Annotation of *Drosophila mojavensis* Fosmid 18

Bill Dirkes

4/20/2008

Overview

Annotation is a process of analyzing the genome sequence of an organism. This includes identification of putative protein-coding sequences, pseudogenes, and repeats. Annotation also includes an analysis of synteny with another species. In this paper I will be presenting my annotation of fosmid 18 in *Drosophila mojavensis*. On this fosmid, I have identified four functional proteins (Figure 1). Three are in the same relative orientation as their orthologs in *Drosophila melanogaster*. The ortholog of the fourth gene is on chromosome 2L in *D. melanogaster*. In terms of repeats, the fosmid is approximately 29% repetitious elements, with the majority of these being DNA elements. Finally, a conserved region not within a repetitious element was identified containing numerous stop codons in all open reading frames. This suggests the element serves some functional role, but is not protein-coding.

mGluRA: 2021-8793
CG32016: 15917-10881
CG11093: 27985-23433
CG5367: 36693-38152 (missing first exon)

Table 1: Summary of genes annotated on fosmid 18.

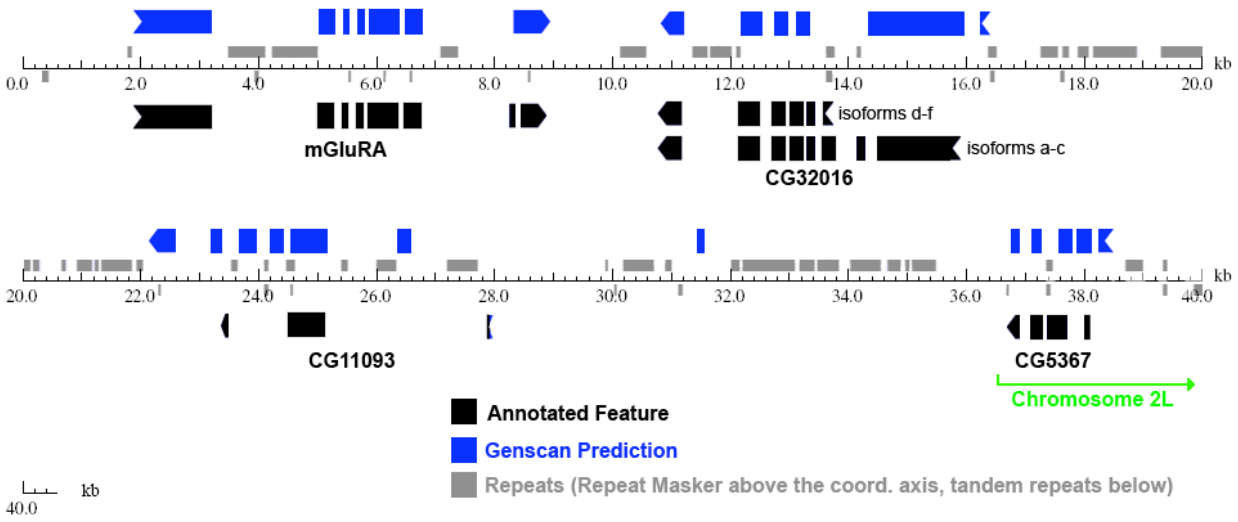


Figure 1: Final feature map.

Genscan Predictions

To begin my analysis, I ran Genscan to identify putative protein-coding sequences. Genscan predicts three features in this fosmid (Figure 1). All three were predicted to have six or more exons. I refer to these features in the order listed in Table 1.

Before running any BLAST searches, I examined my fosmid using the mirror of the UCSC Genome Browser available at gander.wustl.edu (Figure 2). This showed four Refseq genes having high sequence similarity to fosmid 18. In addition, the conservation track shows

numerous regions of high sequence conservation that largely correspond to the Refseq gene matches.

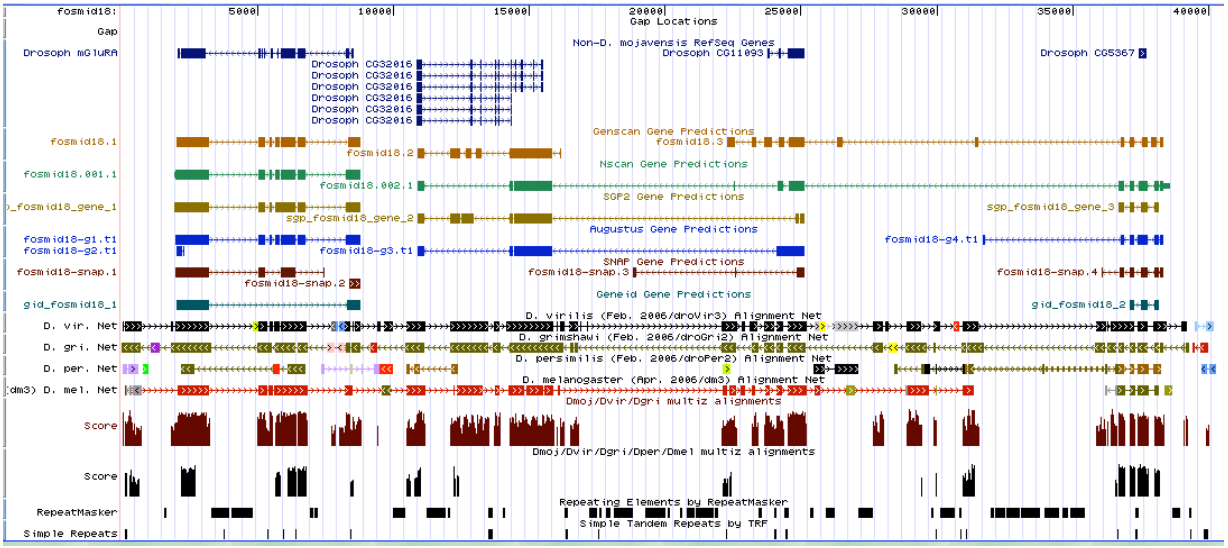
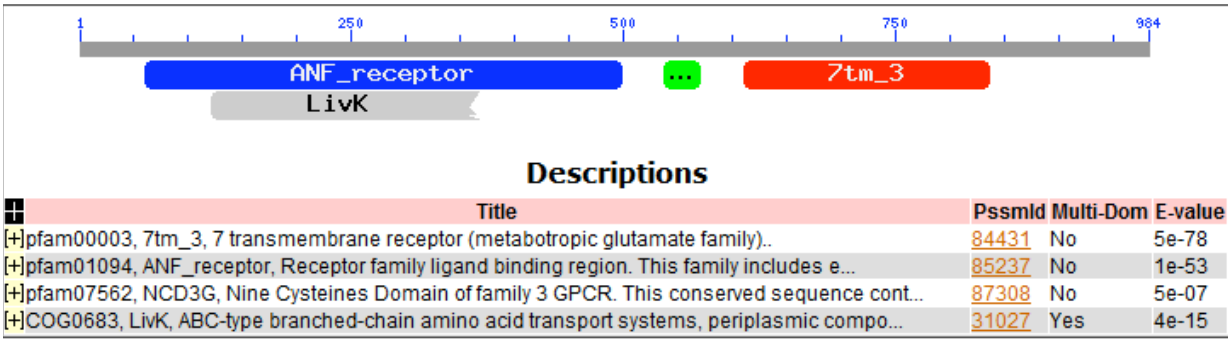


Figure 2: UCSC Overview of Fosmid 18.

Feature 1

I aligned the first peptide sequence prediction from Genscan against the nr database using blastp. BLAST identified four conserved domains, including a 7 transmembrane receptor (Figure 3). The best match was to a protein named *mGluRA* found on the fourth chromosome in *D. melanogaster* and had an expect value of 0.0. This protein is a glutamate receptor precursor with only a single isoform.



Title	Pssmid	Multi-Dom	E-value
[+]pfam00003, 7tm_3, 7 transmembrane receptor (metabotropic glutamate family)..	84431	No	5e-78
[+]pfam01094, ANF_receptor, Receptor family ligand binding region. This family includes e...	85237	No	1e-53
[+]pfam07562, NCD3G, Nine Cysteines Domain of family 3 GPCR. This conserved sequence cont...	87308	No	5e-07
[+]COG0683, LivK, ABC-type branched-chain amino acid transport systems, periplasmic compo...	31027	Yes	4e-15

Figure 3: blastp of Genscan's predicted peptide to nr reveals numerous conserved domains.

A tblastn alignment of the *mGluRA* protein in *D. melanogaster* to the fosmid resulted in a nearly complete match excluding only the end of the protein (Figure 4). I then ran the same search using each exon individually. All exons except the last align with a low E value and high coverage. In contrast, the last exon contains a large gap and only two-thirds coverage of the real protein (Figure 5). Particularly, a region between about 8410 and 8500 in the fosmid sequence appeared to contain a low complexity repeat. Tandem Repeat Finder confirmed that this region is a

repetitious element (Figure 6). Thus, it appears that a tandem repeat inserted into the coding sequence, splitting the exon into two separate exons.

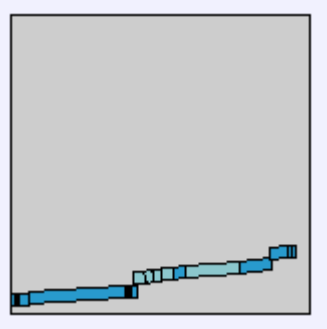


Figure 4: Dot plot of the mGluRA protein (x-axis) against fosmid 18 (y-axis).

```
 Score =  109 bits (272),  Expect = 3e-22
 Identities = 64/111 (57%), Positives = 69/111 (62%), Gaps = 30/111 (27%)
 Frame = +1

Query  1     VQTTTLCISISLSASVALVCLYSPKVYILVFHPDKNVRKLTMNSTVYRRSAAAVA-----  55
             +Q TTLCISISLSASVAL+CLYSPKVYILVFHPDKNVRKLTMNSTVYRRSA   A
Sbjct  8245  IQITTLCISISLSASVALICLYSPKVYILVFHPDKNVRKLTMNSTVYRRSATTGAGQVAG  8424

Query  56    ----------------------QGAPTSSGYSRT----HAPGTSALTGGAVG  81
                                   QG PTSSGYSRT    ++      TGG +G
Sbjct  8425  TGTGTGTGTGTGTGAGAATGIQGVPTSSGYSRTPIVLNSEDACYGTGGGIG  8577
```

Figure 5: tblastn alignment of the last exon in *mGluRA* (query) against fosmid 18 (subject).

I identified the best canonical splice site donor and acceptor for the exons flanking the new intron based on the conservation track along with the tblastn alignments previously described. After identifying the best canonical splice site donor, I was able to more precisely pick the best canonical splice site acceptor based on the phase necessary for the exon.

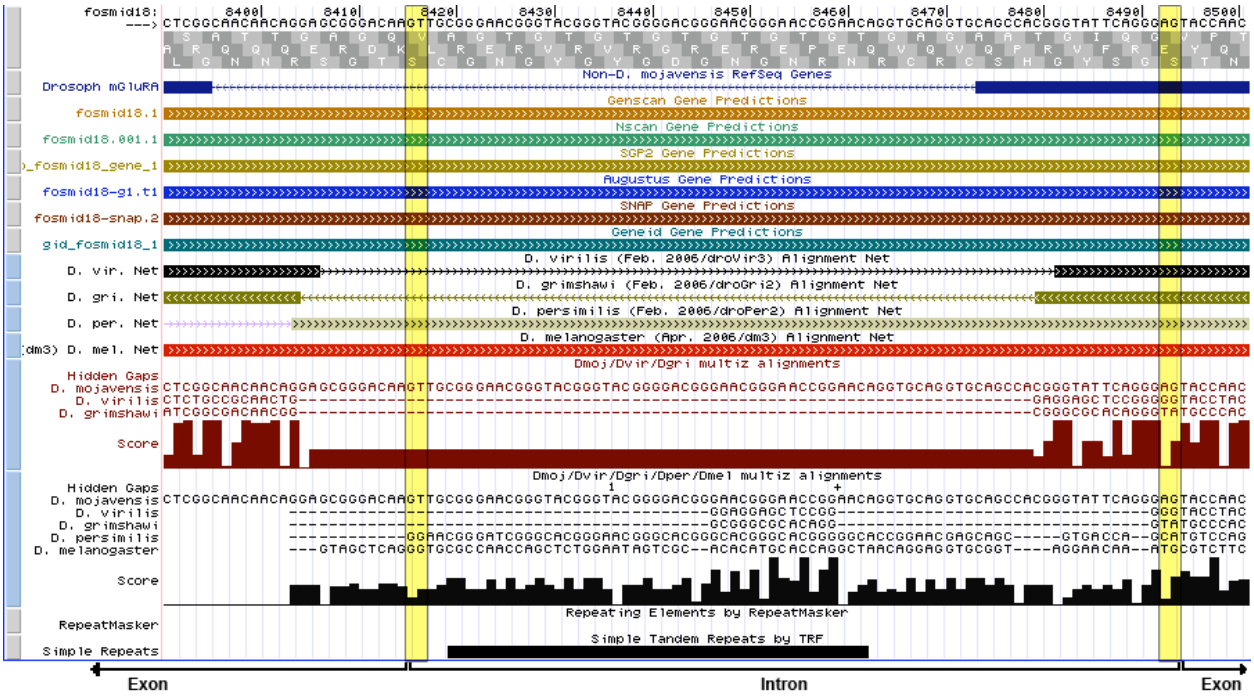


Figure 6: UCSC shows a simple repeat in the middle of the predicted protein sequence (bottom track). Canonical splice site donor and acceptor highlighted for the revised gene model. Note that while numerous other canonical donor and acceptor sites are in the region, only these two maintain the correct phase and frame for the new final exon. Diagram of predicted gene structure in the region is at the bottom.

The final gene model has strong conservation for all exons with the exception of the last exon I identified. Length of the protein was conserved, with my gene model having 984 residues and the ortholog having 976 amino acids.

Feature 2

As a first step in my analysis, I ran a blastp search of the second peptide sequence from Genscan against the nr database. The best alignment was to *NM_166797* (*CG32016* in FlyBase), which is found on the fourth chromosome in *D. melanogaster*. This gene has 6 isoforms (Figure 20); however, only two isoforms are unique based on the coding sequence (all other variants have differing UTRs). Isoforms a, b, and c have the same protein-coding exons. Isoforms d, e, and f have a different start codon which is internal to one of the exons shared by isoforms a-c. All isoforms share the last three exons.

I ran a tblastn alignment of each exon in isoform b of *mGluRA* in *D. melanogaster*, the longest isoform, to fosmid 18. The first exon only contains the start codon (M) and thus no alignment is possible. Nevertheless, a start codon was found at the beginning of the second exon in *D. mojavensis*. Because conservation is low between the two species and there is no means for determining the UTRs in *D. mojavensis*, I annotated the gene starting at the second exon. I will refer to each exon based on the ortholog numbering in *D. melanogaster* for this paper.

The second, sixth, and seventh exons map to regions with approximately 35% identity. Exon three does not have any strong matches; the best match has an expect value of 1675 and aligns only the first 5/8 of the sequence. For exon four, the last half of the alignment is in a repeat, resulting in a run of glutamines in the protein sequence (Figure 7). Thus I believe this exon was divided into two. Exon five has 50% identity but only matches 2/3 the length of the exon. The last exon (exon eight) has 60% identity and matches the entire length.

```
 Score = 61.6 bits (148),  Expect = 6e-08
 Identities = 47/128 (36%), Positives = 64/128 (50%), Gaps = 14/128 (10%)
 Frame = -1

Query  2      IPSVKDLEAQMIKVDMRT--DLINPIAGS-----LCQTVQMEKPIARDTEAFKKLLQQLG  54
              IPSVKDLEAQM KVD+R   D+    A +        T Q +K + RDTEAFK+LLQQLG
Sbjct  13905  IPSVKDLEAQMRKVDLRPECDISKSQANAHQPQQTVHTEQRDKSVTRDTEAFKRLLQQLG  13726

Query  55     SQARQHHPCNDDCRTINLSNIANHVHLESKLHQKINDGHLQQPELSVNVPTMPTSSHVFL  114
              SQ       N+   T  +++    V +  +L Q+    H QQ +             H
Sbjct  13725  SQ-------NNTESTYPITSNMPAVRVVPRLQQQQQQLHQQQQQHQQQQQQQQQQQHHQH  13567

Query  115    QKRLEIQH  122
              Q++ + QH
Sbjct  13566  QQQQQQQH  13543
```

Figure 7: A tblastn search of exon 4 of *CG32016* in *D. melanogaster* against the fosmid 18 sequence aligns partially to a repetitious element in the fosmid.

Despite the poor quality alignment beyond 13719 in the subject sequence, there are no canonical splice site donors until 13675, which is just before the repetitious element. Thus, I annotated this exon end coordinate as 13675.

```
 Score = 19.6 bits (39),  Expect = 268493
 Identities = 12/32 (37%), Positives = 17/32 (53%), Gaps = 2/32 (6%)
 Frame = -1

Query  9      LQQPELSVNVPTMPTSSHVFLQKRLEIQHLIQ  40
              + Q +L  NV  +  S      KR+EIQ L+Q
Sbjct  13407  MPQNQLHPNVHELNISLQT--HKRMEIQQLLQ  13318
```

Figure 8: tblastn of the unaligned half of exon 4 of *CG32016* in *D. melanogaster* to fosmid 18.

I then ran a tblastn search of the unaligned portion of exon 4 of *CG32016* in *D. melanogaster* against fosmid 18. The alignment matched the last three-fourths of the protein sequence with 37% identity. Then, I used the UCSC genome browser to identify the best canonical splice sites based on conservation of length of the exon. Thus, the third and fourth exons in the gene I annotated probably arose from a repeat inserting into the fourth exon of the ortholog gene.

Of further interest, I found that exon 7 has conservation with *Drosophila* species beyond the Genscan prediction (Figure 9). Because the conservation and match to the exon extends beyond the coordinates predicted by Genscan, I used the coordinates predicted by the gene finder SGP2. These coordinates extend about 10bp beyond the tblastn alignment to the *D. melanogaster* exon. However, the conservation beyond the end of the Genscan prediction supports this model.

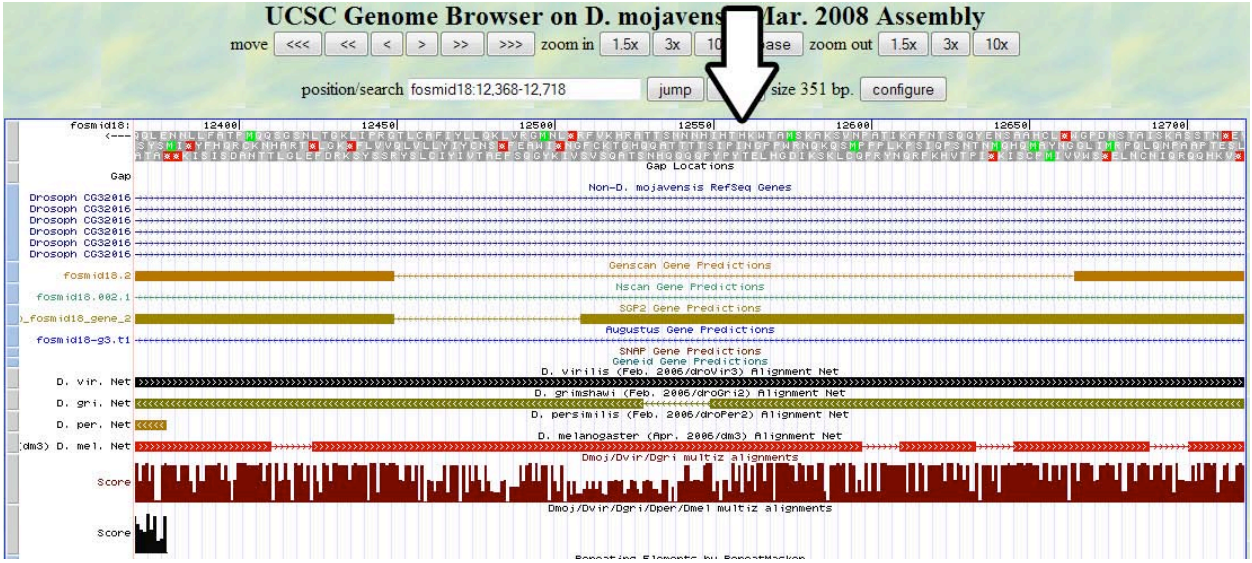


Figure 9: UCSC genome browser in the region of exon seven. Note that the tblastn search of the seventh exon (the arrow points to the left end of the alignment) extends beyond the Genscan prediction.

Feature 3

A blastp search of Genscan's third predicted peptide against the nr database identifies three conserved domains (Figure 10). The best match is to *CG11093* in *D. melanogaster* on the fourth chromosome. This gene has only one isoform and three coding exons. The alignment includes only to the first two-thirds of the first exon, but has 91% sequence similarity (Figure 11).

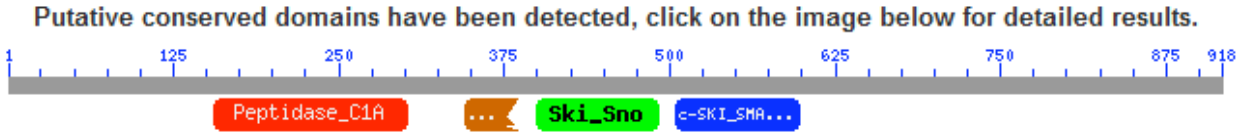
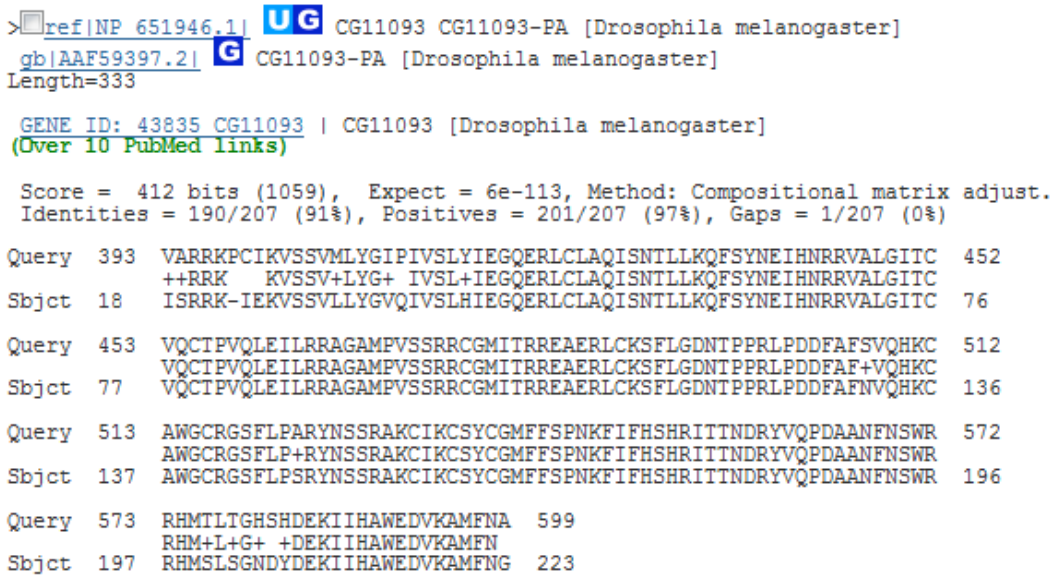


Figure 10: Three conserved domains identified by BLAST during blastp search of the third predicted feature against the nr database.



```
>ref|NP_651946.1| UG CG11093 CG11093-PA [Drosophila melanogaster]
 gb|AAF59397.2| G CG11093-PA [Drosophila melanogaster]
Length=333

 GENE ID: 43835 CG11093 | CG11093 [Drosophila melanogaster]
(Over 10 PubMed links)

 Score =  412 bits (1059),  Expect = 6e-113, Method: Compositional matrix adjust.
 Identities = 190/207 (91%), Positives = 201/207 (97%), Gaps = 1/207 (0%)

Query  393  VARRKPCIKVSSVMLYGIPIVSLYIEGQERLCLAQISNTLLKQFSYNEIHNRRVALGITC  452
            ++RRK   KVSSV+LYG+ IVSL+IEGQERLCLAQISNTLLKQFSYNEIHNRRVALGITC
Sbjct  18   ISRRK-IEKVSSVLLYGVQIVSLHIEGQERLCLAQISNTLLKQFSYNEIHNRRVALGITC  76

Query  453  VQCTPVQLEILRRAGAMPVSSRRCGMITRREAERLCKSFLGDNTPPRLPDDFAFSVQHKC  512
            VQCTPVQLEILRRAGAMPVSSRRCGMITRREAERLCKSFLGDNTPPRLPDDFAF+VQHKC
Sbjct  77   VQCTPVQLEILRRAGAMPVSSRRCGMITRREAERLCKSFLGDNTPPRLPDDFAFNVQHKC  136

Query  513  AWGCRGSFLPARYNSSRAKCIKCSYCGMFFSPNKFIFHSHRITTNDRYVQPDAANFNSWR  572
            AWGCRGSFLP+RYNSSRAKCIKCSYCGMFFSPNKFIFHSHRITTNDRYVQPDAANFNSWR
Sbjct  137  AWGCRGSFLPSRYNSSRAKCIKCSYCGMFFSPNKFIFHSHRITTNDRYVQPDAANFNSWR  196

Query  573  RHMTLTGHSHDEKIIHAWEDVKAMFNA  599
            RHM+L+G+ +DEKIIHAWEDVKAMFN
Sbjct  197  RHMSLSGNDYDEKIIHAWEDVKAMFNG  223
```

Figure 11: blastp of the 3rd Genscan peptide against the nr database shows strong conservation.

Blat of the real protein sequence against my fosmid also reveals high conservation, but only one alignment (one exon). Regions on both sides also contain conservation. In addition, the gene predictors are all identifying different gene models in the region.

Neither the first nor third (last) exon in the real gene match the unmasked fosmid sequence using a tblastn search, even when the maximum expect value is raised to one million. However, after raising the maximum E value further, I found numerous matches. Because the previous tblastn of the second exon showed a strong alignment on the minus strand around 25kb, we know the first exon must occur at a position >25kb on the minus strand, and the 3rd exon must occur at <25kb on the minus strand. The best alignment for the 1st exon is E>14,000,000 at about 38kb and the best alignment for the third exon is E=3,700,000 at 19,400. There is no conservation around the 3rd exon. The third exon, based on the first upstream splice site acceptor and the first downstream stop codon, is larger than in *D. melanogaster*.

Tblastn of the second exon in *CG11093* to the fosmid sequence reveals a high degree of similarity for the first two-thirds of the protein (Figure 12). The remaining region contains an N repeat and is not strongly conserved. When annotating this region, I found donor and acceptor splice sites near the coordinates of the tblastn alignment ends. Based on the evidence of conservation and the canonical splice sites corresponding to this region, I was able to define the most probable coordinates for my annotation.

```
 Score =  432 bits (1112),  Expect = 6e-119
 Identities = 222/311 (71%), Positives = 245/311 (78%), Gaps = 30/311 (9%)
 Frame = -1

Query  1      VSSVLLYGVQIVSLHIEGQERLCLAQISNTLLKQFSYNEIHNRRVALGITCVQCTPVQLE  60
              VSSV+LYG+ IVSL+IEGQERLCLAQISNTLLKQFSYNEIHNRRVALGITCVQCTPVQLE
Sbjct  25128  VSSVMLYGIPIVSLYIEGQERLCLAQISNTLLKQFSYNEIHNRRVALGITCVQCTPVQLE  24949

Query  61     ILRRAGAMPVSSRRCGMITRREAERLCKSFLGDNTPPRLPDDFAFNVQHKCAWGCRGSFL  120
              ILRRAGAMPVSSRRCGMITRREAERLCKSFLGDNTPPRLPDDFAF+VQHKCAWGCRGSFL
Sbjct  24948  ILRRAGAMPVSSRRCGMITRREAERLCKSFLGDNTPPRLPDDFAFSVQHKCAWGCRGSFL  24769

Query  121    PSRYNSSRAKCIKCSYCGMFFSPNKFIFHSHRITTNDRYVQPDAANFNSWRRHMSLSGND  180
              P+RYNSSRAKCIKCSYCGMFFSPNKFIFHSHRITTNDRYVQPDAANFNSWRRHM+L+G+
Sbjct  24768  PARYNSSRAKCIKCSYCGMFFSPNKFIFHSHRITTNDRYVQPDAANFNSWRRHMTLTGHS  24589

Query  181    YDEKIIHAWEDVKAMFNGGTRKRLV--------GSSINPRSPGSPTQSTAVSSGVGSSSC  232
              +DEKIIHAWEDVKAMFNGGTRKRLV         S+ N  +  S + + +  S  G+++C
Sbjct  24588  HDEKIIHAWEDVKAMFNGGTRKRLVSCNNNNNNNSNNNNNNNNSTSNNNSSRSNSGNTNC  24409

Query  233    ----------------GSPTSFERDTNEVDGKSGSSLPQRSQIDSQYNYGTVAAAAAV--  274
                                S  S E  T   +    ++   +  ++ QYNY TVAAAAA
Sbjct  24408  NPNQNSPALDQTPIHNSSSISRETKTTLEETLCTTTTMSKRPMNQQYNYSTVAAAAAAAV  24229

Query  275    ----VGVTAVA  281
                  VGVTAVA
Sbjct  24228  VGVGVGVTAVA  24196
```

Figure 12: tblastn of the 2nd exon from the ortholog to the fosmid sequence.

C. Schaffer annotated the gene using gene predictors for guidance, but had no other evidence supporting his predictions for the first and third exons. In addition, he also annotated the N repeat in the second exon as an intronic region, rather than as part of the exon. The predicted coordinates for the second exon are based on the tblastn alignment, a comparison with the gene predictors, and nearby canonical splice sites.

```
 Score = 18.9 bits (37),  Expect = 456687
 Identities = 9/15 (60%), Positives = 9/15 (60%), Gaps = 0/15 (0%)
 Frame = -3

Query  65     AAAAAVVGVTAVAAA  79
              A AAA    TA AAA
Sbjct  24070  ATAAATASATAAAAA  24026
```

Figure 13: After annotating the second exon, I used the remaining portion of the exon in the *D. melanogaster* ortholog and tried to align it to somewhere else in the fosmid. This is the best hit given that the next exon in the gene must be on the minus strand downstream of the second exon.

In addition, a tblastn of the unused portion of the second exon in the ortholog (the part that contains the repeat of Ns) did not align well to any other regions downstream of the second exon I annotated (E=456,687).

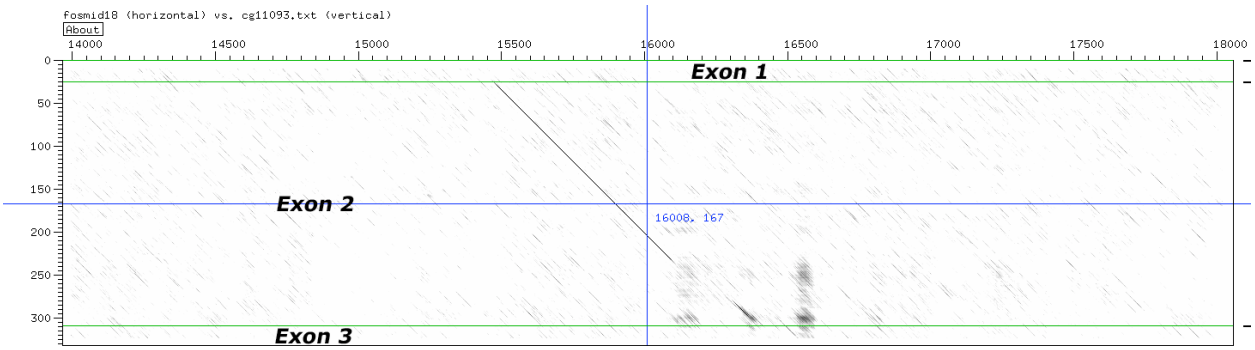


Figure 14: Dotter software aligning protein sequence against the fosmid sequence. The x-axis is the fosmid sequence, and the y-axis is the protein sequence. Green lines demarcate the exon boundaries of the protein. The solid diagonal line represents the region of the second exon with highest conservation.

Because manual evaluation of all tblastn matches between the first exon sequence and the fosmid sequence did not reveal any regions meeting the criteria described previously for a valid gene model, I used the Dotter program for identification of all alignments of the protein sequence to the fosmid sequence (Figure 14). This software creates a dot plot of my fosmid sequence against the protein sequence. The software is more powerful than BLAST because it generates all possible alignments. Even using this software, I could not identify any exons that aligned properly to the protein and contained a start codon. Thus, my prediction of the proper exon was based on the previously specified criteria. In addition, I assume that no conservation exists for the exon sequence, but that the length of the sequence is relatively conserved. Using these criteria, I identified a putative first exon at 27985-27908 (Figure 15).

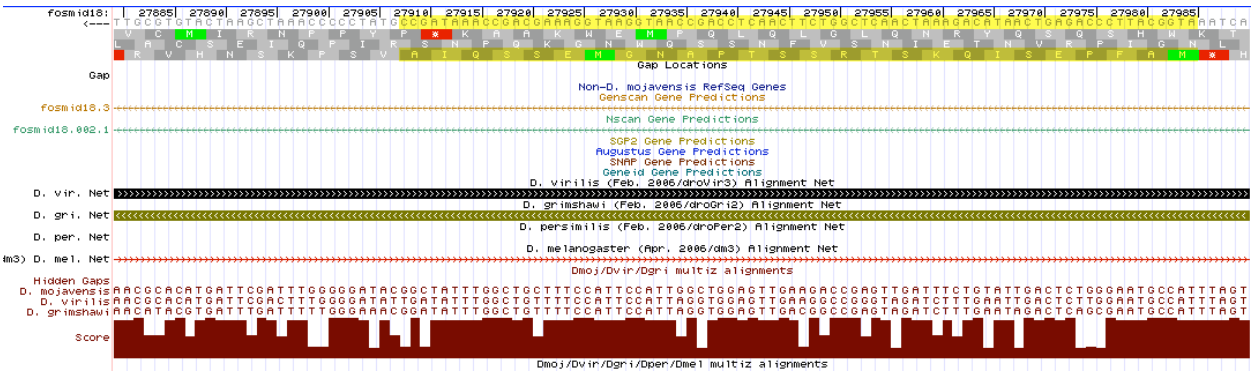


Figure 15: Putative 1st exon for CG11093.

Because this region had conservation with other *Drosophila* species, I extracted the DNA from the conserved region and ran a blastx search of this nucleotide sequence against the nr database. The only alignment is to an experimental protein in *D. pseudoobscura*. This alignment starts at the start codon I predicted, and ends shortly after my predicted splice site.

Dotter allowed me to find a potential final exon around 23kb. Upon analysis using the mirror of the UCSC Genome Browser, I found that a potential third exon exists at 23506-23433 (Figure 16). This exon contains some matched amino acids, a canonical splice site donor, a stop codon at the end, and conserves the length of the ortholog's exon.

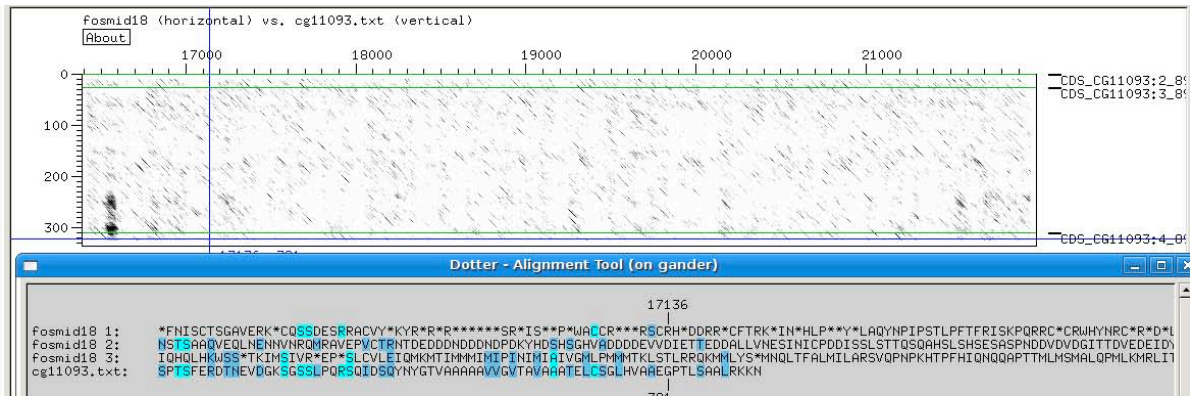


Figure 16: Dotter visualization of the alignment in the putative exon 3 region. The second frame of the fosmid sequence (fosmid18 2) has a long open reading frame with some conservation to the protein sequence (cg11093.txt) as evidenced by the colored boxes.

Feature 4

Unlike the other features, this feature was not predicted by Genscan. However, a Refseq alignment in the region matched to *CG5367* in *D. melanogaster*. I searched fosmid 18 using each exon from *CG5367* in a tblastn search. All exons except the first aligned with low E value, about 75% identity, and almost complete coverage. For the first exon, no matches could be identified even after raising the E value to 10 million. Thus, I searched the *D. mojavensis* scaffold for the last 1kb of DNA in my fosmid using BLAT at the gander server, and found an identical match to a region on fosmid 17. This indicates that fosmid 17 and fosmid 18 overlap.

I then ran a tblastn search of the first exon protein sequence for *CG5367* against fosmid 17. The best alignment ends at approximately 19121. Repeating the process previously described for exon 2, I found that the second exon starts at 25332. This indicates that the length of the intron is approximately 6.2kb, which is beyond the end of fosmid 18. Besides missing the first exon, the gene model is complete with canonical splice sites, a stop codon at the end of the sequence, and no stop codons within the open reading frame.

ClustalW of mGluRA

A ClustalW alignment of the *mGluRA* ortholog amino acid sequence from six species revealed relatively high conservation (Figure 17). The species included four fly species (including *D. mojavensis*) and mouse. The protein sequence was not highly conserved, however usually replacements with similar amino acids occurred. In addition, the ends of the alignment had much lower conservation.

```
D.melanogaster      ---------MKQKNNNGTILVVVMVLSWSRVVDLKSPSNTHTQDSVSVSLPGDIILGGLF 51
D.yakuba            ---------MKQKKKIANILVVVMVLFWSPVVDLKGPSNAHTQDSVSVSLPGDIILGGLF 51
D.grimshawi         MREAFKANSNNYGSNIPIRLMLMVVTLWSCLPQLSGAGSSQSHDSVSVFLPGDIILGGLF 60
D.willistoni        MREAFKANSNNYGSNIPIRLMLMVVTLWSCLPQLSGAGSSQSHDSVSVFLPGDIILGGLF 60
D.mojavensis        -------------------MLMSATCW--LSQLDIAGSTQVQDTVSVSLPGDIILGGLF 38
M.musculus          ------------MKMLTRLQVLMLALFSKGFLVSLGDHNFMRR--EIKIEGDLVLGGLF 45
                                        ::    :     :.  .    :   .: : **::*****
```

Figure 17: ClustalW alignment of the mGluRA protein sequence shows low conservation at the beginning of the protein.

Then, I used ClustalW to try to identify a 5' UTR for *mGluRA*. I used the UCSC genome browser to align the *D. melanogaster* nucleotide sequence for the first exon against the other species. I then extracted the DNA of the best alignment and 2kb upstream of the alignment. I included the DNA of the first exon because it will help anchor the alignment. Finally, I used ClustalW to align the nucleotide sequences. While the first exon is aligned properly, there are no stretches of conserved bases upstream of the coding sequence. Thus, the alignment does not provide any evidence of a UTR.

Analysis of Conservation

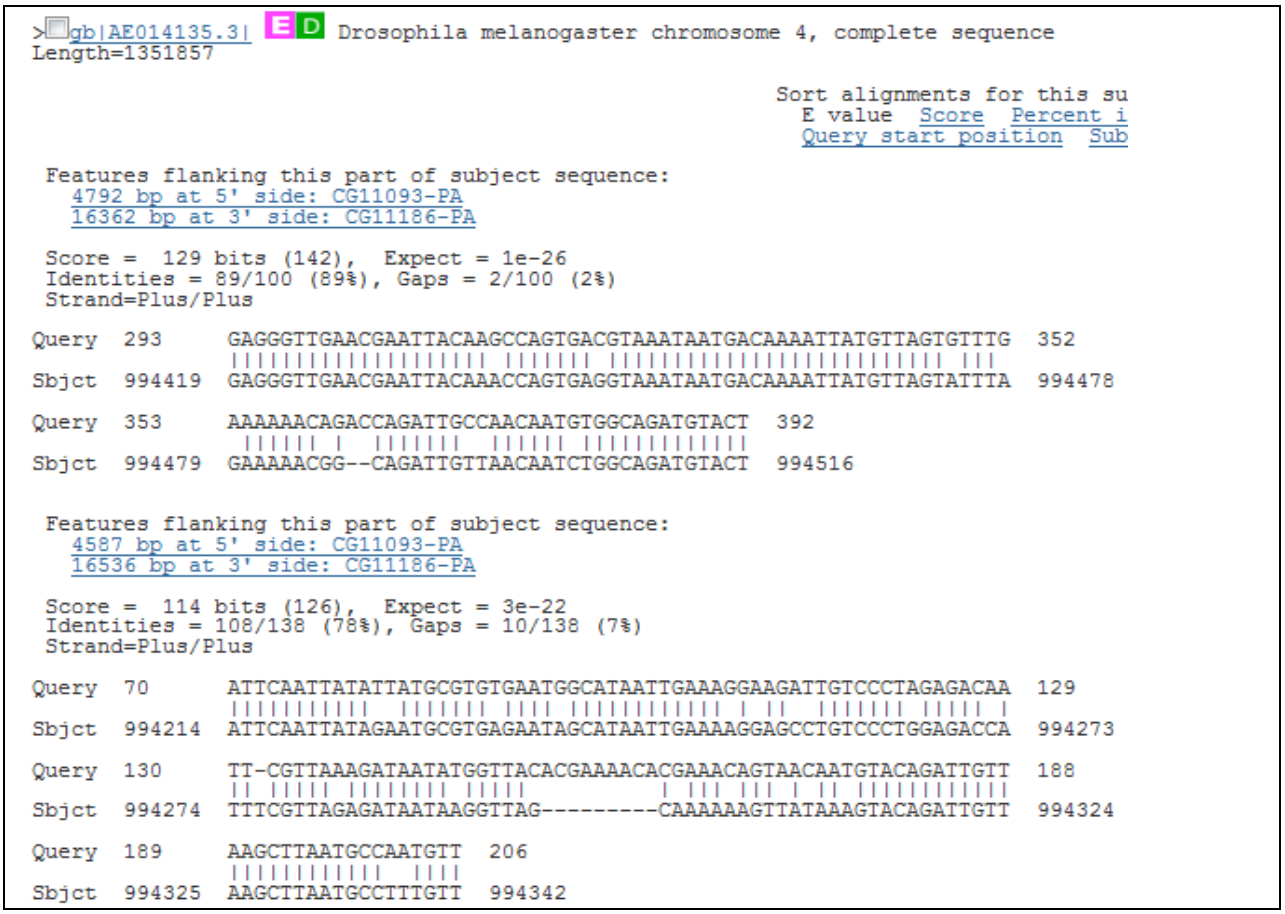
Finally, I evaluated two regions of high conservation not associated with any protein-coding sequences or repetitious elements. Inspection of these loci revealed stop codons in all six reading frames, indicating that these regions cannot be protein-coding. I extracted the DNA in each region of high conservation, which is approximately from 27670-28040 and 28900-29420. Then, using blastn I searched the extracted DNA from 27670-28040 to the nt database (Figure 18). No sequences of high similarity and coverage were identified, although some conservation exists with *Drosophila virilis*.

```
>gb|DQ378288.1| Drosophila virilis strain Tucson 15010-1001.10 chromosome 4 fosmid
44I5, complete sequence
Length=39850

 Score =  169 bits (186),  Expect = 1e-38
 Identities = 142/169 (84%), Gaps = 4/169 (2%)
 Strand=Plus/Minus

Query  196    TTTTTGGTAAAAGTCAACGCACATGATTCGATTTGGGGGATACGGCTATTTGGCTGCTTT  255
              || |||||||| || ||||||||||||||| || ||||||||  | |||||||||| |||
Sbjct  11107  TTGTTGGTAAA-GTTAACGCACATGATTCGACTTTGGGGATATTGATATTTGGCTGTTTT  11049

Query  256    CCATTCCATTGGCTGGAGTTGAAGACCGAGTTGATTTCTGTATTGACTCTGGGAATGCCA  315
              |||||||||| | |||||||||| ||| || ||| | || ||||||||| |||||||||
Sbjct  11048  CCATTCCATTAGGTGGAGTTGAAGGCCGGGTAGATCTTTGAATTGACTCTTGGAATGCCA  10989

Query  316    TTTAGTGCAATTTCTGaaaaaagaaaaccaaaaaTC-TATAAAGATTAT  363
              |||||||| ||||||| ||| ||||||| || | ||  |||||| ||||
Sbjct  10988  TTTAGTGCTATTTCTG-AAAGAGAAAACAAAGAGTCGAATAAAG-TTAT  10942
```

Figure 18: The best alignment from a blastn search of the nucleotide sequence from 27670-28040 from fosmid 18 against the nt database. Note that while the result has a low E value, not enough evidence is available to determine if this is a functional region.

The blastn alignment of the DNA sequence extracted from 28900-29420 in fosmid 18 against the nt database reveals conservation of this sequence and the fourth chromosomes of both *D. virilis* and *D. melanogaster* (Figure 19). A blastn search of the Flybase EST database for all Drosophila species using the same query as before found only matches of short lengths and low coverage. Thus, the region is probably not expressed, but may be a regulatory element. Due to the distance of this region from neighboring genes, this feature is unlikely to be a UTR. Further experimentation is necessary at this locus to determine potential functional roles.

```
>gb|DQ378288.1| Drosophila virilis strain Tucson 15010-1001.10 chromosome 4 fosmid
44I5, complete sequence
Length=39850

 Score =  363 bits (402),  Expect = 4e-97
 Identities = 371/479 (77%), Gaps = 39/479 (8%)
 Strand=Plus/Minus

Query  3      TTAAAGATTGAAATAAATGTATTCAATCTGCGCGAGGAACGCCCATTCTGTATGGAAACG  62
              ||||| ||||  | ||||| ||||||||||  | ||   ||||| || | |   |||| |
Sbjct  10637  TTAAA-ATTGGTAAAAATGCATTCAATCTGGAC-AGACTCGCCCCTTATAT---GAAAGG  10583

Query  63     CTGGCACATTCAATTATATTATGCGTGTGAATGGCATAATTGAAAGGAAGATTGTCCCTA  122
                 |||||||||||||||||||||||||||||||||||||||||| ||||||| |||| |
Sbjct  10582  TC--CACATTCAATTATATTATGCGTGTGAATGGCATAATTGAAAAGAAGATTCTCCCCA  10525

Query  123    GAGACAATTCGTTAAAGATAATATGGTTACACGAAAACACGAAACAGTAACAATGTACAG  182
              ||||||||||||||||||||||||||||||           ||| ||||||||   ||||
Sbjct  10524  GAGACAATTCGTTAAAGATAATATGGTTACA----------AAAAAGTAACAAAACACAG  10475

Query  183    ATTGTTAAGCTTAATGCCAATGTTTCCCTTATGGGAAACGAAACAAATGATTGCAGGACT  242
              ||||||||||||||||||||||||||| | |  ||||  ||||||||||||||| |  ||||
Sbjct  10474  ATTGTTAAGCTTAATGCCAATGTTTCTCCTGCGGGAGGCGAAACAAATGATTACGAGACT  10415

Query  243    TCTTAACTTCGCTA----------GCGAGAGA--CGAAGAGGGGCGAGGGCGCAAGCCAC  290
              |||||||||    |          ||||||||  | ||| | || |  ||||  ||| ||
Sbjct  10414  TCTTAACTTGTAAAGGGAGACACGGCGAGAGAGCCAAAGGGAGGAGGAGGCGGGAGCAAC  10355

Query  291    CAGAGGGTTGAACGAATTACAAGCCAGTGACGTAAATAATGACAAAATTATGTTAGTGTT  350
              |||||||||||||||||||| ||||||||||||||||||||||||||||||||||||| ||
Sbjct  10354  CAGAGGGTTGAACGAATTACAAACCAGTGACGTAAATAATGACAAAATTATGTTAGTATT  10295

Query  351    TGAAAAAACAGACCAGATTGCCAACAATGTGGCAGATGTACTTAGTTAGTCAGCGAACAA  410
                 ||||||||  |||||||  ||||||||||||||||||||||  ||| ||  || ||
Sbjct  10294  AAGAAAAACAG--CAGATTGTTAACAATGTGGCAGATGTACTTAGCAAGTTAG-TAAAAA  10238

Query  411    TAAAGTCGGGGGAGGCGGGCTTTGCCCAAGGCTGCCGATTTCTAAGCGGGGTTGTTTTC  469
              |||| ||||||  |      |||| || ||| |    | |  ||| ||||||||||||
Sbjct  10237  TAAA-TCGGGGATG------TTTGTCCGAGGTTAGGAAATATTAAACGGGGTTGTTTTC  10186
```



```
>gb|AE014135.3|  Drosophila melanogaster chromosome 4, complete sequence
Length=1351857

                                                    Sort alignments for this su
                                                      E value  Score  Percent i
                                                      Query start position  Sub

 Features flanking this part of subject sequence:
   4792 bp at 5' side: CG11093-PA
   16362 bp at 3' side: CG11186-PA

 Score =  129 bits (142),  Expect = 1e-26
 Identities = 89/100 (89%), Gaps = 2/100 (2%)
 Strand=Plus/Plus

Query  293     GAGGGTTGAACGAATTACAAGCCAGTGACGTAAATAATGACAAAATTATGTTAGTGTTTG  352
               ||||||||||||||||||| ||||||| |||||||||||||||||||||||||| |||
Sbjct  994419  GAGGGTTGAACGAATTACAAACCAGTGAGGTAAATAATGACAAAATTATGTTAGTATTTA  994478

Query  353     AAAAAACAGACCAGATTGCCAACAATGTGGCAGATGTACT  392
                ||||||  |  |||||||  |||||| ||||||||||||
Sbjct  994479  GAAAAACGG--CAGATTGTTAACAATCTGGCAGATGTACT  994516


 Features flanking this part of subject sequence:
   4587 bp at 5' side: CG11093-PA
   16536 bp at 3' side: CG11186-PA

 Score =  114 bits (126),  Expect = 3e-22
 Identities = 108/138 (78%), Gaps = 10/138 (7%)
 Strand=Plus/Plus

Query  70      ATTCAATTATATTATGCGTGTGAATGGCATAATTGAAAGGAAGATTGTCCCTAGAGACAA  129
               ||||||||||  ||||||| |||| |||||||||||| | ||  ||||||| ||||| |
Sbjct  994214  ATTCAATTATAGAATGCGTGAGAATAGCATAATTGAAAAGGAGCCTGTCCCTGGAGACCA  994273

Query  130     TT-CGTTAAAGATAATATGGTTACACGAAAACACGAAACAGTAACAATGTACAGATTGTT  188
               || ||||| |||||||| |||||          | ||| ||| | || ||||||||||||
Sbjct  994274  TTTCGTTAGAGATAATAAGGTTAG---------CAAAAAAGTTATAAAGTACAGATTGTT  994324

Query  189     AAGCTTAATGCCAATGTT  206
               ||||||||||||  ||||
Sbjct  994325  AAGCTTAATGCCTTTGTT  994342
```

Figure 19: (Top) Best match resulting from a blastn search using the nucleotide sequence from 28900-29420 in fosmid 18 against the nt database. (Bottom) The second best match from the previously-described search. Note the high conservation to both *D. virilis* and *D. melanogaster*.

Synteny

Regions for features 1-3 appear syntenic to the *D. melanogaster* ortholog, as evidenced by Figure 20 compared to the orientation of the genes predicted by Genscan. However, the ortholog to feature 4 is on chromosome 2L in *D. melanogaster*. Inspection of the neighboring fosmid (fosmid 17) is necessary to determine more.

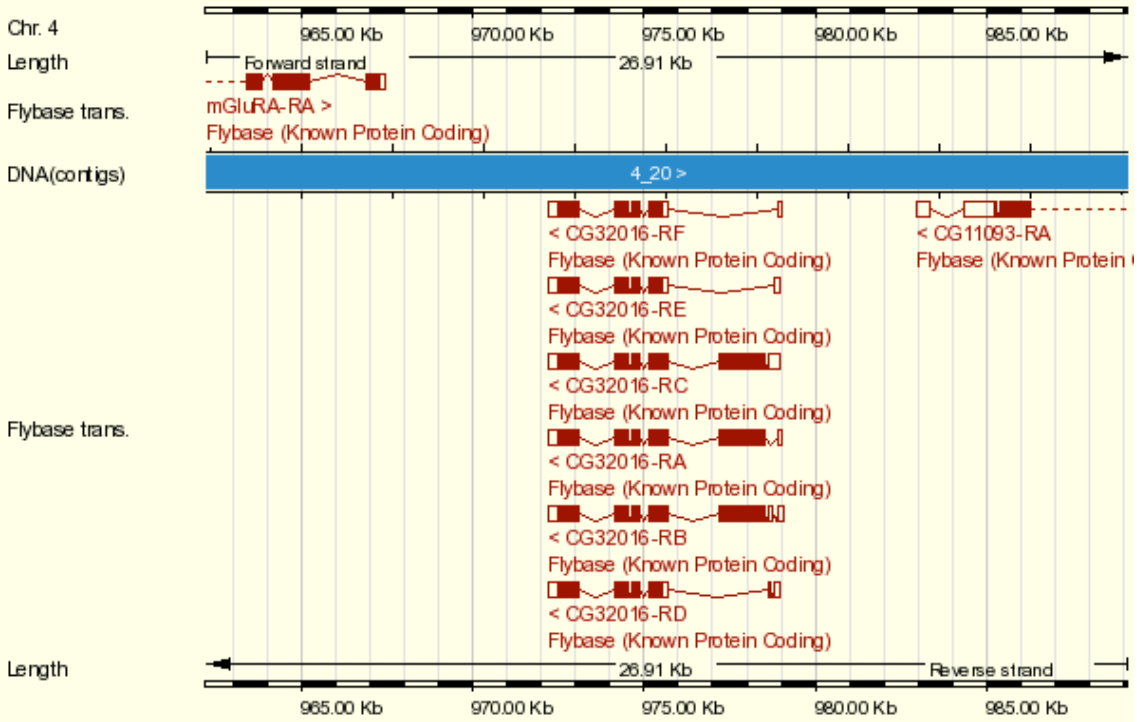


Figure 20: Flybase annotation shows synteny of the genes found on Fosmid 18 with the fourth chromosome of *D. melanogaster*.

Repeat Masker Results

First, I ran Repeat Masker to identify repetitious elements in the fosmid (Table 2, Table 3). The fosmid was only 29.5% repetitious sequences based on the parameters used. Very few *LINE* elements were found in the fosmid. Four repetitious elements were over 500bp long, as listed in Table 3.

```
total length:      40602 bp  (40602 bp excl N/X-runs)
GC level:         38.61 %
bases masked:      11983 bp ( 29.51 %)
==================================================
               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:                4         1024 bp    2.52 %
      LINE1           0            0 bp    0.00 %
      LINE2           0            0 bp    0.00 %
      L3/CR1          0            0 bp    0.00 %

LTR elements:         1           26 bp    0.06 %
      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:	27	7568 bp	18.64 %
MER1_type	0	0 bp	0.00 %
MER2_type	0	0 bp	0.00 %
Unclassified:	2	328 bp	0.81 %
Total interspersed repeats:		8946 bp	22.03 %
Small RNA:	0	0 bp	0.00 %
Satellites:	19	2955 bp	7.28 %
Simple repeats:	7	699 bp	1.72 %
Low complexity:	0	0 bp	0.00 %

Table 2: Repeat Masker output.

SW score	position begin	position end	matching repeat	repeat class/family	length
1737	21222	21768	dmoj.7.20.centroid	DNA	546
3904	3309	3951	dmoj.0.14.centroid	Satellite	642
5300	19271	20021	dmoj.1.77.centroid	DNA	750
5031	4029	4837	dmoj.1.77.centroid	DNA	808

Table 3: Repetitious elements longer than 500bp.

Conclusion

Fosmid 18 is 30% repetitious sequences. In addition, the fosmid contains three functional genes syntenic with the fourth chromosome in *D. melanogaster*. The ortholog to *mGluRA* and *CG32016* both had one exon split by a repeat. The ortholog to *CG11093* has strong conservation in exon 2, which contains three functional domains. However, the other two exons are not conserved and were determined based on limited evidence. Finally, I identified a gene model with 75% sequence conservation to *CG5367* in *D. melanogaster*, which was not predicted by Genscan. Further work to verify these annotations is necessary, particularly with *CG11093*. In addition, synteny analysis in the neighboring fosmid is necessary for feature 4, which has moved from chromosome 2L in *D. melanogaster* to chromosome 4 in *D. mojavensis*.

mGluRA: 2021-8793
CG32016: 15917-10881
CG11093: 27985-23433
CG5367: 36693-38152 (missing first exon)

Table 4: Summary of genes annotated on fosmid 18.

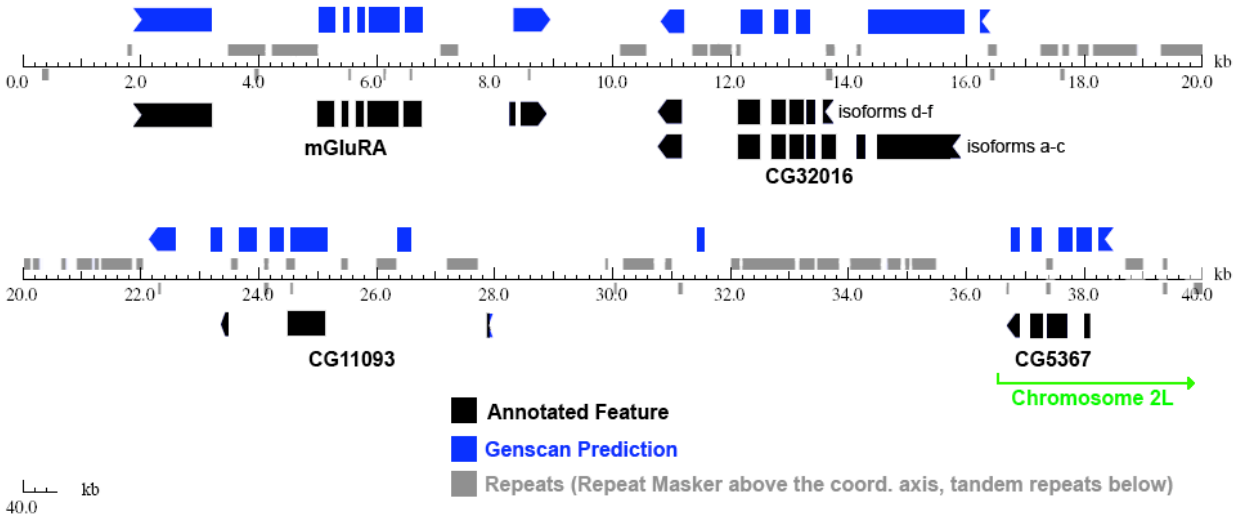


Figure 21: Final feature map.

Appendix

Feature 1

>Dmoj3_fosmid18_mGluRA_NM_079900_pep
MLMSATCWLSQLDIAGSTQVQDTVSVSLPGDIILGGLFPVHEKGEGTPSC
GPKVYNRGVQRLEAMLYAIDRVNNDTNILPGITIGVHILDTCSRDTYALN
QSLQFVRASLNNLDTSVFECSDTSIPQLRKNATSGPVFGVIGGSYSSVSL
QVANLLRLFHIPQISPASTAKTLSDKSRFDLFARTVPPDTFQSVALVDII
KNFNWSYVSTIHSEGSYGEYGIEAFHKEATERHVCIAAAEKVPSAADDKV
FDAIIGKLLKKPNARGVILFTRAEDARRILLAAKRANLSQPFHWVASDGW
GKQQKLLEGLEDIAEGAITVELQSEIIEDFDRYMMQLTPTSNKRNPWFAE
YWEDTFNCVLPTSSASDKINKGDDSNQKEKVVCDDSFRLSDKVGYEQESK
TQFVVDAVYAFAHALHNLHNDLCTQSDQNVGPQHHQHSESVWYRRPADRQ
SQACPNMASYDGKDFYNNYLLNVSFVDLAGSEVKFDRQGDGLARYDILNY
QRLDNSSGYQYKVIGKWFNSLELNLDTLVWNKANEMPTSACSLPCEAGMI
KKQQGDTCCWICDSCEAYEYVYDEQTCRDCGPGLWPYPDKLSCFALNIQY
MRWNSLFALVPLAIAVLGITMTIIVIVLFARNHDTPLVRASGRELSYTLL
FGILVCYCNTFALIAKPTIGSCVLQRFGIGVGFSIIYSALLTKTNRISRI
FHSASKSAQRLKYISPQSQVVITTSLIAIQVLITMIWMIVEPPGTRFYYP
DRTEVILKCKIQDMSFLFSQLYNMILITICTIYAIKTRKIPENFNESKFI
GFTMYTTCIIWLAFVPIYFGTGNSYEIQITTLCISISLSASVALICLYSP
KVYILVFHPDKNVRKLTMNSTVYRRSATTGAGQGVPTSSGYSRTPIVLNS
EDACYGTGGGIGVGVEIGVGTITMSATERPSPSISQKSRNCDLDRNQSVV
IHRNDECYNNCASTSEKCDCGLGIGDPACIGIED*

>Dmoj3_fosmid18_mGluRA_NM_079900_cds
ATGCTCATGTCAGCCACATGTTGGCTTTCGCAACTGGATATAGCCGGCTC
AACGCAAGTGCAGGACACTGTGTCCGTGTCGTTGCCAGGTGATATAATAC
TAGGTGGACTGTTTCCGGTCCACGAGAAAGGCGAGGGTACCCCGTCATGC
GGACCAAAAGTCTATAATCGCGGTGTTCAGCGTTTGGAGGCGATGCTCTA
TGCCATTGATCGTGTTAATAACGATACAAATATATTGCCCGGGATAACAA
TAGGTGTGCATATACTCGACACCTGCTCGCGGGATACGTATGCCCTGAAT
CAATCGCTGCAATTTGTGAGAGCCTCACTCAACAACTTGGACACATCCGT
GTTCGAATGCTCCGACACATCAATACCTCAGCTCAGAAAAAATGCGACAT
CTGGCCCCGTGTTTGGTGTTATTGGCGGATCGTACAGTTCGGTATCACTC
CAAGTGGCTAATCTATTGCGACTCTTTCATATTCCGCAAATATCGCCGGC
ATCAACTGCAAAAACTCTATCGGATAAATCCAGATTTGATCTATTTGCAC
GTACCGTGCCACCCGATACATTCCAATCGGTTGCTTTGGTGGACATCATC
AAGAACTTTAATTGGTCGTATGTATCGACAATACATTCCGAGGGCTCCTA
TGGCGAATATGGCATTGAGGCCTTCCATAAGGAGGCCACCGAGAGACACG
TTTGCATTGCTGCCGCCGAGAAAGTGCCGAGCGCTGCCGACGACAAAGTC
TTTGATGCCATCATAGGGAAACTCCTTAAAAAACCAAATGCCCGCGGAGT
TATCCTTTTCACACGCGCCGAAGACGCCAGACGCATTCTTCTGGCTGCTA
AGCGTGCTAATTTATCCCAGCCCTTCCATTGGGTGGCCAGCGATGGTTGG
GGTAAGCAGCAAAAACTGCTAGAGGGCCTGGAGGACATCGCCGAGGGAGC
CATAACTGTGGAACTGCAGTCGGAAATAATTGAGGATTTCGATAGATATA

TGATGCAACTGACGCCCACATCCAATAAGCGCAACCCCTGGTTTGCTGAG
TACTGGGAGGACACCTTCAATTGTGTGTTGCCAACGAGCTCCGCAAGTGA
CAAGATCAATAAGGGAGATGACAGCAACCAGAAGGAGAAAGTTGTTTGTG
ATGATAGTTTCAGGCTATCCGATAAAGTTGGATACGAACAGGAATCGAAA
ACGCAATTCGTTGTGGATGCTGTGTACGCGTTTGCCCACGCCCTTCATAA
TTTGCACAACGACTTGTGCACACAAAGTGATCAAAATGTGGGACCACAAC
ATCATCAGCACAGTGAATCCGTATGGTACCGTAGGCCAGCGGATCGCCAA
TCACAGGCATGCCCCAATATGGCTAGTTATGACGGCAAGGACTTCTACAA
CAATTATTTGTTGAATGTGTCATTTGTAGATTTGGCTGGCAGTGAAGTGA
AGTTTGATCGTCAGGGTGATGGACTGGCTAGATATGACATTTTAAATTAT
CAACGCCTTGATAACTCTTCGGGATATCAGTACAAGGTGATTGGCAAATG
GTTTAATTCCCTCGAGCTTAATTTGGATACCCTTGTGTGGAACAAGGCCA
ACGAAATGCCAACATCAGCATGCTCCTTGCCATGTGAAGCTGGAATGATT
AAAAAACAACAGGGGGACACTTGCTGCTGGATCTGTGACAGCTGCGAAGC
GTACGAATACGTTTATGACGAGCAGACATGTAGGGACTGTGGCCCGGGAC
TGTGGCCATATCCCGATAAGTTATCCTGCTTTGCCCTCAACATTCAGTAT
ATGCGTTGGAATTCGCTGTTCGCACTTGTTCCACTGGCGATTGCCGTCCT
TGGCATCACGATGACGATAATCGTTATTGTGCTATTTGCGAGAAATCATG
ACACGCCTCTAGTGCGCGCATCTGGACGAGAGCTGAGCTATACGCTGCTG
TTCGGCATCCTGGTCTGTTACTGCAACACATTCGCCTTGATCGCTAAGCC
AACCATTGGTTCATGCGTACTGCAAAGATTTGGCATCGGAGTCGGGTTCT
CCATCATTTATAGTGCTTTGCTAACAAAAACAAATCGCATATCTCGCATA
TTTCATTCAGCATCCAAATCGGCGCAGCGCCTCAAATACATTAGTCCACA
GTCTCAGGTTGTGATAACGACATCCCTGATCGCAATACAAGTTTTGATAA
CAATGATCTGGATGATTGTGGAACCACCAGGAACTAGATTTTATTATCCC
GACCGCACCGAGGTGATCTTAAAATGCAAGATACAGGATATGTCGTTTCT
CTTTTCTCAGCTCTATAATATGATACTCATTACAATTTGCACAATTTACG
CTATTAAAACACGTAAGATACCAGAAAACTTTAACGAGTCCAAATTTATT
GGATTTACAATGTATACAACCTGCATTATTTGGCTCGCTTTTGTACCAAT
TTACTTCGGCACCGGTAACTCCTACGAGATCCAAATAACTACGCTCTGCA
TCTCTATCAGCTTGAGTGCTTCGGTTGCCCTCATCTGCCTTTACTCGCCC
AAGGTGTATATATTGGTGTTTCATCCTGACAAAAATGTCCGCAAGTTGAC
AATGAACAGCACTGTGTATAGGCGCTCGGCAACAACAGGAGCGGGACAAG
GAGTACCAACTAGTTCCGGATATAGTCGCACTCCAATTGTATTGAATTCT
GAAGACGCATGCTATGGTACTGGAGGCGGAATCGGGGTCGGAGTTGAAAT
CGGAGTCGGTACTATTACTATGTCAGCAACGGAGCGACCGTCCCCAAGTA
TCAGCCAGAAAAGTCGAAATTGCGACTTGGATCGAAATCAAAGTGTTGTC
ATTCATAGGAACGATGAATGCTACAACAATTGTGCTTCCACATCTGAGAA
ATGCGATTGCGGCTTGGGTATTGGTGATCCAGCCTGCATTGGCATTGAAG
AT

```
fosmid18    GEP   exon  2021  3201  .     +     0     gene_id "mGluRA";
trans_id "NM_079900"
fosmid18    GEP   exon  5036  5283  .     +     0     gene_id "mGluRA";
trans_id "NM_079900"
fosmid18    GEP   exon  5433  5539  .     +     0     gene_id "mGluRA";
trans_id "NM_079900"
```

```
fosmid18    GEP   exon  5673  5798  .     +     0     gene_id "mGluRA";
trans_id "NM_079900"
fosmid18    GEP   exon  5872  6391  .     +     0     gene_id "mGluRA";
trans_id "NM_079900"
fosmid18    GEP   exon  6480  6775  .     +     0     gene_id "mGluRA";
trans_id "NM_079900"
fosmid18    GEP   exon  8245  8415  .     +     0     gene_id "mGluRA";
trans_id "NM_079900"
fosmid18    GEP   exon  8491  8793  .     +     0     gene_id "mGluRA";
trans_id "NM_079900"
fosmid18    GEP   CDS   2021  3201  .     +     0     gene_id "mGluRA";
trans_id "NM_079900"
fosmid18    GEP   CDS   5036  5283  .     +     1     gene_id "mGluRA";
trans_id "NM_079900"
fosmid18    GEP   CDS   5433  5539  .     +     2     gene_id "mGluRA";
trans_id "NM_079900"
fosmid18    GEP   CDS   5673  5798  .     +     0     gene_id "mGluRA";
trans_id "NM_079900"
fosmid18    GEP   CDS   5872  6391  .     +     0     gene_id "mGluRA";
trans_id "NM_079900"
fosmid18    GEP   CDS   6480  6775  .     +     2     gene_id "mGluRA";
trans_id "NM_079900"
fosmid18    GEP   CDS   8245  8415  .     +     0     gene_id "mGluRA";
trans_id "NM_079900"
fosmid18    GEP   CDS   8491  8793  .     +     0     gene_id "mGluRA";
trans_id "NM_079900"
fosmid18    GEP   start_codon 2021  2023  .     +     .     gene_id "mGluRA";
trans_id "NM_079900"
fosmid18    GEP   stop_codon  8794  8796  .     +     .     gene_id "mGluRA";
trans_id "NM_079900";
```

Feature 2 isoform 1

>Dmoj3_fosmid18_CG32016_NM_166797.1_cds
ATGATAAGGAAACGCAACACATCTACATGGTTTATTTTTTTTTGTTTGTT
CAGGTACTCACGATCTGATTTACTTGCTTTGCGCTTCGAAGGAAGTGCAC
GTCAACGACCAAATTGTGCAAATCGAGCTGATCTGCAGACGTTAAATTTT
TGGAAAAGTAATGGGGCTCCACTGAGTGTGAGCAGCAACAGTTGTGGAAA
TAAAAGCGGATTGTCACCCGAAAGGGAAAGTTCAAGTCTAACGAGTTCAA
ATGGACTCTTGAGTTCGAGGCGAGCTTTACGAAATCGCGAACGAGCCCAC
AATTATTATCAGCGCTTTTCATCCGGCGATCAAATTGCTGGAGAAGAATT
GCAGCAGCAGAGCAGCTTAGGCTCAGCCAGTGTCAATCTGTGTGGCTCCT
CCACATATAAATCACCTAATATAGATCATCGTAGTATATCCTCAGCACAT
TTAATGCCAGCGTTTGCGAAAAAGCGTTTTGTAGTTGAGCAGGCCGAATC
TCAACATGAGGATGGAAATGCAAAGGAGAGTAGTGCACCAGCCGCTTCGG
CTACTCAAAGCAATCTTGACCTGAAAAGTCACAGTGAAGTGCGCGCCATC
AGCCGAGCCTCTAAAACGGCGGCTGCTTCATCCCCAGATTGGGAGCGTGG
GGAGAAACATTTAAACCCACCACATAATGAACAATCGATGTCGCCGACTG
CATTTTTGGCTGGGCGTCAAGTGGCGAGCGGCTTGAATTTGCAAGAAAGA
CGGATTGGCAGCGGCCGCCTATTGCCACGTAATGATAATTGGGAATATAC

AAAAAACAAGGAGGCGGATAATTTATTAAATAATCTTGATAAGGAGAGAC
CTCAAAATGGAAACCTTACCAGCCAGTATCAGCCGCAGCAGCAGCAACAT
CAGCGTCTGCGTACATCGTCTGGCAAACATATCGATAGTTTAGTAGATCG
CGATCGTGACCGTGAGAGAGATCGTCGTAAGGCTGTCGATAATCATCGTG
ATCAAAATGATGGAAAAAAAAATGTTCAAAGTGGACGTAGTCGCACTGCA
ACCAACAAGGATAAATTCAATTTCGGTGATTGTCAGTCGGGGGCTATGCA
AAATCGTGGCAAGCGCATGAATTTGTATCAAACTCATGAACCGGAGTGGT
TTAGTGCTGGACCTACCTCCCAGCATGAAACGATTGATCTACATGGCTTT
GAAGATTACGAAAGTGAAAGCTCATCGGCAGATTTGAAGAACATCAAAGA
TGACATTTGTAGAGAATCAGATGTTAAGCCGACTCAACCCCCATCTCGAA
GTAGTAGCCTAGCCAGTTTGAATAACAACAAAAACGATGCGGAGGAACAA
CACGCTGCAACAAATAATAAACAATTGAACTACAGTGTTAACTCAGTCCC
AAAATCTGATGTTGAGTTTAATTTCGATGCTTATCTTAACATGGATCCAA
TGGATCACGCTCTCATGGGTAATGACAGCGAGGTGCATGATGTCCAAAGT
ACATCGCGTTTCAGTCAATGGTTTGCATGCAAAGAGCCAACAAAGAACGA
GCGCGTATTTATTGGAGACACTACAGCGCTAGATACGCAAGAAATCCCGA
GTGTTAAGGACTTAGAAGCACAAATGAGAAAAGTGGACTTGCGCCCAGAA
TGTGATATAAGTAAATCGCAGGCGAACGCACATCAACCACAACAAACTGT
GCATACAGAGCAGAGAGACAAATCTGTTACCAGAGATACGGAAGCTTTTA
AAAGACTTTTGCAACAGTTGGGCTCGCAGAATAATACTGAATCAACATAC
CCTATAACTTCAAATATGCCAGCACTATATCAGCAACATCGTCCTAAAAA
GAATGATTCTATGTCGCATAGACATATGGAGATGATGCCACAAAATCAAT
TACATCCTAATGTTCATGAGCTGAATATATCATTGCAAACGCATAAACGA
ATGGAGATTCAACAATTATTACAAGGTGTCACTCGGGGCGATGTTTCGGT
TGAAATTTTAGAGAAGGAGTTGAATAATCCGAATGCAACGCCTCTATCTA
GGGAAGTAATTGCAGCAGTACTTTTTGAGTTAACAAATTCTGGGCGTAGT
ACTCTCCACCAACAACAACATCTAACGTATGCAAACGATATAAATTCCCA
AAACTTTATGAAAAATAATCAATTCTTTAACAAGAATGTGAACCCTGATG
ATATTTACACACATCAAAACACAACGAGCCAAGGCCTACATCAAAATCCA
AGGCATACTAATTCACCGACACCTTTAGCATTTACGCCCACTTCAGTTTT
AAGAAAAATGACGGCTGATAAGGAGACAACGGCGTCATCCACTTTGAATA
ACAGTAATATATTCAATCAACAACAGCAACAACAATATGTTCAGCAATTT
AAACTAACAACGCCAGCACAGCCGAAACTGGATAATCAGCGGATAGGGCA
TAATATGTTGAGTGAAACACCAGCAGCGCCAAATCAATTGCAACCTCGAA
TGATCCTGGGTGGTAATTATGCCATGCAGCATCAAATGAATACCAACAGT
CCACAAATTTCGCCAAAATTACCGCCCCCAATGTCTCAAAAGCAAAATCG
ATGGCCACCTGGAAACATACCCATATCCACAACAACAACAGCACAACAGC
ACGGCACAAAGTGTTTCGGTCGTCCTATTCTAAAAGGGACTTTGAACTCG
GGCTCACAACAAATGCCGACAGCATTTCTATTAAATAATGAGCTACAGCA
ACATATGCACCAACAAAAGCTAAAGTCAATGCAAACTACTTCTTCGGATA
CAATAAATAATGATAGCATTCACATGGGATCAACAGCACTTCAAATGGTT
CCATATAACGAAGGAATCTCTTCCCAGCTTCAACAACAGCCACAACCACA
TTATATCCAACAGCAGAGAGCCCAGCCACAACCCCAGCCCGTTCCTGGAC
ATCGATTATTGCATGTAGACTTACATCGTCCGAATAATTTAACGGTACAA
CCAGTTACTATATCGTCAGGTGGTCCCGATGGCAATGGTGATTCGTCGAC
AAATTTAATGAAAAATAACAACCTCAGGCCGATAAACAACTATCATCAAA
GAGACGAAAGATTGTCACCAACATCAAATCAATTGGCACAATGGTTTTCG
CCAGAATTACTTGCCAGAGCATCGGCTGGAAAATTGCCTCTATTGAATAT

GAATCAGGCGCTGAGCTTGGAAGAGTTCGAACGTAGCATGCAACATTCCT
CAGCCACAGTGCTCAAT

>Dmoj3_fosmid18_CG32016_NM_166797.1_pep
MIRKRNTSTWFIFFCLFRYSRSDLLALRFEGSARQRPNCANRADLQTLNF
WKSNGAPLSVSSNSCGNKSGLSPERESSSLTSSNGLLSSRRALRNRERAH
NYYQRFSSGDQIAGEELQQQSSLGSASVNLCGSSTYKSPNIDHRSISSAH
LMPAFAKKRFVVEQAESQHEDGNAKESSAPAASATQSNLDLKSHSEVRAI
SRASKTAAASSPDWERGEKHLNPPHNEQSMSPTAFLAGRQVASGLNLQER
RIGSGRLLPRNDNWEYTKNKEADNLLNNLDKERPQNGNLTSQYQPQQQQH
QRLRTSSGKHIDSLVDRDRDRERDRRKAVDNHRDQNDGKKNVQSGRSRTA
TNKDKFNFGDCQSGAMQNRGKRMNLYQTHEPEWFSAGPTSQHETIDLHGF
EDYESESSSADLKNIKDDICRESDVKPTQPPSRSSSLASLNNNKNDAEEQ
HAATNNKQLNYSVNSVPKSDVEFNFDAYLNMDPMDHALMGNDSEVHDVQS
TSRFSQWFACKEPTKNERVFIGDTTALDTQEIPSVKDLEAQMRKVDLRPE
CDISKSQANAHQPQQTVHTEQRDKSVTRDTEAFKRLLQQLGSQNNTESTY
PITSNMPALYQQHRPKKNDSMSHRHMEMMPQNQLHPNVHELNISLQTHKR
MEIQQLLQGVTRGDVSVEILEKELNNPNATPLSREVIAAVLFELTNSGRS
TLHQQQHLTYANDINSQNFMKNNQFFNKNVNPDDIYTHQNTTSQGLHQNP
RHTNSPTPLAFTPTSVLRKMTADKETTASSTLNNSNIFNQQQQQQQYVQQF
KLTTPAQPKLDNQRIGHNMLSETPAAPNQLQPRMILGGNYAMQHQMNTNS
PQISPKLPPPMSQKQNRWPPGNIPISTTTTAQQHGTKCFGRPILKGTLNS
GSQQMPTAFLLNNELQQHMHQQKLKSMQTTSSDTINNDSIHMGSTALQMV
PYNEGISSQLQQQPQPHYIQQQRAQPQPQPVPGHRLLHVDLHRPNNLTVQ
PVTISSGGPDGNGDSSTNLMKNNNLRPINNYHQRDERLSPTSNQLAQWFS
PELLARASAGKLPLLNMNQALSLEEFERSMQHSSATVLN*

```
fosmid18    GEP   exon  14451 15917 .     -     0     gene_id "CG32016";
trans_id "NM_166797.1"
fosmid18    GEP   exon  14264 14387 .     -     0     gene_id "CG32016";
trans_id "NM_166797.1"
fosmid18    GEP   exon  13675 13907 .     -     0     gene_id "CG32016";
trans_id "NM_166797.1"
fosmid18    GEP   exon  13317 13467 .     -     0     gene_id "CG32016";
trans_id "NM_166797.1"
fosmid18    GEP   exon  13039 13263 .     -     0     gene_id "CG32016";
trans_id "NM_166797.1"
fosmid18    GEP   exon  12509 12976 .     -     0     gene_id "CG32016";
trans_id "NM_166797.1"
fosmid18    GEP   exon  12099 12449 .     -     0     gene_id "CG32016";
trans_id "NM_166797.1"
fosmid18    GEP   exon  10881 11128 .     -     0     gene_id "CG32016";
trans_id "NM_166797.1"
fosmid18    GEP   CDS   14451 15917 .     -     0     gene_id "CG32016";
trans_id "NM_166797.1"
fosmid18    GEP   CDS   14264 14387 .     -     0     gene_id "CG32016";
trans_id "NM_166797.1"
fosmid18    GEP   CDS   13675 13907 .     -     2     gene_id "CG32016";
trans_id "NM_166797.1"
fosmid18    GEP   CDS   13317 13467 .     -     0     gene_id "CG32016";
trans_id "NM_166797.1"
```

```
fosmid18    GEP   CDS   13039 13263 .     -     2     gene_id "CG32016";
trans_id "NM_166797.1"
fosmid18    GEP   CDS   12509 12976 .     -     2     gene_id "CG32016";
trans_id "NM_166797.1"
fosmid18    GEP   CDS   12099 12449 .     -     2     gene_id "CG32016";
trans_id "NM_166797.1"
fosmid18    GEP   CDS   10881 11128 .     -     2     gene_id "CG32016";
trans_id "NM_166797.1"
fosmid18    GEP   start_codon 15915 15917 .     -     .     gene_id
"CG32016"; trans_id "NM_166797.1"
fosmid18    GEP   stop_codon  10880 10878 .     -     .     gene_id
"CG32016"; trans_id "NM_166797.1"
```

Feature 2 isoform 2

>Dmoj3_fosmid18_CG32016_NM_166799.1_pep
MRKVDLRPECDISKSQANAHQPQQTVHTEQRDKSVTRDTEAFKRLLQQLG
SQNNTESTYPITSNMPALYQQHRPKKNDSMSHRHMEMMPQNQLHPNVHEL
NISLQTHKRMEIQQLLQGVTRGDVSVEILEKELNNPNATPLSREVIAAVL
FELTNSGRSTLHQQQHLTYANDINSQNFMKNNQFFNKNVNPDDIYTHQNT
TSQGLHQNPRHTNSPTPLAFTPTSVLRKMTADKETTASSTLNNSNIFNQQ
QQQQYVQQFKLTTPAQPKLDNQRIGHNMLSETPAAPNQLQPRMILGGNYA
MQHQMNTNSPQISPKLPPPMSQKQNRWPPGNIPISTTTTAQQHGTKCFGR
PILKGTLNSGSQQMPTAFLLNNELQQHMHQQKLKSMQTTSSDTINNDSIH
MGSTALQMVPYNEGISSQLQQQPQPHYIQQQRAQPQPQPVPGHRLLHVDL
HRPNNLTVQPVTISSGGPDGNGDSSTNLMKNNNLRPINNYHQRDERLSPT
SNQLAQWFSPELLARASAGKLPLLNMNQALSLEEFERSMQHSSATVLN*

>Dmoj3_fosmid18_CG32016_NM_166799.1_cds
ATGAGAAAAGTGGACTTGCGCCCAGAATGTGATATAAGTAAATCGCAGGC
GAACGCACATCAACCACAACAAACTGTGCATACAGAGCAGAGAGACAAAT
CTGTTACCAGAGATACGGAAGCTTTTAAAAGACTTTTGCAACAGTTGGGC
TCGCAGAATAATACTGAATCAACATACCCTATAACTTCAAATATGCCAGC
ACTATATCAGCAACATCGTCCTAAAAAGAATGATTCTATGTCGCATAGAC
ATATGGAGATGATGCCACAAAATCAATTACATCCTAATGTTCATGAGCTG
AATATATCATTGCAAACGCATAAACGAATGGAGATTCAACAATTATTACA
AGGTGTCACTCGGGGCGATGTTTCGGTTGAAATTTTAGAGAAGGAGTTGA
ATAATCCGAATGCAACGCCTCTATCTAGGGAAGTAATTGCAGCAGTACTT
TTTGAGTTAACAAATTCTGGGCGTAGTACTCTCCACCAACAACAACATCT
AACGTATGCAAACGATATAAATTCCCAAAACTTTATGAAAAATAATCAAT
TCTTTAACAAGAATGTGAACCCTGATGATATTTACACACATCAAAACACA
ACGAGCCAAGGCCTACATCAAAATCCAAGGCATACTAATTCACCGACACC
TTTAGCATTTACGCCCACTTCAGTTTTAAGAAAAATGACGGCTGATAAGG
AGACAACGGCGTCATCCACTTTGAATAACAGTAATATATTCAATCAACAA
CAGCAACAACAATATGTTCAGCAATTTAAACTAACAACGCCAGCACAGCC
GAAACTGGATAATCAGCGGATAGGGCATAATATGTTGAGTGAAACACCAG
CAGCGCCAAATCAATTGCAACCTCGAATGATCCTGGGTGGTAATTATGCC

ATGCAGCATCAAATGAATACCAACAGTCCACAAATTTCGCCAAAATTACC
GCCCCCAATGTCTCAAAAGCAAAATCGATGGCCACCTGGAAACATACCCA
TATCCACAACAACAACAGCACAACAGCACGGCACAAAGTGTTTCGGTCGT
CCTATTCTAAAAGGGACTTTGAACTCGGGCTCACAACAAATGCCGACAGC
ATTTCTATTAAATAATGAGCTACAGCAACATATGCACCAACAAAAGCTAA
AGTCAATGCAAACTACTTCTTCGGATACAATAAATAATGATAGCATTCAC
ATGGGATCAACAGCACTTCAAATGGTTCCATATAACGAAGGAATCTCTTC
CCAGCTTCAACAACAGCCACAACCACATTATATCCAACAGCAGAGAGCCC
AGCCACAACCCCAGCCCGTTCCTGGACATCGATTATTGCATGTAGACTTA
CATCGTCCGAATAATTTAACGGTACAACCAGTTACTATATCGTCAGGTGG
TCCCGATGGCAATGGTGATTCGTCGACAAATTTAATGAAAAATAACAACC
TCAGGCCGATAAACAACTATCATCAAAGAGACGAAAGATTGTCACCAACA
TCAAATCAATTGGCACAATGGTTTTCGCCAGAATTACTTGCCAGAGCATC
GGCTGGAAAATTGCCTCTATTGAATATGAATCAGGCGCTGAGCTTGGAAG
AGTTCGAACGTAGCATGCAACATTCCTCAGCCACAGTGCTCAAT

```
fosmid18    GEP   exon  13675 13875 .     -     0     gene_id "CG32016";
trans_id "NM_166799.1"
fosmid18    GEP   exon  13317 13467 .     -     0     gene_id "CG32016";
trans_id "NM_166799.1"
fosmid18    GEP   exon  13039 13263 .     -     0     gene_id "CG32016";
trans_id "NM_166799.1"
fosmid18    GEP   exon  12509 12976 .     -     0     gene_id "CG32016";
trans_id "NM_166799.1"
fosmid18    GEP   exon  12099 12449 .     -     0     gene_id "CG32016";
trans_id "NM_166799.1"
fosmid18    GEP   exon  10881 11128 .     -     0     gene_id "CG32016";
trans_id "NM_166799.1"
fosmid18    GEP   CDS   13675 13875 .     -     0     gene_id "CG32016";
trans_id "NM_166799.1"
fosmid18    GEP   CDS   13317 13467 .     -     0     gene_id "CG32016";
trans_id "NM_166799.1"
fosmid18    GEP   CDS   13039 13263 .     -     2     gene_id "CG32016";
trans_id "NM_166799.1"
fosmid18    GEP   CDS   12509 12976 .     -     2     gene_id "CG32016";
trans_id "NM_166799.1"
fosmid18    GEP   CDS   12099 12449 .     -     2     gene_id "CG32016";
trans_id "NM_166799.1"
fosmid18    GEP   CDS   10881 11128 .     -     2     gene_id "CG32016";
trans_id "NM_166799.1"
fosmid18    GEP   start_codon 13873 13875 .     -     .     gene_id
"CG32016"; trans_id "NM_166799.1"
fosmid18    GEP   stop_codon  10880 10878 .     -     .     gene_id
"CG32016"; trans_id "NM_166799.1";
```

Feature 3

>Dmoj3_fosmid18_CG11093_NM_143689.1_pep
MAFPESIQKSTRSSTPANGMESSQIAVSSVMLYGIPIVSLYIEGQERLCL

AQISNTLLKQFSYNEIHNRRVALGITCVQCTPVQLEILRRAGAMPVSSRR
CGMITRREAERLCKSFLGDNTPPRLPDDFAFSVQHKCAWGCRGSFLPARY
NSSRAKCIKCSYCGMFFSPNKFIFHSHRITTNDRYVQPDAANFNSWRRHM
TLTGHSHDEKIIHAWEDVKAMFNAIVGMLPMMMTKLSTLRRQKMMLYS*

>Dmoj3_fosmid18_CG11093_NM_143689.1_cds
ATGGCATTCCCAGAGTCAATACAGAAATCAACTCGGTCTTCAACTCCAGC
CAATGGAATGGAAAGCAGCCAAATAGCCGTAAGCTCCGTAATGCTTTATG
GTATACCCATAGTATCCCTGTACATTGAAGGTCAAGAGCGCCTGTGCTTG
GCCCAGATTTCCAACACACTTTTGAAACAGTTCAGCTATAATGAGATCCA
CAATCGACGCGTGGCACTGGGTATAACGTGCGTGCAGTGCACGCCCGTCC
AACTGGAGATACTACGCAGAGCTGGTGCTATGCCCGTCAGTTCCCGACGC
TGCGGGATGATAACGCGGCGAGAGGCGGAACGGCTCTGCAAGAGCTTTCT
GGGTGATAATACACCGCCTCGGCTGCCCGATGATTTTGCATTTAGCGTAC
AGCACAAATGCGCGTGGGGCTGTCGCGGATCGTTTCTGCCCGCACGTTAT
AATTCGTCGCGTGCGAAGTGCATTAAGTGCAGCTATTGTGGAATGTTCTT
CTCGCCAAATAAATTCATTTTCCACTCGCATCGCATTACCACCAACGATC
GATATGTACAGCCGGATGCAGCTAACTTCAATTCCTGGCGTCGTCATATG
ACTTTGACTGGCCACAGTCATGATGAGAAGATAATTCACGCTTGGGAAGA
TGTTAAGGCAATGTTTAATGCCATAGTGGGCATGTTGCCGATGATGATGA
CGAAGTTGTCGACATTGAGACGACAGAAGATGATGCTTTACTCG

```
fosmid18    GEP   exon  27908 27985 .     -     0     gene_id "CG11093";
trans_id "NM_143689.1"
fosmid18    GEP   exon  24537 25128 .     -     0     gene_id "CG11093";
trans_id "NM_143689.1"
fosmid18    GEP   exon  23433 23506 .     -     0     gene_id "CG11093";
trans_id "NM_143689.1"
fosmid18    GEP   CDS   27908 27985 .     -     0     gene_id "CG11093";
trans_id "NM_143689.1"
fosmid18    GEP   CDS   24537 25128 .     -     0     gene_id "CG11093";
trans_id "NM_143689.1"
fosmid18    GEP   CDS   23433 23506 .     -     2     gene_id "CG11093";
trans_id "NM_143689.1"
fosmid18    GEP   start_codon 27983 27985 .     -     .     gene_id
"CG11093"; trans_id "NM_143689.1"
fosmid18    GEP   stop_codon  23432 23430 .     -     .     gene_id
"CG11093"; trans_id "NM_143689.1";
```

Feature 4

>Dmoj3_fosmid18_CG5367_NM_135543.1_pep
EANNRSYIRAHDELRSYKAYEENHQIVEAHNKDYDSGRSSFRLAANTLAD
MSTDSYLKGFLRLLRSPPISASDNMVDIVGATLMDNVPDSLDWRKKGFAT
PSYNQQSCGSCYAFSIAQSIEGQVFKRTGRILSLSEQQIVDCSISHGNQG
CTGGSLRNTLRYLQATGGLMRSVDYKYASKKGACQFVSELAVVNVTSWAI

LPANDENAIQAAVAHIGPVAVSINATPKTFQLYSSTSVNHAMLLIGYDKD
YWILKNWWGEKWGESGYMRMRKGINLCGIANYAAYAIV*

>Dmoj3_fosmid18_CG5367_NM_135543.1_cds
GAAGCCAACAACAGGAGCTACATACGCGCGCATGACGAATTGCGTAGCTA
CAAGGCCTACGAGGAGAACCACCAGATCGTGGAAGCACACAACAAGGACT
ACGACTCGGGGAGGAGCAGCTTTCGCCTGGCAGCGAATACCCTGGCAGAC
ATGAGCACCGACTCGTATCTAAAGGGCTTCTTACGCCTGCTACGGAGTCC
CCCGATATCGGCATCCGATAACATGGTCGATATAGTCGGAGCAACGCTAA
TGGATAATGTGCCAGATAGTTTGGACTGGCGTAAAAAGGGATTCGCGACT
CCGTCGTACAATCAACAGAGCTGCGGCTCGTGCTACGCTTTCAGCATTGC
CCAGAGCATTGAGGGCCAAGTATTCAAGCGCACCGGCAGAATCCTCTCTC
TGAGTGAGCAGCAGATTGTAGATTGCAGCATATCACACGGAAATCAGGGC
TGTACGGGCGGCTCATTGCGCAACACTCTCCGCTATCTACAGGCTACCGG
GGGTCTCATGCGATCCGTCGACTACAAGTATGCCTCAAAGAAAGGAGCTT
GCCAGTTCGTGAGCGAGTTGGCCGTGGTCAACGTCACCTCGTGGGCCATT
TTGCCGGCCAATGACGAGAATGCGATTCAGGCAGCTGTGGCGCATATTGG
CCCAGTTGCAGTGTCTATCAATGCGACGCCGAAAACTTTTCAACTATATA
GTTCGACTTCGGTTAATCATGCGATGCTGCTGATTGGATATGATAAAGAC
TACTGGATCCTTAAGAATTGGTGGGGAGAGAAGTGGGGTGAATCGGGCTA
CATGAGAATGCGTAAGGGCATTAACTTGTGTGGGATTGCCAATTATGCAG
CTTATGCAATAGTT

```
fosmid18    GEP   exon  38000 38152 .     -     0     gene_id "CG5367";
trans_id "NM_135543.1"
fosmid18    GEP   exon  37376 37762 .     -     0     gene_id "CG5367";
trans_id "NM_135543.1"
fosmid18    GEP   exon  37095 37255 .     -     0     gene_id "CG5367";
trans_id "NM_135543.1"
fosmid18    GEP   exon  36693 36855 .     -     0     gene_id "CG5367";
trans_id "NM_135543.1"
fosmid18    GEP   CDS   38000 38152 .     -     0     gene_id "CG5367";
trans_id "NM_135543.1"
fosmid18    GEP   CDS   37376 37762 .     -     0     gene_id "CG5367";
trans_id "NM_135543.1"
fosmid18    GEP   CDS   37095 37255 .     -     0     gene_id "CG5367";
trans_id "NM_135543.1"
fosmid18    GEP   CDS   36693 36855 .     -     1     gene_id "CG5367";
trans_id "NM_135543.1"
fosmid18    GEP   stop_codon  36692 36690 .     -     .     gene_id "CG5367";
trans_id "NM_135543.1";
```

Tandem Repeat Finder Output

Fosmid	*Start*	*End*	*Sequence*	*Length*
fosmid18	39819	39966	TTGATATTTTTTCAAAACTTTATCGATACGATATTTT	147

			AAAAATATTCGACAGCCCTAGTCAGTACACTACCT	
fosmid18	13513	13650	TGCTGCTGTTGT	137
fosmid18	13510	13617	TGCTGT	107
fosmid18	13510	13611	TGCTGCTGCTGTTGCTGTTGATGTTGGTGGTGCTGATGT	101
fosmid18	16342	16413	AT	71
fosmid18	24010	24078	TGC	68
fosmid18	107	171	TA	64
fosmid18	37302	37366	ATGTAT	64
fosmid18	24442	24504	TTG	62
fosmid18	17526	17587	TAATAATGAAGAATAGTAAT	61
fosmid18	17525	17584	ATAATAATGTAGA	59
fosmid18	39306	39365	CT	59
fosmid18	24458	24508	TGTTGTTGTTAT	50
fosmid18	3762	3806	ACGGACAGACGG	44
fosmid18	8419	8462	CGGGAA	43
fosmid18	30847	30890	AT	43
fosmid18	5355	5397	TC	42
fosmid18	31070	31109	AGAGAGAGAA	39
fosmid18	36630	36666	GA	36
fosmid18	3769	3804	GACGGACAGACAGACG	35
fosmid18	6397	6432	ATGAG	35
fosmid18	22220	22255	TATTTAAATAAGGT	35
fosmid18	31066	31099	AG	33
fosmid18	176	208	T	32
fosmid18	29949	29976	TGT	27
fosmid18	5941	5966	GGGACTGTGGCCC	25

Repeat Masker Output

query sequence	*posit begin*	*ion end*	*matching repeat*	*repeat class/family*	*ID*	
fosmid18	1567	1637	dmoj.1.54.centroid	DNA	1	
fosmid18	3309	3951	dmoj.0.14.centroi	Satellite	2	
fosmid18	4029	4837	dmoj.1.77.centroid	DNA	3	
fosmid18	6919	7064	dmoj.0.64.centroi	Satellite	4	
fosmid18	7093	7166	dmoj.0.83.centroi	Satellite	5	*
fosmid18	7103	7198	dmoj.8.25.centroid	DNA	6	
fosmid18	7142	7213	dyak.0.22.centroid	DNA	7	*
fosmid18	10005	10083	dmoj.5.80.centroid	DNA	8	
fosmid18	10080	10191	dmoj.7.58.centroid	DNA	9	*

fosmid18	10187	10230	dmoj.7.20.centroid	DNA	10	*
fosmid18	10223	10331	dmoj.7.58.centroid	DNA	9	
fosmid18	10328	10438	dmoj.7.20.centroid	DNA	11	*
fosmid18	11226	11285	dmoj.0.61.centroi	Satellite	12	*
fosmid18	11233	11512	dmoj.5.28.centroid	Unknown	13	
fosmid18	11526	11927	dmoj.1.12.centroid	DNA	14	
fosmid18	11988	12076	dmoj.8.5.centroid	DNA	15	
fosmid18	13511	13653	(CTG)n	Simple_repeat	16	
fosmid18	14040	14119	dmoj.0.51.centroi	Satellite	17	
fosmid18	16305	16436	(TA)n	Simple_repeat	18	
fosmid18	17187	17477	dmoj.7.58.centroid	DNA	19	
fosmid18	17453	17510	dmoj.7.58.centroid	DNA	9	*
fosmid18	17584	17680	dmoj.7.58.centroid	DNA	9	
fosmid18	17817	18002	G2_DM	LINE/Jockey	20	
fosmid18	18106	18579	G2_DM	LINE/Jockey	20	
fosmid18	18580	18822	G_DM	LINE/Jockey	21	
fosmid18	19271	20021	dmoj.1.77.centroid	DNA	22	
fosmid18	20085	20180	dmoj.7.20.centroid	DNA	23	
fosmid18	20558	20638	dmoj.7.20.centroid	DNA	9	
fosmid18	20814	21099	dmoj.0.96.centroi	Satellite	24	
fosmid18	21108	21207	dmoj.5.80.centroid	DNA	25	
fosmid18	21222	21768	dmoj.7.20.centroid	DNA	9	
fosmid18	21838	21908	dmoj.1.12.centroid	DNA	9	*
fosmid18	21907	21958	dmoj.8.5.centroid	DNA	26	
fosmid18	23442	23559	(CAT)n	Simple_repeat	27	
fosmid18	24008	24082	(CTG)n	Simple_repeat	28	
fosmid18	24393	24406	dmel.8.13.centroid	LTR	29	
fosmid18	24407	24513	TART_DV	LINE/telomeric	30	
fosmid18	24514	24525	dmel.8.13.centroid	LTR	29	
fosmid18	25330	25439	dmoj.1.50.centroi	Satellite	31	
fosmid18	25919	26249	dwil.16.121.centroi	d DNA	32	
fosmid18	27104	27175	dmoj.1.12.centroid	DNA	33	*
fosmid18	27173	27633	dmoj.1.12.centroid	DNA	33	
fosmid18	29802	29856	dmoj.0.64.centroi	Satellite	34	
fosmid18	30117	30190	dmoj.1.54.centroid	DNA	35	
fosmid18	30193	30591	dmoj.1.50.centroi	Satellite	36	
fosmid18	30526	30608	dmoj.1.16.centroid	DNA	37	*
fosmid18	30609	30622	ISFUN1	LINE/Penelope	38	
fosmid18	30835	30920	(TA)n	Simple_repeat	39	
fosmid18	31969	32027	dmoj.0.99.centroi	Satellite	40	
fosmid18	32017	32106	dana.6.5.centroid	Helitron	41	*
fosmid18	32132	32192	dmoj.7.10.centroi	Satellite	42	*
fosmid18	32144	32532	dmoj.7.58.centroid	DNA	43	

fosmid18	32556	33050	dmoj.7.58.centroid	DNA	43	
fosmid18	33113	33375	dmoj.7.10.centroi	Satellite	44	
fosmid18	33399	33604	dmoj.7.20.centroid	DNA	43	
fosmid18	33605	33655	dmoj.31.44.centroid	DNA	45	*
fosmid18	33634	33708	dmoj.0.90.centroi	Satellite	46	
fosmid18	33718	33799	dmoj.0.64.centroi	Satellite	47	
fosmid18	33966	34059	dmoj.1.12.centroid	DNA	48	*
fosmid18	34054	34481	dmoj.1.12.centroid	DNA	48	
fosmid18	34610	34666	INV2j_DB	UnknownRepbaseI	49	*
fosmid18	34663	34787	dmoj.1.77.centroid	DNA	50	
fosmid18	34779	34821	dmoj.8.5.centroid	DNA	51	*
fosmid18	34926	34983	dmoj.5.80.centroid	DNA	52	
fosmid18	35006	35434	dmoj.0.83.centroi	Satellite	53	*
fosmid18	35135	35435	dmoj.8.5.centroid	DNA	51	
fosmid18	37303	37366	(TATATG)n	Simple_repeat	54	
fosmid18	37367	37391	dmoj.0.64.centroi	Satellite	55	
fosmid18	38652	38772	dmoj.7.10.centroi	Satellite	56	*
fosmid18	38753	38954	dmoj.7.20.centroid	DNA	43	
fosmid18	39290	39370	(TC)n	Simple_repeat	57	