Craig Wilen

Professor Elgin

Bio 4342

May 2, 2006

Annotation of 7G17

Overview

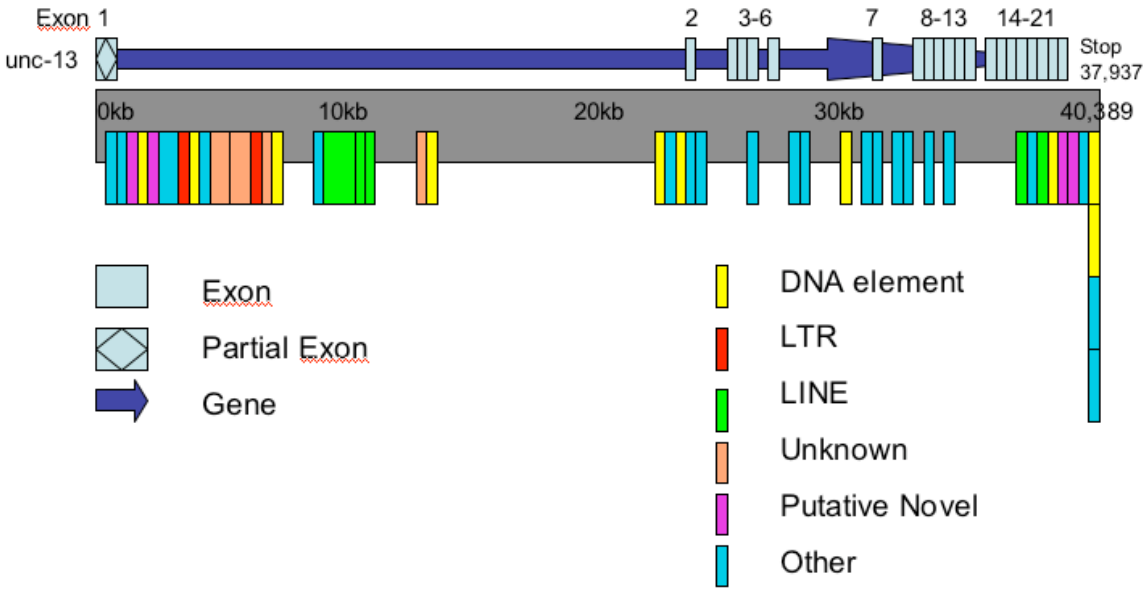

Figure 1. Map of 7G17

Fosmid 7G17 contains the partial sequence of only one gene, unc-13. This protein, critical in neurotransmitter release, is highly conserved between drosophila and mammals suggesting its important role in the nervous system. 37,937bp of the 39,270bp gene representing 20 of 21 exons are located on 7G17. The 40,389bp fosmid contains 45 repetitive elements comprising 27.5% of the entire fosmid including 4 putative novel repeats. Retroelement family repeats and DNA element repeats make up 14.1% and 4.5% of the fosmid, respectively.

Gene

I began the annotation of 7G17 by analyzing the Genscan output (Figure 2).

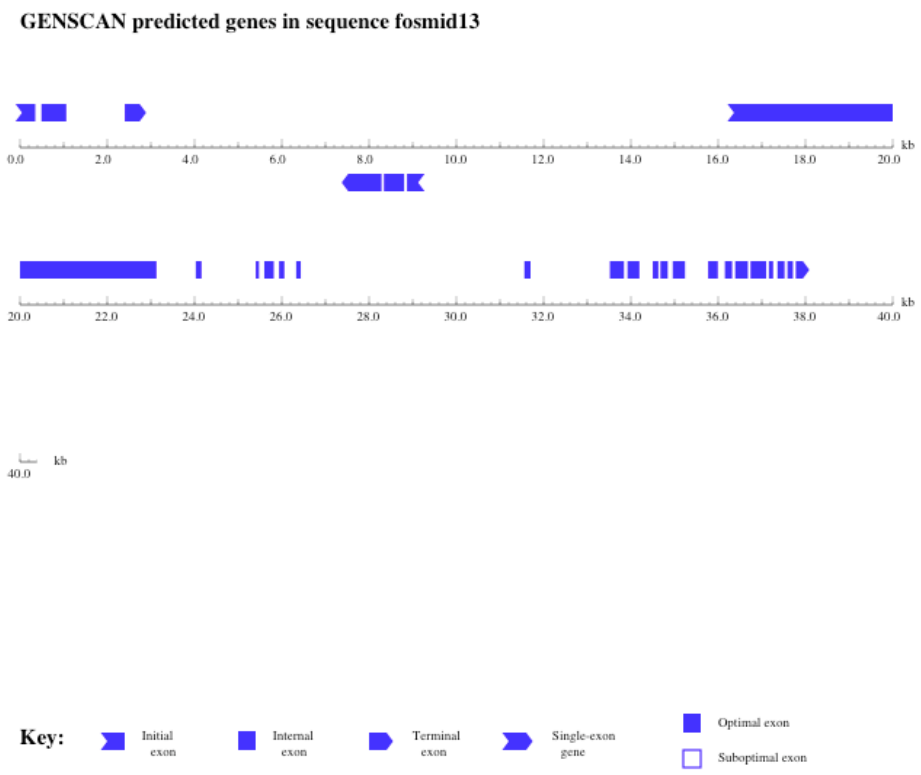

Figure 2. Genscan output of 7G17

The Genscan data predicts that 7G17 contains three genes with the second gene in opposite orientation. All predicted gene features were then compared to *D. melanogaster* by using Blat on the UCSC Genome Browser. Predicted gene feature 1 corresponds to part of the N-terminus of unc-13. The output showing the position of unc-13 on chromosome 4 of *D. melanogaster* as well as its alignment with 7G17 is shown below in Figure 3.

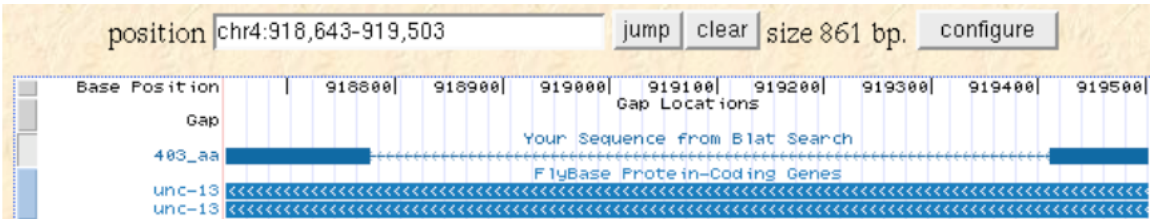

Figure 3. Predicted gene region 1.

Predicted gene region 2 was similarly analyzed. The results shown below in Figure 4 show that predicted gene region 2 matches to an unordered chromosome comprised of all unassembled *D. melanogaster* contigs. Thus, it was determined the predicted gene region 2 was not a gene.

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	493_aa	297	26	391	493	72.0%	U	+-	3115568	3118396	2829
browser details	493_aa	112	363	437	493	78.2%	2h	+-	445835	446058	224
browser details	493_aa	45	109	147	493	69.3%	3R	++	387567	387683	117
browser details	493_aa	36	372	437	493	59.1%	U	+-	2252115	2252312	198
browser details	493_aa	30	372	437	493	57.6%	3h	++	853369	853566	198

Figure 4. Blat hits of predicted gene region 2

Predicted gene region 3 also aligned to unc-13 in *D. melanogaster*. The results are shown below in Figure 5.

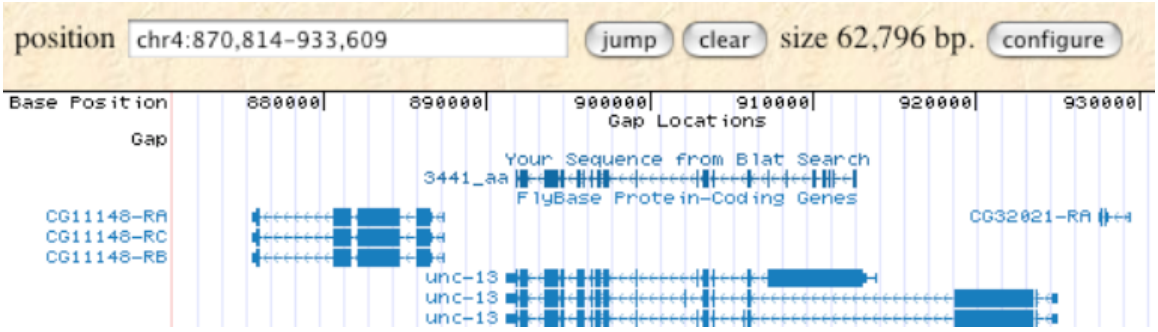

Figure 5. Blat results from predicted gene region 3 against *D. melanogaster*

Thus, the Blat data suggests that only one gene is present in 7G17 and that the genes and exons predicted by Genscan are incorrect. This was confirmed by using Blastn to compare the predicted gene regions on 7G17 with all known sequences. Predicted

gene regions 1 and 3 had highly significant hits to *D. melanogaster* unc-13 (e-value less than 10^-5). Predicted gene region 2 had no significant hits (e-value greater than 10^-5). Thus, both the Blat and Blast data confirm that 7G17 contains only one gene.

Fosmid 7G17 contains the partial sequence of only unc-13. This highly conserved and exclusively neural 200kD protein participates in the regulation of neurotransmitter release. The chemical messengers that transmit information from the pre-synaptic neuron to the post-synaptic neuron are neurotransmitters. The exocytotic release of neurotransmitters from the pre-synaptic neuron is tightly regulated, and the cellular machinery, including the SNARE complex and unc-13, needed for such release is extraordinarily conserved.

Unc-13, short for uncoordinated mutant 13, was first discovered by classical mutagenesis studies that screened for uncoordinated movements in *C. elegans*. Despite the viability of unc-13 knockouts in *C. elegans,* the gene is essential in both drosophila and mice. It was later learned that unc-13 was essential for proper synaptic vesicle fusion. Although the precise function of unc-13 remains elusive, the C-terminus of the protein is known to interact directly with syntaxin, a key component of the SNARE complex. It is thought that unc-13 may bind calcium and the internal messenger diacylglycerol (DAG) as part of its role in the neuron, but more research is needed to identify its precise mechanism of action.

Three unc-13 isoforms have been detected in *D. melanogaster*; however, only isoform A, the largest of the three, was annotated here. The variations between isoforms can be seen below in Figure 6. The differences between isoforms of the large N-terminal

exons at 910kb and 920kb are irrelevant to the annotation of 7G17 because the sequence of these exons is not fully contained in 7G17.

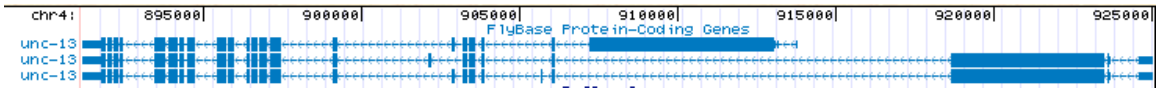

Figure 6. The three isoforms of unc-13 from *D. melanogaster*

To annotate unc-13, Blastx (filter off) was used to compare the entire 40kb of 7G17 with the amino acid sequence of each of the 21 unc-13 isoform-a exons from *D. melanogaster*.

Only 1,060bp of exon 1 are contained in 7G17. The initial 1,333bp of the exon extend beyond my fosmid. However, since the complete exon 1 was annotated last year, we have now completely annotated unc-13.

Exon 2 from *D. melanogaster* is the first complete exon contained in 7G17. A detailed description of its annotation, which follows, serves as a representative example for each of the 20 other exons annotated. Ensembl was used to extract the peptide sequence of exon 2 from *D. melanogaster* (NP_726614). This sequence was then compared to 7G17 by Blastx. The output is below in Figure 7.

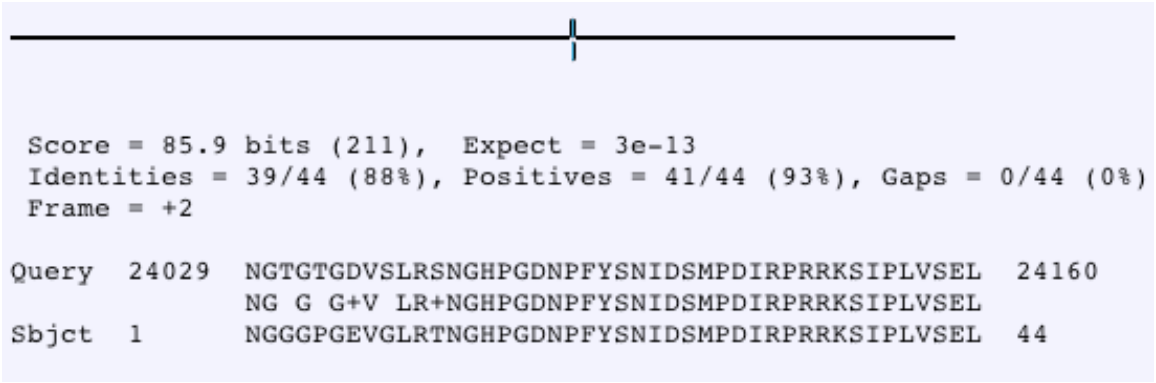

Figure 7. Blastx output of 7G17 and exon 2.

Next, to define the exact intron/exon boundaries for exon 2, the DNA sequence immediately adjacent to 24,029bp and 24,160bp was extracted by using "Get DNA" from 7G17 on the goose.wustl.edu site. The AG end of intron 1 was located at 24,027-8bp and the GT beginning of intron 2 was located at 24,161-2bp. Exon 2 has no phase shift, and its boundaries are 24,029 and 24,160. This method was successfully used to identify the exon boundaries of all 21 exons. The high conservation and limited number of insertions and deletions in unc-13 facilitated this analysis by reducing the search time for AG and GT splice sites. The same analysis on a less conserved protein may have been significantly more challenging and less accurate.

Finally, the 5' UTR extends beyond 7G17 and the conservation of the 3' UTR was too low to produce significant Blast hits; thus, neither can be reported on here.

Clustal Analysis[1]

ClustalW was used to align the region of unc-13 contained in 7G17 from *D. virilis*, with *D. mojavensis*, *D. melanogaster*, *Homo sapien* (human), *Canis familiaris* (dog), and *Mus musculus* (mouse). Blast was used to align the translated partial *D. virilis* sequence with the genbank nr database. The *D. melanogaster*, human, dog, and mouse sequences were then extracted. The *D. mojavensis* sequence was found on the UCSC genome browser under gene mapper predictions (NM_143692).

Unc-13 is highly conserved across all analyzed organisms as expected due to its important role in neurotransmitter release. The most highly conserved region lies between *D. virilis* residues 14 and 836. Although 7G17 does not contain the N-terminus

[1] No promoter region was present on 7G17, so clustal analysis on a promoter region was not possible.

of unc-13, this region appears to posses the greatest variability. Figure 8 shows the N-terminal sequence of unc-13 and a representative portion of the highly conserved middle region.

```
mus_musculus          ------------------------------EEEEGQEGQDAMSR-----AKANWLRAFN 585
homo_sapien           ------------------------------EDEEGQEGQDSMSR-----AKANWLRAFN 485
Canis_familiaris      ------------------------------EDEEGQEGQDSMSR-----AKANWLRAFN 48
partial_virilis       ------------------------------------------------------------
melanogaster          ENLS----DHVSDLTDLSKLISQKKKTLLRGETEEVVGGHMQVLRQTEITARQRWHWAYN 1625
Mojavensis            DNLSVDDEDKLTDLLPTRSKSQKKKKVLLRGETEEVVSGHMQIIRKPEITAKQRWHWAYN 1667


mus_musculus          KVRMQLQELNSQP--------MGSPGGG--LIIIDSMPDIR-KRKPIPLVSDL--SLVQS 632
homo_sapien           KVRMQLQEARGEGEMSKSLWFKGGPGGG--LIIIDSMPDIR-KRKPIPLVSDLAMSLVQS 542
Canis_familiaris      KVRLQLQEARGEGEMSKSLWFKGGPGGG--LIIIDSMPDIR-KRKPIPLVSDL--SLVQS 103
partial_virilis       -------NGTGTGDVS--LRSNGHPGDNPFYSNIDSMPDIRPRRKSIPLVSELT-MAATK 50
melanogaster          KIIMQLNNGGGPGEVG--LRTNGHPGDNPFYSNIDSMPDIRPRRKSIPLVSELT-MAATK 1682
Mojavensis            KIILQLNNGTGTGDVG--LRSNGHPGDNPFYSNIDSMPDIRPRRKSIPLVSELT-MAATK 1724
                             :  .          * **..     ******** :**.*****:*    . .

mus_musculus          TGSSDPYVTVQVGKTKKRTKTIYGNLNPVWEENFHFECHNSSDRIKVRVWDEDDDIKSRV 872
homo_sapien           TGSSDPYVTVQVGKTKKRTKTIYGNLNPVWEENFHFECHNSSDRIKVRVWDEDDDIKSRV 782
Canis_familiaris      TGSSDPYVTVQVGKTKKRTKTIYGNLNPVWEENFHFECHNSSDRIKVRVWDEDDDIKSRV 343
partial_virilis       SGTSDPYVTVQVSKVKKRTRTMPQELNPVWNEKFHFECHNSSDRIKVRVWDEDNDLKSKL 290
melanogaster          SGTSDPYVTVQVSKVKKRTRTMPQELNPVWNEKFHFECHNSSDRIKVRVWDEDNDLKSKL 1922
Mojavensis            SGTSDPYVTVQVSKVKKRTRTMPQELNPVWNEKFHFECHNSSDRIKVRVWDEDNDLKSKL 1964
                      :*:*********.*.****:*:  :*****:*:**********************:*:**::

mus_musculus          KQRFKRESDDFLGQTIIEVRTLSGEMDVWYNLDKRTDKSAVSGAIRLHISVEIKGEEKVA 932
homo_sapien           KQRFKRESDDFLGQTIIEVRTLSGEMDVWYNLDKRTDKSAVSGAIRLHISVEIKGEEKVA 842
Canis_familiaris      KQRFKRESDDFLGQTIIEVRTLSGEMDVWYNLDKRTDKSAVSGAIRLHISVEIKGEEKVA 403
partial_virilis       RQKLTRESDDFLGQTIIEVRTLSGEMDVWYNLEKRTDKSAVSGAIRLHISVEIKGEEKVA 350
melanogaster          RQKLTRESDDFLGQTIIEVRTLSGEMDVWYNLEKRTDKSAVSGAIRLHISVEIKGEEKVA 1982
Mojavensis            RQKLTRESDDFLGQTIIEVRTLSGEMDVWYNLEKRTDKSAVSGAIRLHISVEIKGEEKVA 2024
                      :*::.**************************:**************************
```

Figure 8. ClustalW alignment of unc-13 region

As expected, *D. virilis* aligns more closely with the other two drosophila species than the three mammalian species, but the cladogram of the three drosophila species based on unc-13 does not correspond to the actual evolutionary history of the three species. *D. mojavensis* and *D. virilis* are actually more closely related than *D. virilis* and *D. melanogaster*, but the unc-13 alignment suggests otherwise (Figure 9). This does not appear to be statistically significant since the p-value corresponding to the position of the

D. mojavensis branch is 0.13, which is greater than the .05 cutoff. This demonstrates the importance of using evidence from multiple sources to determine accurate evolutionary relationships.

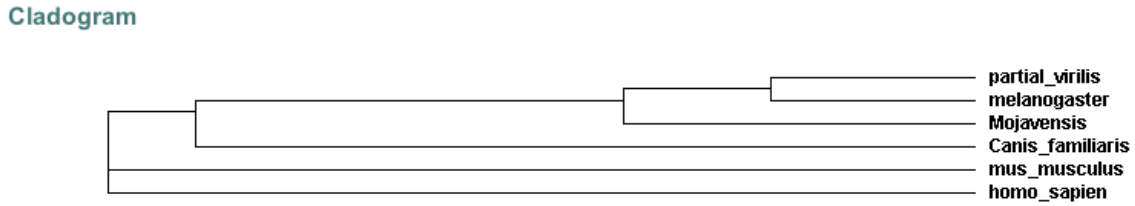

Figure 9. Cladogram from Clustal analysis

Repeats

RepeatMasker was run on 7G17 to identify and mask repeats. Also, a pair-wise Blastn search of 7G17 against the Bio4342 *D. virilis* sequence was conducted. The results are below in Figure 10.

Sequences producing significant alignments:	Score (bits)	E Value
7G17	1.693e+04	0.0
37A19	4224	0.0
47I5	4155	0.0
48K11	115	1e-24
XAAA73	78	3e-13
XAAA63	78	3e-13
26D14	52	2e-05
XAAA112	50	7e-05
45P4	48	3e-04

Figure 10. Blastn output of 7G17 against Bio4342 *D. virilis* sequence

Fosmids 37A19 and 47I5 overlap with 7G17. Only five other hits have e-values less than the e-05 threshold. These five hits correspond to four different regions. Each region was investigated, but only that of 48K11 will be discussed since it is representative of the other three regions.

The putative novel repeat from 48K11 corresponds to 3319-3485bp in 7G17. This can be seen in the alignment below in Figure 11.

```
>48K11
          Length = 39341

 Score =  115 bits (58), Expect = 1e-24
 Identities = 143/171 (83%), Gaps = 4/171 (2%)
 Strand = Plus / Minus


Query: 3319  ggcattcaaattttctacctacttgcgccgagactcggctctgcagcaacaacattacta 3378
              ||||||| |||||||||||||||||| |||||||||||||||| |||||||||| |
Sbjct: 13626 ggcattcgaattttctacctacttgcaccgagactcggctctgcaacaacaacattaatg 13567


Query: 3379  cccaa----ttgtgatgttgttgtttctacgcgctaccttgtaggaacgacaaagaaacc 3434
              |||||    |  || ||||| || || || | ||| | |||||||||||||| |||||||
Sbjct: 13566 cccaaatcgtattgttgttgctggttttatgagctgctttgtaggaacgacagagaaacc 13507


Query: 3435  cgcaaagaacacagtggcattatttgtccctttttctttgcaccgagtttct 3485
               ||||  |||||||||||| || ||||| ||||||||| ||| ||||||||
Sbjct: 13506 agcaacaaacacagtggcagtaattgtctcttttttctttacactgagtttct 13456
```

Figure 11. Repeat alignment of 48K11 and 7G17

It was confirmed that this putative novel repeat was not actually part of another previously masked repeat by examining its position relative to the other repeats listed in Table 1 below. A 400bp region from 3,200-3,600bp was extracted and Blastn was used to compare it with the library of known repeats. The output is below in Figure 12.

Sequences producing significant alignments:	Score (bits)	E Value
TART#LINE/Telomere	30	0.23
DM176_I#LTR/Gypsy Drosophila melanogaster element 17.6	28	0.92
CIRCE#LINE/Composite	28	0.92
TABOR_I#LTR/Gypsy	26	3.6
OSVALDO_I#LTR/Gypsy	26	3.6
INVADER2_I#LTR/Gypsy	26	3.6
HMSBEAGLE_I#LTR/Gypsy	26	3.6
GYPSY_I#LTR/Gypsy	26	3.6
GYPSY9_I#LTR/Gypsy	26	3.6
GYPSY7_I#LTR/Gypsy	26	3.6
BLOOD_LTR#LTR/Gypsy	26	3.6
G7_DM#LINE/Jockey	26	3.6
G3_DM#LINE/Jockey	26	3.6
DOC5_DM#LINE/Jockey	26	3.6
I_DM#LINE/I D.melanogaster (W-IR1 mutation) I factor DNA, compl...	26	3.6

Figure 12. Blastn output of 3,200-3,600bp against repeat library

The lack of significant hits suggests that this repeat is novel. The best match is to a LINE repeat suggesting that this repeat belongs to the retroelement family. To confirm

that this is not a low complexity repeat, the extracted region was compared to the *D. virilis* nucleotide sequence on fly base. The results suggest that this is not a low complexity repeat and that there are similar repeats across the *D. virilis* genome. The output is below in Figure 13.

Sequences producing significant alignments:		Score (bits)	E Value
gnl|dvir|scaffold_13052	freeze 1 assembly	795	0.0
gnl|dvir|scaffold_13174	freeze 1 assembly	470	e-131
gnl|dvir|scaffold_13050	freeze 1 assembly	430	e-119
gnl|dvir|scaffold_12937	freeze 1 assembly	391	e-107
gnl|dvir|scaffold_13324	freeze 1 assembly	351	1e-95
gnl|dvir|scaffold_12238	freeze 1 assembly	335	8e-91
gnl|dvir|scaffold_13049	freeze 1 assembly	309	5e-83
gnl|dvir|scaffold_12100	freeze 1 assembly	238	1e-61
gnl|dvir|scaffold_12958	freeze 1 assembly	224	2e-57
gnl|dvir|scaffold_12936	freeze 1 assembly	188	1e-46
gnl|dvir|scaffold_13036	freeze 1 assembly	174	2e-42
gnl|dvir|scaffold_12970	freeze 1 assembly	172	7e-42
gnl|dvir|scaffold_12728	freeze 1 assembly	157	4e-37
gnl|dvir|scaffold_13045	freeze 1 assembly	121	2e-26
gnl|dvir|scaffold_12799	freeze 1 assembly	117	3e-25
gnl|dvir|scaffold_12723	freeze 1 assembly	107	3e-22
gnl|dvir|scaffold_12963	freeze 1 assembly	98	3e-19
gnl|dvir|scaffold_12954	freeze 1 assembly	92	2e-17
gnl|dvir|scaffold_12875	freeze 1 assembly	66	1e-09
gnl|dvir|scaffold_12967	freeze 1 assembly	58	3e-07
gnl|dvir|scaffold_9800	freeze 1 assembly	50	7e-05
gnl|dvir|scaffold_13047	freeze 1 assembly	46	0.001

Figure 13. Blastn output of 3,200-3,600 against *D. virilis*

The RepeatMasker output was combined with the pair-wise Blast data to generate the complete repeat table below (Table 1). The repeat types as a percentage of 7G17 are summarized in Table 2 below. For comparison, the repeat data for *D. melanogaster* for the region containing unc-13 is in Table 3 below.

Begin	End	Length	Repeat type
1233	1461	229	DNA
2678	2707	30	Putative novel LTR
3319	3485	167	Putative novel LINE
3614	4568	955	TRF
4569	4618	50	LTR
4618	4685	68	DNA
4681	4801	121	TRF

4795	5417	623	Unknown
5414	6300	887	Unknown
6301	6511	211	LTR
6663	6753	91	Unknown
6753	6989	237	DNA
9175	9204	30	Simple_repeat
9205	10140	936	LINE
10135	10190	56	LINE
10184	13094	2911	LINE
13479	13633	155	Unknown
13655	14187	533	DNA
23195	23225	31	Low_complexity
23356	23758	403	DNA
23359	23806	448	LINE
23773	23845	73	DNA
24264	24350	87	Simple_repeat
24709	24738	30	Low_complexity
26587	26670	84	Simple_repeat
28611	28634	24	Low_complexity
28794	28880	87	Low_complexity
30112	30166	55	DNA
31231	31256	26	Low_complexity
31871	31891	21	Low_complexity
32371	32488	118	TRF
32797	32824	28	Low_complexity
33415	33500	86	Simple_repeat
34213	34233	21	Low_complexity
38352	38501	150	LINE
38502	38539	38	Simple_repeat
38540	38788	249	LINE
38755	38827	73	DNA
39010	39092	83	Putative novel LTR
39344	39368	25	Putative novel LINE
39703	40080	378	LINE
40038	40119	82	DNA
40114	40183	70	DNA
40233	40255	23	Low_complexity
40351	40373	23	Low_complexity

Table 1. Individual repeat data for 7G17

	% DNA elements	% Retro elements	% Total repeats
#bp	1823	5694	11106
% of total	4.51%	14.10%	27.50%

Table 2. Summary of repeat types in 7G17

	%DNA element	%Retro element	% total repeats
#bp	1823	5694	11106
% of total	4.51%	14.10%	27.50%

Table 3. Summary of repeat types in unc-13 region of *D. melanogaster*

Fosmid 7G17 has 45 repeats of which 28 are greater than 100bp and 6 are greater than 500bp. The percentage of retroelement repeats including LTRs and LINEs is nearly 3 times that of DNA element repeats. Also, four putative novel repeats which are most likely retroelement family repeats were detected. No readily discernible repeat pattern is evident by examining 7G17 but it appears that the repeats cluster together in certain regions particularly around 5kb, 10kb, and 39kb.

Synteny

Unc-13 is on Chr 4 of *D. melanogaster* as expected. Since 7G17 contains only one gene, it is not possible to evaluate synteny, by definition. The two genes that flank unc-13 in *D. melanogaster* are CG11148 and CG32021. The UCSC Genome Browser output of the *D. melanogaster* unc-13 gene is shown in Figure 14.

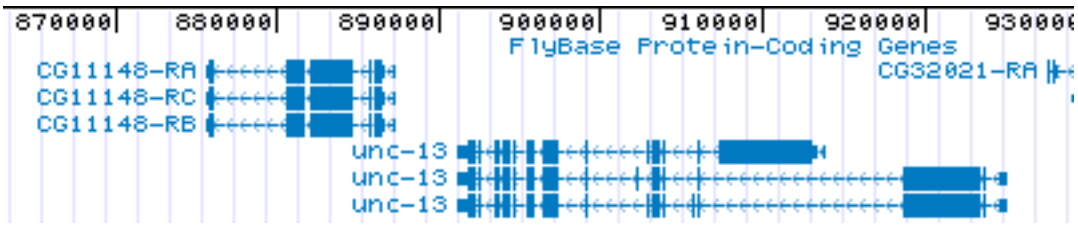

Figure 14. UCSC Genome Browser output.

Annotated *D. virilis* vs *D. melanogaster*

The final annotation of *D. virilis* unc-13 is aligned against *D. melanogaster* unc-13 in Figure 15 below.

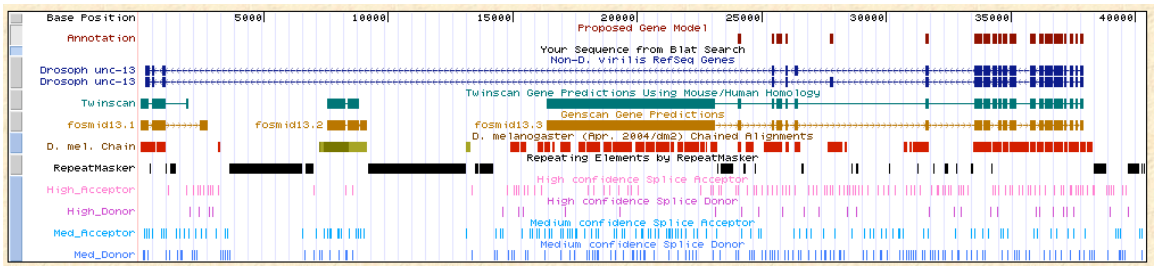

Figure 15. Annotated *D. virilis* and *D. melanogaster*

The second drosophila refseq is isoform A. It does not align perfectly because BLAT is not as sensitive as Blast. Ensembl was used to annotate unc-13, but this output shows refseq.

Summary

7G17 spans the gap between fosmids 37A19 and 47I5. It encodes unc-13, a protein important in synaptic vesicle fusion and neurotransmitter release. Unc-13 is encoded by 37kb and occupies the majority of the 40kb fosmid. Repeats are dispersed throughout 7G17 and the total repeat percentage is 27.5% which is consistent with that seen for other dot chromosome fosmids annotated in this class.