Annotating *Drosophila grimshawi* Fosmid DGA05E01

Jimmy Ma
Bio 434W
Professor Elgin
May 3, 2010

Overview

Annotation is the process of describing and analyzing all the features in a genomic sequence including any genes, pseudogenes, and repeats. Here, I describe the preliminary steps towards annotating DGA05E01, a 40 kB fosmid on the dot chromosome of *Drosophila grimshawi*. Because the dot chromosome expresses genes despite being largely heterochromatic, it represents an especially interesting subject for comparative studies among different *Drosophila* species to better understand genomic controls of gene expression [1]. GENSCAN predicts five potential coding sequences in this region (Figure 1a) including orthologs to *eIF4G*, *mGluRA*, *CG32016*, *mtt*, and *CG4847*. A total of four annotated features were found including orthologs to *mGluRA*, *CG32016, CG11093*, and *CG5367*. Of the predicted features, *eIF4G*, *mGluRA*, and *CG32016* are located on the dot chromosome in *D. melanogaster* while *mtt* and *CG4847* are located on chromosome 2R in *D. melanogaster*. The *eIF4G* ortholog may not be a real ortholog based on a possible duplication event. There is relatively low repeat content in the region totaling about 4.31%. Synteny is conserved here between *D. grimshawi* and *D. melanogaster* with the same relative order and orientation of genes except for the presence of *CG5367* which is on chromosome 2L in *D. melanogaster*. ClustalW analysis showed that mGluRA has well conserved domains in five species of Drosophila and in mouse [3]. However, there was no conserved regulatory sequence within the first 2kB of the respective start sites.

Feature – Location – Gene Span Size (bp)
eIF4G: spans off fosmid-5962 – 5962 bp in fosmid DGA05E01
CG32016: 19680-15637 – 4044 bp
mGluRA: 10136-13817 – 3682 bp
CG5367: 28810-27510 – 1301 bp
CG11093: 24324-21242 – 3083 bp

Table 1. Summary of features.

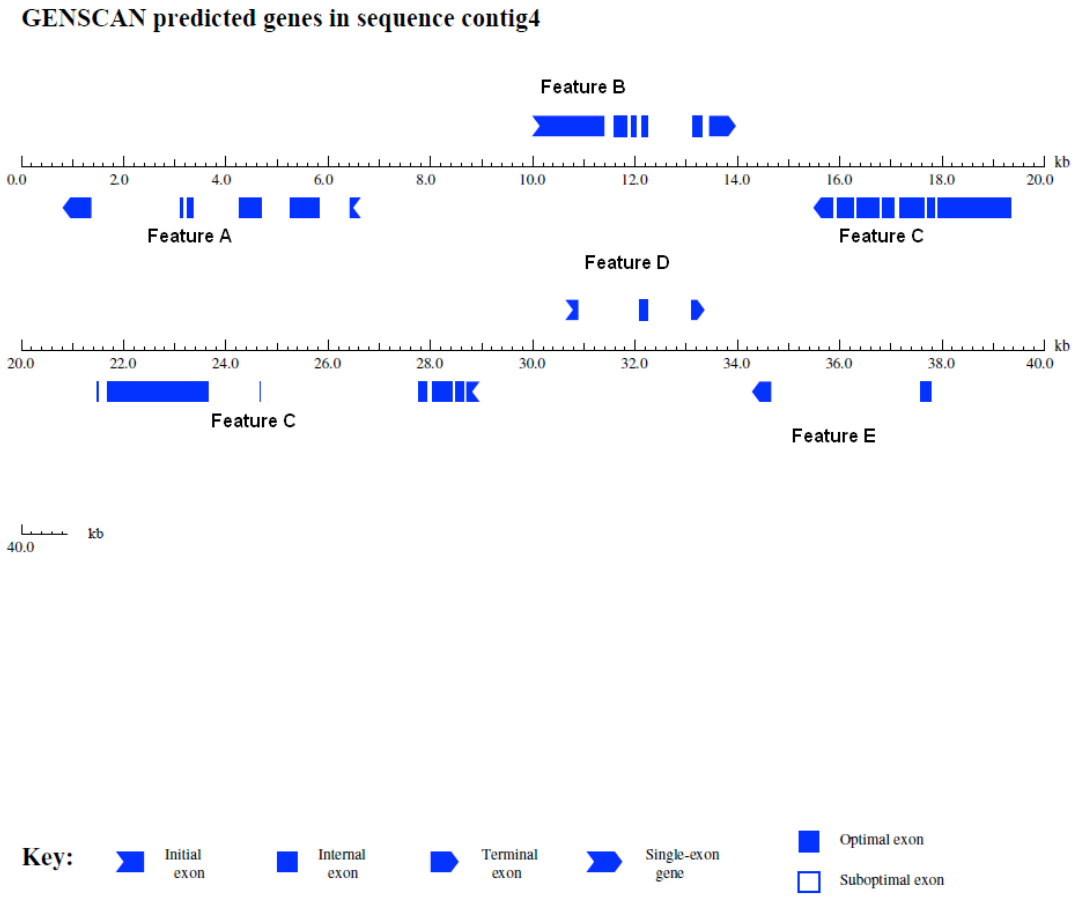


Figure 1a. Initial GENSCAN feature predictions for DGA05E01

.

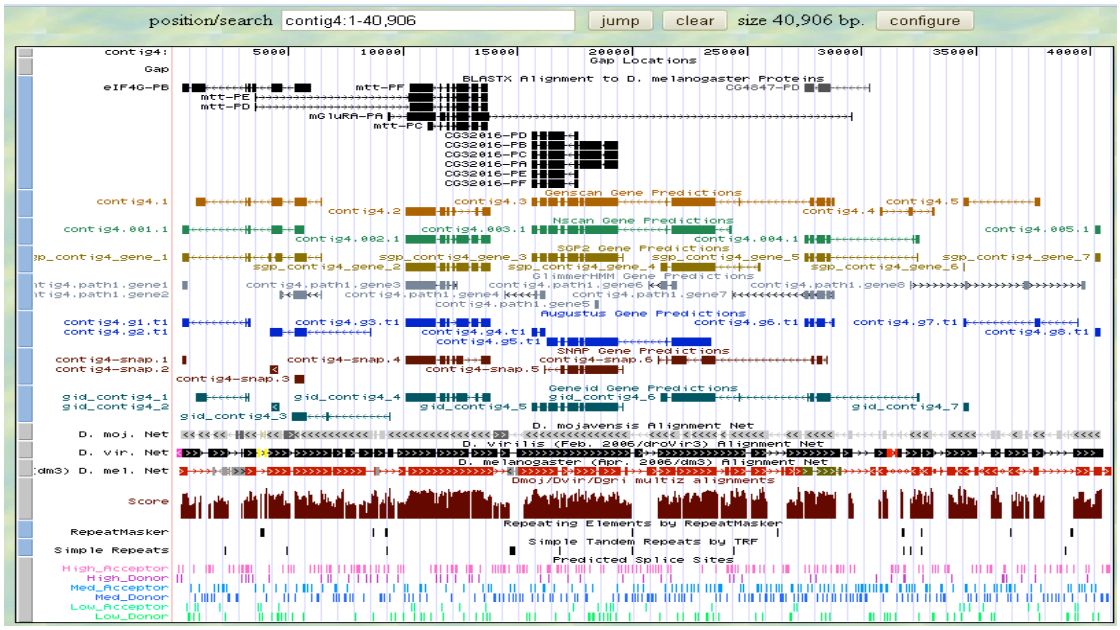


Figure 1b. UCSC Browser image of DGA05E01.

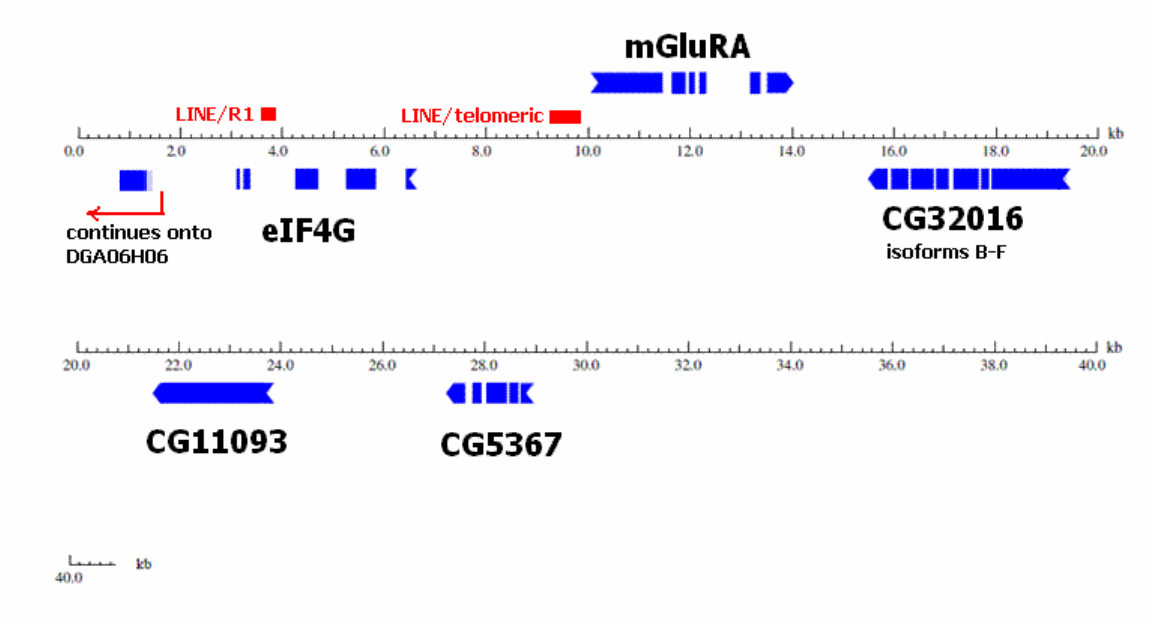


Figure 2. Final map of annotated features.

Genes

Feature A

In this paper, I will refer to each of the *D. grimshawi* orthologous exons in order using a numerical notation corresponding to the order from the start codon (e.g., exon A.1 refers to feature A's first exon with the start codon). I will refer to exons in *D. melanogaster* orthologs using an abbreviated form of the exon name as found in Gene Record Finder (e.g., exon eIF4G:FBgn0023213:15 is exon 15) [4]. The *D. melanogaster* exons will be introduced with the first mention of the corresponding *D. grimshawi* exon.

Starting with the features found on the *blastx* track on the local UCSC Genome Browser, I aligned the full fosmid DNA sequence against all the translated *eIF4G* exons found in *D. melanogaster* on the Gene Record Finder using *blastx* to see if all the exons truly were represented on the browser view (*eIF4G* accession #: NP_001096852) (Figure 3a) [5,6,7]. Of the fifteen exons in *D. melanogaster* ortholog of *eIF4G*, only the first seven were completely or partially in this fosmid with only four of those seven having significant results in this *blastx* search. Because the most significant hits map to the start of my fosmid and the feature runs in the negative direction, much of the putative coding sequence can be expected to be outside of my fosmid boundaries (Figure 3b).

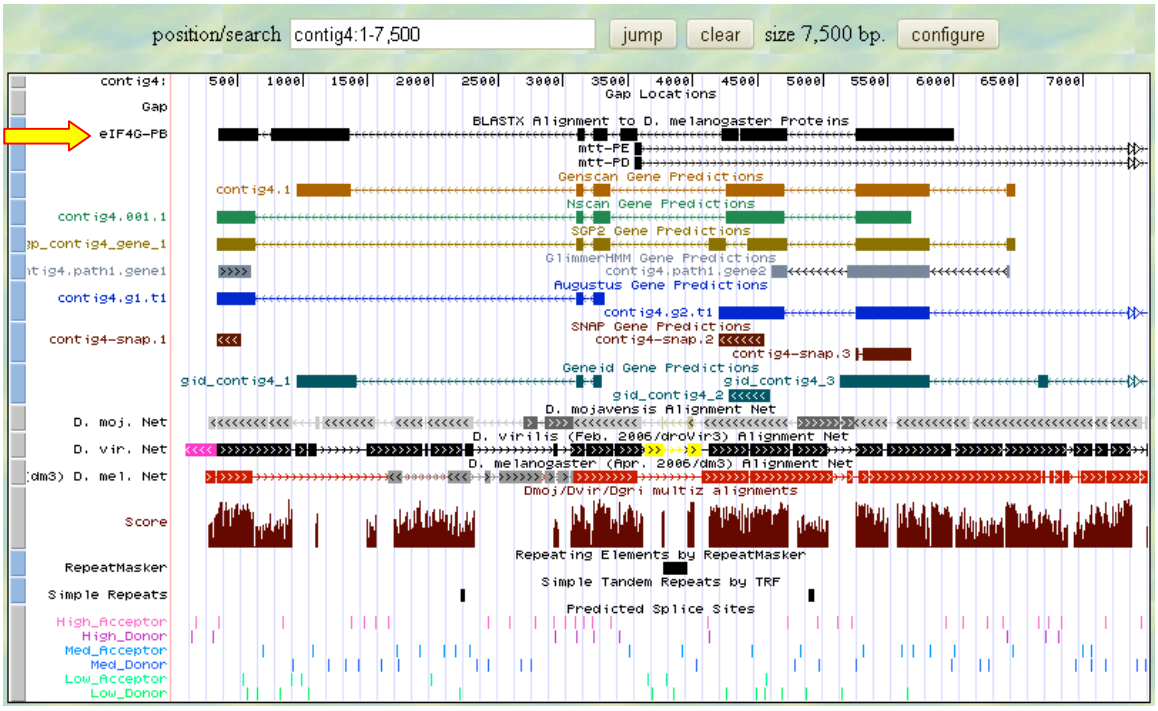


Figure 3a. *Blastx* track on UCSC Genome Browser. The window view covers bp 1-7500 of the fosmid. Arrow points to feature A which potentially runs past the start of the fosmid.

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
50165	eIF4G:CDS_FBgn0023213:15_924	17.7	49.3	0%	1.8	50%	
50166	eIF4G:CDS_FBgn0023213:14_924	112	319	2%	4e-28	62%	
50167	eIF4G:CDS_FBgn0023213:13_924	95.9	485	3%	5e-52	72%	
50168	eIF4G:CDS_FBgn0023213:12_924	17.3	64.3	0%	2.3	71%	
50169	eIF4G:CDS_FBgn0023213:11_924	49.3	64.7	0%	5e-10	66%	
50170	eIF4G:CDS_FBgn0023213:10_924	24.3	86.6	0%	0.014	83%	
50171	eIF4G:CDS_FBgn0023213:9_924	124	235	2%	3e-31	70%	
50172	eIF4G:CDS_FBgn0023213:8_924	22.3	414	2%	0.11	87%	
50173	eIF4G:CDS_FBgn0023213:7_924	16.9	111	0%	3.0	66%	
50174	eIF4G:CDS_FBgn0023213:6_924	20.0	132	1%	1.5	50%	
50175	eIF4G:CDS_FBgn0023213:5_924	18.9	157	1%	1.1	57%	
50176	eIF4G:CDS_FBgn0023213:4_924	17.3	83.1	0%	2.3	55%	
50177	eIF4G:CDS_FBgn0023213:3_924	20.0	150	1%	0.48	75%	
50178	eIF4G:CDS_FBgn0023213:2_924	17.3	49.6	0%	2.3	80%	
50179	eIF4G:CDS_FBgn0023213:1_924	18.1	48.9	0%	1.1	85%	

Figure 3b. *Blastx* hits of DGA05E01 DNA against all *D. melanogaster eIF4G* exons. Boxed hits have significant E-values that are used as anchors to find the exons that lie in between the significant exons. Note the high E-values for the exons lower on the list. The high E-values along with large coordinates outside those of earlier exons suggest that those corresponding exons (exons 1-8) do not exist within the region.

Each of these hits was investigated in more detail on an exon-by-exon basis using the local UCSC Genome Browser, Gene Record Finder, and BLAST analysis. Because *blastx* results do not show any significant hit for exon A.1 (*eIF4G:15* in *D. melanogaster*), I first found the exon boundaries for exon A.2 (*eIF4G:14*) using its significant *blastx* hit and then extrapolated upstream to locate the best match for exon A.1. However, though the two best hits for exon A.2 together cover residue 2 to 247 of exon 14, they have overlapping coordinates, lie in different frames, and the best hit has a premature stop codon (Figure 3c). Even when conserving length and using the second hit as an anchor for the rest of the exon, the predicted exon would have premature stop

codons in the sequence. When accounting for a possible intron insertion into this region, even the smallest possible intron of 47 bp from bp 5867 to 5821 overlaps with regions that match well with the *D. melanogaster* sequence. This does not mean that an intron insertion is not possible; it only points to a less likely explanation. Furthermore, it seems unlikely that such a large insertion would maintain the high level of conservation in the overlap region between these two hits. Together, these results suggest that the *D. grimshawi* ortholog of isoform B with exon A.2 may be a pseudogene and highly evolving (Figure 3d).

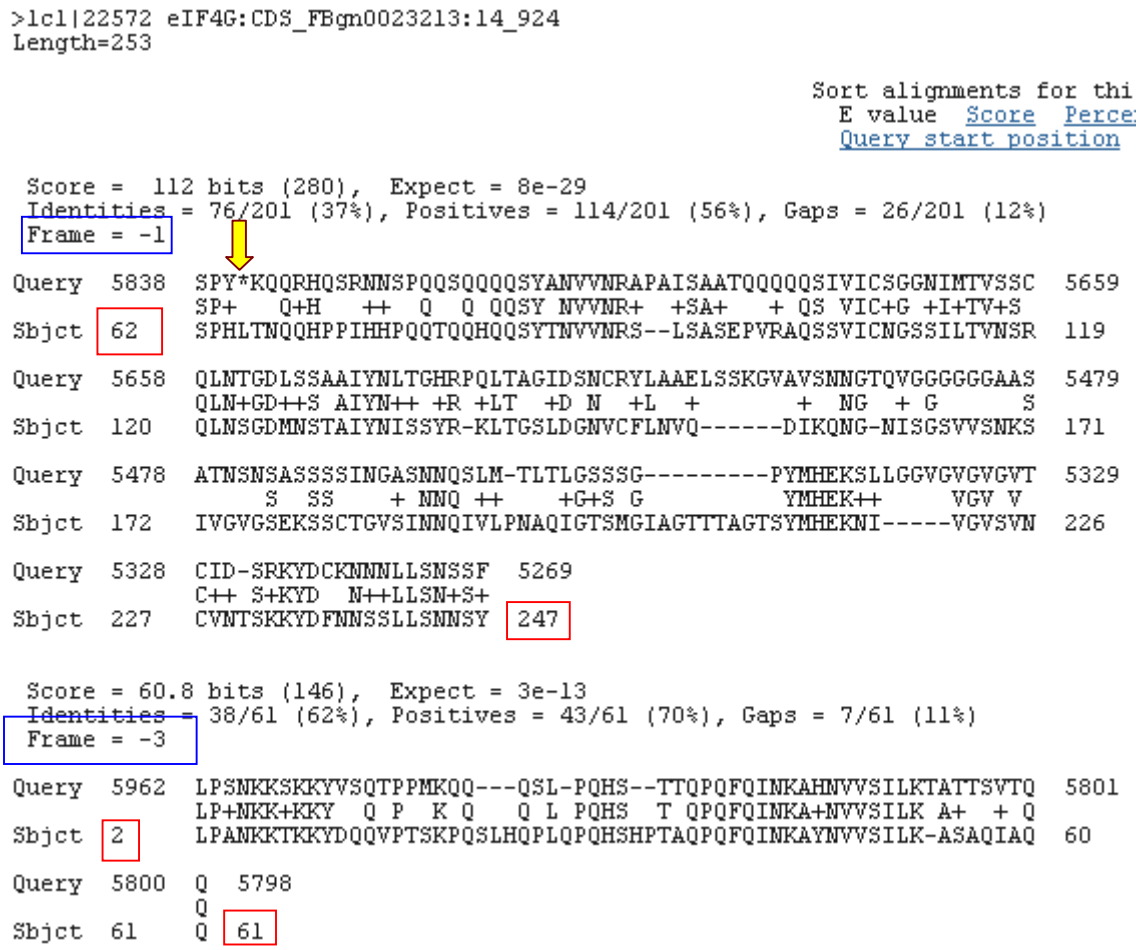


Figure 3c. *Blastx* results of DGA05E01 against *eIF4G:14_924* (exon A.2). Red boxes show high coverage of subject exon 14. Note the difference in frame (blue boxes), the overlapping coordinates of the matches, and the stop codon in the best hit (arrow).

Isoforms / ID	CDS_FBgn0023213:15_924	CDS_FBgn0023213:14_924	CDS_FBgn0023213:13_924	CDS_FBgn0023213:12_924	CDS_FBgn0023213:11_9
Unique	X	X	X	X	X
eIF4G-RA	X		X	X	X
eIF4G-RB	X	X	X	X	X
eIF4G-RC	X		X	X	X

Figure 3d. *eIF4G* isoforms in *D. melanogaster*. Note only isoform B has exon 14. In *D. grimshawi*, isoform B may be a pseudogene because of the premature stop codon found in Figure 3c.

When examining the remaining significant *blastx* hits, the same fragmented coverage also occurs in exon A.3 (*eIF4G:13*). Like exon A.2, exon A.3's hits occur in different frames and have overlapping coordinates (Figure 3e). Though there are no premature stop codons in the initial *blastx* alignments, I still found premature stop codons in every frame using any of the four hits as anchors while trying to conserve the length of exon A.3. These premature stop codons occurred fairly close to each match and usually constricted the putative exon to almost the length of the hit. The only anchor that had a longer extension was hit 3 (Figure 3f). However, the predicted exon's residues would be radically different from those of exon 13 and with the presence of significant unmodified portions of exon 13 in other frames, it seems more likely that exon A.3 is part of a pseudogene.

```
>lcl|27203 eIF4G:CDS_FBgn0023213:13_924
Length=217

                                                          Sort alignments for thi
                                                            E value  Score  Perce
                                                            Query start position

 Score = 95.9 bits (237),  Expect(4) = 5e-52
 Identities = 45/85 (52%), Positives = 56/85 (65%), Gaps = 9/85 (10%)
 Frame = -3

Query  4604  RHLHMPAMYP-----NVVLQQYSQYGQRQPTFQTPQIQYGPAPIPYYPYQYIPSLQQQPP  4440
              RH+H+  MY       N+VLQQY+QY  RQ TF     +QY PAP+PYY YQY+P+LQQQPP
Sbjct  41    RHVHVQPMYSQPLHQNMVLQQYTQYNPRQQTFPASHLQYAPAPMPYYQYQYVPTLQQQPP  100

Query  4439  PPPQHSRSGVATNASVKIGGNTMPV  4365
                  H+RS V  N +V +G N  PV
Sbjct  101   ----HTRSAVTVNTNVNVGNNLQPV  121


 Score = 59.7 bits (143),  Expect(4) = 5e-52
 Identities = 29/40 (72%), Positives = 32/40 (80%), Gaps = 1/40 (2%)
 Frame = -2

Query  4722  YVSSGNNSAGNTRSNQQSG-IFRGPPSTANAPRGTGAVGT  4606
              YVS+GNN++GNTRSN QSG IFRGPPST NAPRG     T
Sbjct  1     YVSTGNNNSGNTRSNPQSGGIFRGPPSTPNAPRGASGGAT  40


 Score = 54.3 bits (129),  Expect(4) = 5e-52
 Identities = 27/43 (62%), Positives = 33/43 (76%), Gaps = 2/43 (4%)
 Frame = -1

Query  4348  PGATTSQLQLFTGSVQTGANTALGVSSTGS--GQVGVPPMVNV  4226
              PGA++SQ+QL T +VQ GA+T +GV   GS  GQVGVPPMV V
Sbjct  132   PGASSSQIQLLTSTVQPGASTVMGVGGPGSTMGQVGVPPMVGV  174


 Score = 44.3 bits (103),  Expect(4) = 5e-52
 Identities = 21/30 (70%), Positives = 23/30 (76%), Gaps = 0/30 (0%)
 Frame = -2

Query  4218  VAMASRRRHRHSLQIIHPETKKNILGELDK  4129
              V  ASRRRH+H LQII P TKKNIL + DK
Sbjct  188   VQPASRRRHQHRLQIIDPTTKKNILDDFDK  217
```

Figure 3e. *Blastx* alignment for exon A.3 (DGA05E01 against exon 13). As in exon A.2 (Figure 3c), the coverage is fragmented over different frames with overlapping coordinates. From top to bottom, hit 1, 2, 3, and 4. These hit numbers correspond to those used in Figure 3f.

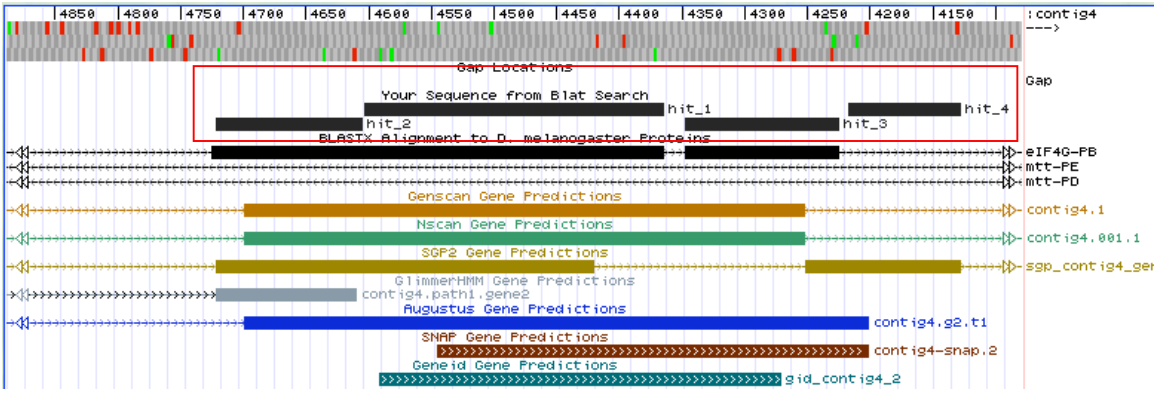


Figure 3f. Boxed area shows hits 1-4 and where they align against the fosmid. Note that in every single hit, any extension using the hit as an anchor runs into a stop codon except for hit 3 in frame -1. Figure 3e shows the various frames of each hit. Note also that the

GENSCAN prediction overlaps partially with hit 3 but does not completely cover it.

If both exon A.2 and A.3 contribute to pseudogenes, then according to Figure 3d, none of the orthologous isoforms of *eIF4G* represent real genes in DGA05E01. This is supported further after considering the possibility that this orthologous *eIF4G* lies in a putatively duplicated region which completely overlaps with the previously annotated fosmid DGA06H06. In the Repeats section, I consider the subject of a possible insertion or duplication in this area more closely. For now, the important fact is that after combining the DNA sequence from this region with that of DGA06H06, the *blastx* results against *eIF4G* exons show that every exon has a significant hit. These hits are all downstream of the original DGA05E01 matches and each covers the full length of its respective *D. melanogaster* counterpart. Jeanette Wong, who annotated DGA06H06, had previously annotated these hits as the *D. grimshawi* ortholog to *eIF4G*. The original DGA05E01 matches show up as less significant hits and still have the same problems as before (Figure 3g). When searching only DGA06H06 against the *D. melanogaster* exons, *blastx* returned the same results. The combined sequence was constructed by aligning the two fosmids and replacing the overlapping region with sequence from my fosmid. As such, because both searches return the same results, it seems highly likely that the best ortholog to *eIF4G* lies in DGA06H06 rather than in DGA05E01.

Even though there is already a better (and complete) ortholog to *eIF4G* in *D. grimshawi* further downstream (on the negative strand) of the matches in DGA05E01, it does not say whether or not the hits in DGA05E01 are part of a pseudogene or not. It only provides support for the fact that hits in DGA05E01 are present in addition to a recognizable ortholog to *eIF4G*. Further exploration of *eIF4G* orthologous exons in DGA05E01 using BLAST did not yield any better results to answer this question. For example, searching for exon A.1 provided no results after narrowing the *blastx* search region to 2 kB upstream of bp 5962, the earliest match point of exon A.2, and increasing the E-threshold to 1e+40. A full Needleman-Wunsch alignment of the same region from *D. grimshawi* and the coding sequence of exon 15 suggests that the best predicted location for exon A.1 would begin at bp 6484 and end at bp 6410 in DGA05E01 (Figure 3h) [8]. This alignment along with the remaining *blastx* alignments for DGA05E01 do not give better results than those from DGA06H06. The only indication of an incomplete ortholog is the lack of good *blastx* matches following those in DGA05E01 going into DGA06H06. All matches following exon 9 in DGA06H06 other than those in the *eIF4G* ortholog do not form a continuous *eIF4G* ortholog based on their coordinates.

```
>lcl|42123 eIF4G:CDS_FBgn0023213:13_924
Length=217

                                                    Sort alignments for this s
                                                      E value  Score  Percent
                                                      Query start position  Su

 Score =  271 bits (693),  Expect = 4e-76
 Identities = 142/222 (63%), Positives = 165/222 (74%), Gaps = 15/222 (6%)
 Frame = -1

Query  17490  YVSSGNNSAGNTRSNQQSG-IFRGPPPTANAPRGTGAGGPRHLHMPAMYP-----NVVLQ  17329
              YVS+GNN++GNTRSN QSG IFRGPP T NAPRG   G  RH+H+  MY      N+VLQ
Sbjct  1      YVSTGNNNSGNTRSNPQSGGIFRGPPSTPNAPRGASGGATRHVHVQPMYSQPLHQNMVLQ  60

Query  17328  QYSQYGQRQPTFQTPHIQYGPAPIPYYPYQYLPSLQQQPPPPPQHSRSGVATNASVNIGG  17149
              QY+QY  RQ TF   H+QY PAP+PYY YQY+P+LQQQPP    H+RS V  N +VN+G
Sbjct  61     QYTQYNPRQQTFPASHLQYAPAPMPYYQYQYVPTLQQQPP----HTRSAVTVNTNVNVGN  116

Query  17148  NTMPVQTGPNGQLAGPGATTSQLQLLTGSVQTGANTVIGVGATGS--GQVGVPPM--VSV  16981
              N  PVQ+GPNG L  PGA++SQ+QLLT +VQ GA+TV+GVG  GS  GQVGVPPM  V V
Sbjct  117    NLQPVQSGPNGPLPVPGASSSQIQLLTSTVQPGASTVMGVGGPGSTMGQVGVPPMVGVGV  176

Query  16980  MPPNVAQQPVQVVAPASRRRHQHRLQIIHPETKKNILDELDK  16855
              M  +V  QPVQ V PASRRRHQHRLQII P TKKNILD+ DK
Sbjct  177    MTTSVQPQPVQ-VQPASRRRHQHRLQIIDPTTKKNILDDFDK  217


 Score = 95.9 bits (237),  Expect(4) = 4e-52
 Identities = 45/85 (52%), Positives = 56/85 (65%), Gaps = 9/85 (10%)
 Frame = -3

Query  25612  RHLHMPAMYP-----NVVLQQYSQYGQRQPTFQTPQIQYGPAPIPYYPYQYIPSLQQQPP  25448
              RH+H+  MY      N+VLQQY+QY  RQ TF    +QY PAP+PYY YQY+P+LQQQPP
Sbjct  41     RHVHVQPMYSQPLHQNMVLQQYTQYNPRQQTFPASHLQYAPAPMPYYQYQYVPTLQQQPP  100

Query  25447  PPPQHSRSGVATNASVKIGGNTMPV  25373
                  H+RS V  N +V +G N  PV
Sbjct  101    ----HTRSAVTVNTNVNVGNNLQPV  121


 Score = 59.7 bits (143),  Expect(4) = 4e-52
 Identities = 29/40 (72%), Positives = 32/40 (80%), Gaps = 1/40 (2%)
 Frame = -2

Query  25730  YVSSGNNSAGNTRSNQQSG-IFRGPPSTANAPRGTGAVGT  25614
              YVS+GNN++GNTRSN QSG IFRGPPST NAPRG     T
Sbjct  1      YVSTGNNNSGNTRSNPQSGGIFRGPPSTPNAPRGASGGAT  40
```

Figure 3g. Representative alignment after combining fosmid DGA06H06 with DGA05E01. The new best hit for exon 13 is in DGA06H06 and aligns completely with moderate identity. The second hit is the best hit from DGA05E01 which also exists in DGA06H06 (and lies in the overlap region). The other hits from DGA05E01 (Figure 3e) are the subsequent hits after the second hit.

```
Dgri4_dna       1501 GATCGGTCCTTTCGGAATGCAACCAGCTGCTTCAGCACTATCAACACAAT   1550
                                 |||||||..|||....||.|..||.|||||||||
15                 1 ----------------ATGCAACAGGCTATACCAACTTTACCAACACAAT     34

Dgri4_dna       1551 TTTATATCGCTAAAAACATGCAGCCC----------CAAAATTTGGTATGT   1591
                     ...||||.|.||||..||||||||||         ||||||.||
15                35 CAGATATAGATAAAGCCATGCAGCCCCATTCAGCACAAAATATG------     78
```

Figure 3h. Alignment of bp 5962-8000 of DGA05E01and nucleotide sequence of exon 15 using *needle*, an EMBOSS tool from the Genomics Education Partnership (GEP) website. *Needle* performs a full Needleman-Wunsch global alignment with multiple sequences.

In the end, because of the premature stop codons in exon A.2 and A.3, the low identity between *eIF4G,* the presence of the better ortholog in DGA06H06, and the possible paralogous duplication event, I conclude that the *eIF4G* ortholog in DGA05E01 is most likely part of a pseudogene. If this area in *D. grimshawi* truly is part of a duplication, it does provide a possible explanation for how a pseudogene may have arisen. The differences between the paralog and ortholog would become more noticeable because the region would be under less purifying selection. However, because these conclusions are heavily based on whether or not there is a duplicated region here, further investigation should be done to confirm them.

Feature B

The *blastx* track on the UCSC Genome Browser suggests that Feature B is a potential ortholog of *D. melanogaster* protein mGluRA (accession #: NP_524639.2). To get the proper alignment data and coordinates, I aligned the entire fosmid DNA sequence against all the translated *D. melanogaster* exons using *blastx*. Results show eight high matching hits corresponding to each of the eight *D. melanogaster* exons for mGluRA found in the Gene Record Finder (Figure 4b).

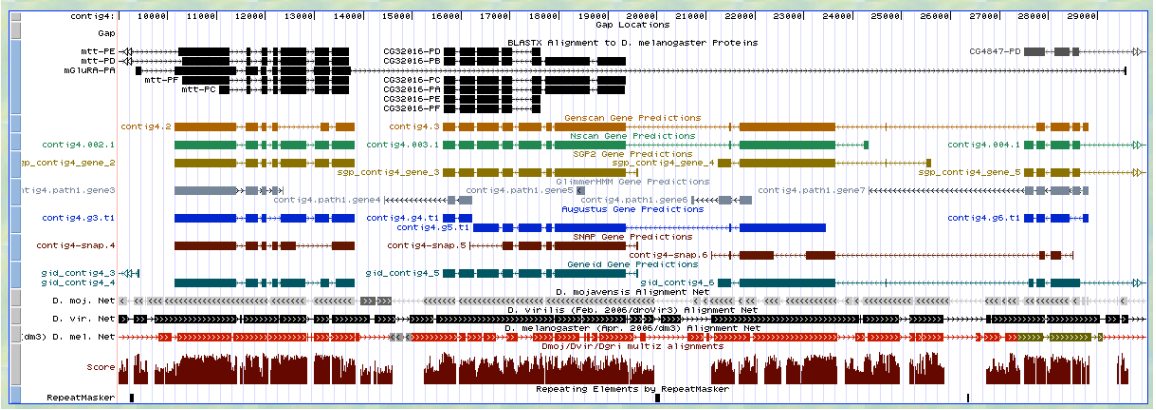


Figure 4a. Local UCSC Genome Browser view of mGluRA.

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
57820	mGluRA:CDS_CG11144:2_917	374	487	2%	5e-107	84%	
57821	mGluRA:CDS_CG11144:3_917	289	408	1%	1e-81	80%	
57822	mGluRA:CDS_CG11144:4_917	128	334	1%	1e-33	83%	
57823	mGluRA:CDS_CG11144:5_917	71.2	136	0%	1e-16	94%	
57824	mGluRA:CDS_CG11144:6_917	79.7	228	1%	4e-19	85%	
57825	mGluRA:CDS_CG11144:7_917	336	599	2%	1e-95	100%	
57826	mGluRA:CDS_CG11144:8_917	203	457	2%	5e-56	96%	
57827	mGluRA:CDS_CG11144:9_917	126	359	3%	1e-32	75%	

Figure 4b. *Blastx* hits of fosmid DNA against all *D. melanogaster* mGluRA exons.

Each of these hits was investigated in more detail. The *blastx* results show that exon B.1 in *D. grimshawi* in reality represents two *D. melanogaster* exons *(mGluRA:2* and *mGluRA:3).* Because the end of exon 2 is only 3 bp away from the start of exon 3 in the fosmid, it is unlikely that these two exons are far enough apart to have an intronic region in between. Furthermore, the two exons are in the same +2 frame, allowing for continuous translation to occur between the two exons (Figure 4c). These two *D. melanogaster* exons were annotated as a single exon in my fosmid, matching the results of all gene predictors. The start of the gene model occurs at bp 10193, just outside of the

matched region on *blastx* at the 11th residue (bp 10193) (Figure 4d). The predicted start site had fairly good matches based on a Needleman-Wunsch alignment using *needle* from the GEP website. The remaining exons (B.2-B.7) all had high coverage and identity of at least 80% or greater. They matched well with their *D. melanogaster* counterparts and the gene model was completed with 7 exons between bp 10136 and 13817 (Figure 4e).

```
>lcl|57820 mGluRA:CDS_CG11144:2_917
Length=229

                                                    Sort alignments for this sub
                                                      E value  Score  Percent id
                                                      Query start position  Subj

 Score =  374 bits (960),  Expect = 5e-107
 Identities = 186/220 (84%), Positives = 203/220 (92%), Gaps = 1/220 (0%)
 Frame = +2

Query  10193  LMLMVVTLWSCLPQLSGAGSSQSHDSVSVFLPGDIILGGLFPVHEKGEGAPPCGPKVYNR  10372
              L++++V  WS +  L    ++ + DSVSV LPGDIILGGLFPVHEKGEGAP CGPKVYNR
Sbjct  11     LVVVMVLSWSRVVDLKSPSNTHTQDSVSVSLPGDIILGGLFPVHEKGEGAP-CGPKVYNR  69

Query  10373  GVQRLEAMLYAIDRVNNDTNLLPGITIGVHILDTCSRDTYALNQSLQFVRASLNNMDTSV  10552
              GVQRLEAMLYAIDRVNND N+LPGITIGVHILDTCSRDTYALNQSLQFVRASLNN+DTS
Sbjct  70     GVQRLEAMLYAIDRVNNDPNILPGITIGVHILDTCSRDTYALNQSLQFVRASLNNLDTSG  129

Query  10553  FECSDTSTPQLRKNATSGPVFGVIGGSYSSVSLQVANLLRLFHIPQISPASTAKTLSDKS  10732
              +EC+D S+PQLRKNA+SGPVFGVIGGSYSSVSLQVANLLRLFHIPQ+SPASTAKTLSDK+
Sbjct  130    YECADGSSPQLRKNASSGPVFGVIGGSYSSVSLQVANLLRLFHIPQVSPASTAKTLSDKT  189

Query  10733  RFDLFARTVPPDTFQSVALVDIIKNLNWSYVSTIHSEGSY  10852
              RFDLFARTVPPDTFQSVALVDI+KN NWSYVSTIHSEGSY
Sbjct  190    RFDLFARTVPPDTFQSVALVDILKNFNWSYVSTIHSEGSY  229


>lcl|57821 mGluRA:CDS_CG11144:3_917
Length=180

                                                    Sort alignments for this sub
                                                      E value  Score  Percent id
                                                      Query start position  Sub

 Score =  289 bits (740),  Expect = 1e-81
 Identities = 146/181 (80%), Positives = 156/181 (86%), Gaps = 1/181 (0%)
 Frame = +2

Query  10856  EYGIEAFHKEATERHVCIAAAEKVPSASDDKIFDSIISKLLKKPNARGVILFTRAEDARR  11035
              EYGIEA HKEATER+VCIA AEKVPSA+DDK+FDSIISKL KKPNARGV+LFTRAEDARR
Sbjct  1      EYGIEALHKEATERNVCIAVAEKVPSAADDKVFDSIISKLQKKPNARGVVLFTRAEDARR  60

Query  11036  ILLAAKRANLSQPFHWVASDGWGKQQKLLEGLEEIAEGAITVELQSEIIEDFDRYMMQLT  11215
              IL AAKRANLSQPFHW+ASDGWGKQQKLLEGLE+IAEGAITVELQSEII DFDRYMMQLT
Sbjct  61     ILQAAKRANLSQPFHWIASDGWGKQQKLLEGLEDIAEGAITVELQSEIIADFDRYMMQLT  120

Query  11216  PRSNQRNPWFAEYWEDTFNCVLSLGSVDAKLAIDGEDGHSKSGNEKEKTICDDSLRLSEK  11395
              P +NQRNPWFAEYWEDTFNCVL+  SV    +       +K G  K KT CDDS RLSEK
Sbjct  121    PETNQRNPWFAEYWEDTFNCVLTSLSVKPDTSNSANSTDNKIG-VKAKTECDDSYRLSEK  179

Query  11396  V  11398
              V
Sbjct  180    V  180
```

Figure 4c. Exon B.1 represents exon 2_917 and 3_917 from *D. melanogaster*. The start and end sites are only 3 bp apart, much less than the 43 bp cutoff for introns. Note that both exons are in the same frame.

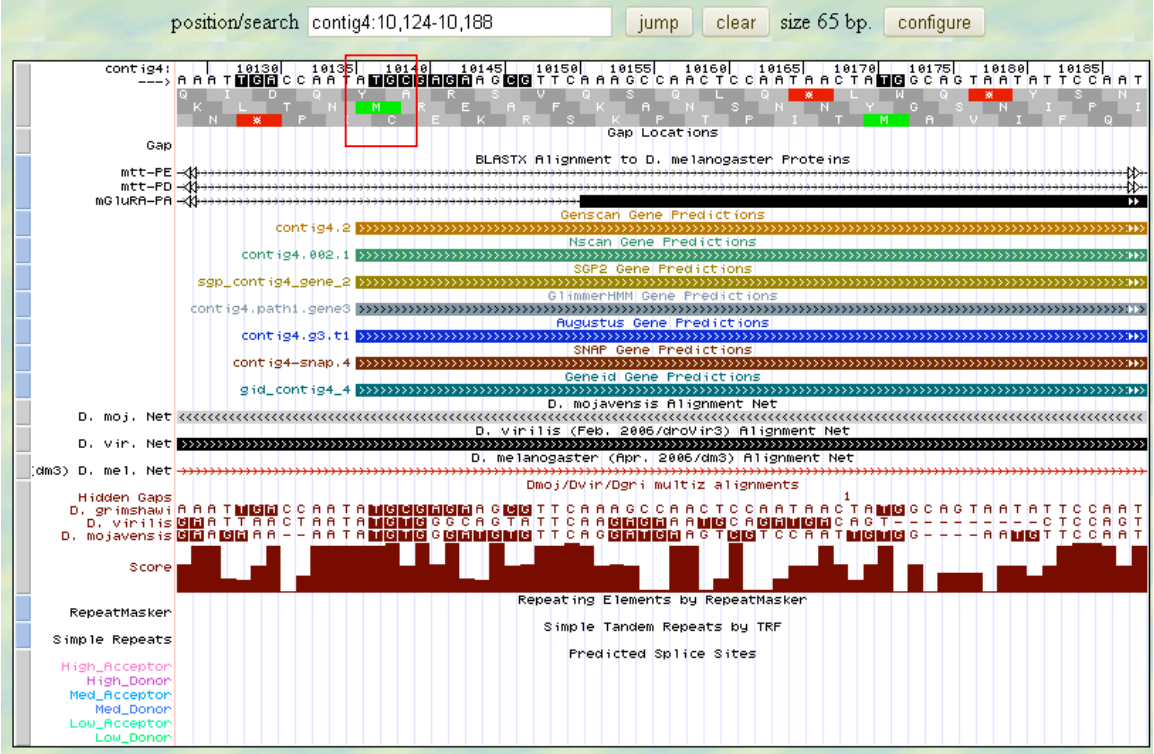


Figure 4d. Red box denotes start site of gene model for mGluRA.

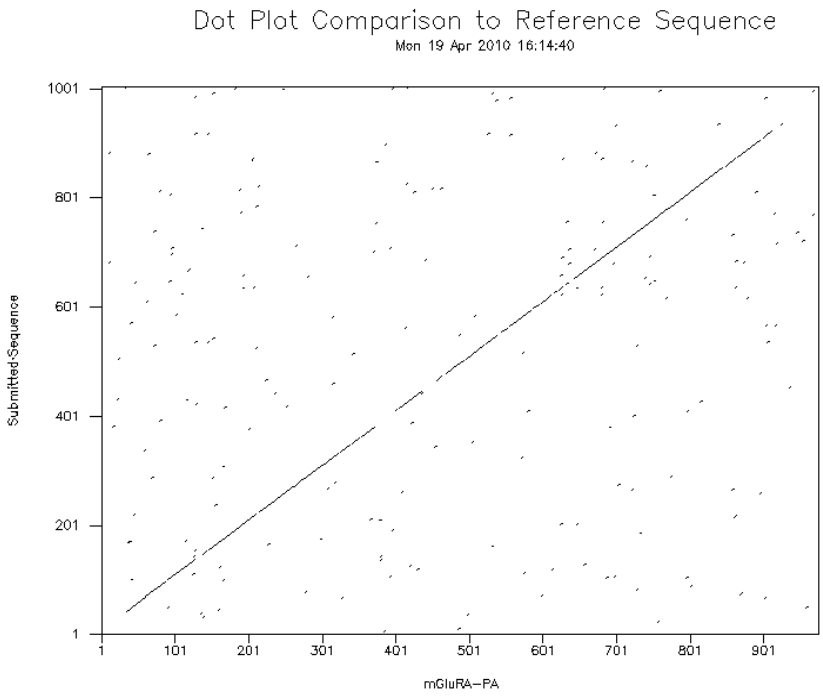


Figure 4e. Dot plot of gene model in *D. grimshawi* shows fairly good homology and alignment to *mGluRA* in *D. melanogaster*.

Based on the *blastx* track on the local UCSC Genome Browser, GENSCAN feature B also corresponds to the *mtt* gene in *D. melanogaster* (accession #: NP_610386,

NP_724683, NP_995780, NP_001137619). One possible explanation for *mtt* is simply an incorrect association with this region because of a good match in sequence. Based on a literature search, *mGluRA* is the gene coding for the metabotropic glutamate receptor. It is found on chromosome four in *D. melanogaster* and has many conserved domains throughout its sequence (Figure 4g). Similarly, *mtt* (mangetout) is found on chromosome 2R in *D. melanogaster,* has associated G-protein coupled receptor activity, and has many of the same conserved domains as *mGluRA*. Because both of these proteins have similar functionality and conserved domains, I did a *blast2seq* alignment between the two peptide sequences to find which areas match between the two proteins and found matching up to 80% identity in certain regions (Figure 4i). Comparing the dot plot and the conserved domain diagrams in Figures 4g and 4h, the only areas that match between *mtt* and *mGluRA* are the conserved domains. Furthermore, the rest of *mtt* does not match significantly at all in DGA05E01 (roughly 40% of the protein). Based on these observations, it seems likely that *mtt* was incorrectly matched to this region because of its close identity with *mGluRA* at those conserved domains.

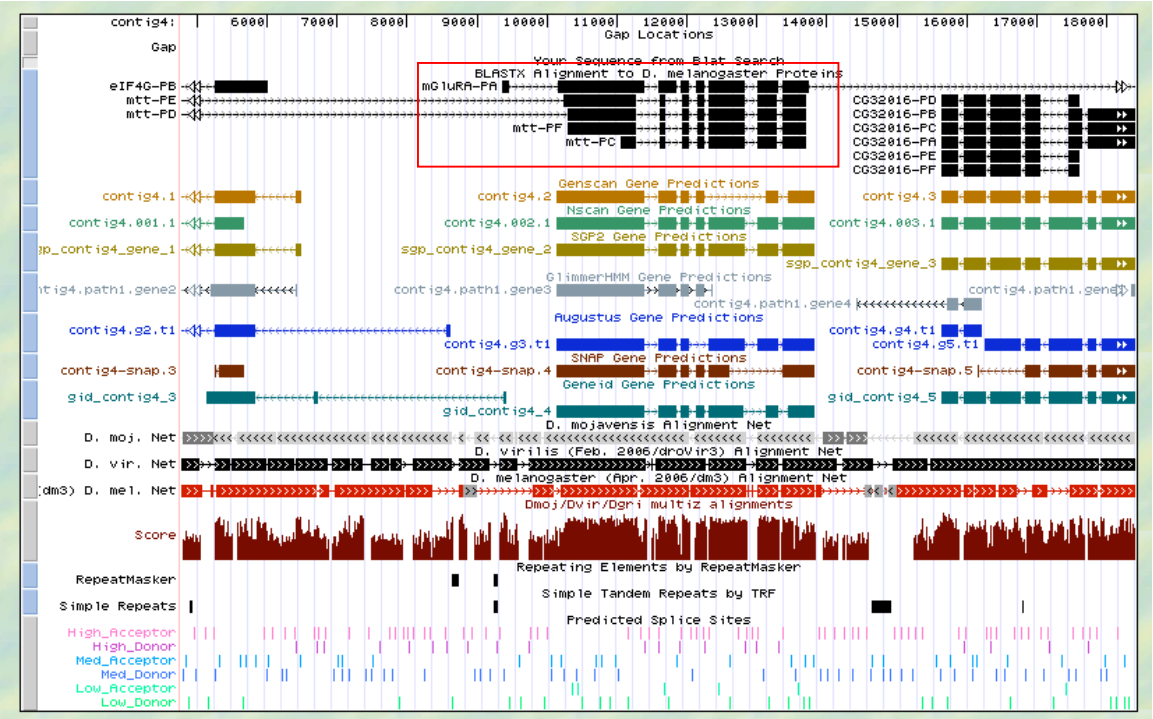


Figure 4f. Boxed region shows *blastx* hits for both *mtt* and *mGluRA* corresponding to the same region and same GENSCAN prediction.

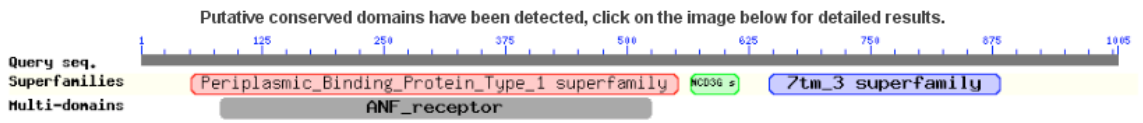


Figure 4g. Conserved domains in mGluRA.

Figure 4h. Conserved domains in mtt isoform D. Note the similarity in conserved domains as that in mGluRA.

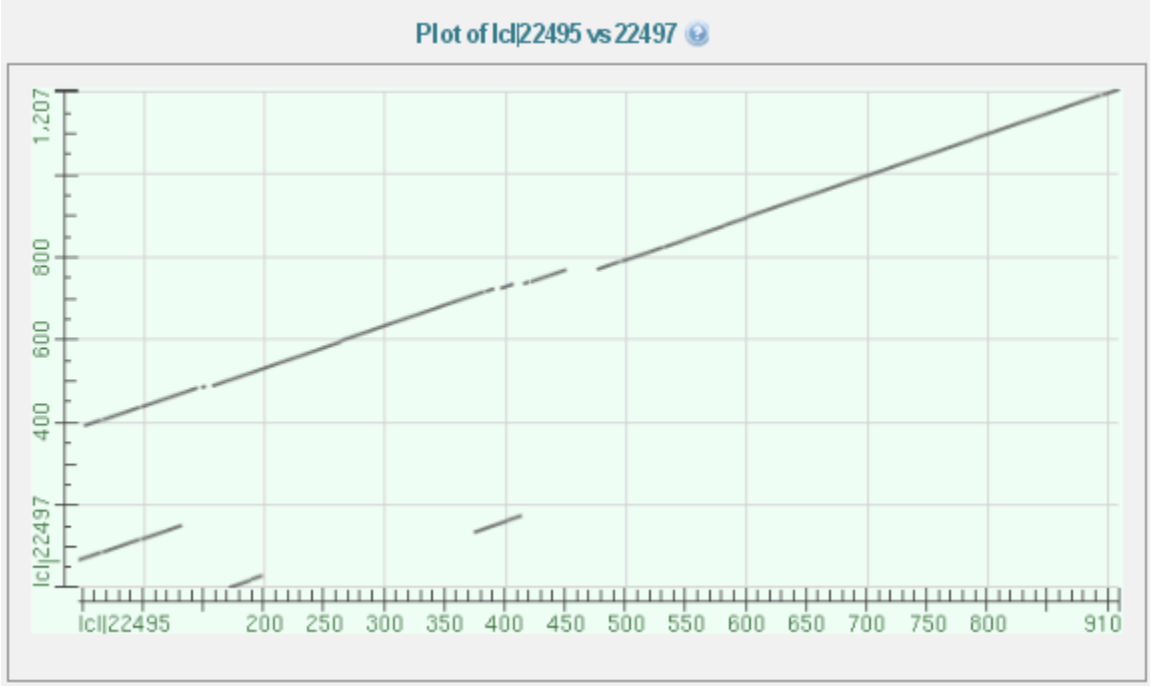


Figure 4i. Dot plot of mtt isoform D (vertical axis) and mGluRA (horizontal axis) peptide sequence. Note that the high matching occurs in the region of the conserved domains (~400-1207 bp in mtt isoform D and ~75-910 bp in mGluRA) that correspond to the PBP1/ANF receptor and 7tm_3 superfamily regions.

Feature C

The *blastx* track from the local UCSC Genome Browser suggests that Feature C represents *D. grimshawi* ortholog to *CG32016* in *D. melanogaster*. According to the Gene Record Finder, there are a total of six different isoforms of *CG32016* with three different start sites. *Blastx* results for fosmid DNA against CG32016 show hits corresponding to seven of the nine different exons in *D. melanogaster* (Figure 5a).

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
35692	CG32016:CDS_CG32016:3_914	274	335	3%	2e-76	53%	
35693	CG32016:CDS_CG32016:4_914	35.0	67.4	0%	1e-05	41%	
35694	CG32016:CDS_CG32016:5_914	57.8	204	2%	5e-12	75%	
35695	CG32016:CDS_CG32016:5_915	39.3	185	2%	2e-06	75%	
35696	CG32016:CDS_CG32016:6_914	48.5	112	0%	1e-09	75%	
35697	CG32016:CDS_CG32016:7_914	99.0	269	2%	2e-24	58%	
35698	CG32016:CDS_CG32016:8_914	54.3	376	3%	4e-11	87%	
35699	CG32016:CDS_CG32016:9_914	108	195	1%	9e-28	72%	

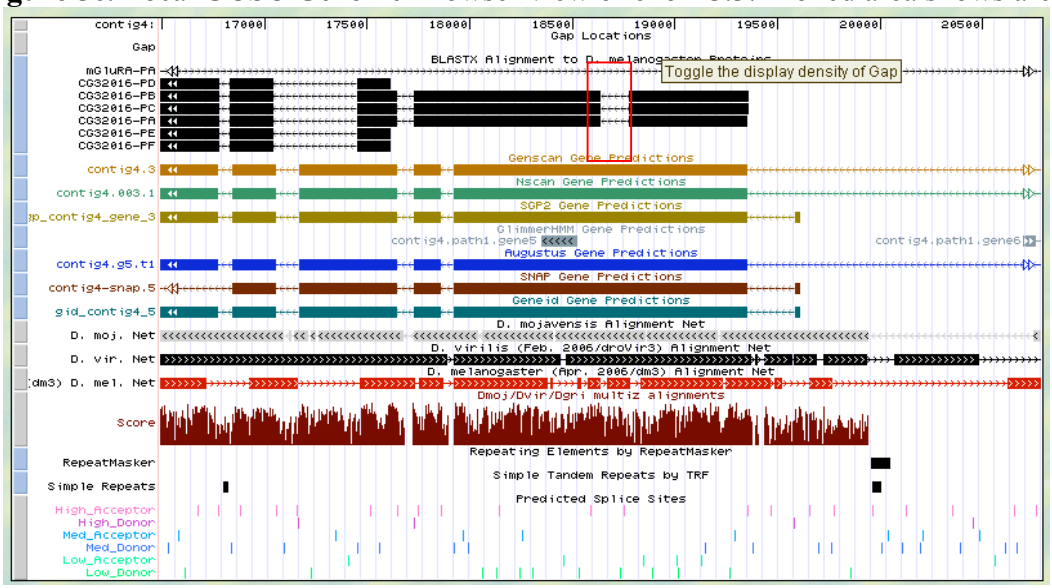
Figure 5a. *Blastx* hits of fosmid DNA against CG32016 peptide sequence.

Exons C.1 (*CG32016:10_915*) and C.2 (*CG32016:13_914*) were initially not included with the *blastx* results. In order to find the locations of these two exons, I first determined the exon boundaries for exon C.3 (*CG32016:3_914*) using the *blastx* results (Figure 5b). Based on these results and agreement with several gene predictors on the local UCSC Genome Browser, the boundaries for exon C.3 were set between bp 19355 and 17921 (Figure 5c). Although the *blastx* track in Figure 5c show a region of low similarity in the middle of exon C.3, six gene predictors, results from NCBI *blastx*, and the lack of repetitious sequence suggest that this is the best model for the exon. Finally, the moderate level of identity (40%) suggests some divergence between the two sequences somewhat explains the slightly shorter predicted exon (443 residues instead of 448 residues).

```
>lcl|35692 CG32016:CDS_CG32016:3_914
Length=448

                                                          Sort alignments for this s
                                                            E value  Score  Percent
                                                            Query start position  Su

 Score =  274 bits (700),  Expect = 2e-76
 Identities = 207/511 (40%), Positives = 266/511 (52%), Gaps = 101/511 (19%)
 Frame = -1

Query  19354  YSRVDLLALRYEDSSRRRPSCANRTELQKLNFWKINGLPSNLSNNNSNSCSSYGNKTGLS  19175
              YS+VDLLALRYE  SR+RP C+ R ELQ L FWKIN    N +    +S  S  NK  LS
Sbjct  1      YSKVDLLALRYEGKSRQRPQCSTRLELQTLGFWKIN---LNTAALTVSSAYSNQNKNRLS  57

Query  19174  PERENVSLTSSN-GLLSSRRALRNRERAHNYYQRFTAGDQI---GEDAQASLASLSGGNL  19007
              PE +N SL  SN   +SSRRA+RNRERA+NYYQRF   D +   GED      +LS G
Sbjct  58     PEADNSSLICSNSSSISSRRAMRNRERANNYYQRFVPTDSLLISGEDKDKD--ALSHGQ-  114

Query  19006  CSASYKSANIDHRSISSSHLMPAFAKRRFVVAIGPNETQSEKANYEDGLGSSAKENTSIS  18827
                  YK   IDHRSISSSHLMPAFAKRRFV++ G N   SE++N  +G+ + A +
Sbjct  115    ---PYKLNIIDHRSISSSHLMPAFAKRRFVISKGSN---SEESN--EGINTCASKG----  162

Query  18826  CQANADLKNTTRSTLSSPMRRTPAASAEWERSEKHSNF-QPDSEQLLALSPTFLSGKQGT  18650
                          SSP R+     +E + +E   NF QPD +Q ++ SPTF + +
Sbjct  163    ------------KAASSPSRK----GSELDTAETCLNFVQPDHDQCMSSSPTFSTSR---  203

Query  18649  NLNVQERRIGSGRLLPRNDNWEYKN-KDADSSLANLEKDRPQNGGIV--------IQQRQ  18497
                  QERRIGSGRLLPR+DNW+YKN K  ++S+ N ++  P   G             Q R
Sbjct  204    ----QERRIGSGRLLPRSDNWDYKNEKTVEASIENEKETSPNGSGSTSSLNQHNQSQHRS  259

Query  18496  RTCSTKHSDRSNDILGDRDRDRDRERDRDRDRRINERSRDPNEGKKNMLSGRRATNRDKF  18317
              RT S +  +R  ++  DR    D ++  DR                 ++ RR + ++ F
Sbjct  260    RTFSGRLVERVPEVT-DRRFQYDSKKSFDRQG----------------INNRRISGKEPF  302

Query  18316  NFGDSTMQNRGKRMNMYQTHDRHEPEWFSAGPTSQHETIDLHGFEDYENDSAADVKPAKE  18137
                   + Q+R KR N Y  H+  EPEWFSAGP SQ ETIDLHGFED E +    V    E
Sbjct  303    -----STQSRSKRGNSYLIHE--EPEWFSAGPKSQLETIDLHGFEDLEKNEERSV---TE  352

Query  18136  DVNKEFVALTNQIGSPTSR--------SSSIASLHVIDSS----NDSNEMTSSN-----G  18008
              D N +   L   + +  S+        + S+    VI S      D N +TS       G
Sbjct  353    DKNNQIQQLDKNLDAQASKDEASMRNSNDSLNFREVIPSDEKKHTDENVVTSIQNSTDLG  412

Query  18007  PPSKRS--NDNPIPKSEVEFNFDAFLNMDPM  17921
               P+K       P    E EFNFDAFLNM P+
Sbjct  413    HPNKNKPIQMQPSQNPESEFNFDAFLNMHPL  443
```

Figure 5b. Best alignment between DGA05E01 and exon 3_914.

Figure 5c. Local UCSC Genome Browser view of exon C.3. Boxed area shows area of low similarity in *blastx* track.



After using the Genome Record Finder to find the distance from the start of exon C.3, I ran the DNA sequence from about 500 bp upstream of exon C.3 in my fosmid against exon 13_914 and found no significant hits even at 1e+20 E-value which is most likely because exon 13_914 is very short (MDTSKISA). Figure 5d shows the ClustalW alignment of the same region against the translated portion of exon 13_914 which suggests exon C.2 ends about 300 bp upstream of the exon C.3. However, Figure 5e shows the ClustalW alignment against DGA05E01 and a region downstream that corresponds to three gene predictors. Although both the ClustalW predicted sequence (MCISKIVFK) and the software predicted sequence (MDTTEDNS) have about the same level of similarity to the *D. melanogaster* sequence and matching phase, the software predicted sequence was used in my final gene model because of higher conservation of length and grouping of amino acids (MDT at the start). The shortness of exon 13_914 contributed to the difficulty in concluding which match to use in the model. From this prediction, possible exon donor/acceptor sites, and complementary phase information from C.3, I used the UCSC Genome Browser and set the boundaries of exon C.2 between bp 19610 and 19585.

```
Dgri4_dna       ACAAACAGCACATATATACATATATATATGAGCGTGTATGTATGTATGTACGTATATGTG 180
CG32016_13      ------------------------------------------------------------ATGGA 5
                                                                             ***

Dgri4_dna       TATATCCAAAATTGTTTTTAAAGTGCAGCAAACAAGCGAGGGCTACACATTGATAAAATC 240
CG32016_13      TACATCAAAGATTAGTGCCAG---------------------------------------- 26
                ** *** ** ***  *   *
```

Figure 5d. ClustalW alignment between exon 13_914 and a region 500 bp upstream of

the start site for exon C.3. Alignment shows that exon C.2 ends about 300 bp upstream of the start of exon C.3.

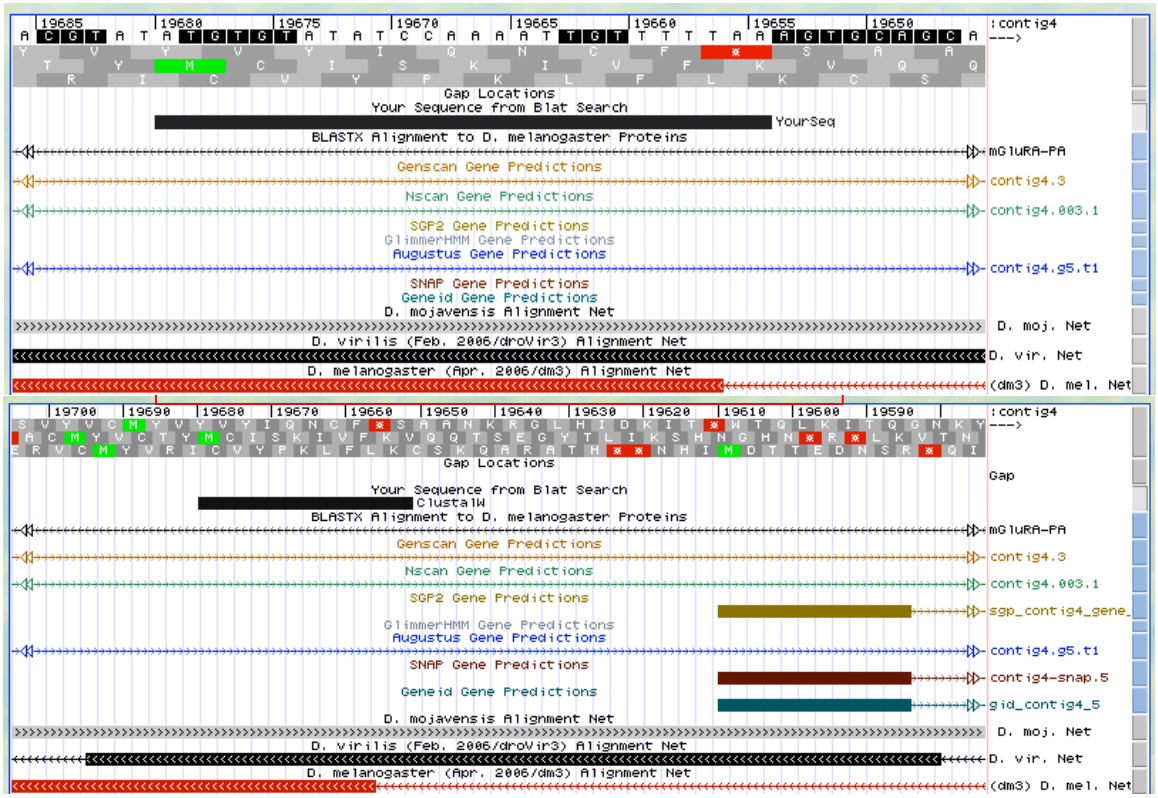


Figure 5e. BLAT alignment of ClustalW results for exon C.2 against DGA05E01. Red boxed area shows site of consensus of three gene predictors and blue boxed area shows the ClustalW results. Short exon length of gene predictor expectations (MDTTEDNS) matches exon 13_914's short length (MDTSKISA). Clustering of MDT at start of both suggest better model than ClustalW alignment. Note the methionine start that corresponds to start sites in certain isoforms. The short length and high level of mismatches in residues makes any *blastx* alignment difficult to find.

Using exon C.2's boundaries and exon distances from the Gene Record Finder, I tried using a localized BLAST analysis and ClustalW alignment to find the boundaries for exon C.1. Interestingly, exon 10_915's peptide sequence, which corresponds to exon C.1, is only X. This suggests that the exon itself does not code for a complete amino acid until it is spliced with the next exon. As a result, exon C.1 needs to splice together with exon C.3 in order to form a complete codon. The initial *blastn* alignment of the 500 bp region upstream of exon C.2 against the full nucleotide sequence of exon 10_915 did not produce any significant results even at an E-value threshold of 1e+300. Exon 10_915 was then aligned with DNA from 500 bp upstream of exon C.2 in ClustalW (Figure 5f). From inspection, it seems likely that the last AT in exon C.1 should be spliced together with a G from exon C.3 (Figures 5f and 5g). Though these two exons complement each other in phase, the first base in exon C.3 is an A rather than G, creating an incomplete methionine codon (Figure 5h). The next closest acceptor site at bp 19319 does border a G but if this represents the actual acceptor site, the start of exon C.3 will be truncated.

Because of the six gene predictors that match the longer exon C.3, the high similarity of the original exon C.3 *blastx* hit, and the lower reliability of ClustalW for exon discovery, I conclude that exon C.3 most likely exists in its longer form. However, this does not mean that exon C.1 and the ortholog to *CG32016* isoform A do not exist in *D. grimshawi*. Because of the lower reliability of ClustalW alignment, the location and existence of exon C.1 should be further investigated.

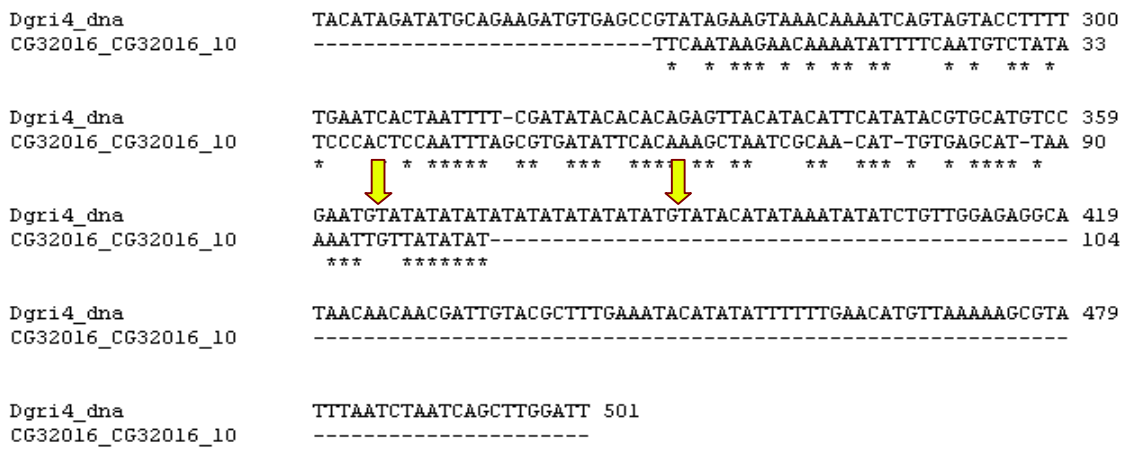


Figure 5f. ClustalW alignment of exon 10_915 with the DNA sequence 500 bp upstream of exon C.2 to find the location of exon C.3. Arrows point to closest donor sites to alignment. Regardless of the donor site, the most probable spliced bases are AT to start the orthologous exon and peptide (*CG32016* isoform A). This does not produce a viable methionine start based on the predicted start for exon C.3 (Figure 5g).

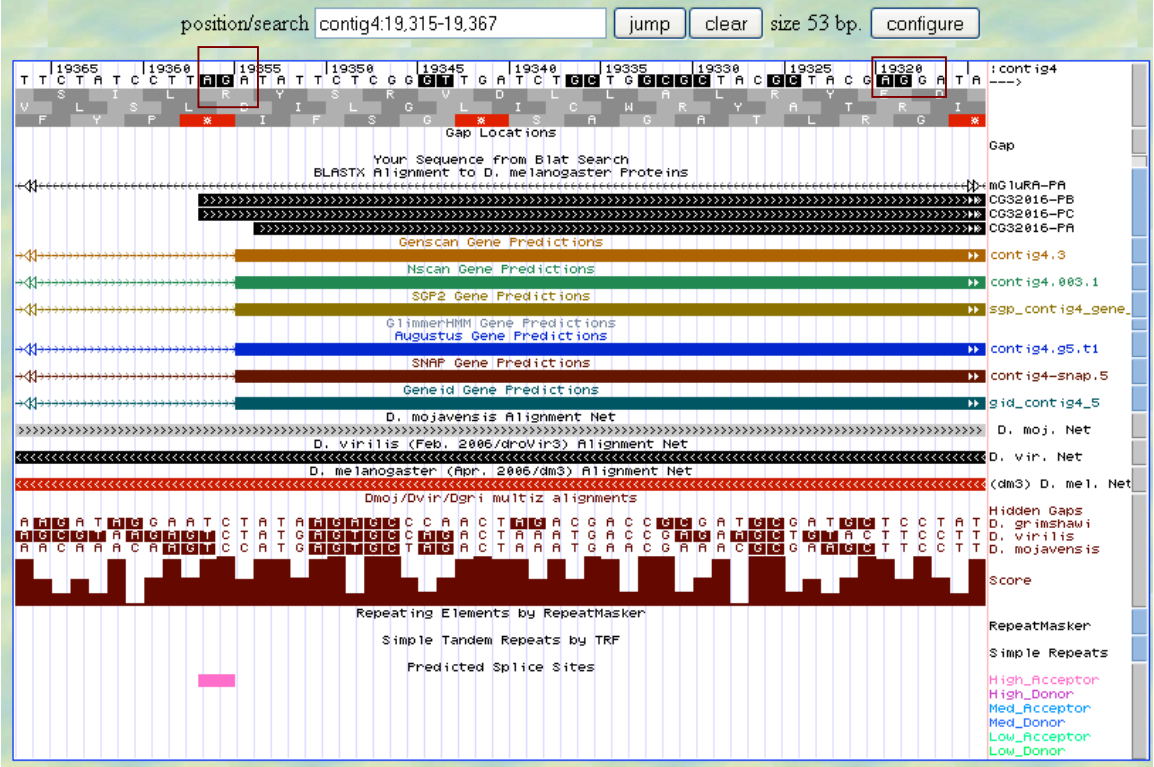


Figure 5g. Boxed areas show possible acceptor sites for exon C.3. The first start site has complementary phase with exons C.1 and C.2 but does not have the necessary G to complete the methionine start codon. Note the six gene predictors that match at the first site. The second start site has the necessary G but lacks the correct phase to maintain the frame that results in the most conserved sequence.

The remaining exons (C.4-C.10) all matched fairly well in the initial *blastx* alignment and when examined on the local UCSC Genome Browser, showed agreement with almost all gene predictors. The exon donor sites were matched with GTs or GCs and acceptor sites were matched with AGs. Furthermore, at almost all sites, the phases of donors and acceptors complemented completely, so that the sum of trailing nucleotides at donor/acceptor sites equals three (a complete codon). Exons C.2 and C.6 (*CG32016:15_915*) both have methionine start sites and correspond to different starting exons for two groups of isoforms. Orthologs of isoforms B and C start with exon C.2 and orthologs of isoforms D, E, and F start with exon C.6. Exon C.6 starts in the middle of exon C.5 (*CG32016:5_914*) and ends at the same donor site (Figure 5f). All isoforms end at the same site in exon C.10 (*CG32016:9_914*) and range between bp 19680 and 15637 (isoforms B and C) or bp 17613 and 15637 (isoforms D, E, and F) (Figure 5g, Table 2).

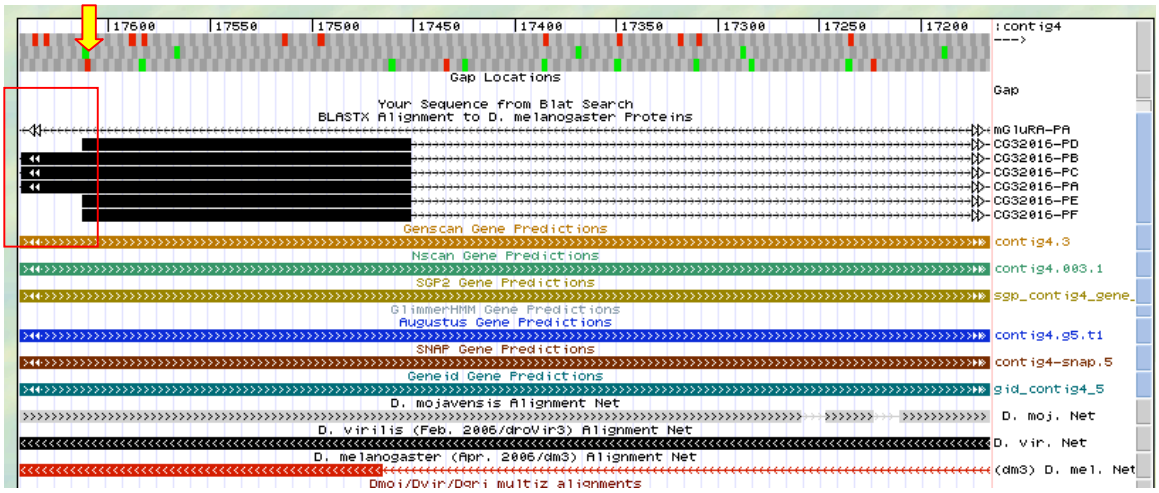


Figure 5f. Boxed region shows the difference in length of C.5 and C.6. Arrow points to methionine at the start of exon C.6.

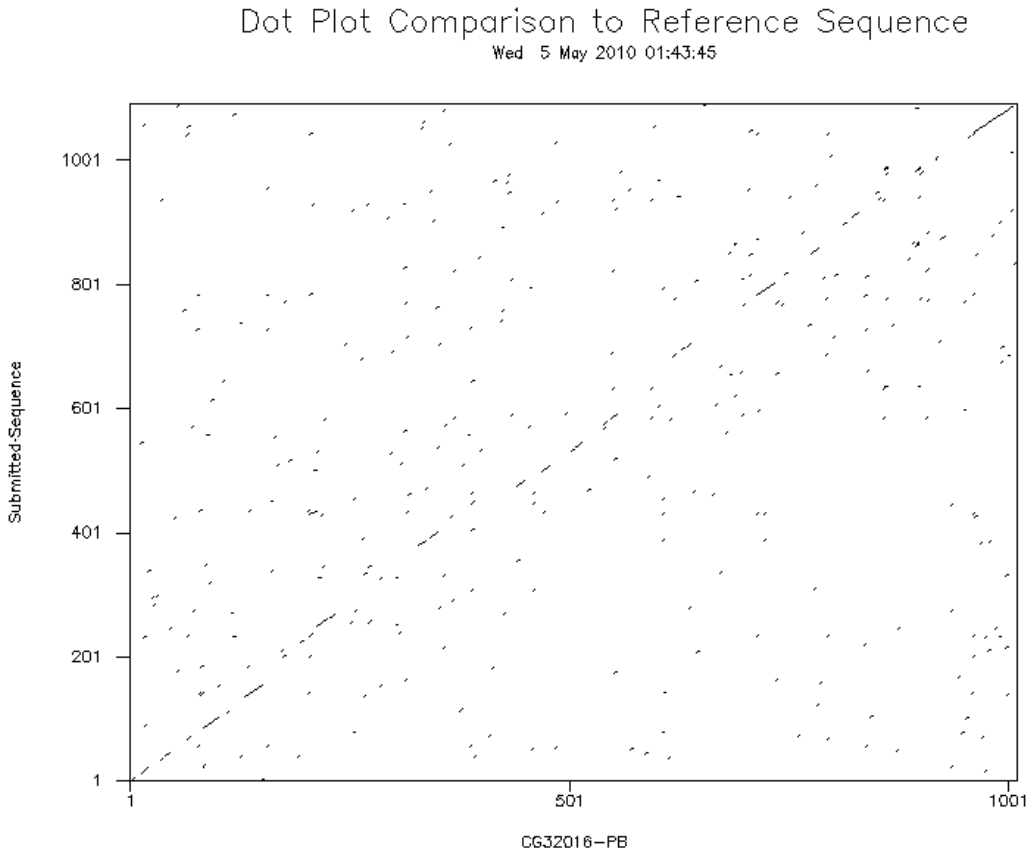


Figure 5g. Dot plot of gene model for the *D. grimshawi* ortholog of CG32016 isoform B against the *D. melanogaster* CG32016 isoform B.

Isoform	Accession Numbers
CG32016-PA	NP_726618
CG32016-PB	NP_726619

CG32016-PC	NP_726620
CG32016-PD	NP_726621
CG32016-PE	NP_726622
CG32016-PF	NP_726623

Table 2. Isoforms of CG32016 in *D. melanogaster* and their accession numbers.

Feature D

Blastx incorrectly predicted feature D on the local UCSC Genome Browser as the ortholog to the *CG4847* gene (accession #: NP_611221, NP_725686, NP_725687, NP_725688, NP_995874). Thanks to work by Matthew Kwong, feature D corresponds more closely with the *CG5367* gene (accession #: NP_609387). Confirmed by both BLAT (Figure 6a) and NCBI *blastx* analysis with unmasked sequence, feature D seems to match best with *CG5367* instead of *CG4847* [9]. Figure 6b shows all the *blastx* hits corresponding to this region in local BLAST Viewer. All of these hits were filtered in the UCSC Genome Browser view in Figure 6a using an algorithm developed by Wilson Leung that incorporates alignment and overlap information to output the "best" hit. In this case, *CG5367* was incorrectly filtered from the initial browser view. When the *D. grimshawi* ortholog to *CG5367* is aligned against the *D. melanogaster* counterpart, it has fairly high matching (Figure 6c).

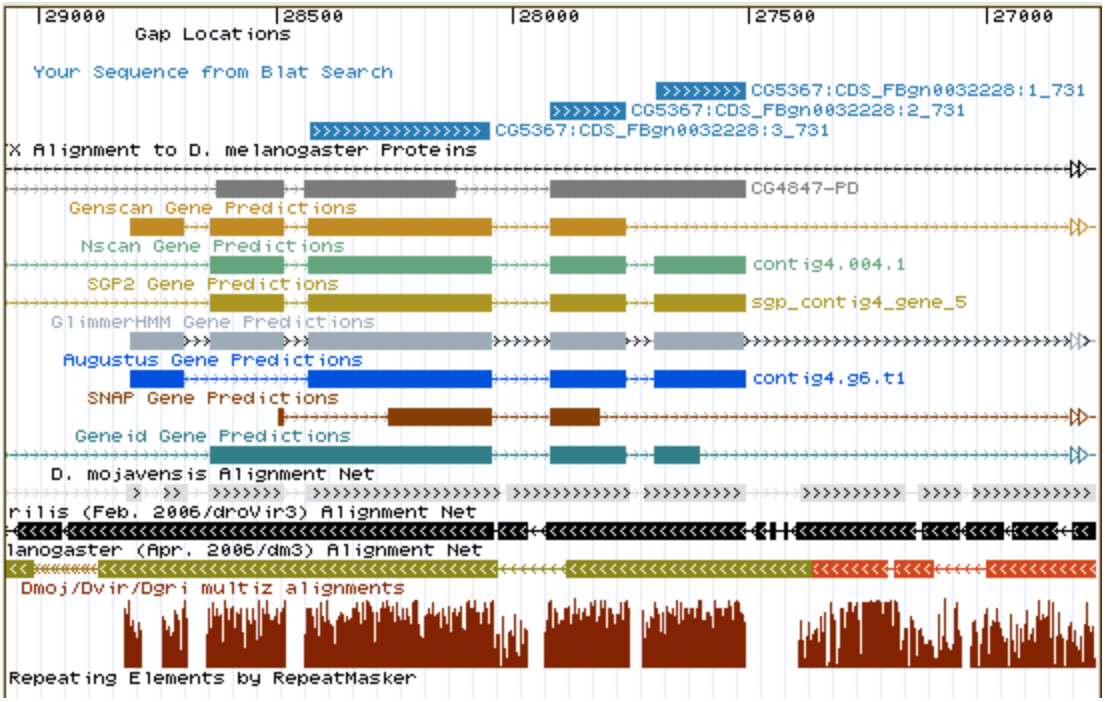


Figure 6a. BLAT analysis shows *CG5367* matching where the predicted *CG4847* ortholog lies. BLAT found *CG5367* exons 3 through 5. Exons 1 and 2 matched with location of GENSCAN prediction but the alignment for exon 1 is poor.

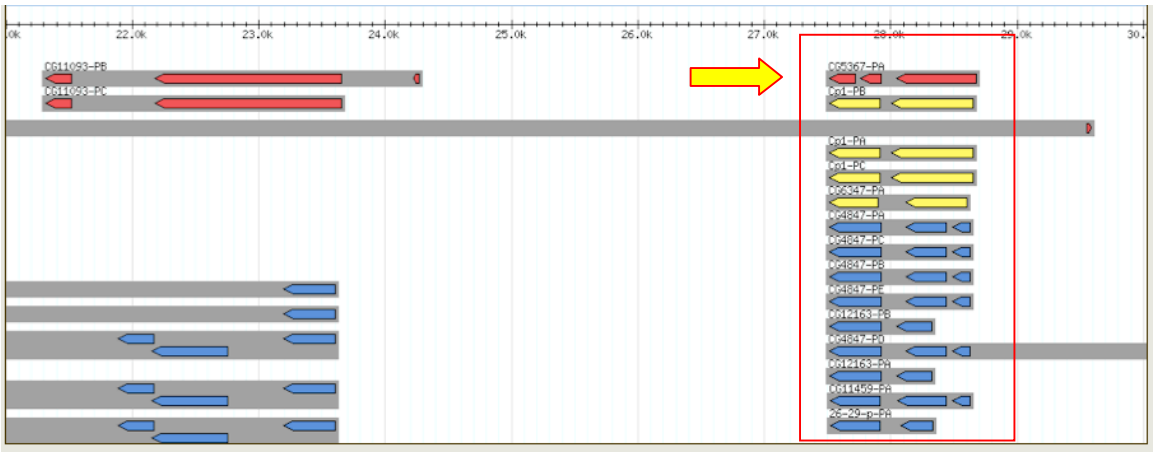


Figure 6b. Local BLAST Viewer of *blastx* output. *CG5367-PA* had highest quality alignment but was masked and not called by *blastx* possibly due to gene overlap. As a result, *CG4847* was incorrectly called instead. Arrow points to *CG5367*. Match quality corresponds to color: red is the highest, then yellow, and finally blue.

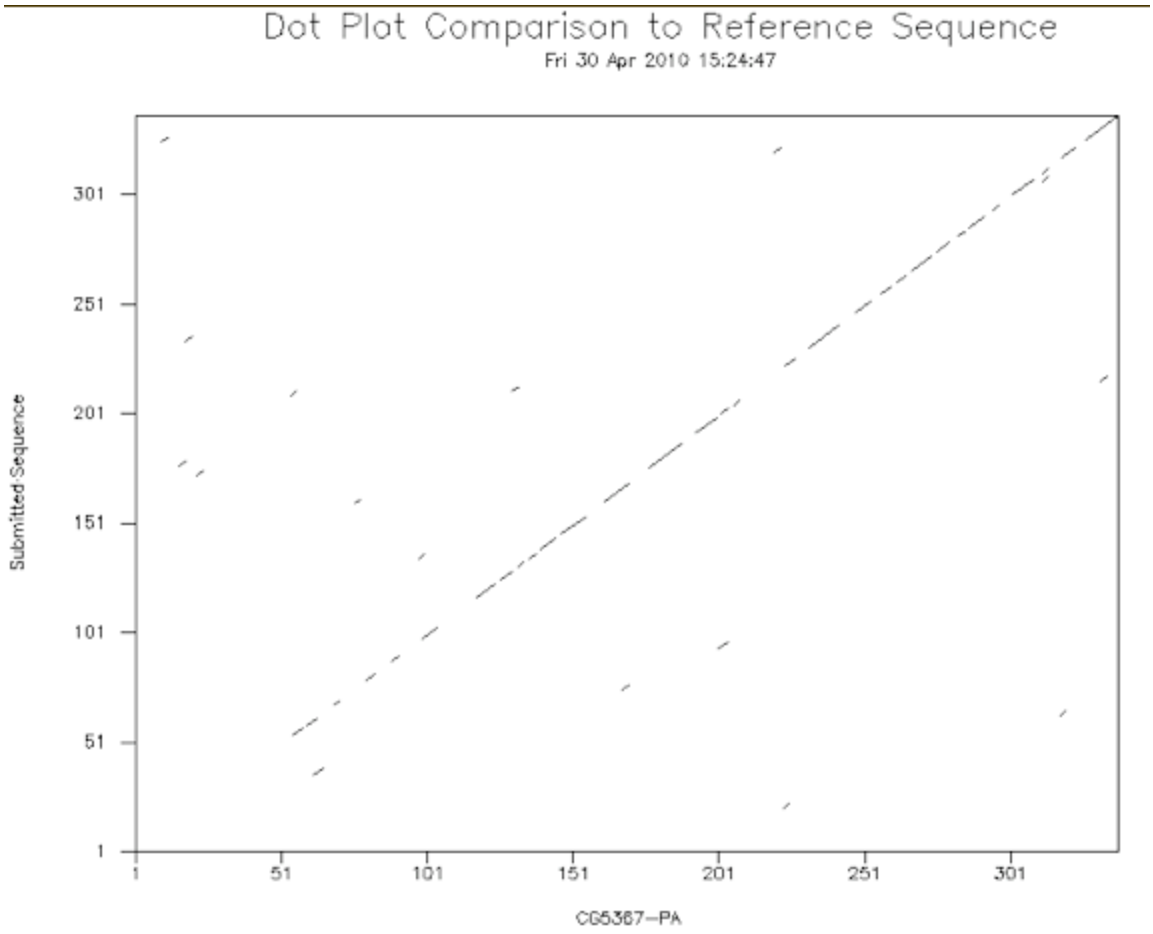


Figure 6c. Dot plot of *D. grimshawi* ortholog of *CG5367* against the *D. melanogaster* form. The first exon is not conserved well as seen by the lack of linearity at the start of the dot plot and the Needleman-Wunsch alignment (Figure 6f).

Based on the NCBI *blastx* output, the *D. grimshawi* ortholog to *CG5367* has good alignment with four of the five exons (Figure 6d). Only exon D.1 (*CG5367:5_731*) in the protein does not have good alignment as shown by the dot plot in Figure 6c. After using the Gene Record Finder to narrow the search window to 200 bp upstream of exon D.2, a *blastx* search of the region failed to produce any significant matches. A Needleman-Wunsch alignment of the nucleotide sequence of this same region using *needle* (Figure 6f) followed by a BLAT of the aligned sequence against the fosmid (Figure 6g) helped find the boundaries for exon D.1. The remaining exon boundaries in this gene ortholog could be found from the initial *blastx* alignment, exon donor/acceptor splice site, and complementary phase information.

```
>lcl|52503 CG5367:CDS_FBgn0032228:3_731
Length=129

                                                    Sort alignments for this :
                                                      E value  Score  Percent
                                                      Query start position  S

 Score =  207 bits (527),  Expect = 4e-57
 Identities = 96/129 (74%), Positives = 112/129 (86%), Gaps = 0/129 (0%)
 Frame = -2

Query  28431  NTDSYLQGFLRLLRSPPNSTTDNIADIVGSPLMNNVPESFDWRKKGFNTPPYNQQSCGSC  28252
              +TD YL+GFLRLL+S    + DN+A+IVGSPLM NVPES DWR KGF TPPYNQ SCGSC
Sbjct  1      STDGYLKGFLRLLKSNIEDSADNMAEIVGSPLMANVPESLDWRSKGFITPPYNQLSCGSC  60

Query  28251  YAFSVAQSIEGQVFKRTGKLLALSEQQIVDCSVSHGNHGCIGGSLRNTLTYLQATGGLMR  28072
              YAFS+A+SI GQVFKRTGK+L+LS+QQIVDCSVSHGN GC+GGSLRNTL+YLQ+TGG+MR
Sbjct  61     YAFSIAESIMGQVFKRTGKILSLSKQQIVDCSVSHGNQGCVGGSLRNTLSYLQSTGGIMR  120

Query  28071  SLDYKYAAK  28045
                DY Y A+
Sbjct  121    DQDYPYVAR  129
```

Figure 6d. Representative highly matching *blastx* hit for DGA05E01 DNA against *CG5367 D. melanogaster* exon DNA. This hit is for the third of five exons in *CG5367.*

```
Dgri4_dna        1 AGTTTACTTCTCCAATACAGTTGCAGCAATGCGGACCGATGTATTATTAT     50
5                1 ----------------------------ATGTGGA------AATTAATTT     16

Dgri4_dna       51 TACACTGCTTTTATTGTTT-----GCTTAA--TGTTAATTT---------     84
5               17 T------CTTTGGTTGTTTATGTGGCCTAAATTGTCAAATTGTAACATCC     60

Dgri4_dna       85 AATTAAATGCCAACAAAA-----------CGAATT--AAGACCAATAAAC    121
5               61 AATTTGA-GCGAAGGAAATTCTTCTTCTGCAAATTGCAAGAGC-------    102

Dgri4_dna      122 AATGAACAGTTTGAAAATTTTAAGGTATGATAAGTAACACAGTTGTTTGA    171
5              103 ---GAA---TTCGAAAAATTTAAG--------------------------    120

Dgri4_dna      172 TTTTTTCTAGTAATTTATTTCGTTTACAGA    201
5              120 ------------------------------    120
```

Figure 6f. Needleman-Wunsch alignment of the region 28635 to 28835 in DGA05E01 with the coding region of exon 5_731 to find the boundaries of exon D.1. Arrows point to the methionine start and the exon donor site at the end of the alignment.

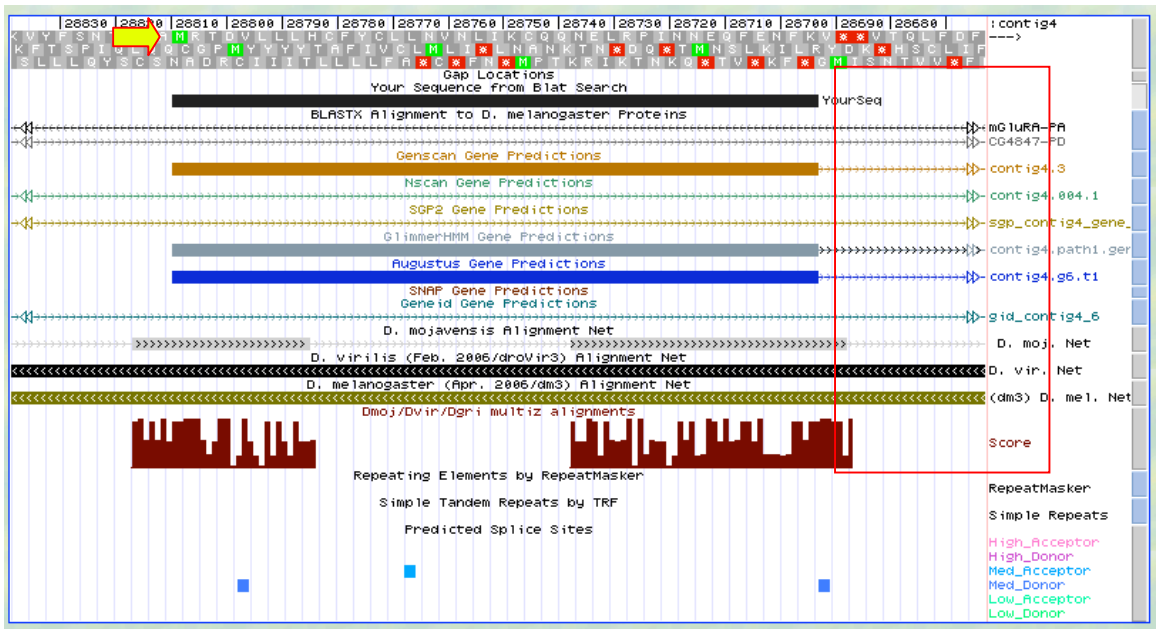


Figure 6g. BLAT alignment of the Needleman-Wunsch results against the fosmid. Arrow points to the methionine start. Note that the Needleman-Wunsch results match exactly with three gene predictors and splice predictors, suggesting a probable starting exon.

Feature E

Already seen on the *blastx* track of the local UCSC Genome Browser, GENSCAN predicted feature E (seen as contig4.4 and contig4.5 on the browser view) does not correspond to any *blastx* match to *D. melanogaster* (Figure 7a). Later *blastp* analysis of the predicted exon peptide sequences against the nonredundant database of *D. melanogaster* proteins yielded no significant hits (Figure 7b). As a result, feature E does not have any *D. melanogaster* orthologs. A FlyBase.org *blastp* analysis of the peptide sequences against the annotated protein database of all other *Drosophila* species also did not return any significant hits (Figure 7c). Interestingly, several *Drosophila* species had peptide fragments that matched the same portion of contig4.4 (Figure 7d). This suggests that contig4.4 may have a conserved domain and warrants a closer look for a possible gene in this area of DGA05E01.

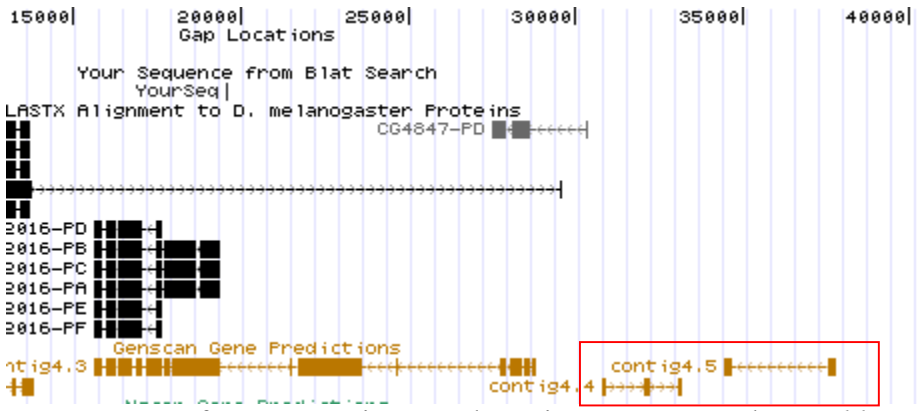


Figure 7a. GENSCAN features contig4.4 and contig4.5 correspond to no *blastx*

predictions.

BLAST Hit Summary				
☑	Description	Species	Score	E value
☑	CG4238-PF	Dmel	29.261	0.642449
☑	CG4238-PD	Dmel	29.261	0.65326
☑	CG4238-PE	Dmel	29.261	0.692551
☑	if-PC	Dmel	28.8758	1.03368
☑	CG15890-PA	Dmel	27.335	2.86069
☑	lok-PC	Dmel	26.1794	6.70012
☑	lok-PB	Dmel	26.1794	6.70012
☑	lok-PA	Dmel	25.7942	7.53033
☑	CG4725-PA	Dmel	25.409	9.67215

BLAST Hit Summary				
☑	Description	Species	Score	E value
☑	CG13252-PA	Dmel	27.335	3.89204
☑	CG18437-PB	Dmel	26.5646	6.86412
☑	CG1909-PB	Dmel	26.1794	8.96482
☑	CG1909-PA	Dmel	26.1794	8.96482

Figure 7b. *Blastp* results of GENSCAN predicted exon sequence against nr database of *D. melanogaster* proteins. Top table corresponds to contig4.4 hits and bottom table correponds to contig4.5 hits. Note the high E-values.

BLAST Hit Summary				
☑	Description	Species	Score	E value
☑	Dgri\GH18332-PA	Dgri	30.8018	1.48643
☑	Dwil\GK15042-PA	Dwil	29.6462	3.12356
☑	Dgri\GH10869-PA	Dgri	29.6462	3.48142
☑	Dper\GL19671-PA	Dper	29.6462	3.59958
☑	Dana\GF14596-PA	Dana	29.6462	3.59958
☑	Dpse\GA18053-PA	Dpse	29.261	3.62974
☑	CG4238-PF	Dmel	29.261	3.69081
☑	CG4238-PD	Dmel	29.261	3.75292
☑	Dgri\GH25212-PA	Dgri	29.261	3.78437
☑	Dsim\GD23127-PA	Dsim	29.261	3.81607
☑	CG4238-PE	Dmel	29.261	3.97864
☑	Dere\GG24824-PA	Dere	29.261	4.01198
☑	Dsec\GM16848-PA	Dsec	29.261	4.01198
☑	Dyak\GE17889-PA	Dyak	29.261	4.18289
☑	Dmoj\GI18244-PA	Dmoj	28.4906	6.95856

BLAST Hit Summary				
☑	Description	Species	Score	E value
☑	Dpse\GA24356-PA	Dpse	30.8018	2.07742

Figure 7c. *Blastp* results of contig4.4 against the annotated protein database of all 12 *Drosophila* species on FlyBase.org. Top table corresponds to contig4.4 hits and bottom table correponds to contig4.5 hits. Note the high E-values.

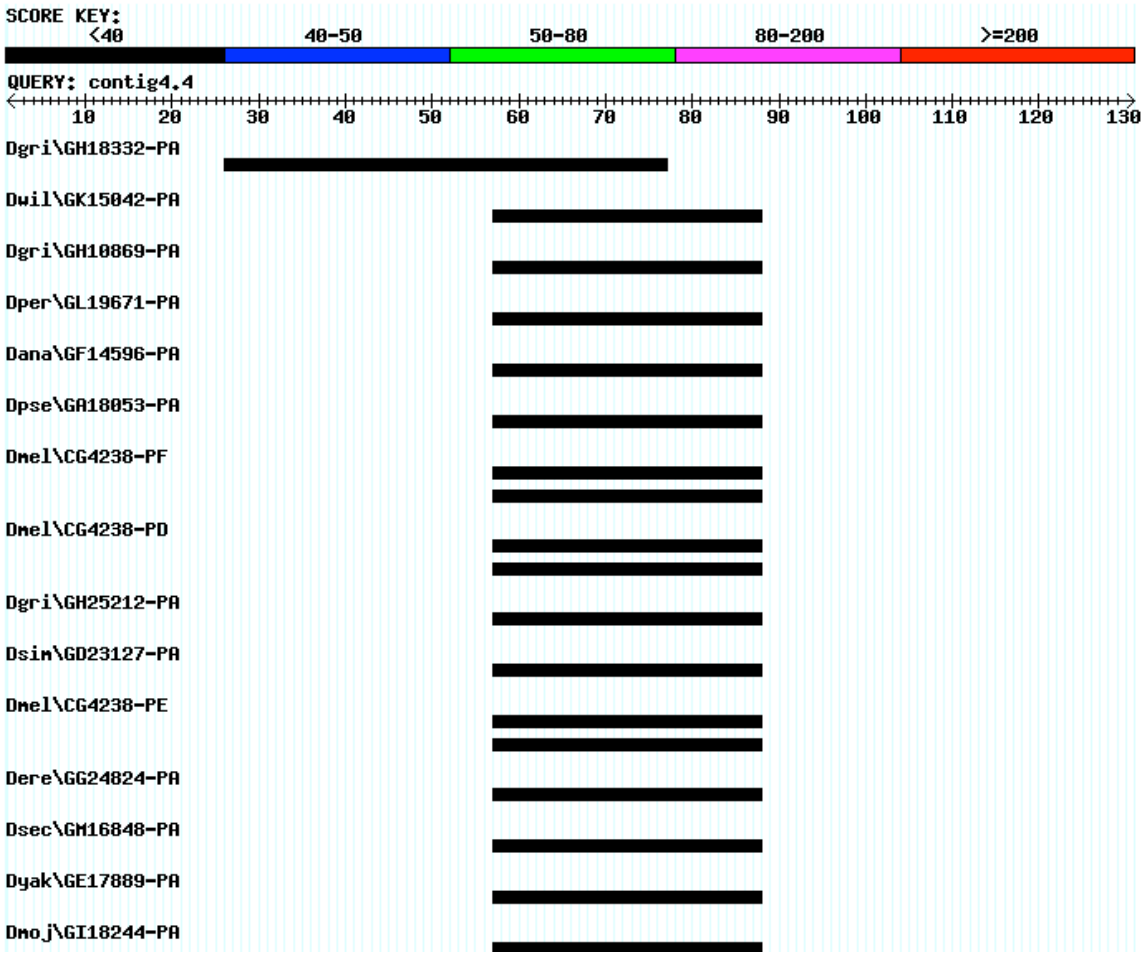


Figure 7d. The same *blastp* results from Figure 7c presented in graphical form. Note that most hits match at the same region of the peptide. This may be a conserved domain in predicted peptide contig4.4 and warrants further investigation to check if contig4.4 represents a part of a real gene.

Other Features Unpredicted by GENSCAN

From a *blastx* of the unmasked fosmid sequence against annotated proteins in *D. melanogaster* as well as from Matthew Kwong's analysis on *CG5357*, *CG11093* (accession #: NP_001162829, NP_001162830) was found with high level of matching (Figure 6b and 8a). *CG11093* is on chromosome four and has two different isoforms in *D. melanogaster* and five exons.

From an initial *blastx* analysis of fosmid DNA against *CG11093* exons, I only found three significant hits corresponding to the last three exons. Using these as anchors, I extended each search trying to conserve exon length. Exon E.3 (*CG11093:3_907*) and exon E.4 (*CG11093:2_907*) had almost complete coverage of their *D. melanogaster* counterparts. However, exon E.5 (*CG11093:1_907*) ended very prematurely with an early stop codon about 45 bp upstream of where it was expected if length were conserved (Figure 8b). Because of the good conservation of all these exons, gene predictor agreement, and high quality sequence in Consed (Figure 8c), I decided to choose bp

21242 as the end of the gene model.

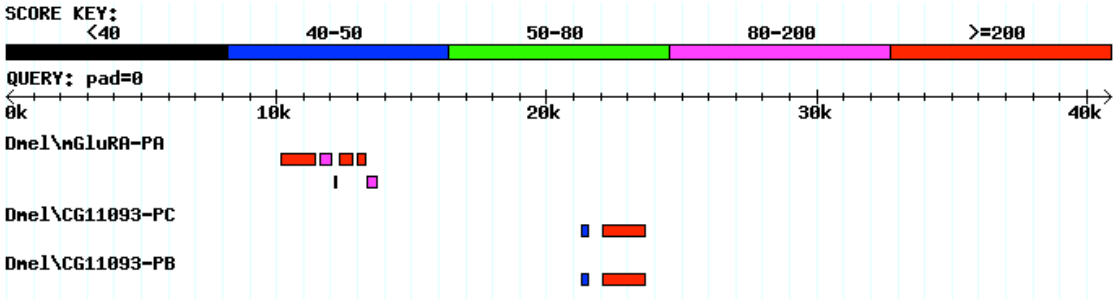


Figure 8a. Red denotes high level of matching. Note the high level of matching for *CG11093*.

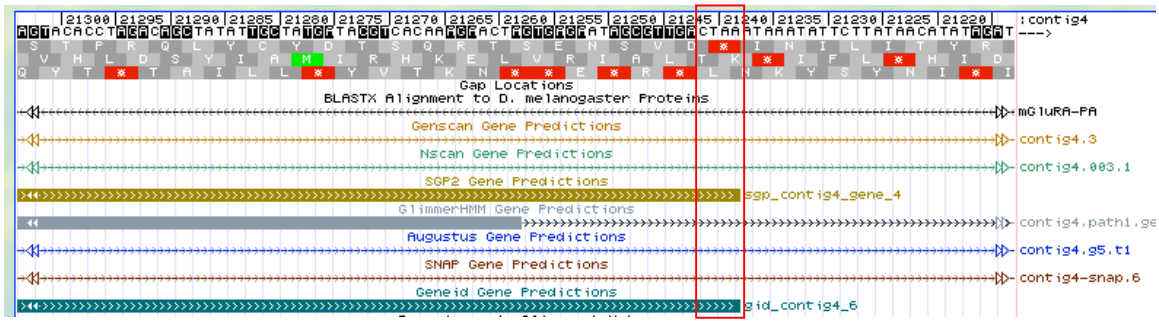


Figure 8b. Boxed area shows the stop codon in frame -1 corresponding to exon E.5. Note that several gene predictors predict this to be the end of the exon as well. However, this makes exon E.5 significantly shorter than its orthologous exon 1_907. Note Figure

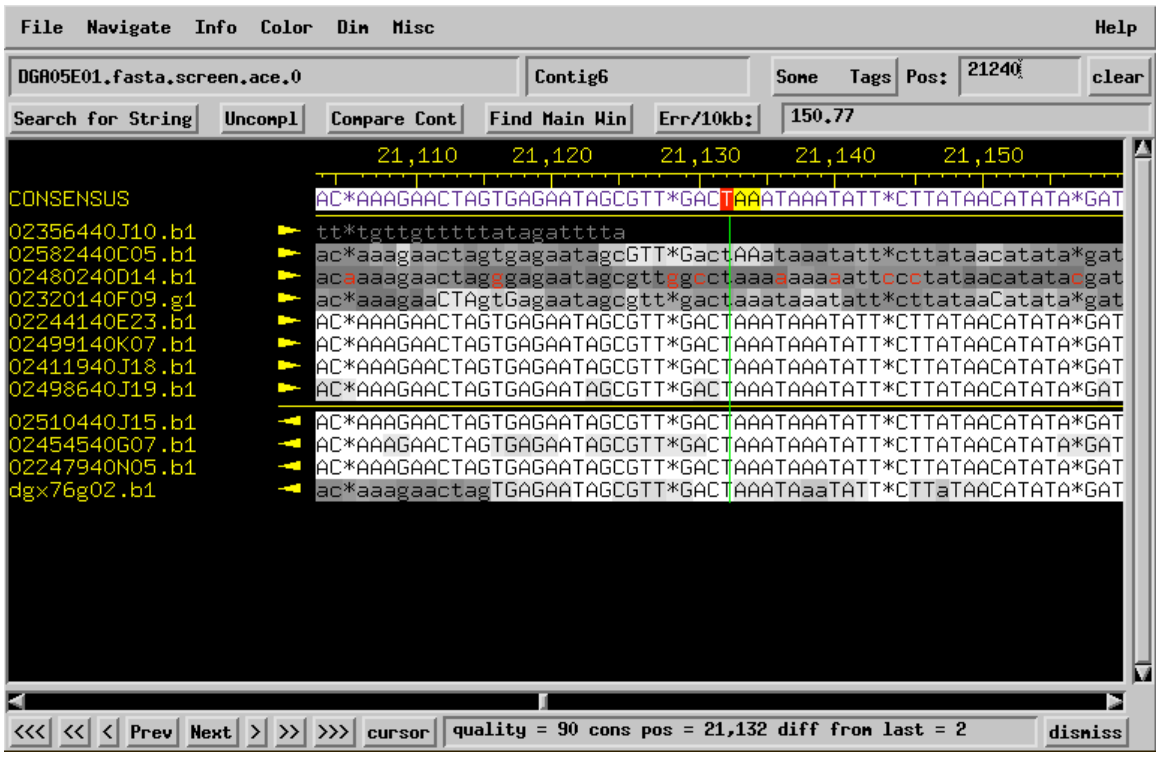


Figure 8c. Current highlighted consensus base (T) has a quality value of 90. The surrounding bases all have similarly high quality values which suggest sequencing errors are unlikely in this region. The highlighted TAA corresponds to the stop codon in Figure 8b.

For exons E.1 (*CG11093:6_907*) and E.2 (*CG11093:4_907*), no significant *blastx* alignment came from the initial results. Even after narrowing the search region based on distances from Gene Record Finder, *blastx* found no significant results in the area. ClustalW gave good alignments for both in their respective narrowed search regions (Figure 8d). However, the best alignment for exon E.1 is closer to exon E.3 than E.2 is, suggesting that exon E.2's coordinates either are incorrect or an inversion event occurred in this area. From the local UCSC Genome Browser, there is no exon predicted in the region covered by E.2. Furthermore, there are several premature stop codons in frame -2 in which exon E.2 goes through, shortening the putative exon's length in half and preventing any fusion of exons E.2 and E.1 into one exon based on this model. For exon E.1, NSCAN predicts the exact coordinates as the ClustalW alignment while maintaining the conserved length of exon 6_907. From these values, it seems unlikely that exon E.2 exists in the predicted region and the ortholog of *CG11093* isoform B exists in *D. grimshawi*. This does not preclude the existence of exon E.2, the fusion of exons E.1 and E.2, or *CG11093* isoform A; it only represents the best predicted model based on the available evidence. However, the protein still is significantly shorter than in *D. melanogaster* (Figure 8e). Because of the odd location of exon E.2 and this shortening, further investigation should be done to confirm these conclusions.

```
Dgri4_dna       CCAGAAATTGCACTAAATGGCATTCGCTGAGTCTATTCAAAGATCTACTCGGCCGTCAAC 6720
exon_6          ----------------ATGGATTTAAATGAAAATTTTAAAATGGCTTCTATATCTGTGGC 44
                               ****  **   ***   * ** ***   ** **    *     *

Dgri4_dna       TCCACCTAATGGAATGGAAAACAGCCAAATATCCGTTTCCCAAAAATCAAATCACGTATG 6780
exon_6          TCACTTTGATAGGATGGAAAACTGCCAAACGTCT------CAAAGGTCAAATCAC----- 93
                **    * ** * ********* ******  **       ****  *********

Dgri4_dna       TCAATAGCCTATCGCATACAGGCAAATGCCTATATGTAGATGTTGATGTGCTTGTATTTA 2340
exon_4          ---------------------------------------------ATGTCTAAATACTTT 15
                                                             ****     ** **

Dgri4_dna       GATTGTTGCCCATGTGCATGTCATACTTTTGTTTTTTAACGGGAAATATGAGAAGTGGGT 2400
exon_4          TGTCACCGCCCAAATCGTTGTC-TGTCTAGCTTTCTTATCAGCAGA-AGGAAAATTGAAA 73
                  *    *****  *   **** *   *   *** *** * * * * * ** ** **

Dgri4_dna       AAGATCCAGGCGTGGACACGGGCAGAACACCCTATATAAACAAAGCACTTAAATTAATTT 2460
exon_4          AG------------------------------------------------------- 75
                *
```

Figure 8d. ClustalW alignments for exons E.1 and E.2 in narrowed search regions based on distances calculated from Gene Record Finder results in *CG11093*. Top alignment refers to exon_6 for E.1 and the bottom alignment refers to exon_4 for E.2.

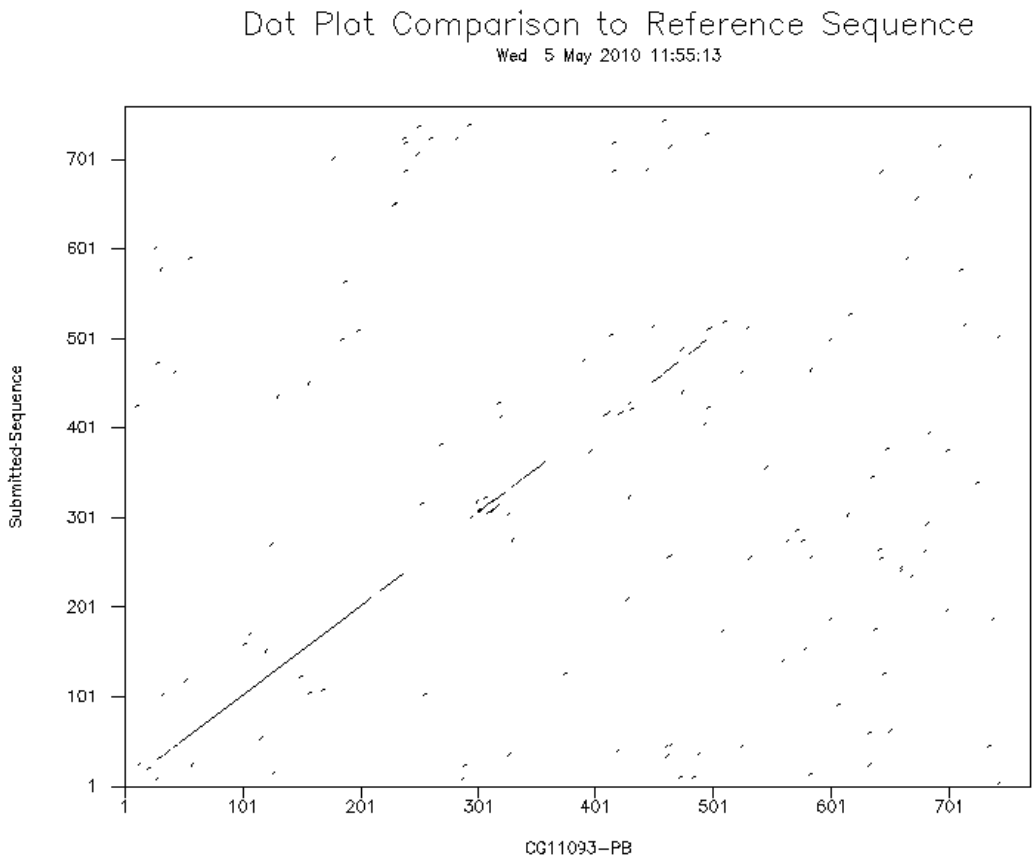


Figure 8e. Dot plot of *D. grimshawi* ortholog against CG11093-PB in *D. melanogaster*. Note the lack of homology at the end of the protein. This is due partially to the truncated sequence of the exon E.5.

Clustal Analysis

ClustalW alignment was done on the gene ortholog of *mGluRA* in *D. grimshawi* among four species of *Drosophila* and also between mouse (*Mus musculus*) as an outgroup to explore how a basic neural receptor evolves between different species and differing complexities of nervous systems. Figure 9a shows that most of the alignment matches fairly well, especially in conserved domain regions like the PBP1 binding/ANF receptor and 7tm_3 regions (Figure 9b). This can be expected because of the importance of glutamate and its receptors in communication within nervous systems in organisms from flies to mammals. Heavy purifying selection at these highly conserved regions can be expected because any change in a key region of a protein can disrupt an optimized configuration and lead to a dysfunctional protein. From the alignments, the start and end of the genes show lower conservation compared to the central portions of the gene in which the functional domains like the transmembrane and receptor binding domains are found. The ends can be expected to change somewhat especially because these locations represent sites that become modified the most whether at transcript or final protein stage through chemical additions or cleavages. This initial analysis suggests that despite increasing complexity of nervous system between *Drosophila* species and *Mus musculus*, the basic parts of the system do not change significantly as seen in the high level of

conservation between fly and mouse metabotropic glutamate receptors.

```
Grimshawi_          FTMYTTCIIWLAFVPIYFGTGNSYEIQITTLCISISLSASVALICLYSPKVYILVFHPDK 890
Virilis_            FTMYTTCIIWLAFVPIYFGTGNSYEIQITTLCISISLSASVALVCLYSPKVYILVFHPDK 881
Mojavensis_         FTMYTTCIIWLAFVPIYFGTGNSYEIQITTLCISISLSASVALICLYSPKVYILVFHPDK 851
Melanogaster_       FTMYTTCIIWLAFVPIYFGTGNSYEVQTTTLCISISLSASVALVCLYSPKVYILVFHPDK 880
Yakuba_             FTMYTTCIIWLAFVPIYFGTGNSYEVQTTTLCISISLSASVALVCLYSPKVYILVFHPDK 880
[Mus                FTMYTTCIIWLAFLPIFYVTSSDYRVQTTTMCISVSLSGFVVLGCLFAPKVHIVLFQPQK 832
                    **************:**:: *...*.:* **:***:***. *.* **::***:*::*:*:*

Grimshawi_          NVRKLTMNSTVYRRSATTAG-------------------------AQGMPTSSGYSRTPV 925
Virilis_            NVRKLTMNSTVYRRSAATGG-------------------------APGVPTSSGYGRTPI 916
Mojavensis_         NVRKLTMNSTVYRRSATTGAGQVAGTGTGTGTGTGTGTGAGAATGIQGVPTSSGYSRTPI 911
Melanogaster_       NVRKLTMNSTVYRRSAAAVA-------------------------QGAPTSSGYSRTHA 914
Yakuba_             NVRKLTMNSTVYRRSAAAGA-------------------------QGAPTSSGYSRTQA 914
[Mus                NVVTHRLHLNRFSVSG-----------------------------------TATTYSQSSA 858
                    ** .  :: . :  *.                                    *:: *.::

Grimshawi_          GLNATDVGVGTG----------VAATIATPSDRPSQNGSQNGSPCSELDQNQTVIIHKNE 975
Virilis_            GLNAADAGIGP--------------AAKAVSERQPQSDCEN-SPCSELDQNQTAVIHRNE 961
Mojavensis_         VLNSEDACYGTGGGIGVGVEIGVGTITMSATERPSPSISQK-SRNCDLDRNQSVVIHRND 970
Melanogaster_       PGTSALTGGAVG-------------TNASSSTLPTQN-----SPHLDEASAQTNVAHKTN 956
Yakuba_             PGTTALTGGAVG-------------TTASSSALPTQN-----SPTLDDASGQSNVVHKSN 956
[Mus                STYVPTVCNGRE-----------------------------------VLDSTTSSL----- 879
                          .  .                                        : :

Grimshawi_          ECFNGAAPTSGKCDCVIGIVEPSCIAKTQD 1005
Virilis_            ECVNG-APTSDKCDCRLGAADPSCTRIND- 989
Mojavensis_         ECYNNCASTSEKCDCGLGIGDPACIGIED- 999
Melanogaster_       ---GEFLPEVGE------RVEPICHIVNK- 976
Yakuba_             ---GEFLPEEGE------FVGSICNRINK- 976
[Mus                ------------------------------
```

Figure 9a. Part of the ClustalW alignment. Note the ends are not as well conserved and the simple repeat in *D. mojavensis* (boxed).

Figure 9b. Conserved domains in mGluRA. These correspond to most of the regions that have high conservation among species.

To examine how regulation of *mGluRA* evolves, I aligned with ClustalW the 2 kB region upstream of the mGluRA ortholog in each of the five species of *Drosophila* used in the previous alignment. Interestingly, no regulatory pattern emerged from the alignment; only sparsely distributed bases showed any consensus (Figure 9c). None of the conserved sequences from J Kadonaga's lab page were seen. As such, if there is regulation of *mGluRA* expression in the first 2 kB upstream of the gene, it may either be outside of the 2 kB region or differs enough among these species to be difficult to detect through conservation.

```
dgrim           CGCCGCC-GTTGCCATCATCAGCA-TCATCGGTA--CCATTCTTGTATTGGTA--CTT-C 1246
dmel_           CTACGTT-ATCGAGTATATAAATA-TGTTAATTT--TCTTAAAAGAGAATGTG--TTTTC 1236
dmoj_           ACATAGC-AGTCCAATGATAGGCAGCCATAGTTAAGCCGACTCAGTGGAGACGGCCTGAA 1252
dvir_           TCCTGTCGAGAGTCATGCCAAGGTATTTTCTCGTCGTCTCTTCTGGTATAACAGCTTCGC 1232
dyak_           GTATGTA-ATAGGAACGAGAAGGCGATTTGGACT--CTATCCTGGATCAGGATCACGAGC 1249
                                              *              *

dgrim           TGGTAAAGTTCTGGCATCGAAGCTTTTTCTGGGTCGCCATTGT-TACTTTCCACACA-TA 1304
dmel_           CTATATGTATGTGCTGACGTA-CCTCATCTGGATAATTGTTTTATATTTTGAGGATGGTA 1295
dmoj_           CAATAAAACAAATACATTATATATCATTGTATAAAGCCATTCCACACCAT---TATACTA 1309
dvir_           CTAAATAGACAGGTGGGCAGTTAGCAGGCATGAGAGAGAATTTGGTGGCTGTGGATTTTT 1292
dyak_           CGAGGCGATCTAGCCATGCCTCGACAGTTGACACTAAGGACTAAGAAGGCGAAGCACTTG 1309
                                                                        *

dgrim           CACACACACGCACACACAAATATATG-AGTG-TGTATATAGTCGGACTCTGGTT-CTGCA 1361
dmel_           TATA-ACTTTCTTAATTTAAGATCTGGAATA-TATCGATAGTTCAAATTTCTGT-AGTTA 1352
dmoj_           TCTATTTAGATTGGAGTTGGAAACTCAAGTG-TGGCGACA----AACTTGAACT-TTAGA 1363
dvir_           CATTGTTCCCCAAGATGTTCCATCTCCTCTG-CTATTCGTTAAGTATTTCCAAT-TGGCC 1350
dyak_           ATTGTTTTAGAACGTTTCCACGCCCACCCTAACGCCCACGTGACAAGATTGTCTACTGTA 1369
                                              *              *       *
```

Figure 9c. Representative ClustalW alignment of first 2 kB upstream of each *mGluRA* ortholog.

Repeats

Table 3 shows the RepeatMasker output for fosmid DGA05E01. The overall amount of repetitious sequence is fairly low at 4.31% (1762 bp). The only larger repetitious elements in the region are two LINE elements, one R1 and one telomeric element. However, even these LINE elements are less than 500 bp. The vast majority of repetitious elements are much smaller sequences corresponding to simple repeats (2.24%) and low complexity regions (1.57%).

```
total length:      40906 bp  (40906 bp excl N/X-runs)
GC level:         37.09 %
bases masked:       1762 bp ( 4.31 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
Retroelements            2          203 bp    0.50 %
   SINEs:                0            0 bp    0.00 %
   Penelope              0            0 bp    0.00 %
   LINEs:                2          203 bp    0.50 %
    CRE/SLACS            0            0 bp    0.00 %
     L2/CR1/Rex          0            0 bp    0.00 %
     R1/LOA/Jockey       1          191 bp    0.47 %
     R2/R4/NeSL          0            0 bp    0.00 %
     RTE/Bov-B           0            0 bp    0.00 %
     L1/CIN4             0            0 bp    0.00 %
   LTR elements:         0            0 bp    0.00 %
     BEL/Pao             0            0 bp    0.00 %
     Tyl/Copia           0            0 bp    0.00 %
     Gypsy/DIRS1         0            0 bp    0.00 %
       Retroviral        0            0 bp    0.00 %

DNA transposons          0            0 bp    0.00 %
   hobo-Activator        0            0 bp    0.00 %
   Tc1-IS630-Pogo        0            0 bp    0.00 %
   En-Spm                0            0 bp    0.00 %
   MuDR-IS905            0            0 bp    0.00 %
   PiggyBac              0            0 bp    0.00 %
   Tourist/Harbinger     0            0 bp    0.00 %
   Other (Mirage,        0            0 bp    0.00 %
    P-element, Transib)

Rolling-circles          0            0 bp    0.00 %

Unclassified:            0            0 bp    0.00 %

Total interspersed repeats:          203 bp    0.50 %


Small RNA:               0            0 bp    0.00 %

Satellites:              0            0 bp    0.00 %
Simple repeats:         16          918 bp    2.24 %
Low complexity:         12          641 bp    1.57 %
```

Table 3. Summary table describing RepeatMasker results for DGA05E01.

Previously while finishing this fosmid, I found a tandem repeat region at the start of my assembly (Figure 10a). During finishing, the tandem repeat region was collapsed using a force join. Because this corrected the fragment sizes in the *in silico* restriction digest analysis, the force join seemed to be the proper solution to the tandem repeat. However, though the force join most likely was the correct solution to the tandem repeat in DGA05E01, the first 13 kB of DGA05E01 seems to be part of a larger repeat region in fosmid DGA06H06, a fosmid previously annotated by Jeanette Wong (Figure 10b). When aligned against each other, the dot plot shows the first 13 kB of DGA05E01 overlaps with DGA06H06 and also suggests that there may be a duplicated region in DGA06H06 from 13 kB to that includes the overlapping region (Figure 10c).

To see if this long repeat is a local duplication between DGA05E01 and DGA06H06, the starting fosmid end sequence in DGA05E01 was aligned against the DGA06H06 DNA sequence in *blastn*. If the fosmid end lies outside of the repeat region in DGA06H06, the shorter length of DGA05E01 will suggest that there is real difference between the two fosmids. This difference could occur because the two fosmids came from different sources with an duplication polymorphism in one of them. However, the results from the fosmid end alignment lie within the repeat region. As a result, the overlap is too short to determine if there is an duplication polymorphism difference between the two fosmids.

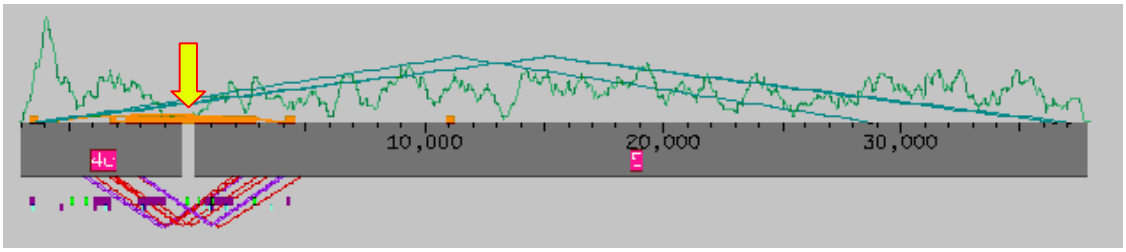


Figure 10a. Finishing view of DGA05E01 from Consed. Arrow points to orange region which denotes a long tandem repeat in Consed.

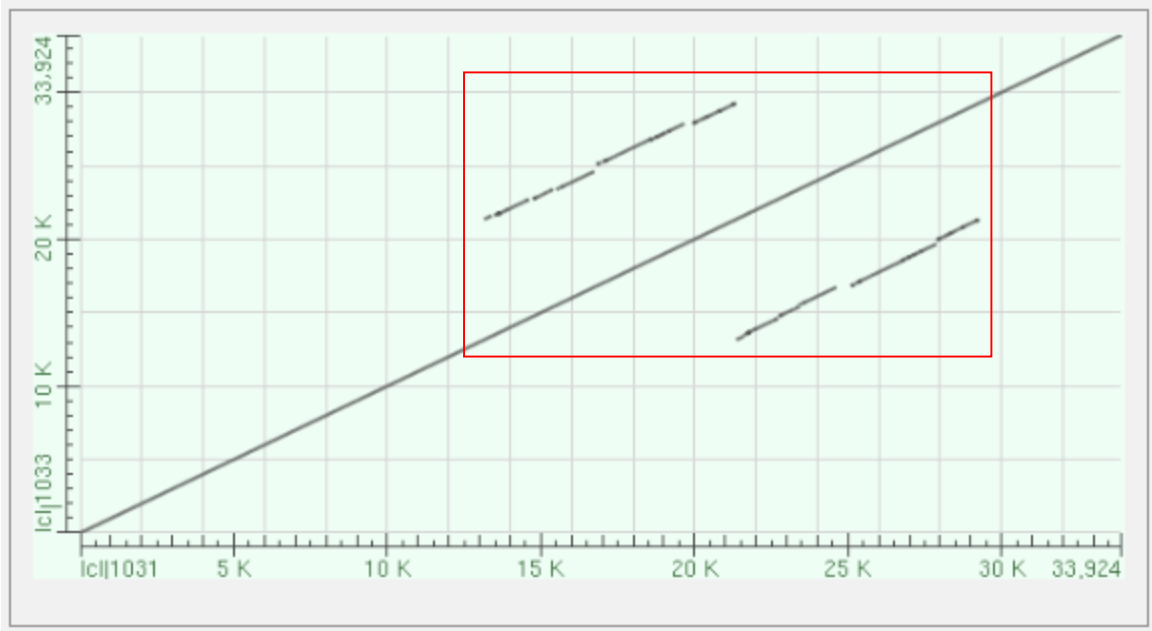


Figure 10b. Dotplot of DGA06H06 aligned against itself. The boxed region shows a repetitious region that goes from 13 kB to 21 kB and then repeats again from 22 kB to 29 kB.

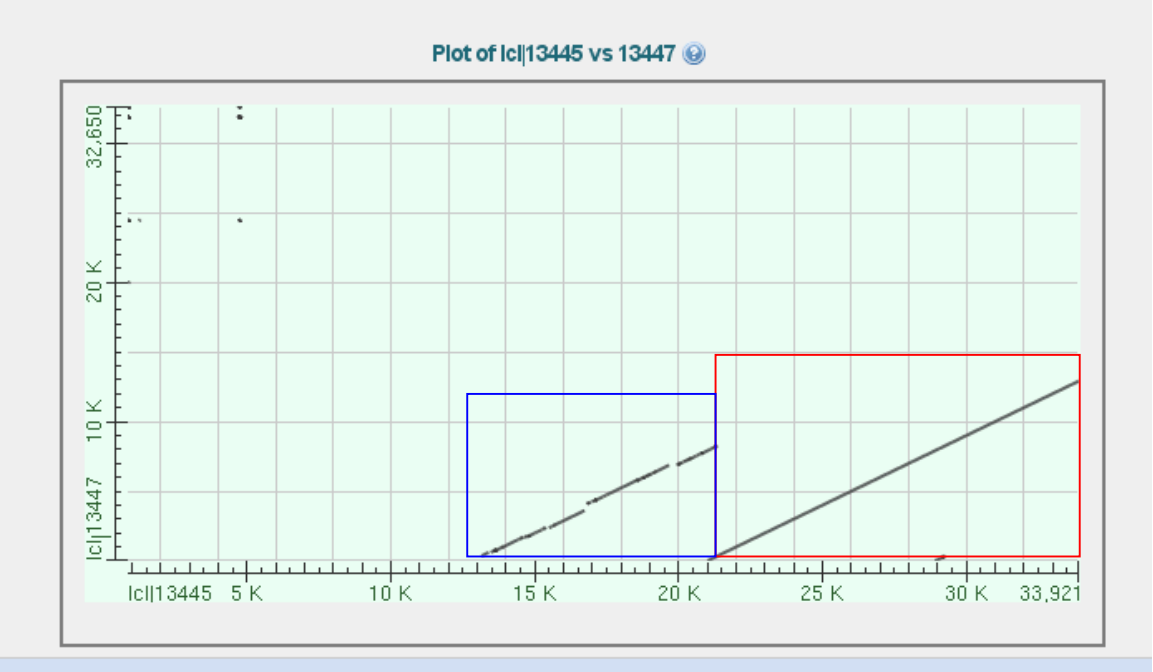


Figure 10c. Dot plot of DGA05E01 (vertical axis) alignment against DGA06H06 (horizontal axis). Comparing to Figure 10b, the blue boxed area show that the first 13 kB of DGA05E01 overlaps with the second repeat in DGA06H06. The red boxed area between 21 kB and 33921 kB represents the region that DGA05E01 overlaps with DGA06H06.

Though these results are inconclusive about the duplication polymorphism, they do not reject the possibility of duplication in this area. If this repetitious region is a result of duplication, it helps explain some of the features seen in the annotation process, especially those of the *D. grimshawi* ortholog for *eIF4G* which is discussed in more detail in the section on Feature A.

Synteny

Figure 11a shows the order and orientation of genes in *D. grimshawi*. These genes maintain their relative orientation when compared against their orthologous counterparts in *D. melanogaster*. All genes except *CG5367* come from chromosome four in *D. melanogaster*. *CG5367* instead comes from chromosome 2L. From the Repeats analysis, there are no real indicators for how the ortholog of *CG5367* became adjacent to genes that are all found on chromosome four in *D. melanogaster*. Further investigation in other fosmids may produce the necessary information to provide an explanation.

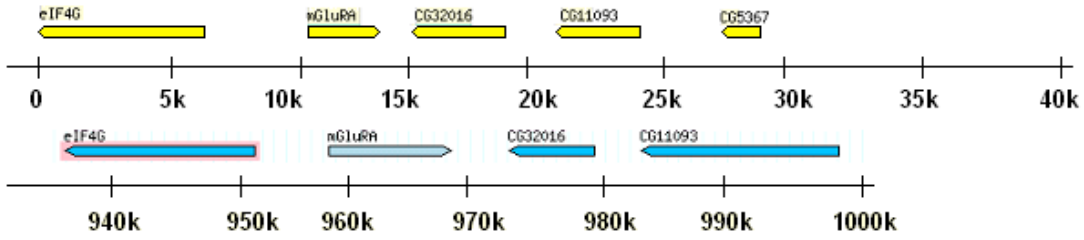


Figure 11a. Map of gene order and orientation in *D. grimshawi* (top row) against that of *D. melanogaster* (bottom row).

Conclusion

Of the five features predicted by GENSCAN, three represent fairly accurate predictions. The three accurately predicted orthologous features—*eIF4G*, *mGluRA*, and *CG32016*—all have multiple exons and for CG32016, multiple isoforms. GENSCAN mistakenly predicted the ortholog of *mtt* to be in the same region as *mGluRA* but this was found to be a mismatch due to high conservation. Similarly, *CG4847* was mistakenly predicted where *CG5367* in reality was. Finally, *CG11093* was mistakenly considered to be part of feature C. In total, four gene models were established for *D. melanogaster* orthologs in *D. grimshawi* with one possible pseudogene (*eIF4G*). Repetitious elements were fairly low at 4.31% but the first 13 kB of DGA05E01 may be part of a larger tandem repeat region. Synteny of most of the features is conserved with the same relative order and orientation of all genes except for *CG5367* which instead comes from chromosome 2L of *D. melanogaster*. ClustalW analysis shows that mGluRA is well conserved through many Drosophilids and even in mammals. However, there was no conserved regulatory sequence evidence found in the first 2 kB upstream of the start site.

Feature – Location – Gene Span Size (bp)
eIF4G: spans off fosmid-5962 – 5962 bp in fosmid DGA05E01
CG32016: 19680-15637 – 4044 bp
mGluRA: 10136-13817 – 3682 bp
CG5367: 28810-27510 – 1301 bp
CG11093: 24324-21242 – 3083 bp

Table 4. Summary of features.

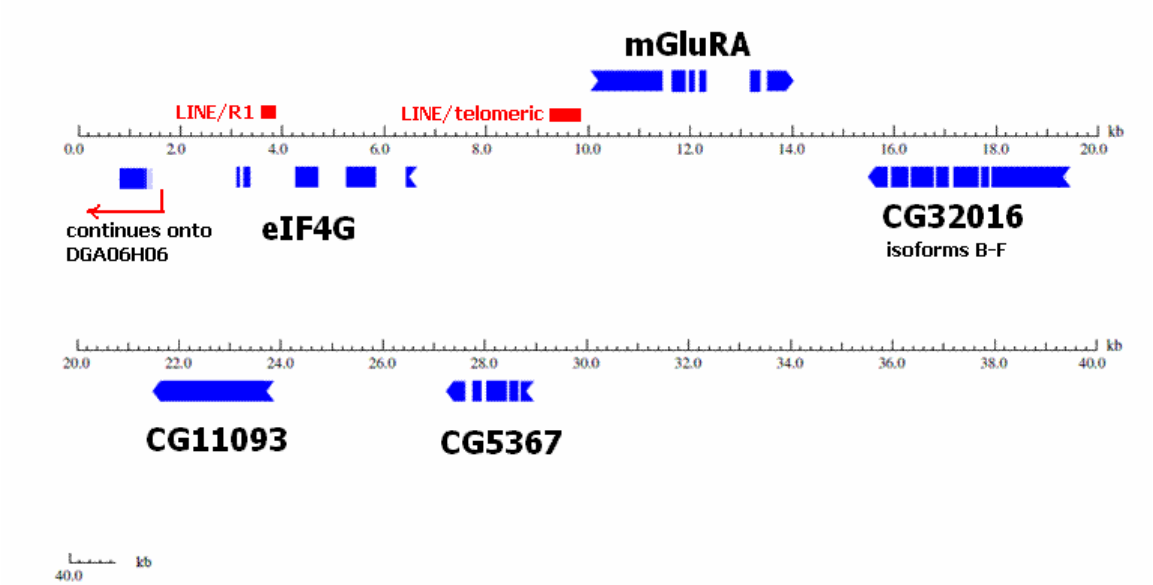


Figure 12. Final map of features.

Appendix

Feature Files

Please see DGA05E01.fasta for DNA sequence, DGA05E01.pep for peptide sequences, and DGA05E01.gff for all GFF data.

Works Cited

[1] Drosophila 12 Genomes Consortium (2007). Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203-218.

[2] Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78-94.

[3] Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H.*, Valentin F.*, Wallace I.M., Wilm A., Lopez R.*, Thompson J.D., Gibson T.J. and Higgins D.G. (2007) ClustalW and ClustalX version 2. Bioinformatics 2007 23(21): 2947-2948.

[4] S. Tweedie, M. Ashburner, K. Falls, P. Leyland, P. McQuilton, S. Marygold, G. Millburn, D. Osumi-Sutherland, A. Schroeder, R. Seal, H. Zhang, and The FlyBase Consortium. FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Research (2009) 37: D555-D559; doi:10.1093/nar/gkn788.

[5] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D.

The human genome browser at UCSC. *Genome Res.* 2002 Jun;12(6):996-1006.

[6] Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita P, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ. The UCSC Genome Browser database: update 2010. *Nucleic Acids Res*. 2010 Jan;38(Database issue):D613-9. Epub 2009 Nov 11.

[7] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.

[8] *EMBOSS: The European Molecular Biology Open Software Suite* (2000) Rice,P. Longden,I. and Bleasby,A. Trends in Genetics 16, (6) pp276--277

[9] Kent WJ. BLAT - the BLAST-like alignment tool. *Genome Res.* 2002 Apr;12(4):656-64.