Annotation of 16B18
Fine Song
May 3, 2006

I. Overview

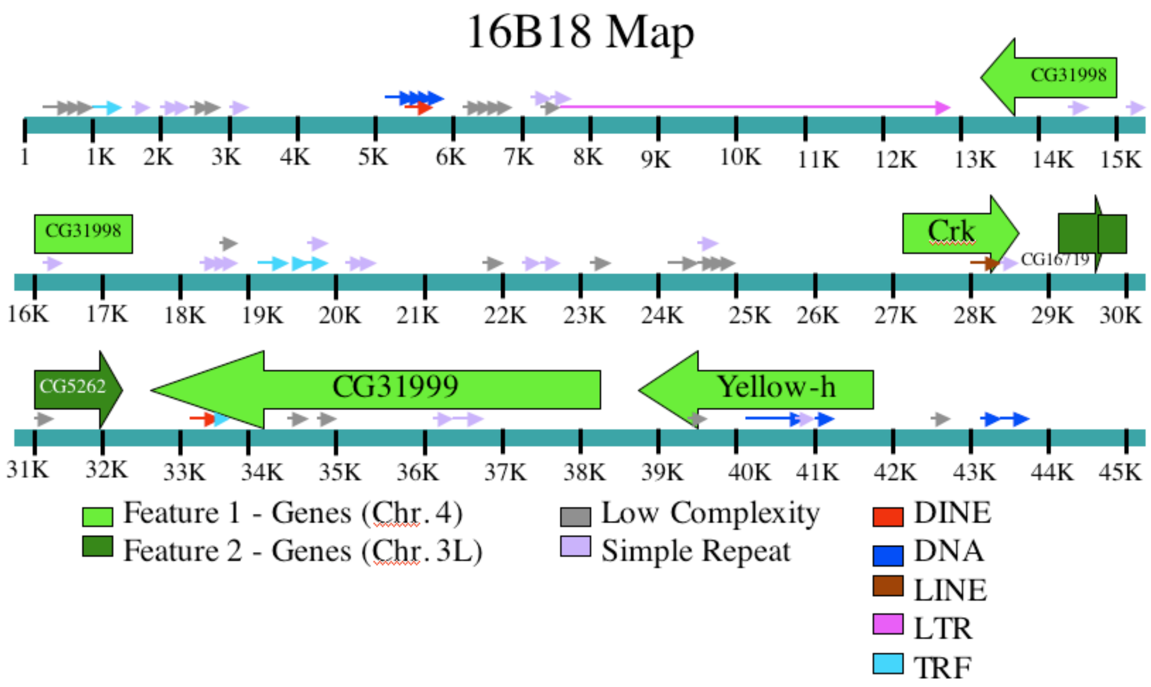


Figure 1: Annotated map of my fosmid

For my annotation project, I received 16B18, a 45,417 bp (37.6% GC content) fosmid in the fourth ("dot") chromosome of Drosophila virilis. *Initially Genscan gave me nine predicted features, but I narrowed that number down to six genes, two mispredictions, and one incomplete, truncated gene. CG31998 is a 4-exon anonymous gene with unknown function and protein family that spans from 13,116-18,991. CG1587, also known as Crk, is a 6-exon gene with three isoforms, thought to be involved with SH3/SH2 adaptor activity, protein binding, intercellular signaling cascade, development, myoblast fusion, positive regulation of JNK cascade, and imaginal disc fusion (thorax closure). It belongs to the Proto Oncogene C Crk P38 family and spans from 27,071-28,859. CG16719 is a 1-exon anonymous gene that is associated with protein binding and mesoderm development, belongs to the PA P protein family, and spans from 29,261-29,935. CG5262 is a 4-exon anonymous gene spanning from 29,711-32,261 whose function is associated with amino-acid-polyamine transport and aromatic amino acid permease and whose protein family is currently listed as ambiguous in Ensembl. CG31999 is a 12-exon anonymous gene whose function is EGF-like and related to calcium ion binding, protein binding, aspartic acid and asparagine hydroxylation sites. It belongs to the Latent Transforming Growth Factor Beta Binding Precursor LTBP protein family and spans from 32,675-38,128. CG1629, known as yellow-h, is a 3-exon gene with association with the major royal jelly protein, a part of the yellow precursor protein family, and spans from 38,908-41,784. No novel repeats were found. There was lack of synteny between my fosmid and the genes associated with chromosome 3L of* D.

melanogaster *(CG16719, CG5262), but other genes in my fosmid had synteny with chromosome 4 of* D. melanogaster *(CG31998, CG1587, CG31999, CG1629).*

II. Genes

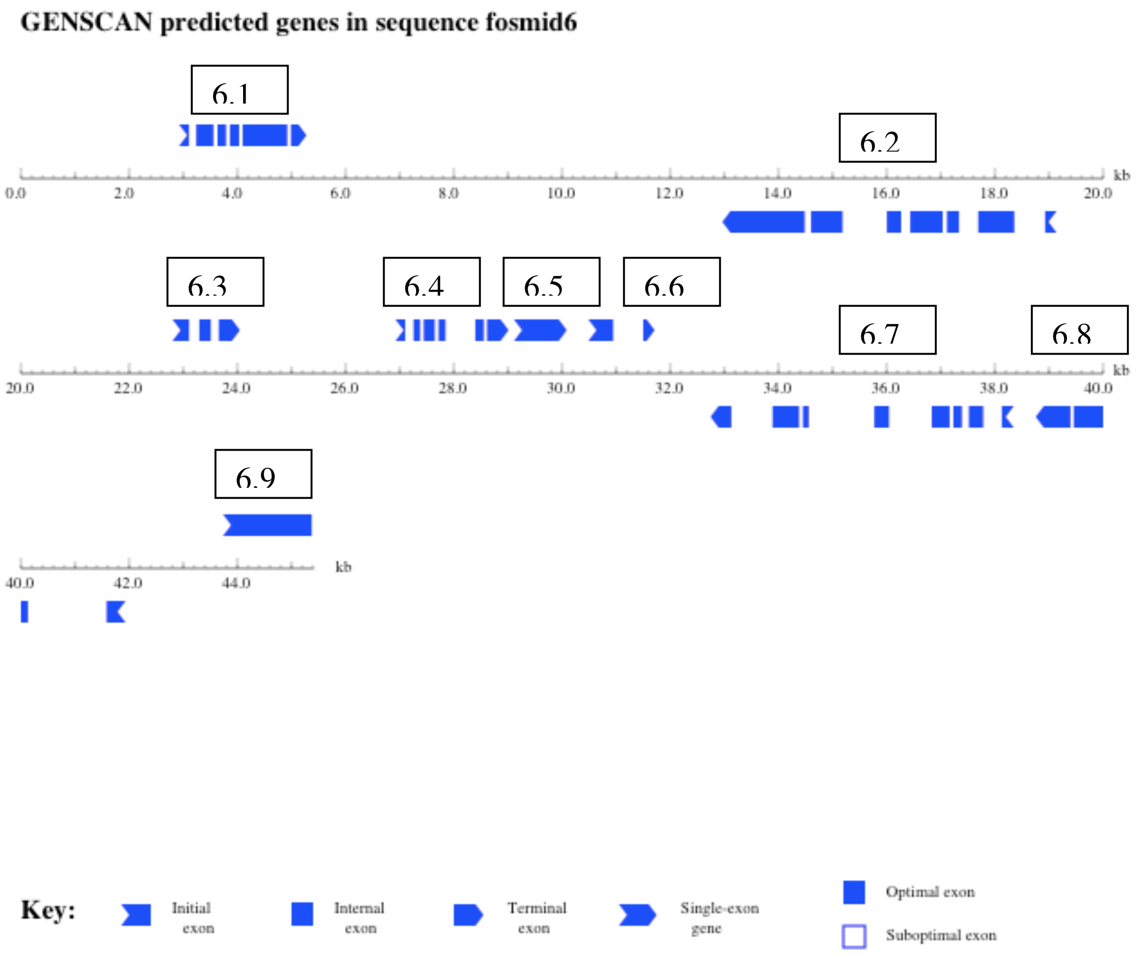


Figure 2: Genscan predictions for my fosmid

Initially, Genscan predicted nine features for my fosmid (Fig. 2). To determine the validity of these predictions, I used the UCSC genome browser as a tool to visualize and locate each feature (Fig. 3). In approaching each feature, I used essentially the same process to analyze and determine the validity of each prediction. I will use my examination of the fourth Genscan-predicted feature to show the steps involved in that process.

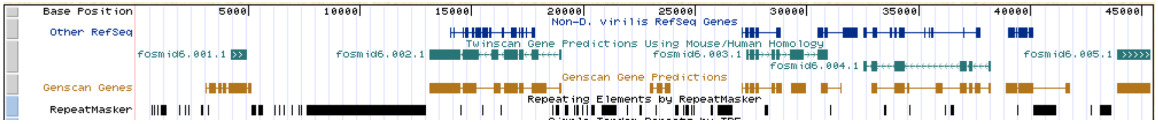
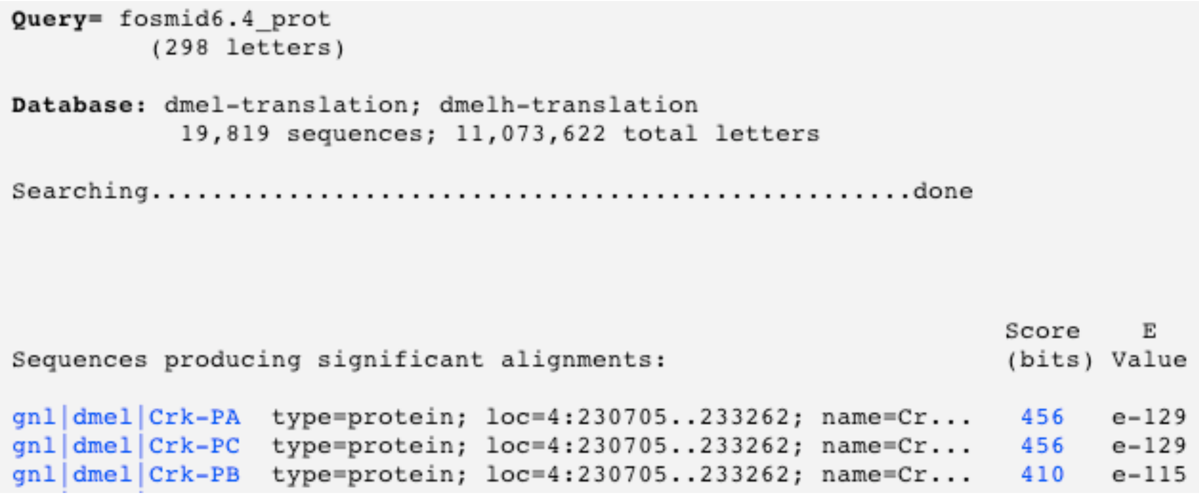


Figure 3: USCS output on goose server for my fosmid

6.4 I first took the translated protein sequence from Genscan for the fourth feature and performed a Blastp in FlyBase against an annotated amino acid *D. melanogaster* database. The best match in the Blast search was the Crk gene (NM_143651; CG1587), which has "A", "B", and "C" isoforms (Fig. 4).



```
Query= fosmid6.4_prot
         (298 letters)

Database: dmel-translation; dmelh-translation
           19,819 sequences; 11,073,622 total letters

Searching..................................................done



                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

gnl|dmel|Crk-PA  type=protein; loc=4:230705..233262; name=Cr...   456   e-129
gnl|dmel|Crk-PC  type=protein; loc=4:230705..233262; name=Cr...   456   e-129
gnl|dmel|Crk-PB  type=protein; loc=4:230705..233262; name=Cr...   410   e-115
```

Figure 4: FlyBase Blastp matches with predicted feature 4 protein

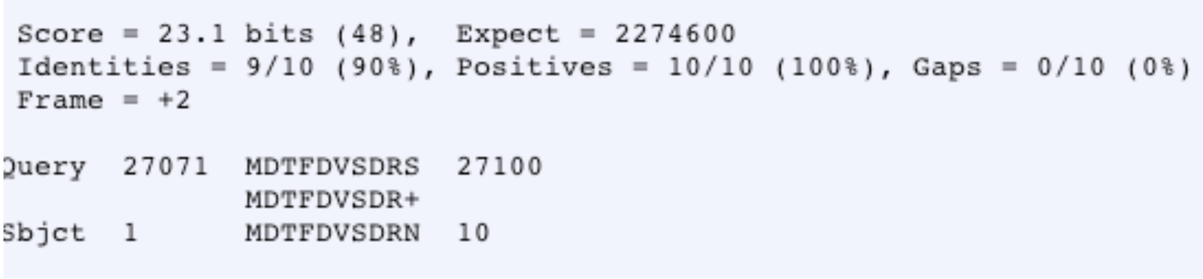
To obtain more information about this Crk protein, I utilized the *Drosophila melanogaster* division of the Ensembl website. In there, not only did I find the transcript structure (among other useful information) of this gene and its isoforms, but there was also the peptide sequence of each individual exon, which I used to perform Blast2 (blastx) against my fosmid to get the coordinates for Crk or Crk-related sequences in *D. virilis*.

Isoforms A and C seem to cover the same exon regions and are more extensive in coverage of the third exon than isoform B. Isoform C seems to have a longer and separated or split UTR compared to A. Based on the evidence we have so far, either isoform A or C could be used for further investigation of this gene, since the "extra" exon in C was really a UTR and both A and C encode for the same number of amino acids. I decided to choose A because its exon transcript spans longer (Table 1).

Isoform	Number of CDS	Transcript Length	Protein Length
A	6	1,129 bps	271 residues
B	6	1,028 bps	253 residues
C	6	1,093 bps	271 residues

Table 1: The different isoforms of Crk

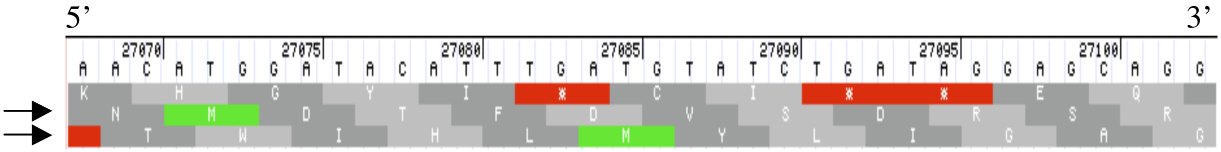
I then performed Blast2 (blastx) on individual exons from isoform A protein sequence and my entire fosmid DNA. For exon 1, I had to raise the E-value significantly because the protein sequence for it was short with only 10 amino acids (Fig. 5). Raising the E-value was justified in giving me a starting point to locate the exon using UCSC genome browser in my fosmid.



```
Score = 23.1 bits (48),  Expect = 2274600
Identities = 9/10 (90%), Positives = 10/10 (100%), Gaps = 0/10 (0%)
Frame = +2

Query  27071   MDTFDVSDRS  27100
               MDTFDVSDR+
Sbjct  1       MDTFDVSDRN  10
```

Figure 5: Blast2 (blastx) output of exon 1 of Crk gene

As I plugged in the approximate values for this positively oriented exon, the UCSC genome browser indicated both the second and third frames of translation to be free of stop codons, which is usually what is desired in identifying a fully translatable exon. Had I not been provided with approximate coordinates through Blast2, I would have favored the second frame over the third frame due to the Met start codon that is more upstream and therefore able to include more amino acids.



5'
3'
27070
27075
27080
27085
27090
27095
27100
A A C A T G G A T A C A T T T G A T G T A T C T G A T A G G A G C A G G

Figure 6: Region of my fosmid matching to Crk exon 1

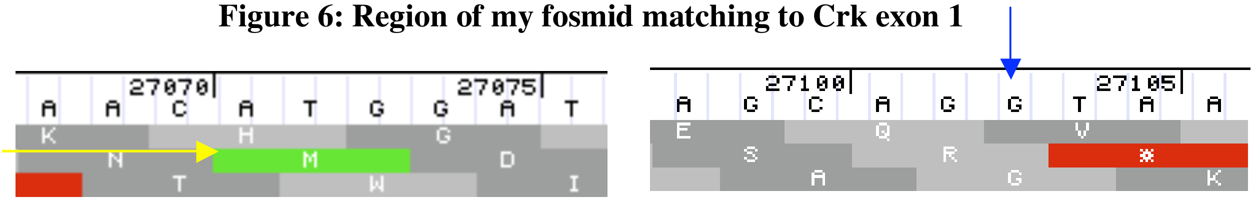
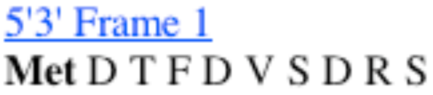


Figure 7: Close-up views of the 5' and 3' ends of proposed Crk exon 1

Zooming in to each end, I found the start codon on the 5' end of the fosmid and a GT donor site on the 3' end, as expected (Fig. 7). For internal exons and the terminal exon, I would have looked for the AG acceptor site on the 5' end. Also, in examining the splice sites of multiple exons, it is important to make sure the frames match, i.e. any base overhang must be matched by the next exon to make sure all the exons are aligned properly. Splice site tracker on the UCSC genome browser for our fosmids was extremely helpful in that regard.

To confirm this exon for my gene model, I translated this DNA region into a protein sequence using ExPASy (Fig 8). Upon seeing that there were no stop codons in the translation, I concluded that exon 1 spanned bases 27,071-27,102 in my *D. virilis* fosmid in the positive orientation.

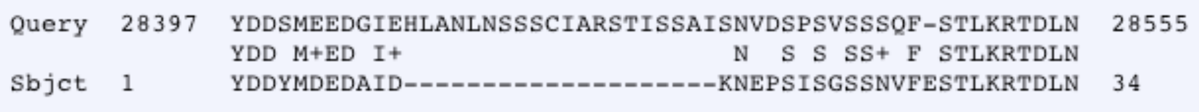


5'3' Frame 1

Met D T F D V S D R S

Figure 8: Translated confirmation of Crk exon 1 in my fosmid

Using the same process and logic, I found the boundaries for the five other exons in this Crk gene model, which can be found in the GTF files in the Appendix as well as the summary table below. One particularly interesting exon was the fifth exon, which had a large insertion in the Query (my *D. virilis* fosmid) sequence, compared to *D. melanogaster* in Blast 2 (blastx) (Fig. 9). One possibility for the gap is gene insertion in *D. virilis* through evolution.



```
Query  28397  YDDSMEEDGIEHLANLNSSSCIARSTISSAISNVDSPSVSSSQF-STLKRTDLN  28555
              YDD M+ED I+                     N  S S SS+ F STLKRTDLN
Sbjct  1      YDDYMDEDAID--------------------KNEPSISGSSNVFESTLKRTDLN  34
```

Figure 9: Blast2 results from exon 5 of Crk gene

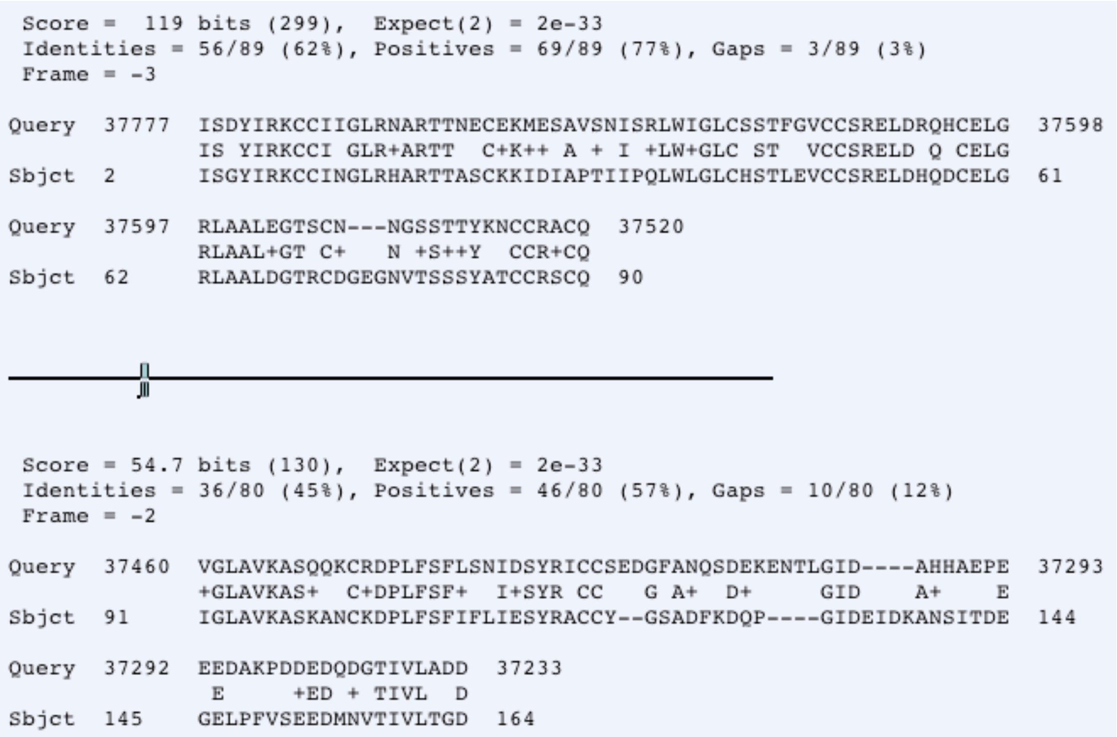
The Crk gene can be found on the fourth chromosome of *D. melanogaster*. According to Gene Ontology (GO) in Ensembl, Crk is thought to be involved with SH3/SH2 adaptor activity, protein binding, intercellular signaling cascade, development, myoblast fusion, positive regulation of JNK cascade, and imaginal disc fusion (thorax closure). Crk belongs to the Proto Oncogene C Crk P38 family.

Using the same methods and logic as for Crk, I validated other features as genes in my fosmid including the second, seventh, and eight features predicted by Genscan. Out of those, the second, seventh, and eight features can be found on chromosome 4 of *D. melanogaster*, while the fifth and sixth features, which overlap each other by 225 bp, are found in chromosome 3L of *D. melanogaster*.

6.8 The eighth feature predicted by Genscan is predicted to be the 3-exon yellow-h gene (NM_143655, CG1629) after performing FlyBase Blastp (blastx) on predicted protein sequence from Genscan. The only major orthologous function for this gene is as part of the major royal jelly protein. It is part of the yellow precursor protein family.
6.7 The seventh feature is predicted to be a 12-exon anonymous CG31999 (NM_166745) whose function is EGF-like and related to calcium ion binding, protein binding, aspartic acid and asparagine hydroxylation sites. It belongs to the Latent Transforming Growth Factor Beta Binding Precursor LTBP protein family.
6.6 The sixth feature is predicted to be a 4-exon anonymous CG5262 gene (NM_140966) whose function is associated with amino acid-polyamine transport and aromatic amino acid permease. Its protein family is currently listed as ambiguous.
6.5 The fifth feature is predicted to be a 1-exon anonymous CG16716 gene (NM_140087), which is associated with protein binding and mesoderm development and belongs to the PA P protein family.
6.2 The second feature is a 4-exon CG31998 gene (NM_166742) both an unknown function and protein family.

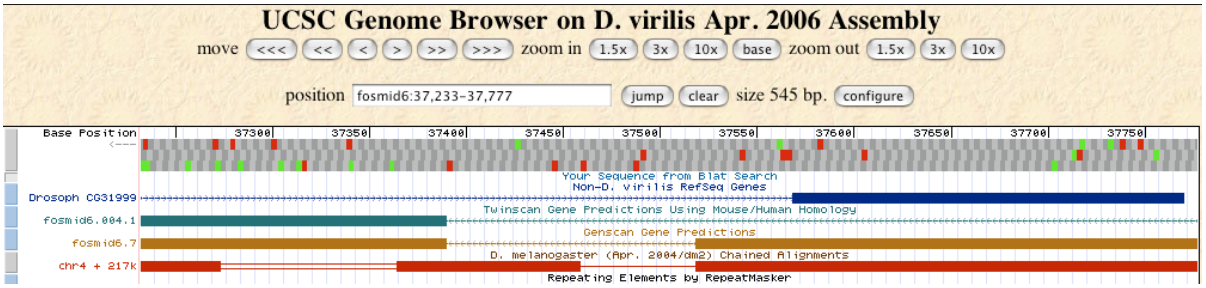
Interesting regions that I came across among the features that have been mentioned include exon 2 in feature 6.7. Looking at the Blast2 (blastx) output performed between my fosmid and its amino acid sequence from Ensembl, exon 2 seems to have a gap that extends from 37,460-37,520 (Fig. 10). I initially hypothesized that perhaps the gap is due to a transposable element or an insertion of an intron, but looking at the UCSC genome browser, stop codons riddle all three frames in the negative direction, which is the orientation of this exon (Fig. 11). Hence, the stop codons are not restricted to the gap.

A possible reason for this to occur is that a gene duplication event allows this exon to be "disposable" and have greater flexibility in gaining mutations than the other copy. Since 11 of 12 exons for this feature is well conserved, it is safe to say that this is not a processed mRNA that was transposed back into the genome and that it is a real gene. Since this exon did not seem to be a coding exon, I did not include it as "exon 2" in my annotation. Instead, I designated the third exon that I examined as "exon 2" in my annotation of feature 6.7.



```
 Score =  119 bits (299),  Expect(2) = 2e-33
 Identities = 56/89 (62%), Positives = 69/89 (77%), Gaps = 3/89 (3%)
 Frame = -3

Query  37777   ISDYIRKCCIIGLRNARTTNECEKMESAVSNISRLWIGLCSSTFGVCCSRELDRQHCELG  37598
               IS YIRKCCI GLR+ARTT  C+K++ A + I +LW+GLC ST  VCCSRELD Q CELG
Sbjct  2       ISGYIRKCCINGLRHARTTASCKKIDIAPTIIPQLWLGLCHSTLEVCCSRELDHQDCELG  61

Query  37597   RLAALEGTSCN---NGSSTTYKNCCRACQ  37520
               RLAAL+GT C+   N +S++Y  CCR+CQ
Sbjct  62      RLAALDGTRCDGEGNVTSSSYATCCRSCQ  90


 Score = 54.7 bits (130),  Expect(2) = 2e-33
 Identities = 36/80 (45%), Positives = 46/80 (57%), Gaps = 10/80 (12%)
 Frame = -2

Query  37460   VGLAVKASQQKCRDPLFSFLSNIDSYRICCSEDGFANQSDEKENTLGID----AHHAEPE  37293
               +GLAVKAS+  C+DPLFSF+  I+SYR CC   G A+  D+     GID    A+    E
Sbjct  91      IGLAVKASKANCKDPLFSFIFLIESYRACCY--GSADFKDQP----GIDEIDKANSITDE  144

Query  37292   EEDAKPDDEDQDGTIVLADD  37233
                E     +ED + TIVL  D
Sbjct  145     GELPFVSEEDMNVTIVLTGD  164
```

Figure 10: Blast2 (blastx) output for exon 2 of feature 6.7

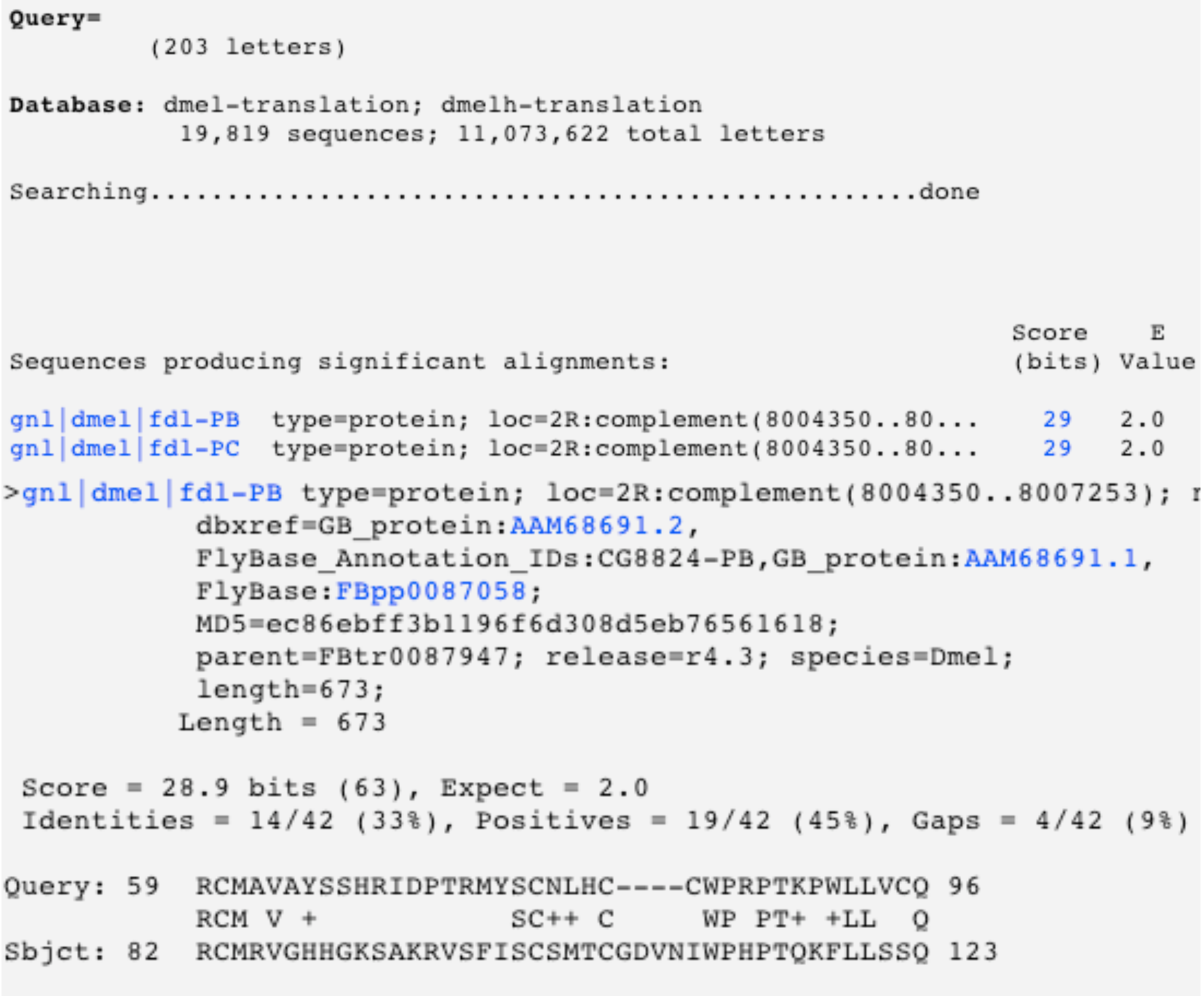


UCSC Genome Browser on D. virilis Apr. 2006 Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
jump clear configure
Base Position
37300
37350
37400
37450
37500
37550
37600
37650
37700
37750
<---
Your Sequence from Blat Search
Non-D. virilis RefSeq Genes
Drosoph CG31999
Twinscan Gene Predictions Using Mouse/Human Homology
fosmid6.004.1
Genscan Gene Predictions
fosmid6.7
D. melanogaster (Apr. 2004/dm2) Chained Alignments
chr4 + 217k
Repeating Elements by RepeatMasker

position fosmid6:37,233-37,777 size 545 bp.

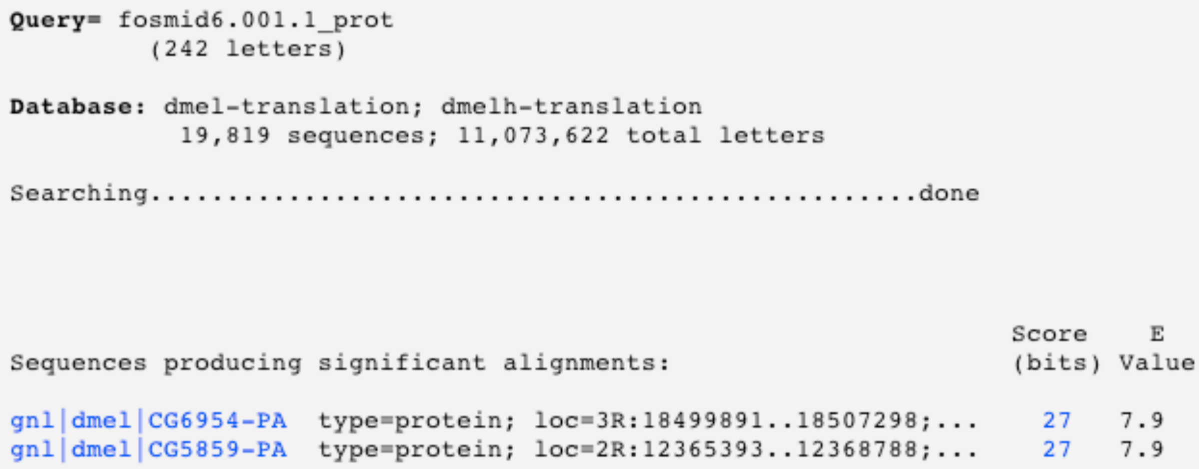
Figure 11: UCSC genome browser for exon 2 of feature 6.7

It should also be mentioned that features 6.1 and 6.3 predicted by Genscan were mispredictions, based on lack of significant matches (only two matches each, with high E-values) in FlyBlast Blastp (Figs 12-14).

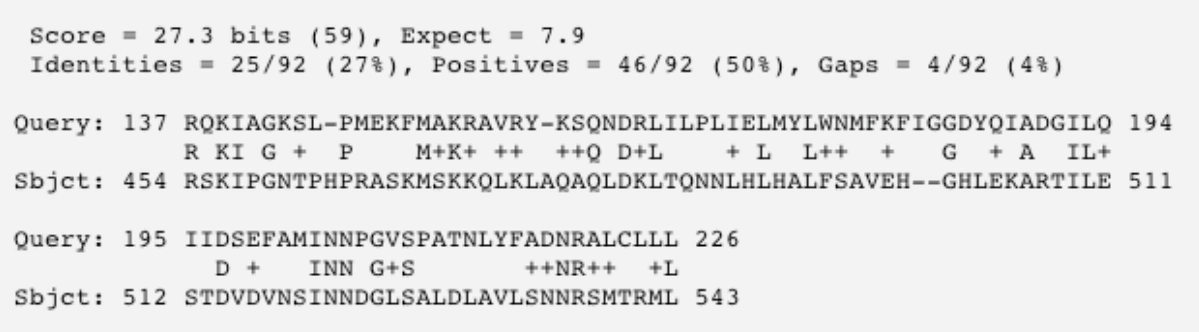


```
Query=
         (203 letters)

Database: dmel-translation; dmelh-translation
           19,819 sequences; 11,073,622 total letters

Searching..................................................done



                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

gnl|dmel|fdl-PB  type=protein; loc=2R:complement(8004350..80...    29    2.0
gnl|dmel|fdl-PC  type=protein; loc=2R:complement(8004350..80...    29    2.0

>gnl|dmel|fdl-PB type=protein; loc=2R:complement(8004350..8007253); r
          dbxref=GB_protein:AAM68691.2,
          FlyBase_Annotation_IDs:CG8824-PB,GB_protein:AAM68691.1,
          FlyBase:FBpp0087058;
          MD5=ec86ebff3b1196f6d308d5eb76561618;
          parent=FBtr0087947; release=r4.3; species=Dmel;
          length=673;
         Length = 673

 Score = 28.9 bits (63), Expect = 2.0
 Identities = 14/42 (33%), Positives = 19/42 (45%), Gaps = 4/42 (9%)

Query: 59   RCMAVAYSSHRIDPTRMYSCNLHC----CWPRPTKPWLLVCQ 96
            RCM V +              SC++ C      WP PT+ +LL  Q
Sbjct: 82   RCMRVGHHGKSAKRVSFISCSMTCGDVNIWPHPTQKFLLSSQ 123
```

Figure 12: FlyBase Blastp results for feature 6.3



```
Query= fosmid6.001.1_prot
         (242 letters)

Database: dmel-translation; dmelh-translation
           19,819 sequences; 11,073,622 total letters

Searching..................................................done



                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

gnl|dmel|CG6954-PA  type=protein; loc=3R:18499891..18507298;...    27    7.9
gnl|dmel|CG5859-PA  type=protein; loc=2R:12365393..12368788;...    27    7.9
```



```
 Score = 27.3 bits (59), Expect = 7.9
 Identities = 25/92 (27%), Positives = 46/92 (50%), Gaps = 4/92 (4%)

Query: 137 RQKIAGKSL-PMEKFMAKRAVRY-KSQNDRLILPLIELMYLWNMFKFIGGDYQIADGILQ 194
           R KI G +  P    M+K+ ++  ++Q D+L    + L  L++   +  G  + A  IL+
Sbjct: 454 RSKIPGNTPHPRASKMSKKQLKLAQAQLDKLTQNNLHLHALFSAVEH--GHLEKARTILE 511

Query: 195 IIDSEFAMINNPGVSPATNLYFADNRALCLLL 226
             D +   INN G+S       ++NR++  +L
Sbjct: 512 STDVDVNSINNDGLSALDLAVLSNNRSMTRML 543
```

Figure 13: FlyBase Blastp results for feature 6.1

Even when I widened opportunities to get matches, including using the Genscan-predicted mRNA sequence over the predicted protein sequence and using tBlastx (nr database). A Blast2 search between feature 6.3, which was predicted to be a one-exon gene, and my fosmid yielded no results. Only near E-value of 1e6 (1 million) there were matches, which were terrible in their short length and lack of homology (Fig. 14). Performing NCBI blastn (nr database) for feature 6.1 produced a terrible match (Fig. 15).

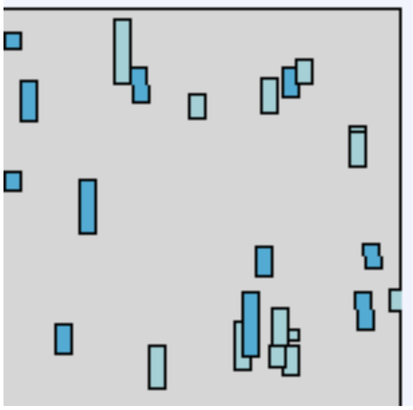


Figure 14: Blast2 between my fosmid and a possible feature 6.3 region

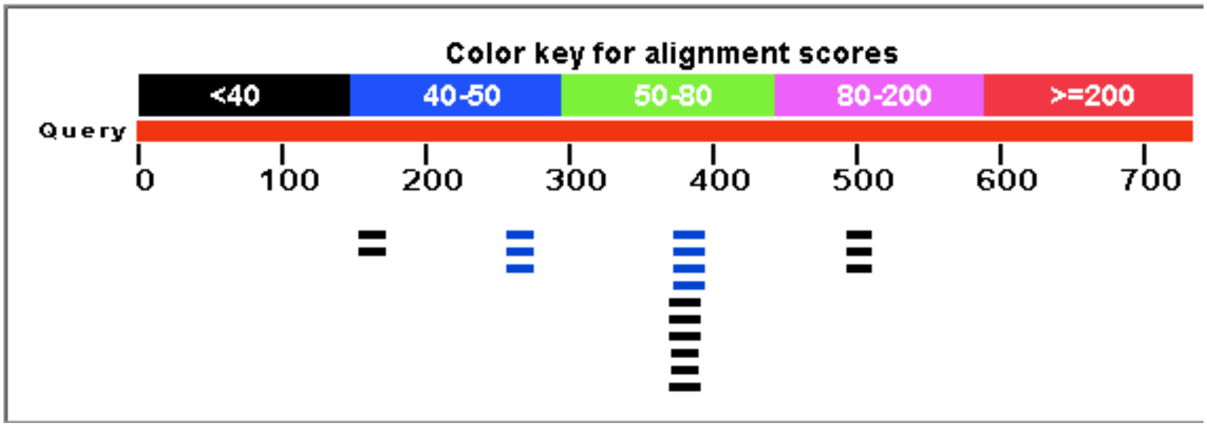


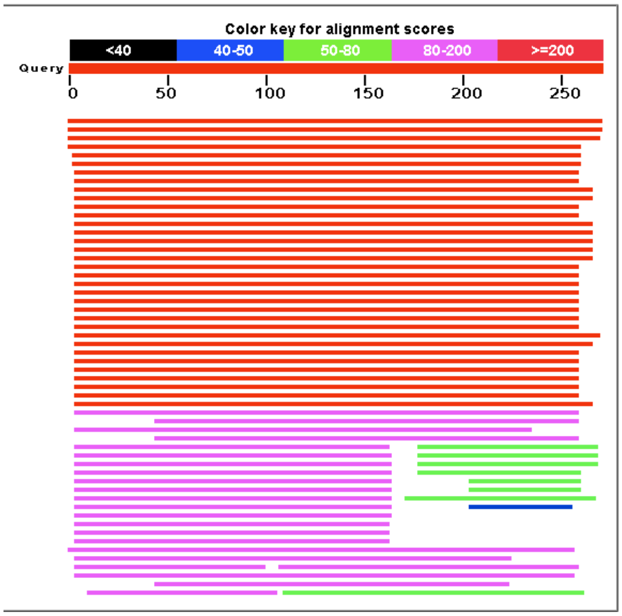
Figure 15: NCBI Blastn (nr) results of my fosmid feature 6.1

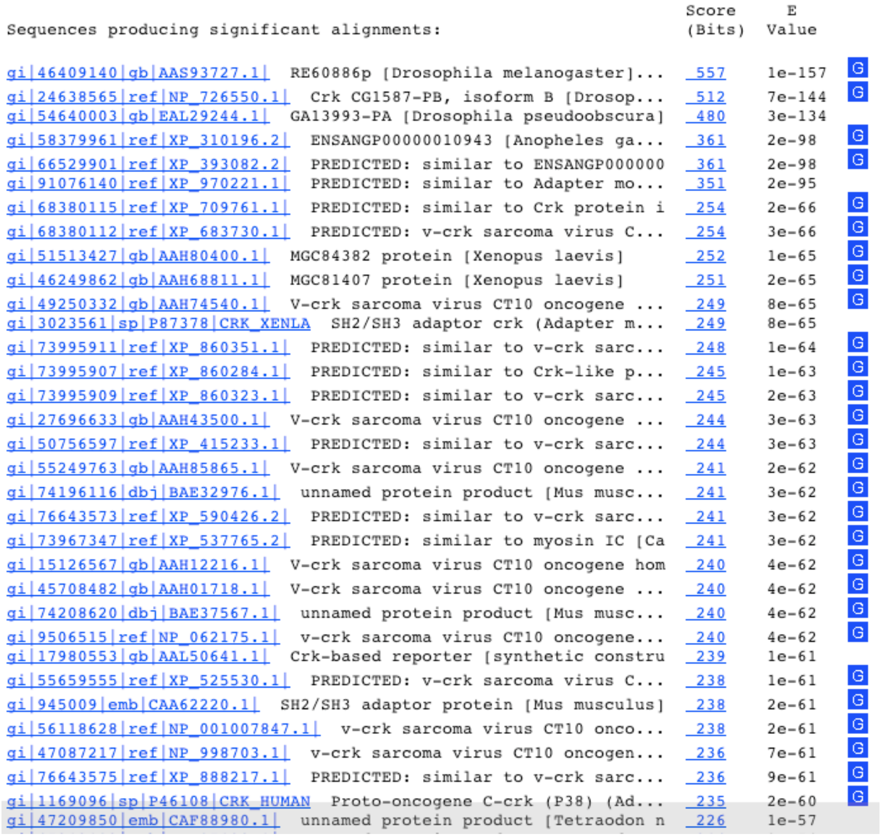
Additionally, the ninth (last) feature predicted by Genscan was most likely a rho-5 gene (NM_205957, CG33304) based on strong FlyBase Blastp results. However, this is a six-exon gene and only part of the first exon is on the 3' end of my fosmid. Therefore, I did not annotate this gene, which can be found in chromosome 2L and is a rhomboid-like protein.

Overall, producing all the GTF formatted files were successful for the six annotated genes. The only problematic region was on or around exon 7 of feature 6.7, because the translated protein sequence for this exon had amino acid sequences from the second frame instead of the third frame. This may be due to a problematic, predicted AG acceptor site, which lacked any degree of confidence in the splice site tracker (in contrast to the high confidence sites in the associated GT donor site). The AG site has a two base overhang instead of the one base overhang needed to match with its corresponding, high confidence GT donor site. At the same time, going downstream (relative to the fosmid) leads to stop codons, and going upstream to a site with a one base overhang leads to the loss of 22 nucleotides for the exon and leads to complications with a failure in one of the splice site tests in the Annotation Check program. This is something that should be addressed by another annotator.

III. **Clustal Analysis**

For the first part of my Clustal analysis, I used the Crk gene to compare against other species. Putting in the protein sequence of this gene into NCBI blastp (nr database), I noticed that this gene was well conserved and has multiple low E-value matches (Fig. 16).





```
                                                                 Score     E
Sequences producing significant alignments:                      (Bits)  Value

gi|46409140|gb|AAS93727.1|  RE60886p [Drosophila melanogaster]...  557    1e-157  G
gi|24638565|ref|NP_726550.1|  Crk CG1587-PB, isoform B [Drosop...  512    7e-144  G
gi|54640003|gb|EAL29244.1|  GA13993-PA [Drosophila pseudoobscura]  480    3e-134
gi|58379961|ref|XP_310196.2|  ENSANGP00000010943 [Anopheles ga...  361    2e-98   G
gi|66529901|ref|XP_393082.2|  PREDICTED: similar to ENSANGP000000  361    2e-98   G
gi|91076140|ref|XP_970221.1|  PREDICTED: similar to Adapter mo...  351    2e-95
gi|68380115|ref|XP_709761.1|  PREDICTED: similar to Crk protein i  254    2e-66   G
gi|68380112|ref|XP_683730.1|  PREDICTED: v-crk sarcoma virus C...  254    3e-66   G
gi|51513427|gb|AAH80400.1|  MGC84382 protein [Xenopus laevis]      252    1e-65   G
gi|46249862|gb|AAH68811.1|  MGC81407 protein [Xenopus laevis]      251    2e-65   G
gi|49250332|gb|AAH74540.1|  V-crk sarcoma virus CT10 oncogene ...  249    8e-65   G
gi|3023561|sp|P87378|CRK_XENLA  SH2/SH3 adaptor crk (Adapter m...  249    8e-65
gi|73995911|ref|XP_860351.1|  PREDICTED: similar to v-crk sarc...  248    1e-64   G
gi|73995907|ref|XP_860284.1|  PREDICTED: similar to Crk-like p...  245    1e-63   G
gi|73995909|ref|XP_860323.1|  PREDICTED: similar to v-crk sarc...  245    2e-63   G
gi|27696633|gb|AAH43500.1|  V-crk sarcoma virus CT10 oncogene ...  244    3e-63   G
gi|50756597|ref|XP_415233.1|  PREDICTED: similar to v-crk sarc...  244    3e-63   G
gi|55249763|gb|AAH85865.1|  V-crk sarcoma virus CT10 oncogene ...  241    2e-62   G
gi|74196116|dbj|BAE32976.1|  unnamed protein product [Mus musc...  241    3e-62   G
gi|76643573|ref|XP_590426.2|  PREDICTED: similar to v-crk sarc...  241    3e-62   G
gi|73967347|ref|XP_537765.2|  PREDICTED: similar to myosin IC [Ca  241    3e-62   G
gi|15126567|gb|AAH12216.1|  V-crk sarcoma virus CT10 oncogene hom  240    4e-62   G
gi|45708482|gb|AAH01718.1|  V-crk sarcoma virus CT10 oncogene ...  240    4e-62   G
gi|74208620|dbj|BAE37567.1|  unnamed protein product [Mus musc...  240    4e-62   G
gi|9506515|ref|NP_062175.1|  v-crk sarcoma virus CT10 oncogene...  240    4e-62   G
gi|17980553|gb|AAL50641.1|  Crk-based reporter [synthetic constru  239    1e-61
gi|55659555|ref|XP_525530.1|  PREDICTED: v-crk sarcoma virus C...  238    1e-61   G
gi|945009|emb|CAA62220.1|  SH2/SH3 adaptor protein [Mus musculus]  238    2e-61   G
gi|56118628|ref|NP_001007847.1|  v-crk sarcoma virus CT10 onco...  238    2e-61   G
gi|47087217|ref|NP_998703.1|  v-crk sarcoma virus CT10 oncogen...  236    7e-61   G
gi|76643575|ref|XP_888217.1|  PREDICTED: similar to v-crk sarc...  236    9e-61   G
gi|1169096|sp|P46108|CRK_HUMAN  Proto-oncogene C-crk (P38) (Ad...  235    2e-60   G
gi|47209850|emb|CAF88980.1|  unnamed protein product [Tetraodon n  226    1e-57
```

Figure 16: Conservation of Crk gene through NCBI blastp (nr) output

Since Crk in *D. melanogaster* is well conserved, it was safe to utilize Ensembl's Orthologue Prediction list instead sorting through the NCBI blastp results one by one. I chose to compare my *D. virilis*' Crk region with *Mus* musculus (mouse), *Caenorhabditis elegans* (worm), *Xenopus tropicalis* (frog), and *Monodelphis domestica* (opossum). A Cladogram is provided by the ClustalW output (Fig. 17).

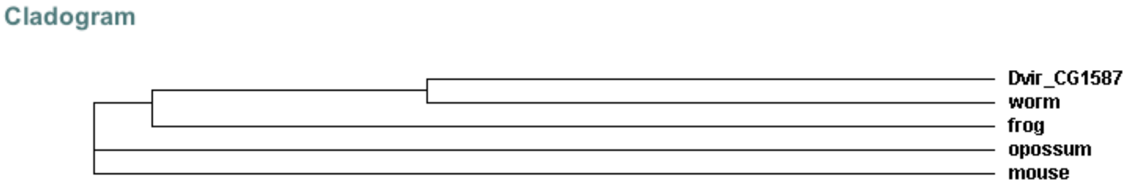
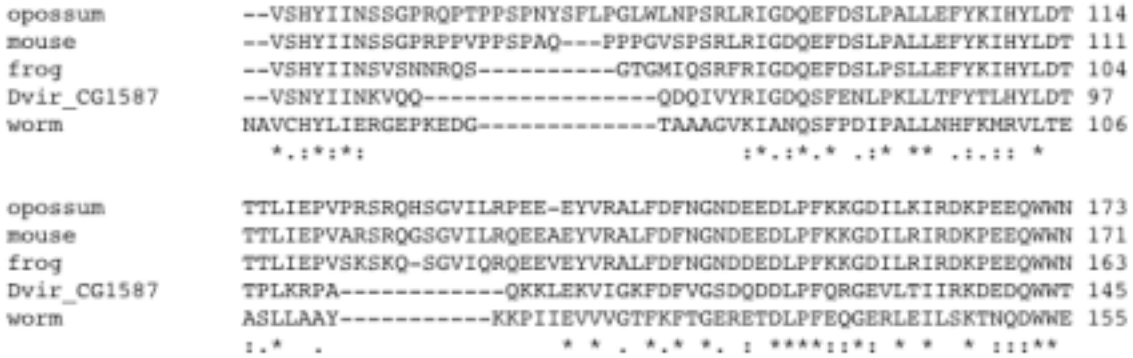


Figure 17: ClustalW Cladogram

One would expect that since the worm is closer in evolutionary distance it would have more similar patterns of conservation to my fosmid than the other species, which is exactly what one finds with the Clustal results   In the first row of Fig. 18, there is a gap that seems to be correlated to evolutionary distance from my fosmid, i.e. the closer the species, the larger (and closer in size) the gap. Also, in the second row, there is a gap that is present only in worm and my fosmid. Overall, there seems to be significant conservation throughout the entire ClustalW multiple sequence alignment, which leads one to think that this gene is an important one throughout species.



```
opossum         --VSHYIINSSGPRQPTPPSPNYSFLPGLWLNPSRLRIGDQEFDSLPALLEFYKIHYLDT 114
mouse           --VSHYIINSSGPRPPVPPSPAQ---PPPGVSPSRLRIGDQEFDSLPALLEFYKIHYLDT 111
frog            --VSHYIINSVSNNRQS----------GTGMIQSRFRIGDQEFDSLPSLLEFYKIHYLDT 104
Dvir_CG1587     --VSNYIINKVQQ-----------------QDQIVYRIGDQSFENLPKLLTFYTLHYLDT 97
worm            NAVCHYLIERGEPKEDG-------------TAAAGVKIANQSFPDIPALLNHFKMRVLTE 106
                  *.:*:*:                           :*.:*.* .:* ** .:.:: *

opossum         TTLIEPVPRSRQHSGVILRPEE-EYVRALFDFNGNDEEDLPFKKGDILKIRDKPEEQWWN 173
mouse           TTLIEPVARSRQGSGVILRQEEAEYVRALFDFNGNDEEDLPFKKGDILRIRDKPEEQWWN 171
frog            TTLIEPVSKSKQ-SGVIQRQEEVEYVRALFDFNGNDDEDLPFKKGDILRIRDKPEEQWWN 163
Dvir_CG1587     TPLKRPA------------QKKLEKVIGKFDFVGSDQDDLPFQRGEVLTIIRKDEDQWWT 145
worm            ASLLAAY-----------KKPIIEVVVGTFKFTGERETDLPFEQGERLEILSKTNQDWWE 155
                :.*   .                * * . *.* *. : ****::*: * *  * :::**
```

Figure 18: Segment of ClustalW output for Crk

To analyze a promoter region, I took the anonymous gene CG1629, or feature 6.8, and extracted the 1,000 bases in front of its first Met start codon (41,783-42,784). I compared this region with other *Drosophila* species, due to the rapidly changing nature of promoter regions and the need to keep evolutionary distance short so that analysis of conservation is contained within a reasonable temporal boundary. I chose *D. mojavensis, D. grimshawi, D. persimilis,* and *D. ananassae* to compare to my *D. virilis* fosmid (Fig. 19).

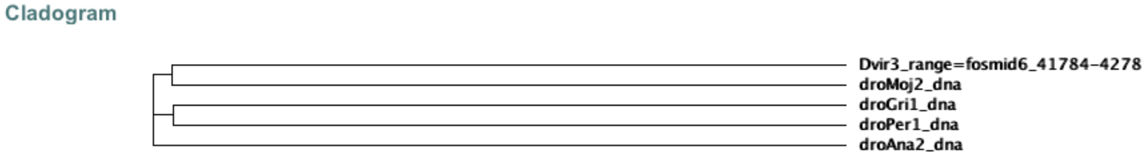
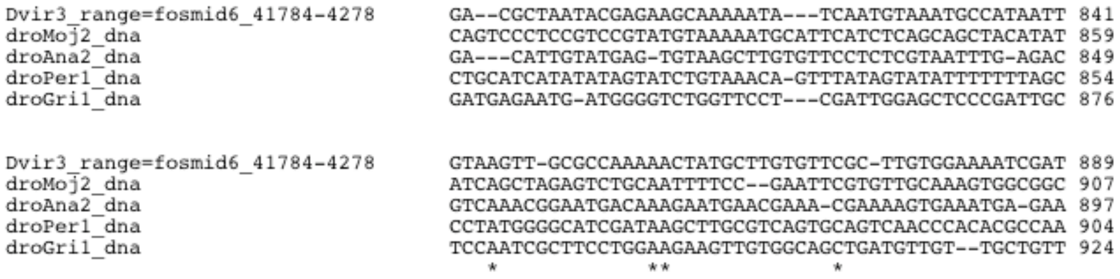


Figure 19: ClustalW Cladogram

Examining the multiple sequence alignment results produced from ClustalW, there does not seem to be a lot of conservation for this putative gene (Fig. 20). This may be due to the gene so rapidly mutating that even a multiple sequence alignment within a species is not enough to detect significant alignments. Or, perhaps this gene is not very important and thus there is little need for an organism to conserve an unneeded gene.



```
Dvir3_range=fosmid6_41784-4278      GA--CGCTAATACGAGAAGCAAAAATA---TCAATGTAAATGCCATAATT 841
droMoj2_dna                         CAGTCCCTCCGTCCGTATGTAAAAATGCATTCATCTCAGCAGCTACATAT 859
droAna2_dna                         GA---CATTGTATGAG-TGTAAGCTTGTGTTCCTCTCGTAATTTG-AGAC 849
droPer1_dna                         CTGCATCATATATAGTATCTGTAAACA-GTTTATAGTATATTTTTTTAGC 854
droGri1_dna                         GATGAGAATG-ATGGGGTCTGGTTCCT---CGATTGGAGCTCCCGATTGC 876


Dvir3_range=fosmid6_41784-4278      GTAAGTT-GCGCCAAAAACTATGCTTGTGTTCGC-TTGTGGAAAATCGAT 889
droMoj2_dna                         ATCAGCTAGAGTCTGCAATTTTCC--GAATTCGTGTTGCAAAGTGGCGGC 907
droAna2_dna                         GTCAAACGGAATGACAAAGAATGAACGAAA-CGAAAAGTGAAATGA-GAA 897
droPer1_dna                         CCTATGGGGCATCGATAAGCTTGCGTCAGTGCAGTCAACCCACACGCCAA 904
droGri1_dna                         TCCAATCGCTTCCTGGAAGAAGTTGTGGCAGCTGATGTTGT--TGCTGTT 924
                                         *          **             *
```

Figure 20: Segment of ClustalW output for CG1629

IV. Repeats

Repetitive elements comprise 25.4% of my fosmid. Repeat Masker ran with the "-no low" option off, allowing low complexity and simple repeats to be detected. A breakdown of the main repeat families can be seen in Table 2.

Type of Repeat Family	Length (b.p.)	% Genome (45417 b.p.)
DINE	186	0.4
DNA	1820	4.0
LINE	120	0.26
Low Complexity	822	1.8
LTR	5331	11.7
Simple Repeats	1313	2.9
TRF	503	1.1
Unknown	1454	3.2
TOTAL	*11549*	*25.4*

Table 2. Repeats Summary

To see if there is any novel repeats, I performed a blastn against all fosmids on my masked fosmid in goose and subsequently used the Herne viewer to visualize the repeats superimposed on the fosmid with regions of conservation (Fig. 21). At first I thought I had two potential regions of conservation that were each flanked by repeat regions. These regions could have been inserted repeats or a repeats that Repeat Masker missed. I proceeded in getting two types of extracts from these regions to perform blastn (using repeat superlibrary): one getting only the in-between sequences and the other including the repeats surrounding the in-between sequences. Not only did I not get any significant hits, but also when I went back to the individual transcripts (red arrows), which indicated that these two regions were each coming from one region (or fosmid). So I ended up not finding any novel repeats.

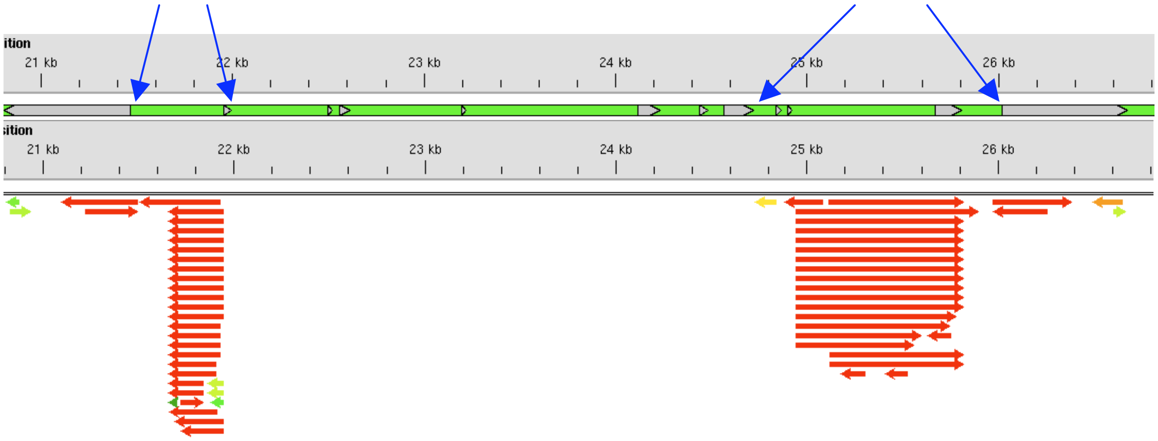


Figure 21: Herne output of blastn for finding repeats

More details on repeats can be seen in Table 3 below.

Start	End	Length	Repeat Family	Repeat
681	720	40	Low_complexity	AT_rich
759	805	47	Low_complexity	AT_rich
931	959	29	Low_complexity	AT_rich
1105	1314	210	TRF	dvir.11.33.centroid
1913	1971	59	Simple_repeat	(TA)n
2151	2184	34	Simple_repeat	(TATATG)n
2313	2371	59	Simple_repeat	(CATATA)n
2862	2926	65	Low_complexity	AT_rich
2939	2969	31	Low_complexity	AT_rich
3177	3226	50	Simple_repeat	(TAAA)n
5158	5275	118	DNA	dvir.16.2.centroid
5182	5306	125	DINE	yakuba_cons
5285	5362	78	DNA	dvir.16.17.centroid
5433	5524	92	DNA	dvir.16.2.centroid
5551	5649	99	DNA	dvir.16.2.centroid
6164	6198	35	Low_complexity	AT_rich
6364	6412	49	Low_complexity	AT_rich
6576	6608	33	Low_complexity	AT_rich
6957	6988	32	Low_complexity	AT_rich
7222	7279	58	Simple_repeat	(TA)n
7372	7406	35	Low_complexity	AT_rich
7593	7624	32	Simple_repeat	(TA)n
7625	12955	5331	LTR	dvir.3.94.centroid
14511	14555	45	Simple_repeat	(ATG)n
15441	15476	36	Simple_repeat	(CTG)n
16278	16322	45	Simple_repeat	(CTG)n
18578	18617	40	Simple_repeat	(CATG)n
18772	18809	38	Simple_repeat	(TA)n
18810	18850	41	Simple_repeat	(CAGT)n
18860	18889	30	Low_complexity	AT_rich
19159	19292	134	TRF	dvir.11.33.centroid
19601	19644	44	TRF	dvir.11.33.centroid
19689	19779	91	TRF	dvir.11.33.centroid
19911	19951	41	Simple_repeat	(CATA)n
20130	20188	59	Simple_repeat	(CATATA)n
20359	20482	124	Simple_repeat	(CATA)n
20808	21469	662	Unknown	dvir.22.25.centroid
21957	21993	37	Low_complexity	AT_rich
22500	22522	23	Simple_repeat	(TATTG)n
22560	22613	54	Simple_repeat	(TA)n
23198	23219	22	Low_complexity	AT_rich
24117	24231	115	Low_complexity	AT_rich

24436	24483	48	Low_complexity	AT_rich
24565	24722	158	Simple_repeat	(CTG)n
24840	24865	26	Low_complexity	AT_rich
24899	24919	21	Low_complexity	AT_rich
25671	25806	136	Unknown	dvir.22.25.centroid
26018	26673	656	Unknown	dvir.22.25.centroid
28095	28214	120	LINE	PENELOPE
28217	28243	27	Simple_repeat	(TA)n
31044	31080	37	Low_complexity	AT_rich
33423	33483	61	DINE	DNAREP1_DM
33484	33507	24	TRF	dvir.11.23.centroid
34575	34596	22	Low_complexity	AT_rich
36164	36233	70	Simple_repeat	(TATG)n
36410	36583	174	Simple_repeat	(TATG)n
39416	39451	36	Low_complexity	AT_rich
40148	40952	805	DNA	dvir.16.2.centroid
40953	40998	46	Simple_repeat	(CGGA)n
40999	41152	154	DNA	dvir.16.2.centroid
42711	42742	32	Low_complexity	AT_rich
43126	43190	65	DNA	dvir.16.17.centroid
43195	43603	409	DNA	dvir.16.2.centroid

Table 3. Detailed Repeats Location

V. Synteny

In terms of synteny, there is a mix of preservation and non-preservation. For the chromosome 3L genes, synteny is lacking for the most part (Fig. 22). CG16719 is in the same orientation and spans about the same number of bases, but taking the two genes together, they are much farther apart in *D. melanogaster* than in *D. virili*. Perhaps there is some biological mechanism in place, like recombinations, that allows the separation of two closely located genes over evolutionary time. Also, CG5262, which I found stronger evidence for being a gene than CG16719, has an opposite orientation in *D. melanogaster* compared to *D. virilis*.

As for chromosome 4, there was strong synteny for three of four genes (Fig. 23). CG1629 (yellow-h) had opposite orientation in *D. melanogaster* compared to *D. virilis*. Nevertheless, the four genes are aligned in similar placement in the two species and span similar number of bases, leading me to conclude that overall, there is synteny for these genes.

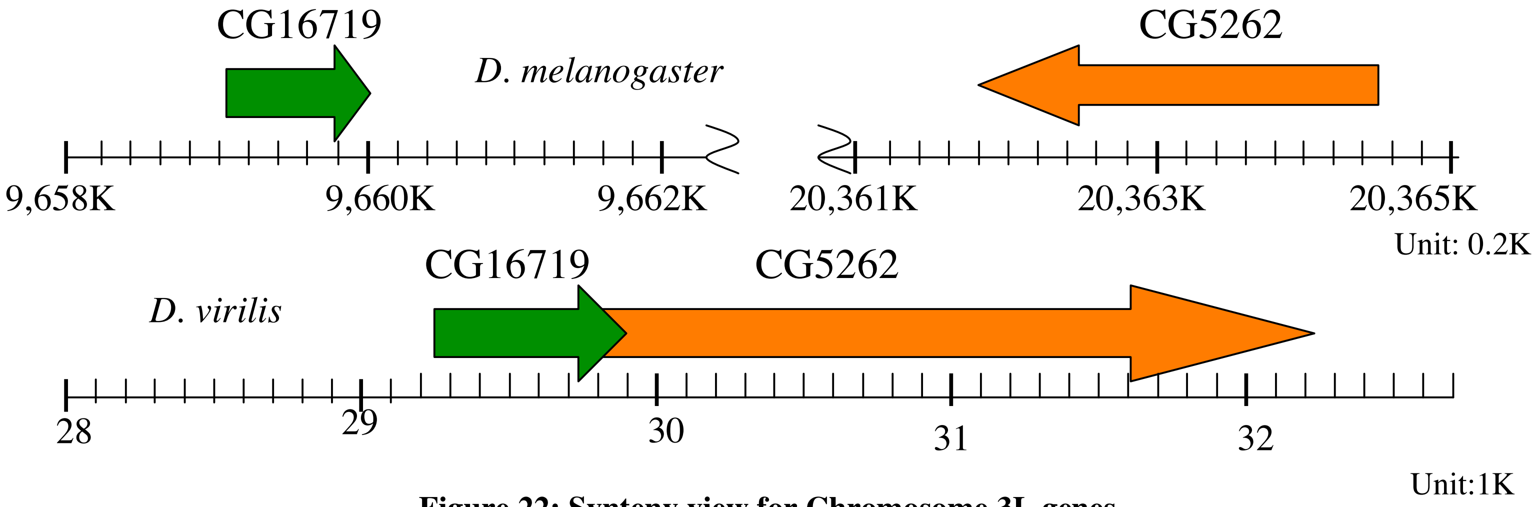


Figure 22: Synteny view for Chromosome 3L genes

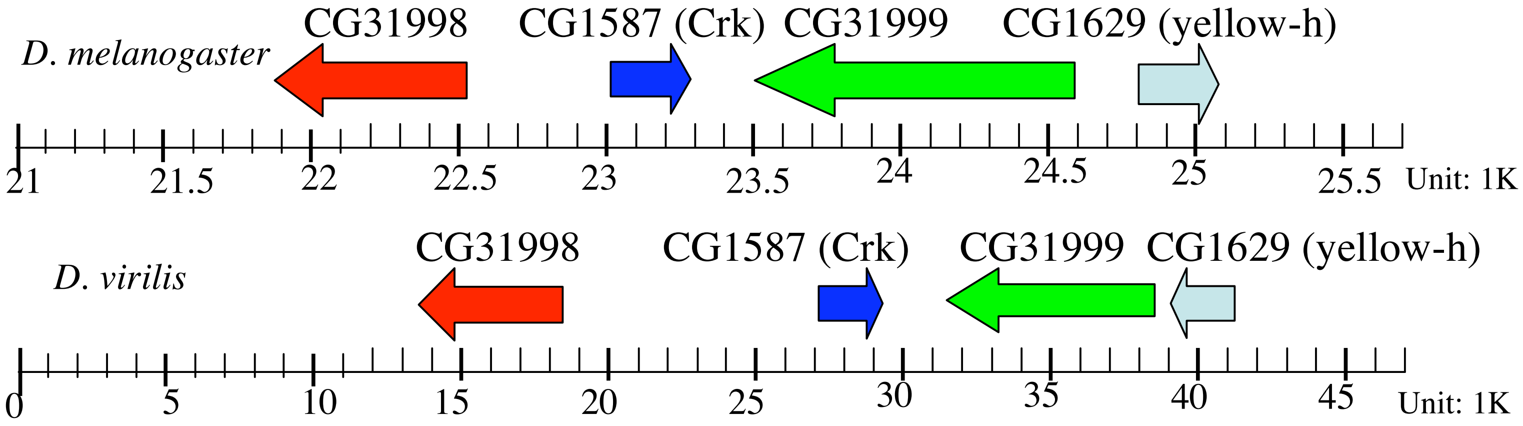


Figure 23: Synteny view for Chromosome 4 genes

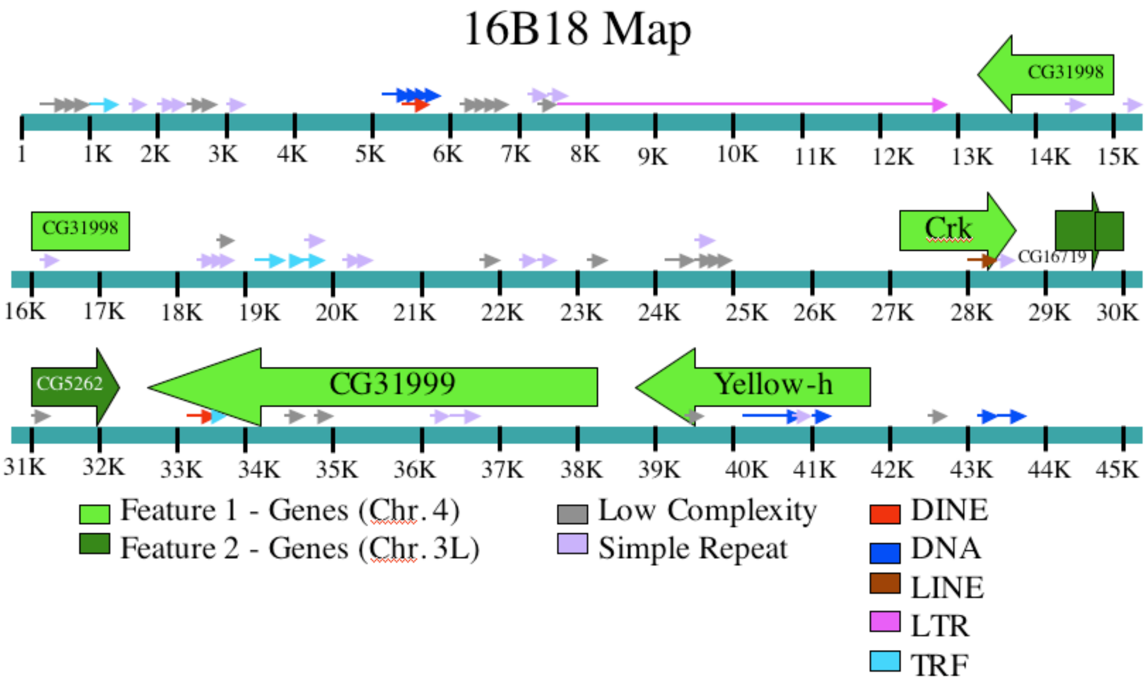


Figure 24: Map of my fosmid (reprise)

Appendix

- FASTA files goes here (see electronic copies)
  - Translated protein sequences for genes in my *D. virilis* fosmid
  - Nucleic sequences which code for the proteins from genes in my *D. virilis* fosmid
  - Genomic region around each gene, 500 bp upstream and downstream of the coding sequence
  - ClustalW input for Crk (CG1587) gene and CG1629 promoter region

Feature	Transcript ID	Strand	Exon	Start	Stop	Phase
6.2	CG31998	Minus (-)	1	18,928	18,991	0
6.2	CG31998	Minus (-)	2	17,692	18,358	2
6.2	CG31998	Minus (-)	3	17,114	17,337	1
6.2	CG31998	Minus (-)	4	13,116	17,035	2
6.4	CG1587	Plus (+)	1	27,071	27,102	0
6.4	CG1587	Plus (+)	2	27,256	27,378	1
6.4	CG1587	Plus (+)	3	27,444	27,648	1
6.4	CG1587	Plus (+)	4	27,718	27,846	0
6.4	CG1587	Plus (+)	5	28,397	28,555	0
6.4	CG1587	Plus (+)	6	28,614	28,859	0
6.5	CG16719	Plus (+)	1	29261	29935	0
6.6	CG5262	Plus (+)	1	29,711	29,764	0
6.6	CG5262	Plus (+)	2	30,483	30,946	0
6.6	CG5262	Plus (+)	3	31,211	31,434	1
6.6	CG5262	Plus (+)	4	31,501	32,261	2
6.7	CG31999	Minus (-)	1	38,128	38,200	0
6.7	CG31999	Minus (-)	2	36,834	37,163	0
6.7	CG31999	Minus (-)	3	35,772	36,044	0
6.7	CG31999	Minus (-)	4	35,593	35,658	2
6.7	CG31999	Minus (-)	5	35,451	35,508	0
6.7	CG31999	Minus (-)	6	35,011	35,316	2
6.7	CG31999	Minus (-)	7	34,427?	34,560?	2?
6.7	CG31999	Minus (-)	8	33,891	34,356	2
6.7	CG31999	Minus (-)	9	33,683	33,815	2
6.7	CG31999	Minus (-)	10	33,531	33,619	2
6.7	CG31999	Minus (-)	11	32,899	33,129	1
6.7	CG31999	Minus (-)	12	32,511	32,675	1
6.8	CG1629	Minus (-)	1	41,582	41,784	0
6.8	CG1629	Minus (-)	2	39,457	40,129	1
6.8	CG1629	Minus (-)	3	38,908	39,396	0

Table 4: Exon boundaries of each gene found in my fosmid