Micah Rickles-Young
Bio 434W
Spring 2017
Dr. Elgin, Dr. Shaffer, Dr. Bednarski

Annotation of *Drosophila eugracilis* Contig 24

Introduction

The chromatin in a eukaryotic cell is not packaged uniformly in the nucleus. While all DNA is packaged in nucleosome arrays, some regions are more densely packaged and some are more loosely packaged. The more densely packed chromatin, heterochromatin, contains DNA bound tightly around nucleosomes and is generally associated with silencing of transcriptional activity. The more loosely packed chromatin, euchromatin, is associated with actively transcribed genes and regions with more DNA-protein interactions. While tight packaging tends to prevent transcription of heterochromatin, some organisms may still express genes found in heterochromatic regions, including many members of the fruit fly genus, *Drosophila*.

The fourth chromosome in *Drosophila*, also referred to as the F-element, is composed mainly of heterochromatin. It is abundant in histone 3 lysine 9 di- and tri-methylation (H3K9me2/3) and is highly repeat dense (30% repeats), two common features of transcriptionally silent, heterochromatic DNA. However, the F-element contains ~80 genes in the distal 1.3 Mb of the chromosome that are fully transcriptionally active. The exact mechanism by which genes on the *Drosophila* F-element are transcribed is not fully understood.

This paper examines *D. eugracilis* contig 24, a 57 kb segment from the fourth chromosome, details gene and transcription start site (TSS) annotations, and describes the function of *unc-13,* a gene whose ortholog in *D. eugracilis* was identified. Gene annotation of contig 24 used the well-annotated *D. melanogaster* genome as a reference. Both *D. eugracilis* and *D. melanogaster* are members of the *melanogaster* group in the *Sophophora* subgenus of *Drosophila. D. melanogaster* and *D. eugracilis* share a most recent common ancestor 10-15

million years ago, allowing for a noticeable level of genetic difference while keeping most important genomic motifs recognizably similar. Annotation of TSSs also used another member of the *melanogaster* group, *D. biarmipes,* as a reference.

This study seeks to annotate the *D. eugracilis* F-element, adding to the existing body of analyzed *Drosophila* species. By examining the conserved elements of the F element across the *Drosophila* phylogenetic tree, we hope to gain an understanding of the genomic features that contribute to this unique chromatin state and the specific details of the relationship between chromatin structure and gene expression.

Contig 24 Overview

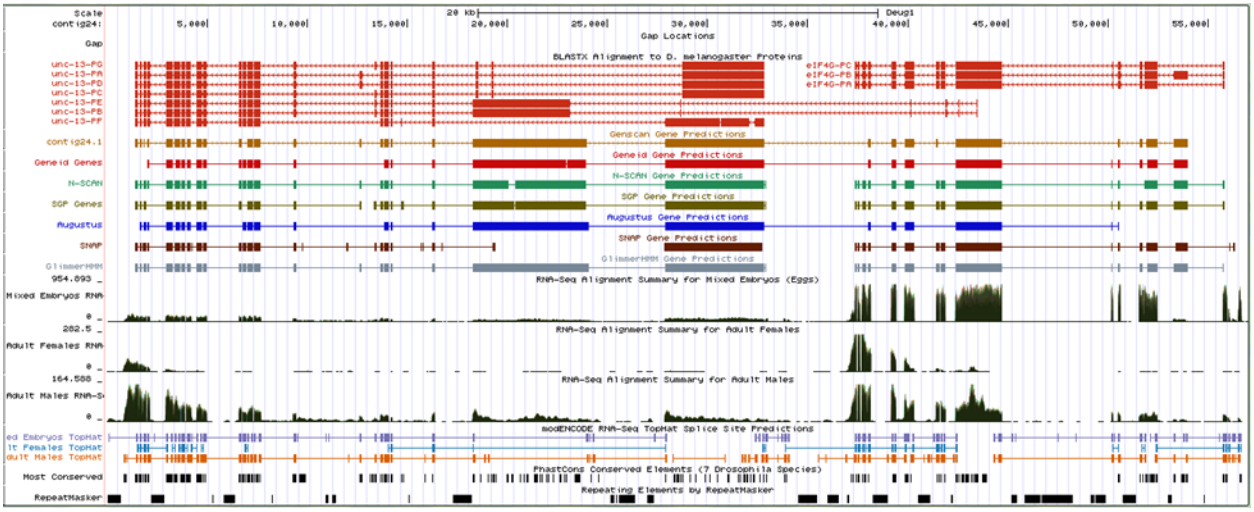


Figure 1: UCSC Genome browser overview of* D. eugracilis *contig 24*. This contig contains two features orthologous to* D. melanogaster *representing the genes* unc-13 *and* eIF4G.

The UCSC Genome browser is a web-based application that compiles and displays evidence tracks for gene annotation, including model-based gene prediction, BLAST alignment, and RNA sequencing (RNA-seq) data. The GEP mirror of the UCSC genome browser was used for annotating contig 24. BLAST alignment to *D. melanogaster* identified two potential orthologous features in *D. eugracilis,* both on the minus strand. The two orthologs tentatively identified by the BLAST track of the UCSC genome browser are *eIF4G*, a gene that codes for a

translation initiation factor and *unc-13*, which codes for a protein responsible for synaptic vesicle maturation (Fig. 1).

D. eugracilis ortholog of *eIF4G*

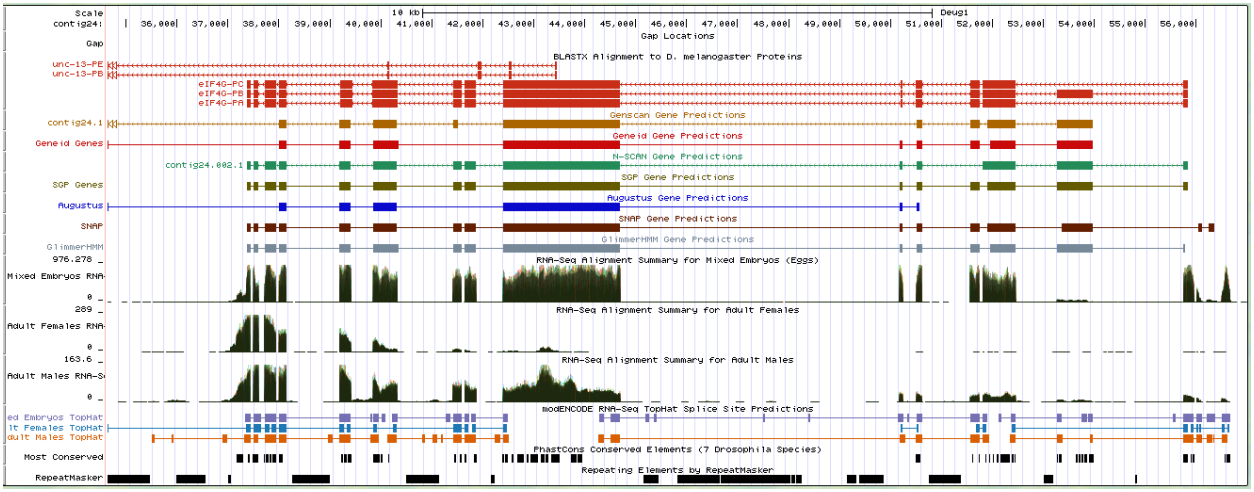
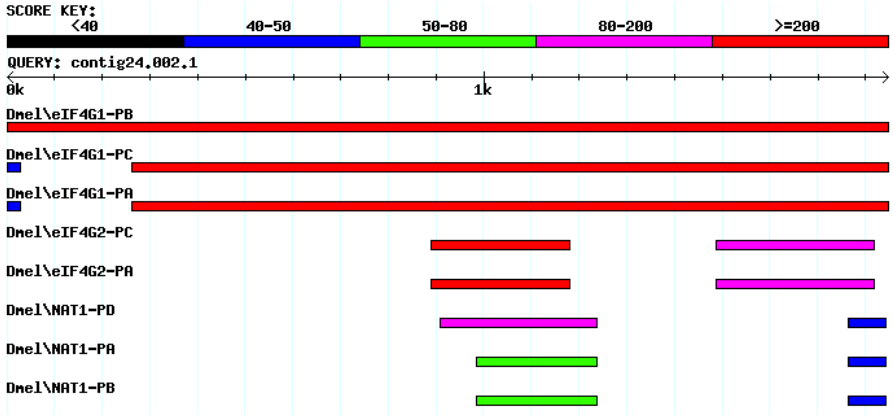


Figure 3: UCSC Genome Browser view of the first feature*. BLAST alignment suggests that the* D. melanogaster *ortholog is eIF4G.*

Query	
Description	Length
contig24.002.1 length=1847	1847

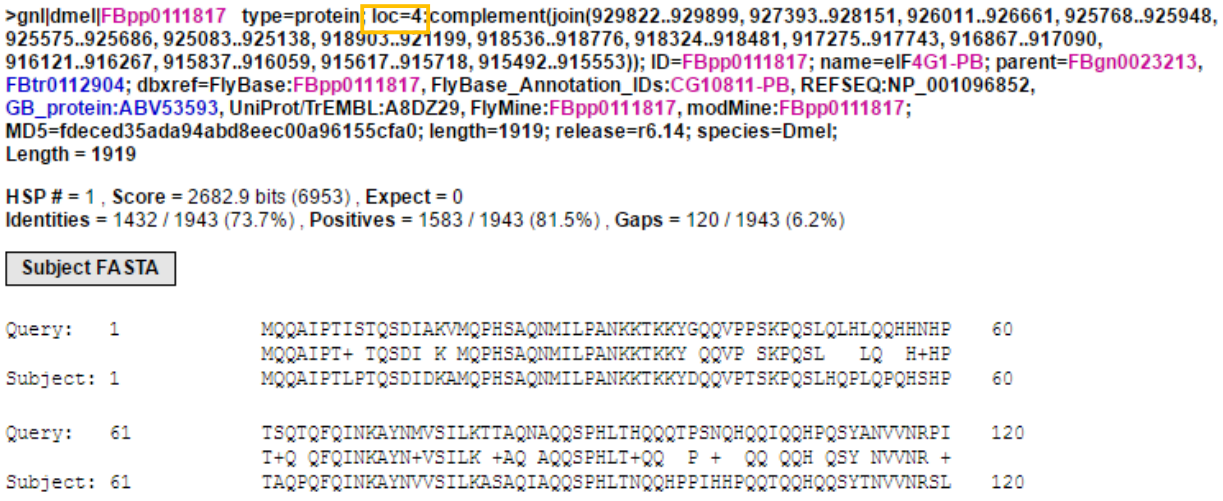


Send BLAST Hits to HitList

BLAST Hit Summary				
	Description	Species	Score	E value
☑	eIF4G1-PB	Dmel	2682.9	0
☑	eIF4G1-PC	Dmel	2378.59	0
☑	eIF4G1-PA	Dmel	2378.59	0
☑	eIF4G2-PC	Dmel	240.736	1.03771e-62
☑	eIF4G2-PA	Dmel	240.736	1.03771e-62
☑	NAT1-PD	Dmel	134.806	8.43124e-31
☑	NAT1-PA	Dmel	72.4034	4.76708e-12
☑	NAT1-PB	Dmel	72.4034	4.76708e-12

***Figure 2: blastx alignment of N-SCAN gene prediction to* D. melanogaster.** eIF4G *is referred to as "*eIF4G1*" in the FlyBase blastx results, but all other references to this gene from sources used for annotation refer to it as* eIF4G*. Alignment strongly suggests* eIF4G *as the* D. melanogaster *ortholog of this feature.*

To identify any other possible orthologs this feature could represent, the N-SCAN gene model predicted amino acid sequence was analyzed for homology to the *D. melanogaster* reference genome using FlyBase annotated proteins database blastp (Fig. 3). The gene prediction matched with *eIF4G-PB* with a BLAST score of 2682.9 and to *eIF4G-PA* and *eIF4G-PC* with a score of 2378.59. Both had an expect (E) value of 0, meaning that it was almost impossible for the alignment to have been the result of random chance. The second highest scoring alignments were to *eIF4G2-PA* and *eIF4G2-PC,* each with a score of 240.736 and an E value of 1.03771e-62 (Fig. 3). The alignment to *eIF4G2* was likely due to the similarity between *eIF4G2* and *eIF4G*, two related translation initiation factors. The analysis report showed that *D. melanogaster eIF4G2,* as well as *NAT1*, the third sequence that matched to the predicted polypeptide, are not located on the fourth chromosome. It would be expected that, in the absence of a major evolutionary event, orthologs would retain similar locations in the genome. As shown in Figure 4, *eIF4G* is located on the fourth chromosome.



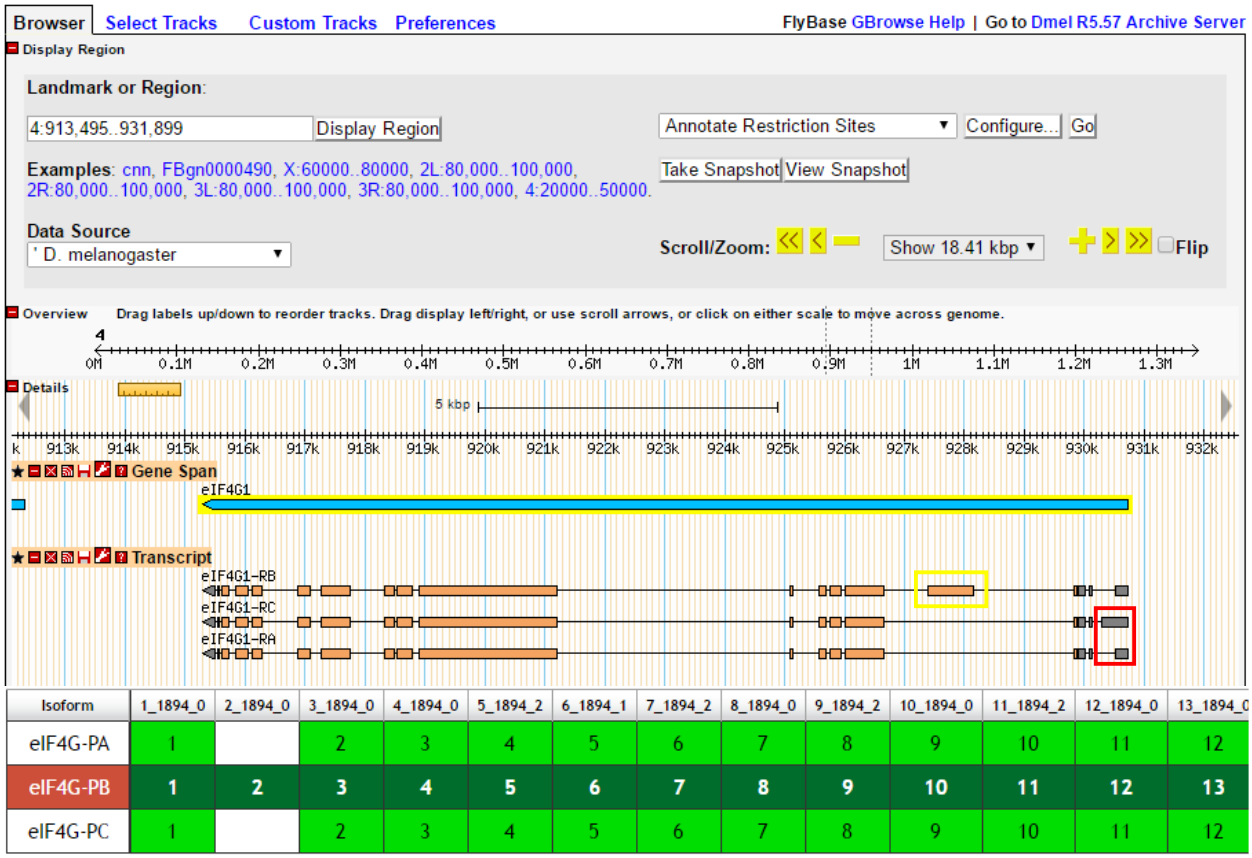
>gnl|dmel|FBpp0111817 type=protein loc=4;complement(join(929822..929899, 927393..928151, 926011..926661, 925768..925948, 925575..925686, 925083..925138, 918903..921199, 918536..918776, 918324..918481, 917275..917743, 916867..917090, 916121..916267, 915837..916059, 915617..915718, 915492..915553)); ID=FBpp0111817; name=eIF4G1-PB; parent=FBgn0023213, FBtr0112904; dbxref=FlyBase:FBpp0111817, FlyBase_Annotation_IDs:CG10811-PB, REFSEQ:NP_001096852, GB_protein:ABV53593, UniProt/TrEMBL:A8DZ29, FlyMine:FBpp0111817, modMine:FBpp0111817; MD5=fdeced35ada94abd8eec00a96155cfa0; length=1919; release=r6.14; species=Dmel;
Length = 1919

HSP # = 1 , Score = 2682.9 bits (6953) , Expect = 0
Identities = 1432 / 1943 (73.7%) , Positives = 1583 / 1943 (81.5%) , Gaps = 120 / 1943 (6.2%)

Subject FASTA

```
Query:   1           MQQAIPTISTQSDIAKVMQPHSAQNMILPANKKTKKYGQQVPPSKPQSLQLHLQQHHNHP   60
                     MQQAIPT+ TQSDI K MQPHSAQNMILPANKKTKKY QQVP SKPQSL   LQ  H+HP
Subject: 1           MQQAIPTLPTQSDIDKAMQPHSAQNMILPANKKTKKYDQQVPTSKPQSLHQPLQPQHSHP   60

Query:   61          TSQTQFQINKAYNMVSILKTTAQNAQQSPHLTHQQQTPSNQHQQIQQHPQSYANVVNRPI   120
                     T+Q QFQINKAYN+VSILK +AQ AQQSPHLT+QQ  P +  QQ QQH QSY NVVNR +
Subject: 61          TAQPQFQINKAYNVVSILKASAQIAQQSPHLTNQQHPPIHHPQQTQQHQQSYTNVVNRSL   120
```

Figure 4: Highest scoring blastp alignment of N-SCAN prediction to* D. melanogaster *protein data base*.*** *Alignment confirms that the suspected ortholog is on the fourth chromosome*



Isoform	1_1894_0	2_1894_0	3_1894_0	4_1894_0	5_1894_2	6_1894_1	7_1894_2	8_1894_0	9_1894_2	10_1894_0	11_1894_2	12_1894_0	13_1894_0
eIF4G-PA	1		2	3	4	5	6	7	8	9	10	11	12
eIF4G-PB	1	2	3	4	5	6	7	8	9	10	11	12	13
eIF4G-PC	1		2	3	4	5	6	7	8	9	10	11	12

***Figure 5: Isoforms of eIF4G in* D. melanogaster.** *The B isoform contains one exon (shown in the yellow box) not present in the A or C isoforms. The A and C isoforms share the same exon configuration with the only differences being in the 5' UTR (show in in the red box). The Gene Record Finder CDS map shows the differences between PB and PA/PC. Thirteen of the fifteen exons are shown in the table.*

Gene Record Finder was used to analyze all coding sequences (CDSs). Gene Record Finder and the associated GBrowse graphical viewer showed the locations and sequences of the exons in the *D. melanogaster* ortholog. *eIF4G* has three isoforms in *D. melanogaster*: *eIF4G-PA, eIF4G-PB,* and *eIF4G-PC* (the A, B, and C isoforms, Fig. 5). The B isoform contains 15 exons, while the A and C isoforms each contain 14 exons. The A and C isoforms have identical coding sequences but distinct 5' untranslated regions (5' UTR). The second exon present in the B isoform is absent from the A and C isoforms. All other exons are present in all isoforms.

For each CDS of *D. melanogaster eIF4G*, the amino acid sequence provided by Gene Record Finder (subject) was aligned to *D. eugracilis* contig 24 (query) using NCBI blastx (Table

1). For all alignments, the low-complexity filter was disabled, compositional adjustments were turned off, and the E value threshold was set to 1e-2. Each *D. melanogaster* CDS aligned with contig 24 in the same order as in *D. melanogaster*. Exon 6_1894_1 showed the lowest scoring alignment with an E value of 6e-4, showing alignment of 13/18 amino acids (Fig. 6); however, inclusion of this exon is supported by RNA-seq data including 851 reads from mixed embryos.

Flybase_ID	Query start-end	Subject Length	E value	Identities	Positives	Gaps	Frame
1_1894_0	55827-55750	26	9e-12	22/26	23/26	0/26	-1
2_1894_0	53965-53264	253	5e-93	159/256	189/256	25/256	-3
3_1894_0	52455-51808	217	5e-118	181/219	197/219	5/219	-1
4_1894_0	51744-51565	60	6e-22	38/60	46/60	0/60	-1
5_1894_2	50605-50507	36	2e-16	30/33	31/33	0/33	-3
6_1894_1	50239-50201	18	6e-4	11/13	12/13	0/13	-3
7_1894_2	44696-42393	765	0.0	565/773	628/773	13/773	-2
8_1894_0	41870-41649	80	5e-34	61/79	65/79	5/79	-2
9_1894_2	41576-41412	52	5e-21	42/55	45/55	3/55	-2
10_1894_0	40317-39847	156	7e-77	122/157	138/157	2/157	-1
11_1894_2	39410-39189	74	1e-43	64/74	69/74	0/74	-2
12_1894_0	38146-38000	49	4e-28	44/49	47/49	0/49	-3
13_1894_0	37941-37720	74	5e-45	69/74	72/74	0/74	-1
14_1894_2	37597-37499	33	6e-19	32/33	33/33	0/33	-3
15_1894_2	37435-37376	20	7e-10	20/20	20/20	0/20	-3

Table 1: Exon by exon NCBI blastx search of* D. melanogaster *exons in* D. eugracilis *contig 24. *All exons present in* D. melanogaster *align to* D. eugracilis.

eIF4G:Dmel_exon_1

Sequence ID: Query_177997 **Length:** 26 **Number of Matches:** 3

Range 1: 1 to 26 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
47.8 bits(112)	9e-12	22/26(85%)	23/26(88%)	0/26(0%)	-1

```
Query  55827  MQQAIPTISTQSDIAKVMQPHSAQNM  55750
              MQQAIPT+ TQSDI K MQPHSAQNM
Sbjct  1      MQQAIPTLPTQSDIDKAMQPHSAQNM  26
```

eIF4G:2_1894_0

Sequence ID: Query_102835 **Length:** 253 **Number of Matches:** 1

Range 1: 1 to 253 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
291 bits(745)	5e-93	159/256(62%)	189/256(73%)	25/256(9%)	-3

```
Query  53965  ILPANKKTKKYGQQVPPSKPQSLQLHLQQHHNHPTSQTQFQINKAYNMVSILKTTAQNAQ  53786
              ILPANKKTKKY QQVP SKPQSL   LQ  H+HPT+Q QFQINKAYN+VSILK +AQ AQ
Sbjct  1      ILPANKKTKKYDQQVPTSKPQSLHQPLQPQHSHPTAQPQFQINKAYNVVSILKASAQIAQ  60

Query  53785  QSPHLTHQQQTPSNQHQQIQQHPQSYANVVNRPISASAPVGAQQSTVMCNGSNIMTVNSC  53606
              QSPHLT+QQ  P +  QQ QQH QSY NVVNR +SAS PV A QS+V+CNGS+I+TVNS
Sbjct  61     QSPHLTNQQHPPIHHPQQTQQHQQSYTNVVNRSLSASEPVRA-QSSVICNGSSILTVNSR  119

Query  53605  QLNSGDLNTTAIYNLSSHQALAGSLDEHVRFLNVPDIKKNGNNIGNATVVSNSSNAIVGN  53426
              QLNSGD+N+TAIYN+SS++ L GSLD +V FLNV DIK+NGN  G  +VVSN S   VG+
Sbjct  120    QLNSGDMNSTAIYNISSYRKLTGSLDGNVCFLNVQDIKQNGNISG--SVVSNKSIVGVGS  177

Query  53425  GTTSCTGVSTNSQIML----------------------HDKNIVGVSVNCVDNNRKYDFK  53312
                +SCTGVS N+QI+L                      H+KNIVGVSVNCV+ ++KYDF
Sbjct  178    EKSSCTGVSINNQIVLPNAQIGTSMGIAGTTTAGTSYMHEKNIVGVSVNCVNTSKKYDFN  237

Query  53311  NSSLVLNNNFQTSTPE  53264
              NSSL+ NN++  ST E
Sbjct  238    NSSLLSNNSYPASTAE  253
```

eIF4G:Dmel_exon_3

Sequence ID: Query_71997 **Length:** 217 **Number of Matches:** 6

Range 1: 1 to 217 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
361 bits(926)	5e-118	181/219(83%)	197/219(89%)	5/219(2%)	-1

```
Query  52455  YVSTGNSSSGNSRSNPQSGGIFRGPPPTASAPRGSSGGATRHVHVQQMYSQPLHQNVVLQ  52276
              YVSTGN++SGN+RSNPQSGGIFRGPP T +APRG+SGGATRHVHVQ MYSQPLHQN+VLQ
Sbjct  1      YVSTGNNNSGNTRSNPQSGGIFRGPPSTPNAPRGASGGATRHVHVQPMYSQPLHQNMVLQ  60

Query  52275  QYTQY-PRQQTFPSTHLQYASAPMPYYPYQYVPTLQQQPSPHTRNAVTVNTNVNVGNTLQ  52099
              QYTQY PRQQTFP++HLQYA APMPYY YQYVPTLQQQP PHTR+AVTVNTNVNVGN LQ
Sbjct  61     QYTQYNPRQQTFPASHLQYAPAPMPYYQYQYVPTLQQQP-PHTRSAVTVNTNVNVGNNLQ  119

Query  52098  PVQSGPNGPLPGPGTNSSQLQLLTSAVQPGSSSVMGVSGSPG--MGQVGVSPMVGVGVMP  51925
              PVQSGPNGPLP PG +SSQ+QLLTS VQPG+S+VMGV G PG  MGQVGV PMVGVGVM
Sbjct  120    PVQSGPNGPLPVPGASSSQIQLLTSTVQPGASTVMGV-GGPGSTMGQVGVPPMVGVGVMT  178

Query  51924  TSVQTQSVQVQPSSRRRHQHRLPIIDPATQKNILEDLDK  51808
              TSVQ Q VQVQP+SRRRHQHRL IIDP T+KNIL+D DK
Sbjct  179    TSVQPQPVQVQPASRRRHQHRLQIIDPTTKKNILDDFDK  217
```

eIF4G:Dmel_exon_4

Sequence ID: Query_125399 **Length:** 60 **Number of Matches:** 5

Range 1: 1 to 60 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
78.6 bits(192)	6e-22	38/60(63%)	46/60(76%)	0/60(0%)	-1

```
Query  51744  TNLNTDKDFSEQATLTNTSAAVLSEGPIRIPQQDGVAISNLSNIISQGSESRINASFSPI  51565
              T  NTD +FS+Q T TNT A VLSEGPIRIPQQ+ V ++NL++  SQGSESR NA + PI
Sbjct  1      TKSNTDNEFSDQVTSTNTPATVLSEGPIRIPQQESVGLNNLTSTSSQGSESRTNAPYIPI  60
```

eIF4G:Dmel_exon_5

Sequence ID: Query_245705 **Length:** 36 **Number of Matches:** 7

Range 1: 4 to 36 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
61.6 bits(148)	2e-16	30/33(91%)	31/33(93%)	0/33(0%)	-3

```
Query  50605  ISRQDVGQTPIVSAMSDAPSVEILPTPQRGRSK  50507
              ISR DVG TPIVSAM+DAPSVEILPTPQRGRSK
Sbjct  4      ISRTDVGPTPIVSAMTDAPSVEILPTPQRGRSK  36
```

eIF4G:6_1894_1

Sequence ID: Query_19103 **Length:** 18 **Number of Matches:** 1

Range 1: 1 to 13 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
25.0 bits(53)	6e-04	11/13(85%)	12/13(92%)	0/13(0%)	-3

```
Query  50239  IPIVSPKKVSDSI  50201
              IPIVSPK VS+SI
Sbjct  1      IPIVSPKNVSESI  13
```

eIF4G:Dmel_exon_7

Sequence ID: Query_222625 **Length:** 765 **Number of Matches:** 3

Range 1: 1 to 765 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
1065 bits(2753)	0.0	565/773(73%)	628/773(81%)	13/773(1%)	-2

```
Query  44696  ETDDAVSKAIAATPEEPLQPNP-HQKLLTSELSEQKQAALKTDITKEVEK-VTEKANTEV  44523
              ETDDA S  I+   EE + PN  H  LL S+ S+ KQA   ++I+K+    + E    E+
Sbjct  1      ETDDAGSTPISRATEESMPPNQTHPNLLISDSSKHKQAVSNSEISKDAPTGLKEMHVAEL  60

Query  44522  NFGVFSGNHQLEILTATETNSQPSKNFSVQPEISNIDELPPVDSVADSAKIHTILDNYDS  44343
                V S NH   IL      S  S +FS  PEI +I  + P    A S  +H +LDN +
Sbjct  61     ESSVASENHPAGILDVNVNKSSQSLDFSADPEIDSIGAISPTSPNAVSPILHEVLDNTEL  120

Query  44342  TQSNEMEKSIVENFKDNQCEEQTVHGEISLVSVPDEIELSSMALKKGTCLDDSYIETLDI  44163
               S ++E S  E FKD+Q  E+  H E+SL +  DE E+S+MAL+    LD++ IE
Sbjct  121    --SKKLENSTTERFKDSQSVEKPTHQELSLRNATDETEISAMALQDVNSLDNNQIEKTYP  178

Query  44162  IETKNNGDVLDEVATNESPTETNSTNNTDIDLNLQFPLKSDEMSETDLEDKPKSTSPSDE  43983
               + K N DV +++++ ES  ++ ST NT +D+ LQ    SD   ET L DK  ST
Sbjct  179    SKPKLNVDVSEDISSRESAIKSTSTKNTGVDVGLQ----SDSKPETTLNDKQDSTDLK-V  233

Query  43982  KTGITVSSVINYNEGQWSPSNPSGKKQYNRDQLLQLREVKASRIQPEVKNVSILPQPNLM  43803
              K    +SS+INYNEGQWSP+NPSGKKQY+R+QLLQLREVKASRIQPEVKNVSILPQPNLM
Sbjct  234    KVSAKISSIINYNEGQWSPNNPSGKKQYDREQLLQLREVKASRIQPEVKNVSILPQPNLM  293

Query  43802  PAFIRNNNNNKRVQSMVGMIGNRSSDSGGNYIGKQISMSGVMGGGSRNSMKGMIHVNLSL  43623
              P+FIRNNNNNKRVQSMVG+IGNRS++S GNYIGKQ+SMSGV  GG R+SMKGMIHVNLSL
Sbjct  294    PSFIRNNNNNKRVQSMVGIIGNRSNESAGNYIGKQMSMSGVQSGGGRSSMKGMIHVNLSL  353

Query  43622  NQDVKLSENENAWRPRVLNKSDVDSDTKST---QELVRRVRGILNKLTPERFDTLVKEII  43452
              NQDVKLSENENAWRPRVLNKSD DSD KS     ELVRRVRGILNKLTPERFDTLV+EII
Sbjct  354    NQDVKLSENENAWRPRVLNKSDGDSDAKSALEKDELVRRVRGILNKLTPERFDTLVEEII  413

Query  43451  KLKIDTPEKMDEVIVLVFEKAIDEPNFSVSYARLCHRLISEVKGRDERMESGTKSNLAHF  43272
              KLKIDTP+K+DEVIVLVFEKAIDEPNFSVSYARLC RL +EVK  DERMES TKSN AHF
Sbjct  414    KLKIDTPDKVDEVIVLVFEKAIDEPNFSVSYARLCQRLAAEVKVIDERMESETKSNSAHF  473

Query  43271  RNALLDKTEREFTQNVSQSTAKEKKLQPIVDKIKKSTDANEKAELEAFLEEEERKIRRRS  43092
              RNALLDKTE+EFTQNVSQSTAKEKKLQPIVDKIKK TDANEKAELEAFLEEEERKIRRRS
Sbjct  474    RNALLDKTEQEFTQNVSQSTAKEKKLQPIVDKIKKCTDANEKAELEAFLEEEERKIRRRS  533

Query  43091  GGTVRFIGELFKISMLTGKIIYSCIDTLLNPHSEDMLECLCKLLTTVGAKFEQTPVNSKE  42912
              GGTVRFIGELFKISMLTGKIIYSCIDTLLNPHSEDMLECLCKLLTTVGAKFE+TPVNSK+
Sbjct  534    GGTVRFIGELFKISMLTGKIIYSCIDTLLNPHSEDMLECLCKLLTTVGAKFEKTPVNSKD  593

Query  42911  PGRCYSLEKSITRMQAIASKTDKDGAKVSSRVRFMLQDVIDLRKNKWQTSRNEAPKTMGQ  42732
              P RCYSLEKSITRMQAIASKTDKDGA+VSSRVRFMLQDVIDLRKNKWQTSRNEAPKTMGQ
Sbjct  594    PSRCYSLEKSITRMQAIASKTDKDGARVSSRVRFMLQDVIDLRKNKWQTSRNEAPKTMGQ  653

Query  42731  IEKEAKNEQISAQYFGTLSSNTLVGTQGGSGKRDDRGNARYGESRSGSGYGGSHSQRGDN  42552
              IEKEAKNEQ+SAQYFGTLSS T  G+QGGSGKRDDRGN+RYGESRS S YGGSHSQRGDN
Sbjct  654    IEKEAKNEQLSAQYFGTLSSTTPGGSQGGSGKRDDRGNSRYGESRSSSAYGGSHSQRGDN  713

Query  42551  GNLRHQQQSNIGGVSGSGGAAHSNGNNDDNTWHVQTSKGSRSQAVDSNKLEGL  42393
              GNLRHQQQ+N+GG + SGGA HSNGNND+NTWHVQTSKGSRS AVDSNKLEGL
Sbjct  714    GNLRHQQQNNVGG-NVSGGAGHSNGNNDENTWHVQTSKGSRSLAVDSNKLEGL  765
```

eIF4G:Dmel_exon_8

Sequence ID: Query_227659 **Length:** 80 **Number of Matches:** 5

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

Score	Expect	Identities	Positives	Gaps	Frame
115 bits(289)	1e-34	62/80(78%)	66/80(82%)	5/80(6%)	-2

```
Query  41870  SKLSDQNLETKKMGGLGQFIW---NPAKQSSVPTATPSNPFAVLSSLNDKNSSEWDR--A  41706
              SKLSDQNLETKKMGGL QFIW   +  + SS PT TPSNPFAVLSSL DKNS+E DR  +
Sbjct  1      SKLSDQNLETKKMGGLTQFIWISSDTTRLSSAPTPTPSNPFAVLSSLIDKNSNERDRDRS  60

Query  41705  VPRNKGSYNKGSMERDRYDR  41646
               PRNKGSYNKGSMERDRYDR
Sbjct  61     GPRNKGSYNKGSMERDRYDR  80
```

eIF4G:Dmel_exon_9

Sequence ID: Query_63447 **Length:** 52 **Number of Matches:** 6

Range 1: 1 to 52 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
75.5 bits(184)	5e-21	42/55(76%)	45/55(81%)	3/55(5%)	-2

```
Query  41576  IHSRTGSSQGSRENSSSRSGQQGHGRSLLSTSVQKSASQSKYTQQQVVPGRHTAK  41412
              +HSRTGSSQGSRENSSSR GQQ  GR+LLS+SVQKS S SKYT QQ  P RHT K
Sbjct  1      MHSRTGSSQGSRENSSSRGGQQ--GRTLLSSSVQKSTSHSKYT-QQAPPTRHTVK  52
```

eIF4G:Dmel_exon_11

Sequence ID: Query_174819 **Length:** 74 **Number of Matches:** 22

Range 1: 1 to 74 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
140 bits(354)	1e-43	64/74(86%)	69/74(93%)	0/74(0%)	-2

```
Query  39410  PLIVKKILTVSDLWNKNLKDNSPLKVAKKFLKTYLIYCTQEVGPNFARSMWIKFNLKWSD  39231
              PLIVKKILT+SDLWN NLK+NSP  VAKKFLKTYLIYCTQEVGPNFAR+MWIKFNLKWSD
Sbjct  1      PLIVKKILTISDLWNNNLKENSPSNVAKKFLKTYLIYCTQEVGPNFARNMWIKFNLKWSD  60

Query  39230  FMPENEIEDFIKCN  39189
              FMPE+E+ DFIK N
Sbjct  61     FMPESEVADFIKFN  74
```

eIF4G:10_1894_0

Sequence ID: Query_174287 **Length:** 156 **Number of Matches:** 1

Range 1: 2 to 156 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
240 bits(612)	7e-77	122/157(78%)	138/157(87%)	2/157(1%)	-1

```
Query  40317  QPSLGSSTVNAGGLYRGSEQPSPASLPFSQSTRCVAPAAVFNEASETDLKNIKSVVSEMI  40138
              Q S+GSS VN G LYRGSEQ + A+  FSQ+TR VAP AVF EASETDLK IKSVVSE++
Sbjct  2      QSSVGSSNVNTGPLYRGSEQQTSAT--FSQTTRSVAPVAVFIEASETDLKLIKSVVSEIV  59

Query  40137  ELASASETVTPGVVACIKRVPEELRCSFLYYILTDYLHLADVGKQYRRYLAITVSLLIQQ  39958
              +L++AS+ VTPG V+CIKRVPE+LRCSF+YYILTDYLHLA+VGKQYRRYL+I VS LIQQ
Sbjct  60     DLSAASKEVTPGAVSCIKRVPEKLRCSFIYYILTDYLHLANVGKQYRRYLSIAVSQLIQQ  119

Query  39957  NYISVDHFRLAYNEFSEYANDLIVDIPELWLYILQFA  39847
              NYIS DH RLAYNEF+ YANDLIVDIPELWLYILQFA
Sbjct  120    NYISADHLRLAYNEFTVYANDLIVDIPELWLYILQFA  156
```

eIF4G:Dmel_exon_12

Sequence ID: Query_194959 **Length:** 49 **Number of Matches:** 7

Range 1: 1 to 49 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
95.5 bits(236)	4e-28	44/49(90%)	47/49(95%)	0/49(0%)	-3

```
Query  38146  RLEYVENESMSPVIEQRESPEKHVKNVIDHIDHLLKEGTTADCIIDYSN  38000
              RLEYVENES SPVI+ RE+PEKHVKNVIDHI+HLLKEGTTADCIIDYSN
Sbjct  1      RLEYVENESKSPVIDHRETPEKHVKNVIDHIEHLLKEGTTADCIIDYSN  49
```

eIF4G:Dmel_exon_13

Sequence ID: Query_200355 **Length:** 74 **Number of Matches:** 13

Range 1: 1 to 74 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
145 bits(365)	5e-45	69/74(93%)	72/74(97%)	0/74(0%)	-1

```
Query  37941  GNIMVVDKLFIRGLTETLSNFSILYKENSYKLETETFQKFCIPVLLRYIDSNEDHQLECL  37762
              GNIMVVDKLFIRGLTETLSNFSI YK+NSYKLE+ETFQKFCIPVL RYIDSNEDHQLECL
Sbjct  1      GNIMVVDKLFIRGLTETLSNFSIHYKDNSYKLESETFQKFCIPVLQRYIDSNEDHQLECL  60

Query  37761  YTMQLLVHGLEHPR  37720
              YT+QLLVHGLEHPR
Sbjct  61     YTLQLLVHGLEHPR  74
```

eIF4G:Dmel_exon_14

Sequence ID: Query_46867 **Length:** 33 **Number of Matches:** 4

Range 1: 1 to 33 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
68.6 bits(166)	6e-19	32/33(97%)	33/33(100%)	0/33(0%)	-3

```
Query  37597  LLSELIGELYDAYVIQKESLCKWRDSKDQSAGK  37499
              LLSELIGELYDA+VIQKESLCKWRDSKDQSAGK
Sbjct  1      LLSELIGELYDAFVIQKESLCKWRDSKDQSAGK  33
```

eIF4G:Dmel_exon_15

Sequence ID: Query_84067 **Length:** 20 **Number of Matches:** 5

Range 1: 1 to 20 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
42.0 bits(97)	7e-10	20/20(100%)	20/20(100%)	0/20(0%)	-3

```
Query  37435  VAVKSLNPFFNSLLNDDAN*  37376
              VAVKSLNPFFNSLLNDDAN*
Sbjct  1      VAVKSLNPFFNSLLNDDAN*  20
```

Figure 6: blastx alignments of all* D. melanogaster eIF4G *exons to* D. eugracilis *contig 24. *Subject is* D. melanogaster *exons and Query is* D. eugracilis *contig 24. All exons in* D. melanogaster *aligned to contig 24.*

Start Codon

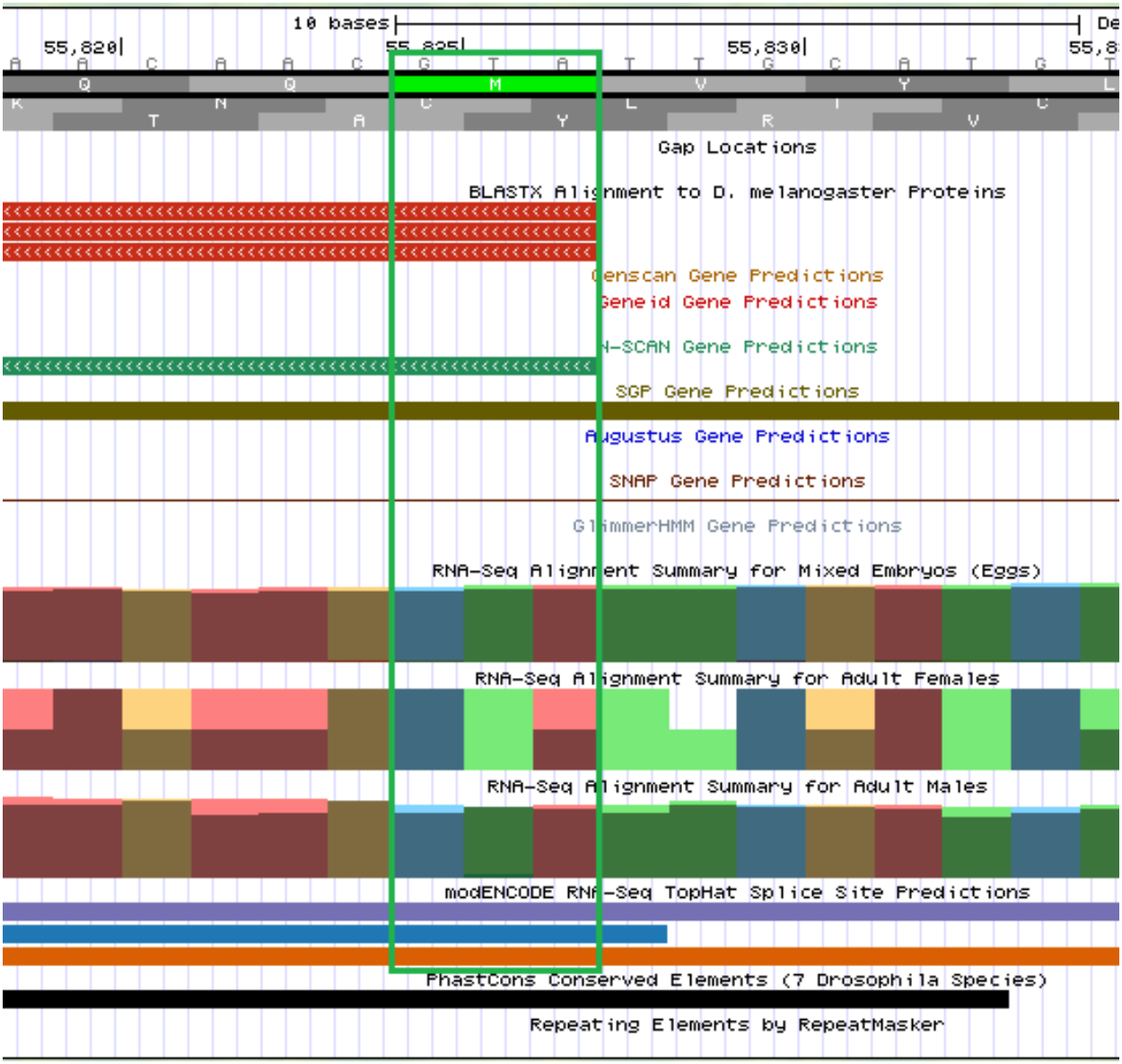


Figure 7: Proposed start codon. *The methionine at the position shown in the green box is in frame with the first CDS and agrees with the N-SCAN gene prediction. The black box in this and subsequent figures indicates the frame of the CDS (-1 in this case).*

A start codon is present at the beginning of the BLAST alignment to the first CDS starting at 55827 bp in frame -1 (Fig. 7). This is the translation start site for all isoforms of the gene. RNA-seq reads span upstream of the start codon, corresponding to the 5' UTR.

***eIF4G* Splice Sites**

The gene BLAST alignment track of the GEP mirror of the UCSC genome browser was used to identify where the *D. eugracilis* exons were likely to be based on the match to the *D. melanogaster* exons. The specific coordinates of the *D. eugracilis* exons were identified by using model-based gene predictions, RNA-seq coverage, and presence of TopHat Junctions to determine intron splice sites. Evidence to suggest the end of an exon are the end of gene model-predicted exons, sudden drops in the level of RNA-seq coverage, and TopHat junctions. Splice sites were found by identifying a splice donor sequence (GT or potentially GC) just after the end of one exon and a splice acceptor site (AG) just before the start of the next exon. When an intron is removed during mRNA processing, both splice sites are removed, directly connecting the bases in the exons that are adjacent to either splice site.

Splice sites do not necessarily occur after a complete codon. Splice sites can occur in phase 0, 1, or 2, corresponding to immediately after a complete codon, after one base following a codon, and after two bases following a codon, respectively. Corresponding splice acceptor and donor sites must not result in incomplete codons for the gene model to be valid. For this to be the case, the phase of the splice donor site and splice acceptor site must add up to zero or to three. For example, exon 10 has a GT sequence at 39845-39844 bp (Fig. 8A) directly next to where six model-based gene predictors, RNA-seq coverage, and TopHat all indicate a splice site. Given that exon 10 is in reading frame -1, if this is the splice donor site, the splice donor would be in phase 1 because splicing the RNA at this sequence would have left one base after the last complete codon. It is therefore required that the phase of the splice acceptor site on exon 11 is phase 2. Exon 11 contains an AG sequence at 39414-39413 bp. Seven model-based gene

predictors, RNA-seq reads, and TopHat support this position as a splice acceptor site. Additionally, this splice site is in phase 2 in frame -2, the frame that exon 11 is in (Fig. 8B). These sites were concluded to be a donor/acceptor pair.

Figure 8A: Tenth exon splice donor site. *This splice site is in phase 1 because the exon is in frame -1*

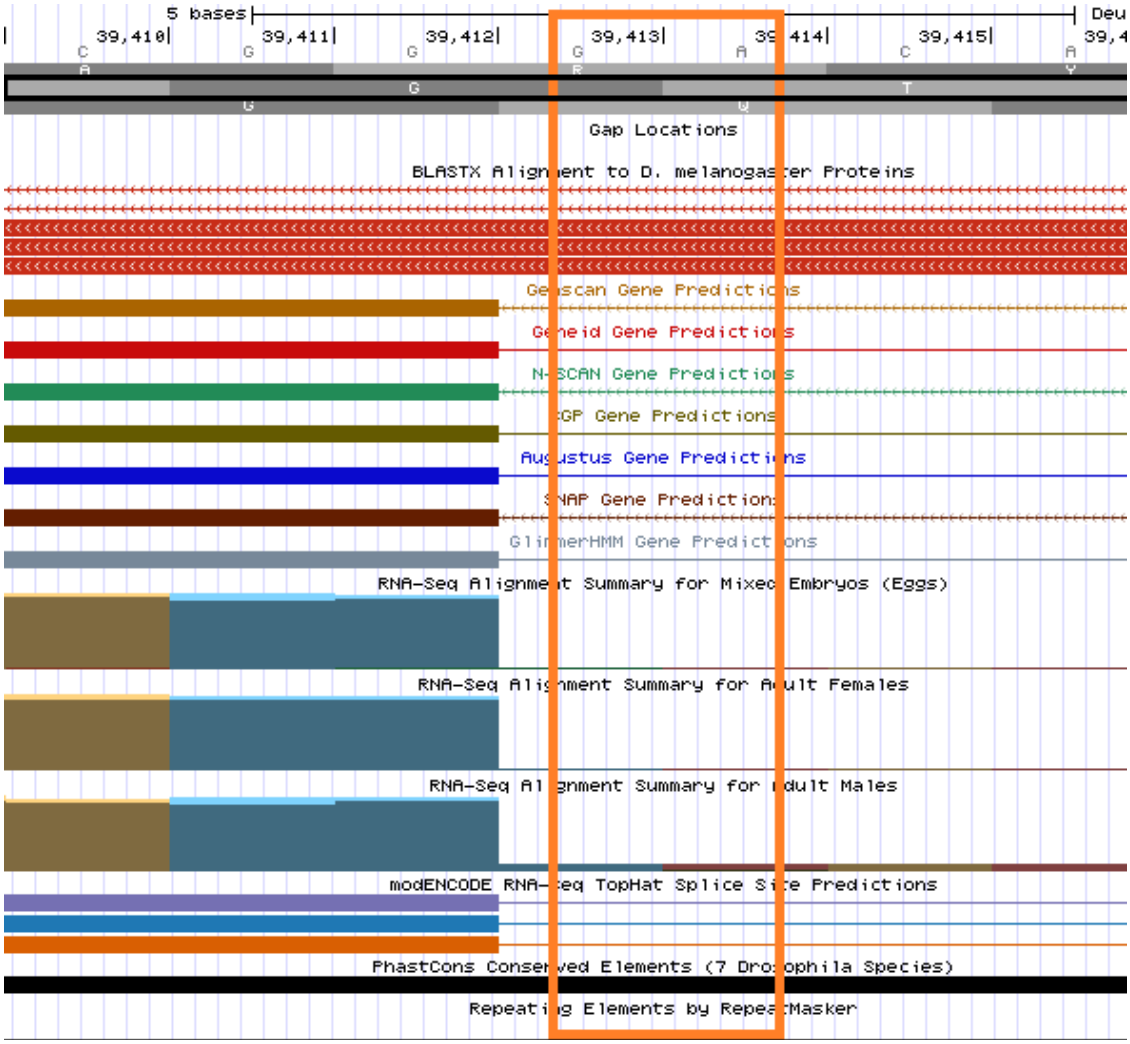


Figure 8B: Eleventh exon splice acceptor site. *This splice site is in phase 2 because the exon is in frame -2*

Five model-based gene predictors, RNA-seq coverage, and TopHat suggest that the AG sequence at 51746-51745 bp is the splice acceptor site for exon 4 (Fig. 9A). This proposed splice site is in phase 0, meaning that a potential splice donor must also be in phase 0. Looking at the end of exon 3, there is a GT sequence at 51803-51802 bp. However, the only evidence track that supports this splice site is the N-SCAN gene prediction (Fig. 9A). Furthermore, this splice site is in phase 1, not in phase with the identified splice acceptor site. RNA-seq data including 835

mixed-embryo reads as well as the TopHat junctions supports a splice donor site in phase 0 at 51807-51806 bp. This position contains the alternative splice donor sequence, GC. The sequence of this splice donor site in *D. melanogaster* was found, using GBrowse, to also be GC, lending further support to this being the splice donor site (Fig. 9C).

A stop codon (TAA) in frame with the fifteenth exon is present at 37378-37376 bp. This stop codon is supported by four model-based gene predictors. RNA-seq reads continuing downstream of this stop codon represent the 3' UTR. This stop codon also aligns with the stop codon in *D. melanogaster* (Fig. 10).

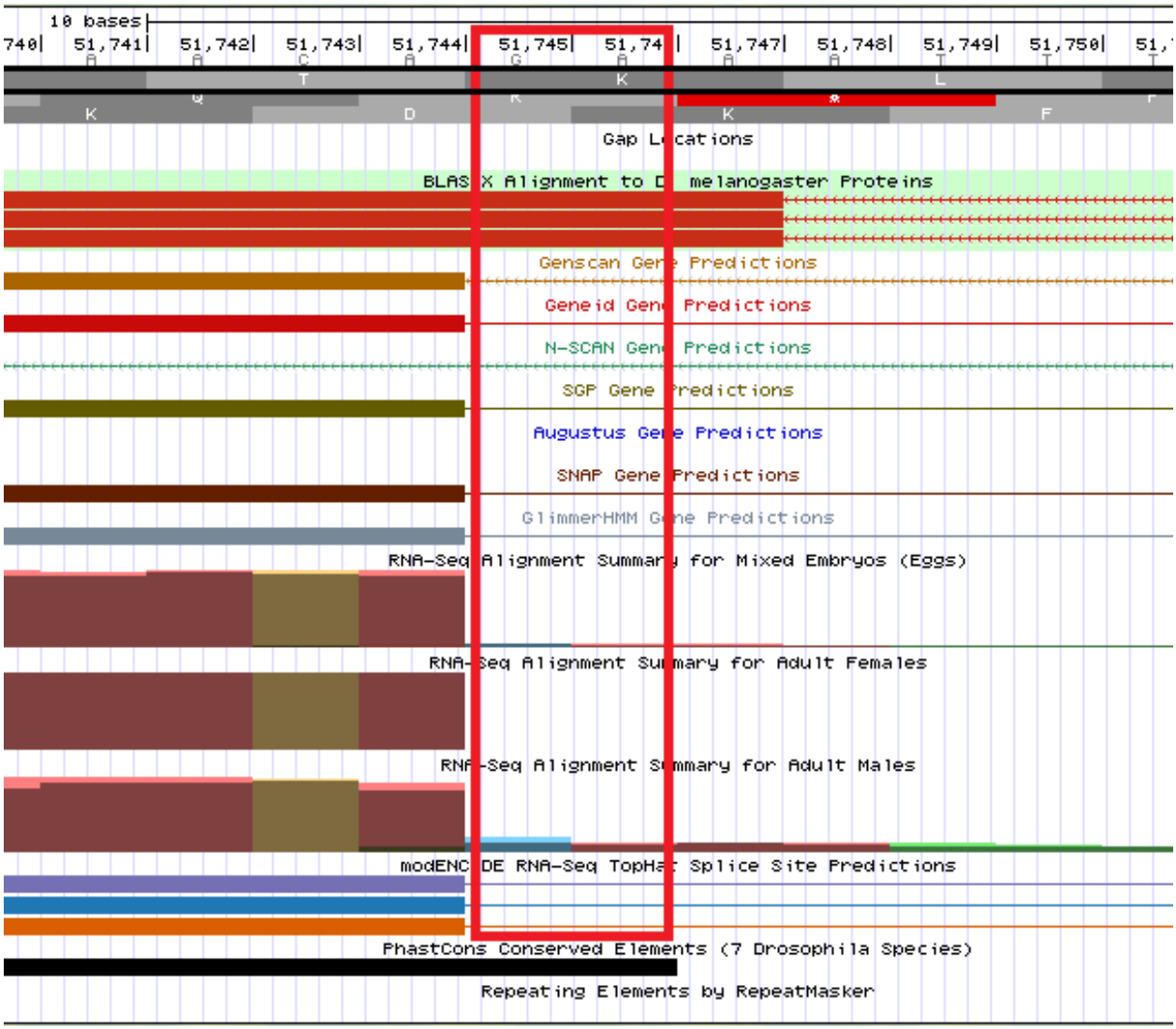


Figure 9A: Fourth exon splice acceptor site. *This splice site is in phase 0 because the exon is in frame -1*

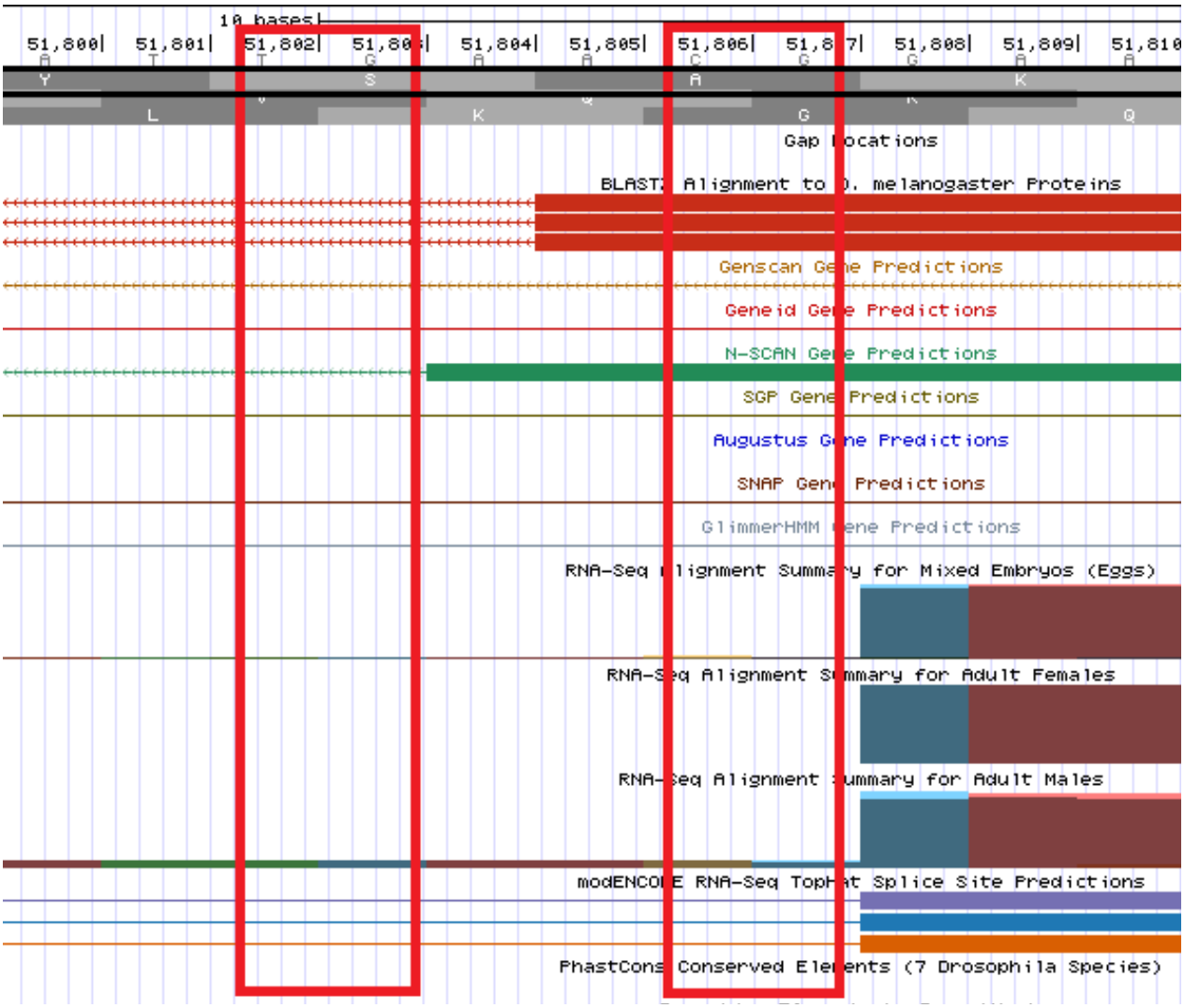


Figure 9B: Third exon splice donor site. *To agree with phase 0 splice acceptor site, the splice donor must be the alternative GC donor sequence*

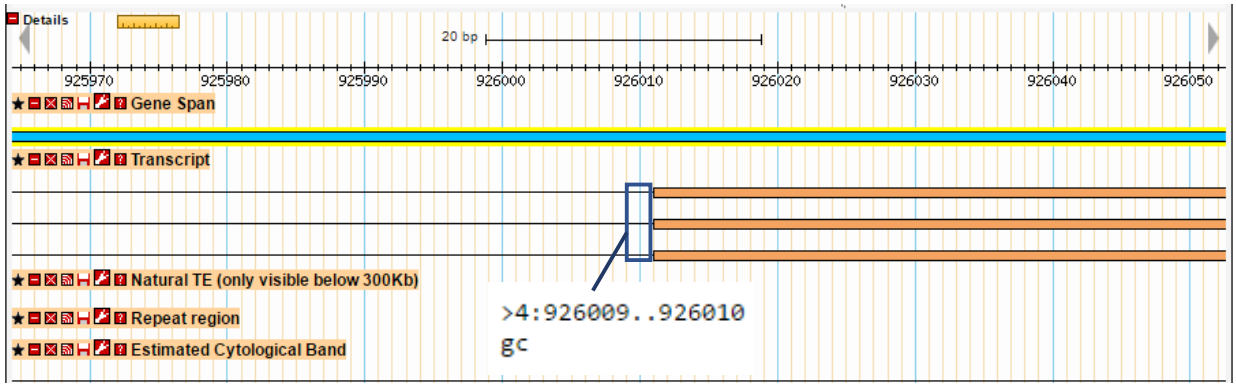
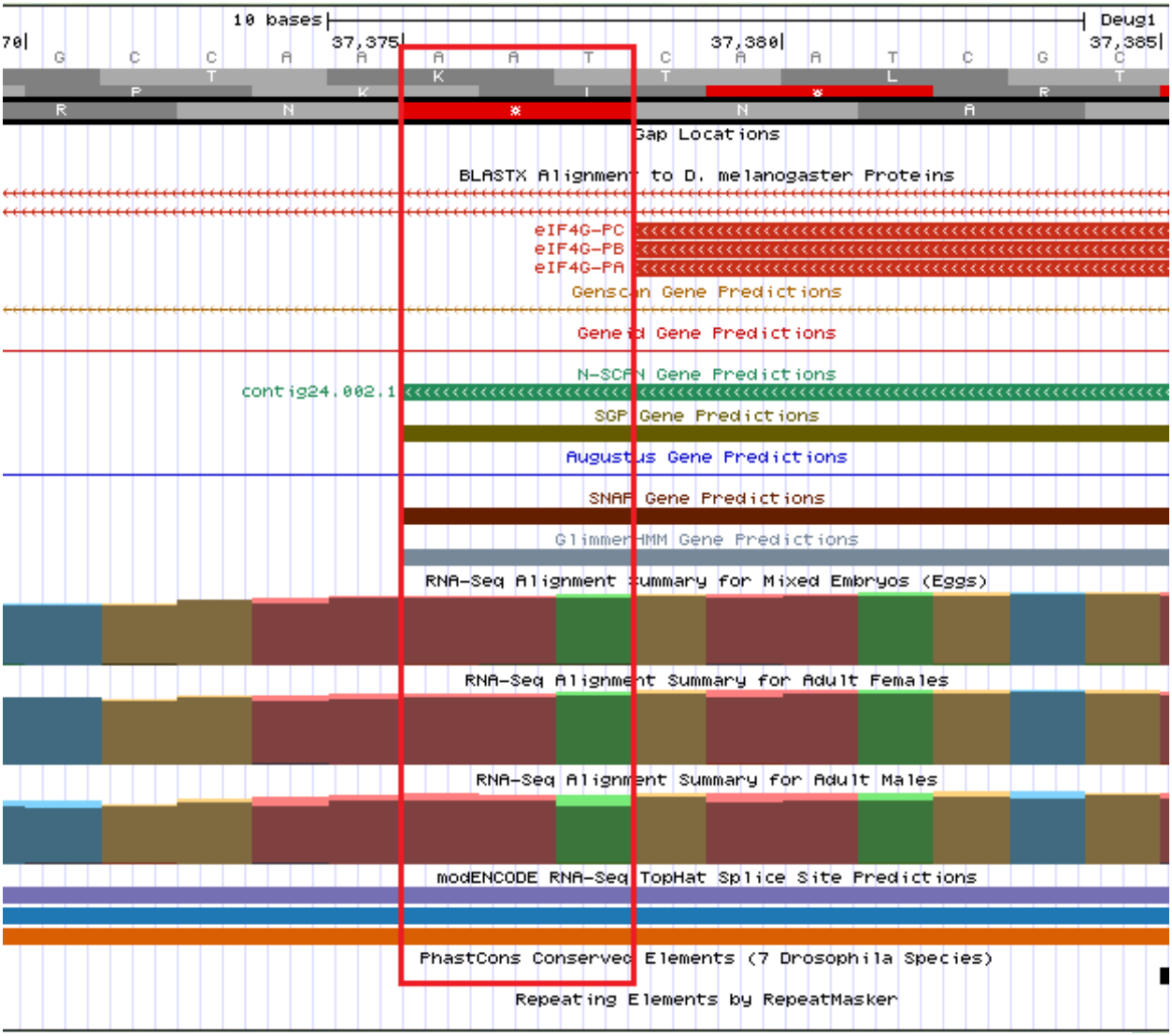


Figure 9C: Third exon splice donor site in* D. melanogaster*. The* D. melanogaster *sequence also contains the alternative GC splice donor sequence. Note that the reported sequence by GBrowse is in the + strand while the gene is in the – strand.



eIF4G:Dmel_exon_15

Sequence ID: Query_84067 **Length:** 20 **Number of Matches:** 5

Range 1: 1 to 20 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
42.0 bits(97)	7e-10	20/20(100%)	20/20(100%)	0/20(0%)	-3

```
Query  37435  VAVKSLNPFFNSLLNDDAN*  37376
              VAVKSLNPFFNSLLNDDAN*
Sbjct  1      VAVKSLNPFFNSLLNDDAN*  20
```

Figure 10: Proposed stop codon. *The stop codon at the position shown in the red box is in frame with the final CDS and agrees with the four gene prediction tracks. The final exon of* D. eugracilis *aligns perfectly with the* D. melanogaster *ortholog at the amino acid sequence level (see Fig. 6).*

Verification of *eIF4G* Gene Model

Flybase ID	Splice Acceptor	Splice Donor	Acceptor Phase	Donor Phase	Range	Frame
1_1894_0	Start Codon	55749-55748		0	55827-55750	-1
2_1894_0	53967-53966	53263-53262	0	0	53965-53264	-3
3_1894_0	52457-52456	51807-51806	0	0	52455-51808	-1
4_1894_0	51746-51745	51563-51562	0	2	51744-51564	-1
5_1894_2	50618-50617	50504-50503	1	1	50616-50505	-3
6_1894_1	50242-50241	50172-50171	2	1	50240-50173	-3
7_1894_2	44700-44699	42392-42391	2	0	44698-42393	-2
8_1894_0	41872-41871	41644-41643	0	1	41870-41645	-2
9_1894_2	41580-41579	41411-41410	2	0	41578-41412	-2
10_1894_0	40322-41321	39845-39844	0	1	40320-39846	-1
11_1894_2	39414-39413	39188-39187	2	0	39412-39189	-2
12_1894_0	38148-38147	37999-37998	0	0	38146-38000	-3
13_1894_0	37943-37942	37718-37717	0	1	37941-37719	-1
14_1894_2	37601-37600	37497-37496	2	1	37599-37498	-3
15_1894_2	37439-37438	Stop Codon	2		37437-37379	-3

Table 2: Proposed* D. eugracilis *exons and splice sites. *All exons passed Gene Model Checker plausibility criteria.*

Table 2 shows the proposed gene model for the *D. eugracilis* ortholog of *D. melanogaster eIF4G*. The second exon is only present in the B isoform, while all other exons are found in all isoforms. The phases of all corresponding splice donor and acceptor sites add up to three or zero. Biologically, introns must be at least 40 nucleotides long. No exon identified is

within 40 nucleotides of another, supporting the plausibility of this model. The plausibility of the splice sites was assessed using Gene Model Checker, which takes an input of the exon ranges and verifies that basic biological criteria are met for all proposed splice sites, the start codon, and the stop codon. The proposed exons all passed Gene Model Checker's criteria. Gene Model Checker also returns a dot plot and protein alignment of the orthologous *D. melanogaster* gene and the proposed *D. eugracilis* gene. Gene Model Checker was used to verify the gene model for both the B and A isoforms. The C isoform has an identical coding sequence to the A isoform and therefore is described by the same Gene Model Checker report as isoform A (Fig. 11).

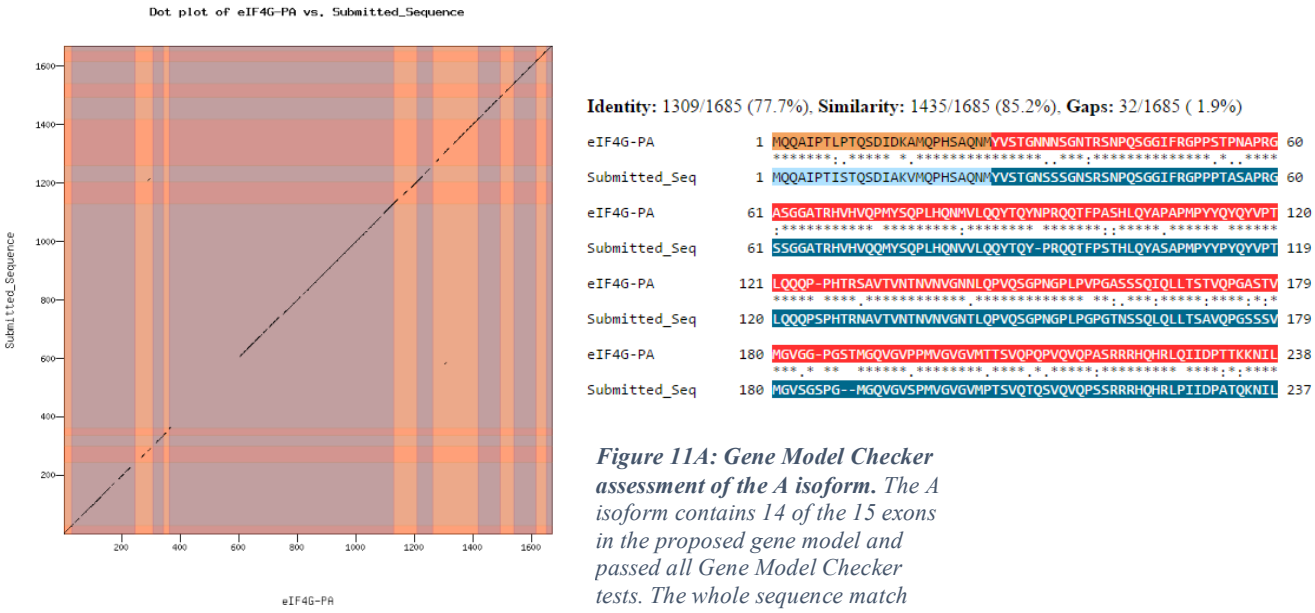


Figure 11A: Gene Model Checker assessment of the A isoform. *The A isoform contains 14 of the 15 exons in the proposed gene model and passed all Gene Model Checker tests. The whole sequence match can be found in the supplemental figures (S2).*

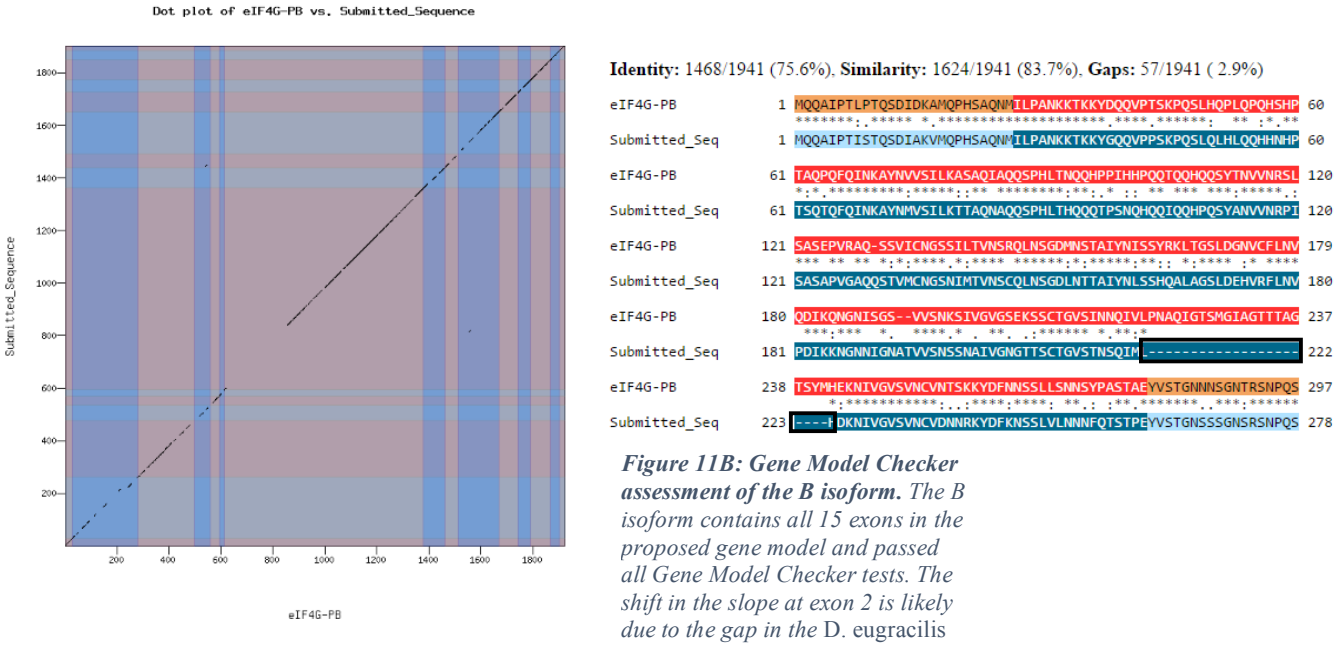


Figure 11B: Gene Model Checker assessment of the B isoform. *The B isoform contains all 15 exons in the proposed gene model and passed all Gene Model Checker tests. The shift in the slope at exon 2 is likely due to the gap in the* D. eugracilis *exon 2 shown in the black box above. Full protein alignment can be found in supplemental figures (S3)*

Approximately the first third of exon 7 in the B isoform (sixth exon in the A and C isoforms) represents a region of low homology between *D. eugracilis* and *D. melanogaster*, as shown in Figure 11B by a break in the alignment. Figure 12 shows this region in the UCSC Genome Browser along with the *Drosophila* conservation track. This track shows that the lack of homology between *D. melanogaster* and *D. eugracilis* is in a region of generally low homology among *Drosophila* species. The second exon (only present in the B isoform) shows several areas of low homology along with a gap (shown in a black box in Fig. 11B) in the *D. eugracilis* gene, represented by a shift over (rightward) of the dot plot alignment. Figure 13 shows that exon 2 of the B isoform of *eIF4G* contains multiple regions that show low conservation and one region with no conservation identified in the conservation track of the UCSC Genome Browser, consistent with the fragmented dot plot.

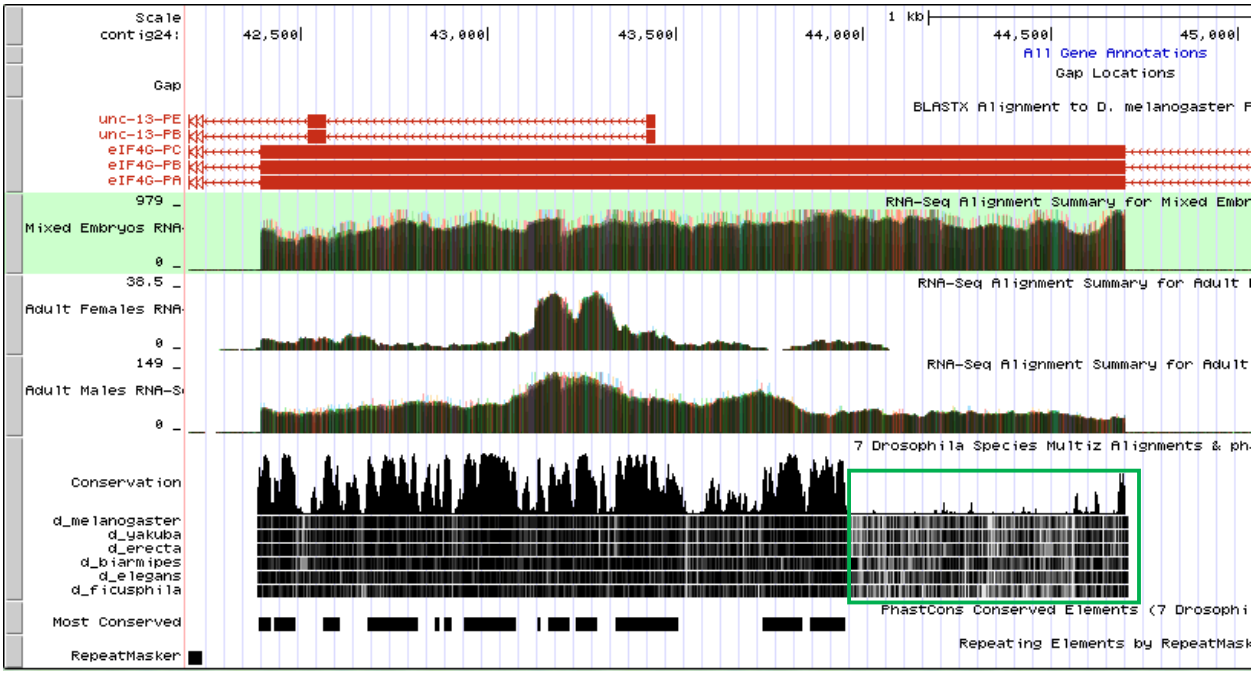


***Figure 12: UCSC Genome Browser view of* eIF4G *exon 7 in* D. eugracilis.** *The region shown in the green box is less conserved across* Drosophila *species than the rest of the exon.*

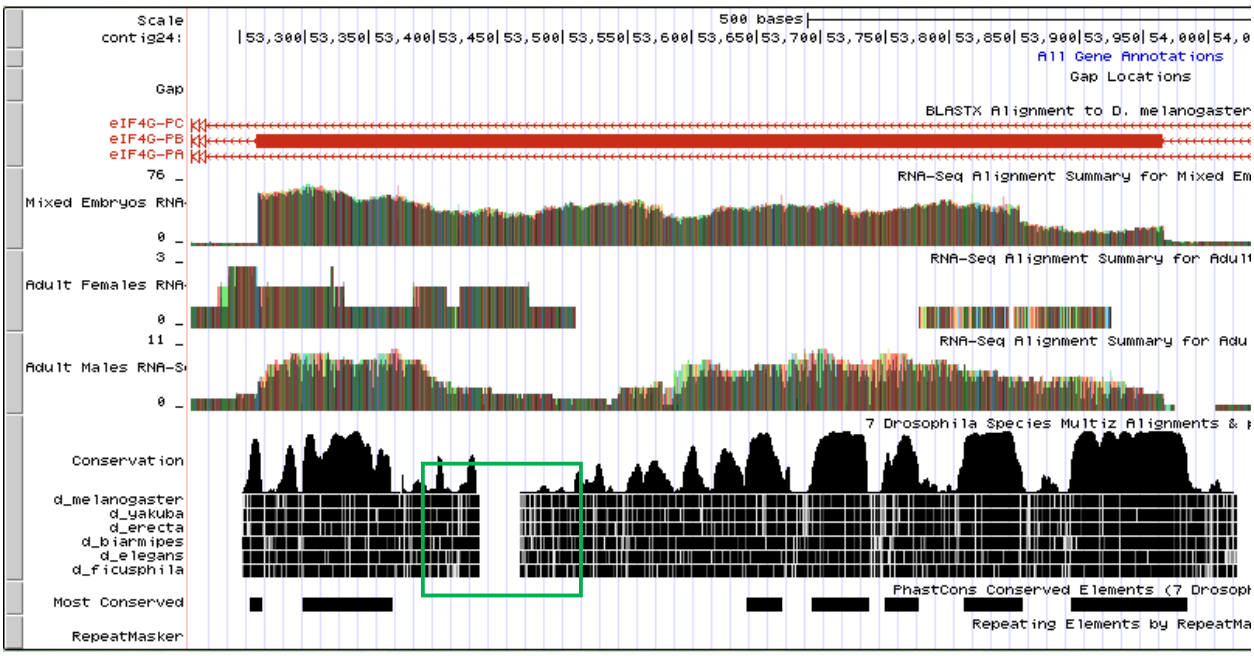


***Figure 13: UCSC Genome Browser view of* eIF4G-PB *exon 2 in* D. eugracilis.** *The region shown in the green box is not conserved across* Drosophila *species.*

eIF4G Transcription Start Site

One potential reason genes can be expressed on the heterochromatin-rich F element is unique core promoter motif patterns surrounding the transcription start sites (TSSs) of F element genes shared across *Drosophila* species. Using *D. melanogaster* TSS annotations and *D. biarmipes* RNA-polymerase II Chip-seq as references, the TSSs of the *D. eugracilis* F element are being annotated for this project. This project will add to the current base of knowledge for future comparative genomics analyses of *Drosophila* TSSs.

All three isoforms of *eIF4G* likely share the same TSS because all 5' UTRs begin at the same position. Isoforms A and B in *D. melanogaster* have identical 5' UTRs while the first 5' UTR exon of the C isoform is longer than the other two isoforms (Fig. 14).

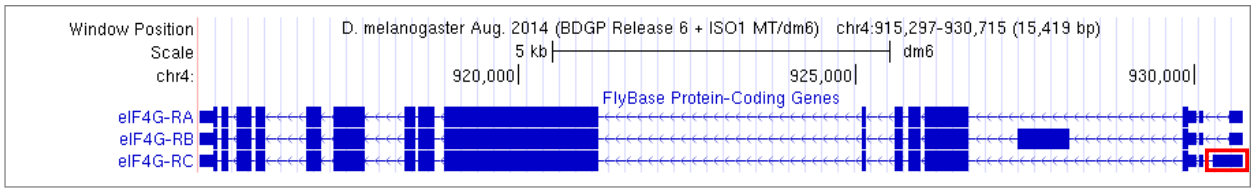


Figure 64: Gene Record Finder view of D. melanogaster eIF4G. *The 5' UTR of all three isoforms (shown in the box) start at the same position.*

The UCSC Genome Browser GEP Mirror contains DNase I hypersensitivity site (DHS) and Celniker TSS annotation tracks for *D. melanogaster*. Comparison between these data for *D. melanogaster eIF4G* can be used as evidence for locating the TSS of *D. eugracilis eIF4G*. Figure 15 shows the UCSC Genome Browser overview of *D. melanogaster eIF4G* and Figure 16 shows a zoomed in view of the region around the TSS. The sensitivity to DNase I is tracked in three cell lines and the DHS peaks are marked in the S2 and Kc lines. The Celniker TSS annotation track shows two TSSs immediately upstream of the first 5' UTR exon. A promoter can be described as peaked, intermediate, or broad based on the number of DHS peaks and the number of annotated TSSs. This region contains one DHS peak and two annotated TSSs, classifying it as an intermediate promoter. RAMPAGE data uses recorded RNA expression to locate a TSS and is

useful in supporting the classification of the core promoter. The CAGE and RAMPAGE data for *D. melanogaster eIF4G* (Fig. 17) show several spread-out peaks rather than a peak at one or two bases, further supporting an intermediate promoter.

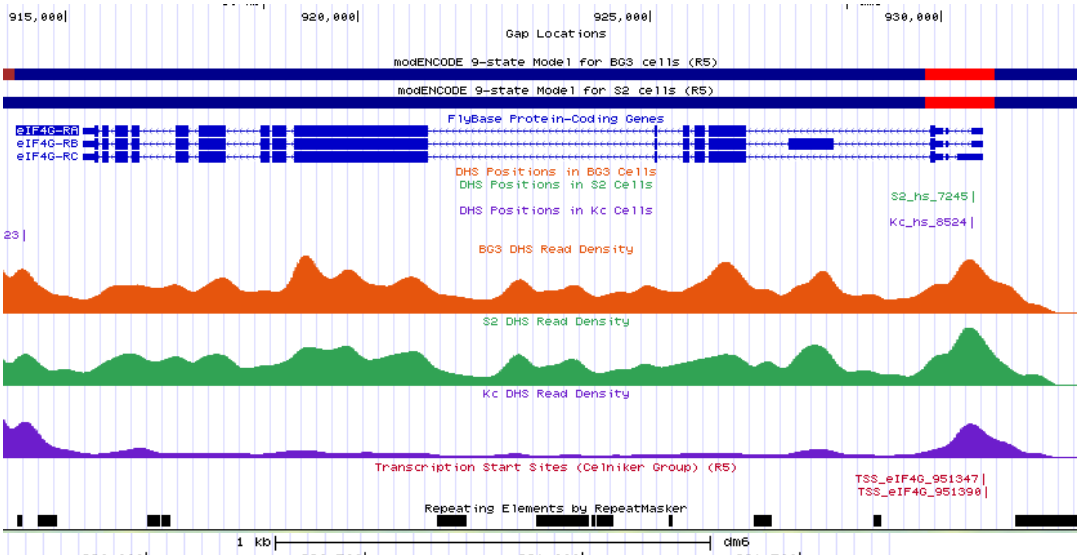


Figure 15:* D. melanogaster eIF4G *DHS overview. *Known DHS positions were obtained experimentally from three cell lines.*

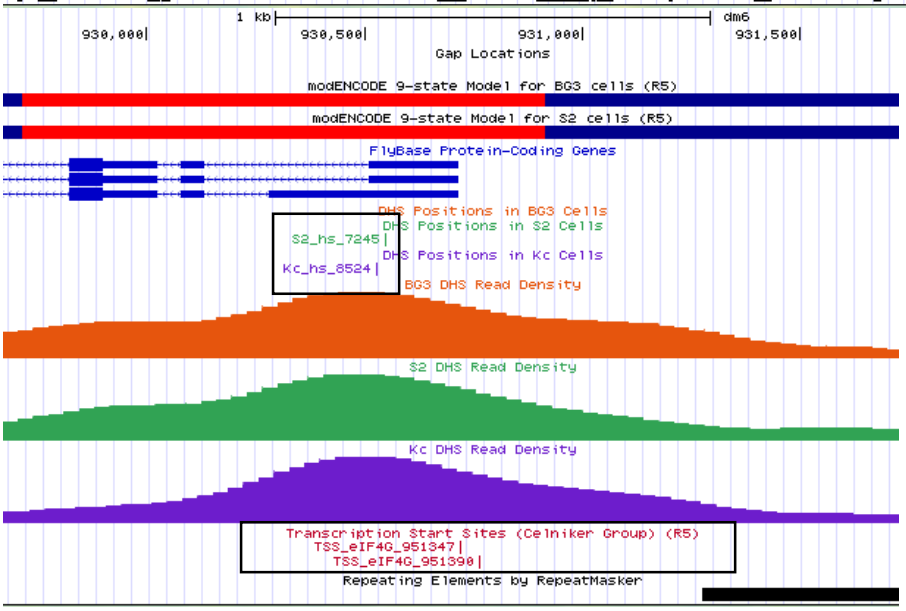


***Figure 16: DHS peaks and annotated TSSs of* D. melanogaster eIF4G.** *The read density shows one peak in DHS while having two celniker-identified TSSs (black boxes), classifying this as an intermediate*

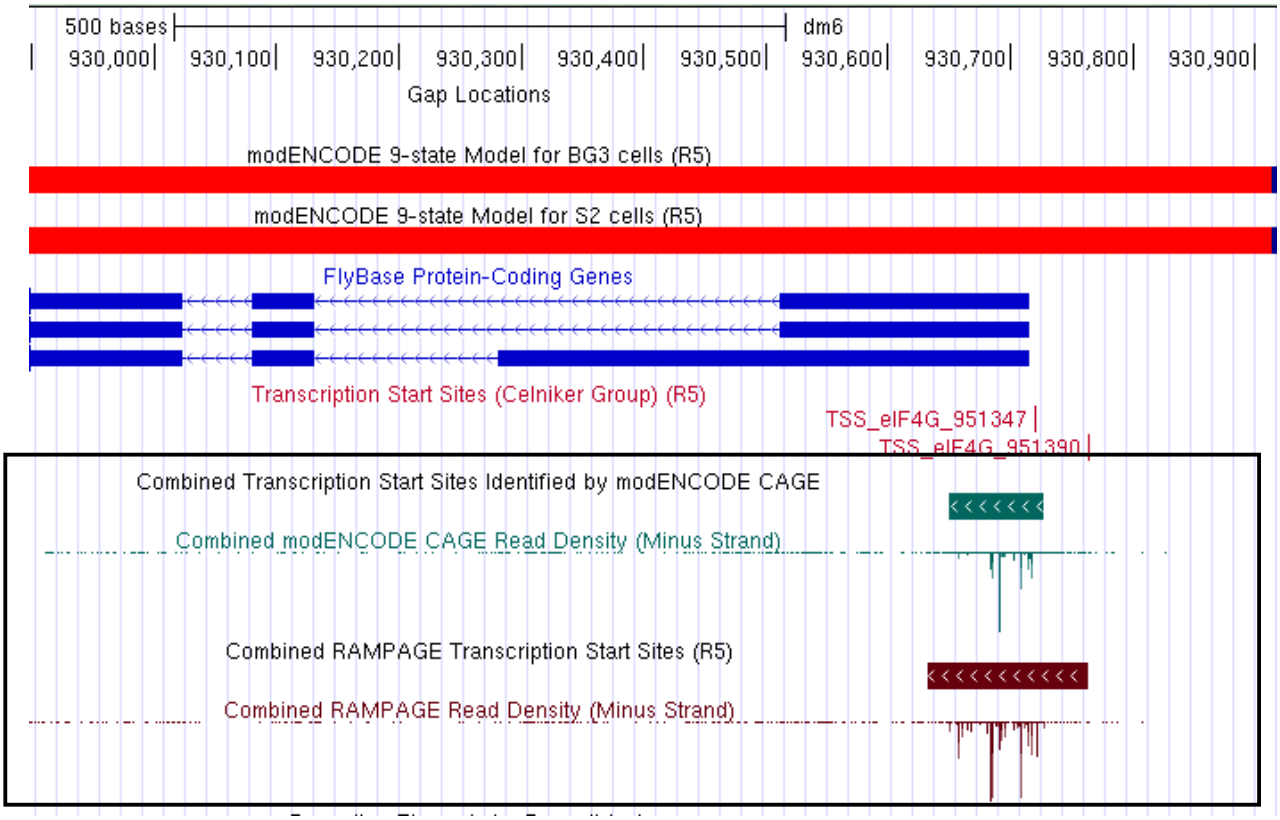


Figure 17: CAGE and RAMPAGE data. *The CAGE and RAMPAGE data (black box) are spread out rather than having a single peak, while still having several distinct peaks, supporting an intermediate promoter.*

The TSSs annotated in *D. melanogaster* are located immediately upstream of the 5' UTR of *eIF4G.* To locate the corresponding region in *D. eugracilis*, the DNA sequence of the first 5' UTR exon of the C isoform of *D. melanogaster eIF4G* (Fig. 18) was obtained from Gene Record Finder. The sequence was aligned to *D. eugracilis* contig 24 using blastn (Fig. 19). Because of the relatively low levels of conservation in UTRs compared to coding sequences, the parameters of blastn were adjusted to allow for matches of less similar sequences. "Match/Mismatch Scores" were set to "1, -1," "Gap Costs" were set to "Existance: 2 Extension: 1," and the "Low complexity regions" filter was turned off. Out of the 434 bp sequence of the 5' UTR exon, a 433 bp long alignment was obtained. The sequences matched with 68% identity. The first base in the *D. melanogaster* subject corresponds to position 56686 in *D. eugracilis* contig 24, making this position a candidate for the *D. eugracilis* TSS.

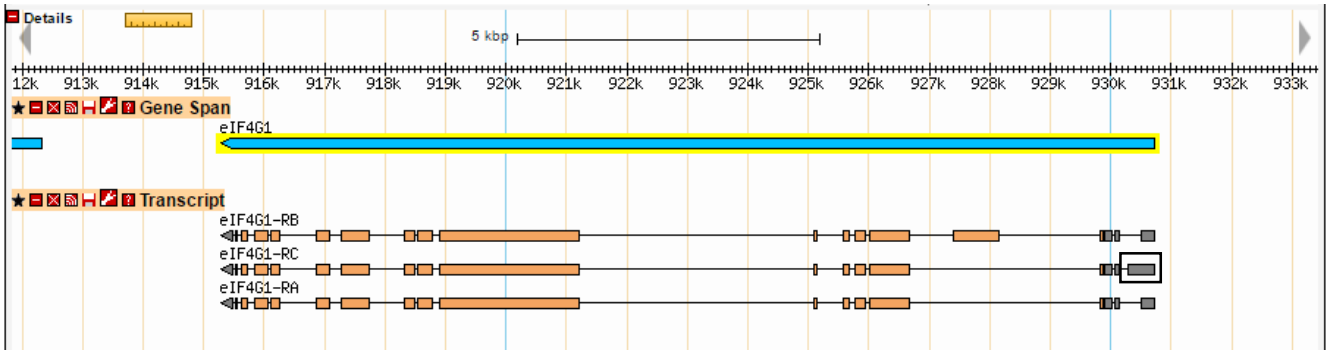


Figure 18: GBrowse view of* D. melanogaster *eIF4G. *The boxed 5'UTR exon in the C isoform was used for blastn alignment to D. eugracilis.*

Range 1: 1 to 433 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
199 bits(138)	3e-53	312/460(68%)	53/460(11%)	Plus/Minus

```
Query  56253  TATTTGCCGACGTTGCTGTTGTGACGATAGACGTACATATGTATAATGTGCAAACGCAAA  56312
Sbjct  433    TATTTGCCGACGTTGCCGTTGTGAC------------TTTGGAGA--TTTCGTTCTCACA  388

Query  56313  ATCGCAATCGCAAAGCGCAGACGGAGAAAACGTGTCACCGAGTCCACCGTCTGCTTATAT  56372
Sbjct  387    ATCACAAT-----GGTGCATATGGTGAAAACGTGCCACGAATTCCATCGTTTACAT-CAT  334

Query  56373  TTTTGAAGGATTCCCAAGAATGATAACAACGTGTT-GTCGTTTAATTATATTT--TTTCT  56429
Sbjct  333    TTTTAAAAAGTT---AAGACACATTTCAGAATGTTAGTGGTAAAATCTTTTTTCCCTTCA  277

Query  56430  AACTGCGAAATCGACTTGCAGAAAGTCATGTAA----TGTTTCCCAGTCCTCATTAAAGA  56485
Sbjct  276    CTGTGCACAGTAAACTTGCAAAGTATCATGTAATATTTGTTGCCCAGTGCTCATTATTAA  217

Query  56486  TCTCCAAATTACCTGGACA--------TATGTGCCTCAAAGATGCGATTATTTACAAAAA  56537
Sbjct  216    CCATCAATTTACCTGAGCAAATGTTTGTATGTGCCTATCCGACGTGATTATGTACAAAA-  158

Query  56538  CAATAGGACTGTTAAAGATTCCACTTTAAGTGTATTTTCACA--GCA---------ATAA  56586
Sbjct  157    -AATACAAATGTTAAGAATTCCACTTTAAGTGTATTTTAAAAATGTATTTATTTATAAAA  99

Query  56587  ATGAAGCAGCACTTTGAGCTTTAAATCTTTACAATCAAAAACTTTTTTTTGTTAATTTCC  56646
Sbjct  98     ATGAAGCAGTAATTCCAGCTTT--ATCTTTACATTCAAAAACTATTTTTTAATAATTTCC  41

Query  56647  ACATGCAATGTGTTCTGTTCATAGGCACAGAGTTGCATTT  56686
Sbjct  40     ACATGCAATGTGTTCTGTTCATACGAACAGAGTTGCATTT  1
```

Figure 19: blastn alignment of* D. melanogaster eIF4G *C isoform first 5' UTR exon (subject) and* D. eugracilis *contig 24 (query). *Out of 434 bp, 433 bp aligned to* D. eugracilis. D. melanogaster *was the subject and* D. eugracilis *contig 24 was the query.*

D. biarmipes is a more closely related species to *D. eugracilis* than *D. melanogaster*. The UCSC genome browser contains a track for RNA polymerase II ChIP-sequencing (RNApolII ChIP-seq) in *D. biarmipes*. This track represents physical evidence of localization of RNApolII to a promoter. Figure 6 shows two RNApolII ChIP-seq peaks near the end of the 5' UTR of *D.*

biarmipes eIF4G. The further upstream peak does not show any noticeable sequence similarity to *D. eugracilis*, however the region surrounding the first peak (Fig. 20) shows a high level of homology. The sequence for the region shown in Figure 20 was obtained from the UCSC genome browser and aligned to *D. eugracilis* contig 24 using the same blastn parameters as the previous BLAST search (Fig. 21). The alignment shows that the position corresponding to the RNApolII ChIP-seq peak in *D. biarmipes* is 56447 bp in *D. eurgacilis* contig 24. This position is another potential TSS in *D. eugracilis*. However, due to RNA-seq peaks upstream of this position, this is unlikely to be a possible TSS.

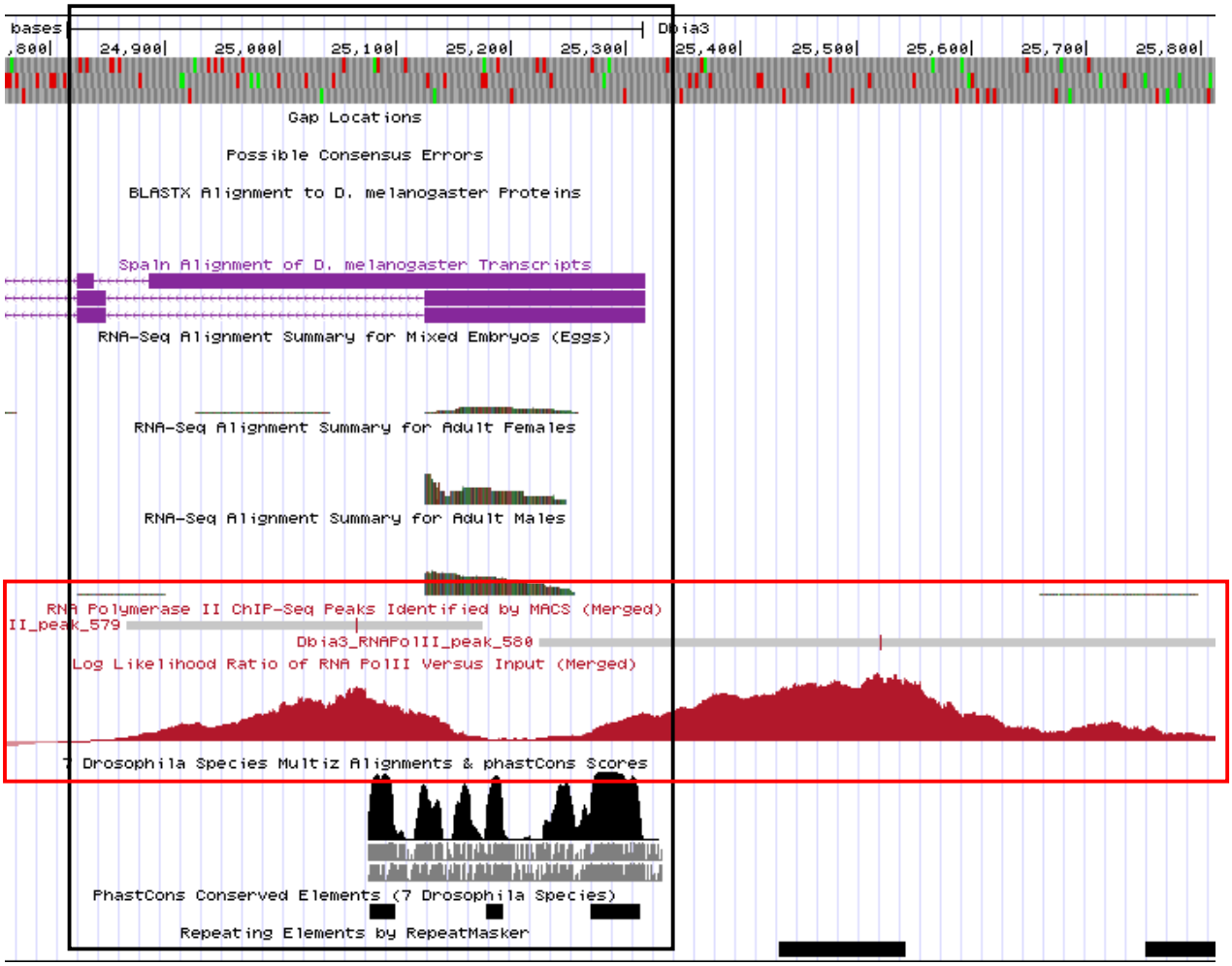


Figure 20:* D. biarmipes *RNA PolII ChIP-seq data. *The peaks in the red box indicate where RNA-pol is binding, functional evidence of a TSS. The region in the black box shows the region in* D. biarmipes *that was used as the subject in a blastn alignment against* D. eugracilis*. The righthand peak is within a repeat-dense region and did not align to* D. eugracilis *contig 24.*

Dbia3_dna range=contig51:24867-25362 5'pad=0 3'pad=0 strand=+ repeatMasking=none
Sequence ID: Query_122371 Length: 496 Number of Matches: 24

Range 1: 1 to 468 Graphics ▼ Next Match ▲ Previous Match

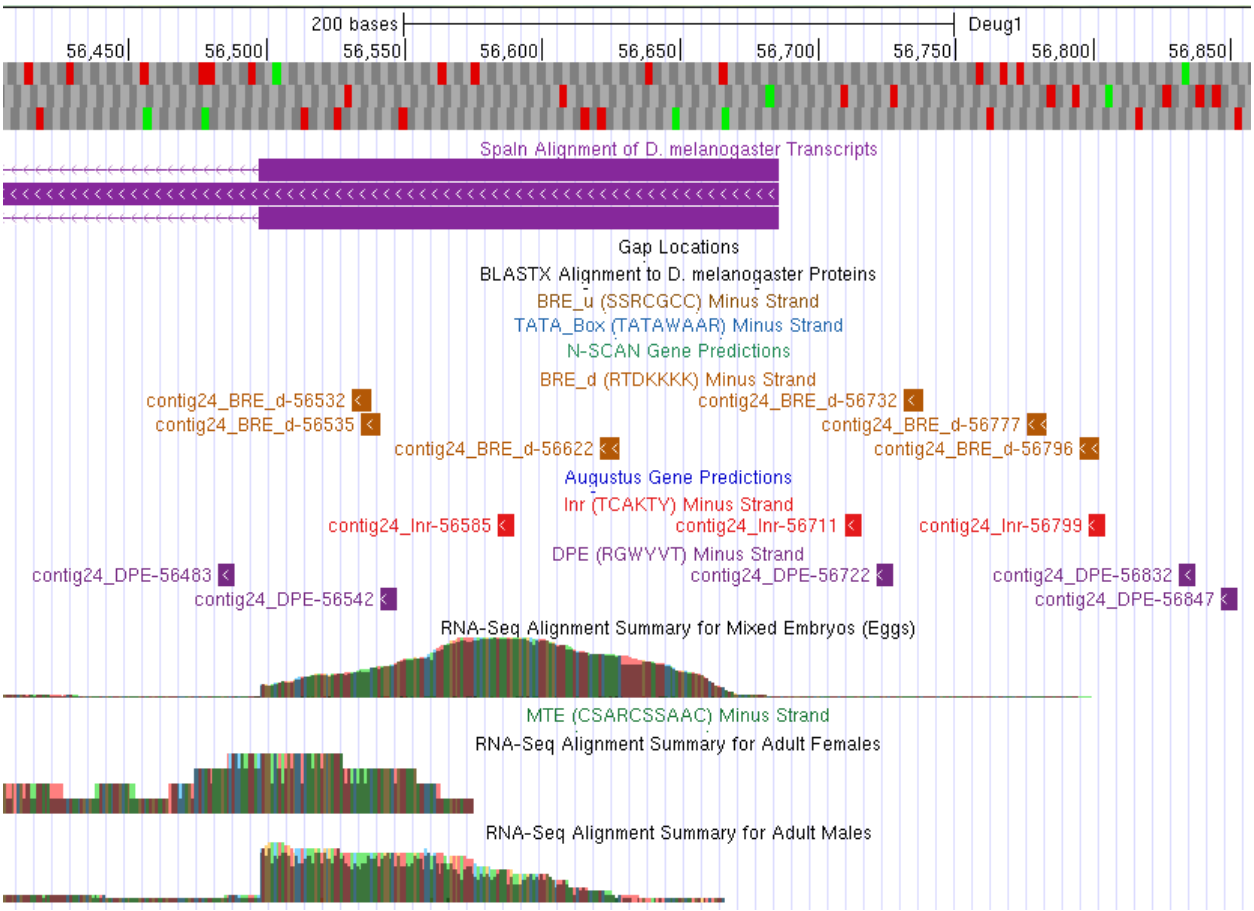
Score	Expect	Identities	Gaps	Strand
268 bits(186)	5e-74	349/488(72%)	38/488(7%)	Plus/Plus

```
Query  56232  ATTAAGCCTGTTAACTTTACATATTTGCCGACGTTGCTGTTGTGACGATAGACGTACATA  56291
              || |||||| ||| |||| ||||||||||||||||||| |||| ||| | |||| | |
Sbjct  1      ATCAAGCCTTTTAGCTTTGCATATTTGCCGACGTTGCTGCTGTGCAGATGGCCGTAAA-A  59

Query  56292  TGTATAATGTGCAAACGCAAAATCGCAAT-CGCAAAGCGCAGACGGAGAAAACGTGTCAC  56350
              ||| ||  | ||||| |||||     || | |||       | || ||||||  ||  |
Sbjct  60     TGTGCAAGATTTAAACGTAAAATTAATATTCACAAAT-----AAGGTGAAAACATGCAAA  114

Query  56351  CGAGTCCACCGTCTGCTTATATTTTTGAAGGATTCCCAAGAATGATAACAACGTGTTGTC  56410
              | ||  | | |||| ||   ||| | ||  || | | |||  ||||||||||
Sbjct  115    TGTGTTTATCATCTGCGTAA--TTTCGGAG--TTGCTACGAATATTAACAACGTGTCAAT  170

Query  56411  GTTTAATTATATTTTTTCTAACTGCGAAATCGACTTGCAGAAAG-----TCATGTAATGT  56465
              ||||    || ||       | |  || || | ||||  |||     |||||||| | |
Sbjct  171    GTTTTCCCTTAATT-------CGGTAAAGTCAATTTGCACCAAGAGTCATCATGTAATAT  223

Query  56466  TT----CCCAGTCCTCATTAAAGATCTCCAAATTACCTGGACATATGTGCCTCAAAGATG  56521
              ||    |||||||| |||||||   | |||||| |||||||  |||||||||| ||  ||
Sbjct  224    TTGTTGCCCAGTCTTCATTAACACTTTCCAATTTACCTGAGCATATGTGCCTTAACAATC  283

Query  56522  CGATTATTTACAAAAACAATAGGACTGTTAAAGATTCCACTTTAAGTGTATTTTCA-CAG  56580
              ||||||| ||||||||   |  ||||||||| |||| ||||||||||||||| | |||
Sbjct  284    CGATTATATACAAAAA---TTCGACTGTTAATGATTTCACTTTAAGTGTATTTTAAACAG  340

Query  56581  CAATAAATGAAGCAGCACTTTGAGCTTTAAATCTTTACAATCAAAAAC-----TTTTTTT  56635
                ||| || |||  | | ||  |||||| |||| |||||| ||||||      || ||||
Sbjct  341    TTATATATCAAGGCGTAATTCTAGCTTTTAATCATTACAACCAAAAAAACAATTTTCTTTT  400

Query  56636  TGTTAATTTCCACATGCAATGTGTTCTGTTCATAGGCACAGAGTTGCATTTCTTGT--GG  56693
              |||||||||||||||||||||||||||| |||||||||||||||| || ||  |  ||
Sbjct  401    TGTTAATTTCCACATGCAATGTGTTCTGTCCATAGGCACAGAGTTGCAGTTGTTATTCGA  460

Query  56694  TATATCGA  56701
              ||||||||
Sbjct  461    TATATCGA  468
```

***Figure 21: blastn alignment of* D. biarmipes *and* D. eugracilis.** *The boxed 'G' is the position of the ChIP-seq RNA polII peak in* D. biarmipes*, corresponding to 56447 bp in* D. eugracilis*. D. biarmipes is the subject and* D. eugracilis *contig 24 is the query. The peak in RNA-seq does not necessarily indicate the exact position of the TSS because RNApolII is distributed across an active gene.*

Transcription of genes requires binding of transcription factors at promoters, and therefore protein-binding DNA motifs can often be found throughout promoters. Figure 22 shows the locations of all *Drosophila* core promoter motifs surrounding the first RNA-seq data for *eIF4G* in *D. eugracilis*. In *Drosophila*, core promoter elements are not found at every TSS. In the region shown in Figure 22, only the Initiator (Inr), Downstream Promoter Element (DPE) and Downstream B Recognition Element (BRE[d]) motifs are found. Table 1 shows the location of core promoter elements in *D. eugracilis* and in the corresponding region in *D. melanogaster*. There is no clear conservation in these elements and none of these motifs together suggest the

same TSS, indicating that these motifs are not required for a functional TSS. TopHat junctions end just downstream of the first RNA-seq data for *D. eugracilis eIF4G*, supporting the notion



that this region contains the TSS (Fig. 23).

***Figure 22: Core promoter motifs in* D. eugracilis eIF4G.** *While many motifs can be found, none can clearly be attributed to a distinct TSS due to lack of conservation with* D. melanogaster. The red line indicates position 56686.

Core promoter motif	*D. eugracilis*	*D. melanogaster*
BRE^{u}	NA	NA
TATA Box	NA	NA
BRE^{d}	-56532, -56535, -56622, -56732, -56777, -56796	-930651, -930654
Inr	-56585, -56711, -56799	-930754
MTE	NA	NA
DPE	-56483, -56542, -56722	-930752
Ohler_motif1	NA	NA
DRE	NA	NA
Ohler_motif5	NA	NA

***Table 3: Core promoter motif locations in the putitive TSS region of* D. eugracilis *and surrounding the TSS of* D. melanogaster.** *No two of the motifs found supported the same TSS position, and it is unlikely that core promoter motifs play a role in the TSS.*

Ohler_motif6	NA	NA
Ohler_motif7	NA	NA
Ohler_motif8	NA	NA

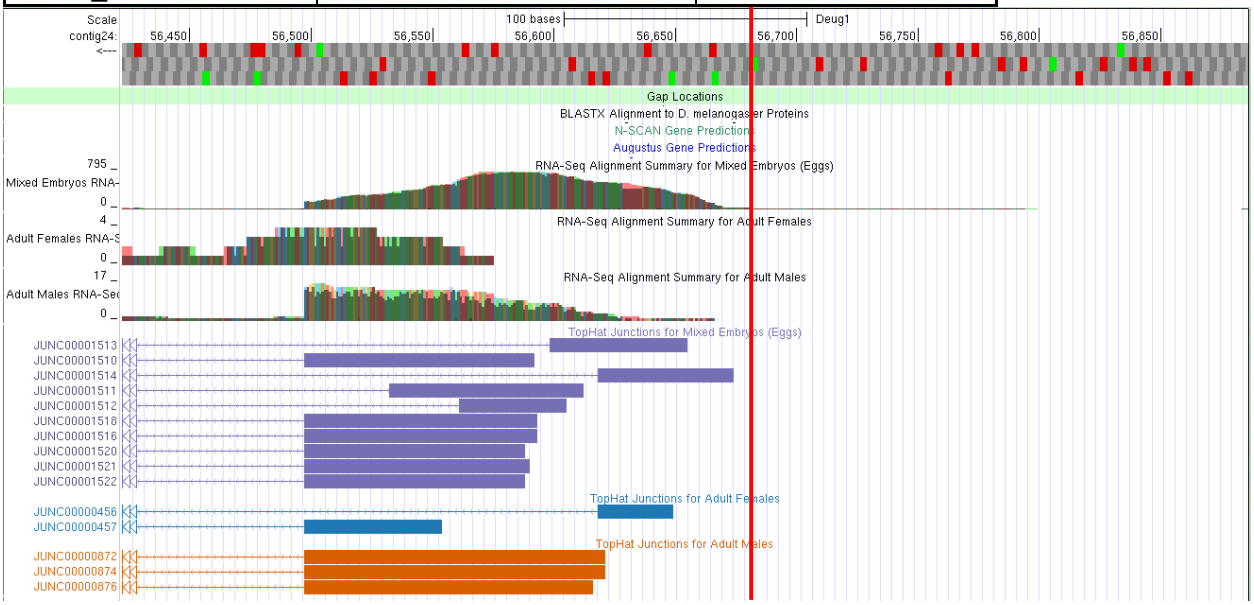


***Figure 23: TopHat tracks for* D. eugracilis.** *No TopHat junctions are found past the end of the RNA-seq data, supporting this region as containing the TSS. The red line indicates position 56686. RNA-seq data begins at position 56684.*

Evidence for the location of the TSS of *D. eugracilis eIF4G* obtained from *D. melanogaster* Celniker TSS annotation and BLAST alignment place the TSS near position 56686

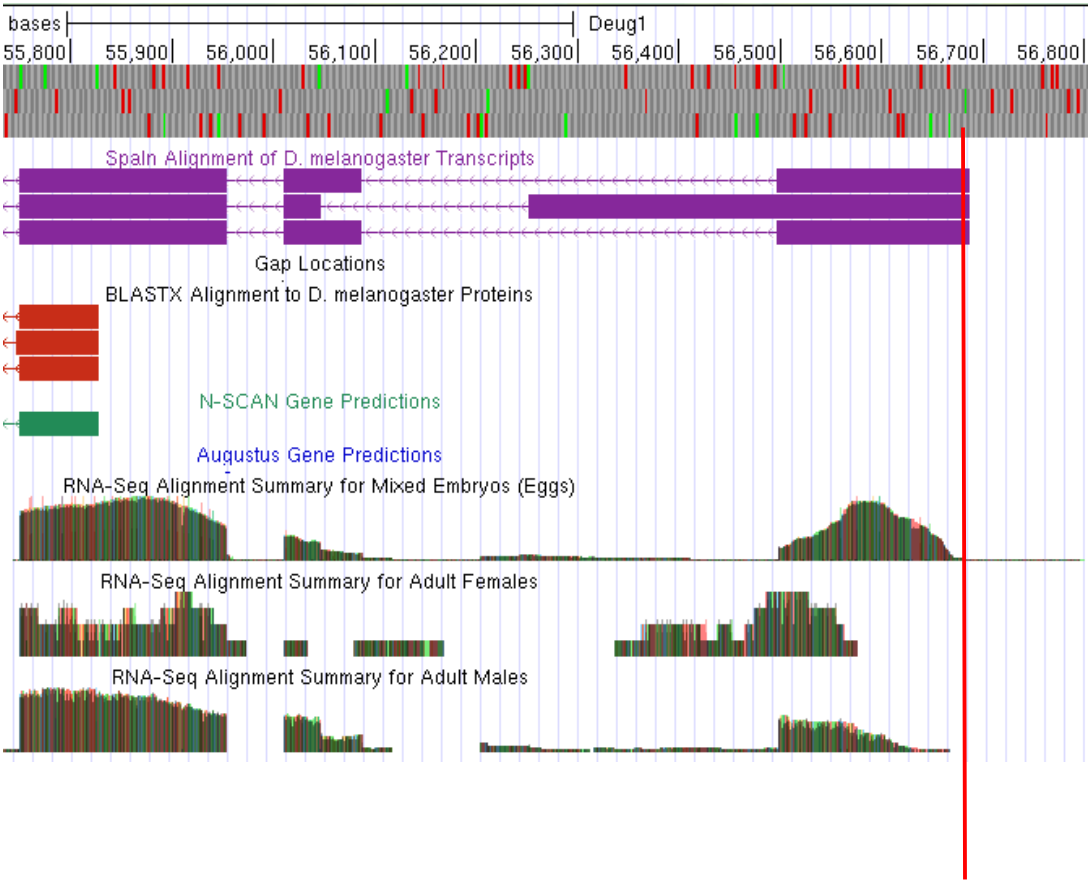


Figure 24: Proposed TSS in* D. eugracilis *is shown by the red line. *The red line indicates position 56686, which aligns closely to the end of the RNA-seq coverage.*

in contig 24, which is consistent with RNA-seq reads which begin just after this position. The peak in *D. biarmipes* RNApolII ChIP-seq corresponds to position 56447 in *D. eugracilis.*

D. eugracilis Ortholog of *unc-13*

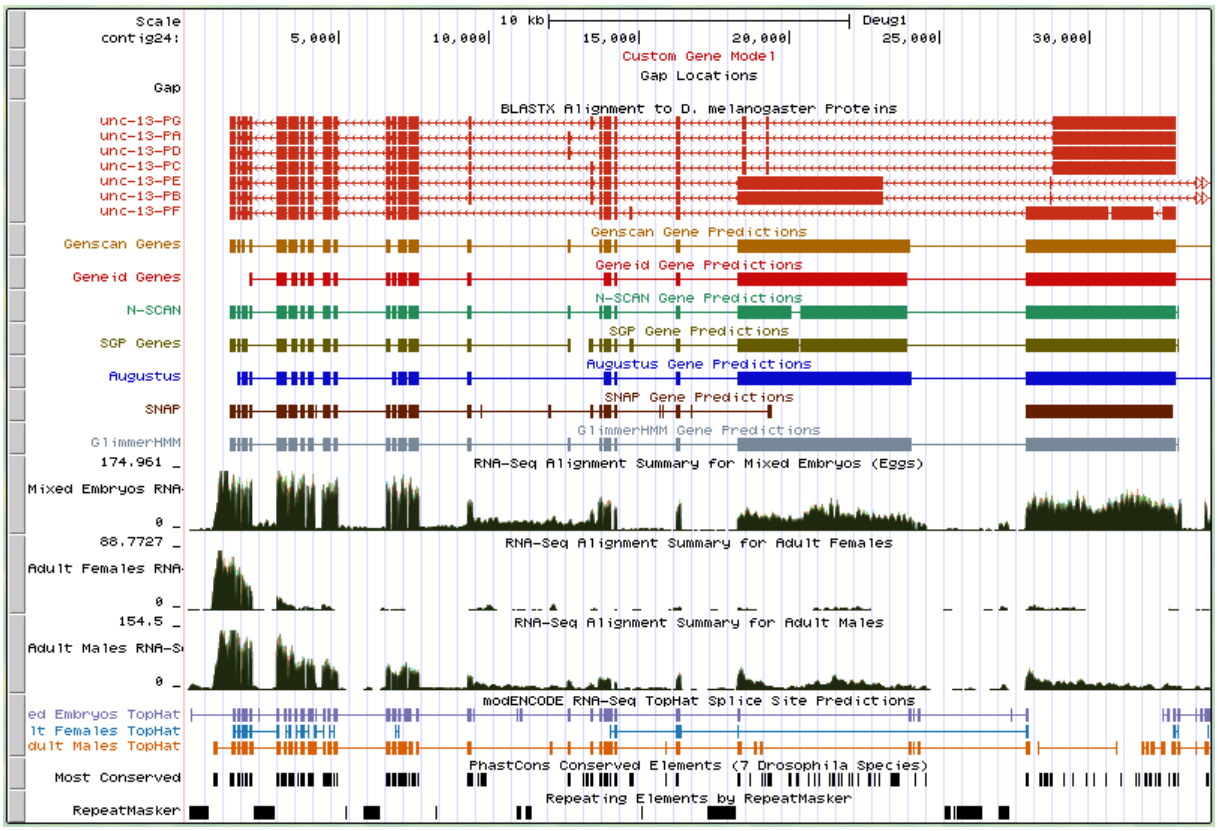
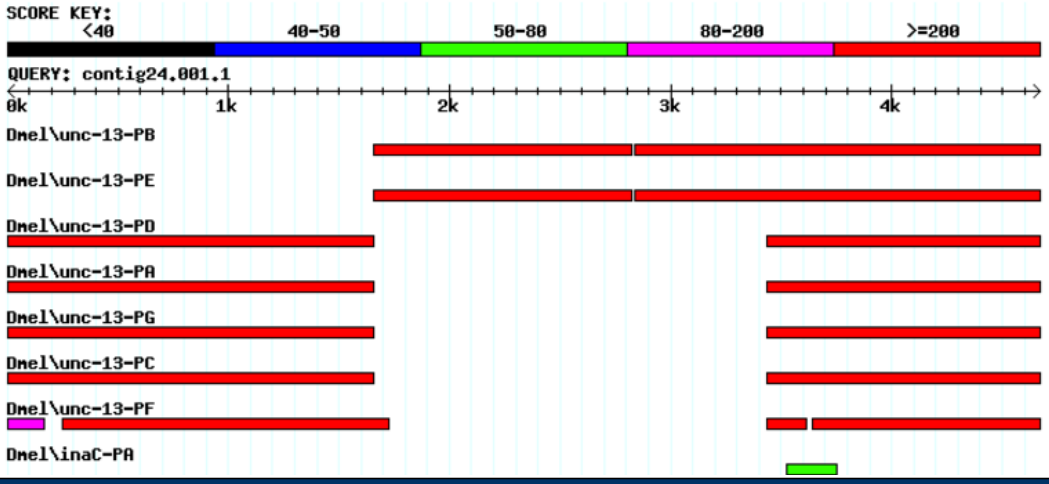


Figure 25: UCSC Genome browser overview of second feature in* D. eugracilis *contig 24*. This feature is orthologous to the* D. melanogaster *gene* unc-13.



BLAST Hit Summary

Description	Species	Score	E value
unc-13-PB	Dmel	3084.66	0
unc-13-PE	Dmel	3078.5	0
unc-13-PD	Dmel	2620.5	0
unc-13-PA	Dmel	2620.5	0
unc-13-PG	Dmel	2575.82	0
unc-13-PC	Dmel	2570.42	0
unc-13-PF	Dmel	2058.11	0
inaC-PA	Dmel	73.559	5.98564e-12

***Figure 26: blastx alignment of N-SCAN gene prediction to* D. melanogaster.** *Alignment strongly suggests* unc-13 *as the* D. melanogaster *ortholog of this feature.*

A FlyBase Annotated Proteins blastp search of the N-SCAN gene prediction of the second feature in contig 24 was performed (Fig. 26). The top 7 results, all with E values of 0 and alignment scores ranging from 2058.11 to 3084.66, correspond to the 7 isoforms of *unc-13* in *D. melanogaster*. The result with the second highest alignment score, *inaC-PA*, has a score of 73.559 and an E value of 5.98564e-12. Further, this gene is found on chromosome 2R while *unc-13* is found on chromosome 4. This feature was concluded to be the *D. eugracilis* ortholog of *unc-13* (Fig. 27).

>gnl|dmel|FBpp0088307 type=protein; loc=4:complement(join(886965..892796, 885760..885891, 883567..883641, 883143..883362, 882961..883081, 882617..882723, 878851..878992, 876870..877194, 876517..876789, 876296..876454, 876114..876239, 875543..875710, 875203..875484, 874246..874467, 873994..874160, 873635..873926, 873214..873571, 872127..872225, 871912..872074, 871720..871846, 871493..871654)); ID=FBpp0088307; name=unc-13-PB; parent=FBgn0025726, FBtr0089247; dbxref=FlyBase_Annotation_IDs:CG2999-PB, FlyBase:FBpp0088307, GB_protein:AAN06593.1, REFSEQ:NP_726615, GB_protein:AAN06593, UniProt/TrEMBL:Q8IM86, FlyMine:FBpp0088307, modMine:FBpp0088307; MD5=ff1af0224b8362cae6c96ee6c6d78605; length=3183; release=r6.15; species=Dmel;
Length = 3183

HSP # = 1 , Score = 3084.66 bits (7996) , Expect = 0
Identities = 1536 / 1890 (81.3%) , Positives = 1636 / 1890 (86.6%) , Gaps = 80 / 1890 (4.2%)

Subject FASTA

```
Query:   2838        SPKMNNLMQKAKTYKRHNFVLRRGCNLPNTELDTPDFTSSDNGKSSTFKGKIVINNYEDG   2897
                     S KM+ LMQKAKTYKRH+FVLRRGCN+ ++EL+ PDF SS N  SS    +I++N   +
Subject: 1317        SSKMSGLMQKAKTYKRHSFVLRRGCNMSDSELEMPDFVSSGNDNSSISTREILLNQSIEV   1376

Query:   2898        VDK--NLNDQNDYGIKKVLSG-------NELNNLFPVVGDLKKTQSPISVAGSTAIHISE   2948
                      D+  + N +N     K VL G        N  NNLFP+VGDLKK QSP+ +A  T I
Subject: 1377        EDEQEDFNYKNRCDSKSVLGGSIEKLNGNLTNNLFPIVGDLKKIQSPLPLAVLTEI----   1432
```

Figure 27: Highest scoring blastp alignment of N-SCAN prediction to* D. melanogaster *protein data base. *Alignment confirms that the suspected ortholog is on the fourth chromosome*

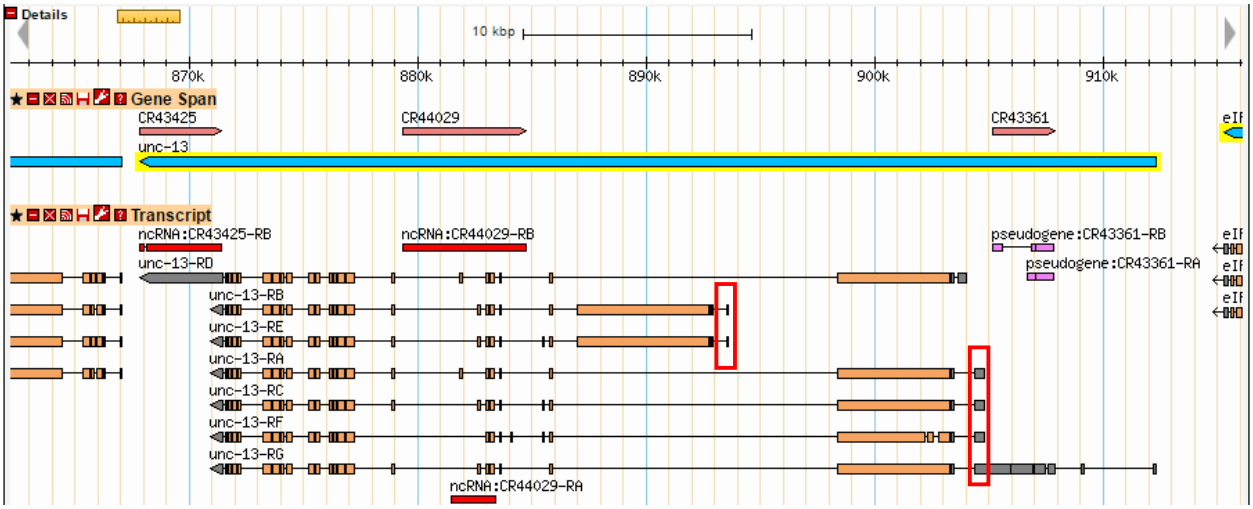


***Figure 28: Isoforms of* unc-13 *in* D. melanogaster.** unc-13 *has 29 unique exons across 7 isoforms with 6 unique coding sequences. The separate start codons are shown in the red boxes.*

unc-13 has 7 isoforms in *D. melanogaster* representing 6 unique coding sequences. The A and D isoforms have identical coding sequences, differing only in the 5' UTR. The A, C, D, F, and G isoforms share a start codon while the B and E isoforms share a different start codon upstream of the first (Fig. 28). The A, D, E, and G isoforms contain 22 CDSs, the B isoform contains 21 CDSs, the C isoform contains 23 CDSs, and the F isoform contains 24 CDSs.

Using the same settings as for *eIF4G,* each *D. melanogaster unc-13* exon was aligned to *D. eugracilis* contig 24 using NCBI blastx (Fig. 30). Of the 29 unique exons in *D. melanogaster*,

27 aligned to *D. eugracilis* contig 24 (Table 3). The two exons that did not align to contig 24, 1_2147_0 and 8_2147_0, have lengths of 2 AA and 3 AA respectively. Both exons were identified in *D. eugracilis* using the UCSC Genome Browser.

Flybase_ID	Query start-end	E value	Size	Identities	Positives	Gaps	Frame
1_2147_0							
5_2147_0	32833-27875	0		873/1694	1103/1694	105/1694	-3
2_2147_0	32833-32381	1e-47		81/152	98/152	1/152	-3
3_2147_2	32101-31853	4e-34		59/83	63/83	0/83	-3
4_2147_2	31777-27875	0		663/1340	853/1340	99/1340	-3
6_2147_0	23998-18278	0		1048/2009	1313/2009	168/2009	-3
7_2147_0	16381-16250	1e-26		41/44	42/44	0/44	-3
8_2147_0							
9_2147_0	14766-14698	1e-5		12/23	17/23	0/23	-1
10_2147_0	14263-14189	1e-12		24/25	25/25	0/25	-3
11_2147_0	14066-13848	1e-51		73/73	73/73	0/73	-2
12_2147_2	13786-13670	8e-22		38/39	39/39	0/39	-3
13_2147_1	13489-13385	5e-19		33/35	34/35	0/35	-3
14_2147_1	12734-12630	7e-22		35/35	35/35	0/35	-2
15_2147_2	9451-9314	3e-29		46/46	46/46	0/46	-3
16_2147_1	7687-7364	5e-72		107/108	108/108	0/108	-3
17_2147_0	7287-7015	3e-60		91/91	91/91	0/91	-1
18_2147_0	6949-6791	6e-32		53/53	53/53	0/53	-3
19_2147_0	6733-6608	1e-26		42/42	42/42	0/42	-3
20_2147_0	4990-4823	1e-35		55/56	55/56	0/56	-3
21_2147_0	4766-4485	3e-66		94/94	94/94	0/94	-2
22_2147_0	4205-3984	1e-46		71/74	74/74	0/74	-2
23_2147_0	3887-3723	2e-32		55/55	55/55	0/55	-2

24_2147_1	3660-3370	2e-59		97/97	97/97	0/97	-1
25_2147_0	3314-2958	1e-74		117/119	118/119	0/119	-2
26_2147_2	2136-2041	5e-18		31/32	31/32	0/32	-1
27_2147_2	1976-1818	8e-35		53/53	53/53	0/53	-2
28_2147_1	1753-1628	3e-26		42/42	42/42	0/42	-3
29_2147_0	1574-1413	4e-33		53/54	54/54	0/54	-2

Table 4: Exon by exon NCBI blastx search of* D. melanogaster *exons in* D. eugracilis *contig 24. *Exons 8 and 1 in* D. melanogaster *failed to align to contig 24.*

unc-13:2_2147_0

Sequence ID: Query_211975 **Length:** 152 **Number of Matches:** 1

Range 1: 1 to 152 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
156 bits(395)	1e-47	81/152(53%)	98/152(64%)	1/152(0%)	-3

```
Query  32833  HYARHEYFHNTQNDALASDTGISTYNHMNYGTQNIREYFIEPCNLSNQGPDDYSSEGQYA  32654
              HY RH+YFHNTQN AL+SDT   +Y+ ++Y TQ  REYF E   LSNQGP++ S
Sbjct  1      HYVRHDYFHNTQNGALSSDTSRISYSQISYETQPSREYFSESYALSNQGPEECSRVSHLN  60

Query  32653  SDTIPTTVDKSNNSYINDYIEQFGAQEQHEAEEESFDNWNEN-STVTTPYEVEFVNQKRT  32477
              SDT+ TTVD SNNSY  DY+E +GA  Q + EE+S DNWNEN S V   Y +   N   T
Sbjct  61     SDTVLTTVDNSNNSYGYDYLECYGANIQCDPEEDSVDNWNENTSVVADQYGLGHNNLNCT  120

Query  32476  SQKLLPKLPSNINGSASSKPCPPHMDINLKTE  32381
              S KLLPKLP+  NG  SS  C P MD+   T+
Sbjct  121    SSKLLPKLPNIENGRGSSNACAPQMDVKFNTK  152
```

unc-13:3_2147_2

Sequence ID: Query_46149 **Length:** 91 **Number of Matches:** 1

Range 1: 9 to 91 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
114 bits(286)	4e-34	59/83(71%)	63/83(75%)	0/83(0%)	-3

```
Query  32101  FAMGQTAAMDSGSGTYDVFENMSRPYTSMLPLGYSDFEECYYNIDNLSTYSDTPQINNAQ  31922
              F   QT AM S S TY+V+E M RPYTSMLPL YSD++E  YN DNLSTYSDTP  NN Q
Sbjct  9      FDTDQTDAMGSESSTYEVYEKMQRPYTSMLPLDYSDYQEGCYNTDNLSTYSDTPPSNNTQ  68

Query  31921  LNLQKQRKFSLMMAMTTASVIAS  31853
              L  Q QRK SLMMAMTTASVIAS
Sbjct  69     LKRQMQRKISLMMAMTTASVIAS  91
```

unc-13:4_2147_2

Sequence ID: Query_82667 Length: 1280 Number of Matches: 1

Range 1: 1 to 1280 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
1131 bits(2925)	0.0	677/1341(50%)	868/1341(64%)	101/1341(7%)	-3

```
Query  31777  SSIITNAATRASERMLEPKCCGGIVTPGDTGAVASTPSVNTATISKTRKLPKVLPTPQNK  31598
              ++I TNAA R  +R L  + C  IV   D+G+V S PS     I+KTRKLPKVLPTP  K
Sbjct  1      NTISTNAAARDLDRCLATESCEVIVDTRDSGSVTSFPSSAVTAITKTRKLPKVLPTPLCK  60

Query  31597  LSLHTKTTGTEALNFS---NPIAEKIHRSKQLPKLPTSLSQTKPYSVPNSNFSTYCALEA  31427
               S H  T  T+AL+ S   +P+ EK HR KQLPKLP SL Q+   +  NSN++T  A +A
Sbjct  61     SSRHPITIATDALSSSYTSDPLPEKSHRPKQLPKLPISLPQSNDRASLNSNWATPPAPDA  120

Query  31426  LSFNLLDNKSAS-PSPTEITVT----TTSFPTVTSYEDSPKPFAYSYESKDPIEVFLSKS  31262
              L FN  D+KSAS P+PT  T+T    TTS+   T +  +     Y Y+SK+P  VF  KS
Sbjct  121    LPFNSFDHKSASSPTPTT-TITKDTETTSYLVETDFIGARHNALYQYDSKEPNIVFSDKS  179

Query  31261  IEADPAPSF-------SIEKQCVSVELSNNIIQENSPPCHIPEISTANPDIKPDHCDSFI  31103
              +EA+ +P++       S +  C  V L +NI+Q  S  CH+PEI     DI+ +   S I
Sbjct  180    VEAEHSPTWTPLSPIQSKQSPCPPVALPSNIMQNVSLTCHLPEIEATRSDIEREPESSSI  239

Query  31102  DSIFQNEKSPDSYINPESALFNISEYLKPYTLEKITFTGEKPNQITDAASISTTAHPSNG  30923
              + I + EK  D Y  P SALFNISEYLKPYTL K   + EK N I +AAS STT  P N
Sbjct  240    EPILEIEKLADPYSGPGSALFNISEYLKPYTLNKPILSEEKKNHIANAASTSTTT-PLN-  297

Query  30922  KLSGEINVLKSTLSSGDALVPYSNTCPTSCVNYKPLEVESGSNILMQTNLNTNPAEFIIL  30743
                          ++S D    YSN   ++CVN++PL+VES  NI ++ N  TN AE ++
Sbjct  298    ------------ITSDDEFSSYSNKWTSTCVNFQPLDVESSLNISLKVNAGTNQAELLMT  345

Query  30742  PSSSCIPALSLSSNLISDHLTDADIFPVIGFPPDKDVTVKDSFQNNAFTTTVYINNNEVV  30563
              P  S  P L +SSN  SD+            PPD           +AFTTTV +N+ E V
Sbjct  346    PLKSSTP-LFISSNGTSDNFNLRK-----SSPPD-----------SAFTTTVNVNSFETV  388

Query  30562  PVTKSQNATPT--------SEAPIISYSDYMKKFELPDLPQPIMILSEKLPVTQSVSGAV  30407
               V+ SQ A+P+        S AP++SYSDYMK+FELP+LPQPIM LSE    TQS S  V
Sbjct  389    LVSGSQTASPSPSNLKSPPSIAPLLSYSDYMKQFELPELPQPIMDLSENDTATQSDSFNV  448

Query  30406  ------TADS-EAEKELDVENRCSELLPSYFTKLFSEYNAPISSSEIRASPINDEIDN-S  30251
                    AD+  +  ++DVE++ S  LPSY ++ F   + P  S + +   I +++D+ S
Sbjct  449    INNTLTNADNLNSYNQMDVESKSSLQLPSYSSESFDPCSVPSFSIKNKEYKIVEKLDSLS  508

Query  30250  FYMKTESSIPITTTVCSAGPPSYFPENEAKEKNHLTFDDTFYDSFNVDIIELTASVAQVE  30071
                   ES   + + V         P  E+   N + FDDTFYDSFNVDI ELTA V  V
Sbjct  509    NVESVESPKTLVSPVNPLNCSKLLPGTESIVSNDVAFDDTFYDSFNVDIKELTAFVDHVA  568

Query  30070  SENDLNNAPIESLNNSINIENETSFEFSIEKTLDTRDMNQNVSLGQGGYYRPSQAQQKPN  29891
               E+ L N P           ++TS EFS +KT DT DMNQN+S G+ GYY+PSQAQQK +
Sbjct  569    PEDGLYNFP----------NDKTSVEFSFDKTEDTIDMNQNLSSGECGYYKPSQAQQKAS  618

Query  29890  VVASAATSVLDGISKGLKGGLDGVFSNVSSSVEATQTVNATRKAFSFNLASKLVPSVGGL  29711
              VVASAA+SVLDGISKGLKGGLDGVFS VSS+V+ TQ+  ++++ FSFNLASK+VPSVGGL
Sbjct  619    VVASAASSVLDGISKGLKGGLDGVFSGVSSTVDVTQSNPSSKRGFSFNLASKIVPSVGGL  678
```

```
Query  29536  PN-TLKGETHLNNVEMVETSSTLLENVCDNYYENYDETLLTDKMMNACMLDRRSEYGLIE  29360
              N + K  ++ N  E+ E SSTL+ NVCD+Y  +YDE +LT++M+N  MLD  SE+GLIE
Sbjct  736    HNKSTKSNSYYN--EVGEISSTLVRNVCDSYDNSYDEMILTNEMVNIGMLDSESEFGLIE  793

Query  29359  NSYSYHVSDEGQLATFNTQSTLLKDVPNESELGIENLKIKNNPTYWHEPMTKKGSTSSSG  29180
              NSYSY V D  Q+ + N+ +   ++V N    GIE    KN P   H+P TKK ST   G
Sbjct  794    NSYSYQVPDNEQIDSVNSYNNKTQNVTNN---GIEKANTKNKPVPLHDPPTKKAST--VG  848

Query  29179  MLGSILGKAAAAVQSATHAVNQGASTVVSVVGQKQTLLPATISIHDMDGISSTTTTIKRE  29000
              M GSILGKAAAAVQSAT AVNQ AS+V SVV QK T++P T ++  +  + S     +
Sbjct  849    MFGSILGKAAAAVQSATQAVNQSASSVASVVAQKPTIVPRTNNVLLLSSVCSPNEIKRNS  908

Query  28999  SNVD-----SYQLTNEESLSSPYKNTIDEFENTNIKMQEYSTYIEKETFVNYQSNGNQHQ  28835
              S+V+      YQ+ + ESLSS Y NT  +++N+N+K+ E+ TY +   + +Y +NGNQ Q
Sbjct  909    SSVEFDSEYGYQMPDVESLSSHYANTGGDYDNSNMKIHEFGTYADDRPYADYHTNGNQSQ  968

Query  28834  FRNDSVLSEQSQVISNVSKALPTVPPSGSTGKKLPTVNGKSGLLIKQMPTEIYDDESDLD  28655
              F+ ++V+  + +VI+  +  LP  P   +TGKKLPTVNGKS LLIKQMPTE+YDDESD D
Sbjct  969    FKEEAVIPGEPEVIN--TNILPIGPQ--ATGKKLPTVNGKSALLIKQMPTEVYDDESDTD  1024

Query  28654  DLDVNPSIGKEPSYRIDGEQDDYYMDLQQTTPSNQINGYYEHVNNGYDYREDYFNEEDEY  28475
              +LDV+PS GK PSY I  EQ+DYYMDLQQTTPS Q NG+YE VNNGYDYREDYFNEEDEY
Sbjct  1025   ELDVSPSTGKVPSYSIYSEQEDYYMDLQQTTPSIQPNGFYEQVNNGYDYREDYFNEEDEY  1084

Query  28474  KYLEQQREQEQH-EPKIKKYVKQTNSMLLTCQAQSSLDFIDEGQDDDFIYDSYHSEEDSG  28298
              KYLEQQREQE+H +PK KKY+KQ     ++     SLDFID GQDDDFIYD+YHSE+DSG
Sbjct  1085   KYLEQQREQEEHNQPKNKKYLKQAK---ISKIQPPSLDFIDVGQDDDFIYDNYHSEDDSG  1141

Query  28297  NYLDESSSGSVGPSEGRNLKMDTNGEVALTSTSIQIKSDSLAPIKNHIQKHDSLICQPTA  28118
              NYL+ SSSGSVGP EG  +K+D+N E +    S+  KSDS  P  + +QKHD++I + T
Sbjct  1142   NYLEGSSSGSVGPIEGSIIKVDSNIEASF--ASLNKKSDSFTPTNDSLQKHDTVIGESTT  1199

Query  28117  KKTSFNEEKTCPDLDEREEDIHDQLSDLTDLNNLLPQKKKTLLRGETEEVVGGNMQMIRQ  27938
              K T    EK CPD+DE +E++ D +SDLTDL+ L+ QKKKTLLRGETEEVVGG+MQ++RQ
Sbjct  1200   KLTRLRTEKMCPDVDEEDENLSDHVSDLTDLSKLISQKKKTLLRGETEEVVGGHMQVLRQ  1259

Query  27937  PEITAGQRWHWAYNKIIMQLN  27875
               EITA QRWHWAYNKIIMQLN
Sbjct  1260   TEITARQRWHWAYNKIIMQLN  1280
```

unc-13:5_2147_0

Sequence ID: Query_166469 **Length:** 1630 **Number of Matches:** 1

Range 1: 1 to 1630 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
1496 bits(3873)	0.0	873/1694(52%)	1103/1694(65%)	105/1694(6%)	-3

```
Query  32833  HYARHEYFHNTQNDALASDTGISTYNHMNYGTQNIREYFIEPCNLSNQGPDDYSSEGQYA  32654
              HY RH+YFHNTQN AL+SDT   +Y+ ++Y TQ  REYF E   LSNQGP++ S
Sbjct  1      HYVRHDYFHNTQNGALSSDTSRISYSQISYETQPSREYFSESYALSNQGPEECSRVSHLN  60

Query  32653  SDTIPTTVDKSNNSYINDYIEQFGAQEQHEAEEESFDNWNEN-STVTTPYEVEFVNQKRT  32477
              SDT+ TTVD SNNSY  DY+E +GA  Q + EE+S DNWNEN S V   Y +   N   T
Sbjct  61     SDTVLTTVDNSNNSYGYDYLECYGANIQCDPEEDSVDNWNENTSVVADQYGLGHNNLNCT  120

Query  32476  SQKLLPKLPSNINGSASSKPCPPHMDINLKTEGMCIKREQRHGGCKAKAHECDRHFFPGD  32297
              S KLLPKLP+  NG  SS  C P MD+   T+GMCIK +  +G C AKAH+      P D
Sbjct  121    SSKLLPKLPNIENGRGSSNACAPQMDVKFNTKGMCIKIDHSYGVCMAKAHDFVGRLSPSD  180

Query  32296  DQNIYADNINGYTGFAFSSAFHNNVVSAAPLRTLAQTARSNFYLSQDFFELNADVDQAGE  32117
               QNI  +N+NGY G A+SS F +N +S+APLR L Q+ R + YL ++    NAD   A +
Sbjct  181    YQNILGNNLNGYAGCAYSSTFLDNAMSSAPLRVLPQSPRCSSYLGRNIIGFNAD---AAQ  237

Query  32116  LSQCDFAMGQTAAMDSGSGTYDVFENMSRPYTSMLPLGYSDFEECYYNIDNLSTYSDTPQ  31937
                   F   QT AM S S TY+V+E M RPYTSMLPL YSD++E  YN DNLSTYSDTP
Sbjct  238    RDGRGFDTDQTDAMGSESSTYEVYEKMQRPYTSMLPLDYSDYQEGCYNTDNLSTYSDTPP  297

Query  31936  INNAQLNLQKQRKFSLMMAMTTASVIASGETRVPVLSKHSKKPTETRTDSLLGSSIITNA  31757
               NN QL  Q QRK SLMMAMTTASVIASGE RVPV SK SKK TE +TDS++G++I TNA
Sbjct  298    SNNTQLKRQMQRKISLMMAMTTASVIASGEIRVPVHSKQSKKSTEIQTDSIIGNTISTNA  357

Query  31756  ATRASERMLEPKCCGGIVTPGDTGAVASTPSVNTATISKTRKLPKVLPTPQNKLSLHTKT  31577
              A R  +R L  + C  IV   D+G+V S PS     I+KTRKLPKVLPTP  K S H  T
Sbjct  358    AARDLDRCLATESCEVIVDTRDSGSVTSFPSSAVTAITKTRKLPKVLPTPLCKSSRHPIT  417

Query  31576  TGTEALN---FSNPIAEKIHRSKQLPKLPTSLSQTKPYSVPNSNFSTYCALEALSFNLLD  31406
                T+AL+    S+P+ EK HR KQLPKLP SL Q+   +  NSN++T  A +AL FN  D
Sbjct  418    IATDALSSSYTSDPLPEKSHRPKQLPKLPISLPQSNDRASLNSNWATPPAPDALPFNSFD  477

Query  31405  NKSA-SPSPTEITVT----TTSFPTVTSYEDSPKPFAYSYESKDPIEVFLSKSIEADPAP  31241
              +KSA SP+PT  T+T    TTS+   T +  +     Y Y+SK+P  VF  KS+EA+ +P
Sbjct  478    HKSASSPTPT-TTITKDTETTSYLVETDFIGARHNALYQYDSKEPNIVFSDKSVEAEHSP  536

Query  31240  SF-------SIEKQCVSVELSNNIIQENSPPCHIPEISTANPDIKPDHCDSFIDSIFQNE  31082
              ++       S +  C  V L +NI+Q  S  CH+PEI     DI+ +   S I+ I + E
Sbjct  537    TWTPLSPIQSKQSPCPPVALPSNIMQNVSLTCHLPEIEATRSDIEREPESSSIEPILEIE  596

Query  31081  KSPDSYINPESALFNISEYLKPYTLEKITFTGEKPNQITDAASISTTAHPSNGKLSGEIN  30902
              K  D Y  P SALFNISEYLKPYTL K   + EK N I +AAS STT  P N
Sbjct  597    KLADPYSGPGSALFNISEYLKPYTLNKPILSEEKKNHIANAASTSTTT-PLN--------  647

Query  30901  VLKSTLSSGDALVPYSNTCPTSCVNYKPLEVESGSNILMQTNLNTNPAEFIILPSSSCIP  30722
                   ++S D    YSN   ++CVN++PL+VES  NI ++ N  TN AE ++ P  S  P
Sbjct  648    -----ITSDDEFSSYSNKWTSTCVNFQPLDVESSLNISLKVNAGTNQAELLMTPLKSSTP  702
```

```
Query  30721  ALSLSSNLISDHLTDADIFPVIGFPPDKDVTVKDSFQNNAFTTTVYINNNEVVPVTKSQN  30542
               L +SSN  SD+                      K S  ++AFTTTV +N+ E V V+ SQ
Sbjct  703    -LFISSNGTSDNFN----------------LRKSSPPDSAFTTTVNVNSFETVLVSGSQT  745

Query  30541  ATPT--------SEAPIISYSDYMKKFELPDLPQPIMILSEKLPVTQSVSGAV------T  30404
               A+P+        S AP++SYSDYMK+FELP+LPQPIM LSE    TQS S  V
Sbjct  746    ASPSPSNLKSPPSIAPLLSYSDYMKQFELPELPQPIMDLSENDTATQSDSFNVINNTLTN  805

Query  30403  ADS-EAEKELDVENRCSELLPSYFTKLFSEYNAPISSSEIRASPINDEIDN-SFYMKTES  30230
               AD+  +  ++DVE++ S  LPSY ++ F   + P  S + +   I +++D+ S     ES
Sbjct  806    ADNLNSYNQMDVESKSSLQLPSYSSESFDPCSVPSFSIKNKEYKIVEKLDSLSNVESVES  865

Query  30229  SIPITTTVCSAGPPSYFPENEAKEKNHLTFDDTFYDSFNVDIIELTASVAQVESENDLNN  30050
                 + + V         P  E+   N + FDDTFYDSFNVDI ELTA V  V  E+ L N
Sbjct  866    PKTLVSPVNPLNCSKLLPGTESIVSNDVAFDDTFYDSFNVDIKELTAFVDHVAPEDGLYN  925

Query  30049  APIESLNNSINIENETSFEFSIEKTLDTRDMNQNVSLGQGGYYRPSQAQQKPNVVASAAT  29870
               P           ++TS EFS +KT DT DMNQN+S G+ GYY+PSQAQQK +VVASAA+
Sbjct  926    FP----------NDKTSVEFSFDKTEDTIDMNQNLSSGECGYYKPSQAQQKASVVASAAS  975

Query  29869  SVLDGISKGLKGGLDGVFSNVSSSVEATQTVNATRKAFSFNLASKLVPSVGGLLSSSNSS  29690
               SVLDGISKGLKGGLDGVFS VSS+V+ TQ+  ++++ FSFNLASK+VPSVGGLL+S++S+
Sbjct  976    SVLDGISKGLKGGLDGVFSGVSSTVDVTQSNPSSKRGFSFNLASKIVPSVGGLLTSTSST  1035

Query  29689  STKQTQAPKSVTSPTLI-TSVENDSTDSCTYETTISPP-LYKTAENHFYPATLPN-TLKG  29519
               S KQT    S T+PTLI  S EN S+ +  Y  T SP    K  E + Y AT+ N + K
Sbjct  1036   SIKQT---GSETNPTLILISPENVSSRNSNYIPTTSPSCTQKNGEENLYSATVHNKSTKS  1092

Query  29518  ETHLNNVEMVETSSTLLENVCDNYYENYDETLLTDKMMNACMLDRRSEYGLIENSYSYHV  29339
                ++ N  E+ E SSTL+ NVCD+Y  +YDE +LT++M+N  MLD  SE+GLIENSYSY V
Sbjct  1093   NSYYN--EVGEISSTLVRNVCDSYDNSYDEMILTNEMVNIGMLDSESEFGLIENSYSYQV  1150

Query  29338  SDEGQLATFNTQSTLLKDVPNESELGIENLKIKNNPTYWHEPMTKKGSTSSSGMLGSILG  29159
               D  Q+ + N+ +   ++V N    GIE    KN P   H+P TKK ST   GM GSILG
Sbjct  1151   PDNEQIDSVNSYNNKTQNVTNN---GIEKANTKNKPVPLHDPPTKKAST--VGMFGSILG  1205

Query  29158  KAAAAVQSATHAVNQGASTVVSVVGQKQTLLPATISIHDMDGISSTTTTIKRESNVD---  28988
               KAAAAVQSAT AVNQ AS+V SVV QK T++P T ++  +  + S     +  S+V+
Sbjct  1206   KAAAAVQSATQAVNQSASSVASVVAQKPTIVPRTNNVLLLSSVCSPNEIKRNSSSVEFDS  1265

Query  28987  --SYQLTNEESLSSPYKNTIDEFENTNIKMQEYSTYIEKETFVNYQSNGNQHQFRNDSVL  28814
                  YQ+ + ESLSS Y NT  +++N+N+K+ E+ TY +   + +Y +NGNQ QF+ ++V+
Sbjct  1266   EYGYQMPDVESLSSHYANTGGDYDNSNMKIHEFGTYADDRPYADYHTNGNQSQFKEEAVI  1325

Query  28813  SEQSQVISNVSKALPTVPPSGSTGKKLPTVNGKSGLLIKQMPTEIYDDESDLDDLDVNPS  28634
                 + +VI+  +  LP  P   +TGKKLPTVNGKS LLIKQMPTE+YDDESD D+LDV+PS
Sbjct  1326   PGEPEVIN--TNILPIGPQ--ATGKKLPTVNGKSALLIKQMPTEVYDDESDTDELDVSPS  1381

Query  28633  IGKEPSYRIDGEQDDYYMDLQQTTPSNQINGYYEHVNNGYDYREDYFNEEDEYKYLEQQR  28454
                GK PSY I  EQ+DYYMDLQQTTPS Q NG+YE VNNGYDYREDYFNEEDEYKYLEQQR
Sbjct  1382   TGKVPSYSIYSEQEDYYMDLQQTTPSIQPNGFYEQVNNGYDYREDYFNEEDEYKYLEQQR  1441

Query  28453  EQEQH-EPKIKKYVKQTNSMLLTCQAQSSLDFIDEGQDDDFIYDSYHSEEDSGNYLDESS  28277
               EQE+H +PK KKY+KQ     ++      SLDFID GQDDDFIYD+YHSE+DSGNYL+ SS
Sbjct  1442   EQEEHNQPKNKKYLKQAK---ISKIQPPSLDFIDVGQDDDFIYDNYHSEDDSGNYLEGSS  1498


Query  28276  SGSVGPSEGRNLKMDTNGEVALTSTSIQIKSDSLAPIKNHIQKHDSLICQPTAKKTSFNE  28097
               SGSVGP EG  +K+D+N E +    S+  KSDS  P  + +QKHD++I + T K T
Sbjct  1499   SGSVGPIEGSIIKVDSNIEASF--ASLNKKSDSFTPTNDSLQKHDTVIGESTTKLTRLRT  1556

Query  28096  EKTCPDLDEREEDIHDQLSDLTDLNNLLPQKKKTLLRGETEEVVGGNMQMIRQPEITAGQ  27917
               EK CPD+DE +++ D +SDLTDL+ L+ QKKKTLLRGETEEVVGG+MQ++RQ EITA Q
Sbjct  1557   EKMCPDVDEEDENLSDHVSDLTDLSKLISQKKKTLLRGETEEVVGGHMQVLRQTEITARQ  1616

Query  27916  RWHWAYNKIIMQLN  27875
               RWHWAYNKIIMQLN
Sbjct  1617   RWHWAYNKIIMQLN  1630
```

unc-13:6_2147_0

Sequence ID: Query_235151 Length: 1944 Number of Matches: 1

Range 1: 2 to 1944 Graphics ▼ Next Match ▲ Pre

Score	Expect	Identities	Positives	Gaps	Fra
1734 bits(4492)	0.0	1056/2001(53%)	1320/2001(65%)	152/2001(7%)	-3

```
Query  23998  MNTSLKPVPEDPEKKILFRKQELKTNTEDKLIFAENALKSQIQIKEQLRQQQQPIMYTSS  23819
              MNTS   V  D EKK    K+ELK NT++KLIFAENALKSQI+IKEQLR QQQ  +Y SS
Sbjct  2      MNTSQLQVTGDTEKKSQLLKKELKINTQEKLIFAENALKSQIKIKEQLRLQQQSTIYASS  61

Query  23818  --SNCPAGAPVRAESLIQS-ANRPDYPLKINSVRTPNQKQLSNMQPTIRKSPTFLDSSFS  23648
                S+  AG+ VRA  L Q   N   + +  +  +     Q+  MQP + KSP  LD S+
Sbjct  62     LLSSSAAGS-VRAPLLSQGHLNSIQHNMDFDLAKA----QIPEMQPPMSKSPNGLDFSYL  116

Query  23647  SY--LNTNGSMISIKSEQQLSEIYNSQQHSDYIISDYMDKIATRISLLETELKFAWRALD  23474
              SY  +NTN SMISIKSEQQL + YNS+QHSDYIISDYMDKIATRISLLETELKFAWRALD
Sbjct  117    SYPSINTNESMISIKSEQQLCQSYNSEQHSDYIISDYMDKIATRISLLETELKFAWRALD  176

Query  23473  LLSTEYGKIWVRLEKLENISIEQQSVVSNLMDLIGATNHEQQKGNIDVLEFPSFRDENKL  23294
              LLSTEYGKIW+RLEKLENISIEQQSVV NL+DLIGA+  E QK +I+ ++ P ++DE++L
Sbjct  177    LLSTEYGKIWIRLEKLENISIEQQSVVGNLVDLIGASKKELQKVDIERMKVPLYQDEDQL  236

Query  23293  LPLVIKDTLGLDIVLEAEPSNPDFEKELRFRDHDHTSVSKLAQVRDSDLFFNSACAFGVH  23114
              LPL ++DTL +DI    + SN DF+K L F +H+ T V+K  Q   S+   NSA A   H
Sbjct  237    LPLEMEDTLDIDI----QSSNRDFDKNLTFENHEKTFVTKHTQATKSEDLMNSAYAIDSH  292

Query  23113  DPYLHFEDDALNRQSLNIEQMKMVLEKPDEPNSKCFQSDIEVSRKESSIQLYSDSDVQFY  22934
                 +FE+   N ++L+I  +K   EK  EP  K  QSD E + KES  +LYSDSD+  Y
Sbjct  293    P---NFENIDFNGKNLDIGIIKFGFEKGYEPKQKGNQSDFE-AYKESRTKLYSDSDLMLY  348

Query  22933  EKQHFLANNARTELMKEFINGRRVLNEISESSGVSSKGSNISQHKRGSKNDQSFMDFDGD  22754
              E+Q FLAN+AR ELMKEF+NGRRVLNEIS SSG  SK S   ++    + +    + D
Sbjct  349    ERQQFLANSARAELMKEFLNGRRVLNEISASSGALSKSSEKDKNLVKPR-EPLLSEIDDT  407

Query  22753  NIRQTSNELDNLIEDVKFFRLTASKTSEGGIQVLPIEDTGGTEISANVNEINESFYKKLN  22574
               +R+++ EL +   +  FF++ ASKT EGG QVLPIEDTGGTE  AN+NE++ESFYK LN
Sbjct  408    YVRRSTCELVSPSSN-NFFQIAASKTGEGGNQVLPIEDTGGTESPANINEVDESFYKNLN  466

Query  22573  EAYRDNDLSNEIFKVDALLYQSEVAHEEQSIPTIRGHSASSSVLRTPDERDITDGQSFET  22394
              EAYRDN+LS+EIFKVDALL+QSE  H+  S   IR  S SS VL       IT  Q  +T
Sbjct  467    EAYRDNELSSEIFKVDALLHQSEATHDHISF--IRNQSTSSPVLNNQRVNAITGTQPLQT  524

Query  22393  AETSLT---KQSKNKESRRYRKKKHHKNEMDMINKLKCILTQAHSSHIINGDTKELDTDL  22223
              A+  L     Q  NKE R++RKKKHH NEMDMINKLKCIL QA S+ II  D +EL + L
Sbjct  525    AKACLVTEPNQHLNKEPRKHRKKKHHTNEMDMINKLKCILAQAQSTEIIKRDIEELGSKL  584

Query  22222  VDIVTGTGTEQEEVGSENIHNGLKNQRSISRITNFNDDIRIILDSMTETLLQEINKIPGL  22043
               D          ++  +EN     K+   IS  +   DDIR +L+++ +TLL EINKI GL
Sbjct  585    SDACV------KQTETEN-----KSDNHIST-SYLKDDIRKMLNNVIQTLLGEINKIKGL  632

Query  22042  QSLSASHISQLRKTVWAEEKFFQKLSQVDKKLTLLLLNPITVTEELQRLHITNTNEMFVL  21863
              QSLSA+ +SQL+  VW+EE+FFQK+S VDKKLTLLLLNP+TVTEELQRL I+NTNE FVL
Sbjct  633    QSLSATQLSQLQNAVWSEERFFQKISLVDKKLTLLLLNPLTVTEELQRLCISNTNEKFVL  692

Query  21862  VMKKFKRNIDTLKKLVGSSLNDFKIVKTCASRFSPKSLDSSFKSNVCSLSAHLLRNNSDL  21683
              V+KK K+NIDTLKKLVGSSL+DFKI  + A+  +  SL+ S  SN  S SAHLLRNNS+L
Sbjct  693    VIKKLKKNIDTLKKLVGSSLDDFKIDNSHANMTTFHSLNPSHNSNDSSFSAHLLRNNSNL  752
```

```
Query  21682  DEQLKILETQEIEINRKKKIDEIVSELNENAHKTFNPYIENESNYSYTNSKEDFRRSTTN  21503
              DEQLKILETQEIEI+RKKKIDEIVS L+E  + TFNP +E +S +S TNS ++    + N
Sbjct  753    DEQLKILETQEIEIHRKKKIDEIVSGLSEETY-TFNPDMEYKSQHSCTNSTDNIIGFSKN  811

Query  21502  IYNEDEYIKSLKKSLERHNSMIFLLHLQNPEKNKVSSDINFTQMDLNRASLSPPPPAPND  21323
              IYNEDEYIKSL+KSLERHNSMIFLLHLQNPEK+KV +DIN  QMD NRASLSPPPPAP D
Sbjct  812    IYNEDEYIKSLRKSLERHNSMIFLLHLQNPEKHKVLADINDAQMDSNRASLSPPPPAPTD  871

Query  21322  NFSFNETVNQFCGEIDSHKNAKSESRLSSMNCWEPNSPIAVGMQHSN--FNNIIEKRGKK  21149
                  N+ +NQ   +    +N KS+S LSSM+ W  NS ++VG+Q+ +   NNI+E    +
Sbjct  872    TVYLNDAINQITYQ---QRNGKSDSGLSSMSGWSANSQVSVGLQNYDPACNNILE-NCSE  927

Query  21148  RIQAYPPLPVA-SSDFRSLPATYDKMNTSKDNDYVITEENLNYIQELSKNLPICSAYENK  20972
              R+Q Y   P+A +S+F S P ++++  T+   ++VI EENLNYI ELSKNLPICSAYENK
Sbjct  928    RLQTYHSFPLAENSNFHSFPLSHEQAQTTNQTEFVILEENLNYIYELSKNLPICSAYENK  987

Query  20971  SIFD-----TNSDSVETFPTVDEILHWDHVNEKDKNNVSMEQLNTTIMPDLLSAQIRKSA  20807
              SIFD      +S+ +  F TVDE+L WD +NE DK N S ++LN+ +MPDLL+AQI  S
Sbjct  988    SIFDMKYEICDSNEIGKFSTVDEMLEWDQLNEPDKQNFSGKRLNSNLMPDLLTAQIPNSI  1047

Query  20806  TINPPSFRKKYQNANGGNVESLEDKNFSRSSHTE---LNKDCKTTPNKPNLTDRLVFYPS  20636
                   S   K  + N   V   E+K   R S  E    N   +    +P LTD+LVFYPS
Sbjct  1048   -----SKHNKNTSVNIEYVRQKENKGMDRRSIIEPNIYNGKSEDQICRPCLTDKLVFYPS  1102

Query  20635  YNSIADYNSITNLDYVGEQNQITDITDSQYLPNLEVPPVHLDFTANTNSEAHLCVSEKKT  20456
               NSI D+NS  + + + +Q+Q   I +       + P  +LD+T+ T S+A   V   +T
Sbjct  1103   SNSITDHNSSHDFNCLSQQDQTRIIKEFGSAHLNQDPTNYLDYTSGTYSKAPPEVLTHET  1162

Query  20455  SQQELGLNHR---IHIKKHHHSY-----------AKSPKVWKKLNNILTDNLKLKRMSKF  20318
              +   L  NH    +    +  SY           AK  KVWK+LN IL DNLKLKR+SKF
Sbjct  1163   NSSHLEFNHESESLFNSPNTSSYCKQKFVPGTSPAKPSKVWKRLNTILADNLKLKRVSKF  1222

Query  20317  NRSQSLPGDVQSQGNQSQGRGQAGSCPSISRKPPNASERQINLSKRVQKLPMR-------  20159
              NRS SLPGDVQSQG Q Q RGQAGSCP I  K  N +   + LSKR+QKLP+R
Sbjct  1223   NRSLSLPGDVQSQGLQRQPRGQAGSCPFI-HKRSNLAGSPVQLSKRIQKLPIRFIGRAKG  1281

Query  20158  --IVRRLSHPNTTDLLDPADHEPCSSDTGVYTNYLSSKMNNLMQKAKTYKRHNFVLRRGC  19985
                 VRR S P++   LD A  +  SS+ G+    +SSKM+ LMQKAKTYKRH+FVLRRGC
Sbjct  1282   VPFVRRSSSPDSAVSLDSAADKRFSSEKGLKKKTISSKMSGLMQKAKTYKRHSFVLRRGC  1341

Query  19984  NLPNTELDTPDFTSSDNGKSSTFKGKIVIN-------NYEDGVDKNLNDQNDY--GIKKV  19832
              N+ ++EL+ PDF SS N  SS    +I++N         ED   KN  D      G  +
Sbjct  1342   NMSDSELEMPDFVSSGNDNSSISTREILLNQSIEVEDEQEDFNYKNRCDSKSVLGGSIEK  1401

Query  19831  LSGNELNNLFPVVGDLKKTQSPISVAGSTAIHISEPANNKEIVNIKNGTIKMPEIFLETT  19652
              L+GN  NNLFP+VGDLKK QSP+ +A  T I   +   + +  +IKN  I+MP+I LET
Sbjct  1402   LNGNLTNNLFPIVGDLKKIQSPLPLAVLTEIPSYKDEYSNKSDSIKNSPIEMPKILLET-  1460

Query  19651  SQSCLQNLNSVYTDDIGDESILRNSVNIDVNVPTLNIFKTIGEVGDIAITTFLTTTTTAT  19472
                +C Q LN  ++DD  D++IL NS    VN PT +I KT+ +  +  +T   TTTT +
Sbjct  1461   --ACNQELNLAHSDDDVDKNILANSAKDYVNAPTFSILKTVEDASEPTMTPLHTTTTTNS  1518

Query  19471  TTIACRSLSSSGGLLVTQQCLDLPNKYEYGSRDDDDNRSQHSSRTLSSSRRQSTEDSIDT  19292
                    SL+ +  L VTQQCLDLPN   +GSR+DDDNRSQHS+RTLSSSRRQSTEDSIDT
Sbjct  1519   ------SLNVTSALWVTQQCLDLPNYPGWGSREDDDNRSQHSARTLSSSRRQSTEDSIDT  1572

Query  19291  DDEYFYYELRQLEEQEQQRTESSGITSSERLEDNEVLFSQIGLLLQNDT-----FRSFDV  19127
              DDEYFYYELRQLEEQE+QR  +S I S ER  DN+VLFSQIG LLQND      FR  +
Sbjct  1573   DDEYFYYELRQLEEQEKQRAHNSAIPSCERQNDNDVLFSQIGQLLQNDVNGGDGFRHSNG  1632
```

```
Query  19126  CNDVEDNMNYSPSESVKLRMSEVFGELKSVVKLHPEVCPYSEIHTAFEDIVDANAFKKPT  18947
              CND ED + +SPSESVKLRMSEVF ELKSVV L+P V         FE +       KPT
Sbjct  1633   CNDGEDAIIFSPSESVKLRMSEVFKELKSVVSLNPSV----NNDATFEGV----PIVKPT  1684

Query  18946  GEKFTTVCDMHKAWQDLNVDLKTSEKIPNLN--------------NKQQKSRRLEKKTFD  18809
               EK  TV D+H AWQD+N DL+ +    + N              NKQ+K RRL+KKT D
Sbjct  1685   YEKLETVSDLHSAWQDVNGDLQIAASDIDSNEDLVGNKGRETPTYNKQRKLRRLKKKTRD  1744

Query  18808  RNVISSNDEASSSSSHSENEC----VECSPKLSAEM-------KSSASSETSGPDTPNE  18662
              R +  S +  SSSSS HSENEC     +C+ K  AE        KS ASSETSGPDTP E
Sbjct  1745   RKINISKNATSSSSSCHSENECNTPLGQCTQKSVAEKDTNDISNKSEASSETSGPDTPAE  1804

Query  18661  MSDVEINETECSSKAEVDLHNLYLHGNLSEI----QNSILEFDKDHT------NEDQSST  18512
              +SDV+I+ETE   +A+   +   + GN   +    +  +L+FD DH+         Q +T
Sbjct  1805   LSDVDISETENGLRADDGQNIDNMRGNSGSLLKVNRKYLLQFDVDHSLINQPLETSQYNT  1864

Query  18511  HVLESISNGSVQS---PFETKNVISPKSDGDGSQAVGSNGSAAGLGSSKWKLLKTLKERK  18341
              H+LE+I++ S+ S     ++K ++S  S  DGSQAVG N +A GL SSKWKLLKTLKERK
Sbjct  1865   HMLENITSASIPSQNRQIDSKTLMSQSSHADGSQAVG-NETAVGLSSSKWKLLKTLKERK  1923

Query  18340  IEEKNNQEKMKEDEITKDKEK  18278
              IEEKNNQ+K+KEDE+ KD++K
Sbjct  1924   IEEKNNQDKIKEDEMIKDRDK  1944
```

unc-13:7_2147_0

Sequence ID: Query_227301 **Length:** 44 **Number of Matches:** 1

Range 1: 1 to 44 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
90.9 bits(224)	1e-26	41/44(93%)	42/44(95%)	0/44(0%)	-3

```
Query  16381  NGGGNGEVGIRNNGHPGDNPFYSNIDSMPDIRPRRKSIPLVSEL  16250
              NGGG GEVG+R NGHPGDNPFYSNIDSMPDIRPRRKSIPLVSEL
Sbjct  1      NGGGPGEVGLRTNGHPGDNPFYSNIDSMPDIRPRRKSIPLVSEL  44
```

unc-13:9_2147_0

Sequence ID: Query_242027 **Length:** 25 **Number of Matches:** 1

Range 1: 3 to 25 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
30.4 bits(67)	1e-05	12/23(52%)	17/23(73%)	0/23(0%)	-1

```
Query  14766  YTSIPTKKEQCNLLSLHDNFQTT  14698
              Y + P+K +QC LLSL +N +TT
Sbjct  3      YPNAPSKMDQCYLLSLQENVETT  25
```

unc-13:10_2147_0

Sequence ID: Query_230785 **Length:** 25 **Number of Matches:** 1

Range 1: 1 to 25 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
50.1 bits(118)	1e-12	24/25(96%)	25/25(100%)	0/25(0%)	-3

```
Query  14263  TMAATKRNAGLTSAVPRATLNDEEL  14189
              TMAATKRNAGLTSAVPRATLNDE+L
Sbjct  1      TMAATKRNAGLTSAVPRATLNDEDL  25
```

unc-13:11_2147_0

Sequence ID: Query_27585 Length: 73 Number of Matches: 1

Range 1: 1 to 73 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
164 bits(414)	1e-51	73/73(100%)	73/73(100%)	0/73(0%)	-2

```
Query  14066  KMHVYKKALQALIYPISSTTPHNFLLWTATSPTYCYECEGLLWGIARQGVRCTECGVKCH  13887
              KMHVYKKALQALIYPISSTTPHNFLLWTATSPTYCYECEGLLWGIARQGVRCTECGVKCH
Sbjct  1      KMHVYKKALQALIYPISSTTPHNFLLWTATSPTYCYECEGLLWGIARQGVRCTECGVKCH  60

Query  13886  EKCKDLLNADCLQ  13848
              EKCKDLLNADCLQ
Sbjct  61     EKCKDLLNADCLQ  73
```

unc-13:12_2147_2

Sequence ID: Query_236811 Length: 39 Number of Matches: 1

Range 1: 1 to 39 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
77.0 bits(188)	8e-22	38/39(97%)	39/39(100%)	0/39(0%)	-3

```
Query  13786  AAEKSSKHGAEDKANSIITAMKERMKQREREKPEIFELI  13670
              AAEKSSKHGAEDKANSIITAMK+RMKQREREKPEIFELI
Sbjct  1      AAEKSSKHGAEDKANSIITAMKDRMKQREREKPEIFELI  39
```

unc-13:13_2147_1

Sequence ID: Query_86443 Length: 35 Number of Matches: 2

Range 1: 1 to 35 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
68.9 bits(167)	5e-19	33/35(94%)	34/35(97%)	0/35(0%)	-3

```
Query  13489  AVFSVEEKSHTGHMKAMKQSVLDGTSKWSAKIAIT  13385
              AVFSVEEKSH GHMKA+KQSVLDGTSKWSAKIAIT
Sbjct  1      AVFSVEEKSHAGHMKAVKQSVLDGTSKWSAKIAIT  35
```

unc-13:14_2147_1

Sequence ID: Query_117243 Length: 35 Number of Matches: 2

Range 1: 1 to 35 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
77.0 bits(188)	7e-22	35/35(100%)	35/35(100%)	0/35(0%)	-2

```
Query  12734  MTFGVDPDTHIDSLEQAEHATVEGTSKWSCKLTIT  12630
              MTFGVDPDTHIDSLEQAEHATVEGTSKWSCKLTIT
Sbjct  1      MTFGVDPDTHIDSLEQAEHATVEGTSKWSCKLTIT  35
```

unc-13:15_2147_2

Sequence ID: Query_37195 **Length:** 46 **Number of Matches:** 1

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

Score	Expect	Identities	Positives	Gaps	Frame
98.6 bits(244)	3e-29	46/46(100%)	46/46(100%)	0/46(0%)	-3

```
Query  9451  ICAQGLIAKDKSGTSDPYVTVQVSKVKKRTRTMPQELNPVWNEKFH  9314
             ICAQGLIAKDKSGTSDPYVTVQVSKVKKRTRTMPQELNPVWNEKFH
Sbjct  1     ICAQGLIAKDKSGTSDPYVTVQVSKVKKRTRTMPQELNPVWNEKFH  46
```

unc-13:16_2147_1

Sequence ID: Query_12485 **Length:** 108 **Number of Matches:** 1

Range 1: 1 to 108 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
224 bits(570)	5e-72	107/108(99%)	108/108(100%)	0/108(0%)	-3

```
Query  7687  ECHNSSDRIKVRVWDEDNDLKSKLRQKLTRESDDFLGQTIIEVRTLSGEMDVWYNLEKRT  7508
             ECHNSSDRIKVRVWDEDNDLKSKLRQKLTRESDDFLGQTIIEVRTLSGEMDVWYNLEKRT
Sbjct  1     ECHNSSDRIKVRVWDEDNDLKSKLRQKLTRESDDFLGQTIIEVRTLSGEMDVWYNLEKRT  60

Query  7507  DKSAVSGAIRLHISVEIKGEEKVAPYHVQYTCLHENLFHYLCEENSGM  7364
             DKSAVSGAIRLHISVEIKGEEKVAPYHVQYTCLHENLFHYLCEEN+GM
Sbjct  61    DKSAVSGAIRLHISVEIKGEEKVAPYHVQYTCLHENLFHYLCEENTGM  108
```

unc-13:17_2147_0

Sequence ID: Query_105339 **Length:** 91 **Number of Matches:** 1

Range 1: 1 to 91 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
189 bits(480)	3e-60	91/91(100%)	91/91(100%)	0/91(0%)	-1

```
Query  7287  VKLPTQKGDDAWKLYFDEIPEEIVDEFSMRYGIENIYQAMTHFHCLSAKYLCPGVPAVMS  7108
             VKLPTQKGDDAWKLYFDEIPEEIVDEFSMRYGIENIYQAMTHFHCLSAKYLCPGVPAVMS
Sbjct  1     VKLPTQKGDDAWKLYFDEIPEEIVDEFSMRYGIENIYQAMTHFHCLSAKYLCPGVPAVMS  60

Query  7107  TLLANINAYYAHTTASSAVSASDRFAASNFG  7015
             TLLANINAYYAHTTASSAVSASDRFAASNFG
Sbjct  61    TLLANINAYYAHTTASSAVSASDRFAASNFG  91
```

unc-13:18_2147_0

Sequence ID: Query_62159 **Length:** 53 **Number of Matches:** 1

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

Score	Expect	Identities	Positives	Gaps	Frame
106 bits(265)	6e-32	53/53(100%)	53/53(100%)	0/53(0%)	-3

```
Query  6949  KEKFVKLLDQLHNSLRIDLSMYRNNFPASSPEKLMDLKSTVDLLTSITFFRMK  6791
             KEKFVKLLDQLHNSLRIDLSMYRNNFPASSPEKLMDLKSTVDLLTSITFFRMK
Sbjct  1     KEKFVKLLDQLHNSLRIDLSMYRNNFPASSPEKLMDLKSTVDLLTSITFFRMK  53
```

unc-13:19_2147_0

Sequence ID: Query_76833 Length: 42 Number of Matches: 1

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

Score	Expect	Identities	Positives	Gaps	Frame
90.9 bits(224)	1e-26	42/42(100%)	42/42(100%)	0/42(0%)	-3

```
Query  6733  VQELSSPPRASTVVKDCVKACLRSTYQFLFENCYELYNREFQ  6608
             VQELSSPPRASTVVKDCVKACLRSTYQFLFENCYELYNREFQ
Sbjct  1     VQELSSPPRASTVVKDCVKACLRSTYQFLFENCYELYNREFQ  42
```

unc-13:20_2147_0

Sequence ID: Query_96419 Length: 56 Number of Matches: 1

Range 1: 1 to 56 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
117 bits(293)	1e-35	55/56(98%)	55/56(98%)	0/56(0%)	-3

```
Query  4990  VDPNETKRAPDDHEPKLDSVDFWHKLIALIVSVIDEDKNSYGTVLNQFPQELNIGQ  4823
             VDPNE KRAPDDHEPKLDSVDFWHKLIALIVSVIDEDKNSYGTVLNQFPQELNIGQ
Sbjct  1     VDPNEAKRAPDDHEPKLDSVDFWHKLIALIVSVIDEDKNSYGTVLNQFPQELNIGQ  56
```

unc-13:21_2147_0

Sequence ID: Query_190077 Length: 94 Number of Matches: 1

Range 1: 1 to 94 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
206 bits(525)	3e-66	94/94(100%)	94/94(100%)	0/94(0%)	-2

```
Query  4766  LSASSMWHLFAVDMKYALEEHEQHRLCKSSAYMNLHFRVKWLYSNYVKEVPPYKGAVPDY  4587
             LSASSMWHLFAVDMKYALEEHEQHRLCKSSAYMNLHFRVKWLYSNYVKEVPPYKGAVPDY
Sbjct  1     LSASSMWHLFAVDMKYALEEHEQHRLCKSSAYMNLHFRVKWLYSNYVKEVPPYKGAVPDY  60

Query  4586  PAWFEPFVMQWLNENDDVSLEYLHGAFKRDKKDG  4485
             PAWFEPFVMQWLNENDDVSLEYLHGAFKRDKKDG
Sbjct  61    PAWFEPFVMQWLNENDDVSLEYLHGAFKRDKKDG  94
```

unc-13:22_2147_0

Sequence ID: Query_125245 Length: 74 Number of Matches: 1

Range 1: 1 to 74 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
149 bits(377)	1e-46	71/74(96%)	74/74(100%)	0/74(0%)	-2

```
Query  4205  FQKSSEHALFSNSVVDVFTQLTQCFDVVSKLECPDPEIWKRYMKRFAKTIVKVLIAYADI  4026
             FQKSSEHALFSNSVVDVFTQLTQCFDVVSKLECPDPEIWKRYM+RFAKTIVKVLIAYADI
Sbjct  1     FQKSSEHALFSNSVVDVFTQLTQCFDVVSKLECPDPEIWKRYMRRFAKTIVKVLIAYADI  60

Query  4025  VKIEFPDHMKDERI  3984
             VK+EFP+HMKDERI
Sbjct  61    VKLEFPEHMKDERI  74
```

unc-13:23_2147_0

Sequence ID: Query_139253 **Length:** 55 **Number of Matches:** 1

Range 1: 1 to 55 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
108 bits(269)	2e-32	55/55(100%)	55/55(100%)	0/55(0%)	-2

```
Query  3887  ACILMNNIQQLRVQLEKMFESMGGDKLEEDAANILKELQQNLNSALDDLASQFAI  3723
             ACILMNNIQQLRVQLEKMFESMGGDKLEEDAANILKELQQNLNSALDDLASQFAI
Sbjct  1     ACILMNNIQQLRVQLEKMFESMGGDKLEEDAANILKELQQNLNSALDDLASQFAI  55
```

unc-13:24_2147_1

Sequence ID: Query_176483 **Length:** 97 **Number of Matches:** 1

Range 1: 1 to 97 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
187 bits(475)	2e-59	97/97(100%)	97/97(100%)	0/97(0%)	-1

```
Query  3660  LEPRITQSVRELGDMLLSIKGGSGTLAAGNLAAQRNAVAVEADEVLRPLMDLLDGSLTLY  3481
             LEPRITQSVRELGDMLLSIKGGSGTLAAGNLAAQRNAVAVEADEVLRPLMDLLDGSLTLY
Sbjct  1     LEPRITQSVRELGDMLLSIKGGSGTLAAGNLAAQRNAVAVEADEVLRPLMDLLDGSLTLY  60

Query  3480  AQSCEKTVLKRLLKELWKIVMRILEKTIVLPPMTDKT  3370
             AQSCEKTVLKRLLKELWKIVMRILEKTIVLPPMTDKT
Sbjct  61    AQSCEKTVLKRLLKELWKIVMRILEKTIVLPPMTDKT  97
```

unc-13:25_2147_0

Sequence ID: Query_233175 **Length:** 119 **Number of Matches:** 1

Range 1: 1 to 119 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
231 bits(590)	1e-74	117/119(98%)	118/119(99%)	0/119(0%)	-2

```
Query  3314  MMFKHLTDNAKNLASNAKIEDMGRLFKSHMAGKQDVKSALSGVMDISKEVEKNLSPKQCA  3135
             MMFKHLTDNAKNLASNAKIEDMGRLFKSHMAGKQDVKSALSGVMDISKEVEKNLSPKQCA
Sbjct  1     MMFKHLTDNAKNLASNAKIEDMGRLFKSHMAGKQDVKSALSGVMDISKEVEKNLSPKQCA  60

Query  3134  VLDVALDTIKQYFHAGGNGLKKTFLEKSSELQSLRYALSLYTQMTDTLIKTFISSQIHE  2958
             VLDVALDTIKQYFHAGGNGLKKTFLEKS ELQSLRYALSLYTQMTDTLIKTFISSQ+HE
Sbjct  61    VLDVALDTIKQYFHAGGNGLKKTFLEKSPELQSLRYALSLYTQMTDTLIKTFISSQVHE  119
```

unc-13:26_2147_2

Sequence ID: Query_245081 **Length:** 32 **Number of Matches:** 1

Range 1: 1 to 32 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
65.9 bits(159)	5e-18	31/32(97%)	31/32(96%)	0/32(0%)	-1

```
Query  2136  DAENSEESVGEISVQIDLFSHPGTGEHKVNVK  2041
             D ENSEESVGEISVQIDLFSHPGTGEHKVNVK
Sbjct  1     DLENSEESVGEISVQIDLFSHPGTGEHKVNVK  32
```

unc-13:27_2147_2

Sequence ID: Query_7299 **Length:** 53 **Number of Matches:** 1

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

Score	Expect	Identities	Positives	Gaps	Frame
114 bits(286)	8e-35	53/53(100%)	53/53(100%)	0/53(0%)	-2

```
Query  1976  VAANDLKWQIPSGMFRPFVDINLIGPHLQEKKRKFATKSKSNNWSPKYNESFS  1818
             VAANDLKWQIPSGMFRPFVDINLIGPHLQEKKRKFATKSKSNNWSPKYNESFS
Sbjct  1     VAANDLKWQIPSGMFRPFVDINLIGPHLQEKKRKFATKSKSNNWSPKYNESFS  53
```

unc-13:28_2147_1

Sequence ID: Query_219879 **Length:** 42 **Number of Matches:** 1

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

Score	Expect	Identities	Positives	Gaps	Frame
89.7 bits(221)	3e-26	42/42(100%)	42/42(100%)	0/42(0%)	-3

```
Query  1753  TIGNEEQLDFFELHICVKDYCFARDDRLVGVAVIPLKDISEK  1628
             TIGNEEQLDFFELHICVKDYCFARDDRLVGVAVIPLKDISEK
Sbjct  1     TIGNEEQLDFFELHICVKDYCFARDDRLVGVAVIPLKDISEK  42
```

unc-13:29_2147_0

Sequence ID: Query_237609 **Length:** 54 **Number of Matches:** 1

Range 1: 1 to 54 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
110 bits(274)	4e-33	53/54(98%)	54/54(100%)	0/54(0%)	-2

```
Query  1574  GSVACWLPLMRRIEMDETGWTILRILSQRNNDDVAKEFVKLKSEIRQEPTMGT*  1413
             GSVACWLPLMRRIEMDETGWTILRILSQRNND+VAKEFVKLKSEIRQEPTMGT*
Sbjct  1     GSVACWLPLMRRIEMDETGWTILRILSQRNNDEVAKEFVKLKSEIRQEPTMGT*  54
```

Figure 29: blastx alignments of all* D. melanogaster *exons to* D. eugracilis *contig 24. *Subject is* D. melanogaster *exons and Query is* D. eugracilis *contig 24.*

Start Codons

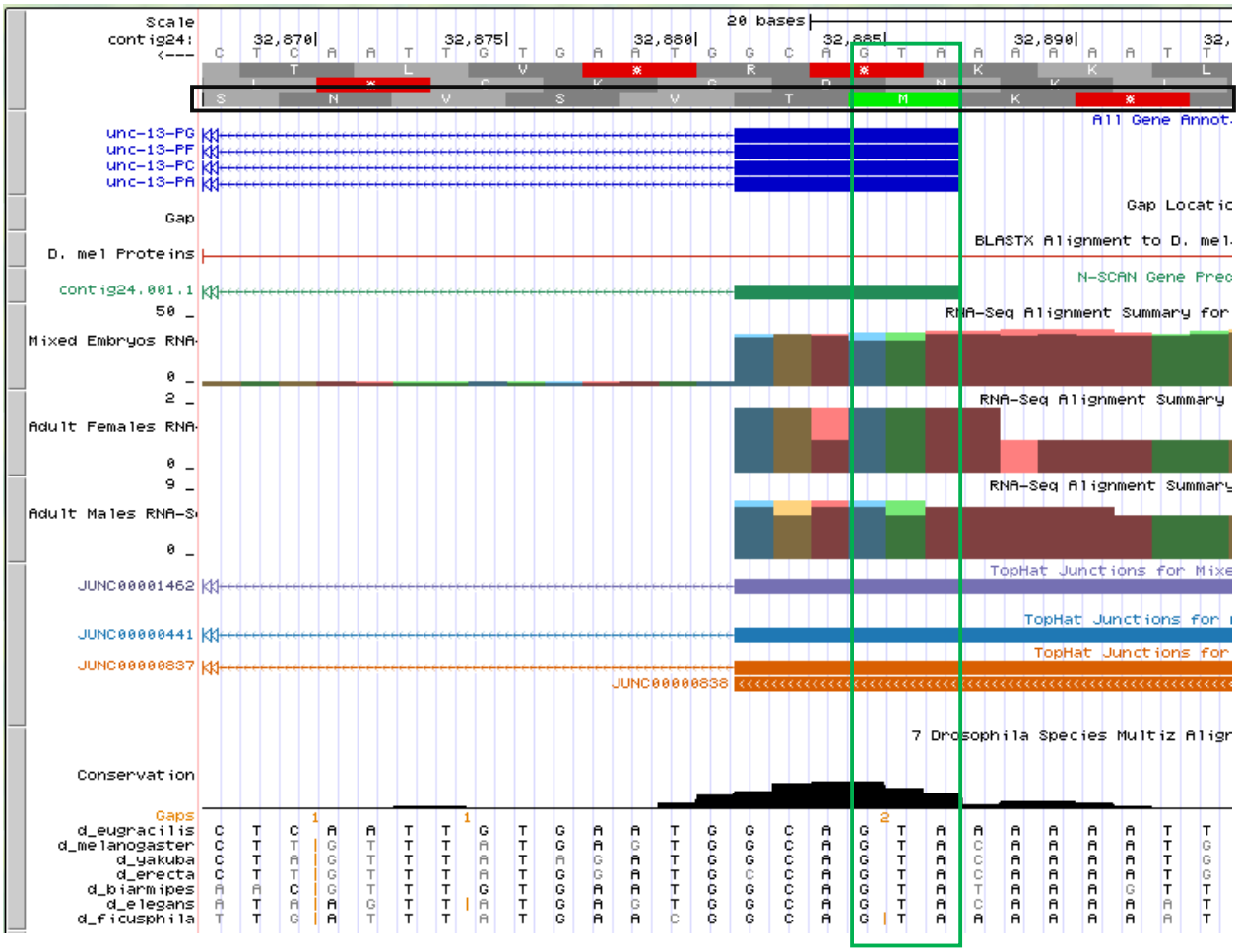


Figure 30: Proposed first start codon. *This start codon corresponds to the A, C, D, F, and G isoforms of* unc-13.

Among the 7 isoforms of *unc-13*, there are two unique start codons. The first annotated *D. melanogaster* exon, 1_2147_0, failed to align to *D. eugracilis* contig 24 by NCBI blastx. However, the exon was identified in *D. eugracilis* using RNA-seq and TopHat junction tracks to locate the splice donor site corresponding to the splice acceptor site of the second exon. The 2 AA sequence upstream of the splice donor site is MT, identical to the first annotated exon in *D. melanogaster* (Fig. 30). The A, C, D, F, and G isoforms share this start codon.

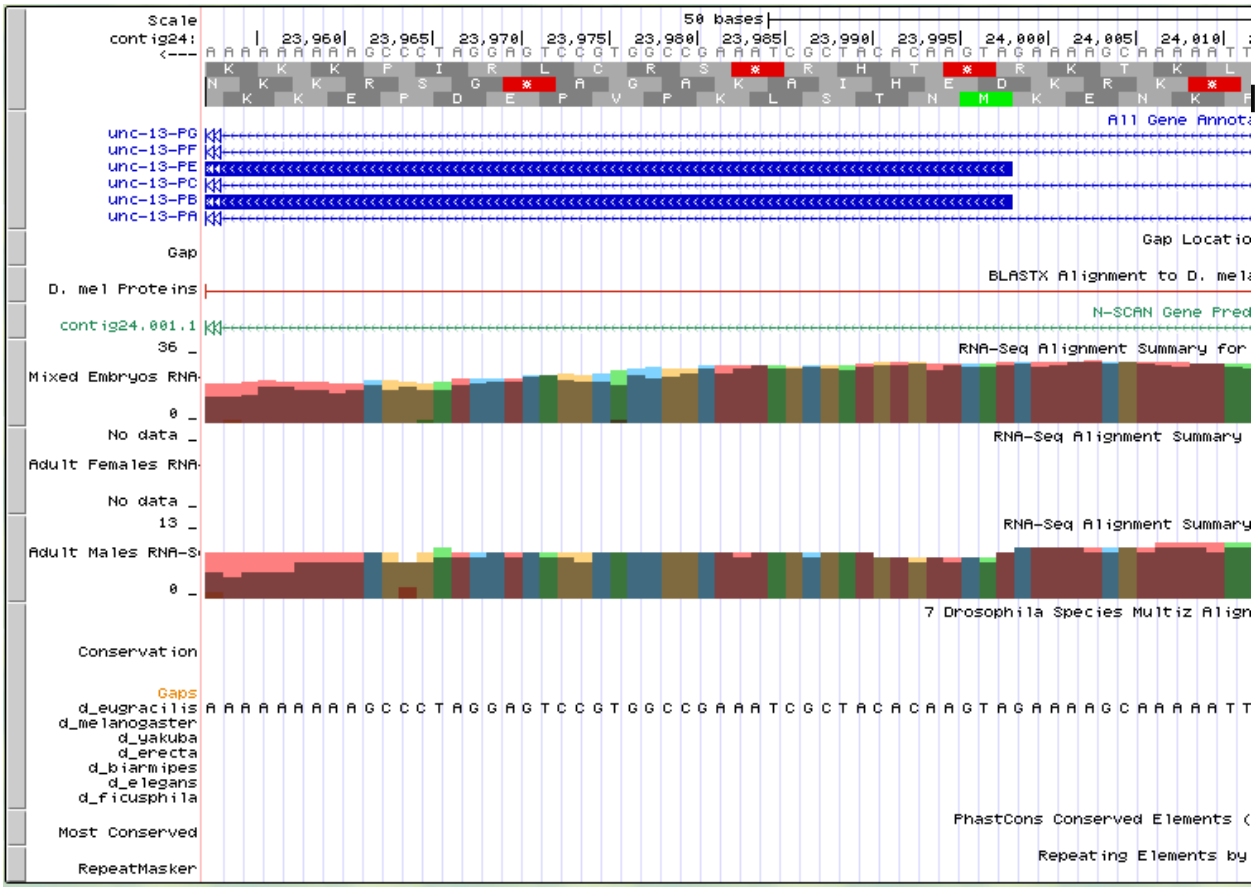


Figure 31: Proposed second start codon. *This start codon corresponds to the B and E isoforms of* unc-13.

The first exon of both *unc-13-PB* and *unc-13-PE* in *D. melanogaster* is 6_2147_0, which aligned by blastx search to position 23998 in contig 24 (Fig. 29). The methionine at this position was identified as the start codon for the B and E isoforms of *unc-13* (Fig. 31).

Identification of *unc-13* exons

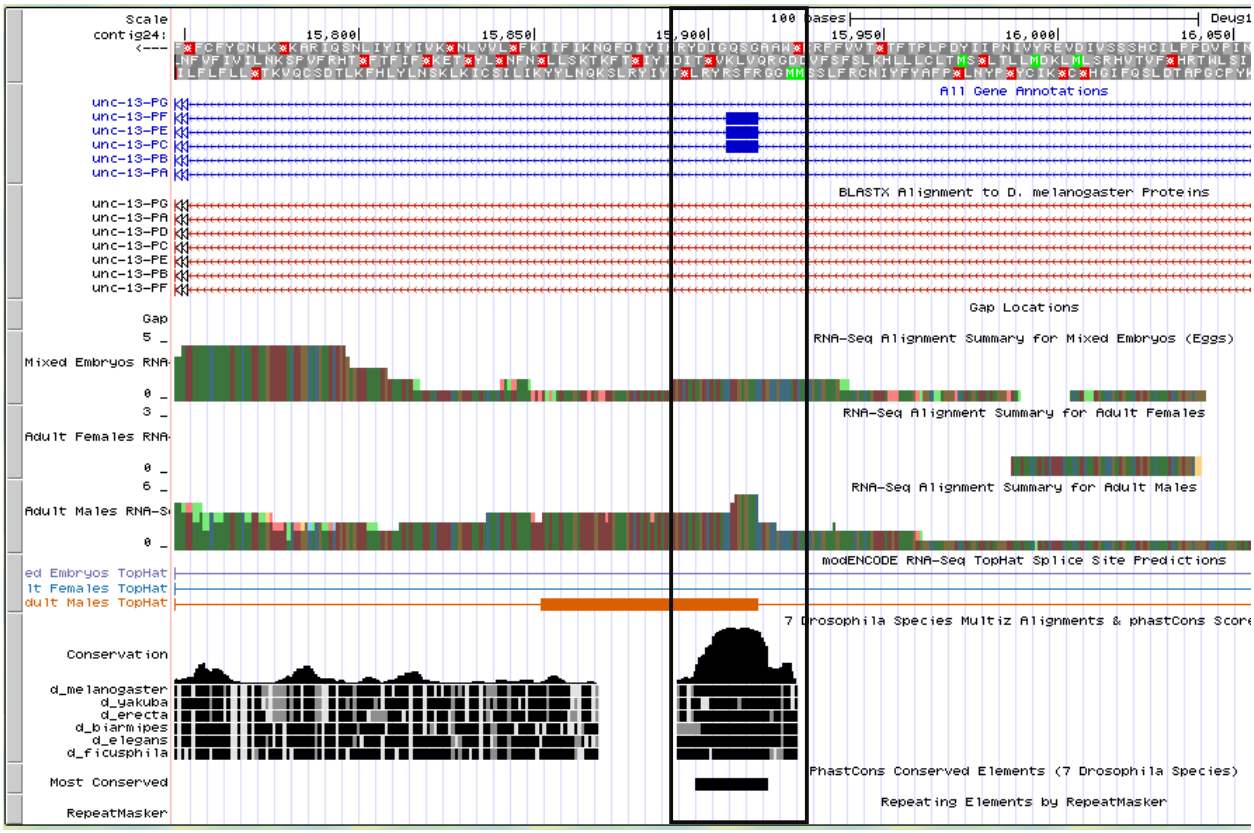


Figure 32: Identification of* D. eugracilis *ortholog of 8_2147_0. *The boxed region shows a highly conserved region with no BLAST track gene prediction. Both splice sites are in phase 0, which is the appropriate splice donor and acceptor phase.*

To form a complete gene model of *unc-13* in *D. eugracilis*, all CDSs must be correctly identified. While blastx alignment of annotated *D. melanogaster* exons to *D. eugracilis* contig 24 was unable to locate an exon corresponding to exon 8_2147_0, a corresponding exon was found using UCSC Genome Browser evidence tracks. The two exons adjacent to 8_2147_0 in *D. melanogaster* are 7_2147_0 and 9_2147_0. The first base of 9_2147_0 aligns by blastx to position 14766 in *D. eugracilis* while the last base of 7_2147_0 aligns to position 16250 (Fig. 33). Within the region between these two genes, a peak in conservation is located around position 15900. As shown in Figure 34, this region contains the highly-conserved sequence VLK flanked by conserved splice acceptor and splice donor sequences. This sequence is identical to 8_2147_0 in *D. melanogaster*.

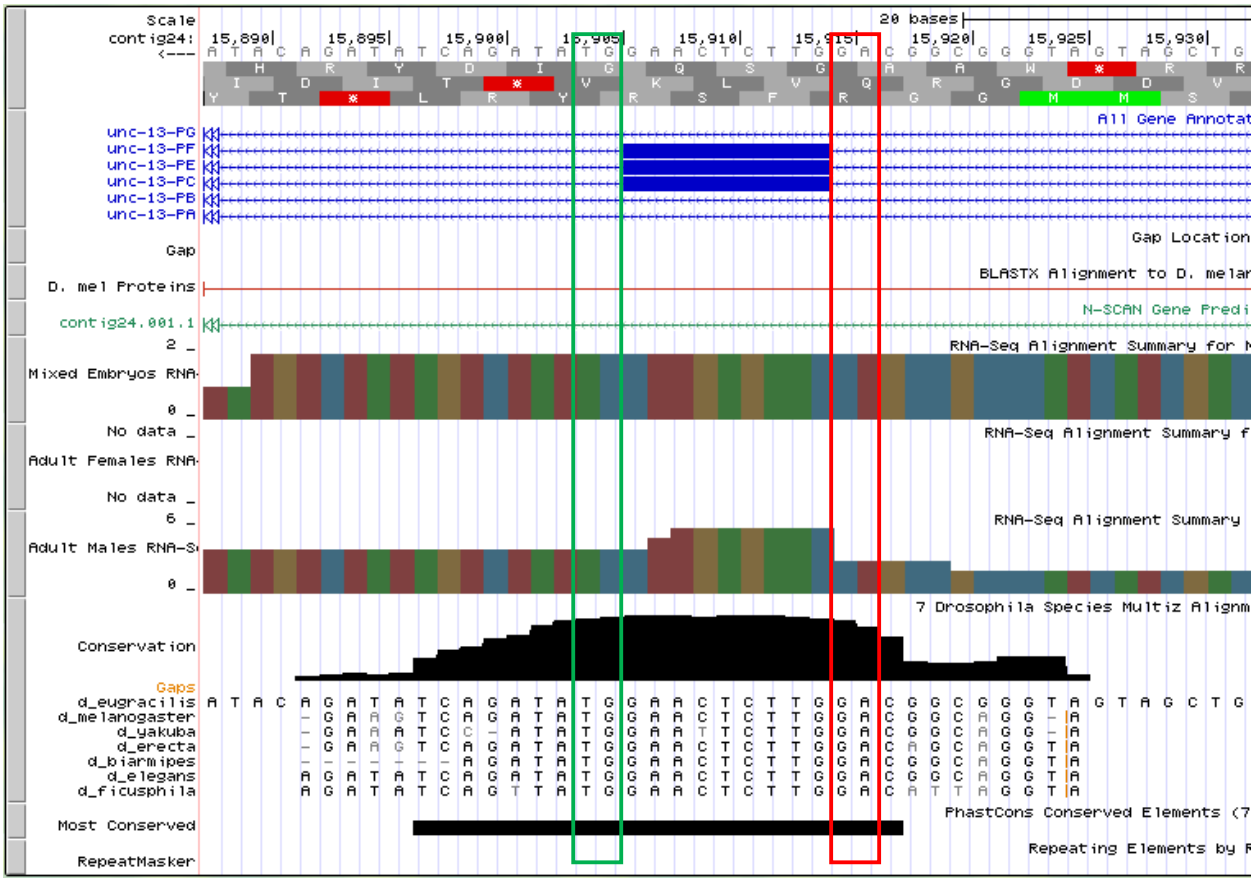


Figure 33: Identification of* D. eugracilis *ortholog of 8_2147_0. *Splice acceptor (red box) and splice donor (green box) sites are both conserved across* Drosophila.

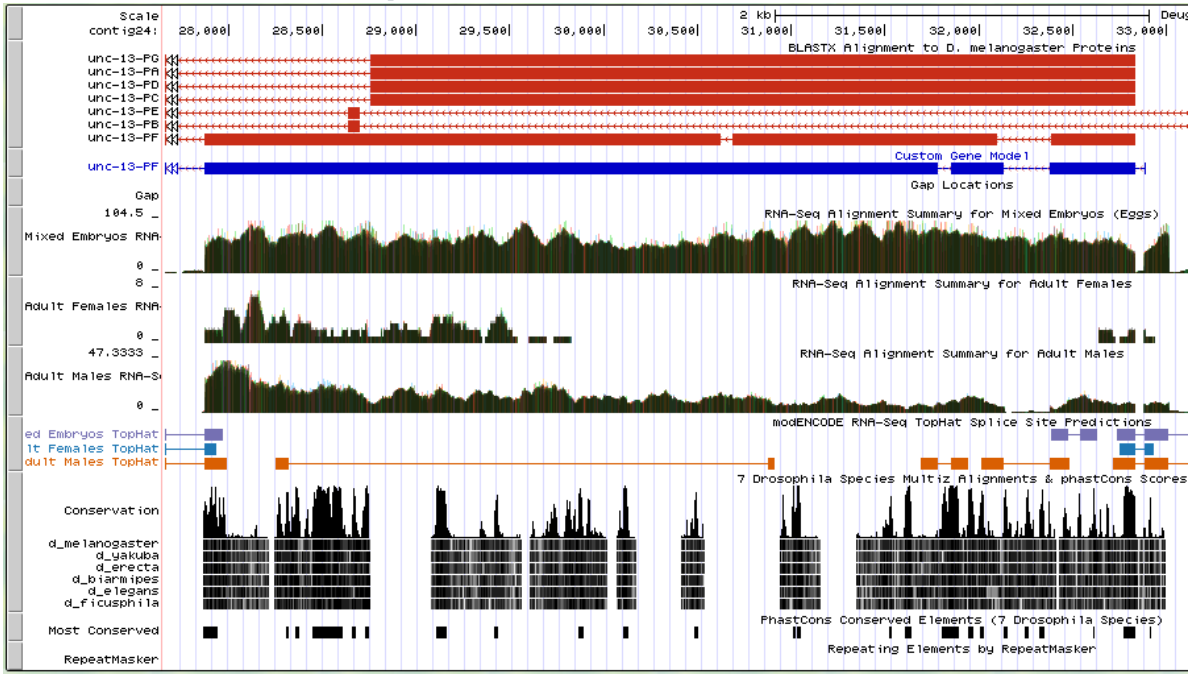


***Figure 34: Annotation of the F isoform in* D. eugracilis unc-13.** *The BLAST track of the GEP mirror of the UCSC Genome Browser did not correctly identify the F isoform.*

Exons 2_2147_0, 3_2147_2, and 4_2147_2 in *D. melanogaster* are all unique to *unc-13-PF*. The three exons appear to result from alternative splicing of 5_2147_0 because the first base of 2_2147_0 and the last base of 4_2147_2 are identical to the first and last base of 5_2147_0. Due to being spliced from the same genomic region as 5_2147_0, which spans across the introns between 2_2147_0, 3_2147_2, and 4_2147_2, RNA-seq summary data is not very informative for the F isoform. Additionally, the BLAST alignment track of the Genome Browser did not appear to correctly show the exon position (Fig. 34). The blastx alignment of these three exons to *D. eugracilis* contig 24 were the primary evidence used to add these exons to the gene model. The splice donor site of exon 2 of *unc-13-PF* in *D. eugracilis* was found immediately downstream of the alignment of the blastx alignment of 2_2147_0 at position 32381 (Fig. 35). Exon 3_2147_2 aligned to *D. eugracilis* starting from the ninth amino acid at position 32101 and a corresponding splice acceptor site was found upstream of positon 32136, 8 AA upstream of the alignment (Fig. 36). The splice donor site for exon 3 and splice acceptor site for exon 4 of *D. eugracilis unc-13-PF* were both adjacent to the blastx alignment of the last amino acid of 3_2147_ 2 and the first amino acid of 4_2147_2, respectively. The splice donor site downstream of position 31853 and the splice acceptor site upstream of 31777 were both also supported by TopHat junctions (Fig. 37).

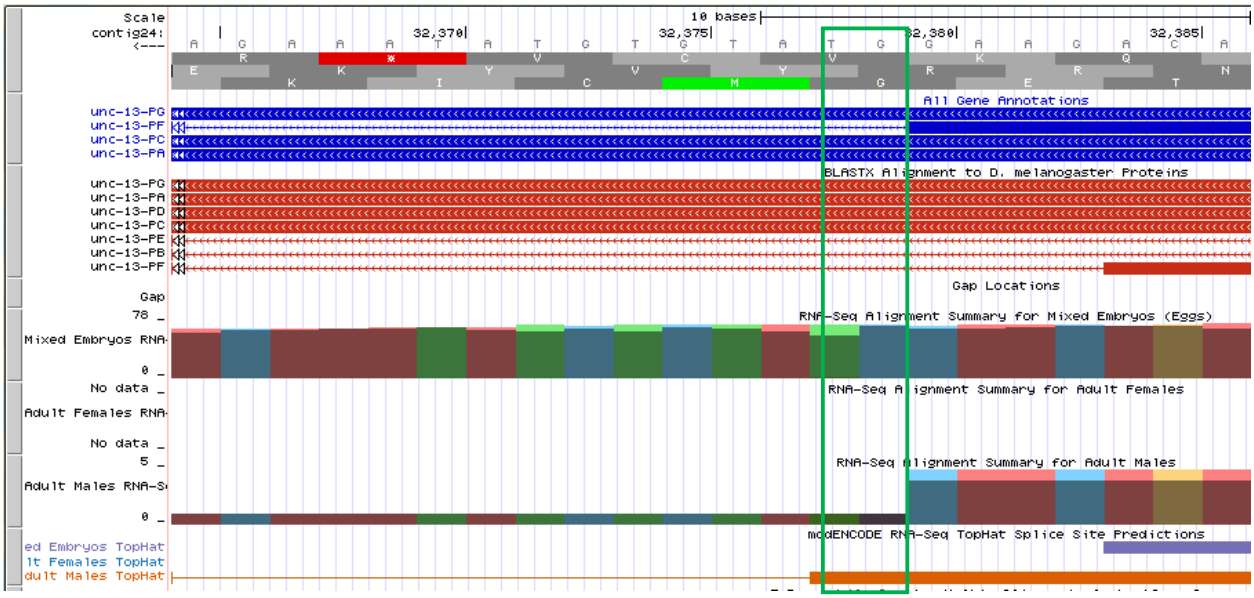


***Figure 35: Splice donor site of exon 2 in* unc-13-PF.** *The splice donor site (green box) was located based on blastx alignment.*

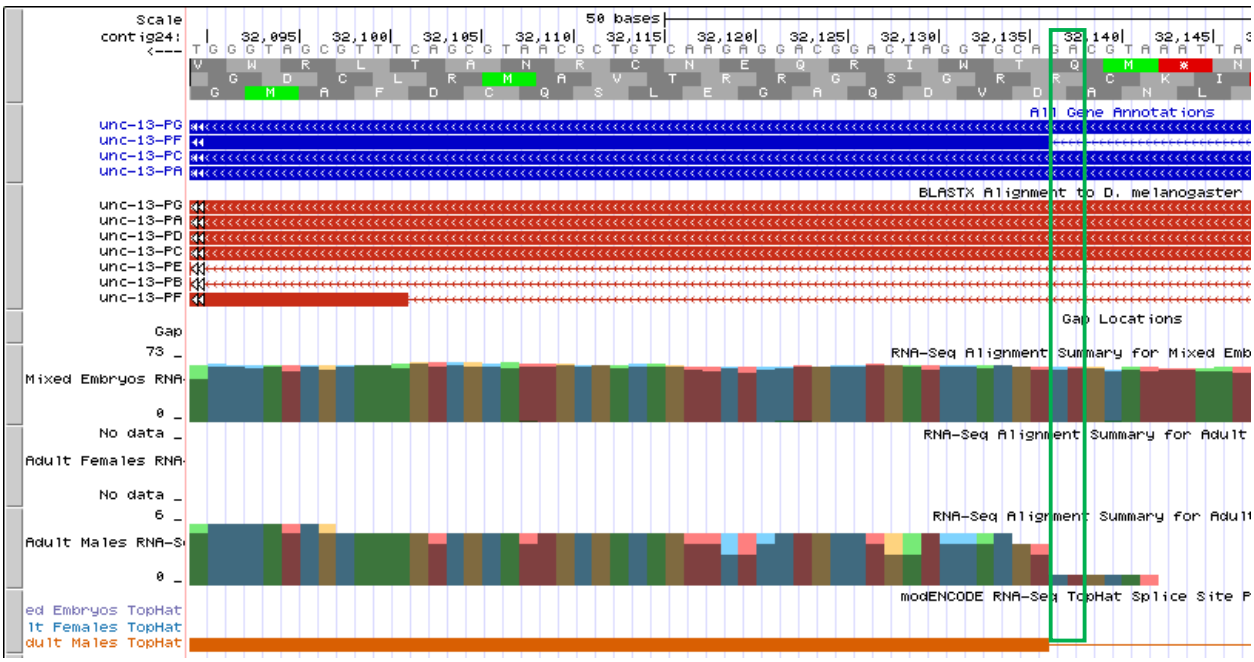


***Figure 36: Splice acceptor site of exon 3 in* unc-13-PF.** *The splice acceptor site (green box) was located based on blastx alignment.*

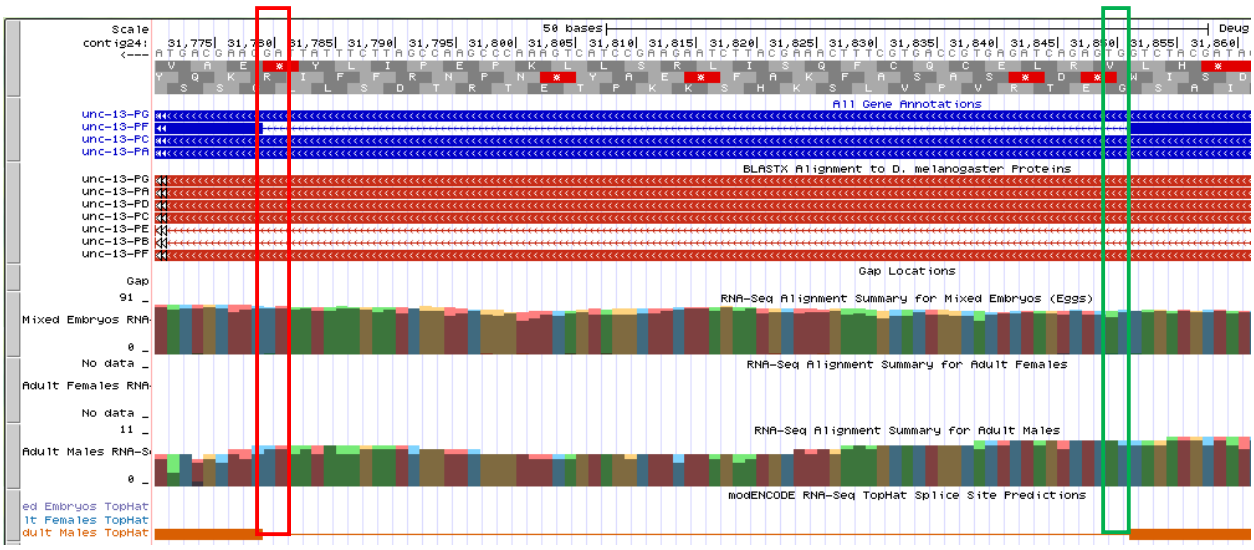


***Figure 37: Splice donor site of exon 3 and splice acceptor site of exon 4 in* unc-13-PF.** *The splice donor site (green box) and splice acceptor site (red box) were located based on blastx alignment. The splice sites are also supported by adult male TopHat junctions.*

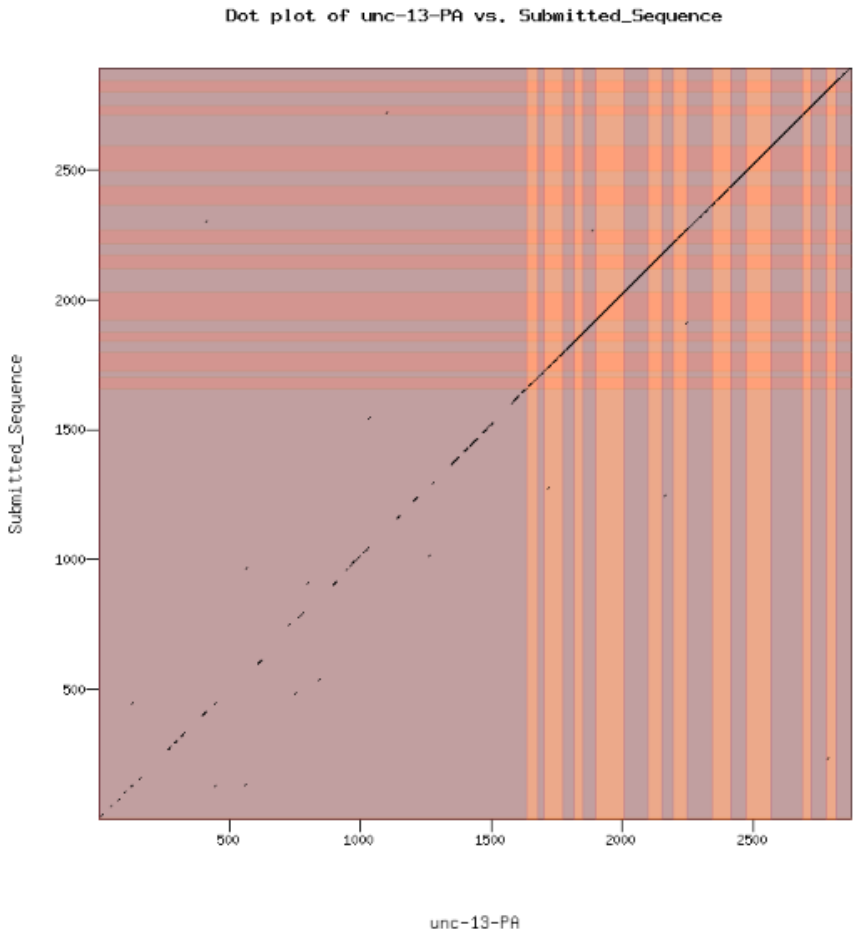
Verification of *unc-13* Gene Model

Flybase_ID	3' splice site	3' Phase	5' splice site	5' Phase	Range	Frame
1_2147_0	Start Codon		32881-32880	0	32887-32882	-3
2_2147_0	32835-32834	0	32379-32378	2	32833-32380	-3
3_2147_2	32138-32137	1	31851-31850	2	32136-31852	-3
4_2147_2	31781-31780	1	27874-27873	0	31779-27875	-3
5_2147_0	32835-32834	0	27874-27873	0	32833-27875	-3
6_2147_0	24051-24050	0	18277-18276	0	23998-18278	-3
7_2147_0	16383-16382	0	16249-16248	0	16381-16250	-3
8_2147_0	15916-15915	0	15905-15904	0	15914-15906	-2
9_2147_0	14768-14767	0	14697-14696	0	14766-14698	-1
10_2147_0	14265-14264	0	14188-14187	0	14263-14189	-3
11_2147_0	14068-14067	0	13846-13845	2	14066-13847	-2
12_2147_2	13790-13789	1	13667-13666	1	13788-13668	-3
13_2147_1	13492-13491	2	13383-13382	2	13490-13384	-3
14_2147_1	12737-12736	2	12628-12627	2	12735-12629	-2
15_2147_2	9455-9454	1	9311-9310	1	9453-9312	-3
16_2147_1	7690-7689	2	7363-7362	0	7688-7364	-3
17_2147_0	7289-7288	0	7014-7013	0	7287-7015	-1
18_2147_0	6951-6950	0	6790-6789	0	6949-6791	-3
19_2147_0	6735-6734	0	6607-6606	0	6733-6608	-3
20_2147_0	4992-4991	0	4822-4821	0	4990-4823	-3
21_2147_0	4768-4767	0	4484-4483	0	4766-4485	-2

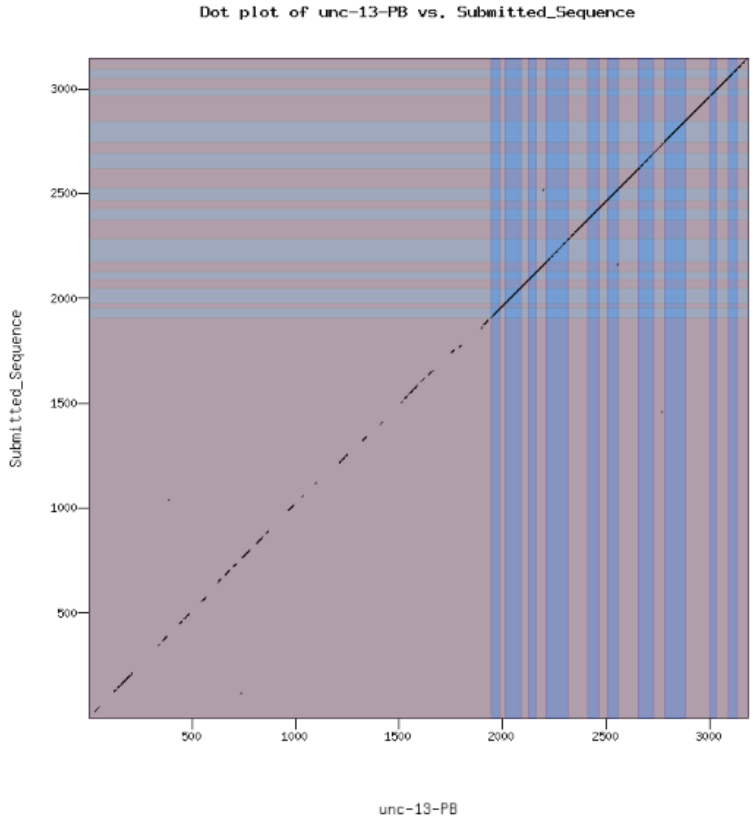
22_2147_0	4207-4206	0	3983-3982	0	4205-3984	-2
23_2147_0	3889-3888	0	3720-3719	1	3887-3721	-2
24_2147_1	3663-3662	2	3369-3368	0	3661-3370	-1
25_2147_0	3316-3315	0	2956-2955	2	3314-2957	-2
26_2147_2	2140-2139	1	2039-2038	2	2138-2040	-1
27_2147_2	1980-1979	1	1815-1814	1	1978-1816	-2
28_2147_1	1756-1755	2	1627-1626	0	1754-1628	-3
29_2147_0	1576-1575	0	Stop Codon		1574-1416	-2

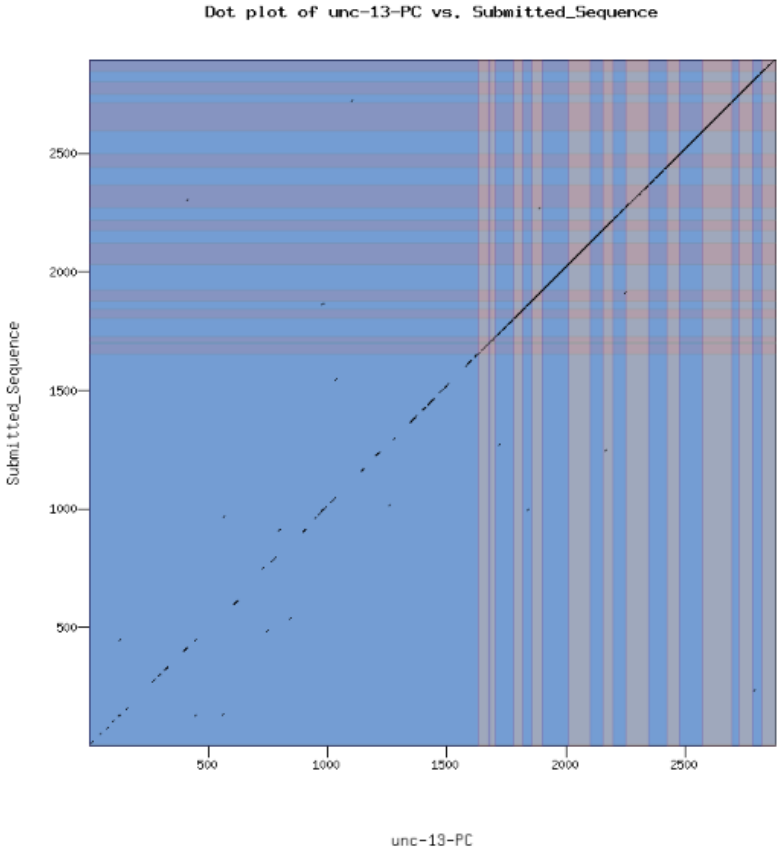
Table 5: Proposed* D. eugracilis unc-13 *exons and splice sites. *Exons 13_2147_1 (blue box) and 14_2147_1 (yellow box) are not present in the same isoforms. All exons passed Gene Model Checker plausibility criteria.*

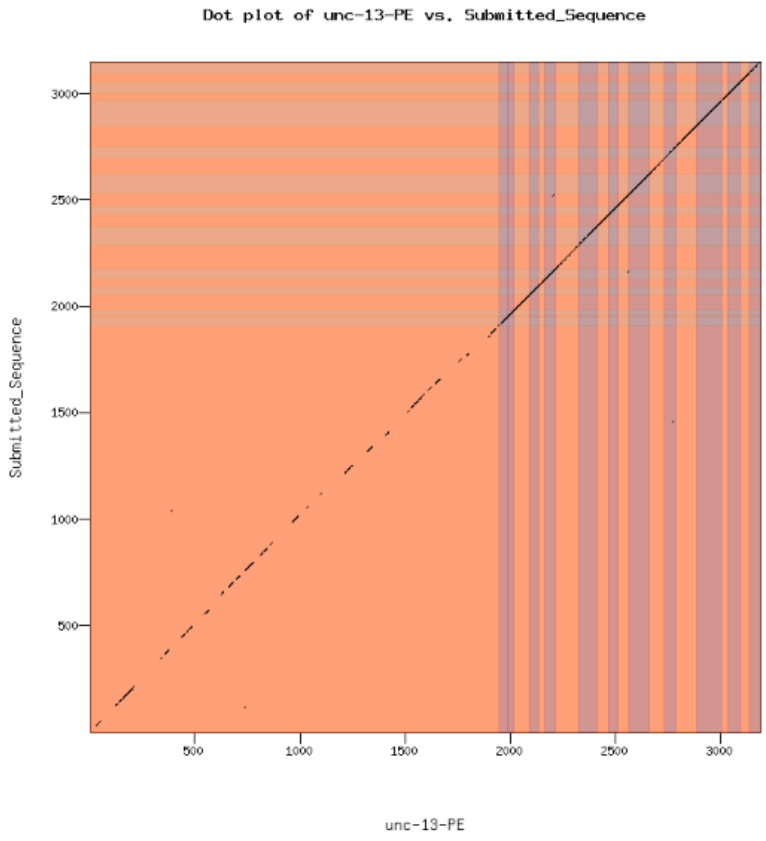
Table 5 shows the proposed exons locations for *D. eugracilis unc-13*. The exon positions achieve the same criteria for biological plausibility as the exons in *eIF4G.* All isoforms pass Gene Model Checker's criteria. Isoforms A and D have identical coding sequences and therefore are both described by the Gene Model Checker results of the A isoform. For all isoforms, the C

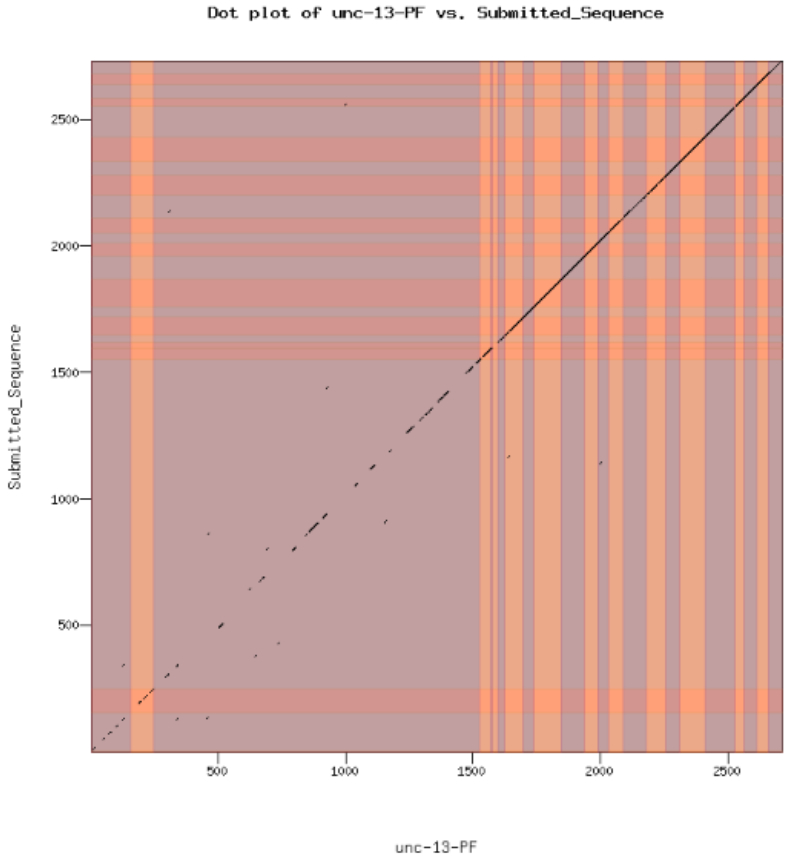


terminal half is more highly conserved, indicating that this region contains the functional domain of the protein (Fig. 38).


Dot plot of unc-13-PB vs. Submitted_Sequence
Submitted_Sequence
3000
2500
2000
1500
1000
500
500
1000
1500
2000
2500
3000
unc-13-PB


Dot plot of unc-13-PC vs. Submitted_Sequence
Submitted_Sequence
2500
2000
1500
1000
500
500
1000
1500
2000
2500
unc-13-PC


Dot plot of unc-13-PE vs. Submitted_Sequence
Submitted_Sequence
3000
2500
2000
1500
1000
500
500
1000
1500
2000
2500
3000
unc-13-PE


Dot plot of unc-13-PF vs. Submitted_Sequence
Submitted_Sequence
2500
2000
1500
1000
500
500
1000
1500
2000
2500
unc-13-PF

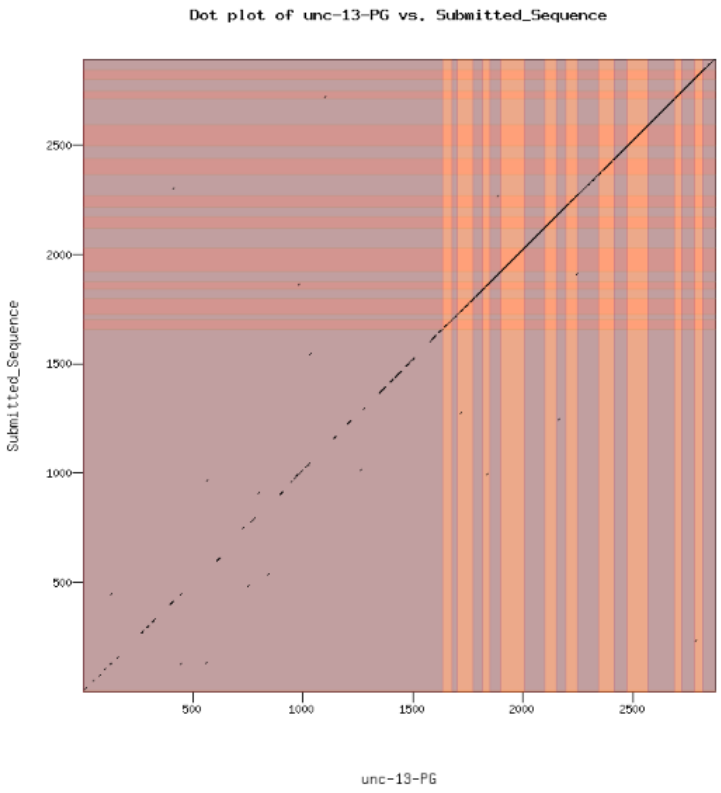


Figure 38: Gene model checker protein alignment dot plots of all* unc-13 *isoforms. *Exons 15-29 are shared by all isoforms and correspond to the highest homology between* D. melanogaster *and* D. eugracilis. *The isoforms represented by each dot plot is shown in the title of each dot plot. Amino acid alignments can be found in the supplemental figures.*

unc-13 Transcription Start Sites

Two unique transcription start sites were identified in *D. eugracilis unc-13.* The B and E isoforms share one TSS while the A, C, and F isoforms share a separate TSS. The TSS of the D and G isoforms could not be definitively identified in *D. eugracilis*. Because no two core promoter motifs found predicted the same TSS position and there was no conservation of core promoter motifs between *D. melanogaster* and *D. eugracilis* in the promoter region of *unc-13*, no

conclusion could be drawn from the core promoter motifs. Tables of the core promoter motif positons in *D. eugracilis* and *D. melanogaster* can be found in supplemental figures.

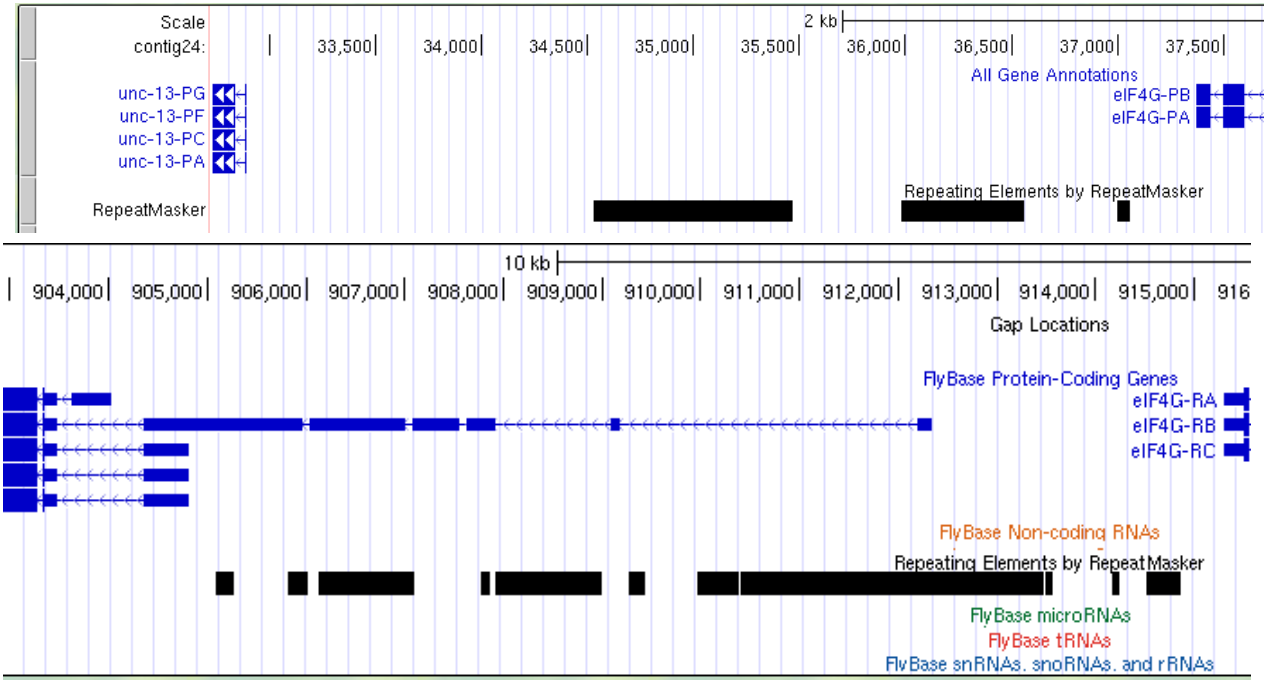


Figure 39: Region between* unc-13 *and* eIF4G *in* D. melanogaster *(bottom) vs* D. eugracilis *(top). *The region in* D. eugracilis *is ~4500 bp while in* D. melanogaster, *it is ~12000 bp. The region is very repeat dense.*

The first exon annotated for the G isoform in *D. melanogaster* is located 9000 bp upstream of the start codon. The region between *unc-13* and *eIF4G* in *D. eugracilis* is only 4500 bp (Fig. 39). This region in *D. biarmipes* is 6000 bp and contains only one RNApolII ChIP-seq peak, consistent with roughly where the TSS for the A, C, and F isoforms would be (Fig. 40). NCBI blastn alignment does not return any significant match in this region in *D. eugracilis* to the first annotated 5' UTR exon in *D. melanogaster unc-13-PG*.

Alignment by NCBI blastn region surrounding the RNApolII ChIP-seq peak in *D. biarmipes* to *D. eugracilis* contig 24 does not yield any significant results. While not suggesting any specific position, the *D. biarmipes* RNApolII ChIP-seq data indicates that an active TSS is present upstream of the A, C, and F isoforms of *unc-13* (Fig. 40).

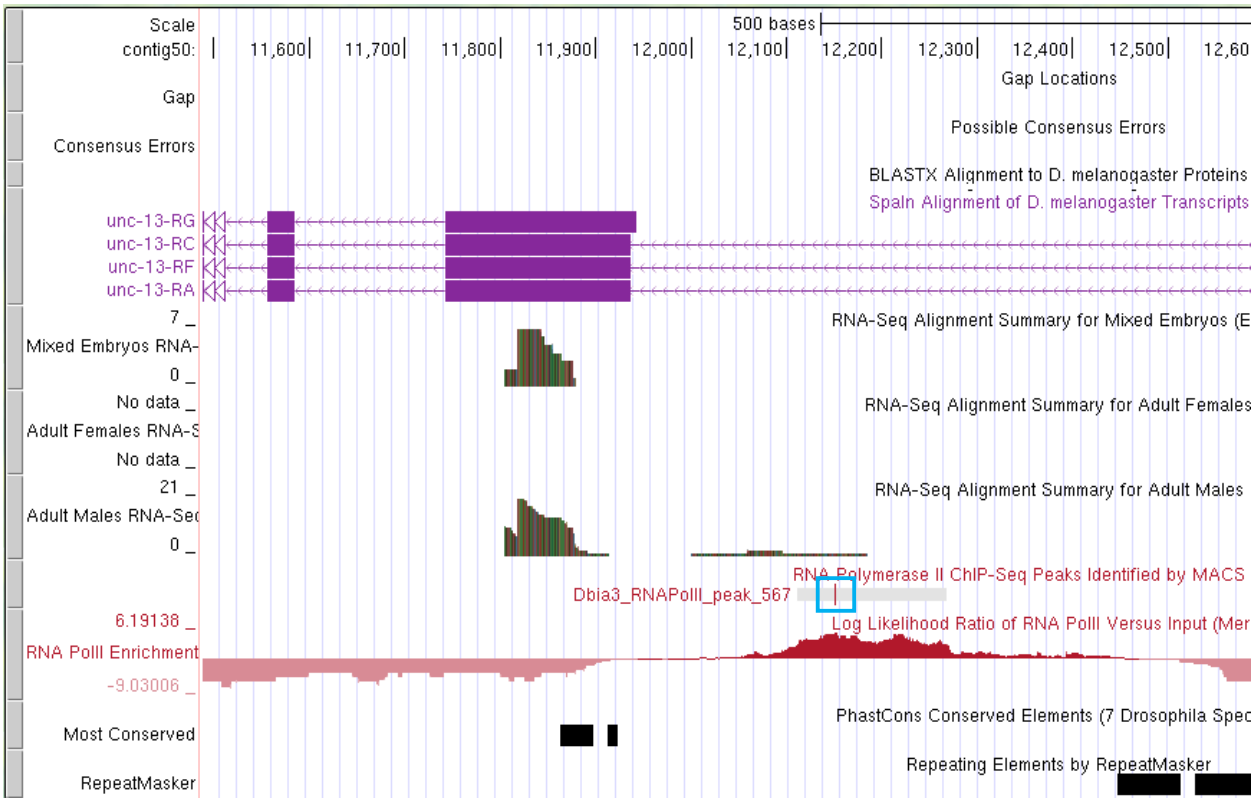


Figure 40:* D. biarmipes *ChIP-seq data. *The region in* D. biarmipes *surrounding the peak shown in the blue box does not align by NCBI blastn to* D. eugracilis*, however it does indicate that this gene is transcribed.*

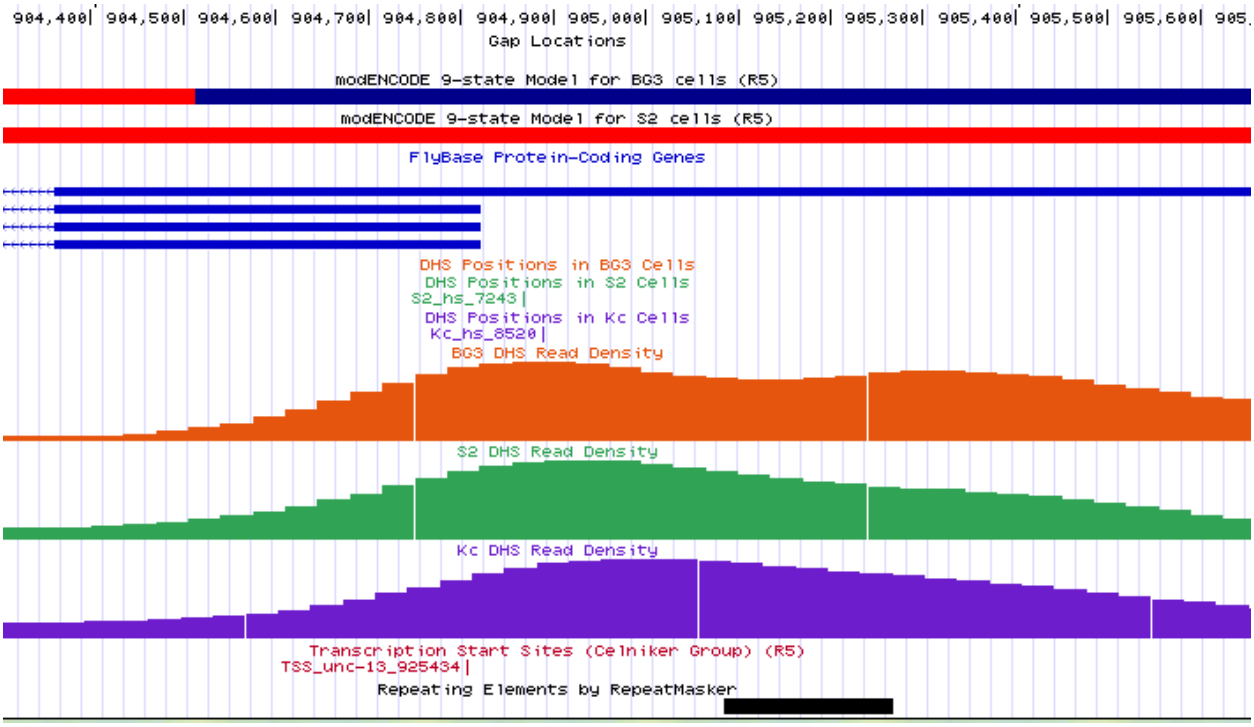


Figure 41: DHS peaks and annotated TSSs of* D. melanogaster unc-13 *A, C, and F isoforms*. There is one DHS read peak and one Celniker annotated TSS, indicating a peaked promoter.*

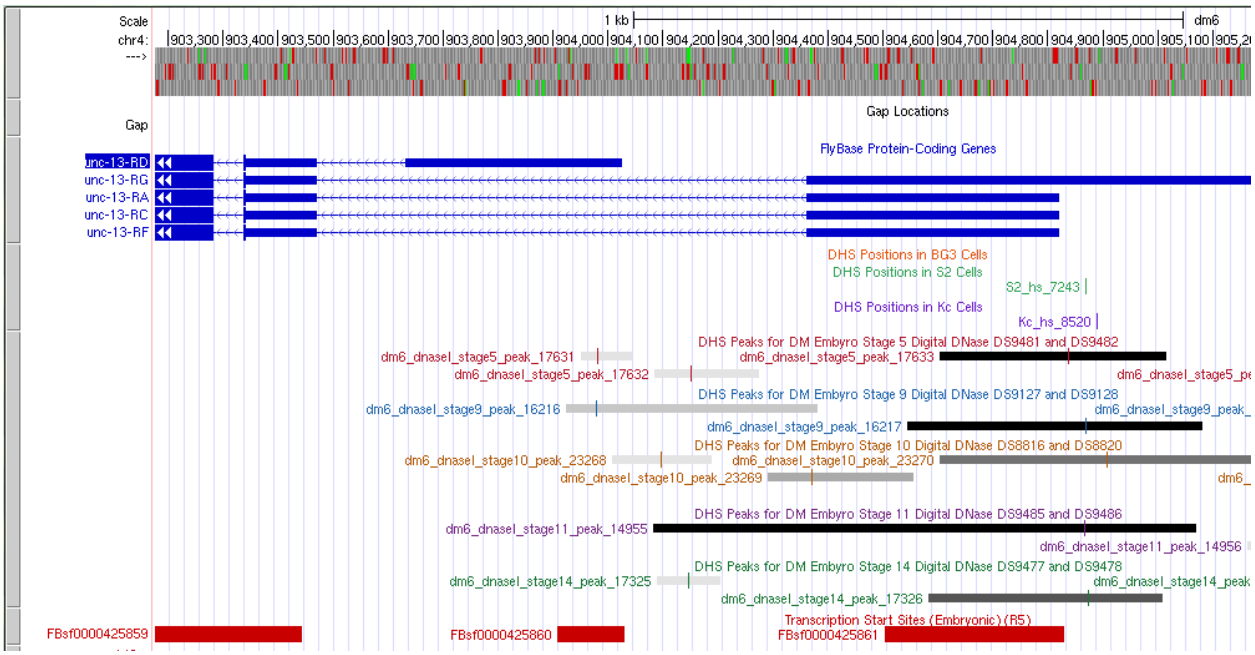


Figure 42: **D. melanogaster *embryonic DHS positions and embryonic TSS annotations.*** *The region in the box corresponds to where a TSS for the D isoform of* unc-13 *would be, however the lack of homology prevents definitively identifying a TSS in* D. eugracilis.

The region in *D. melanogaster* shown in Figure 41 shows one DHS peak and one Celniker TSS annotation, classifying this region as a peaked promoter. RAMPAGE data (Fig. 41) is consistent with a slightly broad promoter with some defined peaks. In Figure 42, embryonic TSS annotations places a separate TSS for the D isoform compared to the A, C, and F isoforms. Due to the failure of the first 5' UTR exon of the D isoform to align by NCBI blastn to *D. eugracilis* contig 24 and lack of any distinct embryonic RNAseq peaks, the precise location of the TSS of the D isoform could not be identified. Alignment of the first 5' UTR in the A isoform of *D. melanogaster* to *D. eugracilis* contig 24 using NCBI blastn is shown in Figure 43. Of 461 nucleotides, bases 175-332 aligned to *D. eugracilis*, with base 175 aligning to position 33929 in contig 24. Using the assumption that the 5'UTR in *D. eugracilis* would be a similar length, the first base of the 5' UTR in *D. eugracilis* was determined to be at position 34103. Because the Celniker annotation of the TSS in *D. melanogaster* is 13 bp downstream of the start of the first 5' UTR, this suggests position 34090 as the TSS. RNAseq data begins to drop near this position,

but because there are RNAseq reads extending until position 34120, the concluded TSS range for the A, C, and F isoforms of *unc-13* is within the region 34090-34120 or a bit to the right in *D. eugracilis* contig 24 (Fig. 44).

unc-13:6

Sequence ID: Query_69005 Length: 461 Number of Matches: 6

Range 1: 175 to 322 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
99.7 bits(68)	2e-23	114/154(74%)	9/154(5%)	Plus/Minus

```
Query  33779  ATATTCTAATTAGTTAAAAATCTGTAAAT---CACTATAAACTTACCTTTGATTTACAAC  33835
              |||||  ||  || |  ||  ||| | | |   ||||  ||||    |||  |||||  ||
Sbjct  322    ATATTTAAACTAATCCAATTTCTTTTATTAGTCACTTAAAACGCTTCTT--ATTTAGTAC  265

Query  33836  ACAGTTAATGAAAATTGGTTATTTTGTATTGTATAATTCCAATTCAATAAAAATCACTGA  33895
               ||||  |||||||| ||||||||||||||||||||||||||||| |||  |||||  ||||  ||
Sbjct  264    GCAGTAAATGAAAAATGGTTATTTTGTATTGTATAATTCCAGTTCCTTAAAATTCACAGA  205

Query  33896  TCATGCGCAAAGAAAAGAGAGAAACGATAGCACC  33929
               ||||||||||  |||||||||    ||||  ||||
Sbjct  204    GCATGCGCAAATCAAAGAGAG----GATAACACC  175
```

***Figure 43: blastn alignment of* D. melanogaster unc-13-PA *first 5' UTR exon and* D. eugracilis.** D. melanogaster *is the subject and* D. eugracilis *contig 24 is the query.*

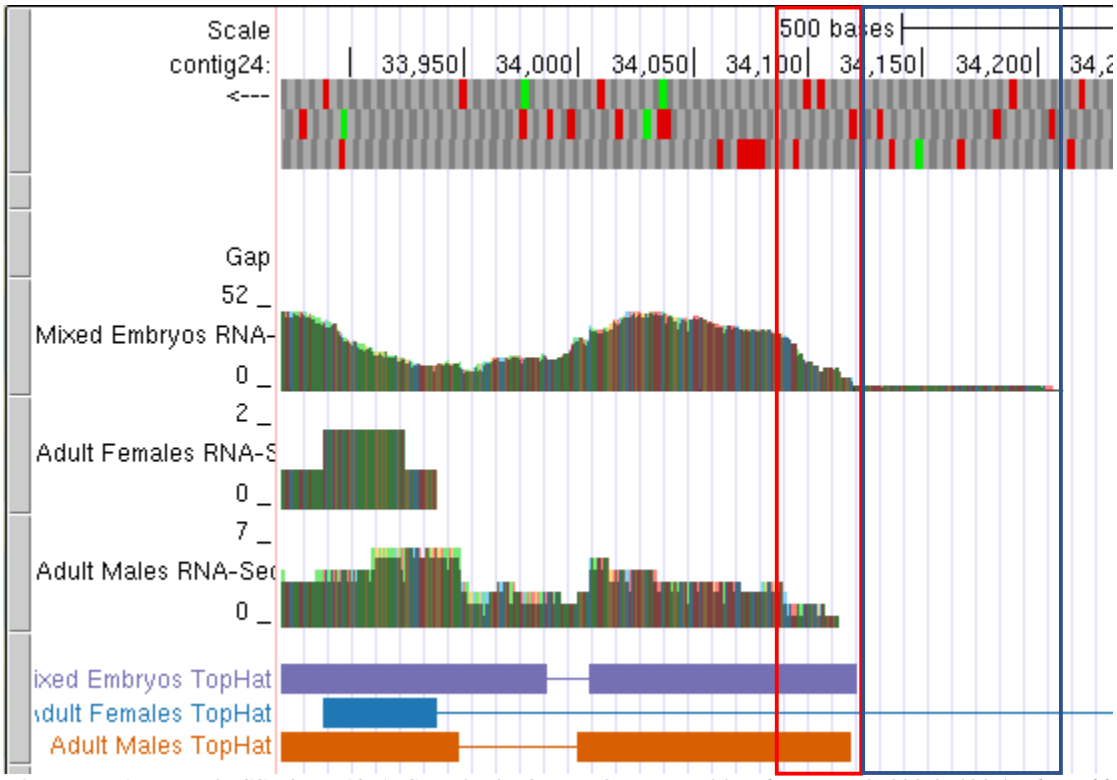


***Figure 44: Annotated TSS of* unc-13 *A, C, and F isoforms of* D. eugracilis.** *The region 34090-34120 (in the red box) corresponds to the approximate position extrapolated from blastn alignment to* D. melanogaster *and the end of RNAseq reads. The region in the blue box contains a low level of RNAseq reads in the Mixed Embryos track, indicating transcription upstream of this point.*

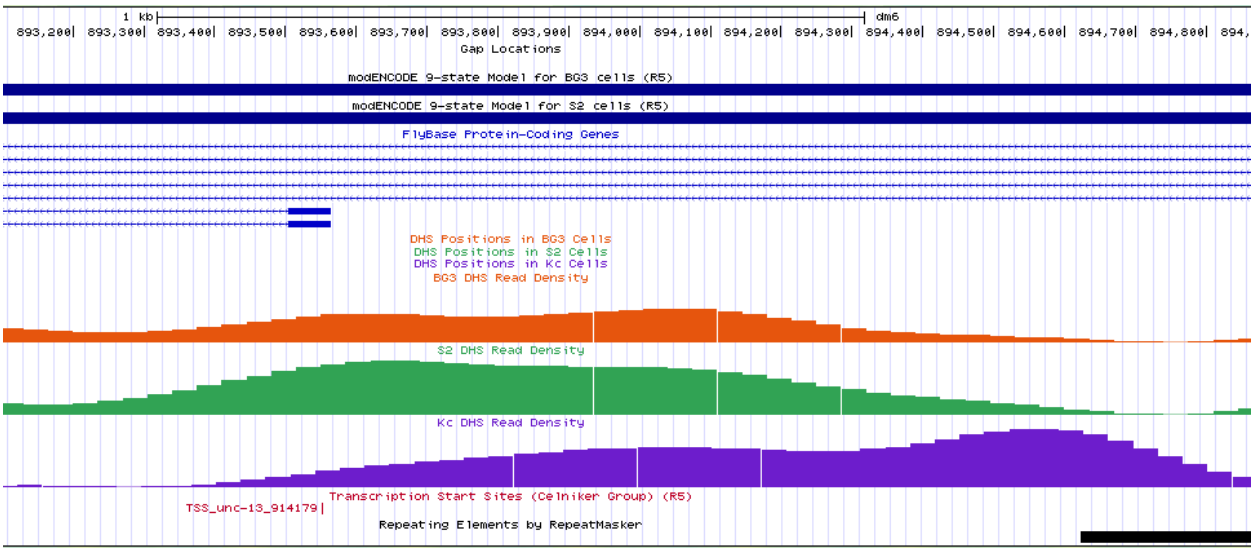


Figure 45: DHS peaks and annotated TSSs of* D. melanogaster unc-13 *B and E isoforms*. There are no DHS read peaks and one Celniker annotated TSS, indicating a peaked promoter.*

The region surrounding the first 5' UTR exon of the B and E isoforms of *D. melanogaster unc-13* contains one Celniker TSS annotation and no annotated DHS peaks, classifying it as a peaked promoter (Fig. 45). RAMPAGE data indicates an intermediate promoter, however there is no strongly defined RAMPAGE peak. All 61 nucleotides of the first 5' UTR exon in *D. melanogaster* align to *D. eugracilis* contig 24 using NCBI blastn (Fig. 46). The first base of the first 5' UTR exon aligns to position 24579 in *D. eugracilis* contig 24. (Fig. 47)

unc-13:14

Sequence ID: Query_144491 Length: 61 Number of Matches: 4

Range 1: 1 to 61 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
61.1 bits(41)	7e-13	52/61(85%)	1/61(1%)	Plus/Minus

```
Query  24520  AACTTAGTTAATTCTTTGTTGCGCCTAGATTCGTTCTCATGAAGCTA-CAGAAAATAATC  24578
              || |||||| |||||||||||||||||||| | |||| |||||||||||| ||| || ||| |
Sbjct  61     AATTTAGTAAATTCTTTGTTGCGCCTAGCTCCGTTTTCATGAAGCTACCAGCAAGTAAGC  2

Query  24579  T  24579
              |
Sbjct  1      T  1
```

***Figure 46: blastn alignment of* D. melanogaster unc-13-PA *first 5' UTR exon and* D. eugracilis.** D. melanogaster *is the subject and* D. eugracilis *contig 24 is the query.*

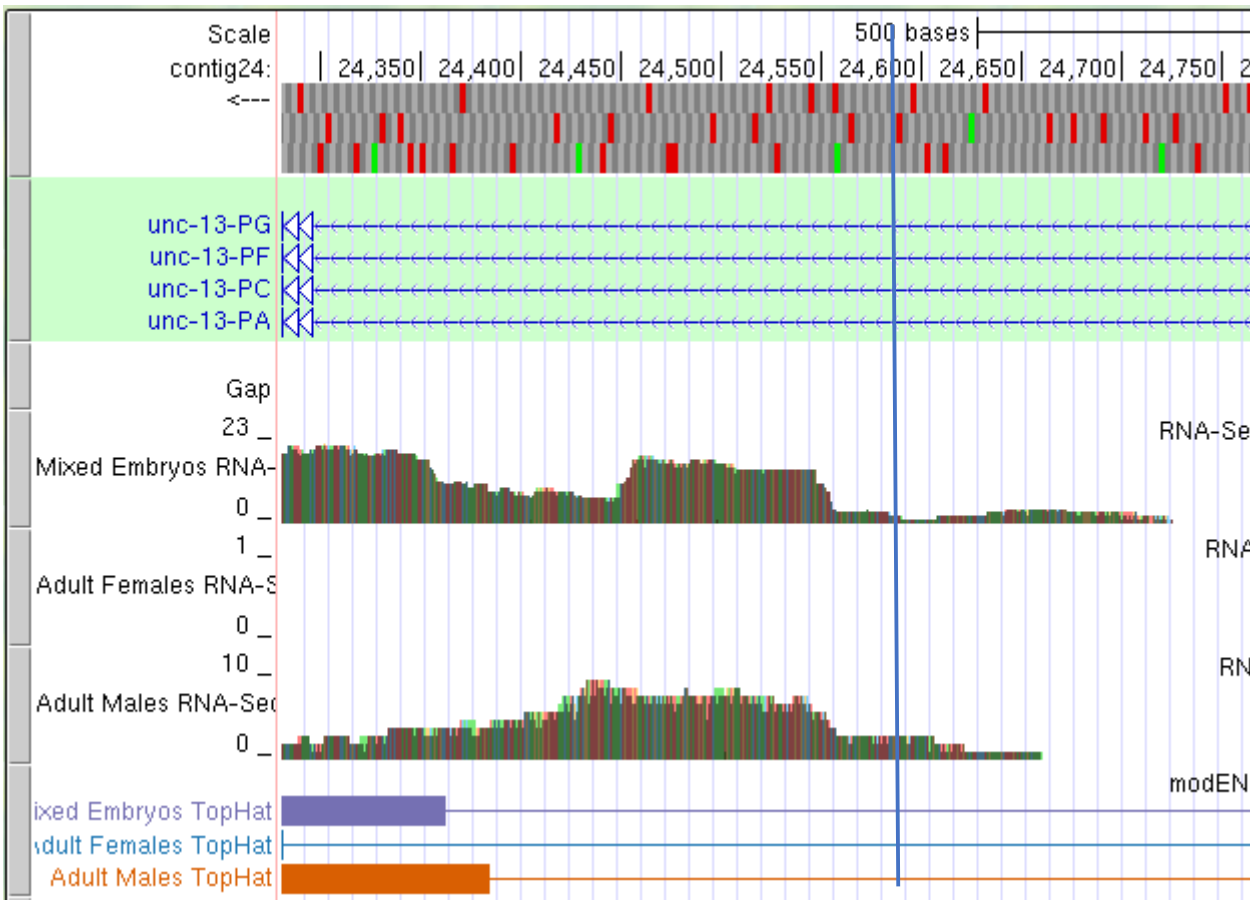


***Figure 47: Proposed TSS for the B and E isoforms of* D. eugracilis unc-13.** *The TSS at position 24579 (blue line) is consistent with blastn alignment to* D. melanogaster *as well as being just upstream of a drop in RNA-seq reads.*

Structure, Function, and Conservation of *unc-13*

After completing annotation of the coding sequences and TSSs in *D. eugracilis* contig 24, analysis of the structure, function, and conservation of *unc-13* was performed. Common features of the specific genes found on the F-element may be informative for determining why a heterochromatin-rich chromosome can have normal gene expression.

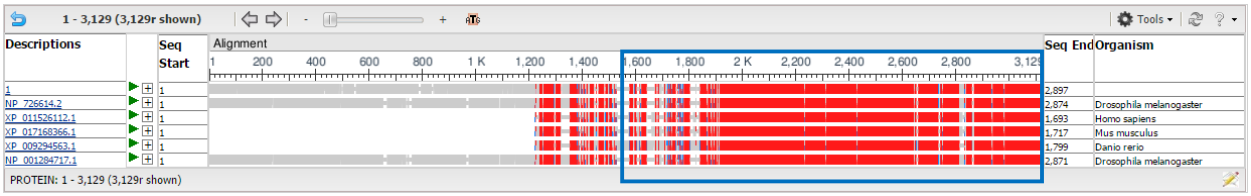


***Figure 48: NCBI SmartBlast alignment of* D. eugracilis unc-13 *to reference genomes of* Homo sapiens, Mus musculus, D. melanogaster, *and* Danio rerio.** *The C-terminal half of the protein (shown in the blue box) is highly conserved across all model organisms, indicating an important function.*

NCBI SmartBlast performs a multiple sequence alignment (MSA) of a submitted sequence to four reference genomes, *D. melanogaster, Homo sapiens, Mus musculus,* and *Danio rerio*. Figure 48 shows that the C-terminal half of the protein coded by *unc-13* is conserved across all species shown, suggesting that it may have an important biological function. The whole MSA is shown in the supplemental figures.

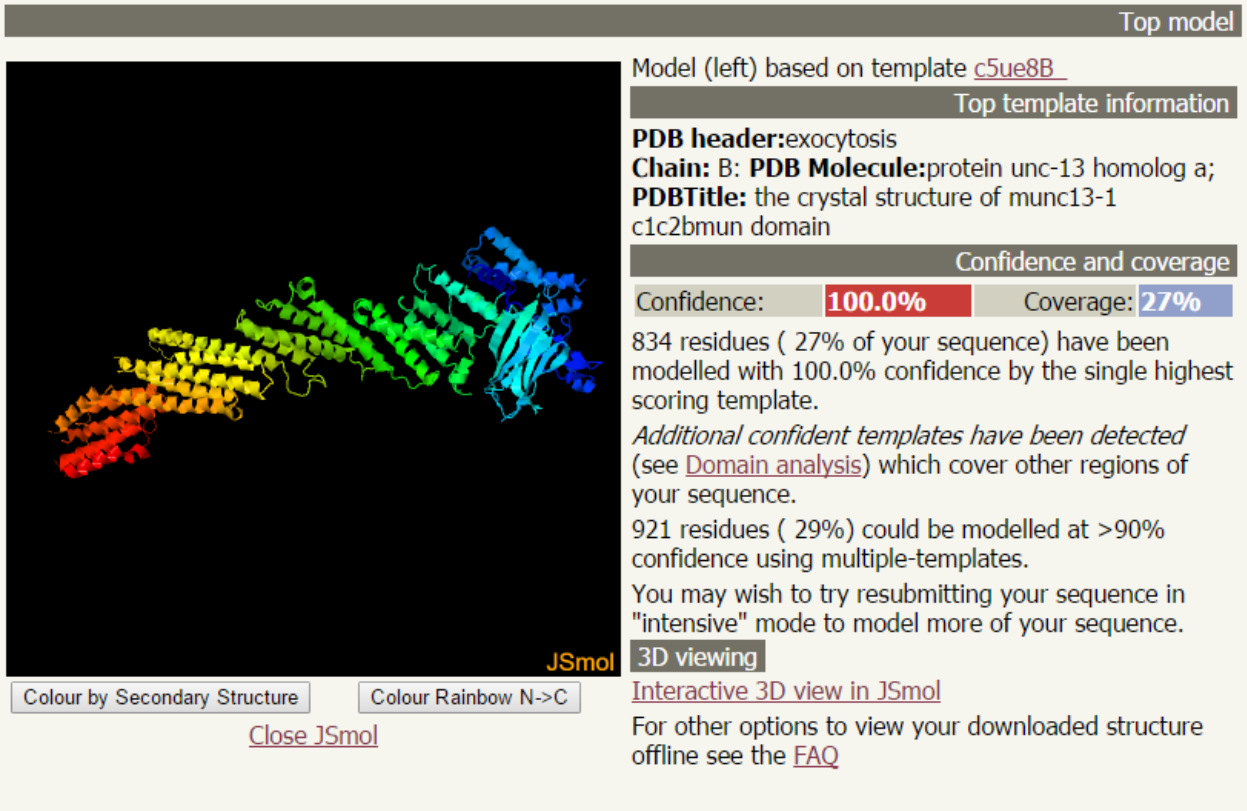


***Figure 49: Phyre2 gene model structure prediction for* D. eugracilis unc-13.** *Phyre2 predicts that this gene contains the munc13-1 domain. This domain is identified from the* mammalian unc-13 homolog A *gene and is associated with release of neurotransmitters from synaptic vesicles*

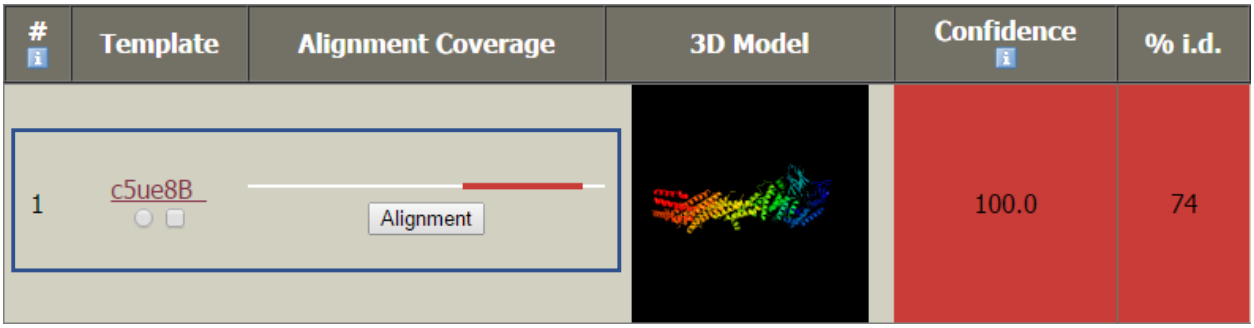


***Figure 49: Phyre2 reported the location of the identified structure in* D. eugracilis unc-13-PA.** *The boxed alignment corresponds to the highly-conserved region found in Figure 48.*

Phyre2 is a homology-based gene structure prediction tool which uses known gene structures to build a model for unknown structures with similar sequences. Phyre2 identified the functional domain of UNC13 protein in *D. eugracilis* to be a munc13-1 c1c2mun domain (Fig. 49). This domain is associated with transfer of synaptic vesicles between neurons. This domain accounts for 27% of the protein (isoform A) and the location of this domain (Fig. 50) falls within the highly-conserved region shown in Figure 48.

A search of the literature on *D. melanogaster unc-13* further elucidated the function of the protein. UNC13 is responsible for transferring information between the Ca^{2+} channels and synaptic vesicle transport proteins to allow for synaptic vesicle maturation. Null mutations of *unc-13* result in complete paralysis, mimicking the phenotype of removal of core synaptic vesicle transport proteins. UNC13 is an example of a protein coded for on the F-element that is necessary for survival.

Non-Coding RNA

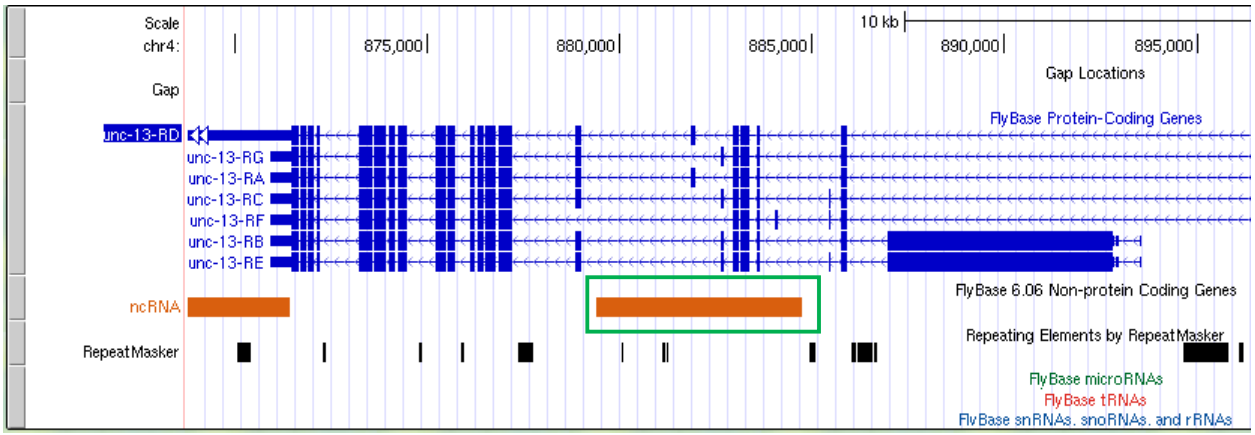


***Figure 50: GEP mirror of the UCSC Genome Browser view of* D. melanogaster.** *The green box shows a non-coding RNA, CR44029, found in* D. melanogaster *that may potentially also exist in* D. eugracilis.

An annotated non-coding RNA (ncRNA) called CR44029 is found in the region between exons 8_2147_0 and 15_2147_2 in *D. melanogaster* (Fig. 50). The sequence of this ncRNA was obtained from the UCSC genome browser and aligned to *D. eugracilis* contig 24 using NCBI

blastn (Fig. 52). Although CR44029 in *D. melanogaster* spans across multiple exons, only the region between exons 14 and 15 in *D. eugracilis* contain raised RNA-seq read counts (Fig. 51), and therefore this was the only region aligned to CR44029 using blastn.

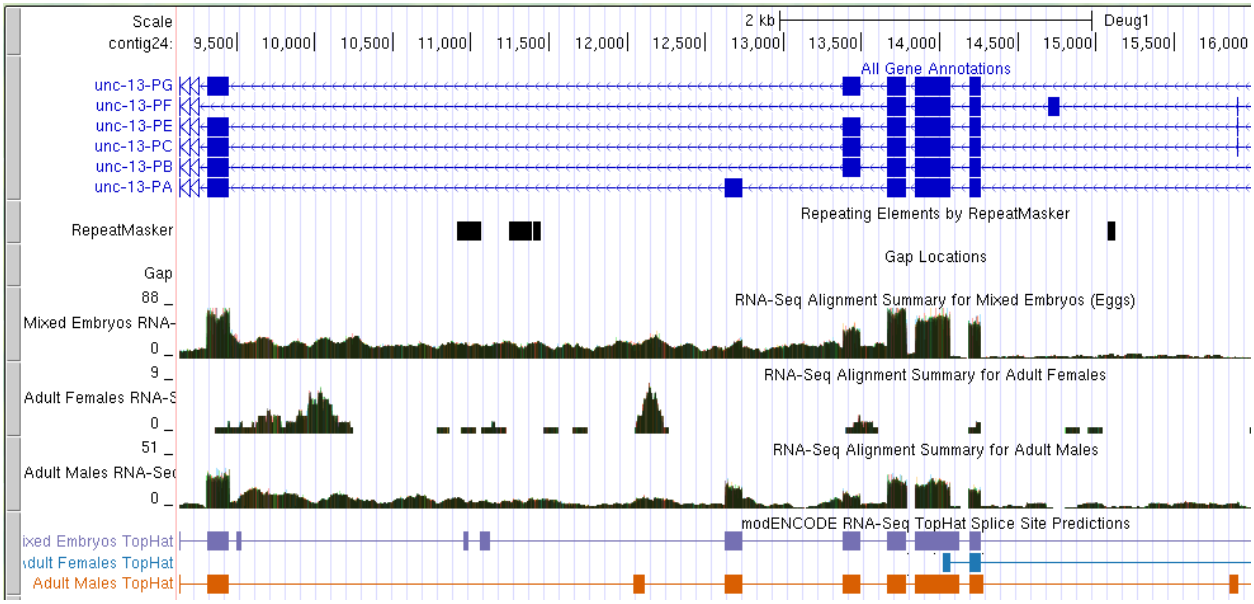


***Figure 51: RNA-seq reads extend throughout a non-repeat dense intron in* D. eugracilis.** *The black box indicates the region that was tested for alignment with CR44029. While some portions aligned, there was no indication of one feature in this region.*

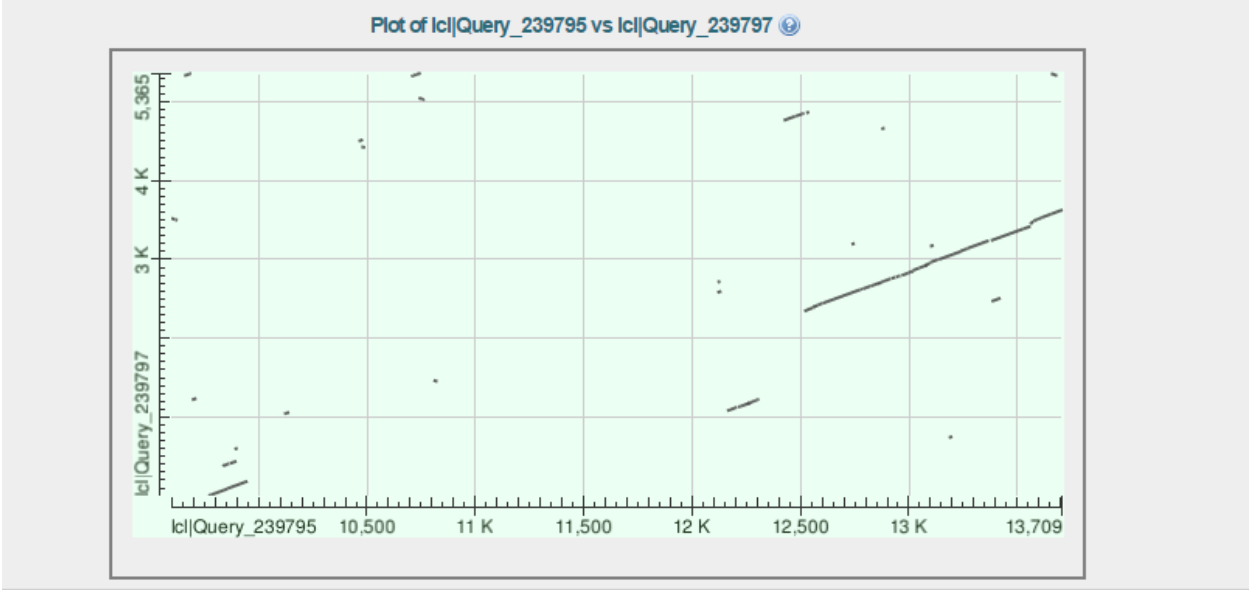


Figure 52: blastn alignment of* CRR44029 *to* D. eugracilis *contig 24. *Several matches were found, but none suggested one continuous feature orthologous to CRR44029 in this region.*

Synteny

If the genes in orthologous regions between two species are in the same relative orientation to each other, the regions are said to be syntenic. Contig 24 in *D. eugracilis* is fully syntenic with the corresponding region in *D. melanogaster* (867740-931515). Both species contain *unc-13* and *eIF4G* both on the minus strand.

Repeats

A distinct feature of heterochromatic regions of the genome including the F-element is higher repeat density relative to the whole genome. The repeat content of *D. eugracilis* contig 24 and the orthologous region in *D. melanogaster* was examined. Contig 24 contains 54 recognized repeats with an average length of 236.41 bp, meaning that contig 24 is 22.4% repeats. The corresponding region in *D. melanogaster*, 867740-931515 bp, contains 51 repeats with an average length of 274.18 bp, meaning that this region is 21.9% repeats. Helitron repeats are the most common repeat type in both species. Both species have 6 repeats longer than 500 bp, though these repeats are not the same between species (Table 6). The repeats in contig 24 were mapped to their location in the genome using a custom track on the GEP mirror of the UCSC genome browser (Fig. 53). Although the specific repeats are not shared between *D. eugracilis* and *D. melanogaster*, the repeat density and frequency of large repeats is similar between the two species. Characteristics of these repeats could potentially contribute to the unique chromatin state, but future comparative genomic analysis of repeats across more *Drosophila* species is necessary.

Start	Length	Family
***D. melanogaster* chr4: 867740-931515**		
894775	1049	TcMar-Tc1
906402	698	CMC-Transib
907916	1076	P
910744	2390	Gypsy
921363	508	Helitron
923061	899	TcMar-Mariner
***D. eugracilis* contig 24**		
17272	942	RC/Helitron
25578	583	RC/Helitron
34533	937	RC/Helitron
46088	572	LINE
46662	1373	LINE
50747	640	Unknown

Table 6: Repeats in* D. eugracilis *contig 24 and corresponding region in* D. melanogaster *that are >500 bp in length. *Consistent with repeats in* D. melanogaster *being longer on average than those in* D. eugracilis *contig 24, the average length of repeats >500 bp is 1103.3 bp in* D. melanogaster, *but only 841.2 bp in contig 24.*

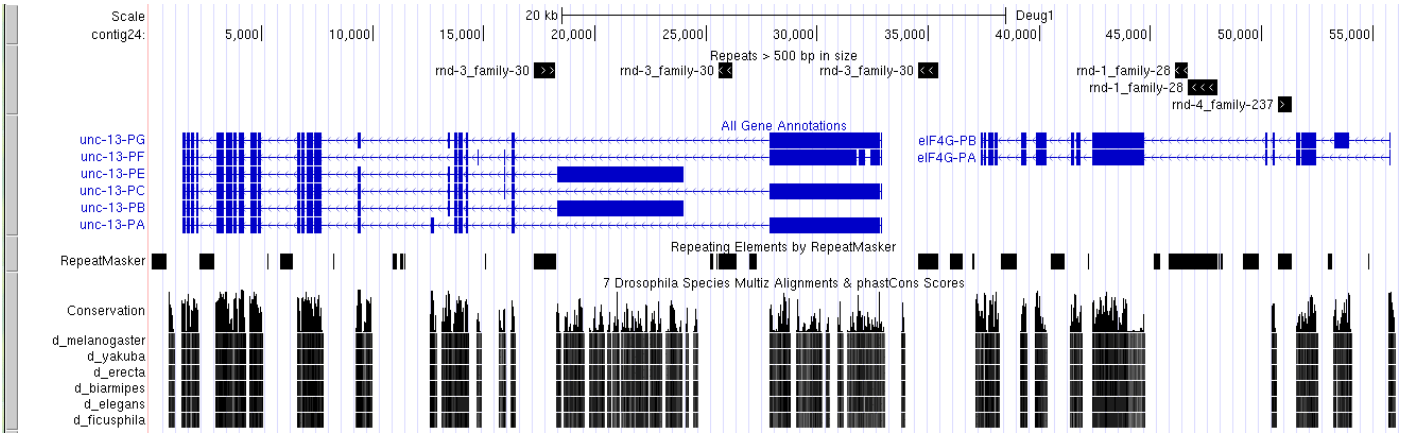


Figure 53: Locations of large repetitions elements in* D. eugracilis *contig 24. *A custom track was used to show the locations of identified repeats with length >500 bp. Of the six identified repeats, three are in the introns of* eIF4G, *two are in the introns of* unc-13, *and one is in the intergenic region.*

Conclusion

Drosophila eugracilis contig 24 contains two genes, *unc-13* and *eIF4G*. *eIF4G* has three isoforms representing two unique coding sequences. All 15 unique exons and all isoforms of

eIF4G have been annotated. The TSS of all isoforms of *eIF4G* was determined to be position 56686 in contig 24. *unc-13* has seven isoforms representing six unique coding sequences. All 29 unique exons and all isoforms of *unc-13* have been annotated. The TSS of the A, C, and F isoforms of *unc-13* was determined to be within the region 34090-34120 bp. The TSS of the B and E isoforms was determined to be position 24567. The TSS of the G and D isoforms could not be determined. The function of *unc-13* is an example of a biologically vital gene that is expressed within the heterochromatic F-element. Gene and TSS annotation of *D. eugracilis* contig 24 is complete and ready for future *Drosophila* comparative genomic studies.

References

Mathias A Bohme *et al.* "Active zone scaffolds differentially accumulate Unc13 isoforms to tune Ca2+ channel–vesicle coupling" *Nature Neuroscience* 19 (2016): 1311–1320

Aravamudan, B *et al.* "*Drosophila unc-13* is Essential for Synaptic Transmission" *Nature Neuroscience* 2 (1999): 965 – 971

Acknowledgements

I would like to thank the instructors and TAs for Bio4342/434W, Dr. Sarah Elgin, Dr. Chris Shaffer, Dr. April Bednarski, Dr. Jeremy Buhler, Wilson Leung, Kailong Mao, Emily Chi, and Ryan Friedman.

Supplemental Figures

Supplemental figures have been submitted as a digital file.