Annotation of *Drosophila grimshawi* contig7

A

Andrew Stein
May 10, 2009

Abstract

In this paper, I present information regarding the annotation of *Drosophila grimshawi* contig7. This contig contains sequence from the fourth chromosome of *D. grimshawi.* I began work on annotating this contig by first examining the GenScan output information. GenScan shows one predicted feature oriented along the plus strand of contig7 that consists of sixteen exons. This feature represents the *D. grimshawi* ortholog to the *D. melanogaster* gene *unc-13*. In *D. melanogaster*, *unc-13* has three different isoforms, *unc-13-RA, unc-13-RB, and unc-13-RC*. Furthermore, it consists of a total of 25 translated exons that are shared across the three isoforms. All three of the isoforms are represented in *D. grimshawi* contig7; however, this contig only has ten of the translated exons (including the translation start site). All of the exon boundaries for the ten exons present in contig7 have been appropriately noted (Table 1 and Figure 0).

Drosophila grimshawi			
***Unc-13* Isoforms RA, RB, RC**			
Exon	**Start**	**End**	**Exon Size**
2	3294	3299	2
3	3371	8437	**1689**
4	19872	20003	44
5	29665	29739	25
6	29805	30024	73
7	30081	30201	40
8	32310	32416	36
28 (4 and 5)	20751	20759	3
32 (7 and 9)	30759	30865	36
50 (4)	13650	19520	**1957**
Exons 2-8 are in isoform RA.			
Exons 50, 4-7, 32 are in isoform RB.			
Exons 2-4, 28, 5-7, 32 are in isoform RC.			

Table 1: Summary of the exon boundaries for the *D. grimshawi* ortholog of *unc-13*

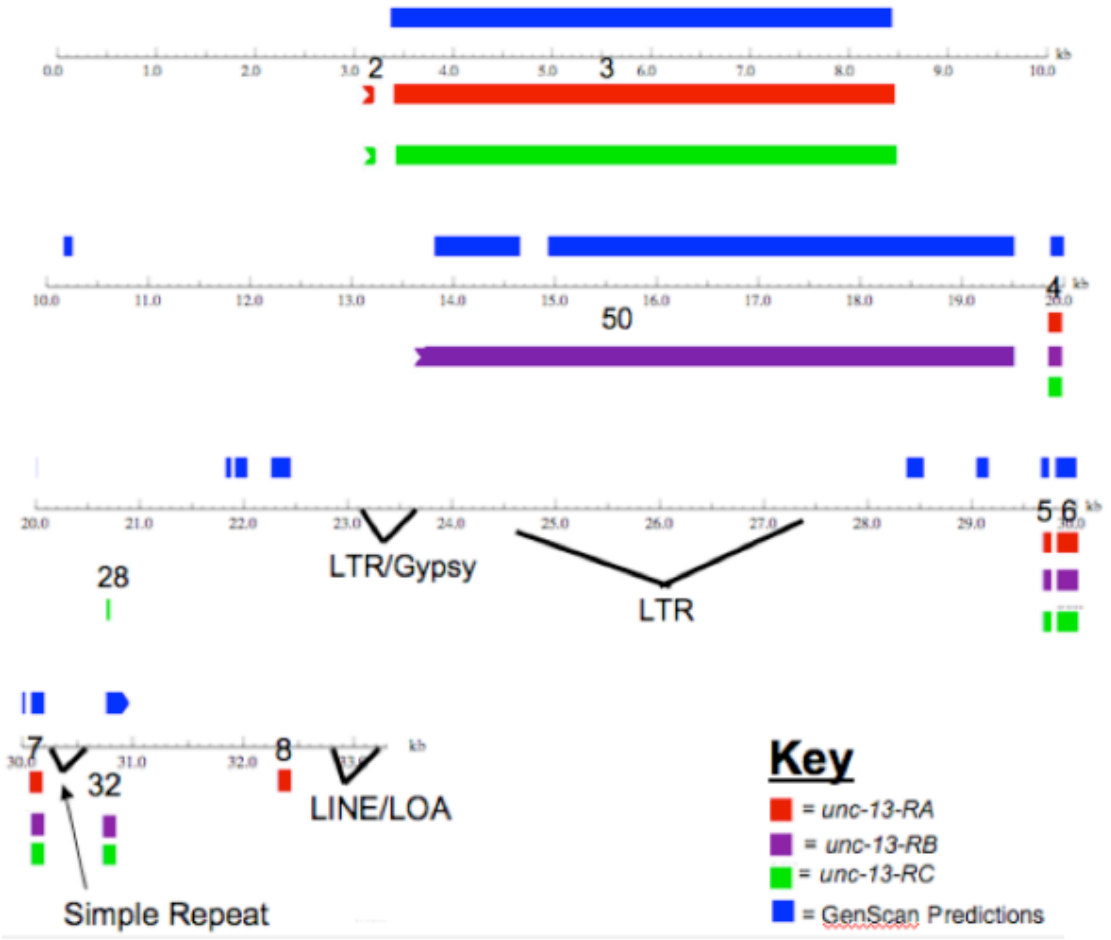

Figure 0: Final map showing the relevant features and repetitive elements present in *D. grimshawi* contig7

Introduction

Annotation is an important process that reveals the interesting aspects hidden behind DNA sequence information. This entire process identifies the genes, pseudogenes, noncoding regions, repetitive regions, and all other important aspects of a genomic sequence. In this project, I used information from previously annotated *D. melanogaster* proteins to elucidate the relevant genes present in contig7. Although *D. melanogaster* and *D. grimshawi* diverged from one another around 40-60 million years ago, many of the proteins are shared between these two species. Therefore, it is reasonable to use *D. melanogaster* information to find biologically important facets of *D. grimshawi* sequences. In this paper, I discuss the steps that I took to determine the exon boundaries of the *D. grimshawi unc-13* ortholog that is present in contig7. Furthermore, I examined the repetitive regions present in contig7 along with the conservation of the *unc-13* gene across many species.

Project Work Flow

Initial GenScan Results

To begin work on annotating contig7, I surveyed the initial GenScan ouput. GenScan predicts one feature oriented along the plus strand of contig7 (Figure 1). Furthermore, the output indicates that this feature contains a total of sixteen exons. GenScan shows that the initial exon of this feature is not located in contig7, but the terminal exon is present. The GenScan predictions were a useful starting point, but some proved to be incorrect after further analysis.

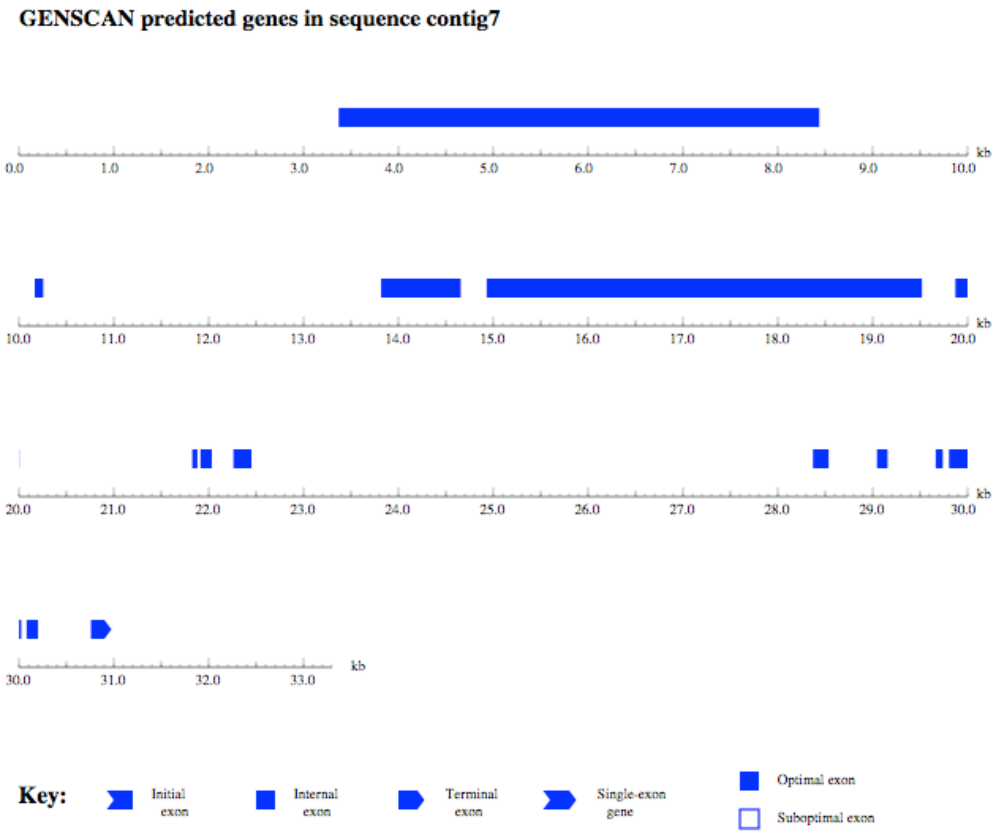

Figure 1: Initial GenScan output showing one feature present in contig7

Annotation of Predicted Feature

First, I went to the UCSC browser to determine the proteins from *D. melanogaster* that align to any portion of contig7 (Figure 2). This browser shows that three isoforms of the *unc-13* gene from *D. melanogaster* align to my contig.

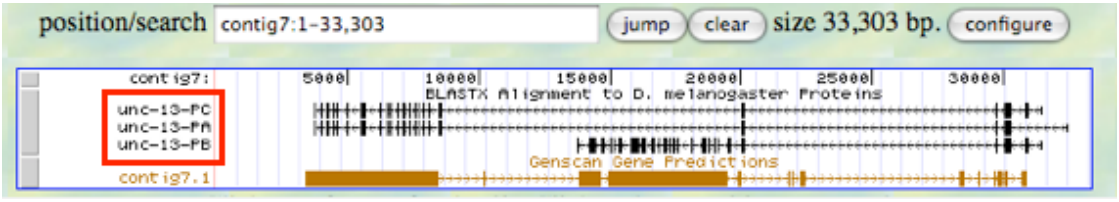

Figure 2: UCSC browser shows an alignment between *unc-13* from *D. melanogaster* and contig7

To provide further evidence that the *unc-13* ortholog is present in contig7, I extracted the predicted protein sequence of my feature from the GenScan output and ran a BLASTP search with this sequence against the annotated proteins from *D. melanogaster*. Once again, the best BLAST hit was to the three *unc-13* isoforms. Therefore, I investigated this gene in further detail and discovered that it is located on the fourth chromosome in *D. melanogaster,* and it plays a role in vesicle maturation during exocytosis. It is also involved in neurotransmitter release by acting in synaptic vesicle priming prior to vesicle fusion.

Once I determined that the *D. grimshawi* ortholog to *unc-13* is present in contig7, I moved on to find the locations of the exons. I used the Gene Record Finder website to find the protein sequences for each exon from *unc-13* in *D. melanogaster.* This website shows that *unc-13* is comprised of 25 total translated exons that are shared among the three isoforms: *unc-13 RA*, *RB*, and *RC*. I investigated the exons present in *unc-13-RA* first. Exon number two, the first translated exon, only codes for two amino acids (M and T); therefore, I decided to annotate exon three first. To begin this process, I extracted the protein sequence of exon three from *D. melanogaster* and used the TBLASTN align two sequences analysis to align this protein sequence (query) to the masked contig7 sequence (subject) (Figure 3).

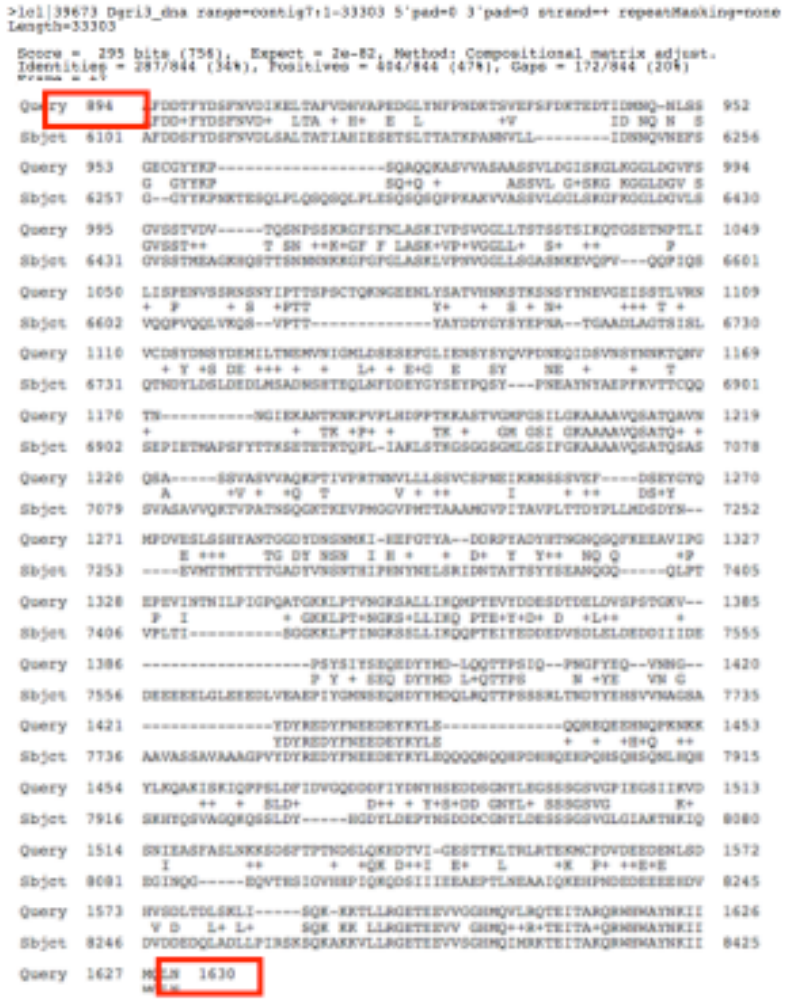

Figure 3: Alignment of protein sequence for exon three with masked contig7 sequence

As Figure 3 displays, amino acids 894-1630 from exon three align with contig7 between base pairs 6101 and 8437 in frame +2. Next, I relaxed the parameters for the TBLASTN alignment in an attempt to align the remainder of the protein sequence from exon three with contig7. However, relaxing the parameters did not help me to discover any new hits. I wanted to determine the boundaries of this exon in contig7, so I navigated to base pair 8437 using the UCSC browser. As Figure 4 shows, all of the gene prediction models indicate the end of an exon at this base pair location. Furthermore, a "GT" splice donor is present immediately after base pair 8437, and there is strong sequence conservation at this location among the other *Drosophila* species (Figure 4). This gives exon three a donor phase of 0.

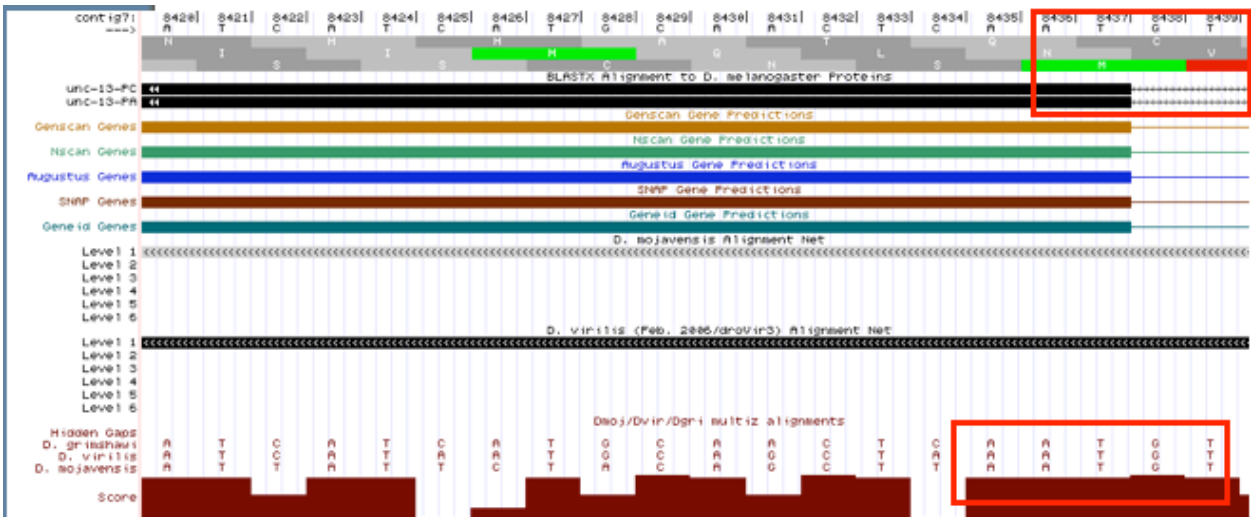

Figure 4: UCSC browser view showing the end of exon three within contig7 around base pair 8437

Based upon the information from the UCSC browser and the TBLASTN output, I am confident that exon 3 of *unc-13-RA* ends at base pair 8437 in contig7. Since I did not know where exon three began, I decided to estimate its start point using the assumption that exon length is conserved between *D. melanogaster* and *D. grimshawi*. From the Gene Record Viewer website, I learned that exon three is 1630 amino acids long in *D. melanogaster*. Using this value, I determined that base pair 3547 of contig7 is far enough away from base pair 8437 for a total of 1630 amino acids to be present in this exon. Therefore, I navigated to base pair 3547 of contig7 using the UCSC browser and surveyed the gene prediction models to find the start of exon three. All of the gene prediction models indicate that this exon begins at base pair 3371 (Figure 5). Furthermore, as Figure 5 displays, there is a high probability splice acceptor site that is conserved between all of the other species of *Drosophila*. All of these pieces of information suggest that exon three does indeed begin at base pair position 3371. Thus, exon three is most likely located between base pairs 3371 and 8437 in contig7.

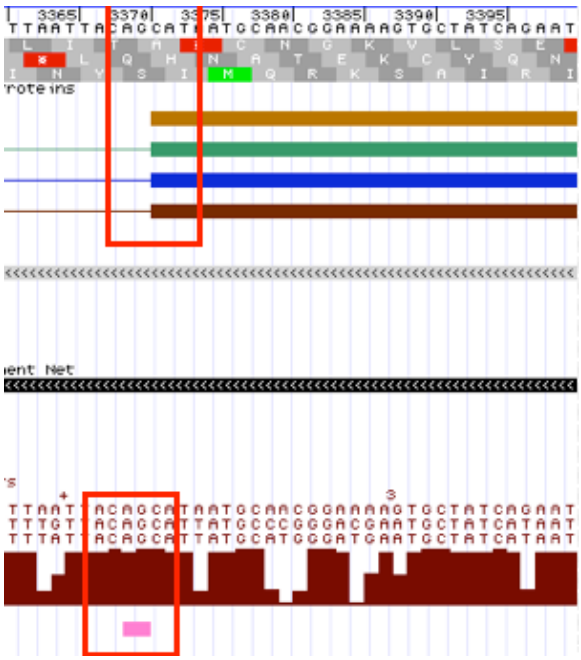

Figure 5: UCSC browser showing the start of exon three

Once I determined the boundaries of exon three, I knew that the small second exon must be located 5' of this region. I scanned all of the positive frames (since the gene is located on the plus strand) between base pair 3371 and the start of contig7 while taking note of all the "MT" amino acid pairs. I found a total of five possibilities after completing this scan. After investigating all five of these possibilities in greater detail, I realized that only two of them had "GT" splice donors within close enough range to remain as an exon solely consisting of amino acids M and T (Figure 6).

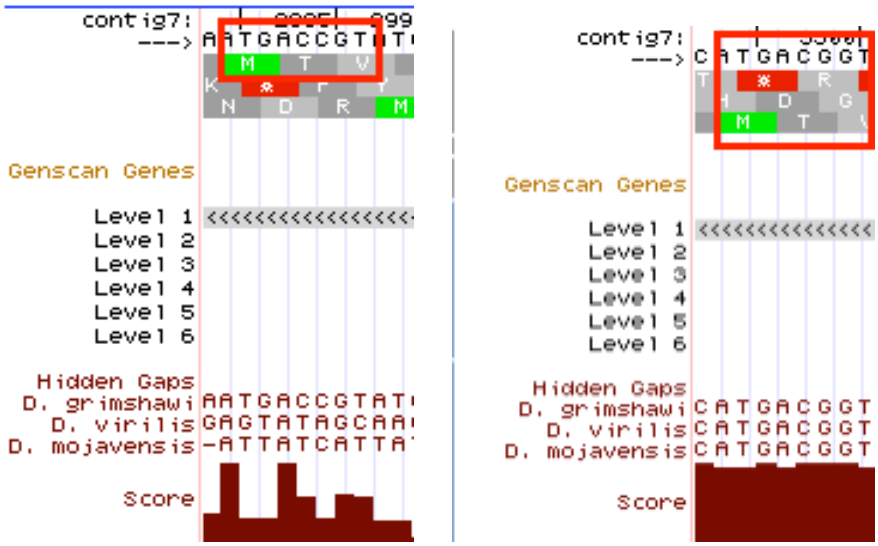

Figure 6: Two locations where amino acids MT are located and flanked by "GT"

Figure 6 shows that one of the two possibilities has much stronger conservation across the other species of *Drosophila*. This possibility is located between base pairs 3294 and 3299 of contig7.

I wanted to gather more information about what is present within this portion of contig7, so I examined the gene prediction models in this region (Figure 7).

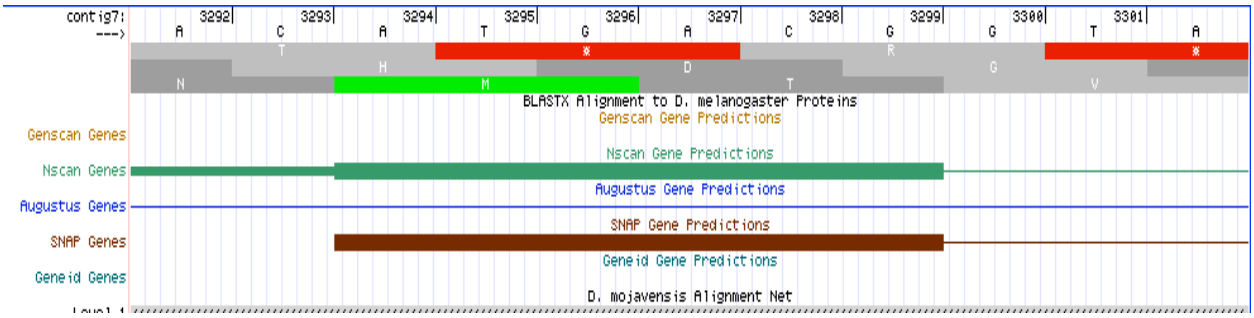

Figure 7: Gene prediction models between base pairs 3294 and 3299

Figure 7 reveals that both NScan and SNAP predict an exon between base pairs 3294 and 3299. Based upon location, conservation, and other gene prediction models, I am confident that exon two is located between base pairs 3294 and 3299. Additionally, exon two has a donor phase of 0, which matches up perfectly to exon three since it has an acceptor phase of 0.

The annotation of exons two and three proved to be more complicated than the annotation of the other exons for the *unc-13-RA* isoform. I followed a general procedure for annotating the remainder of the exons. Therefore, I will explain in detail how I annotated exon four, but I will not go into detail for the other exons since I followed the same basic protocol.

From the Gene Record Viewer website, I extracted the protein sequence for *D. melanogaster unc-13-RA* exon four. Then, I used the TBLASTN align two sequences analysis to align exon four to the masked contig7 sequence. This BLAST analysis shows that all of exon four aligns to contig7 between base pairs 19872 and 20003 in frame +3 (Figure 8).

```
>lcl|57905 contig7
Length=33303

 Score = 84.0 bits (206),  Expect = 2e-20, Method: Compositional matrix adjust.
 Identities = 40/44 (90%), Positives = 42/44 (95%), Gaps = 0/44 (0%)
 Frame = +3

Query  1      NGGGPGEVGLRTNGHPGDNPFYSNIDSMPDIRPRRKSIPLVSEL  44
              NG  PG+VGLR+NGHPGDNPFYSNIDSMPDIRPRRKSIPLVSEL
Sbjct  19872  NGTAPGDVGLRSNGHPGDNPFYSNIDSMPDIRPRRKSIPLVSEL  20003
```

Figure 8: BLAST alignment between exon four and contig7 masked DNA sequence

Next, I went to the UCSC browser to investigate the intron donor and acceptor sites to appropriately define the exon boundaries. I examined the browser around base pair 19872 in frame +3 and uncovered a strong acceptor of phase 0 immediately before base pair 19872 (Figure 9). This acceptor phase matches to that from exon three since both are of phase 0. Furthermore, there is strong conservation among the other species of *Drosophila,* and all of the gene predictor models show the start of an exon at this location. From all of these sources, I am confident that exon four begins at base pair 19872 in contig7.

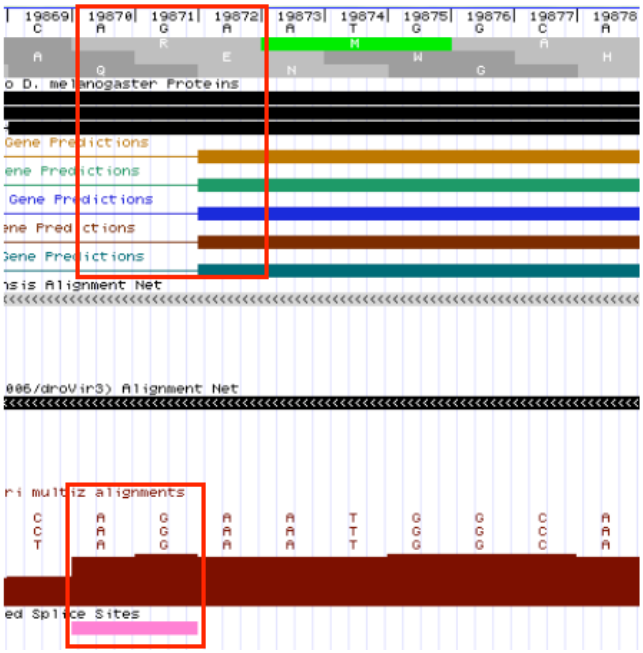

Figure 9: UCSC browser showing the start of exon four in contig7

After determining the first base pair of exon four, I moved on to annotate the 3' end of exon four. I navigated to base pair position 20003 in the UCSC browser, and I noticed that there is a high quality donor splice site immediately after base pair 20003 (Figure 10). In a similar fashion to the start of this exon, there is high conservation among the other species of *Drosophila*. Also, all of the gene predictor models show the end of an exon at this location. From all of this information, I am confident that the end of exon four is at base pair 20003. Therefore, exon four spans from base pairs 19872 to 20003 in contig7 (donor phase of 0).

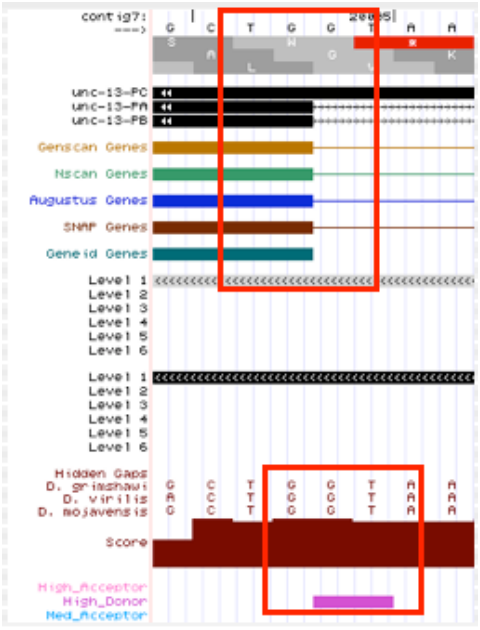

Figure 10: UCSC browser showing the end of exon four in contig7

As explained earlier, I used the same basic procedure to annotate the remaining exons from the *unc-13-RA* isoform. Overall, I was able to find and annotate exons two through eight in contig7 (Table 2). This means that not all of *unc-13-RA* is present in contig7 since exons nine through twenty-three are missing.

Exon	Start	End	Exon Size	% Identity
2	3294	3299	2	100
3	3371	8437	**1689**	?
4	19872	20003	44	90
5	29665	29739	25	96
6	29805	30024	73	98
7	30081	30201	40	97
8	32310	32416	36	100

Table 2: Summary of the exon boundaries for exons two through eight of *unc-13-RA* in contig7

In order to ensure that the exon boundaries were appropriately defined, I used the Gene Model Checker program. The results of the Gene Model Checker analysis show that all of the exons are valid and have been annotated correctly. Thus, I was able to show that exons two through eight of the *D. grimshawi* ortholog to the *D. melanogaster* gene *unc-13-RA* are present in contig7.

Although many of the exons are shared between the three isoforms of *unc-13,* there are some exons that are specific to certain isoforms. Therefore, after annotating all of the exons specific to *unc-13-RA*, I moved on to annotate the exons present in isoforms *RB* and *RC*. From the Gene Record Viewer website, I discovered that exons 28, 32 and 50 should also be present in contig7. Exon 28 has only three amino acids (VLK), so I needed to approach the annotation of this small exon in a different fashion. Exon 28 is located between exons 4 and 5, so I extracted the DNA sequence of contig7 from between exons 4 and 5. Then I used the EMBOSS explorer to translate the DNA sequence into all three plus frames. I surveyed the translated sequences and searched for amino acids VLK. I only found one location for this amino acid sequence, so I determined the base pair position of this sequence (20751) and navigated to this location on the UCSC browser (Figure 11). As Figure 11 displays, there is high conservation among the *Drosophila* species at this location. Additionally, there are phase 0 acceptor and donor sites surrounding these three amino acids in frame +3, which is appropriate for the phases of the flanking exons. Although none of the gene predictor models show an exon at this location, I am confident that exon 28 is located here due to the presence of the donor and acceptor sites along with the conservation among all of the *Drosophila* species. In short, exon 28 is located between base pair positions 20751 and 20759 in contig7.

Figure 11: UCSC browser showing amino acids VLK present in contig7 (exon 28)

After annotating exon 28, I moved on to determine the exon boundaries of exon 32. Fortunately, I was able to employ the normal procedure (used to annotate exon four) to find this exon in contig7. Exon 32 is present from base pairs 30759 to 30865 in contig7.

The final exon that I annotated in contig7 was exon 50. This exon is the first translated exon for *unc-13-RB* and is 1944 amino acids long in *D. melanogaster*. I needed to approach the annotation of this exon in a novel manner as well due to its size. I started by performing a TBLASTN align two sequences analysis to align exon 50 (query) to the masked contig7 sequence (subject). The BLAST output revealed four major hits that align amino acids 126 through 1944 to base pairs 13905 through 19520 in contig7 in frame +3 (Figure 12).

```
 Score =  163 bits (413),  Expect = 1e-42, Method: Compositional matrix adjust.
 Identities = 76/89 (85%), Positives = 83/89 (93%), Gaps = 0/89 (0%)
 Frame = +3

Query   126    SMISIKSEQQLCQSYNSEQHSDYIISDYMDKIATRISLLETELKFAWRALDLLSTEYGKI  185
               S+ SI+SEQQLCQ YN++QHSDYIISDYMDKIATRISLLETELKFAWRALDLLSTEYGKI
Sbjct   13905  SLASIRSEQQLCQLYNAQQHSDYIISDYMDKIATRISLLETELKFAWRALDLLSTEYGKI  14084

Query   186    WIRLEKLENISIEQQSVVGNLVDLIGASK  214
               W RLEKLENIS+EQQSVVGNL+ LI  +K
Sbjct   14085  WTRLEKLENISVEQQSVVGNLMGLIVVAK  14171
```

Figure 12: Two out of the four BLAST hits showing that amino acids 126 through 1944 align to contig7

I navigated to base pair 19520 on the UCSC browser to define the end of exon 50. At base pair 19520, there is strong conservation among the species of *Drosophila* and a phase 0 donor site (which matches to the phase of the next exon in isoform *RB*). Therefore, I believe that exon 50 ends at base pair 19520. Defining the start of exon 50 required more work owing to the fact that the TBLASTN alignment did not align the first 125 amino acids. I surveyed all of the amino acids 5' of base pair 13905 and noted all of the methionines present until the first stop codon in frame +3. I found five methionines that could possibly represent the start of exon 50. Surveying each of these methionines in greater detail showed that only the methionine present at base pair 13650 has conservation across the other species of *Drosophila*. Furthermore, one of the gene predictor models shows the start of an exon at base pair 13650. All of the evidence seems to point to the fact that exon 50 begins at base pair 13650. Thus, I am confident that exon 50 is located between base pairs 13650 and 19520 in contig7.

Overall, I annotated exons 2 through 8, 28, 32, and 50 in the contig7 sequence (Table 3). Exons 2-8 are found in isoform *RA* while exons 50, 4-7, and 32 are in isoform *RB*. Isoform *RC* contains exons 2-4, 28, 5-7, and 32. Using the exon boundaries I defined, I also ran the Gene Model Checker program on isoforms *RB* and *RC*. The results from Gene Model Checker for both of these isoforms indicate that all exons have been appropriately defined.

Drosophila grimshawi			
***Unc-13* Isoforms RA, RB, RC**			
Exon	**Start**	**End**	**Exon Size**
2	3294	3299	2
3	3371	8437	**1689**
4	19872	20003	44
5	29665	29739	25
6	29805	30024	73
7	30081	30201	40
8	32310	32416	36
28 (4 and 5)	20751	20759	3
32 (7 and 9)	30759	30865	36
50 (4)	13650	19520	**1957**
Exons 2-8 are in isoform RA.			
Exons 50, 4-7, 32 are in isoform RB.			
Exons 2-4, 28, 5-7, 32 are in isoform RC.			

Table 3: Summary of all exons of *unc-13* present in contig7

Search for Additional Features

After completing the annotation of the *unc-13* ortholog in contig7, I wanted to see if there were any other features present at the start of the contig since *unc-13* does not begin until base pair 3294. Therefore, I ran a BLASTX search using the first 3293 base pairs of contig7 against the protein database for *D. melanogaster*. BLAST did not reveal any hits (Figure 13).

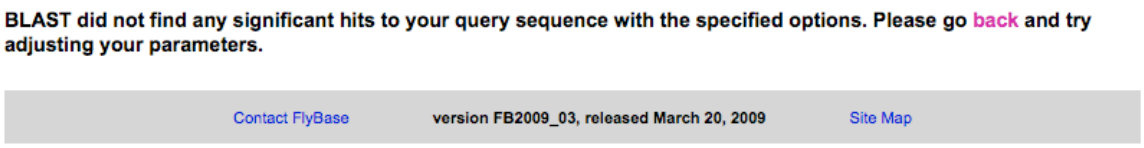

Figure 13: BLASTX output screen indicating that no significant hits were found

I also ran a BLASTX search with the same region of contig7 against all of the *Drosophila* species on Flybase. Once again, this did not turn up any significant hits. Based upon the results of these two BLASTX searches, I concluded that there are no additional features aside from *unc-13* present in contig7. Furthermore, this means that some of the initial GenScan predictions (such as the small exons around 22 kb) were miscalled.

Clustal Analysis

ClustalW is an important tool used in annotation to examine the alignment between sequences. This process can help uncover conserved domains in proteins across a wide array of species.

First, I used ClustalW to align the protein coding sequence for exon three of *unc-13* from *D. melanogaster* with *D. grimshawi*. I chose this exon since I needed to investigate it in more detail to define the exon boundaries owing to its low level of conservation between *D. melanogaster* and *D. grimshawi*. The clustal output shows that there is low conservation of exon three between these two species. However, there are conserved domains at both the start and end of the exon (Figure 14).

Figure 14: Clustsalw output showing conservation of exon three between *D. melanogaster* and *D. grimshawi* at the start and end of the exon

Next, I performed a similar ClustalW alignment between the protein sequences for exon 50 in both *D. melanogaster* and *D. grimshawi*. Once again, my reason for performing this Clustal alignment was to deduce the level of conservation of this exon between these two species. The results were very similar to that of the Clustal alignment for exon three, as there was high conservation at the start and end of the exon, but there was a low level of conservation overall.

Then, I moved on to analyze the conservation of the *unc-13* protein and its orthologs between the following four species of *Drosophila*: *melanogaster*, *virilis*, *mojavensis*, and *grimshawi*. I went to FlyBase and searched for the predicted gene models for *D. mojavensis* and *D. virilis* in order to find the protein sequence for their *unc-13* orthologs. Then, I extracted the protein sequence from each of these species and independently performed a BLAST2 analysis between the *unc-13* protein sequence of *D. mojavensis* or *D. virilis* against exon eight from *D. grimshawi* (according to my annotation). Then, I determined the location of exon eight in both of these species and extracted the protein sequence from the beginning until this position. Furthermore, I extracted the protein sequence for exons two through eight for *D. melanogaster* and *D. grimshawi.* Finally, I aligned the extracted protein sequences from all four species of *Drosophila* using ClustalW. The alignment shows very strong conservation at the end of the alignment with almost all identical amino acids across the four species (Figure 15). At the start of the alignment there is some conservation but mainly between *D. virilis, mojavensis,* and *grimshawi* (see boxes at start of alignment). It is apparent from this clustal analysis that *D. melanogaster* is the most distant species evolutionarily among the four.

Figure 15: Clustal output showing the start and end of the alignment between protein sequences for *unc-13* and its orthologs in four different species of *Drosophila*

Due to the fact that there was fairly good conservation among the four species of *Drosophila* for their *unc-13* orthologs, I investigated the conservation of *unc-13* between human, mouse, *C. elegans*, and *D. grimshawi* using ClustalW. I found the protein sequence for the *unc-13* orthologs in all of these species from Ensembl and then aligned them using ClustalW. However, the results of this clustal alignment were not very informative as most of the alignment was determined by the mouse and human sequence due to their high degree of similarity. Thus, this clustal alignment did not help me understand the degree of conservation of *unc-13* across all of these species (output not shown).

The final clustal analysis that I performed was to try to find the location of a 5' untranslated region (UTR). According to Gene Record Viewer, there are three 5' UTRs, exons 1, 48, and 49. I analyzed exon 1 first by extracting the DNA sequence for four species of *Drosophila* (*melanogaster*, *mojavensis*, *virilis, grimshawi*) that included the first large exon along with two thousand base pairs upstream of the start of this exon. I included the first large exon in this analysis in order to anchor the alignment. Once I ran ClustalW, I searched for the exon 1 DNA sequence from *D. melanogaster*. I have highlighted this region in Figure 16 below. Overall, there does not appear to be a greater degree of conservation within this region as compared to any other random region of this alignment. In this manner, I was unable to uncover the location of exon 1 within the *D. grimshawi* sequence. However, in the future, it would be worthwhile to repeat this alignment by using the small second exon (MT) to anchor the sequence instead of the first large exon.

Figure 16: Clustal output showing the exon 1 sequence of *D. melanogaster*. The other species of *Drosophila* do not show conservation in this region.

Repeat Analysis

Another aspect of annotation is to determine the quantity and type of repetitive elements present in the DNA sequence. First, I examined the output table from the RepeatMasker analysis (Figure 17). According to Table 4, contig7 is comprised of 17.06% repetitious elements. The bulk of these repetitious elements are LTRs.

```
file name: contig7.fasta
sequences:             1
total length:      33303 bp  (33303 bp excl N/X-runs)
GC level:         36.74 %
bases masked:       5681 bp ( 17.06 %)
===============================================
               number of      length   percentage
               elements*    occupied  of sequence
-----------------------------------------------

SINEs:                0           0 bp    0.00 %
      ALUs            0           0 bp    0.00 %
      MIRs            0           0 bp    0.00 %

LINEs:                2         381 bp    1.14 %
      LINE1           0           0 bp    0.00 %
      LINE2           0           0 bp    0.00 %
      L3/CR1          0           0 bp    0.00 %

LTR elements:         2        3288 bp    9.87 %
      MaLRs           0           0 bp    0.00 %
      ERVL            0           0 bp    0.00 %
      ERV_classI      0           0 bp    0.00 %
      ERV_classII     0           0 bp    0.00 %

DNA elements:         0           0 bp    0.00 %
      MER1_type       0           0 bp    0.00 %
      MER2_type       0           0 bp    0.00 %

Unclassified:         3         486 bp    1.46 %

Total interspersed repeats:     4155 bp   12.48 %


Small RNA:            0           0 bp    0.00 %

Satellites:           6         670 bp    2.01 %
Simple repeats:       5         678 bp    2.04 %
Low complexity:       1         178 bp    0.53 %
```

Figure 17: Output from RepeatMasker

Although RepeatMasker performs a thorough job in determining the identified repetitive elements within a given DNA sequence, it sometimes misses novel repetitive elements. Therefore, I decided to analyze contig7 in further detail to see if RepeatMasker missed any important repetitious elements. I used the masked contig7 DNA sequence and ran a nucleotide BLAST with this sequence against the nucleotide/nonredundant database. I used the BLAST viewer to analyze the significant hits to see if any were adjacent to previously defined repeats. However, the only significant hit is shown in Figure 18, and it simply represents a random hit to human cDNA sequence. Thus, the additional repeat analysis did not turn up any new results. However, in the future, it would be beneficial to compare contig7 to the entire *D. grimshawi* genome to look for duplications. These duplications could indicate the presence of novel repeats in contig7.

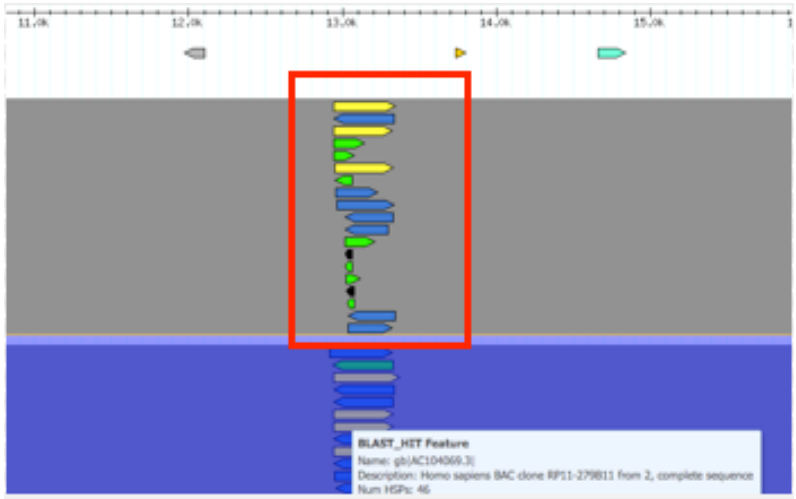

Figure 18: The only significant BLAST hit for the masked contig7 sequence as shown by the BLAST viewer

Due to the fact that there are no additional repetitive regions in contig7 aside from those determined by RepeatMasker, I investigated the RepeatMasker results in greater detail. I discovered that only four of the repetitive elements are greater than 300 base pairs in length (Table 4). I will annotate these four repetitive elements and include them in my final contig map.

Repeat Class	Start	End	Repeat Size
LTR	24768	27433	2665
LTR/Gypsy	23220	23690	470
Simple_repeat	30343	30701	358
LINE/LOA	32977	33291	314

Table 4: List of the four repetitive elements present in contig7 that I will annotate

Synteny

Last, I investigated the synteny of contig7 as compared to *D. melanogaster*. However, since contig7 only contains a portion of one gene, it is not possible to conduct a synteny analysis solely using contig7. Instead, I used the information on FlyBase to examine the genes around the *unc-13* ortholog in *D. grimshawi*. Additionally, I compared this region of *D. grimshawi* to the orthlogous region of the fourth chromosome in *D. melanogaster* to determine whether or not these regions are syntenic. As Figure 19 shows, the *unc-13* ortholog in *D. grimshawi* is not flanked by the same genes that flank *unc-13* in *D. melanogaster*. Looking 30 kb upstream or downstream, only the Sox102 gene is present in both species of *Drosophila*. Thus, synteny is not maintained within this region of the *D. grimshawi* genome.

Figure 19: Region of *D. grimshawi* that contains the *unc-13* ortholog as compared to the orthologous region in *D. melanogaster*

Conclusion

Contig7 contains one gene that represents the *D. grimshawi* ortholog to the *D. melanogaster* gene *unc-13*. In *D. melanogaster*, *unc-13* has three different isoforms that contain a total of 25 translated exons distributed among all three isoforms. Overall, I showed that 10 out of the 25 exons are present in contig7. Also, all three isoforms, *unc-13 RA, RB,* and *RC* are present in this contig (Figure 20). Furthermore, contig7 is comprised of 17.06% repetitious elements. However, only four of the repetitive elements span over 300 base pairs of sequence (Figure 20). The region of the *D. grimshawi* genome where the *unc-13* ortholog is located is not syntenic with the orthologous region of the *D. melanogaster* genome. In conclusion, all of the relevant features and repetitive elements have been appropriately documented in contig7.

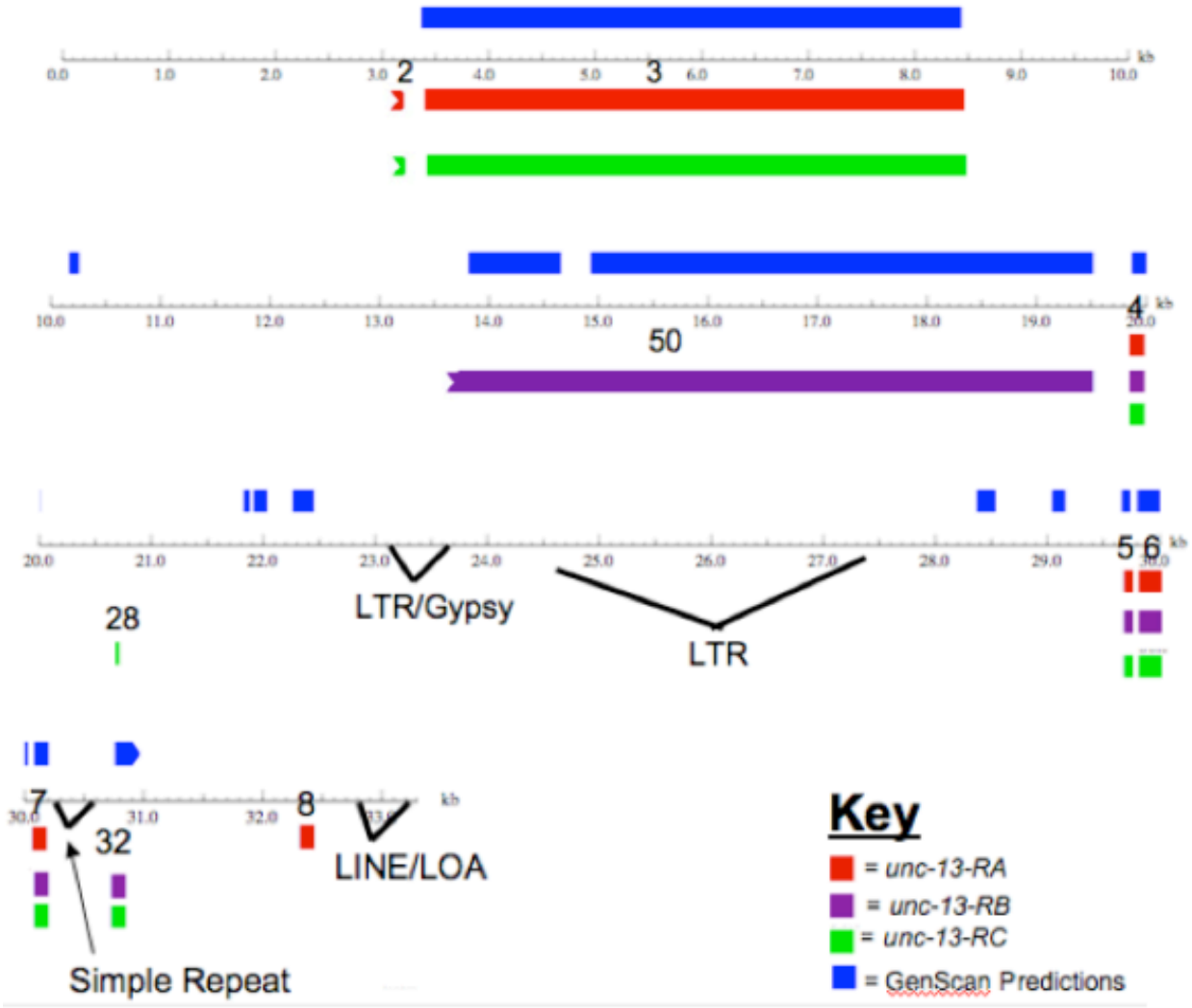

Figure 20: Final map showing the relevant features and repetitive elements present in *D. grimshawi* contig7

Annotation of contig3

I also annotated one of the features located on contig3. This feature is highlighted in Figure 21, and it is oriented along the minus strand of contig3. GenScan predicts that this feature contains a total of three exons.

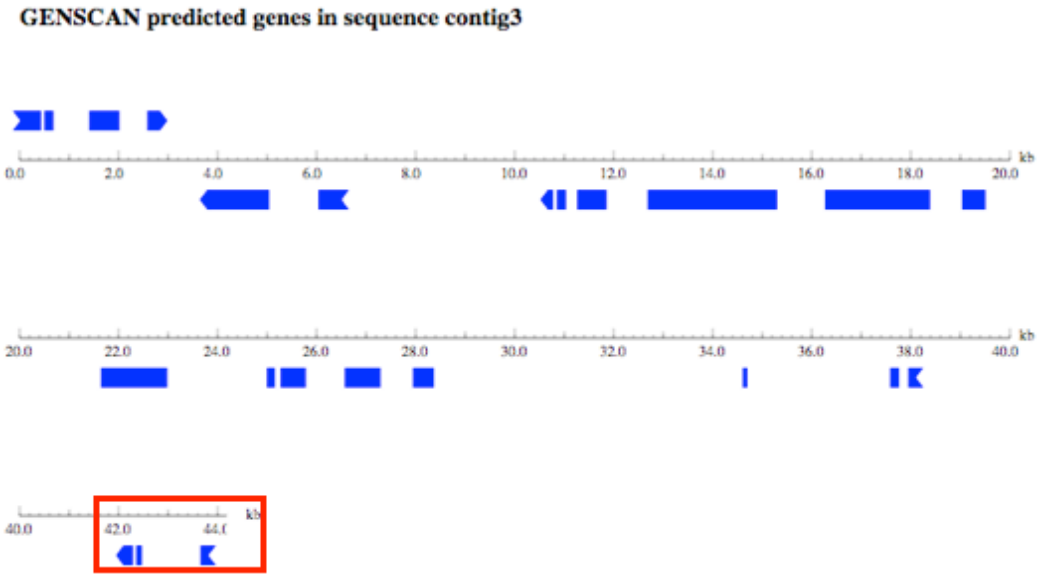

Figure 21: Initial GenScan predictions for contig3

First, I wanted to determine the *D. melanogaster* ortholog that this feature represented. Therefore, I began the annotation process by performing a BLASTX search with base pairs 40000 to 44000 of contig3 (query) against the nonredundant protein database (subject). The best experimentally supported BLAST hit is to the *D. melanogaster* protein *CG31997* (Figure 22).

ref|XP_001996521.1|	GH23991 [Drosophila grimshawi] >gb|EDV909...	217	9e-54
ref|XP_002059720.1|	GJ14631 [Drosophila virilis] >gb|EDW71118...	199	2e-48
ref|XP_002011462.1|	GI14118 [Drosophila mojavensis] >gb|EDW10...	191	7e-46
ref|XP_001844380.1|	conserved hypothetical protein [Culex qui...	179	2e-42
ref|NP_726539.1|	CG31997 [Drosophila melanogaster] >gb|AAL490...	178	4e-42
ref|XP_002044493.1|	GM23232 [Drosophila sechellia] >ref|XP_00...	177	8e-42
ref|XP_001982750.1|	GG16462 [Drosophila erecta] >gb|EDV45269....	176	1e-41
ref|XP_002099577.1|	GE14528 [Drosophila yakuba] >gb|EDW99289....	174	8e-41
ref|XP_001656438.1|	hypothetical protein AaeL_AAEL000445 [Aed...	171	5e-40
ref|XP_001352360.2|	GA16601 [Drosophila pseudoobscura pseudoo...	171	7e-40

Figure 22: BLAST output showing that the best experimentally supported BLAST hit is to the *D. melanogaster* protein *CG31997*

According to the Gene Record Viewer website, this gene consists of three translated exons and has only one isoform in *D. melanogaster*. Also, it is located on chromosome 4 in *D. melanogaster*. The annotation of this feature proved to be quite simplistic since all of the exons from *D. melanogaster* aligned well to the contig3 sequence. Thus, I was able to employ the generic method (as explained for exon four from *unc-13*) using TBLASTN analyses and perusal of the UCSC browser for intron donor and acceptor sites to determine the boundaries of each exon. Overall, I defined the boundaries for all three exons of *CG31997* in contig3 (Table 5).

Drosophila grimshawi			
CG31997			
Exon	**Start**	**End**	**Exon Size**
1	43655	43814	53
2	42354	42474	40
3	42098	42290	64
Located on the minus strand			

Table 5: Summary of all exon boundaries of the *D. grimshawi* ortholog to the *D. melanogaster* gene *CG31997*

After defining the exon boundaries, I checked to make sure that all of the boundaries were appropriately defined by using the Gene Model Checker analysis. This analysis showed that all exon boundaries are correct. Overall, I annotated the *D. grimshawi* ortholog to the *D. melanogaster* gene *CG31997* in contig3.

Appendix

1.) *unc-13-RA, RB,* and *RC* protein sequence fasta files
2.) *unc-13-RA, RB,* and *RC* nucleic acid sequence fasta files
3.) *unc-13-RA, RB,* and *RC* GFF files
4.) *CG31997* protein sequence fasta file
5.) *CG31997* nucleic acid sequence fasta file
6.) *CG31997* GFF file
7.) Sequence files used for Clustal analyses