Annotation of Fosmid 2 in **D. virilis**

Alan Tseng
3/5/07
Bio4342W

Overview

For my annotation project, I was given fosmid 2, a 61,074 base pair fosmid located on the dot chromosome of *Drosophila virilis*. After using RepeatMasker on the fosmid to eliminate repeat structures in my sequence, I proceeded to generate a Genscan prediction of possible genes in the fosmid. Initially, Genscan produced four predicted features (shown in Figure 1). After through investigation of each prediction using BLAST, Ensembl, and the UCSC Genome Brower, my final annotation of fosmid 2 includes a total of 3 genes: *pan* (pangolin, CG17964), CG32005, and *caps* (calcium-dependent secretion activator, CG33653). The genes *pan* and *caps* are only partially present in fosmid 2, with the 5' ends of each gene extending off the end of the fosmid. The final fosmid map is shown in Figure 2, with all of the intron/exon boundaries of each gene isoform defined.

In addition to formulating these gene models, I also performed a ClustalW analysis on the peptide sequence of the *pan* protein among the orthologs of various species. This ClustalW analysis revealed highly conserved regions in the protein, which may indicate functional domains. I also ran ClustalW for the 5' upstream region of CG32005 with two other species of *Drosophila*. Conservation in this alignment indicated upstream regulatory and untranslated regions common to all three species. Furthermore, analysis of repetitive elements in the fosmid was done by looking at the RepeatMasker output. This analysis revealed numerous LINEs and DNA elements within the fosmid. Last of all, the synteny of the genes on my fosmid relative to the genes on *Drosophila melanogaster* was determined. The orientation of the *D. melanogaster* genes was collected using the UCSC Browser, then matched to the genes on fosmid 2.

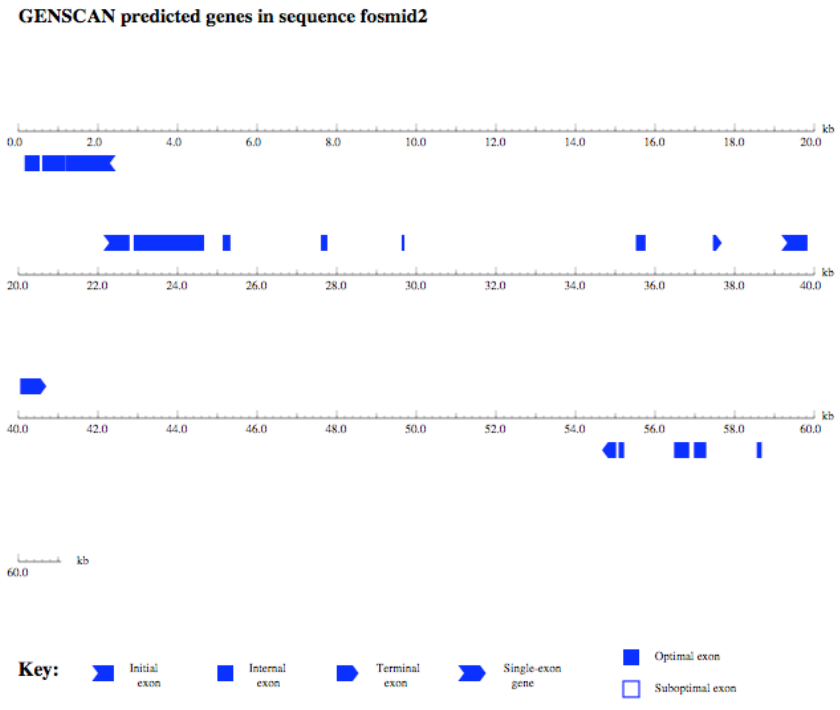

Figure 1: Initial Genscan predictions

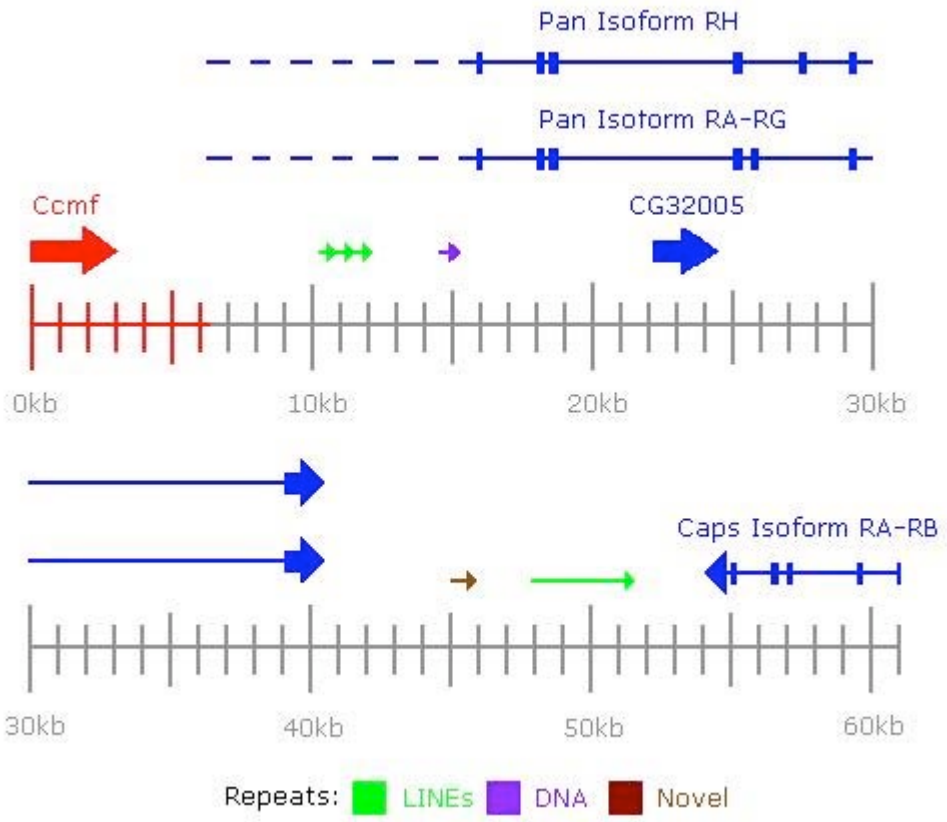

Figure 2: Fosmid 2 final map, with dashed lines indicating uncertainty of boundary

Gene Models

Feature 1

Looking at the first feature from my Genscan predictions, I first tried to determine which protein matched the prediction by running BLASTX with the region in question. Doing so produced matches to a cytochrome c biogenesis factor in *E. coli*, also known as *CcmF* (Figure 3). This was an extremely good match; an E value of almost 0 and an identity of 100% suggested that this gene did not in fact occur in *D. virilis*, but rather appeared in my fosmid as a result of *E. coli* contamination. This contamination most likely originated from the bacterial vectors that carried the *Drosophila* DNA, but how such a large segment of contamination (more than 3kb) got through the sequencing pipeline is a mystery. In addition, I checked to see if RepeatMasker caught the contaminated region and masked it. However, looking at the RepeatMasker output file revealed that RepeatMasker did not mask the region.

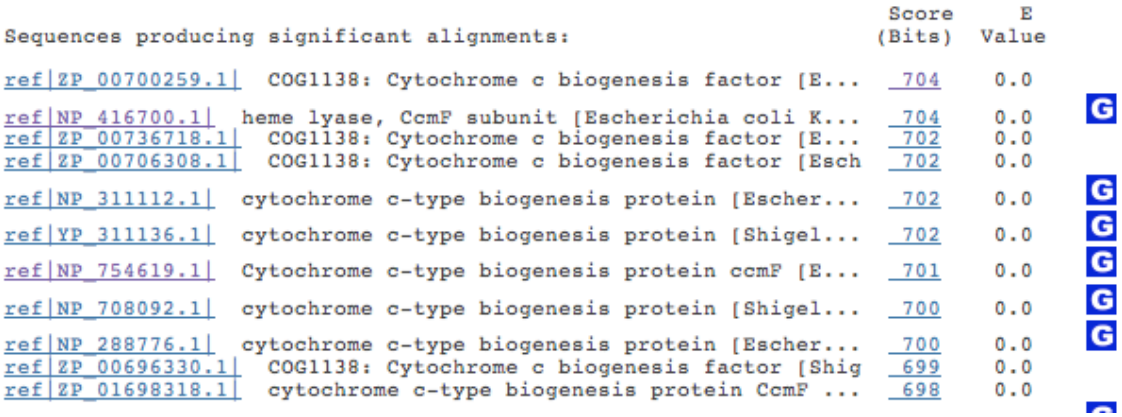

Figure 3: BLASTX of feature 1

With the discovery of this contamination, a large portion of my fosmid became unusable, since it could not be trusted. From looking at the UCSC Genome Browser, I was not confident in the fosmid sequence until position 6351, the point at which an alignment to the *D. melanogaster* genome starts (Figure 4).

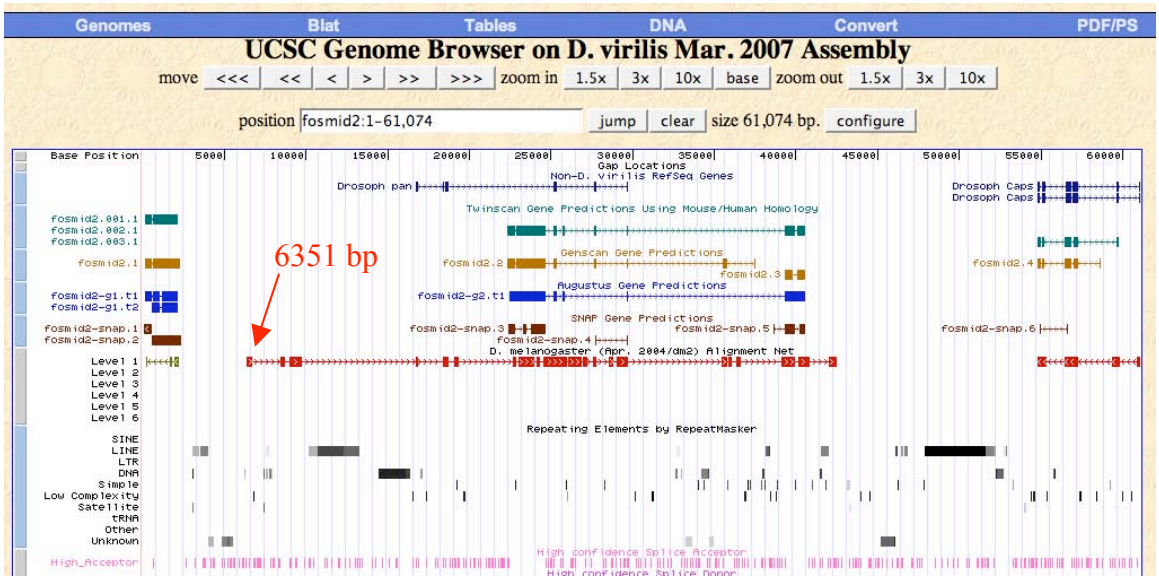

Figure 4: Position of *D. melanogaster* alignment

However, the only way to locate the exact starting point of the *virilis* fosmid would be to find the region of overlap between this fosmid and other fosmids in the entire assembly of the *D. virilis* genome.

Feature 2 & 3

Running a BLASTX analysis of feature 2 produced a high scoring hit to *pan* (pangolin, CG17964), a 11-exon gene with 9 different isoforms on *D. melanogaster*. Performing a BLAT alignment of the *D. melanogaster* peptide revealed that the gene

covered both features 2 and 3, suggesting that a Genscan misprediction interpreted the *pan* gene as two genes.

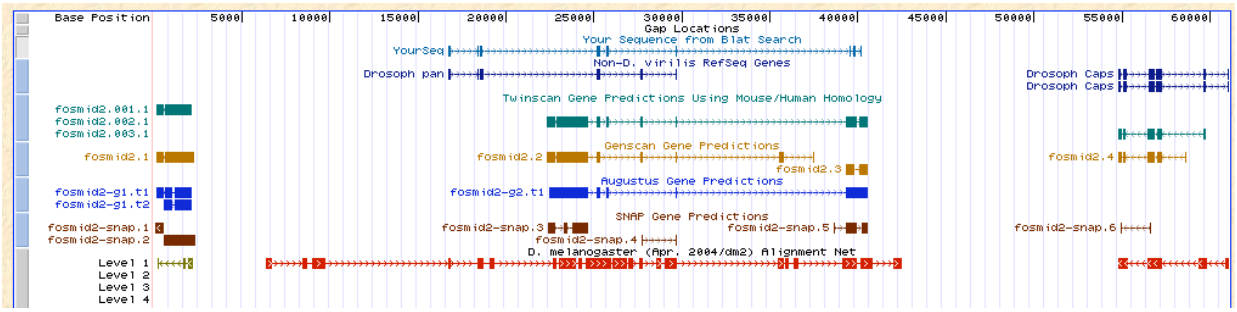

Figure 5: BLAT of *D. melanogaster* peptide against fosmid 2

However, this BLAT alignment did not align all of the exons to the fosmid. To see which exons were missing from the fosmid, a BLAST2seq alignment was carried out between the fosmid sequence and the *D. melanogaster* peptide sequence. The result showed high scoring alignments for the last 8 exons of isoforms RA-RH, but the first 3 exons were not found. In addition, the last exon of isoform RI was also not found in the fosmid.

To address these questions, I first looked at the orientation of the gene relative to my fosmid. Seeing that exon 4 was relatively close to the end of my sequence confidence boundary, I hypothesized that the first 3 exons extended off my fosmid and thus could not be found in the BLAST2seq alignment. To confirm this, I looked at the intron and exon information listed by Ensembl. What I discovered was that the intron between exons 3 and 4 was 21,052 base pairs in length, which extended well beyond my fosmid in the 5' direction. Thus it seems very probable that my fosmid only contains exons 4 to 11.

Next, to determine the cause of the absence of the last exon in isoform RI, I compared the intron and exon sequences of isoforms RI and RH, which are quite similar in *D. melanogaster*. After looking through the exonic and intronic sequences, I found that isoform RI was a variation of isoform RH. This variation in isoform RI was an extention of exon 10 past the exon boundary defined by isoform RH and into the neighboring intronic region. Looking at Figure 6, the outlined red intronic region in isoform RH is identical to the outlined red exonic region in Isoform RI. However, when investigating this region in my *D. virilis* fosmid, I found that this region was littered with stop codons, signifying that isoform RI could not be a viable isoform in *D. virilis*. As a result, it can be concluded that isoform RI is not a putative isoform in *virilis*, and will not be annotated as such.

Annotation of the exons of isoforms RA to RG was simple, since these isoforms differed only in their 5' untranslated regions. Because annotation of UTRs are beyond the scope of this paper, I have grouped isoforms RA-RG together. The exon/intron boundaries of isoforms RA-RG, and isoform RH can be found in the appendix at the end of this paper.

a) isoform RH

```
3,206 gtaagatatttaaaaaaatttttat.........aacatcgcgtttatcccactcccag

  154 TGGCATGCCTTGGGGCGTGAGGAGCAGGCGAAGTATTACGAGTTGGCGCGAAGGGAGCGA
      CAACTGCACATGCAAATGTATCCTGATTGGAGCTCTCGTACAAATGCCTCTCGCGGCAAA
      AAACGGAAGCGGAAGCAAGATACCAATGATGGAG

1,054 gttagatacattctcttaaactaag.........attataattgcctctttttccttag

   70 GCAATAATATGAAAAAATGCCGCGCCCGTTTCGGACTAGATCAACAGAGTCAGTGGTGCA
      AACCGTGCAG

7,927 gtatgttttaccctatttttgttt.........ttatgtgttcttctttttacatag

1,402 GCGCAAAAAGAAATGTATTCGTTATATGGAAGCCCTGAACGGCAATGGGCCCGCCGAAGA
      CGGCAGCTGTTTTGATGAGCACGGAAGTCAGCTAAGTGATGACGACGAAGATGACTACGA
      TGATGATAAACTAGGCGGAAGTTGTGGAAGCGCCGACGAAACCAATAAAATAGAAGATGA
      GGACTCGGAATCCCTTAATCAATCTATGCCAAGCCCAGGCTGTCTTAGTGGATTGTCCAG
      TTTACAGAGCCCATCGACAACAATGAGCTTGGCAAGCCCACTTAACATGAATGCCAATTC
      AGCAACCAATGTTATATTTCCTGCTTCTTCTAATGCACTTTTAATTGTCGGCGCAGACCA
      GCCGACAGCACAGCAACGACCCACATTGGTGTCAACCTCGGGATCGAGCAGCGGTTCAAC
      CAGTAGCATAAGTACAACCCCAAATACGTCGAGTACAGTTTCGCCAGTTACATGTATGAC
      CGGTCCGTGTCTCGGTTCATCTCAGGAACGAGCCATGATGCTTGGAAATCGGTTCAGTCA
      CTTGGGAATGGGGCTAAGTCCACCAGTAGTTAGCACGAGCACCAGTAAATCTGAACCATT
      TTTTAAGCCTCATCCCACAGTTTGCAATAATCCTATATTTGCATTGCCATCAATTGGTAA
      TTGTAGTTTAAATATTTCTTCAATGCCAAACACATCCCGAAACCCTATTGGTGCTAACCC
      ACGAGATATTAATAATCCCCTTAGCATCAATCAGCTGACTAAAAGACGTGAATATAAAAA
      TGTTGAGTTAATTGAAGCTAGTGAGTCAAAGACTATAGTTGCCCATGCCGCTACATCCAT
      TATTCAACATGTAGCAGTGAACGGCTATCATGCTAATCACTCGCTTTTAAATAGCAACTT
      GGGCCACCTTCATCACCAATTAAATAATCGTACAGAAAACCCGAATAGAAGTGAGCAGAC
      AATGCTGTCCGTAAGTAATCATTCTGTAAATAGCAGTGAATGCCATAAAGAATCTGATTC
      GCAGGCTATTGTATCTAGCAATCCTCCAAACGCTGGATCTTCCGATAACGGCGTTATTAG
      CGTTTCATAACTGATCGCCATGGATTTGTAGAATTGATTAATCAAATTTTAATGTTAATG
      TGTGAGAAAATAATTAAAATAAAAACATAAGTGCAATAATTATTATTCTCGCCAGCCCAA
      AGCCATCTTATGAAACAAAAGGCCATCAAAACTAGTGCCACGACTCACATCGCAATATTT
```

b) isoform RI

```
  154 TGGCATGCCTTGGGGCGTGAGGAGCAGGCGAAGTATTACGAGTTGGCGCGAAGGGAGCGA
      CAACTGCACATGCAAATGTATCCTGATTGGAGCTCTCGTACAAATGCCTCTCGCGGCAAA
      AAACGGAAGCGGAAGCAAGATACCAATGATGGAG

1,054 gttagatacattctcttaaactaag.........attataattgcctctttttccttag

2,015 GCAATAATATGAAAAAATGCCGCGCCCGTTTCGGACTAGATCAACAGAGTCAGTGGTGCA
      AACCGTGCAGGTATGTTTTACCCCTATTTTTGTTTACAATGCTTTACATTTTATTGCAAA
      TGGATTTACAAGATTCACGCAAGAATTAATTGACAAAAAAATTAAATTTTCATATTCGCT
      GTTTTCATAAAAACATTTAAAAAATCCTCGTGACCCTAAAACATGGCTACTCTGTAACTA
      GTGATAACTATATTTAAAAGGAAATTTTTCTGTTTGTTCATTAATTATTTAAACATTATA
      GTTTTCTGAGTTATGATTCATGAAAAAGGGTGAGGGTATAAGAAACCATATTTCGTTATC
      TATCGTAAAATACCCTAACATATCGTTGAACGTATTTTTCTTGAACTATCTAAACAAATT
      TTGCAAAACTTTTCATCAATTTATAAATGAATTTCGACAAATTAAATTCTTATGTGACGA
      TCTAATAAAAACATTACATGGACACTAAAAAAACGGTTGGAACATAAATTTTAACCAAAT
      AAGCAAACAAGGGCTTTACGACTCAACCCGGCTGTTGATACTGATAAAGAATATATAAAT
      TTTATATGGTCGCAAAGGTTTTCTTTATTACGTTGCAGACTTCAGAAAAAAGTCAATATA
      TAGTCAAAAATGGTATTAAAATGTTTAAGATGCAATGCAATAAGAGTTTAGCAAGAGAGA
      ATGCTATAGTCAATTTTCCCGTTACTCAGTAAAATTGAGAGAGTGGAATTGCACAATTTT
      TCTGGGATATCAATAGATATTGGTCAATAAAATAAAAAAAAAAAATTGTTTAAAAGTGTGG
      GAGTGGTGGGGCGTGACCAAAGCGTTTTCGGCAAATCGAGACAAGCCTAATAAAAATATG
      AAAAAATAACAAATTCAATTCAATCATTTTCAAACGGTTTGTGGGCGAAAGTGTTTTGAC
      ACATTTTGACACACCAATTCGAACGCCCAAAACTGCCACGCCCAGACTTTTAAAAAATGT
      ATTAATATTTTTTCACATTTTATTAGATTTGTAAATTGCCACGTTTGTAAAGATTTGTAA
      AAATTAGGCTGCGCTAACTCTAAACGGCCATTTTTTCTCATTTCATTCCTAACATCTATC
      GATATCCCAGAAAAAATATGAAATTTCGCGTTCGCATTCACACTAGCTGAGTAACGGGTA
      TCTGATAGTCGGGGAACTAGACTATAGCATTCTGTCTTGTTTTATTTATTTTATAAATAA
      TAACTTTAAAGAATGTGTGCTTGTATTATTTTAAAATGTATTTTCTTTTTTTCAATTTAT
```

(c)

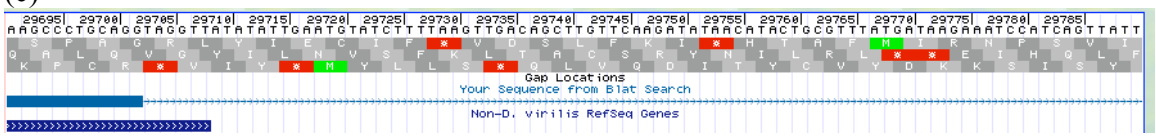

Figure 6: Similarities between the two outlined regions indicate splicing variation in *D. melangaster* (a and b), and stop codons in *D. virilis* fosmid (c)

CG32005

To investigate my fosmid further, I looked into the region that showed alignment between my fosmid and the *D. melanogaster* genome. When I clicked "Open D. melanogaster browser at position" in the UCSC Browser, I discovered a small gene nested in between two *pan* exons in the *melanogaster* genome (Figure 7). This gene, CG32005, was a one exon gene with one isoform.

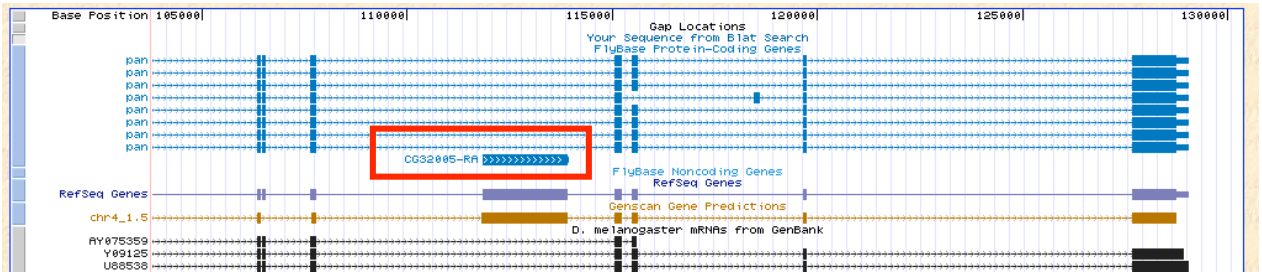

Figure 7: CG32005 gene, outlined in red

To determine whether or not this gene was in my fosmid, I ran a BLAST2seq alignment between my fosmid sequence and the *melanogaster* peptide. The resulting alignment was highly conserved, matching 678 amino acids out of 687. However, determining the exact location of the start codon of this gene was difficult, since the region in which the start codon should be contains three start codons, all of which are possible candidates for the beginning of the gene. These three start codons can be seen in Figure 8.

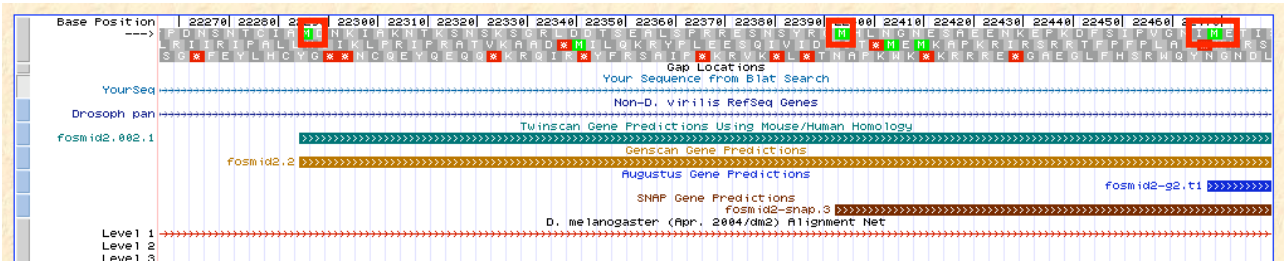

Figure 8: locations of the three putative start codons

In order to pick the correct start codon for this gene, I first checked to make sure there were no stop codons before any of them. When I found no stop codons, I checked for sequence similarity between sequences before the start codons and *melanogaster* sequences. By running a BLAST2seq alignment between all of the possible combinations, I found that the middle start codon had the best match and the highest similarity. Thus, I used this start codon in my final annotation. The BLAST2seq alignment for these start codons can be shown in Figure 9.

(a)

```
 Score = 12.3 bits (20),  Expect = 8770860, Method: Composition-based stats.
 Identities = 4/7 (57%), Positives = 6/7 (85%), Gaps = 0/7 (0%)

Query  1   MHLNGNE  7
           ++LN NE
Sbjct  11  INLNSNE  17
```

(b)

Sequence 1: lcllseq_1
Length = 25

Sequence 2: lcllseq_2
Length = 687

No significant similarity was found

Figure 9: BLAST2seq of start codons in CG32005. Similarity found in middle codon (a) versus no similarity in upstream codon (b).

Feature 4

The last feature of my fosmid occurs at the 3' end of my fosmid. To begin the investigation, I first used the region predicted by Genscan to perform a BLASTX search of putative genes. What I discovered from this BLAST search was a match to a *caps* (calcium-dependent secretion activator, CG33653) protein in *D. melanogaster*. An Ensembl search for this gene revealed that the gene contains 23 exons. However, due the fact that the gene was located near the end of my fosmid, only 6 exons could be annotated (exons 18-23), with one exon truncated at the 3' end of the fosmid. This is feature is shown in Figure 10.

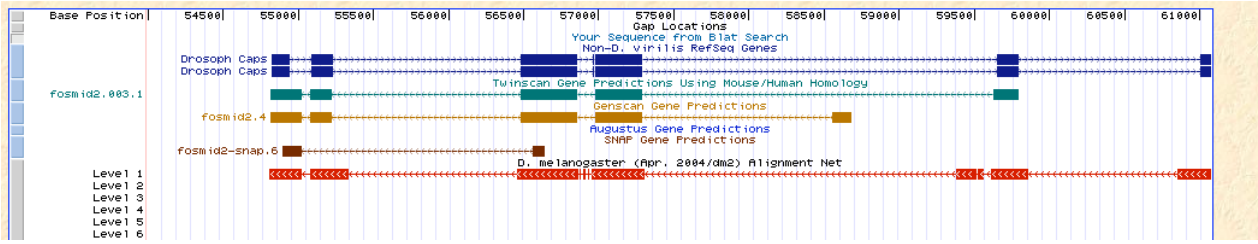

Figure 10: *caps* gene

The *caps* gene in *melanogaster* contains 3 isoforms, RA to RC. However, because isoform RC is shorter and ends before exon 18 of isoforms RA and RB, it does not exist in my fosmid and will not be included in this annotation. In addition, exons 18-23 of isoforms RA and RB are identical and cannot be distinguished in this fosmid. Thus, the two isoforms are grouped together in this annotation.

To find the intron and exon boundaries of the *caps* gene in my fosmid, I again used the *melanogaster* peptide sequence and the DNA sequence from my fosmid in a BLAST2seq alignment. The resulting alignment displayed high identity, as shown in

Figure 11. Based on this data, I was able to produce a good gene model for the *caps* gene in *D. virilis*.

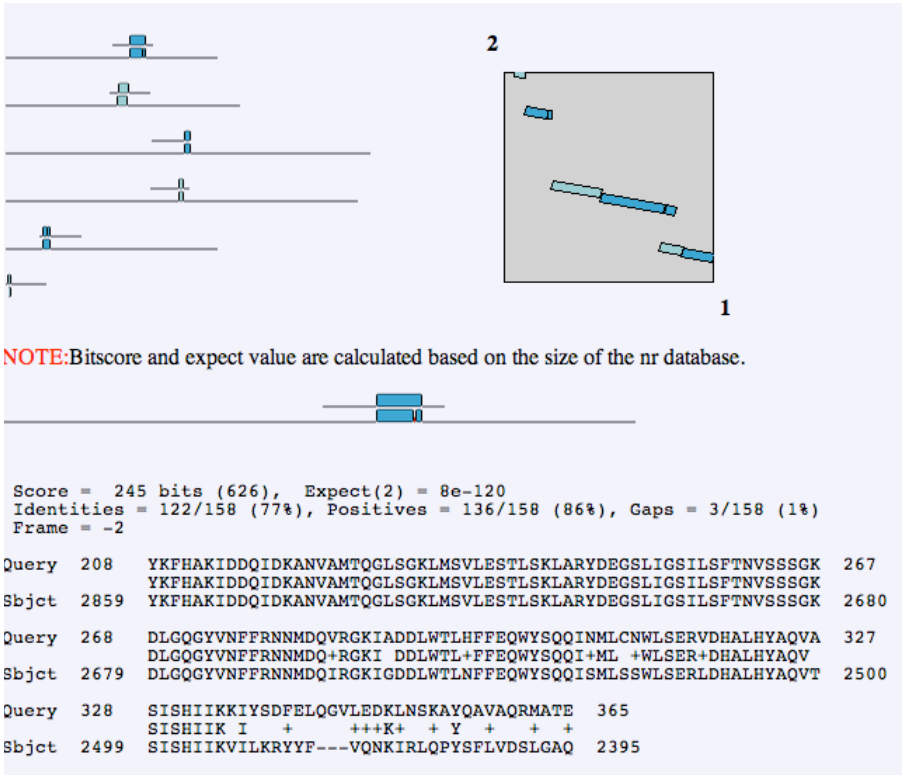

Figure 11: BLAST2seq alignment of *D. melanogaster caps* sequence to fosmid sequence

Summary of genes and functions:

***Pan*, CG17964** – codes for the *pangolin* protein, which functions as a transcription regulator and induces the Wingless pattern in embryonic and adult cells. Also plays a role as a segment polarity protein.

CG32005 – molecular function unknown. This gene was annotated as a part of the *D. melanogaster* genome project done in Berkeley.

***Caps*, CG33653** – codes for a calcium binding protein that functions in neurotransmitter secretion, synaptic vesicle exocytosis, transmission of nerve impulses and other vesicle-mediated transport mechanisms.

ClustalW

In my ClustalW analysis, I used the CG32005 gene for the 5' upstream region comparison since this was the only gene in my fosmid with a 5' region. To perform this comparison, I extracted a 2kb region upstream of the CG32005 start codon and compared this to the orthologous regions of two other *Drosophila* species I found to contain the CG32005 gene. Because this gene is newly discovered and relatively unknown, I was

only able to find the CG32005 gene in *D. melanogaster* and *D. pseudoobscura* using methods such as a BLAT alignment with the *melanogaster* peptide in the UCSC Browser, a BLASTP search for similar genes, and a Flybase search. Once I found the genes for both *melanogaster* and *pseudoobscura*, I extracted the same 2kb upstream sequence of the start codons and performed the ClustalW alignment. To make sure the *D. pseudoobscura* gene was correct, I checked for synteny by comparing the position of the gene to *melanogaster*. The results showed a nested gene within *pan*, which was consistent with all three *Drosophila* species.

The resulting Clustal alignment produced two regions in particular that contained high conservation. These two regions are shown in Figure 12. Since data from Ensembl combined with EST data of the 5' region does not show an upstream UTR region in *melanogaster*, these two regions of conservation are most likely promoter regions or other upstream regulatory regions for CG32005. However, to ascertain that these two regions are in fact functional, additional experiments must be done before these intergenic sequences can be annotated. Experiments such as ChIP assays would be greatly beneficial in determining the presence of transcription factor binding sites in these regions. An alternative experiment could find out whether these regions influence gene expression by placing these sequences upstream of a reporter.

A)

```
dp3_dna         AGAGATGTAATGAGTTTATGTCATTACAAATAAACAATTAAAT-TAATTATTTCATG--C 1240
dm2_dna         AATTTTGCGAAAAATCTATCTAA--AGGAATCAATGCCTAAAAATAGTTGAACAATTTCC 1260
Dvir4_dna       ---TTTAAAAAAAATCCACACGAG-AAGAATATAGATTTAC---TAACTACTCAGT---C 1268
                     *   *  * *  *    *  *  ***  *    **    **  *      *   *

dp3_dna         GATTGAAAAAGTAT-GAGAATATGAAATGTTAAATACATTATTACATTACGAAGTATTAA 1299
dm2_dna         GGAAAGAACAATTT-ATGAACAC-AAATGAAAAAAATAATCTT-TATT-CGATTTGTCCA 1316
Dvir4_dna       AAGGAAATTACTTCAGTAAATCTCAATTGTGGCAGAGGGTGGTGTAGG-TCAAGAATAAA 1327
                      *  * *      **    ** **    * *   *  *  *     *    *  *

dp3_dna         TAATAATAATAATTATTAAATAA-AGACTCCATATAATAAATGT---TGCTTTTCGGCTA 1355
dm2_dna         TTAGAACACAAGTAGAATAATAACACGAAGTGTATGATGAAATCGAGTGTTTACATATAA 1376
Dvir4_dna       AATTCATCATAGCAGGTCAA-AGTGGAATCTATACAAATACAAATTTGGCTCTCCCATAG 1386
                     *    *       ** *          **  *  *        * *

dp3_dna         ATTT-TTGTTATTTATTAAATGCAAATCTTAAAATATTTGAGCAAAAAGAAATATGGGTT 1414
dm2_dna         AGTG-TATTTATAAATTAAATATAAGTTTTAAAA-AT---AGGATGTCATCTTTTGTTTC 1431
Dvir4_dna       ATTTATTTTTATTTGTTGTGC-CAAAACCCAAAGAAT-----CAAAATAGAATATGGCAG 1440
                * *  *  ****   **      **     ***  **      *        * **
```

B)

```
dp3_dna         ATCGGCTTCGGCCTCGGTCTCTCCAGCGTGGAGT--GT-CGATAAAAAACGACGTTCACA 325
dm2_dna         ACGAACATAATTATGAGCAAGTAAACAGTTGAATTGGC-CAATTAAGATTAGTGATAAGA 327
Dvir4_dna       AGGAGTGTGCCCTTCCACAACTCATATTTGGGACTACTACGACTACAATTCATACAAATA 358
                *     *    *       *     * *       * *  *  *         * *

dp3_dna         GTCATCGTAAACTTTTAATTTTAAGGCGAGTGTAATTCGTCAAAATGTTGTG----GCAG 381
dm2_dna         AGTTTTACACATCTTTAATTTAGAATCAACCG--ATTTTTTCAACCGTTTTC----AGTG 381
Dvir4_dna       ACTGTGCTTAACCTGCA--TTTGATGTATATCCCATTAATAAAATTTCTATATAAAATGG 416
                    *     *  *  *  **  *          ***  *  **    * *        *

dp3_dna         TATATACACATTTCTGAAAATCGTTATAA---CCAAAACCTATAGGGCATTTAAAACCTT 438
dm2_dna         AATAAATAAATTTCTAATATATATCCTCG---AAAAATGTAATTTTCAAGTACTAATTTA 438
Dvir4_dna       AGTGCGATAATGTGCCAAAATCATTAATAATACAAAATTCAAAATTGAATTTTCAATTTG 476
                  *      ** *   * *    *          ***    *      * *   **  *

dp3_dna         TGAAGC-AAATAAACT-ATAATTTAATTGAAAC--AAAT-TCAAATTATACCCGTACACA 493
dm2_dna         TAAA---AAATATTTTGATAATATTTTTTAAATCAAAATGTTAAATTATTTGCTTTAAGT 495
Dvir4_dna       GACAGCTGTATCATTTTGTATATTGGTGAAAAT--AGTTTGAAAATAATGCTCGGCCATA 534
                   *     **    *  **   *  *  ***   *  *   **** **   *    *
```

Figure 12: Two conserved regions upstream of CG32005

The second ClustalW analysis involves the comparison of a protein of a *virilis* gene relative to orthologous genes of various other species. To perform this analysis, I chose the *pan* gene. To find the *pangolin* protein in other species, I used the Ensembl predictions to guide my decisions. However, many of the predictions made by Ensembl either did not have enough similarity to the *virilis* protein to be useful, or were not fully discovered and existed only as a fragment in the database. A search using Expasy produced similar results, but I was able to find 4 other species with suitable *pangolin* proteins for the ClustalW analysis. Species I used in this analysis are: *Drosophila melanogaster, Aedes aegypti, Tribolium castaneum, Oncopeltus fasciatus, and Drosophila simulans*. Matches to only insect species suggests that this protein is important in insect function, but that this function has been lost in other organisms.

The ClustaW alignment showed two specific regions of conservation that were particularly significant (Figure 13). These two regions suggest functional domains within the protein that have been conserved throughout evolution. As expected, *Drosophila* species are closer in relation to the other species, as shown in the cladogram in Figure 14.

A)

```
[ Dmelanogaster]      --------PFFCHNADPLSTPPPAHCGIPPYQLDPKMG----LTRPALYPFAGGQYPYPM 158
[ Dvirilis]           ----------------------------------------LTRPALYPFAATQYPYPM 18
[ Aedes               --------PFFCHNGDHLSSPPPAHCGILPYQLDPKT---MGLTRPPLYSFPTSQYPYPM 154
[ Tribolium           TGHPGAALPFFCHNGDPLSQPPPAHMGIPPYQLDSKTAGSMGLTRPPMYPFPAGQYPYPI 176
                                                          ****.:*.*.  *****:

[ Dmelanogaster]      LSSDMSQVAS-WHTPSVYS---ASSFRTPYPSSLPINTTLASDFPFRFSPSLLPSVHATS 214
[ Dvirilis]           LSPDMSQVAS-WHTPSVYS---ASSFRTPYPSSLPINTTLPSDFPFRFSPSLLPSVHATS 74
[ Aedes               LSPDMPLVASSWHTPSMY--SPATGFRNPYPASLQIYTS-LPSDFYRYSPSLLPSVHS-- 209
[ Tribolium           LSPEMSQVAASWHTPSMYPLSPGAGFRSPYPSALPISTSSLPSDFYRFSPTGLIPPHPGL 236
                      **.:*. **: *****:*    .:.**.***::* * *:  ..  :*:**: * . *.
```

B)

```
[ Dmelanogaster]      MRAKVVAECTLKESAAINQILGRRWHELSREEQSKYYEKARQERQLHMELYPGWSARDNY 60
[ Aedes               MRSQVVAECTLKESAAINQILGRKWHSLTREQQSVYYDKARQERQLHMELYPGWTARDNY 60
[ Tribolium           MRAKVVAECTLKESAAINQILGRRWHALGREEQAKYYELARRERQLHMQLYPDWSSRAN- 59
[ Oncopeltus          MRAKVVAECTLKESAAINQILGRRWHALGREEQAKYYELARRERQLHMQLYPDWSSRAN- 59
[ Dvirilis]           VTCQKRKSVKKTDRQRIREVLHKIWHALGREEQAKYYELARRERQLHMQMYPDWSSRTN- 59
                      : .:   . . .:   *.::* : ** * **:*: **: **:******::**.*::* *

[ Dmelanogaster]      GYVSKKKKRKKDR-STTDSGGNNMKKCRARFGLDQQSQWCKPCRRKKKCIRYMEALN--- 116
[ Aedes               GYGAKKKKRKKDR-SPADPGGNSMKKCRARYGLDQQNQWCKPCRRKKKCIRYKEMTDSEE 119
[ Tribolium           --ATRGKKRKRKQ-DPAD-GGNSMKKCRARYGLDQQNQWCKPCRPREQLTWRR------- 108
[ Oncopeltus          --TTRGKKRKRKQ-ETND-GANSMKKCR------------------------------- 83
[ Dvirilis]           --ASRGKKRKRKQDASSDGGGNNMKKCRARFGLDQQNQWCKPCR--------------- 101
                        :: ****:.:   . * *.*.*****
```

Figure 13: Conserved regions in *pangolin*

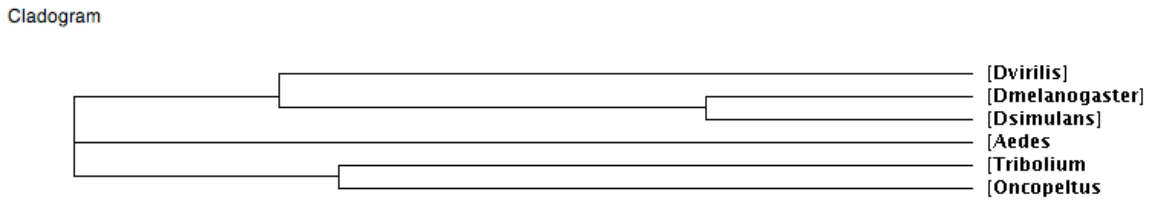

Figure 14: ClustalW cladogram

Repeats

Repetitive elements comprise 30.42% of my entire fosmid. A table of the major repeat families is shown in Table 1.

	number of elements*	length occupied	percentage of sequence
SINEs:	0	0 bp	0.00 %
ALUs	0	0 bp	0.00 %
MIRs	0	0 bp	0.00 %
LINEs:	19	9973 bp	16.33 %
LINE1	0	0 bp	0.00 %
LINE2	0	0 bp	0.00 %
L3/CR1	0	0 bp	0.00 %
LTR elements:	0	0 bp	0.00 %
MaLRs	0	0 bp	0.00 %
ERVL	0	0 bp	0.00 %
ERV_classI	0	0 bp	0.00 %
ERV_classII	0	0 bp	0.00 %
DNA elements:	14	3956 bp	6.48 %
MER1_type	0	0 bp	0.00 %
MER2_type	0	0 bp	0.00 %
Unclassified:	0	0 bp	0.00 %
Total interspersed repeats:		13929 bp	22.81 %
Small RNA:	0	0 bp	0.00 %
Satellites:	4	288 bp	0.47 %
Simple repeats:	20	959 bp	1.57 %
Low complexity:	24	850 bp	1.39 %

Table 1: Summary of repeats

To detect any significant repeats, I looked for any masked repeat in the RepeatMasker output that had a length of more than 500 base pairs. These repeats are outlined in Table 2.

Table 2: List of significant repeat elements

Pos Begin	Pos End	Matching Repeat	Repeat Class	Length
3057	3462	PENELOPE	LINE/Penelope	405
10180	10888	dvir.0.42.centroid	LINE	708
10679	11633	dvir.0.5.centroid	LINE	954

11632	12318	dvir.0.85.centroid	LINE	686
12323	13241	dvir.0.85.centroid	LINE	918
14675	15504	dvir.14.34.centroid	DNA	829
45174	46009	dvir.13.51.centroid	Novel	835
47900	51610	dvir.0.85.centroid	LINE	3710

One important repeat to note is the LINE at position 47900-51610. Because this LINE is so long (3.7kb) it could have very likely played a role in some translocation event in the *virilis* genome. To find out what kind of event might have occurred, it would be useful to discuss the synteny of the genes in fosmid 2 relative to the genes in *D. melanogaster.*

Table 3: Repeats

pos begin	pos end	matching repeat	repeat class/family	length
3010	3072	dmoj.0.99.centroi	Satellite	62
3018	3090	dvir.16.2.centroid	DNA	72
3057	3462	PENELOPE	LINE/Penelope	405
3505	4022	dvir.16.17.centroid	LINE	517
4028	4368	dvir.11.33.centroid	TRF	340
4852	5157	dvir.11.33.centroid	TRF	305
5158	5344	dvir.11.33.centroid	TRF	186
5342	5525	dvir.11.33.centroid	TRF	183
6209	6292	dvir.16.2.centroid	DNA	83
6803	6852	AT_rich	Low_complexity	49
7367	7431	dmoj.0.99.centroi	Satellite	64
7377	7450	dvir.16.2.centroid	DNA	73
7541	7684	dvir.16.2.centroid	DNA	143
7565	7767	PENELOPE	LINE/Penelope	202
7711	7912	dmoj.8.25.centroid	DNA	201
7763	7908	dvir.16.2.centroid	DNA	145
10180	10888	dvir.0.42.centroid	LINE	708
10679	11633	dvir.0.5.centroid	LINE	954
11632	12318	dvir.0.85.centroid	LINE	686
12323	13241	dvir.0.85.centroid	LINE	918
14433	14674	dvir.16.2.centroid	DNA	241
14675	15504	dvir.14.34.centroid	DNA	829
15503	15861	dvir.15.21.centroid	DNA	358
15862	16072	dvir.16.2.centroid	DNA	210
16067	16365	dvir.16.2.centroid	DNA	298
16492	16526	AT_rich	Low_complexity	34
16974	17103	dvir.16.2.centroid	DNA	129
17353	17383	AT_rich	Low_complexity	30
19241	19272	(TA)n	Simple_repeat	31
19617	19654	AT_rich	Low_complexity	37
19679	19715	AT_rich	Low_complexity	36
22825	22859	(CAG)n	Simple_repeat	34
25922	25947	(TTTTTG)n	Simple_repeat	25

25982	26056	A-rich	Low_complexity	74
28247	28276	(TATG)n	Simple_repeat	29
30133	30178	AT_rich	Low_complexity	45
31158	31182	AT_rich	Low_complexity	24
31215	31238	AT_rich	Low_complexity	23
32582	32728	dvir.16.2.centroid	DNA	146
32684	32867	dvir.16.17.centroid	LINE	183
32951	33019	dvir.16.2.centroid	DNA	68
33220	33379	dana.5.140.centroid	TRF	159
33252	33621	dvir.11.23.centroid	TRF	369
34026	34083	(TA)n	Simple_repeat	57
34184	34336	dvir.16.2.centroid	DNA	152
34337	34370	(TCCG)n	Simple_repeat	33
34371	34594	dvir.16.2.centroid	DNA	223
34599	34663	dvir.16.2.centroid	DNA	64
34672	34938	dvir.11.33.centroid	TRF	266
35781	35827	(TGG)n	Simple_repeat	46
36823	36858	AT_rich	Low_complexity	35
36907	36935	AT_rich	Low_complexity	28
37044	37084	(TATATG)n	Simple_repeat	40
37159	37206	(TTG)n	Simple_repeat	47
37855	37914	(TTATA)n	Simple_repeat	59
37931	38080	dvir.16.2.centroid	DNA	149
38081	38115	(TCTG)n	Simple_repeat	34
38116	38382	PENELOPE	LINE/Penelope	266
38349	38412	dvir.16.17.centroid	LINE	63
38417	38458	AT_rich	Low_complexity	41
38764	38800	AT_rich	Low_complexity	36
38812	38843	(TTA)n	Simple_repeat	31
39961	39991	(CAG)n	Simple_repeat	30
40991	41032	(CATA)n	Simple_repeat	41
41473	41560	dvir.16.2.centroid	DNA	87
41527	41797	PENELOPE	LINE/Penelope	270
41798	41827	(CGGA)n	Simple_repeat	29
41828	41976	PENELOPE	LINE/Penelope	148
43132	43316	(TATATG)n	Simple_repeat	184
43266	43359	dmoj.0.64.centroi	Satellite	93
43908	43936	AT_rich	Low_complexity	28
44479	44501	AT_rich	Low_complexity	22
45174	46009	dvir.13.51.centroid	Novel	835
46007	46049	dvir.13.0.centroid	FB	42
46050	46238	dvir.0.42.centroid	LINE	188
46240	46260	(TAAA)n	Simple_repeat	20
46426	46576	dvir.0.85.centroid	LINE	150
46604	46812	dvir.0.85.centroid	LINE	208
47806	47828	(CAGT)n	Simple_repeat	22
47845	47896	dvir.0.85.centroid	LINE	51
47900	51610	dvir.0.85.centroid	LINE	3710
51637	51873	dvir.0.85.centroid	LINE	236

51889	52093	dvir.0.10.centroid	LINE	204
52079	52209	dvir.0.5.centroid	LINE	130
52199	52272	dvir.16.2.centroid	DNA	73
52277	52686	dvir.16.2.centroid	DNA	409
52687	52833	dvir.0.85.centroid	LINE	146
52864	52956	dvir.0.14.centroid	LINE	92
53258	53343	(TATATG)n	Simple_repeat	85
53979	54095	dana.5.27.centroi	Satellite	116
54372	54407	AT_rich	Low_complexity	35
54541	54575	AT_rich	Low_complexity	34
54604	54632	AT_rich	Low_complexity	28
55347	55384	AT_rich	Low_complexity	37
55722	55816	dvir.16.2.centroid	DNA	94
55805	55835	dvir.16.2.centroid	DNA	30
57317	57346	AT_rich	Low_complexity	29
57426	57451	AT_rich	Low_complexity	25
58412	58432	AT_rich	Low_complexity	20
59230	59292	(TATG)n	Simple_repeat	62
59979	60021	AT_rich	Low_complexity	42
60507	60541	AT_rich	Low_complexity	34

Synteny

Looking at synteny in fosmid 2, there is a preservation of synteny in the region around *pan* and CG32005. The two genes occur on chromosome 4 in both species, and there is a conservation of gene orientation and placement. This leads me to conclude that there is synteny for the pan and CG32005 genes.

However, when looking at the *caps* gene, I found that the orientation of the gene has been reversed relative to the *melanogaster* gene. In addition, while the gene is found on the 4th chromosome in *melanogaster*, it is located more than 1 Mb away from *pan* and CG32005. To investigate this event, I looked at the flanking regions surrounding the *caps* gene in both *melanogaster* and *virilis*. What I found were large regions of LINES that existed at the 5' end of the gene in both species. These LINES were large (2-4kb), and could very well have participated in translocating and inverting the *caps* gene after divergence of the *melanogaster* and *virilis* species.

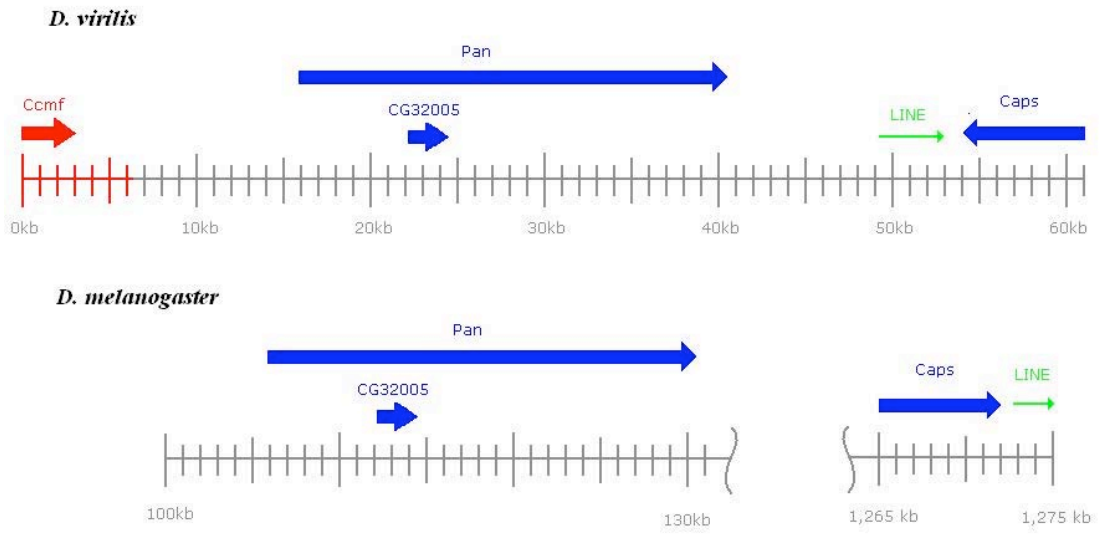

Figure 15: Synteny map

Appendix

Pan

Isoform RA-RG

Exon	Position (phase)
exon5:	(2)16744-16831(2)
exon6:	(1)18392-18472(2)
exon7:	(1)18539-18688(2)
exon8:	(1)25136-25328(0)
exon9:	(0)25690-25855(1)
exon10:	(2)29632-29701(2)
exon11:	(1)39297-40551(0)

>pan Isoform RA-RG
LTRPALYPFAATQYPYPMLSPDMSQVASWHTPSVYSASSFRTPYPSSLPI
NTTLPSDFPFRFSPSLLPSVHATSHHVLNSHPSIVTSNSKQDCGVQDSTT
NNRYSRNLDTKSSSNSQANDCKDSSNDKKKPHIKKPLNAFMLYMKEMRAK
VVAECTLKESAAINQILGRRWHELSREEQSKYYEKARQERQLHMELYPGW
SARDNYGYVSKKKKRKKDRSTADSGGNNMKKCRARFGLDQQNQWCKPCRR
KKKCIRYMEALHGNGALANGSGMDDAGNMSQLSDDDDDDEELGGASCGSG
DETETNKMADNDDTESMSQSLSSPGCLSGLSSVQSPSTTTSLASPLNMNM
LTSPATPALPSATSNIGPLPVNNSNEQAASSSQSRSAPGSGTGSSSGSTC
SISNTPNTSSTASPVTSATGTAPTSVSERAMMLGTRFSHLGMGLTLPVCG
SNPEQLFQSHTHLAAVSLAGSSGSGSSSSSTPTSLATNGFNSYTGSVTAG
GSIKTIVPNAAASVSAPAPATGPATSLHRNPIGANPRDINNPLSINQLTK
RREDNNIVILGSCDPQSASVILRHNAAHNPYTHTHPHPHPHPHHALFSSS
FSQHFQQQLNNHLSATSSSGSSIGSVVQPIDTPALNSMKRRNTSDSPAAT
GNSATPNATETGAISVS*

>pan Isoform RA-RG
GTCTTACTCGTCCTGCCTTATATCCATTTGCTGCCACCCAATATCCTTAT
CCGATGCTAAGTCCTGATATGTCTCAAGTAGCTTCATGGCACACGCCGTC
GGTTTACTCAGCATCTAGTTTTAGAACACCCTATCCCTCCTCATTACCAA
TTAATACAACGCTACCGAGCGACTTCCCATTTCGATTCTCACCGAGTCTG
TTGCCTTCCGTACATGCGACATCTCACCACGTTTTAAATTCCCATCCATC
GATTGTGACGTCTAATTCCAAACAGGATTGCGGCGTACAGGATTCCACAA
CGAACAATCGATATTCAAGAAACTTGGATACCAAAAGCTCATCAAATTCG
CAAGCAAACGACTGTAAAGATAGTTCAAATGATAAAAAGAAGCCACATAT
CAAGAAACCTCTGAACGCATTTATGCTGTACATGAAAGAGATGAGAGCCA
AAGTTGTAGCTGAATGCACCCTGAAAGAGTCAGCGGCGATTAATCAAATA
CTGGGGCGACGGTGGCACGAACTTTCCCGCGAAGAGCAAAGCAAATATTA
CGAAAAAGCGCGACAAGAGCGCCAATTGCATATGGAGTTGTATCCGGGGT
GGAGCGCACGAGATAACTACGGTTACGTGTCAAAAAAGAAAAAGCGTAAA
AAAGACAGATCGACAGCGGATTCGGGAGGTAACAACATGAAAAAGTGTCG
AGCGCGATTTGGACTGGACCAGCAGAATCAATGGTGCAAGCCCTGCAGAC
GCAAAAAGAAATGCATTCGCTACATGGAGGCATTACATGGAAACGGCGCG
CTGGCGAACGGTAGCGGCATGGATGATGCTGGCAATATGAGCCAGTTGAG
TGATGATGACGACGATGACGAGGAATTGGGTGGTGCCAGCTGTGGCAGCG
GGGATGAGACCGAGACAAATAAAATGGCCGACAACGACGACACAGAGTCA
ATGAGCCAGTCGCTGTCGAGCCCCGGGTGCCTGAGCGGACTGTCCAGTGT
GCAAAGTCCCTCGACAACGACGAGCCTGGCCAGTCCACTAAATATGAACA
TGCTAACTAGTCCAGCAACGCCCGCTCTACCTTCTGCAACCAGCAATATA
GGCCCTCTCCCTGTTAACAACAGCAATGAGCAGGCAGCGTCGAGCAGTCA
ATCACGATCAGCGCCTGGGTCTGGGACTGGGTCCAGCAGCGGATCAACAT
GTAGCATTAGTAACACACCTAATACCTCTAGCACAGCGTCGCCAGTGACG
TCTGCGACCGGAACAGCGCCGACCTCTGTCAGTGAGCGCGCCATGATGCT
TGGCACCCGCTTTAGCCATCTAGGCATGGGGCTTACTCTACCCGTATGTG
GCAGTAATCCAGAGCAACTATTCCAATCGCACACCCACCTGGCTGCTGTA
AGCCTCGCTGGAAGCAGTGGCAGCGGCAGCAGCAGCAGCAGCACTCCAAC
ATCTTTGGCAACAAATGGCTTCAACAGTTATACGGGCTCTGTAACAGCGG
GTGGCAGCATAAAGACAATAGTACCAAATGCTGCAGCATCTGTATCAGCG
CCGGCGCCGGCTACTGGACCAGCAACAAGCTTACATCGCAACCCGATTGG
CGCTAACCCACGTGACATTAACAATCCGTTGAGTATCAATCAGTTGACCA
AGCGGCGAGAGGATAATAACATTGTGATACTCGGCAGCTGCGACCCGCAA
TCAGCTTCAGTTATTTTGCGCCATAACGCGGCCCATAATCCATACACGCA
CACTCACCCGCATCCGCATCCTCATCCGCATCATGCGCTCTTCAGCAGCA
GCTTTAGTCAACACTTCCAGCAACAGTTGAACAACCATTTGTCAGCCACA
AGCAGCAGCGGCAGCAGCATCGGGTCAGTCGTGCAACCGATTGACACGCC
CGCCCTGAATAGTATGAAACGCCGCAATACATCCGATTCCCCAGCAGCGA
CCGGAAATAGCGCAACTCCCAACGCAACCGAAACTGGAGCCATTAGTGTT
TCG

Isoform RH

Exon	Position (phase)
exon5:	(2)16744-16831(2)
exon6:	(1)18392-18472(2)

exon7: (1)18539-18688(2)
exon8: (1)25136-25328(0)
exon9: (0)27604-27763(1)
exon10: (2)29632-29701(2)
exon11: (1)39297-40551(0)

>pan Isoform RH
GTCTTACTCGTCCTGCCTTATATCCATTTGCTGCCACCCAATATCCTTAT
CCGATGCTAAGTCCTGATATGTCTCAAGTAGCTTCATGGCACACGCCGTC
GGTTTACTCAGCATCTAGTTTTAGAACACCCTATCCCTCCTCATTACCAA
TTAATACAACGCTACCGAGCGACTTCCCATTTCGATTCTCACCGAGTCTG
TTGCCTTCCGTACATGCGACATCTCACCACGTTTTAAATTCCCATCCATC
GATTGTGACGTCTAATTCCAAACAGGATTGCGGCGTACAGGATTCCACAA
CGAACAATCGATATTCAAGAAACTTGGATACCAAAAGCTCATCAAATTCG
CAAGCAAACGACTGTAAAGATAGTTCAAATGATAAAAAGAAGCCACATAT
CAAGAAACCTCTGAACGCATTTATGCTGTACATGAAAGAGATGAGAGCCA
AAGTTGTAGCTGAATGCACCCTGAAAGAGTCAGCGGCGATTAATCAAATA
CTGGGGCGACGGTGGCATGCCTTGGGGCGTGAGGAGCAGGCCAAATATTA
CGAGTTGGCACGACGGGAACGCCAGCTGCACATGCAGATGTATCCCGATT
GGAGCTCACGCACGAACGCGTCACGCGGCAAGAAGCGGAAACGGAAGCAA
GATGCGAGCAGCGACGGAGGAGGTAACAACATGAAAAAGTGTCGAGCGCG
ATTTGGACTGGACCAGCAGAATCAATGGTGCAAGCCCTGCAGACGCAAAA
AGAAATGCATTCGCTACATGGAGGCATTACATGGAAACGGCGCGCTGGCG
AACGGTAGCGGCATGGATGATGCTGGCAATATGAGCCAGTTGAGTGATGA
TGACGACGATGACGAGGAATTGGGTGGTGCCAGCTGTGGCAGCGGGGATG
AGACCGAGACAAATAAAATGGCCGACAACGACGACACAGAGTCAATGAGC
CAGTCGCTGTCGAGCCCCGGGTGCCTGAGCGGACTGTCCAGTGTGCAAAG
TCCCTCGACAACGACGAGCCTGGCCAGTCCACTAAATATGAACATGCTAA
CTAGTCCAGCAACGCCCGCTCTACCTTCTGCAACCAGCAATATAGGCCCT
CTCCCTGTTAACAACAGCAATGAGCAGGCAGCGTCGAGCAGTCAATCACG
ATCAGCGCCTGGGTCTGGGACTGGGTCCAGCAGCGGATCAACATGTAGCA
TTAGTAACACACCTAATACCTCTAGCACAGCGTCGCCAGTGACGTCTGCG
ACCGGAACAGCGCCGACCTCTGTCAGTGAGCGCGCCATGATGCTTGGCAC
CCGCTTTAGCCATCTAGGCATGGGGCTTACTCTACCCGTATGTGGCAGTA
ATCCAGAGCAACTATTCCAATCGCACACCCACCTGGCTGCTGTAAGCCTC
GCTGGAAGCAGTGGCAGCGGCAGCAGCAGCAGCAGCACTCCAACATCTTT
GGCAACAAATGGCTTCAACAGTTATACGGGCTCTGTAACAGCGGGTGGCA
GCATAAAGACAATAGTACCAAATGCTGCAGCATCTGTATCAGCGCCGGCG
CCGGCTACTGGACCAGCAACAAGCTTACATCGCAACCCGATTGGCGCTAA
CCCACGTGACATTAACAATCCGTTGAGTATCAATCAGTTGACCAAGCGGC
GAGAGGATAATAACATTGTGATACTCGGCAGCTGCGACCCGCAATCAGCT
TCAGTTATTTTGCGCCATAACGCGGCCCATAATCCATACACGCACACTCA
CCCGCATCCGCATCCTCATCCGCATCATGCGCTCTTCAGCAGCAGCTTTA
GTCAACACTTCCAGCAACAGTTGAACAACCATTTGTCAGCCACAAGCAGC
AGCGGCAGCAGCATCGGGTCAGTCGTGCAACCGATTGACACGCCCGCCCT
GAATAGTATGAAACGCCGCAATACATCCGATTCCCCAGCAGCGACCGGAA
ATAGCGCAACTCCCAACGCAACCGAAACTGGAGCCATTAGTGTTTCG

>pan isoform RH
LTRPALYPFAATQYPYPMLSPDMSQVASWHTPSVYSASSFRTPYPSSLPI
NTTLPSDFPFRFSPSLLPSVHATSHHVLNSHPSIVTSNSKQDCGVQDSTT
NNRYSRNLDTKSSSNSQANDCKDSSNDKKKPHIKKPLNAFMLYMKEMRAK
VVAECTLKESAAINQILGRRWHALGREEQAKYYELARRERQLHMQMYPDW
SSRTNASRGKKRKRKQDASSDGGGNNMKKCRARFGLDQQNQWCKPCRRKK
KCIRYMEALHGNGALANGSGMDDAGNMSQLSDDDDDDEELGGASCGSGDE
TETNKMADNDDTESMSQSLSSPGCLSGLSSVQSPSTTTSLASPLNMNMLT
SPATPALPSATSNIGPLPVNNSNEQAASSSQSRSAPGSGTGSSSGSTCSI
SNTPNTSSTASPVTSATGTAPTSVSERAMMLGTRFSHLGMGLTLPVCGSN
PEQLFQSHTHLAAVSLAGSSGSGSSSSSTPTSLATNGFNSYTGSVTAGGS
IKTIVPNAAASVSAPAPATGPATSLHRNPIGANPRDINNPLSINQLTKRR
EDNNIVILGSCDPQSASVILRHNAAHNPYTHTHPHPHPHPHHALFSSSFS
QHFQQQLNNHLSATSSSGSSIGSVVQPIDTPALNSMKRRNTSDSPAATGN
SATPNATETGAISVS*

CG32005

Exon	Position (phase)
Exon1	(0)22393-24666(0)

>CG32005
ATGCACCTAAATGGAAATGAAAGCGCCGAAGAGAATAAGGAGCCGAAGGA
CTTTTCCATTCCCGTTGGCAATATAATGGAAACGATCTCGCAATCGTCAT
CATCATCATCGACAACAACAACTGTCCCGAATGCACAATTCTTTGATCAT
CAACAGACCACGAATGAGGCAGTTACCAAGGTCAACTCGGATGTGATGGA
GCACATATTTAATACACCCACAAGCAATGCGGTGGTGCCGACTACAACAT
CCACAATAATGTCGGAAATTCTATTGCGGCAGCAGCACGACATGCTCCAG
CACTGGCCAAAGCTGGAGGACCCTCAGCGCAAATGGAATCGGGAGCAATT
GGATACAAGCTGCAATCGTGATGCATTACTGGCGCGCATATACAACAATA
AAGTGCTCTCCCAGCACAAGAGCACACAACGCCAGCAACAGCAGCAACAG
CAGCAGCAGCAGCAGCAAAGACAGTTTCTAGCAACACAATTGCTTTATGC
ACAGTTTCTACATCAGCCCCATTTGTCATTGGAGGTCGATCAGGAGCGAG
AACGTAGGCTACATATATTGGAGTTTTCTGGTTTGCGAAAATCTTCCGTT
AGCAATGAGCCAAAAGCAACACCATTATTGTCAACATCATCACTATCAAT
TGGCTTATCACACGATAACAACAATAATCGAACACAAAGCGATATCGAAC
AGAGTGTTCGATATGATCACGACCACATCCGGAGAGATGGAACAGATGCA
CACAACTCGAATTCGAACTCATCGACTAAAGACAAATGCAAGTTAAGCAA
ACGCACAATAGCTCAATGCGCTTCAAGCTCGAAGGGCTTCTTTTTGCGCG
AGCAGAAGAACGTATTGTTTGATATAAAAAAGGGCCTAGAGCAATTGACA
TTACTTTGCAGTCAATTTCAGGATAGTCTAAGTATGCCAGATGAAAGGGC
AGACAGACCTGATGATGACAATACAAACACCTCGACCTCGACCTCGACCC
TTGAAGTTACCGTACAACGCGAATTGGAAAACAATCGATTGAAAATCGAA
GGCATGTTGGTCCAAGTGAGAAGTCTTTATCGTCAGTGGAGCAGTGCTGA
ACTCTATTATCTACGGAGCTTGCAGAGACTAGGCCTTACTCCTGGCAGCA
AGCCCGATTCTCCCACTCATGATGTTATGGCCCTTGCTGCGATTGCCTTG
TCTGCCGAATGCGATTCAGGGTTAAGAAACAAAACTGATGCCGCCGAAGT

TAGCCCAAAAAACTGCTATAGGCAAGAGAACAACTGCTCAAGTAAACCCG
AAGTACATTCATCGCCAGAGAAACTGCGTACACTGCAAGATATTGAAAAT
ATTATACTGCAGCAAGCAGCAGCCGCTGCTGCTAGTCAACAGAAACAATC
TGACCCAGGTTTCGTGAATGGCCACAGCGAAAGTGCAGCGAGCAGTGATA
GCGACGAAGAAAAAGAGTCATCATCACCCGCTTGTATTTGGCATCCAAAG
AACAATGTTCATCCAGCCGATGACGGTGACACCTGCAGCACTGCCGCCGA
AATTCTTTTGGAATATACTTCGTTATCGTCGGTGCAGGCGAAAATTAATG
CATTGATGGGCAGTGGCAACAGCTTGCCTCTGCCTGGCTCTGGTTCAAGT
ACTAACAAAATCGCCGAGAACTGTTCGGTCTTGAATCTATCACCAGGGAC
GCAGAGAGCACCGTTTAATCGAGAGTCGGCCACTTCGGATCCCTTATACT
TGGCCGATATTCCGTTCCCTCCAACGCCAGCCACACCACCAACAAGTAGC
AACAGCAGCTGTTCGACAGTTACGGGCCCGCTTGGCTGCACATCGTATTC
TTTGAGCACGAGTAAATATAACCACCGGCGAAAATCTAGATATGCGCGAC
GCATCGAAACTCCCACATCCTCATCATCGGCCAGCACCTGCCAGCCGGAG
TACTCGGCCAAGGAACTCAACGGGGACGCATATAAGGTTGTCGCTATTGC
CAGTATGCCAATGTGCCCATCCTCGGTATCATCGACAACCGCGTCAGCCT
CCGGCTCCGCCTCAGCCGCAGCGGCCCCCATTCAAGATTCACCCTCAGCC
GTTGCGGCGGTAGCCGCATTCCAGGAACGTGCCATTACGGACATGTTTAA
GGCCCAGTTTAGCGCTTTGACAGTGGCGGCTGGAATTGCAACTGGATGTG
GAAGCGAATCAGGACTTGGAGCTGGATCTGACACTCCCTACGACTTAAGC
ATTGGAACGAGGCTAAAAAAAATG

>CG32005
MHLNGNESAEENKEPKDFSIPVGNIMETISQSSSSSSTTTTVPNAQFFDH
QQTTNEAVTKVNSDVMEHIFNTPTSNAVVPTTTSTIMSEILLRQQHDMLQ
HWPKLEDPQRKWNREQLDTSCNRDALLARIYNNKVLSQHKSTQRQQQQQQ
QQQQQQRQFLATQLLYAQFLHQPHLSLEVDQERERRLHILEFSGLRKSSV
SNEPKATPLLSTSSLSIGLSHDNNNNRTQSDIEQSVRYDHDHIRRDGTDA
HNSNSNSSTKDKCKLSKRTIAQCASSSKGFFLREQKNVLFDIKKGLEQLT
LLCSQFQDSLSMPDERADRPDDDNTNTSTSTSTLEVTVQRELENNRLKIE
GMLVQVRSLYRQWSSAELYYLRSLQRLGLTPGSKPDSPTHDVMALAAIAL
SAECDSGLRNKTDAAEVSPKNCYRQENNCSSKPEVHSSPEKLRTLQDIEN
IILQQAAAAAASQQKQSDPGFVNGHSESAASSDSDEEKESSSPACIWHPK
NNVHPADDGDTCSTAAEILLEYTSLSSVQAKINALMGSGNSLPLPGSGSS
TNKIAENCSVLNLSPGTQRAPFNRESATSDPLYLADIPFPPTPATPPTSS
NSSCSTVTGPLGCTSYSLSTSKYNHRRKSRYARRIETPTSSSSASTCQPE
YSAKELNGDAYKVVAIASMPMCPSSVSSTTASASGSASAAAAPIQDSPSA
VAAVAAFQERAITDMFKAQFSALTVAAGIATGCGSESGLGAGSDTPYDLS
IGTRLKKM*

Caps

Isoform RA-RB

Exon	Position (phase)
Exon 18	(-)61074-61004(2)
Exon 19	(1)59793-59623(2)
Exon 20	(1)57286-56977(0)

Exon 21	(0)56858-56478(0)
Exon 22	(0)55222-55083(2)
Exon 23	(1)55021-54820(0)

>caps isoform RA-RB
GATCTAATGGAATCCTCAATAGCTCAATCGATTCATAAGGGCTTTGAGAA
GGAACGTTGGGAAAGCAAAGGGATAAACGCTGCTTTGAATCCAGCCGCAT
TAAACAATGCCGCCCAGGCGCTAAACACGGCAGCTCTCAACCCAGCGGGG
CTGCTGAGCGGCAAGAAAGATCAAGTCAATTTTTATGTACCGAAGCTACC
AAGGCGCACAGCTGCACCTGCGGCAGCCGCTGATGAAATGAGAAATGGTT
GTGCAACTTCCGAAGACCTTTTCTGGAAACTAGACGCACTGCAATCCTTT
ATTAGAGATCTGCATTGGCCAGATGCCGAGTTTCGTCAGCATTTGGAGCA
GAGACTTAAGATGATGGCTGTTGATATGATTGAGCAGTGCATACAGCGCA
CAGACTCTTCCTTTCAGTCTTGGTTAAAAAAAAATGTTGCCTTTATATCG
ACCGATTATATTATACCATCAGAGATGTGCGCTATGGTAAATGTGATATT
AGATGCTAAAAACCAAAGCTTTAAATTAACCACCATCGATGGTATCGATC
TGTACAAATTCCACGCAAAAATCGATGATCAAATAGACAAAGCGAATGTA
GCGATGACACAAGGATTAAGTGGCAAACTTATGTCAGTGCTAGAGTCGAC
TTTGTCGAAGTTGGCGCGTTACGATGAAGGTAGCCTCATTGGCTCGATTC
TTAGTTTTACAAACGTGTCCAGCTCAGGAAAAGATCTGGGACAAGGCTAT
GTGAATTTTTTTCGTAACAATATGGATCAAATACGCGGCAAGATAGGTGA
TGACCTATGGACACTGAACTTCTTTGAGCAGTGGTACTCTCAGCAAATAA
GTATGCTCTCCAGCTGGCTATCAGAACGTTTGGACCACGCCCTACACTAT
GCACAGGTCACGTCTATCTCTCATATTATAAAGAAAATTTATTCAGACTT
TGAGCTTCAAGGTGTTCTCGAGGACAAGCTAAATTCTAAAGCATATCAGA
CAGTTATACAACGAATGACTGCTGAAGAGGCAACCTGTGCGCTGAGCATG
CCAGAAGGCACCGATGGTGATGAAAATTGTGACGATATCCGCAACGGTGA
CGAAGAGGATGGTGGAGATGATTCTGGCGCTCACGCAGCGCGGAGCCTCC
CAAAGCCAAAGACGGGAGCTGCCCAAGCTGCGGCTGTCACCAACGTAGTC
GCTGGTCGGGTTGGTAATTTACTTGGTAAAGGCATTGGCGGACTTAGCTC
AAAGCTGGGAAGCGGCGGTTGGTTT

>caps isoform RA-RB
DLMESSIAQSIHKGFEKERWESKGINAALNPAALNNAAQALNTAALNPAG
LLSGKKDQVNFYVPKLPRRTAAPAAAADEMRNGCATSEDLFWKLDALQSF
IRDLHWPDAEFRQHLEQRLKMMAVDMIEQCIQRTDSSFQSWLKKNVAFIS
TDYIIPSEMCAMVNVILDAKNQSFKLTTIDGIDLYKFHAKIDDQIDKANV
AMTQGLSGKLMSVLESTLSKLARYDEGSLIGSILSFTNVSSSGKDLGQGY
VNFFRNNMDQIRGKIGDDLWTLNFFEQWYSQQISMLSSWLSERLDHALHY
AQVTSISHIIKKIYSDFELQGVLEDKLNSKAYQTVIQRMTAEEATCALSM
PEGTDGDENCDDIRNGDEEDGGDDSGAHAARSLPKPKTGAAQAAAVTNVV
AGRVGNLLGKGIGGLSSKLGSGGWF*