Mindy Tittiger

Dr. Elgin

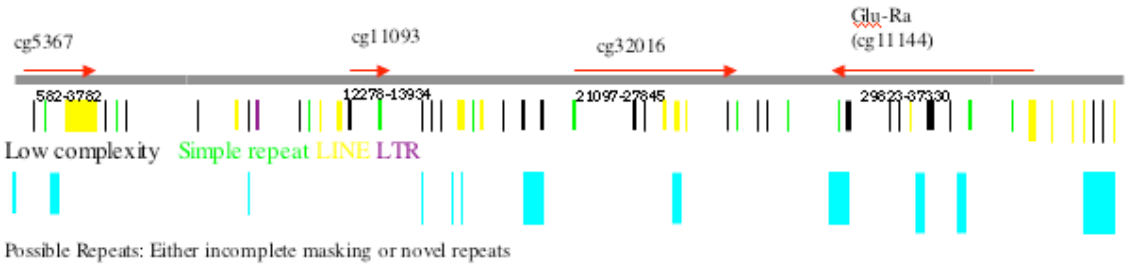
Biol 4342

XAAA103 Annotation

Overview

In-situ hybridization places fosmid XAAA103 on the dot chromosome of D. littoralis. This fosmid, 39850bp in length, contains 4 features: homologous to part of gene CG5367 from chromosome 2L of D. melanogaster; and genes CG 11093, CG 32016, and CG 11144 from chromosome 4 of D. melanogaster. This is evidence of chromosomal rearrangement. While the order of the three 4th chromosome genes is conserved in D. littoralis, the exon boundaries are not always conserved. The fosmid contains several repetitive elements: LINEs, simple repeats, and one LTR. A blast of fosmid XAAA103 to other D. littoralis sequence revealed possible novel repeats as well as incomplete masking of repetitive elements.

D. littoralis fosmid



D. melanogaster

Genes

Fosmid XAAA103 contains four features. Protein coding sequence match provides evidence for the genes. Nucleotide sequence is not conserved between D. Melanogaster and D. littoralis, so no mRNA evidence exists for UTRs. Only coding exon evidence exists.

CG5367 Except for the first 42 amino acids, the CG5367 protein matches the entire length (amino acid 43 to 338) to the fosmid. CG5367 covers 5 exons of coding sequence on chromosome arm 2L in D. melanogaster. My fosmid contains 4 exons of coding sequence for the gene. In order to remain in frame and to exclude a stop codon, exon 3 does not include matching amino acids 261 to 273.

CG5367 Exon Boundaries in Fosmid XAAA103

Exon	CG5367 Start	End
1	582	734
2	1038	1424
3	2434	2553
4	3591	3782

The coding sequence for the missing 42 amino acids likely extends beyond the length of my fosmid--as the gene is located at the beginning of the sequence. Exon boundaries and sizes are not conserved between D. melanogaster and D. littoralis. The table below shows what CG5367 amino acids are contained in each exon of both species. In D. littoralis, exon 3 contains additional amino acids not found in the CG5267 protein.

Exon	CG5367 D. melanogaster	D. littoralis
1	1-64aa	43-91aa
2	65-117aa	92-220aa
3	118-246aa	221-273aa
4	247-297aa	276-338aa
5	298-338aa	

This gene sequence contains a peptidase_c1 domain and is involved cathepsin L activity—"catalysis of the hydrolysis of peptide bonds" (Flybase).

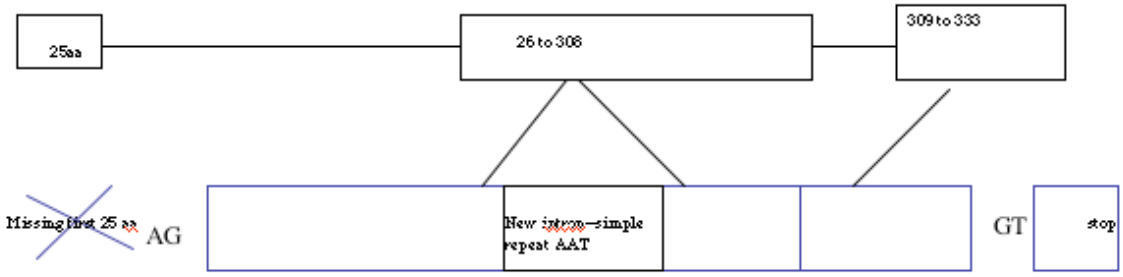
CG11093 The second feature is similar to CG11093 in D. melanogaster. Base positions 12778 to 13802 of my fosmid contain coding sequence for amino acids 26 to 316 of the CG11093 protein sequence (333 amino acids in length).

CG11093 Exon Boundaries in Fosmid XAAA103

Exon	CG11093 Start	End
1	12778	13389
2	13572	13802
3	13872	13934

The gene "encodes a product with putative transcription regulator activity" (Flybase). The transforming protein ski domain is only partially conserved in the fosmid. The first exon in D. melanogaster contains the first 25 amino acids. I could not locate this exon, or the start codon, in my fosmid. This may be evidence that the first exon was lost in the chromosomal rearrangement. The remaining amino acid sequence is conserved, though the exon boundaries are not. A simple repeat (AAT) covers the length of a new intron between coding exon 1 and exon 2 (splitting coding exon2 in D. melanogaster).

CG11093: Changes in Exon Boundaries Between D. melanogaster and D. littoralis



Exon1: 12778-13389 Exon2: 13572-13802 Exon3:13872-13934

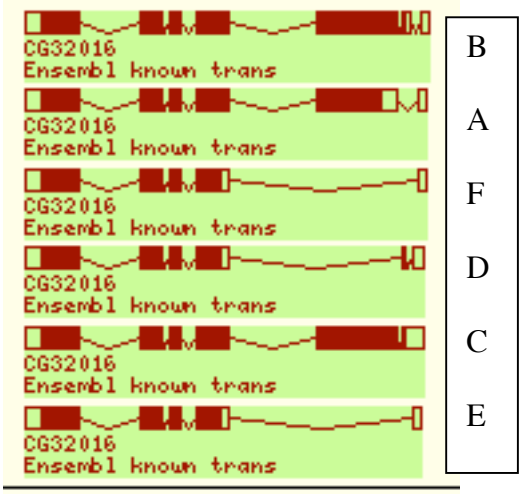
The following table shows only what amino acids from the CG 11093 protein are present in each exon. Exon 2 in D. littoralis ends with 20 additional amino acids not found in the D. melanogaster protein.

	CG11093	
Exon	D. melanogaster	D. littoralis
1	1-25aa	26-229aa
2	26-308aa	242-316aa
3	309-333aa	

It is not surprising that a feature homologous to this gene is contained within my fosmid; a blast to the probe containing part of the CG11093 sequence returned a high E-value match.

CG32016 CG32016 has 6 associated protein forms—a result of alternative splicing and the use of alternative start codons in D. melanogaster. The protein products are putatively involved in cell communication. The coding sequence for all forms are present in my fosmid. Amino acids 216 to 233, of the gene in D. melanogaster, contain a pfscan domain.

Structure of the six forms of CG32016



Protein form B covers every coding exon of the gene in D. melanogaster. Forms A and B differ only in the length of the first exon amino acid sequence in my fosmid: protein A matches from base position 21319 to 22581, whereas protein B matches from 21094 to 22581. This indicates to me that the proteins translate from different start codons—just as they do in D. melanogaster. However, I could not locate a second start near 21319. Forms E and F also appear to utilize a different start codon.

The following table shows that forms A and B are contained within the same coding exons and the same reading frame, but use different start codons. Exon 1 contains the first 449 amino acids of form B, and only the first 370 amino acids of form A. The table also shows conserved protein sequence corresponding to exons in both D. melanogaster and D. littoralis. Exons contain coding sequence for approximately the same amino acids, but boundaries are not entirely conserved.

CG32016 Conserved Protein Sequence Forms A and B

Exon	A-D. melanogaster	A-D. littoralis	B- D. melano	B-D. littoralis
1	1-370aa	1-370aa	1-9aa	9-457aa
2	371-415aa	376-414aa	10-458aa	463-501aa
3	416-540aa	415-520aa	459-503aa	502-607aa
4	541-609aa	553-582aa	504-628aa	640-669aa
5	610-738aa	608-737aa*	629-697aa	695-824aa*
6	739-846aa	734-845aa	698-826aa	821-932aa
7	847-923aa	845-923aa	827-934aa	932-1010aa
8			935-1010aa	

*boundaries undefined

Exon boundaries corresponding to areas of protein match in the appropriate frame were difficult to find. As a result, the exons include coding sequence for additional amino acids not found in the D. melanogaster gene. The exons also exclude some conserved amino acids: exon1 excludes 3 amino acids at the end, exon 2 excludes 4 amino acids at the beginning, and exon 4 excludes 10 amino acids at the beginning. I could not determine boundaries for exon 5—the sequence length listed corresponds to protein match. Exons 3 and 6 contain a stop in D. littoralis. If the predicted boundaries are correct, this feature may be a pseudogene.

CG32106 Exon boundaries in XAAA103

	CG32016	
Exon	Start	End
1	21020	22572
2	24305	24436
3	25609	26094
4	26354	26440
5	26677	27099
6	27167	27628
7	27609	27848

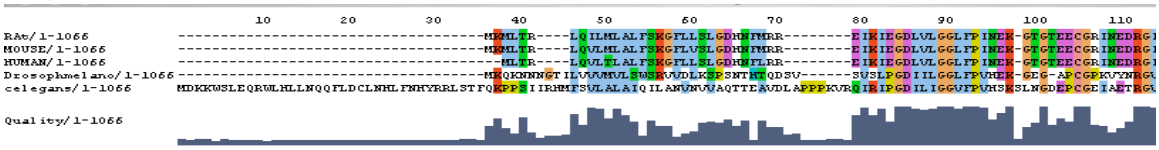
CG11144 My fosmid contains coding sequence for the full length (976aa) of the CG11144 associated protein starting with amino acid 11. The gene is also known as Glu-ra, which is involved in metabotropic glutamate receptor-like activity. Except for a combining of coding exons 1 and 2 in D. melanogaster, exons are conserved for this gene in D. littoralis. In D. melanogaster, this gene is translated off of the opposite strand; the same is true for the feature in D. littoralis. I searched for the reverse compliments of the normal AG and GT to find the exon boundaries for this feature. A clustal analysis of this protein compared the Glu-ra proteins in M. musculus, H. sapiens, and C. elegans revealed strong conservation except for the ends. So it is not surprising that the first 10 amino acids are not conserved between D. melanogaster and D. littoralis.

CG11144 Exon Boundaries in Fosmid XAAA103

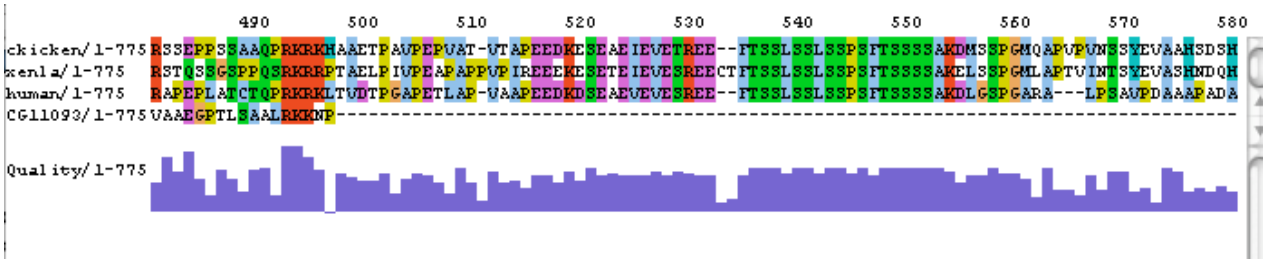
	CG11144	
Exon	Start	End
1	37330	36083
2	34893	34624
3	34531	34448
4	33526	33401
5	33341	32769
6	32033	31695
7	30161	29811

Clustal Analyses

A clustal analysis was performed on Glu-ra. The Glu-ra protein was highly conserved in C. elegans, H. sapiens, and M. musculus. The entire length was conserved except for the beginning and end sequences.



In a blast to the swissprot database, CG11093 hit to a transforming ski protein that is similar to only a portion of the entire protein. In a comparison of the Anopholes protein to the D. melanogaster protein, only amino acids 30 to 214 (out of 333) matched. A clustal was not useful in providing any information about conservation in the remaining protein sequence.



Blast did not reveal any protein with similarity to the CG32016 protein. A clustal of upstream sequence from CG11093 and the ski gene from humans and mice did not reveal any putative promoter regions—sporadic sequence match covered no more than 3 bases.

Repeats

The fosmid sequence contains several LINEs and simple repeats as well as low complexity sequence that comprise 14.55% of the fosmid length. One LTR is present in the sequence between the chromosome 2L gene and the 4th chromosome genes. See the attached RepeatMasker table for locations of each type of repeat. A blast of XAAA103 to other D. littoralis fosmids revealed possibly novel repeats as well as incomplete masking of sequence. Incomplete masking appeared in two ways: (1) when my unmasked fosmid sequence matched to already masked sequence in another fosmid, and (2) when unmasked sequence in both fosmids matched immediately next to matching masked

sequence. Possible new repeats are not located next to any masked sequence. Some of these possibly novel repeats match well to D. virilis sequence. The repeats may be specific to these species. Other possible repeats show partial match to distant species, such as H. sapiens. These may be repeats specific to D. littoralis.

Possible Repeats

XAA103	Blast hit to	Hit length	Length of sequence
1 to 75	D. virilis(Lpg gene) glue protein	57	76
1506 to 1818	D. virilis(Lpg gene) glue protein	159	313
1481 to 1548	D. virilis clone	35	68
1611 to 2113	D. virilis clone	361	503
11499 to 11553	Human DNA clone	20	55
16993 to 17118	Stronglocentrotus PM27 gene	29	126
17254 to 17405	Mus musculus chrm 5 clone	20	152
17825 to 17958	Oryza sativa chrm 12 BAC	21	134
30182 to 30388	D. virilis clone	169	207
30521 to 30809	D. virilis(Lpg gene) glue protein	195	289
31102 to 31182	D. lummei part of L6 transposon	54	81

Incomplete masking

1950-2136	17539-17799	35868-35909
2115-2180	24433-24609	38615-38680
15387-15452	24742-24762	38730-38757
16223-16373	24901-24940	38771-39216
16385-16420	30462-30515	39443-39517
16817-16869	35395-35424	39545-39578

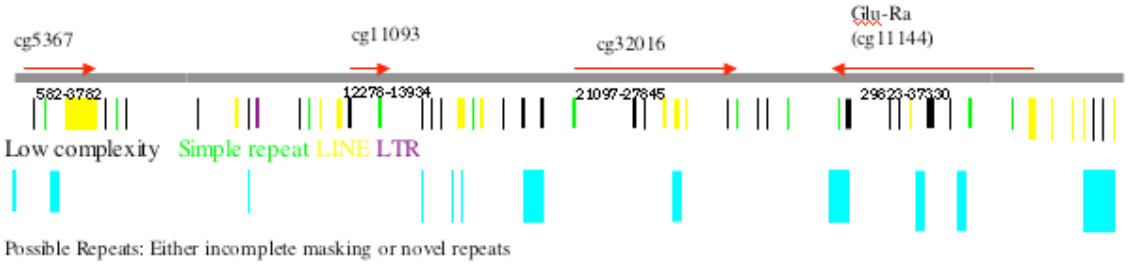
Including possible new repeats and incompletely masked repeats, 24%(9634bp) of my fosmid is repetitive. No repeats are present in coding sequence, only gene introns.

Synteny

Fosmid XAAA103 contains three features homologous to genes on the 4th chromosome of D. melanogaster. The order of the genes is conserved between species. The approximate genomic length of the genes and spacing between the genes is similar. The fosmid also contains a feature homologous to a gene on chromosome 2L of D.

melanogaster. Thought the order of genes is conserved between species, the coding exon boundaries are not always conserved. I have no evidence for conservation of repetitive sequence between the two species. Further inquiry may reveal repetitive sequence surrounding these genes in D. melanogaster. There is no evidence of transposable elements in my D. littoralis sequence.

XAAA103



D. melanogaster

Repeat Masked Sequence

LTR	low complexity	LINE	simple repeat
9389-9538	344-380	2668-3105	547-575
	3271-3314	8450-8573	3555-3582
	4401-4476	11712-11780	11113-11166
	7240-7302	11980-12107	13399-13561
	8802-8822	16421-16688	16689-16727
	10794-10830	16728-16816	17149-17179
	12559-12675	17422-17517	20975-21034
	14884-14921	23927-24078	26573-26617
	15472-15498	24610-24741	29329-29365
	15569-15609	24763-24900	30810-30836
	18165-18189	25114-25235	35425-35595
	18861-18919	31229-31497	35574-35595
	19830-19888	32465-32540	37502-37526
	22877-22901	33556-33977	
	23539-23566	35596-35867	
	26231-26253	37819-38582	
	27555-27590	38681-38729	
	28119-28153	39217-39345	
	32095-32123	39350-39442	
	32158-32190	39731-39843	
	33987-34026		
	34404-34446		
	39518-39545		
	39666-39693		
150bp	992bp	4095bp	560bp
			5797bp total=14.55%