Ben French

Bio 434W

Dr. Elgin

Annotation of contig12 in *Drosophila eugracilis*

Abstract

Genes found in the F element of *Drosophila* are transcriptionally active despite their presence in a heterochromatic environment. A comparative genomics approach may be used to identify features of the transcription start sites and associated regulatory regions of these genes that allow for expression in heterochromatin. In order to perform these comparative analyses, careful manual annotations of coding spans and transcription start sites of F element genes and genes found in euchromatic reference regions in the D element are required. To scale up the annotation process, these genomic regions are broken into projects, or contigs, and assigned to students for annotation of genes. This report consists of the annotation of the coding spans (CDSs) of all the genes found in contig12 of *Drosophila eugracilis,* and the annotation of the TSSs of *CG33521, PIP4K,* and *Mitf*. Annotation of the CDSs of contig12 genes involved the consolidation of several lines of evidence, including conservation of amino acid sequences to *D. melanogaster*, RNA-Seq data, and computer-based gene predictors. This data was used to produce the most parsimonious annotation of the CDSs of these genes. Likewise, annotation of the genes' TSSs also involved the consolidation of multiple lines of evidence, including comparison of nucleotide sequences with the initial transcribed exon of *D. melanogaster*, RNA-Seq data, TopHat splice junction predictions, and potential core promoter motifs. Using this evidence, putative TSSs were identified, and search regions were defined where evidence was

ambiguous. Annotations of the F and D elements will be used in a search for regulatory motifs by phylogenetic footprinting.

Introduction

In eukaryotes, DNA is packaged in two distinct forms: euchromatin and heterochromatin. The former is loosely packaged in nucleosome arrays and rich in expressed genes. The latter uses a relatively condensed nucleosome array and tends to be gene poor. This model of chromosome organization appears to be used to silence unwanted parts of the genome (*e.g.* repetitious elements) while ensuring that important genes are available to the transcriptional machinery. Like other eukaryotic organisms, *Drosophila* use this system of euchromatin and heterochromatin to organize their genomes. However, the dot chromosome (Muller F element) of *Drosophila* appears to violate previous assumptions about the purpose of heterochromatin. This small chromosome appears to be almost entirely heterochromatic, yet it contains approximately 80 genes that are expressed at levels comparable to those of euchromatic genes. Furthermore, heterochromatic genes appear to be specifically adapted to operate in a heterochromatic environment, as they are often poorly expressed when moved into euchromatic regions (Elgin and Reuter 2013).

To understand this phenomenon in the *Drosophila* F element, the Genomics Education Partnership (GEP) is improving selected genomic regions of several *Drosophila* species and manually annotating the genes therein to perform comparative genomic analyses. This comparative genomic analysis (phylogenetic footprinting) will be used to reveal more information about the evolutionary history of these F element genes as well as to search for conserved motifs that may contribute to their ability to remain expressed in a heterochromatic environment. To obtain annotations of F element and euchromatic reference D element genes, a

two-step protocol is followed. First, the chromosomal regions to be annotated must be "finished." That is, the consensus assembled genomic sequence of the region must be manually checked for inconsistencies such as misassembiles or gaps, and improved where possible. Second, the coding spans and TSSs in these "finished" sequences must be carefully annotated. More information on this project can be found at www.gep.wustl.edu. Most of the evidence used for gene annotation is found in the GEP mirror of the UCSC (University of California, Santa Cruz) Genome Browser. This report covers the annotation of the CDSs of all five genes found in contig12 of the *D. eugracilis* F element: *CG33521, PIP4K, Mitf, Arf102F,* and *Dyrk3* and the annotation of the TSSs of *CG33521, PIP4K,* and *Mitf.*

CDS annotation is a multi-step process that involves the identification of the ortholog in *Drosophila melanogaster*, the reference species; an exon-by-exon annotation of the coding exons; determination of splice sites; and a confirmation of the proposed gene model using the GEP Gene Model Checker, which includes a comparison to the *D. melanogaster* ortholog. Similarly, TSS annotation is a multi-step process that requires annotators to perform several different searches for evidence, including sequence similarity between the untranslated exons of *Drosophila melanogaster* and the *D. eugracilis* contig, placement of untranslated exons using RNA-Seq data, and searching for core promoter motifs. This CDS and TSS annotation workflow uses parsimony to genes in *D. melanogaster* and expression data (in the form of RNA-Seq data from *D. eugracilis*) as evidence to create specific models of *D. eugracilis* coding exons and TSSs.

CG33521

Identification of the Ortholog

The project (contig12) is shown in Fig. 1. BLASTx (Basic Local Alignment Search Tool) alignments suggest that there are five genes in this project (designated as Features 1-5). The feature in question appears to correspond to *CG33521*.

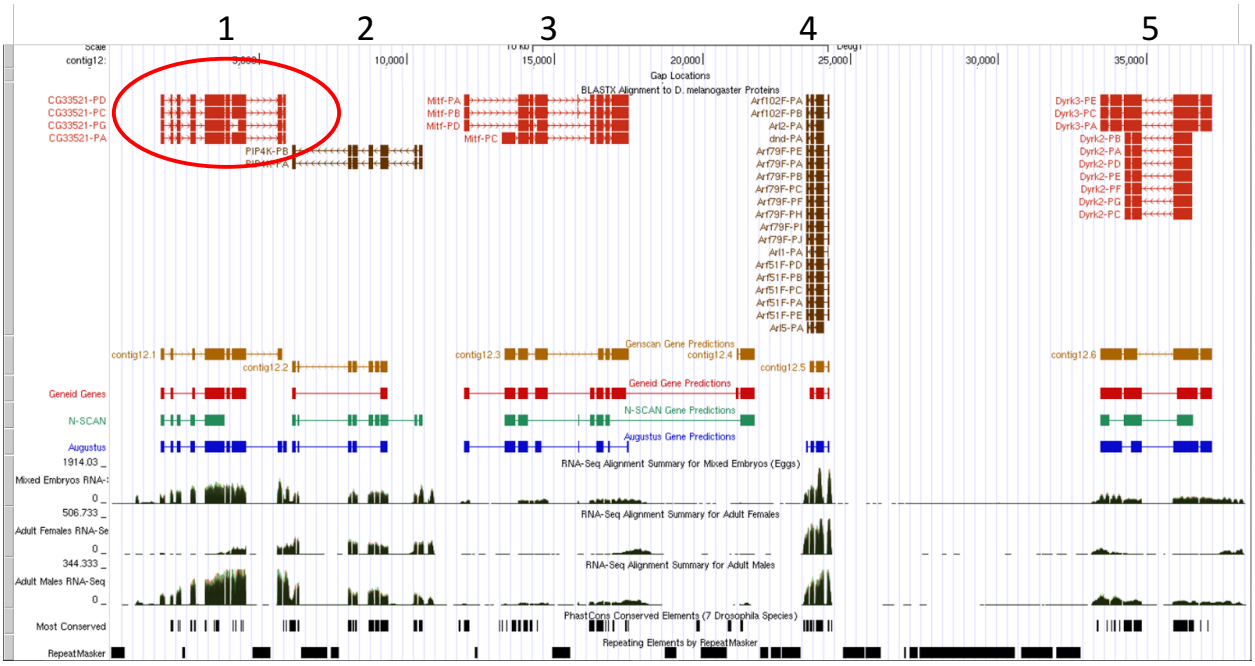


Figure 1: The project region and selected feature. This 38,500 base pair (bp) region is contig12. Shown in the red circle is Feature 1, which was selected for annotation as displayed in the Genome Browser.

Examination of the computer-based *ab initio* gene predictions corresponding to Feature 1 reveals several predictions that correlate reasonably well with this feature. The Genscan prediction, contig12.1, was selected for BLASTp analysis (Fig. 2).

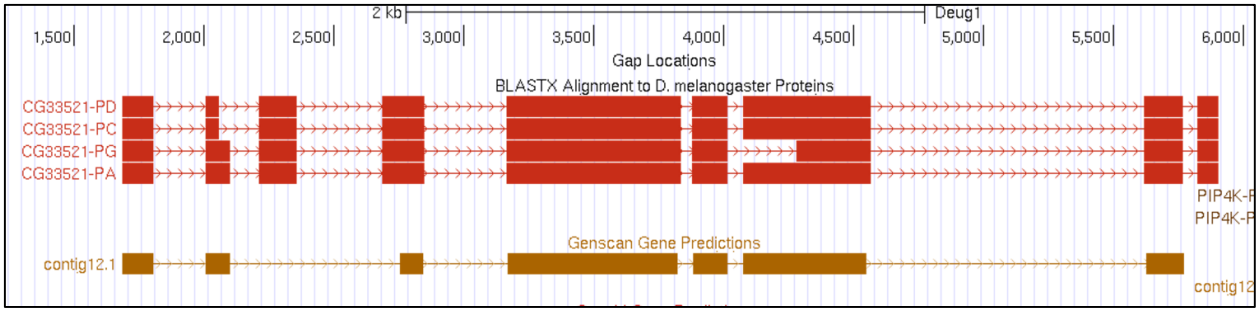


Figure 2: Gene prediction used to obtain predicted protein sequence. The Genscan gene prediction, contig12.1, corresponds reasonably well with the BLASTx alignment.

The predicted protein sequence was obtained from the Genscan gene prediction. This sequence was used in a BLASTp analysis through FlyBase to search for similarities to *D. melanogaster* proteins. The results obtained provide very strong evidence that *CG33521* is the *D. melanogaster* ortholog of this feature, as the only matches obtained were to the protein products of *CG33521*. The e-scores of the matches did not exceed 1e-100 (Fig.3), which is not an unexpected result, given that the Genscan model may have missed two exons.

BLAST Hit Summary				
☑	Description	Species	Score	E value
☑	CG33521-PD	Dmel	627.861	6.77829e-180

Figure 3: Summary of BLASTp search. The FlyBase BLASTp program provided this summary of the BLASTp alignments obtained when searching with the Genscan predicted protein (query) against the *Drosophila* Annotated Proteins Database (subject). The alignment to the isoforms of *CG33521* was the only alignment produced.

Along with the summary shown in Figure 3, the BLASTp search provided the following alignment of PD (Fig.4); similar alignments are available for the other three isoforms of *CG33521*. Note that *CG33521* is found on the 4th chromosome of *D. melanogaster*, providing further evidence that it is indeed the orthologous gene.

>gnl|dmel|FBpp0090953 type=protein; loc=4 join(1186518..1186634, 1187522..1187572, 1187735..1187880, 1188189..1188349, 1189054..1189689, 1189881..1190008, 1190067..1190549, 1192020..1192162, 1192225..1192306, 1192357..1192359); ID=FBpp0090953; name=CG33521-PD; parent=FBgn0250819, FBtr0091473; dbxref=FlyBase:FBpp0090953, FlyBase_Annotation_IDs:CG33521-PD, GB_protein:AAX53604.1, REFSEQ:NP_001014702, GB_protein:AAX53604, UniProt/TrEMBL:Q59DP3, FlyMine:FBpp0090953, modMine:FBpp0090953; MD5=321799940563a26c65f02514e73e592f; length=649; release=r6.14; species=Dmel;
Length = 649

HSP # = 1 , **Score =** 627.861 bits (1618) , **Expect =** 6.77829e-180
Identities = 366 / 631 (58%) , **Positives =** 444 / 631 (70.4%) , **Gaps =** 72 / 631 (11.4%)

Subject FASTA

```
Query:   1          MDTLNTNITFESSENSLPSKKGKSKKSKTTKYDSQNINMKKSFTKFDELQKRNLIH----   56
                    MD+LN    +  S  N LP KK K+KKSKT +YD+QNINMKKSFTKFD+LQKRN+IH
Subject: 1          MDSLNPQSSKISFGNGLPLKKKKTKKSKTNEYDNQNINMKKSFTKFDDLQKRNVIHSTIP   60

Query:   57         ----------------------------------VRSRDSKKAQEV--------------   68
                                                     +R +D  K  +
Subject: 61         EKVENCHQCKKPVYKMEEVILSLKTATTIFHKTCLRCKDCGKHLKFDSYNVHDGSLYCSM   120

Query:   69         -FRIV------YEEITTRKPELIIRENQPIELPPDVARASDKPSLGLDELQQLDVRSKFK   121
                     F+++      YEE T RK ELIIRENQPI+LPPDVA+ASDKPSLGLDELQ+L++RSKFK
Subject: 121        HFKLIFAPKVVYEEFTPRKAELIIRENQPIKLPPDVAKASDKPSLGLDELQELNLRSKFK   180

Query:   122        VFENGCKEHNNNLQERQDNITHCNAITQNKSIRSTLTKLQKLGITNSEPRKLSDINTRND   181
                    VFENG +EHNNNL+ERQD      AIT +KSI+STLTK   LGI NSE  KL D N+ N
Subject: 181        VFENGYEEHNNNLRERQD-----IAITHSKSIQSTLTKFHGLGIPNSELTKLDDKNSDN-   234

Query:   182        LNTDDESDTDILYSRKDIERERPQGLGDAMNDIRSKFEHGQAMLKEERREERKQELQSIR   241
                     N+D + D + +  +K+IERE P GLG+AMNDIRSKFE G  M KE RREERKQE+Q+IR
Subject: 235        -NSDGDGDMNFMCLKKEIERETPVGLGEAMNDIRSKFEQGDLMAKEGRREERKQEIQNIR   293

Query:   242        SRLFLGKQAKIKEMYQLAVAESEQRGNSVGKTSDINVITATQQIKDRFENGDVFNDNRIQ   301
                    SRLFLGKQAKIKEMY+LAVAESEQ   SVGKT DI  I  TQ+IK+RFENG+V+ D++I
Subject: 294        SRLFLGKQAKIKEMYKLAVAESEQPVTSVGKTPDICAIKPTQEIKNRFENGEVYKDSKIL   353

Query:   302        SKEPIFGMQADENVFESAISKSSRSIFMEMDANISSISSKSCVKNVQSDKKILYHNQMCQ   361
                    S E   G+  D +VFES ISK+SR+IFM++DANI S  S   V+    DKK   HNQ  Q
Subject: 354        SSEVSCGIHEDADVFESGISKASRNIFMKLDANIKSGLSNH-VQYTLPDKKYQIHNQKVQ   412

Query:   362        ENSNVDVIKSDSKQEEVKVTTEELAKRFKFFEEYSPDRKKKKKFCMTPPREDVQNLLAID   421
                    ENS+V+++KSDSK EEVKV TEEL+K+FKFFE YSP   KK+ F MTPPRE V      D
Subject: 413        ENSDVEIVKSDSKPEEVKVATEELSKKFKFFETYSPAENKKRIFRMTPPREGVVMFPPPD   472

Query:   422        LDTD--VSHDLFDDTVLNNTKTTTTILTKFREMEEQKLNDKNKERNPKPLKCFTPPPEID   479
                     +T+  +S  LF+D +L  TKTT+TIL KFREMEEQK++D+ K++NPKPLKCFTPPPEI
Subject: 473        SETNQQISTTLFNDNILQKTKTTSTILNKFREMEEQKMSDQQKKKNPKPLKCFTPPPEIS   532

Query:   480        NH-LKSDTEEDNYSDSEQNSDDDEEEFENVPPNSYHKDEALYEAQNSARAKQLRAKFEKW   538
                    +  ++SDTEE++ SD EQNS++DEE   N P NSY+ D+AL EAQ+ ARAKQLRAKFEKW
Subject: 533        HQFIRSDTEEESDSDYEQNSENDEESEIN-PSNSYYNDKALLEAQSVARAKQLRAKFEKW   591

Query:   539        QINEIERELNEGRENV-SKLISNESIESAKT   568
                    Q NEIE+E+ EGR +V S+LISNESIESAK
Subject: 592        QNNEIEQEIKEGRIDVYSQLISNESIESAKA   622
```

Figure 4: BLASTp alignment of the predicted protein from Genscan against *D. melanogaster CG33521-PD.* Because *CG33521-PD* and *CG33521-PC* contain the same CDS, an identical alignment was produced by matching this query to the *CG33521-PC* amino acid sequence; appropriate similar results were obtained for the other isoforms. Note that *CG33521* is located on the 4th chromosome of *D. melanogaster* (red box). The start of this feature aligns exactly to the start of *CG33521*-PD in *D. melanogaster* (yellow box), however, there are a few amino acids from the end of *CG33521*-PD in *D. melanogaster* that were not included in this alignment (shown in the blue boxes).

From the BLASTp evidence, *CG33521* was determined to be the *D. melanogaster* ortholog of this feature. FlyBase reports that this gene has three isoforms with unique CDS's in *D. melanogaster* and is predicted to have a function related to zinc ion binding. Figure 5 shows all four isoforms of *CG33521* as they occur in *D. melanogaster*.

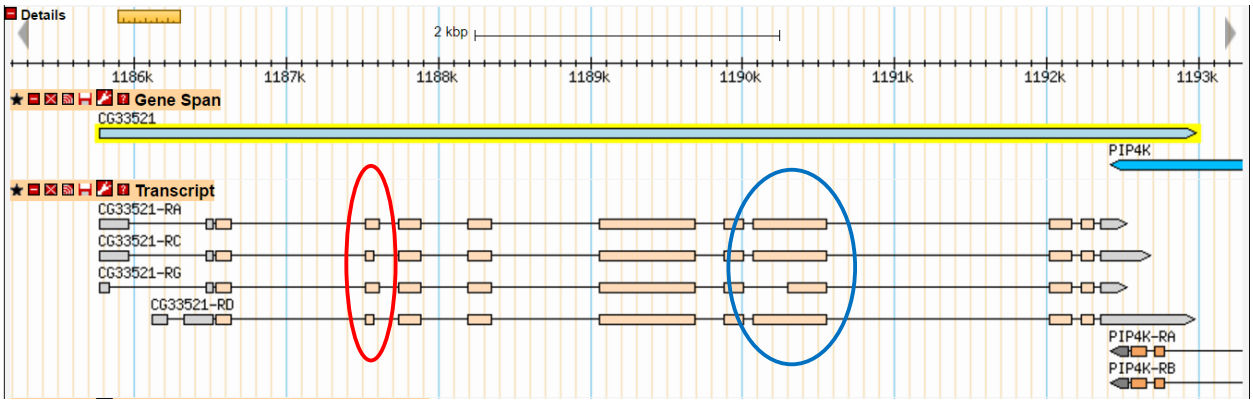


Figure 5: GBrowse view of *CG33521.* In *D. melanogaster, CG33521*-PC and *CG33521*-PD only differ in their 5' and 3' untranslated regions (UTRs). *CG33521*-PA and *CG33521*-PG differ from PC and PD in that they have a shorter second exon (red circle). *CG33521*-PG differs even further in that its 7th exon is shorter (blue circle).

Exon-by-Exon Annotation of *CG33521-PC*

Because *CG33521*-PC and *CG33521*-PD aligned best to the predicted protein in the BLASTp search, these isoforms were annotated first. More specifically, *CG33521*-PC was annotated, as all of the annotated coding exons can be attributed to *CG33521*-PD as well. As seen on Figure 7, the 5th exon of *CG33521*-PC is the largest, and would therefore be expected to produce the strongest BLASTx alignment. Using the amino acid sequence of the 5th exon of *CG33521*-PC from the *D. melanogaster* Gene Record Finder model as the subject and the DNA sequence of contig12 as the query produced the BLASTx output shown in Figure 6.

CG33521:6_12124_2
Sequence ID: Query_675 Length: 211 Number of Matches: 1

Range 1: 1 to 211 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
263 bits(673)	3e-84	143/218(66%)	169/218(77%)	7/218(3%)	+2

```
Query  3170  SDKPSLGLDELQQLDVRSKFKVFENGCKEHNNNLQERQDNITHCNAITQNKSIRSTLTKL  3349
             SDKPSLGLDELQ+L++RSKFKVFENG +EHNNNL+ERQD      AIT +KSI+STLTK
Sbjct  1     SDKPSLGLDELQELNLRSKFKVFENGYEEHNNNLRERQDI-----AITHSKSIQSTLTKF  55

Query  3350  QKLGITNSEPRKLSDINTRNDLNTDDESDTDILYSRKDIERERPQGLGDAMNDIRSKFEH  3529
               LGI NSE  KL D N+ N  N+D + D + +  +K+IERE P GLG+AMNDIRSKFE
Sbjct  56    HGLGIPNSELTKLDDKNSDN--NSDGDGDMNFMCLKKEIERETPVGLGEAMNDIRSKFEQ  113

Query  3530  GQAMLKEERREERKQELQSIRSRLFLGKQAKIKEMYQLAVAESEQRGNSVGKTSDINVIT  3709
             G  M KE RREERKQE+Q+IRSRLFLGKQAKIKEMY+LAVAESEQ   SVGKT DI  I
Sbjct  114   GDLMAKEGRREERKQEIQNIRSRLFLGKQAKIKEMYKLAVAESEQPVTSVGKTPDICAIK  173

Query  3710  ATQQIKDRFENGDVFNDNRIQSKEPIFGMQADENVFES  3823
              TQ+IK+RFENG+V+ D++I S E   G+  D +VFES
Sbjct  174   PTQEIKNRFENGEVYKDSKILSSEVSCGIHEDADVFES  211
```

Figure 6: BLASTx search of contig12 against the 5th exon of *CG33521*-PC. The *D. melanogaster* exon amino acid sequence (subject) was searched against contig12 (query) This search produced a very strong match (e-score = 3e-84). The aligned sequence is in frame +2.

After a BLASTx placement of this exon was obtained, the region of the contig where the alignment began was investigated. Because BLASTx only reports similarity to the subject amino acid sequence, it is useful in identifying the region where splicing of the exon occurs, but similarity is not the most accurate evidence available to annotate the splice sites of exons. Available evidence tracks in the Genome Browser used to determine the precise location of the splice donor and acceptor sites included RNA-Seq, TopHat junctions, and *ab initio* gene predictors. The splice acceptor site of exon 5 was annotated (Fig.7) by using the NCBI BLASTx alignment to target the region at the start of this exon where RNA-Seq data, TopHat junctions, conservation, and computer-based gene predictions were used to determine the most well-supported splice acceptor site. The region of contig12 which aligned to the end of the exon in the BLASTx results was used to locate the region where the splice donor site was expected to occur. By viewing this region in the Genome Browser, RNA-Seq data tracks, TopHat junctions, and

computer-based gene prediction tracks could be shown to support a single putative GT splice donor site. (Fig.8).

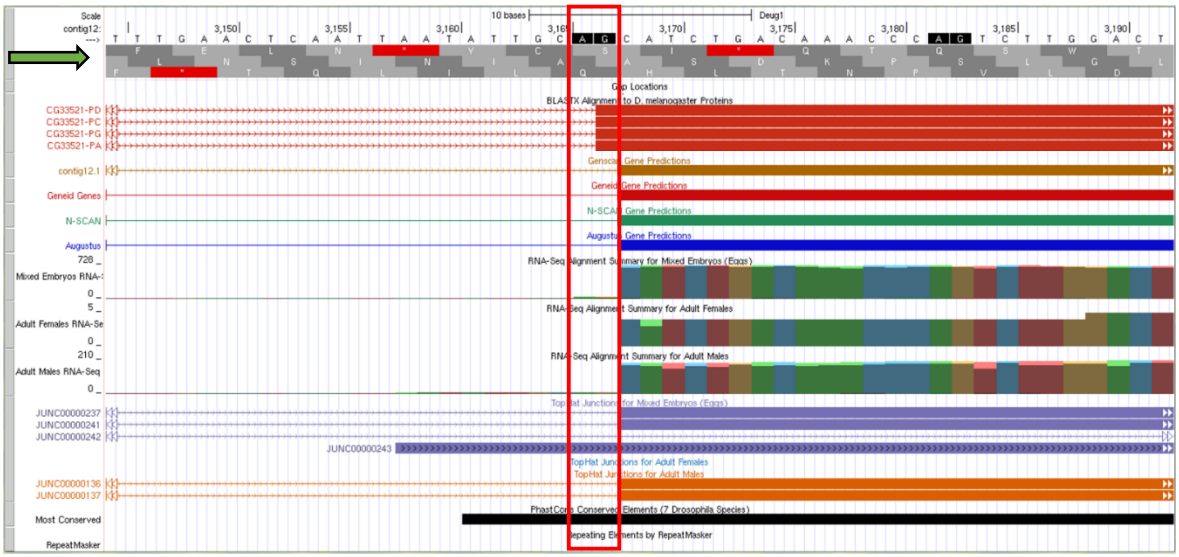


Figure 7: Exon five splice acceptor site. This proposed splice acceptor site (red box) is supported by the computer-based gene predictions, RNA-Seq data, the conservation track, and all but one TopHat junction. The AG sequence that corresponds with this site is in phase two, given that the exon is in reading frame 2 (green arrow). This was made note of, along with the position of the splice site. Note that this splice acceptor site (3,168 bp) is nearby, but not exactly the position of the start of the NCBI BLASTx alignment (3,170 bp).

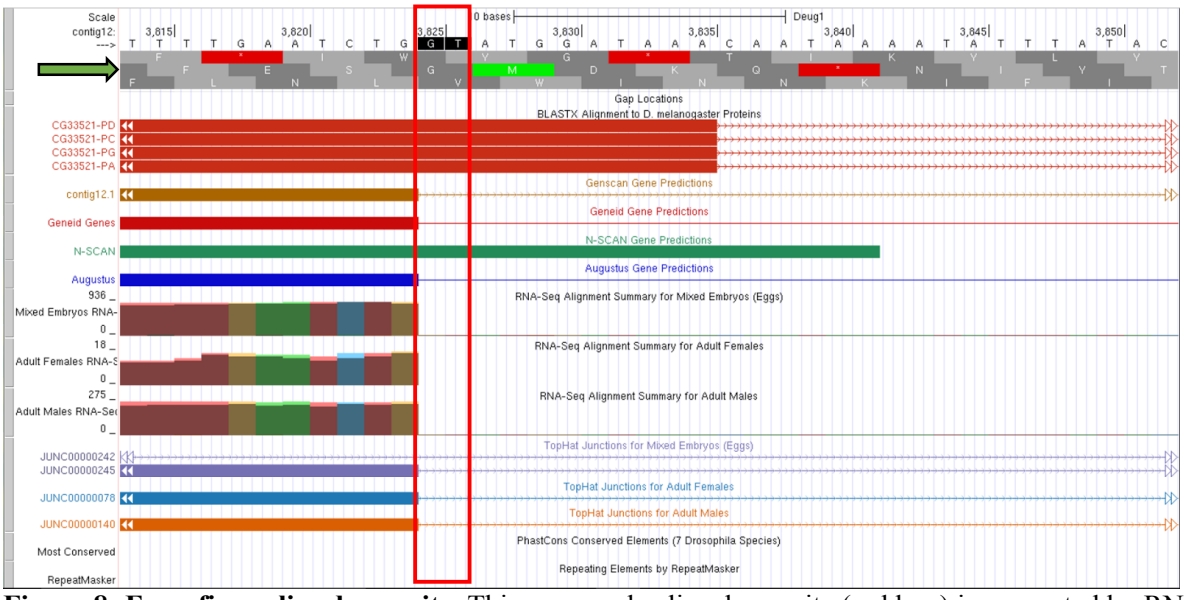


Figure 8: Exon five splice donor site. This proposed splice donor site (red box) is supported by RNA-Seq data, TopHat junctions, and all but one computer-based gene prediction. The GT sequence that corresponds with this splice site is the only available GT in the immediate area. Because this exon is in frame +2 (indicated by the green arrow), this donor splice site is in phase one. Note that, once again, this splice donor site (3,824 bp) is nearby, but not exactly the position of the end of the NCBI BLASTx alignment (3,823 bp).

Looking at the GBrowse view of this *CG33521*-PC in *D. melanogaster* shows that the other large exon in this isoform (exon seven) is only separated from the now annotated exon 5 by a single exon. Since exon five and exon seven can serve as anchors for the annotation of smaller exons, the next exons annotated were those flanking these two large exons. Thus, the next exon annotated was exon six. Because it is not as large as exon five, its NCBI BLASTx alignment did not produce as low an e-score (Fig.9); however, taking the size of the exon into consideration, the e-score of 1e-08 was sufficiently low to use the alignment as a tool to locate the approximate position of the splice acceptor and splice donor sites (Figs. 10 and 11).

CG33521:7_12124_2

Sequence ID: Query_106503 **Length:** 42 **Number of Matches:** 1

Range 1: 1 to 42 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
39.3 bits(90)	1e-08	23/43(53%)	28/43(65%)	1/43(2%)	+2

```
Query  3887  ISKSSRSIFMEMDANISSISSKSCVKNVQSDKKILYHNQMCQE  4015
             ISK+SR+IFM++DANI S  S   V+    DKK   HNQ  QE
Sbjct  1     ISKASRNIFMKLDANIKSGLSNH-VQYTLPDKKYQIHNQKVQE  42
```

Figure 9: BLASTx alignment of the amino acid subject sequence of *CG33521*-PC exon six from *D. melanogaster* (subject) against contig12 (query). The e-score of this alignment is 1e-08 and the aligned sequence in contig12 is in frame +2.

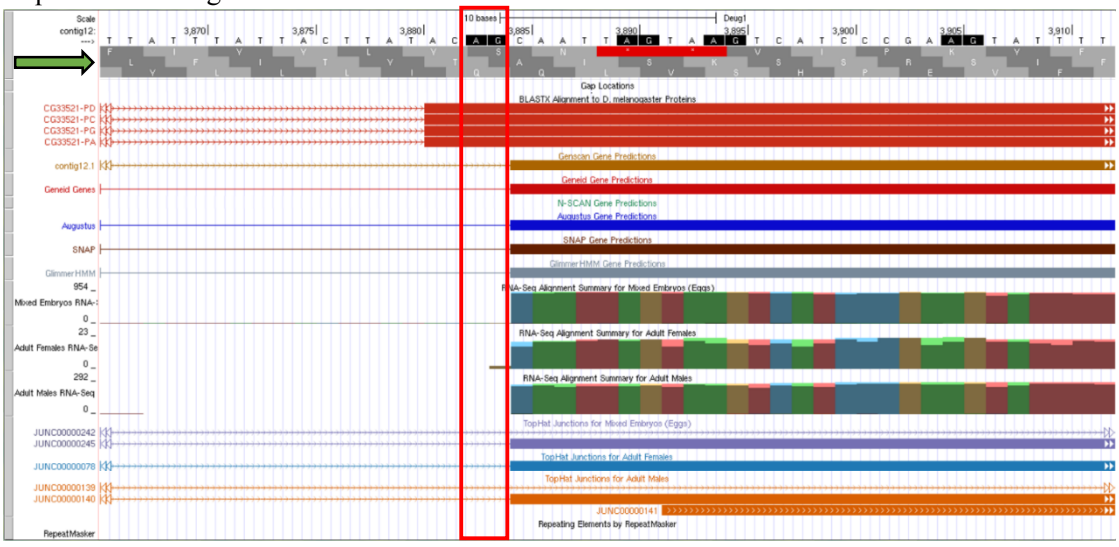


Figure 10: Exon six splice acceptor site. Given that the BLASTx alignment is in reading frame +2 (green arrow), the only AG splice site in the immediate region in the correct phase (phase two) is shown in the red box. This proposed splice acceptor site is supported by RNA-Seq data, TopHat junctions, and computer-based gene predictions.

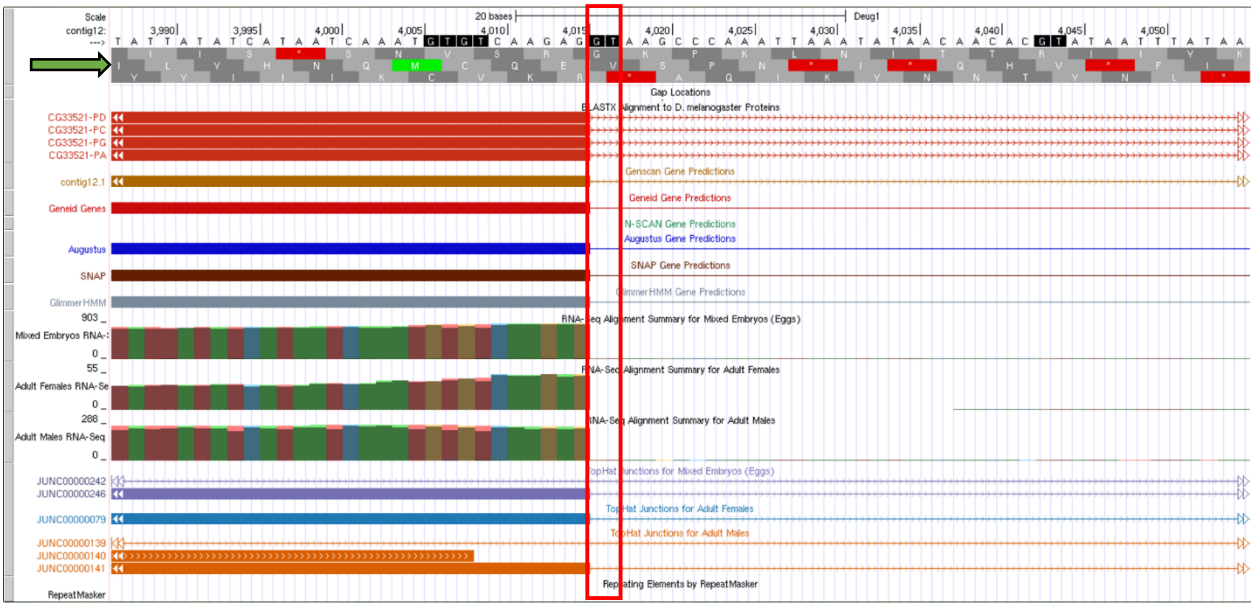


Figure 11: Exon six splice donor site. This proposed GT splice donor site of this exon is identified with the red box. This annotation is supported the BLASTx alignment track, RNA-Seq data, TopHat junctions, and computer-based gene predictions. Because this exon is in frame +2 (green arrow), this splice site is in phase zero.

As shown by the convergence of many different evidence tracks supporting both putative splice sites of exon six, its shorter length was not an issue in annotation. After the annotation of exon six, exon seven was used in an NCBI BLASTx search (Fig.12) and its splice sites were subsequently annotated (Fig.13 and 14). Exon seven in *CG33521*-PC is large and therefore expected to provide a low e-score. This is confirmed in Figure 12.

CG33521:8_12124_0

Sequence ID: Query_213343 **Length:** 161 **Number of Matches:** 1

Range 1: 1 to 161 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
190 bits(483)	1e-59	98/162(60%)	125/162(77%)	4/162(2%)	+1

```
Query  4075  NSNVDVIKSDSKQEEVKVTTEELAKRFKFFEEYSPDRKKKKKFCMTPPREDVQNLLAIDL  4254
             NS+V+++KSDSK EEVKV TEEL+K+FKFFE YSP   KK+ F MTPPRE V      D
Sbjct  1     NSDVEIVKSDSKPEEVKVATEELSKKFKFFETYSPAENKKRIFRMTPPREGVVMFPPPDS  60

Query  4255  DTD--VSHDLFDDTVLNNTKTTTTILTKFREMEEQKLNDKNKERNPKPLKCFTPPPEIDN  4428
             +T+  +S  LF+D +L  TKTT+TIL KFREMEEQK++D+ K++NPKPLKCFTPPPEI +
Sbjct  61    ETNQQISTTLFNDNILQKTKTTSTILNKFREMEEQKMSDQQKKKNPKPLKCFTPPPEISH  120

Query  4429  H-LKSDTEEDNYSDSEQNSDDDEEEFENVPPNSYHKDEALYE  4551
               ++SDTEE++ SD EQNS++DEE   N P NSY+ D+AL E
Sbjct  121   QFIRSDTEEESDSDYEQNSENDEESEIN-PSNSYYNDKALLE  161
```

Figure 12: BLASTx search of contig12 (query) against the 7th exon of *CG33521*-PC (subject). This BLASTx alignment has a strong e-score of 1e-59. The aligned sequence in contig12 is in frame +1.

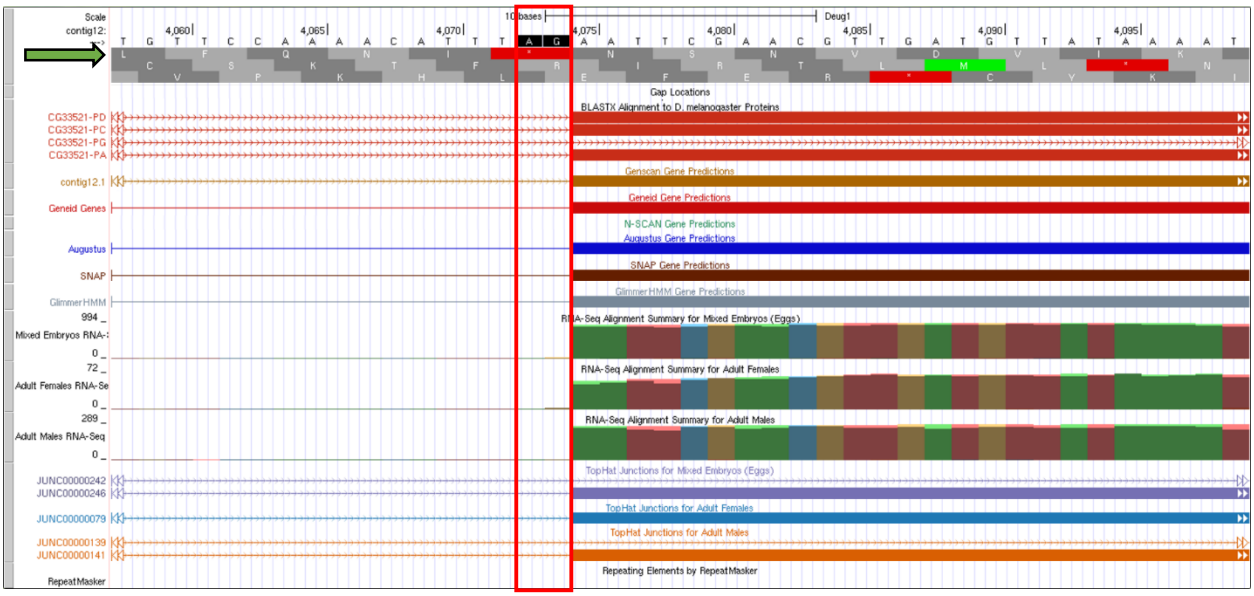


Figure 13: Exon seven splice acceptor site. This AG acceptor splice site (red box) is supported by the NCBI BLASTx alignment as well as the BLASTx alignment track, RNA-Seq data, TopHat junctions, and computer-based gene predictions. This exon is in frame +1 (green arrow), which means that this acceptor splice site is in phase zero. This corresponds to the phase zero splice donor site of exon six.

Figure 14: Exon seven splice donor site. There is a single GT splice donor motif in the region. This GT proposes a splice donor site at position 4,551. This position is supported by the NCBI BLASTx alignment, RNA-Seq data, TopHat predictions, and most computer-based gene predictions. Thus, this GT corresponds to the splice donor site of exon seven (identified with the red box). Because this exon is in frame +1 (green arrow), this splice donor site is in phase zero.

The next exon annotated was exon eight of *CG33521*-PC. A BLASTx search was performed using the amino acid sequence of this exon from *D. melanogaster* as the subject and contig12 as the query (Fig.15). This alignment was used to confirm the presence of this exon and help determine its position. The splice sites were then analyzed and annotated using evidence in the GEP Genome Browser (Figs. 16 and 17).

CG33521:10_12124_0

Sequence ID: Query_199293 **Length:** 47 **Number of Matches:** 1

Range 1: 1 to 47 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
68.6 bits(166)	7e-19	37/47(79%)	42/47(89%)	1/47(2%)	+1

```
Query  5629  AQNSARAKQLRAKFEKWQINEIERELNEGRENV-SKLISNESIESAK  5766
             AQ+ ARAKQLRAKFEKWQ NEIE+E+ EGR +V S+LISNESIESAK
Sbjct  1     AQSVARAKQLRAKFEKWQNNEIEQEIKEGRIDVYSQLISNESIESAK  47
```

Figure 15: BLASTx search of contig12 (query) against the 8th exon of *CG33521*-PC in *D. melanogaster* (subject). This exon was well-conserved and had an e-score of 7e-19, which is a strong score considering the size of this exon. The aligned sequence in contig12 is in frame +1.

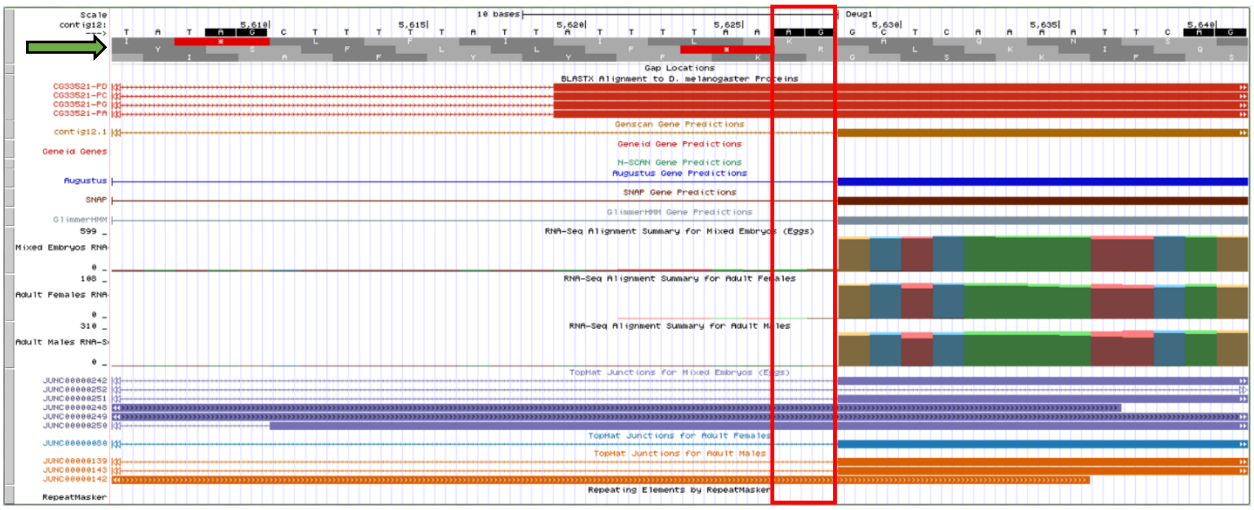


Figure 16: Exon eight splice acceptor site. There are two AG splice acceptor motifs highlighted in black that indicate possible splice sites in the correct phase (phase zero) in frame +1 (green arrow). Looking at RNA-Seq data, TopHat junctions, and computer-based gene predictions confirms that the downstream AG corresponds to the correct splice acceptor site at position 5,629. This AG splice acceptor site is indicated by the red box.

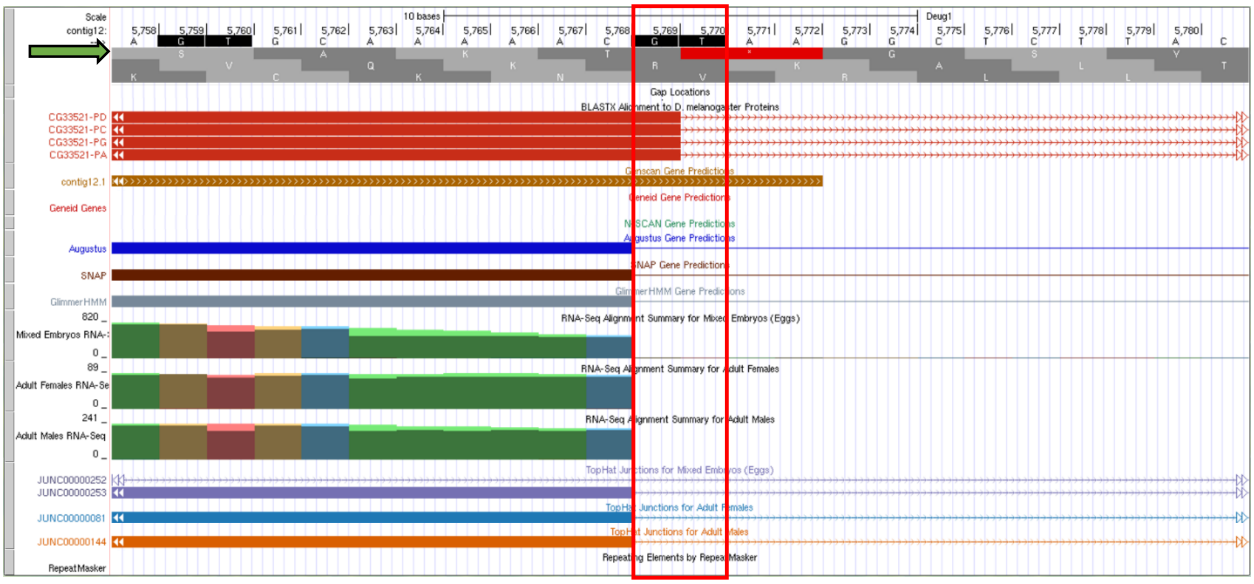


Figure 17: Exon eight splice donor site. The downstream GT is supported by RNA-Seq data, TopHat junctions, and most computer-based gene predictions. Given that this exon is in reading frame +1 (green arrow), this GT splice donor site is identified by the red box and is in phase two.

All the subsequent splice sites were annotated following this workflow. There were no abnormalities of note, except for the splice acceptor site of the last exon, which will be addressed

with the discussion of the annotation of the stop codon. Every interior exon of *CG33521*-PC in *D. melanogaster* produced a BLASTx alignment with an e-score of 1e-08 or lower.

Annotation of the Start Codon

The start codon of this isoform was confirmed by several lines of evidence. The first coding exon produced a BLASTx alignment with an e-score of 4e-11 (Fig.18), which is a reasonable value given the shorter length of the exon. After obtaining the BLASTx alignment of this exon, it was not difficult to obtain the start codon, as the BLASTx match aligned to a methionine that was supported by the Genome Browser's computer-based gene prediction tracks and the starting amino acid sequence (Figs. 18 and 19).

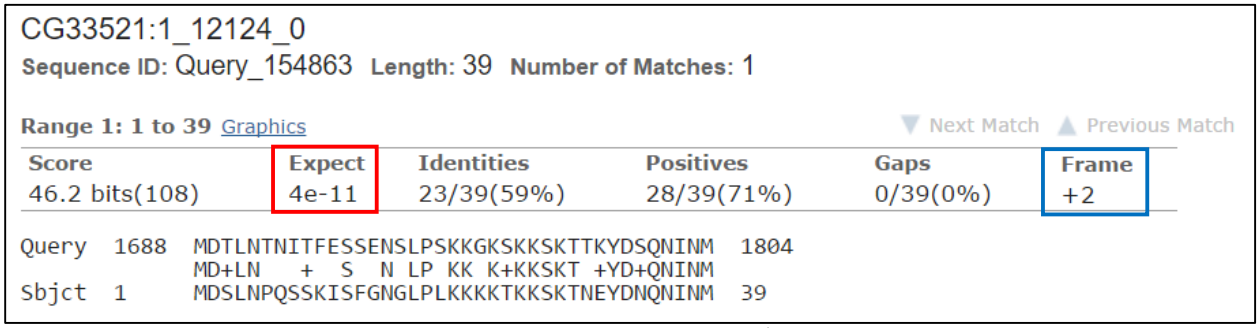


Figure 18: BLASTx search of contig12 (query) against the 1st exon of *CG33521*-PC in *D. melanogaster* (subject). The methionine predicted to serve as the start codon of this isoform can be seen at the start of this alignment. This alignment has an e-score of 4e-11 (red box) and the aligned sequence from contig12 is in frame +2 (blue box).

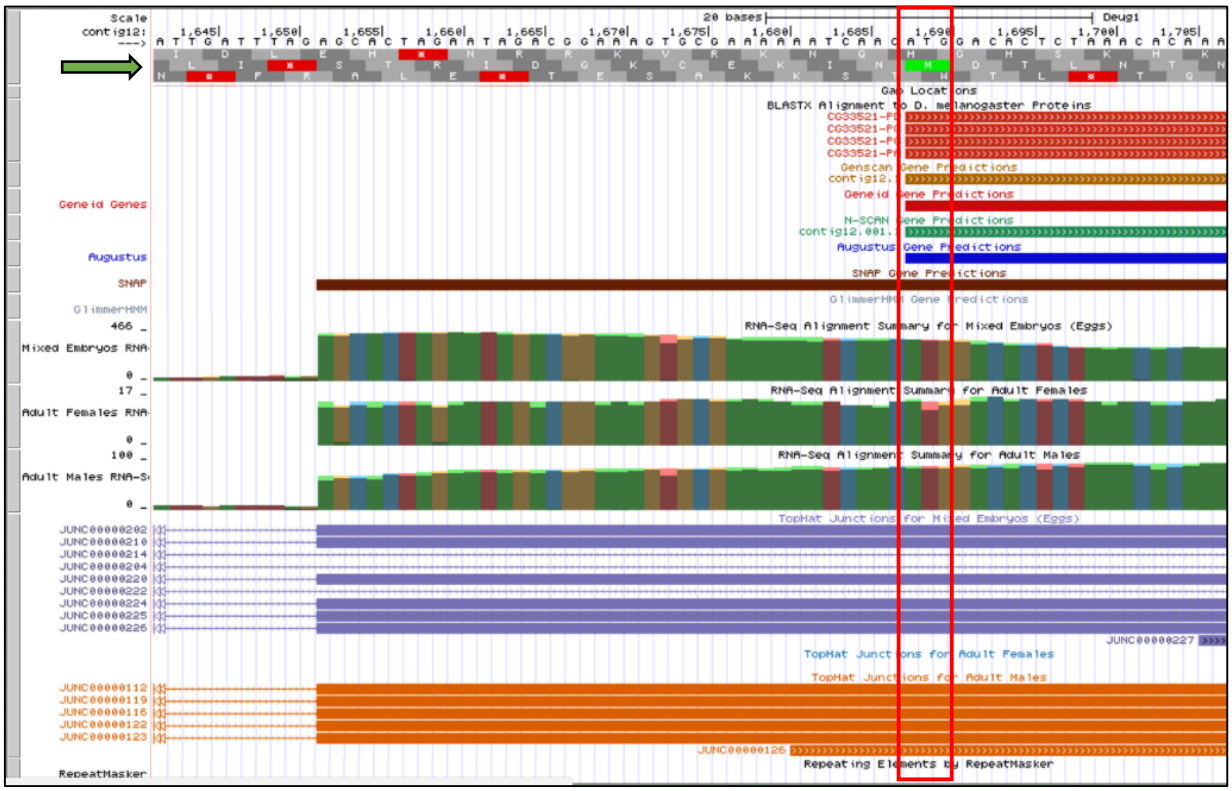


Figure 19: Annotation of the start codon of *CG33521*-PC. The only nearby methionine is in frame +2, which can be confirmed as the reading frame by the stop codons in frame +1 and +3. The RNA-Seq reads all begin upstream of the start codon, corresponding to the 5" UTR. This annotation (shown in the red box) is supported by the NCBI BLASTx alignment, the BLASTx alignment track and the computer-based gene predictions.

Annotation of the Stop Codon and Final Exon

Annotation of the stop codon as well as the final exon of this isoform provided a challenge, as the final exon of this isoform in *D. melanogaster* contains only the stop codon (Fig.20). As it is impossible to locate a unique 3 bp sequence with a BLASTx search, a different approach was required to locate this exon.

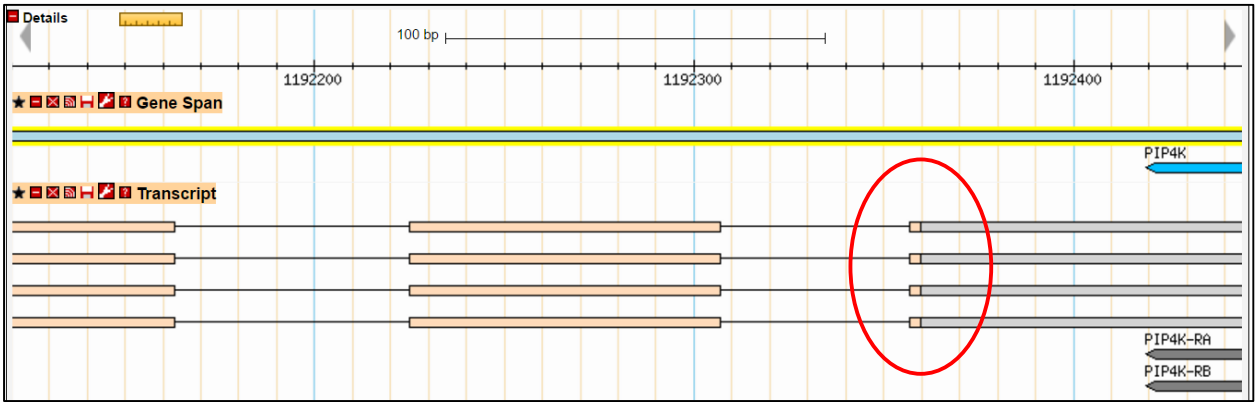


Figure 20: The final exon of *CG33521-PC* is a three bp stop codon. This exon will require a different annotation strategy, as it cannot be located with a BLASTx alignment.

The first step taken was to examine the region in contig12 where the exon could occur, looking for TopHat junctions, which could predict a possible exon (Fig.21). The analysis of TopHat junctions did not provide any evidence of a final intron. The next step taken was to use the 3' UTR of *CG33521*-PC to perform a BLASTn search against contig12 (Fig.22). Generally, untranslated exons are not as well conserved as coding exons, but in the absence of a coding exon, the 3' UTR could serve as an anchor to help locate the final exon.

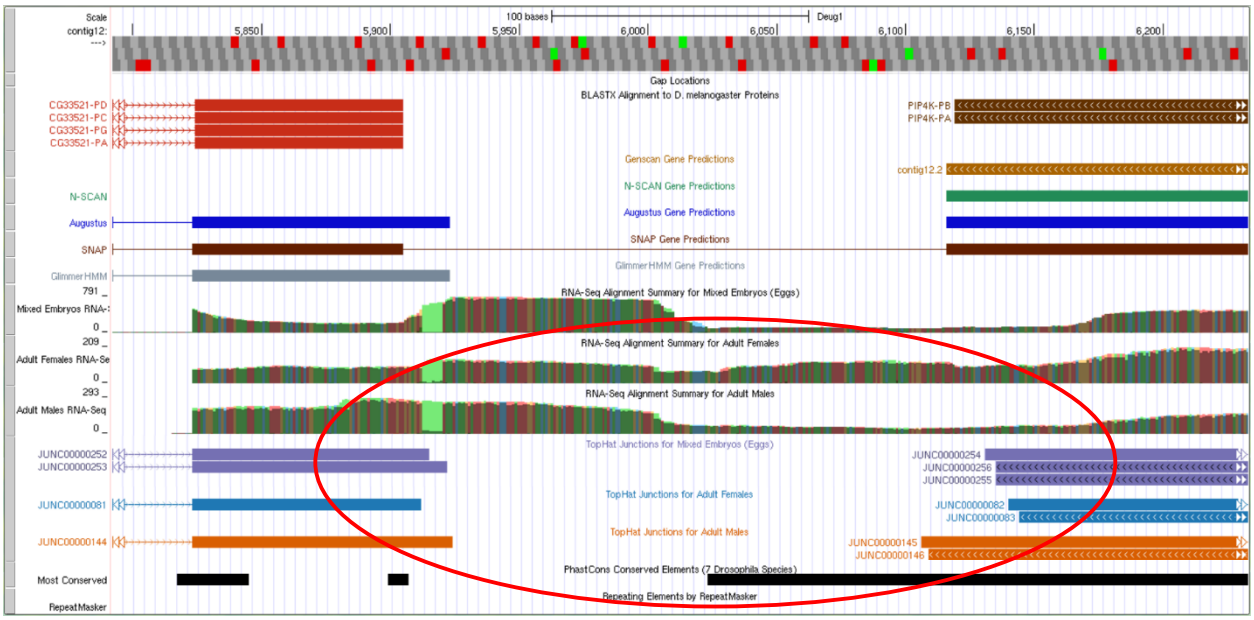


Figure 21: TopHat junctions do not identify a possible final exon. The search region for the final exon (identified by the red circle) extends from the end of the penultimate coding exon to the start of the next predicted gene.

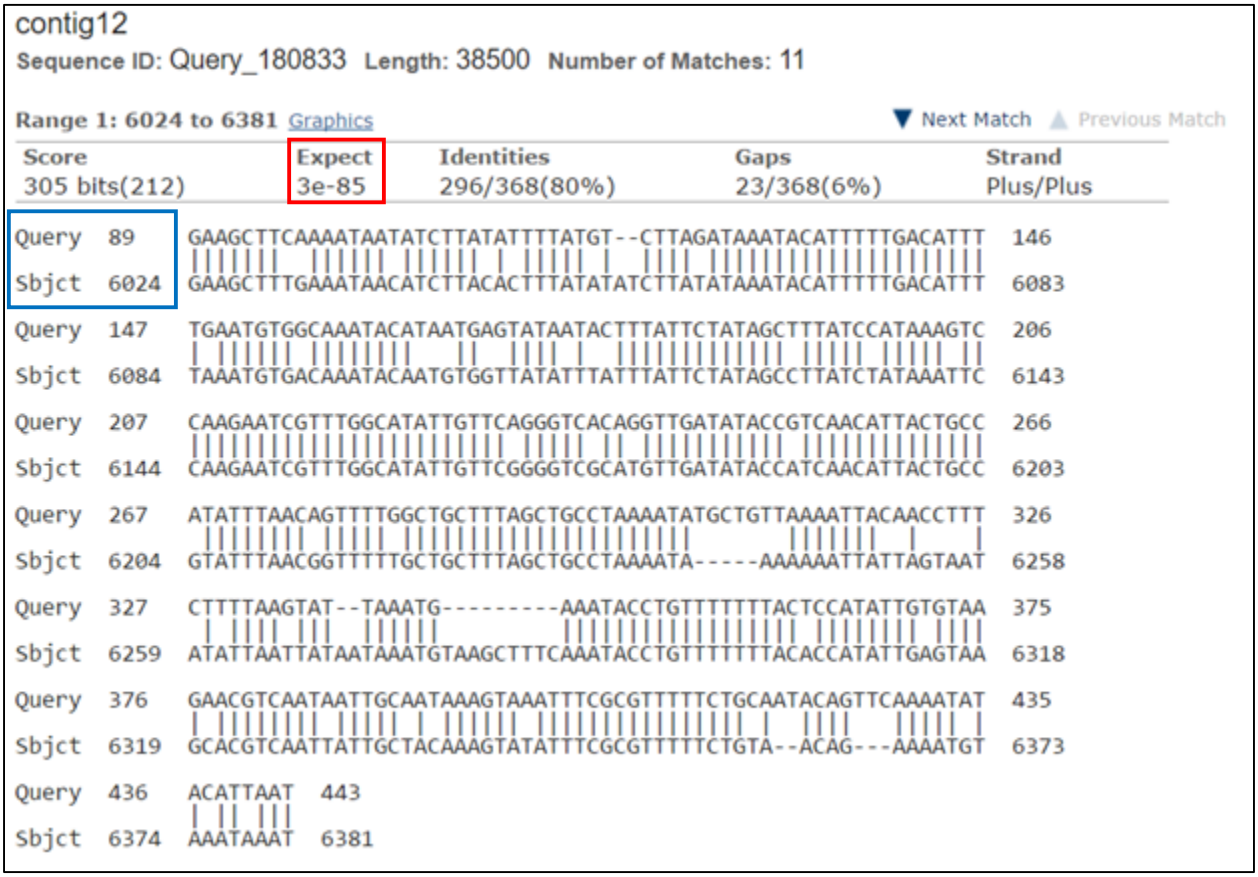


Figure 22: BLASTn alignment of the 3' UTR of *CG33521*-PC from *D. melanogaster* (subject) to contig12 (query). This search provided an alignment with a strong e-score of 3e-85 (red box); however, the first 88 bases of the query remained unaligned (blue box).

The BLASTn alignment did not precisely place the final coding exon of this isoform. The position of the exon was extrapolated by subtracting 88 bases from the start of the subject alignment, but a putative stop codon with canonical splice sites could not be found in the region. The next step was to use the Small Exons Finder (Fig.23) to search for possible stop codons. In order to create a broad search region with well-define boundaries, the search region between the 9th exon and the start of the next putative gene were set as the upstream and downstream boundaries in the Small Exons Finder.

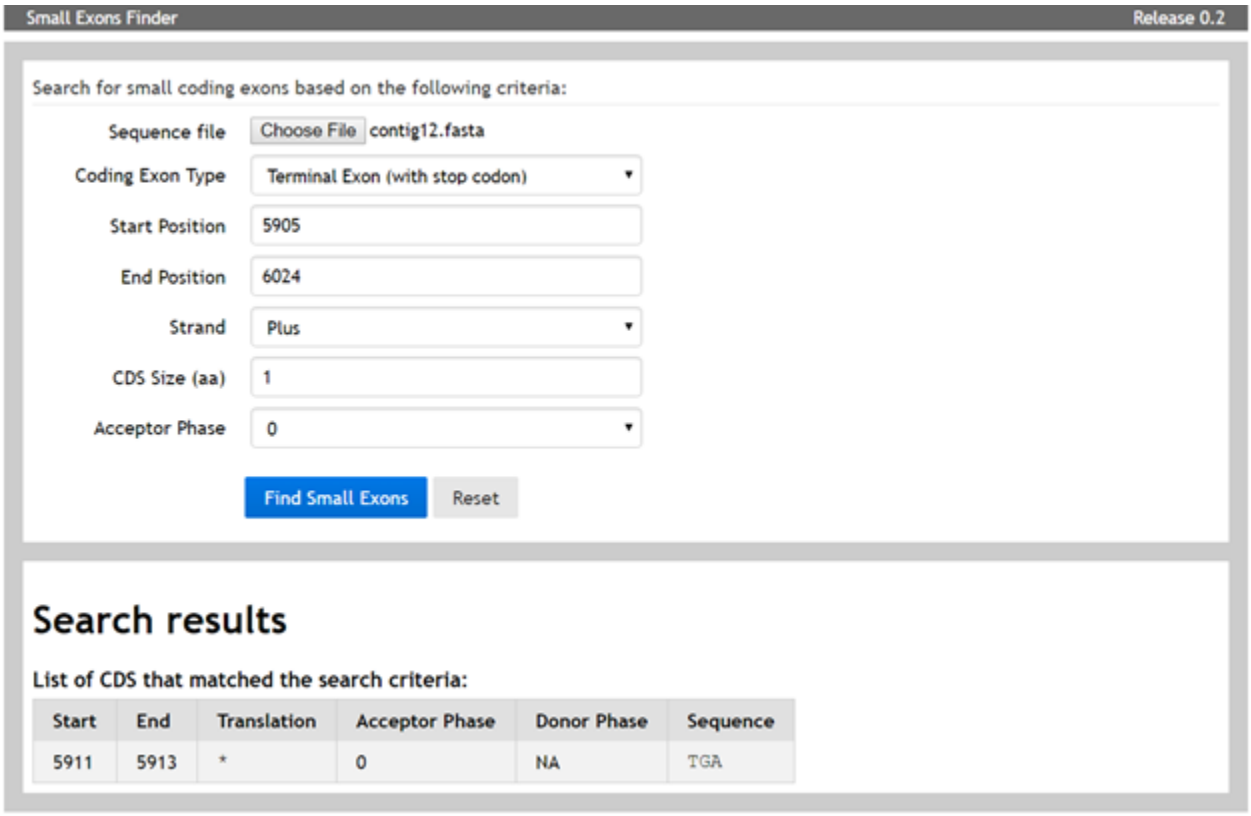


Figure 23: Small Exons Finder search results. The Small Exons Finder found one possible candidate for the final exon *CG33521*-PC at position 5911.

Looking at the position of the candidate exon identified by the Small Exons Finder reveals that it is located 6 bp downstream of the end of the 9th, penultimate exon. This violated the minimum 40 bp intron rule.

These lines of evidence suggest that this small, 10th exon does not exist in *D. eugracilis*. This hypothesis was supported by looking at the *D. eugracilis* RNA-Seq reads track, which are continuous and show no evidence of the existence of an intron (Fig.24). After discarding the small 10th exon hypothesis, the next most parsimonious hypothesis is that the 9th exon is the final exon. Incorporating this hypothesis into the annotated gene model would require extending the end of the exon until a reasonable stop codon is found. After adopting this hypothesis, it becomes

important to examine the low-quality RNA-Seq data that occurs immediately downstream of the end of the BLASTx alignment to the 9th exon (Fig.25).

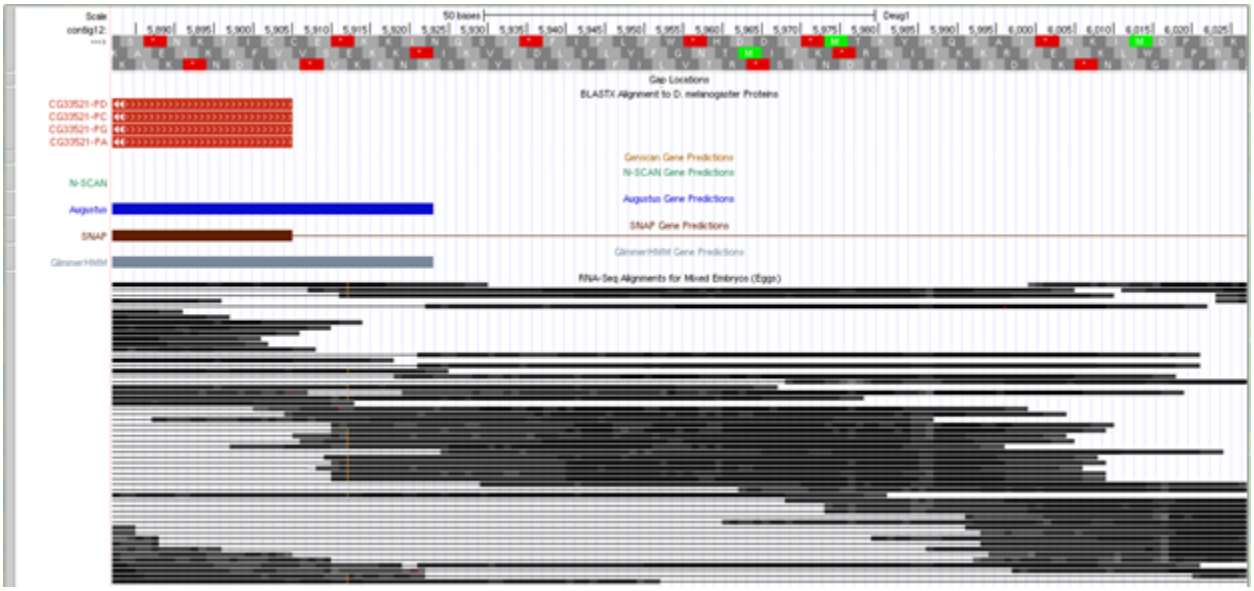


Figure 24: Analysis of *D. eugracilis* RNA-Seq reads. The *D. eugracilis* RNA-Seq reads occurring after the penultimate exon are continuous, without evidence of an intron. This supports the theory that there is no 10th exon of this isoform.

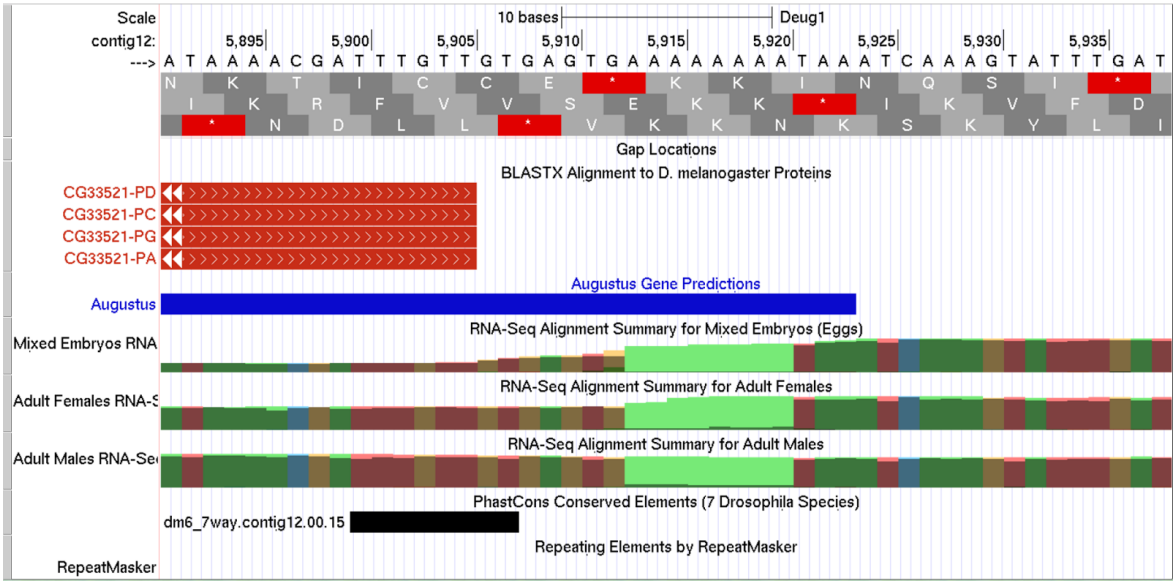


Figure 25: Low-quality RNA-Seq data. The light green bases indicate low-quality RNA-Seq data. This region contains a mono-A run.

There is a mononucleotide A run in this region. It is important to confirm that the correct number of As are in the consensus sequence. Mononucleotide runs are particularly deserving of close examination because the *D. eugracilis* genome assembly was generated using 454 sequencing, a technology prone to errors at mononucleotide runs larger than six bp With the previous hypothesis, this mono-A run would have been in an intron and the accuracy of the consensus sequence here would have been irrelevant to the final CDS annotation. However, given that the coding span of the 9th exon may include this mono-A run, an incorrect number of As in the consensus sequence would cause a frameshift, which would significantly affect the final gene model. The reads were analyzed and it was determined that an A was missing from the consensus sequence (Fig.26).

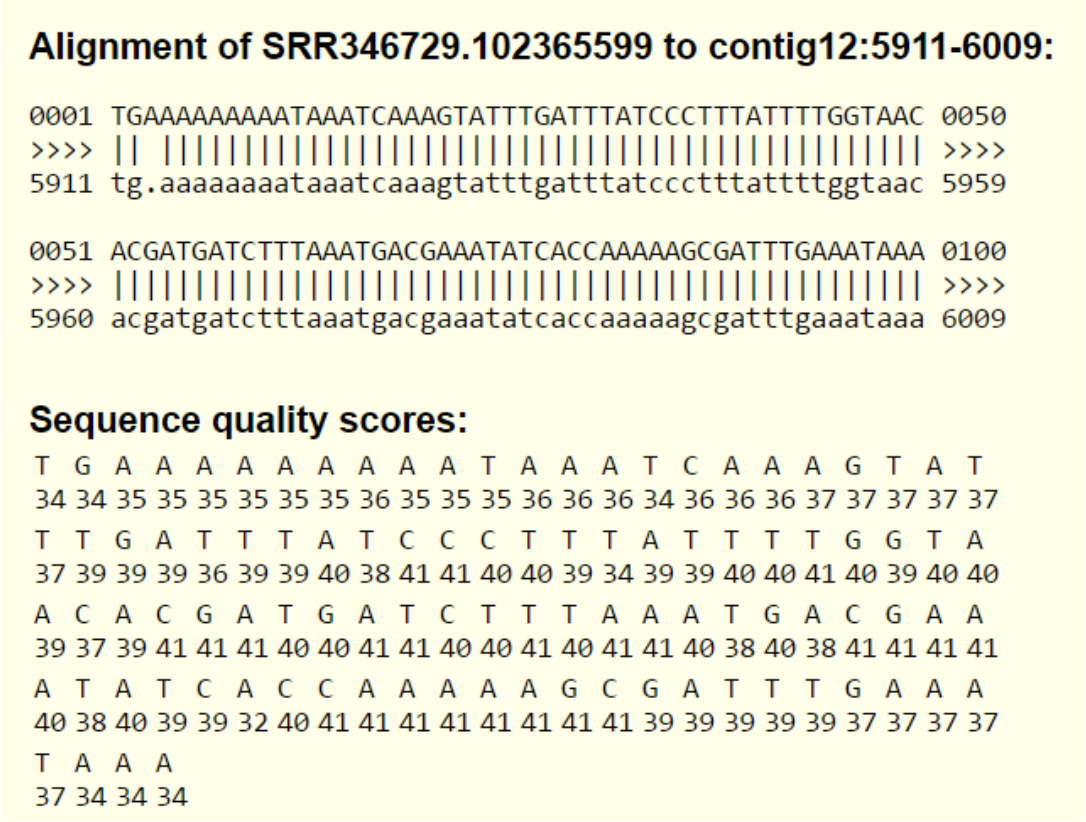


Figure 26: Analysis of *D. eugracilis* RNA-Seq reads in the mono-A run. Examination of this RNA-Seq data shows that there are nine high-quality A's in this read and confirms that there is an A missing from the consensus sequence (A). Each A has a quality score of at least 35 (B).

The missing A was added to the consensus sequence with the GEP Sequence Updater. The Genome Browser was updated with this corrected sequence using the Annotation Files Merger. This allowed for the visualization of the improved sequence. Using ExtractSeq, the end of the 9th exon and 200 bp of downstream sequence was extracted from the corrected sequence. This extracted sequence was translated using TranSeq. Because the reading frame of the amino acid sequence is now in frame +1, the first stop codon in the frame +1 translation of the corrected sequence becomes the stop codon of the 9th exon of *CG33521*-PC. The new amino acid sequence is visualized in Figure 27.

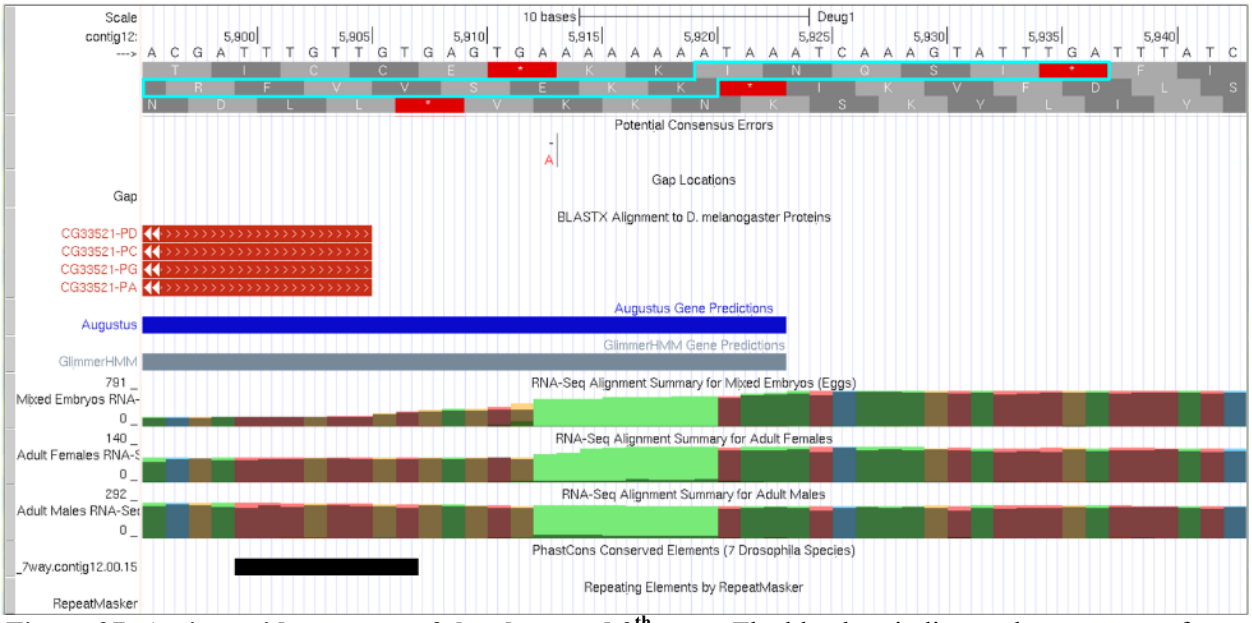


Figure 27: Amino acid sequence of the elongated 9th exon. The blue box indicates the sequence of amino acids of the 9th exon. The reading frame shifts from frame +2 to +1 at the end of the corrected mono-A run.

In this corrected sequence, the first stop codon that occurs in the correct frame is found at position 5935. Extending the 9th exon of *CG33521*-PC to this codon is the most parsimonious annotation of this gene. After finishing the annotation of this isoform, other species found near *D. eugracilis* on the *Drosophila* phylogenetic tree were examined to determine if they possess this annotation or if they had a small 10th exon, as *D. melanogaster* does (Fig.28 and 29).

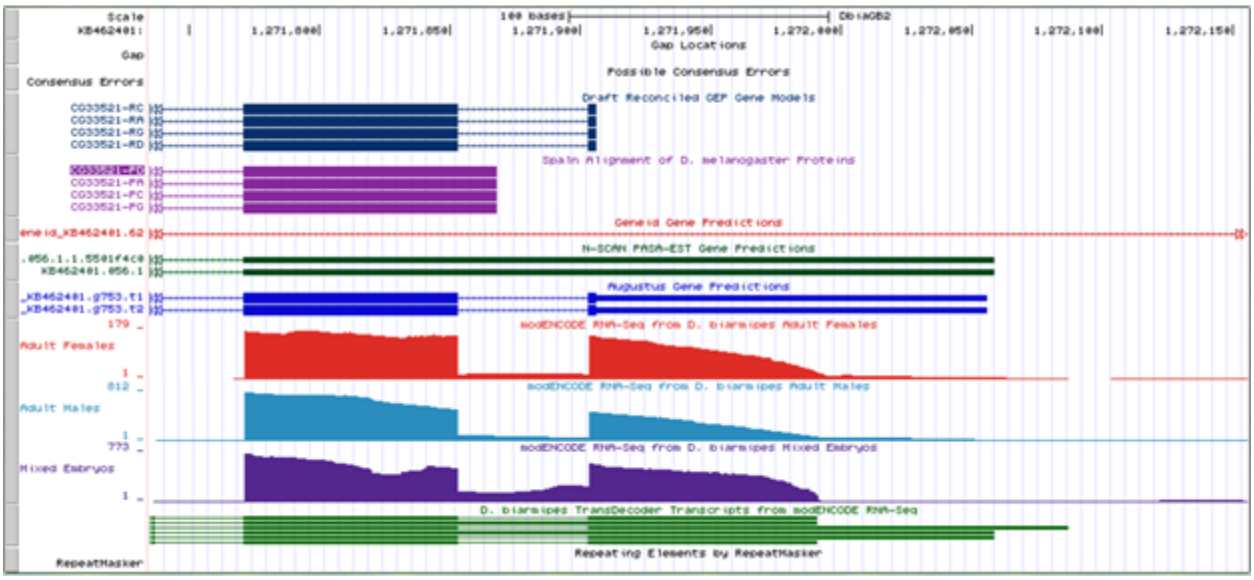


Figure 28: View of the stop codon exon of *CG33521* in *D. biarmipes*. This species is similar to *D. melanogaster* as it contains a three-bp-long final coding exon, as predicted by Augustus and RNA-Seq data and confirmed by GEP annotation.

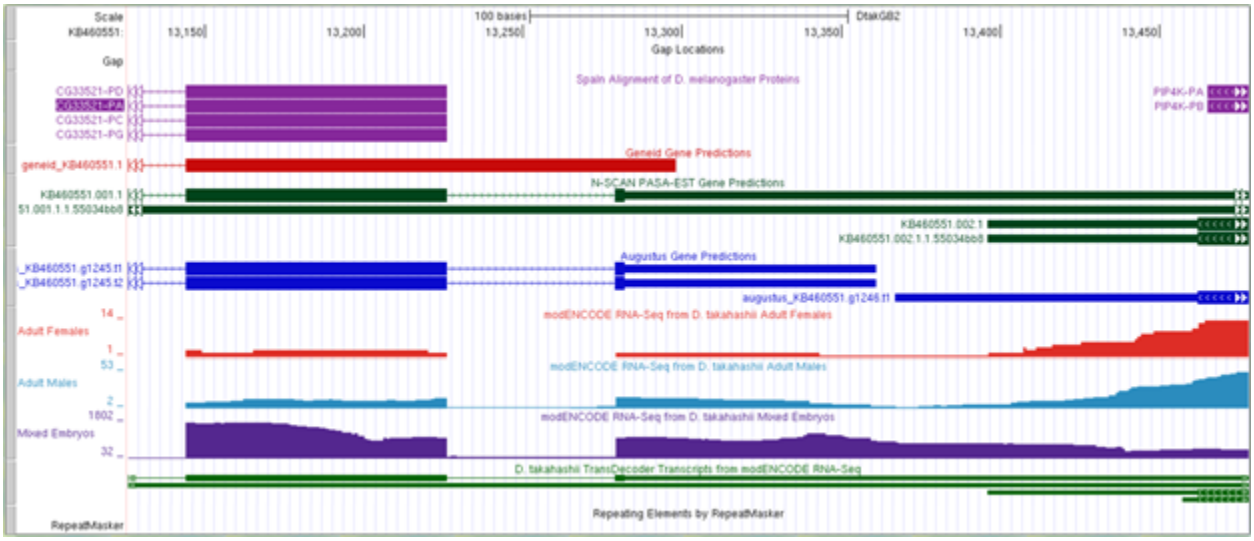


Figure 29: View of the stop codon exon of *CG33521* in D. *takahashii*. This species is similar to *D. melanogaster* as it appears to contain a three-bp-long final coding exon, and predicted by Augustus, N-Scan, and the RNA-Seq data.

Looking at the evidence tracks in the BCM-HGSC (Baylor College of Medicine Human Genome Sequencing Center) assemblies of *D. biarmipes* and *D. takahashii* reveals that neither of these species appear to contain the extended 9th exon that was annotated in *D. eugracilis*.

Looking at the *Drosophila* conservation track in the *D. melanogaster* Genome Browser reveals

that the mutation of the AG splice acceptor site of the stop codon exon is unique to *D. eugracilis* (Fig. 30).

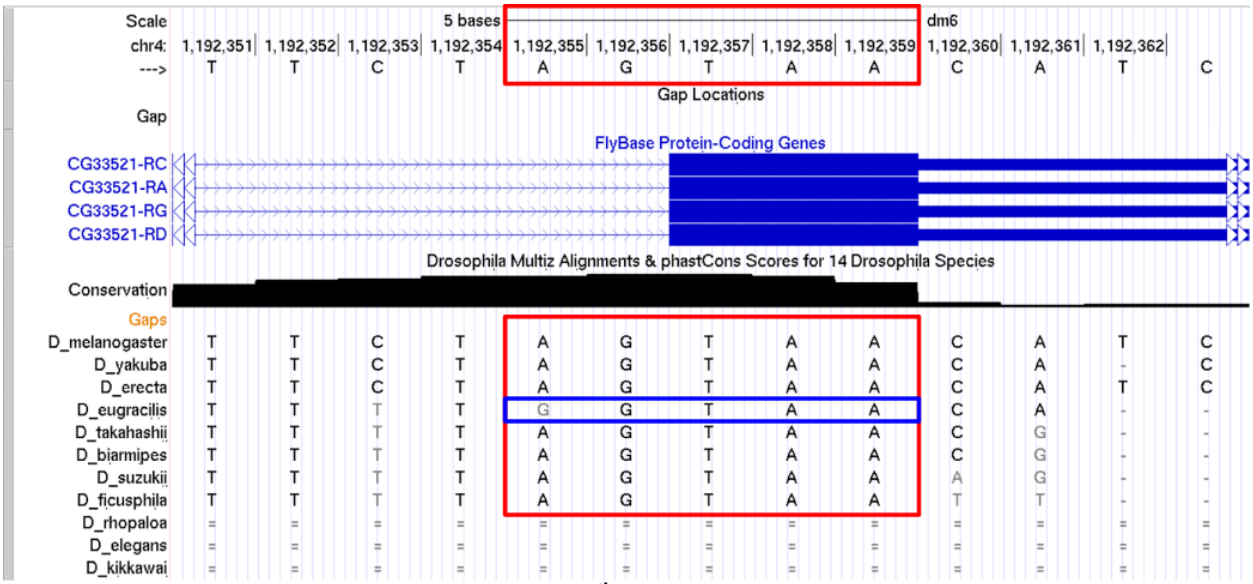


Figure 30: The AG splice acceptor site of the 10th exon is not conserved in *D. eugracilis*. The AG motif has mutated to a GG. This confirms the observation that this small exon is not found in *D. eugracilis* but is found in nearby species.

After examining the aforementioned data, the nine-exon model of *CG33521*-PC was confirmed as the best-supported hypothesis. The ninth exon's coordinates, as well as the coordinates of the first eight exons of *CG33521*-PC were organized into Table 1.

Name of exon in *D. melanogaster*	Start/splice donor	Stop/splice acceptor	Frame	Acceptor Phase	Donor Phase	Exon BLASTp e-score
1_12124_0	1688	1804	+2	NA	0	4e-11
3_12124_0	2006	2056	+2	0	0	4e-08
4_12124_0	2210	2355	+2	0	2	6e-22
5_12124_1	2682	2842	+1	1	1	3e-27
6_12124_2	3168	3824	+2	2	1	3e-84
7_12124_2	3885	4015	+2	2	0	1e-08
8_12124_0	4075	4551	+1	0	0	1e-59
10_12124_0	5629	5768	+1	0	2	7e-19
11_12124_1	5824	5937	+2 -> +1	1	NA	4e-08

Table 1: Final coding exon annotations of *CG33521*-PC. Using the previously described strategies, the positions of the coding exons of *CG33521*-PC were annotated. The 9th exon's frameshift from +2 to +1 is depicted in the corresponding box in the "Frame" column.

Checking the *CG33521*-PC Gene Model

After gathering the CDS annotations of *CG33521*-PC, the model was tested using the GEP Gene Model Checker (GMC) to confirm that this annotation follows the basic biological rules of CDS annotation. The exons all passed the GMC, and the resulting dot plot and amino acid alignment are shown in Figure 31 and Figure 32.

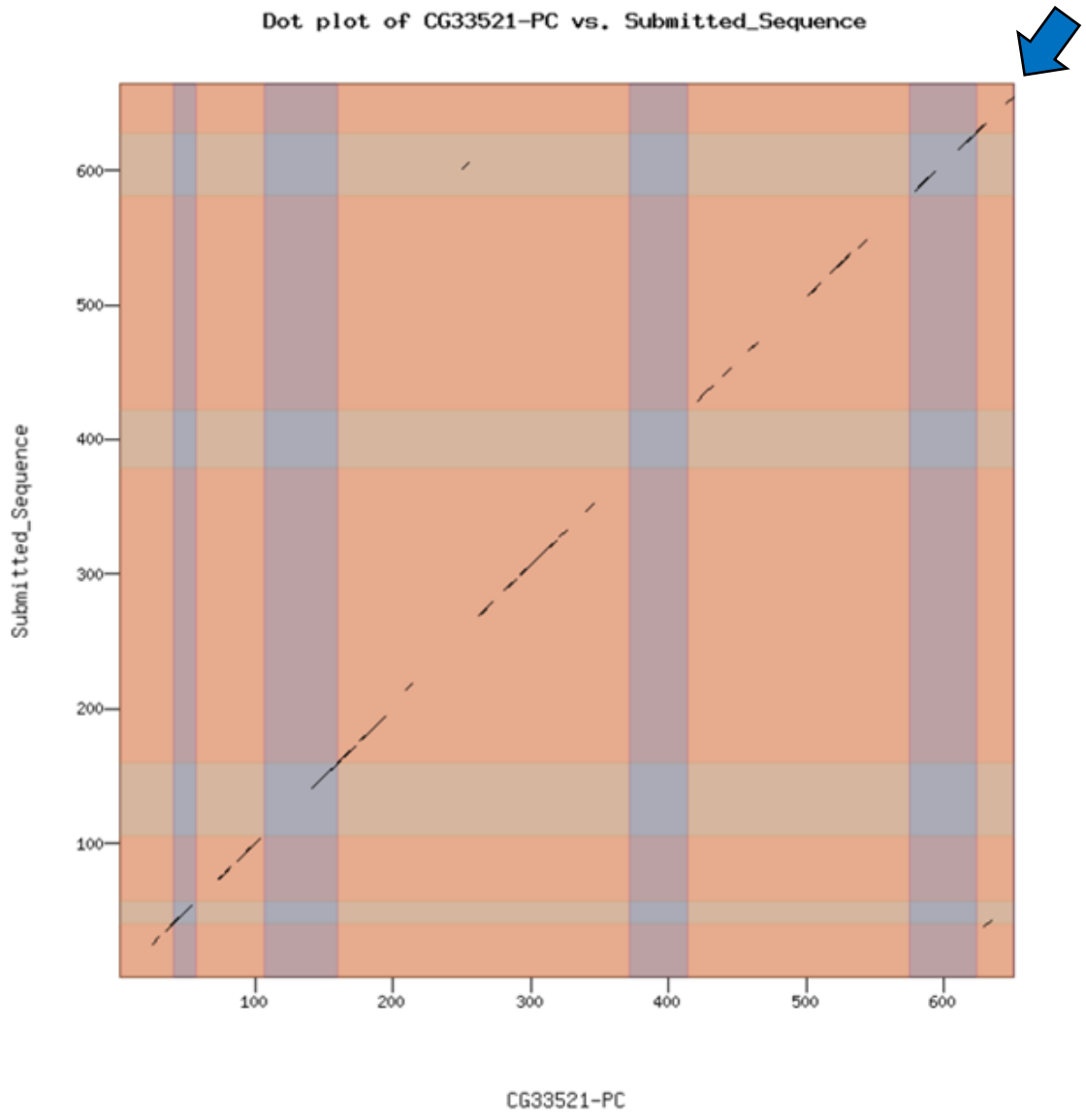


Figure 31: Dot Plot of the CDS annotation of *CG33521*-PC from *D. melanogaster* against the gene model for Feature 1 of *D. eugracilis* contig12. This gene is relatively well conserved. At the top right of the plot, the line does not reach exactly the top right corner of the plot (blue arrow). This is to be expected due to the addition of several amino acids to the end of the 9th exon of the *D. eugracilis* gene.

Figure 32: Protein Alignment of *CG33521*-PC annotation. The additional amino acids added to the 9th exon can be seen at the end of this alignment.

Annotation of *CG33521*-PA

After confirming the annotation of *CG33521*-PC (and *CG33521*-PD, having an identical CDS), *CG33521*-PA was examined for annotation (Fig.33). The splice sites of the 2nd exon were determined with the standard annotation strategy used to annotate the coding exons of *CG33521*-PC (Figs. 34 and 35). The exon table was then updated with the annotation of this longer second exon (Table 2).

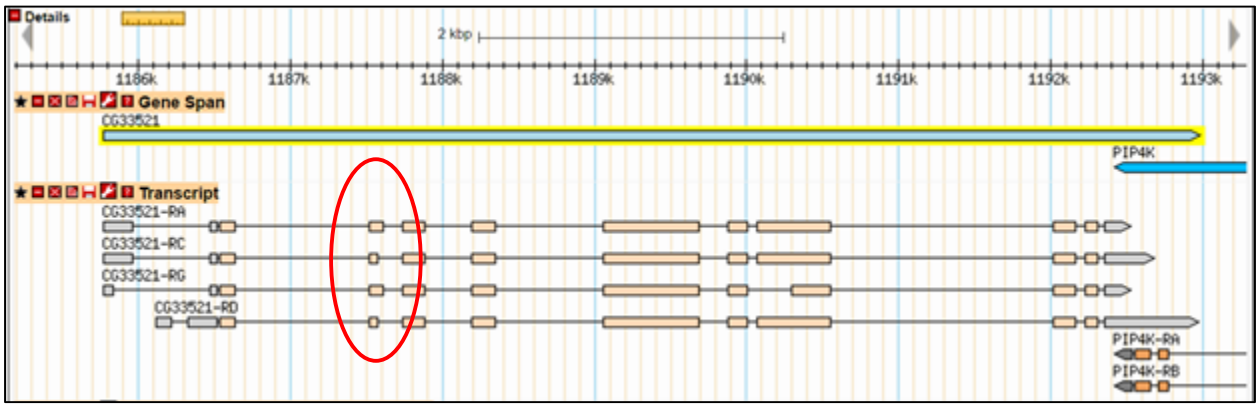


Figure 33: Comparison of *CG33521*-PC/*CG33521-PD* to *CG33521*-PA. The CDS of *CG33521*-PA only differs from the CDS of *CG33521*-PC/*CG33521*-PD at the second exon (shown in the red circle).

CG33521:2_12124_0

Sequence ID: Query_75995 **Length:** 31 **Number of Matches:** 1

Range 1: 1 to 31 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
58.9 bits(141)	9e-16	27/31(87%)	30/31(96%)	0/31(0%)	+2

```
Query  2006  KKSFTKFDELQKRNLIHVRSRDSKKAQEVFR  2098
             KKSFTKFD+LQKRN+IHVRSRDSKKAQ +FR
Sbjct  1     KKSFTKFDDLQKRNVIHVRSRDSKKAQSLFR  31
```

Figure 34: BLASTx search of the second exon of *CG33521*-PA (subject) against contig12 (query). This exon is well-conserved, with an e-score of 9e-16. This is an acceptably low score considering the short size of this exon. The aligned sequence in contig12 is in frame +2.

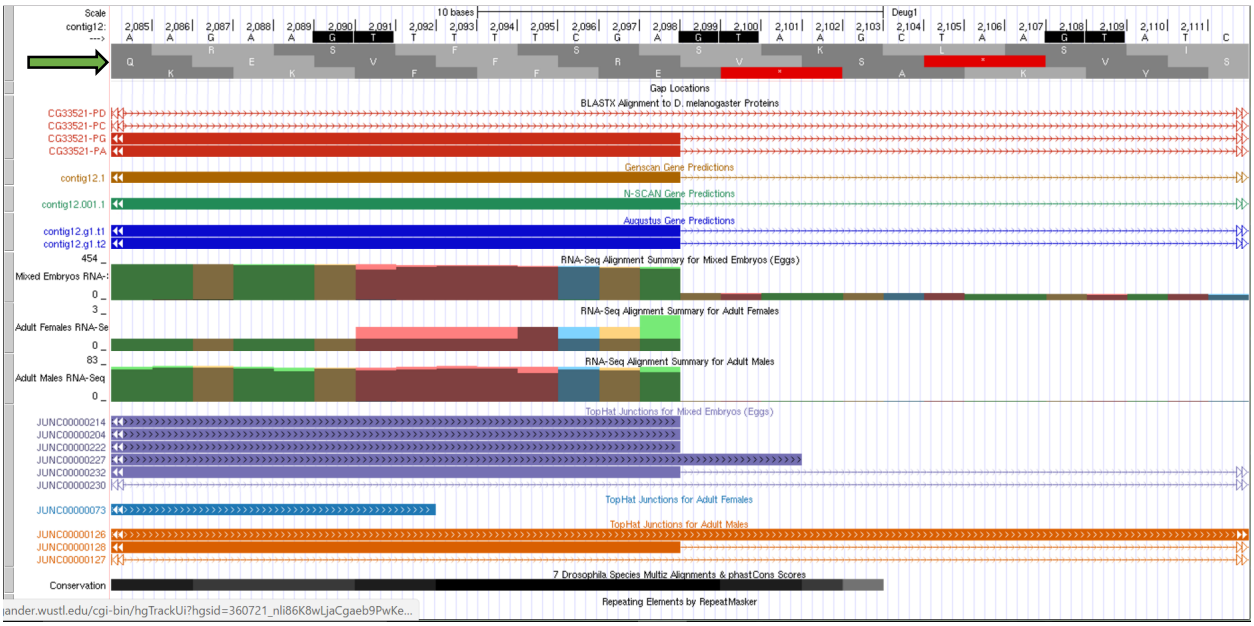


Figure 35: *CG33521*-PA second exon splice donor site. This exon is in frame +2 (indicated by the green arrow). There are three GT's that predict splice sites in the correct phase (phase zero). The most upstream of these is not supported by any of the evidence tracks. The most downstream of these predicts an in frame stop codon. Thus, the middle GT is most likely corresponds to the putative splice donor site. It is supported by the BLASTx alignment track, computer-based gene predictions, RNA-Seq data, and Tophat junctions.

Name of exon in *D. melanogaster*	Start/splice donor	Stop/splice acceptor	Frame	Acceptor Phase	Donor Phase	Exon BLASTp e-score
1_12124_0	1688	1804	+2	NA	0	4e-11
2_12124_0	2006	2098	+2	0	0	9e-16
4_12124_0	2210	2355	+2	0	2	6e-22
5_12124_1	2682	2842	+1	1	1	3e-27
6_12124_2	3168	3824	+2	2	1	3e-84
7_12124_2	3885	4015	+2	2	0	1e-08
8_12124_0	4075	4551	+1	0	0	1e-59
10_12124_0	5629	5768	+1	0	2	7e-19
11_12124_1	5824	5937	+2 -> +1	1	NA	4e-08

Table 2: Final coding exon annotations of *CG33521*-PA. This table is identical to the table of coding exons of *CG33521*-PC except for the longer second exon in *CG33521*-PA.

The coding exons of *CG33521*-PA were tested using the GMC where they passed the provided checks of the basic biological rules of annotation. The GMC produced the Dot Plot shown in Figure 36 and the Protein Alignment shown in Figure 37.

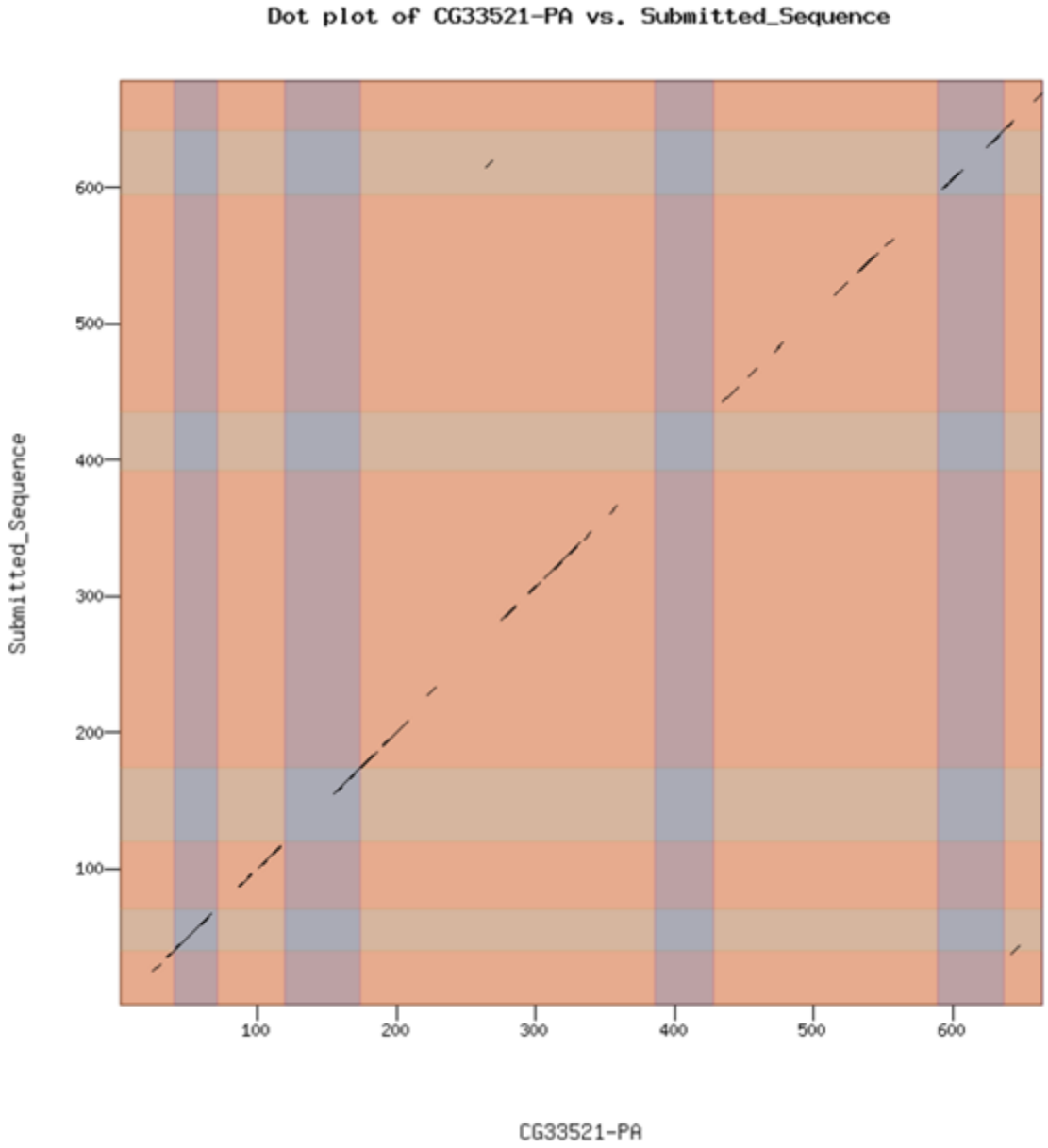


Figure 36: Dot Plot of *D. melanogaster CG33521*-PA vs. the proposed model of *D. eugracilis CG33521*-PA.

Alignment of CG33521-PA vs. Submitted_Seq

View plain text version

Identity: 444/682 (65.1%), **Similarity:** 527/682 (77.3%), **Gaps:** 23/682 (3.4%)

```
CG33521-PA          1 MDSLNPQSSKISFGNGLPLKKKKTKKSKTNEYDNQNINMKKSFTKFDDLQKRNVIHVRSR 60
                      **:**.: :  *  *.** ** *:*****.:**.*************:*****:******
Submitted_Seq       1 MDTLNTNITFESSENSLPSKKGKSKKSKTTKYDSQNINMKKSFTKFDELQKRNLIHVRSR 60

CG33521-PA         61 DSKKAQSLFRSTIPEKVENCHQCKKPVYKMEEVILSLKTATTIFHKTCLRCKDCGKHLKF 120
                      ******.:**** .*: :.*:**:*.*********.***.****** *********:**
Submitted_Seq      61 DSKKAQEVFRSTTAEHPDRCRQCNKTVYKMEEVILQLKTGTTIFHKMCLRCKDCGKQLKS 120

CG33521-PA        121 DSYNVHDGSLYCSMHFKLIFAPKVVYEEFTPRKAELIIRENQPIKLPPDVAKASDKPSLG 180
                      ****.***:.***.*** **:**.****:*.**.**********.******.********
Submitted_Seq      121 DSYNIHDGTFYCSIHFKSIFSPKIVYEEITTRKPELIIRENQPIELPPDVARASDKPSLG 180

CG33521-PA        181 LDELQELNLRSKFKVFENGYEEHNNNLRERQDI-----AITHSKSIQSTLTKFHGLGIPN 235
                      *****:*::********** :******:****      ***:.***:*****:: ***.*
Submitted_Seq      181 LDELQQLDVRSKFKVFENGCKEHNNNLQERQDNITHCNAITQNKSIRSTLTKLQKLGITN 240

CG33521-PA        236 SELTKLDDKNSDN--NSDGDGDMNFMCLKKEIERETPVGLGEAMNDIRSKFEQGDLMAKE 293
                      **  **.* *: *  *:*.::* ::::  :*:**** * ***:**********.*: * **
Submitted_Seq      241 SEPRKLSDINTRNDLNTDDESDTDILYSRKDIERERPQGLGDAMNDIRSKFEHGQAMLKE 300

CG33521-PA        294 GRREERKQEIQNIRSRLFLGKQAKIKEMYKLAVAESEQPVTSVGKTPDICAIKPTQEIKN 353
                       ********:*.****************:********  .*****.** .*..**:**:
Submitted_Seq      301 ERREERKQELQSIRSRLFLGKQAKIKEMYQLAVAESEQRGNSVGKTSDINVITATQQIKD 360

CG33521-PA        354 RFENGEVYKDSKILSSEVSCGIHEDADVFESGISKASRNIFMKLDANIKSGLSNH-VQYT 412
                      *****:*::*.:* *.*   *:: * :****.***:**.***::****.*  *:  *: .
Submitted_Seq      361 RFENGDVFNDNRIQSKEPIFGMQADENVFESAISKSSRSIFMEMDANISSISSKSCVKNV 420

CG33521-PA        413 LPDKKYQIHNQKVQENSDVEIVKSDSKPEEVKVATEELSKKFKFFETYSPAENKKRIFRM 472
                      .***   ***  ****:*:::***** *****:****:*:***** *** .:**: * *
Submitted_Seq      421 QSDKKILYHNQMCQENSNVDVIKSDSKQEEVKVTTEELAKRFKFFEEYSPDRKKKKKFCM 480

CG33521-PA        473 TPPREGVVMFPPPDSETNQQISTTLFNDNILQKTKTTSTILNKFREMEEQKMSDQQKKKN 532
                      *****.*  : . * :*:  :*  **:*.:*::.****:***.********:.*::*::*
Submitted_Seq      481 TPPREDVQNLLAIDLDTD--VSHDLFDDTVLNNTKTTTTILTKFREMEEQKLNDKNKERN 538

CG33521-PA        533 PKPLKCFTPPPEISHQFIRSDTEEESDSDYEQNSENDEESEIN-PSNSYYNDKALLEAQS 591
                      *************.:: ::*****:. ** ****::***.  * *.***::*:** ***.
Submitted_Seq      539 PKPLKCFTPPPEIDNH-LKSDTEEDNYSDSEQNSDDDEEEFENVPPNSYHKDEALYEAQN 597

CG33521-PA        592 VARAKQLRAKFEKWQNNEIEQEIKEGRIDVYSQLISNESIESAKAIRERFENMKKSETAM 651
                       ************** ****:*::*** :* *.***********:*******::: :* :
Submitted_Seq      598 SARAKQLRAKFEKWQINEIERELNEGRENV-SKLISNESIESAKTIRERFENLNNFDTHL 656

CG33521-PA        652 ENPSKTQIKRFV---------- 663
                      **..*::*****
Submitted_Seq      657 ENTQKAEIKRFVVSEKKINQSI 678
```

Figure 37: Protein Alignment of *CG33521*-PA annotation.

Annotation of *CG33521*-PG

After confirming the annotation of the *CG33521*-PA ortholog, the last remaining isoform, *CG33521*-PG, was examined for annotation. The only coding exon in this isoform that differs from *CG33521*-PA is exon seven. This can be seen in Figure 38.

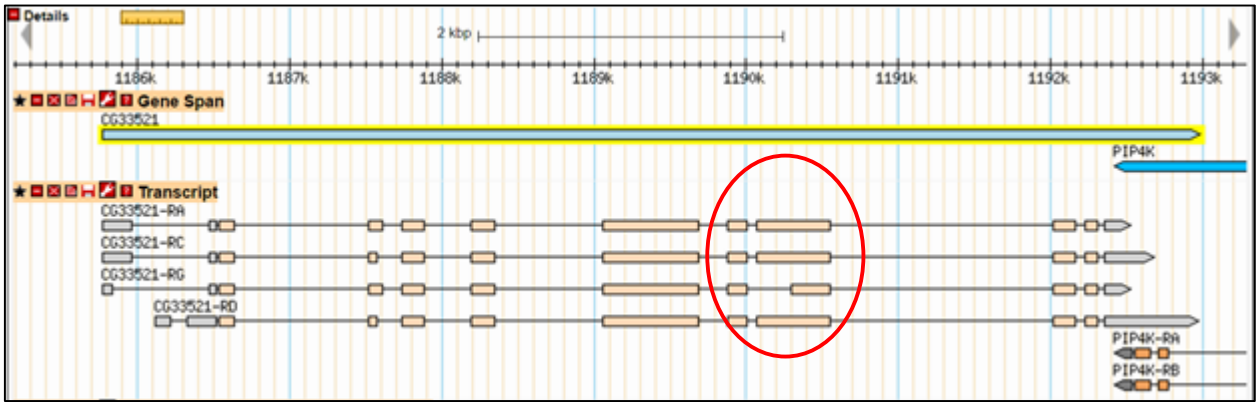


Figure 38: Comparison of *CG33521*-PA to *CG33521*-PG in *D. melanogaster*. *CG33521*-PG only differs from *CG33521*-PA in that its 7th exon is shorter (shown in red circle).

CG33521:9_12124_0

Sequence ID: Query_218421 **Length:** 84 **Number of Matches:** 1

Range 1: 2 to 84 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
112 bits(279)	2e-33	55/84(65%)	70/84(83%)	2/84(2%)	+1

```
Query  4303  TKTTTTILTKFREMEEQKLNDKNKERNPKPLKCFTPPPEIDNH-LKSDTEEDNYSDSEQN  4479
             TKTT+TIL KFREMEEQK++D+ K++NPKPLKCFTPPPEI +  ++SDTEE++ SD EQN
Sbjct  2     TKTTSTILNKFREMEEQKMSDQQKKKNPKPLKCFTPPPEISHQFIRSDTEEESDSDYEQN  61

Query  4480  SDDDEEEFENVPPNSYHKDEALYE  4551
             S++DEE   N P NSY+ D+AL E
Sbjct  62    SENDEESEIN-PSNSYYNDKALLE  84
```

Figure 39: BLASTx alignment of *D. melanogaster CG33521*-PG exon seven (subject) against contig12 (query). This alignment has an e-score of 2e-33 and the aligned sequence is in frame +1.

The BLASTx search performed to annotate this isoform produced a relatively strong alignment with an e-score of 2e-33 (Fig.39). Viewing the start of the BLASTx alignment in the Genome Browser shows a region with two possible AG splice acceptor sites in the correct phase (Fig.40). Because the putative splice site occurs in a region belonging to another coding exon, RNA-Seq data cannot be used to confirm the position of the splice acceptor site. Furthermore, there are no TopHat junctions predicting a splice site in this region. With no other evidence available, the preferred annotation strategy is to avoid truncation of conserved sequences. Therefore, the upstream AG splice site was selected as the splice site for this exon to avoid losing the conserved amino acids that occur starting at 4,282, before the next AG splice site.

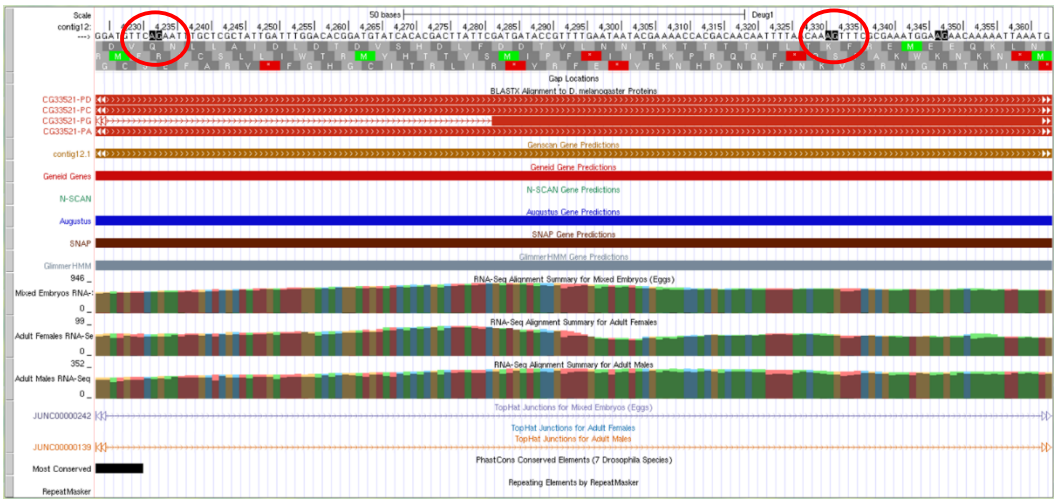


Figure 40: There are two AG splice sites on either side of the BLASTx alignment that are in the correct phase (phase zero). These AG splice sites are indicated by the red circles.

This annotation was confirmed to follow basic biological rules in the GEP Gene Model Checker. It is also possible that this isoform is not expressed in *D. eugracilis*. Examination of the annotation of this gene in the *D. biarmipes* BCM-HGSC Genome Browser showed that this isoform exists in *D. biarmipes,* a nearby species. Furthermore, the preservation of the current

number of isoforms is a more parsimonious hypothesis than the removal of *CG33521*-PG. As there is no positive evidence that the isoform does not exist in *D. eugracilis,* this isoform was retained in the final gene model of *CG33521* in *D. eugracilis,*. This annotation of *CG33521*-PG produced the set of coding exons found in Table 3. The Dot Plot and Protein Alignment generated by this annotation are shown in Figure 41 and Figure 42.

Name of exon in *D. melanogaster*	Start/splice donor	Stop/splice acceptor	Frame	Acceptor Phase	Donor Phase	Exon BLASTp e-score
1_12124_0	1688	1804	+2	NA	0	4e-11
2_12124_0	2006	2098	+2	0	0	9e-16
4_12124_0	2210	2355	+2	0	2	6e-22
5_12124_1	2682	2842	+1	1	1	3e-27
6_12124_2	3168	3824	+2	2	1	3e-84
7_12124_2	3885	4015	+2	2	0	1e-08
9_12124_0	4234	4551	+1	0	0	2e-33
10_12124_0	5629	5768	+1	0	2	7e-19
11_12124_1	5824	5937	+2 -> +1	1	NA	4e-08

Table 3: Final coding exon annotations of *CG33521*-PG. This table is identical to the table of coding exons of *CG33521*-PA except for the variation in the splice acceptor site of the seventh exon.

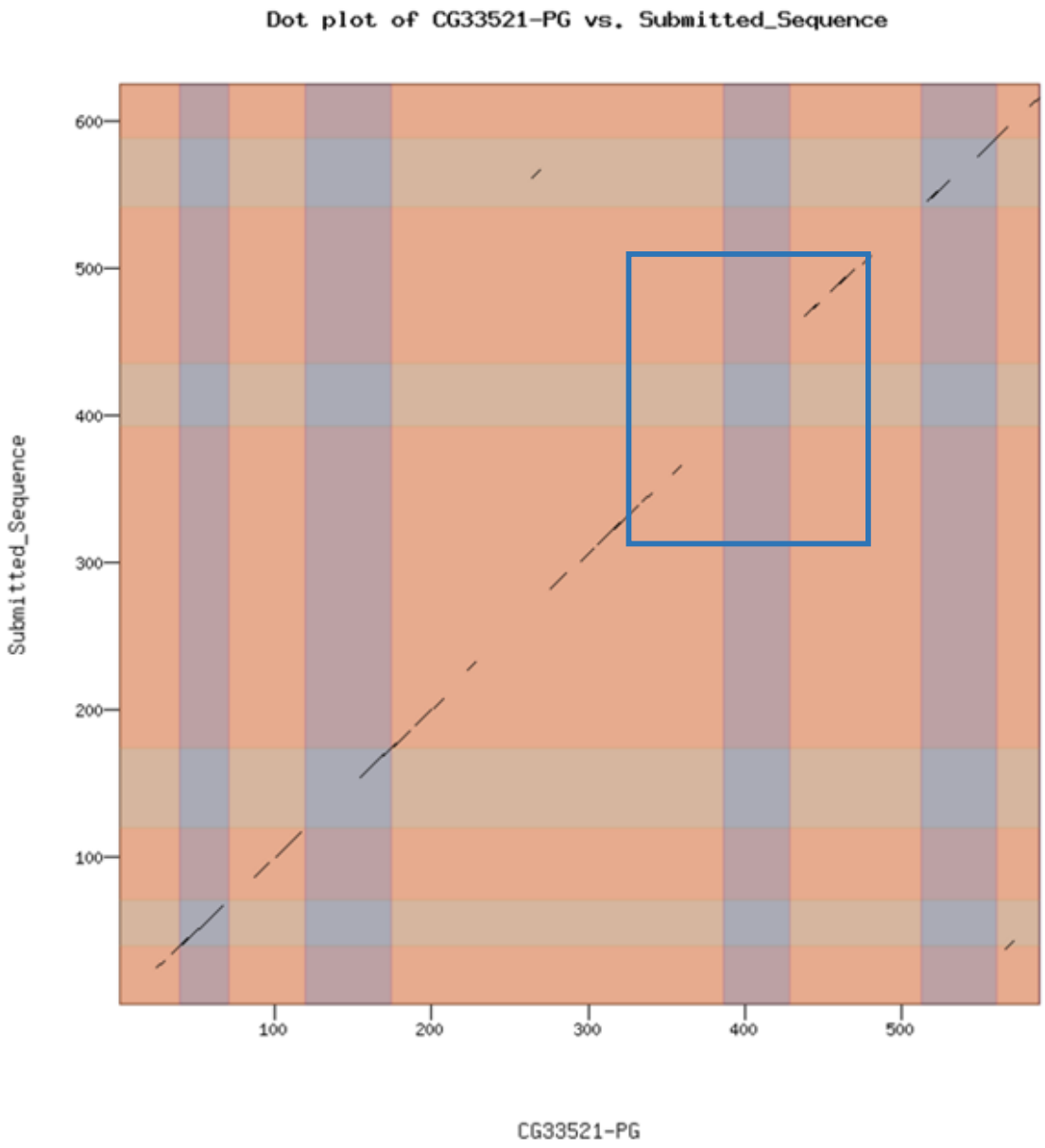


Figure 41: Dot Plot of *D. melanogaster CG33521*-PG amino acids vs. the proposed gene model of *D. eugracilis CG33521*-PG . The 7th exon of this amino acid alignment can be seen to be shifted slightly out of position (off diagonal, see blue box). This is expected, due to the additional bases incorporated into the start of the 7th exon in the *D. eugracilis* model.

Alignment of CG33521-PG vs. Submitted_Seq

View plain text version

Identity: 401/627 (64.0%), **Similarity:** 472/627 (75.3%), **Gaps:** 43/627 (6.9%)

```
CG33521-PG          1 MDSLNPQSSKISFGNGLPLKKKKTKKSKTNEYDNQNINMKKSFTKFDDLQKRNVIHVRSR 60
                      **:**.: :  *  *.** ** *:*****.:**.************:*****:******
Submitted_Seq       1 MDTLNTNITFESSENSLPSKKGKSKKSKTTKYDSQNINMKKSFTKFDELQKRNLIHVRSR 60

CG33521-PG         61 DSKKAQSLFRSTIPEKVENCHQCKKPVYKMEEVILSLKTATTIFHKTCLRCKDCGKHLKF 120
                      ******.:****  .*: :.*:**:*.********.***.***** ********:**
Submitted_Seq      61 DSKKAQEVFRSTTAEHPDRCRQCNKTVYKMEEVILQLKTGTTIFHKMCLRCKDCGKQLKS 120

CG33521-PG        121 DSYNVHDGSLYCSMHFKLIFAPKVVYEEFTPRKAELIIRENQPIKLPPDVAKASDKPSLG 180
                      ****:***.::***:*** **.**:****:*.**.**********:******.********
Submitted_Seq      121 DSYNIHDGTFYCSIHFKSIFSPKIVYEEITTRKPELIIRENQPIELPPDVARASDKPSLG 180

CG33521-PG        181 LDELQELNLRSKFKVFENGYEEHNNNLRERQDI-----AITHSKSIQSTLTKFHGLGIPN 235
                      *****:*::********** :******:****      ***:.***:*****:: ***.*
Submitted_Seq      181 LDELQQLDVRSKFKVFENGCKEHNNNLQERQDNITHCNAITQNKSIRSTLTKLQKLGITN 240

CG33521-PG        236 SELTKLDDKNSDN--NSDGDGDMNFMCLKKEIERETPVGLGEAMNDIRSKFEQGDLMAKE 293
                      **  **.* *: *  *:*.:.* :::  :*:**** * ***:**********:*: * **
Submitted_Seq      241 SEPRKLSDINTRNDLNTDDESDTDILYSRKDIERERPQGLGDAMNDIRSKFEHGQAMLKE 300

CG33521-PG        294 GRREERKQEIQNIRSRLFLGKQAKIKEMYKLAVAESEQPVTSVGKTPDICAIKPTQEIKN 353
                       ********:*.****************:********  .*****.** .*..**:**:
Submitted_Seq      301 ERREERKQELQSIRSRLFLGKQAKIKEMYQLAVAESEQRGNSVGKTSDINVITATQQIKD 360

CG33521-PG        354 RFENGEVYKDSKILSSEVSCGIHEDADVFESGISKASRNIFMKLDANIKSGLSNH-VQYT 412
                      *****:*::*.:* *.*   *:: * .**** ***:** ***::****.*  *:  *:.
Submitted_Seq      361 RFENGDVFNDNRIQSKEPIFGMQADENVFESAISKSSRSIFMEMDANISSISSKSCVKNV 420

CG33521-PG        413 LPDKKYQIHNQKVQEK----------------------TKTTSTILNKFREMEEQKMSDQ 450
                       .***   **   **:                     ****:**.*********:.*:
Submitted_Seq      421 QSDKKILYHNQMEQENLLAIDLDTDVSHDLFDDTVLNNTKTTTILTKFREMEEQKLNDK 480

CG33521-PG        451 QKKKNPKPLKCFTPPPEISHQFIKSDYEEESDSDYEQNSENDEESEIN-PSNSYYNDKAL 509
                      :*::**************.:: ::*****:. ** ****::***.  * *.***::*:**
Submitted_Seq      481 NKERNPKPLKCFTPPPEIDNH-LKSDTEEDNYSDSEQNSDDDEEEFENVPPNSYHKDEAL 539

CG33521-PG        510 LEAQSVARAKQLRAKFEKWQNNEIEQEIKEGRIDVYSQLISNESIESAKAIRERFENMKK 569
                       ***. ************* ****:*::*** :* *:**********:*******:::
Submitted_Seq      540 YEAQNSARAKQLRAKFEKWQINEIERELNEGRENV-SKLISNESIESAKTIRERFENLNN 598

CG33521-PG        570 SETAMENPSKTQIKRFV---------- 586
                       :* :**..*::*****
Submitted_Seq      599 FDTHLENTQKAEIKRFVVSEKKINQSI 625
```

Figure 42: Protein Alignment of *CG33521*-PG annotation. The gap resulting from the addition of amino acids in the annotation can be seen in the alignment of the 7th exon (black circle).

After confirming the annotations of all four isoforms, work was begun to annotate the TSSs of *CG33521*.

Identification of *D. melanogaster* Transcription Start Sites

The first step in the TSS annotation workflow is to survey the TSSs of the orthologous gene in *D. melanogaster*. Previous work done in contig12 of *D. eugracilis* has confirmed that the furthest upstream gene in this project is *CG33521* and has produced annotations of the coding DNA sequences for all four isoforms of this gene. In *D. melanogaster,* the untranslated exons of this gene have been annotated as well. Examination of the *D. melanogaster* records at FlyBase GBrowse and the Gene Record Finder show that the *CG33521* gene has two unique TSSs (Figs. 43 and 44).

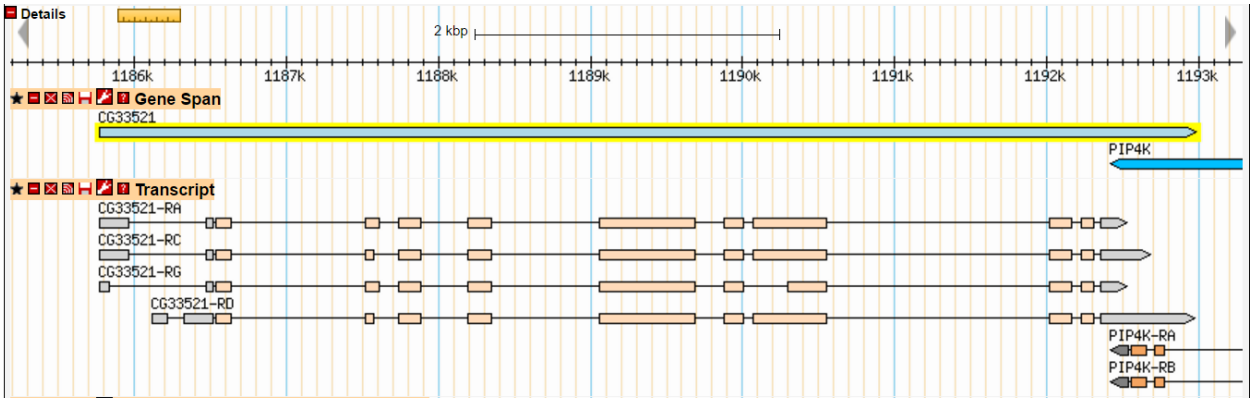


Figure 43: The annotation of *CG33521* in *D. melanogaster* indicates two unique TSSs. The *CG33521* gene in *D. melanogaster* generates four isoforms. Visual inspection of the 5' untranslated exons in FlyBase GBrowse suggests that the first three isoforms share a TSS and that *CG33521*-RD has a unique TSS.

A

Exon usage map:

Isoform	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
CG33521-RC	1				2		3	4	5	6	7	8		9	10		11	
CG33521-RA	1				2	3		4	5	6	7	8		9	10			11
CG33521-RG		1			2	3		4	5	6	7		8	9	10			11
CG33521-RD			1	2			3	4	5	6	7	8		9	10	11		

Select a row to display the corresponding exon sequence:

FlyBase ID	5' Start	3' End	Strand	Size (bp)
1	1,185,768	1,185,966	+	199

B

Exon usage map:

Isoform	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
CG33521-RC	1				2		3	4	5	6	7	8		9	10		11	
CG33521-RA	1				2	3		4	5	6	7	8		9	10			11
CG33521-RG		1			2	3		4	5	6	7		8	9	10			11
CG33521-RD			1	2			3	4	5	6	7	8		9	10	11		

Select a row to display the corresponding exon sequence:

FlyBase ID	5' Start	3' End	Strand	Size (bp)
2	1,185,768	1,185,834	+	67

Figure 44: *CG33521*-RC, *CG33521*-RA (A), and *CG33521*-RG (B) all share the same TSS at 1,185,768. Examination of the initial 5' transcribed exons of these isoforms confirms that they possess an identical TSS in *D. melanogaster* (red circles).

Classification of the TSS of *CG33521*-RC in *D. melanogaster*

CG33521-RC and *CG33521*-RA have a longer initial untranslated exon than *CG33521*-RG, making those exons better candidates for use in a BLASTn search. For this reason, *CG33521*-RC was selected from among these three as the isoform to be used to find the TSS shared by all three isoforms. The initial untranslated exon of *CG33521*-RC is also longer (and thus more likely to produce a useful BLASTn alignment) than *CG3352*-RD. Therefore, this TSS was selected for annotation first.

Examination of *CG33521*-RC and the corresponding evidence tracks (annotated TSSs and DHS positions) in the *D. melanogaster* Genome Browser identified the promoter of this TSS as intermediate (Fig. 45). There are no DHS positions in the available cell types and two Celniker TSSs identified that correspond to the start of this exon.

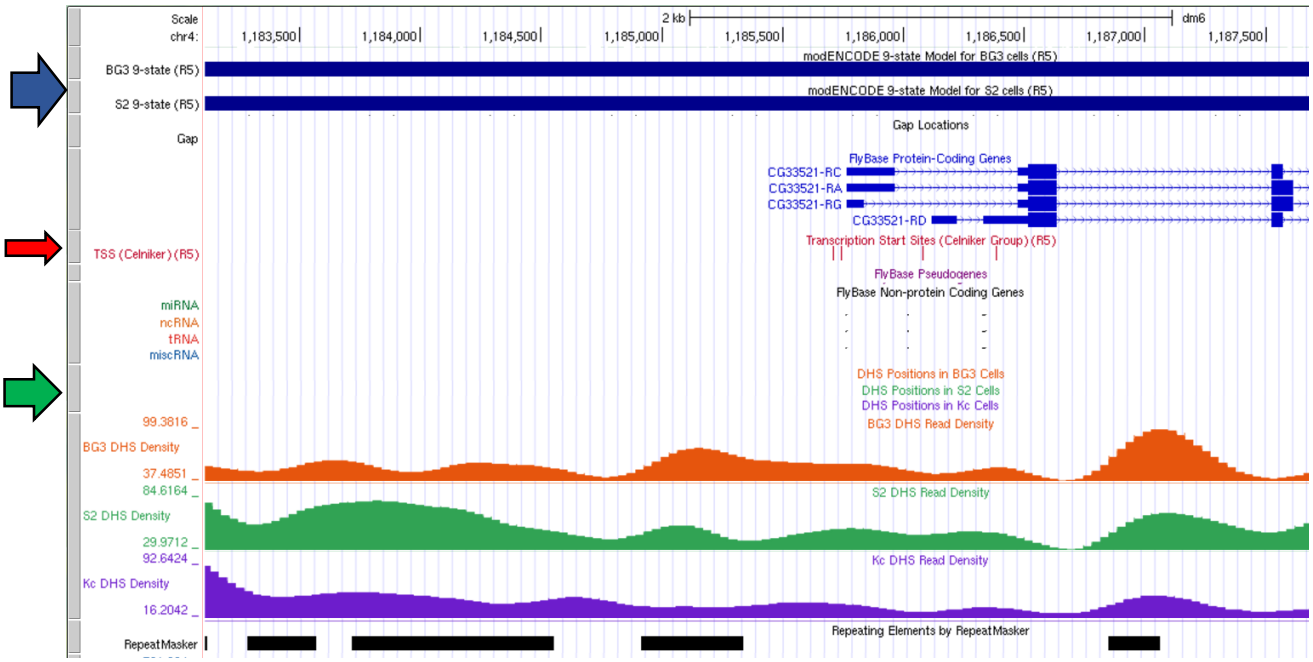


Figure 45: Identification of the *CG33521*-RC promoter as intermediate. There are no DHS positions that correspond to this TSS in *D. melanogaster* (green arrow), however there are two annotated TSSs that appear to correspond to the 5' end of the first exon of *CG33521*-RC (red arrow). Having multiple annotated TSSs with no DHS positions meets the criteria for an intermediate promoter. Interestingly, the two 9-state chromatin models suggest continuous heterochromatin instead of the expected TSS region (blue arrow). This heterochromatic region in the 9-state model suggests that this gene may be inactive in these cell types and is consistent with the lack of a DHS positions in the three displayed cell lines.

While enough evidence is available on the Genome Browser to classify the promoter of this TSS, the evidence tracks show two unexpected results. The first is the increase in DHS read density downstream of the TSS; however, this is associated with the repetitious element identified by RepeatMasker and is most likely to be spurious. This element was classified as a DNA Helitron by performing a search of the sequence in RepBase, a database of repeats found in eukaryotic DNA. The other unexpected result shown on the Genome Browser is the lack of a red region on 9-state models for BG3 cells and S2 cells. One would expect to find a red region

indicating an active promoter if this gene is active; yet, the entire region around the TSS of this gene is blue, indicating that the region is in a heterochromatic state in the cell types used, BG3 and S2. Due to the abundance of data gathered on the *D. melanogaster* genome, there are other tracks that were consulted to provide evidence that *CG33521*-RC is active in *D. melanogaster* in other cell types (Figs. 46 and 47).

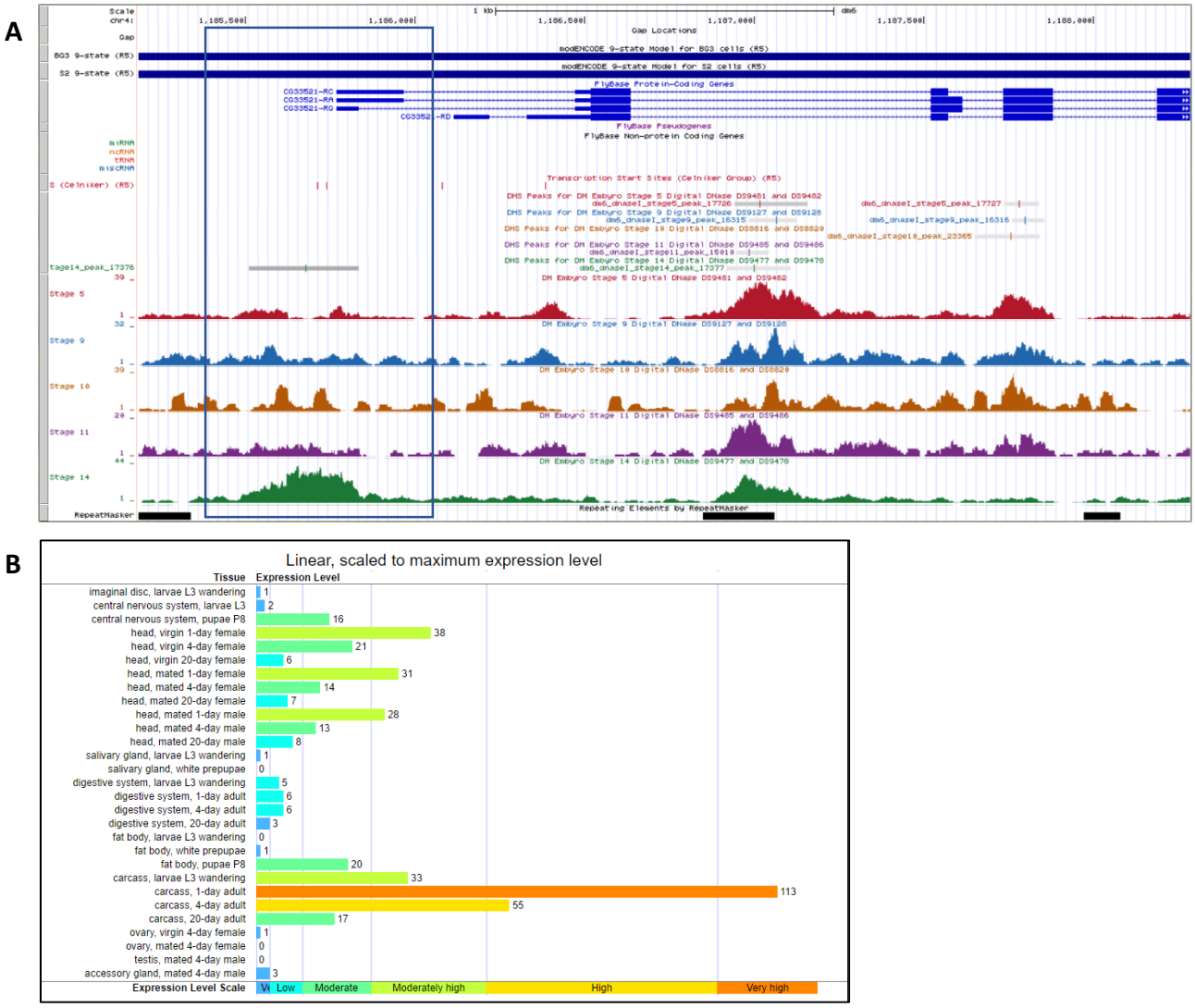


Figure 46: CG33521 appears to be selectively expressed. Analysis of the DHS read density of *D. melanogaster* at different stages of development reveals a DHS peak in stage 14 that appears to correspond to the annotated TSS of *CG33521*-RC (blue box). This suggest that the promoter of this TSS is most highly accessible in developmental stage 14 (A). Examination of *CG33521* expression data from the FlyBase record shows that this gene is highly expressed in the carcasses of young adults and is not highly expressed elsewhere (B). If the selected cell types do not correspond to these tissues and developmental stages, their chromatin structures corresponding to *CG33521* would not be expected to indicate an actively transcribed gene.

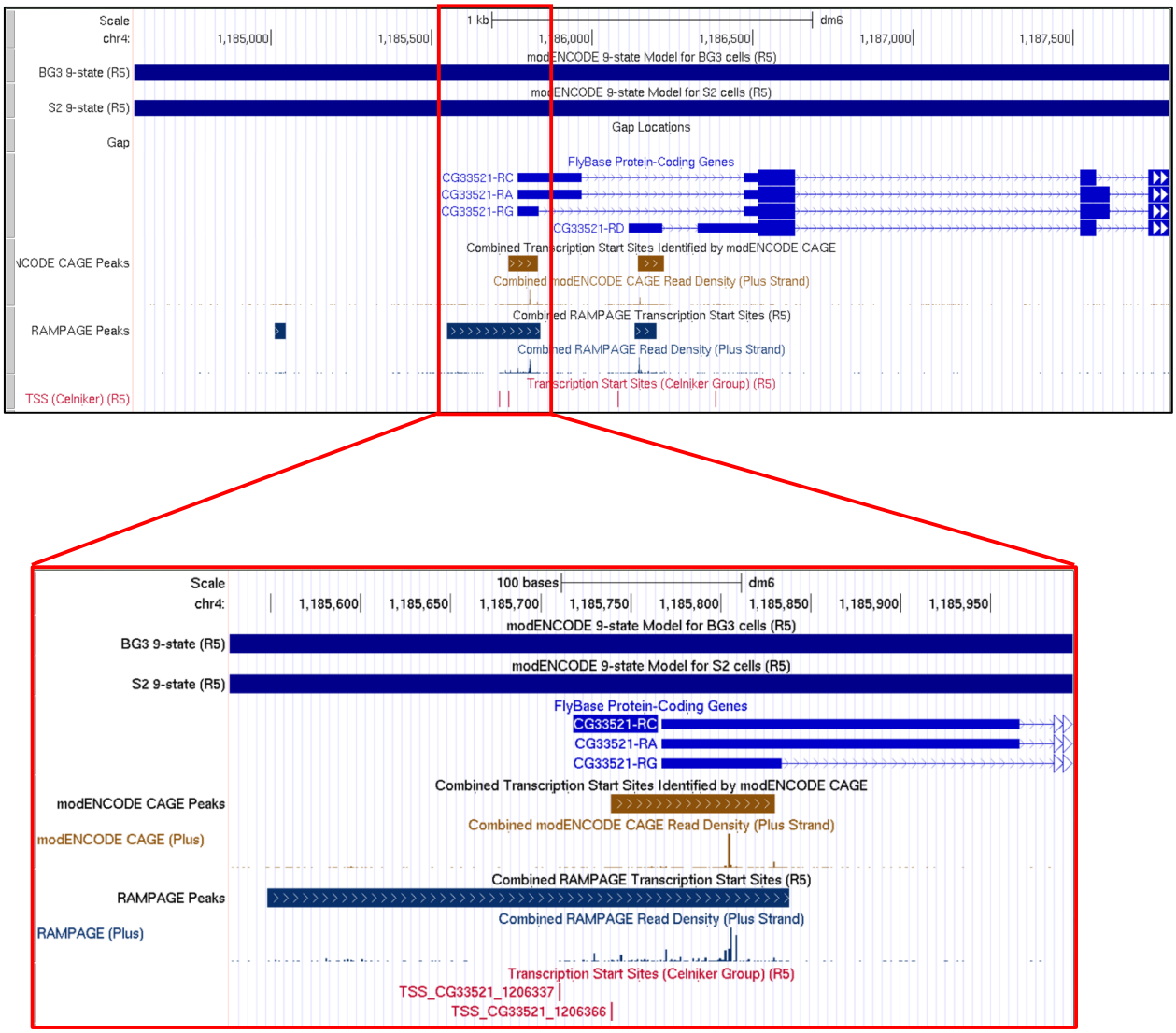


Figure 47: CAGE and RAMPAGE data support the annotated TSS. The annotated plus-strand CAGE and RAMPAGE peaks both suggest a TSS at the start of the annotated initial 5' exon.

CAGE (Cap Analysis Gene Expression) and RAMPAGE (RNA Annotation and Mapping of Promoters for Analysis of Gene Expression) are two methods used to identify transcription start sites. The annotated CAGE and RAMPAGE peaks that align with the annotated TSS of *CG33521*-RC as well as the DHS peak found in stage 14 of development provide evidence that this truly is an active promoter in *D. melanogaster*. Furthermore, there is a single dominant peak in the CAGE sample and two dominant peaks in the RAMPAGE sample, which is consistent with the intermediate classification.

Definition of a Search Region and Putative TSS for *CG33521*-RC in *D. eugracilis*

After obtaining classifying the promoter of *CG33521*-RC as intermediate, a BLASTn search was performed using the initial untranslated exon of *CG33521*-RC as the query and the sequence of contig12 as the subject. Recall that *CG33521*-RC was selected because it contains the longest initial untranslated exon, and thus is the candidate most likely to produce a significant BLASTn alignment; however, no BLASTn alignments were produced from this search when using a threshold e-value of 10. Even without a BLASTn alignment, it is possible to locate the general position of the putative first untranslated exon using RNA-Seq data and TopHat junctions in the *D. eugracilis* Genome Browser (Figs. 48 and 49). RNA polymerase II data has been collected in *D. biarmipes*, a neighboring species of *D. eugracilis*; however, there were no RNA polII peaks found that correspond to the TSS of this gene. In addition, the *Drosophila* multiple sequence alignment of the *Drosophila* conservation track was examined. The sequence of *D. eugracilis* DNA that corresponds to the TSS of *CG33521*-RC in *D. melanogaster* was identified in contig12 with the Short Match function at position 781 (Fig. 50).

Figure 48: *D. eugracilis* RNA-Seq reads map for initial untranslated exon of *CG33521*-RC. Examination of the *D. eugracilis* RNA-Seq reads track shows a transcribed region that suggests the presence of an untranslated exon (red box).

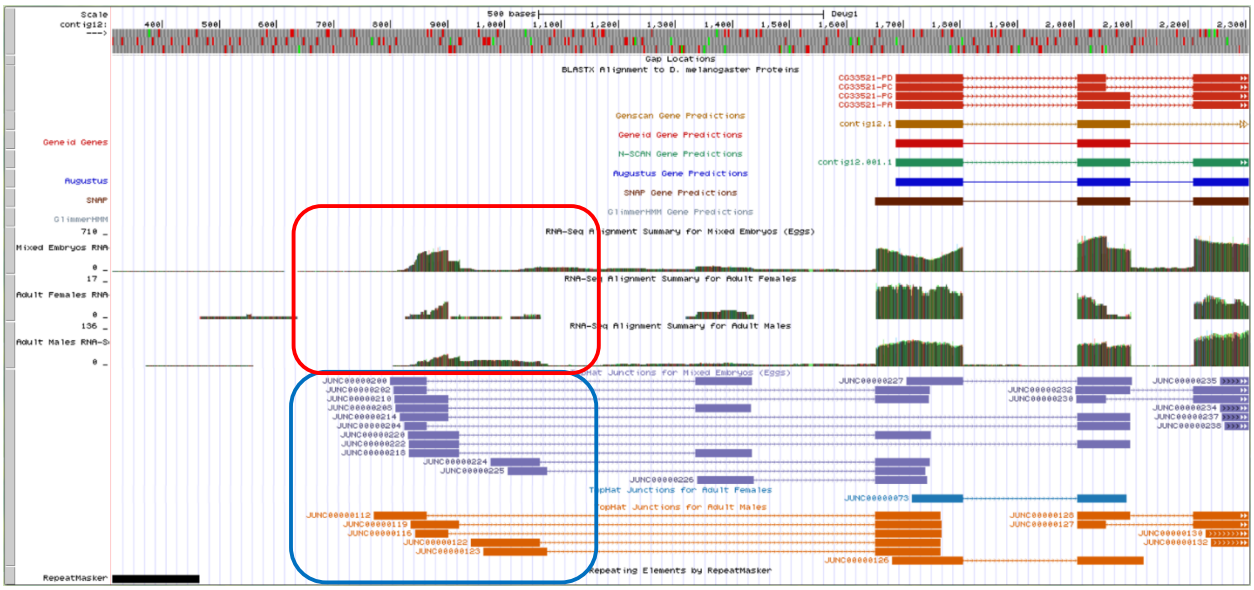


Figure 49: *D. eugracilis* RNA-Seq data and TopHat junctions (score > 9). These evidence tracks show that there is a feature being transcribed in the region where the initial untranslated exon is expected to occur. However, there is not a consensus among these junctions as to where the exon splice occurs (blue box). Furthermore, drop-offs in RNA-Seq data suggest multiple splice sites in this exon (red box).

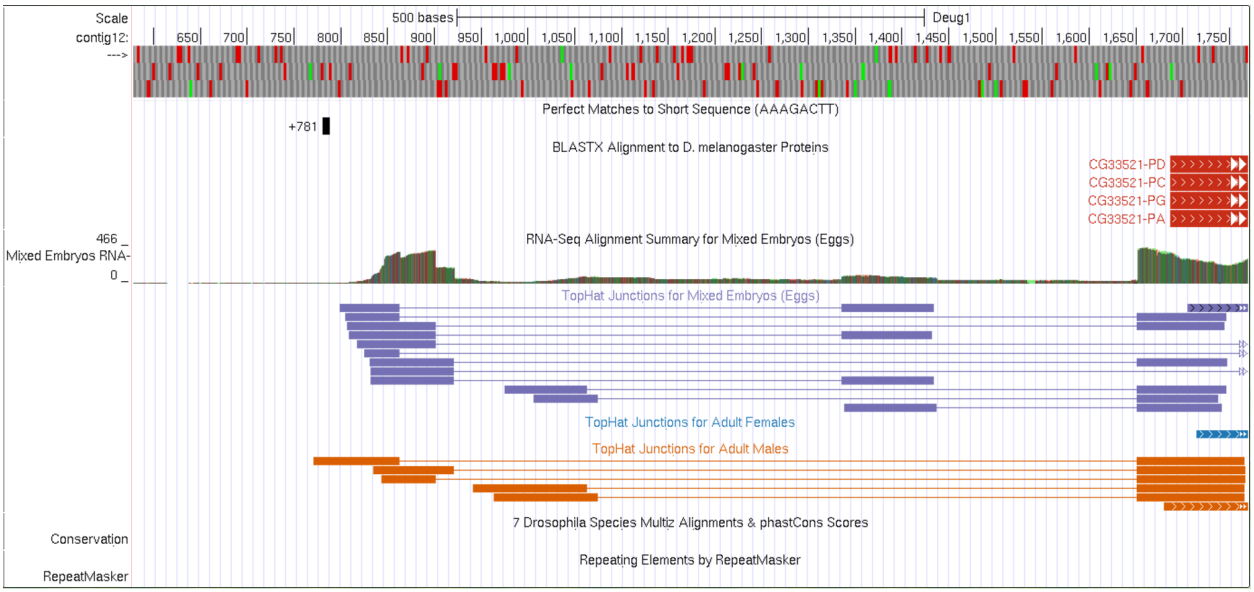


Figure 50: Identification of a putative TSS of *CG33521*-RC with Short Match. With no BLASTn alignment available, the short sequence, AAAGACTT, which is the *D. eugracilis* sequence that corresponds to the start of the initial untranslated exon of *CG33521*-RC in *D. melanogaster,* was mapped back onto contig12 using the Short Match function.

Even though the RNA-Seq data and TopHat junctions predict features occurring in the same general position in the contig, they do not form a consensus as to the donor splice site of the initial untranslated exon. It is important to note that all TopHat junctions with a score less than 10 are not shown in the above screenshot (as well as every image taken from the GEP Genome Browser in this report). One possible explanation for this is that there are alternative splice sites that are unannotated in the *D. melanogaster* model of *CG33521*. This hypothesis was supported by examination of RNA-Seq exon junctions (through GBrowse) in *D. melanogaster* (Fig. 51) as well as by the comments on the gene model found at FlyBase (Fig. 52). If there are additional unannotated exons in *D. melanogaster*, they can provide an explanation for irregularities in RNA-Seq data supporting the putative first exon. Furthermore, some additional unannotated exons could contain novel TSS (such as the one suggested by the exon junction circled in red in Fig. 51), which may be worth investigating at a later time.

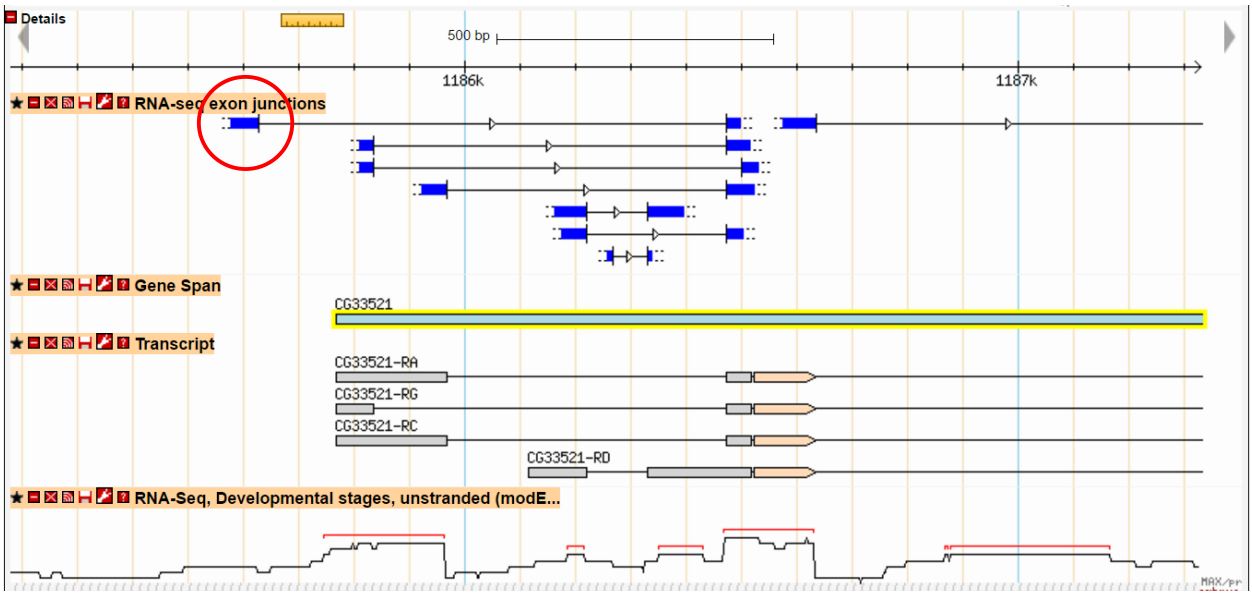


Figure 51: *D. melanogaster* RNA-seq exon junctions. Examination of these exon junctions reveals at least one exon junction (15 reads) near the TSS of *CG33521*-RC (red circle) that does not correspond to the annotated 5' end of *CG33521*.

Comments on Gene Model
Gene model reviewed during 5.55
Gene model reviewed during 5.47
Annotated transcripts do not represent all possible combinations of alternative exons and/or alternative promoters. Low-frequency RNA-Seq exon junction(s) not annotated.
Gene model reviewed during 5.39

Figure 52: Comments on the *D. melanogaster CG33521* gene model. The comments made on this gene model in FlyBase state that there may be unannotated isoforms and that low-frequency RNA-Seq exon junctions were not annotated (blue box).

The splice junctions provide evidence that there is at least one unannotated splice site near the first exon of *CG33521*-RC, which could be contributing to the ambiguity in the RNA-Seq data and TopHat junctions. After taking this information into account, a narrow search region in *D. eugracilis* was defined from the start of the RNA-Seq reads to the first set of 5' splice sites proposed by the TopHat junctions (772-1075); as the putative TSS determined by the Short Match function is only based on the conservation of a few bases, it was not used to define the TSS search region. This search region encompasses the area upstream of the start of the RNA-Seq data, where the TSS would be most likely to occur. Because there was no putative TSS

position available from a BLASTn alignment, the narrow search region was scanned for *Drosophila* core promoter motifs using the "Core Promoter Motifs (Plus)" track in the *D. eugracilis* Genome Browser (Fig. 53).

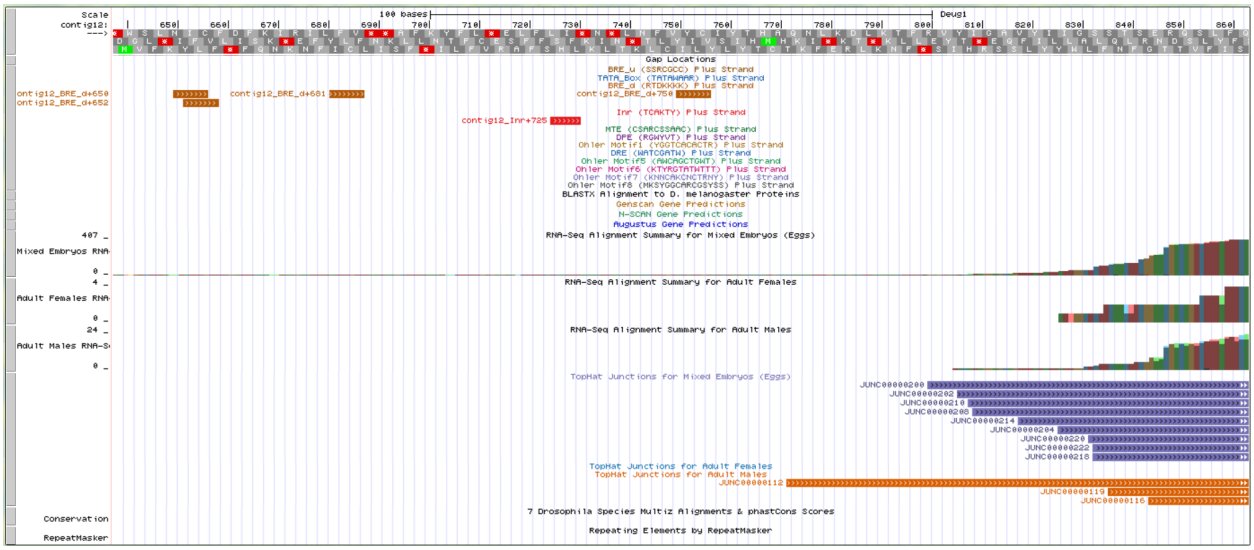


Figure 53: *Drosophila* core promoter motifs in the narrow search region. The "Core Promoter Motifs (Plus)" track in the Genome Browser reveals 5 core promoter motifs that occur in this region. There are four BRE^{d} motifs (brown) and one Inr motif (red). The first 5' splice site predicted by the TopHat junctions serves as the downstream boundary of the search region, thus the splice site cannot be visualized by examining the search region alone in the Genome Browser. The splice sites can be seen in the narrow search region defined in Figure 54.

The four BRE^{d} motifs identified in the *CG33521*-RC narrow search region are likely spurious occurrences of the motif in the sequence, because their complementary motif, the TATA box, is not found in this search region. The Inr motif at 725 may correspond to the TSS, as these motifs are strongly associated with transcription start sites in *D. melanogaster*; however, the motif alone is not enough evidence to declare a putative TSS position and it does not correspond to the putative TSS identified with short match at 781. A note was made of the position of this motif, but there is still not enough available evidence to support a specific position over a TSS search region. A search for the same set of *Drosophila* core promoter motifs was performed in the *D. melanogaster* Genome Browser. Given that there is an annotated TSS of

CG33521-RC in *D. melanogaster,* a motif search region consisting of 300 bp upstream and 300 bp downstream of this start site was used. Because there is no track in the *D. melanogaster* Genome Browser that simultaneously identifies all core promoter motifs, the short match function was used to search for each motif individually. The motif sequences were obtained from http://gander.wustl.edu/~wilson/core_promoter_motifs.html. All the identified core promoter motifs from both species were organized into a table (Table 4). A BRE^d motif was identified that is consistent with a promoter at position 784, three bases away from the putative TSS of 781. However, BRE^d promoters work in tandem with TATA boxes. With no corresponding TATA box identified, one can conclude that this motif is unlikely to correspond to the putative TSS of *CG33521*-RC in *D. eugracilis*.

Core promoter motif	*D. eugracilis*	*D. melanogaster*
BRE^u	NA	NA
TATA Box	NA	NA
BRE^d	650, 652, 681, 750	1185730, 1185780, 1185911, 1185976, 1186010, 1186012
Inr	725	1185564, 1185713, 1185728, 1186022
MTE	NA	NA
DPE	NA	1185562, 1185595
Ohler_motif1	NA	NA
DRE	NA	NA
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	NA	NA
Ohler_motif8	NA	NA

Table 4: Core promoter motifs found in the TSS search regions of *CG33521*-RC in *D. eugracilis* and *D. melanogaster*. Several instances of different motifs were found in these search regions; however, none of the *D. melanogaster* motifs are consistent with the annotated TSS of *CG33521*-RC at 1185768, and thus none are highlighted. The BRE^d motif at position 750 in contig12 is consistent with a TSS that occurs near (position 784) the putative TSS of *CG33521*-RC at position 781. However it is unlikely that this motif is truly relevant to this putative TSS as there is no corresponding TATA box, and BRE^d motifs are highly overrepresented in the *Drosophila* genome.

After completing the search for core promoter motifs, a wide search region was defined to complement the narrow search region. As there is an abundance of evidence suggesting that the initial untranslated exon of *CG33521*-RC exists in *D. eugracilis*, but there is no BLASTn alignment or other data available to define a definitive TSS position, the narrow (638-863) and wide (638-1075) search regions were submitted as the final annotations of this TSS (Fig. 54). Recall that *CG33521*-RC shares a TSS with *CG33521*-RA and *CG33521*-RG. As a result, the defined TSS search region applies to the TSS of all three isoforms.

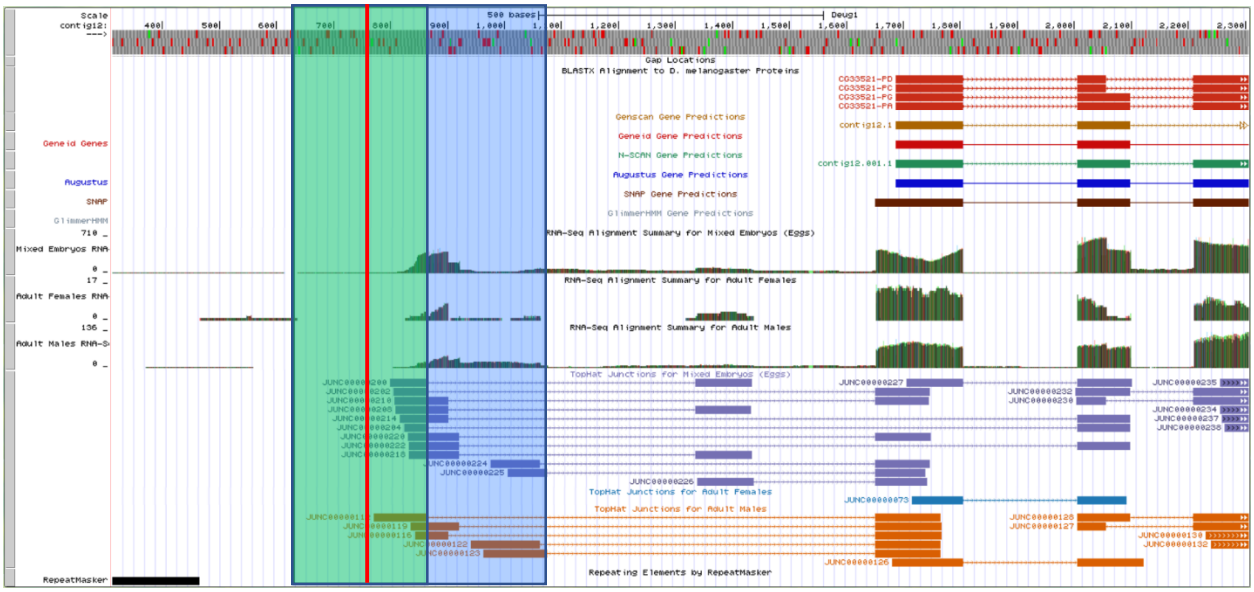


Figure 54: Narrow and broad TSS search regions of *CG33521*-RC. The narrow search region has coordinates 638-863 and extends from the start of RNA-Seq data to the first 5' splice site predicted by the TopHat junctions (green box). The broad search region has coordinates 638-1075 and extends from the start of the RNA-Seq data to the end of the last TopHat junctions corresponding to the exon suggested by RNA-Seq data (blue box). The putative TSS identified with the Short Match function is at position 781 (red line).

Classification of the TSS of *CG33521*-RD in *D. melanogaster*

The fourth isoform of this gene, *CG33521*-RD, uses a TSS that is unique from the other three isoforms. Therefore, it was annotated independently of the other TSS. The first step taken in the annotation of this TSS was to classify the promoter of *CG33521*-RD in *D. melanogaster*.

Examination of the annotated TSS track and the DHS positions tracks in the *D. melanogaster* Genome Browser identified the promoter of this isoform as intermediate (Fig. 55). After classifying the promoter as intermediate (no DHS positions and two Celniker TSS) the annotated TSS, CAGE, RAMPAGE, and RNA-Seq data were used to confirm that the gene is expressed in *D. melanogaster*.

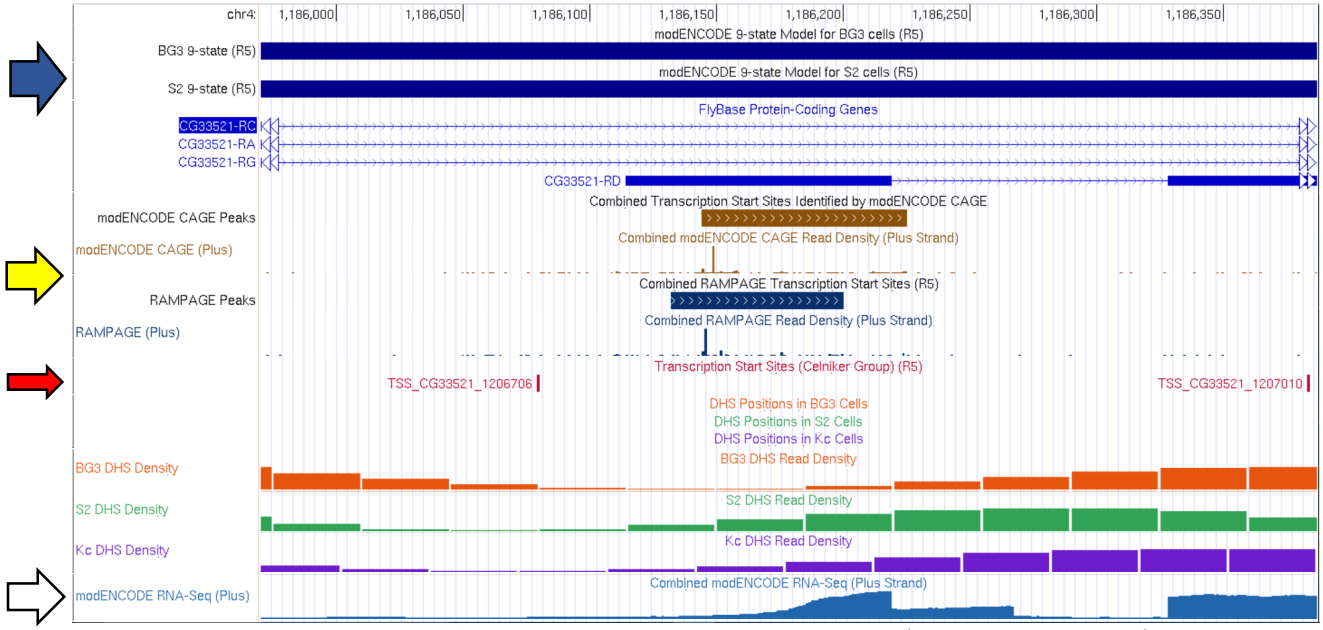


Figure 55: Identification of the *CG33521*-RD promoter as intermediate. There are two annotated TSSs corresponding to *CG33521*-RD, resulting in a classification as intermediate (red arrow). As with the other TSS of *CGG33521*, the BG3 and S2 9-state models are all blue, indicating heterochromatin in these two cell types (blue arrow). However, there is evidence that this promoter is active in both the CAGE and RAMPAGE data (yellow arrow) and the plus-strand RNA-Seq data indicates the transcription of this exon (white arrow). Interestingly, the CAGE and RAMPAGE data have a single TSS peak, suggesting a peaked promoter. However, the GEP protocol is to classify promoters by their DHS sites and Celniker annotated TSSs. Thus, this promoter will remain classified as intermediate.

Annotation of the Putative TSS Position of *CG33521*-RD in *D. eugracilis*

After classifying the promoter of *CG33521*-RD in *D. melanogaster* as intermediate, a BLASTn search was performed using the initial 5' untranslated exon from *CG33521*-RD in *D. melanogaster* as the query and the entirety of contig12 as the subject. The search produced an alignment that occurred in the expected region of contig12 (Figs. 56 and 57).

Deug1_dna range=contig12:1-38500 5'pad=0 3'pad=0 strand=+ repeatMasking=none
Sequence ID: Query_42283 **Length:** 38500 **Number of Matches:** 65

Range 1: 1423 to 1472 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
25.4 bits(16)	0.062	43/62(69%)	12/62(19%)	Plus/Plus

```
Query  1     ATATAAATATCCTAGGTAGAGATCGGTAATAGGTATCATAAATAAATTTGCTATTTTTCC  60
             ||||||| |||| |||||         ||| ||  |||||||||| ||||  ||| | | |
Sbjct  1423  ATATAAACATCCCAGGTA--------TAA-AG--ATCATAAATATATTT-TTATCTATTC  1470

Query  61    CA  62
             ||
Sbjct  1471  CA  1472
```

Figure 56: BLASTn alignment of the initial untranslated exon of *CG33521*-RD in *D. melanogaster* (query) against contig12 (subject). This alignment does not have a very low e-score (red box), but this is a fairly short sequence, contributing to this high score. Furthermore, this sequence aligns to the first base of the *D. melanogaster* exon (blue box). This alignment places the putative TSS at position 1423 in contig12.

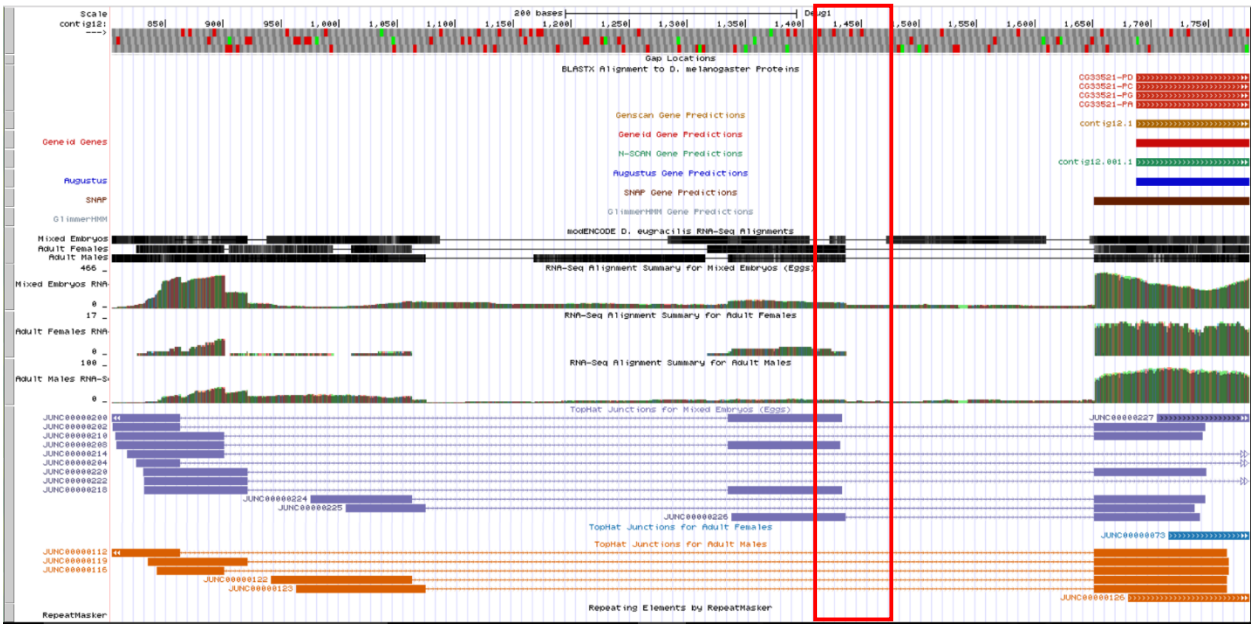


Figure 57: Position of the BLASTn alignment in contig12. The BLASTn alignment maps to approximately the expected position in contig12 (red box). However, the start of the alignment occurs near the end of the TopHat predictions and RNA-Seq data. Furthermore, there is RNA-Seq read coverage upstream that suggests that this untranslated exon might be longer.

While the BLASTn match generally aligned to the expected region in contig12, the exon predicted by the alignment did not correspond with the TopHat junctions or RNA-Seq data. The start of the alignment was found to occur near the end of the increase in RNA-Seq reads and TopHat alignments. The RNA-Seq expression from different developmental stages as well as the RNA-Seq exon junctions were examined in in *D. melanogaster* GBrowse (Fig. 58). This analysis confirmed that there are unannotated splice sites near this exon as well as RNA transcription at certain developmental stages that do not correspond to the annotated exon. Looking at the strand-specific RNA-Seq data corresponding to *CG33521*-RD in the *D. melanogaster* Genome Browser reveals a clear splice site that does not match the annotated exon (Fig. 59).

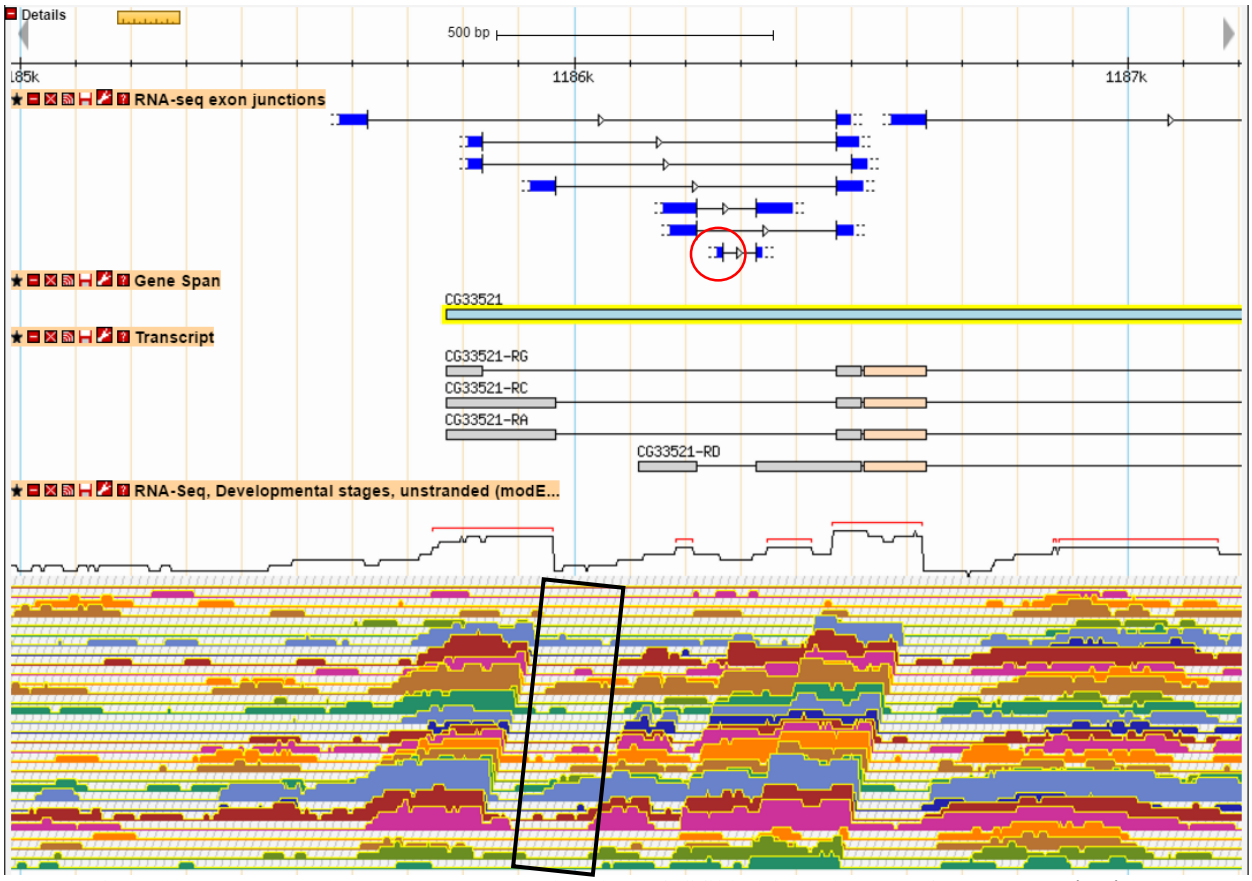


Figure 58: Evidence supporting unannotated splice sites of *CG33521*-RD in *D. melanogaster*. Viewing *CG33521*-RC's initial 5' exons in FlyBase GBrowse reveals that there is at least one RNA-Seq exon junction that does not correspond to a splice site in the *CG33521* gene model (red circle). There also appears to be transcription that does not align to any annotated exons in the gene model (black box).

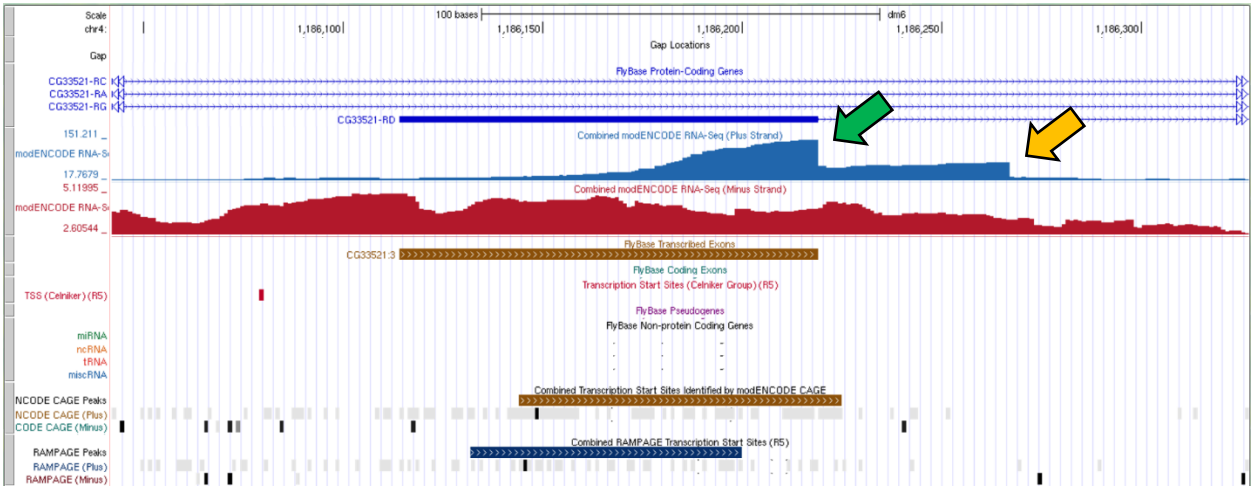


Figure 59: Strand-specific RNA-Seq data indicates an unannotated exon in *CG33521*-RD in *D. melanogaster*. The first drop-off in plus-strand RNA-Seq data corresponds to the annotated 5' splice site of the initial untranslated exon of *CG33521*-RD (green arrow). The second drop off in plus-strand RNA-Seq data appears to correspond to an unannotated splice site in this gene (yellow arrow).

The above data is evidence that there is at least one unannotated splice site near this exon. This data supports the hypothesis that the BLASTn alignment corresponds to the orthologous first exon of *CG33521*-RD in *D. eugracilis* and that the extra RNA-Seq data corresponds to unannotated exons. With this hypothesis, the first base of the BLASTn alignment (1423) was set as the putative TSS. A search region consisting of 300 upstream and 300 downstream bases was created to search for *Drosophila* core promoter motifs from position 1123-1723 (Fig. 60). A similar search region was used to search for motifs around the annotated TSS of *CG33521*-RD in *D. melanogaster* using the short match function in the *D. melanogaster* Genome Browser. The results of both of these motif searches were organized into a table (Table 5). No motifs that support the TSS were found.

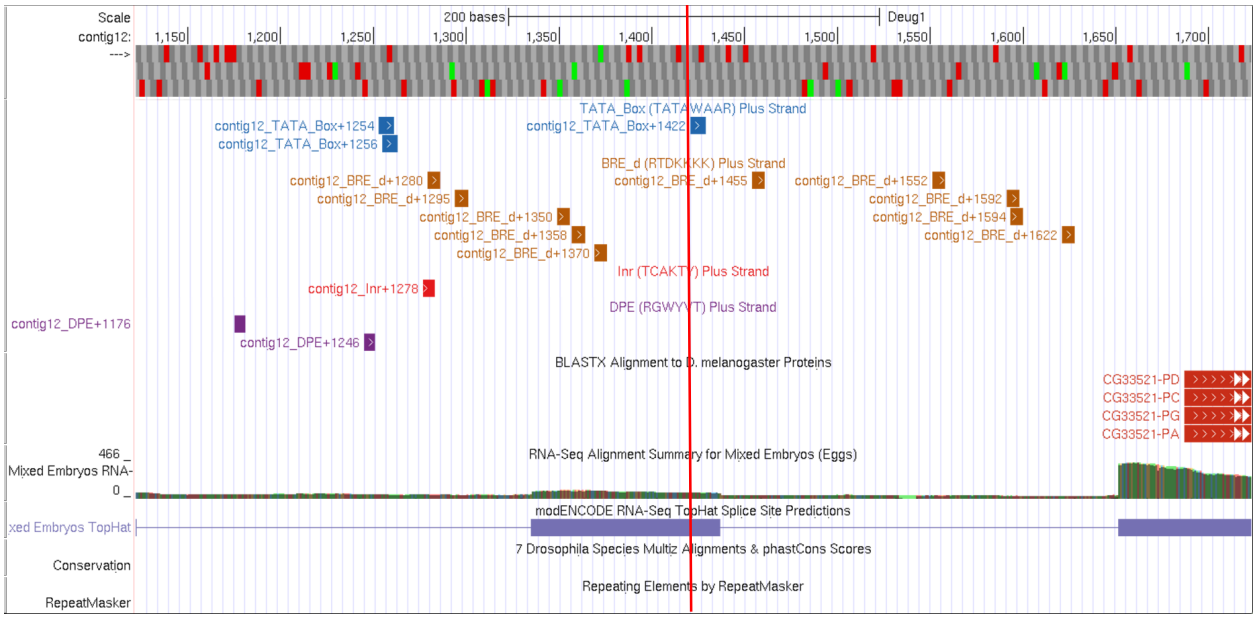


Figure 60: Search for *Drosophila* core promoter motifs in *D. eugracilis*. None of these motifs, identified by the "Core Promoter Motifs (Plus)" track, correspond to the putative TSS at 1423 (red line). Blue boxes indicate TATA boxes, brown boxes indicate BRE^{d} motifs, red boxes indicates Inr motifs, and purple boxes indicate DPE motifs.

Core promoter motif	*D. eugracilis*	*D. melanogaster*
BRE^{u}	NA	NA
TATA Box	1254, 1256, 1422	1186114
BRE^{d}	1280, 1295, 1350, 1358, 1370, 1455, 1552, 1592, 1594, 1622	1185911, 1185976, 1186010, 1186012, 1186375
Inr	1278	1186022
MTE	NA	NA
DPE	1176, 1246	NA
Ohler_motif1	NA	NA
DRE	NA	NA
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	NA	NA
Ohler_motif8	NA	NA

Table 5: Core promoter motifs found in the TSS search region of *CG33521*-RD in *D. eugracilis* and *D. melanogaster*. None of the these motifs correspond to the putative TSS in *D. eugracilis* or the annotated TSS in *D. melanogaster*.

With the search for *Drosophila* core promoter motifs complete, 1423 was established as the putative TSS for *CG33521*-RD in *D. eugracilis* (Fig. 61). As *D. biarmipes* is a closely related species to *D. eugracilis,* the *D. biarmipes* Genome Browser was examined to determine if this additional transcribed region occurs in *D. biarmipes* as well. However, there was no conservation found between the species that would allow for the positioning of the putative TSS position using the Short Match function.

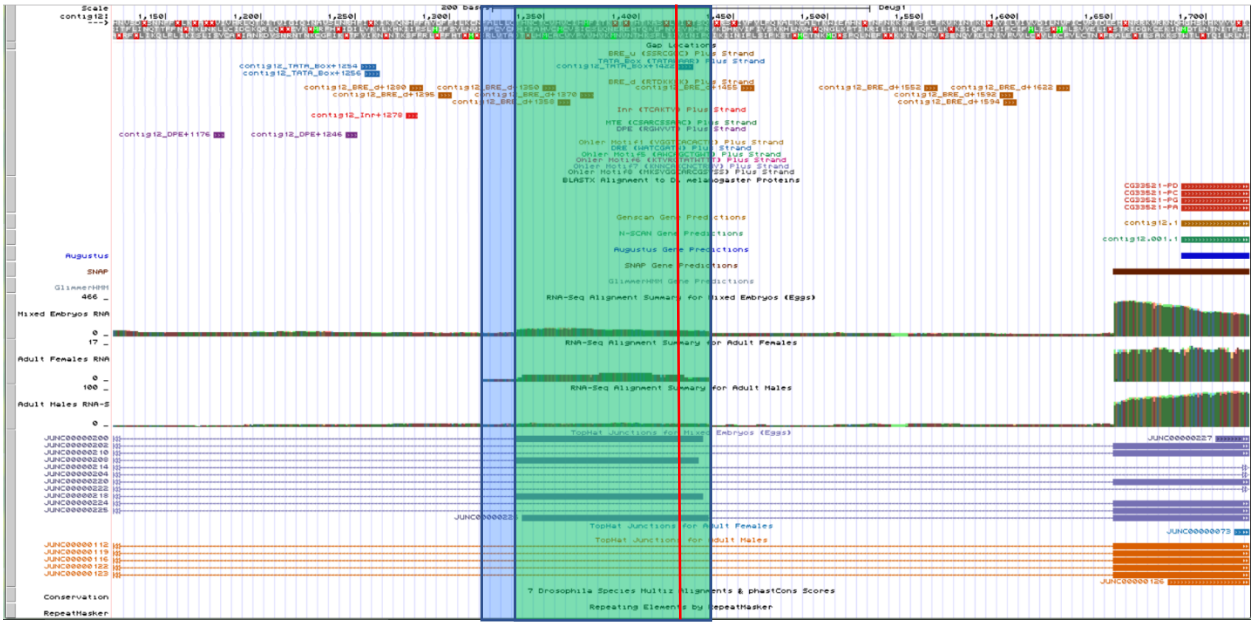


Figure 61: The putative TSS of *CG33521*-RD in *D. eugracilis* occurs at position 1423. This position is indicated by the red line on the Genome Browser. In addition, search regions were defined using RNA-Seq data and TopHat junctions. The narrow search region (1336-1437) contains the TopHat junctions (green box). The broad search region (1318-1437) encompasses the increase in RNA-Seq expression data in embryos and males (blue box).

PIP4K

Identification of the Ortholog

The next feature in contig12 is identified in Figure 62. Examination of the computer-based gene predictions corresponding to Feature 2 reveals several predictions that correlate reasonably well with this feature. The N-Scan prediction, contig12.002.1, was selected for BLASTp analysis (Fig. 63).

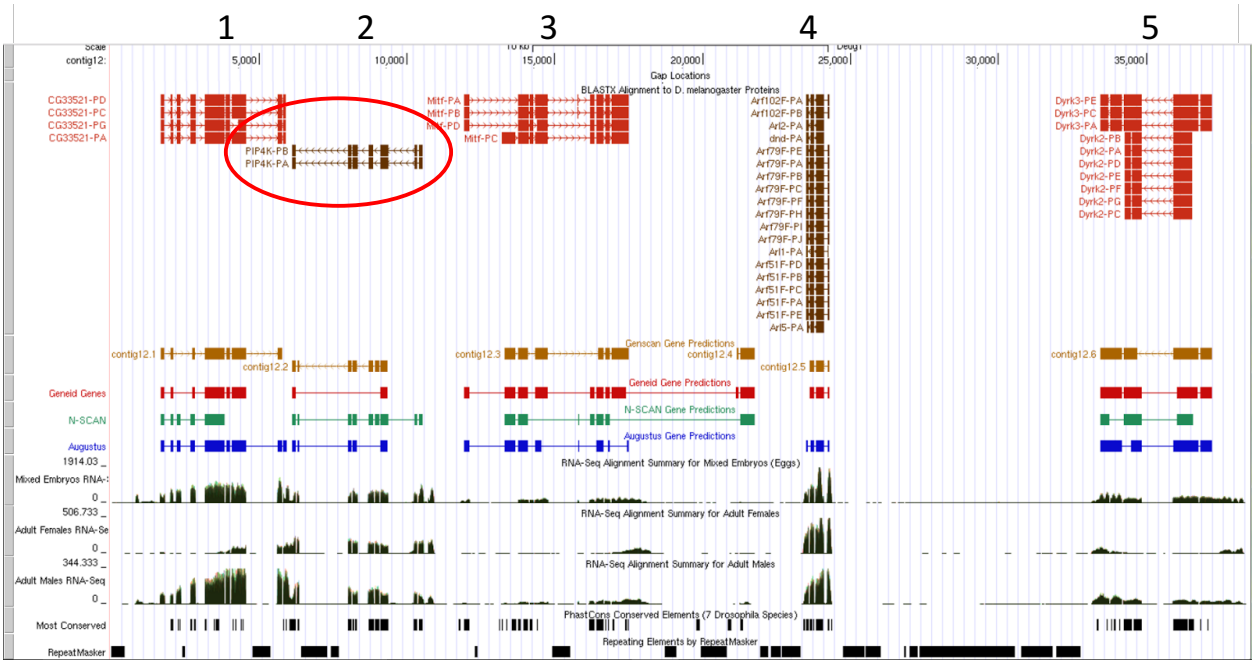


Figure 62: The position of the next feature in contig12. Shown in the red circle is feature 2, which is the next feature downstream from *CG33521*.

BLAST Hit Summary				
☑	Description	Species	Score	E value
☑	PIP4K-PB	Dmel	791.186	0
☑	PIP4K-PA	Dmel	791.186	0
☑	PIP5K59B-PI	Dmel	151.754	9.77271e-37
☑	sktl-PC	Dmel	149.443	5.73096e-36
☑	sktl-PB	Dmel	149.443	5.73096e-36
☑	sktl-PA	Dmel	149.443	5.73096e-36
☑	PIP5K59B-PA	Dmel	148.288	1.27673e-35
☑	PIP5K59B-PF	Dmel	147.902	1.39945e-35
☑	PIP5K59B-PE	Dmel	147.902	1.45907e-35
☑	PIP5K59B-PH	Dmel	147.517	1.78256e-35

Figure 63: Summary of the BLASTp search. The FlyBase BLASTp program provided this summary of BLASTp alignments obtained when searching the N-Scan predicted protein (query) against the *Drosophila* Annotated Proteins Database (subject). The alignment to *PIP4K* produced the lowest e-score of 0. The following alignments all produced similar e-scores that were all significantly lower than the e-score of *PIP4K*.

The alignment of the N-Scan prediction to *PIP4K*-PB produced by the BLASTp search is shown in Figure 64. An identical score was produced from the alignment of the prediction to *PIP4K*-PA (data not shown). *PIP4K* is found on the 4th chromosome of *D. melanogaster*, which is further evidence that it is the orthologous gene to the second feature of *D. eugracilis* contig12.

Figure 64: BLASTp alignment of the N-Scan predicted protein against *D. melanogaster PIP4K*-PB. An alignment identical to this was produced by matching this query against the *PIP4K*-PA amino acid sequence. *PIP4K* is located on the 4th chromosome of *D. melanogaster* (red box). The entire *PIP4K*-PB isoform aligns to the second feature. This is confirmed by observing the start of the alignment at position 1 (yellow box), and the position at the end of the alignment, which matches the length of the *PIP4K*-PB subject sequence (blue boxes).

From the BLASTp evidence, *PIP4K* was determined to be the *D. melanogaster* ortholog of the second feature. FlyBase reports that the full name of this gene is *Phosphatidylinositol 5-phosphate 4-kinase*, and that the enzyme it produces mediates the conversion of phosphatidylinositol 4 phosphate to phosphatidylinositol 4,5 bisphosphate. This enzyme is also implicated in the regulation of mTOR (mammalian target of rapamycin) signaling and control of

cell size. Examination of *PIP4K* in the Gene Record Finder confirmed that *PIP4K*-PB and *PIP4K*-PA have identical CDSs. Figure 65 shows both isoforms of *PIP4K* as they occur in *D. melanogaster*. Because *PIP4K*-PA and *PIP4K*-PB have identical CDSs, the remainder of the report which covers the CDS annotation of this gene will refer to the annotation of both isoforms as *PIP4K*.

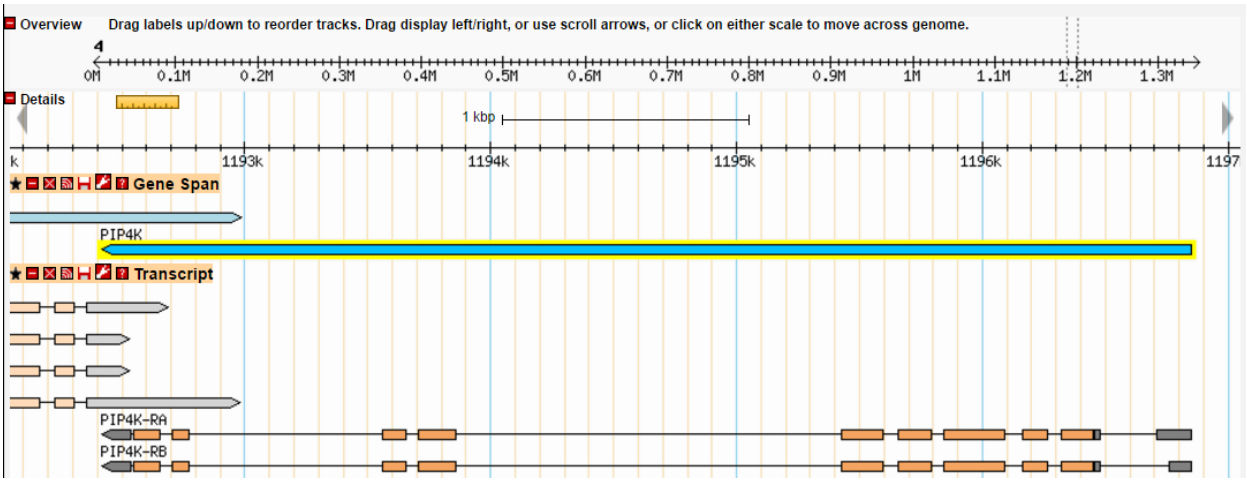
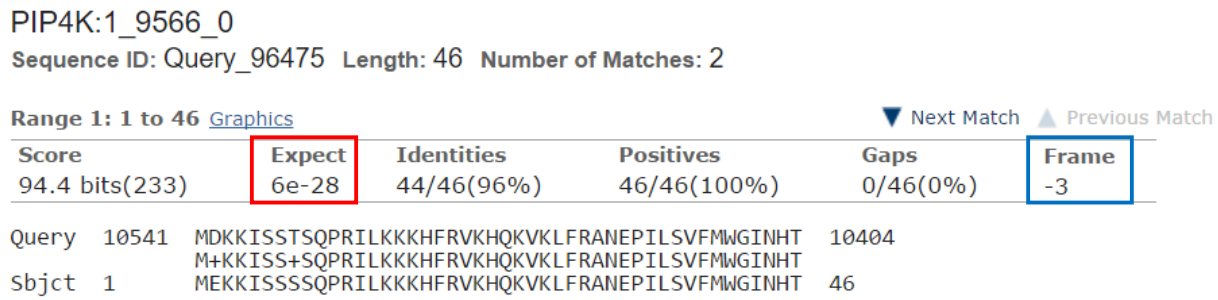


Figure 65: FlyBase GBrowse view of *PIP4K*. In *D. melanogaster*, *PIP4K*-PA and *PIP4K*-PB have identical CDSs. The two isoforms only differ in their initial untranslated exons.

Exon-by-Exon Annotation of *PIP4K*

Standard CDS exon annotation protocol was followed to annotate these coding exons. The *PIP4K* exon in *D. melanogaster* was taken from the Gene Record Finder and used as the subject in a BLASTx search against a query of contig12. The resulting alignment was used to determine the general location of the exon. Using that location as a guideline, knowledge of biological rules regarding splicing and evidence tracks in the *D. eugracilis* Genome Browser were used to determine the exact boundaries of these exons in *D. eugracilis*. The positions of these annotated exons were compiled in Table 6. No unexpected results (*e.g.* GC splice donor

sites or novel exons) were encountered in the annotation of the CDS of this gene. As this project places particular emphasis on the annotation of the initial methionine of the coding spans of these genes, the BLASTx output used to place the initial coding exon is included in Figure 66. The methionine suggested by this alignment is the same methionine that was confirmed to correspond to the start of the first coding exon by viewing *PIP4K* in the *D. eugracilis* Genome Browser (Fig. 67).


PIP4K:1_9566_0

Sequence ID: Query_96475 **Length:** 46 **Number of Matches:** 2

Range 1: 1 to 46 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
94.4 bits(233)	6e-28	44/46(96%)	46/46(100%)	0/46(0%)	-3

```
Query  10541  MDKKISSTSQPRILKKKHFRVKHQKVKLFRANEPILSVFMWGINHT  10404
              M+KKISS+SQPRILKKKHFRVKHQKVKLFRANEPILSVFMWGINHT
Sbjct  1      MEKKISSSSQPRILKKKHFRVKHQKVKLFRANEPILSVFMWGINHT  46
```

Figure 66: BLASTx search of contig12 (query) against the 1st exon of *PIP4K* in *D. melanogaster* (subject). The methionine predicted to serve as the start codon of this gene can be seen at the start of this alignment. This output has an e-score of 6e-28 (red box) and the aligned sequence from contig12 is in frame -3 (blue box).

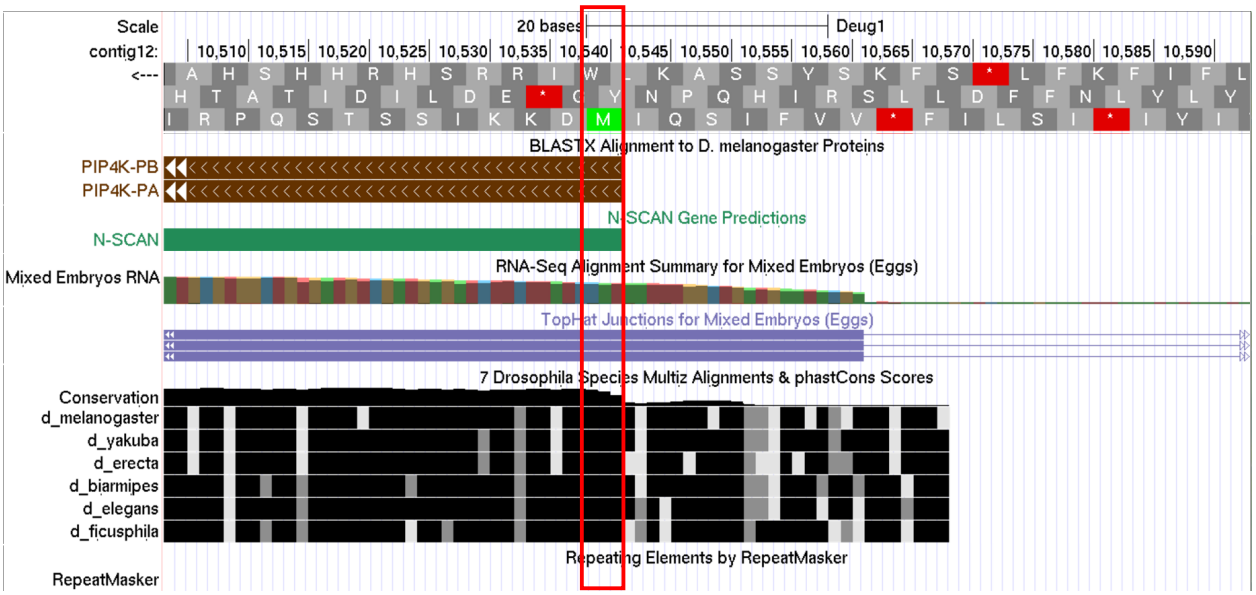


Figure 67: Annotation of the start codon of *PIP4K* (red box). This is the only methionine in the region. It is in the correct frame (-3) and aligns with the N-Scan prediction. This methionine occurs shortly after the start of RNA-Seq data, which is what one would expect given that *PIP4K* has a short untranslated component to it's first coding exon.

Name of exon in *D. melanogaster*	Start/splice donor	Stop/splice acceptor	Frame	Acceptor Phase	Donor Phase	Exon BLASTp e-score
PIP4K:1_9566_0	10541	10404	-3	NA	0	6e-28
PIP4K:2_9566_0	10344	10247	-2	0	2	2e-19
PIP4K:3_9566_1	9362	9113	-1	1	0	1e-53
PIP4K:4_9566_0	9059	8929	-3	0	2	9e-27
PIP4K:5_9566_1	8869	8701	-2	1	0	5e-31
PIP4K:6_9566_0	8322	8171	-2	0	2	1e-28
PIP4K:7_9566_1	8113	8019	-2	1	1	4e-20
PIP4K:8_9566_2	6357	6293	-1	2	0	8e-11
PIP4K:9_9566_0	6233	6117	-3	0	NA	4e-22

Table 6: Final coding exon annotations of *PIP4K*. This table contains the most parsimonious annotations of the coding exons of *PIP4K* in *D. eugracilis*.

Checking the *PIP4K* Gene Model

The gene model of *PIP4K* proposed in Table 6 was tested using the GMC. All exons follow the basic biological rules of CDS annotation and thus, passed the GMC. The resulting dot plot and amino acids alignment are shown in Figure 68 and Figure 69.

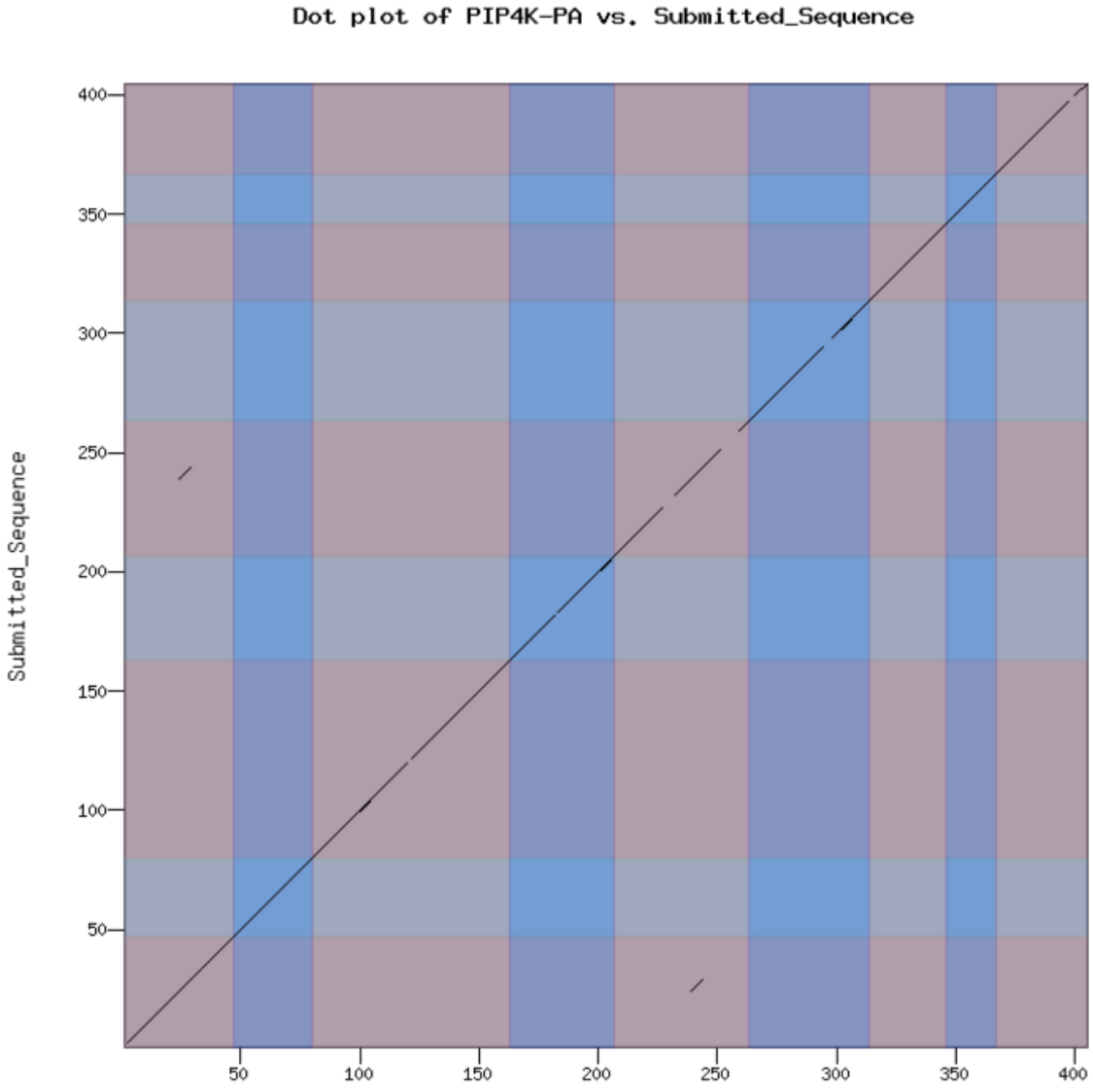


Figure 68: Dot Plot of the CDS annotation of *PIP4K* from *D. melanogaster* against the gene model for Feature 2 of *D. eugracilis*. While this alignment states that it uses the coding exons of *D. melanogaster PIP4K*-PA, the Dot Plot produced by aligning the model of Feature 2 against *PIP4K*-PB is identical to the one shown. The continuity of this line indicates that this gene is highly conserved.

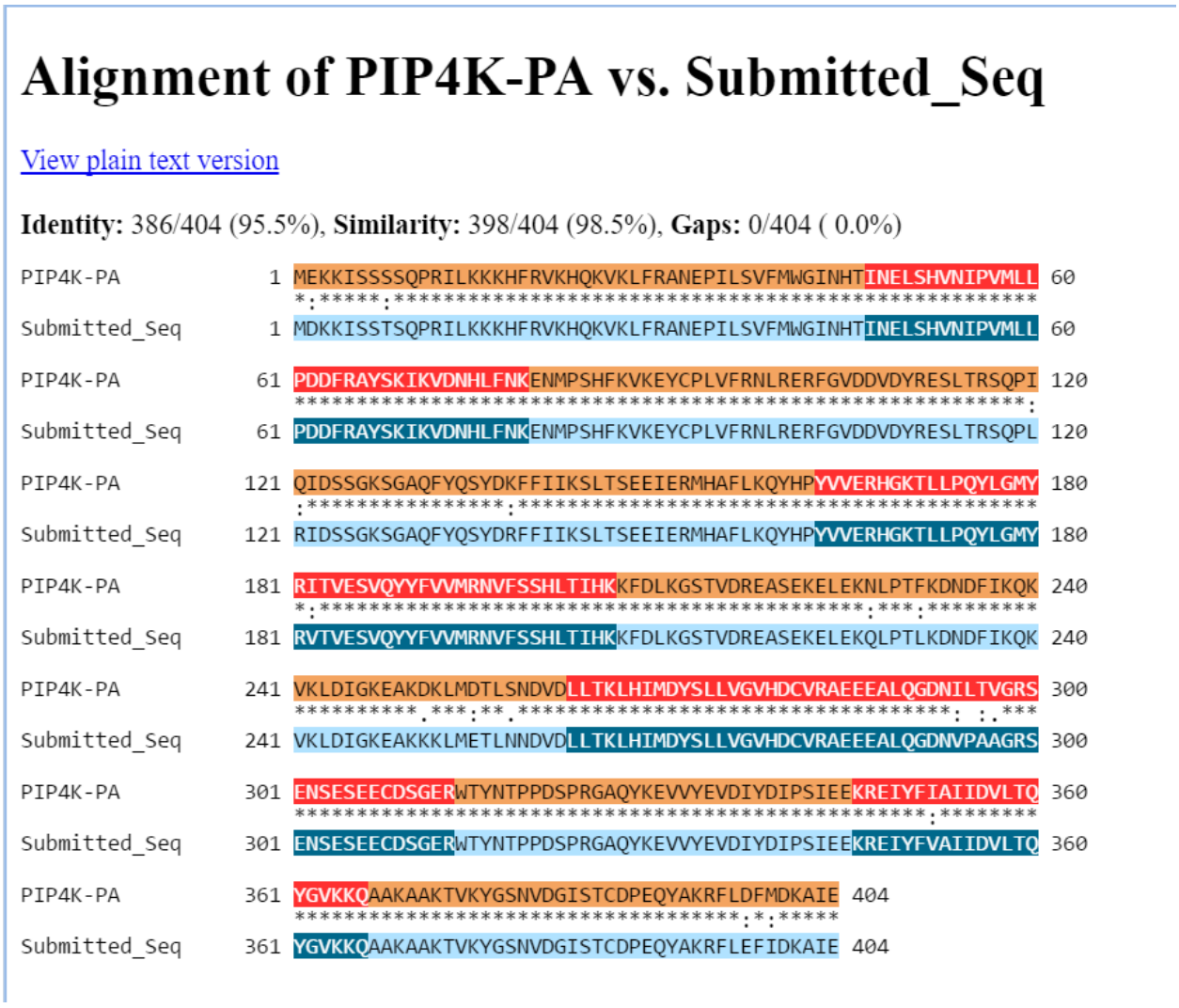


Figure 69: Protein Alignment of *PIP4K* annotation. Just as in the Dot Plot, this protein alignment of *PIP4K*-PA against the model of Feature 2 is identical to the alignment of *PIP4K*-PB against the model of Feature 2.

Identification of the Transcription Start Site of *PIP4K* in *D. melanogaster*

The TSS of *PIP4K* was annotated after confirming the CDS annotation to be the most parsimonious model to the *D. melanogaster* ortholog. The untranslated exons of *PIP4K* have already been annotated in *D. melanogaster.* These untranslated exons were examined in the Gene

Record Finder and it was determined that both isoforms of *PIP4K* share a single TSS in *D. melanogaster* (Fig. 70).

A

Exon usage map:

Isoform	2	1	3	4	5	6	7	8	9	10	11
PIP4K-RB		1	2	3	4	5	6	7	8	9	10
PIP4K-RA	1		2	3	4	5	6	7	8	9	10

Select a row to display the corresponding exon sequence:

FlyBase ID	5' Start	3' End	Strand	Size (bp)
1	1,196,848	1,196,758	-	91

B

Exon usage map:

Isoform	2	1	3	4	5	6	7	8	9	10	11
PIP4K-RB		1	2	3	4	5	6	7	8	9	10
PIP4K-RA	1		2	3	4	5	6	7	8	9	10

Select a row to display the corresponding exon sequence:

FlyBase ID	5' Start	3' End	Strand	Size (bp)
2	1,196,848	1,196,708	-	141

Figure 70: *PIP4K*-PB (A) and *PIP4K* (B) share the same TSS. The 5' start position of the first transcribed exons of both isoforms confirms that they possess an identical TSS in *D. melanogaster* (red circles).

Classification of the TSS of *PIP4K*-RA in *D. melanogaster*

Both isoforms of *PIP4K* have identical TSSs. However, the initial untranslated exon of *PIP4K*-RA is longer than that of *PIP4K*-RB, and thus it is a better candidate for a BLASTn search. For this reason, the initial untranslated exon of *PIP4K*-RA was used to annotate the TSS.

The TSS found to correspond to this exon was then annotated as the TSS of *PIP4K*-RB as well, since the two isoforms share a TSS in *D. melanogaster*.

The evidence tracks (Celniker TSSs and DHS positions) corresponding to *PIP4K-RA* in the *D. melanogaster* Genome Browser were used to classify the promoter as peaked (Fig. 71). There is one Celniker TSS and one DHS position that correspond to the start of this exon.

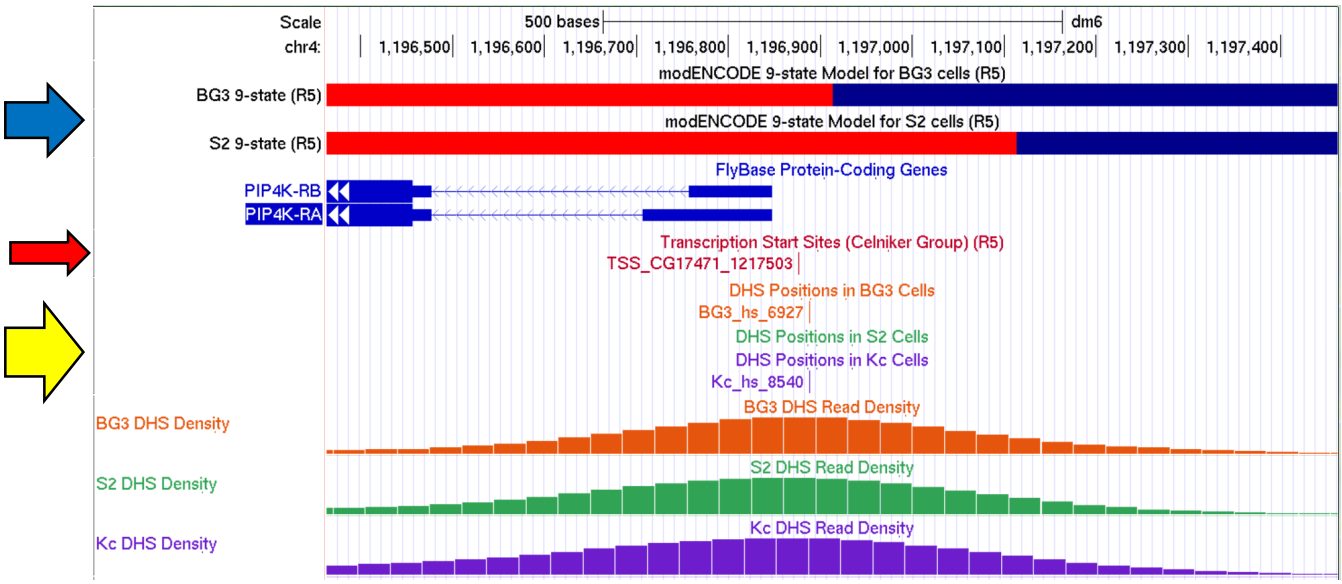


Figure 71: Identification of the *PIP4K*-RA promoter as peaked. There is a Celniker annotated TSS that corresponds to *CG17471* (red arrow). Looking up *CG17471* in FlyBase reveals that it is the annotation symbol for *PIP4K*. BG3 cells and Kc cells share a DHS position (yellow arrow). Both 9-state models are red in the region corresponding to the TSS, indicating an active promoter or TSS (blue arrow).

Definition of a Putative TSS for *PIP4K*-RA in *D. eugracilis*

After classifying the promoter of *PIP4K*-RA as peaked, a BLASTn search was performed using the initial untranslated exon of *PIP4K*-RA as the query and the sequence of contig12 as the subject. Recall that *PIP4K*-RA was selected because it contains the longest initial untranslated exon, and thus is the most likely candidate to produce a significant BLASTn alignment. The resulting alignment is shown in Figure 72. This alignment occurs in the expected region of contig12 (Fig. 73).

contig12

Sequence ID: Query_156493 **Length:** 38500 **Number of Matches:** 10

Range 1: 10894 to 10931 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
45.4 bits(30)	9e-08	34/38(89%)	0/38(0%)	Plus/Minus

```
Query  15     TAAATTTATGTAAACTACAAATTTGTCTGTTCAAAAGC  52
              ||||||||||||||||||| ||||  |||  ||||||||
Sbjct  10931  TAAATTTATGTAAACTACACATTTTTCTCATCAAAAGC  10894
```

Figure 72: BLASTn alignment of the initial untranslated exon of *PIP4K*-RA in *D. melanogaster* (query) against contig12 (subject). This alignment has a relatively low e-score when the length of the sequence is taken into consideration (red box). The sequence of contig12 aligns to the 15th base of *PIP4K*-RA (blue box). The position of the putative TSS was extrapolated from this alignment by adding 14 bases contig12 sequence. Using this method, the putative TSS was determined to be position 10945 in *D. eugracilis*.

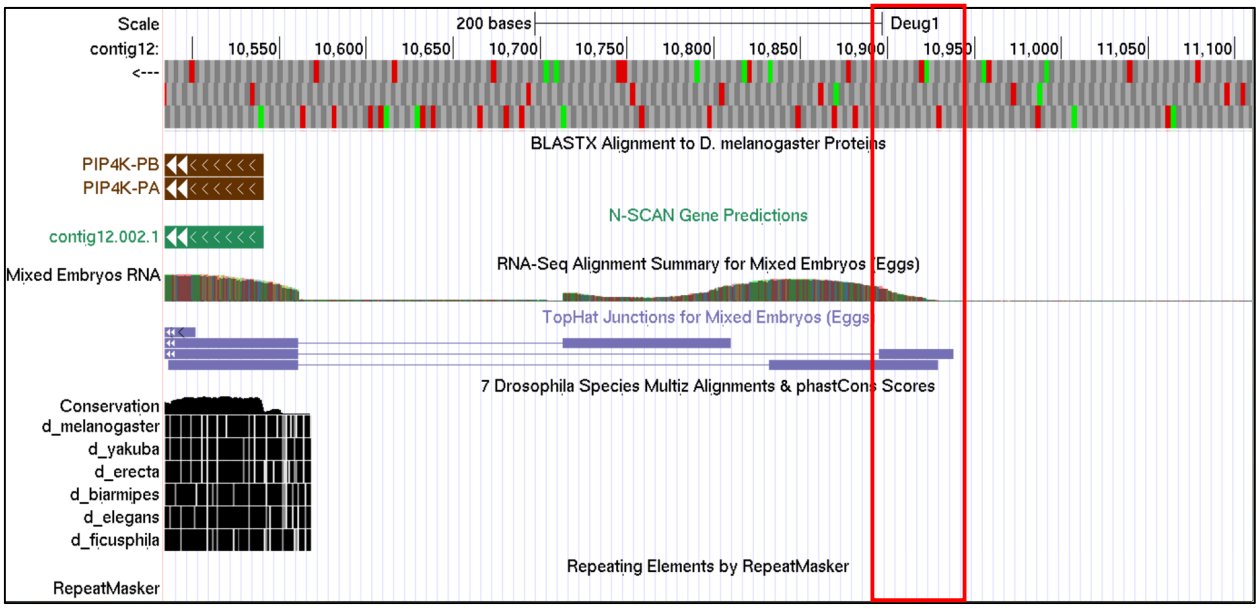


Figure 73: Position of the BLASTn alignment in contig12. The BLASTn alignment maps to approximately the expected position in contig12 (red box). RNA-Seq data and Top Hat junctions suggest that the initial untranslated exon begins around this region.

There is no RNA polII data available for *D. eugracilis*. However, there is an RNA polII data track available in *D. biarmipes*, a neighboring species. The RNA polII peaks associated with the expected TSS of *PIP4K* were examined in the BCM-HGSC assembly in the *D. biarmipes* Genome Browser (Fig.

74). The strongest peak was selected as the query for a BLASTn search against a subject of contig12. The resulting alignment is shown in Figure 75. The putative TSS at 10945 falls within the alignment, providing strong evidence that this TSS is in the correct region.

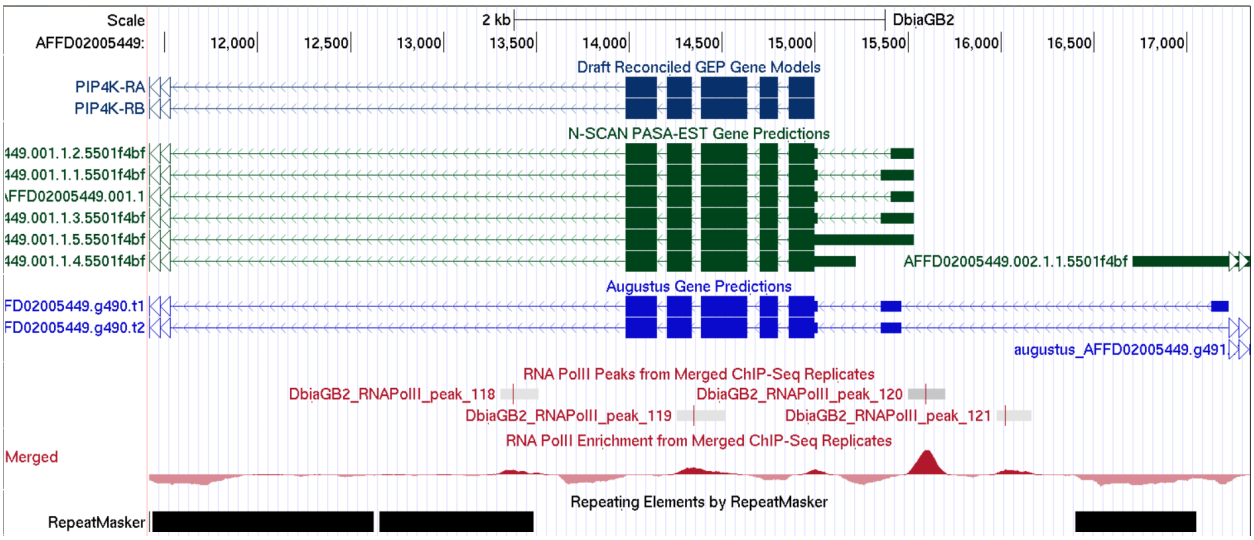


Figure 74: Analysis of RNA polII data associated with the putative TSS of *PIP4K* in *D. biarmipes*. Using the N-Scan predictions and Augustus predictions (which reflect the RNA-Seq data), DbiaGB2_RNAPolII_peak_120 was selected as the RNA polII peak most closely associated with the putative TSS. The DNA sequence of this peak was extracted for use in a BLASTn search.

Deug1_dna range=contig12:1-38500 5'pad=0 3'pad=0 strand=+ repeatMasking=none

Sequence ID: Query_61831 **Length:** 38500 **Number of Matches:** 28

Range 1: 10930 to 10982 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
32.6 bits(21)	0.001	40/57(70%)	4/57(7%)	Plus/Plus

```
Query  1      TATAAAAAGAAAATGTGTATCGATTATTTCAGATATGTATTTGGTGTTTAACATATG  57
              || | | |||||| |||||||||   || |||    |||||||| ||  | |||||
Sbjct  10930  TACAGACAGAAAAAATGTATCGATACATTTAGA----TATTTGGTATTATATATATG  10982
```

Figure 75: BLASTn alignment of the *D. biarmipes* RNA polII peak (query) against contig12 (subject). This alignment has a fairly low e-score, which is not unexpected given the short length of the sequence. The putative TSS at 10945 falls within the region of contig12 that aligns to this RNA polII peak.

After using the BLASTn alignment to establish a putative TSS at 10945, a search region consisting of the 300 upstream and 300 downstream bases was defined to search for *Drosophila* core promoter motifs from position 10645-11245 (Fig. 76). A similar search region (1196548-1197148) was used to search for core promoter motifs around the annotated TSS of *PIP4K*-RA in *D. melanogaster* using the short match function in the *D. melanogaster* Genome Browser. The results of both of these core promoter motif searches were organized into a table (Table 7). No motifs that support the TSS were found. With the search for *Drosophila* core promoter motifs complete, 10945 was established as the putative TSS for *PIP4K* in *D. eugracilis*.

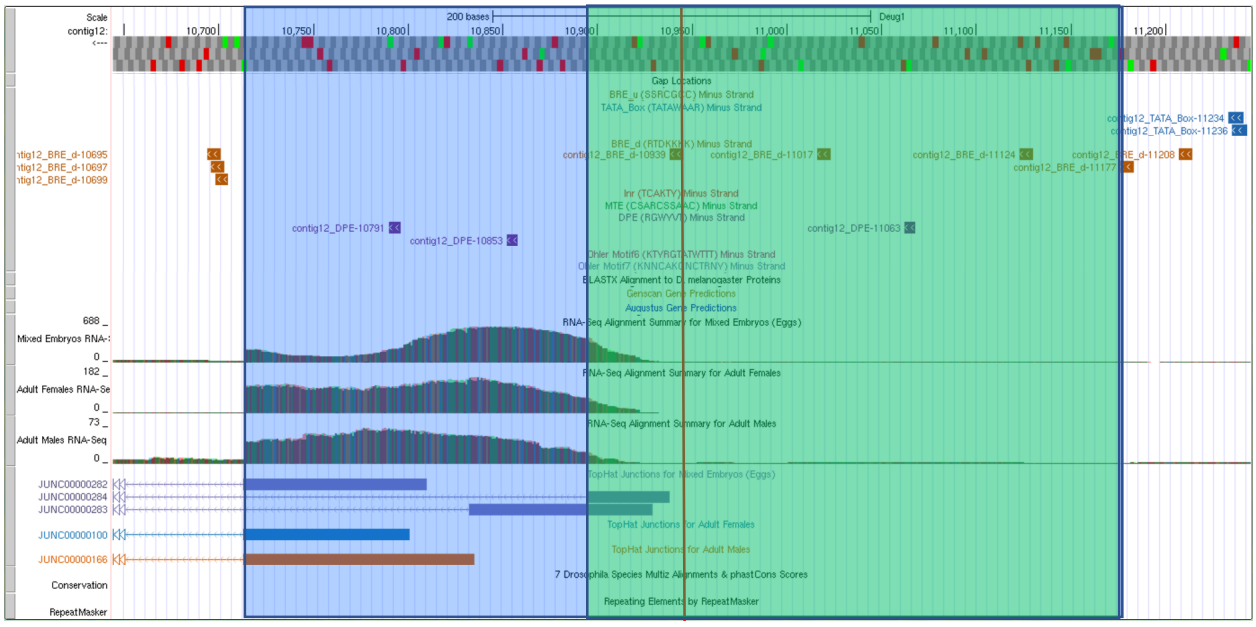


Figure 76: Search for *Drosophila* core promoter motifs in *D. eugracilis*. None of these motifs, identified by the "Core Promoter Motifs (Minus)" track, correspond to the putative TSS at 10945 (red line). Blue boxes indicate TATA boxes, brown boxes indicate BRE^{d} motifs, and purple boxes indicate DPE motifs. In addition RNA-Seq data and TopHat junctions were used to define TSS search regions. The narrow search region (10896-11177) extends from the first TopHat splice donor site to the closest break in RNA-Seq data (green box). The broad search region (10714-11177) extends from the splice donor site suggested by the RNA-Seq data and three TopHat junctions to the closest break in RNA-Seq data (blue box.)

Core promoter motif	*D. eugracilis*	*D. melanogaster*
BRE^{u}	NA	NA
TATA Box	-11234, -11236	-1196604, -1196880
BRE^{d}	-10695, -10679, -10699, -10939, -11017, -11124, -11177, -11208	-1196551, -1196620, -1196630, -1196932, -1196935, -1197041
Inr	NA	NA
MTE	NA	NA
DPE	-10791, -10853, -11063	-1196609, -1196692, -1196726, -1196786, -1196689
Ohler_motif1	NA	NA
DRE	NA	NA
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	NA	NA
Ohler_motif8	NA	NA

Table 7: Core promoter motifs found in the TSS search region of *PIP4K*-RA in *D. eugracilis* and *D. melanogaster*. None of these motifs correspond to the *D. eugracilis* putative TSS or *D. melanogaster* annotated TSS.

Mitf

Identification of the Ortholog

The next feature in contig12 is identified in Figure 77. Examination of the computer-based gene predictions corresponding to Feature 3 reveals several predictions that correlate reasonably well with this feature. The Genscan prediction, contig12.3, was selected for BLASTp analysis (Fig. 78).

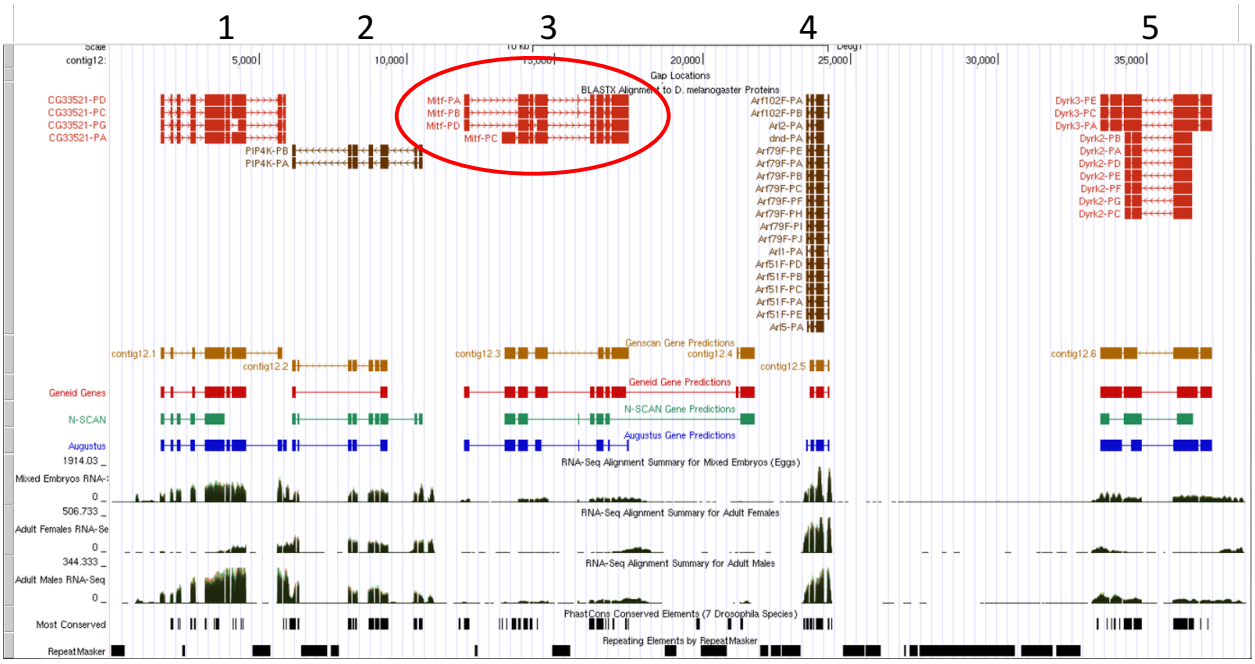


Figure 77: The position of the next feature in contig12. Shown in the red circle is feature 3, which is the next feature downstream from *PIP4K*.

BLAST Hit Summary				
☑	Description	Species	Score	E value
☑	Mitf-PB	Dmel	462.996	4.50137e-130
☑	Mitf-PA	Dmel	462.996	4.50137e-130
☑	Mitf-PC	Dmel	461.455	1.14602e-129
☑	Mitf-PD	Dmel	461.455	1.29882e-129
☑	sdt-PL	Dmel	30.4166	7.1535

Figure 78: Summary of the BLASTp search. The FlyBase BLASTp program provided this summary of BLASTp alignments obtained when searching the Genscan predicted protein (query) against the *Drosophila* Annotated Proteins Database (subject). The alignments to *Mitf* isoforms produced e-scores significantly lower than the next aligned gene, *sdt*.

Along with the summary shown in Figure 78, the BLASTp search provided the following alignment of *Mitf*-PB (Fig. 79). The alignment to *Mitf*-PA was identical to Figure 79. Similar alignments were produced for the other two isoforms of *Mitf*-PC and *Mitf*-PD. Note that *Mitf* is found on the 4th chromosome of *D. melanogaster*, providing further evidence that it is indeed the orthologous gene.

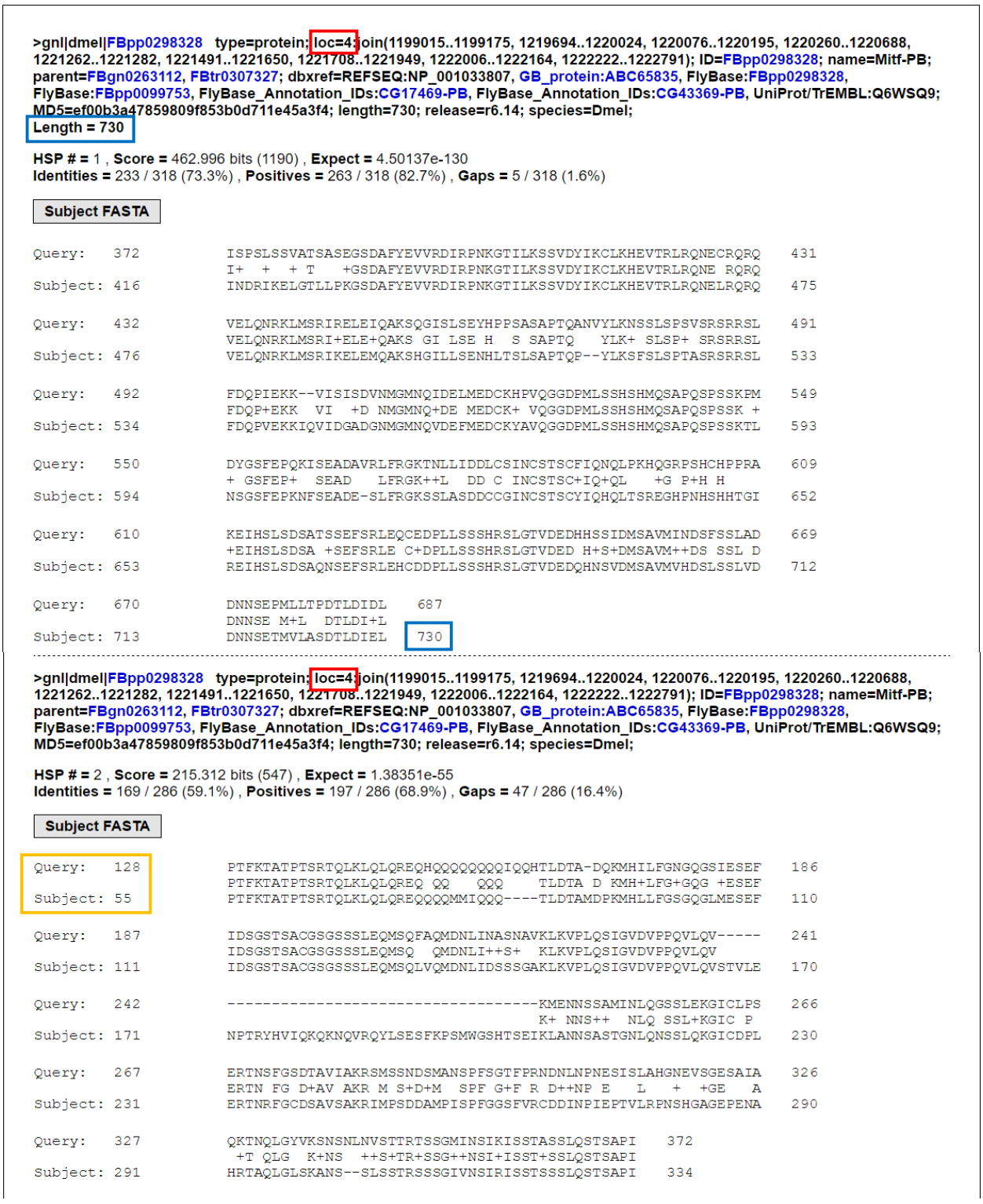


Figure 79: BLASTp alignment of the predicted protein from Genscan against *D. melanogaster Mitf*-PB. Because *Mitf-PB* and *Mitf*-PA contain the same CDS, an identical alignment was produced by matching this query to the *Mitf*-PA (subject) amino acid sequence; appropriate similar results were obtained for the other isoforms. Note that *Mitf* is located on the 4th chromosome of *D. melanogaster* (red boxes). The first 55 bases of *Mitf*-PB did not align to the Genscan predicted protein (yellow box) as well as the sequence between the two fragments. The last base of *Mitf*-PB aligns to the predicted protein (blue boxes).

From the BLASTp evidence, *Mitf* was determined to be the *D. melanogaster* ortholog to this feature. FlyBase reports that the full name of this gene is also *Mitf*, and it encodes a protein that plays a role through sequence-specific DNA binding. Examination of *Mitf* in the Gene Record Finder confirmed that *Mitf*-PB and *Mitf*-PA have identical CDSs. Furthermore, *Mitf*-PC and *Mitf*-PD are two distinct isoforms. Figure 80 shows all four isoforms of *Mitf* as they occur in *D. melanogaster*. The *D. melanogaster* model of *Mitf* contains a very large intron. Examination of this intron in the *D. melanogaster* Genome Browser reveals that this intron spans a gap.

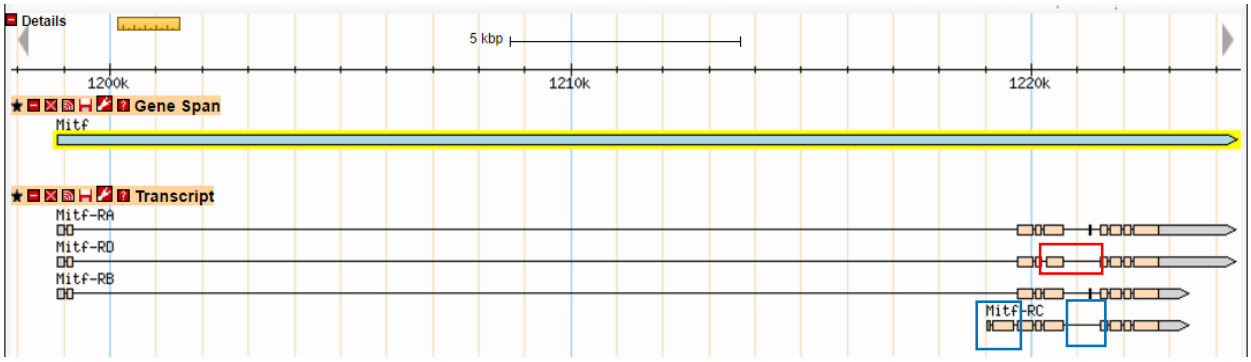


Figure 80: Analysis of *Mitf* in FlyBase GBrowse. In *D. melanogaster*, *Mitf*-PB and *Mitf*-PA have identical CDSs. *Mitf*-PD differs from these two isoforms in that is has a shorter fourth exon and does not have a small fifth exon (red box). *Mitf*-PC also does not contain the small fifth exon and also has a different initial coding exon (blue boxes). Analysis of the region reveals a very large (approximately 20kb) intron, which is reported to span a gap.

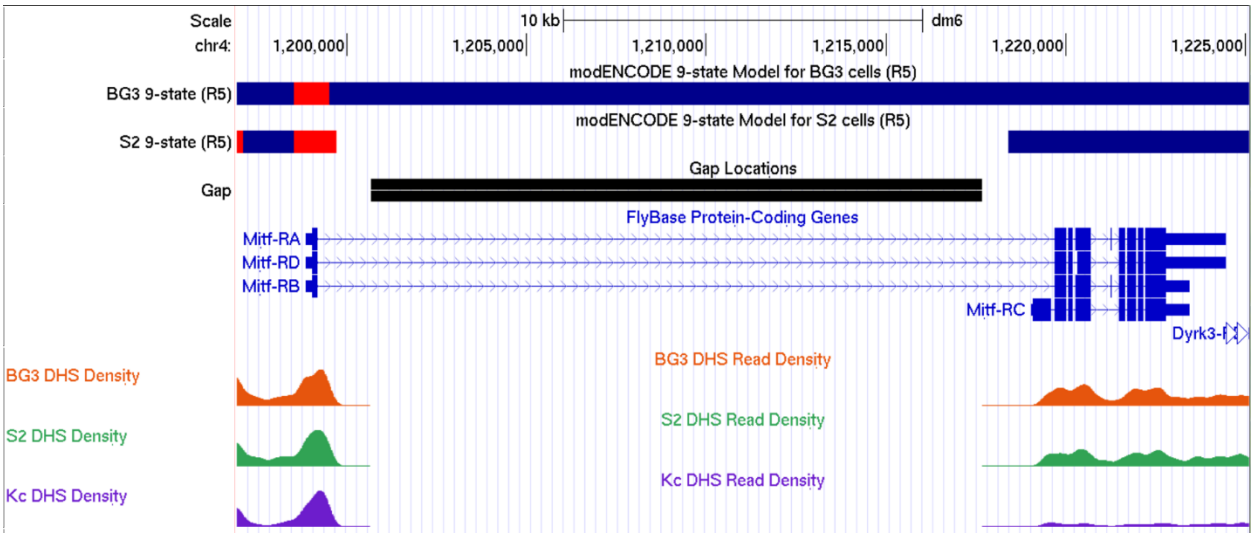
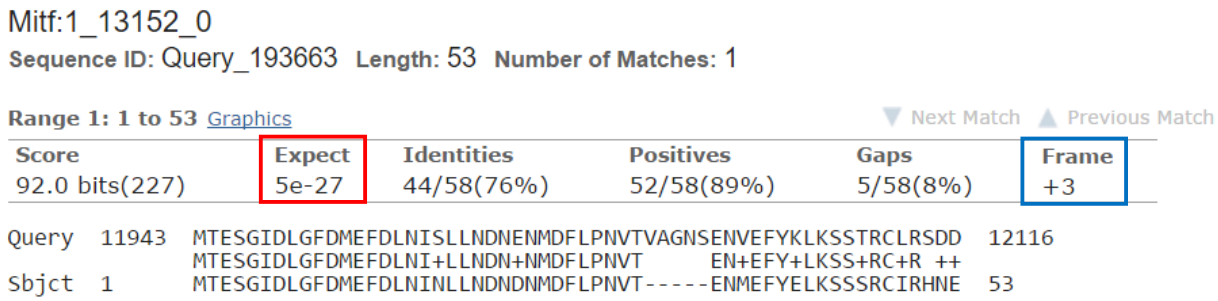


Figure 81: Examination of the large intron in the *D. melanogaster* Genome Browser. This intron is listed as approximately 20 kb long, and it encompasses a gap in the sequence identified by the "Gap" track. Since this is a gap, this 20 kb is probably an arbitrary length, as gaps do not have known lengths.

Exon-by-Exon Annotation of *Mitf*-PB

With the large intron accounted for, an exon-by-exon annotation was conducted on the CDS of the third feature. Standard CDS exon annotation protocol was followed to annotate these coding exons. The *Mitf* exon in *D. melanogaster* was taken from the Gene Record Finder and used as the subject in a BLASTx search against a query of contig12. The resulting alignment was used to determine the general location of the exon. Using that location as a guideline, knowledge of biological rules regarding splicing and evidence tracks in the *D. eugracilis* Genome Browser were used to determine the exact boundaries of these exons in *D. eugracilis*. The positions of these annotated exons were compiled in Table 8. Since *Mitf*-PB (and thus, *Mitf*-PA, since they have identical CDSs) produced the best BLASTp alignment, it was annotated first. As this project places particular emphasis on the annotation of the initial methionine of the coding spans of these genes, the BLASTx output used to place the initial coding exon is included in Figure 82. The methionine suggested by this alignment is the same methionine that was confirmed to correspond to the start of the first coding exon by viewing *Mitf*-PB in the *D. eugracilis* Genome Browser (Fig. 83).


Mitf:1_13152_0

Sequence ID: Query_193663 **Length:** 53 **Number of Matches:** 1

Range 1: 1 to 53 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
92.0 bits(227)	5e-27	44/58(76%)	52/58(89%)	5/58(8%)	+3

```
Query  11943  MTESGIDLGFDMEFDLNISLLNDNENMDFLPNVTVAGNSENVEFYKLKSSTRCLRSDD  12116
              MTESGIDLGFDMEFDLNI+LLNDN+NMDFLPNVT     EN+EFY+LKSS+RC+R ++
Sbjct  1      MTESGIDLGFDMEFDLNINLLNDNDNMDFLPNVT-----ENMEFYELKSSSRCIRHNE  53
```

Figure 82: BLASTx search of contig12 (query) against the 1st exon of *Mitf*-PB (subject). The methionine predicted to serve as the start codon of this isoform can be seen at the start of this alignment. This match has an e-score of 5e-27 (red box) and the aligned sequence from contig12 is in frame +3 (blue box).

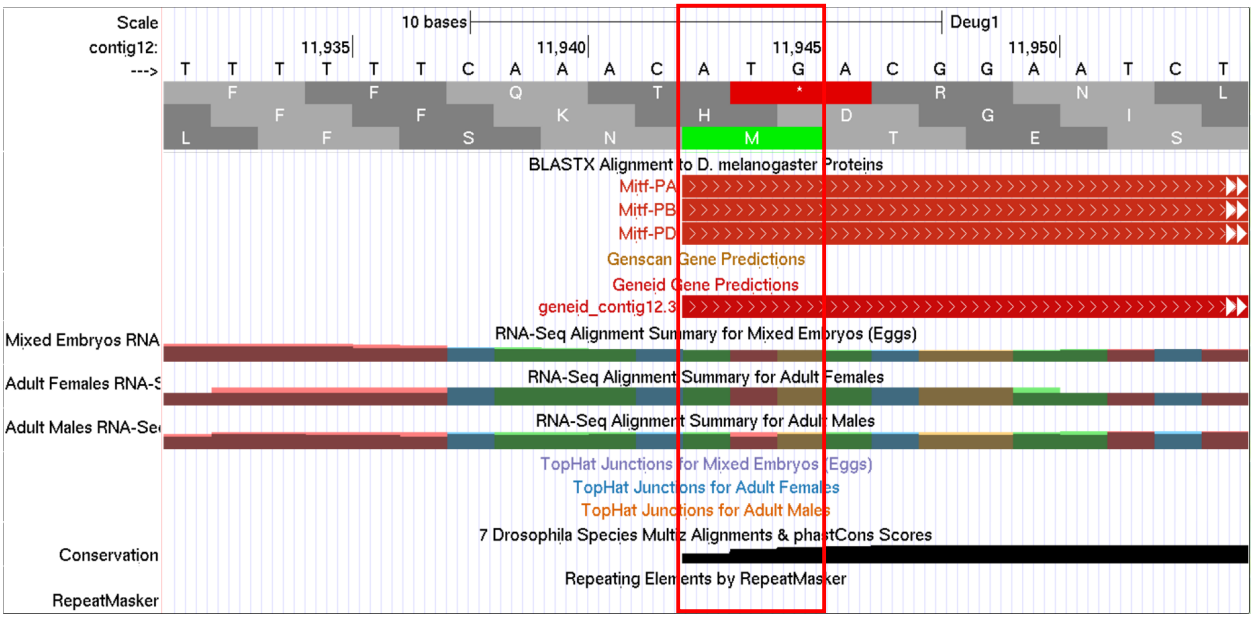


Figure 83: Annotation of the start codon of *Mitf*-PB. The methionine in frame +3 aligns with the BLASTx prediction. The RNA-Seq reads all begin upstream of the start codon, corresponding to the 5' UTR. This annotation (red box) is also supported by the Geneid prediction and by conservation data.

The first four exons of *Mitf*-PB were annotated by following the previously defined standard annotation protocol. The fifth exon of *Mitf*-PB is seven amino acids long and thus, did not produce a BLASTx alignment. The Small Exons Finder was used to locate two putative exons in the search region defined by the splice donor site of the fourth exon and the splice acceptor site of the sixth exon. The amino acid sequence of the first exon identified it as the ortholog of the 5th exon of *Mitf*-PB (Fig. 84). Examination of the position of this predicted exon in the *D. eugracilis* Genome Browser reveals that it is supported by RNA-Seq data and TopHat junctions (Fig. 85).

Sequence file: Choose File contig12.fasta
Coding Exon Type: Internal Exon
Start Position: 14767
End Position: 16182
Strand: Plus
CDS Size (aa): 7
Donor Site: GT
Acceptor Phase: 0
Donor Phase: 0
Find Small Exons | Reset

Search results

List of CDS that matched the search criteria:

Start	End	Translation	Acceptor Phase	Donor Phase	Sequence
15784	15804	LPSFDSD	0	0	CTACCATCATTCGACAGCGAT
15826	15846	KEEQNKE	0	0	AAGGAGGAACAAAATAAGGAA

Figure 84: Two putative exons predicted by the Small Exons Finder. Two exons of the expected size and phases were found in the region between the 4th and 6th exons of *Mitf*-PB. The amino acid sequence of the first putative exon, LPSFDSD, is identical to the amino acid sequence of the 5th exon of *Mitf*-PB in *D. melanogaster*. This is good evidence that this exon prediction is the ortholog to the 5th exon of *Mitf*-PB.

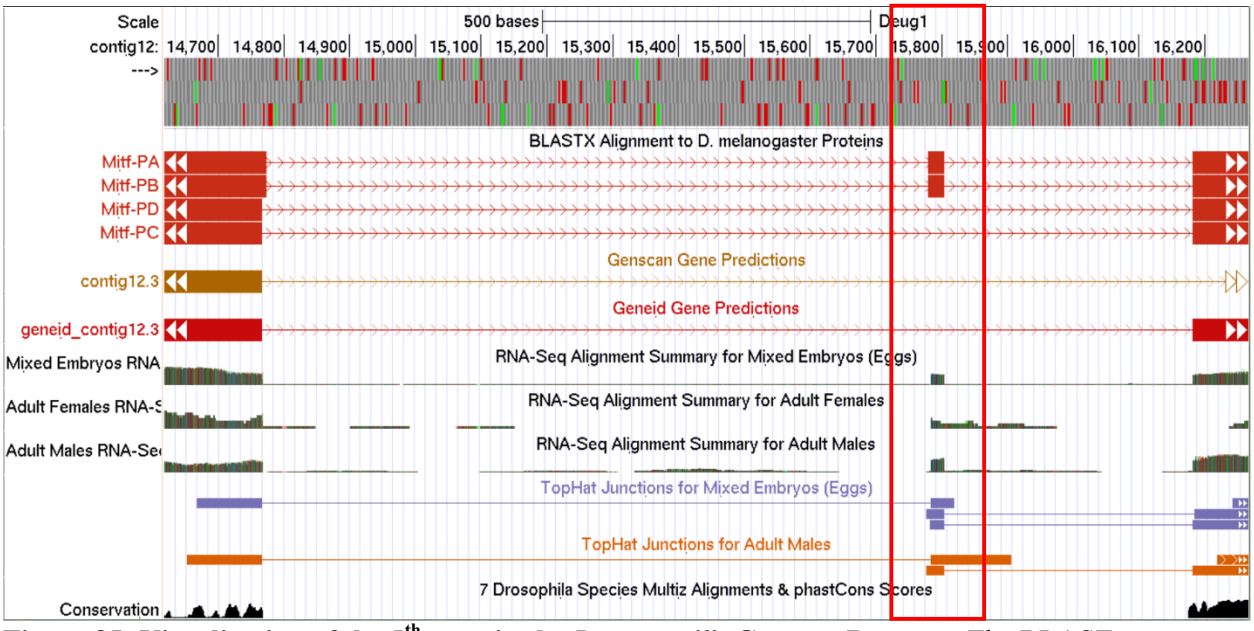


Figure 85: Visualization of the 5th exon in the *D. eugracilis* Genome Browser. The BLASTx alignment track identified the same sequence as the Small Exons Finder as the putative 5th exon (red box). This exon is supported by RNA-Seq data and TopHat junctions.

The BLASTx output for the 6th exon of *Mitf*-PB contained two alignments with reasonably high e-scores (Fig. 86). The first alignment had a lower e-score and was supported by RNA-Seq data and TopHat junctions. The second alignment occurred in a region of DNA with no prominent features downstream of the *Mitf*-PB annotation. While this second alignment was interesting, it did not obstruct the annotation of the true orthologous exon of *Mitf*-PB. However, the 7th exon of *Mitf*-PB also produced two sets of alignments in its BLASTx output, suggesting that there may be a feature worth investigating. BLAT was used to visualize all the alignments to *Mitf* exons in contig12 (Fig. 87).

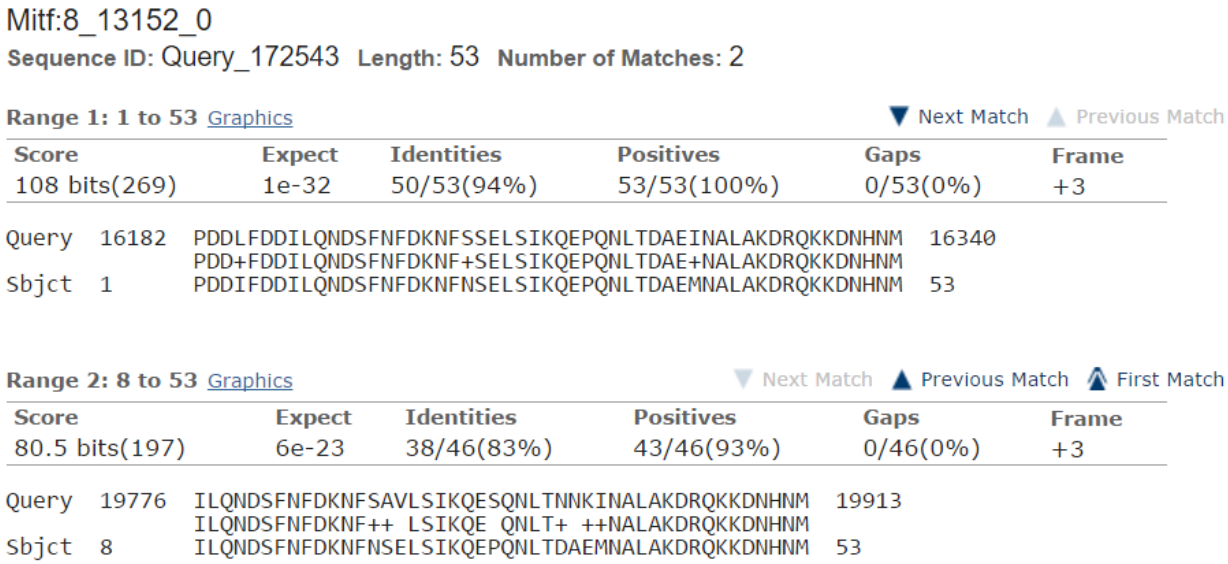


Figure 86: The BLASTx search of the 6th exon of *Mitf*-PB (subject) against contig12 (query) produced two alignments. The first alignment corresponds to an exon in the expected position that is supported by RNA-Seq data and TopHat junctions. The second alignment does not appear to correspond to an expressed feature.

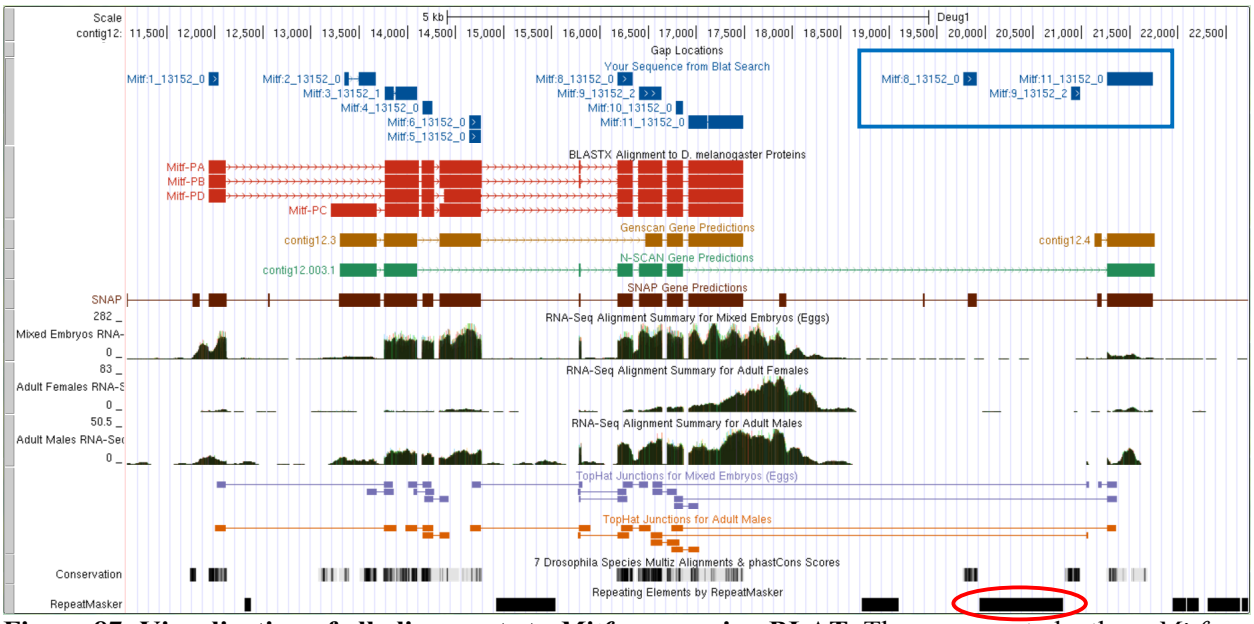


Figure 87: Visualization of all alignments to *Mitf* exons using BLAT. There appear to be three *Mitf* exons located downstream of the putative orthologous gene (blue box). There is some RNA-Seq data that corresponds to at least one of these exons; however it is possible that it was mismapped. Analysis of the repeat near these genes (red circle) reveals that it is a helitron, which uses a rolling circle mechanism to replicate itself. It is possible that, during replication, part of the *Mitf* gene was picked up and duplicated downstream.

There appear to be three exons that occur downstream of the *Mitf* ortholog. The presence of in-frame stop codons in these exons is evidence that this feature is a pseudogene and not actually expressed. The classification of a corresponding repeat as a helitron offers the possibility that these exons may have been duplicated and moved during the replication and transposition of the helitron. After addressing these duplicate exons and confirming that they do not correspond to an actively transcribed feature, no further issues were encountered in the exon-by-exon annotation of *Mitf*-PB. The coordinates of all the *Mitf*-PB exons annotated in *D. eugracilis* were assembled in Table 8.

Name of exon in *D. melanogaster*	Start/splice donor	Stop/splice acceptor	Frame	Acceptor Phase	Donor Phase	Exon BLASTp e-score
>Mitf:1_13152_0	11943	12118	+3	NA	2	5e-27
>Mitf:3_13152_1	13762	14101	+2	1	0	1e-55
>Mitf:4_13152_0	14155	14274	+1	0	0	2e-21
>Mitf:5_13152_0	14333	14767	+2	0	0	3e-48
>Mitf:7_13152_0	15784	15804	+1	0	0	NA
>Mitf:8_13152_0	16182	16341	+3	0	1	1e-32
>Mitf:9_13152_2	16403	16644	+1	2	0	3e-49
>Mitf:10_13152_0	16699	16863	+1	0	0	2e-20
>Mitf:11_13152_0	16923	17486	+3	0	NA	4e-88

Table 8: Final coding exon annotations of *Mitf*-PB. These exons were annotated using the methods described above. As *Mitf*-PA has an identical CDS to *Mitf*-PB, this set of coding exons also describes the CDS of *Mitf*-PA.

Checking the *Mitf*-PB Gene Model

After gathering the CDS annotations of *Mitf*-PB, the model was tested by using the GMC. All exons of the putative model were found to pass the tests of the basic biological rules of CDS annotation. The resulting Dot Plot and Protein Alignment are shown in Figure 88 and Figure 89.

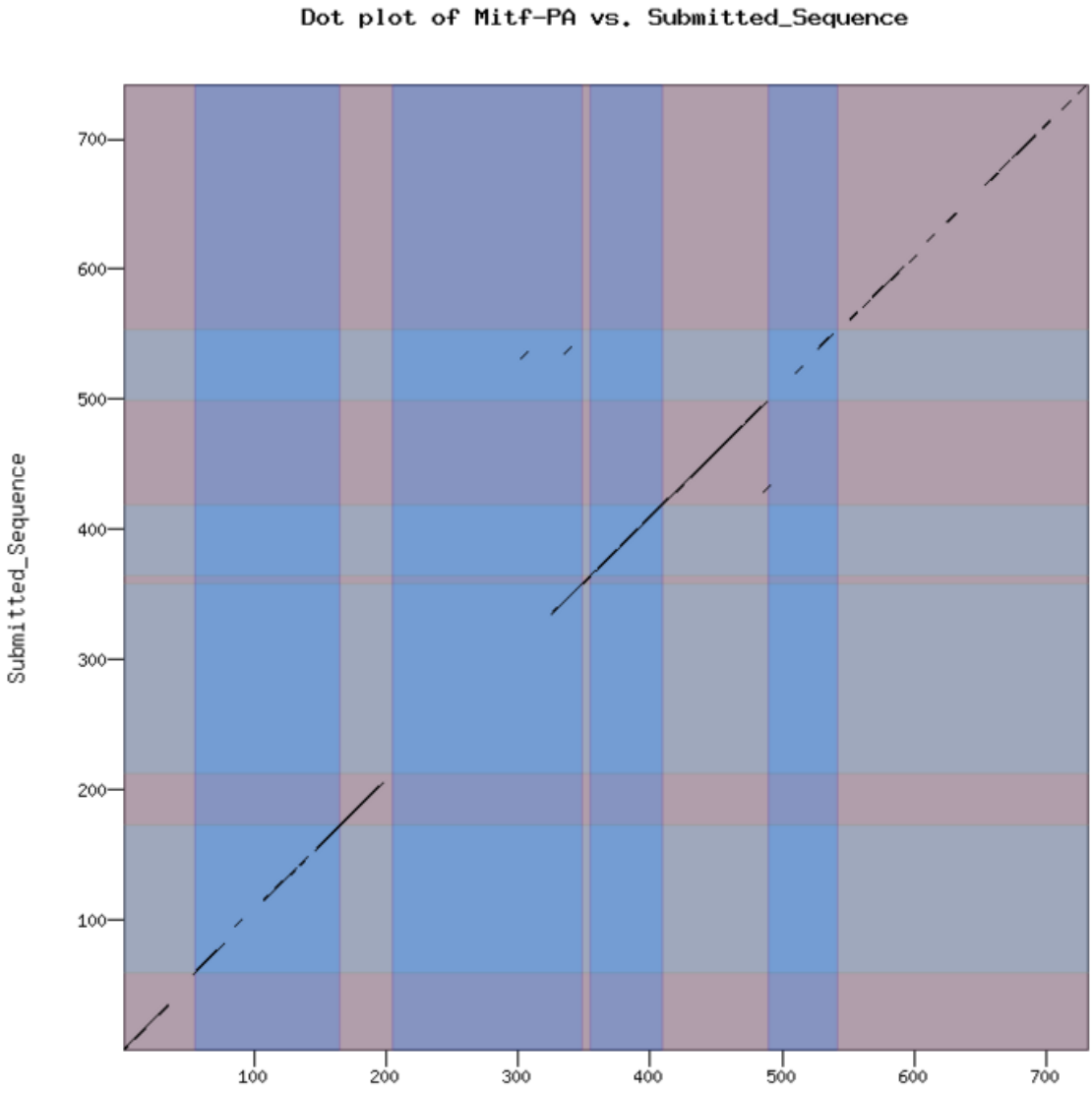


Figure 88: Dot Plot of the CDS annotation of *Mitf*-PA from *D. melanogaster* against the gene model for feature 3 of *D. eugracilis* contig12. This is the Dot Plot produced by the shared model of the CDS of *Mitf*-PA and *Mitf*-PB. The 4th exon of these gene models does not appear to be well conserved. However, analysis of the RNA-Seq data and TopHat junctions corresponding to the 4th putative exon of the *D. eugracilis* model of *Mitf*-PB confirms that there is strong evidence supporting the provided exon annotation.

Alignment of Mitf-PA vs. Submitted_Seq

View plain text version

Identity: 570/744 (76.6%), **Similarity:** 640/744 (86.0%), **Gaps:** 17/744 (2.3%)

```
Mitf-PA           1 MTESGIDLGFDMEFDLNINLLNDNDNMDFLPNVT-----ENMEFYELKSSSRCIRHNEIP 55
                    ******************.*****:*********     **:***:****:**:* ::**
Submitted_Seq     1 MTESGIDLGFDMEFDLNISLLNDNENMDFLPNVTVAGNSENVEFYKLKSSTRCLRSDDIP 60

Mitf-PA          56 TFKTATPTSRTQLKLQLQREQ----QQQMMIQQQTLDTAMDPKMHLLFGSGQGLMESEFI 111
                    ********************    ***  ***:***** * ***:***.*** :*****
Submitted_Seq    61 TFKTATPTSRTQLKLQLQREQHQQQQQQQQIQQHTLDTA-DQKMHILFGNGQGSIESEFI 119

Mitf-PA         112 DSGSTSACGSGSSSLEQMSQLVQMDNLIDSSSGAKLKVPLQSIGVDVPPQVLQVSTVLEN 171
                    ********************:.******::*...************************
Submitted_Seq   120 DSGSTSACGSGSSSLEQMSQFAQMDNLINASNAVKLKVPLQSIGVDVPPQVLQVSTVLEN 179

Mitf-PA         172 PTRYHVIQKQKNQVRQYLSESFKPSMWGSHTSEIKLANNSASTGNLQNSSLQKGICDPLE 231
                    *************************:**.:.:::*: ***::  ***.***:**** * *
Submitted_Seq   180 PTRYHVIQKQKNQVRQYLSESFKPSIWGCNNTDVKMENNSSAMINLQGSSLEKGICLPSE 239

Mitf-PA         232 RTNRFGCDSAVSAKRIMPSDDAMPISPFGGSFVRCDDINPIEPTVLRPNSHGAGEPENAH 291
                    *** **.*:** *** *.*:*:*. ***.*:* * *::** *.  *  ... :**.  *:
Submitted_Seq   240 RTNSFGSDTAVIAKRSMSSNDSMANSPFSGTFPRNDNLNPNESISLAHGNEVSGESAIAQ 299

Mitf-PA         292 RTAQLGLSKANSSL--SSTRSSSGIVNSIRISSTSSSLQSTSAPISPSVSSVATSVSELP 349
                    :* ***  *:**.*  *:**:***::***:****:*************:******.****
Submitted_Seq   300 KTNQLGYVKSNSNLNVSTTRTSSGMINSIKISSTASSLQSTSAPISPSLSSVATSASELP 359

Mitf-PA         350 SFDSDPDDIFDDILQNDSFNFDKNFNSELSIKQEPQNLTDAEMNALAKDRQKKDNHNMIE 409
                    ********:***************.***************:*****************
Submitted_Seq   360 SFDSDPDDLFDDILQNDSFNFDKNFSSELSIKQEPQNLTDAEINALAKDRQKKDNHNMIE 419

Mitf-PA         410 RRRRFNINDRIKELGTLLPKGSDAFYEVVRDIRPNKGTILKSSVDYIKCLKHEVTRLRQN 469
                    ************************************************************
Submitted_Seq   420 RRRRFNINDRIKELGTLLPKGSDAFYEVVRDIRPNKGTILKSSVDYIKCLKHEVTRLRQN 479

Mitf-PA         470 ELRQRQVELQNRKLMSRIKELEMQAKSHGILLSENHLTSLSAPTQP--YLKSFSLSPTAS 527
                    * ****************:***:****:** *** * .* *****.  ***. ****:.*
Submitted_Seq   480 ECRQRQVELQNRKLMSRIRELEIQAKSQGISLSEYHPPSASAPTQANVYLKNSSLSPSVS 539

Mitf-PA         528 RSRRSLFDQPVEKKIQVIDGADGNMGMNQVDEFMEDCKYAVQGGDPMLSSHSHMQSAPQS 587
                    **********:***  **. :* ******:**:*****:.******************
Submitted_Seq   540 RSRRSLFDQPIEKK--VISISDVNMGMNQIDELMEDCKHPVQGGDPMLSSHSHMQSAPQS 597

Mitf-PA         588 PSSKTLNSGSFEPKNFSEADE-SLFRGKSSLASDDCCGINCSTSCYIQHQLTSREGHPNH 646
                    ****.:: *****:::****   *****:.*  ** *.*******:**:**..::*:*.*
Submitted_Seq   598 PSSKPMDYGSFEPQKISEADAVRLFRGKTNLLIDDLCSINCSTSCFIQNQLPKHQGRPSH 657

Mitf-PA         647 SHHTGIREIHSLSDSAQNSEFSRLEHCDDPLLSSSHRSLGTVDEDQHNSVDMSAVMVHDS 706
                    .*  .  :********* .*******:*:*****************:*.*:******::**
Submitted_Seq   658 CHPPRAKEIHSLSDSATSSEFSRLEQCEDPLLSSSHRSLGTVDEDHHSSIDMSAVMINDS 717

Mitf-PA         707 LSSLVDDNNSETMVLASDTLDIEL 730
                    :***.******.*:*:.*****:*
Submitted_Seq   718 FSSLADDNNSEPMLLTPDTLDIDL 741
```

Figure 89: Protein alignment of *Mitf*-PA annotation. This is protein alignment produced by the comparison of *D. melanogaster Mitf*-PA and *Mitf*-PB against the shared model of these isoforms' CDSs in *D. eugracilis*.

Exon-by-Exon Annotation of *Mitf*-PD

Mitf-PD differs from *Mitf*-PB and *Mitf*-PA in that it has a shorter 4th exon and does not contain the short 5th exon that required the Small Exons Finder to locate. Standard CDS annotation protocol was followed to annotate this unique 4th exon. The exon table was then updated to reflect the change in the 4th exon and the omission of the short 5th exon (Table 9).

Name of exon in *D. melanogaster*	Exon	Start/splice donor	Stop/splice acceptor	Frame	Acceptor Phase	Donor Phase	Exon BLASTp e-score
>Mitf:1_13152_0	1	11943	12118	+3	NA	2	5e-27
>Mitf:3_13152_1	2	13762	14101	+2	1	0	1e-55
>Mitf:4_13152_0	3	14155	14274	+1	0	0	2e-21
>Mitf:6_13152_0	4	14414	14767	+2	0	0	9e-42
>Mitf:8_13152_0	5	16182	16341	+3	0	1	1e-32
>Mitf:9_13152_2	6	16403	16644	+1	2	0	3e-49
>Mitf:10_13152_0	7	16699	16863	+1	0	0	2e-20
>Mitf:11_13152_0	8	16923	17486	+3	0	NA	4e-88

Table 9: Final coding exon annotations of *Mitf*-PD. This table is identical to the table of coding exons of *Mitf*-PD except for the shorter 4th exon and the omission of the short 5th exon.

Checking the *Mitf*-PD Gene Model

The proposed model of *Mitf*-PD was tested by using the GMC. All exons of the putative gene model were found to pass the tests of the basic biological rules of CDS annotation. The resulting Dot Plot and Protein Alignment are shown in Figure 90 and Figure 91.

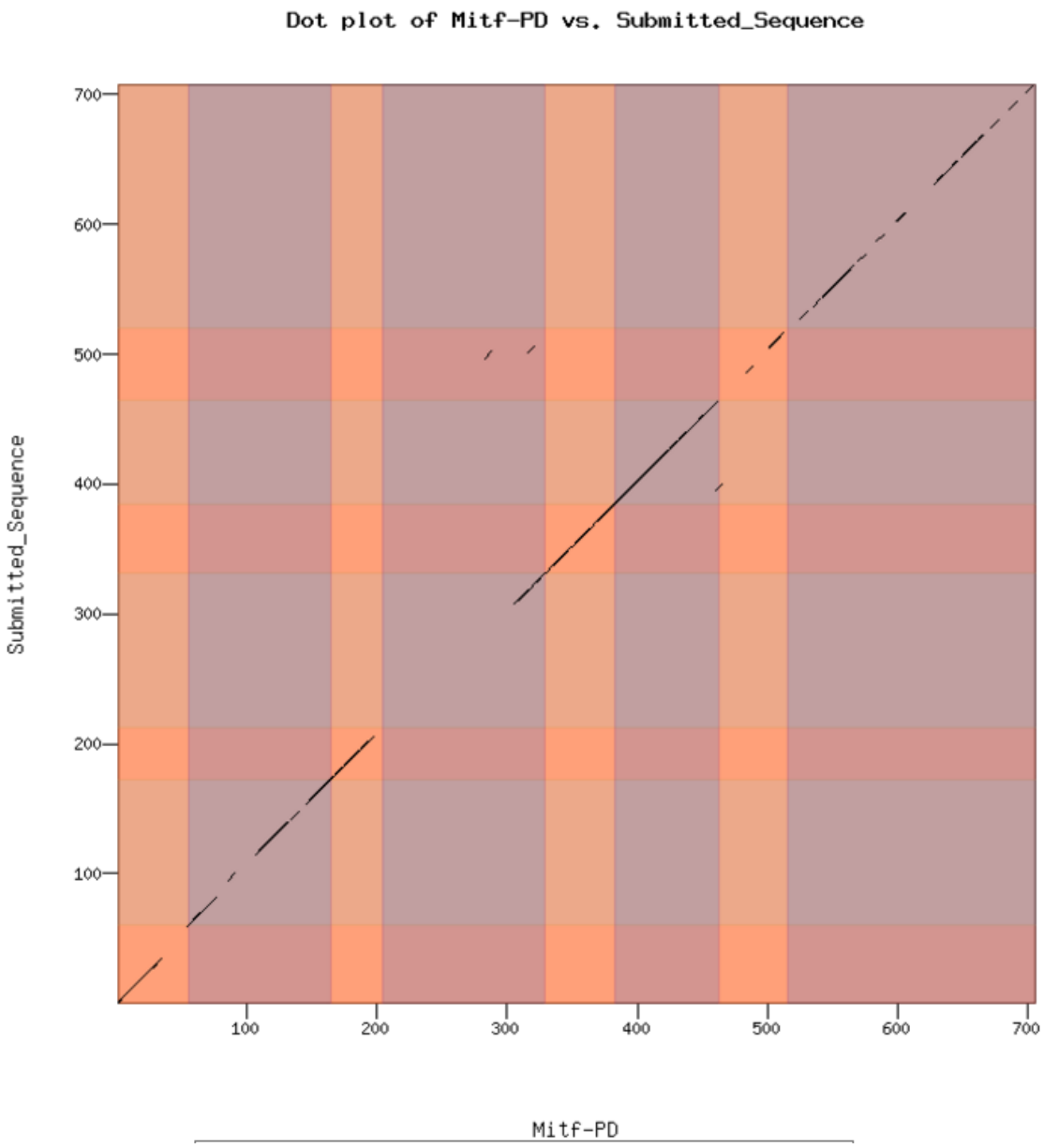


Figure 90: Dot Plot of *D. melanogaster Mitf*-PD vs. the proposed model of *D. eugracilis Mitf*-PD.

Alignment of Mitf-PD vs. Submitted_Seq

View plain text version

Identity: 548/718 (76.3%), **Similarity:** 612/718 (85.2%), **Gaps:** 25/718 (3.5%)

```
Mitf-PD           1 MTESGIDLGFDMEFDLNINLLNDNDNMDFLPNVT-----ENMEFYELKSSSRCIRHNEIP 55
                    ******************.*****:*********     **:***:****:**:* ::**
Submitted_Seq     1 MTESGIDLGFDMEFDLNISLLNDNENMDFLPNVTVAGNSENVEFYKLKSSTRCLRSDDIP 60

Mitf-PD          56 TFKTATPTSRTQLKLQLQREQ----QQQMMIQQQTLDTAMDPKMHLLFGSGQGLMESEFI 111
                    ********************    ***  ***:***** * ***:***.*** :*****
Submitted_Seq    61 TFKTATPTSRTQLKLQLQREQHQQQQQQQQIQQHTLDTA-DQKMHILFGNGQGSIESEFI 119

Mitf-PD         112 DSGSTSACGSGSSSLEQMSQLVQMDNLIDSSSGAKLKVPLQSIGVDVPPQVLQVSTVLEN 171
                    **********************:.******::*...*************************
Submitted_Seq   120 DSGSTSACGSGSSSLEQMSQFAQMDNLINASNAVKLKVPLQSIGVDVPPQVLQVSTVLEN 179

Mitf-PD         172 PTRYHVIQKQKNQVRQYLSESFKPSMWGSHTSEKGICDPLERTNRFGCDSAVSAKRIMPS 231
                    ************************:**        *:  :*** **.*:** *** *.*
Submitted_Seq   180 PTRYHVIQKQKNQVRQYLSESFKPSIWG--------CNNTDRTNSFGSDTAVIAKRSMSS 231

Mitf-PD         232 DDAMPISPFGGSFVRCDDINPIEPTVLRPNSHGAGEPENAHRTAQLGLSKANSSL--SST 289
                    :*:*. ***.*:* * *::** *.  *   ... :**.  *::* ***  *:**.*  *:*
Submitted_Seq   232 NDSMANSPFSGTFPRNDNLNPNESISLAHGNEVSGESAIAQKTNQLGYVKSNSNLNVSTT 291

Mitf-PD         290 RSSSGIVNSIRISSTSSSLQSTSAPISPSVSSVATSVSEPDDIFDDILQNDSFNFDKNFN 349
                    *:***::***:****:************:******.*****:****************.
Submitted_Seq   292 RTSSGMINSIKISSTASSLQSTSAPISPSLSSVATSASEPDDLFDDILQNDSFNFDKNFS 351

Mitf-PD         350 SELSIKQEPQNLTDAEMNALAKDRQKKDNHNMIERRRRFNINDRIKELGTLLPKGSDAFY 409
                    ****************:*****************************************
Submitted_Seq   352 SELSIKQEPQNLTDAEINALAKDRQKKDNHNMIERRRRFNINDRIKELGTLLPKGSDAFY 411

Mitf-PD         410 EVVRDIRPNKGTILKSSVDYIKCLKHEVTRLRQNELRQRQVELQNRKLMSRIKELEMQAK 469
                    ********************************** ***************:***.***
Submitted_Seq   412 EVVRDIRPNKGTILKSSVDYIKCLKHEVTRLRQNECRQRQVELQNRKLMSRIRELEIQAK 471

Mitf-PD         470 SHGILLSENHLTSLSAPTQP--YLKSFSLSPTASRSRRSLFDQPVEKKIQVIDGADGNMG 527
                    *:** *** * .* *****.  ***. ****:.**********:***  **. :* ***
Submitted_Seq   472 SQGISLSEYHPPSASAPTQANVYLKNSSLSPSVSRSRRSLFDQPIEKK--VISISDVNMG 529

Mitf-PD         528 MNQVDEFMEDCKYAVQGGDPMLSSHSHMQSAPQSPSSKTLNSGSFEPKNFSEADE-SLFR 586
                    ***:**:*****:.**********************.:: *****:::****   ***
Submitted_Seq   530 MNQIDELMEDCKHPVQGGDPMLSSHSHMQSAPQSPSSKPMDYGSFEPQKISEADAVRLFR 589

Mitf-PD         587 GKSSLASDDCCGINCSTSCYIQHQLTSREGHPNHSHHTGIREIHSLSDSAQNSEFSRLEH 646
                    **:.*  ** *.*******:**:**.::*:*.*.*  .  :********* .*******.
Submitted_Seq   590 GKTNLLIDDLCSINCSTSCFIQNQLPKHQGRPSHCHPPRAKEIHSLSDSATSSEFSRLEQ 649

Mitf-PD         647 CDDPLLSSSHRSLGTVDEDQHNSVDMSAVMVHDSLSSLVDDNNSETMVLASDTLDIEL 704
                    *:*****************:*.*:******::**:***.******.*:*:.*****.*
Submitted_Seq   650 CEDPLLSSSHRSLGTVDEDHHSSIDMSAVMINDSFSSLADDNNSEPMLLTPDTLDIDL 707
```

Figure 91: Protein alignment of *D. melanogaster Mitf*-PD vs. the proposed model of *D. eugracilis Mitf*-PD.

Exon-by-Exon Annotation of *Mitf*-PC

Mitf-PC differs from *Mitf*-PB and *Mitf*-PA in that it has a unique 1st exon that does not produce a gap-spanning intron between itself and the 2nd exon. *Mitf*-PC also does not contain the short 5th exon. Standard CDS annotation protocol was followed to annotate this unique 1st exon. As this project places particular emphasis on the annotation of the initial methionine of the coding spans of these genes, the BLASTx output used to place the initial coding exon is included in Figure 92. The evidence tracks supporting the start codon annotation in the *D. eugracilis* Genome Browser are shown in Figure 93 The exon table from *Mitf*-PB was then updated to reflect the change in the 1st exon and the loss of the short 5th exon (Table 10).

Mitf:2_13152_0
Sequence ID: Query_35013 **Length:** 167 **Number of Matches:** 1

Range 1: 1 to 167 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
150 bits(380)	1e-45	86/167(51%)	112/167(67%)	11/167(6%)	+1

```
Query  13210  MFTYWSKKSIPSKVDMKKTKTITKCIRPE----EMDKEKPEQPTINTHVLEFQDFLRRVQ  13377
              MF+YW+K++I S VD KK   ++   +      +M+K++   P      LEFQDFL+RVQ
Sbjct  1      MFSYWTKQNISSSVDGKKMDKVSTMRKKNSTKVKMEKKETVVPITEMRALEFQDFLKRVQ  60

Query  13378  TLQMSHIDAND--NQVLTGNRYNIEIGEE---LREANNIGQKSS--VLRSTKCDCMPMPA  13536
              TLQMSH ++ D  N     +  +IE+GEE      A+NI +     VLRS+ CDC+P+
Sbjct  61     TLQMSHNESRDLVNGPNICSSMDIEVGEEPPTTSFADNIDKACPLPVLRSSDCDCVPLCM  120

Query  13537  ANVRVTVGIDKDLEMILEMDPSIVDLGDVGIAEMAGPRLVGLPPLSG  13677
               N+RV+VGIDKDLEMILEMDPSIVDLGD+ I E A PR+VGLPPLSG
Sbjct  121    TNIRVSVGIDKDLEMILEMDPSIVDLGDISIDETAEPRIVGLPPLSG  167
```

Figure 92: BLASTx search of contig12 (query) against the 1st exon of *Mitf*-PC in *D. melanogaster* (subject). The methionine predicted to serve as the start codon of this isoform can be seen at the start of this alignment. This match has an e-score of 1e-45 (red box) and the aligned sequence from contig12 is in frame +1 (blue box).

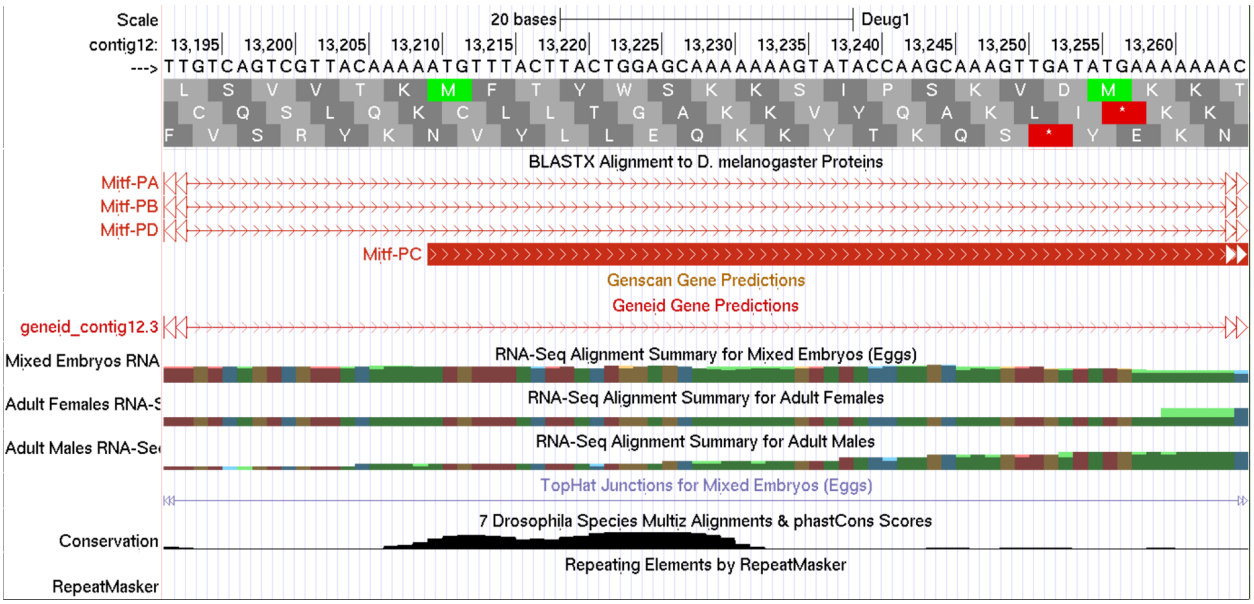


Figure 93: Annotation of the start codon of *Mitf*-PC. There are two candidate methionines in the correct reading frame (+1). However, the BLASTx alignment supports the upstream methionine located at 13210.

Name of exon in *D. melanogaster*	Start/splice donor	Stop/splice acceptor	Frame	Acceptor Phase	Donor Phase	Exon BLASTp e-score
>Mitf:2_13152_0	13210	13679	+1	NA	2	1e-45
>Mitf:3_13152_1	13762	14101	+2	1	0	1e-55
>Mitf:4_13152_0	14155	14274	+1	0	0	2e-21
>Mitf:5_13152_0	14333	14767	+2	0	0	3e-48
>Mitf:8_13152_0	16182	16341	+3	0	1	1e-32
>Mitf:9_13152_2	16403	16644	+1	2	0	3e-49
>Mitf:10_13152_0	16699	16863	+1	0	0	2e-20
>Mitf:11_13152_0	16923	17486	+3	0	NA	4e-88

Table 10: Final coding exon annotations of *Mitf*-PC. This table is identical to the table of coding exons of *Mitf*-PB except for the unique initial coding exon and the loss of the short 5th exon.

Checking the *Mitf*-PC Gene Model

The proposed model of *Mitf*-PC was tested by using the GMC. All exons of the putative gene model were found to pass the tests of the basic biological rules of CDS annotation. The resulting Dot Plot and Protein Alignment are shown in Figure 94 and Figure 95.

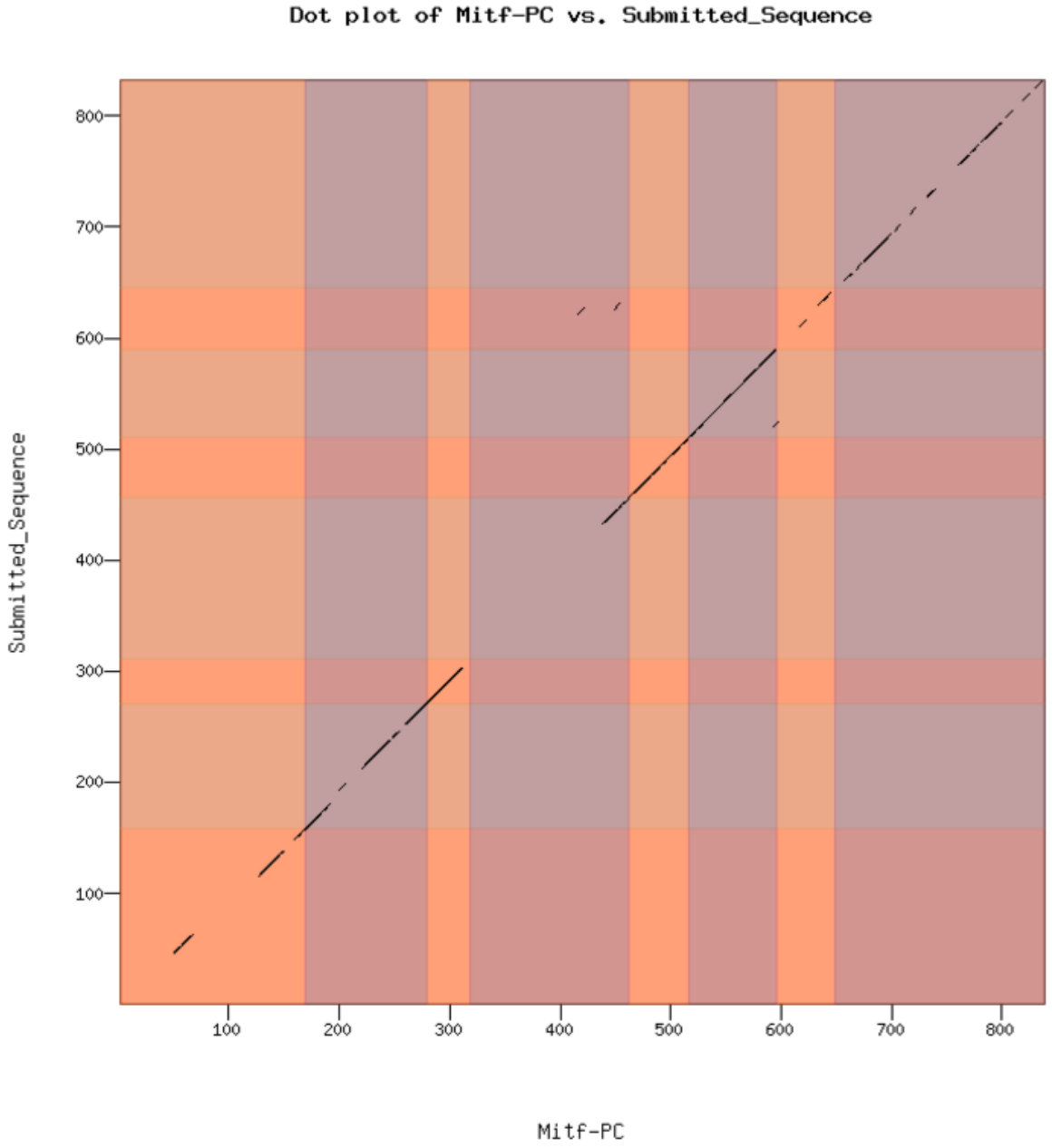


Figure 94: Dot Plot of *D. melanogaster Mitf*-PC vs. the proposed model of *D. eugracilis Mitf*-PC.

Alignment of Mitf-PC vs. Submitted_Seq

View plain text version

Identity: 605/846 (71.5%), **Similarity:** 693/846 (81.9%), **Gaps:** 23/846 (2.7%)

```
Mitf-PC           1 MFSYWTKQNISSSVDGKKMDKVSTMRKKNSTKVKMEKKETVVPITEMRALEFQDFLKRVQ 60
                    **:**:*:.*.*.** ** ..::.  : :    :*:*::.  *  : :.*******:***
Submitted_Seq     1 MFTYWSKKSIPSKVDMKKTKTITKCIRPE----EMDKEKPEQPTINTHVLEFQDFLRRVQ 56

Mitf-PC          61 TLQMSHNESRDLVNGPNICSSMDIEVGEEPPTTSFADNIDKACPLPVLRSSDCDCVPLCM 120
                    ****** ::.*  *      . :**:***      *:**.: ..  ****:.***:*:
Submitted_Seq     57 TLQMSHIDAND--NQVLTGNRYNIEIGEE---LREANNIGQKSS--VLRSTKCDCMPMPA 109

Mitf-PC          121 TNIRVSVGIDKDLEMILEMDPSIVDLGDISIDETAEPRIVGLPPLSGGPTFKTATPTSRT 180
                    :*:**:********************:.* * * **:*******************
Submitted_Seq    110 ANVRVTVGIDKDLEMILEMDPSIVDLGDVGIAEMAGPRLVGLPPLSGGPTFKTATPTSRT 169

Mitf-PC          181 QLKLQLQREQ----QQQMMIQQQTLDTAMDPKMHLLFGSGQGLMESEFIDSGSTSACGSG 236
                    **********    ***  ***:***** * ***.***.*** .***************
Submitted_Seq    170 QLKLQLQREQHQQQQQQQQIQQHTLDTA-DQKMHILFGNGQGSIESEFIDSGSTSACGSG 228

Mitf-PC          237 SSSLEQMSQLVQMDNLIDSSSGAKLKVPLQSIGVDVPPQVLQVSTVLENPTRYHVIQKQK 296
                    *********:.******::*...**********************************
Submitted_Seq    229 SSSLEQMSQFAQMDNLINASNAVKLKVPLQSIGVDVPPQVLQVSTVLENPTRYHVIQKQK 288

Mitf-PC          297 NQVRQYLSESFKPSMWGSHTSEIKLANNSASTGNLQNSSLQKGICDPLERTNRFGCDSAV 356
                    **************.**.:.:::*: ***::  ***.***:**** * **** **.*:**
Submitted_Seq    289 NQVRQYLSESFKPSIWGCNNTDVKMENNSSAMINLQGSSLEKGICLPSERTNSFGSDTAV 348

Mitf-PC          357 SAKRIMPSDDAMPISPFGGSFVRCDDINPIEPTVLRPNSHGAGEPENAHRTAQLGLSKAN 416
                      *** *.*:*:*. ***.*:* * *::** *.  *  ... :**.  *::* ***  *:*
Submitted_Seq    349 IAKRSMSSNDSMANSPFSGTFPRNDNLNPNESISLAHGNEVSGESAIAQKTNQLGYVKSN 408

Mitf-PC          417 SSL--SSTRSSSGIVNSIRISSTSSSLQSTSAPISPSVSSVATSVSEPDDIFDDILQNDS 474
                    *.*  *:**:***::***:****:*************:******.*****:*********
Submitted_Seq    409 SNLNVSTTRTSSGMINSIKISSTASSLQSTSAPISPSLSSVATSASEPDDLFDDILQNDS 468

Mitf-PC          475 FNFDKNFNSELSIKQEPQNLTDAEMNALAKDRQKKDNHNMIERRRRFNINDRIKELGTLL 534
                    *******.***************:**********************************
Submitted_Seq    469 FNFDKNFSSELSIKQEPQNLTDAEINALAKDRQKKDNHNMIERRRRFNINDRIKELGTLL 528

Mitf-PC          535 PKGSDAFYEVVRDIRPNKGTILKSSVDYIKCLKHEVTRLRQNELRQRQVELQNRKLMSRI 594
                    ************************************** ***************
Submitted_Seq    529 PKGSDAFYEVVRDIRPNKGTILKSSVDYIKCLKHEVTRLRQNECRQRQVELQNRKLMSRI 588

Mitf-PC          595 KELEMQAKSHGILLSENHLTSLSAPTQP--YLKSFSLSPTASRSRRSLFDQPVEKKIQVI 652
                    :***:****:** *** * .* *****.  ***. ****:.**********:***  **
Submitted_Seq    589 RELEIQAKSQGISLSEYHPPSASAPTQANVYLKNSSLSPSVSRSRRSLFDQPIEKK--VI 646

Mitf-PC          653 DGADGNMGMNQVDEFMEDCKYAVQGGDPMLSSHSHMQSAPQSPSSKTLNSGSFEPKNFSE 712
                    . .:* ******:**:*****:.**********************.:: *****:::**
Submitted_Seq    647 SISDVNMGMNQIDELMEDCKHPVQGGDPMLSSHSHMQSAPQSPSSKPMDYGSFEPQKISE 706

Mitf-PC          713 ADE-SLFRGKSSLASDDCCGINCSTSCYIQHQLTSREGHPNHSHHTGIREIHSLSDSAQN 771
                    **   *****:.*  ** *.*******:**:**.:::*:*.*.* .  .:********* .
Submitted_Seq    707 ADAVRLFRGKTNLLIDDLCSINCSTSCFIQNQLPKHQGRPSHCHPPRAKEIHSLSDSATS 766

Mitf-PC          772 SEFSRLEHCDDPLLSSSHRSLGTVDEDQHNSVDMSAVMVHDSLSSLVDDNNSETMVLASD 831
                    *******:*:***************:*.*:******::**:***.******.*:*:.*
Submitted_Seq    767 SEFSRLEQCEDPLLSSSHRSLGTVDEDHHSSIDMSAVMINDSFSSLADDNNSEPMLLTPD 826

Mitf-PC          832 TLDIEL 837
                    ****:*
Submitted_Seq    827 TLDIDL 832
```

Figure 95: Protein alignment of *D. melanogaster Mitf*-PC vs. the proposed model of *D. eugracilis Mitf*-PC.

Identification of the Transcription Start Sites of *Mitf* in *D. melanogaster*

The TSSs of *Mitf* were annotated after confirming the CDS annotation of the gene to be the most parsimonious model to the *D. melanogaster* ortholog. The untranslated exons of *Mitf* have already been annotated in *D. melanogaster.* These untranslated exons were examined in the Gene Record Finder and it was determined that *Mitf*-RA, *Mitf*-RD, and *Mitf*-RC have unique TSSs. *Mitf*-RB shares a TSS with *Mitf*-RD. The untranslated exons can be seen in Figure 96. Before work was begun on the annotation of TSSs in *D. eugracilis*, *Mitf* was examined in the *D. biarmipes* Genome Browser. However, there was no RNA polII data associated with the TSS of *Mitf* in *D. biarmipes*, so these TSS annotations were performed without the use of *D. biarmipes* RNA polII data.

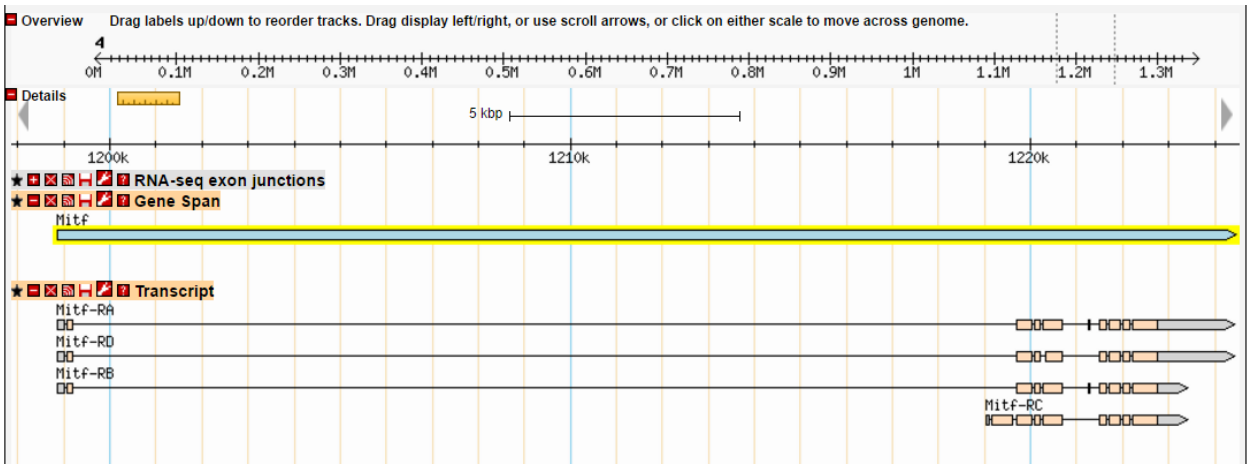


Figure 96: FlyBase GBrowse view of *Mitf*. The unique TSS of *Mitf*-RC can be seen clearly. The TSS of *Mitf*-RA is reported to occur 6 bp upstream of the TSS of *Mitf*-RD and *Mitf*-RB.

Classification of the TSS of *Mitf*-RA in *D. melanogaster*

The evidence tracks corresponding to *Mitf*-RA in the *D. melanogaster* Genome Browser were used to classify this TSS as broad. There are two Celniker TSSs and two DHS positions associated with the TSS of this isoform (Fig. 97).

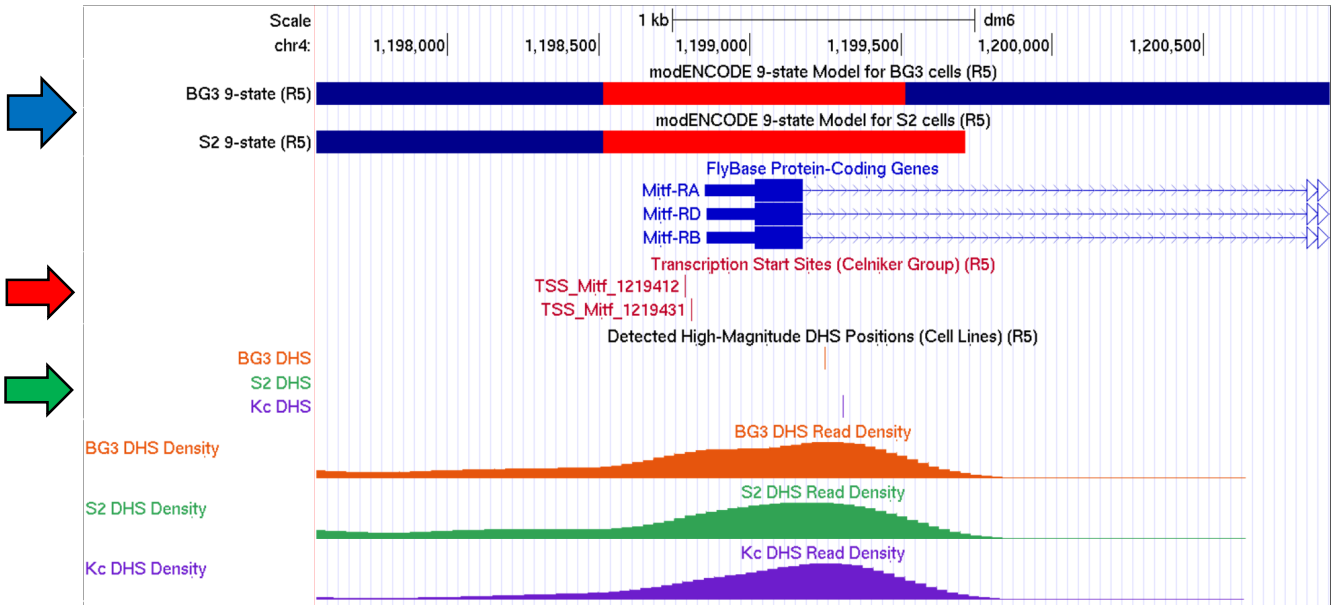


Figure 97: Classification of the *Mitf*-RA promoter as broad. Both 9-state chromatin models are red, indicating an active promoter/TSS region (blue arrow). There are two annotated Celniker TSSs (red arrow) as well as two separate DHS positions (green arrow).

Definition of a Putative TSS for *Mitf*-RA in *D. eugracilis*

After classifying the promoter of *Mitf*-RA as broad, a BLASTn search was performed using the initial untranslated exon of *Mitf*-RA as the query and the sequence of contig12 as the subject. The resulting alignment is shown in Figure 98. This alignment occurs in the expected region of contig12 (Fig. 99).

Deug1_dna range=contig12:1-38500 5'pad=0 3'pad=0 strand=+ repeatMasking=none
Sequence ID: Query_48183 **Length:** 38500 **Number of Matches:** 1

Range 1: 11795 to 12118 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
158 bits(109)	3e-41	234/343(68%)	43/343(12%)	Plus/Plus

```
Query  6      TGTTATGAAACTTTAGGTATTCATTAGTAGTT--CCGAATCTTTAACTGCTGAGGCGAAA  63
              ||||||||||| || || ||||||| |  | ||  |||  | ||   ||||
Sbjct  11795  TGTTATGAAATTTGAGATATTCATAACCACTTATCCGCCTGTTCGGCTGC----------  11844

Query  64     TTATGTGAACTTCATAATTAGTAAAAAATAAAAATTGTTCTATGTATATTGTTAGTTGGG  123
               |||    ||  ||| ||   | || || ||| ||||| ||   || | |||| | | |
Sbjct  11845  --ATG----CTATATAGTTTAAAGAATATTAAATTTGTTTTA---ATTTAGTTATTGGAG  11895

Query  124    TA--TATAAAATTTCCTAAG--TATAAAAAATTTAATCCT---AAATATGACGGAATCTG  176
              |   ||||||  |   |||   |||| | | ||   |  |   ||| ||||||||||||||
Sbjct  11896  TCAGTATAAATATAAATAAACCTATATATAGTTCTCTTTTTTCAAACATGACGGAATCTG  11955

Query  177    GAATCGATTTGGGCTTTGATATGGAGTTTGATCTAAATATAAACTTGTTGAATGATAACG  236
              |||||||||||||||||||||||||||| |||||||||||| | |||| ||||||||||
Sbjct  11956  GAATCGATTTGGGCTTTGATATGGAGTTCGATCTAAATATAAGCCTGTTAAATGATAACG  12015

Query  237    ACAATATGGATTTCTTACCAAATGT---------------TACTGAGAATATGGAGTTCT  281
              | |||||||||| ||||||||||                | |||| ||| ||||||||||
Sbjct  12016  AGAATATGGATTTTTTACCAAATGTAACTGTAGCAGGCAATTCTGAAAATGTGGAGTTCT  12075

Query  282    ATGAATTGAAATCATCATCACGCTGTATACGGCATAATGAAAT  324
              || |||| || ||||||| | |||||  |||||    |||| ||
Sbjct  12076  ATAAATTAAAGTCATCAACGCGCTGCTTACGGTCCGATGACAT  12118
```

Figure 98: BLASTn alignment of the initial untranslated exon of *Mitf*-RA in *D. melanogaster* (query) against contig12 (subject). This alignment has an e-score of 3e-41 (red box). This start of this alignment is to the 6th base of the *Mitf*-RA untranslated exon (blue box). Therefore, the position of the putative TSS in *D. eugracilis* was extrapolated to be 11790.

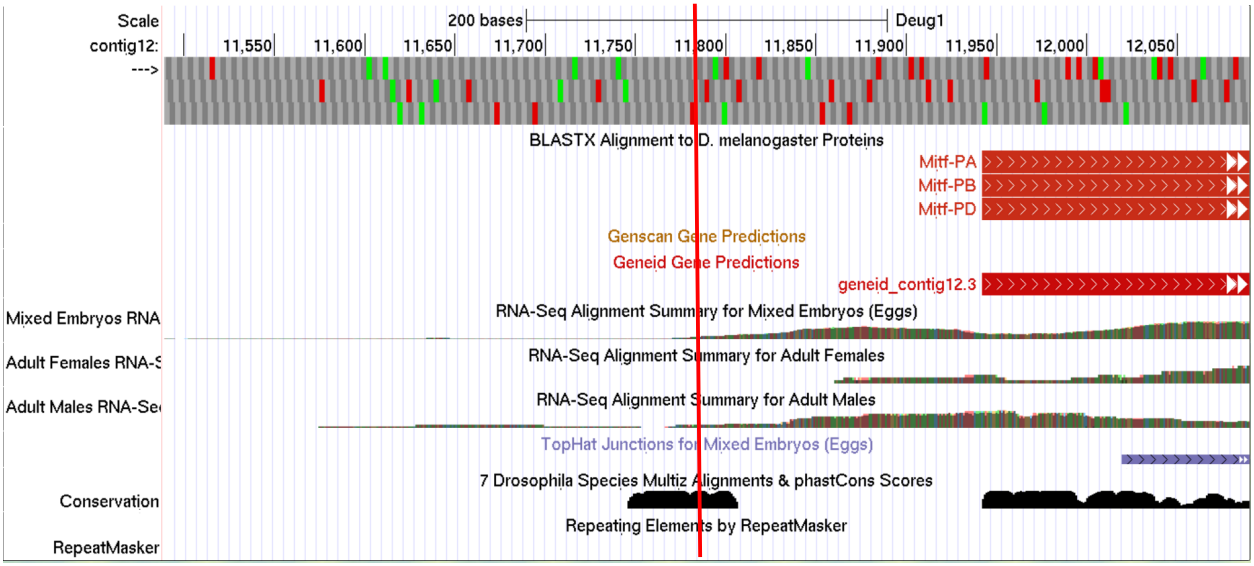


Figure 99: Position of the putative TSS in contig12. The putative TSS of 11790 determined by the BLASTn alignment is indicated by the red line. This putative TSS is supported by conservation data and RNA-Seq data, which appears to suggest that the initial untranslated exon begins downstream of 11790.

After using the BLASTn alignment to establish a putative TSS at 11790, a search region consisting of the 300 bp upstream and 300 bp downstream was defined to search for *Drosophila* core promoter motifs from position 11490-12090 (Fig. 100). A similar search region (1198552-1199152) was used to search for core promoter motifs around the annotated TSS of *Mitf*-RA in *D. melanogaster* using the short match function in the *D. melanogaster* Genome Browser. The results of both of these core promoter motif searches were organized into a table (Table 11). A DRE motif was found in the search region in both *D. eugracilis* and *D. melanogaster.* As DRE motifs are not associated with a specific position, these motifs were highlighted in Table 11. With the search for *Drosophila* core promoter motifs complete, 11790 was established as the putative TSS for *Mitf*-RA in *D. eugracilis*.

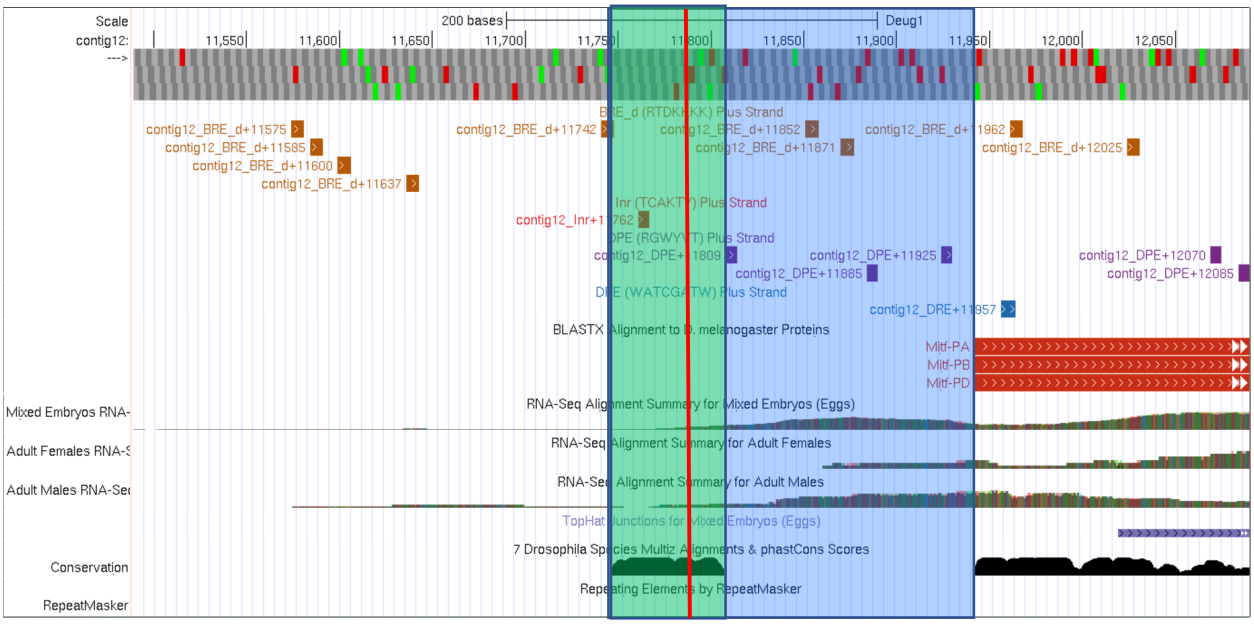


Figure 100: Search for core promoter motifs in *D. eugracilis*. None of the motifs associated with a specific position correspond to the putative TSS at 11790 (red line). There is a DRE motif in this region (shown in the blue evidence track). As DRE motifs do not correspond to a specific position, this motif was highlighted in Table 11, indicating that it is associated with the putative TSS. In addition, narrow and broad TSS search regions were defined using RNA-Seq and conservation data. The narrow search region (11747-11807) encompasses the highly conserved region associated with the putative TSS (green box). The broad search region (11747-11943) extends from the start of the highly conserved region to the start codon of the first coding exon (blue box).

Core promoter motif	*D. eugracilis*	*D. melanogaster*
BRE^{u}	NA	NA
TATA Box	NA	1198975, 1198993
BRE^{d}	11575, 11585, 11600, 11637, 11742, 11852, 118871, 11962, 12025	1198707, 1198804, 1198960, 1198969, 1199034,
Inr	11762	NA
MTE	NA	NA
DPE	11809, 11885, 11925, 12070, 12085	1198582, 1198670, 1198775, 1199127
Ohler_motif1	NA	NA
DRE	11957	1199029
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	NA	NA
Ohler_motif8	NA	NA

Table 11: Core promoter motifs found in the TSS search region of *Mitf*-RA in *D. eugracilis* and *D. melanogaster*. There is a DRE motif in both species that is associated with the TSS.

Classification of the TSS of *Mitf*-RD and *Mitf*-RB in *D. melanogaster*

The evidence tracks corresponding to *Mitf*-RD in the *D. melanogaster* Genome Browser were used to classify this TSS as broad. Since this TSS is so close to the TSS of *Mitf*-RA, the evidence used to classify this promoter is the same as the evidence used to classify the promoter of *Mitf*-RD. There are two Celniker TSSs and two DHS positions associated with the TSS of this isoform (Fig. 101). GEP annotation protocol states that, to create the most parsimonious model to *D. melanogaster,* this TSS should be annotated separately from the TSS of *Mitf*-RA, located six bp upstream.

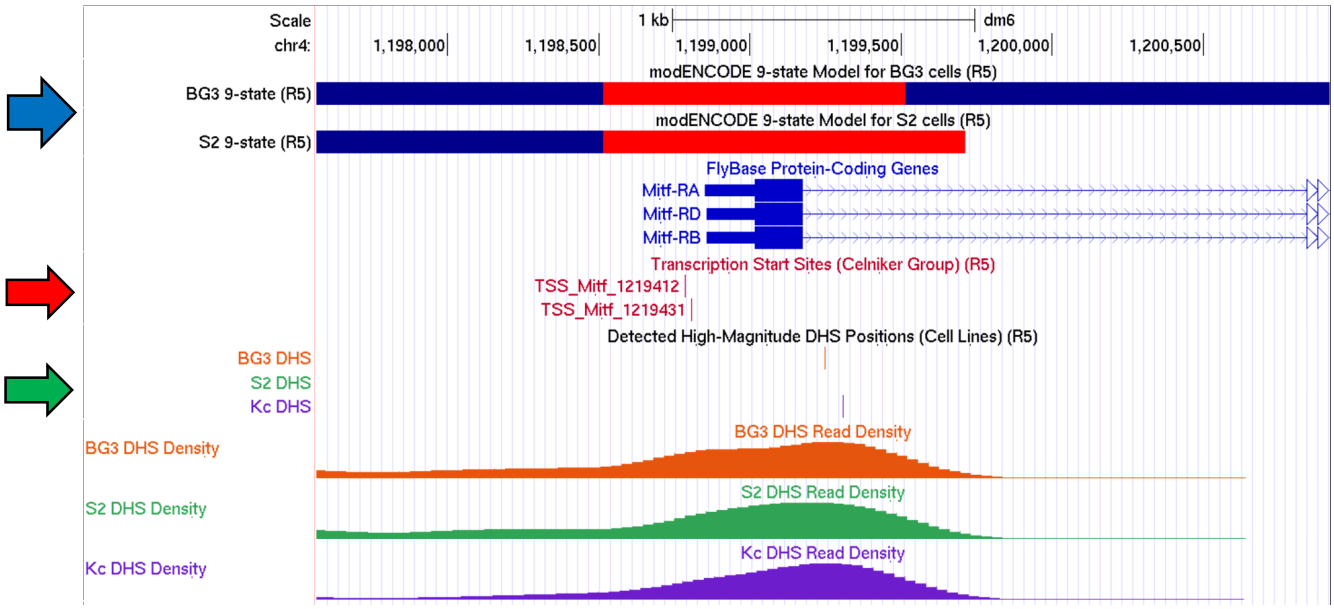


Figure 101: Classification of the *Mitf*-RD promoter as broad. Both 9-state chromatin models are red, indicating an active promoter/TSS region (blue arrow). There are two annotated Celniker TSSs (red arrow) as well as two separate DHS positions (green arrow).

Definition of a Putative TSS for *Mitf*-RD in *D. eugracilis*

After classifying the promoter of *Mitf*-RD as broad, the putative TSS was determined to occur at 11796. This annotation was made because the TSS of *Mitf*-RD is known to occur 6 bp downstream of the TSS of *Mitf*-RA. This TSS occurs in the expected region of contig12 (Fig. 102).

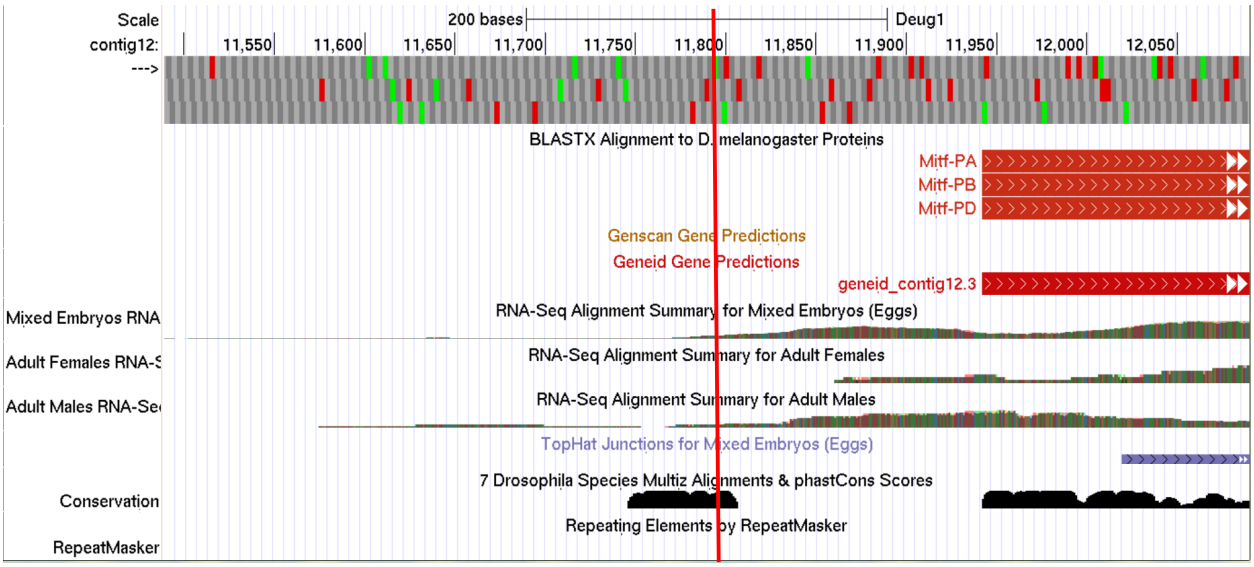


Figure 102: Position of the putative TSS in contig12. The putative TSS of 11796 determined by the BLASTn alignment is indicated by the red line. This putative TSS is supported by conservation data and RNA-Seq data, which appears to suggest the initial untranslated exon begins downstream of 11796.

After determining the location of the putative TSS at 11796, a search region consisting of the 300 bp upstream and 300 bp downstream was defined to search for *Drosophila* core promoter motifs from position 11496-12096 (Fig. 103). A similar search region (1198558-1199158) was used to search for core promoter motifs around the annotated TSS of *Mitf*-RD in *D. melanogaster* using the short match function in the *D. melanogaster* Genome Browser. The results of both of these core promoter motif searches were organized into a table (Table 12). A DRE motif was found in the search region in both *D. eugracilis* and *D. melanogaster.* As DRE motifs are not associated with a specific position, these motifs were highlighted in Table 12. With the search for *Drosophila* core promoter motifs complete, 11796 was established as the putative TSS for *Mitf*-RD and *Mitf*-RB in *D. eugracilis*.

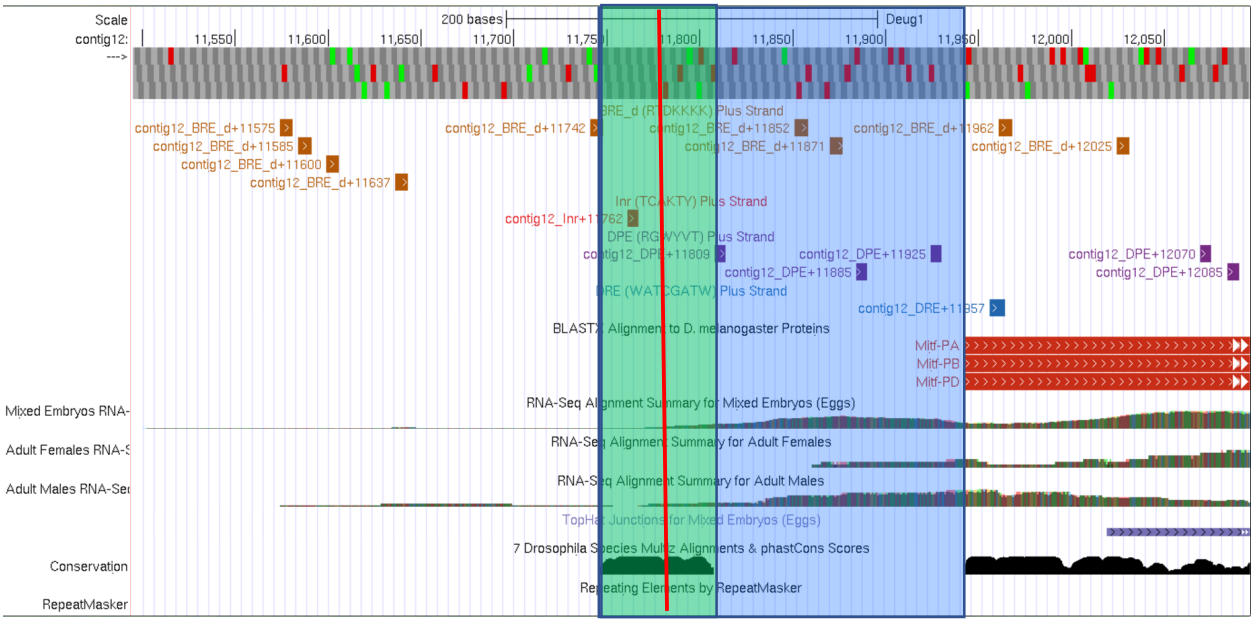


Figure 103: Search for core promoter motifs in *D. eugracilis*. None of the motifs associated with a specific position correspond to the putative TSS at 11796 (red line). There is a DRE motif in this region (shown in the blue evidence track). As DRE motifs do not correspond to a specific position, this motif was highlighted in Table 12, indicating that it is associated with the putative TSS. In addition, narrow and broad TSS search regions were defined using RNA-Seq and conservation data. The narrow search region (11747-11807) encompasses the highly conserved region associated with the putative TSS (green box). The broad search region (11747-11943) extends from the start of the highly conserved region to the start codon of the first coding exon (blue box).

Core promoter motif	*D. eugracilis*	*D. melanogaster*
BRE^{u}	NA	NA
TATA Box	NA	1198975, 1198993
BRE^{d}	11575, 11585, 11600, 11637, 11742, 11852, 118871, 11962, 12025	1198707, 1198804, 1198960, 1198969, 1199034,
Inr	11762	NA
MTE	NA	NA
DPE	11809, 11885, 11925, 12070, 12085	1198582, 1198670, 1198775, 1199127
Ohler_motif1	NA	NA
DRE	11957	1199029
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	NA	NA
Ohler_motif8	NA	NA

Table 12: Core promoter motifs found in the TSS search region of *Mitf*-RA in *D. eugracilis* and *D. melanogaster*. There is a DRE motif in both species that is associated with the TSS.

Classification of the TSS of *Mitf*-RC in *D. melanogaster*

The evidence tracks corresponding to *Mitf*-RC in the *D. melanogaster* Genome Browser were used to classify this TSS as peaked. There is one Celniker TSSs and no DHS positions associated with the TSS of this isoform (Fig. 104).

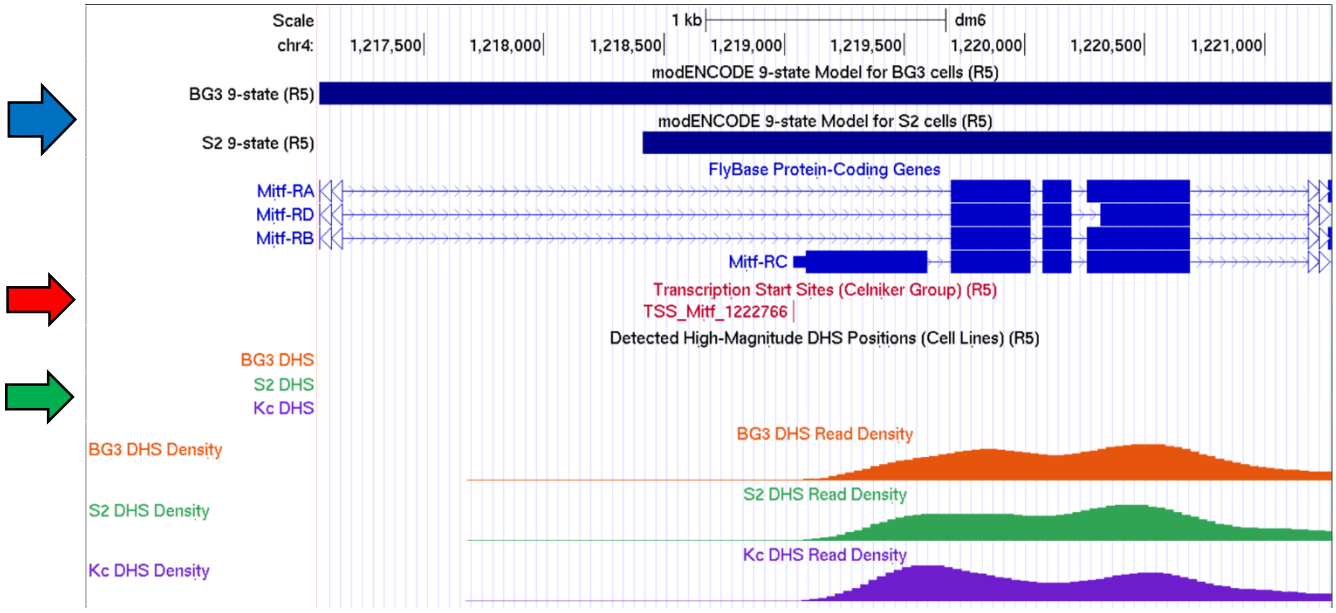


Figure 104: Classification of the *Mitf*-RA promoter as peaked. Both 9-state chromatin models are blue, indicating heterochromatin (blue arrow). However, there is an annotated Celniker TSS (red arrow). There are no DHS position associated with this TSS (green arrow). There is no available data upstream of this TSS due to the gap.

Definition of a Putative TSS for *Mitf*-RC in *D. eugracilis*

After classifying the promoter of *Mitf*-RC as peaked, a BLASTn search was performed using the initial untranslated exon of *Mitf*-RC as the query and the sequence of contig12 as the subject. The resulting alignment is shown in Figure 105. This alignment generally occurs in the expected region of contig12 (Fig. 106); however, there is some RNA-Seq data that occurs upstream of the putative TSS. To address this, a TSS search region will be defined along with a putative TSS at 13158 that includes these RNA-Seq reads. The coordinates of the search region are 12844-13210.

Deug1_dna range=contig12:1-38500 5'pad=0 3'pad=0 strand=+ repeatMasking=none
Sequence ID: Query_182047 **Length:** 38500 **Number of Matches:** 6

Range 1: 13159 to 13679 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
231 bits(160)	5e-63	369/556(66%)	36/556(66%)	Plus/Plus

```
Query  2      CAAGATCAAGGTCTTTAAAATACTCTAAATAGTTTACCGACTGTTAACAAGAATGTTTTC  61
              ||||  |  ||||| || | ||||||||||  ||||  |    |||| ||| ||||||| |
Sbjct  13159  CAAGGTTTAGGTCATTCATATACTCTAAACTGTTTGTCAGTCGTTA-CAAAAATGTTTAC  13217

Query  62     TTACTGGACCAAACAGAATATATCAAGCTCAGTAGATG-GAAAAAAGATGGACAAAGTTT  120
              |||||||| |||| | | |||| |||||  ||| |||  |||||||     |||||
Sbjct  13218  TTACTGGAGCAAAAAAAGTATACCAAGCAAAGTTGATATGAAAAAA-----ACAAAA---  13269

Query  121    CTACCATGAGAAAAAAAAATTCAACAAAAGTTAAAATGGAGAAAAAAGAAACAGTTGTAC  180
                || || |  |||   | |  | || |||   |||||| ||  || |  | |     |
Sbjct  13270  --ACTATAACCAAATGTATTAGACCAGAAG---AAATGGATAAGGAAAAGCCGGAACAGC  13324

Query  181    CGATTACAGAAATGCGTGCTCTGGAATTTCAAGATTTCCTTAAGCGGGTACAGACGTTAC  240
              ||| ||   | |  | ||  ||||| ||||||||||  | | |||||| || || ||||
Sbjct  13325  CGACTATTAATACCCATGTCCTGGAGTTTCAAGATTTTTTGAGGCGGGTTCAAACTTTAC  13384

Query  241    AAATGTCGCACAACGAAAGCCGCGATTTAGTCAATGGACCTAATATCTGCAGTTCCATGG  300
              ||||||| || |  ||       ||||  |  |||  |   ||| |   ||| |
Sbjct  13385  AAATGTCTCATATAGATGCAAACGAT--AACCAA--GTTTTAACAG--GCAATCGGTACA  13438

Query  301    ATATTGAAGTCGGCGAAGAACCTCCTACGACATCTTTTGCCGATAACATAGATAAAGCGT  360
              |||||||| ||||||| ||||    | ||  |         | |||||   |||  |
Sbjct  13439  ATATTGAAATCGGCGAGGAAC----TGCGGGAA--------GCTAACA---ATATTGGCC  13483

Query  361    GTCCTTTACCAGTGCTGCGATCATCTGATTGTGATTGTGTGCCACTTTGTATGACAAATA  420
                | | |||| || |||||  || | |||||||| ||||  |   |    | |||
Sbjct  13484  AAAAATCATCAGTACTACGATCCACTAAATGTGATTGTATGCCTATGCCTGCAGCCAATG  13543

Query  421    TTCGTGTATCAGTTGGTATTGACAAAGACTTAGAAATGATTCTTGAAATGGATCCGAGTA  480
              | |||||| | |||||||| || ||||| ||||||||||||||||||||||||| ||||
Sbjct  13544  TACGTGTAACTGTTGGTATCGATAAAGATTTAGAAATGATTCTTGAAATGGATCCCAGTA  13603

Query  481    TTGTAGATTTGGGTGATATTAGCATTGATGAAACTGCAGAACCGCGCATAGTGGGCTTAC  540
              |||||||||| |||||  || | ||||  ||||  ||||  |||||| ||||||| ||||
Sbjct  13604  TTGTAGATTTAGGTGACGTTGGTATTGCCGAAATGGCAGGGCCGCGCCTAGTGGGTTTAC  13663

Query  541    CTCCACTTTCGGGTGG  556
              |||||||||||| ||
Sbjct  13664  CTCCACTTTCGGGGGG  13679
```

Figure 105: BLASTn alignment of the initial untranslated exon of *Mitf*-RC in *D. melanogaster* (query) against contig12 (subject). This alignment has an e-score of 5e-63 (red box). This match aligns to the second base of *Mitf*-RC (blue box). One base was subtracted from the start of the subject sequence to determine the position of the putative TSS in *D. eugracilis* to be located at 13158.

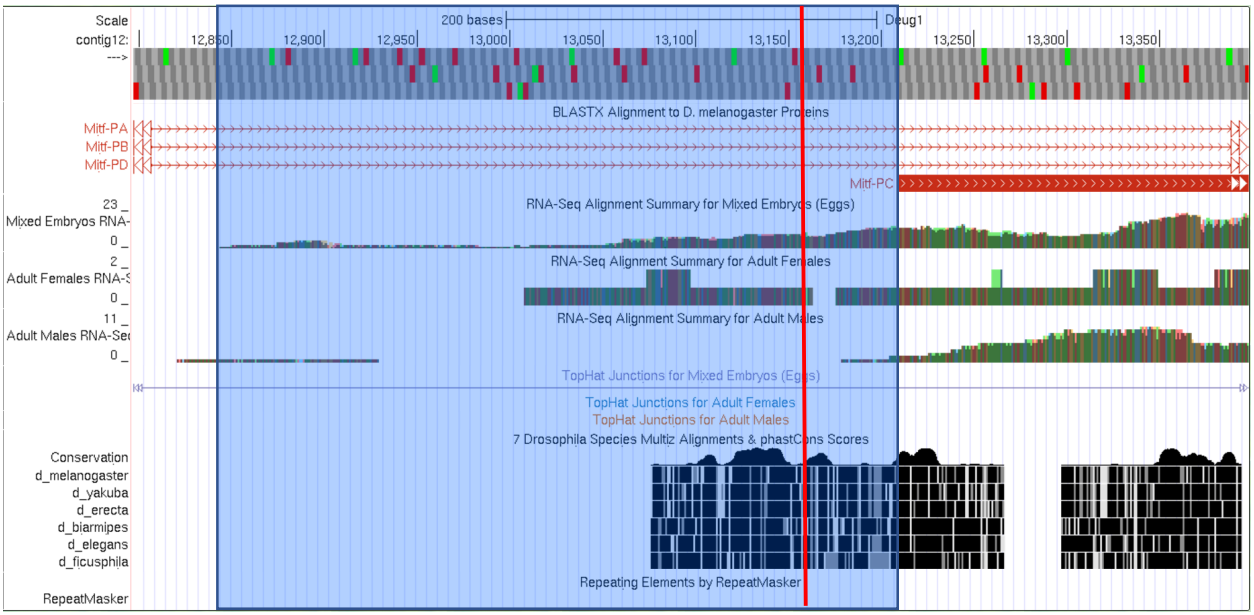
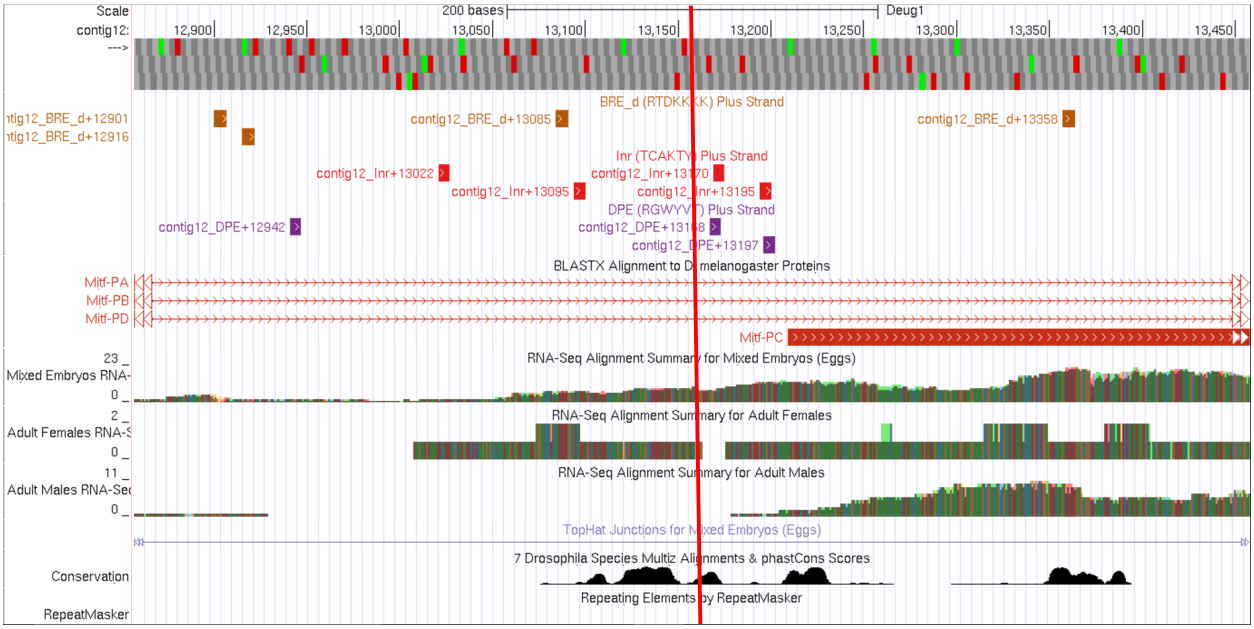


Figure 106: Position of the putative TSS in contig12. The putative TSS of 13158 determined by the BLASTn alignment is indicated by the red line. This putative TSS is supported by conservation data and RNA-Seq data, which appears to suggest the initial untranslated exon begins downstream of 13158. However, there are some RNA-Seq reads that extend beyond the putative TSS. These are encompassed in the broad search region that extends from the start of the embryonic RNA-Seq reads to the start codon of *Mitf*-RC (blue box).

After determining the location of the putative TSS at 13158, a search region consisting of the 300 bp upstream and 300 bp downstream was defined to search for *Drosophila* core promoter motifs from position 12858-13458 (Fig. 107). A similar search region (1218740-1219340) was used to search for core promoter motifs around the annotated TSS of *Mitf*-RC in *D. melanogaster* using the short match function in the *D. melanogaster* Genome Browser. The results of both of these core promoter motif searches were organized into a table (Table 13). None of the core promoter motifs in the search region were found to be associated with the TSSs. There is an Inr motif in *D. eugracilis* associated with a TSS 14 bp downstream of the putative TSS. While this was not a significant enough finding to merit highlighting this motif in Table 13, a note was made of this Inr motif. After finishing the search for core promoter motifs, 13158 was

established as the putative TSS and 12844-13210 was defined as the putative TSS search region (Fig. 108).



Contig 107: Search for core promoter motifs in *D. eugracilis*. None of the motifs associated with a specific position correspond to the putative TSS at 13158 (red line). There is a Inr motif that occurs at position 13170 and suggests a TSS located at 13172. A note was made of this motif, but it is not close enough to the putative TSS at 13158 to be highlighted on Table 13.

Core promoter motif	*D. eugracilis*	*D. melanogaster*
BRE^{u}	NA	NA
TATA Box	NA	1218808, 1218883, 1218885
BRE^{d}	12901, 12916, 13085, 13358	1218753, 1218831, 1218985, 1219069, 1219093
Inr	13022, 13095, 13170, 13195	NA
MTE	NA	NA
DPE	12942, 13168, 13197	1218799, 1218965, 1219032, 1219212, 1219271, 1219315
Ohler_motif1	NA	NA
DRE	NA	NA
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	NA	NA
Ohler_motif8	NA	NA

Table 11: Core promoter motifs found in the TSS search region of *Mitf*-RC in *D. eugracilis* and *D. melanogaster*. None of these motifs are associated with the putative TSS.

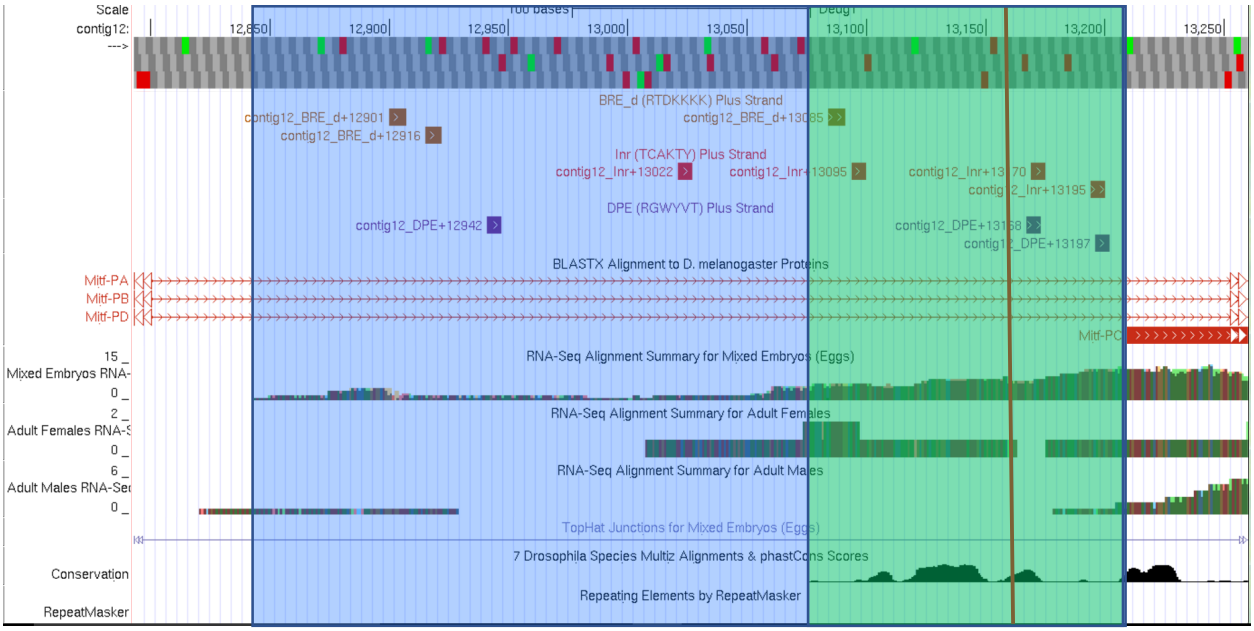


Figure 108: The putative TSS of *Mitf*-RC in *D. eugracilis* occurs at position 13158 (red box). The broad search region occurs at 12844-13210 (green box). The narrow search region (13077-13210) was identified with conservation data (green box).

Arf102F

Identification of the Ortholog

The next feature in contig12 is identified in Figure 109. Examination of the computer-based gene predictions corresponding to Feature 4 reveals several predictions that correlate reasonably well with this feature. The Augustus prediction, contig12.g4.t1, was selected for BLASTp analysis (Fig. 110).

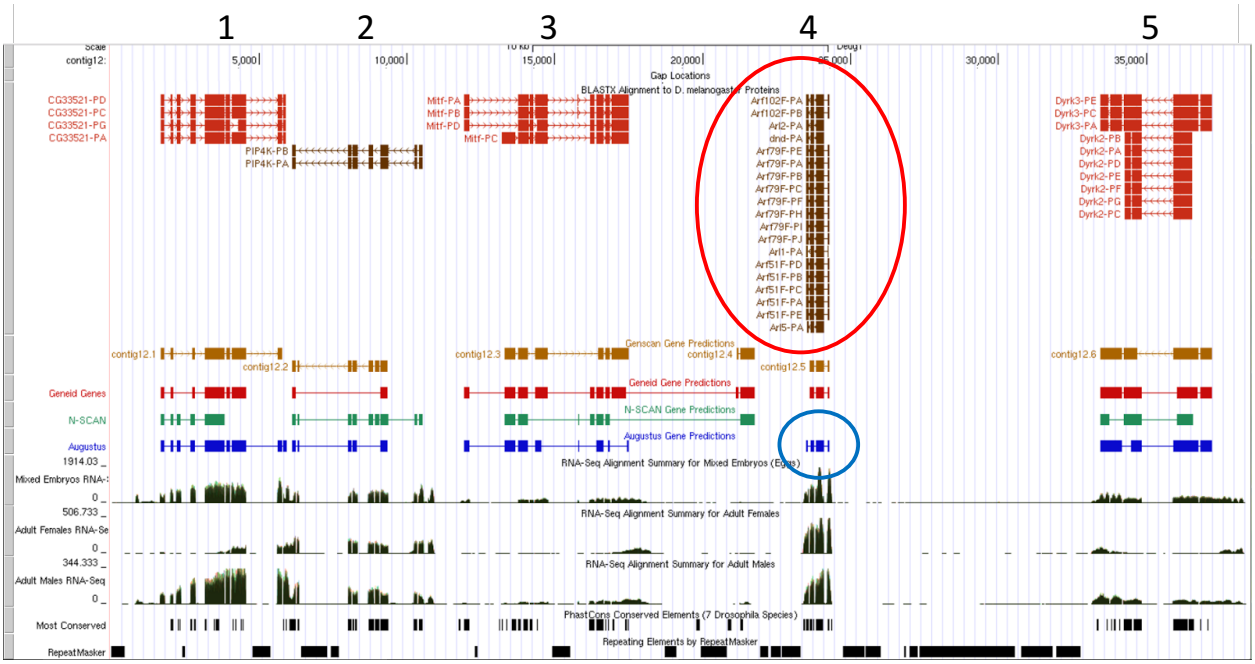


Figure 109: The position of the next feature in contig12. Shown in the red circle is feature 4, which is the next feature downstream from *Mitf*. The Augustus gene prediction selected for BLASTp analysis, contig12.g4.t1, is shown in the blue circle.

BLAST Hit Summary				
☑	Description	Species	Score	E value
☑	Arf102F-PB	Dmel	343.969	4.36794e-95
☑	Arf102F-PA	Dmel	343.969	4.36794e-95
☑	Arf79F-PJ	Dmel	289.271	1.08055e-78
☑	Arf79F-PI	Dmel	289.271	1.08055e-78
☑	Arf79F-PH	Dmel	289.271	1.08055e-78
☑	Arf79F-PF	Dmel	289.271	1.08055e-78
☑	Arf79F-PC	Dmel	289.271	1.08055e-78
☑	Arf79F-PE	Dmel	289.271	1.08055e-78
☑	Arf79F-PB	Dmel	289.271	1.08055e-78
☑	Arf79F-PA	Dmel	289.271	1.08055e-78
☑	Arf51F-PE	Dmel	237.654	3.57861e-63
☑	Arf51F-PA	Dmel	237.654	3.57861e-63
☑	Arf51F-PC	Dmel	237.654	3.57861e-63
☑	Arf51F-PB	Dmel	237.654	3.57861e-63
☑	Arf51F-PD	Dmel	237.654	3.57861e-63
☑	Arl1-PA	Dmel	194.897	2.8201e-50
☑	Arl5-PA	Dmel	172.17	1.81591e-43
☑	dnd-PA	Dmel	160.229	7.82724e-40
☑	Arl2-PA	Dmel	152.14	2.0616e-37
☑	Arl4-PA	Dmel	142.895	1.37095e-34
☑	Arl4-PB	Dmel	141.354	3.98886e-34
☑	Arfrp1-PA	Dmel	106.301	1.33019e-23
☑	Gie-PC	Dmel	103.605	9.21719e-23
☑	Gie-PB	Dmel	103.605	9.21719e-23
☑	Gie-PA	Dmel	103.605	9.21719e-23

Figure 110: Summary of the BLASTp search. The FlyBase BLASTp program provided this summary of BLASTp alignments obtained when searching the Augustus predicted protein (query) against the *Drosophila* Annotated Proteins Database (subject). There was no substantial drop-off in e-scores, but the strongest and most complete alignment was to *Arf102F*.

Along with the summary shown in Figure 110, the BLASTp search provided the following alignment of *Arf102F*-PB (Fig. 111). The alignment to *Arf102F*-PA was identical to Figure 111. Note that *Arf102F* is found on the 4th chromosome of *D. melanogaster*, providing further evidence that it is the orthologous gene.

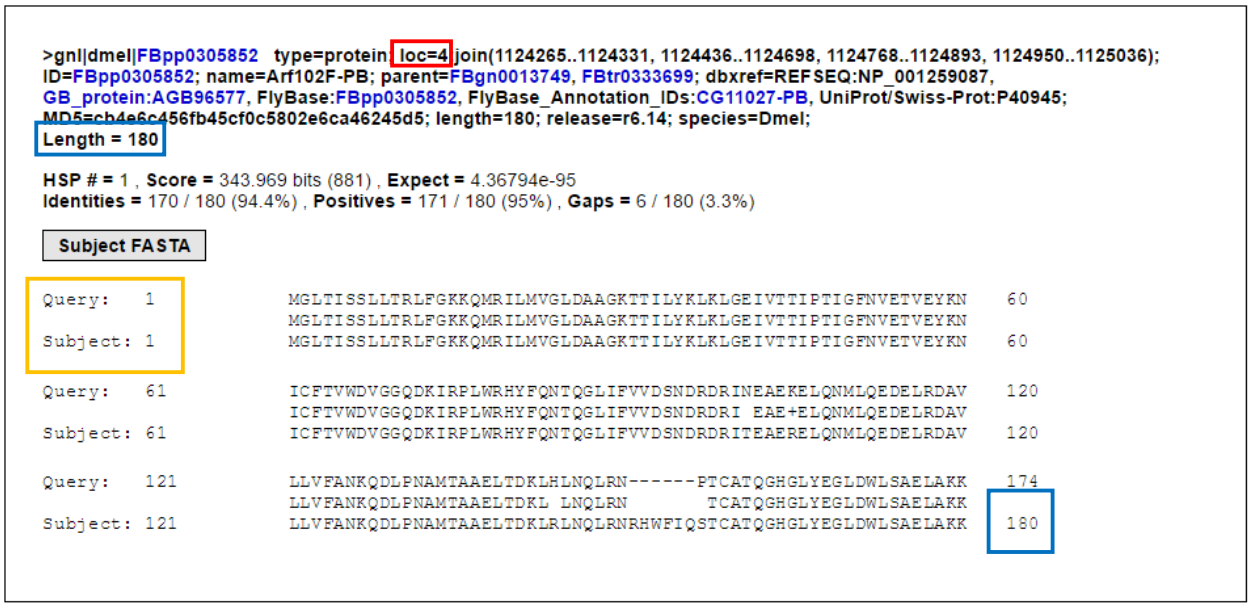


Figure 111: BLASTp alignment of the Augustus predicted protein against *D. melanogaster* *Arf102F*-PB. An alignment identical to this was produced by matching this query against the *Arf102F*-PA amino acid sequence. *Arf102F* is located on the 4th chromosome of *D. melanogaster* (red box). The entire *Arf102F*-PB isoform aligns to the fourth feature. This is confirmed by examination of the start of the alignment (yellow box), and by noting that the position at the end of the alignment matches the length of the *Arf102F*-PB subject sequence (blue boxes).

From the BLASTp evidence, *Arf102F* was determined to be the *D. melanogaster* ortholog of the 4th feature. FlyBase reports that the full name of this gene is *ADP ribosylation factor at 102F*, and that it is associated with NAD(P)+-protein-argenine ADP-ribosyltransferase activity as well as GTP binding. Examination of *Arf102F* in the Gene Record Finder confirmed that *Arf102F*-PB and *Arf102F*-PA have identical CDSs. Figure 112 shows both isoforms of *Arf102F* as they occur in *D. melanogaster*. Because *Arf102F*-PA and *Arf102F*-PB have identical

CDSs, the remainder of this section of the report which covers the CDS annotation of this gene will refer to the annotation of both isoforms as *Arf102F*.

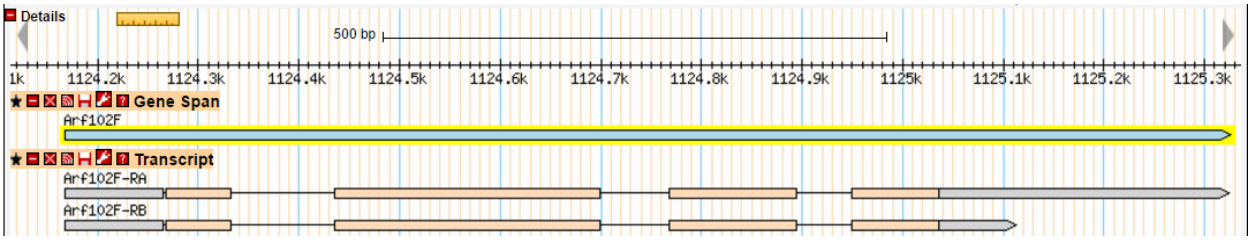


Figure 112: FlyBase GBrowse view of *Arf102F*. In *D. melanogaster*, *Arf102F*-PA and *Arf102F*-PB have identical CDSs. The two isoforms only differ in their 3' untranslated exons.

Exon-by-Exon Annotation of *Arf102F*

With *Arf102F* classified as the ortholog, an exon-by-exon annotation was conducted on the CDS of *Arf102F* in *D. eugracilis*. Standard CDS exon annotation protocol was followed to annotate these coding exons. The positions of these annotated exons were compiled in Table 14. As this project places particular emphasis on the annotation of the initial methionine of the coding spans of these genes, the BLASTx output used to place the initial coding exon is included in Figure 113. The methionine suggested by this alignment is the same methionine that was confirmed to correspond to the start of the first coding exon by viewing *Arf102F* in the *D. eugracilis* Genome Browser (Fig. 114).

Arf102F:1_1329_0

Sequence ID: Query_172991 **Length:** 22 **Number of Matches:** 1

Range 1: 1 to 22 Graphics Next Match Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
44.7 bits(104)	6e-11	22/22(100%)	22/22(100%)	0/22(0%)	-1

```
Query  24289  MGLTISSLLTRLFGKKQMRILM  24224
              MGLTISSLLTRLFGKKQMRILM
Sbjct  1      MGLTISSLLTRLFGKKQMRILM  22
```

Figure 113: BLASTx search of contig12 (query) against the 1st exon of *Arf102F* in *D. melanogaster* (subject). The methionine predicted to serve as the start codon of this isoform can be seen at the start of this alignment. This match has an e-score of 6e-11 (red box) and the aligned sequence from contig12 is in frame -1 (blue box).

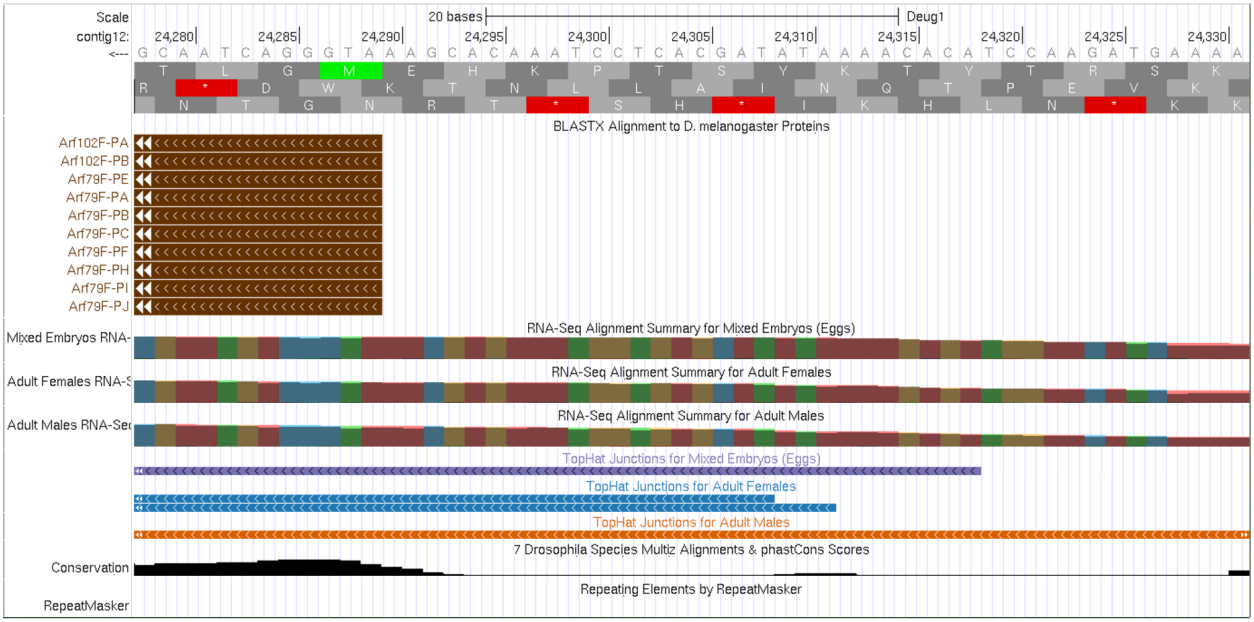


Figure 114: Annotation of the start codon of *Arf102F*. The methioinine predicted by the BLASTx alignment track is in the expected frame and matches the methionine predicted by the BLASTx output from Figure 113. The upstream RNA-Seq read density is indicative of the 5' untranslated exon of *Arf102F*.

Name of exon in *D. melanogaster*	Start/splice donor	Stop/splice acceptor	Frame	Acceptor Phase	Donor Phase	Exon BLASTp e-score
1_1329_0	24289	24223	-1	NA	1	6e-11
2_1329_2	24090	23828	-1	2	0	1e-56
3_1329_0	23760	23635	-2	0	0	6e-24
4_1329_0	23578	23495	-1	0	NA	1e-17

Table 14: Final coding exon annotations of *Arf102F*.

Checking the *Arf102F* Gene Model

The proposed model of *Arf102F* was tested by using the GMC. All exons of the putative gene model were found to pass the tests of the basic biological rules of CDS annotation. The resulting Dot Plot and Protein Alignment are shown in Figure 115 and Figure 116.

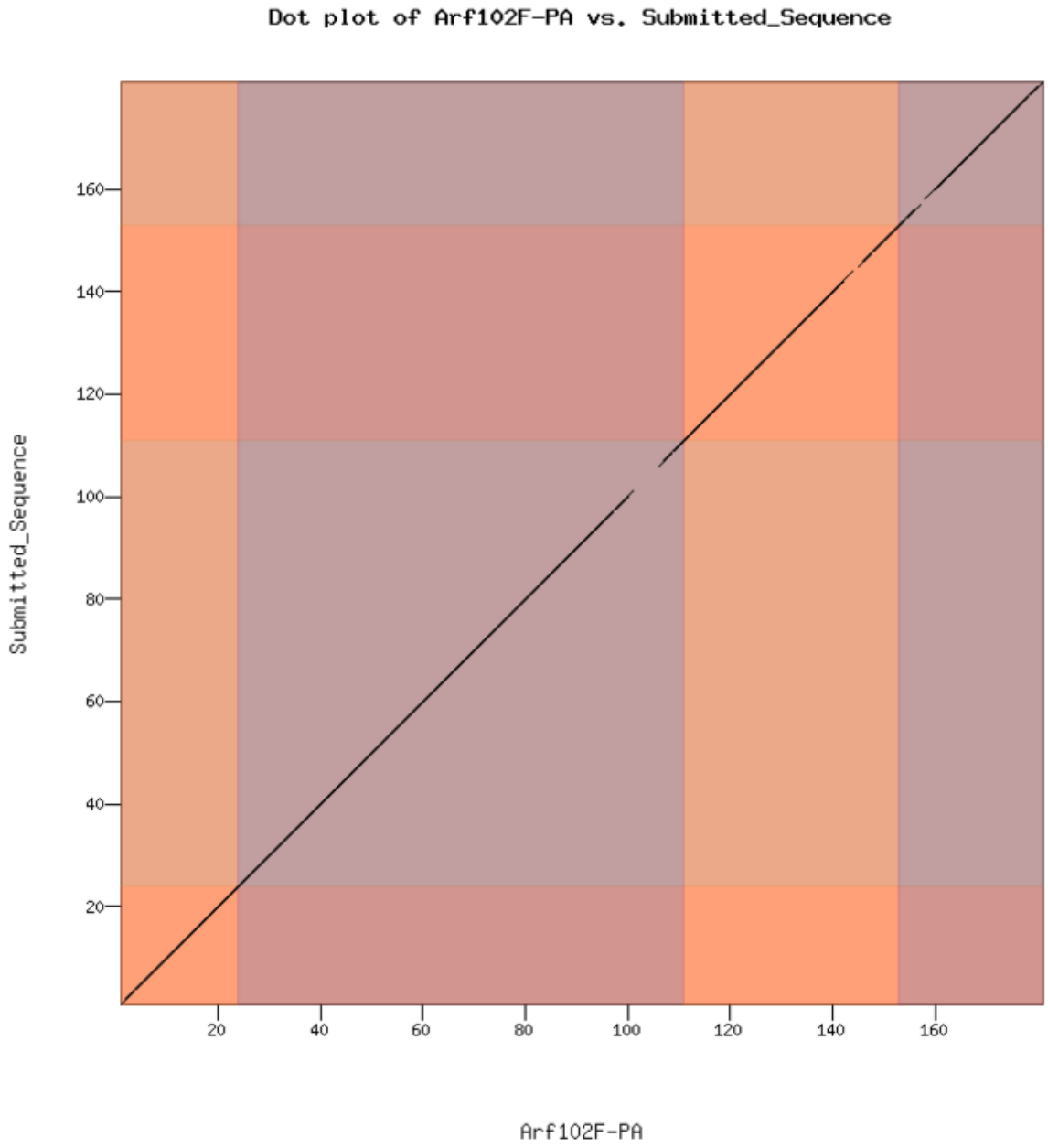


Figure 115: Dot Plot of *D. melanogaster Arf102F*-PA vs. the proposed model of *D. eugracilis Arf102F*.

Alignment of Arf102F-PA vs. Submitted_Seq

View plain text version

Identity: 176/180 (97.8%), **Similarity:** 178/180 (98.9%), **Gaps:** 0/180 (0.0%)

```
Arf102F-PA          1 MGLTISSLLTRLFGKKQMRILMVGLDAAGKTTILYKLKLGEIVTTIPTIGFNVETVEYKN 60
                      ************************************************************
Submitted_Seq       1 MGLTISSLLTRLFGKKQMRILMVGLDAAGKTTILYKLKLGEIVTTIPTIGFNVETVEYKN 60

Arf102F-PA         61 ICFTVWDVGGQDKIRPLWRHYFQNTQGLIFVVDSNDRDRITEAERELQNMLQEDELRDAV 120
                      ************************************.***:**********************
Submitted_Seq      61 ICFTVWDVGGQDKIRPLWRHYFQNTQGLIFVVDSNDRDRINEAEKELQNMLQEDELRDAV 120

Arf102F-PA        121 LLVFANKQDLPNAMTAAELTDKLRLNQLRNRHWFIQSTCATQGHGLYEGLDWLSAELAKK 180
                      ***********************:***********:************************
Submitted_Seq     121 LLVFANKQDLPNAMTAAELTDKLHLNQLRNRHWFIQATCATQGHGLYEGLDWLSAELAKK 180
```

Figure 116: Protein Alignment of *Arf102F*-PA in *D. melanogaster* vs. the proposed model of *D. eugracilis Arf102F*.

Dyrk3

Identification of the Ortholog

The final feature in contig12 is identified in Figure 117. Examination of the computer-based gene predictions corresponding to Feature 5 reveals several predictions that correlate reasonably well with this feature. The Genscan prediction, contig12.6, was selected for BLASTp analysis (Fig. 118).

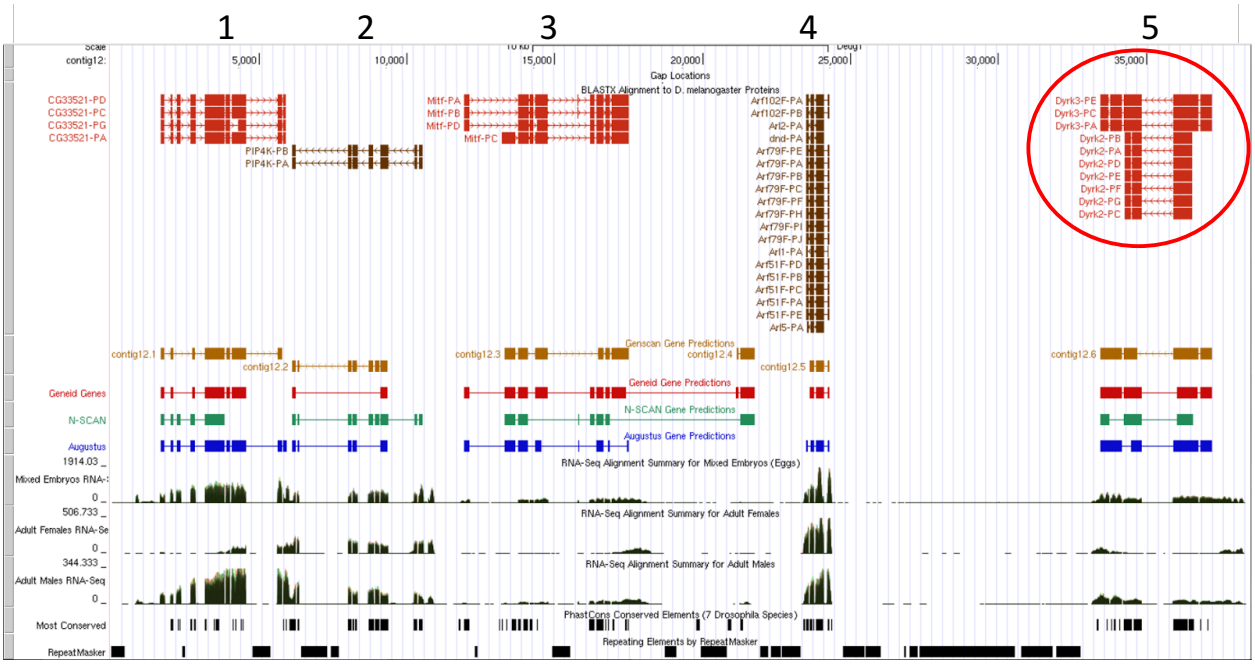
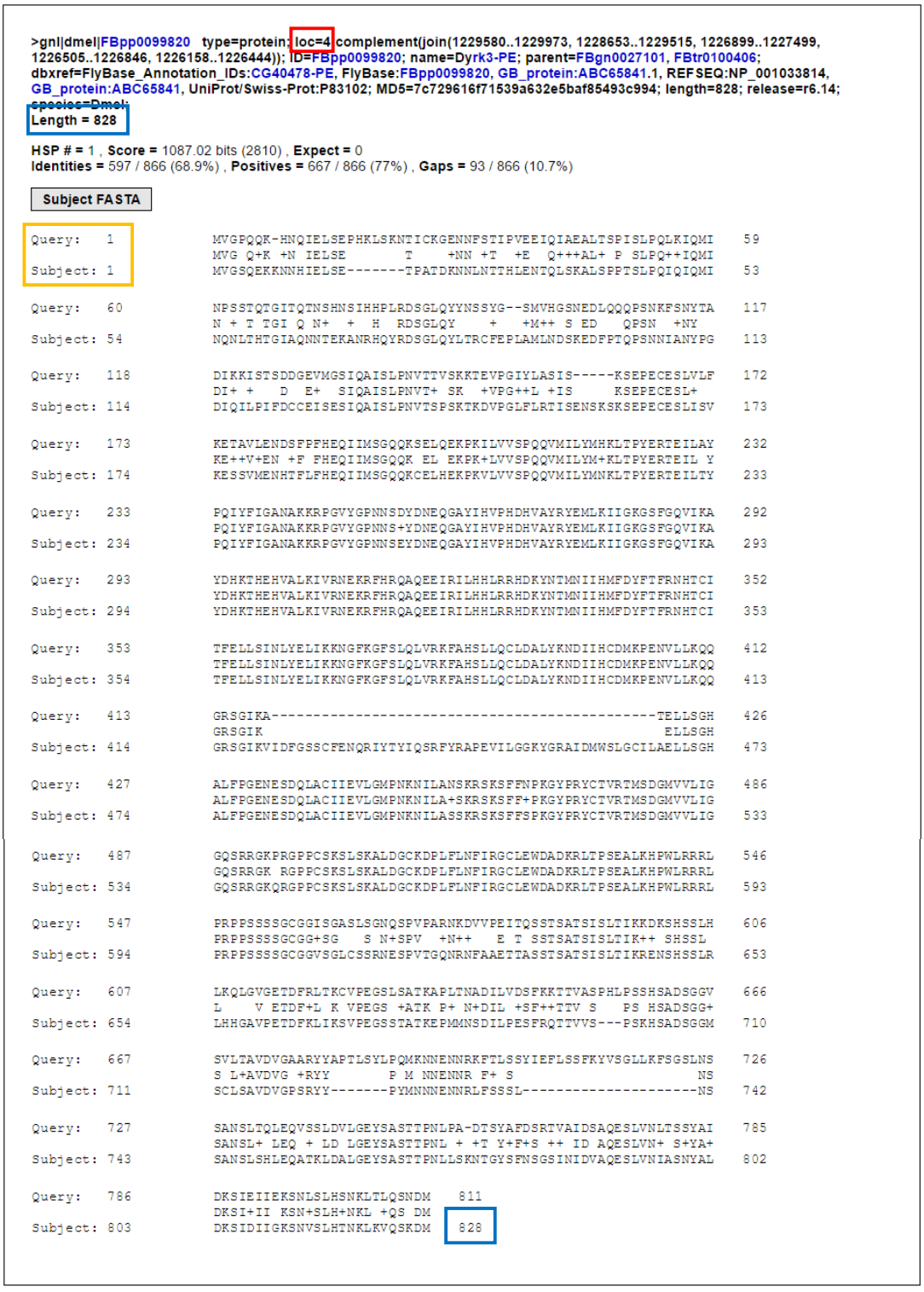


Figure 117: The position of the final feature in contig12. Shown in the red circle is feature 5, which is the next feature downstream from *Arf102F*.

BLAST Hit Summary				
	Description	Species	Score	E value
☑	Dyrk3-PE	Dmel	1087.02	0
☑	Dyrk3-PC	Dmel	1087.02	0
☑	Dyrk3-PA	Dmel	1087.02	0
☑	Dyrk2-PE	Dmel	312.768	7.55836e-85
☑	Dyrk2-PD	Dmel	312.768	7.55836e-85
☑	Dyrk2-PA	Dmel	312.768	7.55836e-85
☑	Dyrk2-PB	Dmel	312.768	7.55836e-85
☑	Dyrk2-PC	Dmel	312.768	7.55836e-85

Figure 118: Summary of BLASTp search. The FlyBase BLASTp program provided this summary of BLASTp alignments obtained when searching with the Genscan predicted protein (query) against the *Drosophila* Annotated Proteins Database (subject). The alignment to the isoforms of *Dyrk3* produced the lowest e-scores, providing evidence that this is the orthologous gene in *D. melanogaster*.

Along with the summary shown in Figure 118, the BLASTp search provided the following alignment of *Dyrk3*-PE (Fig. 119). The alignments to *Dyrk3*-PC and *Dyrk3*-PA were identical to Figure 119. Note that *Dyrk3* is found on the 4th chromosome of *D. melanogaster*, providing further evidence that it is the orthologous gene.



>gnl|dmel|FBpp0099820 type=protein; loc=4 complement(join(1229580..1229973, 1228653..1229515, 1226899..1227499, 1226505..1226846, 1226158..1226444)); ID=FBpp0099820; name=Dyrk3-PE; parent=FBgn0027101, FBtr0100406; dbxref=FlyBase_Annotation_IDs:CG40478-PE, FlyBase:FBpp0099820, GB_protein:ABC65841.1, REFSEQ:NP_001033814, GB_protein:ABC65841, UniProt/Swiss-Prot:P83102; MD5=7c729616f71539a632e5baf85493c994; length=828; release=r6.14; species=Dmel;
Length = 828

HSP # = 1 , Score = 1087.02 bits (2810) , Expect = 0
Identities = 597 / 866 (68.9%) , Positives = 667 / 866 (77%) , Gaps = 93 / 866 (10.7%)

Subject FASTA

```
Query:   1          MVGPQQK-HNQIELSEPHKLSKNTICKGENNFSTIPVEEIQIAEALTSPISLPQLKIQMI   59
                   MVG Q+K +N IELSE       T    +NN +T  +E  Q+++AL+ P SLPQ++IQMI
Subject: 1          MVGSQEKKNNHIELSE-------TPATDKNNLNTTHLENTQLSKALSPPTSLPQIQIQMI   53

Query:   60         NPSSTQTGITQTNSHNSIHHPLRDSGLQYYNSSYG--SMVHGSNEDLQQQPSNKFSNYTA   117
                   N + T TGI Q N+  +  H  RDSGLQY    +   +M++ S ED   QPSN  +NY
Subject: 54         NQNLTHTGIAQNNTEKANRHQYRDSGLQYLTRCFEPLAMLNDSKEDFPTQPSNNIANYPG   113

Query:   118        DIKKISTSDDGEVMGSIQAISLPNVTTVSKKTEVPGIYLASIS-----KSEPECESLVLF   172
                   DI+ +   D  E+  SIQAISLPNVT+ SK  +VPG++L +IS     KSEPECESL+
Subject: 114        DIQILPIFDCCEISESIQAISLPNVTSPSKTKDVPGLFLRTISENSKSKSEPECESLISV   173

Query:   173        KETAVLENDSFPFHEQIIMSGQQKSELQEKPKILVVSPQQVMILYMHKLTPYERTEILAY   232
                   KE++V+EN +F FHEQIIMSGQQK EL EKPK+LVVSPQQVMILYM+KLTPYERTEIL Y
Subject: 174        KESSVMENHTFLFHEQIIMSGQQKCELHEKPKVLVVSPQQVMILYMNKLTPYERTEILTY   233

Query:   233        PQIYFIGANAKKRPGVYGPNNSDYDNEQGAYIHVPHDHVAYRYEMLKIIGKGSFGQVIKA   292
                   PQIYFIGANAKKRPGVYGPNNS+YDNEQGAYIHVPHDHVAYRYEMLKIIGKGSFGQVIKA
Subject: 234        PQIYFIGANAKKRPGVYGPNNSEYDNEQGAYIHVPHDHVAYRYEMLKIIGKGSFGQVIKA   293

Query:   293        YDHKTHEHVALKIVRNEKRFHRQAQEEIRILHHLRRHDKYNTMNIIHMFDYFTFRNHTCI   352
                   YDHKTHEHVALKIVRNEKRFHRQAQEEIRILHHLRRHDKYNTMNIIHMFDYFTFRNHTCI
Subject: 294        YDHKTHEHVALKIVRNEKRFHRQAQEEIRILHHLRRHDKYNTMNIIHMFDYFTFRNHTCI   353

Query:   353        TFELLSINLYELIKKNGFKGFSLQLVRKFAHSLLQCLDALYKNDIIHCDMKPENVLLKQQ   412
                   TFELLSINLYELIKKNGFKGFSLQLVRKFAHSLLQCLDALYKNDIIHCDMKPENVLLKQQ
Subject: 354        TFELLSINLYELIKKNGFKGFSLQLVRKFAHSLLQCLDALYKNDIIHCDMKPENVLLKQQ   413

Query:   413        GRSGIKA----------------------------------------------TELLSGH   426
                   GRSGIK                                               ELLSGH
Subject: 414        GRSGIKVIDFGSSCFENQRIYTYIQSRFYRAPEVILGGKYGRAIDMWSLGCILAELLSGH   473

Query:   427        ALFPGENESDQLACIIEVLGMPNKNILANSKRSKSFFNPKGYPRYCTVRTMSDGMVVLIG   486
                   ALFPGENESDQLACIIEVLGMPNKNILA+SKRSKSFF+PKGYPRYCTVRTMSDGMVVLIG
Subject: 474        ALFPGENESDQLACIIEVLGMPNKNILASSKRSKSFFSPKGYPRYCTVRTMSDGMVVLIG   533

Query:   487        GQSRRGKPRGPPCSKSLSKALDGCKDPLFLNFIRGCLEWDADKRLTPSEALKHPWLRRRL   546
                   GQSRRGK RGPPCSKSLSKALDGCKDPLFLNFIRGCLEWDADKRLTPSEALKHPWLRRRL
Subject: 534        GQSRRGKQRGPPCSKSLSKALDGCKDPLFLNFIRGCLEWDADKRLTPSEALKHPWLRRRL   593

Query:   547        PRPPSSSSGCGGISGASLSGNQSPVPARNKDVVPEITQSSTSATSISLTIKKDKSHSSLH   606
                   PRPPSSSSGCGG+SG   S N+SPV  +N++   E T SSTSATSISLTIK++ SHSSL
Subject: 594        PRPPSSSSGCGGVSGLCSSRNESPVTGQNRNFAAETTASSTSATSISLTIKRENSHSSLR   653

Query:   607        LKQLGVGETDFRLTKCVPEGSLSATKAPLTNADILVDSFKKTTVASPHLPSSHSADSGGV   666
                   L    V ETDF+L K VPEGS +ATK P+ N+DIL +SF++TTV S   PS HSADSGG+
Subject: 654        LHHGAVPETDFKLIKSVPEGSSTATKEPMMNSDILPESFRQTTVVS---PSKHSADSGGM   710

Query:   667        SVLTAVDVGAARYYAPTLSYLPQMKNNENNRKFTLSSYIEFLSSFKYVSGLLKFSGSLNS   726
                   S L+AVDVG +RYY       P M NNENNR F+ S                     NS
Subject: 711        SCLSAVDVGPSRYY-------PYMNNNENNRLFSSSL--------------------NS   742

Query:   727        SANSLTQLEQVSSLDVLGEYSASTTPNLPA-DTSYAFDSRTVAIDSAQESLVNLTSSYAI   785
                   SANSL+ LEQ + LD LGEYSASTTPNL + +T Y+F+S ++ ID AQESLVN+ S+YA+
Subject: 743        SANSLSHLEQATKLDALGEYSASTTPNLLSKNTGYSFNSGSINIDVAQESLVNIASNYAL   802

Query:   786        DKSIEIIEKSNLSLHSNKLTLQSNDM   811
                   DKSI+II KSN+SLH+NKL +QS DM
Subject: 803        DKSIDIIGKSNVSLHTNKLKVQSKDM   828
```

Figure 119: BLASTp alignment of the Genscan predicted protein against *D. melanogaster Dyrk3*-PE. An alignment identical to this was produced by matching this query against the *Dyrk3*-PC and *Dyrk3*-PA amino acid sequences. *Dyrk3* is located on the 4th chromosome of *D. melanogaster* (red box). The entire *Dyrk3*-PE isoform aligns to the fourth feature. This is confirmed by examination of the start of the alignment (yellow box), and by noting that the position at the end of the alignment matches the length of the *Dyrk3*-PE subject sequence (blue boxes).

From the BLASTp evidence, *Dyrk3* was determined to be the *D. melanogaster* ortholog of the 5th feature. FlyBase reports that the full name of this gene is *Dual-specificity tyrosine phosphorylation-regulated kinase 3*, and that it is associated with ATP binding and protein kinase activity. Examination of *Dyrk3* in the Gene Record Finder confirmed that *Dyrk3*-PE, *Dyrk3*-PA and *Dyrk3*-PC have identical CDSs. Figure 120 shows all three isoforms of *Dyrk3* as they occur in *D. melanogaster*. Because *Dyrk3*-PE, *Dyrk3*-PA, and *Dyrk3*-PC, have identical CDSs, the remainder of this section of the report will refer to the annotation of all three isoforms as *Dyrk3*.

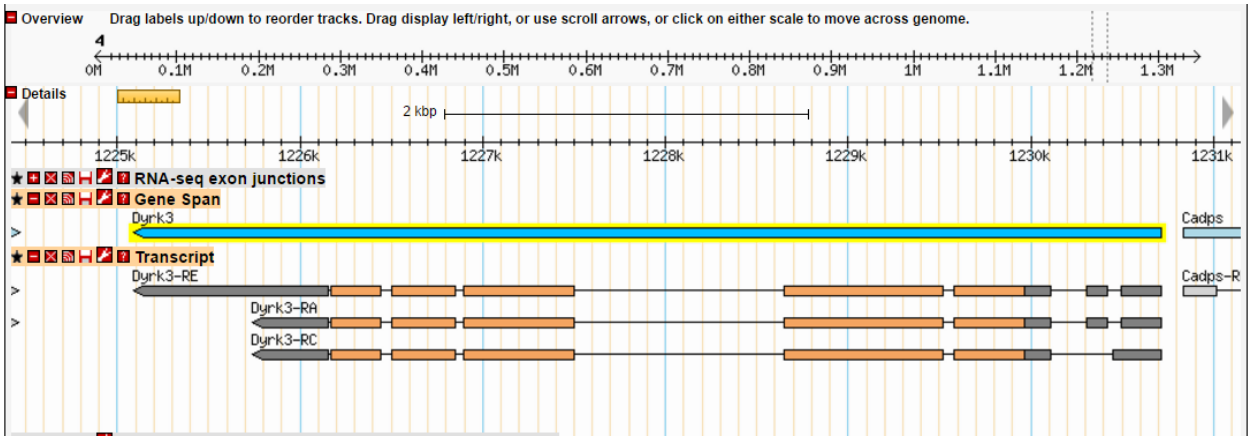


Figure 120: FlyBase GBrowse view of *Dyrk3*. In *D. melanogaster*, *Dyrk3*-PE, *Dyrk3*-PA and *Dyrk3*-PC have identical CDSs. The three isoforms only differ in their untranslated exons.

Exon-by-Exon Annotation of *Dyrk3*

With *Dyrk3* classified as the ortholog, an exon-by-exon annotation was conducted on the CDS of *Dyrk3* in *D. eugracilis.* Standard CDS exon annotation protocol was followed to annotate these coding exons. As this project places particular emphasis on the annotation of the initial methionine of the coding spans of these genes, the BLASTx output used to place the initial coding exon is included in Figure 121. The methionine suggested by this alignment is the same

methionine that was confirmed to correspond to the start of the first coding exon by viewing *Dyrk3* in the *D. eugracilis* Genome Browser (Fig. 122).

Dyrk3:1_2365_0

Sequence ID: Query_106919 **Length:** 131 **Number of Matches:** 1

Range 1: 1 to 131 Graphics Next Match Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
101 bits(251)	8e-29	60/138(43%)	82/138(59%)	10/138(7%)	-2

```
Query  37221  MVGPQQK-HNQIELSEPHKLSKNTICKGENNFSTIPVEEIQIAEALTSPISLPQLKIQMI  37045
              MVG Q+K +N IELSE       T    +NN +T  +E  Q+++AL+ P SLPQ++IQMI
Sbjct  1      MVGSQEKKNNHIELSE-------TPATDKNNLNTTHLENTQLSKALSPPTSLPQIQIQMI  53

Query  37044  NPSSTQTGITQTNSHNSIHHPLRDSGLQYYNSSYG--SMVHGSNEDLQQQPSNKFSNYTA  36871
              N + T TGI Q N+  +  H  RDSGLQY    +   +M++ S ED   QPSN  +NY
Sbjct  54     NQNLTHTGIAQNNTEKANRHQYRDSGLQYLTRCFEPLAMLNDSKEDFPTQPSNNIANYPG  113

Query  36870  DIKKISTSDDGEVMGSIQ  36817
              DI+ +   D  E+  SIQ
Sbjct  114    DIQILPIFDCCEISESIQ  131
```

Figure 121: BLASTx search of contig12 (query) against the 1st exon of *Dyrk3* in *D. melanogaster* (subject). The methionine predicted to serve as the start codon of this isoform can be seen at the start of this alignment. This match has an e-score of 8e-29 (red box) and the aligned sequence from contig12 is in frame -2 (blue box).

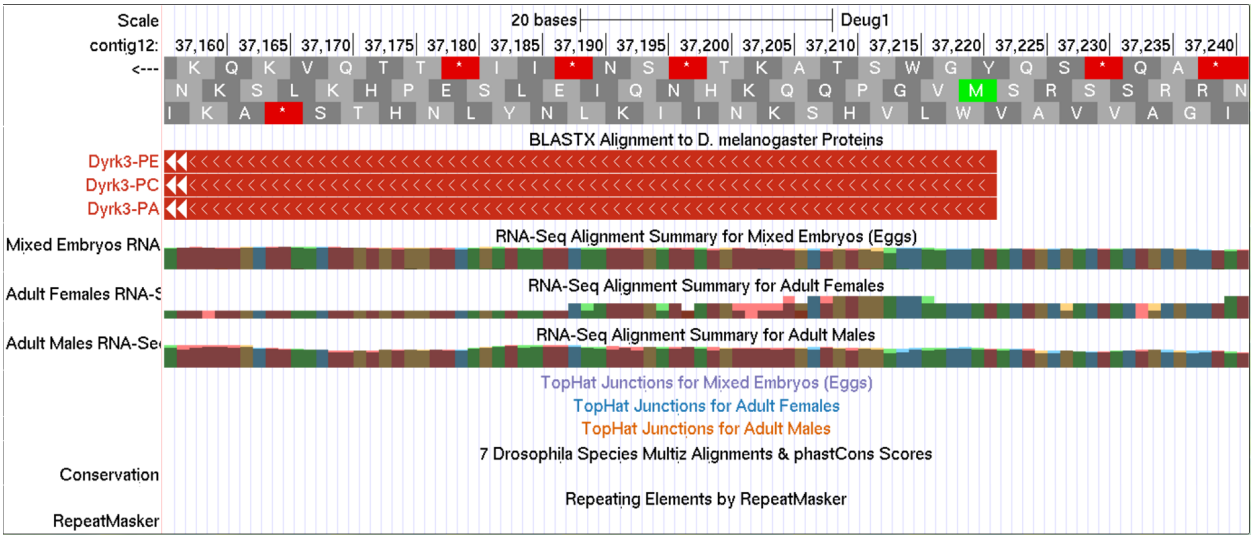


Figure 122: Annotation of the start codon of *Dyrk3*. The methioinine predicted by the BLASTx alignment track is in the expected frame and matches the methionine predicted by the BLASTx output from Figure 121. The upstream RNA-Seq read density is indicative of the 5' untranslated exon of *Arf102F*. In-frame stop codons in frames -1 and -3, as well as the presence of only a single candidate methionine in the expected region, serve as more evidence supporting this annotation.

The BLASTx alignment produced for the 2nd exon of *Dyrk3* was in frame -1 (Fig. 123). When annotating the splice donor site of the 2nd exon of *Dyrk3*, a suitable GT splice donor site in the correct phase could not be found. Analysis of the RNA-Seq data and TopHat junctions was used to confirm that this exon has a GC splice donor site (Fig. 124). For the rest of the exons in *D. eugracilis Dyrk3*, no unusual cases were observed and standard coding exon annotation procedures were followed. All the annotated exons of *Dyrk3* in *D. eugracilis* were compiled in Table 15.

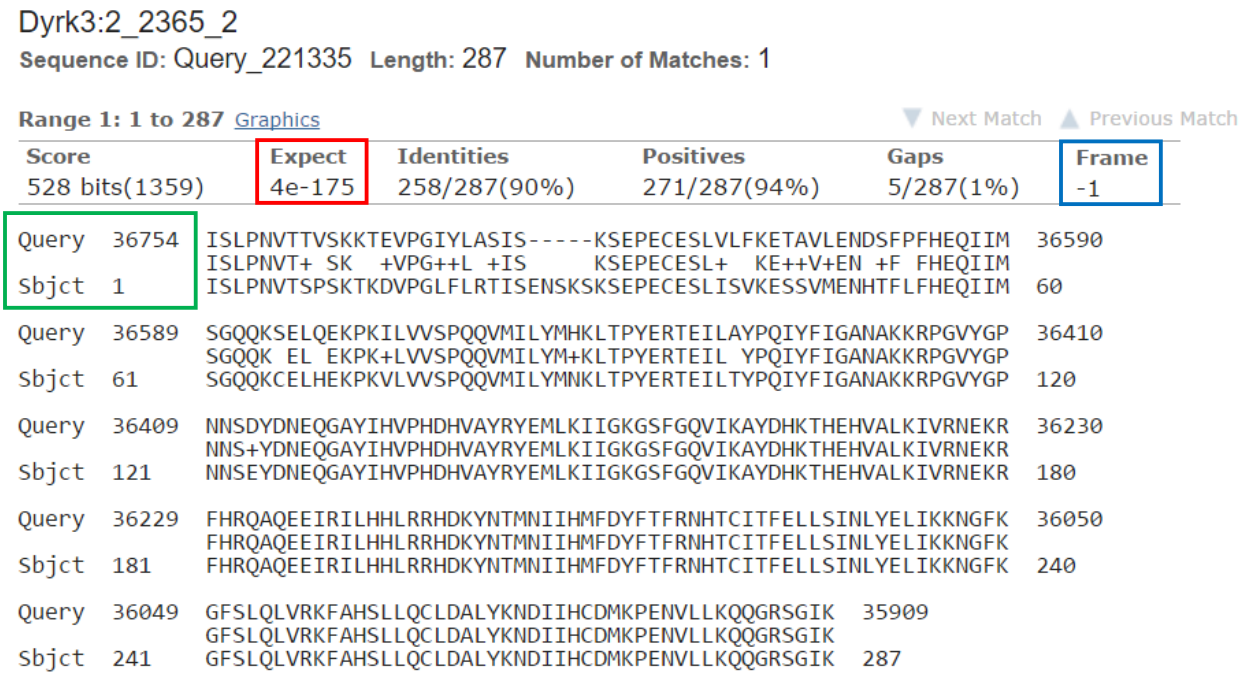


Figure 123: BLASTx of contig12 (query) against the 2nd exon of *Dyrk3* (subject). This alignment predicts the start of the 2nd exon in *Dyrk3* to occur at position 36754 in *D. eugracilis* (green box). This is an alignment made with a lot of confidence, as it has an e-score of 4e-175 (red box). It occurs in frame -1 (blue box).

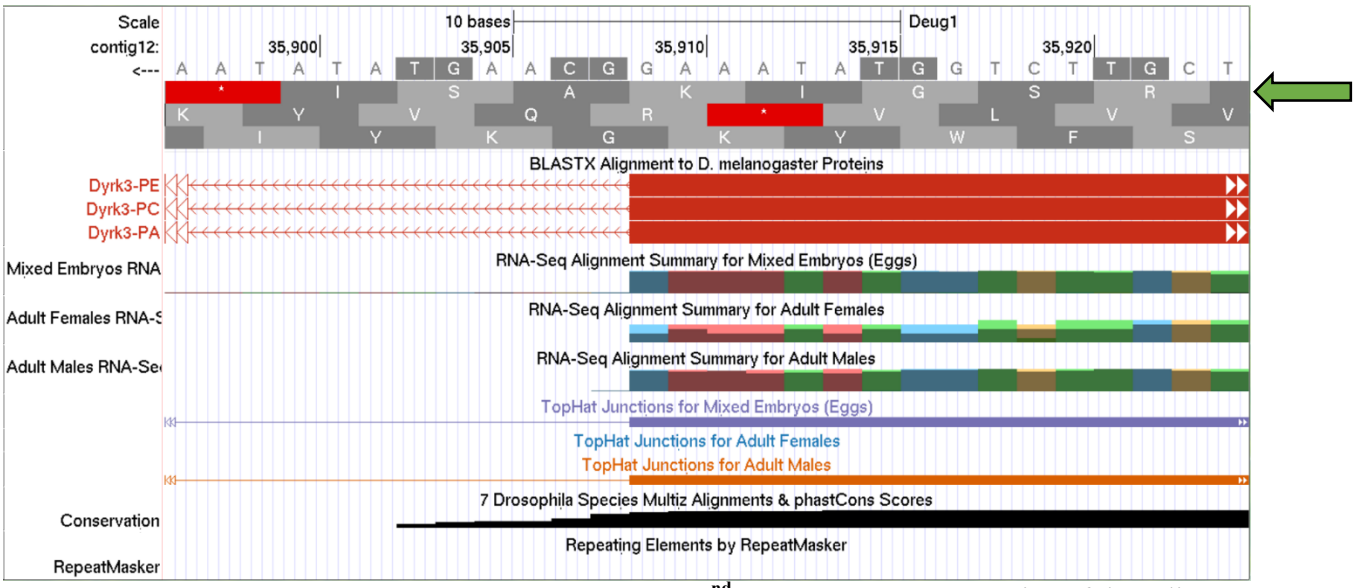


Figure 124: Annotation of a GC splice donor site in the 2nd exon of *Dyrk3*. Annotation of the splice acceptor site of the third exon determined that the splice site is in phase 0. Furthermore, the BLASTx alignment from Fig. 123 of the 2nd exon is in frame -1. Using this information, it was determined that the splice donor site of the 2nd exon of *Dyrk3* must be in frame -1 (green arrow) and be phase 0. The three candidate GTs in the region around the expected splice site are not in the correct phase. However, there is a GC that corresponds perfectly to the BLASTx alignment from Fig. 123 and is supported by RNA-Seq data and TopHat junctions. This is sufficient evidence to support the annotation of the GC splice donor site.

Name of exon in *D. melanogaster*	Start/splice donor	Stop/splice acceptor	Frame	Acceptor Phase	Donor Phase	Exon BLASTp e-score
1_2365_0	37221	36816	-2	NA	1	8e-29
2_2365_2	36756	35909	-1	2	0	4e-175
3_2365_0	34837	34237	-1	0	1	3e-130
4_2365_2	34174	33803	-3	2	1	5e-41
5_2365_2	33745	33459	-3	2	NA	1e-36

Table 15: Final coding exon annotations of *Dyrk3*.

Checking the *Dyrk3* Gene Model

The proposed model of *Dyrk3* was tested by using the GMC. All exons of the putative gene model were found to pass the tests of the basic biological rules of CDS annotation. The resulting Dot Plot and Protein Alignment are shown in Figure 125 and Figure 126.

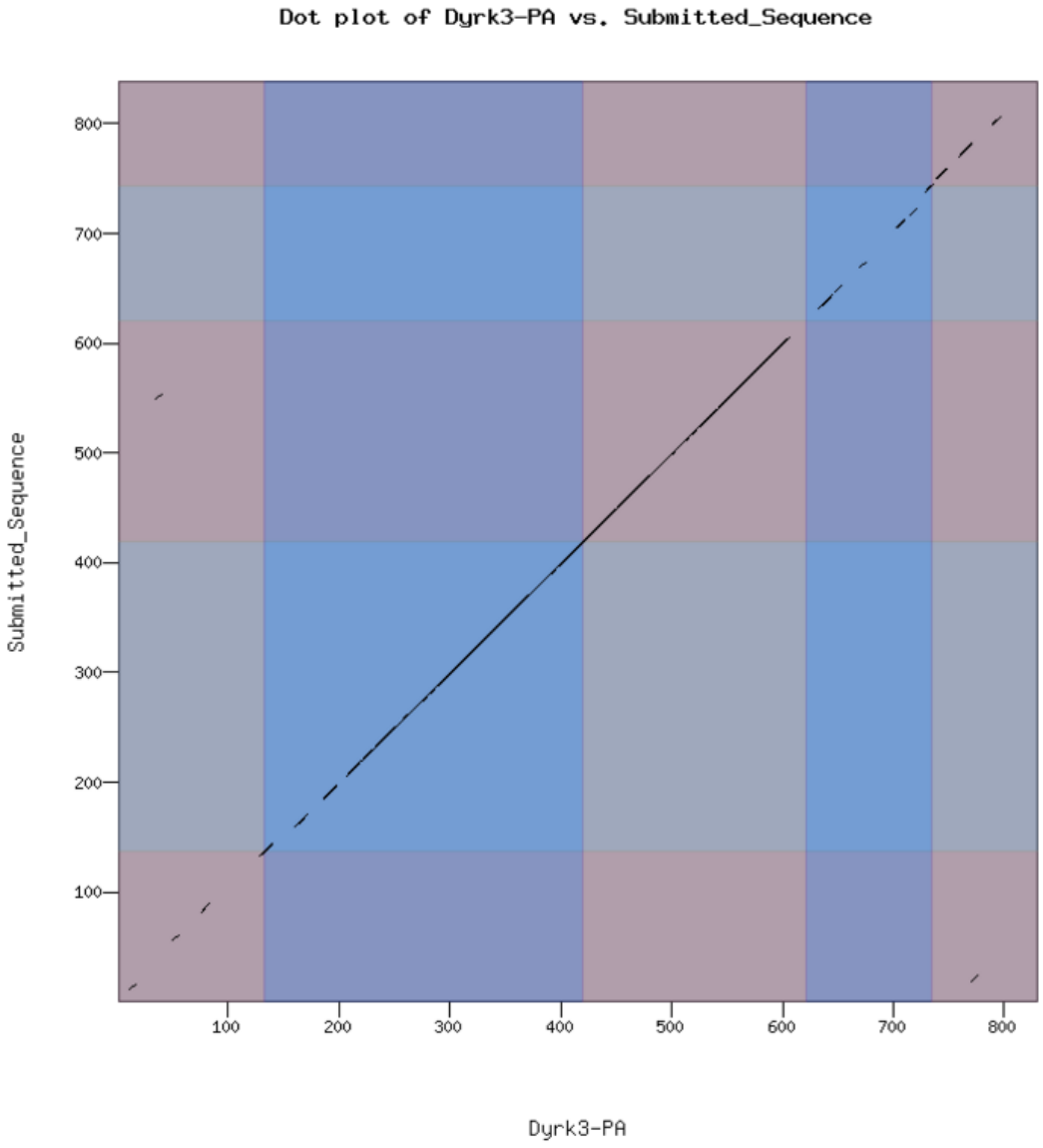


Figure 125: Dot Plot of *D. melanogaster Dyrk3*-PA vs. the proposed model of *D. eugracilis Dyrk3*.

Alignment of Dyrk3-PA vs. Submitted_Seq

View plain text version

Identity: 647/847 (76.4%), **Similarity:** 716/847 (84.5%), **Gaps:** 28/847 (3.3%)

```
Dyrk3-PA          1 MVGSQEKKNNHIELSE-------TPATDKNNLNTTHLENTQLSKALSPPTSLPQIQIQMI 53
                    ***.*:* :*:*****       * ...:**:.*  :*: *:::**:.* ****::****
Submitted_Seq     1 MVGPQQK-HNQIELSEPHKLSKNTICKGENNFSTIPVEEIQIAEALTSPISLPQLKIQMI 59

Dyrk3-PA         54 NQNLTHTGIAQNNTEKANRHQYRDSGLQYLTRCFEPLAMLNDSKEDFPTQPSNNIANYPG 113
                    * . *:***:*.*:.:: :*  ******* . .:   :*::.*:**:  ****:::**..
Submitted_Seq    60 NPSSTQTGITQTNSHNSIHHPLRDSGLQYYNSSYG--SMVHGSNEDLQQQPSNKFSNYTA 117

Dyrk3-PA        114 DIQILPIFDCCEISESIQAISLPNVTSPSKTKDVPGLFLRTISENSKSKSEPECESLISV 173
                    **: :.  *  *:  **********: **..:***::* :**     *********: .
Submitted_Seq   118 DIKKISTSDDGEVMGSIQAISLPNVTTVSKKTEVPGIYLASIS-----KSEPECESLVLF 172

Dyrk3-PA        174 KESSVMENHTFLFHEQIIMSGQQKCELHEKPKVLVVSPQQVMILYMNKLTPYERTEILTY 233
                    **::*:**.:* ***********.**:****:************:**********:*
Submitted_Seq   173 KETAVLENDSFPFHEQIIMSGQQKSELQEKPKILVVSPQQVMILYMHKLTPYERTEILAY 232

Dyrk3-PA        234 PQIYFIGANAKKRPGVYGPNNSEYDNEQGAYIHVPHDHVAYRYEMLKIIGKGSFGQVIKA 293
                    ********************.**************************************
Submitted_Seq   233 PQIYFIGANAKKRPGVYGPNNSDYDNEQGAYIHVPHDHVAYRYEMLKIIGKGSFGQVIKA 292

Dyrk3-PA        294 YDHKTHEHVALKIVRNEKRFHRQAQEEIRILHHLRRHDKYNTMNIIHMFDYFTFRNHTCI 353
                    ************************************************************
Submitted_Seq   293 YDHKTHEHVALKIVRNEKRFHRQAQEEIRILHHLRRHDKYNTMNIIHMFDYFTFRNHTCI 352

Dyrk3-PA        354 TFELLSINLYELIKKNGFKGFSLQLVRKFAHSLLQCLDALYKNDIIHCDMKPENVLLKQQ 413
                    ************************************************************
Submitted_Seq   353 TFELLSINLYELIKKNGFKGFSLQLVRKFAHSLLQCLDALYKNDIIHCDMKPENVLLKQQ 412

Dyrk3-PA        414 GRSGIKVIDFGSSCFENQRIYTYIQSRFYRAPEVILGGKYGRAIDMWSLGCILAELLSGH 473
                    *********************************.**************************
Submitted_Seq   413 GRSGIKVIDFGSSCFENQRIYTYIQSRFYRAPEVILGAKYGRAIDMWSLGCILAELLSGH 472

Dyrk3-PA        474 ALFPGENESDQLACIIEVLGMPNKNILASSKRSKSFFSPKGYPRYCTVRTMSDGMVVLIG 533
                    ****************************.********.**********************
Submitted_Seq   473 ALFPGENESDQLACIIEVLGMPNKNILANSKRSKSFFNPKGYPRYCTVRTMSDGMVVLIG 532

Dyrk3-PA        534 GQSRRGKQRGPPCSKSLSKALDGCKDPLFLNFIRGCLEWDADKRLTPSEALKHPWLRRRL 593
                    ******* ************************************************
Submitted_Seq   533 GQSRRGKPRGPPCSKSLSKALDGCKDPLFLNFIRGCLEWDADKRLTPSEALKHPWLRRRL 592

Dyrk3-PA        594 PRPPSSSSGCGGVSGLCSSRNESPVTGQNRNFAAETTASSTSATSISLTIKRENSHSSLR 653
                    ***********:**  . * *.***..:*:...* * ************:::*****:
Submitted_Seq   593 PRPPSSSSGCGGISGASLSGNQSPVPARNKDVVPEITQSSTSATSISLTIKKDKSHSSLH 652

Dyrk3-PA        654 LHHGAVPETDFKLIKSVPEGSSTATKEPMMNSDILPESFRQTTVVSP---SKHSADSGGM 710
                    *:: .* ****:* *.***** :*** *: *:*** :**::***.**   *.*******:
Submitted_Seq   653 LKQLGVGETDFRLTKCVPEGSLSATKAPLTNADILVDSFKKTTVASPHLPSSHSADSGGV 712

Dyrk3-PA        711 SCLSAVDVGPSRYY-------PYMNNNENNRL--FSSSLNSSANSLSHLEQATKLDALGE 761
                    * *:*****.:***       * *:*******  **.*********::***.:.**.***
Submitted_Seq   713 SVLTAVDVGAARYYAPTLSYLPQMKNNENNRLLKFSGSLNSSANSLTQLEQVSSLDVLGE 772

Dyrk3-PA        762 YSASTTPNLLSKNTGYSFNSGSINIDVAQESLVNIASNYALDKSIDIIGKSNVSLHTNKL 821
                    ******** : :*.*:*:* :: ** *******::*.**:****:** ***:***:***
Submitted_Seq   773 YSASTTPNLPA-DTSYAFDSRTVAIDSAQESLVNLTSSYAIDKSIEIIEKSNLSLHSNKL 831

Dyrk3-PA        822 KVQSKDM 828
                    .:**:**
Submitted_Seq   832 TLQSNDM 838
```

Figure 126: Protein Alignment of *Dyrk3*-PA in *D. melanogaster* vs. the proposed model of *D. eugracilis Dyrk3*.

Additional Analysis of *PIP4K*

The 2nd gene in contig12, *PIP4K*, was selected for additional investigation. As described earlier, the full name of this gene is *Phosphatidylinositol 5-phosphate 4-kinase*, and the enzyme it produces mediates the conversion of phosphatidylinositol 4 phosphate to phosphatidylinositol 4,5 bisphosphate. This enzyme is also implicated in the regulation of mTOR (mammalian target of rapamycin) signaling and control of cell size. Examination of the modENCODE expression data provided for *PIP4K* in reveals that this gene is expressed in a variety of tissue types and developmental stages (Fig. 127).

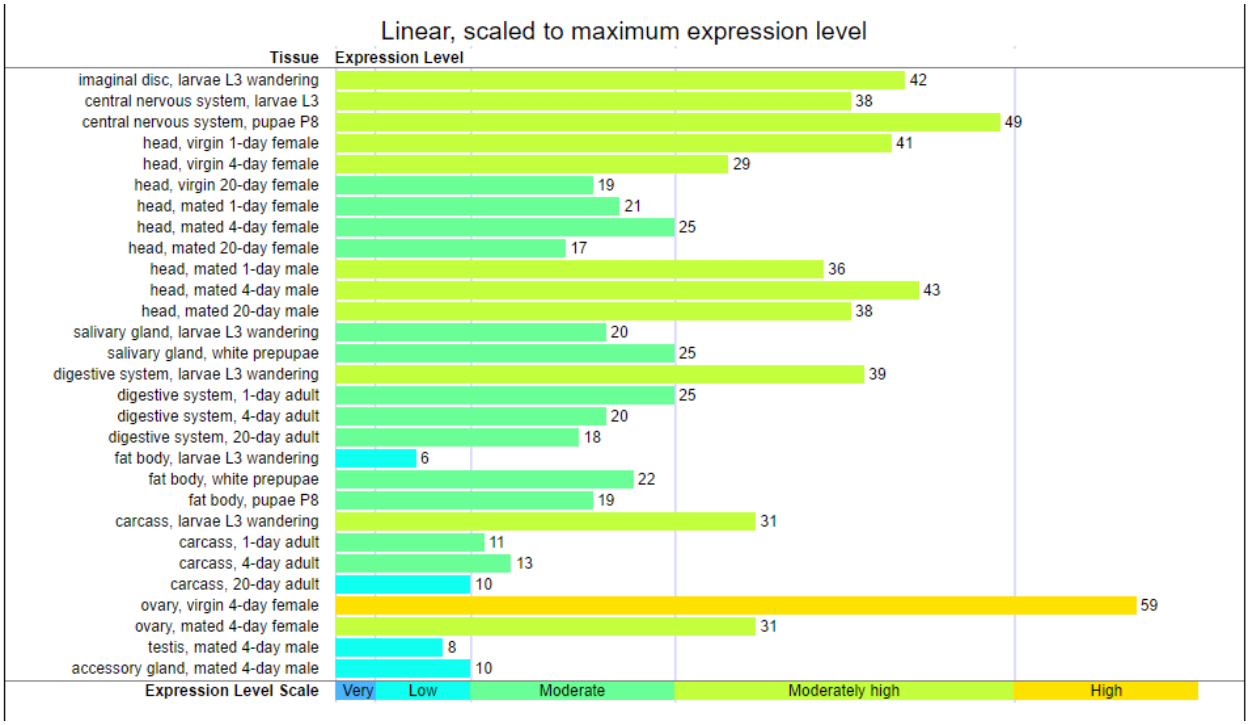


Figure 127: modENCODE Expression data for *PIP4K*. This gene is widely expressed, reaching moderate to high levels of expression in most tissue types and developmental stages.

Phyre^2 Analysis of *PIP4K*

The protein fold recognition server Phyre^2 (Protein Homology/analogy Recognition Engine V 2.0) is a tool that is used to predict 3-dimensional protein structures from their respective amino acid sequences. The amino acid sequence of the annotated CDS of *PIP4K* in *D. eugracilis* was run through a Phyre^2 search. Recall that both *PIP4K*-PA and *PIP4K*-PB have the same CDSs. Thus, in the subsequent descriptions of analyses performed on *PIP4K,* it should be kept in mind that the CDS of both isoforms was analyzed. This shared CDS of the PA and PB isoforms will be referred to as *PIP4K.* The Phyre^2 predicted protein from the search with *D. eugracilis PIP4K* is shown in Figure 128. Along with a putative protein structure for *PIP4K*, Phyre^2 compared the predicted protein to known proteins and provided a summary of the most similar ones (Fig. 129).

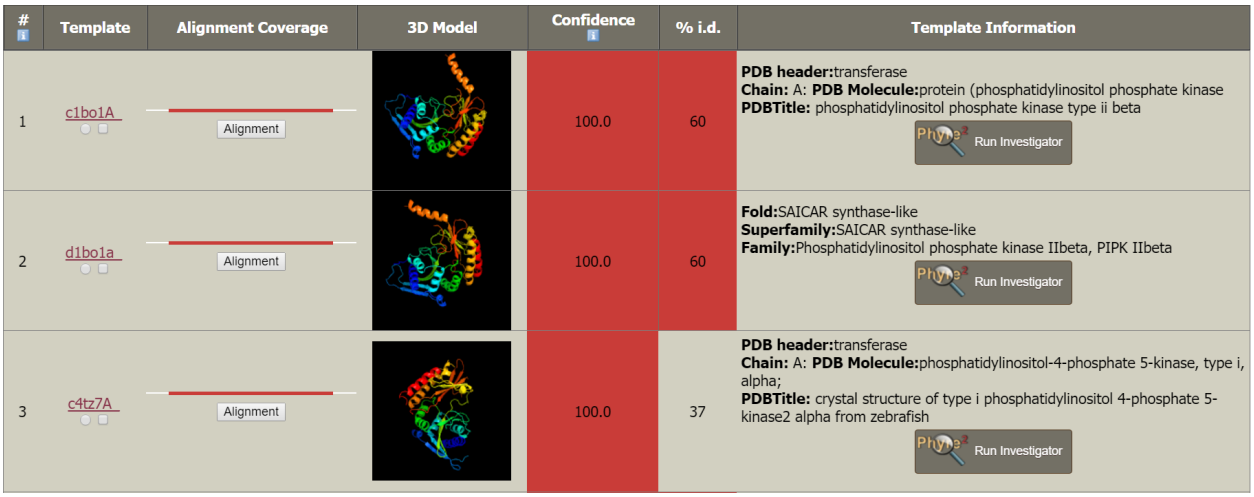
A

B

Confidence and coverage			
Confidence:	100.0%	Coverage:	80%

325 residues (80% of your sequence) have been modelled with 100.0% confidence by the single highest scoring template.

Figure 128: Phyre[2] predicted protein structure. Shown in Figure A is the Phyre[2] predicted protein structure for *PIP4K* in *D. eugracilis.* Figure B describes the confidence and coverage associated with this predicted protein. 80% is a very high level of coverage.



#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c1bo1A	Alignment		100.0	60	**PDB header:**transferase **Chain:** A: **PDB Molecule:**protein (phosphatidylinositol phosphate kinase **PDBTitle:** phosphatidylinositol phosphate kinase type ii beta Run Investigator
2	d1bo1a	Alignment		100.0	60	**Fold:**SAICAR synthase-like **Superfamily:**SAICAR synthase-like **Family:**Phosphatidylinositol phosphate kinase IIbeta, PIPK IIbeta Run Investigator
3	c4tz7A	Alignment		100.0	37	**PDB header:**transferase **Chain:** A: **PDB Molecule:**phosphatidylinositol-4-phosphate 5-kinase, type i, alpha; **PDBTitle:** crystal structure of type i phosphatidylinositol 4-phosphate 5-kinase2 alpha from zebrafish Run Investigator

Figure 129: Comparison of *PIP4K* predicted protein structure to known protein structures. In this figure are the top three proteins that most closely resemble the *PIP4K* predicted protein. Examination of their alignment coverage reveals that the central portion of this protein model appears to be the most well conserved, while the ends are not conserved between the protein models.

The Investigator function was run on the first aligned protein to *PIP4K*. The title of this protein is phosphatidylinositol phosphate kinase type ii beta. The Investigator function allowed for a more in-depth analysis of this protein, which most closely resembles the putative protein structure of *PIP4K*. An analysis of the conservation of phosphatidylinositol phosphate kinase type ii beta was performed (Fig. 130). It was found that the interior of this protein is more highly conserved that the exterior regions. This finding aligns with the pocket detection analysis that was also performed on this protein structure, which found a large pocket in the interior region of this protein (Fig. 131).

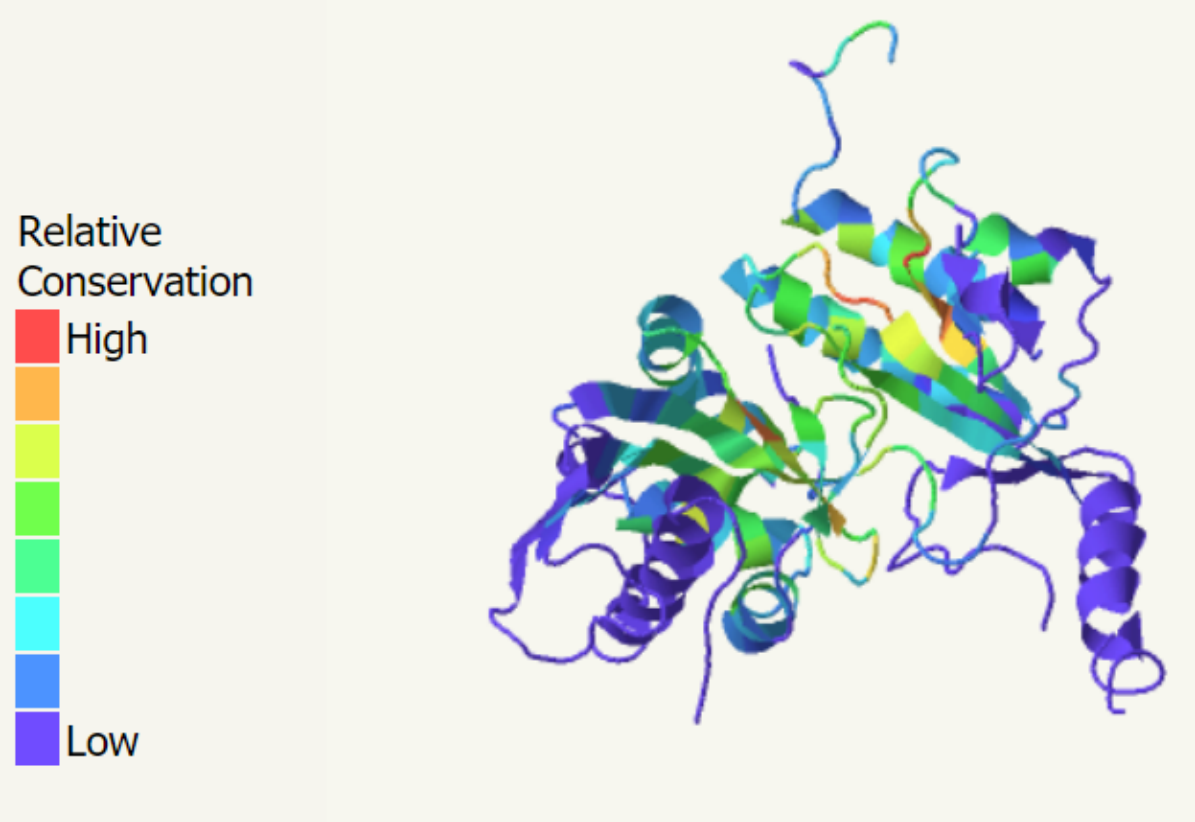


Figure 130: Conservation analysis of phosphatidylinositol phosphate kinase type ii beta. This protein is most highly conserved in the interior part of its structure. The exterior helices are blue, indicating lower levels of conservation.

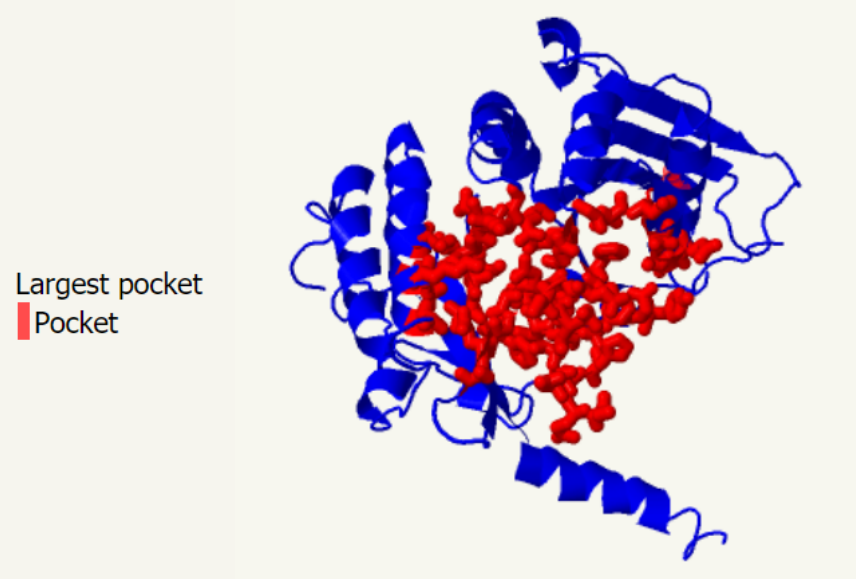


Figure 131: Detected pocket in phosphatidylinositol phosphate kinase type ii beta. Large pockets are frequently associated with active sites. fpocket2 is the program used to detect the largest pocket in this protein model. This protein appears to have a large pocket in its interior region, where conservation was found to be at the highest levels in Figure 130. Referring back to the Dot Plot of this amino acid sequence (Fig. 68) reveals that it is highly conserved within *Drosophila*.

SmartBLAST and Clustal Analysis of *PIP4K*

More comparative analyses were performed on the *D. eugracilis PIP4K* amino acid sequence to characterize this protein. First, a BLASTp search was used to determine that most of the *D. eugracilis PIP4K* amino acid sequence is part of the PIPKc superfamily (Fig. 132). Phosphatidylinositol phosphate kinases in this superfamily catalyze the phosphorylation of phosphatidylinositol phosphate form phosphatidylinositol bisphosphate.

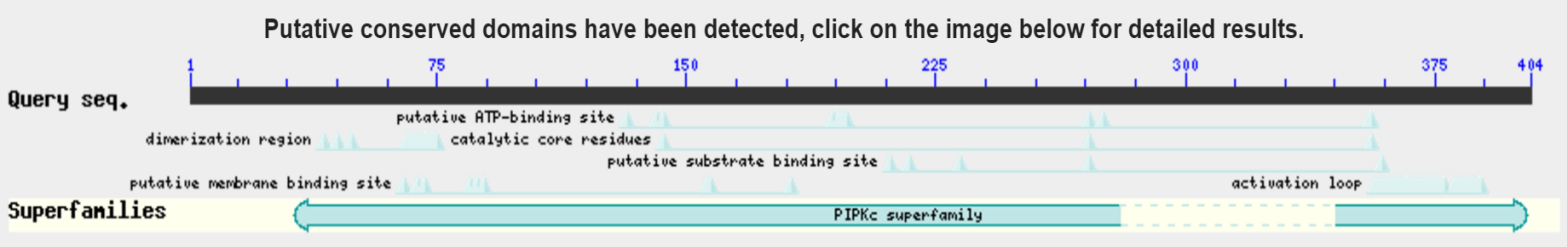
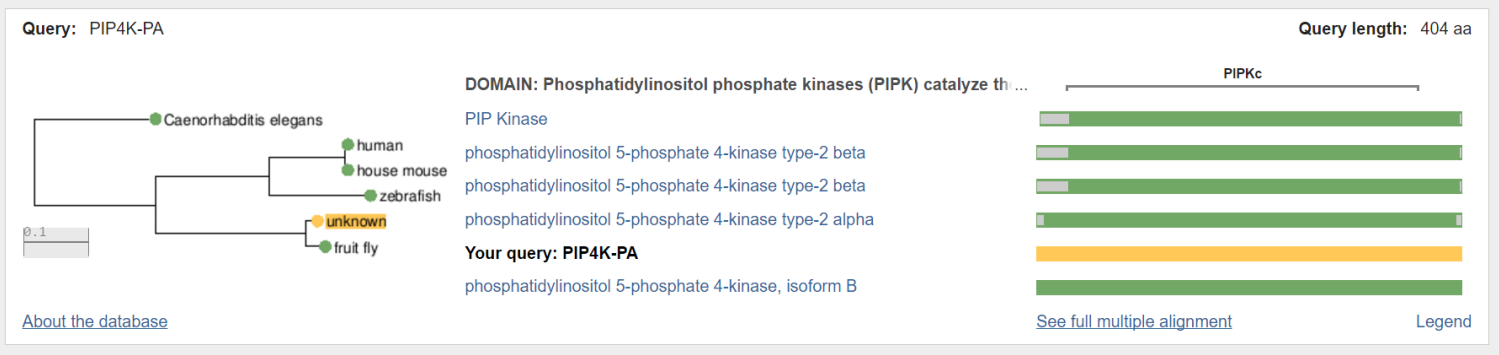


Figure 132: Conserved domains detected in the BLASTp output of a *D. eugracilis PIP4K* query. The detected conserved domains in this amino acid sequence are all part of the PIPKc superfamily. The putative substrate binding domain appears to correspond with the pocket in the Phyre[2] protein model.

The *D. eugracilis PIP4K* amino acid sequence was then entered as the query in an NCBI SmartBLAST search. The output of this search found that this protein is highly conserved across many types of organisms with substantial levels of evolutionary separation, such as humans, house mice, zebrafish, and roundworms (Fig. 133).

A



B

Best hits

Select: All None Selected:0

Alignments GenPept

Description	Max score	Total score	Query cover	E value	Ident	Accession
phosphatidylinositol 5-phosphate 4-kinase, isoform B [Drosophila melanogaster]	791	791	100%	0.0	96%	NP_001033805.1
phosphatidylinositol 5-phosphate 4-kinase type-2 beta [Homo sapiens]	431	431	94%	1e-148	58%	NP_003550.1
PIP Kinase [Caenorhabditis elegans]	423	423	93%	5e-146	55%	NP_497500.1
phosphatidylinositol 5-phosphate 4-kinase type-2 beta [Mus musculus]	424	424	94%	6e-146	58%	NP_473392.1
phosphatidylinositol 5-phosphate 4-kinase type-2 alpha [Danio rerio]	412	412	97%	1e-141	56%	NP_001122174.1

Figure 133: Summary of *D. eugracilis PIP4K* SmartBLAST results. The *PIP4K* query produced alignments to proteins in several organisms with substantial evolutionary distance between them (A). All of these aligned proteins have low e-scores (B).

The five species with similar proteins to the *PIP4K* query identified in the SmartBLAST search were selected as candidates for a Clustal analysis. The orthologous proteins to *PIP4K* were aligned using Clustal Omega (Fig. 134). This Clustal output was found to be consistent with the high levels of conservation described in the BLASTn output. In terms of the evolution of this protein, one hypothesis states this high level of conservation across these species suggests that it emerged early in evolutionary history. This would explain why this protein is shared by so many species. The fact that it is so well conserved suggests that it is important to the reproductive fitness of the organisms and may not have undergone much divergence from its original form.

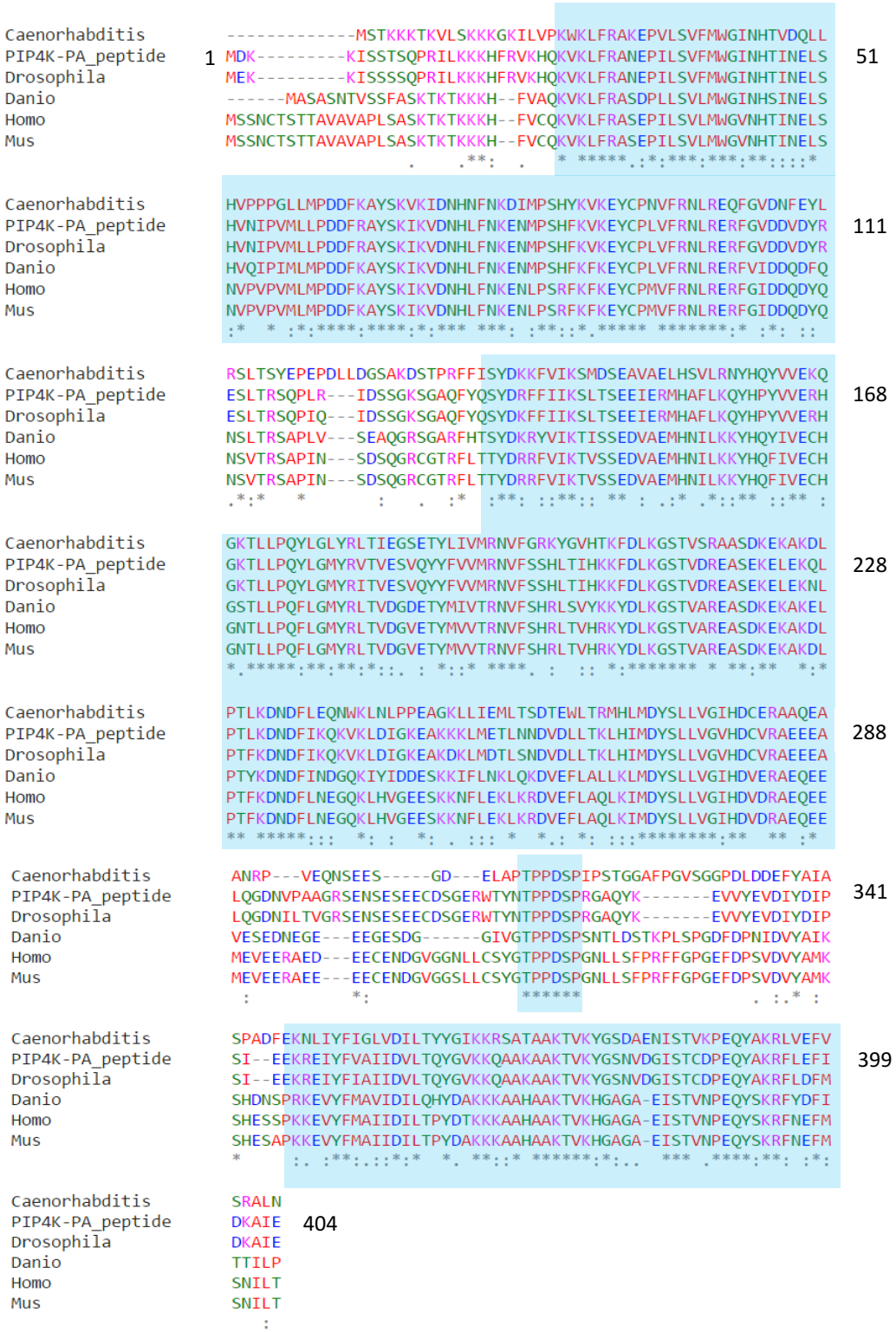


Figure 134: Clustal alignment of orthologous proteins to *PIP4K*. Regions highlighted in blue show high levels of conservation. This alignment provides further evidence that *PIP4K* is highly conserved, even across organisms with substantial evolutionary separation. Most of the amino acid sequence associated with the pocket predicted by Phyre[2] are in highly conserved regions.

Repeats

Repeats in *D. eugracilis*

The entire sequence of *D. eugracilis* contig12 was scanned for repeats by the RepeatMasker program. The total composition of the repeats in contig12 identified by RepeatMasker is summarized in Table 16. The total size of all the repeats in contig12 was found to be 12816 bp. Given that the size of contig12 is 38500 bp, it was determined that the percentage of repetitive DNA in contig12 is 33.3%. This is an expected result, as the repeat density of the F element in most species of *Drosophila* is about 30%. The collection of repeats was then screened by size. All repeats of length 500 bp or greater were assembled into Table 17.

Repeat Name/Class	Repeat Size (bp)	Repeat Occurrences
DNA/TcMar-Mariner		
rnd-1_family-53	552	1
DNA/TcMar-Mariner Total	**552**	**1**
LINE/I		
rnd-1_family-579	202	1
rnd-1_family-588	525	1
rnd-5_family-1962	1912	2
rnd-5_family-2262	233	1
rnd-5_family-2694	144	1
rnd-5_family-2695	413	1
LINE/I Total	**3429**	**7**
LINE/Jockey		
rnd-1_family-24	337	1
rnd-1_family-293	113	1
rnd-1_family-536	120	2
rnd-5_family-2412	104	1
rnd-5_family-701	323	1
LINE/Jockey Total	**997**	**6**
LTR/Pao		
rnd-1_family-16	95	1
rnd-1_family-39	122	1
rnd-1_family-43	976	2
rnd-1_family-78	170	1
LTR/Pao Total	**1363**	**5**
RC/Helitron		
rnd-2_family-9	71	1
rnd-3_family-30	2950	10
RC/Helitron Total	**3021**	**11**
Retroelement		
rnd-1_family-3	832	3
Retroelement Total	**832**	**3**
Unknown		
rnd-1_family-0	128	2
rnd-1_family-396	31	1
rnd-1_family-4	81	1
rnd-1_family-614	133	1
rnd-4_family-237	1977	9
rnd-5_family-1565	91	1
rnd-5_family-176	71	1
rnd-5_family-6968	110	1
Unknown Total	**2622**	**17**
Grand Total	**12816**	**50**

Table 16: *D. eugracilis* contig12 Repeat Summary.

Start Position	End Position	Strand	Repeat Class	Size (bp)
6424	6926	Plus	Unknown	503
6715	7290	Plus	RC/Helitron	576
20308	20807	Minus	RC/Helitron	500
24829	25470	Plus	LTR/Pao	642
27333	27884	Minus	DNA/TcMar-Mariner	552
28861	29389	Minus	Retroelement	529
29897	30421	Minus	LINE/1	525
30762	31837	Minus	LINE/1	1076
31944	32779	Minus	LINE/1	836

Table 17: Large (≥ 500 bp) repeats in *D. eugracilis* contig12. Contig12 contains nine repeats that are greater or equal to 500 bp in length.

Repeats in *D. melanogaster*

Using the *D. melanogaster* Net Track in the *D. eugracilis* Genome Browser, the region of the *D. melanogaster* F element that corresponds to contig12 was identified. This region occurs from 1185654-1230738 in the *D. melanogaster* Genome Browser. Using the "FlyBase Pseudogenes" track and the "FlyBase Non-protein Coding Genes" track, it was determined that there are no pseudogenes and no non-protein coding genes in this region of the *D. melanogaster* F element.

In the *D. melanogaster* Genome Browser, the evidence track ("repeats > 500 bp in size") was created from the repetitious elements identified by RepeatMasker. As the name suggests,

this track presents only those repeats that are greater than 500 bp in size. The region of the *D. melanogaster* F element that corresponds to contig12 only contains a single identified large repeat (Fig. 135). However, as discussed in the annotation of *Mitf*, there is a large intron that occurs in this region that has a gap in the sequence, and it is likely that there are repeats in this intron, some of which may be greater than 500 bp.

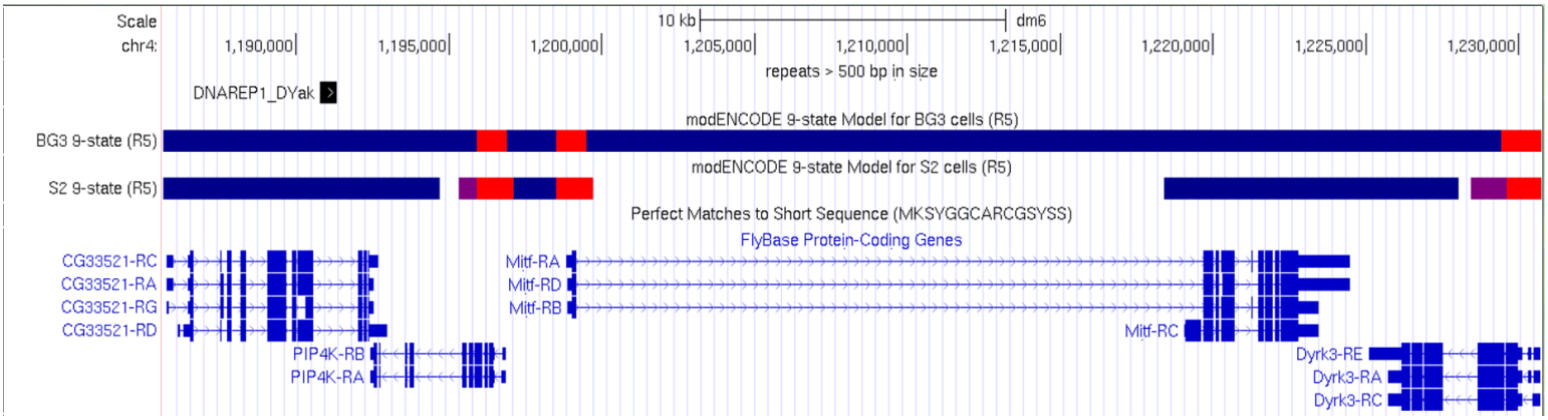


Figure 135: Large (> 500 bp) repeats in the region of the *D. melanogaster* F element that corresponds to *D. eugracilis* contig12. While there is only a single large repeat identified in this region, it is possible that there are more repeats to be found in the gap in the first intron of *Mitf*-RA, which is analyzed in Figure 81.

Synteny

Comparison of Relative Gene Positions Between *D. eugracilis* and *D. melanogaster*

Using the Annotation Files Merger, the GFF files of every annotated gene in contig12 were assembled into a single file. This file was used to produce a custom track in the *D. eugracilis* Genome Browser. With this track, it was possible to view all of the final gene models in *D. eugracilis* contig12 (Fig. 136). The relative gene positions in contig12 were compared to the corresponding region in *D. melanogaster* (Fig. 137).

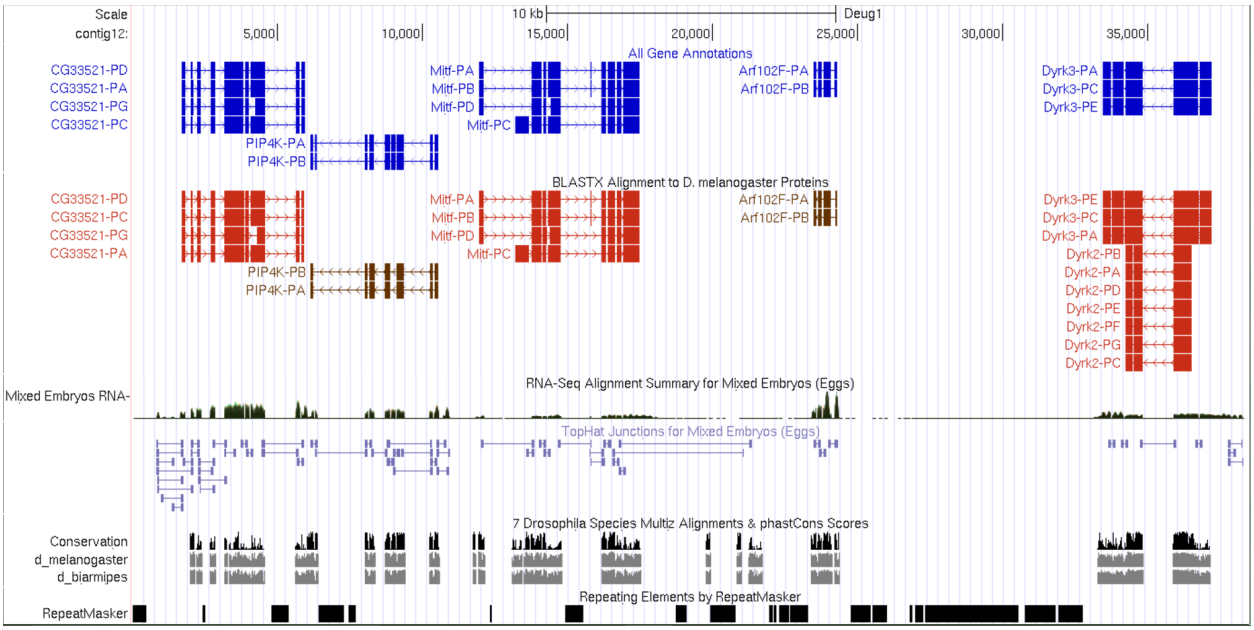


Figure 136: Final gene models of all five genes in contig12. The final gene models are shown in the blue evidence track labelled "All Gene Annotations" at the top of the browser window.

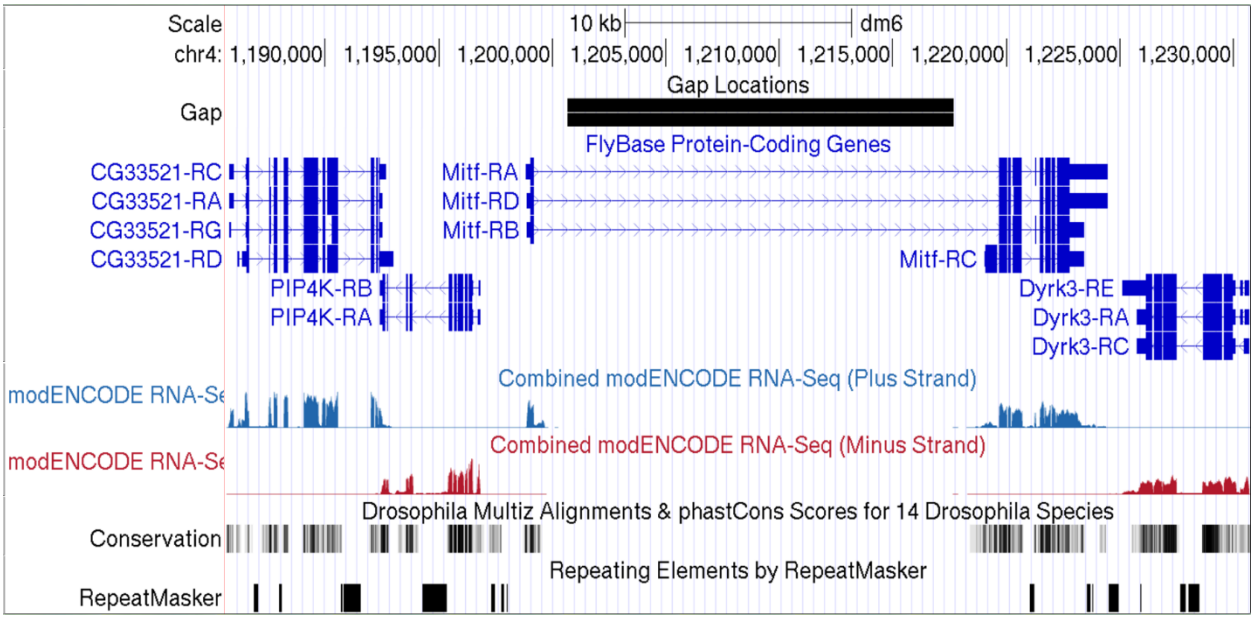


Figure 137: Genes found in the region of *D. melanogaster* corresponding to contig12.

Examination of the two regions reveals that *Arf102F* is present in *D. eugracilis*, but not in *D. melanogaster*. Searching for *Arf102F* in the *D. melanogaster* Genome Browser reveals that it occurs approximately 50 kb from the region corresponding to contig12 (Fig. 138).

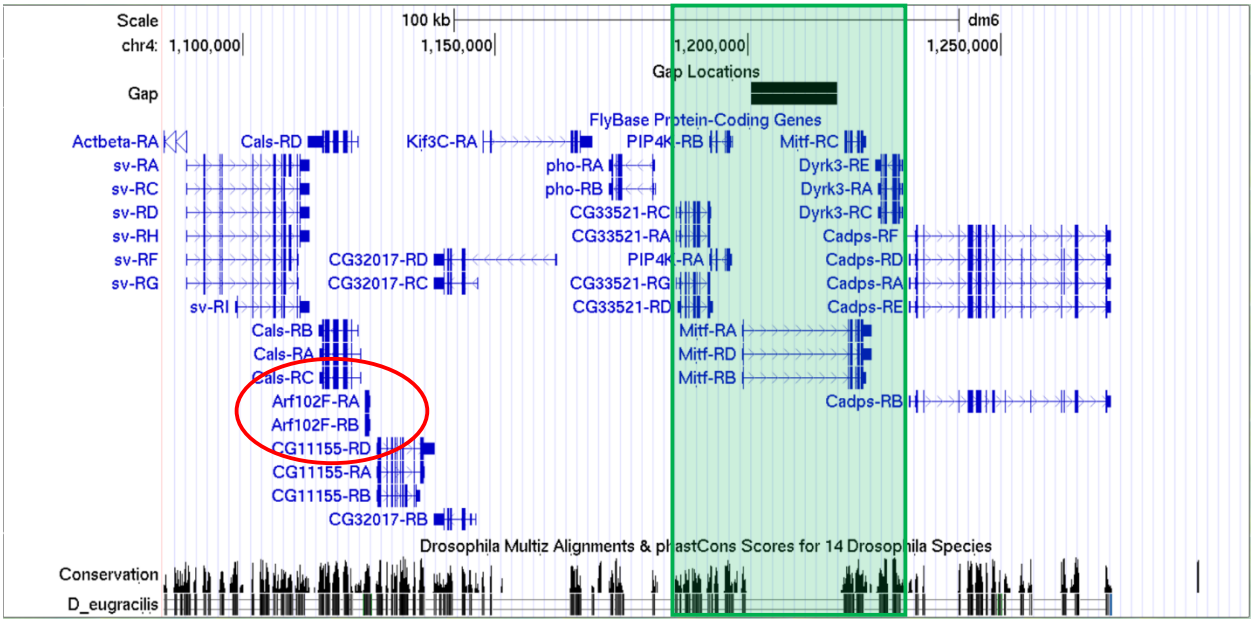


Figure 138: Location of *Arf102F* in *D. melanogaster*. *Arf102F* (red circle) occurs approximately 50 kb upstream of the region corresponding to contig12 in *D. melanogaster* (green box).

The relative orientation of the four genes shared by contig12 and the corresponding region of the *D. melanogaster* F element is conserved. In contig12, *Arf102F* occurs on the minus strand. In *D. melanogaster*, *Arf102F* occurs on the plus strand. However, the orientation of genes in contig12 is arbitrary, so it must be determined if the other four genes have the same or opposite orientations in *D. melanogaster.* All four other genes were found to maintain the same relative orientations between *D. melanogaster* and *D. eugracilis*. Thus, it can be concluded that *Arf102F* has undergone an inversion. Since *Arf102F* has shifted without any apparent change in the flanking genes, one can hypothesize that it was not an inversion that is responsible for this gene's change in position. It is possible that *Arf102F* underwent a transposition and inversion separately. Two large repeats found near *Arf102F* are classified as helitrons. Another hypothesis is that they moved the *Arf102F* gene during replication and transposition.

Discussion

Now that the CDSs and some of the TSSs of the genes in contig12 have been annotated, these annotations will be gathered together with other independent student annotations and reconciled to create final models of these genes. Every contig will be annotated independently by at least two groups for quality control. These CDS and transcription start site annotations will provide a greater set of data to be used in a comparative genomic study of the *Drosophila* F element.

The CDS annotations of these genes could be compared to the CDS annotations of euchromatic genes from the base of the *D. eugracilis* D element, as was done in other *Drosophila* species by Leung *et al.* in "*Drosophila* Muller F elements maintain a distinct set of genome properties over 40 million years of evolution." The TSS annotations of these genes are of particular interest, as the TSS of F element genes are believed to play a role in the genes' expression in heterochromatin. The distributions of core promoter motifs around these *D. eugracilis* F element TSSs will be compared to the distributions of core promoter motifs around euchromatic D element genes, which may identify unique features of F element TSSs. In addition, sequences of regions around the annotated *D. eugracilis* TSSs will be scanned for novel core promoter motifs. Taken together, these comparative analyses will hopefully provide clues as to how these genes are able to function in a heterochromatic environment.

Literature Cited

Elgin, S. C. R., and G. Reuter, 2013 Position-effect variegation, heterochromatin formation, and gene silencing in *Drosophila*. Cold Spring Harb. Perspect. Biol. 5: a017780

Leung W. *et al.,* 2015 *Drosophila* muller f elements maintain a distinct set of genome properties over 40 million years of evolution. G3 (Bethesda). 5:719-40

Appendix

GFF, transcript sequence, and peptide sequence files for each isoform of each gene are submitted electronically.

Acknowledgements

I would like to thank the Bio 434W professors Dr. Sarah Elgin, Dr. Jeremy Buhler, and Dr. Christopher Shaffer, as well as the Bio 434W teaching assistants Wilson Leung, Kailong Mao, Emily Chi, and Ryan Friedman for their guidance and expertise. I would also like to thank Dr. April Bednarski for her instruction and critique of my writing. Lastly, I would like to thank Washington University in St. Louis and the Genomics Education Partnership for the opportunity to perform research with guidance and instruction of such high quality.