Yakov Rubinchik

Fosmid 17

Overview:

I was given a 50k base long sequence from the *D. erecta* and told to annotate all the proteins in the sequence. First I ran my fosmid through RepeatMasker to eliminate all repeats and then I ran the masked sequence through Genscan, Nscan, and comparing it to the homologous region on *D. melanogaster* that has already been annotated. Through these processes I was able to find all the genes and pseudogenes in my fosmid and these are shown below in Overview-Figure 1. I found 3 functional genes and 2 pseudogenes. Through the entire process I also discovered which tools were helpful and which were not. Nscan predicted 2 functional genes with accurate exon/intron borders while Genscan predicted 6 functional genes of which only 2 were functional and real genes. I found that once I had the homologous *D. melanogaster* region, blasting the mRNA of the genes in the homologous region against my fosmid proved to be the most useful technique in finding functional genes and finding accurate exon/intron borders. Sixpack was also helpful in finding the exon/intron borders because it showed all 6 reading frames for my fosmid.

Overview-Figure 1: Genes 1,2, and 3 are 3 genes homologous to 3 genes in *D. melanogaster*. Pseudogene 1 and 2 are 2 pesudogenes homologous to 2 real genes in *D. melanogaster*.

Genes:

Gene 1 is homologous to CG7133(NP_649379.1) in *D. melanogaster*. The gene only has 1 isoform. There is only 1 exon and its sequence is on the minus strand running from 1274-288bp on the fosmid. My starting point for this gene was with Genscan's predicted gene 1. Results for Genscan's gene 1 are in the figure labeled "Gene 1-Figure 1" below.

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.02	PlyA	-	188	183	6							1.05
1.01	Sngl	-	1274	285	990	2	0	42	43	586	0.985	46.37
1.00	Prom	-	2814	2775	40							-7.76

```
>Dere3_dna|GENSCAN_predicted_peptide_1|329_aa
MSNVYKDHYHVLGLARNASDSEIREAFRRLSLQYHPDKNENGAGEFLKINDAYRVLIDHH
KRASYDRRLSFRDLEAIIPSENASGQLSELRIIKTSPGNFHKKLKVAVVIGGVLVGTYVA
YRVFQKSPPIIPVPQPITPPAIPTQELSELHPGYLWTLISGLVTLRSKRILSLGKLATGA
NIRVSPSTLKAPLSSAAEVVAKTVIQGPGGVGSTATSSSSLATPANVVAKMAKTLFKGSR
SGLYTASKTVVVPSTEILHWSKTAAKCTLATLKKGTSYARSPVPSGSVLLSSIPSALRAF
AGTFRPALVKSKNIVVKKASASWIQKKRL
```

Gene 1-Figure 1: Results for Genscan's gene 1

I extracted bp 1-3000 from my fosmid using extractseq and blasted it against *D. melanogaster* in Flybase. I only got 1 hit that had a reasonable e-score. The results are in the figure labeled "Gene 1-Figure 2" below.

>gnl|dmel|3L type=chromosome_arm; loc=3L:1..24543557; ID=3L; dbxref=GB:AE014296;
MD5=ec7148cae3daabbd2a226eaa6e85d7c2; length=24543557; release=r5.17; species=Dmel;
Length = 24543557

HSP # = 1 , Score = 270.093 bits (136) , Expect = 2.05781e-70
Identities = 315 / 378 (83.3%) , Positives = 315 / 378 (83.3%) , Gaps = 15 / 378 (4%)
Strand = Plus / Plus

Genome View | Subject FASTA

```
Query:    593         CCACGTTCGCTGGCGTTGCTAAGGACGAACTGGAAGTTGCAGTTGAACCTACGCCTCCTG   652
                      ||||||||||  |  ||  ||||||||||||||||||||||||||||  |||||||||  |||  |
Subject:              CCACGTTCGCAGCCGATGCTAAGGACGAACTGGAAGTTGCAGAAGAACCTACGGCTCTTT   22066811
22066752
Query:    653         GCCCTTGGATTACTGTTTTGGCTACGACTTCGGCAGCCGAAGATAAGGGTGCTTTCAGAG   712
                      |  |||  |||||||||||||||||  |||||||  ||||||||||||  |||||  ||  |||||||
Subject:              GGCCTCGGATTACTGTTTTGGCCACGACTTTGGCAGCCGAAGAAAAGGGCGCGTTCAGAG   22066871
22066812
Query:    713         TACTAGGAGAAACTCGAATATTTGCTCCTGTAGCCAGTTTTCCGAGACTCAGTATTCTTT   772
                      |  ||||||||||||||||  |||||||  |  ||   ||||||||  |||||  |  ||||||||||||
Subject:              TCCTAGGAGAAACTCGAGTATTTGCCCATGGCGCCAGTTTGCCGAGTCCCAGTATTCTTT   22066931
22066872
Query:    773         TCGATCTCAATGTCACTAGCCCGGAGATCAATGTCCATAGGTATCCGGGATGCAGTTCGC   832
                      ||||||||||||  ||  ||||||||||   ||||||||||||||  |||||  |||||    ||||
Subject:              TCGATCTCAATGCCAATAGCCCGGAGGACAATGTCCATAGGGATCCGAGATGCGAATCGC   22066991
22066932
Query:    833         TTAGTTCCTGCGTTGGAATAGCTGGAGGAGTGATTGGCTGGGGAACTGGGATTATTGGCG   892
                      ||||||||||||||||                  ||||||  ||||||||||||||    |||||
Subject:              TTAGTTCCTGCGTTGGA----------------ATTGGCCGGGGAACTGGGATAGATGGCG   22067036
22066992
Query:    893         GAGATTTCTGGAATACCCGGTACGCCACATAGGTGCCAACCAGCACGCCTCCGATTACGA   952
                      |  |  ||||||||||||||||||||  ||||  ||  ||  ||||||||||  |||||||||  |  |
Subject:              GCGGTTTCTGGAATACCCGGTACCCCACGTATGTACCAACCAGCAAGCCTCCGATGAAGG   22067096
22067037
Query:    953         CAGCCACTTTGAGCTTTT    970
                      ||||  ||||||||||||||
Subject:              CAGCAACTTTGAGCTTTT    22067114
```

Gene 1-Figure 2: Results from nblast of bp 1-3000 from my fosmid and *D. melanogaster* genome.

The region shown in Gene 1-Figure 2 showed a homology between bp 593-970 of my extracted sequence and a region in the gene CG7133 of *D. melanogaster*. I then did a pblast of Genscan's predicted peptide sequence for gene 1 against *D. melanogaster*. Once again I only got 1 hit that had a reasonable e-score. The results are shown below in the figure labeled "Gene 1-Figure 3."

>gnl|dmel|FBpp0078138 type=protein; loc=3L:complement(22066375..22067436); ID=FBpp0078138; name=CG7133-PA;
parent=FBgn0037150, FBtr0078485; dbxref=FlyBase:FBpp0078138, FlyBase_Annotation_IDs:CG7133-PA, GB_protein:AAF51805.2,
REFSEQ:NP_649379, GB_protein:AAF51805; MD5=6cdf10d2532767019459ec3c5a423d60; length=353; release=r5.17;
species=Dmel;
Length = 353

HSP # = 1 , Score = 280.796 bits (717) , Expect = 7.83597e-76
Identities = 161 / 298 (54%) , Positives = 191 / 298 (64.1%) , Gaps = 31 / 298 (10.4%)

Subject FASTA

```
Query:    1           MSNVYKDHYHVLGLARNASDSEIREAFRRLSLQYHPDKNENGAGEFLKINDAYRVLIDHH   60
                      MS+VY+DHY VLGL RNA+DSEI++AFRRLSLQYHPDKNE+GA EFL+IN+A+RVLIDH
Subject:  1           MSDVYEDHYQVLGLPRNATDSEIKDAFRRLSLQYHPDKNEDGAKEFLRINEAHRVLIDHQ   60

Query:    61          KRASYDRRLSFRDLEAIIPSENASGQLSEL------RIIKTSPGNFHKKLKVAVVIGGVL   114
                      +RA YD      D+EAIIP+ENA+GQL EL         +T P +F +KLKVA  IGG+L
Subject:  61          RRALYDCCFQSMDVEAIIPAENANGQLPELGNPFFPMPPETPPASFREKLKVAAFIGGLL   120

Query:    115         VGTYVAYRVFQKSXXXXXXXXXXXXXXXXTQELSELHPGYLWTLISGLVTLRSKRILSLG   174
                      VGTYV YRVFQK                TQELS+ H G LWTL SGL+ LRSKRIL LG
Subject:  121         VGTYVGYRVFQKP-----PPSIPVPRPIPTQELSDSHLGSLWTLSSGLLALRSKRILGLG   175

Query:    175         KLATGANIRVSPSTLKAPLSSAAEVVAKTVIQGPGGVGXXXXXXXXXXXXXNVVAK----   230
                      KLA  AN RVSP TL AP SSAA+VVAKTVI+G   VG             NV  K
Subject:  176         KLAPWANTRVSPRTLNAPFSSAAKVVAKTVIRGQRAVGSSATSSSSLASAANVAVKSLPS   235

Query:    231         ----------MAKTLFKGSRSGLYTASKTVVVPST----EILHWSKTA--AKCTLATL   272
                            +AKTL +GSR+G Y+A KTV   +      +L+W+ T    K T AT+
Subject:  236         KASVNSATETVAKTLSQGSRAGPYSALKTVWSSAVSYLRSLLNWATTPKWGKATPATI   293
```

Gene 1-Figure 3: Results of a pblast of Genscan's predicted peptide sequence for gene 1 against *D. melanogaster*.

The predicted peptide sequence from Genscan was homologous to CG7133 from *D. melanogaster*. Since both the nblast and pblast for Genscan's predicted gene 1 matched up with CG7133 I expected to find a homologous gene to CG7133 in my fosmid. I decided to run a nblast 2 of the mRNA sequence from CG7133 against my fosmid to see where in my fosmid the homologous sequence was. The results are below in the figure labeled "Gene 1-Figure 4."

```
 Score =  520 bits (281),  Expect = 2e-149
 Identities = 580/715 (81%), Gaps = 57/715 (7%)
 Strand=Plus/Minus

Query  184   AAAATGAGCGATGTCTACGAAGATCACTACCAGGTTCTGGGCTTACCGAGAAATGCCACC  243
             ||||||||| | |||||| |||| |||||||| ||||||||||  ||||||| |||| |
Sbjct  1277  AAAATGAGCAACGTCTACAAAGACCACTACCATGTTCTGGGCTTGGCGAGAAACGCCAGC  1218

Query  244   GACAGTGAGATTAAG-GATGCTTTTCGGCGGCTGTCCCTGCAATATCATCCCGACAAAAA  302
             |||||||||| | || || |||||||||||| ||||||||||||||||||||||||||||
Sbjct  1217  GACAGTGAGA-TCAGAGAAGCTTTTCGGCGGTTGTCCCTGCAATATCATCCCGACAAAAA  1158

Query  303   CGAGGATGGAGCGAAGGAGTTCCTTAGAATCAACGAGGCCCATCGCGTCCTGATTGACCA  362
             |||| ||||||||  |||||||||| ||||||||| ||| | ||||| || ||||||||
Sbjct  1158  CGAGAATGGAGCGGGGGAGTTCCTTAAAATCAACGACGCCTACCGCGTGCTAATTGACCA  1099

Query  363   TCAGAGAAGGGCCTT-GTACGATTG-CTGC-T-TCCAGTCCATGGACGTT-GAAGCCATT  417
             ||| | |||||| || ||||||| | | || | ||  ||  | |||| || ||||||||
Sbjct  1098  TCATAAAAGGGC-TTCGTACGATCGTC-GCCTGTC--GTTTAGGGAC-TTAGAAGCCATC  1041

Query  418   ATTCCCG-CTGAGAACGCTAATGGCCAACTGCCTGAATTGGGAAATCCATTCTTCCCAAT  476
             ||||| | | |||||||||| |||||||||| |||||||  ||| || || | |     ||
Sbjct  1043  ATTCC-GTCCGAGAACGCTAGTGGCCAACTGTCTGAATTACGAA-TC-AT-CA----AA-  993

Query  477   GCCACCCGAAACGCCGCCTGCTAGTTTTCGCGAAAAGCTCAAAGTTGCTGCCTTCATCGG  536
                ||   |  ||||   || || ||| | | ||||||||||||| |||| | | |||||
Sbjct  992   ---AC---AT-CGCC---TGGTAATTTCCACAAAAAGCTCAAAGTGGCTGTCGTAATCGG  943

Query  537   AGGCTTGCTGGTTGGTACATACGTGGGGTACCGGGTATTCCAGAAACCGCCGCCATCTA-  595
             |||| |||||||||| || || |||| ||||||||||||||||||| | ||||||  ||
Sbjct  942   AGGCGTGCTGGTTGGCACCTATGTGGCGTACCGGGTATTCCAGAAATCTCCGCCAA-TAA  884

Query  596   TCCCAGTTCCCCGGCCAAT---TCC---A---A--C----GCAGGAACTAAGCGAT-TCG  639
             |||||||||||| ||||||   |||   |   |  |    |||||||||||||||  | |
Sbjct  883   TCCCAGTTCCCCAGCCAATCACTCCTCCAGCTATTCCAACGCAGGAACTAAGCGAACT-G  825

Query  640   CATCTCGGATCCCTATGGACATTG-TCCTCCGGGCTATTGGCATTGAGATCGAAAAGAAT  698
             |||| ||||| |||||||||||||| || ||||||||| || ||||||||||||||||||
Sbjct  824   CATCCCGGATACCTATGGACATTGATC-TCCGGGCTAGTGACATTGAGATCGAAAAGAAT  766

Query  699   ACTGGGACTCGGCAAACTGGCGCCATGG-GCAAATACTCGAGTTTCTCCTAGGACTCTGA  757
             |||| | ||||| |||||||||  || || ||||||| |||||||||||||| |||||||
Sbjct  765   ACTGAGTCTCGGAAAACTGGCTACA-GGAGCAAATATTCGAGTTTCTCCTAGTACTCTGA  707

Query  758   ACGCGCCCTTTTCTTCGGCTGCCAAAGTCGTGGCCAAAACAGTAATCCGAGG-CCAAAGA  816
             | || ||||| |||||||||||| ||||||| |||||||||||||||| ||| |||  ||
Sbjct  706   AAGCACCCTTATCTTCGGCTGCCGAAGTCGTAGCCAAAACAGTAATCCAAGGGCCAG-GA  648

Query  817   GCCGTAGGTTCTTCTGCAACTTCCAGTTCGTCCTTAGCATCGGCTGCGAACGTGG  871
             | |||||||||  |||||||||||||||||||||||| || | ||||||||||
Sbjct  647   GGCGTAGGTTCAACTGCAACTTCCAGTTCGTCCTTAGCAACGCCAGCGAACGTGG  593
```

Gene 1-Figure 4: Results from a nblast 2 of the mRNA sequence from CG7133 against my fosmid

I ran my entire masked fosmid through sixpack to get all the reading frames and then used the results from Gene 1-Figure 4 to see where the first start codon was and where the first stop codon was on the same reading frame. That is where I got the exon borders at 1274-288bp on my fosmid. I then ran gene 1with its coordinates for the exon through Gene checker which verified the gene. I ran a pblast 2 of the peptide sequence from Gene checker for gene 1 against the peptide sequence for CG7133. The results are shown below in the figure labeled "Gene 1-Figure 5".

```
>lcl|42903 FBpp0078138 type=protein; loc=3L:complement(22066375..22067436);
ID=FBpp0078138; name=CG7133-PA; parent=FBgn0037150,FBtr0078485;
dbxref=FlyBase:FBpp0078138,FlyBase_Annotation_IDs:CG7133-PA,GB_protein:AAF51805.2,REFSEQ:NP_649379,GB_protein
MD5=6cdf10d2532767019459ec3c5a423d60; length=353; release=r5.17;
species=Dmel;
Length=353

 Score =  283 bits (725),  Expect = 3e-81, Method: Compositional matrix adjust.
 Identities = 186/338 (55%), Positives = 223/338 (65%), Gaps = 39/338 (11%)

Query  1    MSNVYKDHYHVLGLARNASDSEIREAFRRLSLQYHPDKNENGAGEFLKINDAYRVLIDHH  60
            MS+VY+DHY VLGL RNA+DSEI++AFRRLSLQYHPDKNE+GA EFL+IN+A+RVLIDH
Sbjct  1    MSDVYEDHYQVLGLPRNATDSEIKDAFRRLSLQYHPDKNEDGAKEFLRINEAHRVLIDHQ  60

Query  61   KRASYDRRLSFRDLEAIIPSENASGQLSEL------RIIKTSPGNFHKKLKVAVVIGGVL  114
            +RA YD      D+EAIIP+ENA+GQL EL         +T P +F +KLKVA  IGG+L
Sbjct  61   RRALYDCCFQSMDVEAIIPAENANGQLPELGNPFFPMPPETPPASFREKLKVAAFIGGLL  120

Query  115  VGTYVAYRVFQKSPPIIPVPQPITPPAIPTQELSELHPGYLWTLISGLVTLRSKRILSLG  174
            VGTYV YRVFQK       P    P  IPTQELS+ H G LWTL SGL+ LRSKRIL LG
Sbjct  121  VGTYVGYRVFQKP-----PPSIPVPRPIPTQELSDSHLGSLWTLSSGLLALRSKRILGLG  175

Query  175  KLATGANIRVSPSTLKAPLSSAAEVVAKTVIQGPGGVGSTATSSSSLATPANVVAK----  230
            KLA  AN RVSP TL AP SSAA+VVAKTVI+G   VGS+ATSSSSLA+ ANV  K
Sbjct  176  KLAPWANTRVSPRTLNAPFSSAAKVVAKTVIRGQRAVGSSATSSSSLASAANVAVKSLPS  235

Query  231  ----------MAKTLFKGSRSGLYTASKTVVVPST----EILHWSKT-----AAKCTLAT  271
                    +AKTL +GSR+G Y+A KTV   +      +L+W+ T     A   T+A
Sbjct  236  KASVNSATETVAKTLSQGSRAGPYSALKTVWSSAVSYLRSLLNWATTPKWGKATPATIAG  295

Query  272  LKKG-----TSYARSPVPSGSVLLSSIPSALRAFAGTF  304
            L        +S A++ + +     S I S LRA AG F
Sbjct  296  LVHNSRPVWSSAAKNTLAALMYACSKISSYLRALAGRF  333
```

Gene 1-Figure 5: Results from a pblast 2 of the peptide sequence from Gene checker for gene 1 against the peptide sequence for CG7133.

The protein sequence is not incredibly well conserved as shown by the 55% for identities, 65% for positives, and that the peptide sequence for gene 1 from the fosmid is 329 a.a. and the peptide sequence for CG7133 is 353 a.a. I wanted to find out why there was 24 a.a. difference in peptide length so I ran a clustal2w of the coding sequence for CG7133 and my gene 1. The results are shown below in the figure labeled “Gene 1-Figure 6”.

```
FBtr0078485     ATGAGCGATGTCTACGAAGATCACTACCAGGTTCTGGGCTTACCGAGAAATGCCACCGAC 60
EMBOSS_001      ATGAGCAACGTCTACAAAGACCACTACCATGTTCTGGGCTTGGCGAGAAACGCCAGCGAC 60
                ****** * ****** **** ******** ***********  ******* **** ****

FBtr0078485     AGTGAGATTAAGGATGCTTTTCGGCGGCTGTCCCTGCAATATCATCCCGACAAAAACGAG 120
EMBOSS_001      AGTGAGATCAGAGAAGCTTTTCGGCGGTTGTCCCTGCAATATCATCCCGACAAAAACGAG 120
                ******** *  ** ************ ********************************

FBtr0078485     GATGGAGCGAAGGAGTTCCTTAGAATCAACGAGGCCCATCGCGTCCTGATTGACCATCAG 180
EMBOSS_001      AATGGAGCGGGGGAGTTCCTTAAAATCAACGACGCCTACCGCGTGCTAATTGACCATCAT 180
                 ********  *********** ********* *** * ***** ** ***********

FBtr0078485     AGAAGGGCCTTGTACGATTGCTGCTTCCAGTCCATGGACGTTGAAGCCATTATTCCCGCT 240
EMBOSS_001      AAAAGGGCTTCGTACGATCGTCGCCTGTCGTTTAGGGACTTAGAAGCCATCATTCCGTCC 240
                * ****** * ******* *  ** *   **  * **** * ******** *****  *

FBtr0078485     GAGAACGCTAATGGCCAACTGCCTGAATTGGGAAATCCATTCTTCCCAATGCCACCCGAA 300
EMBOSS_001      GAGAACGCTAGTGGCCAACTGTCTGAATTACGAA--TCAT----------------CAAA 282
                ********** ********** *******  ***   ***                * **

FBtr0078485     ACGCCGCCTGCTAGTTTTCGCGAAAAGCTCAAAGTTGCTGCCTTCATCGGAGGCTTGCTG 360
EMBOSS_001      ACATCGCCTGGTAATTTCCACAAAAAGCTCAAAGTGGCTGTCGTAATCGGAGGCGTGCTG 342
                **  ****** ** *** * * ************* **** * * ********* *****

FBtr0078485     GTTGGTACATACGTGGGGTACCGGGTATTCCAGAAACCGCCGCCATCTATCCCAGTTCCC 420
EMBOSS_001      GTTGGCACCTATGTGGCGTACCGGGTATTCCAGAAATCTCCGCCAATAATCCCAGTTCCC 402
                ***** ** ** **** ******************* * ******   ************

FBtr0078485     CGGCCAAT--------------TCCAACGCAGGAACTAAGCGATTCGCATCTCGGATCC 465
EMBOSS_001      CAGCCAATCACTCCTCCAGCTATTCCAACGCAGGAACTAAGCGAACTGCATCCCGGATAC 462
                * ******              **********************   ***** ***** *

FBtr0078485     CTATGGACATTGTCCTCCGGGCTATTGGCATTGAGATCGAAAAGAATACTGGGACTCGGC 525
EMBOSS_001      CTATGGACATTGATCTCCGGGCTAGTGACATTGAGATCGAAAAGAATACTGAGTCTCGGA 522
                ************  ********** ** ********************** * *****

FBtr0078485     AAACTGGCGCCATGGGCAAATACTCGAGTTTCTCCTAGGACTCTGAACGCGCCCTTTTCT 585
EMBOSS_001      AAACTGGCTACAGGAGCAAATATTCGAGTTTCTCCTAGTACTCTGAAAGCACCCTTATCT 582
                ********  ** * ******* *************** ******** ** ***** ***

FBtr0078485     TCGGCTGCCAAAGTCGTGGCCAAAACAGTAATCCGAGGCCAAAGAGCCGTAGGTTCTTCT 645
EMBOSS_001      TCGGCTGCCGAAGTCGTAGCCAAAACAGTAATCCAAGGGCCAGGAGGCGTAGGTTCAACT 642
                ********* ******* **************** *** * * *** *********  **
```

Gene 1-Figure 6: Results from a clustal2w of the coding sequence for CG7133 and my gene 1. The highlighted hyphens are regions in CG7133 that are not present in my gene. The highlighted "A" of CG7133 and the highlighted "TAG" of my gene show a nonsense mutation.

The results of the clustalw2 show that there are several regions of base pairs in the sequence of CG7133 that are not present in my gene 1. Also, there is a nonsense mutation where there was a point mutation in my gene from an "A" to a "T" resulting in a premature stop codon in my gene 1. Since my gene 1 protein was not very conserved with the CG7133 protein I thought that maybe it may be a pseudogene in my fosmid. First I looked at the function of the CG7133 in *D. melanogaster* and found that it functions in unfolded protein binding and heat shock protein binding. This function seems important so *D. erecta* probably has this functioning gene. I ran a nblast 2 of the mRNA from CG7133 against the entire *D. erecta* genome to find out if maybe the ortholog of CG7133 was found somewhere else in *D. erecta* but I did not get any good hits. Based on the importance of the gene and the fact that in flies pseudogenes are rare, I concluded that this gene is probably a real gene and that though there is not very high conservation between the proteins, the protein coded for by my gene 1 is still functional.

Gene 2 is homologous to CG7130(NP_649380.1) in *D. melanogaster*. The gene only has 1 isoform. There is only 1 exon and its sequence is on the minus strand running from 3823-3443bp on the fosmid. My starting point for this gene was with Nscan's predicted gene 1. Results for Nscan's gene 1 are in the figure labeled "Gene 2-Figure 1" below.

Exon	Strand	Begin	End	Length
1	-	3439	3868	429

127 aa

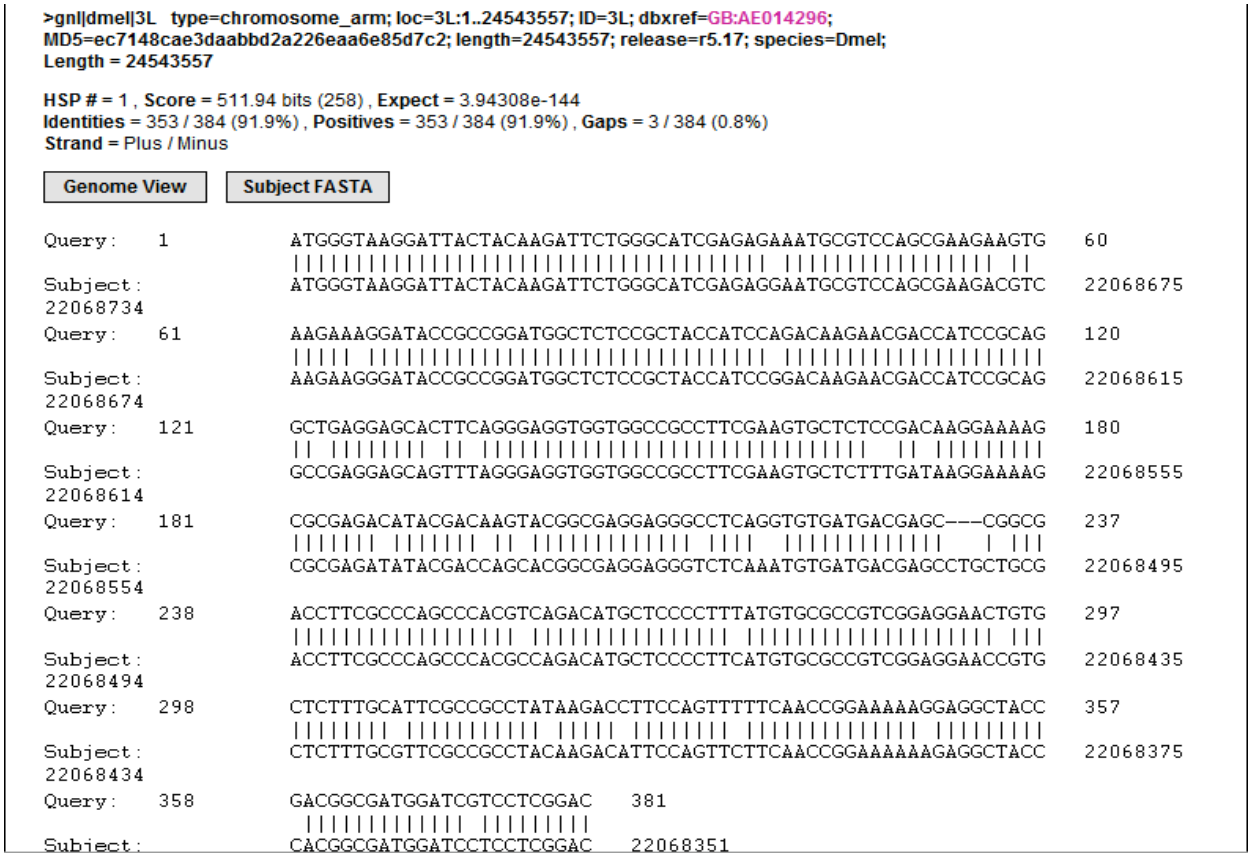
MGKDYYKILGIERNASSEEVKKGYRRMALRYHPDKNDHPQAEEHFREVVA
AFEVLSDKEKRETYDKYGEEGLRCDDEPATFAQPTSDMLPFMCAVGGTVL
FAFAAYKTFQFFNRKKEATDGDGSSSD

Transcript:

ATGGGTAAGGATTACTACAAGATTCTGGGCATCGAGAGAAATGCGTCCAG
CGAAGAAGTGAAGAAAGGATACCGCCGGATGGCTCTCCGCTACCATCCAG
ACAAGAACGACCATCCGCAGGCTGAGGAGCACTTCAGGGAGGTGGTGGCC
GCCTTCGAAGTGCTCTCCGACAAGGAAAAGCGCGAGACATACGACAAGTA
CGGCGAGGAGGGCCTCAGGTGTGATGACGAGCCGGCGACCTTCGCCCAGC
CCACGTCAGACATGCTCCCCTTTATGTGCGCCGTCGGAGGAACTGTGCTC
TTTGCATTCGCCGCCTATAAGACCTTCCAGTTTTTCAACCGGAAAAAGGA
GGCTACCGACGGCGATGGATCGTCCTCGGAC

Gene 2-Figure 1: Results for Nscan's gene 1

I extracted bp 3000-4000 from my fosmid using extractseq and blasted it against *D. melanogaster* in Flybase. I only got 1 hit that had a reasonable e-score. The results are in the figure labeled "Gene 2-Figure 2" below.



>gnl|dmel|3L type=chromosome_arm; loc=3L:1..24543557; ID=3L; dbxref=GB:AE014296;
MD5=ec7148cae3daabbd2a226eaa6e85d7c2; length=24543557; release=r5.17; species=Dmel;
Length = 24543557

HSP # = 1 , Score = 511.94 bits (258) , Expect = 3.94308e-144
Identities = 353 / 384 (91.9%) , Positives = 353 / 384 (91.9%) , Gaps = 3 / 384 (0.8%)
Strand = Plus / Minus

Genome View | Subject FASTA

```
Query:    1          ATGGGTAAGGATTACTACAAGATTCTGGGCATCGAGAGAAATGCGTCCAGCGAAGAAGTG   60
                     ||||||||||||||||||||||||||||||||||||||| |||||||||||||||| ||
Subject:             ATGGGTAAGGATTACTACAAGATTCTGGGCATCGAGAGGAATGCGTCCAGCGAAGACGTC   22068675
22068734
Query:    61         AAGAAAGGATACCGCCGGATGGCTCTCCGCTACCATCCAGACAAGAACGACCATCCGCAG   120
                     ||||| ||||||||||||||||||||||||||||||| ||||||||||||||||||||||
Subject:             AAGAAGGGATACCGCCGGATGGCTCTCCGCTACCATCCGGACAAGAACGACCATCCGCAG   22068615
22068674
Query:    121        GCTGAGGAGCACTTCAGGGAGGTGGTGGCCGCCTTCGAAGTGCTCTCCGACAAGGAAAAG   180
                     || |||||||| || ||||||||||||||||||||||||||||||||  || |||||||||
Subject:             GCCGAGGAGCAGTTTAGGGAGGTGGTGGCCGCCTTCGAAGTGCTCTTTGATAAGGAAAAG   22068555
22068614
Query:    181        CGCGAGACATACGACAAGTACGGCGAGGAGGGCCTCAGGTGTGATGACGAGC---CGGCG   237
                     ||||||| ||||||| || |||||||||||||| ||||  |||||||||||||   | |||
Subject:             CGCGAGATATACGACCAGCACGGCGAGGAGGGTCTCAAATGTGATGACGAGCCTGCTGCG   22068495
22068554
Query:    238        ACCTTCGCCCAGCCCACGTCAGACATGCTCCCCTTTATGTGCGCCGTCGGAGGAACTGTG   297
                     |||||||||||||||||| |||||||||||||||| |||||||||||||||||||| |||
Subject:             ACCTTCGCCCAGCCCACGCCAGACATGCTCCCCTTCATGTGCGCCGTCGGAGGAACCGTG   22068435
22068494
Query:    298        CTCTTTGCATTCGCCGCCTATAAGACCTTCCAGTTTTTCAACCGGAAAAAGGAGGCTACC   357
                     |||||||| |||||||||| ||||| |||||||| ||||||||||||||| |||||||||
Subject:             CTCTTTGCGTTCGCCGCCTACAAGACATTCCAGTTCTTCAACCGGAAAAAAGAGGCTACC   22068375
22068434
Query:    358        GACGGCGATGGATCGTCCTCGGAC   381
                      |||||||||||| |||||||||
Subject:             CACGGCGATGGATCCTCCTCGGAC   22068351
```

Gene 2-Figure 2: Results from nblast of bp 3000-4000 from my fosmid and *D. melanogaster* genome.

The region shown in Gene 2-Figure 2 showed a homology between bp 1-381 of my extracted sequence and a region in the gene CG7130 of *D. melanogaster*. I then did a pblast of Nscan's predicted peptide sequence for gene 1 against *D. melanogaster*. Once again I only got 1 hit that had a reasonable e-score. The results are shown below in the figure labeled "Gene 2-Figure 3."

>gnl|dmel|FBpp0078137 type=protein; loc=3L:complement(22068348..22068734); ID=FBpp0078137; name=CG7130-
parent=FBgn0037151, FBtr0078484; dbxref=FlyBase:FBpp0078137, FlyBase_Annotation_IDs:CG7130-PA,
GB_protein:AAF51806.1, REFSEQ:NP_649380, GB_protein:AAF51806; MD5=fcb3086a827f52f13ec00728065552ab; le
release=r5.17; species=Dmel;
Length = 128

HSP # = 1 , Score = 242.662 bits (618) , Expect = 4.15323e-65
Identities = 118 / 128 (92.2%) , Positives = 122 / 128 (95.3%) , Gaps = 1 / 128 (0.8%)

Subject FASTA

```
Query:    1          MGKDYYKILGIERNASSEEVKKGYRRMALRYHPDKNDHPQAEEHFREVVAAFEVLSDKEK   60
                     MGKDYYKILGIERNASSE+VKKGYRRMALRYHPDKNDHPQAEE FREVVAAFEVL DKEK
Subject:  1          MGKDYYKILGIERNASSEDVKKGYRRMALRYHPDKNDHPQAEEQFREVVAAFEVLFDKEK   60

Query:    61         RETYDKYGEEGLRCDDEP-ATFAQPTSDMLPFMCAVGGTVLFAFAAYKTFQFFNRKKEAT   119
                     RE YD++GEEGL+CDDEP ATFAQPT DMLPFMCAVGGTVLFAFAAYKTFQFFNRKKEAT
Subject:  61         REIYDQHGEEGLKCDDEPAATFAQPTPDMLPFMCAVGGTVLFAFAAYKTFQFFNRKKEAT   120

Query:    120        DGDGSSSD   127
                      GDGSSSD
Subject:  121        HGDGSSSD   128
```

Gene 2-Figure 3: Results of a pblast of Nscan's gene 1 predicted peptide sequence for against *D. melanogaster*.

The predicted peptide sequence from Nscan was homologous to CG7130 from *D. melanogaster*. Since both the nblast and pblast for Nscan's predicted gene 1 matched up with CG7130 I expected to find a homologous gene to CG7130 in my fosmid. I decided to run a nblast 2 of the mRNA sequence from CG7130 against my fosmid to see where in my fosmid the homologous sequence was. The results are below in the figure labeled "Gene 2-Figure 4."

```
>lcl|59535 Dere3_dna range=fosmid17:1-50000 5'pad=0 3'pad=0 strand=+ repeatMasking=none
Length=50000

 Score =  756 bits (838),  Expect = 0.0
 Identities = 534/607 (87%), Gaps = 9/607 (1%)
 Strand=Plus/Minus

Query  94    CAGAACAAGGAGAATACCATTTCACCACAATGGGTAAGGATTACTACAAGATTCTGGGCA  153
             ||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3852  CAGAACAAGGAGAATCCCATTTCACCACAATGGGTAAGGATTACTACAAGATTCTGGGCA  3793

Query  154   TCGAGAGGAATGCGTCCAGCGAAGACGTCAAGAAGGGATACCGCCGGATGGCTCTCCGCT  213
             ||||||| |||||||||||||||| || ||||| ||||||||||||||||||||||||||
Sbjct  3792  TCGAGAGAAATGCGTCCAGCGAAGAAGTGAAGAAAGGATACCGCCGGATGGCTCTCCGCT  3733

Query  214   ACCATCCGGACAAGAACGACCATCCGCAGGCCGAGGAGCAGTTTAGGGAGGTGGTGGCCG  273
             ||||||| ||||||||||||||||||||| |||||||| || ||||||||||||||||||
Sbjct  3732  ACCATCCAGACAAGAACGACCATCCGCAGGCTGAGGAGCACTTCAGGGAGGTGGTGGCCG  3673

Query  274   CCTTCGAAGTGCTCTTTGATAAGGAAAAGCGCGAGATATACGACCAGCACGGCGAGGAGG  333
             ||||||||||||||  || |||||||||||||||| ||||||| || ||||||||||||
Sbjct  3672  CCTTCGAAGTGCTCTCCGACAAGGAAAAGCGCGAGACATACGACAAGTACGGCGAGGAGG  3613

Query  334   GTCTCAAATGTGATGACGAGCCTGCTGCGACCTTCGCCCAGCCCACGCCAGACATGCTCC  393
             | ||||  ||||||||||||||   | ||||||||||||||||||||| ||||||||||||
Sbjct  3612  GCCTCAGGTGTGATGACGAGC---CGGCGACCTTCGCCCAGCCCACGTCAGACATGCTCC  3556

Query  394   CCTTCATGTGCGCCGTCGGAGGAACCGTGCTCTTTGCGTTCGCCGCCTACAAGACATTCC  453
             |||| |||||||||||||||||| |||||||||| |||||||||| ||||| ||||
Sbjct  3555  CCTTTATGTGCGCCGTCGGAGGAACTGTGCTCTTTGCATTCGCCGCCTATAAGACCTTCC  3496

Query  454   AGTTCTTCAACCGGAAAAAAGAGGCTACCCACGGCGATGGATCCTCCTCGGACTGAGCTA  513
             |||| |||||||||||||| ||||||||| |||||||||||| ||||||||||||||||||
Sbjct  3495  AGTTTTTCAACCGGAAAAAGGAGGCTACCGACGGCGATGGATCGTCCTCGGACTGAGCTA  3436

Query  514   AGGATCCAAGGGCTTGATGAAGCAATCTCGGGTACCTAGCGTTCTTCGCTGAATAGTCTT  573
             | ||||  |||||||| |||||||||  |||| | ||||||| |||| ||||||||||||
Sbjct  3435  ACGATCGGAGGGCTTGGTGAAGCAATACCGGGGATCTAGCGTCCTTCACTGAATAGTCTT  3376

Query  574   TAAGATTAATTTATAGGAACTTAATTATTGACTGTTTATCTAATGAATCCTGCGTTACTT  633
             ||||||||||||||||||||||    |||||||||| |||| ||  ||  || | || ||
Sbjct  3375  TAAGATTAATTTATAGGAACTTGTACATTGACTGTTGATCTCATTTATTATGTG-TAGTT  3317

Query  634   ATTGATTAATTTATTTATTTATTTAGTAAGATAAATAAAAATTATGGAATTCGTCCGACT  693
             | ||||  |||  |   ||||||||||||   ||| |||| || |||||| || || |||
Sbjct  3316  ACTGAT--ATTGCTGGTTTTATTTAGTAA---AAACAAAAGTTGTGGAATGCGGCCAACT  3262

Query  694   GTTGCTC  700
              |||||||
Sbjct  3261  CTTGCTC  3255
```

Gene 2-Figure 4: Results from a nblast 2 of the mRNA sequence from CG7130 against my fosmid

I ran my entire masked fosmid through sixpack to get all the reading frames and then used the results from Gene 2-Figure 4 to see where the first start codon was and where the first stop codon was on the same reading frame. That is where I got the exon borders at 3823-3443bp on my fosmid. I then ran gene 2 with its coordinates for the exon through Gene checker which verified the gene. I ran a pblast 2 of the peptide sequence from Gene checker for gene 2 against the peptide sequence for CG7130. The results are shown below in the figure labeled "Gene 2-Figure 5".

```
>lcl|41915 FBpp0078137 type=protein; loc=3L:complement(22068348..22068734);
ID=FBpp0078137; name=CG7130-PA; parent=FBgn0037151,FBtr0078484;
dbxref=FlyBase:FBpp0078137,FlyBase_Annotation_IDs:CG7130-PA,GB_protein:AAF51806.1,REFSEQ:N
MD5=fcb3086a827f52f13ec00728065552ab; length=128; release=r5.17;
species=Dmel;
Length=128

 Score =  242 bits (618),  Expect = 1e-69, Method: Compositional matrix adjust.
 Identities = 118/128 (92%), Positives = 122/128 (95%), Gaps = 1/128 (0%)

Query  1    MGKDYYKILGIERNASSEEVKKGYRRMALRYHPDKNDHPQAEEHFREVVAAFEVLSDKEK  60
            MGKDYYKILGIERNASSE+VKKGYRRMALRYHPDKNDHPQAEE FREVVAAFEVL DKEK
Sbjct  1    MGKDYYKILGIERNASSEDVKKGYRRMALRYHPDKNDHPQAEEQFREVVAAFEVLFDKEK  60

Query  61   RETYDKYGEEGLRCDDEP-ATFAQPTSDMLPFMCAVGGTVLFAFAAYKTFQFFNRKKEAT  119
            RE YD++GEEGL+CDDEP ATFAQPT DMLPFMCAVGGTVLFAFAAYKTFQFFNRKKEAT
Sbjct  61   REIYDQHGEEGLKCDDEPAATFAQPTPDMLPFMCAVGGTVLFAFAAYKTFQFFNRKKEAT  120

Query  120  DGDGSSSD  127
             GDGSSSD
Sbjct  121  HGDGSSSD  128
```

Gene 2-Figure 5: Results from a pblast 2 of the peptide sequence from Gene checker for gene 2 against the peptide sequence for CG7130.

The protein sequence is pretty well conserved as shown by the 92% for identities, 95% for positives, and that the peptide sequence for gene 1 from the fosmid is 127 a.a. and the peptide sequence for CG7130 is 128 a.a. I can conclude from this that gene 2 is a real functional gene whose ortholog is CG7130 from *D. melanogaster*.

Gene 3 is homologous to RpLP0 (NP_524211.1) in *D. melanogaster*. The gene only has 1 isoform. There are 2 exons whose sequences are on the plus strand running from 4450-4503 and 4579-5475bp on the fosmid. My starting point for this gene was with Nscan's predicted gene 2. Results for Nscan's gene 2 are in the figure labeled "Gene 3-Figure 1" below.

Exon	Strand	Begin	End	Length
1	+	4341	4503	162
2	+	4578	5478	900

317 aa

MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL
AVVLMGKNTMMRKAIRGHLENNPQLEKLLPHIKGNVGFVFTKGDLAEVRD
KLLESKVRAPARPGAIAPLHVIIPAQNTGLGPEKTSFFQALSIPTKISKG
TIEIINDVPILKPGDKVGASEATLLNMLNISPFSYGLIVSQVYDSGSIFS
PEILDIKPEDLRAKFQQGVANLAAVCLSVGYPTIASAPHSIANGFKNLLA
IAATTEVEFKEATTIKEYIKDPSKFAAAASVSAAPAAGGAAEKKEEAKKV
ESESEEEDDDMGFGLFD

Transcript:

ATGGTTAGGGAGAACAAGGCAGCATGGAAGGCTCAGTACTTCATCAAGGT
TGTGGAACTGTTCGATGAGTTCCCCAAGTGCTTCATCGTGGGCGCCGACA
ACGTTGGCTCCAAGCAGATGCAGAACATCCGTACCAGCCTGCGTGGACTG
GCCGTCGTGCTTATGGGCAAGAACACCATGATGCGCAAGGCCATCCGCGG
TCATCTGGAGAACAACCCGCAGCTGGAGAAGCTGCTGCCCCACATCAAGG
GTAACGTGGGCTTCGTTTTCACCAAGGGCGATCTCGCCGAGGTGCGTGAC
AAGCTGTTGGAGTCCAAGGTGCGCGCCCCCGCCCGTCCCGGCGCTATTGC
CCCTCTGCACGTCATCATCCCGGCCCAGAACACCGGCTTGGGACCCGAGA
AGACCAGTTTCTTCCAGGCCCTGTCCATCCCGACCAAGATTTCCAAGGGA
ACAATTGAAATCATCAACGATGTGCCCATCCTGAAGCCCGGCGACAAGGT
CGGCGCCTCCGAGGCAACGCTGCTCAACATGTTGAACATCTCGCCCTTCT
CGTACGGTTTGATCGTCAGCCAGGTGTACGACTCCGGCTCGATCTTTTCG
CCTGAGATTCTGGACATTAAGCCCGAGGATCTGCGCGCCAAGTTCCAGCA
GGGAGTGGCCAACCTGGCCGCCGTTTGTTTGTCTGTGGGCTACCCCACCA
TTGCCTCGGCCCCGCACAGCATTGCCAACGGATTCAAGAACCTGCTGGCC
ATTGCTGCCACCACCGAGGTGGAGTTCAAGGAGGCGACCACCATCAAGGA
GTACATCAAGGACCCCAGCAAGTTCGCCGCCGCTGCCTCGGTTTCGGCTG
CCCCCGCCGCCGGCGGAGCTGCCGAGAAGAAGGAGGAGGCCAAGAAAGTC
GAGTCCGAGTCCGAGGAGGAGGACGATGATATGGGCTTCGGTCTGTTCGA
C

Gene 3-Figure 1: Results for Nscan's gene 2

I extracted bp 4000-6000 from my fosmid using extractseq and blasted it against *D. melanogaster* in Flybase. I only got 1 hit that had a reasonable e-score. The results are in the figure labeled "Gene 3-Figure 2" below.

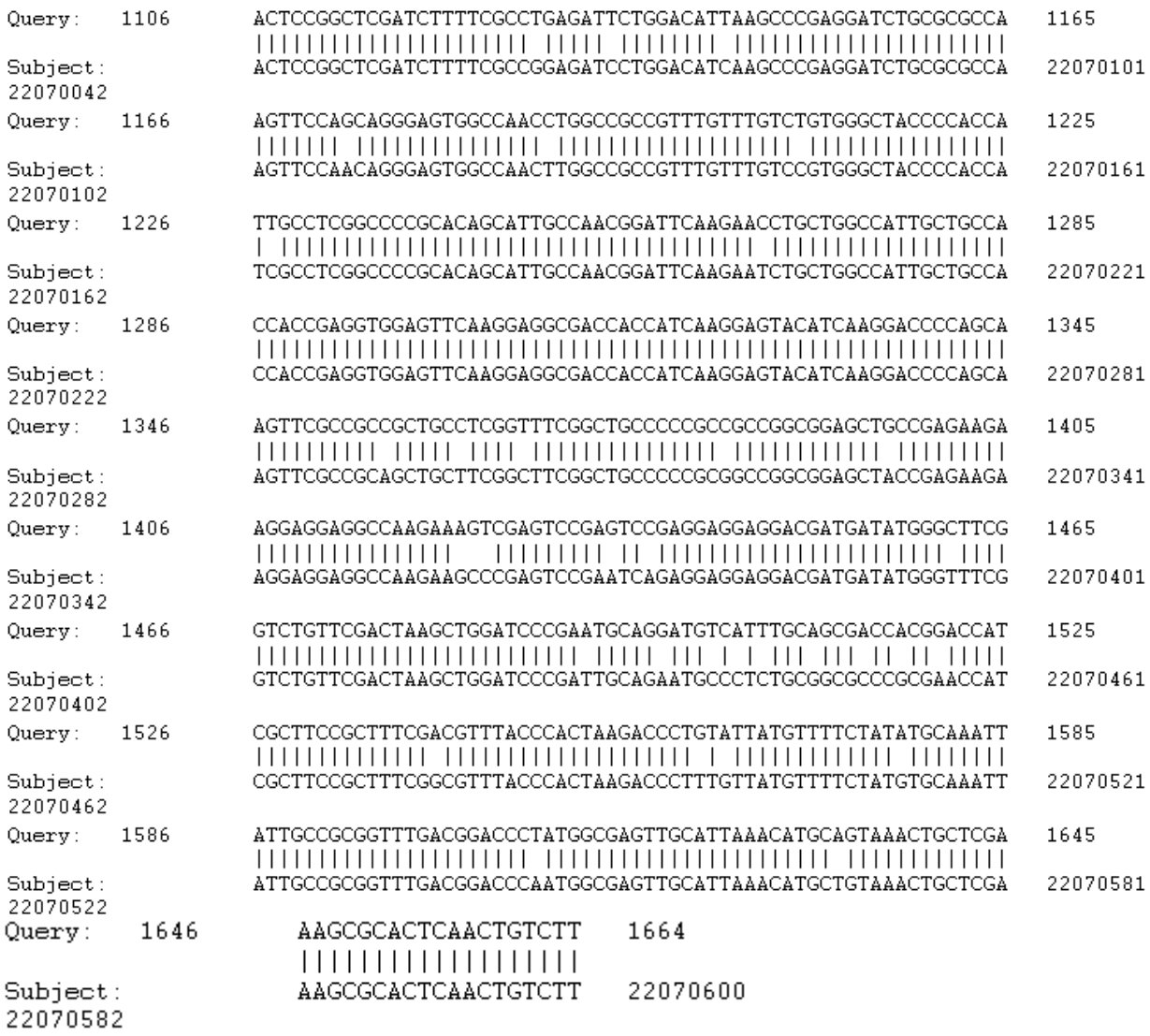
>gnl|dmel|3L type=chromosome_arm; loc=3L:1..24543557; ID=3L; dbxref=GB:AE014296; MD5=ec7148cae3daabbd2a226eaa6e85d7c2; length=24543557; release=r5.17; species=Dmel;
Length = 24543557

HSP # = 1 , **Score** = 2030.42 bits (1024) , **Expect** = 0
Identities = 1348 / 1459 (92.4%) , **Positives** = 1348 / 1459 (92.4%) , **Gaps** = 19 / 1459 (1.3%)
Strand = Plus / Plus

Genome View **Subject FASTA**

```
Query:   215        TCGCTATCGATGTGGTCACACTTGCTTCCGGCGCCAACTTCCCTCTTTCCGTTCTGTGAG   274
                    |||| ||||| | ||||||||| | | ||| |||||||||| ||||||||||||||||||
Subject:            TCGCCATCGAAGCGGTCACACTGGGTGCCGCCGCCAACTTCACTCTTTCCGTTCTGTGAG   22069211
22069152
Query:   275        CGAAAACCGAAAAGTCTGTGCTTTGGTAAGTATTATTAAAAGAGCGAAAAGATGTTGCAT   334
                    ||||||||||||||||||||||||||||||| ||  ||||||  || ||  |||||||||
Subject:            CGAAAACCGAAAAGTCTGTGCTTTGGTAAGTGTTGCTAAAAGTTCGGAATAATGTTGCAT   22069271
22069212
Query:   335        CCCGAGCTTTTTTGGGTGAATAACTGTTGCATGGCGCTGGCCCAGTACCGACTAATCGAG   394
                    ||||||| |||| ||||  ||||||||| || |||| ||| |||| |  |||||||||
Subject:            CCCGAGCATTTTCGGGTACATAACTGTTCCACGGCGGTGGTCCAGCAAAGACTAATCGTT   22069331
22069272
Query:   395        ATCACATCTTCCGCAGTTCTTAAATTCACCCGACGAGTCCCTAATACAAAATCAAAATGG   454
                    |||||  ||| ||||||||||||||||||||||||||||||||||| ||| |||||||
Subject:            ATCACGCCTTTCGCAGTTCTTAAATTCACCCGACGAGTCCCTAATACACAATTAAAATGG   22069391
22069332
Query:   455        TTAGGGAGAACAAGGCAGCATGGAAGGCTCAGTACTTCATCAAGGTTGTGGTAAGTATAG   514
                    |||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||
Subject:            TTAGGGAGAACAAGGCAGCGTGGAAGGCTCAGTACTTCATCAAGGTTGTGGTAAGTATAG   22069451
22069392
Query:   515        AACCG----------CCCTCACTAGCTCGCCCCTGGCTTATGCTCTTAACTAATCCTCGCT   565
                    ||||           | |||||||||| || |||||||||||| |||||| ||||
Subject:            AACCTTATAGAATTCGCTCACTAGCTGGCGCCTGGCTTATGCTGTTAACTGATCC-----   22069506
22069452
```

```
Query:   566        AATTCCTCCTCCAGGAACTGTTCGATGAGTTCCCCAAGTGCTTCATCGTGGGCGCCGACA   625
                         |||||||||||||||||||||||||||| ||||||||||||||||||||||||||
Subject:            -----CTCCTCCAGGAACTGTTCGATGAGTTCCCAAAGTGCTTCATCGTGGGCGCCGACA   22069561
22069507
Query:   626        ACGTTGGCTCCAAGCAGATGCAGAACATCCGTACCAGCCTGCGTGGACTGGCCGTCGTGC   685
                    |||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Subject:            ACGTGGGCTCCAAGCAGATGCAGAACATCCGTACCAGCCTGCGTGGACTGGCCGTCGTGC   22069621
22069562
Query:   686        TTATGGGCAAGAACACCATGATGCGCAAGGCCATCCGCGGTCATCTGGAGAACAACCCGC   745
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Subject:            TTATGGGCAAGAACACCATGATGCGCAAGGCCATCCGCGGTCATCTGGAGAACAACCCGC   22069681
22069622
Query:   746        AGCTGGAGAAGCTGCTGCCCCACATCAAGGGTAACGTGGGCTTCGTTTTCACCAAGGGCG   805
                    ||||||||||||||| |||||||||||||| |||||||| ||||| ||||||||||||||
Subject:            AGCTGGAGAAGCTGCTACCCCACATCAAGGGCAACGTGGGATTCGTGTTCACCAAGGGCG   22069741
22069682
Query:   806        ATCTCGCCGAGGTGCGTGACAAGCTGTTGGAGTCCAAGGTGCGCGCCCCCGCCCGTCCCG   865
                    ||||||||||||||| ||||||||| ||||||||||||||||||||||||||||||||||
Subject:            ATCTCGCCGAGGTGCGCGACAAGCTGCTGGAGTCCAAGGTGCGCGCCCCCGCCCGTCCCG   22069801
22069742
Query:   866        GCGCTATTGCCCCTCTGCACGTCATCATCCCGGCCCAGAACACCGGCTTGGGACCCGAGA   925
                    |||||||||||||||||||||||||||||||| ||||||||||||||||||||||||||
Subject:            GCGCTATTGCCCCTCTGCACGTCATCATCCCGGCGCAGAACACCGGCTTGGGACCCGAGA   22069861
22069802
Query:   926        AGACCAGTTTCTTCCAGGCCCTGTCCATCCCGACCAAGATTTCCAAGGGAACAATTGAAA   985
                    |||||||||||||||||||||||||||||||||||| ||||||||||||||||||||||||
Subject:            AGACCAGTTTCTTCCAGGCCCTGTCCATCCCGACCAAAATTTCCAAGGGAACAATTGAAA   22069921
22069862
Query:   986        TCATCAACGATGTGCCCATCCTGAAGCCCGGCGACAAGGTCGGCGCCTCCGAGGCAACGC   1045
                    |||||||||||||||||||||||||||| |||||||||||||||||||||||||||| || |
Subject:            TCATCAACGATGTGCCCATCCTGAAGCCTGGCGACAAGGTCGGCGCCTCCGAGGCGACAC   22069981
22069922
Query:   1046       TGCTCAACATGTTGAACATCTCGCCCTTCTCGTACGGTTTGATCGTCAGCCAGGTGTACG   1105
                    |||||||||||||||||||||||||||||||||||||||| |||| |||| |||||| ||||
Subject:            TGCTCAACATGTTGAACATCTCGCCCTTCTCGTACGGTCTGATTGTCAACCAGGTCTACG   22070041
22069982
```



Gene 3-Figure 2: Results from nblast of bp 4000-6000 from my fosmid and *D. melanogaster* genome.

The region shown in Gene 3-Figure 2 showed a homology between bp 215-1664 bp of my extracted sequence and a region in the gene RpLP0 of *D. melanogaster*. I then did a pblast of Nscan's predicted gene 2 peptide sequence against *D. melanogaster*. Once again I only got 1 hit that had a reasonable e-score. The results are shown below in the figure labeled "Gene 3-Figure 3."

name=RpLP0-PA; parent=FBgn0000100, FBtr0078481; dbxref=FlyBase:FBpp0078134, FlyBase_Annotation_IDs:CG7490-PA, GB_protein:AAF51807.1, REFSEQ:NP_524211, GB_protein:AAF51807; MD5=86e1796e988a2ee9e406941fb4905ecb; length=317; release=r5.17; species=Dmel;
Length = 317

HSP # = 1 , **Score =** 550.821 bits (1418) , **Expect =** 3.80774e-157
Identities = 270 / 271 (99.6%) , **Positives =** 271 / 271 (100%)

Subject FASTA

```
Query:   1           MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGLAVVLMGKNTM   60
                    MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGLAVVLMGKNTM
Subject: 1           MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGLAVVLMGKNTM   60

Query:   61          MRKAIRGHLENNPQLEKLLPHIKGNVGFVFTKGDLAEVRDKLLESKVRAPARPGAIAPLH   120
                    MRKAIRGHLENNPQLEKLLPHIKGNVGFVFTKGDLAEVRDKLLESKVRAPARPGAIAPLH
Subject: 61          MRKAIRGHLENNPQLEKLLPHIKGNVGFVFTKGDLAEVRDKLLESKVRAPARPGAIAPLH   120

Query:   121         VIIPAQNTGLGPEKTSFFQALSIPTKISKGTIEIINDVPILKPGDKVGASEATLLNMLNI   180
                    VIIPAQNTGLGPEKTSFFQALSIPTKISKGTIEIINDVPILKPGDKVGASEATLLNMLNI
Subject: 121         VIIPAQNTGLGPEKTSFFQALSIPTKISKGTIEIINDVPILKPGDKVGASEATLLNMLNI   180

Query:   181         SPFSYGLIVSQVYDSGSIFSPEILDIKPEDLRAKFQQGVANLAAVCLSVGYPTIASAPHS   240
                    SPFSYGLIV+QVYDSGSIFSPEILDIKPEDLRAKFQQGVANLAAVCLSVGYPTIASAPHS
Subject: 181         SPFSYGLIVNQVYDSGSIFSPEILDIKPEDLRAKFQQGVANLAAVCLSVGYPTIASAPHS   240

Query:   241         IANGFKNLLAIAATTEVEFKEATTIKEYIKD   271
                    IANGFKNLLAIAATTEVEFKEATTIKEYIKD
Subject: 241         IANGFKNLLAIAATTEVEFKEATTIKEYIKD   271
```

Gene 3-Figure 3: Results of a pblast of Genscan's predicted gene 2 peptide sequence against *D. melanogaster*.

The predicted peptide sequence from Nscan was homologous to RpLP0 from *D. melanogaster*. Since both the nblast and pblast for Nscan's predicted gene 2 matched up with RpLP0 I expected to find a homologous gene to RpLP0 in my fosmid. I decided to run a nblast 2 of the mRNA sequence from RpLP0 against my fosmid to see where in my fosmid the homologous sequence was. The results are below in the figure labeled "Gene 3-Figure 4."

```
 Score =  115 bits (62),  Expect = 1e-27
 Identities = 85/96 (88%), Gaps = 2/96 (2%)
 Strand=Plus/Plus

Query  1     GGT-ATCTTATTCGCCATCGAAGCGGTCACACTGGGTGCCGCCGCCAACTTCACTCTTTC  59
             ||| || ||| |||| ||||| | ||||||||| | | ||| |||||||||| |||||||
Sbjct  4204  GGTAATTTTA-TCGCTATCGATGTGGTCACACTTGCTTCCGGCGCCAACTTCCCTCTTTC  4262

Query  60    CGTTCTGTGAGCGAAAACCGAAAAGTCTGTGCTTTG  95
             ||||||||||||||||||||||||||||||||||||
Sbjct  4263  CGTTCTGTGAGCGAAAACCGAAAAGTCTGTGCTTTG  4298


 Score =  161 bits (87),  Expect = 1e-41
 Identities = 93/96 (96%), Gaps = 0/96 (0%)
 Strand=Plus/Plus

Query  95    GTTCTTAAATTCACCCGACGAGTCCCTAATACACAATTAAAATGGTTAGGGAGAACAAGG  154
             |||||||||||||||||||||||||||||||| ||| ||||||||||||||||||||||
Sbjct  4409  GTTCTTAAATTCACCCGACGAGTCCCTAATACAAAATCAAAATGGTTAGGGAGAACAAGG  4468

Query  155   CAGCGTGGAAGGCTCAGTACTTCATCAAGGTTGTGG  190
             |||| ||||||||||||||||||||||||||||||
Sbjct  4469  CAGCATGGAAGGCTCAGTACTTCATCAAGGTTGTGG  4504
```

```
 Score = 1700 bits (920),  Expect = 0.0
 Identities = 1023/1074 (95%), Gaps = 2/1074 (0%)
 Strand=Plus/Plus

Query  189   GGAACTGTTCGATGAGTTCCCAAAGTGCTTCATCGTGGGCGCCGACAACGTGGGCTCCAA  248
             ||||||||||||||||||| |||||||||||||||||||||||||||| ||||||||
Sbjct  4578  GGAACTGTTCGATGAGTTCCCCAAGTGCTTCATCGTGGGCGCCGACAACGTTGGCTCCAA  4637

Query  249   GCAGATGCAGAACATCCGTACCAGCCTGCGTGGACTGGCCGTCGTGCTTATGGGCAAGAA  308
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4638  GCAGATGCAGAACATCCGTACCAGCCTGCGTGGACTGGCCGTCGTGCTTATGGGCAAGAA  4697

Query  309   CACCATGATGCGCAAGGCCATCCGCGGTCATCTGGAGAACAACCCGCAGCTGGAGAAGCT  368
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4698  CACCATGATGCGCAAGGCCATCCGCGGTCATCTGGAGAACAACCCGCAGCTGGAGAAGCT  4757

Query  369   GCTACCCCACATCAAGGGCAACGTGGGATTCGTGTTCACCAAGGGCGATCTCGCCGAGGT  428
             ||| |||||||||||||| |||||||| ||||| ||||||||||||||||||||||||||
Sbjct  4758  GCTGCCCCACATCAAGGGTAACGTGGGCTTCGTTTTCACCAAGGGCGATCTCGCCGAGGT  4817

Query  429   GCGCGACAAGCTGCTGGAGTCCAAGGTGCGCGCCCCCGCCCGTCCCGGCGCTATTGCCCC  488
             ||| |||||||| ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4818  GCGTGACAAGCTGTTGGAGTCCAAGGTGCGCGCCCCCGCCCGTCCCGGCGCTATTGCCCC  4877

Query  489   TCTGCACGTCATCATCCCGGCGCAGAACACCGGCTTGGGACCCGAGAAGACCAGTTTCTT  548
             |||||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||
Sbjct  4878  TCTGCACGTCATCATCCCGGCCCAGAACACCGGCTTGGGACCCGAGAAGACCAGTTTCTT  4937

Query  549   CCAGGCCCTGTCCATCCCGACCAAAATTTCCAAGGGAACAATTGAAATCATCAACGATGT  608
             |||||||||||||||||||||| ||||||||||||||||||||||||||||||||||||||
Sbjct  4938  CCAGGCCCTGTCCATCCCGACCAAGATTTCCAAGGGAACAATTGAAATCATCAACGATGT  4997

Query  609   GCCCATCCTGAAGCCTGGCGACAAGGTCGGCGCCTCCGAGGCGACACTGCTCAACATGTT  668
             |||||||||||||| |||||||||||||||||||||||| || ||||||||||||||||||
Sbjct  4998  GCCCATCCTGAAGCCCGGCGACAAGGTCGGCGCCTCCGAGGCAACGCTGCTCAACATGTT  5057

Query  669   GAACATCTCGCCCTTCTCGTACGGTCTGATTGTCAACCAGGTCTACGACTCCGGCTCGAT  728
             |||||||||||||||||||||||| |||| |||| |||||| ||||||||||||||||||
Sbjct  5058  GAACATCTCGCCCTTCTCGTACGGTTTGATCGTCAGCCAGGTGTACGACTCCGGCTCGAT  5117
```

Gene 3-Figure 4: Results from a nblast 2 of the mRNA sequence from CG7133 against my fosmid. The 3 exons from RpLP0 have homologous regions in the fosmid.

I ran my entire masked fosmid through sixpack to get all the reading frames. This gene was a lot trickier to match up to my fosmid then the other 2 genes. Nscan predicted only 2 exons though there are 3 in RpLP0. I decided to use the data from Gene 3-Figure 4 to find the exon/intron borders. When I used the first exon from RpLP0 to find the first starting codon in my gene 3 and then used the second and third exons to find the other 2 exons I got a peptide sequence that was much shorter then the one for RpLP0. I then looked at where translation starts in RpLP0 by running its mRNA through sixpack and looking for the first start codon. Translation starts at the bp 135 in the middle of exon 2. The first exon is non-coding in *D. melanogaster*, but when I looked at the homologous sequence for exon 1 in my fosmid I got a start codon where there was not one in *D. melanogaster*. When I compared exon 1 of RpLP0 with its homologous region in my fosmid I found that there was a mutation at bp 4224 in my fosmid. There is an "A" at the homologous position in *D. melanogaster* and a "T" in the fosmid. This mutation causes a premature stop codon. I have highlighted the mutation below in figure Gene 3-Figure 5 for clarity.

```
 Score =  115 bits (62),  Expect = 1e-27
 Identities = 85/96 (88%), Gaps = 2/96 (2%)
 Strand=Plus/Plus

Query  1     GGT-ATCTTATTCGCCATCGAAGCGGTCACACTGGGTGCCGCCGCCAACTTCACTCTTTC  59
             ||| || ||| |||| ||||| | ||||||||| | | ||| |||||||||| |||||||
Sbjct  4204  GGTAATTTTA-TCGCTATCGATGTGGTCACACTTGCTTCCGGCGCCAACTTCCCTCTTTC  4262

Query  60    CGTTCTGTGAGCGAAAACCGAAAAGTCTGTGCTTTG  95
             |||||||||||||||||||||||||||||||||||||
Sbjct  4263  CGTTCTGTGAGCGAAAACCGAAAAGTCTGTGCTTTG  4298
```

Gene 3-Figure 5: Exon 1 from RpLP0 and its homologous region in the fosmid. The highlighted section shows the mutation from "A" in RpLP0 to "T" in the fosmid. This causes a premature start codon.

I thought that maybe this gene is a pseudogene then. I ran the mRNA of RpLP0 against the entire genome of *D. erecta* to see if the gene is present any where else in *D. erecta***,** but I did not have any hits. I then read about the function of the RpLP0 gene and found that it is involved in translation, DNA repair, translational elongation, and ribosome biogenesis. The gene serves an important function. I concluded that a functional, real ortholog of RpLP0 has to exist in *D. erecta* based on the idea that the gene has an important function, is not found anywhere else in the genome, and that the transcript and peptide sequence of the predicted gene 2 from Nscan matches so closely to RpLP0. I decided to just use the coding sequence for the RpLP0 protein and find where it was homologous in my fosmid. I found the start codon in my fosmid to start at bp 4450 and for the exon to end at bp 4503 by looking at the homologous region of the end of exon 2 in RpLP0 and looking for a "GT" where 95% of introns start. I found the exon borders for exon 2 of gene 3 on the fosmid to be at 4579-5475 bp. I had to look at the region in the fosmid homologous to the start and end of exon 3 in RpLP0. In order to find the start of the exon I looked for an "AG" where 95% of introns end. I looked for the first stop codon in the reading frame to find the end of the third exon. I then ran gene 3 with its coordinates for the 2 exon through Gene checker which verified the gene. I ran a pblast 2 of the peptide sequence from Gene checker for gene 3 against the peptide sequence for RpLP0. The results are shown below in the figure labeled "Gene 3-Figure 6".

name=RpLP0-PA; parent=FBgn0000100, FBtr0078481; dbxref=FlyBase:FBpp0078134, FlyBase_Annotation_IDs:CG7490-PA, GB_protein:AAF51807.1, REFSEQ:NP_524211, GB_protein:AAF51807; MD5=86e1796e988a2ee9e406941fb4905ecb; length=317; release=r5.17; species=Dmel;
Length = 317

HSP # = 1 , **Score =** 550.821 bits (1418) , **Expect =** 3.80774e-157
Identities = 270 / 271 (99.6%) , **Positives =** 271 / 271 (100%)

Subject FASTA

```
Query:   1     MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGLAVVLMGKNTM   60
               MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGLAVVLMGKNTM
Subject: 1     MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGLAVVLMGKNTM   60

Query:   61    MRKAIRGHLENNPQLEKLLPHIKGNVGFVFTKGDLAEVRDKLLESKVRAPARPGAIAPLH   120
               MRKAIRGHLENNPQLEKLLPHIKGNVGFVFTKGDLAEVRDKLLESKVRAPARPGAIAPLH
Subject: 61    MRKAIRGHLENNPQLEKLLPHIKGNVGFVFTKGDLAEVRDKLLESKVRAPARPGAIAPLH   120

Query:   121   VIIPAQNTGLGPEKTSFFQALSIPTKISKGTIEIINDVPILKPGDKVGASEATLLNMLNI   180
               VIIPAQNTGLGPEKTSFFQALSIPTKISKGTIEIINDVPILKPGDKVGASEATLLNMLNI
Subject: 121   VIIPAQNTGLGPEKTSFFQALSIPTKISKGTIEIINDVPILKPGDKVGASEATLLNMLNI   180

Query:   181   SPFSYGLIVSQVYDSGSIFSPEILDIKPEDLRAKFQQGVANLAAVCLSVGYPTIASAPHS   240
               SPFSYGLIV+QVYDSGSIFSPEILDIKPEDLRAKFQQGVANLAAVCLSVGYPTIASAPHS
Subject: 181   SPFSYGLIVNQVYDSGSIFSPEILDIKPEDLRAKFQQGVANLAAVCLSVGYPTIASAPHS   240

Query:   241   IANGFKNLLAIAATTEVEFKEATTIKEYIKD   271
               IANGFKNLLAIAATTEVEFKEATTIKEYIKD
Subject: 241   IANGFKNLLAIAATTEVEFKEATTIKEYIKD   271
```

Gene 3-Figure 6: Results from a pblast 2 of the peptide sequence from Gene checker for gene 1 against the peptide sequence for CG7133.

The protein sequence is incredibly well conserved as shown by the 99.6% for identities, 100% for positives, and that the peptide sequence for both RpLP0 and gene 3 in the fosmid are both 271 a.a in length. This data supports the claim even more that his gene is a real gene.

Pseudogene 1is homologous to Sfp79B in *D. melanogaster*. I based my conclusion that this gene is a pseudogene after analyzing it in several ways. First I noticed that neither Nscan nor Genscan provided a predicted gene that was an ortholog to Sfp79B which made me first think that it may be a pseudogene. I looked at the function of Sfp79B in *D. melanogaster*. The function is unknown though the protein stands for the seminal fluid

protein 79 so I could not base much of my conclusion on the importance of the gene. Next, I ran a nblast 2 of the mRNA of Sfp79B against my fosmid. I did not get a result until I checked the "More dissimilar sequences" option. The results are shown below in the figure labeled "Pseudogene 1-Figure 1."

```
>lcl|9429 Dere3_dna range=fosmid17:1-50000 5'pad=0 3'pad=0 strand=+ repeatMasking=none
Length=50000

 Score = 84.2 bits (92),  Expect = 3e-19
 Identities = 103/142 (72%), Gaps = 19/142 (13%)
 Strand=Plus/Plus

Query  1     AACTCTTCTCGTTCAGAATGAAGCTCCTTTCAGCCGCATTGGTCCTGCTCATGTCATCGG  60
             ||||||   | |||||||| ||||||| | ||| ||| ||||||||| ||||||| | ||
Sbjct  8679  AACTCTATGCTTTCAGAATAAAGCTCCGTCCAGTCGCTTTGGTCCTGATCATGTCCTTGG  8738

Query  61    CCTTGGCCATGGCCCAGAAGAATACGAACACGAATGAAAACAA-CATCGTTATTGGAAAA  119
             ||| |||| ||||||       ||           ||||| || | |||||||||||||
Sbjct  8739  CCTCGGCCTTGGCCC-------TA-----------GAAAAGAAGCGTCGTTATTGGAAAC  8780

Query  120   GTTTAAAATAAATACATACATA  141
             |||| |||| |||||||| |||
Sbjct  8781  GTTTGAAATCAATACATAAATA  8802
```

Pseudogene 1-Figure 1: Results from a nblast 2 of the mRNA of Sfp79B against the fosmid.

I used the results from Pseudogene 1-Figure 1to find the exon borders of the ortholog on the fosmid to be from 8685-8696bp. The first highlighted region in Pseudogene 1-Figure 1 shows a mutation in the fosmid that causes a premature start codon. The premature stop codon causes a frame shift which causes a premature stop codon as highlighted in the fosmid strand. The second highlighted part on the Sfp79B sequence is where the start codon is in the gene. This mutation that causes a premature start and stop codon means that the protein is only 4 a.a. in length. This is way too short to be a real gene. I also nblasted the mRNA for Sfp79B against the whole *D. erecta* genome to see if there is a functional copy of the gene in *D. erecta*, but I got no hits. This forced me to the conclusion that the ortholog for Sfp79B in the fosmid is a pseudogene and that the function of Sfp79B is not important in *D. erecta*.

Pseudogene 2 is homologous to msopa in *D. melanogaster*. Just like with Sfp79B, there were no Genscan or Nscan predicted genes for an ortholog to msopa in the fosmid so that made me first think that it may be a pseudogene. I looked at the function of msopa in *D. melanogaster* and the molecular function is unknown though biologically it serves as a defense response. Once again, I could not really decide whether the ortholog in the fosmid was a pseudogene based on the importance of msopa. I ran a nblast 2 of the mRNA from msopa against the fosmid, but I did not get a result until I checked the "More dissimilar sequences" option. This made me further believe that the ortholog was a pseudogene. The results for the nblast 2 are below in the figure labeled "Pseudogene 2-Figure 1."

```
>lcl|23297 Dere3_dna range=fosmid17:1-50000 5'pad=0 3'pad=0 strand=+ repeatMasking=none
Length=50000

 Score =  152 bits (168),  Expect = 2e-39
 Identities = 115/135 (85%), Gaps = 3/135 (2%)
 Strand=Plus/Plus

Query  16     CATACTCGCCATGAACTTCATACAGATCGCCGTGCTGTTCGTCCTGGTCGCAGTGGCCTT  75
              |||| |||||||||||||| |||||||||||| ||||| | ||||| || ||||||||||
Sbjct  11742  CATAATCGCCATGAACTTCCTACAGATCGCCTTGCTGGTGGTCCTAGTGGCAGTGGCCTT  11801

Query  76     GGCCAGACCACAGGAAGAT---CCGGCAAATCTGCCAGCTCCAGAGgcagcagcagcacc  132
              ||||||| ||||||| |||   ||| || ||||||||||||||| ||| ||  | ||||
Sbjct  11802  GGCCAGAGCACAGGATGATCCACCGACAGATCTGCCAGCTCCAGACGCAACAAAACCACC  11861

Query  133    accagcagcagcagc  147
              | ||||||||||||
Sbjct  11862  AGCAGCAGCAGCAGC  11876
```

Pseudogene 2-Figure 1: Results of a nblast 2 of the mRNA from msopa against the fosmid.

I ran my fosmid through sixpack and used the results of the nblast 2 to find where the start codon was for the ortholog of msopa in the fosmid. The start codon was at bp 11752 and the stop codon was at bp 11944. I ran a pblast 2 of the peptide sequence for msopa against the peptide sequence for the ortholog in the fosmid and only had the first 24 amino acids with any kind of similarity. Also, the length of the peptide sequence for msopa is 83 a.a. and the length of the peptide sequence in the ortholog is only 63 a.a. Both of these results just provided more evidence that the otholog is a pseudogene. I decided to run a clustal2w of the coding sequence for the translation of msopa against the coding sequence for the translation of the ortholog. The results are shown below in the figure labeled, "Pseudogene 2-Figure 2."

```
FBtr0078482                 ATGAACTTCATACAGATCGCCGTGCTGTTCGTCCTGGTCGCAGTGGCCTT
Dere3_dna                   ATGAACTTCCTACAGATCGCCTTGCTGGTGGTCCTAGTGGCAGTGGCCTT
                            ********* ********** ***** * ***** ** **********

FBtr0078482       GGCCAGACCACAGGAAGATCC---GGCAAATCTGCCAGCTCCAGAGGCAGCAGCAGCACC
Dere3_dna         GGCCAGAGCACAGGATGATCCACCGACAGATCTGCCAGCTCCAGACGCAACAAAACCACC
                  ******* ******* *****   * ** ***************** *** **  * ****

FBtr0078482       ACCAGCAGCAGCAGCAGCACCACCAGCAGCAGCAGCAGC-----ACCACCAGCACCACCA
Dere3_dna         AGCAGCAGCAGCAGCTGGTGCTCCAGCTGGTGTCCCGGGTAAAAATAACCAAAATGTCAA
                  * ************* *   * ***** *  *   * *      *  ****  *   * *

FBtr0078482       GCACCACCAGCTGCAGCACCT---CAA
Dere3_dna         TCACAAC--GTTGTGACCATTGGATAA
                   *** **  * **   *   *    **
```

Pseudogene 2-Figure 2: Results of clustal2w for the coding sequence for the translation of msopa against the coding sequence for the translation of the ortholog.

There seem to be a lot of mutations in the 2 sequences. The highlighted sequences show where a "C" in msopa is mutated to a "T" in the ortholog causing a nonsense mutation. This is the reason that the protein coded for by the ortholo is 20 a.a. shorter. I concluded that the ortholo for msopa in the fosmid is a pseudogene because the sequences for the transcripts are only somewhat similar for the first 135 or so base pairs, the peptide sequences are only somewhat similar for the first 24 a.a., the ortholog codes for a protein that is 20 a.a. shorter than the msopa protein, and the ortholog codes for a protein that is too short to be a real functioning protein.

A chart has also been added that shows all the exon borders for the 3 real genes.

Gene	BP of exon 1	BP of exon 2
Gene1	1274-288	
Gene2	3823-3443	
Gene3	4450-4503	4579-5475

Genes-Figure-1: exon borders for the 3 genes in the fosmid

Clustal Analysis:

I took the peptide sequence from gene 2 of my fosmid and compared it to orthologs found in *D. melanogaster*, *D. yakuba*, and *D. ananassae* using clustal2w. The figure below labeled, "Clustal-Figure-1," shows the results.

```
FBpp0131733      MGKDYYKILGIERNASSEEVKKGYRRMALRYHPDKNDHPQAEEHFREVVAAFEVLSDKEK 60
FBpp0078137      MGKDYYKILGIERNASSEDVKKGYRRMALRYHPDKNDHPQAEEQFREVVAAFEVLFDKEK 60
FBpp0268002      MGKDYYKILGIERNASSEDVKKGYRRMALRYHPDKNDHPQAEEQFREVVAAFEVLSDKEK 60
FBpp0126682      MGKDYYKILGIERNATNEEVKKGYRRMALRYHPDKNDHPQAEEQFKEVVAAFEVLSNKEK 60
                 ***************:.*:************************:*:********* :***

FBpp0131733      RETYDKYGEEGLRCDDEP--ATFAQPTSDMLPFMCAVGGTVLFAFAAYKTFQFFNRKKE- 117
FBpp0078137      REIYDQHGEEGLKCDDEPA-ATFAQPTPDMLPFMCAVGGTVLFAFAAYKTFQFFNRKKE- 118
FBpp0268002      REIYDQYGEDGLKCDDEP--ATFAQPTSDMIPFMCAVGGTVLFAFAAYKTFQFFNRKKK- 117
FBpp0126682      REIYDQFGEEGLRCEDGPDPATFAQPTSDMVPFMCAVGGTVLFAFAAYKTFQFFTRKKEP 120
                 ** **:.**:**:*:* *  *******.**:************************.***:

FBpp0131733      ATDGDGSSSD 127
FBpp0078137      ATHGDGSSSD 128
FBpp0268002      ASDNDGSSSD 127
FBpp0126682      ASNTDGSSSD 130
                 *:. ******
```

Clustal-Figure-1: Results from a clustal2w comparing the protein sequences from gene 2 of the fosmid(first sequence) and the ortholog protein sequences from *D. melanogaster*(second sequence), *D. yakuba*(third sequence), and *D. ananassae*(fourth sequence).

This protein is very well conserved in all 4 species. Except for a few mutated amino acids in each sequence they are all almost identical. Since the protein is so well conserved that means that the function must be very crucial for survival. This gene functions in unfolded protein binding and heat shock protein binding. I then got the first 1k of bp upstream from the start codon for gene 2 and 3 of its orthologs in *D. melanogaster*, *D. yakuba*, and *D. ananassae* and ran them through a clustal2w. The results are below in the figure labeled, "Clustal-Figure-2."

```
FBgn0105459      ------------------------------------------------TCACGCACCTCGGC 14
FBgn0240193      NNNNNNNNNNNNNGGCGGGAGCGCGCACCTTGGACTCCAACAGCTTGTCGCGCACCTCGGC 60
EMBOSS_001       ------------------------------------------------------------ 
FBgn0100484      ------------------------------------------TTTCGCTCACTGAACG 16


FBgn0105459      GAGATCGCCCTTGGTGAAAACGAAGCCCACGTTACCCTTGAT--GTGGGGCAGCAGCTTC 72
FBgn0240193      GAGATCGCCCTTGGTGAACACGAAGCCCACGTTACCCTTGAT--GTGGGGCAGTAGCTTC 118
EMBOSS_001       -----------TGGTGAACACGAATCCCACGTTGCCCTTGAT--GTGGGGTAGCAGCTTC 47
FBgn0100484      GAAAGAGGGTAAGTTGGCGCCGGAAATAGCGATGCGGAAAATATGTGAAATATCGATATT 76
                             * **    ** *     ** * *     **  ***    *      *

FBgn0105459      TCCAGCTGCGGGTTGTTCTCCAGATGACCGCGGATGGCCTTGCGCATCATGGTGTTCTTG 132
FBgn0240193      TCCAGCTGCGGGTTGTTCTCCAGATGACCGCGGATGGCCTTGCGCATCATGGTGTTCTTG 178
EMBOSS_001       TCCAGCTGCGGGTTGTTCTCCAGATGACCGCGGATGGCCTTGCGCATCATGGTGTTCTTG 107
FBgn0100484      TCCTCATGTCAAAAAAT-TCCGGAT-ATCGATTGTATTTTGGAAAATTGTTCCGATACTT 134
                 ***   **        * *** *** * **    *    * *   **  *   * *  *

FBgn0105459      CCCATAAGCACGACGGCCA-GTCC-ACGCAGGCTGGTACGGATGTTCTGCATCTGCTTGG 190
FBgn0240193      CCCATAAGCACGACGGCAA-GTCC-ACGCAGGCTGGTACGGATGTTCTGCATCTGCTTGG 236
EMBOSS_001       CCCATAAGCACGACGGCCA-GTCC-ACGCAGGCTGGTACGGATGTTCTGCATCTGCTTGG 165
FBgn0100484      CCAGGTTATAAAAAAATAATGTCTTGGGCAGACGAGCCTGAACATTAAATATGATTTCTG 194
                 **       *  *     * ***    **** *  *   * *  **    **    *  *

FBgn0105459      AGCCAACGTTGTCGGCGCCCACGATGAA-GCACTTGGGGAACTCATCGAACAGTTCCTGG 249
FBgn0240193      AGCCAACGTTGTCGGCGCCCACGATGAA-GCACTTTGGGAACTCATCGAACAGTTCCTGG 295
EMBOSS_001       AGCCCACGTTGTCGGCGCCCACGATGAA-GCACTTTGGGAACTCATCGAACAGTTCCTGG 224
FBgn0100484      AGAAAAAATGTCTGTCATCAAAATTGATTGTTATTTTGACGCTTATTGT-CATAAACTGG 253
                 **   *  *    * *  * *   ***  *   **  *   ** ** *  **    ****

FBgn0105459      AGGAGGAATTAGCGAGGATTAGTTAAGAGCATAAGCCAGG-GGCGAGCTAGTG-AGGGCG 307
FBgn0240193      AGGACGGATTAGCGAGGATTAGTTGGAAGCATATGCCAGG-GGCGAGGCAGGGGAGAGCG 354
EMBOSS_001       AGGAGGGATCAGTTAA---CAGCATAAGCCAGGCGCCAGCTAGTGAGCGAATTCTATAAG 281
FBgn0100484      CTC-GAAATTAGAAACCAATGATGCCAGACAGGCGTGAATCAGCGATGCTATAAATTTAG 312
                       ** **  *             **   *  *    * **            *

FBgn0105459      GTTCTATACTTACCACAACCTTGATGAAGTACTGAGCCTTCCATGCTGCCTTGTTCTCCC 367
FBgn0240193      GTTCTATACTTACCACAACCTTGATGAAGTACTGAGCCTTCCATGCTGCCTTGTTCTCCC 414
EMBOSS_001       GTTCTATACTTACCACAACCTTGATGAAGTACTGAGCCTTCCACGCTGCCTTGTTCTCCC 341
FBgn0100484      AGATGATACTTTTCGA-----TGAACGAGTATTTTTCTGCCAAATGGGCAATATTTTTCG 367
                      ******  *       ***   **** *   *   * *    **  * ** * *
```

```
FBgn0105459      TAACCATTTTGATTTT-GTATTAGGGACTCGTCGGGTGAATTTAAGAACTGCGGAAGATG 426
FBgn0240193      TAACCATTTTGATCGT-GTATTAGGGACTCGTCGGATGAATTTAAGAACTGCAAAAGACG 473
EMBOSS_001       TAACCATTTTAATTGT-GTATTAGGGACTCGTCGGGTGAATTTAAGAACTGCGAAAGGCG 400
FBgn0100484      ATACCAATAGTATAACCAGACTAAGAATGTAAAGAATGTTTAAATAAATTTTCTGTTGTA 427
                   **** *   **      * ** * *      *  **  *  *  ** *

FBgn0105459      TGATCTCGATTAGTCGGTACTGGGCCAGCGCCATGCAACAGTTATTCACCCAAAAAAGCT 486
FBgn0240193      TCATCTCAATTAGTCAGCACTTGGCCAACACCATTCAACAGTTATTGGACCGAAAATCCT 533
EMBOSS_001       TGATAACGATTAGTCTTTGCTGGACCACCGCCGTGGAACAGTTATGTACCCGAAAATGCT 460
FBgn0100484      ATTTATTCTTAATTACTTTTTGTACCAAGATTCCAATTTTACTTTCCATGCGATACTACT 487
                    *     * * *      *   ***              * *     * * *   **

FBgn0105459      CGGGATGCAACATCTTTTCGCTCTTTTAATAATACTTACCAAAGCACAGACT-TTTCGGT 545
FBgn0240193      CGGGATGCAACAATTTTTCGGGCTTTTAGCATTACTTACCAAAGCACAGACT-TTTCGGT 592
EMBOSS_001       CGGGATGCAACATTATTCCGAACTTTTAGCAACACTTACCAAAGCACAGACT-TTTCGGT 519
FBgn0100484      TTCACCCGAATTGAAGCGCTCTCCGTTCGGTTGAGTGATAGGTGCACAGATTATTTTGAA 547
                         **        *   *  **      * * *     ******* * *** *

FBgn0105459      TTTCGCTCACAGAACGGAAAGAGGGAAGTTGGCGCCGGAAGCAAGTGTGACCACATCGAT 605
FBgn0240193      TTTCGCTCACAGAACGGAAAGAGGGAAGTTGGCGGCGGAAGCCAGTGTGACCGCATCGAT 652
EMBOSS_001       TTTCGCTCACAGAACGGAAAGAGTGAAGTTGGCGGCGGCACCCAGTGTGACCGCTTCGAT 579
FBgn0100484      GTTCAAATCCTTTTTTGTACAAGGAACTTTAAGTACTTAATCGAGCA-GTCTGAATTTAT 606
                  ***     *      * *  **  *  **     *   * * **   * *    *  **

FBgn0105459      AGCGATAAAATTACCATATGGTCTAAAAAAAATACCAATTTCGAAAAGCGACATTAATAT 665
FBgn0240193      TACGATAAAATTACCGCATGGTCTAAAAA--ATACCAATTCCGATAAAAGAAATTATTAT 710
EMBOSS_001       GGCGAATAAGATACCGCACGGTCTGAAAAA-ATACCAAAACCGGTAAGCGATATGAATAT 638
FBgn0100484      TTCTTTAGAAATTC---AAAGTTTGACACAACTACTATTCGTTTTATATTGCATTAAAAC 663
                   *     *  * *   *  ** * * *    *** *       *      ** *  *

FBgn0105459      TTT--ATCTTTGAAATATATAATTCCTTAAAAAATACATGAAAAAAATAATTTAACAAAT 723
FBgn0240193      TTT--TTCTTTGAAATATATAATTCCTTGTGAAATACTTGAAAAAAAAAAC--------- 759
EMBOSS_001       TTTCTATTTTCGAAATGTGTAATTCCTTTTAAAGTACTCAAAAAAGATAGATTTA----- 693
FBgn0100484      AAA------CAAAACTAAATATTTTTCTACAAATTTAAATATAGAAACAAATAAAC---- 713
                            ** *   ** **   *   ** *     * * * * *

FBgn0105459      TATAAATTATTAATAAATGCCAAATGGGAGCTTAACAGTGCAGCTACACAGCTGCTGCGC 783
FBgn0240193      -ATGAATTATTAATTAATG-------GAAGCCTAA-AGTGCC-TTACA------------ 797
EMBOSS_001       -ACAGATCATTAATTACTA-------ATAGGCTAATAGTGCTCCTATACAGCTGCTACGA 745
FBgn0100484      -GTAAATCTGCATTGAACATAA----ATATATTTATTTTATTTATTCATAATCTGTGTGA 768
                      **    * * *           *   * *   *     *  *
```

```
FBgn0105459      TAGACAGAATTTCTTCTGTTAAC--GTGTTAACAGTATTCTGGAGGTAGGGTCACACTTG 841
FBgn0240193      --------ATTACTTCTGTTAATA-GTGTTAACAGTATTTTGAAGGTAGGGTTACTCTTG 848
EMBOSS_001       TCGTCAGAATCTCTTCTGTTAACTAACAGTGCTTCTGGTGGTAAGGTAGGGTCACTCTGG 805
FBgn0100484      TATTT---TTGGGTTTTAATAAATTATTTAAAAAAAAGCAATTGTTTCAAATGGTTCCAT 825
                         *   ** *  ***                       *    *    *

FBgn0105459      T-ACTGGCGGGCAATAATTTATCGATTGGGATGTTAGTTCT-GTTTATTTGCTAAATTTC 899
FBgn0240193      TTACTGGCGGGCAATAATTTATCGATTATC-TGTTAGTTTCAGTTTATTTGTTATATTGT 907
EMBOSS_001       CCACTGGCGGCCGATAATTTATCGATTGAC-----------TGTTTATTTGTAATAATGT 854
FBgn0100484      CCAATGGCGCTGTCGTCTTTACTGGTCCTTATCGATGCATCGATAACTTCATAACTTTCT 885
                   * *****        ****  * *                 *   **    *   *

FBgn0105459      TTC--ATTGGCGAATCAAGTTACCAAAATATAC------------ACAAGCACCAACAAG 945
FBgn0240193      TTTCAATCAGCGAATTGAACTACCGAAATAT----------------------CACCAAG 945
EMBOSS_001       TTTAAATCAGCGAATCGAATCAGCAATATATACTCGCATTACCACATAAGCACCAACAAG 914
FBgn0100484      CTGGAAATCAAAACACAAGTTGAAATAAGATTCT-------------GCAATAAATAAA 931
                  *   *      *    *         * **                      *  **

FBgn0105459      TTAATGCGCTGCCTCACATACTATAG--------------------------------CA 973
FBgn0240193      TTGAGGCGCTGCCTCACATAGCATAC--------------------------------CA 973
EMBOSS_001       TCGAGACGCTGCCTCACATTCCATACACGGCGCAACACCGCTAGACGACCATCTCAGCCA 974
FBgn0100484      TTATTGTACCTGAGCACCCGATTAAGAGAAAAGGAA----------------------CA 969
                 *       *     ***       *                                **

FBgn0105459      GAAC---AAGGAG-AATCCCATTTCACCACA 1000
FBgn0240193      GAAC---AAGGAG-AATCCCATTTAACCACA 1000
EMBOSS_001       GAAC---AAGGAG-AATACCATTTCACCACA 1001
FBgn0100484      AAGCGTTAGGGAGTAACAGTCGTTTATCACC 1000
                  * *   * **** **      ** * ***
```

Clustal-Figure-2: Results from a clustal2w of the first 1k of bp upstream from the start codon for gene 2(top sequence) and 3 of its orthologs in *D. melanogaster*(third sequence), *D. yakuba*(second sequence), and *D. ananassae*(fourth sequence).

I found the TATA box which is highlighted in the figure. Though there are deletions in the other sequences, if you look 1 bp upstream, there is a "T" which still makes the TATA box functional. I could not find the initiators or any DEP's.

Repeats:

The fosmid was checked for repeats using RepeatMasker in order to eliminate any repetitious elements before the fosmid was checked for gene features. RepeatMasker generated 2 tables. The first table below labeled, "Repeats-Figure-1," shows the total amount of each kind of repeat in the fosmid. The second table below labeled, "Repeats-Figure-2," shows the base pair location of every repeat found in the fosmid.

	number of elements*	length occupied	percentage of sequence
Retroelements	0	0 bp	0.00 %
SINEs:	0	0 bp	0.00 %
Penelope	0	0 bp	0.00 %
LINEs:	0	0 bp	0.00 %
CRE/SLACS	0	0 bp	0.00 %
L2/CR1/Rex	0	0 bp	0.00 %
R1/LOA/Jockey	0	0 bp	0.00 %
R2/R4/NeSL	0	0 bp	0.00 %
RTE/Bov-B	0	0 bp	0.00 %
L1/CIN4	0	0 bp	0.00 %
LTR elements:	0	0 bp	0.00 %
BEL/Pao	0	0 bp	0.00 %
Ty1/Copia	0	0 bp	0.00 %
Gypsy/DIRS1	0	0 bp	0.00 %
Retroviral	0	0 bp	0.00 %
DNA transposons	10	1402 bp	2.80 %
hobo-Activator	0	0 bp	0.00 %
Tc1-IS630-Pogo	0	0 bp	0.00 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	0	0 bp	0.00 %
Tourist/Harbinger	0	0 bp	0.00 %
Other (Mirage, P-element, Transib)	3	692 bp	1.38 %
Rolling-circles	0	0 bp	0.00 %
Unclassified:	0	0 bp	0.00 %
Total interspersed repeats:		1402 bp	2.80 %
Small RNA:	0	0 bp	0.00 %
Satellites:	0	0 bp	0.00 %
Simple repeats:	4	212 bp	0.42 %
Low complexity:	5	302 bp	0.60 %
bases masked:		1916 bp	(3.83 %)

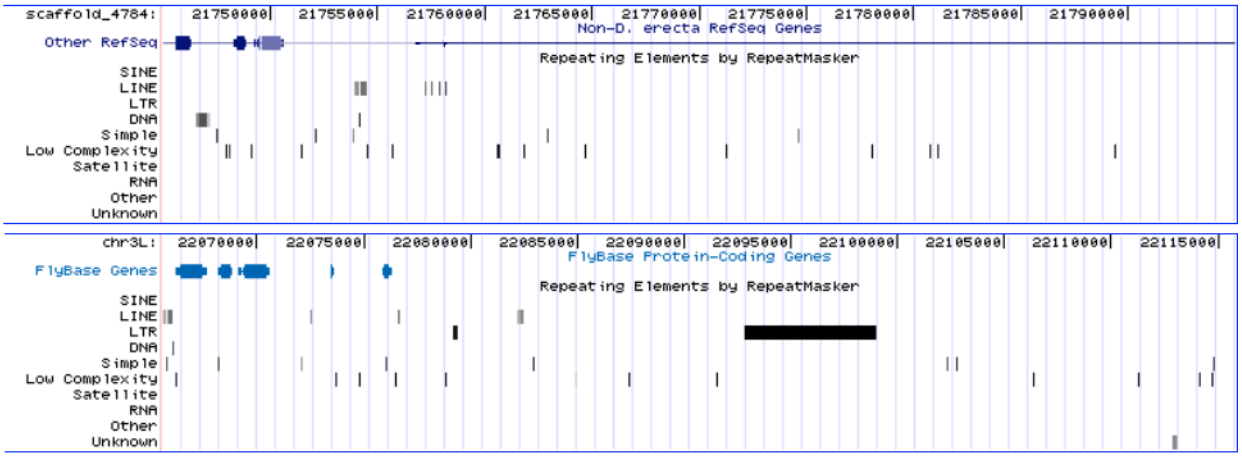
Repeats-Figure-1: total amount of each kind of repeat in the fosmid

+	score	% div.	% del.	% ins.	query sequence	-position in query- begin	end	(left)	C	matching + repeat	repeat class/family	-position in repeat-- (left) begin	end	begin (left)	linkage id/graphic
+	293	19.9	0.0	7.2	Dere3_dna	1572	1690	(48310)	C	PROTOP_A	DNA/P	(296)	809	699	1 *
+	769	11.1	4.4	0.0	Dere3_dna	1662	1796	(48204)	C	PROTOP	DNA/P	(452)	4028	3888	2 *
+	1665	11.8	1.9	0.4	Dere3_dna	1789	2059	(47941)	C	PROTOP	DNA/P	(794)	3686	3412	2
+	226	19.1	9.7	10.3	Dere3_dna	2055	2229	(47771)	C	PROTOP_B	DNA/P	(282)	871	698	3 *
+	495	7.3	0.0	0.0	Dere3_dna	2487	2554	(47446)	+	(TATG)n	Simple_repeat	3	70	(0)	4
+	22	75.4	0.0	0.0	Dere3_dna	3053	3109	(46891)	+	AT_rich	Low_complexity	1	57	(0)	5
+	35	82.1	0.0	0.0	Dere3_dna	4083	4166	(45834)	+	AT_rich	Low_complexity	1	84	(0)	6
+	26	57.7	0.0	0.0	Dere3_dna	6427	6452	(43548)	+	AT_rich	Low_complexity	1	26	(0)	7
+	201	7.4	0.0	0.0	Dere3_dna	7107	7133	(42867)	+	(CATCG)n	Simple_repeat	1	27	(0)	8
+	221	19.5	2.5	3.8	Dere3_dna	8851	8930	(41070)	+	(CTAA)n	Simple_repeat	2	80	(0)	9
+	405	11.1	0.0	1.6	Dere3_dna	8952	9015	(40985)	+	Helitron-1_DYak	DNA/Helitron	49	111	(10735)	10
+	503	14.3	12.4	0.8	Dere3_dna	9018	9130	(40870)	+	DNAREP1_DM	DNA/Helitron	469	594	(0)	11
+	276	5.9	0.0	0.0	Dere3_dna	9162	9195	(40805)	C	PROTOP_A	DNA/P	(564)	541	508	1 *
+	1439	14.3	2.2	5.7	Dere3_dna	9194	9504	(40496)	+	DNAREP1_DM	DNA/Helitron	293	593	(1)	12
+	254	21.4	0.0	0.0	Dere3_dna	12181	12236	(37764)	+	DNAREP1_DM	DNA/Helitron	539	594	(0)	13
+	262	19.6	0.0	0.0	Dere3_dna	12522	12577	(37423)	+	DNAREP1_DM	DNA/Helitron	539	594	(0)	14
+	312	16.1	0.0	0.0	Dere3_dna	12849	12904	(37096)	+	DNAREP1_DM	DNA/Helitron	539	594	(0)	15
+	312	16.1	0.0	0.0	Dere3_dna	13176	13231	(36769)	+	DNAREP1_DM	DNA/Helitron	539	594	(0)	16
+	62	85.9	0.0	2.8	Dere3_dna	15606	15714	(34286)	+	AT_rich	Low_complexity	1	106	(0)	17
+	225	10.8	0.0	0.0	Dere3_dna	17882	17918	(32082)	+	(CA)n	Simple_repeat	2	38	(0)	18
+	26	53.9	0.0	0.0	Dere3 dna	44347	44372	(5628)	+	AT rich	Low complexity	1	26	(0)	19

Repeats-Figure-2: shows the base pair location of every repeat found in the fosmid

Synteny:

This section discusses the synteny between the fosmid and the homologous region in *D. melanogaster*. I looked to see if all the genes in the fosmid were from the same region of the *D. melanogaster* genome. I also looked for a similarity of repetitious elements. UCSC Genome Browser was used to get the area of *D. erecta* that the fosmid came from and to get the homologous area of *D. melanogaster* in order to compare the location of the genes and repetitious elements. Below in the figure labeled "Synteny-Figure-1" are the 2 areas from *D. erecta* and *D. melanogaster* that were compared.



Synteny-Figure-1: 2 homologous areas from *D. erecta*(top) and *D. melanogaster*(bottom) from UCSC genome browser

The genes are generally from about the same region of the genome. *D. melanogaster* of course has the genes, Sfp79B and msopa, which are not present in *D. erecta* because they are pseudogenes. The simple and low complexity repeats seem to be the most conserved between the two genomes. *D. erecta* has a DNA repeat element

right after gene 1. This repeat seems to have caused the other 2 genes that come after gene 1 to shift to the right making those 2 genes farther away from gene 1than the 2 homologous genes in *D. melanogaster* are from the first homologous gene in *D. melanogaster*. *D. melanogaster* also has the long LTR(long terminal repeat elements) that could be a transposon or retroposon.

Tools Used:

RepeatMasker- http://www.repeatmasker.org/

Extractseq- http://gander.wustl.edu/cgi-bin/emboss

Sixpack- http://gander.wustl.edu/cgi-bin/emboss

Flybase- http://flybase.org/

UCSC Genome Browser- http://genome.ucsc.edu/

Clustal2w- http://www.ebi.ac.uk/Tools/clustalw2/index.html

NCBI blast- http://blast.ncbi.nlm.nih.gov/Blast.cgi

References