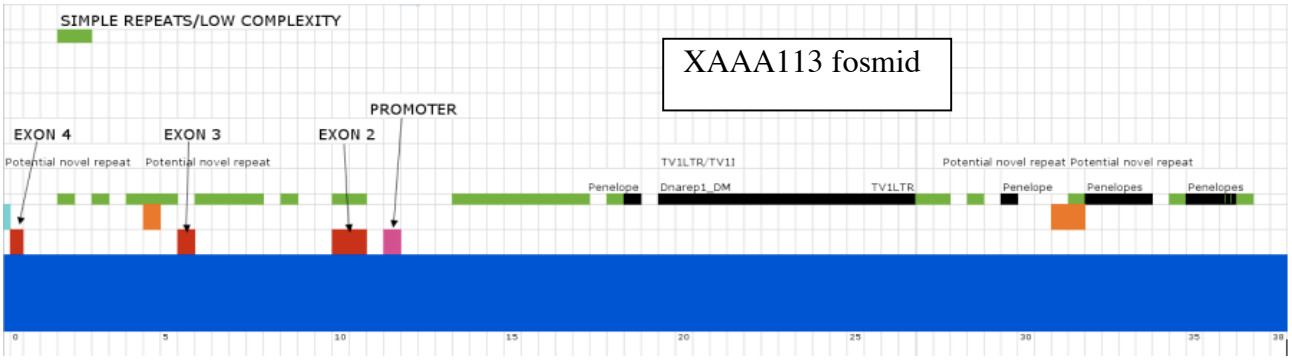
XAAA113 Fosmid Annotation

Andrew Nett

5/7/04

OVERVIEW



The XAAA113 fosmid contains only a partial fragment of one gene. Three discernible exons of this gene – a homologue of the D. *melanogaster* gene CG2052 – are present on the *littoralis* fosmid. The first exon of this gene, if it exists, encodes a UTR, as does the first exon of the *D. melanogaster* gene. The second, third, and fourth exons of the *littoralis* gene exist, respectively, at bases 10337-9482, 5313-5236, and 507-223. Clustal analysis and blastn alignments of multiple Drosophila species suggest that a conserved promoter region containing a TATA box exists at roughly base 11900-11986 of the *littoralis* fosmid. This promoter sequence does not seem to be specific for the CG2052 gene. XAAA113 also contains 57 repetitive elements including two classes of potential novel repeats. The fosmid is syntenic with the fourth chromosome of *D. melanogaster* through at least the first 16,000 bases of XAAA113. Divergence from chromosome four sequence beyond that point may stem from a 40kb-long void of genes, although one cannot rule out that the remaining fosmid sequence better aligns to a different *D. melanogaster* chromosome.

XAAA113 GENES

Preliminary gene exploration

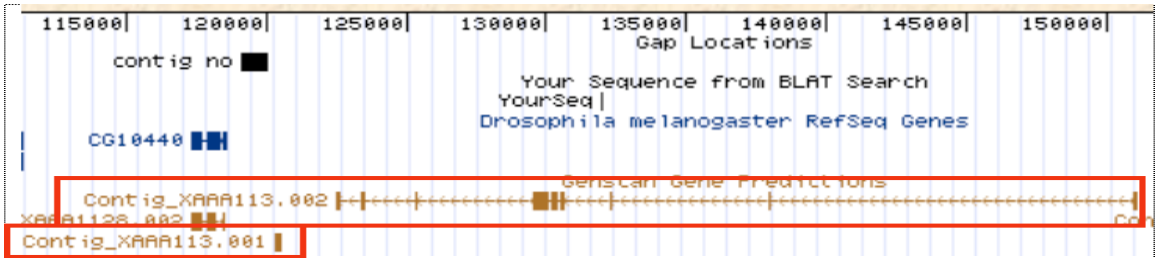


Fig 1. Genescan predictions.

Genescan predicts that the XAAA113 fosmid encodes two genes (Figure 1). The first is a single exon gene beginning at base 486 and ending at 222 with a length of 265 bases. The second predicted gene has nine exons located throughout the contig, spanning from base 31045 to base 2408 (Table 1). Despite this prediction, only one gene actually exists on the contig with exons overlapping predicted exons of both Genescan genes.

Evidence for this hypothesis begins with a blastx query of unmasked XAAA113 against a Drosophila *melanogaster* protein database (Figure 2). This search yields matches to only one region of the *littoralis* fosmid with the best hit occurring to the protein CG2052-PB (Accession #24638609). CG2052, having nine exons, contains a C2H2-type zinc finger domain. It has a role in transcriptional regulation and may bind to DNA promoter or enhancer regions. Initial blastx alignment to this 1,097 aa protein occurs to the contig at bases 217-510 with 86% identity (matching a.a.'s 342-349) and at bases 9458-9595 with 73% identity (matching a.a.'s 279-324). A blastn query of unmasked XAAA113 against the D. *melanogaster* EST database results in a match that overlaps with CG2052 alignment. The EST specified as GH06573. complete AY58304 aligns to fosmid bases 224-384 (84% identity) and to bases 5236-5318 (89% identity).

GENE	EXON	POSITION	LENGTH (bp)
1	1	486-222	265
2	1	31045-30913	132
	2	17967-17885	82
	3	12281-12180	101
	4	10572-1049	153
	5	10337-10142	195
	6	10033-9482	551
	7	5313-5236	77
	8	3434-3335	99
	9	2619-2408	211

Table 1. Genescan predictions.

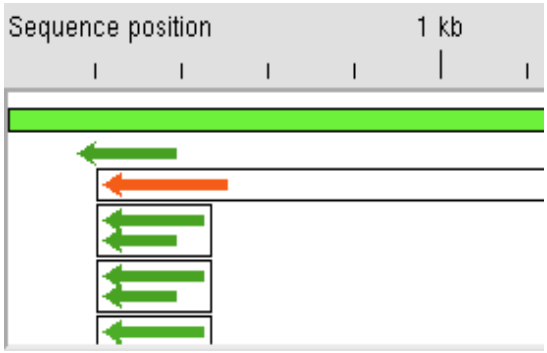


Fig 2. Blastx alignment of CG2052 (aa 342-349) to XAAA113.

Additionally, blat alignment of the first 25,000 bases of the XAAA113 fosmid occurs to a region of the fourth *D. melanogaster* chromosome that contains the CG2052 gene (Figure 3).

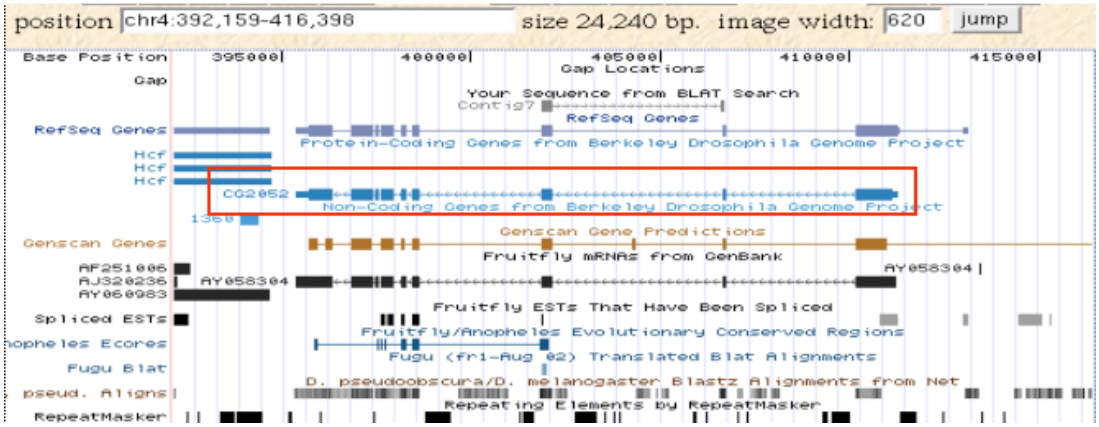


Fig 3. Blat alignment of XAAA113 (base 1-25000) to *D. melanogaster* chromosome 4.

Alignment results suggest the *littoralis* fosmid may contain CG2052, but further analysis is necessary to determine if *littoralis* sequences similar to CG2052 are part of a gene or pseudogene. Blat alignment (Figure 3) and an unfiltered tblastn blast2 alignment of XAAA113 against the CG2052 amino acid sequence (accession #45551180) establishes that the fosmid contains at most only a partial fragment of the CG2052 gene since the end of the contig falls in the middle of the potential gene.

Furthermore, coding of the first 51 aa of the protein is not found by blast2 query. Alignment occurs with only 36% identity to aa 52-377 of CG2052 at fosmid base 10351-9458 (Figures 4,5). CG2052 aa's 369-399 align to base 5316-5224 (90% identity), and aa's 395-492 align to base 510-217 (86% identity).

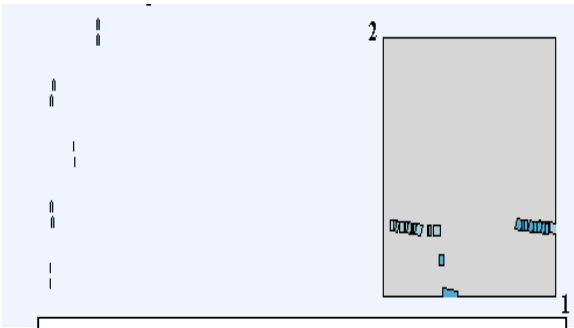


Fig 4. Tblastn blast2 alignment of CG2052 and XAAA113.

Fig 5. XAAA113 alignment to CG2052 (36% identity). * * * *extract numbers??????

Specific Blast2 query (blastx and tblastn) of only the first 51 aa of CG2052 against the unmasked *littoralis* contig yields no results. Alignment does not occur even with an expectancy stringency raised to a value of 1,000,000. Multiple sequence alignment, however, does show relative

conservation of a sequence less than 51 aa long that is upstream of the regions of similarity revealed by the above blast2 query (Figure 6).

Clustal analysis

If the alignment of XAAA113 base 10351-9458 to CG2052 is taken as a potential – though poorly – conserved exon, Clustal analysis of the surrounding fosmid may possibly uncover a nearby upstream exon short enough to escape blast2 alignments of CG2052 and XAAA113.
Indeed, blastn searches of an XAAA113 extract (base 8500-12500) yield matches to *D. melanogaster*, *D. yakuba*, and *D. pseudoobscura* contigs that align to almost the exact same location of the *littoralis* query. Alignment to base 11900-11986 of the *littoralis* fosmid occurs at base 413116-413202 of *D. melanogaster* chromosome 4 (88% identity). This exact region in the *littoralis* contig also aligns to Contig5960_Contig5609 of a genomic *D. pseudoobscura* database at base 75459-75526 with 94% identity. Additionally, base 11900-11987 of the *littoralis* contig matches with 88% identity to base 16885-16799 of Contig 25.42 of the genomic *D. yakuba* database. (A blastn query of the XAAA113 extract masked by Repeatmasker against the *D. yakuba* database gives the same result, ensuring that alignments to the same region of the *littoralis* contig are not the result of some common repeat element existing at that location.)

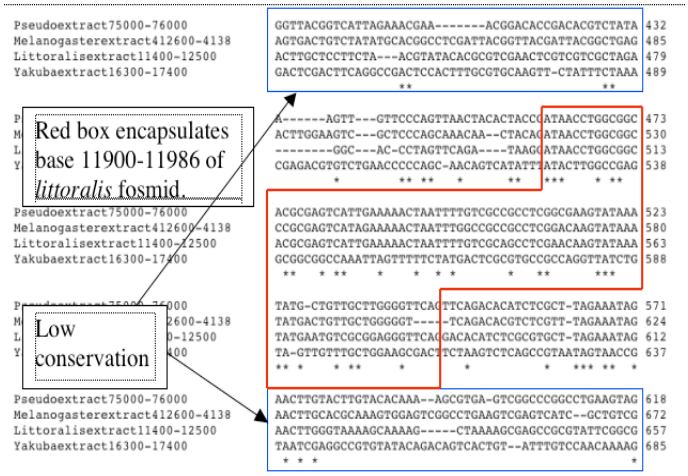


Fig 6. Conservation relatively high around region corresponding to *littoralis* base 11900-11986.

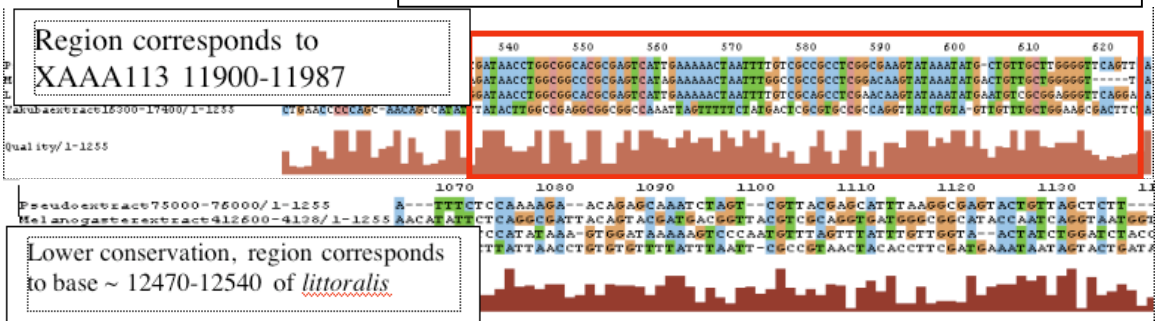


Fig 7. Clustal alignment output; comparison of conservation.

Clustal alignment of sequence extracts (flanking the region of alignment to *littoralis* by 500 bp on each side) shows conservation across the four Drosophila species that appears relatively high around the region targeted

by blastn corresponding to *littoralis* contig base 11900-11986 (Figures 6,7).

At this point, a search for open reading frames in the translated sequence of the *littoralis* fosmid surrounding base 11900-11986 would seem to be the next step in elucidating an exon in the area. Closer examination of the blastn alignment of base 11900-11986 to the fourth *D. melanogaster* chromosome, however, suggests that this *littoralis* region does not actually encode a CG2052 exon, but instead is a promoter region. The region actually matches *D. melanogaster* precisely upstream of the first CG2052 exon (Figure 9). Figure 10 illustrates that the CG2052 gene begins at base 413128 of the *D. melanogaster* chromosome 4. As stated above, base 11900-11986 of XAAA113 (a region with high similarity to sequence in *D. melanogaster, D. pseudoobscura,* and *D. yakuba*) aligns to base 413116 – 413202 of the fourth chromosome in *D. melanogaster.* This alignment only overlaps 12 bases of the CG2052 gene while the rest of the match corresponds to the *D. melanogaster* region directly upstream of the gene.

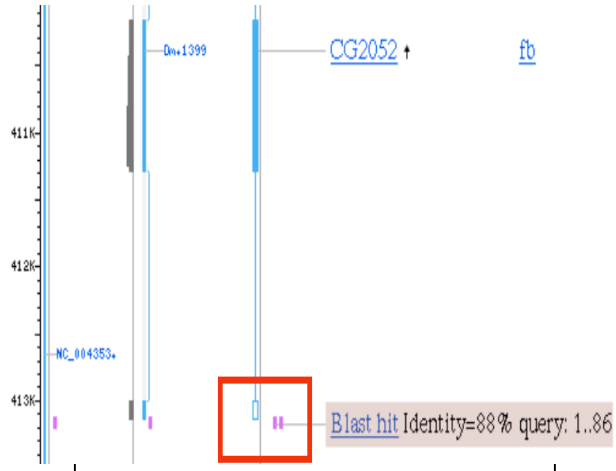


Fig 9. 11900-11986 of XAAA113 hit upstream of CG2052.

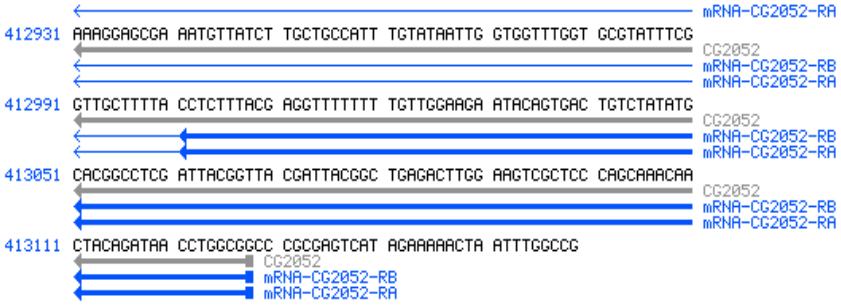


Fig 10. CG2052 position on *melanogaster* chromosome 4.

Thus, bases 11912 through 11986 of the *littoralis* contig may contain a conserved promoter region, which regulates CG2052 transcription. In fact, Clustal alignment reveals what looks like a TATA-box in this region – though this element is approximately 55 bases upstream from the beginning of the *D. melanogaster* CG2052 gene, which is farther than usual (Figure 11). Stretches of GC-dense sequence also appear within this potential promoter region.

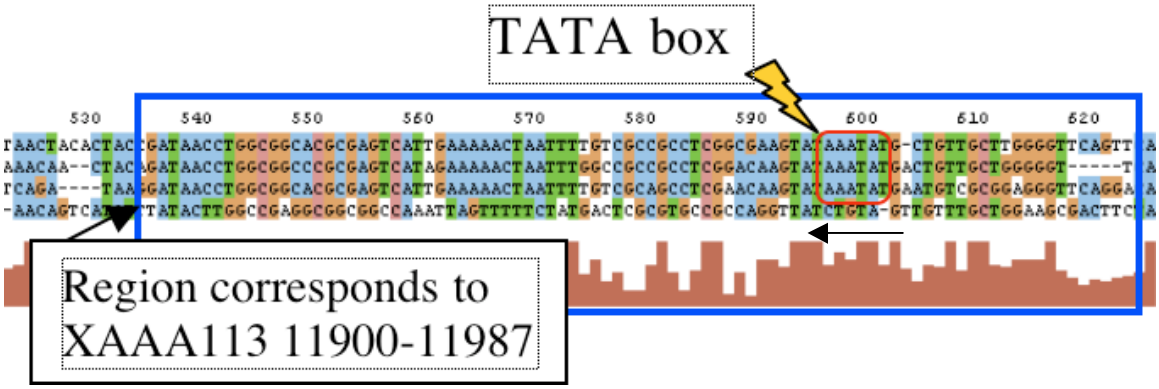


Fig 11. Conserved promoter region @ base 11912-11986 contains a TATA box.

The location of the conserved CG2052 promoter region – which directly neighbors the start of the CG2052 gene in *D. melanogaster* – is now apparent in XAAA113. But the first 52 aa of the protein are still nowhere to be found. Clustal alignment does not retain conservation downstream of the promoter region. The reason for this disappearance becomes obvious upon realization that the entire first exon of the CG2052 gene in *D. melanogaster* is a UTR (Figure 12). The referenced CG2052 aa sequence initially used in blast2 alignment against XAAA113 mistakenly includes this UTR (accession #45551180). This UTR accounts for the first 48 residues of the 51 unaligned aa's of CG2052.

Fig 12. First CG2052 exon in *D. melanogaster* entirely a UTR.

```
CTTTAACCAGGCCCAAACTTCGCACAGTTGGCACCCCAGATCCGTGTTCAGGGTCGTAAG
------------------------------------------------------------

CAGTTGTCCTCCCGATCGAAAAAAACGGCCAAGCCCAGCGAGCTGGCAAATCTGTATTGG
------------------------------------------------------------

CCATCTCCTAACAAATTATCTTGGAATGTTTAAAATGGACTCCGCAGACTTCTGGCAGCA
-------------------------M--F--K--M--D--S--A--D--F--W--Q--Q

AGCGCGTGCTCCGTTTGGTCTGCAAACTGCGCTTCACCAATACTCTTCCCCCCCGAATCA
--A--R--A--P--F--G--L--Q--T--A--L--H--Q--Y--S--S--P--P--N--Q
```

Fig 13. Partial CG2052 sequence. First exon UTR highlighted in yellow.

The location and existence of this CG2052 untranslated exon is uncertain in *littoralis* since it is not conserved.

CG2052 Exon Determination

With no information regarding the first UTR exon, the first amino acid of the *D. melanogaster* CG2052 is now placed at the beginning of translated sequence. Apart from this different numbering, blastx blast2 alignment of CG2052 aa sequence (excluding UTR's) produces results (Table 2) almost

identical to those of the blast2 alignment described previously. These matches overlap regions of Genescan gene predictions.

XAAA113 location	CG2052 location	Alignment Identity	Expectancy	Genescan overlaps
10345-9458	1-324	36%	6.00E-34	Gene 2: exons 5,6
5316-5224	316-346	90%	4.00E-05	Gene 2: exon 7
520-217	342-439	86%	1.00E-44	Single exon gene 1

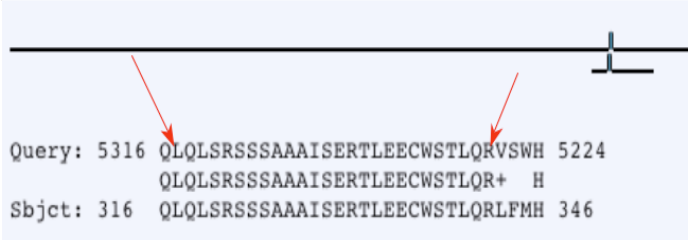
Table 2. Blastx blast2 matches of CG2052 aa sequence to XAAA113 fosmid.

For example, the region from base 10345-9458 of the *littoralis* contig, which aligns to aa 1-324 of CG2052, contains two Genescan exon predictions. These predicted exons, specifically, are exons 5 and 6 of the second Genescan gene (Table 1). If fused, the Genescan prediction spans from base 10337 to 9482. Since the first exon of CG2052 is a UTR, the methionine existing at base 10337 makes a suitable site for initiated translation of the CG2052 homologue in D. *littoralis*. The end site of this second exon should also follow Genescan prediction – and in fact, base 9482 is followed by a "GT" necessary at an intron's beginning (Figure 14). Furthermore, bases 10337 – 9482 align to aa 4-316 of CG2052, almost exactly matching the 316 aa size of the second CG2052 exon.

Fig 14. Blast2 alignment of XAAA113 and CG2052. Second exon boundaries.

The blast2 alignment of CG2052 sequence to XAAA113 occurring at fosmid base 5316-5224 also overlaps a Genescan prediction – exon 7 of the second predicted gene. Here, Genescan expects a 77-bp exon to exist from base 5313 to 5236. This length closely matches the 78-bp 3rd exon of CG2052. The predicted exon starts with a leucine residue following "AG" and ends immediately before a "GT" (Figure 15). Thus, an intron ending in the expected AG precedes the predicted exon, while an intron beginning with the expected GT follows it. This exon aligns to aa 315-342 of CG2052.

Fig 15. Boundaries of third exon of CG2052 homologue.



Blast2 alignment of CG2052 to base 520-217 of XAAA113 underlines the fourth exon of the CG2052 *littoralis* homologue. This alignment is 293 bases long – close to the 286 bp length of the fourth CG2052 exon – and also overlaps a Genescan prediction. This postulated single exon gene spans from base 486-222 with a length of 265 bases. Since Genescan actually predicts a separate gene in this location, demarcation of the fourth exon of the CG2052 homologue does not exactly follow Genescan boundaries. An exon beginning with a leucine residue at base 507 would immediately follow a necessary AG splice site at the intron's end. If ending at base 223 with a threonine residue, the exon immediately precedes "GGT", containing a GT to mark the beginning of an intron (Figure 16). This exon aligns to aa 343-437 of CG2052.

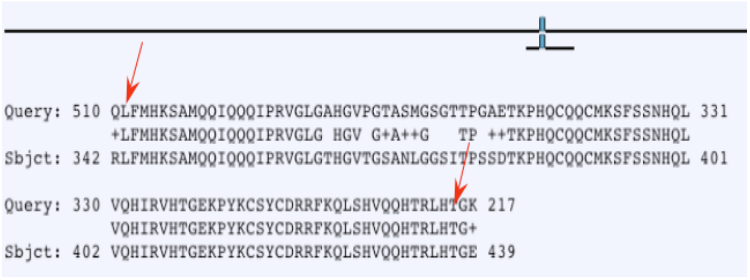


Fig 16. CG2052 homologue 4th exon boundaries.

Thus, the fragment of CG2052 homologue present at the beginning of the XAAA113 fosmid possibly contains 4 exons and one promoter sequence. One such exon (corresponding to the UTR first exon of CG2052), however, may not be present or it may have diverged enough to prevent demarcation of its boundaries. The gene's promoter region exists around base 11912-11986 with a TATA box starting at base 11965. The second exon (corresponding to the second CG2052 exon) occurs at base 10337-9482. The third is at base 5313-5236, and the fourth – at base 507-223 (Table 3).

Exon	Start position	End position
1	UNKNOWN	UNKNOWN
2	10337	9482
3	5313	5236
4	507	223
Promoter	11912	11986

Table 3. CG2052 feature boundaries.

Other XAAA113 Genes

Many of the exons of the second Genescan gene prediction actually overlapped with CG2052, as did the first single exon Genescan gene. The second predicted gene was thus ignored as a few bogus exons linked to real exons of the CG2052 homologue. The absence of blast evidence of any features beyond CG2052 supports disregarding the leftover exon predictions. Additionally, in *D. melanogaster*, a large empty region exists for more than 40kb past the CG2052 gene on chromosome 4 (Figure 17). Though it is not certain if the entire XAAA113 contig corresponds to *D. melanogaster* chromosome 4. Thus, XAAA113 contains only a partial

fragment of a gene, having 3.26% coding DNA, and a gene density of roughly .011 genes/kb (.4 genes/37.265 kb).

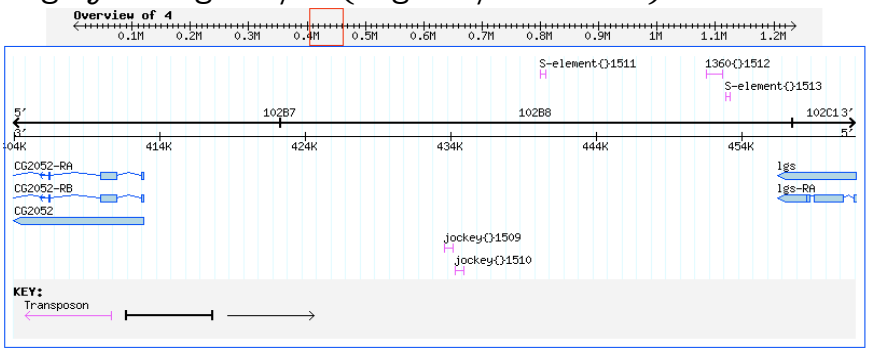


Fig 17. Chromosome 4 empty for >40 kb upstream of CG2052.

XAAA113 REPEAT ANALYSIS

Repeat #	Repeat Class/Family	Repeat	Position on Contig
1	Potential novel repeat	unknown	001-041
2	Low complexity	AT rich	1848-1875
3	Simple Repeat	(CATATA)n	2712-2806
4	Low complexity	AT rich	2894-2929
5	Simple Repeat	(TATG)n	3713-3747
6	Simple Repeat	(TA)n	4059-4106
7	Potential novel repeat	unknown	4121-4207
8	Low complexity	AT rich	4635-4662
9	Simple Repeat	(TATG)n	4670-4728
10	Low complexity	AT rich	5561-5629
11	Simple Repeat	(TAA)n	6064-6104
12	Low complexity	AT rich	6298-6344
13	Simple Repeat	(TA)n	6896-6928
14	Simple Repeat	(TG)n	7149-7294
15	Simple Repeat	(TA)n	8030-8114
16	Simple Repeat	(CTG)n	10069-10139
17	Low complexity	AT rich	10883-10941
18	Low complexity	AT rich	13598-13629
19	Low complexity	AT rich	14010-14040
20	Low complexity	AT rich	14188-14210
21	Simple Repeat	(T)n	14880-14905
22	Simple Repeat	(CAG)n	15152-15211
23	Simple Repeat	(CGA)n	15585-15659
24	Simple Repeat	(CA)n	15820-15869
25	Simple Repeat	(TA)n	16331-16413
26	Simple Repeat	(TATG)n	16417-16574
27	Simple Repeat	(CATATA)n	16461-16584
28	Low complexity	AT rich	16761-16785
29	Low complexity	AT rich	16821-16890

30	Low complexity	AT rich	16916-16944
31	Low complexity	AT rich	17049-17071
32	Low complexity	AT rich	17244-17281
33	Simple Repeat	(TA)n	17434-17596
34	Low complexity	AT rich	18364-18422
35	LINE	PENELOPE	18521 - 18970
36	DNA	DNAREP1_DM	19593 - 19685
37	LTR/GYPSY	TV1LTR	20088 - 20499
38	LTR/GYPSY	TV1I	20500 - 26479
39	LTR/GYPSY	TV1LTR	26480 - 26891
40	Simple Repeat	(TATATG)n	27033-27111
41	Simple Repeat	(TA)n	27187-27244
42	Low complexity	AT rich	27609-27652
43	Simple Repeat	(TG)n	28928-28961
44	LINE	PENELOPE	29898 - 29953
45	Potential novel repeat	unknown	31454-31414
46	Potential novel repeat	unknown	31521-31561
47	Low complexity	T-rich	31780-31852
48	LINE	PENELOPE	32403 - 32538
49	LINE	PENELOPE	32782 - 32829
50	LINE	PENELOPE	33357 - 33774
51	Low complexity	AT rich	34563-34583
52	LINE	PENELOPE	35012 - 35449
53	LINE	PENELOPE	35479 - 35628
54	LINE	PENELOPE	35921 - 36364
55	Simple Repeat	(TCCG)n	36073-36103
56	LINE	PENELOPE	36104-36364
57	Low complexity	AT rich	36760-36789

Table 4. Repeats present in XAAA113 fosmid.

30.63% of the XAAA113 fosmid consists of repetitive DNA including low complexity regions, which themselves comprise 2.05% of the contig. It contains 53 total known repeat elements and regions of low complexity. Of these 53 elements, all but 14 are simple repeats or regions. In addition, the fosmid contains two potential novel repeat elements, one of which is present once (repeat #7) while the other is present in 3 copies (repeats 1, 44, and 45 of Table 4). One of these copies (#44) is inverted. All three copies do not overlap with either EST alignment to XAAA113 or other repetitive elements.

Pairwise blast of the XAAA113 contig to itself reveals this potential repeat element. Bases 31521 through 31561 (repeat #45) align to the region from base 31454 to 31414 (repeat #44) with 87% identity, though the expectancy value of this alignment is rather high at E=0.22 (Figure

18). Although this high value questions the validity of the sequence as a repetitive element, a blast2 alignment of extracted base 31521-31561 shows that the entire potential repetitive element also aligns to bases 1-41 of the contig (Figure 19). This match has 100% identity and an E-value of 2e-13. A blastn of this specific sequence against the nr database and the human genomic database does not yield any significant hits, hinting at nothing about this potential element's origin.

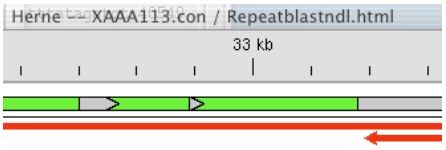
```
Score = 50.7 bits (26), Expect = 0.22
Identities = 36/41 (87%)
 Strand = Plus / Minus

Query: 31521 aaagtattggaaagaaaactagcttgcttagctttgacaag 31561
              ||| ||||||| || ||| ||||||||||||||||| |||||
Sbjct: 31454 aaattattggtaataaacctagcttgcttagctttaacaag 31414
```

Fig 18. Potential novel repeat element.

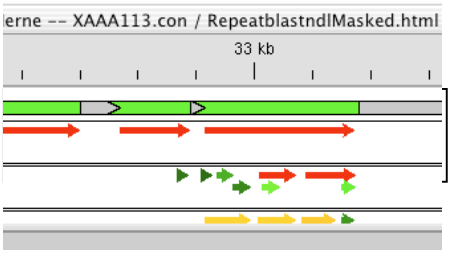
```
Score = 79.5 bits (41), Expect = 2e-13
Identities = 41/41 (100%)
 Strand = Plus / Plus

Query: 31521 aaagtattggaaagaaaactagcttgcttagctttgacaag 31561
              |||||||||||||||||||||||||||||||||||||||||
Sbjct: 1     aaagtattggaaagaaaactagcttgcttagctttgacaag 41

Score = 50.7 bits (26), Expect = 1e-04
Identities = 36/41 (87%)
 Strand = Plus / Minus
```

Though the *littoralis* fosmid does not contain any other likely novel repeat elements, comparison of blastn queries of XAAA113 against the library of masked *littoralis* fosmids a of unmasked fosmids reveals sequence pre ds not masked by Repeatmasker. For example from ~32736-33352 has several hits to other masked *littoralis* fosmids (Figure 20). However, a known PENELOPE LINE element begins at base 33357. A blastn search against the nt database does in fact yield matches indicating that this potential novel repeat is merely part of the same PENELOPE LINE element. Alignment of the potential novel repeat region (base 32736-33352) occurs to bases 2892-3013 of the D. lummei clone L6 transposon Penelope gene sequence at 32737-32854. The adjacent repetitive element masked by RepeatMasker also aligns to the D. lummei clone L6 transposon sequence, indicating that the potential novel repeat region merely escaped masking by RepeatMasker.

Fig 19. Potential novel repeat element.



One additional region of the XAAA113 contig (that does not also align to Drosophila EST's) aligns to other *littoralis* fosmids (repeat #7). Parts of the region of XAAA113 from roughly 4121-4207 align to four other contigs (31,91, 103, and 106), matching two distinct locations of contig 103. This region is only 15 bp away from a simple (TA)n dinucleotide run. Still, the region may be a potential novel repetitive element. Blast2alignment of the region's sequence against XAAA113 indicates that the potential element is present only once in XAAA113. A blastn search against the nt database of this potential element's sequence gives no clues about its origin or composition.



Further search for novel repetive elements by blastn query of XAAA113 against *D. melanogaster, D. pseudoobscura,* and *D. yakuba* datbases comes up empty-handed. Blastn alignment of XAAA113 against these species does not uncover any potential novel repeats, although the alignment against the *D. pseudoobscura* database shows that the exon and possible promoter regions of the *littoralis* CG2052 homologue share at least 71% identity with *D. pseudoobscura.* Alignment of these CG2052 homologue regions all occur to the same *D. pseudoobscura* contig (Contig3205_Contig3739). Small areas across this entire contig also align to several noncoding regions of XAAA113. These alignments are thus syntenic, suggesting that the region of DNA spanning these overlaps (at least bases 216-18874 of XAAA113), and the CG2052 feature it includes, predates the split between D. *littoralis* and *D. pseudoobscura.* (Notice that this *D. pseudoobscura* contig was erroneously not the one used in Clustal analysis of CG2052, suggesting that either the promoter is not specific to CG2052, or that there are other CG2052-like genes.)

Blastn alignment of XAAA113 against the *D. melanogaster* genomic database also does not uncover potential novel repeats though it does give clues about XAAA113 synteny with *D. melanogaster* chromosome 4.

XAAA113/Chromosome 4 Synteny

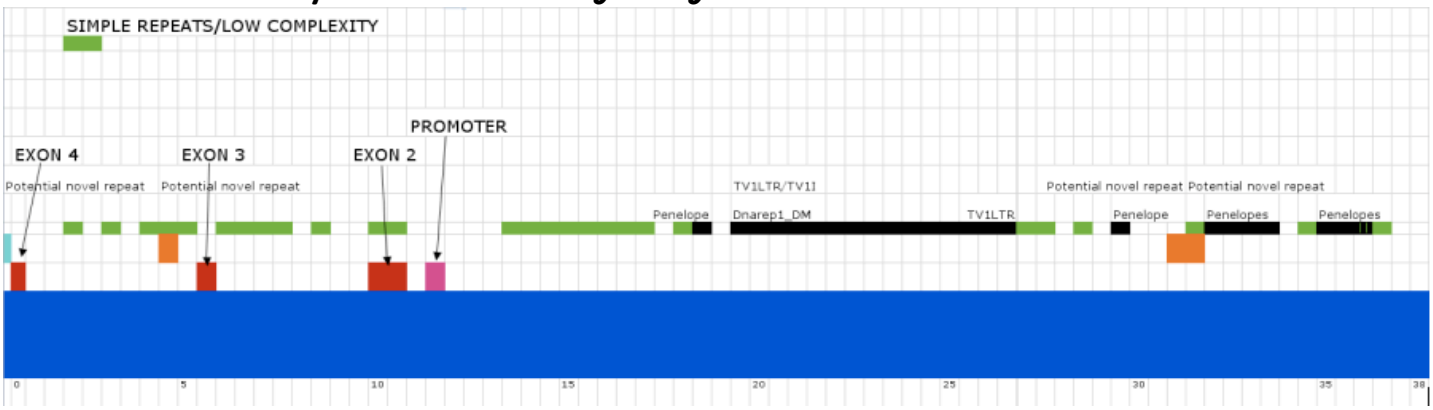


Fig 21. XAAA113 map.

In D. *melanogaster*, the CG2052 gene exists on the fourth chromosome. As expected, Blat alignment of the first 25,000 bases of XAAA113 (containing a CG2052 homologue) occurs to the fourth chromosome of D. *melanogaster*. This alignment does not occur for all 25,000 bases of the XAAA113 extract, however, but instead solely to bases 215—5292, 18790-18873, and 18828-18882. Additionally, blastn alignment of XAAA113 against the nt database yields hits to *melanogaster*, but not beyond base 15011 of the fosmid. (Matches occur to chromosome 4 at base 215-384 with 89% identity, 5183-5319 with 90% identity, 11899-11986 with 88% identity, and at 14937-15011 with 89% identity.) Blastn alignment of XAAA113 against a *D. melanogaster* genomic database yields similar results, though matches

of this blast search best support that XAAA113 and chromosome 4 are syntenic up to at least base 16158 of the littoralis fosmid. Small alignments of chromosome 4 occur to noncoding, nonrepetitive regions of XAAA113 with synteny up to this point. Beyond base 16158, however, matches to chromosome 4 only occur to repetitious areas of XAAA113.

This complete dropoff in XAAA113 similarity to chromosome 4 may merely reflect an absence of coding elements to drive conservation of sequence in this region. As shown previously, there is about a 40kb-long region upstream of CG2052 that contains no genes on the *D. melanogaster* chromosome (Figure 17 – reprinted for convenience). This empty region seems to exist in *littoralis* as well. Thus, lack of functional sequence to suppress mutation may explain the divergence of the XAAA113 fosmid from chromosome four beyond base 16158.

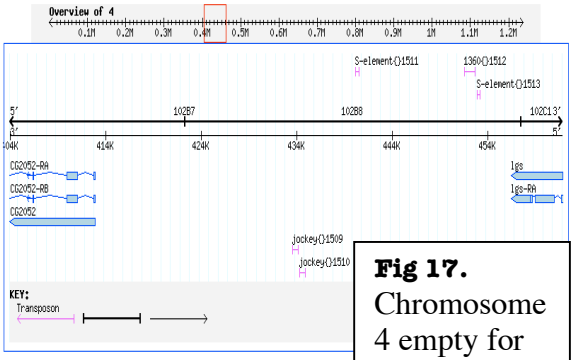


Fig 17. Chromosome 4 empty for >40 kb

Blat alignment of bases 25,001-37,265 of XAAA113 actually occurs as well chromosome 4 as well, although alignment covers only a small fraction of the XAAA113 extracted sequence. Bases 35295-36087, 35500-37265, and 35387-37265 match to chromosome 4. However, all of these regions span repetitive elements. One cannot definitively rule out that the second portion of the XAAA113 contig aligns better to a *D. melanogaster* chromosome other than the fourth. In situ hybridization experiments would help determine where this portion of XAAA113 aligns.