Sarah Spencer
Bio 4342W

29 April 2008

Annotation of 205-H09

Overview

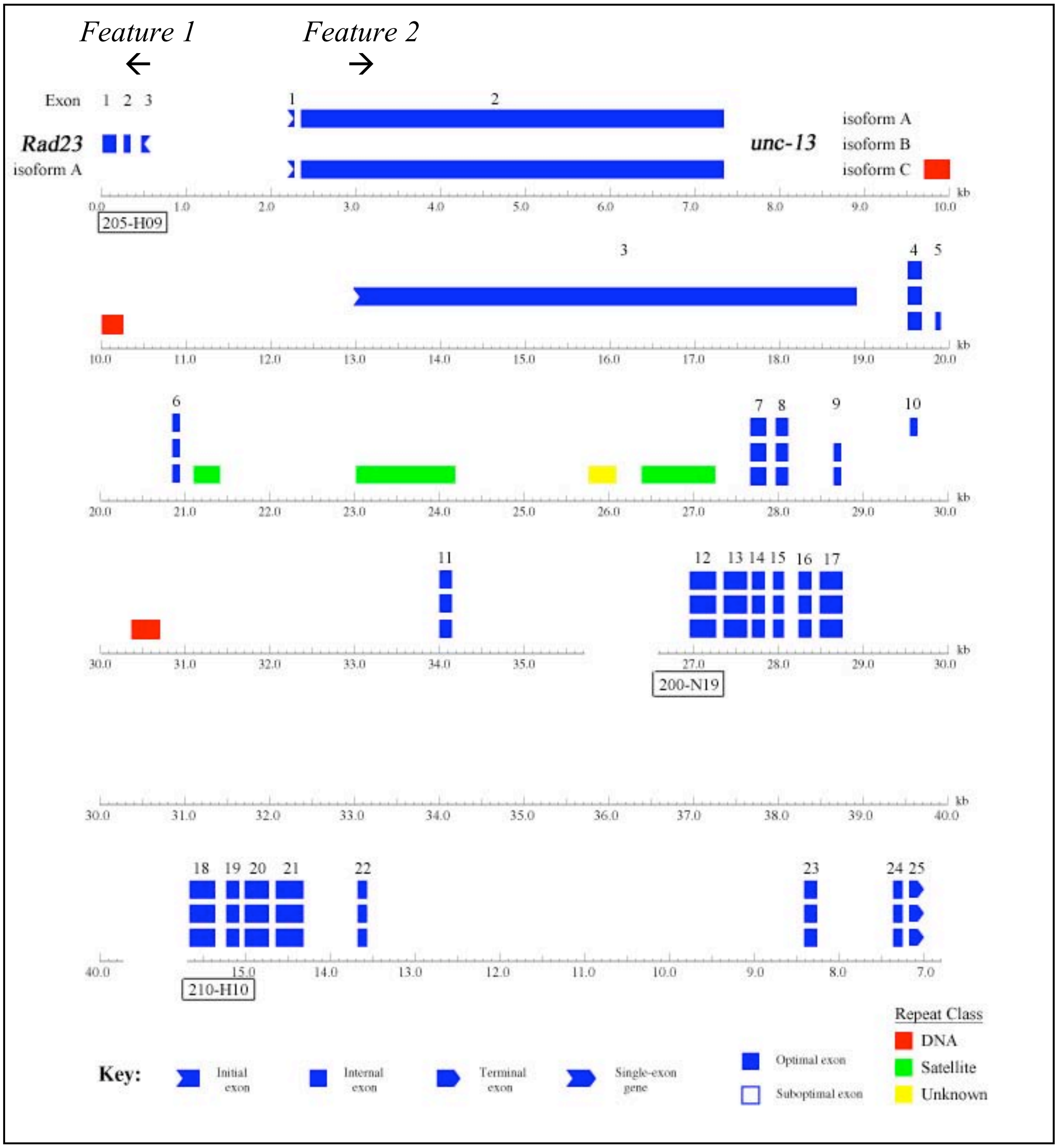


Figure 1. Final annotation map for 205-H09. The fosmid contains two genes, both of which begin in the fosmid and extend out into adjacent sequence. The annotation of unc-13 was continued to completion through two neighboring clones. RepeatMasker was only run on 205-H09, and repeats larger than 300 bp are mapped.

My project involved the annotation of fosmid 205-H09, a 35,708 bp DNA sequence from *Drosophila mojavensis* with 35.45% GC content (Figure 1). This sequence derives from the *D. mojavensis* fourth ("dot") chromosome and is part of a larger effort to finish and annotate the entire chromosome for comparative purposes. I identified and characterized two features in my fosmid. Feature 1 is the gene *Rad23*, a fairly well conserved, two isoform gene involved in damaged DNA binding and repair. Feature 2 was initially predicted by GENSCAN and I confirmed it as *unc-13*, a highly conserved gene involved in neurotransmitter secretion and synaptic signaling. I completed the gene model for *unc-13* by continuing annotation into two neighboring fosmids, 200-N19 and 210-H10. A ClustalW alignment did not identify any non-coding functional sequences upstream of *Rad23*, but *unc-13* protein alignments did identify regions of protein conservation. The RepeatMasker program reported very few repeats in my fosmid despite the long intronic regions of *unc-13*. Finally, a synteny analysis was undertaken that showed all *D.mojavensis* genes on the *D. melanogaster* dot chromosome, but frequently transposed or inverted in relation to each other.

Genes

I initially had evidence for two features. Feature 1 was identified in the UCSC genome browser for my fosmid based on a curated Refseq gene as well as from strong multiple species alignment conservation. Feature 2 was identified in the initial GENSCAN gene predictions for my fosmid (Figure 2). Although Feature 2 was more prominent in my fosmid, defining Feature 1 proved necessary to complete a gene model for Feature 2. Thus, I will begin my description with Feature 1.

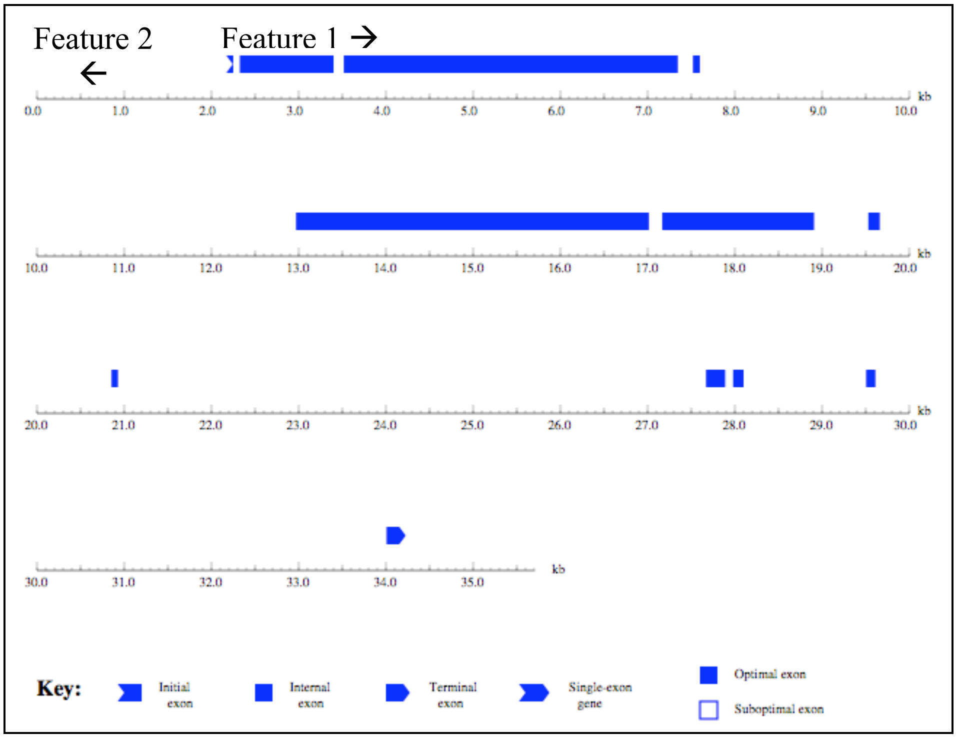

Feature 2 ←
Feature 1 →
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 kb
10.0 11.0 12.0 13.0 14.0 15.0 16.0 17.0 18.0 19.0 20.0 kb
20.0 21.0 22.0 23.0 24.0 25.0 26.0 27.0 28.0 29.0 30.0 kb
30.0 31.0 32.0 33.0 34.0 35.0 kb
Key:
Initial exon
Internal exon
Terminal exon
Single-exon gene
Optimal exon
Suboptimal exon

Figure 2. GENSCAN gene prediction for 205-H09. GENSCAN only identified Feature 2 and did not recognize the partial transcript of Feature 1.

Feature 1 – Rad23

I first identified Feature 1 based on Refseq and multiple species alignment evidence from the UCSC genome browser of 205-H09. The Refseq gene listed is Rad23, and in Gene Record Finder I found that Rad23 in *D. melanogaster* is an eight exon gene with two isoforms. Each isoform consisted of five exons. Suspecting a *Rad23* ortholog in my fosmid, I used the bl2seq BLASTx program to compare individual *D. melanogaster* Rad23 coding exons with the translated DNA sequence of 205-H09. There were no significant alignments to any isoform B exons, but two exons from isoform A aligned to the minus strand at the beginning of my fosmid (the first 520 bases). It thus appeared that my fosmid contained only a part of *Rad23*, two exons from one isoform, and the rest of the gene continued into an adjoining clone.

To investigate my hypothesis, I used the UCSC BLAT program to find my fosmid sequence within the *D. mojavensis* scaffold (Figure 3). My fosmid matched to the scaffold sequence from 784,413-804,189 bps. The BLAT browser showed the *Rad23* Refseq gene partially overlapping my fosmid sequence and partially overlapping the adjacent scaffold sequence. I extracted the scaffold DNA sequence from 803,000-806,000 bps for BLAST analysis and exon determination. This range encompassed the entire putative *Rad23* gene as well as some adjacent DNA sequence; it thus included part of my fosmid DNA and also some unique DNA not included in my fosmid.

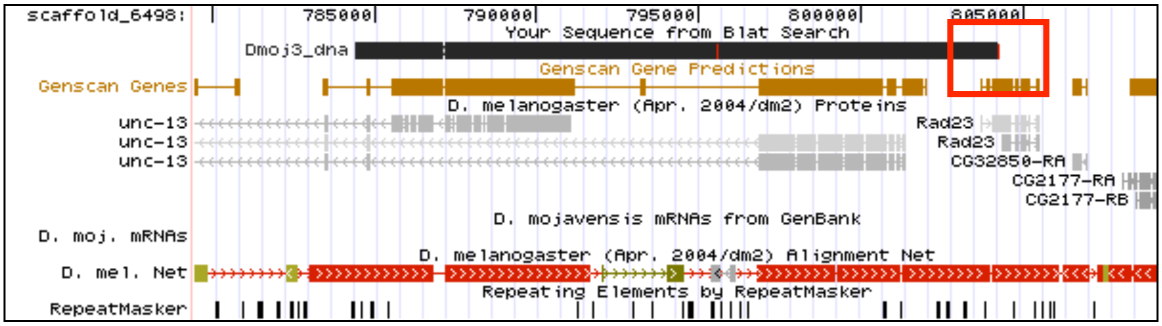
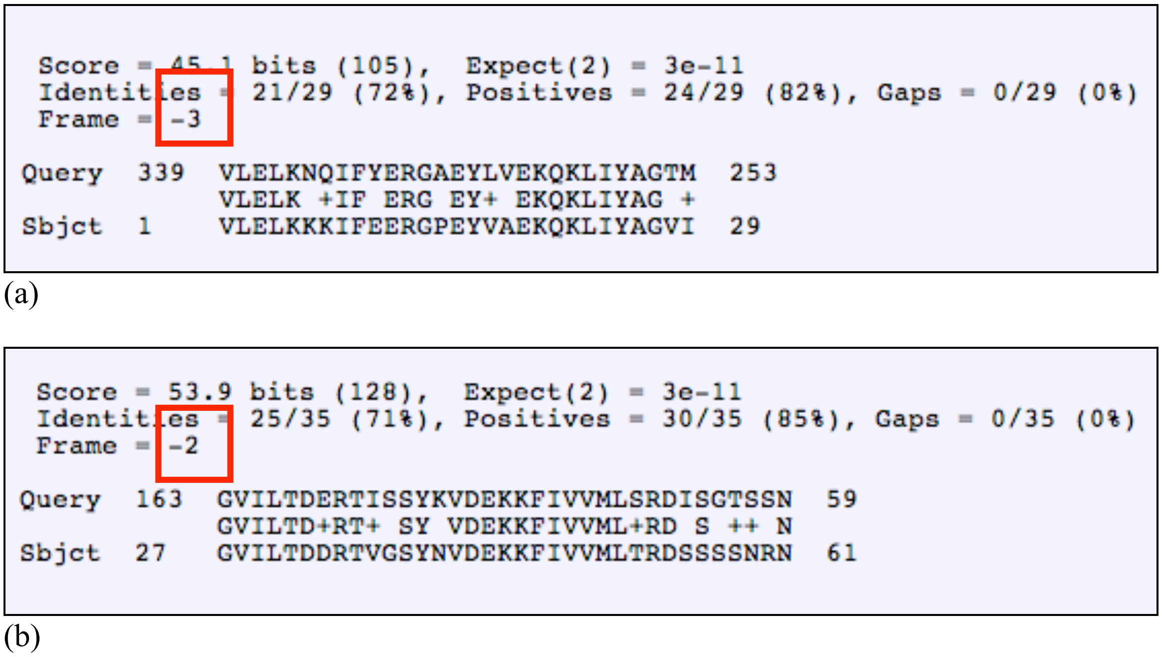


Figure 3. UCSC browser displaying the BLAT results after the *D. mojavensis* assembly was queried with 205-H09. Fosmid 205-H09 is in negative orientation compared to the scaffold sequence. The highlighted box shows how the *Rad23* Refseq gene continues off the end of my fosmid sequence.

Using this extracted sequence, I conducted a bl2seq BLASTx analysis wherein the extracted sequence served as a translated nucleotide query against individual D. melanogaster *Rad23* translated exons from Gene Record Finder. My exon-by-exon search yielded alignments for all exons of both isoforms. I was able to determine which exons mapped to my fosmid because my fosmid DNA overlapped the extracted region from 803,000 to 804,189. Thus, bl2seq matches that involved bases 1-1,189 of the extracted region indicated exons that map to my fosmid.

In my BLAST analysis of the two exons that fall within my fosmid (exons 1 and 2 of isoform A), I found that the alignments to exon 2 fell in two different reading frames (Figure 4). The most likely explanation for this observation was that the *D. mojavensis* *Rad23* gene includes an extra intron compared to the *D. melanogaster* ortholog. To investigate this, I studied the alignment locations and their reading frames in the genome browser of 205-H09 (Figure 5). It was then apparent that exon 2, which begins in the third reading frame of the negative strand, is interrupted by a stop codon and switches to a new exon in the second reading frame after a ~100 bp intron. To describe the division of exon 2, I designated the upstream segment exon 2a and the downstream segment – also downstream of the newly identified intron – exon 2b. According to BLAST alignments in the extracted region, exon 2b continues off my fosmid after the novel intron.



```
 Score = 45.1 bits (105),  Expect(2) = 3e-11
 Identities = 21/29 (72%), Positives = 24/29 (82%), Gaps = 0/29 (0%)
 Frame = -3

Query  339  VLELKNQIFYERGAEYLVEKQKLIYAGTM  253
            VLELK +IF ERG EY+ EKQKLIYAG +
Sbjct  1    VLELKKKIFEERGPEYVAEKQKLIYAGVI  29
```

(a)

```
 Score = 53.9 bits (128),  Expect(2) = 3e-11
 Identities = 25/35 (71%), Positives = 30/35 (85%), Gaps = 0/35 (0%)
 Frame = -2

Query  163  GVILTDERTISSYKVDEKKFIVVMLSRDISGTSSN  59
            GVILTD+RT+ SY VDEKKFIVVML+RD S ++ N
Sbjct  27   GVILTDDRTVGSYNVDEKKFIVVMLTRDSSSSNRN  61
```

(b)

Figure 4. Results of bl2seq BLASTx analysis with the extracted end of 205-H09 used as a translated nucleotide query against the amino acid sequence of the second *Rad23* exon in *D. melanogaster*. Note the change in reading frame that splits the exon 2 amino acid sequence and implies a novel intron. The alignment shows (a) Exon 2a and (b) Exon 2b which continues upstream of my fosmid.

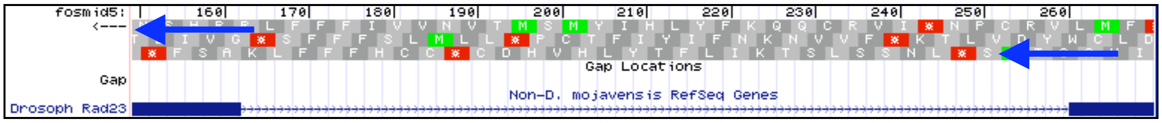


Figure 5. UCSC genome browser display of a *D. mojavensis Rad23* intron that is not present in *D. melanogaster*. The predicted *Rad23* gene is on the minus strand of the 205-H09 sequence and the reading frame between *Rad23* exons 2a and 2b changes from -3 to -2, as indicated by arrows.

I now had a rough concept of the *Rad23* exon locations in my fosmid. To construct a specific gene model, I used the 205-H09 genome browser to analyze the bl2seq alignment boundaries at the level of base pairs and codons. I will use exon 1 of *Rad23* as an example of the process I used for determining exon boundaries.

The bl2seq alignment for exon 1 of *D. melanogaster* showed a match of the entire 22 aa exon sequence to bases 520-455 in fosmid 205-H09. The frame of the match in relation to the fosmid translated nucleotide query sequence was -2. With this information, I proceeded to the 205-H09 genome browser and focused on the 100 bp between bases 440 and 540 (Figure 6). I began by displaying the negative strand so I could see the correct gene sequence. Exon 1 is the first exon in isoform A of *Rad23*, so I knew the upstream end of exon 1 should start with a methionine codon. As seen in Figure 6, the negative frame 2 codes for a methionine that begins at base 520 and is followed by an open reading frame. I thus recorded base 520 as the first base in exon 1 for my *D. mojavensis Rad23* gene model.

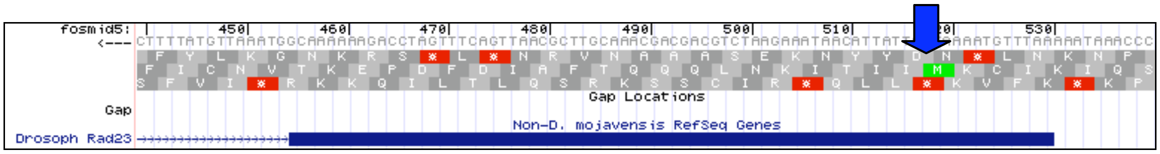

fosmid5:
450
460
470
480
490
500
510
530
<---
CTTTTATGTTAAATGGCAAAAAAAGACCTAGTTTCAGTTAACGCTTGCAAACGACGACGTCTAAGAAATAACATTAT
AAATGTTTAAAAATAAACCC
M
Gap Locations
Gap
Non-D. mojavensis RefSeq Genes
Drosoph Rad23

Figure 6. UCSC genome browser of 205-H09 displaying the region that aligned to exon 1 of the *D. melanogaster Rad23* gene. Note the methionine start site and open reading frame in the reverse oriented second reading frame.

To determine the end of exon 1, I looked for possible GT intron donors near the end of the bl2seq exon alignment. The alignment ended at base 455, and immediately downstream of base 455 in 205-H09 there is an intron donor. The splice site defined by this donor corresponds to the end of a three bp codon in frame -2, demarcating a phase 0 exon end. Another possible intron donor near the region of alignment denotes a splice site after base 448, but includes a base that is not part of a three bp codon with the exon 1 sequence. This single extra base indicates a phase 1 exon boundary. A phase 2 boundary would result if a splice site location included two extra bases after a complete codon.

To determine which intron donor is more probable for my gene model, I had to determine the location of possible AG intron acceptors leading into exon 2. Exon 2 bl2seq alignments began at the first amino acid of the *D. melanogaster* protein and at base 339 in the *D. mojavensis* fosmid sequence (frame -3). I therefore analyzed the region from 300-400 bps for intron acceptors (Figure 7). I found only one AG intron acceptor near the beginning of the alignment that led into an open reading frame. This single intron acceptor designated an exon starting at base 339, perfectly corresponding to the bl2seq alignment. Even more, the acceptor formed a phase 0 exon boundary. Thus, I knew exon 1 ended at base 455 since that position also has a phase 0 boundary.

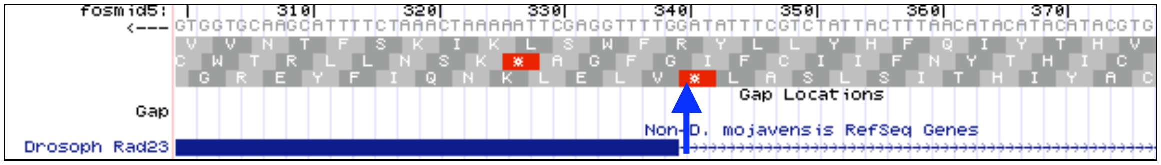
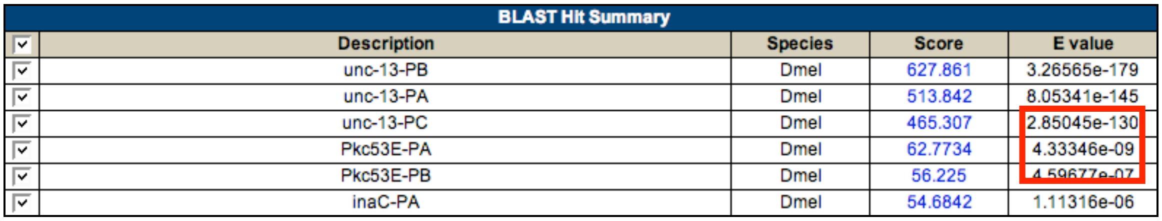

fosmid5:
310
320
330
340
350
360
370
<---
GTGGTGCAAGCATTTTCTAAACTAAAAATTCGAGGTTTTGGATATTTCGTCTATTACTTTAACATACATACATACGTG
Gap Locations
Gap
Non-D. mojavensis RefSeq Genes
Drosoph Rad23

Figure 7. UCSC genome browser of 205-H09 displaying the region where alignment with *D. melanogaster Rad23* exon 2 began. The alignment fell in frame -3 and therefore had to follow the stop codon in order to preserve an open reading frame.

For Rad23, I continued to repeat this method in order to find the end of exon 2 and beginning of exon 3. In each case, I searched for intron donors/acceptors within 30 bps of the alignment boundaries. If the alignments did not encompass the entire D. melanogaster protein sequence, I extrapolated the length and searched for donors/acceptors near the calculated location on my fosmid. I chose donors and acceptors for my gene models based on their position relative to the bl2seq alignments and their phase. For all my gene model determinations, when I found multiple bl2seq alignments from a single *D. melanogaster* exon sequence, I checked any unaligned fosmid sequence for frame changes or stop codons indicating intron insertions. My final gene model for Rad23 in fosmid 205-H09 appears at the end of the paper in Table 3.

Feature 2 – unc-13

Feature 2 dominated my annotation efforts as much as it dominated the length of fosmid 205-H09. I first investigated this feature with the GENSCAN prediction for 205-H09. As mentioned previously, GENSCAN predicted one gene with twelve exons for my fosmid (Figure 2). I used the BLASTp program to search the *D. melanogaster* "Annotated Proteins" database from FlyBase and generate alignments to the GENSCAN predicted protein sequence. The results show significant matches to the A, B, and C isoforms of the gene unc-13 (Figure 8). Following these three matches there is a dramatic rise in E values from e-130 to e-09. This indicates that unc-13 is a strong ortholog candidate. In addition, the top alignments show regions of high similarity interspersed with regions of little or no similarity, an indication of protein domain conservation. Finally, the location of unc-13 is on the dot chromosome (loc = 4) in *D. melanogaster*, an indication of synteny with *D. mojavensis*. All this evidence from the BLAST search suggests unc-13 as a probable ortholog.



BLAST Hit Summary

☑	Description	Species	Score	E value
☑	unc-13-PB	Dmel	627.861	3.26565e-179
☑	unc-13-PA	Dmel	513.842	8.05341e-145
☑	unc-13-PC	Dmel	465.307	2.85045e-130
☑	Pkc53E-PA	Dmel	62.7734	4.33346e-09
☑	Pkc53E-PB	Dmel	56.225	4.59677e-07
☑	inaC-PA	Dmel	54.6842	1.11316e-06

Figure 8. Top BLASTp alignments with the GENSCAN predicted protein sequence (Feature 2) as a query against the *D. melanogaster* "Annotated Proteins" database on FlyBase. Note the significant drop in E value between UNC-13 matches and all other protein alignments.

According to Gene Record Finder, unc-13 in *D. melanogaster* has 29 exons in three isoforms. Its location is confirmed on the dot chromosome (chromosome 4). Assuming a similar gene model in *D. mojavensis*, I began a similar exon-by-exon bl2seq BLASTx analysis using the *D. melanogaster* exons and the 205-H09 fosmid sequence. Many of the exons showed high identity with almost perfect matches to the plus strand of 205-H09 along the entire exon protein length, indicating strong conservation between species. For these exons, I used the same method described for exons 1 and 2 of Rad23 to determine the specific exon boundaries and splice sites. A few exons, however, required a greater amount of effort to characterize.

Exon 2 of isoforms A and B had a well-defined exon end but poor alignment toward the beginning of the exon. The bl2seq alignment began with amino acid 257 and fosmid base 3030 (frame +3). I thus knew that the beginning of exon 2 had to fall between the beginning of the frame 3 open reading frame (~2200 bp) and the beginning of the bl2seq alignment at base 3030 (Figure 9). Evidence from the gene predictions, predicted splice sites, and my own length extrapolation led me to chose base 2325 as the most probable first base of exon 2. This chosen splice site defined a phase 0 boundary.

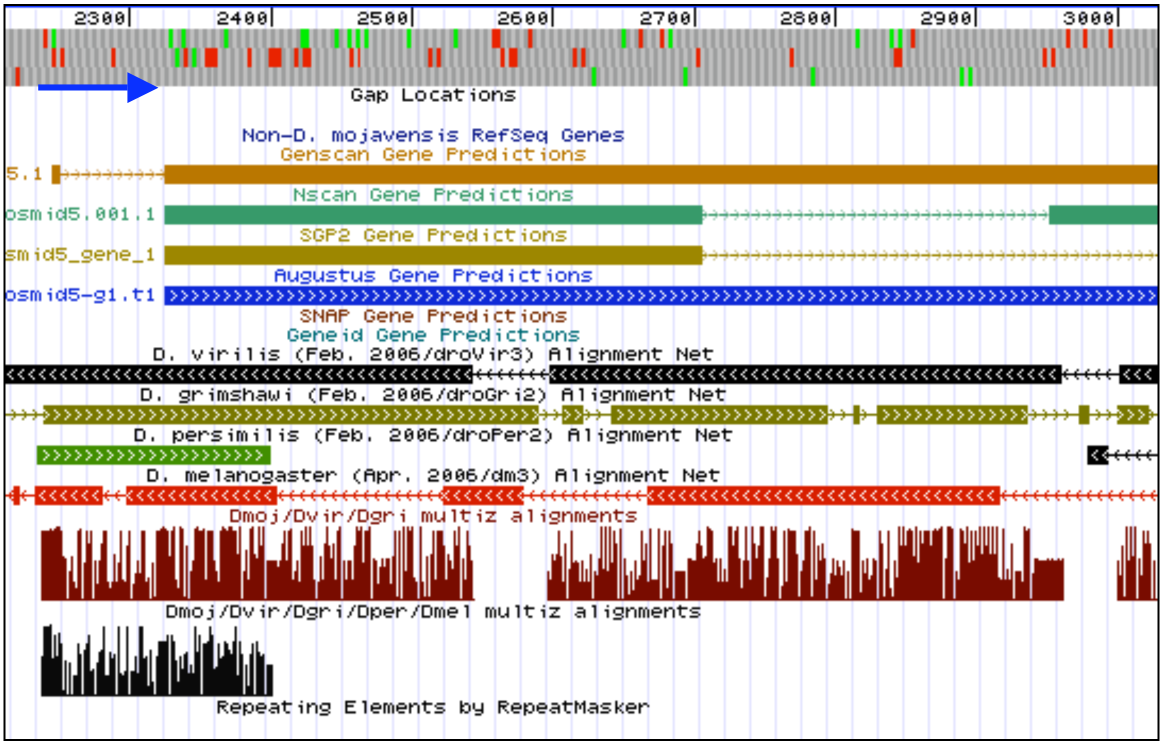
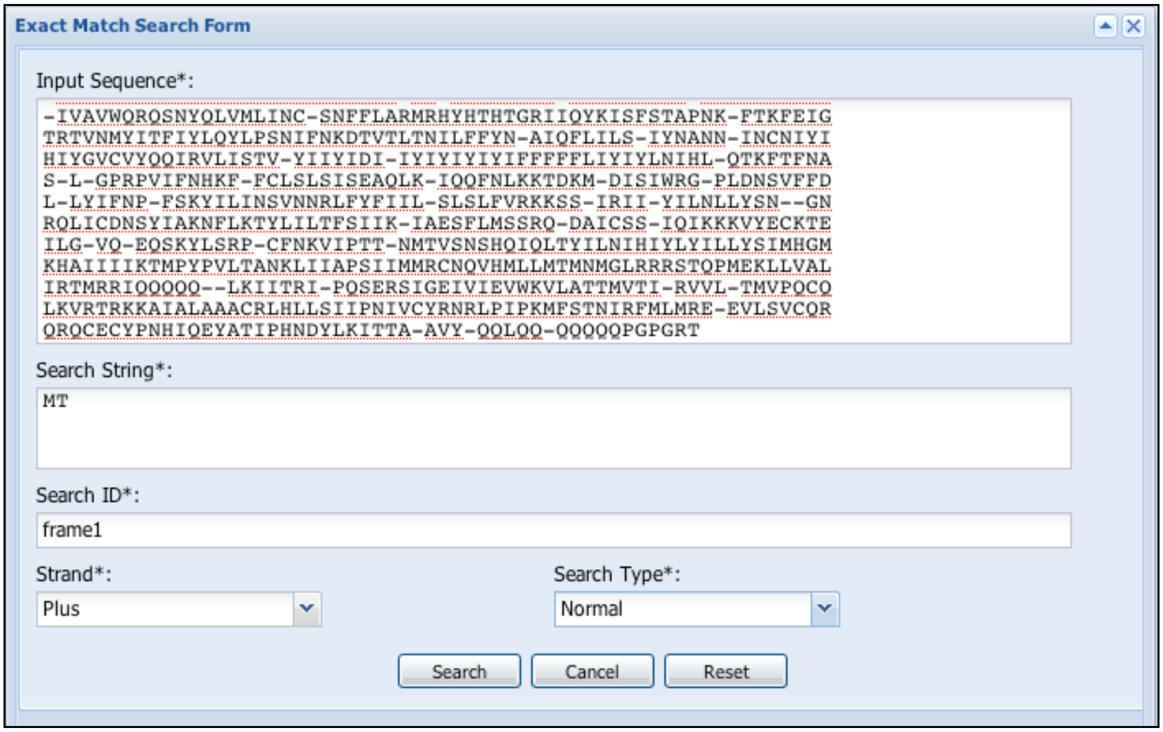

2300
2400
2500
2600
2700
2800
2900
3000
Gap Locations
Non-D. mojavensis RefSeq Genes
Genscan Gene Predictions
5.1
Nscan Gene Predictions
osmid5.001.1
SGP2 Gene Predictions
smid5_gene_1
Augustus Gene Predictions
osmid5-g1.t1
SNAP Gene Predictions
Geneid Gene Predictions
D. virilis (Feb. 2006/droVir3) Alignment Net
D. grimshawi (Feb. 2006/droGri2) Alignment Net
D. persimilis (Feb. 2006/droPer2) Alignment Net
D. melanogaster (Apr. 2006/dm3) Alignment Net
Dmoj/Dvir/Dgri multiz alignments
Dmoj/Dvir/Dgri/Dper/Dmel multiz alignments
Repeating Elements by RepeatMasker

Figure 9. UCSC genome browser of 205-H09 showing the region that must contain an exon 2 start site. The reading frame (+3) begins after a stop codon ~2200 bps as shown by an arrow, but bl2seq alignment to the exon does not begin until base 3030. Note the corresponding gene predictions as well as the multiple species conservation in the region.

The first coding exon in my fosmid, exon 1 of isoforms A and B, was even more difficult to define in 205-H09. In *D. melanogaster*, exon 1 is only two amino acids long: a methionine to start translation followed by a threonine. The bl2seq BLASTx tool could not generate meaningful matches from such a short subject sequence. To locate this exon in my fosmid, I conducted what amounted to a manual BLASTx search using the Find Exact Match program in the GEP toolkit. This involved translating the 205-H09 DNA sequence between the beginning of Rad23 (520 bp) and the beginning of the unc-13 exon 2 alignment (3030 bps). This specified region had to contain exon 1, assuming conservation between species and an absence of gene nesting.

I translated the specified region with the Expasy translate tool. I then copied the three plus strand translations and searched each one for the string MT using the Find Exact Match tool (Figure 10). Since I used the translated fosmid sequence from 1-3030 bps, I had to multiply the resulting matches by three in order to obtain locations on my fosmid DNA sequence (Figure 11). When I investigated these matches in my fosmid, I found only one match with a closely associated intron donor that formed a phase 0 splice site. I thus defined the location of this match, from 2245-2250 bps, as the most probable location for unc-13 exon 1.



Exact Match Search Form

Input Sequence*:

```
-IVAVWQRQSNYQLVMLINC-SNFFLARMRHYHTHTGRIIQYKISFSTAPNK-FTKFEIG
TRTVNMYITFIYLQYLPSNIFNKDTVTLTNILFFYN-AIQFLILS-IYNANN-INCNIYI
HIYGVCVYQQIRVLISTV-YIIYIDI-IYIYIYIYIFFFFFLIYIYLNIHL-QTKFTFNA
S-L-GPRPVIFNHKF-FCLSLSISEAQLK-IQQFNLKKTDKM-DISIWRG-PLDNSVFFD
L-LYIFNP-FSKYILINSVNNRLFYFIIL-SLSLFVRKKSS-IRII-YILNLLYSN--GN
RQLICDNSYIAKNFLKTYLILTFSIIK-IAESFLMSSRQ-DAICSS-IQIKKKVYECKTE
ILG-VQ-EQSKYLSRP-CFNKVIPTT-NMTVSNSHQIQLTYILNIHIYLYILLYSIMHGM
KHAIIIIKTMPYPVLTANKLIIAPSIIMMRCNQVHMLLMTMNMGLRRRSTQPMEKLLVAL
IRTMRRIQQQQQ--LKIITRI-PQSERSIGEIVIEVWKVLATTMVTI-RVVL-TMVPQCQ
LKVRTRKKAIALAAACRLHLLSIIPNIVCYRNRLPIPKMFSTNIRFMLMRE-EVLSVCQR
QRQCECYPNHIQEYATIPHNDYLKITTA-AVY-QQLQQ-QQQQQPGPGRT
```

Search String*:

MT

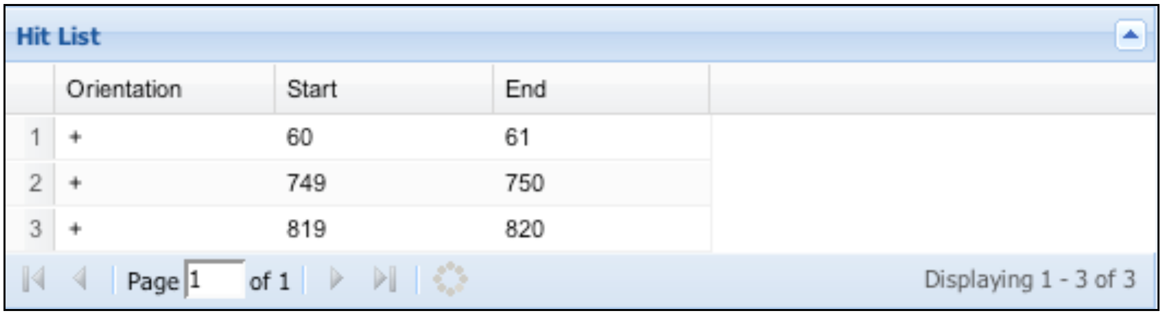
Search ID*:

frame1

Strand*: Plus

Search Type*: Normal

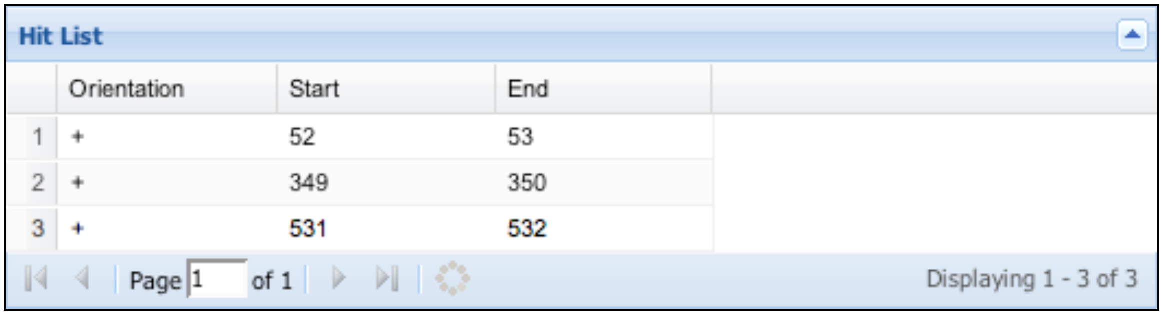
Figure 10. Find Exact Match tool used to search the frame +1 translation of 205-H09 bases 1-3030 for the string "MT".



Hit List

	Orientation	Start	End
1	+	60	61
2	+	749	750
3	+	819	820

(a)



Hit List

	Orientation	Start	End
1	+	52	53
2	+	349	350
3	+	531	532

(b)

Figure 11. Find Exact Match output from a search for "MT" in the (a) frame +1 and (b) frame +3 translations of 205-H09 bases 1-3030. Frame +2 yielded no matches.

Exon 5 in unc-13 isoform C presented a problem similar to exon 1 and thus I used a similar strategy to locate it in my fosmid. According to Gene Record Finder, three amino acids compose exon 5 in *D. melanogaster* (VLK). Also, in isoform C it falls between exons 4 and 6. When I searched the region between my previously defined exons 4 and 6 for exact matches to VLK, I received only one result. This result, when I located it in my fosmid, appeared very promising (Figure 12). It had splice donors and acceptors defining the three amino acid sequence with the necessary phase 0 boundaries. In addition, the high confidence splice sites support my chosen location.

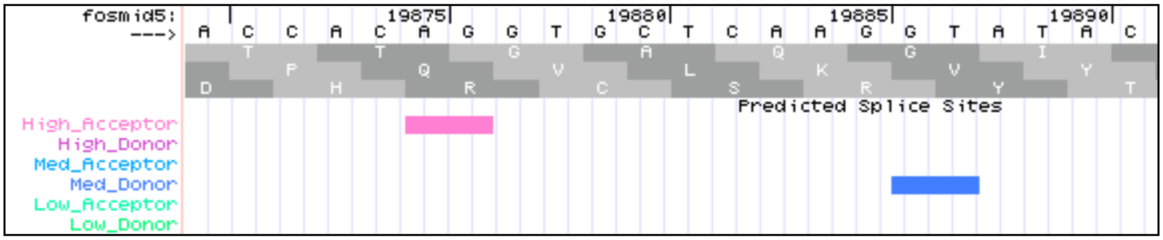


Figure 12. UCSC genome browser of 205-H09 displaying the annotated site for a three amino acid *unc-13* exon (exon 5) identified with the Find Exact Match tool. Note the high confidence splice sites on either side of the frame 2, phase 0 exon.

Exon 3 in isoform B presented a problem similar to the previously discussed exon 2 in isoforms A and B. The bl2seq BLASTx tool using the *D. melanogaster* exon did not align the first 128 amino acids of the exon. Since exon 3 is the first exon in isoform B, I had to find the initial methionine without strong conservation in the beginning of the protein. Figure 13 shows all the possible exon start sites (frame 3+) ranging from the beginning of the open reading frame to the beginning of bl2seq alignment at base 13,431. I assumed that one of the first two methiones is the true translation beginning because of my own length extrapolation as well as the multiple species alignment results from ~13,050-13,250 bps. I was unable to find strong evidence supporting either of these first two methionines over the other, so I chose the methionine providing the longest open reading frame for my unc-13B gene model.

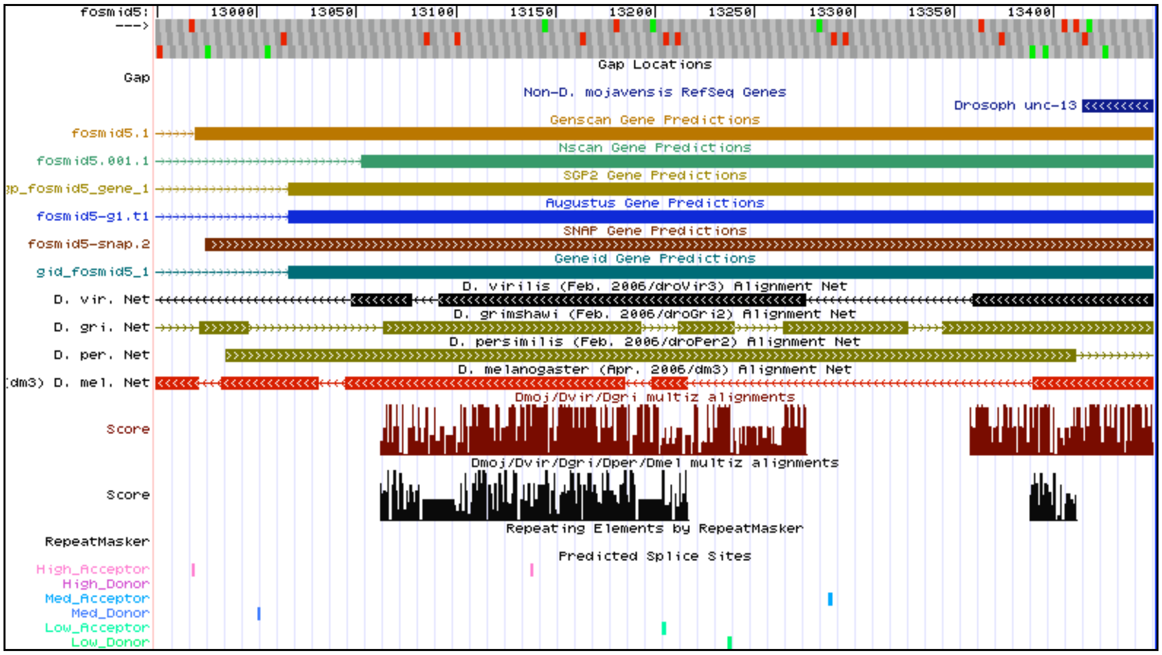


Figure 13. UCSC genome browser of 205-H09 displaying all possible methionine start sites for *unc-13* isoform B. The exon aligns to the third reading frame.

As I continued to define the isoforms of unc-13 in 205-H09, I eventually reached a point where the *D. melanogaster* exons no longer yielded significant alignments regardless of how much I raised the expect value on the bl2seq search. This seemed strange considering the highly conserved exons I had encountered thus far. I assumed this change indicated that the *D. mojavensis* unc-13 gene continued in the fosmid adjacent to mine. To confirm this, I used the BLAT program with the end of 205-H09 (20,000-35,708 bps) to identify the adjacent fosmid in the *D. mojavensis* assembly. The top match corresponded to bases 20,842-26,598 in fosmid 200-N19. I then searched the unique region of 200-N19 (26,599-40,283) for exon 12, the first exon that did not align with 205-H09. Exon 12 mapped along its entire length to the unique region of 200-N19, so I felt confident that my fosmid only contained exons 1-11 and that unc-13 continued in 200-N19.

I continued my work by defining the unc-13 exons present in 200-N19. I used the same methodology described in detail for Rad23 to construct this extension of my gene model. All of the *D. melanogaster* exons present in 200-N19 matched with high identity so that finding intron donors and acceptors to designate exon boundaries was a straightforward process. All the exons I characterized in 200-N19 (exons 12-17) were shared by all three unc-13 isoforms.

Once again, however, I reached a point where exons suddenly stopped yielding significant alignments in my bl2seq searches. I could no longer generate matches with exons 18 and following. Based on my previous experience, I once again assumed the unc-13 gene continued into a fosmid adjacent to 200-N19. To confirm this, I used the same BLAT analysis to identify the adjacent fosmid and check for the unaligned exons in its unique region.

The adjacent fosmid turned out to be 210-H10 from the 2007 *D. mojavensis* annotation. Within this fosmid I was able to finish my annotation of the unc-13 gene.

Exons 18-25 mapped to the fosmid with high conservation and were shared by all three isoforms. The final annotation of isoforms A, B, and C of unc-13 can be found in Table 3 at the end of this report. My annotation spans across three fosmids (205-H09, 200-N19, and 210-H10) in order to encompass the entire unc-13 gene.

Search for Additonal Features

Although my annotations of *Rad23* and *unc-13* involved almost the entire length of 205-H09, I also checked for any possible features between the two genes. The BLASTx results for the region, comparing the fosmid sequence to the non-redundant (nr) protein database, showed only two results (Figure 14). Both matches had high E values, short lengths, and poor identities to a protein sequence that could not be fully coded in the region between Rad23 and unc-13. I therefore concluded that 205-H09 only contains functional coding sequence from the Rad23 and unc-13 genes.

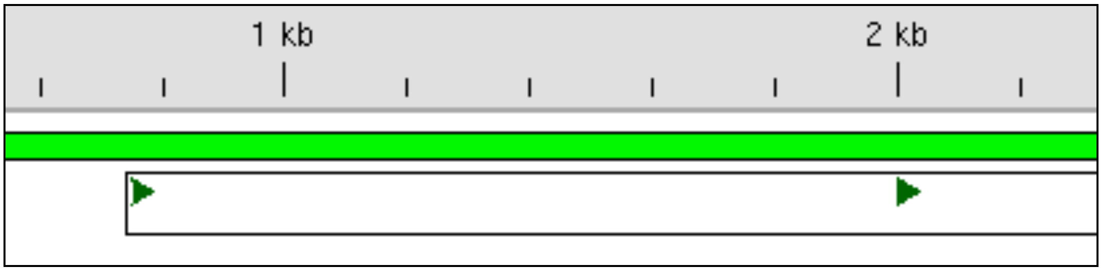


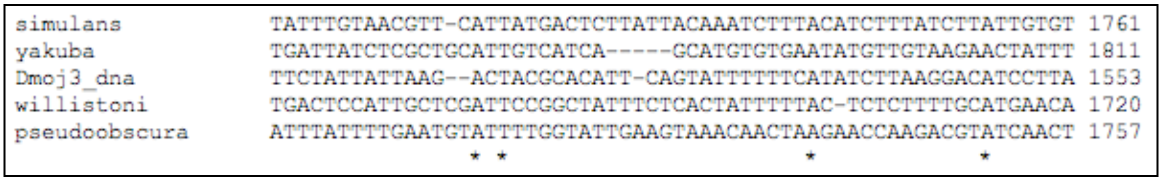
Figure 14. BLASTx output from search of 205-H09 against the nr protein database. Output is displayed in Herne program.

Clustal Analysis

With complete gene models for *Rad23* and *unc-13* in my fosmid, I was able to conduct two separate ClustalW analyses as a part of my annotation procedure. One search was designed to identify conserved non-genic (CNG) elements in a 5' region upstream of *Rad23* and the other was designed to search for protein domains within the protein sequence of *unc-13* (isoform C). The methods and results of this process are summarized below.

Rad23 5' Upstream Region

The 5' upstream region of Rad23 in my fosmid was restricted in length to the region between the first exon of *Rad23* at 520 bps and the first exon of *unc-13* at 2,245 bps. To conduct a ClustalW analysis, I extracted the region from 500-2,245 bps in my fosmid. This 1,746 bp extracted region from the *D. mojavensis* fosmid was aligned with 2,000 bp extractions upstream of *Rad23* in four other *Drosophila* species. I chose to use sequence from species spaced evenly across the Drosophila phylogenetic tree – *D. pseudoobscura*, *D. simulans*, *D. willistoni*, and *D. yakuba* – in the ClustalW alignment because the sequence upstream of *Rad23* was fairly well conserved in closely related species. Despite this careful selection of sequences, the Clustal alignment upstream of Rad23 did not yield any meaningful evidence for CNGs or a 5' UTR (Figure 15). There was no apparent selective pressure and searches for common regulatory element sequences yielded sporadic locations that did not map to a common location between species.

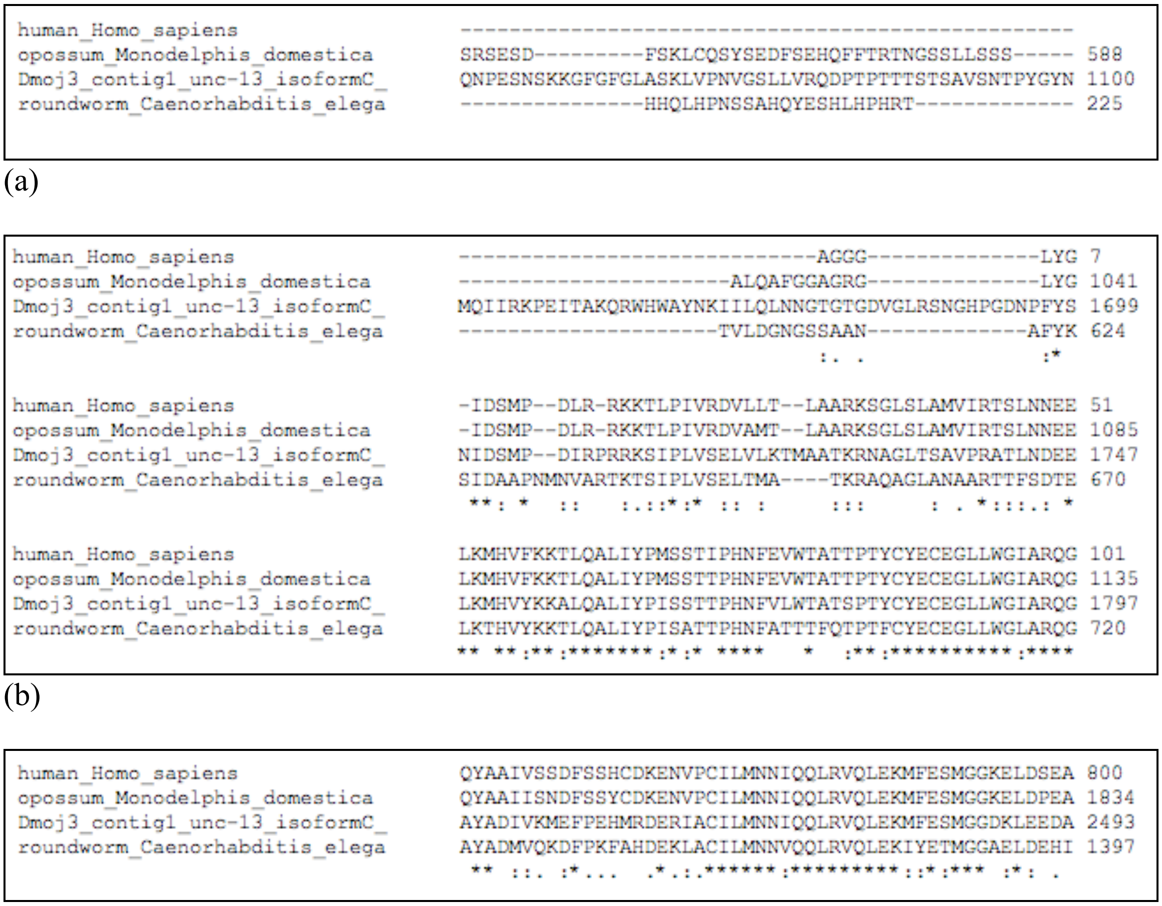


```
simulans         TATTTGTAACGTT-CATTATGACTCTTATTACAAATCTTTACATCTTTATCTTATTGTGT 1761
yakuba           TGATTATCTCGCTGCATTGTCATCA-----GCATGTGTGAATATGTTGTAAGAACTATTT 1811
Dmoj3_dna        TTCTATTATTAAG--ACTACGCACATT-CAGTATTTTTTCATATCTTAAGGACATCCTTA 1553
willistoni       TGACTCCATTGCTCGATTCCGGCTATTTCTCACTATTTTTAC-TCTCTTTTGCATGAACA 1720
pseudoobscura    ATTTATTTTGAATGTATTTTGGTATTGAAGTAAACAACTAAGAACCAAGACGTATCAACT 1757
                                * *                      *            *
```

Figure 15. ClustalW alignment of Rad23 upstream regions in multiple *Drosophila* species. The sequence from fosmid 205-H09 is represented in the line labeled "Dmoj3_dna."

UNC-13 Protein Sequence

I already knew from my gene model annotation that *unc-13* is a very well-conserved gene in *D. mojavensis*. This is logical considering the role of UNC-13 in the core processes of neurotransmitter secretion and synaptic transmission. For a Clustal analysis of protein domains, I found UNC-13 protein sequences from species as distant evolutionarily as possible. The final sequences I used in combination with the annotated *D. mojavensis* sequence were from UNC-13 isoforms in the roundworm *Caenorhabditis elegans*, the opossum *Monodelphis domestica*, and the human *Homo sapiens*. The resulting alignment was interesting because it displayed classic protein domain conservation only in the c-terminal half of the protein (Figure 16). The n-terminal half of UNC-13 had very poor conservation across species, indicating decreased selective pressure and functionality. The transition into the conserved c-terminal region occurs at amino acid 1,696 of the 2,905 amino acid D. mojavensis UNC-13 protein (isoform C).



```
human_Homo_sapiens                  --------------------------------------------------
opossum_Monodelphis_domestica       SRSESD---------FSKLCQSYSEDFSEHQFFTRTNGSSLLSSS----- 588
Dmoj3_contig1_unc-13_isoformC_      QNPESNSKKGFGFGLASKLVPNVGSLLVRQDPTPTTTSTSAVSNTPYGYN 1100
roundworm_Caenorhabditis_elega      ---------------HHQLHPNSSAHQYESHLHPHRT------------- 225
```

(a)

```
human_Homo_sapiens                  -----------------------------AGGG--------------LYG 7
opossum_Monodelphis_domestica       ----------------------ALQAFGGAGRG--------------LYG 1041
Dmoj3_contig1_unc-13_isoformC_      MQIIRKPEITAKQRWHWAYNKIILQLNNGTGTGDVGLRSNGHPGDNPFYS 1699
roundworm_Caenorhabditis_elega      ---------------------TVLDGNGSSAAN-------------AFYK 624
                                                                 :. .              :*

human_Homo_sapiens                  -IDSMP--DLR-RKKTLPIVRDVLLT--LAARKSGLSLAMVIRTSLNNEE 51
opossum_Monodelphis_domestica       -IDSMP--DLR-RKKTLPIVRDVAMT--LAARKSGLSLAMVIRTSLNNEE 1085
Dmoj3_contig1_unc-13_isoformC_      NIDSMP--DIRPRRKSIPLVSELVLKTMAATKRNAGLTSAVPRATLNDEE 1747
roundworm_Caenorhabditis_elega      SIDAAPNMNVARTKTSIPLVSELTMA----TKRAQAGLANAARTTFSDTE 670
                                     **: *  ::   :.::*:* :: :     :::     : . *:::.: *

human_Homo_sapiens                  LKMHVFKKTLQALIYPMSSTIPHNFEVWTATTPTYCYECEGLLWGIARQG 101
opossum_Monodelphis_domestica       LKMHVFKKTLQALIYPMSSTTPHNFEVWTATTPTYCYECEGLLWGIARQG 1135
Dmoj3_contig1_unc-13_isoformC_      LKMHVYKKALQALIYPISSTTPHNFVLWTATSPTYCYECEGLLWGIARQG 1797
roundworm_Caenorhabditis_elega      LKTHVYKKTLQALIYPISATTPHNFATTTFQTPTFCYECEGLLWGLARQG 720
                                    ** **:**:*******:*:* ****   *  :**:**********:****
```

(b)

```
human_Homo_sapiens                  QYAAIVSSDFSSHCDKENVPCILMNNIQQLRVQLEKMFESMGGKELDSEA 800
opossum_Monodelphis_domestica       QYAAIISNDFSSYCDKENVPCILMNNIQQLRVQLEKMFESMGGKELDPEA 1834
Dmoj3_contig1_unc-13_isoformC_      AYADIVKMEFPEHMRDERIACILMNNIQQLRVQLEKMFESMGGDKLEEDA 2493
roundworm_Caenorhabditis_elega      AYADMVQKDFPKFAHDEKLACILMNNVQQLRVQLEKIYETMGGAELDEHI 1397
                                     ** ::. :*...  .*.:.******:*********::*:*** :*: .
```

(c)

Figure 16. ClustalW analysis of UNC-13 showing (a) low conservation in the n-terminal half, (b) the transition from low to high conservation, and (c) high conservation in the c-terminal half (with associated protein domains).

The change in conservation between the n-terminus and c-terminus of UNC-13 is supported by BLAST analysis of the protein sequence. A BLASTx search of the *D. mojavensis* UNC-13 protein sequence (isoform C) against the nr protein database revealed only four high E value alignments to that extended to include amino acids before amino acid 1,650 of the query. No high E value matches aligned to the first 550 aa of the *D. mojavensis* UNC-13 query.

Repeats

Using the RepeatMasker tool, I found that 21.79% of fosmid 205-H09 is composed of repetitive sequence (Table 1). The largest fraction class of repetitive elements was represented by satellite repeats, which compose 8.96% of the fosmid. DNA element repeats were also common, composing 7.60% of the fosmid. The fosmid sequence was composed to a lesser degree of simple repeats (1.80%), unclassified repeats (1.52%), and low complexity DNA (0.27%). RepeatMasker did not identify any SINE, LINE, or LTR elements in the fosmid. Table 2 shows the locations of repeats in 205-H09 above 300 bp in length.

```
GC level:              35.45%
bases masked:          21.79%  (7780 bp)
==========================================
                      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:                   0        0 bp           0.00%
    LINE1                0        0 bp           0.00%
    LINE2                0        0 bp           0.00%
    L3/CR1               0        0 bp           0.00%
LTR elements:            0        0 bp           0.00%
    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:           15        2714 bp        7.60%
    MER1_type            0        0 bp           0.00%
    MER2_type            0        0 bp           0.00%

Unclassified:            3        542 bp         1.52%

Total interspersed repeats:       3256 bp        9.12%

Small RNA:               0        0 bp           0.00%

Satellites:             13        3198 bp        8.96%
Simple Repeats:          7        641 bp         1.80%
Low Complexity:          1        98 bp          0.27%
==========================================
```

Table 1. Summary of Repeatmasker output for 205-H09. The sequence did not include any SINE, LINE, or LTR repetitive elements.

Repeat Class	Position in Query		Length
	Begin	End	
DNA	9707	10238	532
Satellite	21060	21423	364
Satellite	23009	24199	1191
Unknown	25747	26080	334
Satellite	26392	27282	891
DNA	30347	30701	355

Table 2. Repeats greater than 300bp in 205-H09.

I was especially interested in repeat distribution in 205-H09 as it related to the annotated *unc-13* gene. This gene, as evidenced by its gene model construction, is an incredible large structure with some introns longer than 10 kb. Considering the size of the gene alone (55,804 bps) and also its increased size compared to the *D. melanogaster unc-13* gene (33,869 bps), I postulated that the increased gene length was due to repeat insertions and possibly a close association with constitutive heterochromatin. This theory was partially supported in the intron between exons 6 and 7; this intron contains three satellite repeats that are each over 300 bp long (Figure 1). To reach stronger conclusions about the expansion of *unc-13* in *D. mojavensis*, RepeatMasker analysis would have to continue with the *unc-13* annotation into fosmids 200-N19 and 210-H10.

Synteny

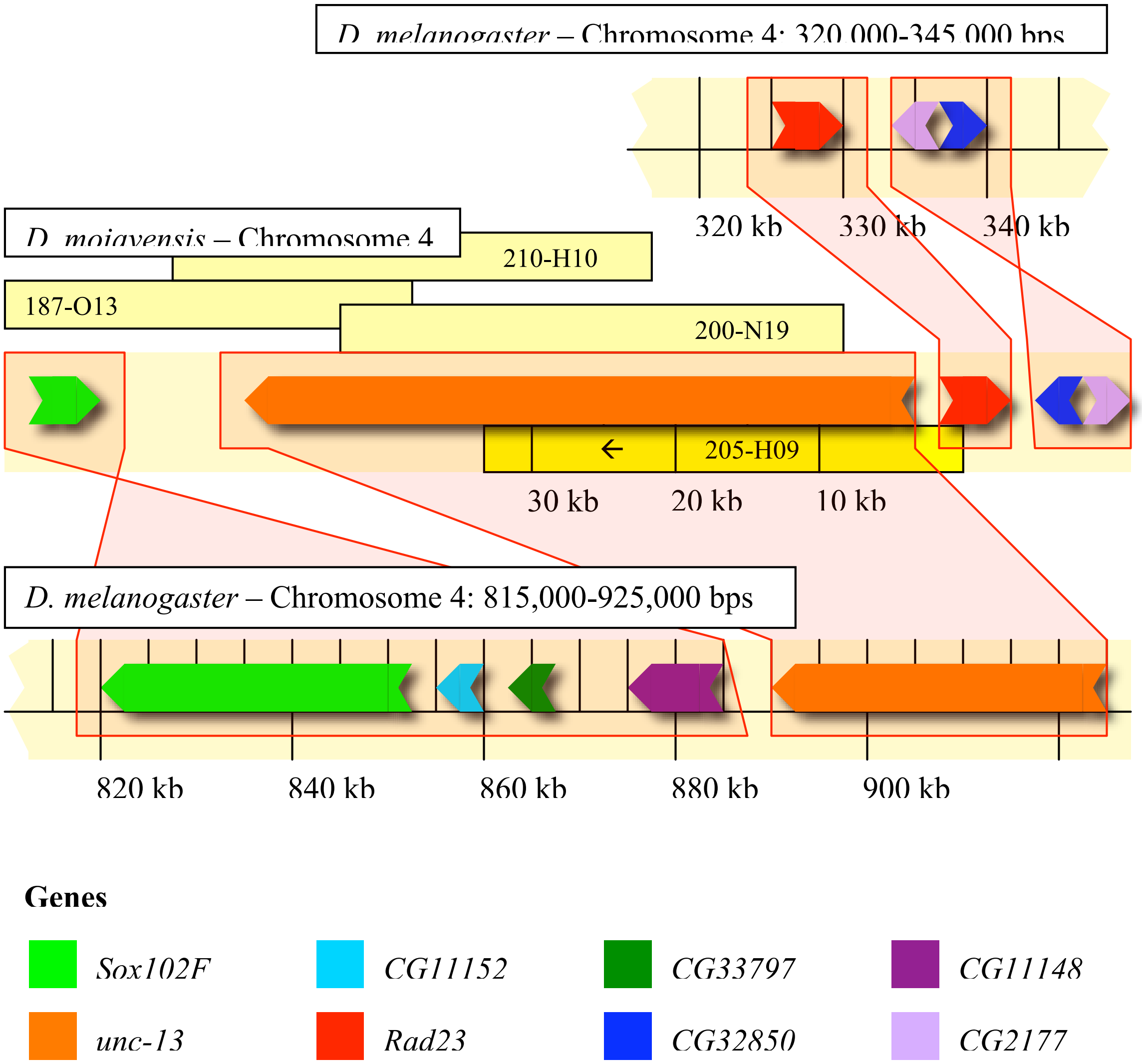


Figure 17. Synteny analysis between *D. mojavensis* 205-H09 and *D. melanogaster*.

Although all the genes within my fosmid and adjacent to my fosmid mapped to the *D. melanogaster* dot chromosome, there was poor evidence for local synteny with *D. melanogaster*. Figure 17 above displays my synteny map, with my fosmid in reverse orientation to match with the *D. mojavensis* and *D. melanogaster* scaffolds. Downstream of my fosmid, the reversed direction of *Sox102F* as well as the apparent loss of three genes between *Sox102F* and *unc-13* indicate a probable inversion. I identified a second inversion upstream of my fosmid that involved a smaller two gene region. The genes in my fosmid, *unc-13* and *Rad23*, had conserved directionality between *D. mojavensis* and *D. melanogaster* unlike the adjacent inverted genes. Nevertheless, the relationship between the two genes is not syntenic because the sequences are separated by approximately 600 kb in *D. melanogaster*.

Discussion

There are some aspects of my entire annotation that I am confident are at least close to if not complete, but there are also aspects of 205-H09 and it's neighboring fosmids which have opportunities for additional investigation. The part of the project I spent the most amount of time on was constructing gene models, especially for the *unc-13* gene with its 25 coding exons and three isoforms across at least three fosmids. My final gene models, which appear in Table 3, are both based on intensive BLASTx searches and careful splice site determinations. Of note in my gene model construction was my identification of unc-13 exons 1 and 5 using Exact String Match in the GEP toolkit.

After I had perfected my gene models, I did my best to use additional analysis tools for further characterization of my fosmid. My ClustalW analysis was interesting as it applied to unc-13 because it identified a strong difference in conservation between the n-terminal and c-terminal halves of the protein. The results of RepeatMasker also had implications for unc-13 and the length of its introns, although analysis in fosmids 200-N19 and 210-H10 would be necessary to confirm this. Finally, I constructed a detailed synteny map illustrating translocations to either side of my fosmid and a clear break in synteny between *unc-13* and *Rad23* (which are 600 kb apart in *D. melanogaster*).

Feature	Transcript ID	Fosmid	Exon*	Isoform	Frame	Start	Stop
1	Rad23	205-H09	1	A	-2	520	455
1	Rad23	205-H09	2a	A	-3	339	261
1	Rad23	205-H09	2b	A	-2	162	1*
2	unc-13	205-H09	1	A, C	+1	2245	2250
2	unc-13	205-H09	2	A, C	+3	2325	7346
2	unc-13	205-H09	3	B	+3	12975	18908
2	unc-13	205-H09	4	A, B, C	+3	19530	19661
2	unc-13	205-H09	5	C	+2	19877	19885
2	unc-13	205-H09	6	A, B, C	+1	20854	20928
2	unc-13	205-H09	7	A, B, C	+1	27670	27889
2	unc-13	205-H09	8	A, B, C	+1	27980	28100
2	unc-13	205-H09	9	B, C	+1	28632	28738
2	unc-13	205-H09	10	A	+1	29505	29611
2	unc-13	205-H09	11	A, B, C	+1	34007	34148
2	unc-13	200-N19	12	A, B, C	+2	26944	27268
2	unc-13	200-N19	13	A, B, C	+1	27337	27609
2	unc-13	200-N19	14	A, B, C	+3	27675	27833
2	unc-13	200-N19	15	A, B, C	+3	27921	28046
2	unc-13	200-N19	16	A, B, C	+1	28216	28383
2	unc-13	200-N19	17	A, B, C	+2	28463	28744
2	unc-13	210-H09	18	A, B, C	-3	15594	15373
2	unc-13	210-H09	19	A, B, C	-3	15252	15086
2	unc-13	210-H09	20	A, B, C	-1	15027	14736
2	unc-13	210-H09	21	A, B, C	-2	14671	14314
2	unc-13	210-H09	22	A, B, C	-1	13675	13577
2	unc-13	210-H09	23	A, B, C	-1	8419	8257
2	unc-13	210-H09	24	A, B, C	-2	7385	7259
2	unc-13	210-H09	25	A, B, C	-3	7197	7039

*Exon numbering begins with the first coding exon.

** Exon 3 of Rad23 isoform A continues into the adjacent fosmid.

Table 3. Final gene annotations for Features 1 and 2.

Appendix

See electronic copy for

- *unc-13* protein sequence fasta file
- *unc-13* nucleic acid sequence fasta file
- *unc-13* GFF file [one for each fosmid annotated]
- *Rad23* protein sequence fasta file
- *Rad23* nucleic acid sequence fasta file
- *Rad23* GFF file
- Fasta sequences for the species used in ClustalW analyses