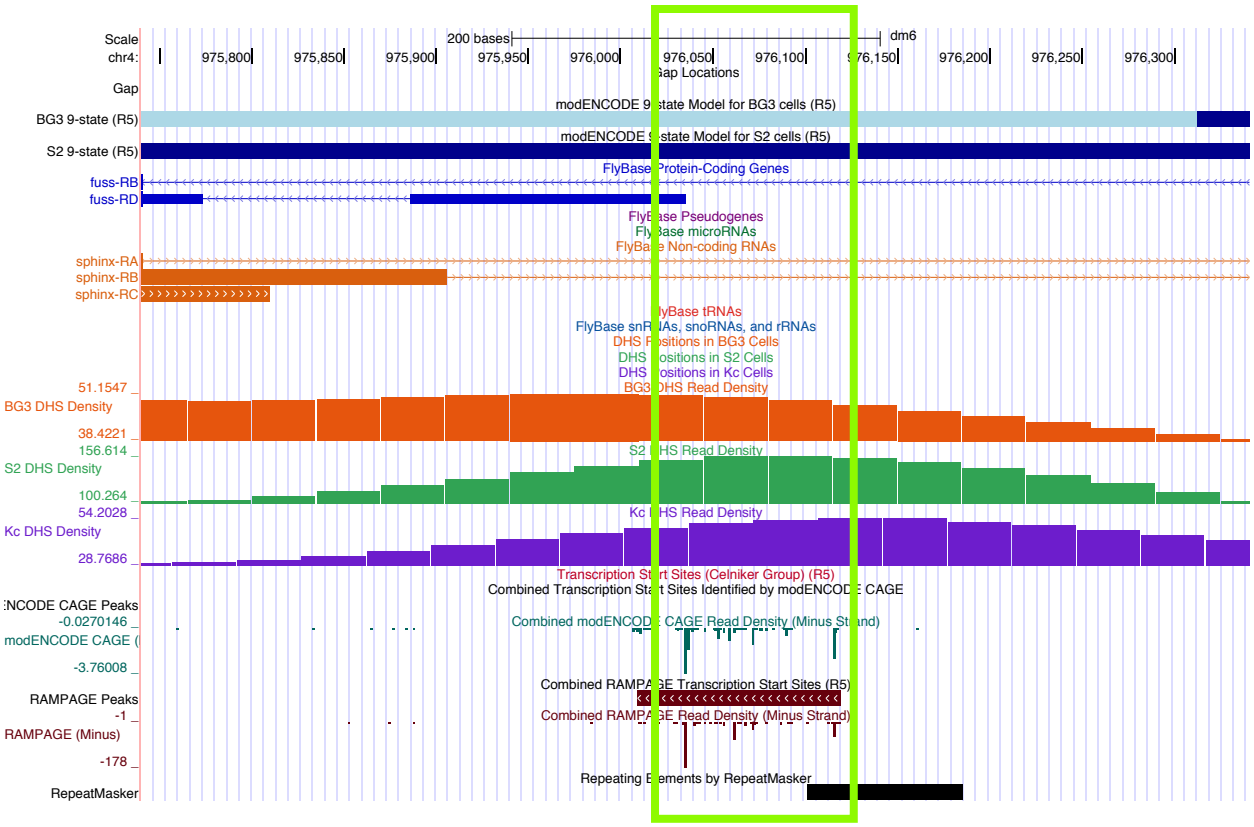


Figure 49: *D. melanogaster fuss-RD*-UTR and promoter region. The fuss-RD isoform is shown in blue. The nine-state epigenetic model defines the promoter region as heterochromatin-like or heterochromatic (light blue and blue tracks). DHS density peaks, CAGE peaks and RAMPAGE peaks all occur around the same region, suggesting a peak promoter (green box). Note the TSS is flanked by a non-coding RNA.

TSSs in *D. melanogaster* and *D. eugracilis* are described in Table 10. Given the weak evidence for a fuss-RC TSS and the lack of confidence in the expression of fuss-PC in *D. eugracilis*, it is possible fuss-RC is not expressed in *D. eugracilis*.

Examination of the fuss-PD transcription start region in *D. melanogaster* demonstrates that this locus is heterochromatin-like or heterochromatic in BG3 and S2 cells, respectively, according to the 9-state epigenetic model (Figure 49). There are no annotated DHS peaks or TSSs, but the DHS read density displays a single broad peak. CAGE and RAMPAGE data corroborates a single prominent peak, thus, the *D. melanogaster* promoter is classified as peaked. Interestingly, the TSS for fuss-RD in *D. melanogaster* lies within an annotated non-coding RNA (ncRNA) on the opposite strand; this will be further discussed in a subsequent section.

To identify the first exon ortholog for *D. eugracilis* fuss-RD, a pairwise blastn alignment was performed as described earlier. There were three alignments to the region between the predicted TSSs for fuss-RB and fuss-RC that displayed correct orientation (Plus/Minus); however, all these alignments were to AT-rich, low-complexity, regions in the contig (Figure 50). These alignments are caused by a low-complexity portion of the exon and demonstrated high E-values, thus, these alignments were discarded as spurious and insignificant. Examination of RNA-seq reads between the TSSs for fuss-RB and fuss-RC did not reveal any stretches of notable transcription not associated with repetitious elements. Taken together, these data suggest there is not an orthologous fuss-RD TSS in *D. eugracilis* and thus the isoform may not be transcribed in *D. eugracilis*. This is evolutionarily plausible, as the coding region produced by isoform D is also produced by isoform B; thus, the production of the functional protein would be retained, although regulation of gene expression might be different.

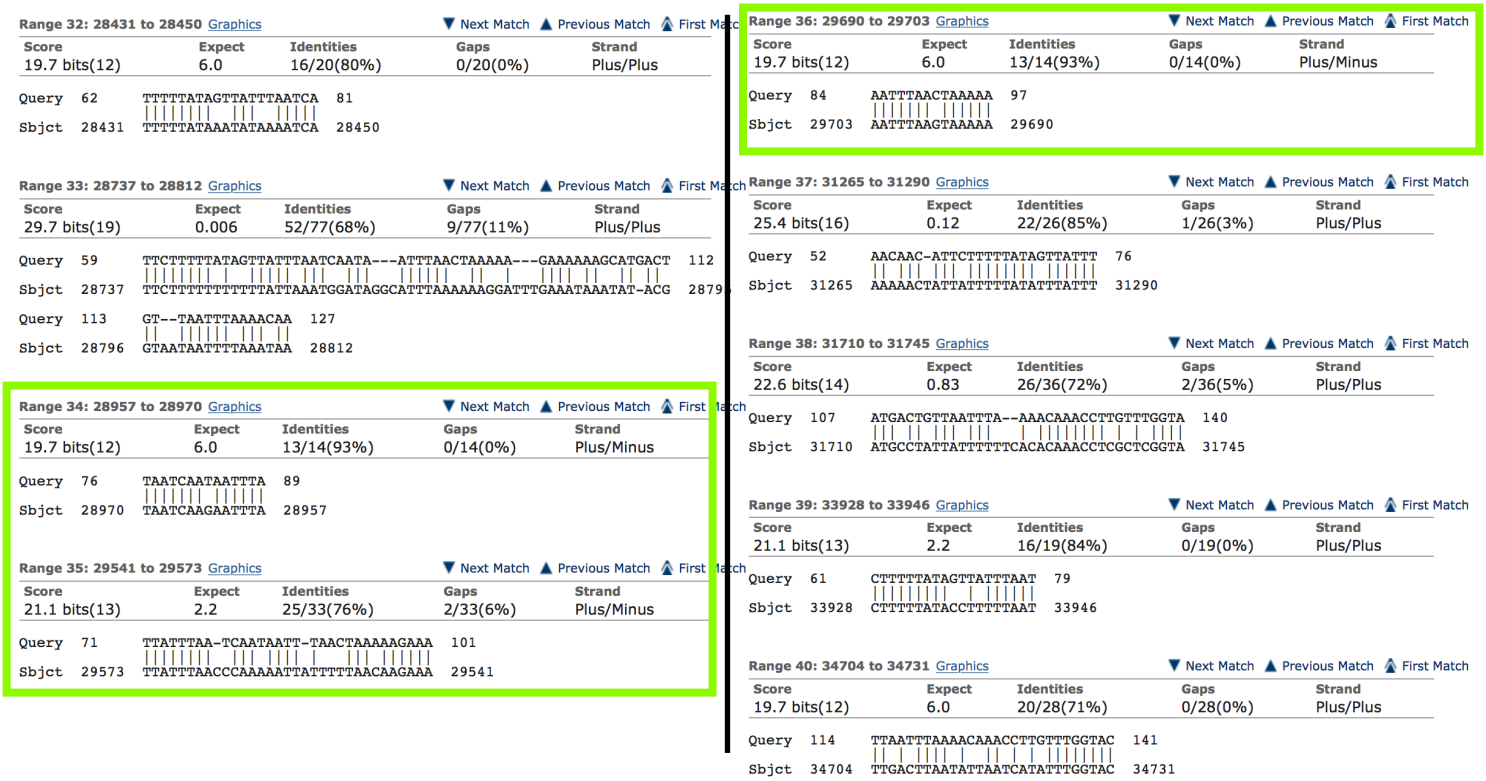


Figure 50: blastn alignments of the *D. melanogaster* fuss-RD first exon (query) against *D. eugracilis* contig22 (subject). All alignments to the region between the B isoform TSS at nucleotide 34191 and the C isoform TSS at nucleotide 28994 are shown. Alignments that display expected strand orientation (Plus/Minus) are boxed in green. Of these three possible alignments, all match only at low-complexity, AT-rich, regions and are likely spurious alignments.

Repetitious Element Analysis

To identify remnants of transposable elements in contig22, the sequence was inspected for repetitious sequences using *RepeatMasker* and a *D. eugracilis*-specific repeat library. *RepeatMasker* output is shown in Figure 51. Repeats larger than 500 nucleotides are likely derived from transposable elements and were of particular interest. These large repeats are summarized in Table 11. Repetitious elements occupy 27.20% (12241/45000) of contig22 (Figure 51). There are eight large repeats that occupy 11.28% (5075/45000) of contig22. These large repeats are 41.46% (5075/12241) of contig22. Qualitatively, there does not appear to be conservation of large repeats between contig22 and the orthologous region in *D. melanogaster* (Figure 52). However, lack of repeat conservation may be due to weaker D. eugracilis sequencing data and sequence assembly errors that incorrectly stack tandem repeats, reducing the quality of the generated repeat library.

```
==================================================
file name: contig22.fasta
sequences:             1
total length:      45000 bp  (45000 bp excl N/X-runs)
GC level:         34.19 %
bases masked:      12241 bp ( 27.20 %)
==================================================
               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:                2          243 bp    0.54 %
      LINE1           0            0 bp    0.00 %
      LINE2           0            0 bp    0.00 %
      L3/CR1          0            0 bp    0.00 %

LTR elements:         0            0 bp    0.00 %
      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:         2          274 bp    0.61 %
     hAT-Charlie      0            0 bp    0.00 %
     TcMar-Tigger     0            0 bp    0.00 %

Unclassified:        48        11736 bp   26.08 %

Total interspersed repeats:    12253 bp   27.23 %


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

Figure 51: *RepeatMasker* output for *D. eugracilis* contig22. *RepeatMasker* identifies 12241 nucleotides of the 45000 nucleotide contig sequence (27.20%) as repetitious.

Contig22 Start	Contig22 End	Repeat Name	Repeat Class	Size (nt)
27073	27663	rnd-3_family-30	RC/Helitron	591
28858	29527	rnd-4_family-237	Unknown	670
31369	31981	rnd-3_family-30	RC/Helitron	613
33117	33738	rnd-4_family-237	Unknown	622
37016	37733	rnd-5_family-291	Unknown	718
38091	38760	rnd-4_family-237	Unknown	670
39208	39791	rnd-4_family-237	Unknown	584
41759	42365	rnd-3_family-30	RC/Helitron	607

Table 11: Repetitious elements greater than 500 nucleotides in contig22. There are eight large repeats in contig22. These repeats are suspected to have been derived from transposable elements.

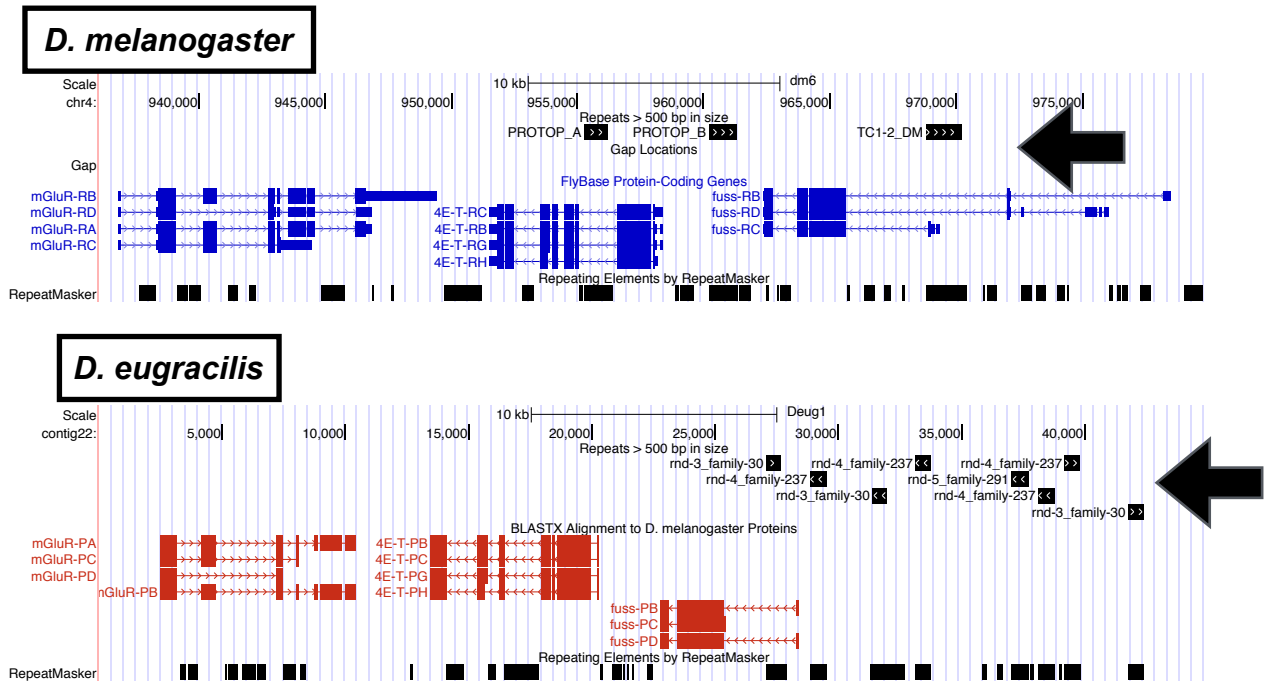


Figure 52: Large repeats in contig22 are distinct from those in the orthologous region in *D. melanogaster*. The three large repeats in *D. melanogaster* and eight large repeats in *D. eugracilis* are shown (black arrows). There is no apparent overlap in location relative to coding genes for *D. melanogaster* repeats.

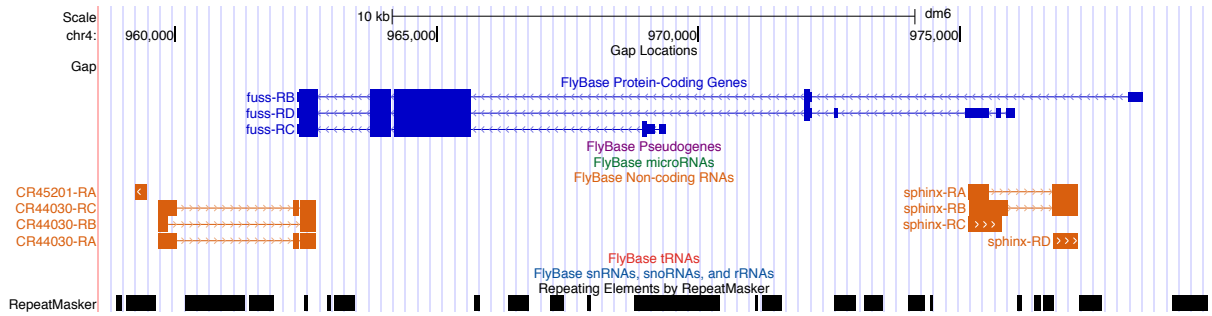


Figure 53: ncRNAs in *D. melanogaster* at the orthologous region to contig22. The orthologous region to contig22 in *D. melanogaster* contains three ncRNAs, *CR45201*, *CR44030*, and *sphinx*. *CR44030* share its final exon with the *fuss* final coding exon but is on the opposite strand. *sphinx* exons flank the annotated TSS for fuss-RD.

Non-coding RNA conservation

The orthologous region to contig22 in *D. melanogaster* contains three annotated ncRNAs, *CR45201*, *CR44030*, and *sphinx* (Figure 53). Interestingly, *CR44030* shares its final exon with the final exon of *fuss*, but utilizes the opposite strand. The *sphinx* sequence overlaps with the TSS for fuss-RD, but is on the opposite strand. To determine if contig22 contains ncRNA orthologs, the FlyBase sequence for each ncRNA was used as query for blastn alignment with contig22 (subject) using the sensitive blast parameters as described earlier. The longest isoform for CR44030 and for sphinx were used as queries.

CR45201 produced no significant alignments in the expected region (downstream of *fuss*) on contig22. *CR44030* only aligned at the coding region for the *fuss* final exon. This is expected, as the *fuss* coding region showed conservation. The *CR44030* exon 1, which does not overlap with *fuss*, was used as separate query for blastn alignment and produced no significant alignments. *sphinx* did not produce significant alignments to the orthologous region on contig22, between the TSS for fuss-RB and fuss-RC. This is unsurprising, as no TSS for fuss-RD was identified in *D. eugracilis*. It is possible that the *sphinx* ncRNA in *D. melanogaster* facilitated transcription initiation from a unique TSS that made possible the fuss-RD isoform. Overall, no

ncRNAs observed in *D. melanogaster* are observed in *D. eugracilis*. This does not mean ncRNA functional orthologs are not present, however, as ncRNAs often demonstrate conservation at the structural level, not the sequence level. Further analysis is necessary to determine if any functional orthologs to the ncRNAs *CR45201*, *CR44030*, and *sphinx* exist in *D. eugracilis*.

Synteny

The region orthologous to contig22 in *D. melanogaster* was inspected to determine synteny to *D. eugracilis*. Homologs to the ncRNAs observed in *D. melanogaster* were not observed in *D. eugracilis*. With respect to the contig22 protein coding genes only, *D. eugracilis* is syntenic with *D. melanogaster* (Figure 54).

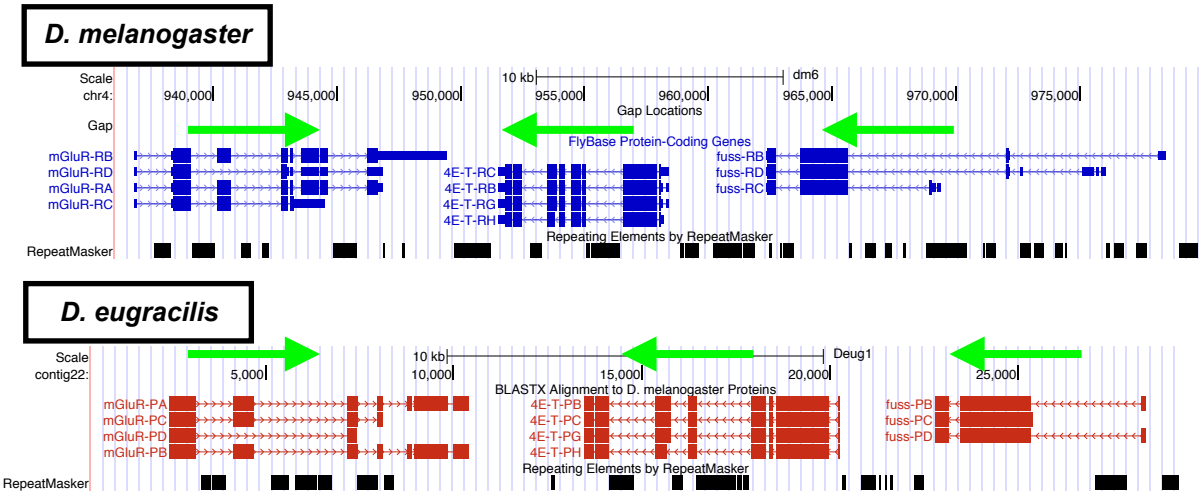


Figure 54: Contig22 demonstrates synteny with the orthologous region in *D. melanogaster*. Gene directionality is indicated with green arrows. Both *D. melanogaster* and *D. eugracilis* show identical ordering and directionality of *mGluR*, *4E-T*, and *fuss*. Both species display repetitious sequences distributed in intergenic and intronic regions.

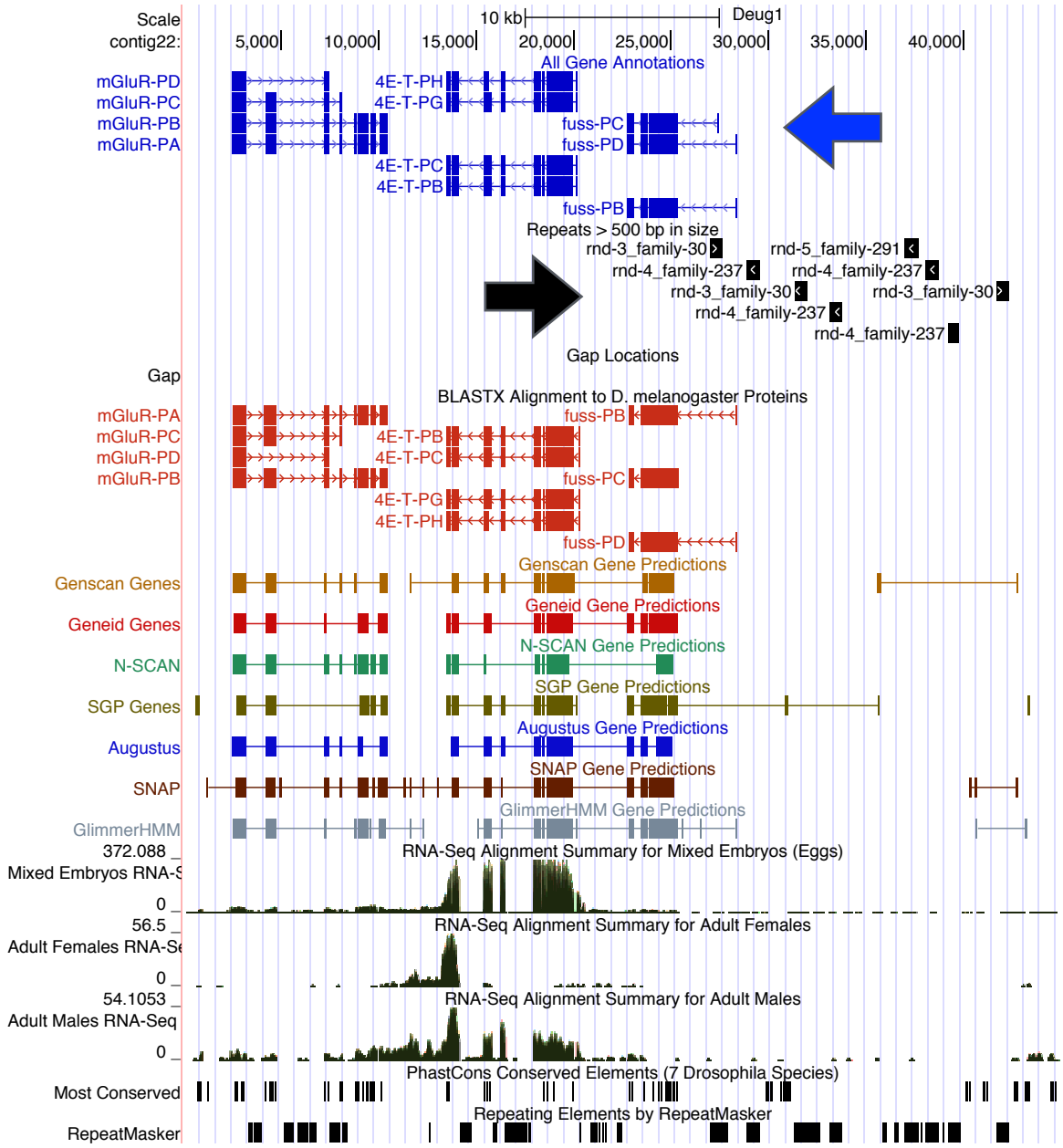


Figure 55: Contig22 final annotation. Annotated genes are displayed as blue tracks (blue arrow)and overlap with blastx predictions. Large repeats (>500 nucleotides) are displayed below the annotation track (black arrow).

Discussion

The final annotation for *D. eugracilis* contig22 is shown in Figure 55. Contig22 contains three putative genes: *mGluR*, *4E*-T, and *fuss*. Isoforms and exons orthologous to *D. melanogaster* were identified for each respective gene. Gene models for each putative *D. eugracilis* ortholog

passed validation by Gene Model Checker. The TSS for 4E-T-PB/C/G was identified at position 20582. No evidence was available to identify the 4E-T-PH TSS. The TSS for all *mGluR* isoforms was identified at position 823. The TSS for fuss-PB was identified at position 34191. Parsimony suggested the TSS for fuss-PC lies at position 28458; however, the evidence supporting fuss-PC annotation is weak and the isoform is likely not expressed in *D. eugracilis*. No TSS could be identified for fuss-PD. However, the absence of evidence for the 4E-T-PH and fuss-PD TSSs does necessarily mean the isoforms are not expressed in *D. eugracilis*. Further experiments must determine if these orthologs are expressed in *D. eugracilis*. *Clustal Omega* multiple sequence alignment of *4E-T* with insect species homologs revealed regions of striking conservation among otherwise divergent amino acid sequences. Further mutagenesis experiments are necessary to characterize the functional relevance, if any, of these conserved domains in *4E-T*. Orthologs to ncRNAs observed in *D. melanogaster* were not identified in contig22. However, since ncRNA function is determined at the secondary structure level, additional analysis is necessary to confirm the lack of orthologous ncRNAs in contig22. Contig22 displayed synteny in coding-gene organization to the reference *D. melanogaster* region. Annotation of contig22 is complete and can be utilized to guide annotation of the total *D. eugracilis* F element. Future comparative genomics and phylogenetic footprinting approaches will incorporate annotation data produced here to better understand *Drosophila* F element gene regulation.

Appendix

Files produced by Gene Model Checker for each isoform (FASTA, PEP, and GFF files) are submitted electronically.

Acknowledgements

I am grateful for the guidance and assistance provided by Dr. Sarah Elgin, Dr. Chris Shaffer, Wilson Leung, Ryan Friedman, and Emily Chi. I am additionally grateful to Dr. April Bednarski for aiding in my improvement as a writer. Finally, I am appreciative of the Genomics Education Partnership and Washington University in St. Louis for supporting my participation in this project.