Stephen McDaniel
May 10, 2007
Elgin, Bio 434W
Annotation Paper, Final

A Report on the Annotation of Fosmid 9

Abstract:

At the beginning of the semester, each student in Biology 4342/434W was given a fosmid, a 40 kilobase section, of the *Drosophila mojavensis* fourth chromosome to finish. Finishing is the process of ensuring that the entire fosmid has high quality sequence without any gaps, discrepancies, or misassemblies. Yet, the finished sequence is simply a long list of base pairs, which has little relevant information.

For the final part of the semester, the Biology 4342/434W students have revisited their fosmids, and are now annotating them. Students are learning how to use a wide variety of tools to parse the informative regions of the genome from the uninformative regions. The most informative regions, of course, are the genes. They carry the information that will be transcribed into mRNA and then translated into proteins. The BLAST suite of programs has been immensely helpful in identifying genes. CLUSTALW, a piece of alignment software, was also used to look for conserved regions in the proteins and untranslated regions upstream of the genes. Finally, the *D. mojavensis* fosmids were examined for synteny with the *D. melanogaster* genome. A synteny analysis looks at what genes are next to each other on the chromosomes and in what orientation they appear. It has been millions of years since *D. mojavensis* and *D. melanogaster* have diverged from each other and it will prove interesting to see how evolution has driven change into their genomes.

Fosmid Overview:

GENSCAN, a predictive program that is used to look for large open reading frames in a particular sequence, was run on the sequence of fosmid 9 (Figure 1).

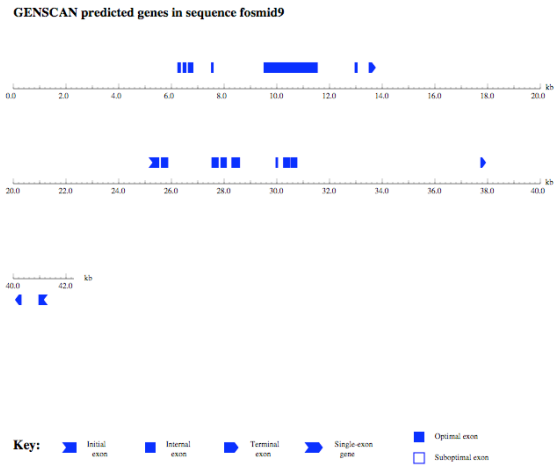


Figure 1: GENSCAN output for fosmid 9.

GENSCAN predicted three features in fosmid 9, though only two of them would actually prove to be genes. GENSCAN also missed two partial genes in the fosmid sequence that were found based on RefSeq data. RefSeq data comes from mRNA evidence, which in turn means that this region of the genome has been found to be transcriptionally active (Figure 2).

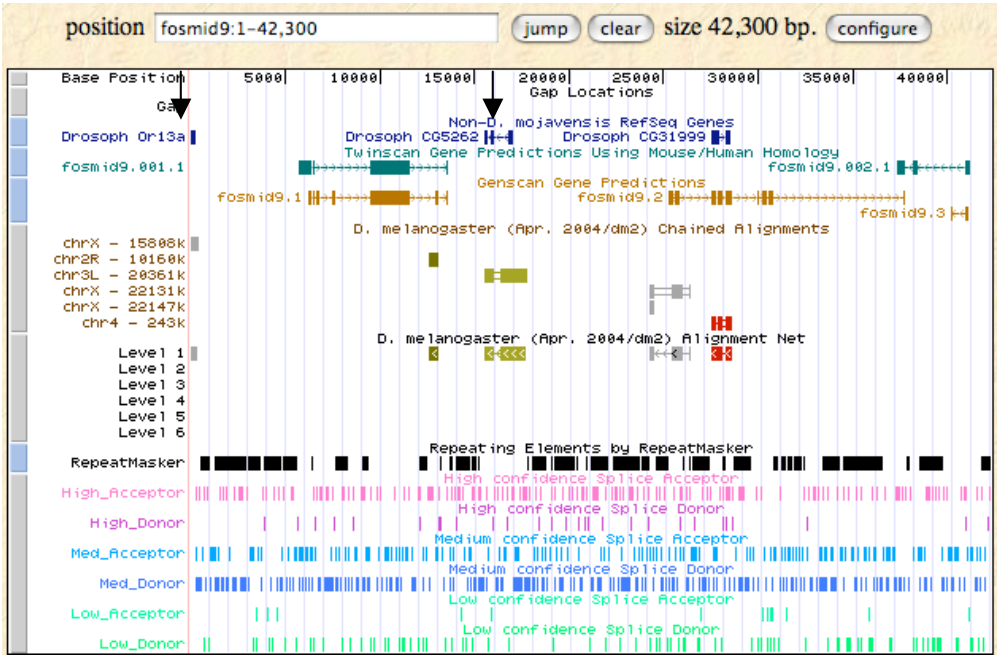


Figure 2: An overview of the fosmid from the goose.wustl.edu website. The Or13a and CG5262 genes were identified due to the RefSeq lines noted above.

Using both the GENSCAN outputs and the Genome Browser on the Goose website, gene models were created for two of the GENSCAN predicted genes, and both of the new genes predicted from the RefSeq data, giving four possible genes in all. Finally RepeatMasker was used to identify repetitive elements in the fosmid. Using these tools a map of the fosmid was produced (Figure 3) and will be explored in depth throughout the rest of the paper.

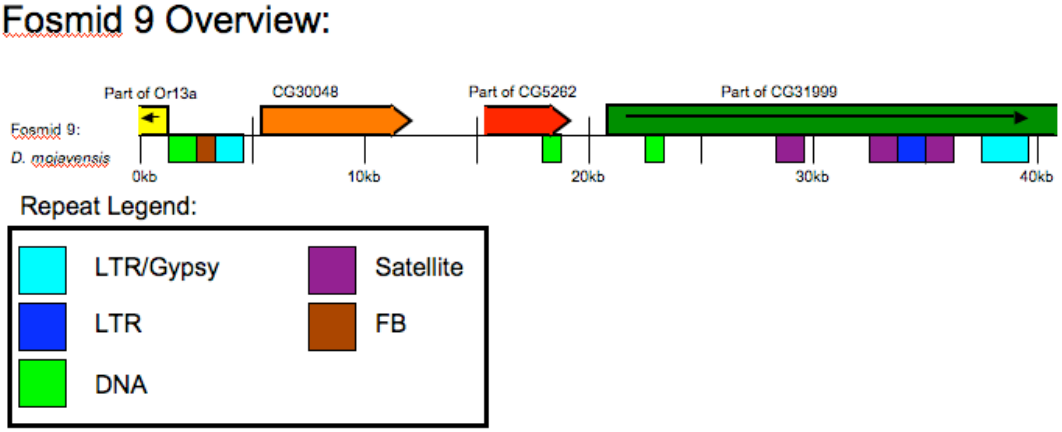


Figure 3: The final annotation of fosmid 9 including the repeats over 500bp.

Feature Analysis:

GENSCAN Feature 1:

This feature was the most difficult because there were three initial blastp results, using the non-redundant protein database, which had very similar E values differing, at most, by one order of magnitude (Figure 4). The first two matches are actually isoforms of the same gene. Their coding sequences are basically the same, but each form has a slightly different splicing pattern for the final coding sequence. A gene model was formed for each of these prospective matches, the two isoforms of CG30048 and CG12636, and a decision was made based on the models formed. What follows is the evidence gathered for each of these possible genes, and the reasoning for the final decision.

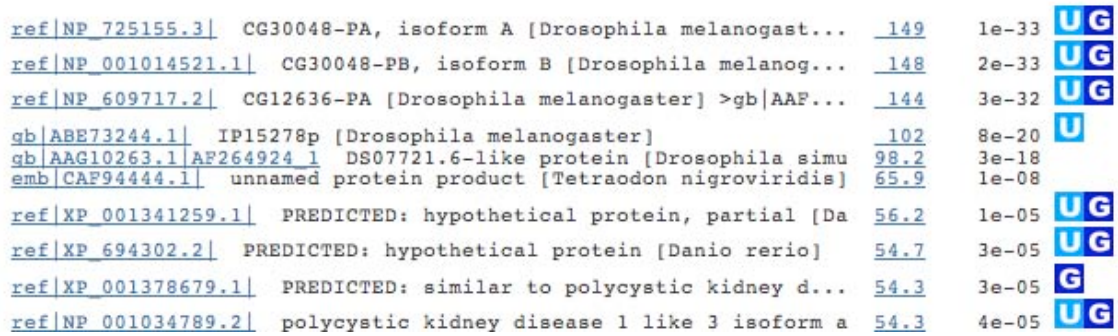


Figure 4: GENSCAN feature 1 blastp results.

CG30048-PA (NP_725155.3):

Once the blastp results came back, the first thing to be done was to find the amino acid sequence of CG30048-PA on the Ensembl website (Figure 5).

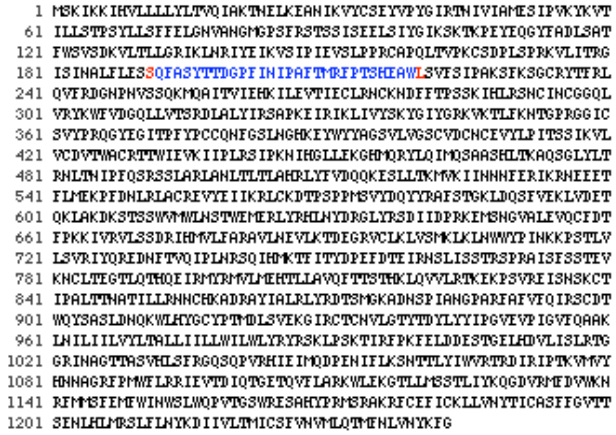
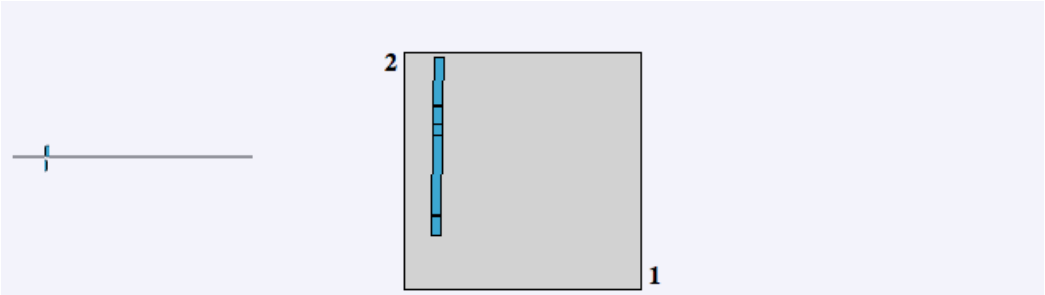


Figure 5: The *D. melanogaster* amino acid sequence for CG30048-PA.

Each exon is illustrated by alternating black and blue letters. The red letters signify amino acids that have coding sequence in neighboring exons. For example, the red D

found between exons one and two has only one or two of the bases, out of three, that code for it in exon one, while the other one or two coding bases are found in exon two.

To begin the process of annotating this gene in fosmid 9, a blast2 alignment was used to try to locate the first exon within the fosmid sequence (Figures 6).



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

```
 Score = 30.8 bits (68),  Expect = 13962
 Identities = 26/146 (17%), Positives = 70/146 (47%), Gaps = 7/146 (4%)
 Frame = +2

Query   5864   VIVTLQSLRGVWFLT-KVSIEERIIDYFQTNASEHREQLLWYHTASLKQTLNFYGIFTQK  6040
               +++ ++S+    + +T  +S    ++ +F+     +     +  T+S+ + L+ YGI ++
Sbjct   46     IVIAMESIPVKYKVTILLSTPSYLLSFFELGNVANGMGPSFRSTSSISEELSIYGIKSKT  105

Query   6041   NLPYQECFYHVSRANFFKSGFHNVSVLITTIG-ITSDKLRSSSYTY-IDLDENVRCEPQF  6214
                  Y++ ++    A F+          ++T +G I  +++     +  I++    RC PQ
Sbjct   106    KPEYEQGYFADLSATFWSVS----DKVLTLLGRIKLNRIYEIKVSIPIEVSLPPRCAPQL  161

Query   6215   SIPQCVNADYPVHYEVSDIIVLRAMY  6292
               ++P+C +    P    ++  I + A++
Sbjct   162    TVPKCSDPLSPRKVLITRGISINALF  187
```

Figure 6: The blast2 alignment for the first exon of CG30048-PA.

The blast2 alignment results gave both a frame, and an area of reference to examine in the fosmid. Using the Goose *Genome Browser* function, a possible start site was found for this protein by looking for the closest methionine to the start of the above alignment (Figure 7).

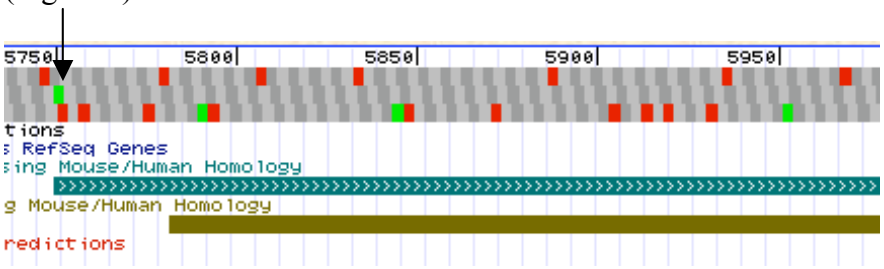


Figure 7: The nearest start site to the predicted one in the blast2 alignment.

This site was both in the correct frame and relatively near to the start site given in the alignment (bp 5864), but because the alignment was so poor, with only a 17% identity with the *D. melanogaster* protein sequence, the difference in start sites was not surprising. However, the blast2 alignment did begin the alignment with the 46th amino

acid, and the predicted start site, while a little earlier in the sequence, does not incorporate all of the missing amino acids. Again, because the amino acid sequence identity between the two species for this exon is so low, this discrepancy is sequence size is not as worrisome as it would be if the sequence identity were much higher. Now that the start site of the protein was found, the end of the exon needed to be found. The alignment suggested that the end of this exon was near bp 6292. This area was examined (Figure 8), and a splice site was found at bp 6299. There are several other sites around the predicted site, but bp 6299 is the closest site that includes the entire alignment.

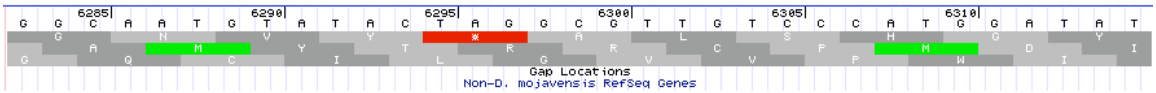


Figure 8: The area surrounding the end of the first exon of CG30048-PA.

Next a blastp search was undertaken for the second exon, which was significantly smaller than the first exon (Figure 9). Again the percent identity was poor, but the

```
 Score = 17.3 bits (33),  Expect = 159748000
 Identities = 6/12 (50%), Positives = 7/12 (58%), Gaps = 0/12 (0%)
 Frame = +3

Query  6894  GPAFTIPPFTQQ  6929
              GP   IP FT +
Sbjct  10    GPFINIPAFTMR  21
```

Figure 9: The blast2 alignment for the second exon of CG30048-PA.

alignment put this exon in the correct region and on the correct strand of DNA. A high quality acceptor site was found at bp 6985, which was close to the beginning of the alignment. This site also coincides with a gene prediction from SGP, a program similar to GENSCAN. The end of this exon, however, was quite a bit more ambiguous. I did not have a great deal of confidence in any of the sites surrounding bp 6929, which was the predicted end site according to blast2. This problem was resolved when the last exon was annotated.

Again, blast2 was used to align exon three with the fosmid sequence (Figures 10 and 11). From these results, it is apparent that a new intron is present in *D. mojavensis* because the two matches from this search are about two kb apart and in separate phases.

```
 Score = 71.2 bits (173),  Expect = 9e-09
 Identities = 56/191 (29%), Positives = 90/191 (47%), Gaps = 23/191 (12%)
 Frame = +3

Query  6987  SSVAMQTILFSSYPRLYIDVIIKCVTNCKGNRFSNYSHVHLKAHCLSCKSMKITKWMWTV  7166
              SS  MQ I    +  L  +V I+C+ NCK + F+  S +HL+++C++C   ++ ++ W V
Sbjct  31    SSQKMQAITVIEHKIL--EVTIECLRNCKNDFFTPSSKIHLRSNCINCGG-QLVRYKWFV  87

Query  7167  DHNLVSSSKRLLYYVSN-----KTSIIVYLNMEAIHRYKPSSKY--QGSSTLRLERNYGP  7325
              D  L+ +S+ L  Y+ +     +  +IVY            SKY   G     L +N GP
Sbjct  88    DGQLLVTSRDLALYIRSAPKEIRIKLIVY------------SKYGIYGRKVKTLFKNTGP  135

Query  7326  VHTACYVSPLDGEAIKTLFSIKCTNENALYKPLKYCLSVNSLLITECITDQLIISRLPAA  7505
                  C V P  G    T F   C N  +L    +Y     S+L+  C+     +  LP
Sbjct  136   RGGICSVYPRQGYEGITPFYPCCQNFGSLNGHKEYWYYAGSVLVGSCVDCNCEV-YLPIT  194

Query  7506  SHIEVQVCNIS  7538
              S I+V VC+++
Sbjct  195   SSIKVLVCDVT  205
```

Figure 10: The blast2 alignment for the first part of the third exon in CG30048-PA.

```
 Score =  192 bits (489),  Expect = 2e-45
 Identities = 183/772 (23%), Positives = 345/772 (44%), Gaps = 58/772 (7%)
 Frame = +1

Query  9694   LVFTEALAVYQPQTPVQLAHLVQILIKILDHIIPLDDMIVNVLARILHIIESTFKIVIAN  9873
              L  T  L     Q+   LA L  + + +   +  +D    ++L +++ II + F+ +  N
Sbjct  257    LYLTRNLTNIPFQSRSSLARLANLTLTLAHRLYFVDQQKESLLTKMVKIINNNFERIKRN  316

Query  9874   AEQNTLMDCYYKKMVEA---MFEILDKFATEWEYIPKS----QCRAESESCLNVDNFRNR  10032
               E+  LM+  +  +  A   ++EI+ +   +    P S      RA S   L+
Sbjct  317    EEETFLMEKPFDNLRLACREVYEIIKRLCKDTPSPPMSVYDQYYRAFSTGKLDQSFVEKL  376

Query  10033  LEQMSTL-NPQVLEHINNWLHAHWKLNNCLFYMGMGMARRIHPDE--NAKKKDCRTFLMT  10203
              +++   L   +       WL++ W++     Y  +   R ++  +  + +K+      +
Sbjct  377    VDETQKLAKDKSTSSWVMWLNSTWEMER--LYRHLNYDRGLYRSDIIDPRKEMSNGVALE  434

Query  10204  MESFDLDLERALVIKSADAMHTLIFTEKLLQELRHLLRNDE--VLISIRSHKHSQYWWYP  10377
              ++ FD   ++ + + S+D +H ++F   +L E+   L+ DE  V + + S K    WWYP
Sbjct  435    VQCFDTFPKKIVRVLSSDRIHMVLFARAVLNEV---LKTDEGRVCLKLVSMKLKLNWWYP  491

Query  10378  -EQDSRTQVLVVNAY------TSTAILEKTQQLSEPF-QYISKLTTKSCTYQYCKGRRRR  10533
               +   T VL V  Y      T    L ++Q   + F  Y  +  T+         R  R
Sbjct  492    INKKPSTLVLSVRIYQREDNFTVQIPLNRSQIHMKTFITYDPEFDTEIRNSLISSTRSPR  551

Query  10534  RKSIEFDDPVEDVENVIHDNVVSS-KEVRMYRTELYGHSVLGVTFTKADIDYRVLLHMTN  10710
               +I F     +V+N + +  + + +E+RMYR  L  H++L V FT +    +V+L
Sbjct  552    --AISFSST--EVKNCLTEGTLQTHQEIRMYRMVLMEHTLLAVQFTTSTHKLQVVLRTKE  607

Query  10711  NPQLNDIETKNTTCLVKSGSKPTALLLRNLCKEARPVYIYIRA--ENILRNWDS--FRET  10878
               P + +I   N+ C + + +    +LLRN C +A   YI +R   +  +   D+
Sbjct  608    KPSVREIS--NSKCTIPALTTNATILLRNNCHKADRAYIALRLYRDTSMGKADNSPIANG  665

Query  10879  GAYYTFSTEMRSCRIWKFARP--EPSWQTILCIPEMNKSVSNGIHCRCNFISDLDADAKP  11052
               A + F  ++RSC  W+++       W    C P M+ SV  GI C CN +
Sbjct  666    PARFAFVFQIRSCDTWQYSASLDNQKWLHYGCYPTMDLSVEKGIRCTCNVLGTYTDYLYY  725

Query  11053  IIAVRMNLKCHLERPVVGRNYEIIICYIVIPIVAIAFLL-IQMHRAAFWDKPLY-----L  11214
              I  V + +       +   N  II+ Y+   ++ + ++L +  +R+    K +      L
Sbjct  726    IPGVEVPIGVFQAAKL---NILIILVYLTALLIILLWILWLYRYRSKLPSKTIRFPKFEL  782

Query  11215  EDVYSGELCRCGDIIIRISFGGRYHSGSSANIFLLLQS-SRGKREIFVHQDPVKTAFNRN  11391
              +D  +GEL    D++I +  GGR ++G++A++ L  +  S+  R I + QDP       N
Sbjct  783    DDESTGEL---HDVLISLRTGGRINAGTTASVHLSFRGQSQPVRHIEIMQDPENIFLKSN  839

Query  11392  CTIFLRLDREFVQLPVRLALGHDNTGTHPHYFCRSIVITDILTEKTQHFRINRWVRTSPG  11571
               T+++ +    +++P ++ + H+N G  P +F R I +TDI T +TQ F   +W+
Sbjct  840    TTLYIWVRTRDIRIPTKVMVYHNNAGRFPMWFLRRIEVTDIQTGETQVFLARKWLEK---  896

Query  11572  AGSKMHLESTMVLDFATTSPKSIYRWPSRFAIAFELCMGKWYLFQSIIGPWRFGINRNSL  11751
                  + + ST++  +     +    W +RF ++FE+    W L+Q + G WR   +   +
Sbjct  897    --GTLLMSSTLI--YKQGDVRMFDVWKNRFMMSFEMFWINWSLWQPVTGSWRESAHYPRM  952

Query  11752  SRWERSCIYVGKNFVAICIVIIFFGRAEPILCDPSPKRYNDFNIVVWLCLIC  11907
              SR +R C ++ K  V   I   FFG              + ++  ++ L +IC
Sbjct  953    SRAKRFCEFICKLLVNYTICASFFGVTTSENLHLMRSLFLNYKDIIVLTMIC  1004
```

Figure 11: The blast2 alignment for the second part of exon three from CG30048-PA.

Not only does it appear that a new intron has been added in *D. mojavensis*, but it also appears that an intron has been lost. The second exon has two possible donor sites at bp 6937 and bp 6946. Neither of these sites have any confidence as donor sites according to the Goose browser. Exon three is predicted to begin at bp 6987, but the nearest acceptor site that incorporates the entire alignment is found at bp 6947. If either of the donor sites is used for exon two along with this acceptor site for exon three, then there are not enough bases between the two exons to form the splicing loop. Therefore, it seems that the intron between exons two and three has been lost and a new intron has formed within the exon three sequence. Also, there are no stop codons in the intervening sequence, so

the protein sequence could be fully translated. A high confidence donor site was found at bp 7523.

The annotation of the final exon is ambiguous. The alignments for exon three are missing about 50 amino acids between them. The blast2 alignment predicted splice site to begin this exon is bp 9694, but there are no splice sites of any confidence in this region (Figure 12). A few hundred base pairs up stream there is a high confidence acceptor site at bp 9499. Not only does this site replace the missing amino acids in the alignment and begin at a high confidence acceptor site, but all three gene prediction programs (SGP, Twinscan, and GENSCAN) call for this site to be the beginning of an exon.

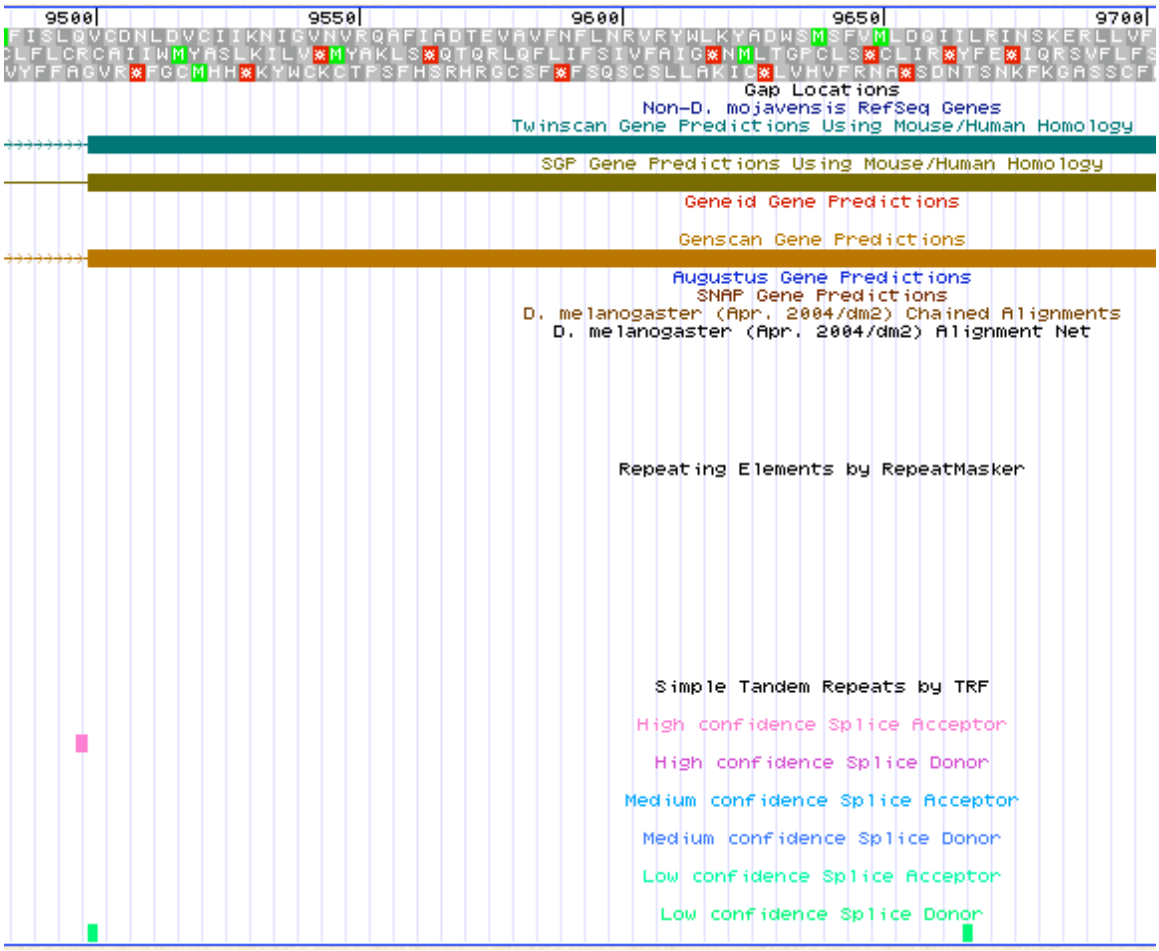


Figure 12: An overview of the region surrounding the beginning of exon three of CG30048-PA.

The exon is proposed to end at bp 11907, and the nearest stop codon is at bp 11973.

CG30048-PB (NP_001014521.1):

This is the second isoform of CG30048. In *D. melanogaster* it only has two exons as opposed to the three in isoform A (Figure 13).

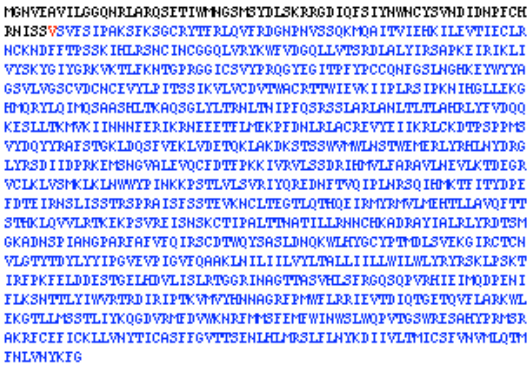


Figure 13: The amino acid sequence of CG30048-PB.

Exon one is the only exon that differs between the two isoforms, but could not be found in fosmid 9. The blast2 program found a plausible alignment beginning at bp 6656, which would put exon one in the correct region (Figure 14), but a methionine could not be found until bp 6824 (Figure 15). The blast 2 alignement terminates at bp 6805, so the entire alignment region would be lost if this methionine were chosen. Also, a stop codon can be found at bp 6944, severely truncating the first exon if one of these methionines were used.

```
 Score = 32.3 bits (72),  Expect =  4799
 Identities = 17/50 (34%), Positives = 27/50 (54%), Gaps = 0/50 (0%)
 Frame = +2

Query  6656  AIIRGGPSRKVFTGNSITIDGSHSRDLSLPKNAKQLLLFEWLCDSAINIE  6805
             A+I GG +R       +I ++GS S DLS  +   Q  ++ W C S  +I+
Sbjct  6     AVILGGQNRLARQSETIWMNGSMSYDLSKRRGDIQFSIYNWNCYSVNDID  55
```

Figure 14: The blastp alignment for exon 1 of CG30048-PB.

Figure 15: The region surrounding the predicted site for exon one (frame 2). The methionines occur after the proposed alignment ends.

The second exon was found to be split in two once again and was annotated identically to the second and third exons in CG30048-PA.

CG12636-PA (NP_609717.2):

This gene is the last of the blastp predictions, and was not used in the final annotation of the fosmid sequence. The amino acid sequence was found for this protein (Figure 16) and the first exon was aligned to the fosmid sequence with blast2 (Figure 17).

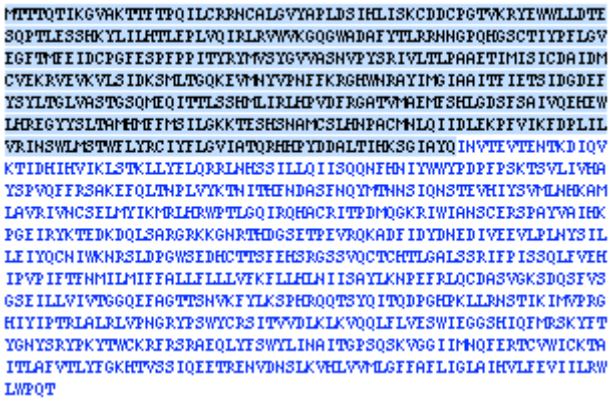


Figure 16: The amino acid sequence for CG12636-PA.

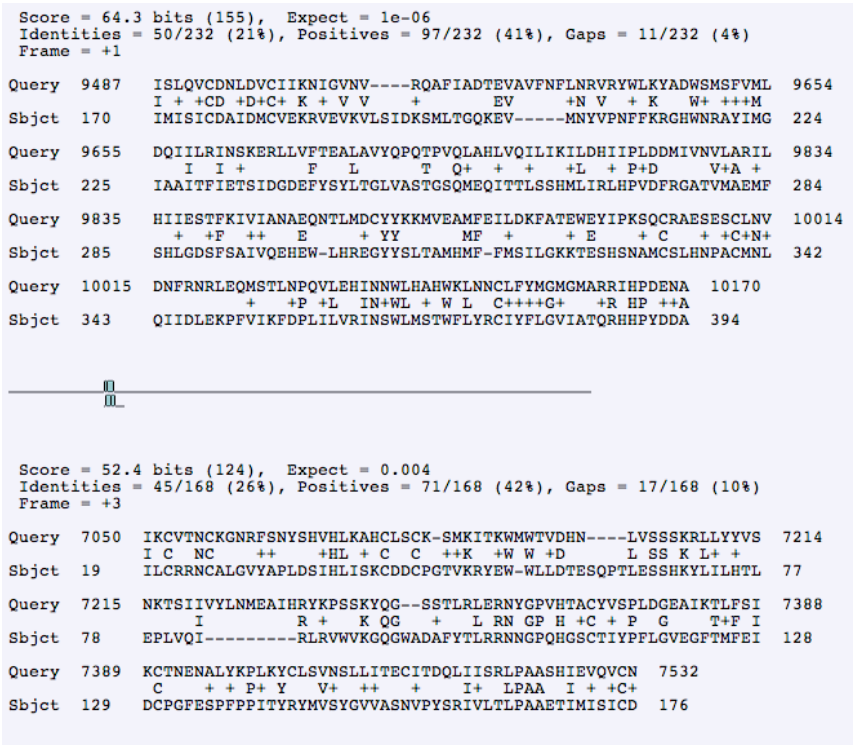


Figure 17: The blast2 alignment for exon 1 from CG12636

Again it appears that a new intron has been introduced to this exon in *D. mojavensis*. Exon one is predicted to begin at bp 7050, but the only methionine that is close enough and doesn't loose any of the alignment is at bp 6999 (Figure 18).

Figure 18: The start of exon one of CG12636-PA.

The exon is proposed to end at bp 7532, but there is a high confidence splice donor at bp 7523. Also, the SGP gene model ends here. Few aligned amino acids would be lost, and only one of them is identical to the *D. melanogaster* sequence.

A similar logic was used in annotating the second exon. The predicted site was at bp 9487, but there is a high confidence splice acceptor at bp 9499. All three gene prediction programs mentioned earlier predict an exon beginning here. The end of this exon was another matter entirely. The alignment ends at bp 10183, but there were no splice sites that had any confidence (Figure 19). Also, this predicted ending site did not match any of the three programs. Therefore, the proposed ending site is not made with any confidence. It does however, include most of the *D. melanogaster* amino acid sequence.

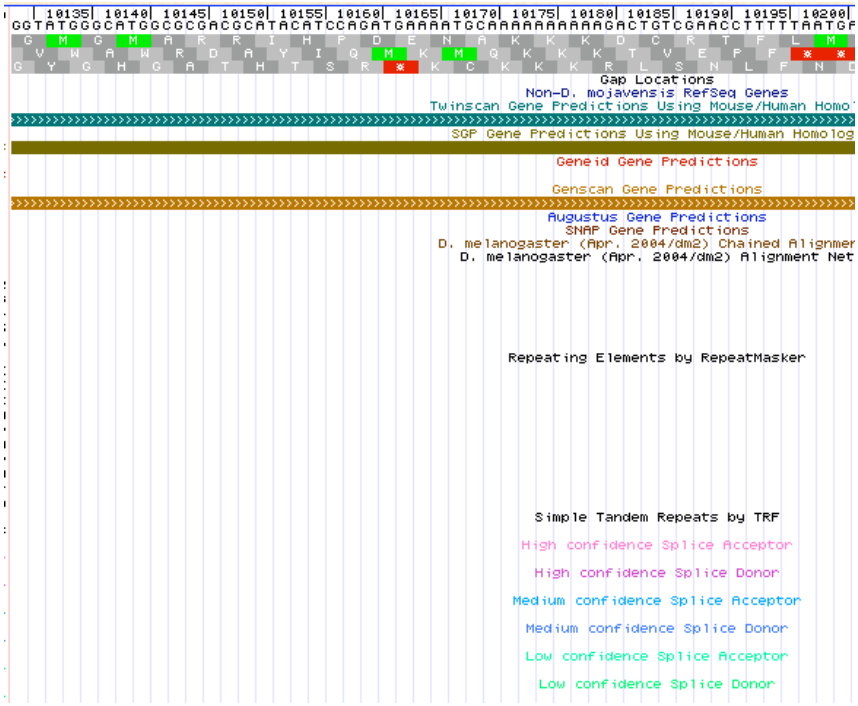


Figure 19: The region surrounding the end of exon 2 of CG12636-PA.

The final exon was not much better. The alignment itself was missing about 40 amino acids from the end (Figure 20), even after some of the boundaries were extended based on a medium confidence splice acceptor site at the beginning of the exon, and the nearest stop codon at the end of the alignment.

```
 Score =  152 bits (383),  Expect = 4e-33
 Identities = 129/582 (22%), Positives = 243/582 (41%), Gaps = 82/582 (14%)
 Frame = +1

Query  10243  IKSADAMHTLIFTEKLLQELRHLLRNDEVLISIRSHK--HSQYWWYPEQ-DSRTQVLVVN  10413
              +K+ D +H +  + KLL EL+  L +  +L+ I S +  H+ YWWYP+   S+T VL+V+
Sbjct  14     VKTIDHIHVIKLSTKLLYELQRRLNHSSILLQIISQQNFHNIYWWYPDPFPSKTSVLIVH  73

Query  10414  AYTSTAILEKTQ--QLSEPFQYISKLTTKSCTYQYCKGRRRRRKSIEFDDPVEDVENVIH  10587
              AY+         +  QL+ P  Y + +T                    F+D   + + + +
Sbjct  74     AYSPVQFFRSAKEFQLTNPLVYKTNIT------------------HFNDASFN-QYMTN  113

Query  10588  DNVVSSKEVRMYRTELYGHSVLGVTFTKADIDYRVLLHMTNNPQLNDIETKNTTCLVKSG  10767
              +++ +S EV +Y   L   ++L V          + + +   P L  I  +   C +
Sbjct  114    NSIQNSTEVHIYSVMLNHKAMLAVRIVNCSELMYIKMRLHRWPTLGQI--RQHACRITPD  171

Query  10768  SKPTALLLRNLCKEARPVYIYIRAENILR------------------------------  10854
               +   + + N C E  P Y+ I     +R
Sbjct  172    MQGKRIWIANSC-ERSPAYVAIHKPGEIRYKTEDKDQLSARGRKKGNRTHDGSETPEVRQ  230

Query  10855  --------NWDSFRET-GAYYTFSTEMRSCRIWKFARPEPSWQTILCIPEMNKSVSNGIH  11007
                     N D   E     Y+   E+  C IWK    +P W    C      S  + +
Sbjct  231    KADFIDYDNEDIVEEVLPLNYSILLEIYQCNIWKNRSLDPGWSEDHCTTSFEHSRGSSVQ  290

Query  11008  CRCNFISDLDADAKPIIAVRMNLKCHLERPVVGRNYEIIICYIVIPIVAIAFLLIQMHR-  11184
              C C+ +  L +   PI +       H+  P+   N  ++I + ++ ++ +    L+ ++
Sbjct  291    CTCHTLGALSSRIFPISS--QLFVEHIPVPIFTFNMILMIFFALLFLLLVFKFLLHLNII  348

Query  11185  AAFWDKPLYLE--DVYSGELCRC----GDIIIRISFGGRYHSGSSANIFLLLQSSRGKRE  11346
              +A+   P +    D   G+  +      +I++ I  GG+  +G+++N+   L+S   ++
Sbjct  349    SAYLKNPEFRLQCDASVGKSDQSFVSGSEILLVIVTGGQEFAGTTSNVKFYLKSPHRQQT  408

Query  11347  IF-VHQDPVKTAFNRNCTIFLRLDREFVQLPVRLALGHDNTGTHPHYFCRSIVITDILTE  11523
               + + QDP      RN TI + + R  + +P RLAL     G +P ++CRSI + D+  +
Sbjct  409    SYQITQDPGHPKLLRNSTIKIMVPRGHIYIPTRLALRLVPNGRYPSWYCRSITVVDLKLK  468

Query  11524  KTQHFRINRWVRTSPGAGSKMHLESTMVLDFATTSPKSIYRWPSRFAIAFELCMGKWYLF  11703
               Q F +  W+      GS +    +   S    Y W  RF    E     WYL
Sbjct  469    VQQLFLVESWIE----GGSHIQFMRSKYFTYGNYSRYPKYTWCKRFRSRAEQLYFSWYLI  524

Query  11704  QSIIGPWRFGINRNSLSRWERSCIYVGKNFVAICIVIIFFGR  11829
               +I GP +  +    ++++ER+C+++ K  + +  V ++FG+
Sbjct  525    NAITGPSQSKVGGIIMNQFERTCVWICKTAITLAFVTLYFGK  566
```

Figure 20: The blast2 alignment for exon three of CG12636-PA.

Feature One Conclusion:

All three potential genes have ambiguities in their models, but CG30048-PA has the fewest. Most of its exon boundaries are marked by high confidence donor or acceptor sites. This was not the case in CG12636-PA, where the end of exon two was quite ambiguous. Also, CG30048-PA had the best blastp E value, albeit by a small margin. It is disconcerting that the other isoform of this gene is missing its first exon. The exon is very small and if the lack of conservation seen in the other exons is present in this one, the blast2 alignment program would not be able to give a predicted alignment with any confidence. This is seen in its prediction. It is quite possible that the exon is present in the sequence, but is not conserved, it could not be recognized. Table 1 details the final annotation boundaries for this feature, though it must be stated that these predictions are not made with a great deal of confidence. There is significant evidence that this feature could be annotated as any of the three genes.

This gene has an ambiguous function in *D. melanogaster*, but in mammalian species such as *Bos taurus, Canis familiaris,* and *Homo sapiens*, this gene is connected with polycystic kidney disease 1. CG30048-PA is located on the right arm of chromosome two. Thus, a rearrangement in the genome has occurred at this location. This is the only gene in the fosmid found to be on the second chromosome in *D melanogaster*.

Table 1: The Exon Boundaries for CG30048-PA in Fosmid 9

Exon Number	Beginning Base Pair	Ending Base Pair	Frame
1	5750	6299	+2
2	6895	7523	+3
3	9499	11973	+1

GENSCAN Feature 2:
CG31999 (NP_726551):

A blastp search was done with the GENSCAN predicted amino acid sequence (Figure 21). CG31999 was returned as the best hit by a margin of 85 orders of magnitude, making it the clear choice to investigate. CG31999, a part of the fibulin precursor family, was found to be the best match. Also the gene for this protein was located on the fourth chromosome of *D. melanogaster*; marking it as the only feature in the fosmid to have this derevation.

Sequences producing significant alignments:		Score (bits)	E Value
gnl|dmel|FBpp0088211	type=protein; loc=4:complement(235731....	348	8e-96
gnl|dmel|FBpp0076512	type=protein; loc=3L:join(7556179..755...	66	6e-11
gnl|dmel|FBpp0076511	type=protein; loc=3L:join(7556179..755...	66	6e-11
gnl|dmel|FBpp0073715	type=protein; loc=X:join(14087412..140...	59	1e-08
gnl|dmel|FBpp0073714	type=protein; loc=X:join(14087412..140...	59	1e-08

Figure 21: The blastp results using the non-redundant protein database for GENSCAN feature two.

The Ensembl database found that this protein has 13 exons, many of them quite small (Figure 22). This posed a problem when looking for these exons with a blast2 alignment. The smaller the exon becomes, the fewer data points the software has to use, and any mismatches that have occurred throughout this protein's evolutionary history would have a greater impact on the blast2 scores. Because of this, the second exon was annotated first because it is significantly longer than the first exon. With the second exon located, the rest of the protein can be built around this location.

```
MFYKLCAFMCFLIVGGHVTTTMASDSISGYIRKCCINGLRMARTTASCKKIDIAPTIIPQ
LWLGLCHSTLEVCCSRILDHQDCILGRLAALDGTRCDGIGHVTSSSYATCCRSCQIGLAV
KASKANCKDPLFSFIFLIESYRACCYGSADFKDQPGIDEIDKANSITDEGILPFVSEEDM
NVTIVLTGDDDICGKIENLCAHICINTFDAYQCKCHPGFMLDNNNVTCSPMKTQICPSGY
NLDKLDNKCIDIDECRIDLHDCKSSQYCHNTNGGYHCLNVKEKICPPGFHYDHDYDACKD
DYKCKDRKCVKIQSCDKGFSLHNGTCSDIDECSHKSLNNCHVNSNQICVNTVGSYSCNCL
PGFNLDATLNKCVDINECSINNHNCLPTQRCDNTIGSYICTRLQSCGTGYTLNAETGNCD
DDDECTLSTHNCPSNYDCHNTRGSFRCYRKISTMLTTRTTSTTVPPLSLENARRSFTSRY
PYPLAVHPIYSQHNDSISTNRRVDCSPGFYRNTLGACIDTNECMEQNPCGNHERCINTNG
HFRCESLLQCSPGYKSTVDGKSCIDIDECDTGEHNCGERQICRNRNGGFVCSCPIGHELK
RSIGGASTCVDTNECALEQRVCPLNAQCFNTIGAYYCECKAGFQKKSDGNNSTQCFDIDE
CQVIPGLCQQKCLNFWGGYRCTCNSGYQLGPDNRTCNDINECEVHKDYKLCMSLCINTPG
SYQCSCPRGYILAADMNTCRDVDECATDSINQVCTGRNDICTNIRGSYKCTTVNCPLGYS
IDPEQKNRCRQNLNFCEGEECYTQPSAFTYNFITFVSKLMIPPDGRTIFTLRGPLWYDNI
IFDLKIVRIQATTNIQKATDGSFDTLQNHNQVNVILKKSLEGPQDIELELSMTVYTNGMP
RGKSVAKLFLFVSQHTF
```

Figure 22: The amino acid sequence for CG31999-PA.

Using the amino acid sequence found in Figure 22, a blast2 alignment was tried using the fosmid sequence and the sequence from exon two (Figure 23). A good match was found and the exon beginning was annotated to bp 27522. There is a problem with this annotation however. There is a large gap in the alignment that was caused by a stop codon found in the middle of the open reading frame (Figure 24). It appears that a small intron has inserted itself in the second exon. A medium confidence splice donor site is at bp 27794. The RefSeq and the GENSCAN predictions both end here as well. The intron

ends at bp 27861. There are no splice sites of any confidence according to the Goose program, but the RefSeq and GENSCAN lines both mark this area as coding sequence.

```
 Score =  167 bits (423),  Expect = 1e-37
 Identities = 89/194 (45%), Positives = 118/194 (60%), Gaps = 29/194 (14%)
 Frame = +2

Query  27521  EQIADYIRKCCISGLRNARTTSVCDKMDSTIVNISNLWLGLCSSTFGVCCSKELDRQNCE  27700
              + I+ YIRKCCI+GLR+ARTT+ C K+D     I  LWLGLC ST  VCCS+ELD Q+CE
Sbjct  1      DSISGYIRKCCINGLRHARTTASCKKIDIAPTIIPQLWLGLCHSTLEVCCSRELDHQDCE  60

Query  27701  LGRLAALEGASCNKGFNLTSTSYTNCCRACQGILEKITYIYL*NVLINVMGFPVGLAVKA  27880
              LGRLAAL+G  C+   N+TS+SY  CCR+CQ                      +GLAVKA
Sbjct  61     LGRLAALDGTRCDGEGNVTSSSYATCCRSCQ----------------------IGLAVKA  98

Query  27881  SQQKCTDPLFSFLLNIDSYRLCCSDDGFSNSELESESGLEGKNKLVLYTDHDEREEEIVE  28060
              S+  C DPLFSF+  I+SYR CC    + +++ + + G++  +K    TD  E
Sbjct  99     SKANCKDPLFSFIFLIESYRACC----YGSADFKDQPGIDEIDKANSITDEGELP---FV  151

Query  28061  SRETVDGTIVLSGD  28102
              S E ++ TIVL+GD
Sbjct  152    SEEDMNVTIVLTGD  165
```

Feature 23: The blast2 alignment for exon two of CG31999-PA. Note the large gap in the alignment along with the * signaling a stop codon.

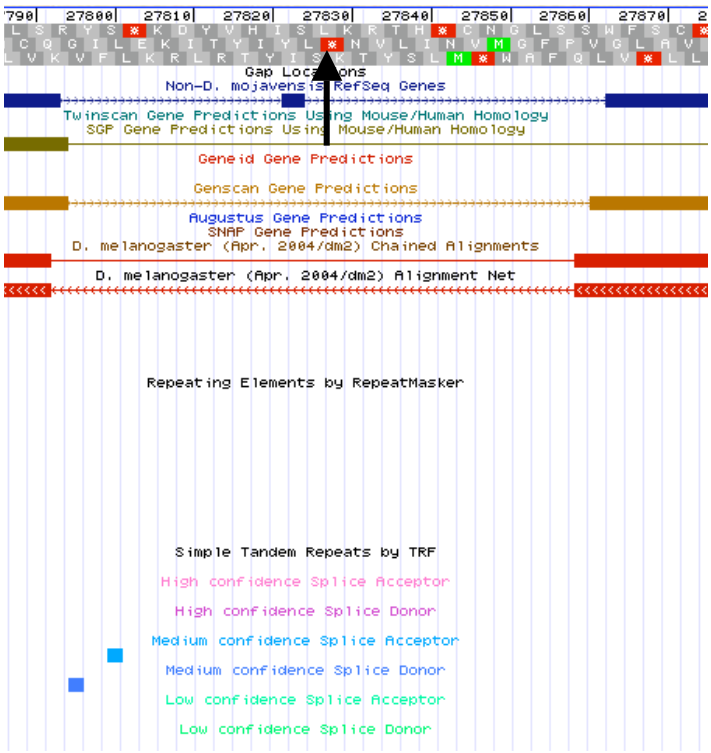


Figure 24: An overview of the region containing a new intron in the second exon of CG31999-PA.

The second part of this exon, now exon three, finishes at bp 28103 at a high confidence splice donor site. The blast2 alignment, SGP, and GENSCAN prediction lines all end here.

With the location of the second and third exons established, the smaller exons surrounding them could be located with more precision. Exon one and exon four were mapped once again using the blast2 sequences alignment tool (Figures 25 and 26).

```
 Score = 11.2 bits (17),  Expect = 11448300000
 Identities = 4/17 (23%), Positives = 10/17 (58%), Gaps = 0/17 (0%)
 Frame = +2

Query  20699  FNKILSNILKIIVNSNM  20749
              F K+ + +  +IV  ++
Sbjct  2      FYKLCAFMCFLIVGGHV  18
```

Figure 25: The blast2 alignment for exon one from CG31999-PA.

```
 Score =  172 bits (436),  Expect = 3e-39
 Identities = 71/110 (64%), Positives = 84/110 (76%), Gaps = 0/110 (0%)
 Frame = +3

Query  28275  DDICGKIPNLCEQVCINTYDAYRCSCHPGYKLNDNNVTCYADKNNICPSGYVLDGNQGKC  28454
              DDICGKI NLC  +C NT+DAY+C CHPG+ L++NNVTC   K  ICPSGY LD    KC
Sbjct  1      DDICGKIENLCAHICENTFDAYQCKCHPGFMLDNNNVTCSPMKTQICPSGYNLDKLDNKC  60

Query  28455  VDIDECQEQLHDCKTSQYCHNTIGGYHCLNIKAKNCPAGYLYNVKSDECE  28604
              +DIDEC+E LHDCK+SQYCHNT GGYHCLN+K K CP G+ Y+   D C+
Sbjct  61     IDIDECREDLHDCKSSQYCHNTNGGYHCLNVKEKECPPGFHYDHDYDACK  110
```

Figure 26: The blast2 alignment for exon four of CG31999-PA.

The first exon was very small, and had many poor alignment possibilities, but none of them were within a reasonable distance from the second exon start site. The nearest match was at bp 20747. It was not only on the correct strand, but it was not in an area that was not marked as coming from a different chromosome. This region also encompasses a small region that is predicted to be an exon by the SGP gene prediction program (Figure 27). The exon also ends with a high confidence splice donor site. This provided the best evidence for the placement of the first exon, but again this conclusion is not made with as much confidence as the placement of the other exons.

The fourth exon, for instance, has a much more confidant placement. It was mapped to bp28276, which corresponds to the starting points for RefSeq data and GENSCAN and SGP predictions. The donor site is mapped to bp 28605, which corresponds to the ending sites for SGP and GENSCAN predictions as well has a high confidence donor site.

None of the remaining exons could be found in fosmid 9, but they were found in two adjacent fosmids further down on the chromosome. This gene has a huge size in *D. mojavensis,* possibly extending for more than 60 kb. However, the exons that were mapped were done with relatively high confidence. Table two details the exon boundaries of CG31999-PA in this fosmid of *D. mojavensis*.

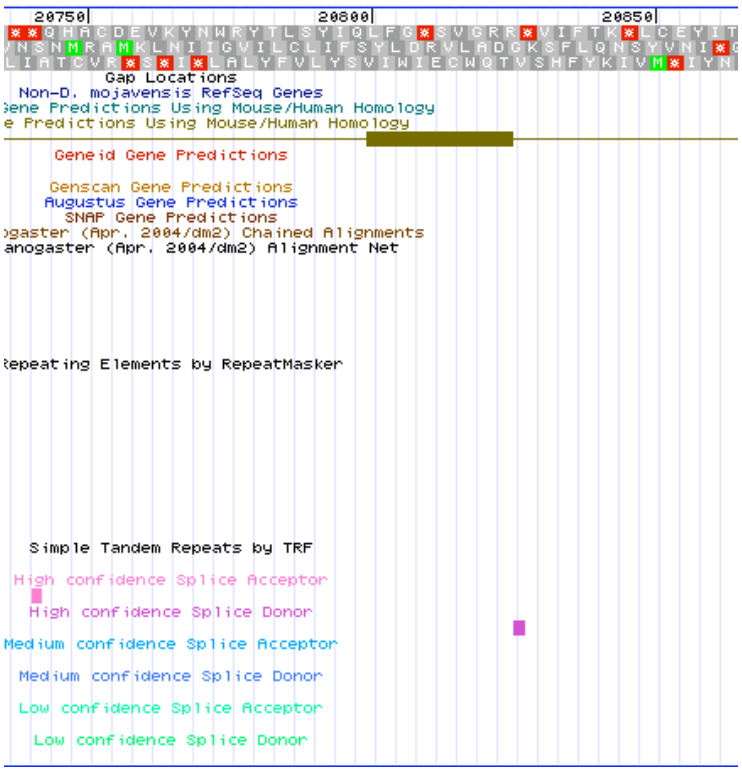


Figure 27: An overview of the region for the first exon of CG31999-PA.

Table 2: The Exon Boundaries for CG31999-PA in Fosmid 9.

Exon Number	Beginning Base Pair	Ending Base Pair	Frame
1	20747	20825	+2
2	27522	27794	+2
3	27861	28103	+2
4	28276	28605	+3

GENSCAN Feature 3:

This is most likely a false prediction by GENSCAN. The GENSCAN predicted amino acid sequence was searched for using the blastp program; there were no matches with acceptable E values (Figure 28). Because of the poor E values, there was no further investigation of this feature.

Sequences producing significant alignments:		Score (bits)	E Value
gnl|dmel|FBpp0071668	type=protein; loc=2R:17954456..1795567...	26	4.7
gnl|dmel|FBpp0071829	type=protein; loc=2R:complement(186103...	26	4.7
gnl|dmel|FBpp0071828	type=protein; loc=2R:complement(186051...	26	4.7
gnl|dmel|FBpp0082691	type=protein; loc=3R:complement(118705...	25	8.1

Figure 28: The blastp results for the GENSCAN prediction of feature three.

RefSeq Predicted Gene One:
CG5262-PA (NP_649223):

The gene for this protein is on chromosome 3L in *D. melaogaster*, and potentially represents another rearrangement event. CG5262-PA is a transmembrane protein in *Homo sapiens* but has an unknown function in most other species. This protein was not found by a GENSCAN prediction, but instead by RefSeq evidence (Figure 29). The corresponding amino acid sequence was found on the Ensembl website (Figure 30) and the exons were all annotated using the blast2 alignment tool (Figures 31, 32, and 33). Unfortunately, the alignments for exon one were far too poor to form a hypothesis on where the first exon would lie. Because the exon is so small, the expect value had to be increased a great deal in order for any alignment to be made. There were many short sequences that were plausible matches, but the evidence for each of them was far too poor to make any definite conclusions. Also, because the rest of the protein was very well conserved, it seemed increasingly unlikely that any of the poor matches that the blast2 alignment was proposing were correct. Therefore, alternate methods were also used. Some of the surrounding regions that were predicted to be coding sequence by various gene predicting proteins were extracted, and an alignment was attempted, but these tests all failed as well. If this gene is functional in *D. mojavensis*, it must have a starting exon that is either so divergent from the *D. melanogaster* sequence that it could not be recognized, or it falls outside of the limits of fosmid 9. Table three enumerates the exon boundaries for CG5262-PA.

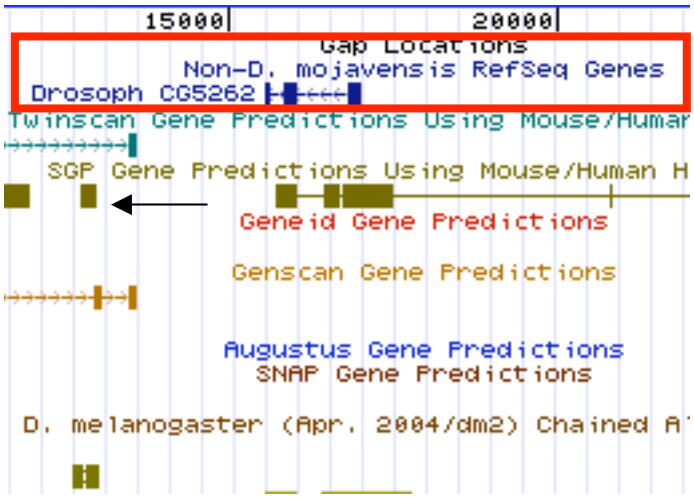


Figure 29: The Genome Brower from Goose showing the RefSeq evidence for CG5262-PA in fosmid 9. The arrow points to a predicted open reading frame that was tested for sequence alignment.

MPRLVNGREAAPTYSHLVGFIFIFNLIVGTGALTLPGVFARAGWMLSLIVIVLLAIISYM
TVTFIIEAMACANAIRNWQTLQALRQSRSSAENSENDDNADDVSLASGAEQVGDFERVPL
TIQNREFHTYQLSHKFELGEMATLFFNEFGRVMEYLCLIVYLYGDLSIYSAAVARSLRDV
VCDQTNGTDTNNLMTWPGDFENNTSLACWKEHTISRLNMTRVLLIGFTLIFGPFVYFNVQ
KTKYLQMLTAAFRWMAFTLMICISLKLLISRGAKGHPATFNVTGIPSLFGACVYSFMCHH
SLPSLLAPIRHKSMVSKILSIDYIIICAFYILLAMTGIFAFERIEDLYTLDFLPYDVAYV
DFWSGLLICIDYFLALFPIFTLSTSFPIVAITLKNNLQSLFLDMSQYESYSVILRLCFPL
LAIIPPFCITYFTESLSSLVAFTGTYAGTGIQYIIPVFLVYFARRTCSELLGSGVVNRFK
SPFKSSAWLVFVFIWSILCVCLVSINLFS

Figure 30: The amino acid sequence for CG5262-PA in *D. melanogaster*.

```
 Score =  220 bits (561),  Expect = 1e-53
 Identities = 111/164 (67%), Positives = 133/164 (81%), Gaps = 12/164 (7%)
 Frame = +2

Query  15575  VGFIFIFNIIVGTGALTLPGVFAKSGWCLSFVVLILLALVSYITVTFVIEAMAGANAIKN  15754
              VGFIFIFN+IVGTGALTLPGVFA++GW LS +V++LLA++SY+TVTF+IEAMA ANAI+N
Sbjct  1      VGFIFIFNLIVGTGALTLPGVFARAGWMLSLIVIVLLAIISYMTVTFIIEAMACANAIRN  60

Query  15755  WQSLQALRYKRNS--------NEEDNELYWNTES----ETVPLTIQNMQFHYYQLTQKFE  15898
              WQ+LQALR  R+S         N +D  L    E     E VPLTIQN +FHYYQL+ KFE
Sbjct  61     WQTLQALRQSRSSAENSENDDNADDVSLASGAEQVGDFERVPLTIQNREFHYYQLSHKFE  120

Query  15899  LGEMAKIFFNDCGRILFYLCFIIYLYGDLSIYSAAVARSLRDVL  16030
              LGEMA +FFN+ GR++FYLC I+YLYGDLSIYSAAVARSLRDV+
Sbjct  121    LGEMATLFFNEFGRVMFYLCLIVYLYGDLSIYSAAVARSLRDVV  164
```

Figure 31: The blast2 alignment for exon two of CG5262-PA.

```
 Score = 88.2 bits (217),  Expect = 7e-14
 Identities = 39/77 (50%), Positives = 55/77 (71%), Gaps = 4/77 (5%)
 Frame = +3

Query  16440  DHSNNSDISISELSSMRLLGGLKDNDNRTCWKEHNVSRLIVYRIILIGFTMLFGPLVLFS  16619
              D +N +D +     + M   G  ++N +  CWKEH +SRL +YR++LIGFT++FGP V F+
Sbjct  2      DQTNGTDTN----NLMYWPGDFENNTSLACWKEHTISRLNMYRVLLIGFTLIFGPFVYFN  57

Query  16620  IQKTKYLQMLTVVFRWL  16670
              +QKTKYLQMLT  FRW+
Sbjct  58     VQKTKYLQMLTAAFRWM  74
```

Figure 32: The blast2 alignment for exon three of CG5262-PA.

```
 Score =  374 bits (961),  Expect = 4e-100
 Identities = 178/253 (70%), Positives = 212/253 (83%), Gaps = 0/253 (0%)
 Frame = +1

Query  16738  AFILMICIALKMLVIDGVKGHPVAVNFYGIPSLFGAGVYSFMCHHSLPSLLAPIKHKSMA  16917
              AF LMICI+LK+L+  G KGHP   N YGIPSLFGA VYSFMCHHSLPSLLAPI+HKSM
Sbjct  1      AFTLMICISLKLLISRGAKGHPATFNVYGIPSLFGACVYSFMCHHSLPSLLAPIRHKSMV  60

Query  16918  KKILSYDYILICSFYIVLAITGIFAFEYVEDLYTLNFLPYHATSESLFSNFLIVIDYFLS  17097
               KILS DYI+IC+FYI+LA+TGIFAFE +EDLYTL+FLPY       +S  LI IDYFL+
Sbjct  61     SKILSIDYIIICAFYILLAMTGIFAFERIEDLYTLDFLPYDVAYVDFWSGLLICIDYFLA  120

Query  17098  LFPVFTLSTSYPLIAITLKNNLQTLFLDMSQYNSYSILIRALFPFLSILLPFCITYFTEN  17277
              LFP+FTLSTS+P++AITLKNNLQ+LFLDMSQY SYS+++R  FP L+I+ PFCITYFTE+
Sbjct  121    LFPIFTLSTSFPIVAITLKNNLQSLFLDMSQYESYSVILRLCFPLLAIIPPFCITYFTES  180

Query  17278  LSILVVITGSYAGVGIQYVIPVCLVYYSRVTCSELLGSGIKNQFQSPFQSNIWLLMVLAW  17457
              LS LV  TG+YAG GIQY+IPV LVY++R TCSELLGSG+ N+F+SPF+S+ WL+ V  W
Sbjct  181    LSSLVAFTGTYAGTGIQYIIPVFLVYFARRTCSELLGSGVVNRFKSPFKSSAWLVFVFIW  240

Query  17458  SVLSVFLVSINLF  17496
              S+L V LVSINLF
Sbjct  241    SILCVCLVSINLF  253
```

Figure 33: The blast2 alignment for exon four of CG5262-PA.

Table 3: The Exon Boundaries for CG5262-PA in Fosmid 9.

Exon Number	Beginning Base Pair	Ending Base Pair	Frame
1	?	?	?
2	15575	16032	+2
3	16433	16671	+3
4	16739	17499	+1

RefSeq Predicted Gene Two:
Or13a (NP_523359.2):

This is the final feature that was examined in fosmid 9. Or13a is found on the X chromosome in *D. melaogaster* and functions as an odorant receptor. Once again, this feature was recognized from RefSeq data as opposed to GENSCAN preditions (Figure 34). The amino acid sequence was found for this protein on the Ensembl website (Figure 35) and the first exon was aligned to fosmid 9 using the blast2 sequences program. It was found at the very beginning of fosmid 9 on the negative strand (Figure 36). This exon finished just before the fosmid ended, and none of the other exons were found. This gene was most likely truncated by the fosmid boundaries and the rest of the protein exists in the neighboring fosmid. Table four gives the exon boundaries for the single exon found in fosmid 9.

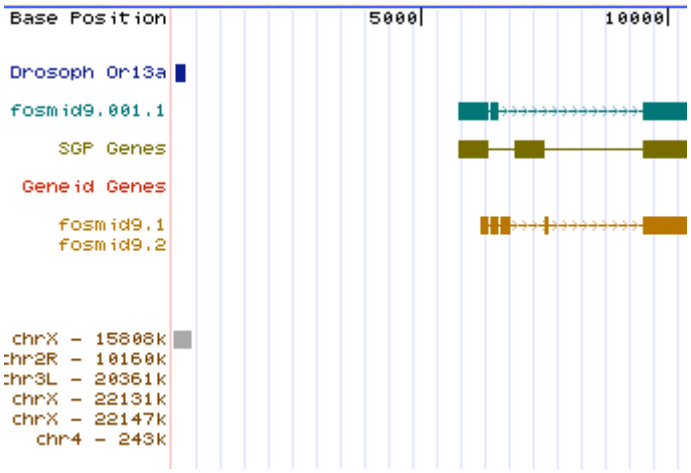


Figure 34: The Genome Brower from Goose showing the RefSeq evidence for Or13a in fosmid 9.

MFYSYPYKALSFPIQCVWLKLNGSWPLTESSRPWRSQSLLATAYIVWAWYVIASVGITIS
YQTAFLLNNLSDIIITTENCCTTFMGVLNFVRLIHLRLNQRKFRQLIENFSYEIWIPNSS
KNNVAAECRRRMVTFSIMTSLLACLIIMYCVLPLVEIFFGPAFDAQNKPFPYKMIFPYDA
QSSWIRYVMTYIFTSYAGICVVTTLFAEDTILGFFITYTCGQFHLLHQRIAGLFAGSNAE
LAESIQLERLKRIVEKHNNIISFAKRLEDFFNPILLANLMISSVLICMVGFQIVTGKNME
IGDYVKFIIYISSALSQLYVLCENGDALIKQSTLTAQILYECQWEGSDRIEIQSFTPTTK
RIRNQIWFMILCSQQPVRITAFKFSTLSLQSFTAILSTSISYFTLLRSVYFDDEKKLD

Figure 35: The amino acid sequence of Or13a in *D. melanogaster*.

```
 Score =  149 bits (376),  Expect = 3e-32
 Identities = 76/122 (62%), Positives = 86/122 (70%), Gaps = 5/122 (4%)
 Frame = -2

Query  371  MFNPLPYKDPSFRMPFQCIWLKLNGSWPLEIKLANQKLLTRSYLFAFIYNVWALYVIISV  192
            MF   PYK  SF  P QC+WLKLNGSWPL     + +      L A  Y VWA YVI SV
Sbjct  1    MFYSYPYKALSF--PIQCVWLKLNGSWPLT---ESSRPWRSQSLLATAYIVWAWYVIASV  55

Query  191  GITISFQTSFLINNFGDIIMTTENCCSTLMGALNFVRLIHLRVNQPKFRKLINQFVLNIW  12
            GITIS+QT+FL+NN  DII+TTENCC+T MG LNFVRLIHLR+NQ KFR+LI  F   IW
Sbjct  56   GITISYQTAFLLNNLSDIIITTENCCTTFMGVLNFVRLIHLRLNQRKFRQLIENFSYEIW  115

Query  11   IP  6
            IP
Sbjct  116  IP  117
```

Figure 36: The blast2 alignment for exon one of Or13a.

Table 4: The Exon Boundaries for Or13a in Fosmid 9.

Exon Number	Beginning Base Pair	Ending Base Pair	Frame
1	371	6	-2

In conclusion, there was evidence for four genes found in fosmid 9: CG30048-PA, CG31999-PA, CG5262-PA, and Or13a. The appendix contains the DNA and amino acid sequences that code for each of these proteins as compiled by Gene Checker.

Clustal Analysis:

Protein Analysis:

CG30048-PA was chosen for the CLUSTALW alignment because it was the only protein in the fosmid that was annotated in its entirety. To gather the species that would be used in comparison, the proposed amino acid sequence from Gene Checker was submitted to a blastp search against the non-redundant protein database. Three other species besides *D. mojavensis* were chosen: *D. melanogaster, Tetraodon nigroviridis* (the green spotted pufferfish), and *Monodelphis domestica* (the gray short tailed opossum). These species all had the top blastp hits when compared to the proposed *D. mojavensis* sequence and represent a wide swath of the evolutionary divergence from *D. mojavensis*. The protein sequence from each of the species was submitted to CLUSTALW, and the results were interpreted. There was very little conservation of this protein across these species, except for one region in the middle, which saw moderate conservation (Figure 37). This suggests that the region of moderate conservation is of importance to this protein and the organism, and therefore, change here is selected against. The rest of the protein, however, can be changed without deleterious consequences. The CLUSTALW analysis might have proven more useful if species more closely related to *D. mojavensis* had been used, but because the protein sequence was so poorly conserved between *D. melanogaster* and *D. mojavensis*, the species with the best blastp hits were chosen instead.

Figure 37: The area of moderate conservation as shown by the CLUSTALW program. Tetraodon stands for *T. nigroviridis,* Monodelphis stands for *M. domestica,* D. stands for *D. mojavensis*, and Drosophila stands for *D. melanogaster.*

5’ Upstream Region:

Next an analysis of the upstream region of Or13a was undertaken. This protein was chosen because the annotation of the first exon was the best. The alignment was excellent (see Figure 36). Two kb worth of DNA was taken from the upstream regions of *D. mojavensis, D. melaogaster, D. erecta,* and *D. virilis*. About one hundred base pairs of sequence from the first exon was also taken as an anchor for the CLUSTALW program to use. This sequence was submitted to CLUSTALW, and the results interpreted (Figure 38). The CLUSTALW program was also run with only the sequences from *D. mojavensis* and *D. virilis* to see if conservation could be seen from these evolutionarily closer relatives (Figure 39).

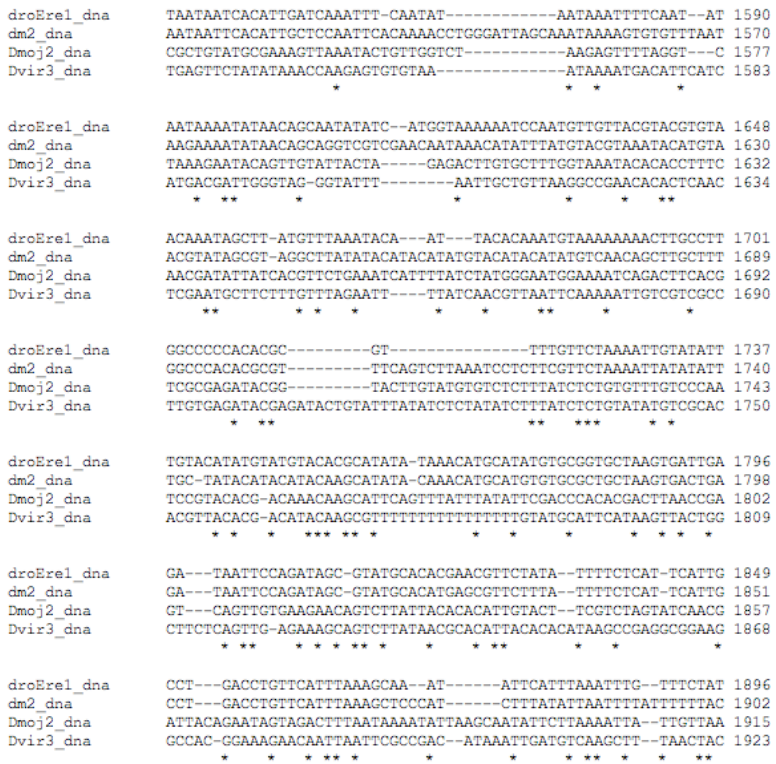


Figure 38: An excerpt from the CLUSTALW alignment for the 5' UTR for the Or13a gene for all four species. The last two to three rows are sequence from the first exon.

In the figure above, it can be seen that the alignment of the sequences is sporadic. There were very few regions that showed long stretches of conservation. This could be due to the fact that in *D. mojavensis* there was a rearrangement event that brought the Or13a gene from the X chromosome to the fourth chromosome. This could have interrupted the 5' UTR region, but this is rather unlikely due the rather extensive conservation between *D. mojavensis* and *D. virilis*. More likely it is simply the fact that the 5' UTR is not under great selective pressure when compared to the gene itself.

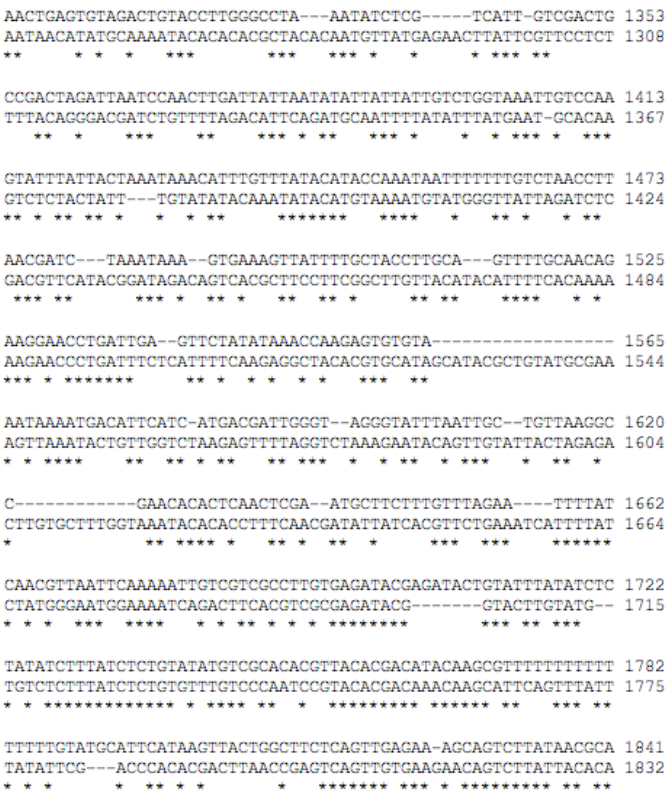


Figure 39: An excerpt from the CLUSTALW alignment for the 5' UTR of the Or13a gene for *D. mojavensis* and *D. virilis* only. Again the last two to three rows represent sequence from the first exon.

When the CLUSTALW alignment was examined for the two species alone, it was found that there is still some degree of conservation between *D. mojavensis* and *D. virilis*. This is not to be unexpected though, because they are closest evolutionarily that we have, only being separated by one common ancestor. It is also of note that most of the discrepancies between the two species derive from in/dels, which are either insertion or deletion events between the two species.

In conclusion, there seems to be only very little conservation between the four species, but a significant amount remains between the two closest species; *D. mojavensis* and *D. virilis.*

Repeat Analysis:

RepeatMasker was run on the fosmid 9 sequence before the annotation process began. Figure 40 summarizes the kinds of repeats found in fosmid 9, while Table 5 details all of the significant repeats found in the fosmid (those over 500 bp). The fosmid is 48.86% repetitive DNA. It is also of note that both high quality discrepancies that were found when finishing the fosmid are contained in repeats. The one positioned at bp 959 is in a DNA element while the discrepancy at position 3126 is in a FB repeat.

```
file name: fosmid9.fasta
sequences:            1
total length:      42300 bp  (42300 bp excl N/X-runs)
GC level:         36.90 %
bases masked:      20667 bp ( 48.86 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:                0           0 bp     0.00 %
      ALUs            0           0 bp     0.00 %
      MIRs            0           0 bp     0.00 %

LINEs:                3         871 bp     2.06 %
      LINE1           0           0 bp     0.00 %
      LINE2           0           0 bp     0.00 %
      L3/CR1          0           0 bp     0.00 %

LTR elements:         5        3028 bp     7.16 %
      MaLRs           0           0 bp     0.00 %
      ERVL            0           0 bp     0.00 %
      ERV_classI      0           0 bp     0.00 %
      ERV_classII     0           0 bp     0.00 %

DNA elements:        29        7665 bp    18.12 %
      MER1_type       0           0 bp     0.00 %
      MER2_type       0           0 bp     0.00 %

Unclassified:         2         116 bp     0.27 %

Total interspersed repeats:    11680 bp    27.61 %


Small RNA:            0           0 bp     0.00 %

Satellites:          27        6511 bp    15.39 %
Simple repeats:      13         705 bp     1.67 %
Low complexity:       8         300 bp     0.71 %
==================================================
```

Figure 40: A summary of the types of repetitious elements found in fosmid 9.

Table 5: A Summary of All the Significant Repeats Found in Fosmid 9.

Number	Beginning Site	Ending Site	Total Length	Repeat Class Family	Repeat Name
1	38563	39781	1218	LTR/Gypsy	TOM_I
2	3844	4750	906	LTR/Gypsy	OSVALDO_LTR
3	28713	29468	755	Satellite	dmoj.0.99.centroi
4	2182	2926	744	DNA	dmoj.1.12.centroid
5	34462	35188	726	LTR	dvir.3.94.centroid
6	35380	36088	708	Satellite	dmoj.0.34.centroi
7	33603	34292	689	Satellite	dmoj.0.34.centroi
8	18019	18647	628	DNA	dmoj.8.25.centroid
9	22917	23541	624	DNA	dmoj.8.25.centroid
10	2982	3487	505	FB	dmoj.35.53.centroid

Synteny Analysis:

Fosmid 9 has a rather interesting construction. Each annotated gene comes from a separate chromosome. As such, every single chromosome is represented in the fosmid sequence. However, when these genes are located using UCSC's August 2005 assembly of *D. mojavensis*, there is only evidence for three of these genes being syntenic: Or13a, CG5262, and CG31999 (Figure 41). CG30048 is found elsewhere in the genome. This evidence further decreases the confidence that I have in annotating feature one as CG30048 in fosmid 9. Finally, Figure 42 looks at fosmid 9 compared to the areas around each gene in *D. melanogaster*. Each window is ~40 kb to mirror the distances seen in fosmid 9.

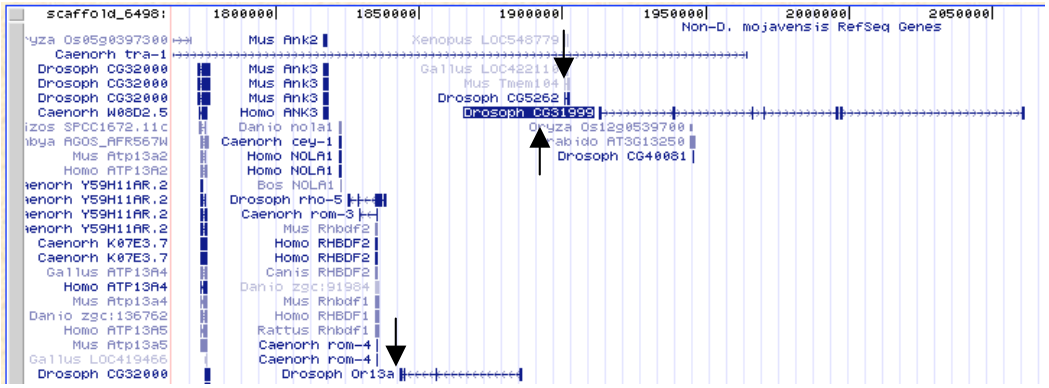


Figure 41: The region containing genes Or13a, CG5262, and CG31999 in the August 2005 assembly of *D. mojavensis*.

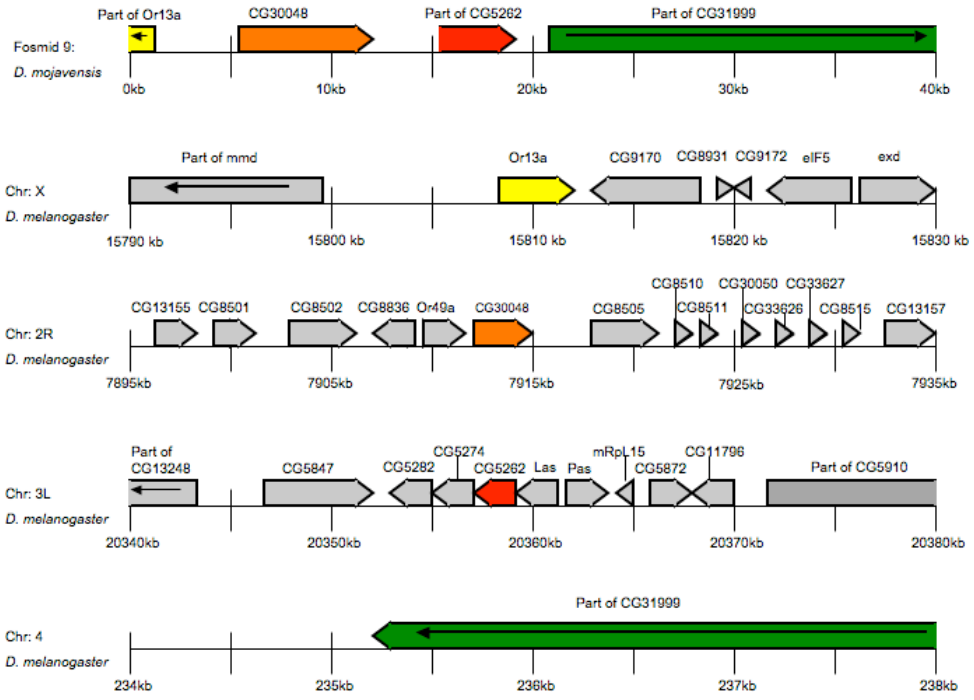


Figure 42: An overview of fosmid 9, and the regions surrounding each of the annotated features in *D. melanogaster*.

Appendix:

CG30048-PA:

DNA Sequence:

>Dmoj2_fosmid9_CG30048-PA_NP_725155_cds

ATGAGAGTGAACCAATTTCTTGTTTTGATTAGCTTGGCAAATGTGACGGC
CTTTATTAGGTACGACAACCCATACGTGTTCGTTAATCGCGTCTATGTGA
ATCACACCATCTTCGTTATTGTAACGCTACAGTCTCTTAGAGGTGTGTGG
TTTCTAACCAAAGTGAGTATAGAAGAACGCATAATTGACTATTTTCAAAC
CAATGCAAGCGAGCACCGTGAACAACTTCTATGGTATCATACAGCAAGCT
TAAAACAAACTCTTAACTTCTATGGTATATTCACCCAAAAGAATCTTCCC
TATCAAGAATGCTTTTATCATGTTTCAAGAGCAAATTTTTTCAAAAGCGG
TTTTCATAATGTTTCTGTCTTGATTACAACGATTGGGATTACATCAGACA
AGCTCAGATCTTCGAGCTATACATACATTGATTTAGATGAAAATGTACGT
TGTGAGCCACAATTTTCTATACCGCAGTGCGTAAATGCAGATTATCCAGT
GCACTATGAAGTTTCTGACATAATTGTGCTTCGGGCAATGTATACTAGGC
GTCCTGCTTTCACGATTCCACCGTTTACTCAACAAGGCGAACAGTCAATT
AGGTTTATTTTAAATATACGTTCCACATTTGATCTTCTTCGGTCATCGGT
TGCTATGCAAACCATATTATTCTCATCATACCCCAGACTTTACATCGATG
TAATTATTAAGTGTGTAACAAACTGCAAAGGAAATAGGTTTTCCAATTAT
TCTCATGTTCACCTTAAGGCCCATTGTCTTTCGTGCAAATCGATGAAAAT
TACCAAGTGGATGTGGACCGTCGACCATAATTTAGTGAGCTCATCGAAAA
GGCTTTTATATTATGTGTCAAATAAAACGTCTATTATAGTATACTTAAAT
ATGGAGGCTATTCATAGATATAAACCATCGTCAAAATACCAAGGATCCTC
CACTCTGCGTCTTGAAAGAAATTATGGTCCTGTACATACCGCTTGTTACG
TAAGCCCACTTGACGGTGAAGCAATCAAAACGTTATTCTCTATTAAGTGT
ACTAATGAGAATGCTTTATATAAGCCTTTGAAGTACTGTTTAAGCGTTAA
TAGCTTGTTAATTACAGAATGTATTACTGATCAGTTGATTATTTCACGAT
TGCCTGCAGCTTCACATATAGAGGTCCAGGTGTGCGATAATTTGGATGTA
TGCATCATTAAAAATATTGGTGTAAATGTACGCCAAGCTTTCATAGCAGA
CACAGAGGTTGCAGTTTTTAATTTTCTCAATCGTGTTCGCTATTGGCTAA
AATATGCTGACTGGTCCATGTCTTTCGTAATGCTTGATCAGATAATACTT
CGAATAAATTCAAAGGAGCGTCTTCTTGTTTTCACAGAGGCACTCGCCGT
GTATCAGCCACAAACTCCTGTACAATTAGCTCACCTTGTACAGATTTTAA
TTAAAATATTGGATCATATTATACCTTTGGATGATATGATAGTCAATGTG
CTTGCTCGAATTCTTCATATTATAGAATCAACTTTCAAAATAGTTATTGC
TAATGCTGAGCAAAATACACTTATGGATTGCTACTATAAAAAGATGGTGG
AAGCTATGTTTGAAATTCTTGATAAATTTGCTACTGAATGGGAGTATATT
CCGAAATCGCAGTGCAGAGCAGAATCTGAGTCCTGCCTAAATGTAGACAA
CTTTCGTAATAGACTAGAACAAATGTCTACCTTGAATCCACAGGTACTAG
AACATATAAATAACTGGTTGCATGCACATTGGAAACTGAATAATTGTTTA
TTTTATATGGGTATGGGCATGGCGCGACGCATACATCCAGATGAAAATGC
AAAAAAAAAAGACTGTCGAACCTTTTTAATGACAATGGAGAGTTTTGACC
TAGATTTAGAGCGCGCATTGGTGATTAAATCCGCAGATGCGATGCATACA
CTAATTTTTACAGAAAAGTTGTTGCAAGAGTTGCGTCACTTACTGCGAAA
TGATGAAGTTCTTATCTCTATTAGAAGCCATAAACATAGTCAGTACTGGT
GGTATCCTGAGCAAGATTCACGGACACAAGTACTAGTTGTAAATGCTTAT
ACATCTACAGCTATCTTGGAGAAAACACAGCAGCTGTCTGAGCCATTTCA
ATATATATCCAAACTCACAACTAAATCCTGTACTTATCAATATTGTAAGG
GTCGACGTCGTAGACGAAAGAGTATTGAATTTGATGATCCTGTAGAAGAT
GTCGAGAATGTCATACACGATAACGTTGTAAGTTCCAAAGAGGTTCGCAT
GTATCGTACAGAACTTTATGGCCATTCTGTGCTTGGGGTTACTTTTACAA
AGGCTGATATAGACTACCGTGTTTTACTTCATATGACAAACAACCCACAG
CTGAATGATATTGAAACAAAAAATACAACGTGCTTAGTGAAATCGGGAAG
TAAACCAACAGCTTTATTACTTCGAAATTTATGTAAGGAAGCACGACCAG
TCTATATATATATTCGCGCTGAAAATATATTAAGGAATTGGGATTCGTTT
CGCGAAACAGGTGCTTATTATACATTCTCCACTGAAATGCGTAGTTGTCG
AATTTGGAAATTTGCACGGCCAGAACCCAGCTGGCAGACTATTCTGTGCA
TTCCCGAAATGAATAAAAGTGTAAGCAATGGTATCCACTGCCGCTGTAAT
TTCATTAGCGACCTTGACGCTGATGCCAAACCAATTATTGCTGTCCGAAT
GAATCTTAAATGCCACTTGGAGCGCCCAGTGGTCGGACGTAATTACGAAA
TTATCATTTGCTATATTGTAATCCCAATAGTGGCAATAGCGTTTCTTTTA
ATTCAAATGCACCGAGCAGCGTTCTGGGACAAGCCCTTATATTTGGAGGA
CGTATACAGTGGCGAGCTTTGTCGCTGTGGTGACATAATAATCCGAATTT
CATTTGGTGGTCGTTATCACTCTGGATCTTCAGCAAACATATTTCTTTTA

TTGCAGTCTTCTCGAGGCAAAAGGGAGATATTCGTTCATCAAGATCCAGT
TAAAACGGCTTTTAATCGCAACTGCACGATTTTTTTGCGACTTGACAGAG
AATTTGTGCAGTTGCCAGTGCGTCTTGCTCTTGGTCATGATAACACTGGC
ACTCATCCACATTATTTTTGTCGCAGCATAGTTATTACTGACATATTAAC
GGAGAAAACACAACACTTTCGTATTAATCGGTGGGTGCGTACATCGCCTG
GGGCAGGAAGTAAAATGCATTTGGAGTCAACTATGGTTCTGGATTTCGCA
ACAACTTCTCCAAAGTCTATATATCGATGGCCATCCCGATTCGCGATTGC
TTTTGAGCTATGCATGGGAAAGTGGTACTTATTTCAATCTATAATAGGTC
CATGGCGTTTCGGAATAAATCGCAATTCCTTAAGCAGATGGGAACGTAGT
TGTATATATGTGGGTAAAAATTTTGTAGCAATATGCATAGTTATCATATT
CTTTGGACGTGCCGAGCCAATATTATGTGACCCAAGTCCAAAACGGTATA
ATGATTTTAATATTGTGGTCTGGCTTTGTCTTATATGTTTAACTGCAAGC
TGTATAACTGAAATTCTTATGATTATATTTCTTAGAATGTTATCAAAATA
TAATTAG

Protein Sequence:

>Dmoj2_fosmid9_CG30048-PA_NP_725155_pep
MRVNQFLVLISLANVTAFIRYDNPYVFVNRVYVNHTIFVIVTLQSLRGVW
FLTKVSIEERIIDYFQTNASEHREQLLWYHTASLKQTLNFYGIFTQKNLP
YQECFYHVSRANFFKSGFHNVSVLITTIGITSDKLRSSSYTYIDLDENVR
CEPQFSIPQCVNADYPVHYEVSDIIVLRAMYTRRPAFTIPPFTQQGEQSI
RFILNIRSTFDLLRSSVAMQTILFSSYPRLYIDVIIKCVTNCKGNRFSNY
SHVHLKAHCLSCKSMKITKWMWTVDHNLVSSSKRLLYYVSNKTSIIVYLN
MEAIHRYKPSSKYQGSSTLRLERNYGPVHTACYVSPLDGEAIKTLFSIKC
TNENALYKPLKYCLSVNSLLITECITDQLIISRLPAASHIEVQVCDNLDV
CIIKNIGVNVRQAFIADTEVAVFNFLNRVRYWLKYADWSMSFVMLDQIIL
RINSKERLLVFTEALAVYQPQTPVQLAHLVQILIKILDHIIPLDDMIVNV
LARILHIIESTFKIVIANAEQNTLMDCYYKKMVEAMFEILDKFATEWEYI
PKSQCRAESESCLNVDNFRNRLEQMSTLNPQVLEHINNWLHAHWKLNNCL
FYMGMGMARRIHPDENAKKKDCRTFLMTMESFDLDLERALVIKSADAMHT
LIFTEKLLQELRHLLRNDEVLISIRSHKHSQYWWYPEQDSRTQVLVVNAY
TSTAILEKTQQLSEPFQYISKLTTKSCTYQYCKGRRRRRKSIEFDDPVED
VENVIHDNVVSSKEVRMYRTELYGHSVLGVTFTKADIDYRVLLHMTNNPQ
LNDIETKNTTCLVKSGSKPTALLLRNLCKEARPVYIYIRAENILRNWDSF
RETGAYYTFSTEMRSCRIWKFARPEPSWQTILCIPEMNKSVSNGIHCRCN
FISDLDADAKPIIAVRMNLKCHLERPVVGRNYEIIICYIVIPIVAIAFLL
IQMHRAAFWDKPLYLEDVYSGELCRCGDIIIRISFGGRYHSGSSANIFLL
LQSSRGKREIFVHQDPVKTAFNRNCTIFLRLDREFVQLPVRLALGHDNTG
THPHYFCRSIVITDILTEKTQHFRINRWVRTSPGAGSKMHLESTMVLDFA
TTSPKSIYRWPSRFAIAFELCMGKWYLFQSIIGPWRFGINRNSLSRWERS
CIYVGKNFVAICIVIIFFGRAEPILCDPSPKRYNDFNIVVWLCLICLTAS
CITEILMIIFLRMLSKYN*

CG31999-PA:

DNA Sequence:

>Dmoj2_fosmid9_CG31999-PA_NP_726551_cds
ATGCGTGCGATGAAGTTAAATATAATTGGCGTTATACTTTGTCTTATATT
CAGTTATTTGGATAGAGTGTTGGCAGACGAACAAATCGCTGACTATATTC
GAAAATGTTGTATAAGTGGCTTGCGTAATGCTCGTACCACAAGTGTATGC
GATAAAATGGATTCAACTATAGTAAATATATCTAATCTTTGGCTTGGATT
GTGCTCCTCGACATTTGGCGTTTGCTGTTCAAAAGAGTTGGACCGTCAGA
ATTGTGAGCTCGGTCGTTTGGCTGCATTGGAGGGTGCTTCGTGTAATAAA
GGCTTTAATTTGACGTCAACCTCGTATACAAATTGCTGCCGAGCTTGTCA
AGTTGGTTTAGCTGTTAAGGCTAGTCAACAAAAGTGTACGGATCCATTGT
TTTCGTTTCTTTTGAATATCGATTCATATCGGTTATGCTGCTCTGATGAC
GGCTTCTCAAATTCTGAGTTGGAATCGGAGTCGGGGTTGGAGGGAAAAAA
TAAACTCGTATTATATACAGATCATGATGAACGAGAGGAGGAAATTGTTG
AGAGCCGCGAAACTGTTGATGGAACAATAGTCCTATCCGGTGATGATGAT
ATCTGTGGAAAAATACCCAATCTTTGTGAGCAAGTATGTATAAACACATA
CGATGCGTACAGGTGCAGCTGTCACCCCGGGTATAAATTGAACGACAATA
ATGTTACCTGTTATGCGGATAAAAACAATATTTGTCCCAGTGGCTACGTT
TTGGATGGTAATCAAGGTAAATGTGTAGACATCGATGAGTGTCAAGAGCA
ACTACATGACTGCAAAACATCGCAATATTGTCACAATACAATCGGAGGAT
ATCACTGTCTCAACATAAAGGCTAAGAATTGCCCAGCAGGCTATTTATAC
AATGTGAAGTCAGATGAATGCGAAGGTA

Protein Sequence:

>Dmoj2_fosmid9_CG31999-PA_NP_726551_pep

MRAMKLNIIGVILCLIFSYLDRVLADEQIADYIRKCCISGLRNARTTSVC
DKMDSTIVNISNLWLGLCSSTFGVCCSKELDRQNCELGRLAALEGASCNK
GFNLTSTSYTNCCRACQVGLAVKASQQKCTDPLFSFLLNIDSYRLCCSDD
GFSNSELESESGLEGKNKLVLYTDHDEREEEIVESRETVDGTIVLSGDDD
ICGKIPNLCEQVCINTYDAYRCSCHPGYKLNDNNVTCYADKNNICPSGYV
LDGNQGKCVDIDECQEQLHDCKTSQYCHNTIGGYHCLNIKAKNCPAGYLY
NVKSDECEG

CG5262-PA:

DNA Sequence:

>Dmoj2_fosmid9_CG5262-PA_NP_649223_cds
GTGGGATTTATTTTCATATTCAATATTATTGTCGGAACTGGAGCGCTAAC
TCTGCCAGGAGTATTTGCCAAATCAGGATGGTGTCTAAGCTTTGTTGTGT
TAATATTATTAGCTCTTGTAAGCTATATCACAGTAACATTTGTTATCGAA
GCCATGGCGGGTGCGAATGCGATAAAAAACTGGCAAAGTCTTCAAGCTTT
ACGTTACAAGCGCAATTCAAATGAAGAAGATAATGAATTATATTGGAATA
CAGAAAGCGAAACCGTTCCCTTGACAATTCAGAATATGCAATTTCATTAC
TATCAACTGACACAAAAGTTTGAATTGGGTGAAATGGCAAAAATATTTTT
TAATGATTGCGGGCGTATATTGTTTTACCTATGCTTTATAATTTACCTAT
ATGGAGACTTGAGCATATATTCCGCAGCGGTGGCTAGAAGTTTGCGTGAT
GTGCTATGCGAACATGATCATTCAAATAACTCGGATATCTCGATTTCTGA
ATTAAGCTCAATGCGTTTGCTTGGCGGGTTGAAAGATAATGATAACCGTA
CATGCTGGAAAGAGCATAATGTATCACGTCTAATAGTGTACAGGATCATA
TTGATTGGATTCACTATGCTTTTTGGACCACTCGTTTTATTCAGCATACA
GAAAACAAAGTATTTGCAAATGCTTACTGTTGTTTTTCGATGGTTGGCGT
TTATTTTAATGATATGCATTGCTCTCAAAATGCTCGTCATTGATGGGGTA
AAAGGACATCCAGTTGCCGTAAATTTTTATGGTATTCCCTCATTGTTTGG
CGCCGGTGTTTACTCGTTCATGTGCCACCATTCGTTGCCCAGTTTATTGG
CACCAATTAAACATAAGTCAATGGCTAAAAAAATTCTTTCATACGATTAC
ATTTTAATTTGTTCGTTTTATATAGTTCTGGCGATTACAGGAATATTTGC
ATTTGAGTACGTTGAAGATCTGTATACTTTGAATTTTCTACCATATCACG
CTACATCTGAAAGTTTATTTTCAAATTTTTTAATTGTTATCGATTATTTC
CTGTCTTTGTTTCCGGTATTTACGTTATCCACTAGTTATCCACTTATCGC
TATCACACTTAAAAACAATCTTCAAACTTTGTTTTTGGACATGTCCCAAT
ATAATTCGTATAGCATATTGATTCGAGCACTTTTCCCCTTTTTATCTATT
CTTTTACCATTTTGTATAACATATTTCACTGAAAACCTATCTATCCTAGT
GGTTATTACGGGGAGCTATGCTGGAGTAGGTATACAATATGTCATTCCTG
TTTGCTTGGTCTATTATTCCCGTGTGACATGTTCTGAACTACTAGGAAGT
GGGATTAAAAATCAATTTCAAAGTCCATTTCAATCAAATATATGGCTATT
AATGGTGTTAGCTTGGTCAGTATTATCCGTATTCTTAGTGTCCATTAATT
TGTTCAGGTAA

Protein Sequence:

>Dmoj2_fosmid9_CG5262-PA_NP_649223_pep
VGFIFIFNIIVGTGALTLPGVFAKSGWCLSFVVLILLALVSYITVTFVIE
AMAGANAIKNWQSLQALRYKRNSNEEDNELYWNTESETVPLTIQNMQFHY
YQLTQKFELGEMAKIFFNDCGRILFYLCFIIYLYGDLSIYSAAVARSLRD
VLCEHDHSNNSDISISELSSMRLLGGLKDNDNRTCWKEHNVSRLIVYRII
LIGFTMLFGPLVLFSIQKTKYLQMLTVVFRWLAFILMICIALKMLVIDGV
KGHPVAVNFYGIPSLFGAGVYSFMCHHSLPSLLAPIKHKSMAKKILSYDY
ILICSFYIVLAITGIFAFEYVEDLYTLNFLPYHATSESLFSNFLIVIDYF
LSLFPVFTLSTSYPLIAITLKNNLQTLFLDMSQYNSYSILIRALFPFLSI
LLPFCITYFTENLSILVVITGSYAGVGIQYVIPVCLVYYSRVTCSELLGS
GIKNQFQSPFQSNIWLLMVLAWSVLSVFLVSINLFR*

Or13a:

DNA Sequence:

>Dmoj2_fosmid9_Or13a_NP_523359_cds
ATGTTCAATCCACTACCGTACAAAGATCCGAGCTTTCGCATGCCATTTCA
GTGCATTTGGCTAAAATTAAATGGCTCTTGGCCTTTGGAAATTAAATTAG
CAAACCAAAAGCTACTTACACGAAGCTATCTCTTTGCATTCATCTACAAT
GTATGGGCTTTGTATGTTATTATATCGGTTGGCATTACCATTAGCTTTCA
AACATCATTTTTGATTAACAATTTCGGAGATATAATAATGACAACTGAAA
ATTGCTGCTCTACGCTGATGGGCGCCCTTAATTTTGTTCGACTAATTCAT
TTACGTGTTAATCAGCCAAAATTTCGTAAACTAATTAATCAATTTGTGCT
TAATATATGGATACCAGAGTA

Protein Sequence:

>Dmoj2_fosmid9_Or13a_NP_523359_pep
MFNPLPYKDPSFRMPFQCIWLKLNGSWPLEIKLANQKLLTRSYLFAFIYN
VWALYVIISVGITISFQTSFLINNFGDIIMTTENCCSTLMGALNFVRLIH
LRVNQPKFRKLINQFVLNIWIPE