FINAL ANNOTATION REPORT:

Drosophila virilis Fosmid 11
(48P14)

Robert Carrasquillo
Bio 4342

2006

TABLE OF CONTENTS

I. Overview

Genscan initially predicted five gene-like features for my assigned *Drosophila virilis* fosmid, fosmid 11. Using BLASTNnt, BLASTPnr, and the UCSC Genome Browser programs, I concluded that the five predicted features corresponded to orthologs of the following five *Drosophila melanogaster* genes: CG11155, Kif3C (CG17461), *pho* (*pleiohomeotic*, CG17743), RpS3A (ribosomal protein S3A, CG2168), and *pan* (*pangolin*, CG17964). The orthologs occurred in that specific order in my fosmid and both CG11155 and *pan* had partial 3' ends that extended beyond the ends of my fosmid. Using TBLASTXnt and TBLASTNnr algorithms to search my predicted CDS sequences and peptide sequences against *D. melanogaster* databases in Flybase, as well as visualization of Genscan and Twinscan gene predictions in the Goose-UCSC Genome Browser, I was able to assign coordinates to start codons, intron-exon boundaries, and stop codons. Gene models were assigned based on the longest isoforms of these genes, those with the most exons.

In addition to gene models, I used ClustalW to align peptide sequences for RpS3A among *D. virilis* and five other species found on the Ensembl website. I also aligned the RpS3A upstream regulatory regions using the first exon as an alignment anchor, and searched for potential regulatory elements. These analyses showed remarkable conservation of the coding sequences but scant conservation in the upstream region. Repeat analysis was conducted in conjunction with output from RepeatMasker. Using BLASTN, I searched for my masked fosmid sequenced against a database of all fosmids. Several significant hits came up that showed consistent alignment to the same four regions and these were resolved by searching for their sequences in a database of known *D. virilis* repeats. Lastly, I determined the relative state of synteny of my fosmid in relation to the *D. melanogaster* dot (Chromosome IV) using the UCSC Genome Browser. This revealed synteny to Chromosome IV for all genes annotated, however there were several large-scale rearrangments.

Figure 1a. Initial Genscan predictions for gene-like features

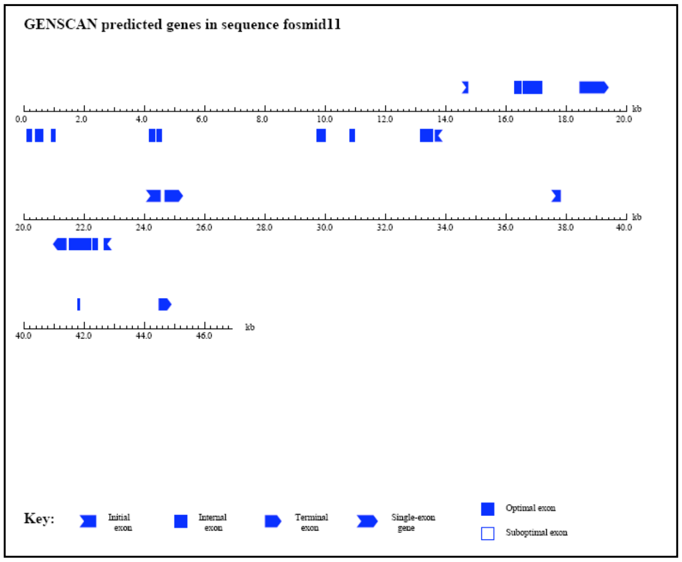
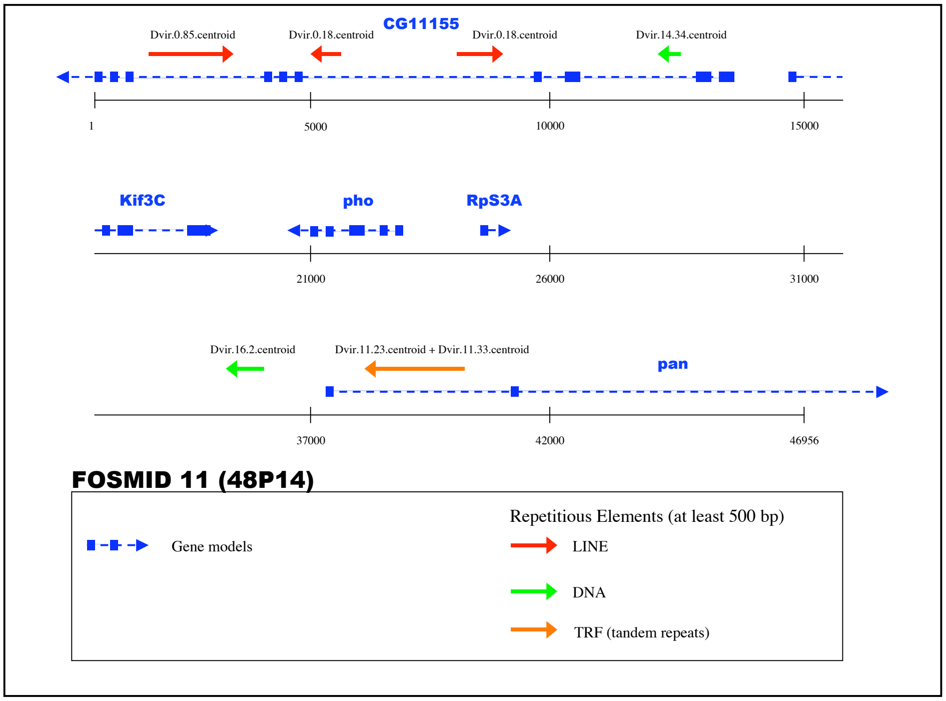


Figure 1b. Final map of Fosmid 11 (48P14) with genes and large repeats



II. Genes

The following list includes the five numbered gene features that I will discuss, the names of their orthologs in *D. melanogaster* with specification of the isoforms used to establish each gene model, and their Genbank accession numbers:

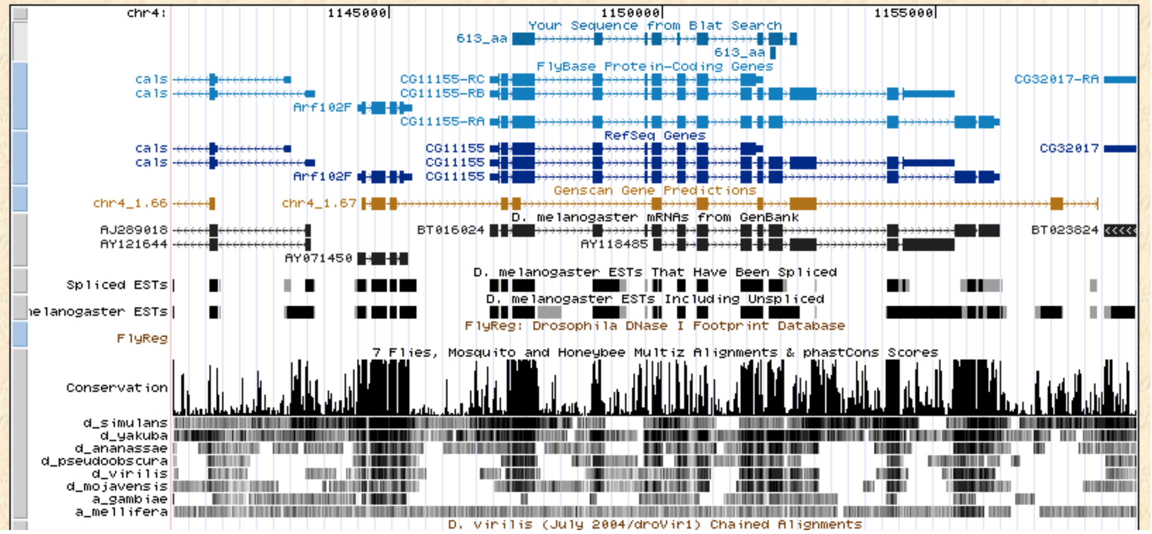
Feature 1:	**CG11155-RA**	**NM_143684**
Feature 2:	**CG17461**	**NM_143682**
Feature 3:	**CG17743-RA**	**NM_079891**
Feature 4:	**CG2168-RA**	**NM_079879**
Feature 5:	**CG17964-RG**	**NM_166723**

I will give an overview on how each gene model was created and detail specific problems I encountered and how they were resolved.

Feature 1 (CG11155-RA, NM_143684):

Using BLASTNnt, I searched using the predicted CDS for feature 1 as a query. The best hits (E values 1e-87 to 5e-22) included Refseqs for three isoforms of *D. melanogaster* CG11155. BLASTPnr with the predicted peptide sequence showed the same hits, this time to the three protein isoforms, with E values of 0.0 for each. Using UCSC *D. melanogaster* BLAT, I found the peptide sequence to align fairly well to *D. melanogaster* CG11155:

Figure 2. Alignment of Feature 1 using UCSC BLAT

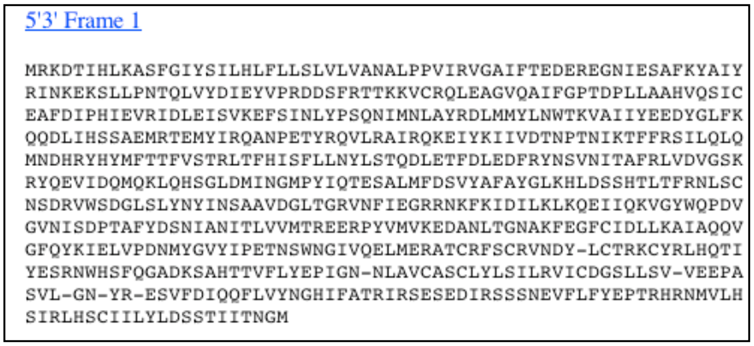


Based on this BLAT output, I did not have alignment to the last exon, which I predicted to extend past my fosmid end. Using TBLASTX2seq, with the *D. melanogaster* exons from Ensembl against a database of an extracted region of the fosmid including Feature

1, I determined the alignment and sequence similarity of each exon between both species. *D. melanogaster* exons 1, 2, 7, 12, 13, and 14 of 14, failed to align well. The first two contain UTR sequences that may be highly diverged and 12-14 may exist off the front end of my fosmid, since the gene occurs in the reverse orientation in the Goose Browser. For exon 7, I used TBLASTX to search for the exon 7 sequence across an extracted region in which it should lie in the fosmid. Alignment did occur in two separate pieces, so it is likely that the sequence may have diverged.

Using the Goose Browser, I zoomed in on the alignment and sequence information to establish the gene model with start and stop codons, and intron-exon boundaries (introns starting GT, ending AG). All of these features were found with certainty except for a proper splice donor at the end of what was the seventh exon in *D. virilis*. I assigned a donor, GT, based on the canonical model, however upon further inspection, I found this to not be the case. A translation of the concatenated exons using the canonical splice donor gave the following:

Figure 3. Error in Feature 1 translation (Stop: -)

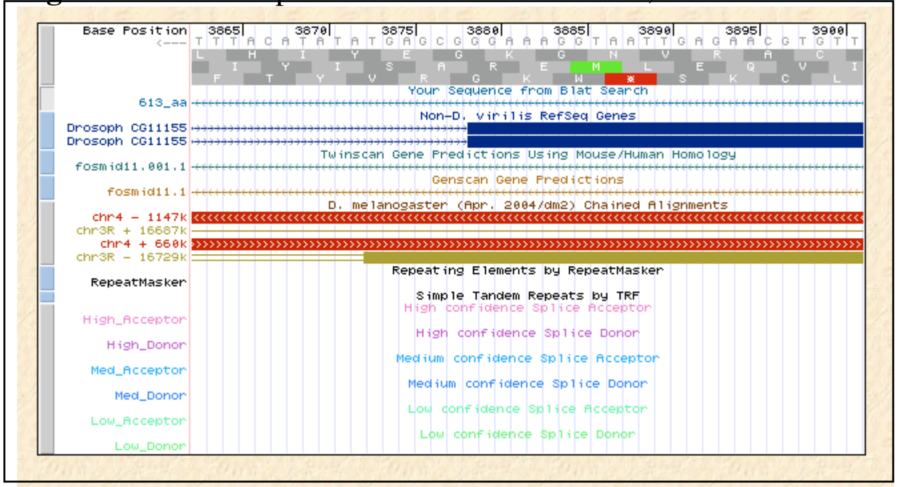


5'3' Frame 1

```
MRKDTIHLKASFGIYSILHLFLLSLVLVANALPPVIRVGAIFTEDEREGNIESAFKYAIY
RINKEKSLLPNTQLVYDIEYVPRDDSFRTTKKVCRQLEAGVQAIFGPTDPLLAAHVQSIC
EAFDIPHIEVRIDLEISVKEFSINLYPSQNIMNLAYRDLMMYLNWTKVAIIYEEDYGLFK
QQDLIHSSAEMRTEMYIRQANPETYRQVLRAIRQKEIYKIIVDTNPTNIKTFFRSILQLQ
MNDHRYHYMFTTFVSTRLTFHISFLLNYLSTQDLETFDLEDFRYNSVNITAFRLVDVGSK
RYQEVIDQMQKLQHSGLDMINGMPYIQTESALMFDSVYAFAYGLKHLDSSHTLTFRNLSC
NSDRVWSDGLSLYNYINSAAVDGLTGRVNFIEGRRNKFKIDILKLKQEIIQKVGYWQPDV
GVNISDPTAFYDSNIANITLVVMTREERPYVMVKEDANLTGNAKFEGFCIDLLKAIAQQV
GFQYKIELVPDNMYGVYIPETNSWNGIVQELMERATCRFSCRVNDY-LCTRKCYRLHQTI
YESRNWHSFQGADKSAHTTVFLYEPIGN-NLAVCASCLYLSILRVICDGSLLSV-VEEPA
SVL-GN-YR-ESVFDIQQFLVYNGHIFATRIRSESEDIRSSSNEVFLFYEPTRHRNMVLH
SIRLHSCIILYLDSSTIITNGM
```

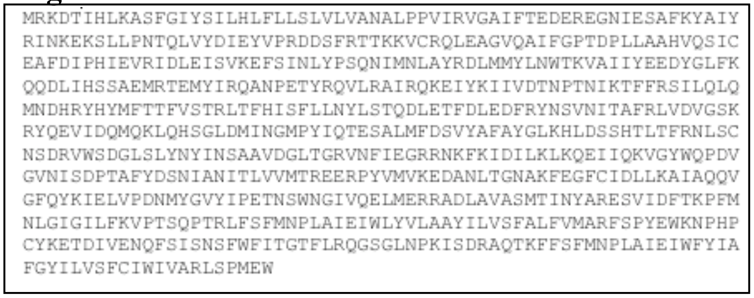
Note the existence of premature stops denoted by hyphens. I made sure that all frames were capable of being read, also accounting for changes in phase. Only one frame was possible for this exon to avoid premature stops. Using the Goose Browser, I found what I thought could be a miscalled base in my consensus, as alignment to *D. melanogaster* Refseqs showed a splice donor site a few bases upstream at GC.

Figure 4. Unusual splice donor GC for Feature 1, exon 7


Base Position
3865
3870
3875
3880
3885
3890
3895
3900
<---
T T T A C A T A T A T G A G C G G G A A A G G T A A T T G A G A A C G T G T T
Your Sequence from Blat Search
613_aa
Non-D. virilis RefSeq Genes
Drosoph CG11155
Drosoph CG11155
Twinscan Gene Predictions Using Mouse/Human Homology
fosmid11.001.1
Genscan Gene Predictions
fosmid11.1
D. melanogaster (Apr. 2004/dm2) Chained Alignments
chr4 - 1147k
chr3R + 16687k
chr4 + 660k
chr3R - 16729k
Repeating Elements by RepeatMasker
RepeatMasker
Simple Tandem Repeats by TRF
High confidence Splice Acceptor
High_Acceptor
High confidence Splice Donor
High_Donor
Medium confidence Splice Acceptor
Med_Acceptor
Medium confidence Splice Donor
Med_Donor
Low confidence Splice Acceptor
Low_Acceptor
Low confidence Splice Donor
Low_Donor

I adjusted the gene model to take that unusual splice donor into consideration and concatenation of the exons produced a reading frame that was read through completely:

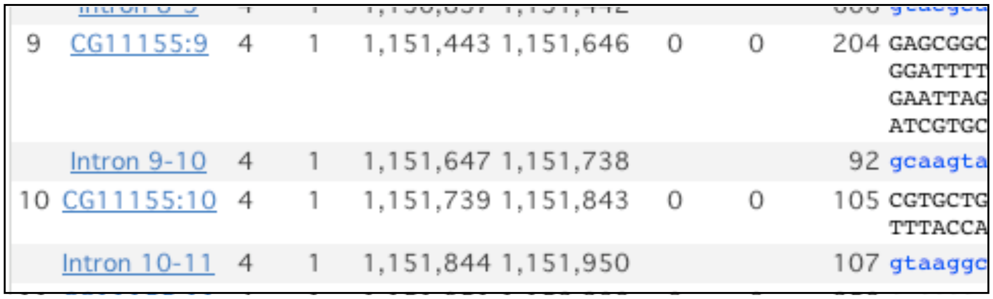
Figure 5. Corrected error in Feature 1 translation



```
MRKDTIHLKASFGIYSILHLFLLSLVLVANALPPVIRVGAIFTEDEREGNIESAFKYAIY
RINKEKSLLPNTQLVYDIEYVPRDDSFRTTKKVCRQLEAGVQAIFGPTDPLLAAHVQSIC
EAFDIPHIEVRIDLEISVKEFSINLYPSQNIMNLAYRDLMMYLNWTKVAIIYEEDYGLFK
QQDLIHSSAEMRTEMYIRQANPETYRQVLRAIRQKEIYKIIVDTNPTNIKTFFRSILQLQ
MNDHRYHYMFTTFVSTRLTFHISFLLNYLSTQDLETFDLEDFRYNSVNITAFRLVDVGSK
RYQEVIDQMQKLQHSGLDMINGMPYIQTESALMFDSVYAFAYGLKHLDSSHTLTFRNLSC
NSDRVWSDGLSLYNYINSAAVDGLTGRVNFIEGRRNKFKIDILKLKQEIIQKVGYWQPDV
GVNISDPTAFYDSNIANITLVVMTREERPYVMVKEDANLTGNAKFEGFCIDLLKAIAQQV
GFQYKIELVPDNMYGVYIPETNSWNGIVQELMERRADLAVASMTINYARESVIDFTKPFM
NLGIGILFKVPTSQPTRLFSFMNPLAIEIWLYVLAAYILVSFALFVMARFSPYEWKNPHP
CYKETDIVENQFSISNSFWFITGTFLRQGSGLNPKISDRAQTKFFSFMNPLAIEIWFYIA
FGYILVSFCIWIVARLSPMEW
```

To be sure of the unusual splice donor, I inspected the exon information in Ensembl for this site in *D. melanogaster* and found there to be a conserved alternative splice donor, GC, at the intron 9-10 boundary as shown.

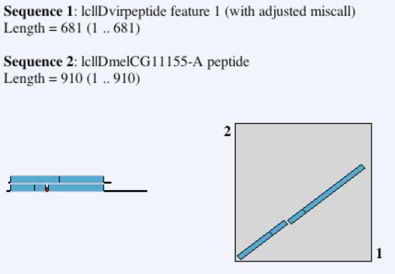
Figure 6a. Presence of alternative splice donor, GC, in *D. melanogaster*



9	CG11155:9	4	1	1,151,443	1,151,646	0	0	204	GAGCGGC GGATTTT GAATTAG ATCGTGC
	Intron 9-10	4	1	1,151,647	1,151,738			92	gcaagta
10	CG11155:10	4	1	1,151,739	1,151,843	0	0	105	CGTGCTG TTTACCA
	Intron 10-11	4	1	1,151,844	1,151,950			107	gtaaggc

Finally, after resolving all alignments, I found that two *D. melanogaster* exons, 5 and 6, are fused in *D. virilis* to make the fourth exon in my gene model. While exon 12 from *D. melanogaster* did not align well to the predicted exon, I found support for its existence from mRNA data, Genscan predictions, and Refseq data in the browser. When concatenated with all other confirmed exons and using BLASTP to search the concatenation against the true *D. melanogaster* peptide, I found that additional exon to be a part of the gene. This corresponds to exon 10 in my gene model. The other exons did not align at all and must exist beyond my fosmid end in *D. virilis*. Based on all this data, I resolved a gene model for the ortholog to *D. melanogaster* CG11155.

Figure 6b. Successful alignment of concatenated exons after splice site correction



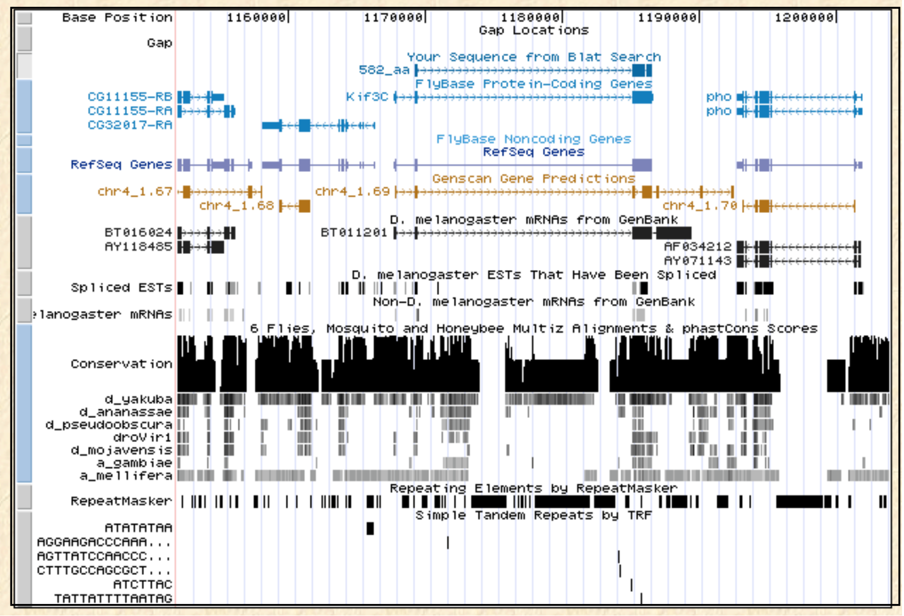
Sequence 1: lcl|Dvirpeptide feature 1 (with adjusted miscall)
Length = 681 (1 .. 681)

Sequence 2: lcl|DmelCG11155-A peptide
Length = 910 (1 .. 910)

Feature 2 (CG17461, NM_143682):

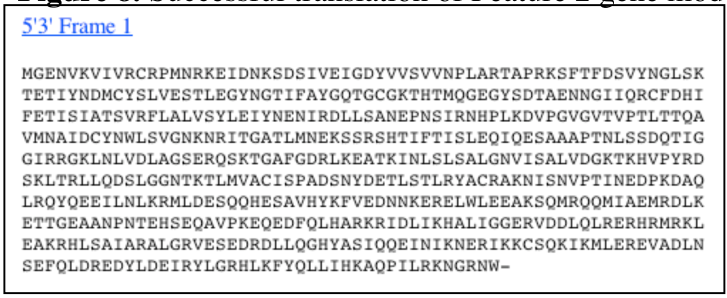
Using the same tools, I found BLAST alignments for the predicted CDS and peptide sequences to *D. melanogaster* KISc_KIF3 protein domain (BLASTNnt 1e-07, BLASTPnr 0.0). *D. melanogaster* BLAT output showed alignment to what is called Kif3C.

Figure 7. *D. melanogaster* BLAT output for Feature 2, alignment to Kif3C



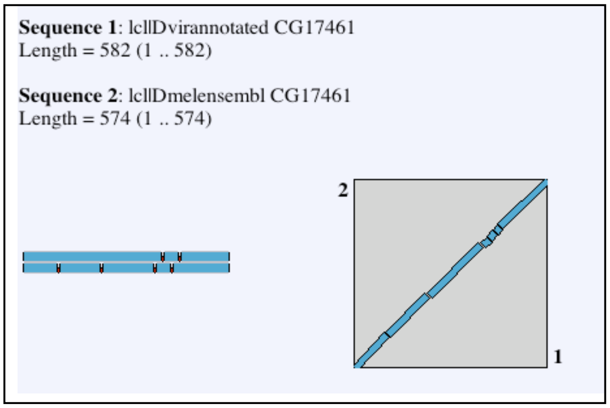
Using TBLASTX2seq with an extracted region including Feature 2 from the Goose Browser, and the *D. melanogaster* individual exons from Ensembl as queries, I found all four *D. melanogaster* exons to align well with some minor base additions apparent in *D. virilis*. Exon boundaries were well supported by Twinscan and Genscan predictions, and the concatenated exons translated perfectly and aligned completely to the *D. melanogaster* peptide (BLASTP2seq).

Figure 8. Successful translation of Feature 2 gene model


5'3' Frame 1

```
MGENVKVIVRCRPMNRKEIDNKSDSIVEIGDYVVSVVNPLARTAPRKSFTFDSVYNGLSK
TETIYNDMCYSLVESTLEGYNGTIFAYGQTGCGKTHTMQGEGYSDTAENNGIIQRCFDHI
FETISIATSVRFLALVSYLEIYNENIRDLLSANEPNSIRNHPLKDVPGVGVTVPTLTTQA
VMNAIDCYNWLSVGNKNRITGATLMNEKSSRSHTIFTISLEQIQESAAAPTNLSSDQTIG
GIRRGKLNLVDLAGSERQSKTGAFGDRLKEATKINLSLSALGNVISALVDGKTKHVPYRD
SKLTRLLQDSLGGNTKTLMVACISPADSNYDETLSTLRYACRAKNISNVPTINEDPKDAQ
LRQYQEEILNLKRMLDESQQHESAVHYKFVEDNNKERELWLEEAKSQMRQQMIAEMRDLK
ETTGEAANPNTEHSEQAVPKEQEDFQLHARKRIDLIKHALIGGERVDDLQLRERHRMRKL
EAKRHLSAIARALGRVESEDRDLLQGHYASIQQEINIKNERIKKCSQKIKMLEREVADLN
SEFQLDREDYLDEIRYLGRHLKFYQLLIHKAQPILRKNGRNW-
```

Figure 9. Excellent alignment of Feature 2 peptide and *D. melanogaster* peptide



Sequence 1: lcl|Dvirannotated CG17461
Length = 582 (1 .. 582)

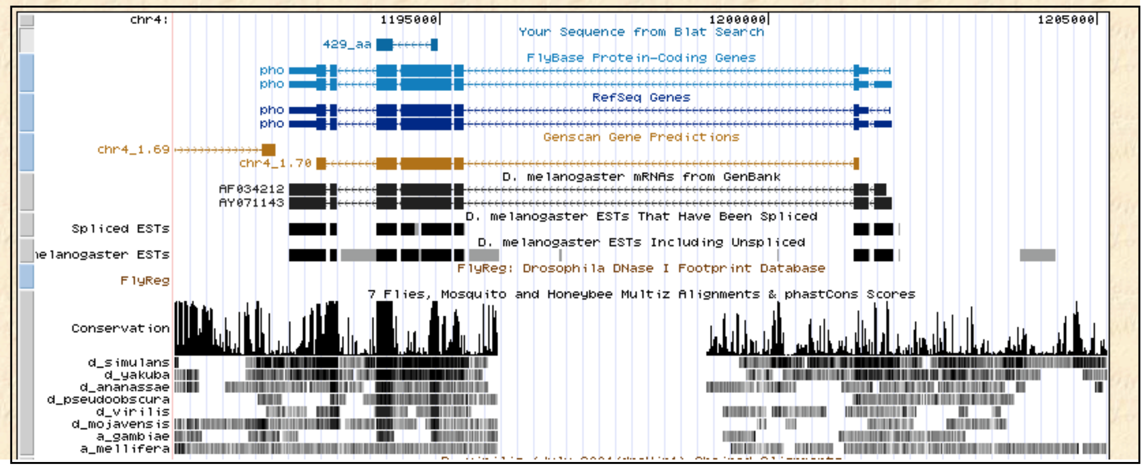
Sequence 2: lcl|Dmelensembl CG17461
Length = 574 (1 .. 574)

These results give me confidence in my gene model for Feature 2, the *D. virilis* ortholog to *D. melanogaster* Kif3C.

<u>Feature 3 (CG17743-RA, NM_079891):</u>

Again, with the same tools and method, I found BLASTNnt and BLASTPnr alignments to *D. melanogaster* Refseqs for *pleiohomeotic*, or *pho* (BLASTNnt 5e-15, BLASTPnr 3e-67). The following shows the UCSC *D. melanogaster* BLAT alignment.

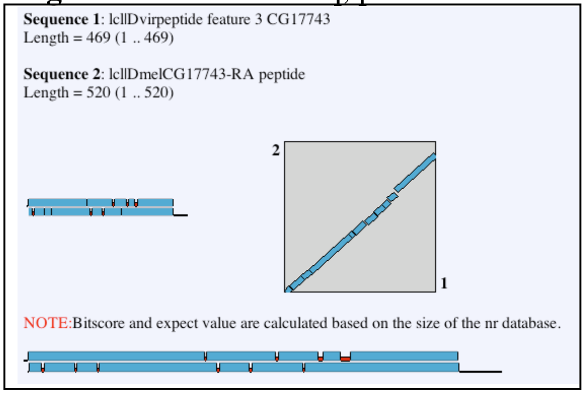
Figure 10. Alignment of Feature 3 in BLAT



chr4:
1195000
1200000
1205000
Your Sequence from Blat Search
429_aa
FlyBase Protein-Coding Genes
pho
pho
RefSeq Genes
pho
pho
Genscan Gene Predictions
chr4_1.69
chr4_1.70
D. melanogaster mRNAs from GenBank
AF034212
AY071143
D. melanogaster ESTs That Have Been Spliced
Spliced ESTs
D. melanogaster ESTs Including Unspliced
melanogaster ESTs
FlyReg: Drosophila DNase I Footprint Database
FlyReg
7 Flies, Mosquito and Honeybee Multiz Alignments & phastCons Scores
Conservation
d_simulans
d_yakuba
d_ananassae
d_pseudoobscura
d_virilis
d_mojavensis
a_gambiae
a_mellifera

In the Goose Browser (*D. virilis*), all exons except the last were predicted in my fosmid. The prediction was prematurely cut off as the browser interpreted the sequence at a splice donor to be a stop codon (3'-<u>AA</u>**TG**-5'). Still, using TBLASTX2seq in the same way described, I was able to confirm all exons from *D. melanogaster* in my fosmid with the exception of those that are most if not completely UTR. Concatenation of my initial gene model showed good translation (no premature stops), however the peptide sequence did not fully align to the *D. melanogaster* peptide, and it seemed as if I was missing another terminal exon.

Figure 11. BLASTP2seq, predicted Feature 3 peptide and *D. melanogaster* peptide



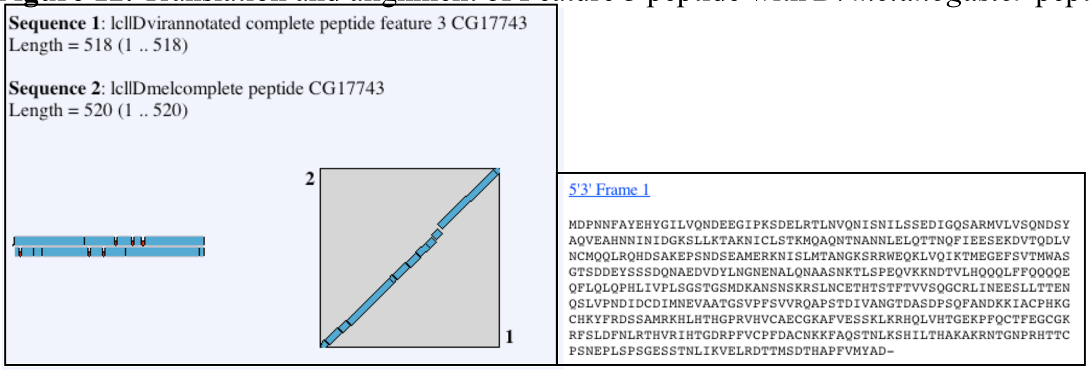
Sequence 1: lcl|Dvirpeptide feature 3 CG17743
Length = 469 (1 .. 469)

Sequence 2: lcl|DmelCG17743-RA peptide
Length = 520 (1 .. 520)

NOTE:Bitscore and expect value are calculated based on the size of the nr database.

Scanning downstream in the Goose Browser, I found a potential reading frame that seemed like a missing last exon. It ended appropriately in a stop codon. This sequence, which was not predicted as part of the gene, was appended. The exons were again concatenated and translated, and the predicted peptide sequence aligned with that of *D. melanogaster*. This time, alignment occurred in full, confirming the last missing exon, and improving my gene model. Again, this was another instance in which a splice donor was interpreted as a stop codon, based upon sequence.

Figure 12. Translation and alignment of Feature 3 peptide with *D. melanogaster* peptide



Sequence 1: lcl|Dvirannotated complete peptide feature 3 CG17743
Length = 518 (1 .. 518)

Sequence 2: lcl|Dmelcomplete peptide CG17743
Length = 520 (1 .. 520)

5'3' Frame 1

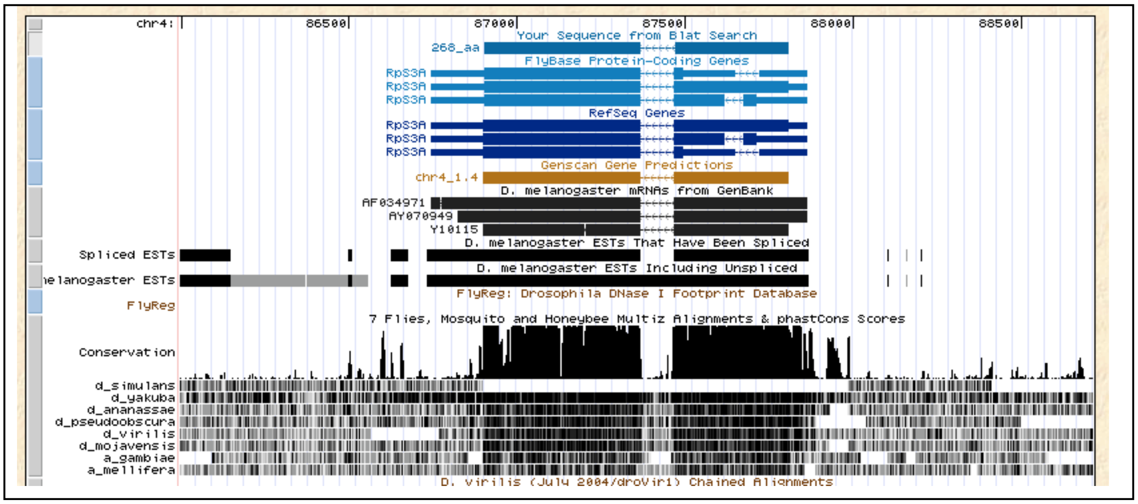
```
MDPNNFAYEHYGILVQNDEEGIPKSDELRTLNVQNISNILSSEDIGQSARMVLVSQNDSY
AQVEAHNNINIDGKSLLKTAKNICLSTKMQAQNTNANNLELQTTNQFIEESEKDVTQDLV
NCMQQLRQHDSAKEPSNDSEAMERKNISLMTANGKSRRWEQKLVQIKTMEGEFSVTMWAS
GTSDDEYSSSDQNAEDVDYLNGNENALQNAASNKTLSPEQVKKNDTVLHQQQLFFQQQQE
QFLQLQPHLIVPLSGSTGSMDKANSNSKRSLNCETHTSTFTVVSQGCRLINEESLLTTEN
QSLVPNDIDCDIMNEVAATGSVPFSVVRQAPSTDIVANGTDASDPSQFANDKKIACPHKG
CHKYFRDSSAMRKHLHTHGPRVHVCAECGKAFVESSKLKRHQLVHTGEKPFQCTFEGCGK
RFSLDFNLRTHVRIHTGDRPFVCPFDACNKKFAQSTNLKSHILTHAKAKRNTGNPRHTTC
PSNEPLSPSGESSTNLIKVELRDTTMSDTHAPFVMYAD-
```

This data confirms the gene model for the *D. virilis* ortholog to *D. melanogaster pho*.

Feature 4 (CG2168-RA, NM_079879):

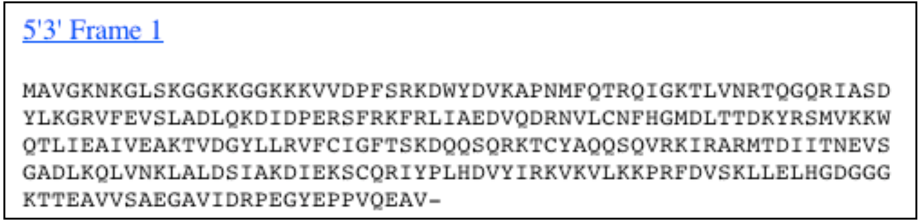
Using the same method and tools, I found highly significant BLASTNnt and BLASTPnr alignments of predicted Feature 4 CDS and peptide sequences to *D. melanogaster* ribosomal protein S3A (BLASTNnt 2e-87, BLASTPnr 6e-133). The following shows the UCSC *D. melanogaster* BLAT output:

Figure 13. Alignment of Feature 4 in UCSC *D. melanogaster* BLAT


chr4:
86500
87000
87500
88000
88500
Your Sequence from Blat Search
268_aa
FlyBase Protein-Coding Genes
RpS3A
RpS3A
RpS3A
RefSeq Genes
RpS3A
RpS3A
RpS3A
Genscan Gene Predictions
chr4_1.4
D. melanogaster mRNAs from GenBank
AF034971
AY070949
Y10115
D. melanogaster ESTs That Have Been Spliced
Spliced ESTs
D. melanogaster ESTs Including Unspliced
elanogaster ESTs
FlyReg: Drosophila DNase I Footprint Database
FlyReg
7 Flies, Mosquito and Honeybee Multiz Alignments & phastCons Scores
Conservation
d_simulans
d_yakuba
d_ananassae
d_pseudoobscura
d_virilis
d_mojavensis
a_gambiae
a_mellifera
D. virilis (July 2004/droVir1) Chained Alignments

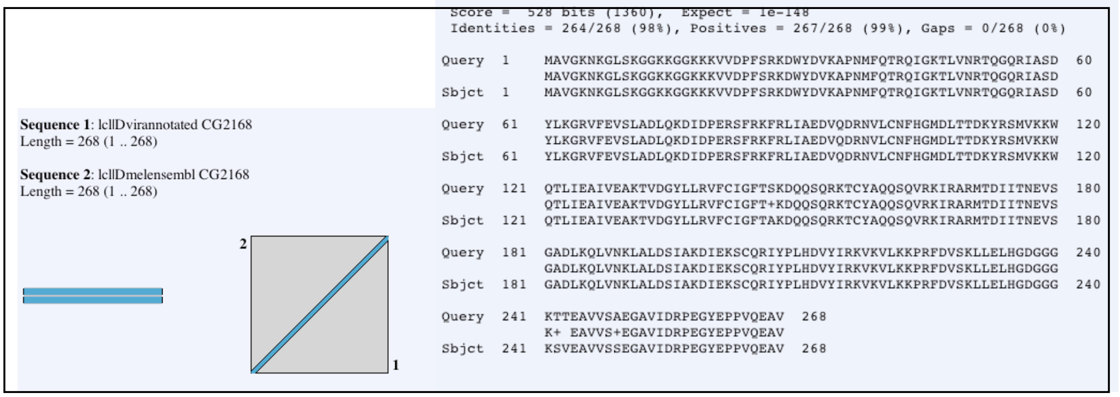
Considering what we know about the conservation of ribosomal RNA sequences and proteins, the BLAT output on conservation and alignment between *D. virilis* and *D. melanogaster* is not surprising. In the Goose Browser, I found it easy to construct an accurate gene model. Using BLASTP2seq, I aligned the concatenated and successfully translated exons to the longest protein isoform of *D. melanogaster* RpS3A from Ensembl. The result was a remarkable conservation, 98 percent amino acid identity and 99 percent positives.

Figure 14. Successful translation of Feature 4 concatenated exons (including stop codon)


5'3' Frame 1

```
MAVGKNKGLSKGGKKGGKKKVVDPFSRKDWYDVKAPNMFQTRQIGKTLVNRTQGQRIASD
YLKGRVFEVSLADLQKDIDPERSFRKFRLIAEDVQDRNVLCNFHGMDLTTDKYRSMVKKW
QTLIEAIVEAKTVDGYLLRVFCIGFTSKDQQSQRKTCYAQQSQVRKIRARMTDIITNEVS
GADLKQLVNKLALDSIAKDIEKSCQRIYPLHDVYIRKVKVLKKPRFDVSKLLELHGDGGG
KTTEAVVSAEGAVIDRPEGYEPPVQEAV-
```

Figure 15. High identity alignment for Feature 4 peptide and *D. melanogaster* peptide


Sequence 1: lcl|Dvirannotated CG2168
Length = 268 (1 .. 268)

Sequence 2: lcl|Dmelensembl CG2168
Length = 268 (1 .. 268)

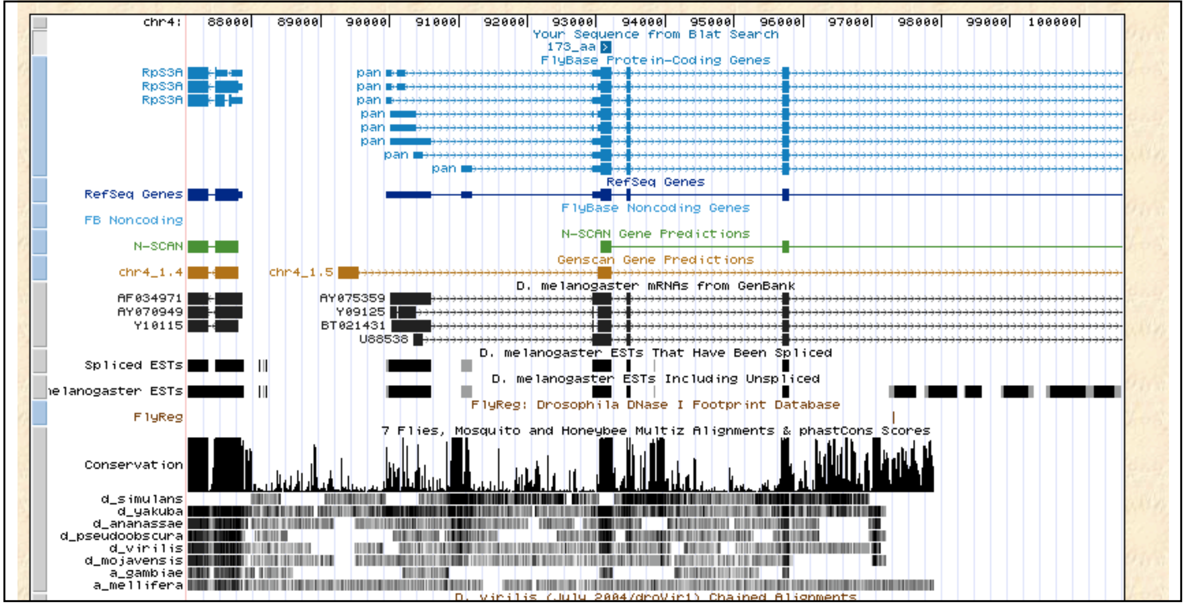
```
 Score =  528 bits (1360),  Expect = 1e-148
 Identities = 264/268 (98%), Positives = 267/268 (99%), Gaps = 0/268 (0%)

Query  1    MAVGKNKGLSKGGKKGGKKKVVDPFSRKDWYDVKAPNMFQTRQIGKTLVNRTQGQRIASD  60
            MAVGKNKGLSKGGKKGGKKKVVDPFSRKDWYDVKAPNMFQTRQIGKTLVNRTQGQRIASD
Sbjct  1    MAVGKNKGLSKGGKKGGKKKVVDPFSRKDWYDVKAPNMFQTRQIGKTLVNRTQGQRIASD  60

Query  61   YLKGRVFEVSLADLQKDIDPERSFRKFRLIAEDVQDRNVLCNFHGMDLTTDKYRSMVKKW  120
            YLKGRVFEVSLADLQKDIDPERSFRKFRLIAEDVQDRNVLCNFHGMDLTTDKYRSMVKKW
Sbjct  61   YLKGRVFEVSLADLQKDIDPERSFRKFRLIAEDVQDRNVLCNFHGMDLTTDKYRSMVKKW  120

Query  121  QTLIEAIVEAKTVDGYLLRVFCIGFTSKDQQSQRKTCYAQQSQVRKIRARMTDIITNEVS  180
            QTLIEAIVEAKTVDGYLLRVFCIGFT+KDQQSQRKTCYAQQSQVRKIRARMTDIITNEVS
Sbjct  121  QTLIEAIVEAKTVDGYLLRVFCIGFTAKDQQSQRKTCYAQQSQVRKIRARMTDIITNEVS  180

Query  181  GADLKQLVNKLALDSIAKDIEKSCQRIYPLHDVYIRKVKVLKKPRFDVSKLLELHGDGGG  240
            GADLKQLVNKLALDSIAKDIEKSCQRIYPLHDVYIRKVKVLKKPRFDVSKLLELHGDGGG
Sbjct  181  GADLKQLVNKLALDSIAKDIEKSCQRIYPLHDVYIRKVKVLKKPRFDVSKLLELHGDGGG  240

Query  241  KTTEAVVSAEGAVIDRPEGYEPPVQEAV  268
            K+ EAVVS+EGAVIDRPEGYEPPVQEAV
Sbjct  241  KSVEAVVSSEGAVIDRPEGYEPPVQEAV  268
```

I feel this is sufficient to confirm my gene model for Feature 4, the *D. virilis* ortholog to *D. melanogaster* ribosomal protein S3A.

Feature 5 (CG17964-RG, NM_166723):

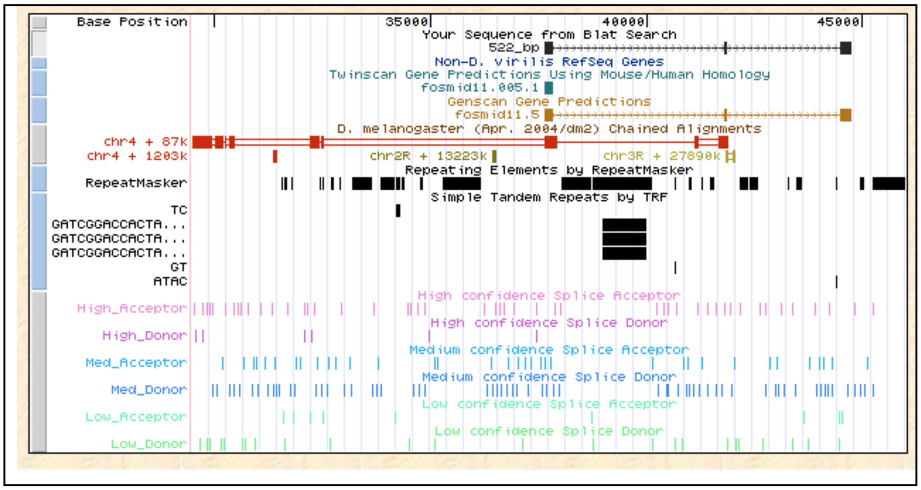
I used BLASTNnt and BLASTPnr again, with the Genscan predictions for CDS and peptide sequences as queries. Both showed hits to *D. melanogaster pangolin* (BLASTNnt 3e-14, BLASTPnr 6e-24). When visualized in UCSC *D. melanogaster* BLAT, however, the prediction only aligned to a single exon of the *D. melanogaster* gene, the first coding exon. Based on the location of this exon (ATG at 37655 of 46956) in my fosmid and the relative locations of the exons in the *D. melanogaster* gene, I predict that the first three exons should be present in my fosmid before the gene runs off my fosmid end.

Figure 16. UCSC *D. melanogaster* BLAT output for predicted Feature 5


chr4: 88000 89000 90000 91000 92000 93000 94000 95000 96000 97000 98000 99000 100000
Your Sequence from Blat Search
173_aa
FlyBase Protein-Coding Genes
RpS3A
RpS3A
RpS3A
pan
pan
pan
pan
pan
pan
pan
pan
RefSeq Genes
RefSeq Genes
FB Noncoding
FlyBase Noncoding Genes
N-SCAN
N-SCAN Gene Predictions
chr4_1.4
chr4_1.5
Genscan Gene Predictions
D. melanogaster mRNAs from GenBank
AF034971
AY070949
Y10115
AY075359
Y09125
BT021431
U88538
D. melanogaster ESTs That Have Been Spliced
Spliced ESTs
D. melanogaster ESTs Including Unspliced
melanogaster ESTs
FlyReg: Drosophila DNase I Footprint Database
FlyReg
7 Flies, Mosquito and Honeybee Multiz Alignments & phastCons Scores
Conservation
d_simulans
d_yakuba
d_ananassae
d_pseudoobscura
d_virilis
d_mojavensis
a_gambiae
a_mellifera
D. virilis (July 2004/droVir1) Chained Alignments

When attempting to create a gene model in Goose BLAT, I noted that BLAT aligned the *D. virilis* sequence to the first three coding exons of *D. melanogaster* shown above in light blue, and Genscan had assumed a complete open reading frame so that the third exon terminated in a stop codon. I found these results questionable.

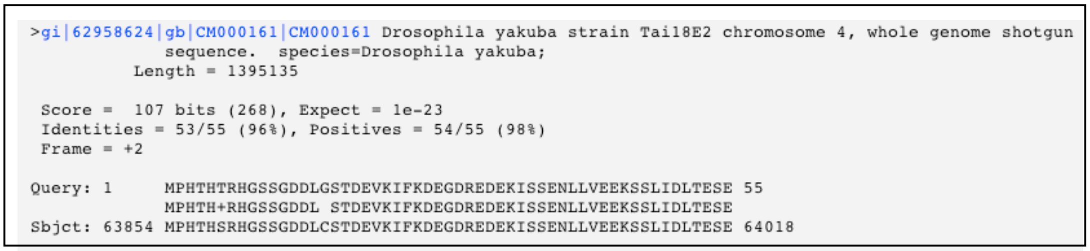
Figure 17. Goose Browser BLAT alignment of Feature 5 in Fosmid 11


Base Position
35000
40000
45000
Your Sequence from Blat Search
522_bp
Non-D. virilis RefSeq Genes
Twinscan Gene Predictions Using Mouse/Human Homology
fosmid11.005.1
Genscan Gene Predictions
fosmid11.5
D. melanogaster (Apr. 2004/dm2) Chained Alignments
chr4 + 87k
chr4 + 1203k
chr2R + 13223k
chr3R + 27890k
Repeating Elements by RepeatMasker
RepeatMasker
Simple Tandem Repeats by TRF
TC
GATCGGACCACTA...
GATCGGACCACTA...
GATCGGACCACTA...
GT
ATAC
High confidence Splice Acceptor
High_Acceptor
High confidence Splice Donor
High_Donor
Medium confidence Splice Acceptor
Med_Acceptor
Medium confidence Splice Donor
Med_Donor
Low confidence Splice Acceptor
Low_Acceptor
Low confidence Splice Donor
Low_Donor

I extracted this region from my fosmid and used it as a database sequence for TBLASTN2seq, with the *D. melanogaster* peptide as a query, and also for TBLASTX2seq with the *D. melanogaster* CDS sequence as a query. Both times, only the first *D. virilis* exon successfully aligned, and with 96 percent identity, even with increasing expect values. To be sure the other two exons were real, I concatenated their sequences and translated. Translation was successful and so I aligned that peptide sequence to the *D. melanogaster* peptide sequence for *pan* via BLASTP2seq. Still, only the amino acids corresponding to the first exon (residues 1-55) aligned at all, even after increasing expect values dramatically.

Next, I used Flybase TBLASTN with *D. yakuba* and *D. mojavensis* genomes to search for the Feature 5 peptide sequence. Since *D. mojavensis* is evolutionarily more closely related to *D. virilis*, the result should tell us with some level of confidence whether these last two exons are indeed real. TBLASTN results for *D. yakuba* show a positive hit to Chromosome 4, which is a positive result for synteny, but the alignment only extended as far as the first exon, yet again.

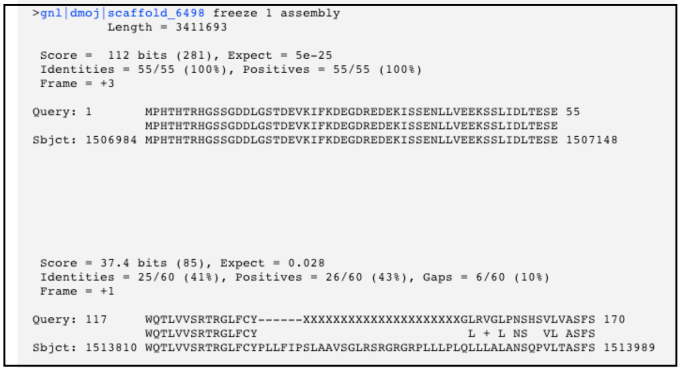
Figure 18. Flybase TBLASTN for Feature 5 peptide against *D. yakuba*



```
>gi|62958624|gb|CM000161|CM000161 Drosophila yakuba strain Tai18E2 chromosome 4, whole genome shotgun
            sequence.  species=Drosophila yakuba;
         Length = 1395135

 Score =  107 bits (268), Expect = 1e-23
 Identities = 53/55 (96%), Positives = 54/55 (98%)
 Frame = +2

Query: 1     MPHTHTRHGSSGDDLGSTDEVKIFKDEGDREDEKISSENLLVEEKSSLIDLTESE 55
             MPHTH+RHGSSGDDL STDEVKIFKDEGDREDEKISSENLLVEEKSSLIDLTESE
Sbjct: 63854 MPHTHSRHGSSGDDLCSTDEVKIFKDEGDREDEKISSENLLVEEKSSLIDLTESE 64018
```

Results for TBLASTN against *D. mojavensis* were no more promising. Again, I saw only significant alignment for the first exon.

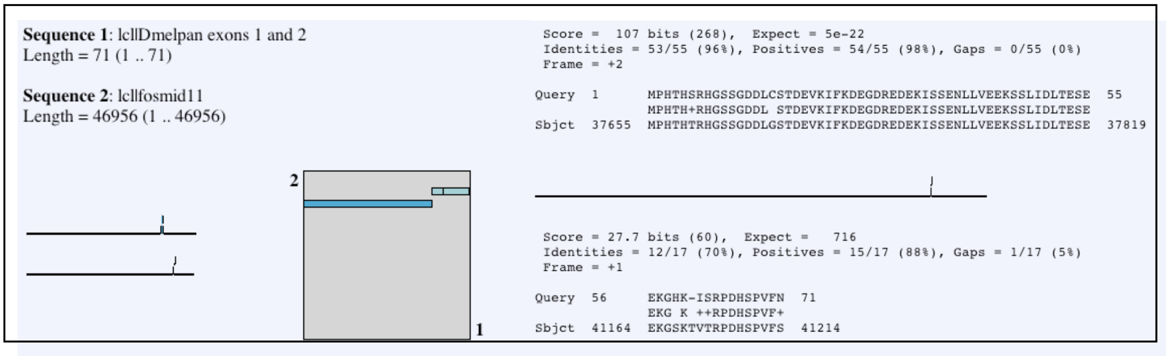
Figure 19. Flybase TBLASTN for Feature 5 peptide against *D. mojavensis*



```
>gnl|dmoj|scaffold_6498 freeze 1 assembly
          Length = 3411693

 Score =  112 bits (281), Expect = 5e-25
 Identities = 55/55 (100%), Positives = 55/55 (100%)
 Frame = +3

Query: 1       MPHTHTRHGSSGDDLGSTDEVKIFKDEGDREDEKISSENLLVEEKSSLIDLTESE 55
               MPHTHTRHGSSGDDLGSTDEVKIFKDEGDREDEKISSENLLVEEKSSLIDLTESE
Sbjct: 1506984 MPHTHTRHGSSGDDLGSTDEVKIFKDEGDREDEKISSENLLVEEKSSLIDLTESE 1507148



 Score = 37.4 bits (85), Expect = 0.028
 Identities = 25/60 (41%), Positives = 26/60 (43%), Gaps = 6/60 (10%)
 Frame = +1

Query: 117     WQTLVVSRTRGLFCY------XXXXXXXXXXXXXXXXXXXXXGLRVGLPNSHSVLVASFS 170
               WQTLVVSRTRGLFCY                               L + L NS   VL ASFS
Sbjct: 1513810 WQTLVVSRTRGLFCYPLLFIPSLAAVSGLRSRGRGRPLLLPLQLLLALANSQPVLTASFS 1513989
```

The above image shows alignment with the Filter ON and so some repetitive sequence was masked (shown by X's). I performed the same alignment with the Filter OFF option, but the results were not significantly different. As a last test, I ran TBLASTN again using *D. melanogaster pan* exons 1 through 4 as my query against my entire fosmid sequence. This showed successful alignment of *D. melanogaster* exons 1 and 2, but nothing more, even after increasing expect values to 10,000. Exon 2 however came up in my fosmid 500 base pairs upstream of the Genscan prediction and may have simply been missed due to its extremely small size, only fifteen residues. The alignment of exons 1 and 2 correspond to 1 through 71 of *D. melanogaster pan*. Note the high identity hit of the *D. melanogaster* exons at 70 percent. This confirms the presence of exon 2 in my fosmid, but not three. I subsequently scanned the Goose Browser in my fosmid for the second exon and determined the location of the appropriate peptide sequence. After consultation with Chris, I learned that there is a gap of about 10 kilobases between this end of my fosmid and the closest of last year's fosmids so that exon 3 may not be present in either fosmid. It is certainly not on my fosmid and it failed to align to the collection of last year's fosmids, despite the annotation of part of the *pangolin* gene by Emiko Morimoto from 2004's class.

Figure 20. Flybase TBLASTN shows alignment of *D. melanogaster* exons 1-2 to fosmid



Sequence 1: lcl|Dmelpan exons 1 and 2
Length = 71 (1 .. 71)

Sequence 2: lcl|fosmid11
Length = 46956 (1 .. 46956)

```
 Score =  107 bits (268),  Expect = 5e-22
 Identities = 53/55 (96%), Positives = 54/55 (98%), Gaps = 0/55 (0%)
 Frame = +2

Query  1      MPHTHSRHGSSGDDLCSTDEVKIFKDEGDREDEKISSENLLVEEKSSLIDLTESE  55
              MPHTH+RHGSSGDDL STDEVKIFKDEGDREDEKISSENLLVEEKSSLIDLTESE
Sbjct  37655  MPHTHTRHGSSGDDLGSTDEVKIFKDEGDREDEKISSENLLVEEKSSLIDLTESE  37819


 Score = 27.7 bits (60),  Expect =   716
 Identities = 12/17 (70%), Positives = 15/17 (88%), Gaps = 1/17 (5%)
 Frame = +1

Query  56     EKGHK-ISRPDHSPVFN  71
              EKG K ++RPDHSPVF+
Sbjct  41164  EKGSKTVTRPDHSPVFS  41214
```

This confirms a two-exon gene model in my fosmid for the *D. virilis* ortholog to *pan*. This also prevents me from attempting to overlap my work with last year's, although the rest of *pan* may have been annotated.

Gene Function Summaries

1. CG11155

Although not officially annotated, Flybase describes CG11155 as a glutamate-gated potassium ion channel. Based on references from Flybase, this protein most likely is involved in the conduction of action potentials at neuromuscular junctions.

2. Kif3C

Kif3C is a motor kinesin. It functions as an ATP-binding structural component of the cytoskeleton with microtubule motor activity and protein targeting.

3. *pho*

Short for *pleiohomeotic*, *pho* is a polycomb transcription factor. It has DNA-binding activity and a putative Zinc-finger domain. It is involved in transcriptional regulation of gene expression.

4. RpS3A

Ribosomal protein S3A is highly conserved among eukaryotes. It has nucleic acid-binding function in the small eukaryotic 40S ribosomal subunit, and thus serves in translation and ribosome structure.

5. *pan*

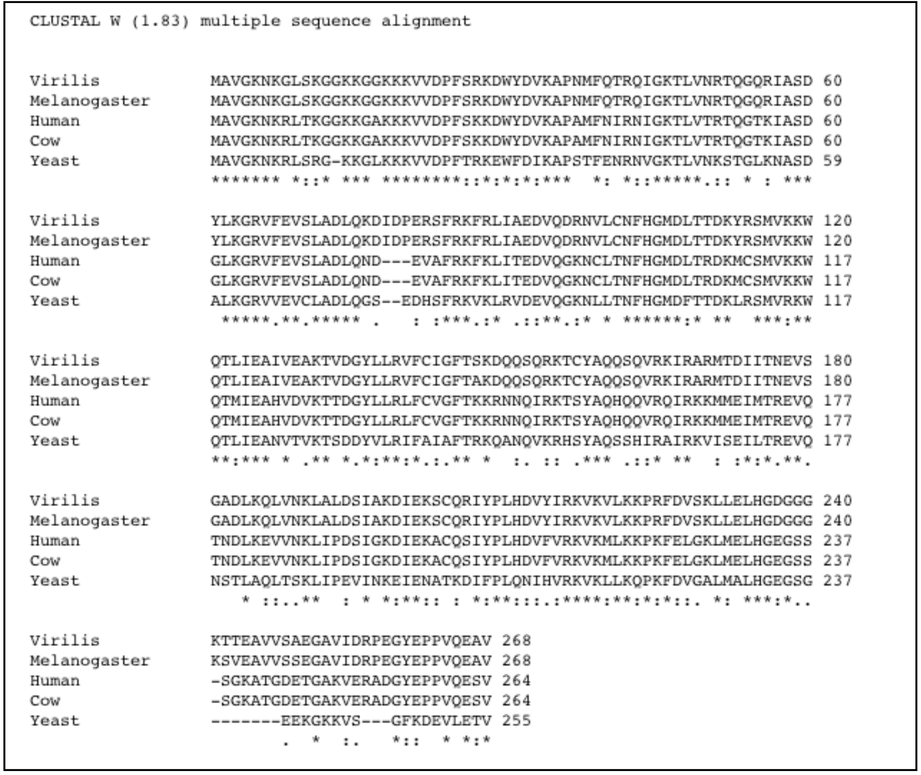
The *pangolin* protein functions in transcriptional regulation, embryonic patterning, cellular polarity, differentiation, and other aspects of development.

III. Clustal Analysis

For my gene, I investigated RpS3A and sought to align my *D. virilis* peptide sequence with peptides from four other species to see alignment at the amino acid level that would speak to the conservation of ribosomal proteins. The RpS3A peptide sequences were collected from Ensembl from the following species (included are the Genbank accession numbers for these peptides):

D. melanogaster	NP_524618
Homo sapiens	NP_000997
Bos taurus	NP_001029210
S. cerevisiae	NP_013648

Figure 21. ClustalW alignment of five RpS3A peptide sequences

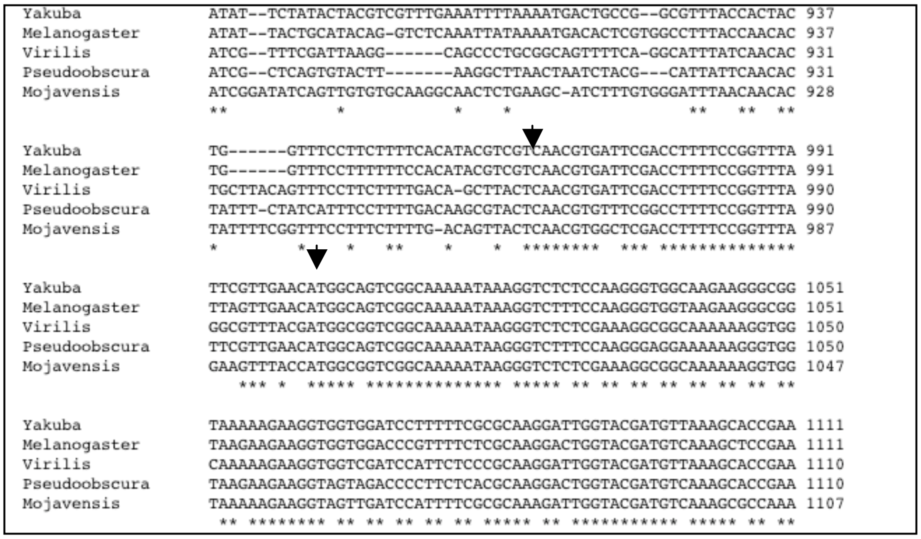


```
CLUSTAL W (1.83) multiple sequence alignment


Virilis           MAVGKNKGLSKGGKKGGKKKVVDPFSRKDWYDVKAPNMFQTRQIGKTLVNRTQGQRIASD 60
Melanogaster      MAVGKNKGLSKGGKKGGKKKVVDPFSRKDWYDVKAPNMFQTRQIGKTLVNRTQGQRIASD 60
Human             MAVGKNKRLTKGGKKGAKKKVVDPFSKKDWYDVKAPAMFNIRNIGKTLVTRTQGTKIASD 60
Cow               MAVGKNKRLTKGGKKGAKKKVVDPFSKKDWYDVKAPAMFNIRNIGKTLVTRTQGTKIASD 60
Yeast             MAVGKNKRLSRG-KKGLKKKVVDPFTRKEWFDIKAPSTFENRNVGKTLVNKSTGLKNASD 59
                  ******* *::* *** ********::*:*:*:***  *: *::*****.:: * : ***

Virilis           YLKGRVFEVSLADLQKDIDPERSFRKFRLIAEDVQDRNVLCNFHGMDLTTDKYRSMVKKW 120
Melanogaster      YLKGRVFEVSLADLQKDIDPERSFRKFRLIAEDVQDRNVLCNFHGMDLTTDKYRSMVKKW 120
Human             GLKGRVFEVSLADLQND---EVAFRKFKLITEDVQGKNCLTNFHGMDLTRDKMCSMVKKW 117
Cow               GLKGRVFEVSLADLQND---EVAFRKFKLITEDVQGKNCLTNFHGMDLTRDKMCSMVKKW 117
Yeast             ALKGRVVEVCLADLQGS--EDHSFRKVKLRVDEVQGKNLLTNFHGMDFTTDKLRSMVRKW 117
                   *****.**.***** .   : :***.:* .::**.:* * ******:* **  ***:**

Virilis           QTLIEAIVEAKTVDGYLLRVFCIGFTSKDQQSQRKTCYAQQSQVRKIRARMTDIITNEVS 180
Melanogaster      QTLIEAIVEAKTVDGYLLRVFCIGFTAKDQQSQRKTCYAQQSQVRKIRARMTDIITNEVS 180
Human             QTMIEAHVDVKTTDGYLLRLFCVGFTKKRNNQIRKTSYAQHQQVRQIRKKMMEIMTREVQ 177
Cow               QTMIEAHVDVKTTDGYLLRLFCVGFTKKRNNQIRKTSYAQHQQVRQIRKKMMEIMTREVQ 177
Yeast             QTLIEANVTVKTSDDYVLRIFAIAFTRKQANQVKRHSYAQSSHIRAIRKVISEILTREVQ 177
                  **:*** * .** *.*:**:*.:.** *  :. :: .*** .::* **  : :*:*.**.

Virilis           GADLKQLVNKLALDSIAKDIEKSCQRIYPLHDVYIRKVKVLKKPRFDVSKLLELHGDGGG 240
Melanogaster      GADLKQLVNKLALDSIAKDIEKSCQRIYPLHDVYIRKVKVLKKPRFDVSKLLELHGDGGG 240
Human             TNDLKEVVNKLIPDSIGKDIEKACQSIYPLHDVFVRKVKMLKKPKFELGKLMELHGEGSS 237
Cow               TNDLKEVVNKLIPDSIGKDIEKACQSIYPLHDVFVRKVKMLKKPKFELGKLMELHGEGSS 237
Yeast             NSTLAQLTSKLIPEVINKEIENATKDIFPLQNIHVRKVKLLKQPKFDVGALMALHGEGSG 237
                     * ::..**  : * *:**:: : *:**:::.:****:**:*:*::. *: ***:*..

Virilis           KTTEAVVSAEGAVIDRPEGYEPPVQEAV 268
Melanogaster      KSVEAVVSSEGAVIDRPEGYEPPVQEAV 268
Human             -SGKATGDETGAKVERADGYEPPVQESV 264
Cow               -SGKATGDETGAKVERADGYEPPVQESV 264
Yeast             -------EEKGKKVS---GFKDEVLETV 255
                         .  *  :.   *::  * *:*
```

As expected, the alignment shows remarkable conservation among these species for RpS3A. Note the four instances of gaps. The first demonstrates a departure of the four multicellular eukaryotes from yeast; all have acquired an additional G in the amino acid sequence. The second shows a unique insertion in the flies (IDP), while the third shows an insertion in all four multicellular species, but with SGKATG for the mammals, and TTEAV/SVEAV for *D. virilis* and *D. melanogaster*, respectively. The fourth insertion event is similar, with RAD for the mammals and RPE for the flies. This shows how the evolutionary divergence and similarity between species in the same clade, such as the flies/insects, the mammals, and the multicellular eukaryotes.

For my promoter region, I thought it interesting to see what conservation would exist for that of the RpS3A gene. Since regulatory regions can diverge in evolution much more than genic sequences, I compared those of five fly species. I extracted the first exon of RpS3A and the regions stretching 1 kilobase upstream. In addition to being aligned in ClustalW, I used this sequence as a query in UCSC BLAT against four other *Drosophila* species (in increasing evolutionary distance from *D. virilis*): *D. mojavensis*, *D. pseudoobscura*, *D. yakuba*, *D. melanogaster*. Once I was able to align the sequence from my fosmid in BLAT to these species, I extracted an identical region including the first exon as an alignment anchor and the 1 kilobase regions upstream of the start codon.

Figure 22. ClustalW alignment of five *Drosophila* species for RpS3A promoter region



```
Yakuba              ATAT--TCTATACTACGTCGTTTGAAATTTTAAAATGACTGCCG--GCGTTTACCACTAC 937
Melanogaster        ATAT--TACTGCATACAG-GTCTCAAATTATAAAATGACACTCGTGGCCTTTACCAACAC 937
Virilis             ATCG--TTTCGATTAAGG------CAGCCCTGCGGCAGTTTTCA-GGCATTTATCAACAC 931
Pseudoobscura       ATCG--CTCAGTGTACTT-------AAGGCTTAACTAATCTACG---CATTATTCAACAC 931
Mojavensis          ATCGGATATCAGTTGTGTGCAAGGCAACTCTGAAGC-ATCTTTGTGGGATTTAACAACAC 928
                    **           *           *    *                  **   **  **

Yakuba              TG------GTTTCCTTCTTTTCACATACGTCGTCAACGTGATTCGACCTTTTCCGGTTTA 991
Melanogaster        TG------GTTTCCTTTTTTCCACATACGTCGTCAACGTGATTCGACCTTTTCCGGTTTA 991
Virilis             TGCTTACAGTTTCCTTCTTTTGACA-GCTTACTCAACGTGATTCGACCTTTTCCGGTTTA 990
Pseudoobscura       TATTT-CTATCATTTCCTTTTGACAAGCGTACTCAACGTGTTTCGGCCTTTTCCGGTTTA 990
Mojavensis          TATTTTCGGTTTCCTTTCTTTTG-ACAGTTACTCAACGTGGCTCGACCTTTTCCGGTTTA 987
                    *        *    *   **    *    *  ********  *** **************

Yakuba              TTCGTTGAACATGGCAGTCGGCAAAAATAAAGGTCTCTCCAAGGGTGGCAAGAAGGGCGG 1051
Melanogaster        TTAGTTGAACATGGCAGTCGGCAAAAATAAAGGTCTTTCCAAGGGTGGTAAGAAGGGCGG 1051
Virilis             GGCGTTTACGATGGCGGTCGGCAAAAATAAGGGTCTCTCGAAAGGCGGCAAAAAAGGTGG 1050
Pseudoobscura       TTCGTTGAACATGGCAGTCGGCAAAAATAAGGGTCTTTCCAAGGGAGGAAAAAAGGGTGG 1050
Mojavensis          GAAGTTTACCATGGCGGTCGGCAAAAATAAGGGTCTCTCGAAAGGCGGCAAAAAAGGTGG 1047
                       *** *  ***** ************** ***** ** ** ** ** ** ** ** **

Yakuba              TAAAAAGAAGGTGGTGGATCCTTTTTCGCGCAAGGATTGGTACGATGTTAAAGCACCGAA 1111
Melanogaster        TAAGAAGAAGGTGGTGGACCCGTTTTCTCGCAAGGACTGGTACGATGTCAAAGCTCCGAA 1111
Virilis             CAAAAAGAAGGTGGTCGATCCATTCTCCCGCAAGGATTGGTACGATGTTAAAGCACCGAA 1110
Pseudoobscura       TAAGAAGAAGGTAGTAGACCCCTTCTCACGCAAGGACTGGTACGATGTCAAAGCACCGAA 1110
Mojavensis          TAAAAAGAAGGTAGTTGATCCATTTTCGCGCAAAGATTGGTACGATGTCAAAGCGCCAAA 1107
                     ** ******** ** ** ** ** ** ***** ** *********** ***** ** **
```

While this image does not represent the entire alignment block, it shows the boundary between the regulatory region and a portion of the first exon. Note the high conservation at the first exon sequence starting with ATG. Upstream of that is the regulatory region. Here, we see a few conserved motifs including one that appears relevant to the TATA box (TCAACGTG) in all five species. Still, the upstream UTR in *D. melanogaster* extends 100 base pairs. This may mean that the transcription start site is further upstream from the region shown above and thus not highly conserved, as essentially no conservation was observed for the rest of the 1 kilobase upstream extracted region.

REPEAT ANALYSIS

To assess what repeats may have been missed by RepeatMasker, I used BLASTNnt with my masked fosmid sequence as a query against a database of all the fosmids to date. Considering only those hits with an E value of Xe-05 or less, I examined the alignments. I used the same E value threshold for the alignments as well and found that all alignments within these significant hits were redundant for four distinct regions:

Repeat Number	Region of Fosmid 11 (masked)
1	1091-1213
2	36318-36507
3	36532-36627
4	40889-40921

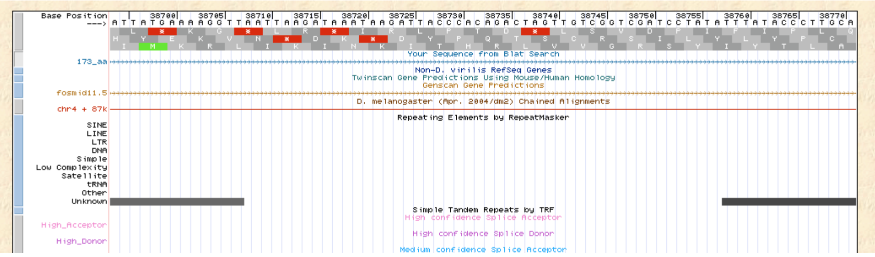
All occur within introns and none of these were close in position to the repeats that were successfully masked by RepeatMasker. Each of these regions was extracted and used as a query against the *D. virilis* repeats database created by Wilson, via BLASTNnt. Repeat 1 came up with no significant hits (all hits E value of 4.1) and was 300 base pairs from the nearest masked repeat. Its size of 120 base pairs, while still considerable, is small. I was unable to make any conclusions about Repeat 1 and assessed the others.

Repeats 2 and 3 each individually had near insignificant hits to LTR retrotransposons (Repeat 2 E value of .095, Repeat 3 E value of .013). However, I thought it best to combine the two as they are only 25 base pairs apart and examination of that sequence in the Goose Browser gave no indication that those 25 bases could not be considered repetitious. I extracted a sequence using the start of Repeat 2 and the end of Repeat 3, and BLASTN yielded a solid hit to an LTR/Gypsy (OSVALDO) retrotransposon (E value 7e-04). Therefore, the coordinates should be taken to be 36318-36627.

Repeat 4 had BLASTN hits to an LTR retrotransposon (E value 0.78) that were fairly poor. It does not occur anywhere near other repetitious elements masked by RepeatMasker or plotted in the Goose Browser and so I concluded that, due to its small size of 40 base pairs, it is likely a fragment of ROO I LTR retrotransposon, which it aligned to, but is too short to consider significant.

While that concluded my search for novel repeats, I did make an additional change while constructing my fosmid map. Two of the largest repeats, each near 1 kilobase each, were separated by 50 base pairs and were of similar lineage: dvir.11.23.centroid and dvir.11.33.centroid. Their coordinates are 38016-38707 and 38758-40122, respectively. Upon inspection of those 50 base pairs in the Goose Browser, the sequence did not distinguish itself from the surrounding repeats and so I joined the two. Both are in the same orientation on the minus strand, both are TRF (tandem repeats) and together create a long repeat of 2107 base pairs. The following shows the regions over which I connected these two repeats:

Figure 23. Two repeats joined to make one whole over a 50 base pair gap


Base Position
38700 38705 38710 38715 38720 38725 38730 38735 38740 38745 38750 38755 38760 38765 38770
---> ATTATGAAAAGGTTAATTAAGATAAATAAGATTACCCACAGACTAGTTGTCGGTCGATCCTATATTTATACCCTTGCA
Your Sequence from Blat Search
173_aa
Non-D. virilis RefSeq Genes
Twinscan Gene Predictions Using Mouse/Human Homology
Genscan Gene Predictions
fosmid11.5
D. melanogaster (Apr. 2004/dm2) Chained Alignments
chr4 + 87k
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
tRNA
Other
Unknown
Simple Tandem Repeats by TRF
High confidence Splice Acceptor
High_Acceptor
High confidence Splice Donor
High_Donor
Medium confidence Splice Acceptor

After this assessment, I report statistics on the repetitive content of my fosmid:

<u>Percent of Fosmid 11 that is repetitious: 36%</u>

Composition of the repetitious DNA:

LINE: 37%
LTR: 5%
DNA: 19%
simple repeats: 7%
low complexity: 6%
TRF (tandem repeats): 26%
unknown: <1%

The following table details the repeats found by RepeatMasker, and was edited to include those I found by pair-wise BLAST.

Column 1: Fosmid name
Column 2: Position (Beginning in fosmid)
Column 3: Position (End in fosmid)
Column 4: Total length
Column 5: Strand
Column 6: Repeat name
Column 7: Repeat class

fosmid11	3082	3155	74	+	dvir.16.2.centroid	DNA
fosmid11	5105	5174	70	C	dvir.16.17.centroid	DNA
fosmid11	5196	5382	187	C	dvir.16.17.centroid	DNA
fosmid11	6294	6365	72	+	dvir.16.2.centroid	DNA
fosmid11	8679	8901	223	C	dvir.16.2.centroid	DNA
fosmid11	8924	8993	70	+	dvir.16.17.centroid	DNA
fosmid11	11291	11359	69	+	dvir.16.2.centroid	DNA
fosmid11	11416	11571	156	+	dvir.16.2.centroid	DNA
fosmid11	12446	12992	547	C	dvir.14.34.centroid	DNA
fosmid11	15850	15965	116	C	dvir.16.17.centroid	DNA
fosmid11	17239	17278	40	C	dvir.16.17.centroid	DNA

fosmid11	18027	18135	109	+	dvir.16.2.centroid	DNA
fosmid11	20264	20381	118	+	dvir.13.20.centroid	DNA
fosmid11	33174	33341	168	C	dvir.16.2.centroid	DNA
fosmid11	35273	36170	898	C	dvir.16.2.centroid	DNA
fosmid11	42382	42477	96	C	dvir.16.2.centroid	DNA
fosmid11	42476	42586	111	C	dvir.16.2.centroid	DNA
fosmid11	45612	45684	73	C	dvir.16.2.centroid	DNA
fosmid11	1514	2831	1318	+	dvir.0.85.centroid	LINE
fosmid11	2803	3088	286	C	dvir.0.10.centroid	LINE
fosmid11	3122	3328	207	+	PENELOPE	LINE
fosmid11	3357	3501	145	+	PENELOPE	LINE
fosmid11	5381	5902	522	C	dvir.0.18.centroid	LINE
fosmid11	5902	6249	348	C	dvir.0.85.centroid	LINE
fosmid11	6250	6305	56	C	dvir.0.18.centroid	LINE
fosmid11	6332	6579	248	+	PENELOPE	LINE
fosmid11	6584	6786	203	C	dvir.0.5.centroid	LINE
fosmid11	6785	7428	644	+	dvir.0.85.centroid	LINE
fosmid11	7427	8680	1254	+	dvir.0.18.centroid	LINE
fosmid11	11332	11415	84	+	PENELOPE	LINE
fosmid11	11614	11762	149	+	PENELOPE	LINE
fosmid11	15440	15476	37	C	PENELOPE	LINE
fosmid11	42166	42360	195	+	dvir.0.14.centroid	LINE
fosmid11	45244	45384	141	C	PENELOPE	LINE
fosmid11	45411	45645	235	C	PENELOPE	LINE
fosmid11	45708	45991	284	+	PENELOPE	LINE
fosmid11	757	779	23	+	AT_rich	Low_complexity
fosmid11	4102	4133	32	+	AT_rich	Low_complexity
fosmid11	4361	4402	42	+	AT_rich	Low_complexity
fosmid11	10355	10375	21	+	AT_rich	Low_complexity
fosmid11	14373	14403	31	+	AT_rich	Low_complexity
fosmid11	14794	14846	53	+	AT_rich	Low_complexity
fosmid11	18396	18426	31	+	AT_rich	Low_complexity
fosmid11	19465	19495	31	+	AT_rich	Low_complexity
fosmid11	20121	20142	22	+	AT_rich	Low_complexity
fosmid11	20650	20677	28	+	AT_rich	Low_complexity
fosmid11	23402	23472	71	+	T-rich	Low_complexity
fosmid11	23507	23530	24	+	AT_rich	Low_complexity
fosmid11	23706	23746	41	+	AT_rich	Low_complexity
fosmid11	23977	24014	38	+	AT_rich	Low_complexity
fosmid11	25205	25233	29	+	AT_rich	Low_complexity
fosmid11	25447	25481	35	+	AT_rich	Low_complexity
fosmid11	25532	25580	49	+	AT_rich	Low_complexity
fosmid11	25655	25680	26	+	AT_rich	Low_complexity
fosmid11	26096	26140	45	+	AT_rich	Low_complexity
fosmid11	26457	26482	26	+	AT_rich	Low_complexity
fosmid11	27105	27150	46	+	AT_rich	Low_complexity

fosmid11	27178	27213	36	+	AT_rich	Low_complexity
fosmid11	27702	27757	56	+	AT_rich	Low_complexity
fosmid11	29264	29308	45	+	AT_rich	Low_complexity
fosmid11	31629	31695	67	+	A-rich	Low_complexity
fosmid11	41265	41301	37	+	AT_rich	Low_complexity
fosmid11	43274	43301	28	+	AT_rich	Low_complexity
fosmid11	12256	12445	190	+	dvir.15.30.centroid	LTR
fosmid11	36318	36627	310	+	dvir.36.76.centroid	LTR/Gypsy OSVALDO
fosmid11	46419	46488	70	+	GYPSY2_I	LTR/Gypsy
fosmid11	46621	46689	69	+	GYPSY6_I	LTR/Gypsy
fosmid11	46698	46954	257	+	GYPSY2_I	LTR/Gypsy
fosmid11	3329	3356	28	+	(CGGA)n	Simple_repeat
fosmid11	8995	9054	60	+	(TTTA)n	Simple_repeat
fosmid11	10001	10037	37	+	(CAAAT)n	Simple_repeat
fosmid11	11572	11613	42	+	(CGGA)n	Simple_repeat
fosmid11	14893	15056	164	+	(CATATA)n	Simple_repeat
fosmid11	22534	22592	59	+	(TA)n	Simple_repeat
fosmid11	31555	31592	38	+	(TG)n	Simple_repeat
fosmid11	31776	31821	46	+	(TA)n	Simple_repeat
fosmid11	32429	32453	25	+	(TCG)n	Simple_repeat
fosmid11	32512	32541	30	+	(CA)n	Simple_repeat
fosmid11	32703	32748	46	+	(TATG)n	Simple_repeat
fosmid11	32891	32922	32	+	(CGGT)n	Simple_repeat
fosmid11	34190	34288	99	+	(TC)n	Simple_repeat
fosmid11	40643	40682	40	+	(TG)n	Simple_repeat
fosmid11	41459	41561	103	+	(TATG)n	Simple_repeat
fosmid11	43471	43610	140	+	(TATATG)n	Simple_repeat
fosmid11	44372	44414	43	+	(CATA)n	Simple_repeat
fosmid11	44955	45048	94	+	(CATA)n	Simple_repeat
fosmid11	45385	45410	26	+	(TCCG)n	Simple_repeat
fosmid11	46265	46316	52	+	(TAA)n	Simple_repeat
fosmid11	3504	3704	201	+	dvir.11.33.centroid	TRF
fosmid11	3724	3815	92	C	dvir.11.33.centroid	TRF
fosmid11	11772	11803	32	+	dvir.11.23.centroid	TRF
fosmid11	11825	12253	429	+	dvir.11.33.centroid	TRF
fosmid11	12993	13111	119	+	dvir.11.33.centroid	TRF
fosmid11	17404	17642	239	+	dvir.11.23.centroid	TRF
fosmid11	17619	17906	288	+	dvir.11.23.centroid	TRF
fosmid11	33235	33649	415	C	dvir.11.33.centroid	TRF
fosmid11	33847	34081	235	C	dvir.11.33.centroid	TRF
fosmid11	34061	34156	96	C	dvir.11.23.centroid	TRF
fosmid11	34324	34404	81	C	dvir.11.33.centroid	TRF
fosmid11	34764	34831	68	+	dvir.11.33.centroid	TRF
fosmid11	38016	40122	2107	C	dvir.11.23.centroid + dvir.11.33.centroid	TRF

fosmid11	40976	41052	77	+	dvir.11.23.centroid	TRF
fosmid11	20147	20263	117	C	dvir.13.51.centroid	Unknown

SYNTENY

All genes annotated in my fosmid from *D. virilis* dot Chromosome IV have orthologs in *D. melanogaster* Chromosome IV. The following figures demonstrate the syntenic blocks present in Fosmid 11, as well as the locations of the genes on *D. melanogaster* Chromosome IV. There are two inversion events as well as the interesting point that these genes, which reside within 47 kilobases of each other on my fosmid, are on opposite polar ends of the *D. melanogaster* Chromosome IV.

Figure 24. Syntenic blocks between Fosmid 11 and *D. melanogaster* Chromosome IV

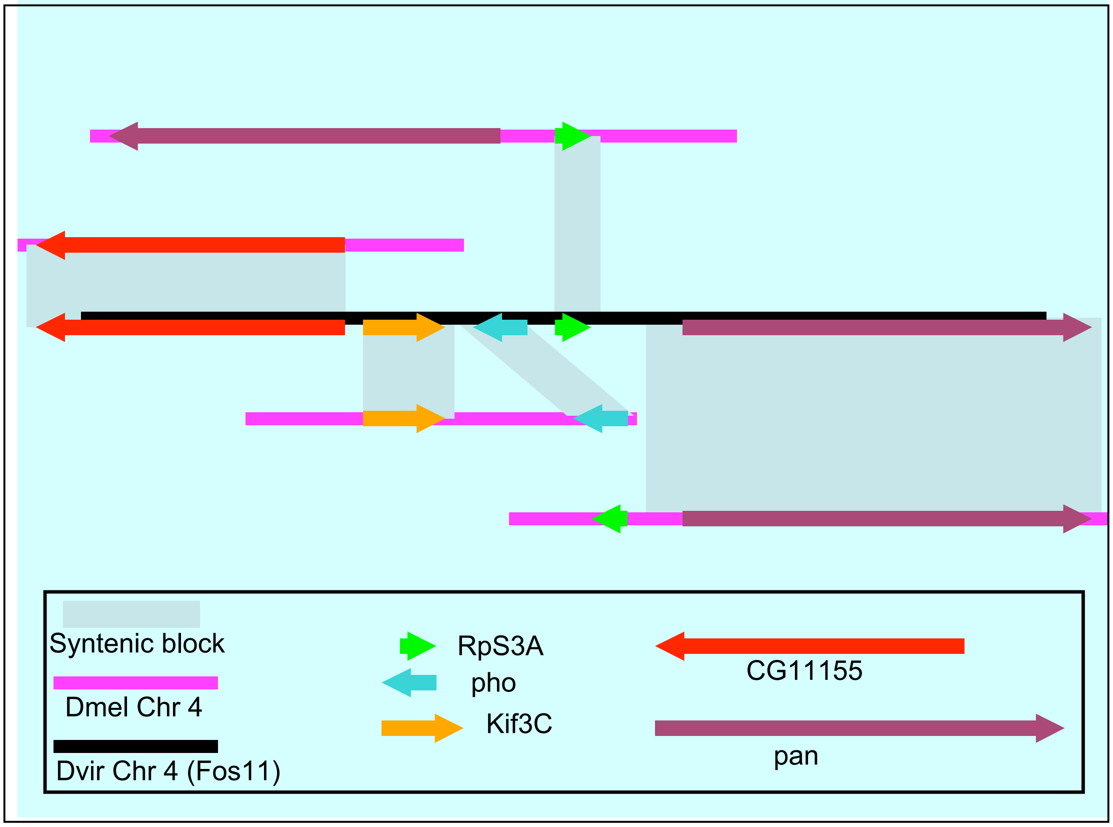
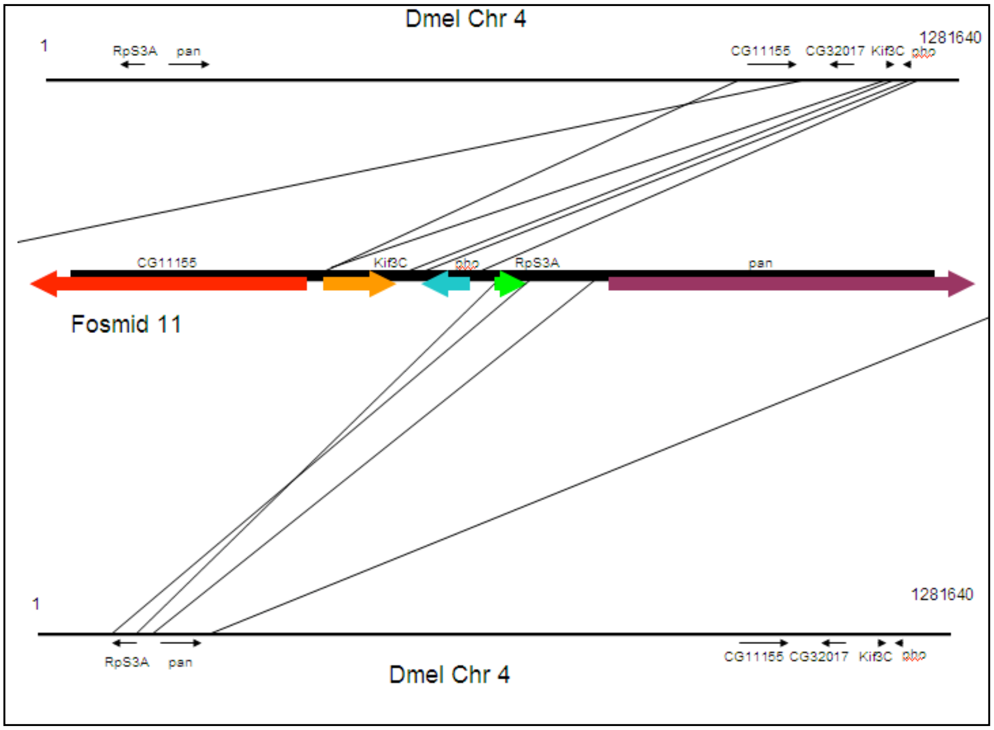


Figure 25. Fosmid 11 genes map to opposite ends of *D. melanogaster* Chromosome IV



CG32017 depicted on the right end of the *D. melanogaster* chromosome exists between CG11155 and Kif3C, but does not appear on my fosmid. It was picked up by Kelli Grim (2006 class) on her Fosmid 7. While chromosomal synteny is preserved, there have obviously been some large-scale rearrangements within Chromosome IV of these *Drosophila* species that merit investigation.