Last Update: 12/24/2023

Annotation Strategy Guide

Jeannette Wong, Washington University in St. Louis, August 2011

Table of Contents

Splice Site Boundaries

Extension of reading frame for length conservation

Found on contig47 of the *D. mojavensis* dot, *rho-5* is a weakly conserved gene compared to its *D. melanogaster* ortholog. We will try to annotate the penultimate exon (CDS 5_9883_0) in this example.

Step 1: Using the amino acid sequence from this exon as the query and the nucleotide sequence of contig47 as the subject, perform a *tblastn* search with the low complexity filter and compositional adjustment turned off (Figure 1).

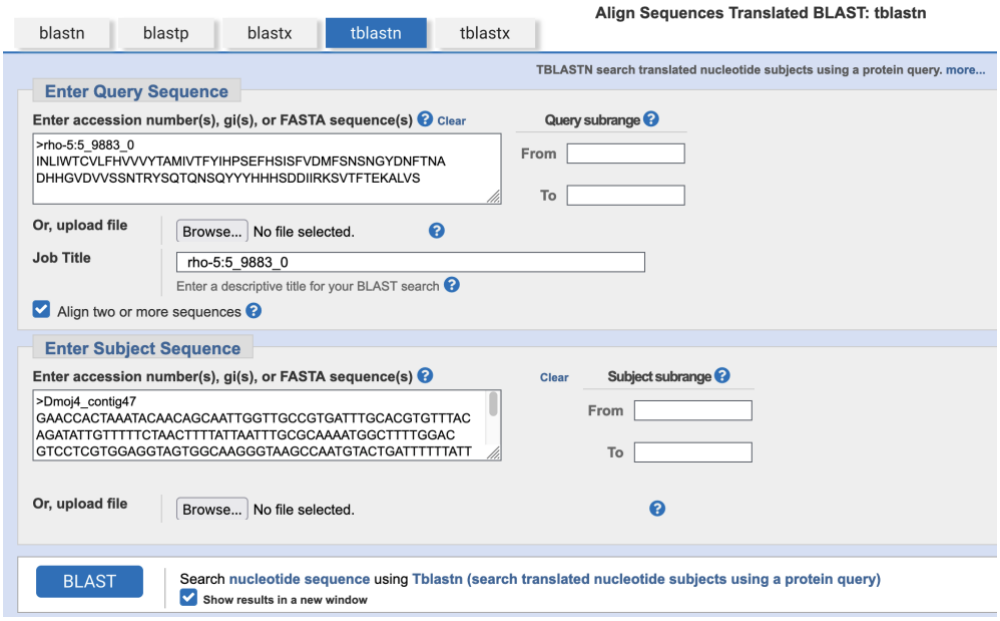

Figure 1 *tblastn* of CDS 5_9883_0 (query) and contig47 nucleotide sequence (subject)

If you cannot find any alignments using these search parameters, you can increase the sensitivity of the BLAST search by changing the following parameters (Figure 2):

1. Increase the Expect value threshold from 0.05 to 1e5 or more.
2. Decrease the word size from 5 to 3 (database searches) or 3 to 2 (*bl2seq* searches).
3. Define a subrange to find the best alignment in a smaller region.

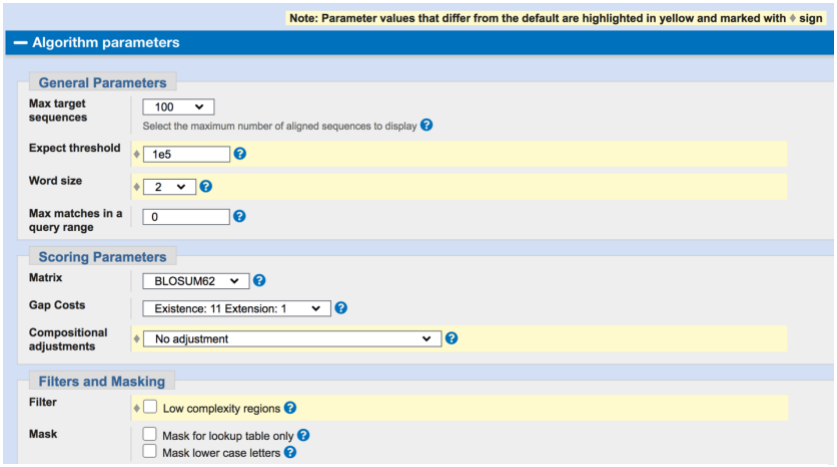

Figure 2 Changing BLAST parameters to increase the sensitivity of your search.

For additional information on word size, compositional adjustment, etc., click on the question mark next to the drop-down box corresponding to the parameter of interests. I was able to find a good alignment for the penultimate exon, and it is found within the expected region of contig47 (Figure 3). While the percent identity is only 55%, the alignment accounts for the first 90 residues of this CDS. However, CDS 5_9883_0 is 95 residues long so we are still missing five amino acids at the 3' end.

Dmoj4_contig47

Sequence ID: **Query_33935** Length: **50000** Number of Matches: **1**

Range 1: 3138 to 3422 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
104 bits(260)	1e-30	53/96(55%)	70/96(72%)	7/96(7%)	-1

```
Query  1     INLIWTCVLFHVVVYTAMIVTFYIHPSEFHSISFV-DMFSNSNG--YDNFTNAD---HHG  54
             INLIWTC+LFH  +Y  +  TFYI+PSEF++ SFV D+F ++NG  Y   TN++   HHG
Sbjct  3422  INLIWTCLLFHFFMYMTLATTFYIYPSEFNTFSFVDDIFGSNNGNKYIVPTNSNIGQHHG  3243

Query  55    VDVVSSNTRYSQTQNSQYYYHHHSDDIIRKSVTFTE  90
              +V S+  RYS+TQ  QYYYHHHS+DIIR +V + E
Sbjct  3242  -EVSSTTRRYSETQKPQYYYHHHSEDIIRNTVAYPE  3138
```

Figure 3 *tblastn* alignment of CDS 5_9883_0 (query) against contig47 (subject)

Step 2: Open a web browser and navigate to contig47 on the Genome Browser (*D. mojavensis* Sep. 2008 assembly) at https://gander.wustl.edu so that we can take a closer look at this region. Figure 4 shows that there are multiple evidence tracks that support the boundaries for this exon. Data from the modENCODE RNA-Seq coverage, TopHat junctions, and donor/acceptor splice site tracks all support this region very well.

While there are several TopHat junctions that suggest the exon should be shorter than the boundaries suggested by the RNA-Seq read coverage data, using these splice sites would remove some of the conserved amino acids of CDS 5_9883_0. For example, if we were to follow the N-SCAN prediction, much of the conserved sequence would be truncated. This model would also contradict the high-scoring donor site predicted right where the RNA-Seq data stops.

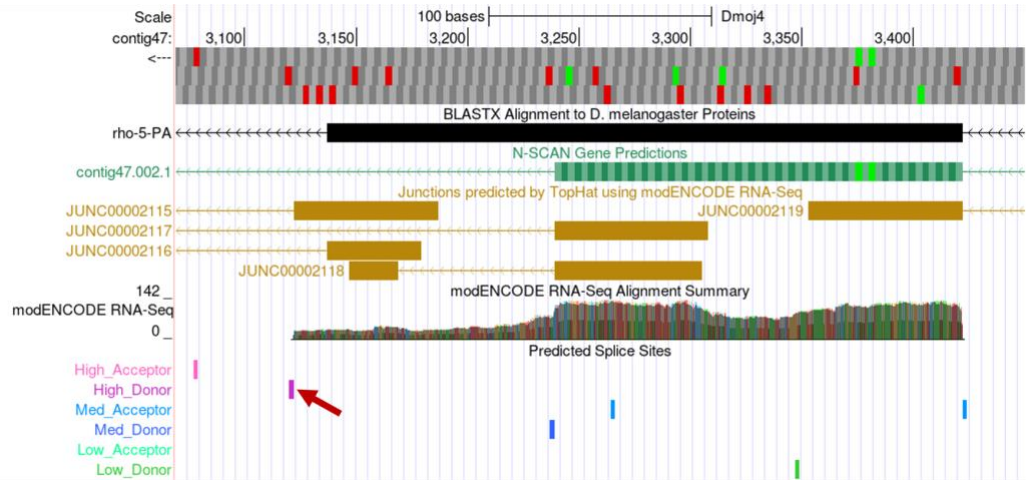

Figure 4 Genome Browser view of the region surrounding CDS 5_9883_0 on contig47

Zooming in on the 3' end, we found the donor site at 3121-3122. This will extend the aligned region and gives us the five amino acids that were missing from the *tblastn* alignment (Figure 5).

Therefore, the final exon coordinates for CDS 5_9883_0 should be 3422-3123 in order to preserve the conserved amino acids and to minimize the number of changes compared to the *D. melanogaster* gene model.

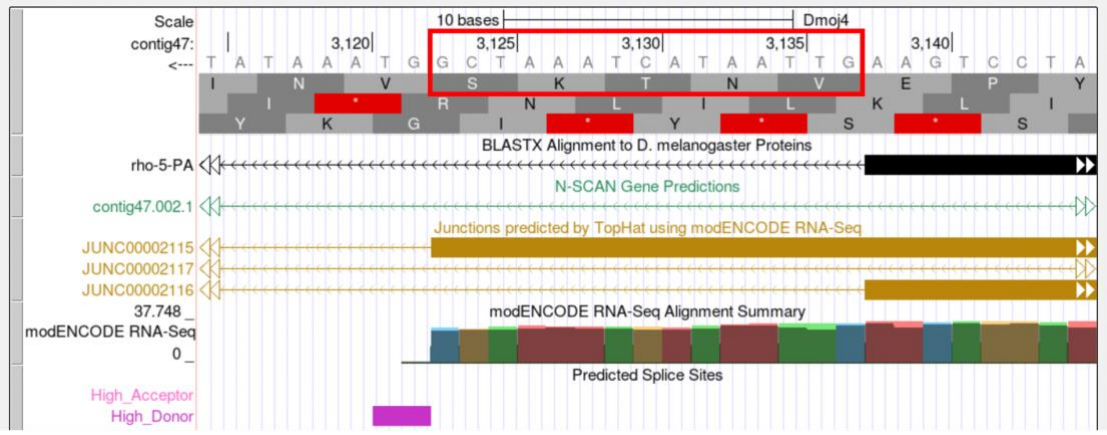

Figure 5 Addition of 5 amino acids to the 3' end of CDS 5_9883_0

Selecting a better supported splice site

While extension of an open reading frame is important for conserving the length of the coding exon, it is also important to consider selecting the best-supported splice site. For example, let's look at the first coding exon (1_2336_0) of *Dyrk3* in *D. mojavensis*.

Found on contig37, the *blastx* track on the Genome Browser reveals an alignment that only covers a small part of the first exon. Consequently, we will need to do a more sensitive *tblastn* search to determine if there is any additional region of conservation to the first exon. The second exon of *Dyrk3* begins at 55,082. Because this gene is oriented in the positive direction, we will look for an exon upstream from this coordinate.

Step 1: Perform a *tblastn* search using the amino acid sequence for the first coding exon of *Dyrk3* (CDS 1_2336_0) as the query and the contig37 nucleotide sequence as the subject with the low complexity filter and compositional adjustments turned off and the Expect threshold set to 10. This *tblastn* search reported a spurious match to the end of the CDS at 14520-14543 (E-value = 7.2).

To increase the sensitivity of the *tblastn* search, I repeated the search using the same parameters except for changing the word size from 3 to 2. This *tblastn* search with the smaller word size was able to locate an alignment that almost covers the entire first exon (Figure 6). The first exon is 131 aa long and this alignment covers the first 129 aa of this exon.

Dmoj4_contig37

Sequence ID: Query_128277 Length: 60000 Number of Matches: 14

Range 1: 54414 to 54809 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
27.3 bits(59)	0.006	32/136(24%)	55/136(40%)	11/136(8%)	+3

```
Query  1      MVGSQEKKNNHIELSETPATDKNNLNTTHLE----NTQLSKALSPP-TSLPQIQIQMINQ  55
              MVG+Q++  N+ +LS T A    +  T  LE    + Q+ K +SP   ++   Q Q  +
Sbjct  54414  MVGTQQQTKNNSQLSTTDAIRDASTFTGDLEAPAIDIQIGKPVSPAMLTMTMSQSQPADG  54593

Query  56     NLTHTGIAQNNTEKANRHQYRDSGLQYLTRCFEPLAMLNDSKEDF--PTQPSNNIANYPG  113
               +  +   Q +    + H        +    +   D K+    P  P   + NY
Sbjct  54594  RVGGSASVQMHPSSLSHHSCMAVPFNF-NSSYGSDRIFQDRKQPAKQPRIPQEQLCNYIA  54770

Query  114    DIQILPIFDCCEISES  129
              D   L   D C +S+S
Sbjct  54771  DSLGL---DACGVSKS  54809
```

Figure 6 *tblastn* of CDS 1_2336_0 of *Dyrk3* (query) against the contig37 sequence (subject)

Step 2: Examining this region in the Genome Browser, we find that this region is located upstream of the second exon and it is well-supported by the RNA-Seq data (Figure 7):

Figure 7 Genome Browser view of the region surrounding the potential first coding exon of *Dyrk3*

Zooming in on this region, we can see there is some support for 54,810 (where the *tblastn* alignment terminates) as the end coordinate for the first exon (Figure 8). However, looking a little further downstream, you can see there is much stronger support to place the end of the exon at 54,816. While there are TopHat junctions suggesting both of these splice donor sites are possible (and both junctions have the same splice acceptor site), the splice site at 54,817-54,818 is supported by the SGP and Genscan gene predictions, the modENCODE RNA-Seq data, and a medium quality splice donor site predicted by GeneSplicer.

Collectively, the available evidence provides better support for the donor site at 54,817-54,818 than the donor site located near the end of the *tblastn* alignment at 54,811-54,812. Selecting this site will also account for the two amino acids length difference between the *tblastn* alignment and the *D. melanogaster* CDS. Therefore, we will annotate the first exon for *Dyrk3* so that it spans from 54,414 to 54,816 (Figure 8).

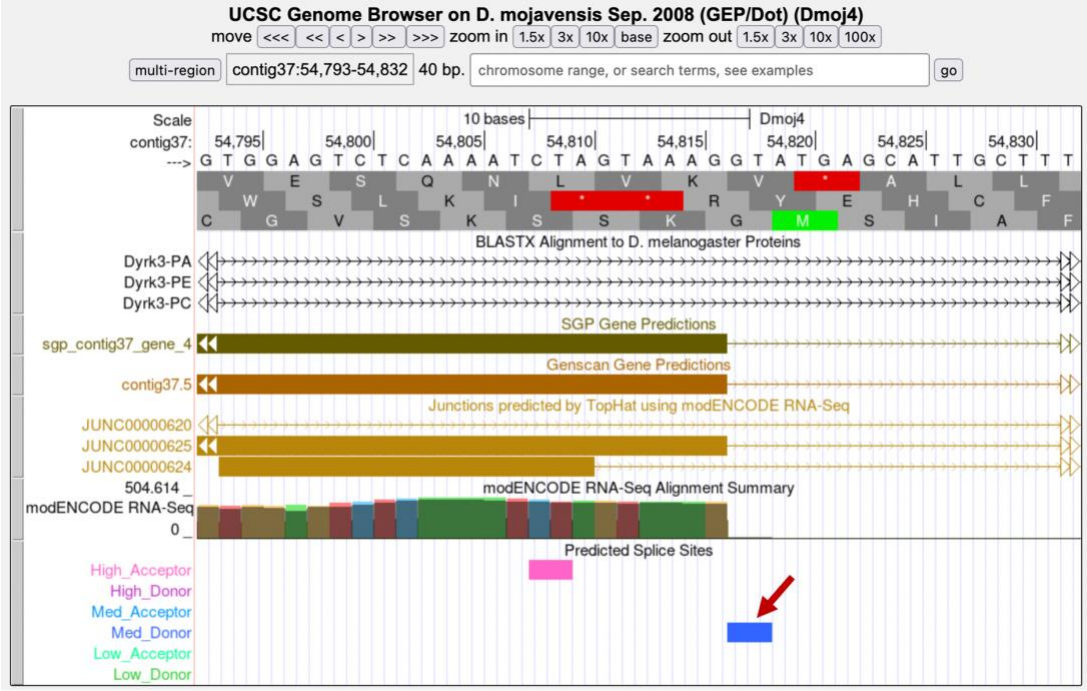

Figure 8 Splice donor site for the first coding exon of *Dyrk3* at 54,817-54,818

Missing/Extra Exons

The first and last exons of a gene model are often less conserved. Depending on the conservation of the internal exons, there might be some conservation at the 3' end of the first exon and at the 5' end of the last exon (Figure 9). Sometimes, there will be very high conservation throughout the entire gene model, but this will not always be the case.

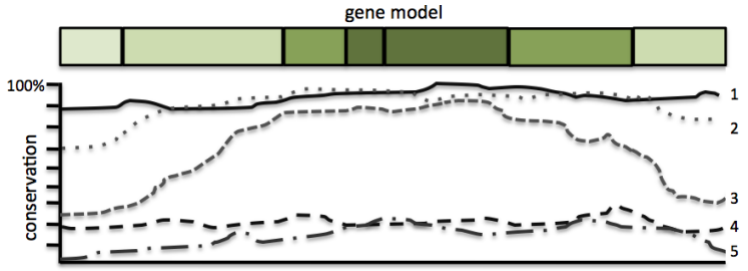

Figure 9 Schematic of gene model conservation by exon. (1): high conservation in all exons, (2): internal exons are highly conserved, flanking exons are slightly less conserved, (3): internal exons are highly conserved, flanking exons show little conservation, (4) low conservation in all exons, (5): very low conservation with flanking ends barely conserved.

Exons found on adjacent contigs

This example will demonstrate the step-by-step strategy for annotating the first coding exon (1_3482_0) of *Or13a* on contig47 in *D. mojavensis*. According to the *blastx* track on the GEP UCSC Genome Browser, *Or13a* is oriented on the minus strand in contig47. The CDS 1_3482_0 has a total length of 117 amino acids in *D. melanogaster*.

Step 1: I performed a *tblastn* search with the CDS 1_3482_0 from *Or13a* as the query against the nucleotide sequence from contig47 with the low complexity filter and compositional adjustments turned off and the default Expect threshold of 0.05. This search did not produce any significant hits in the correct orientation (i.e., on the minus strand) that contains a large open reading frame with a start codon. Even after modifying the BLAST parameters (i.e., decrease the word size, and increase the E-value threshold) and restricting the subject range to where the first exon should be located, there were only partial low quality matches.

Step 2: Because we cannot place the CDS 1_3482_0 based on the *tblastn* search results, we will need to rely on the RNA-Seq data to annotate this CDS. Looking at the Genome Browser for contig47, the RNA-Seq data and TopHat junctions suggest that the orthologous CDS might not exist in *D. mojavensis*.

The TopHat junction JUNC00002159 suggests an intron that spans from 34,425 to 34,227. There is an open reading frame and a start codon (at 34,509-34,507) upstream of 34,425 in frame -3 (Figure 10). However, the splice donor site at 34,424-34,425 is in phase 0 relative to frame -3 and this splice donor site is incompatible with the phase 1 acceptor site at 34,228-34,227 relative to frame -2. Consequently, the open reading frame at 34,509-34,426 cannot be the first CDS of *Or13a* in *D. mojavensis*. In order to use this TopHat splice junction, we would need to propose a change in the gene structure so that the second CDS of *Or13a* (2_3482_1) in *D. melanogaster* becomes the initial CDS of the *Or13a* ortholog in *D. mojavensis* and the regions upstream of the start codon would correspond to the 5' UTR.

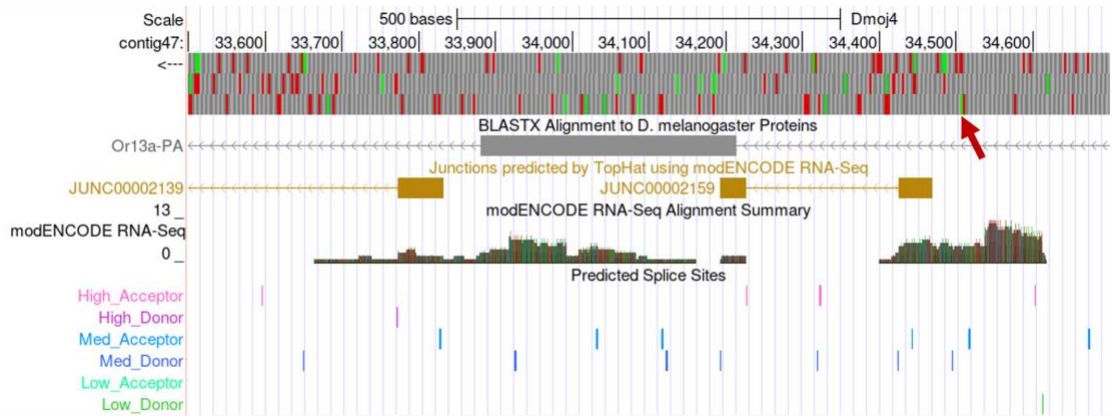

Figure 10 Potential start codon for the *Or13a* model on contig47

While this is a possibility, we would like to create a gene model that minimizes the number of changes (i.e., construct the most parsimonious model) between *D. melanogaster* and *D. mojavensis*. Consequently, we will continue to search for a region within contig47 that could be the first exon of *Or13a*.

Step 3: We will examine the other evidence tracks on the Genome Browser more closely to see if there are any clues as to the location of the first coding exon. Looking at the region upstream of the start of the second exon, we find a small *blastx* alignment block for *Or13a* between 36,500-37,000 (Figure 11):

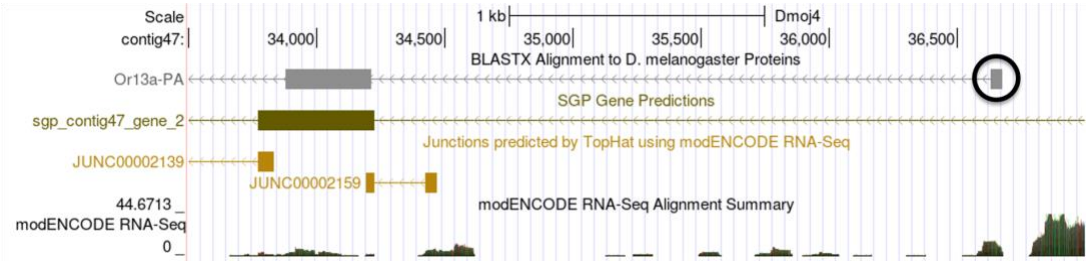

Figure 11 Small *blastx* alignment block located upstream from second exon

Step 4: Clicking on this *blastx* block reveals the alignment details for this feature. Organizing the HSP summaries by subject start in ascending order shows that this alignment block spans from 36,679-36,632 in the -2 frame, and it corresponds to residues 79-94 of the CDS (Figure 12). Recall that the first exon of *Or13a* is 117aa long. This alignment suggests that there is conservation near the end of the first exon with *D. melanogaster*.

```
 Score = 20.2 bits (36), Expect = 3.2e-113, P = 3.2e-113
 Identities = 5/16 (31%)
 Frame = -2

Query: 36679 KCCYDYKVPPNYLRMI 36632
          +CC  +    N++R+I
Sbjct: 79    NCCTTFMGVLNFVRLI 94
```

Figure 12 First alignment block for *Or13a* from the *blastx* track on the Genome Browser

Step 5: Before continuing to examine this region on contig47, it is possible that the first exon is not located within this sequence but could be found on the adjacent contig. A BLAT search of the last 10kb of contig47 against the collection of *D. mojavensis* dot projects reveals that there is a 20kb overlap between contig47 and contig48. Hence, we will search for the CDS in contig48. If this search did not yield any results, then I would construct the model for the first CDS by extending the open reading frame identified in step 4 to the furthest start codon possible.

However, performing a *tblastn* search of the first CDS against contig48 with only the low complexity filter and compositional adjustments turned off revealed the following alignment (Figure 13):

Dmoj4_contig48

Sequence ID: Query_60219 Length: 40000 Number of Matches: 1

Range 1: 31306 to 31671 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
149 bits(376)	6e-46	76/122(62%)	86/122(70%)	5/122(4%)	-2

```
Query  1      MFYSYPYKALSF--PIQCVWLKLNGSWPLT---ESSRPWRSQSLLATAYIVWAWYVIASV  55
              MF   PYK  SF  P QC+WLKLNGSWPL     + +      L A  Y VWA YVI SV
Sbjct  31671  MFNPLPYKDPSFRMPFQCIWLKLNGSWPLEIKLANQKLLTRSYLFAFIYNVWALYVIISV  31492

Query  56     GITISYQTAFLLNNLSDIIITTENCCTTFMGVLNFVRLIHLRLNQRKFRQLIENFSYEIW  115
              GITIS+QT+FL+NN  DII+TTENCC+T MG LNFVRLIHLR+NQ KFR+LI  F   IW
Sbjct  31491  GITISFQTSFLINNFGDIIMTTENCCSTLMGALNFVRLIHLRVNQPKFRKLINQFVLNIW  31312

Query  116    IP  117
              IP
Sbjct  31311  IP  31306
```

Figure 13 *tblastn* alignment of CDS 1_3482_0 (query) against the contig48 nucleotide sequence (subject)

The percent identity for this alignment and the length of sequence conservation is better than all the matches we have previously found on contig47. In addition, there is a medium quality splice donor site at the end of this alignment and an SGP gene prediction (Figure 14). Therefore, we will annotate this region as the first exon of *Or13a* and make a note in the GEP Annotation Report that the model for this gene on contig47 is a partial annotation that is missing its first exon.

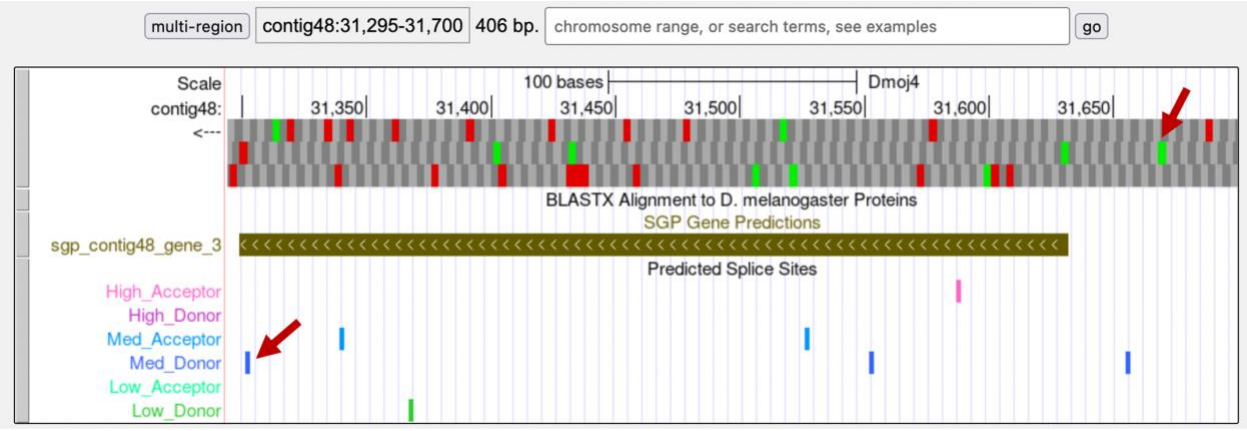

Figure 14 Region of contig48 containing the first coding exon of the *Or13a* ortholog in *D. mojavensis*

Exons with no sequence conservation

Small exons can be quite difficult to annotate. Below is a step-by-step strategy for annotating the first exon of *CG31999* on contig50 in *D. mojavensis*. *CG31999* in *D. mojavensis* spans five contigs on the forward strand with most of the exons demonstrating moderately strong conservation to the *D. melanogaster* ortholog.

Step 1: We will begin our analysis by performing a BLAT search of the *D. melanogaster* CG31999-PA protein against the *D. mojavensis* projects from the dot chromosome (Sep. 2008 assembly). The BLAT result shows that contig50 only contains the first half of *CG31999* and *CG5262* is the gene that is located closest to the start of *CG31999* (Figure 15). A *tblastn* search of the last CDS of *CG5262* (4_7507_2) against contig50 placed this CDS at 18,230-18,985. Examination of the end of the *tblastn* alignment using the Genome Browser placed the end of the coding region of *CG5262* at 18,991.

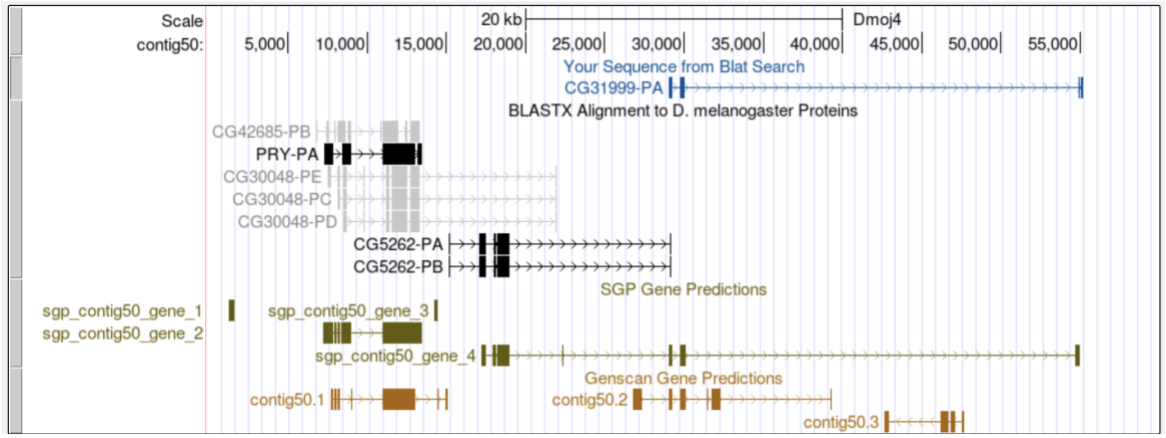

Figure 15 BLAT alignment shows that *CG31999* is located next to *CG5262* on contig50

A *tblastn* search mapped the second CDS of *CG31999* (2_10722_2) to 29,016-29,591 in frame +3. (Note that the first amino acid is missing from the alignment, and there is a novel intron at 29,284-29,349.) Examination of this region in the Genome Browser shows the CDS begins at 29,011 and the acceptor site is in phase 2 relative to frame +3 (Figure 16).

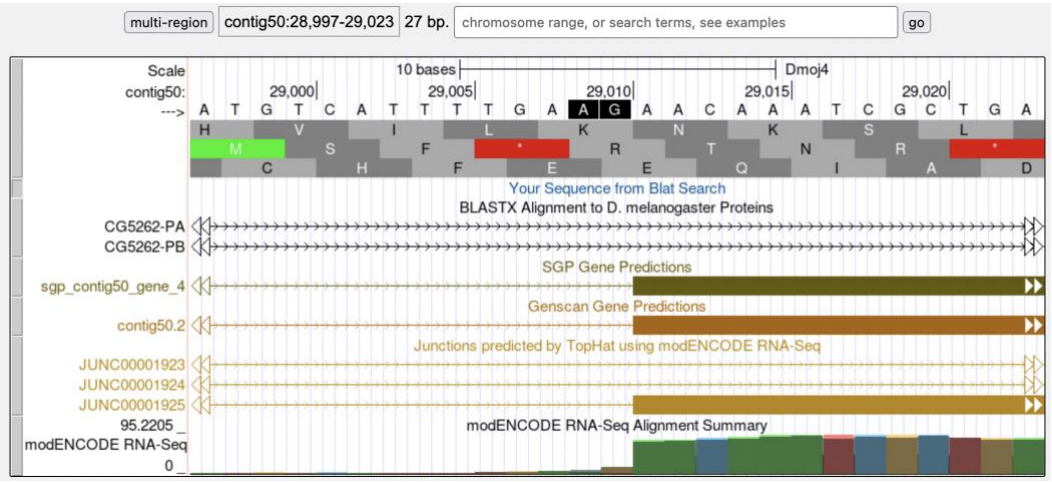

Figure 16 Phase 2 splice acceptor site at the beginning of CDS 2_10722_2

The analysis above allows us to narrow down the location of the first CDS (1_10722_0) to the region between 18,992-29,010. Performing a *tblastn* search within this subject subrange with the low complexity filter and compositional adjustments turned off and an Expect threshold of 1e5 results in 10 hits to the CDS 1_10722_0 sequence. However, 8 out of the 10 hits are found on the minus strand. The best match (with an E-value of 7.7) on the positive strand is located at 21,124-21,153 in frame +1 (Figure 17).

Range 8: 21124 to 21153 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Identities	Positives	Gaps	Frame
12.7 bits(21)	7.7	5/10(50%)	7/10(70%)	0/10(0%)	+1

```
Query  5      LCAFMCFLIV  14
              +CA+   LIV
Sbjct  21124  VCAYYENLIV  21153
```

Figure 17 *tblastn* hit to CDS 1_10722_0 in the positive strand with high E-value

Examination of this region in the Genome Browser shows an open reading frame in frame +1 that is substantially larger (Figure 18) than the expected size of the CDS 1_10722_0 (i.e., 24 amino acids). Consequently, while this region might contain a novel CDS in the *D. mojavensis* ortholog of *CG31999*, it is likely not the orthologous CDS of 1_10722_0.

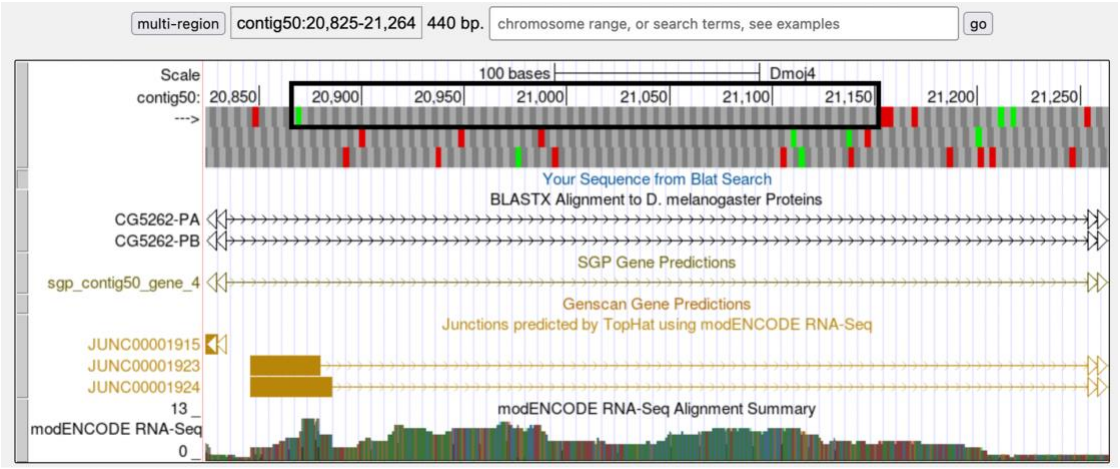

Figure 18 *tblastn* alignment of CDS 1_10722_0 against contig50 identified a large open reading frame in frame +1

Step 2: Since there is no sequence conservation, perhaps there is some evidence of the first exon in the Genome Browser view of the region. However, unlike contig47, there are very few indicators as to where we could find the CDS 1_10722_0 (Figure 19). While there are some predicted exons from gene predictors, RNA-Seq read coverage, and splice donor site predictions in this region, there is not overwhelming evidence that would favor one specific candidate over other candidates as the putative location of the first exon.

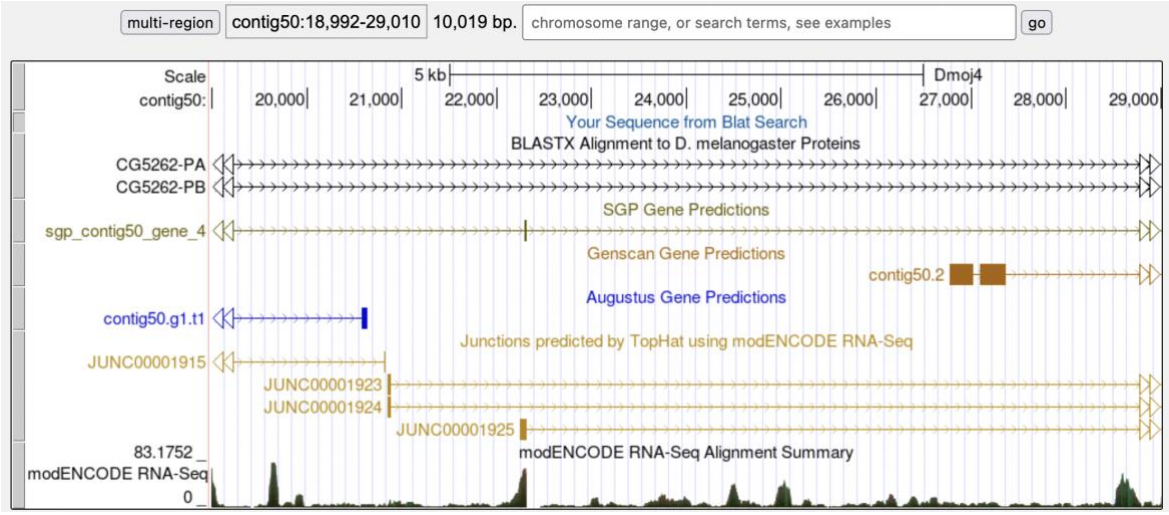

Figure 19 Genome Browser view of the search region for CDS 1_10722_0 in contig50

Step 3: With little evidence available in *D. mojavensis*, *D. virilis* might provide some additional information. While there is no conservation to the *D. melanogaster* CDS 1_10722_0, it might be possible to find some conservation with the annotated *D. virilis CG31999* model. Going back to the GEP UCSC Genome Browser, I navigated to the region dvir_dot_finished:833,500-839,500 in the *D. virilis* Manuscript (GEP/2010) assembly (Figure 20). I then click on the *CG31999* feature in the GEP Gene Annotations track and then click on the "Translated Protein" link to retrieve the *D. virilis CG31999* protein sequence.

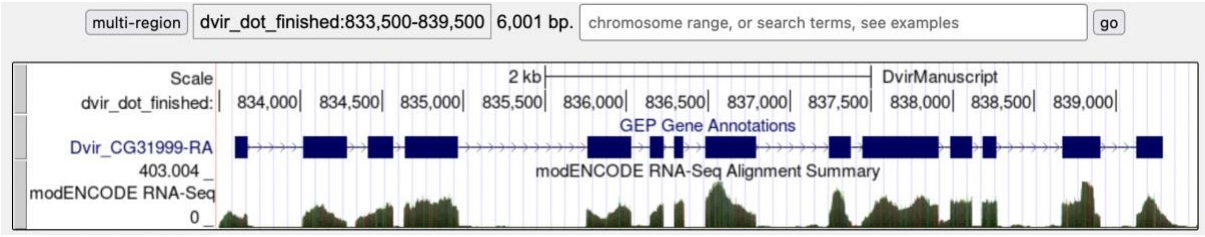

Figure 20 *D. virilis* orthologous region for *CG31999*

Use the translated *D. virilis* protein sequence in a *blastp* alignment against the *D. melanogaster* amino acid sequence to determine the sequence of the first exon in *D. virilis*. Figure 21 highlights the region that corresponds to this sequence, which we will use to compare to the *D. mojavensis* assembly. Unfortunately, this search did not produce any highly significant alignment either.

>Dvir_CG31999-RA_prot length=922
MRSPHIINIIGILFFLSWGRVLADEEQISDYIRKCCIIGLRNARTTNECE
KMESAVSNISRLWIGLCSSTFGVCCSRELDRQHCELGRLAALEGTSCNNG
SSTTYKNCCRACQDSYRICCSEDGFANQSDEKENTLGIDAHHAEPEEEDA
KPDDEDQDGTIVLADDDDICGKIPNLCAHICENTFDAYKCSCNPGYTLDD
NNVTCSENKDHLCPSGYVLDAKQGKCVDIDECKEQLHECKSHQYCHNTIG
GYHCLNVKAKSCPPGYLYNVKSEECEDQDECIQSPCEKGYKCSNYRGGYD
CNPVDSQACGTGFYLKEGSCADIDECDYNVTNSCKTEMHQECKNTVGSYH
CDCLPGYSLDVTQNECVDINECSINNHNCLSTQRCDNTIGSYICTRWTSC
GTGYTLNAETGNCDDDDECALNTHNCPLNYDCYNTKGSFRCHRKTTTTTS
TTSTTTIPLPATSTTAAPRRDVNPQFVSQLAYNRMDYAPSYLTNYGSIAQ
RPCTSGFYRNNLGACVDINECIEYKPCMNHERCINTNGSYRCESLIQCPA
GLRSTPDGTSCVDINECETGDHNCGEKQICRNRHGGYICACPPGHQVTRL
PDGVNSCEDINECAQDQPVCSSNAHCFNTIGSYYCECKSGFQKKPLNGPD
QDNWQTHSQCFDVDECQSIPGLCQQKCVNFWGGYRCTCNSGYELSQDNRT
CNDIDECEVHKDYKLCMGYCINTAGSYQCSCPRGYTLAADKNTCRDIDEC
ETKNENHVCTGRNDICTNIRGSFKCTTINCPNGYINDQDQKNRCRQTNNF
CEGDECYTKPSAYTYNFITFVSKLMIPPDGRTIFTLRGPIWYDDIDFELN
VVRVQAAPNVERASEIYFDTLKSNNQVNLVLKKALEGPQDVELDLSMTVF
TNGMPRGKSVAKLFLFVSQYTY

Figure 21 *D. virilis CG31999* amino acid sequence with the first exon highlighted

Step 4: Because there is no sequence conservation at the amino acid level, we will try to find conservation at the nucleotide level through a MUSCLE multiple sequence alignment analysis with *D. mojavensis*, *D. virilis*, and *D. grimshawi*. Any conserved blocks of nucleotide sequence could indicate a potential first exon region.

I first extracted the nucleotide sequence from *D. mojavensis* containing some of the *CG5262* and *CG31999* flanking sequence in order to anchor the ends of the multiple sequence alignment. I next extracted some of the *CG31999* flanking sequence from the *D. virilis* Manuscript assembly. However, because the 5' end of the orthologous region is flanked by *yellow-h* instead of *CG5262*, I cannot use *CG5262* to anchor the multiple sequence alignment. (This lack of synteny is also observed in *D. grimshawi*.)

When I attempt to search for *yellow-h* on the *D. mojavensis* dot, there are no hits to this *D. melanogaster* gene. A FlyBase *tblastn* search against the *D. mojavensis* assembly revealed that the putative ortholog of *yellow-h* in *D. mojavensis* is found in scaffold_6308. Most of the genes on this *D. mojavensis* scaffold are orthologous to genes that are found on the *D. melanogaster* Muller A element (X chromosome).

All of the extracted sequences are saved in a FASTA file (Figure 22) and analyzed by the MUSCLE program. Unfortunately, the MUSCLE alignment revealed no strongly conserved blocks of sequences among the three species.

```
>Dmoj4_contig50:18228-29283
CGTTTATTTTAATGATATGCATTGCTCTCAAAATGCTCGTCATTGATGGG
GTAAAAGGACATCCAGTTGCCGTAAATTTTTATGGTATTCCCTCATTGTT
TGGCGCCGGTGTTTACTCGTTCATGTGCCACCATTCGTTGCCCAGTTTAT
TGGCACCAATTAAACATAAGTCAATGGCTAAAAAAATTCTTTCATACGAT
TACATTTTAATTTGTTCGTTTTATATAGTTCTGGCGATTACAGGAATATT
... ...
>DvirCAF1_scaffold_13052:1412061-1413937
GAATTTATGGTTCCGATAGACATACTATTAAATGAATCTCTTTGGACCAA
TGGCACAATAGATACGTCCAAACTTTTTATTTCGATTGGCGATCGCGGAT
TCAACGGTCAATCTTCGACATCTGGAATCGCCAGAAATGGTGTAATGTTC
TTTACCCAAGTACATCGGGATAACATCGGTTGTTGGGATACAACTAAACC
ATACTCTCGATCAAATTTGGGAATACTCCTCGATGCTAACAAATTCCCGA
... ...
>DgriCAF1_scaffold_14822:432808-436082
GAATTTATGGTGCCCCTTGACTCGCTCTTGAATGAATCCCTTTGGATCGG
CAACAGCAGCGTGGACACCTCGCAATTATTTGTTCCAATTGGGGATCGTG
GCTTTGGGGGGCAATCTTCGACATCGGGCATAGCCCGAAATGGTGTTATG
TTCTACACACAAGTGCATCGTGATAATATCGGCTGTTGGGACACCAAAAA
GCCCTATAATCGTGCAAATCTCGGTATGCTTTTGGAACCAGATAATGCTT
```

Figure 22 FASTA file for MUSCLE analysis

Step 5: With no conservation at any level, we will now try using the Small Exons Finder program to search the contig50 sequence for open reading frames that match the length of the first exon of *CG31999*. The Small Exons Finder is available through the "Resources & Tools" section of the F Element project page on the GEP website.

Based on our previous analyses, we know that the acceptor site for the second exon is in phase 2, so the donor site for the first exon must be in phase 1. The Small Exons Finder identified two regions that match the search criteria: 19,346-19,418 and 22,579-22,651 (Figure 23). However, these two regions are only weakly supported by the RNA-Seq data. Therefore, I am not inclined to use either of these open reading frames as part of the annotated model.

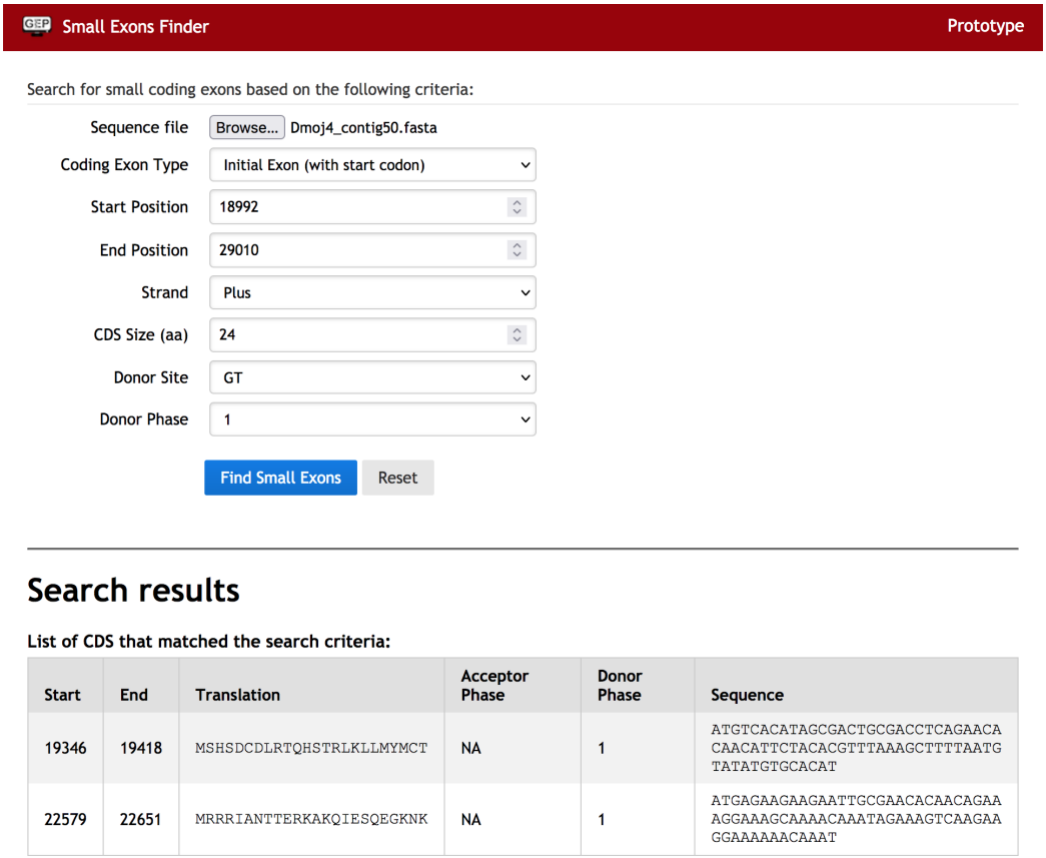

Figure 23 Small Exons Finder identified two regions with correct splice donor phase within the region of interests

Step 6: Given the limited amount of evidence available, we will utilize any gene prediction tracks on contig50 that align well with the rest of the exons for *CG31999*. While the Genscan gene prediction track has two predicted exons upstream from the second exon, there is also an SGP gene prediction that is much better supported by the RNA-Seq data and TopHat splice junction predictions (Figure 24).

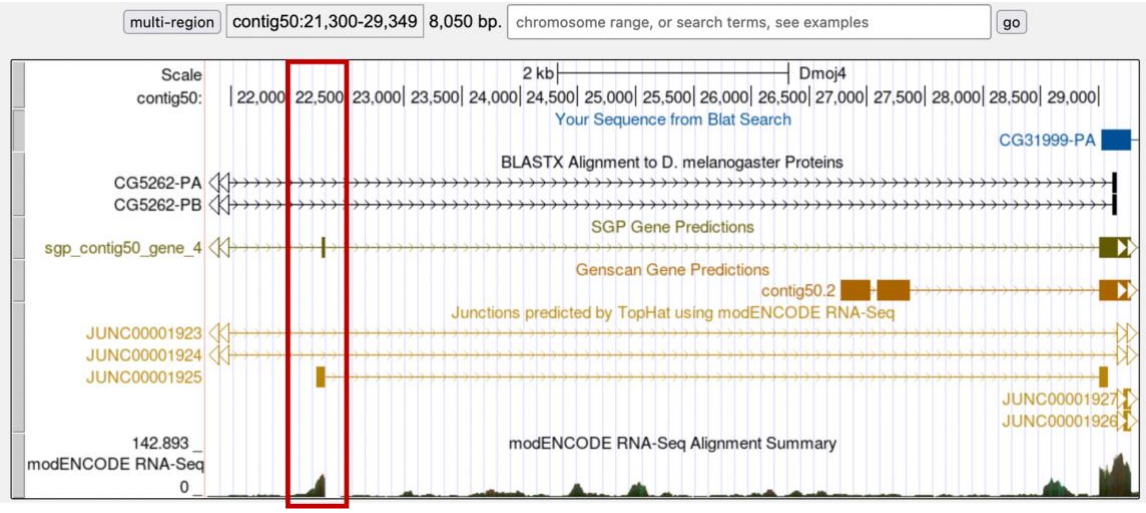

Figure 24 Alignment of gene predictions and RNA-Seq data upstream of the second exon

The Cufflinks transcripts, Oases transcripts, and spliced RNA-Seq reads all show an intron between the 5' end of the second exon and the 3' end of the predicted exon from SGP. The modENCODE RNA-Seq Alignment Summary track shows substantially more reads aligning to this region than the region that corresponds to the Genscan prediction. In addition, there is a high quality splice donor site at the end of the SGP predicted exon. While this prediction is not for an initial exon, the evidence is stronger here than the location suggested by the Genscan prediction. Therefore, we will use the +3 open reading frame from this SGP gene prediction to create the first coding exon for *CG31999* (Figure 25).

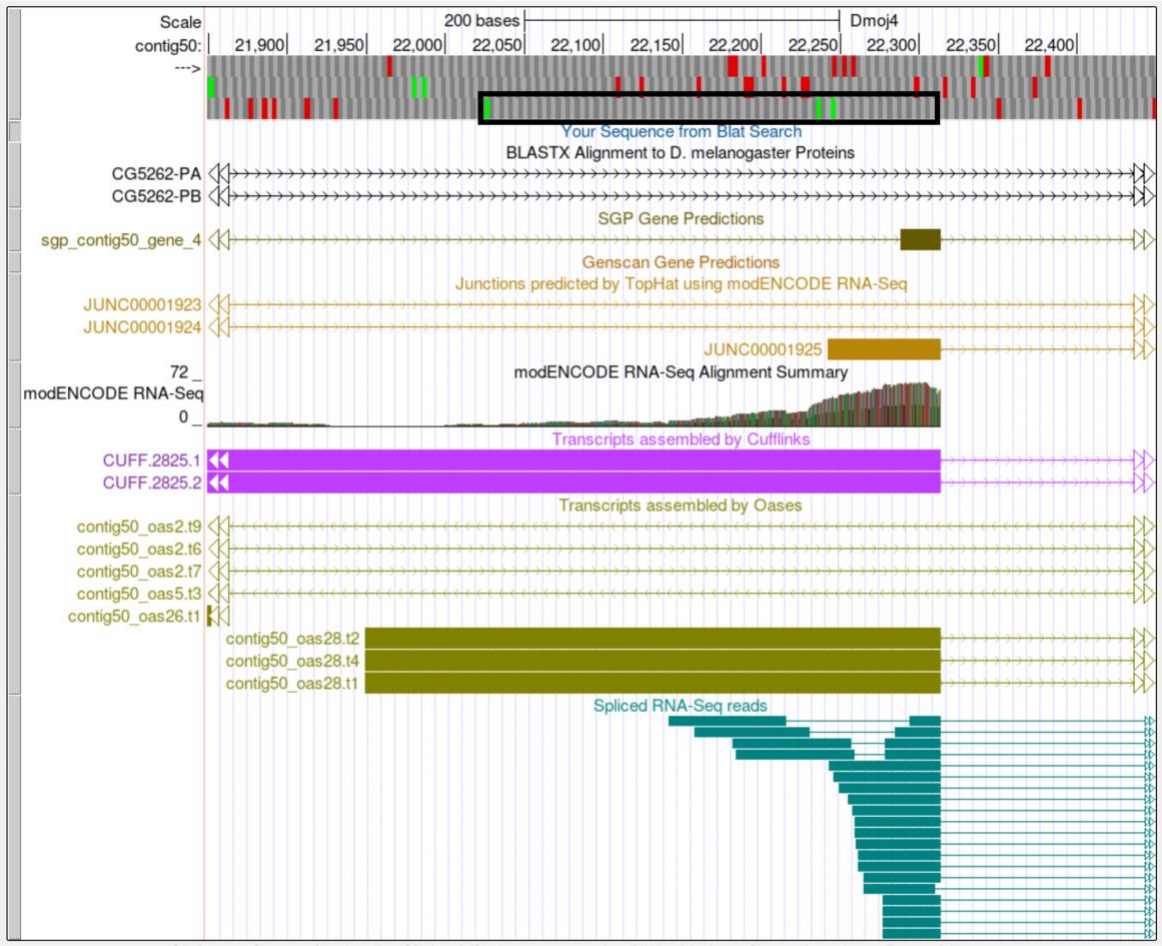

Figure 25 Region surrounding the best candidate for the initial coding exon of *CG31999*

The coordinates for the initial exon of *CG31999* will be 22,026-22,314. Use of frame +3 creates a matching phase 1 splice donor (22,315-22,316). While the length of this initial exon is much larger than the corresponding exon from *D. melanogaster* (96 vs. 24 amino acids), we have exhausted all other possibilities for finding this CDS based on sequence or length conservation. Based on the available computational evidence shown in the Genome Browser, this is our best guess as to where the initial exon of *CG31999* is located in *D. mojavensis*. Until additional evidence becomes available, we will not be able to prove or disprove the existence of this exon in *D. mojavensis*.

Note: if RNA-Seq data is unavailable or if it does not support a computational gene prediction, the next best option is to use the gene prediction closest to the adjacent annotated exon (after you have verified that the splice sites for the predicted exon have the correct phase). If there are no open reading frames that have the correct phase or lack the start codon (when you are looking for an initial exon) or stop codon (when you are looking for a terminal exon), then the isoform might not exist in the species that you are trying to annotate. However, before making this decision, you should attempt all the annotation techniques described above and consult with Wilson or Chris.

Exons of small length

Small exons might be conserved, but due to their size, are difficult to locate with just a simple BLAST analysis. We will use CDS 1_2121_0, the initial exon for isoforms A, C, D, F, and G of *unc-13*, to demonstrate the step-by-step strategy for annotating small exons below.

Step 1: The Gene Record Finder entry for *unc-13* shows that CDS 1_2121_0 only consists of 2 amino acids (Figure 26). We will try to find this CDS on contig12 of *D. mojavensis*.

Select a row to display the corresponding CDS sequence:

FlyBase ID	5' Start	3' End	Strand	Phase	Size (aa)
1_2121_0	903,342	903,337	-	0	2
5_2121_0	903,282	898,393	-	0	1630
7_2121_0	885,891	885,760	-	0	44

Sequence viewer for unc-13: unc-13:1_2121_0

```
>unc-13:1_2121_0
MT
```

Figure 26 The initial CDS 1_2121_0 contains only 2 amino acids.

The CDS usage map shows that there is a single CDS (5_2121_0) between CDS 1_2121_0 and 7_2121_0 in the A, C, D, and G isoforms. By contrast, there are three CDS's (2_2121_0, 3_2121_2, and 4_2121_2) between CDS 1_2121_0 and 7_2121_0 in the F isoform (Figure 27). However, closer examination of CDS coordinates and the JBrowse view shows that these three CDS's in the F isoform actually overlaps completely with the much larger CDS in the A, C, D, and G isoforms (Figure 28).

CDS usage map:

Isoform	1_2121_0	5_2121_0	2_2121_0	3_2121_2	4_2121_2	6_2121_0	7_2121_0
unc-13-PG	1	2					3
unc-13-PA	1	2					3
unc-13-PC	1	2					3
unc-13-PF	1		2	3	4		5
unc-13-PD	1	2					3
unc-13-PB						1	2
unc-13-PE						1	2

Figure 27 CDS usage map for the 7 isoforms of *unc-13*

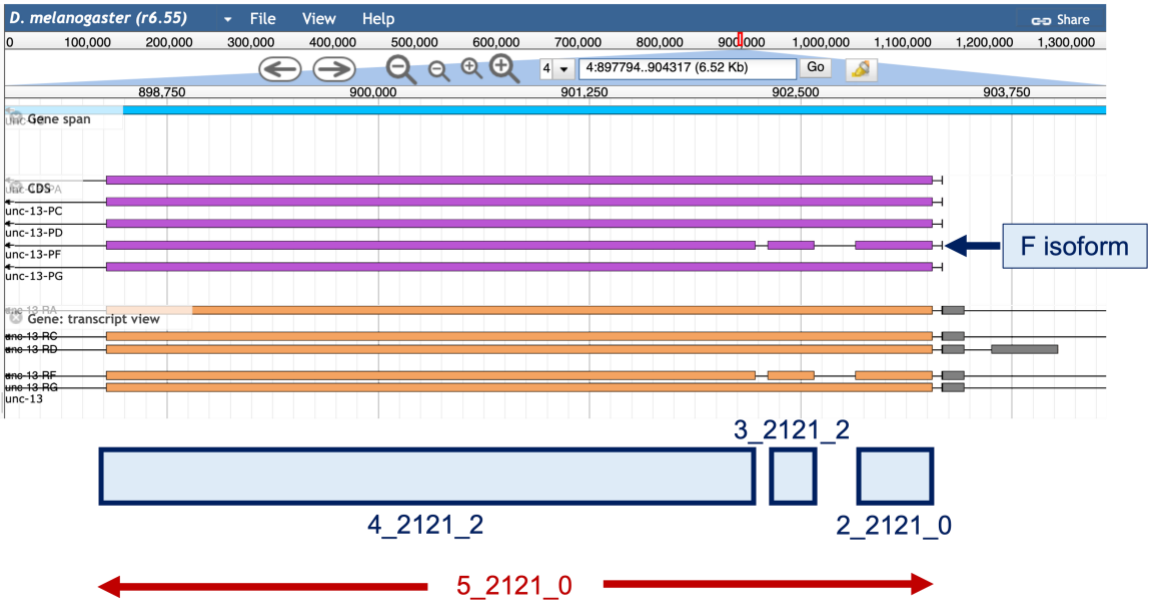

Figure 28 The CDS 5_2121_0 overlaps with three smaller CDS's in the F isoform (2_2121_0, 3_2121_2, and 4_2121_2)

To narrow down the search region, we will use BLAT to locate CDS 5_2121_0 (i.e., the second CDS of the A, C, D, and G isoforms). BLAT found matches in both contigs 11 and 12. Because the match to contig11 is at the end of the contig (Figure 29), we will focus our efforts searching for the first CDS in contig12.

D. mojavensis (Dmoj4) BLAT Results

BLAT Search Results

Go back to contig1 on the Genome Browser.

Custom track name: blat unc-13:5_2121_0

Custom track description: blat on unc-13:5_2121_0

Build a custom track with these results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
browser details	unc-13:5_2121_0	277	266	1630	1630	81.8%	contig12	+-	16894	21186	4293
browser details	unc-13:5_2121_0	206	1198	1630	1630	80.2%	contig11	+-	56894	58282	1389

Figure 29 BLAT search of CDS 5_2121_0 found matches to contig11 and contig12 in the *D. mojavensis* assembly

Step 2: Examination of the *blastx* track for contig12 shows that *Rad23* is located next to *unc-13* (Figure 30). Because nested genes are rare in *Drosophila*, we can define one end of the search boundary based on the start of the first CDS of *Rad23*. The *tblastn* alignment of the first CDS of *Rad23* (1_2303_0) against contig12 shows that the CDS begins at 23,720.

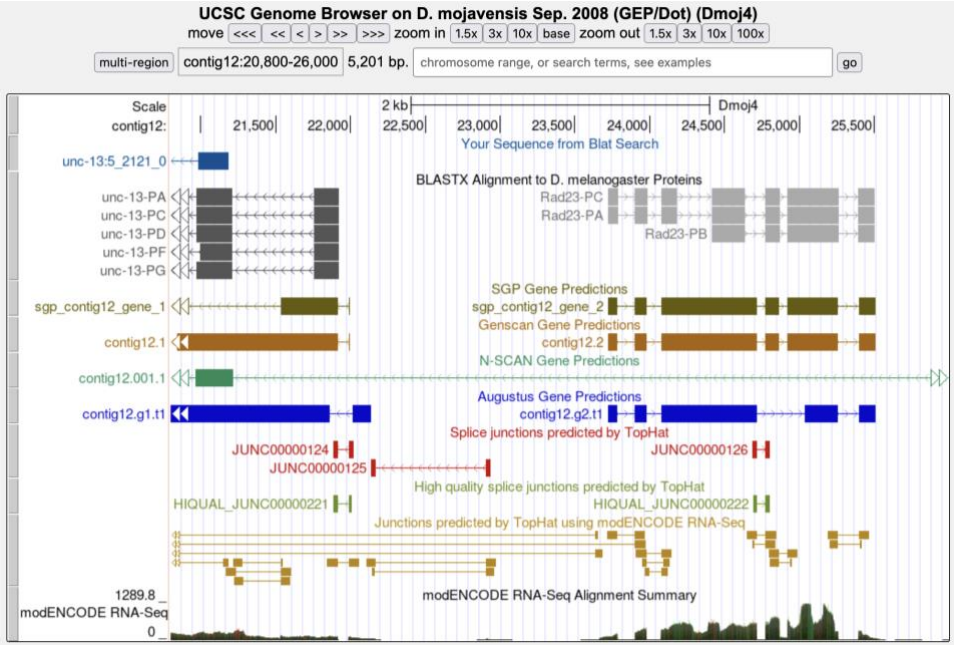

Figure 30 Use the location of the nearby gene (*Rad23*) to help define the search boundary

The results in Figure 29 shows that the first 265 residues of CDS 5_2121_0 are missing from the BLAT alignment to contig12. Extrapolating the size of the missing region from the beginning of the alignment, we expect to find the beginning of CDS 5_2121_0 at around 21,981 (i.e., 21186 + (265*3)). A closer examination of this region indicates that there is a splice acceptor site at 21,917-21,916 that is strongly supported by the Genscan and SGP gene predictions as well as the TopHat junctions and RNA-Seq data (Figure 31).

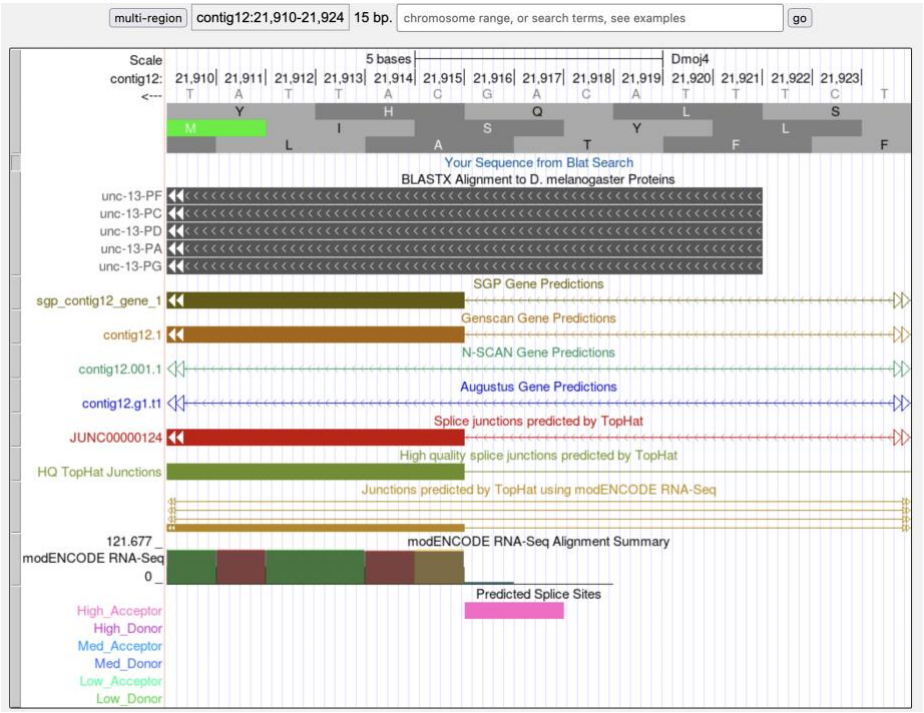

Figure 31 The phase 0 splice acceptor site for CDS 5_2121_0 is supported by multiple gene predictions and RNA-Seq data

Based on the results of the *tblastn* search of CDS 5_2121_0 against contig12, we know that the CDS is in frame -1, which means that the splice acceptor site at 21,917-21,916 is in phase 0. This also means that the phase of the donor site for the first CDS of *unc-13* (CDS 1_2121_0) is in phase 0.

Collectively, our analysis helps narrow down the first CDS to the region 21,916-23,719. In addition, this initial CDS must begin with a methionine and ends with a phase 0 donor site.

Step 3: Looking at this region in the Genome Browser, we can see that the region around 21,990-22,168 have strong evidence from RNA-Seq expression data and computational predictions: a few gene and splice junction predictors have demarcated the same CDS boundaries, and the RNA-Seq data fits these boundaries as well (Figure 32).

Figure 32 Multiple gene predictors and RNA-Seq data identified the putative location of a small exon upstream of the *unc-13* CDS 5_2121_0 (red box)

Zooming into this region, we see that the amino acid sequence is MT (in frame -2), the same as the *D. melanogaster* ortholog. This coding exon has a phase 0 splice donor, which is compatible with the phase 0 acceptor we have previously identified (Figure 33).

According to the RNA-Seq read coverage, there are regions upstream of this CDS that is part of this gene. However, given the presence of in-frame stop codons (and the gene model in *D. melanogaster*), these upstream regions are likely part of the 5' UTR's of *unc-13*.

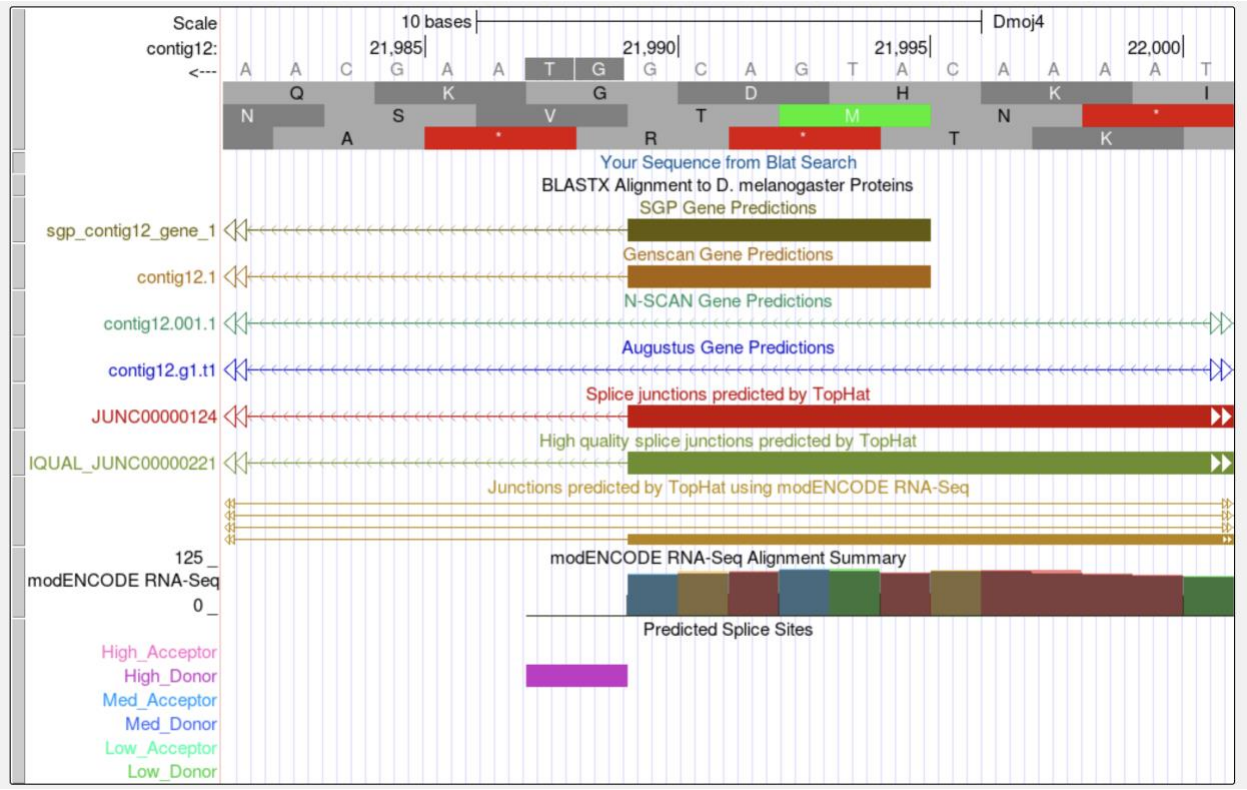

Figure 33 orthologous *unc-13* CDS 1_2121_0 on contig12

Alternate approach: Using the search boundary we have previously defined, we can also search for candidate open reading frames using the Small Exons Finder. With our search criteria, as outlined in Figure 34, the Small Exons Finder identified three candidates. Only one of these candidates (at 21,990-21,995) contains the same amino acid sequence as the orthologous CDS in *D. melanogaster* (MT). When we investigate the three possible candidates using the Genome Browser, we find that the only feature that is supported by multiple gene predictors and RNA-Seq data spans from 21,990 to 21,995. Consequently, the best candidate for the orthologous *unc-13* CDS 1_2121_0 is located at 21,990-21,995.

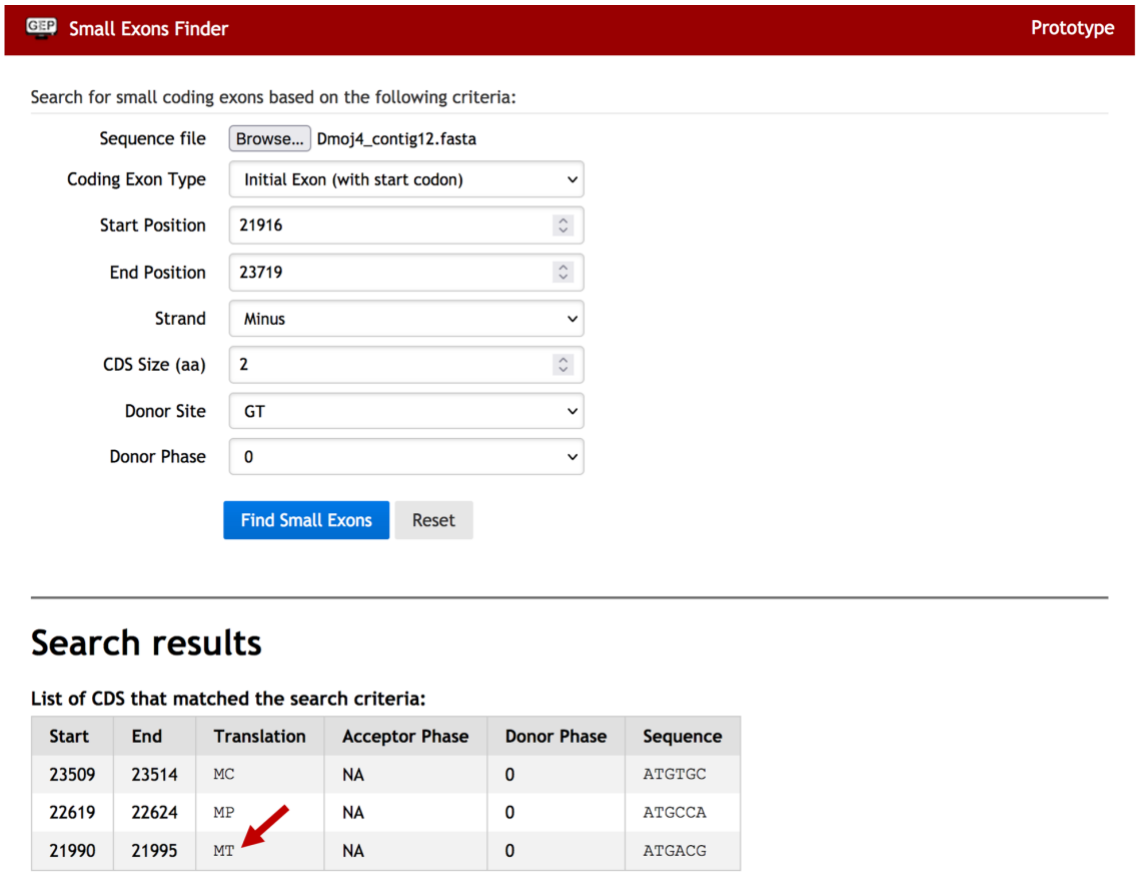

Figure 34 Small Exons Finder analysis for CDS 1_2121_0