Annotation of *D. virilis* fosmid 1

5/1/07

Genomics 4342W

Lisa Sudmeier

I. Overview

After finishing the genome sequence of organisms that data must be interpreted via the process of annotation. The main goal of annotation is to locate and characterize intron and exon boundaries of genes in a particular sequence. Annotation also includes the analysis of repeats within a sequence, which is of particular importance in *Drosophila* annotation because it gives insights into the mysteries of heterochromatin and euchromatin.

Before beginning to annotate fosmid 1 of the 4th chromosome of *Drosophila virilis,* RepeatMasker was used to mask repetitive and low complexity regions in the sequence. To begin annotation, the BlastX program was used to compare the entire fosmid sequence with the *Drosophila melanogaster* annotated peptide sequences in FlyBase. Five significant alignments with genes on the *Drosophila melanogaster* 4th chromosome were obtained. These five genes were PlexB, Mitf, Arf102F, ci, and CG17471. The UCSC Genome Browser also showed RefSeq data that suggested the presence of two additional genes in the fosmid. The amino acid sequences of these two genes were obtained from ensembl and then compared to the whole fosmid sequence using BlastX. Significant alignments suggested that the fosmid contained CG31997 and CG41130, both 4th chromosome genes in *D. melanogaster*. It was also noted that certain regions of the fosmid lacked RefSeq data (see figure below). To determine whether or not these regions contained genes, the DNA was extracted and compared to the annotated peptide sequences of *D. melanogaster* using BlastX. A significant alignment was obtained with CG33521, another 4th chromosome gene in *D. melanogaster*. Further investigation showed that only parts of PlexB and CG33521 were present in the fosmid, suggesting that this fosmid contains 6 complete genes and 2 partial genes in total, which are all located on the 4th chromosome in *D. melanogaster* (see figures below).

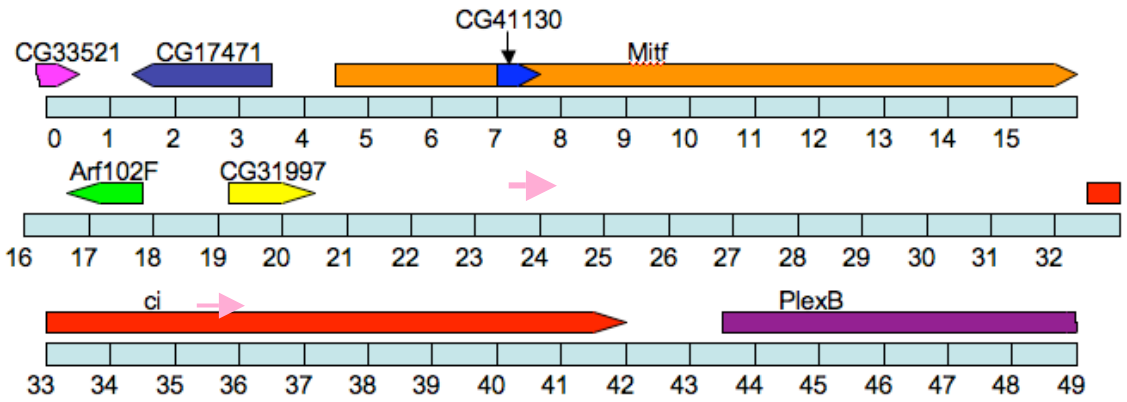


Figure 1: Map of genes located in fosmid 1. Pink arrows are repeats over 500bp. Scale: each gray rectangle = 1kb

Gene	Accesssion Number	Location	Isoforms
CG33521	NP_001014702 - 5	? – 659 bp	4
CG17471	NP_001033805 - 6	1,373 - 3,409 bp	2

Mitf	NP_001033807 - 8	4,533 – 16,024 bp	2
CG41130	NM_001015076	7,168 – 7,617 bp	1
Arf102F	NP_524631	16,858 – 17,761 bp	1
CG31997	NP_726539	19,195 – 20,393 bp	1
Ci	NP_524617	32,615 – 41,986 bp	1
plexB	NP_524616	43,527 bp - ?	1

Figure 2: Summary of genes in fosmid 1

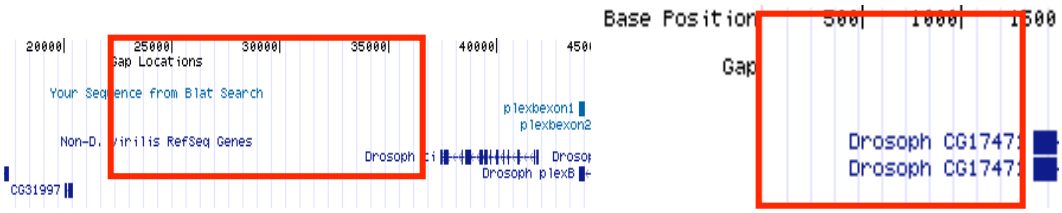


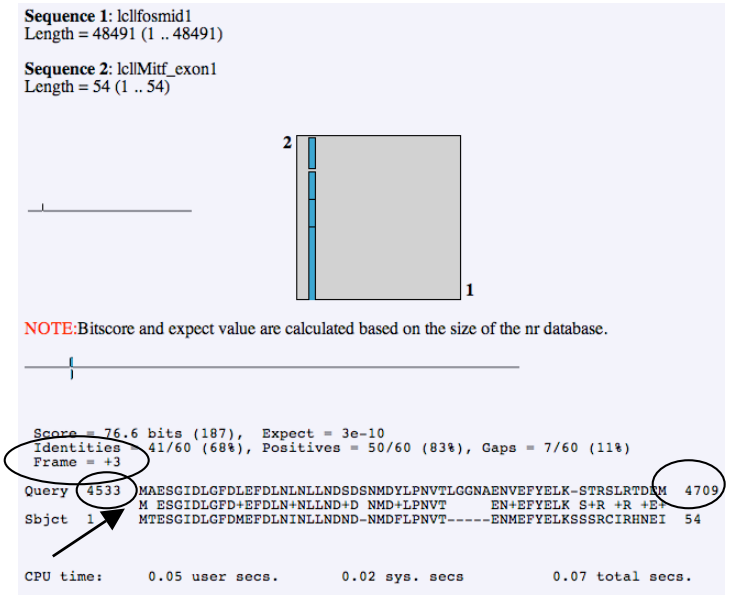
Figure 3: DNA from the regions in the red boxes was extracted and compared to annotated *D. melanogaster* proteins to see if genes were present in the regions despite the lack of RefSeq data. CG33521 was found in the region to the right.

II. Genes

After identifying the possible genes in the fosmid, ensembl was used to identify the amino acid sequence of each of these genes and to determine whether or not they had multiple isoforms. BlastX with an average expect value of 10,000 was then used to compare the amino acid sequence of each exon of each gene to the whole sequence of fosmid 1 to determine approximate intron/exon boundaries and the frame of each exon. Knowing these locations, the start and stop codons as well as each splice site donor and acceptor could be found using the UCSC Genome Browser. A gene checker, which looked for stop codons in open reading frames and checked that splice site donors and acceptors were in the same phase was then used to confirm each gene model.

Annotating Mitf

Ensembl was used to obtain the amino acid sequence of Mitf and identify isoforms. Mitf has two isoforms (A and B), but comparing the amino acid sequences of each of these isoforms showed that they were identical. Therefore, the difference between these two isoforms was not in their translated sequence. Each exon of Mitf was then individually compared to the fosmid sequence using BlastX. The start codon location was identified from the Blast search with the first exon of Mitf (see figure below). The frame of each alignment was recorded (circled in figures below). The Blast search with exon 5 required a large expect value (100 million) before it aligned with a region of the fosmid because the exon was only 7 amino acids long (refer to figure below). The last coding base of the Mitf open reading frame could be identified from the Blast search with exon 9 (the last exon). In some cases, as with exon 4 (shown below), the Blast output did not predict the exact splice site locations. However, the Blast prediction was never off by more than 6 bases.



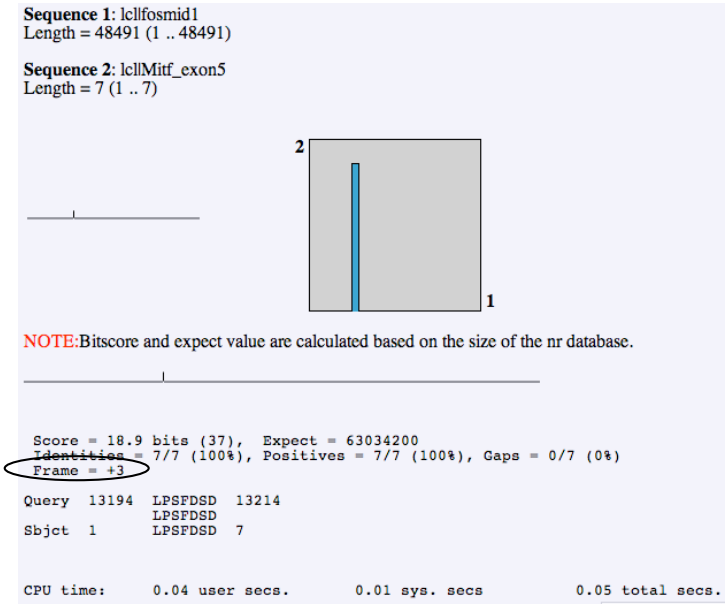
Sequence 1: lcllfosmid1
Length = 48491 (1 .. 48491)

Sequence 2: lcllMitf_exon1
Length = 54 (1 .. 54)

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

```
 Score = 76.6 bits (187),  Expect = 3e-10
 Identities = 41/60 (68%), Positives = 50/60 (83%), Gaps = 7/60 (11%)
 Frame = +3

Query  4533  MAESGIDLGFDLEFDLNLNLLNDSDSNMDYLPNVTLGGNAENVEFYELK-STRSLRTDEM  4709
             M ESGIDLGFD+EFDLN+NLLND+D NMD+LPNVT     EN+EFYELK S+R +R +E+
Sbjct  1     MTESGIDLGFDMEFDLNINLLNDND-NMDFLPNVT-----ENMEFYELKSSSSRCIRHNEI  54

CPU time:     0.05 user secs.       0.02 sys. secs          0.07 total secs.
```

Figure 4: Exon 1 of Mitf compared to fosmid 1 (BlastX). Arrow points to Methionine (start codon). Frame and base numbers of beginning and end of exon are circled.



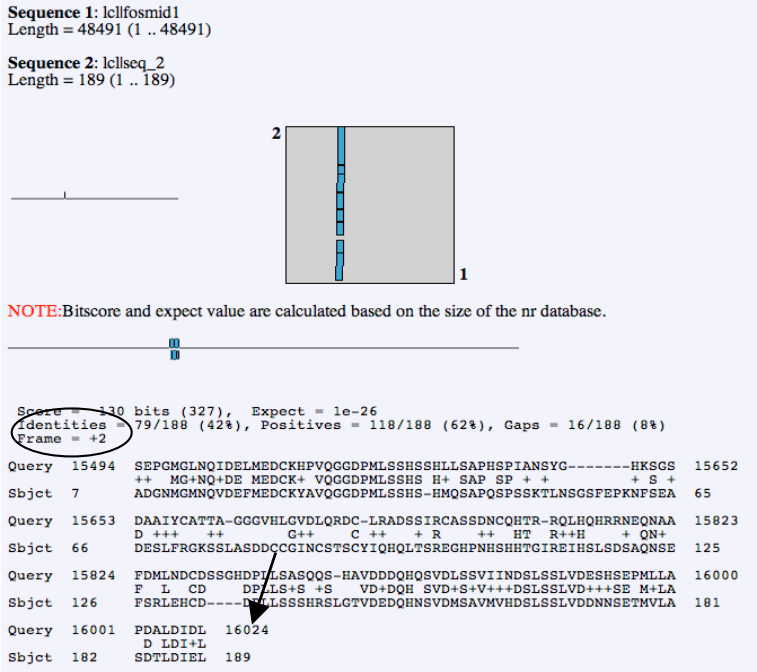
Sequence 1: lcllfosmid1
Length = 48491 (1 .. 48491)

Sequence 2: lcllMitf_exon5
Length = 7 (1 .. 7)

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

```
 Score = 18.9 bits (37),  Expect = 63034200
 Identities = 7/7 (100%), Positives = 7/7 (100%), Gaps = 0/7 (0%)
 Frame = +3

Query  13194  LPSFDSD  13214
              LPSFDSD
Sbjct  1      LPSFDSD  7

CPU time:     0.04 user secs.       0.01 sys. secs          0.05 total secs.
```

Figure 5: Exon 5 of Mitf. An expect value of 100 million was used to find this small exon.



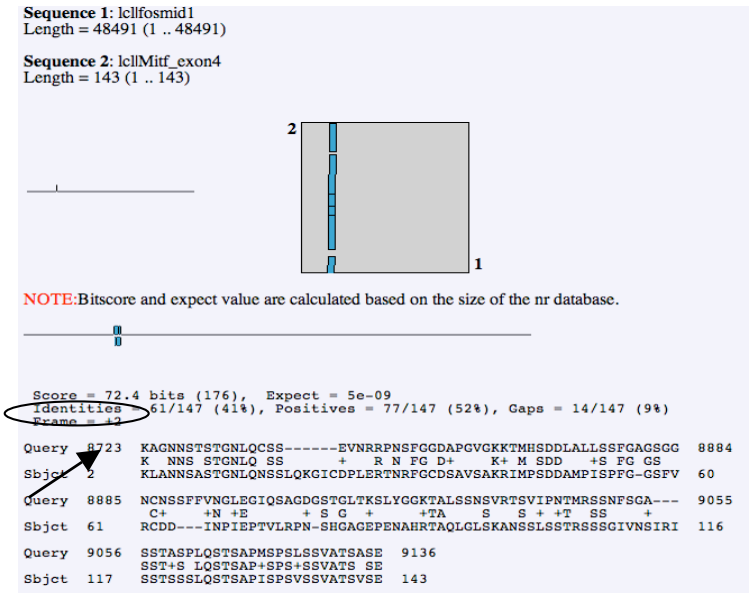
Sequence 1: lcllfosmid1
Length = 48491 (1 .. 48491)

Sequence 2: lcllseq_2
Length = 189 (1 .. 189)

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

```
 Score =  130 bits (327),  Expect = 1e-26
 Identities = 79/188 (42%), Positives = 118/188 (62%), Gaps = 16/188 (8%)
 Frame = +2

Query  15494  SEPGMGLNQIDELMEDCKHPVQGGDPMLSSHSSHLLSAPHSPIANSVG-------HKSGS  15652
              ++   MG+NQ+DE MEDCK+ VQGGDPMLSSHS H+ SAP SP + +          + S +
Sbjct  7      ADGNMGMNQVDEFMEDCKYAVQGGDPMLSSHS-HMQSAPQSPSSKTLNSGSFEPKNFSEA  65

Query  15653  DAAIYCATTA-GGGVHLGVDLQRDC-LRADSSIRCASSDNCQHTR-RQLHQHRRNEQNAA  15823
              D +++    ++          G++    C ++    + R      ++   HT   R++H     + QN+
Sbjct  66     DESLFRGKSSLASDDCCGINCSTSCYIQHQLTSREGHPNHSHHTGIREIHSLSDSAQNSE  125

Query  15824  FDMLNDCDSSGHDPLLSASQQS-HAVDDDQHQSVDLSSVIINDSLSSLVDESHSEPMLLA  16000
              F  L  CD     DPLLS+S +S    VD+DQH SVD+S+V+++DSLSSLVD+++SE M+LA
Sbjct  126    FSRLEHCD----DLLSSSHRSLGTVDEDQHNSVDMSAVMVHDSLSSLVDDNNSETMVLA  181

Query  16001  PDALDIDL  16024
               D LDI+L
Sbjct  182    SDTLDIEL  189
```

Figure 6: Final exon (9) of Mitf. Arrow points to number of last coding base.



Sequence 1: lcllfosmid1
Length = 48491 (1 .. 48491)

Sequence 2: lcllMitf_exon4
Length = 143 (1 .. 143)

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

```
 Score = 72.4 bits (176),  Expect = 5e-09
 Identities = 61/147 (41%), Positives = 77/147 (52%), Gaps = 14/147 (9%)
 Frame = +2

Query  8723  KAGNNSTSTGNLQCSS------EVNRRPNSFGGDAPGVGKKTMHSDDLALLSSFGAGSGG  8884
              K  NNS STGNLQ SS      +   R N FG D+    K+ M SDD   +S FG GS
Sbjct  2      KLANNSASTGNLQNSSLQKGICDPLERTNRFGCDSAVSAKRIMPSDDAMPISPFG-GSFV  60

Query  8885  NCNSSFFVNGLEGIQSAGDGSTGLTKSLYGGKTALSSNSVRTSVIPNTMRSSNFSGA---  9055
               C+    +N +E      + S G  +     +TA    S   S + +T  SS    +
Sbjct  61    RCDD---INPIEPTVLRPN-SHGAGEPENAHRTAQLGLSKANSSLSSTRSSSGIVNSIRI  116

Query  9056  SSTASPLQSTSAPMSPSLSSVATSASE  9136
              SST+S LQSTSAP+SPS+SSVATS SE
Sbjct  117   SSTSSSLQSTSAPISPSVSSVATSVSE  143
```

Figure 7: BlastX with exon 4. Note that the base predicted to be the start of this exon (arrow) does not correspond to the splice site chosen when looking at the genome browser.

The Blast searches above along with those for the other exons of Mitf identified the locations to look for splice sites in the genome browser. A splice site donor is almost always 'GT' and a splice site acceptor is 'AG.' The exon boundaries were mapped by navigating to the predicted exon boundaries in the Genome Browser and locating the 'GT' and 'AG' that could serve as splice site acceptors and donors. It was important in this process to pay attention to the phase of each splice site. In order for a splice site donor and acceptor to be a pair, the exons must splice together so that each exon is in the correct frame. For example, if one exon has a splice site donor that would leave the exon with one extra base on its end (phase 1), the neighboring exon must provide a splice site acceptor in its frame that leaves that exon with two bases at the beginning that form one

codon with the last base of the other exon. In the Mitf example, both the splice site acceptor and donor are in phase 0, so each exon is spliced so that all bases form complete codons on the neighboring ends of each exon (see figures below).

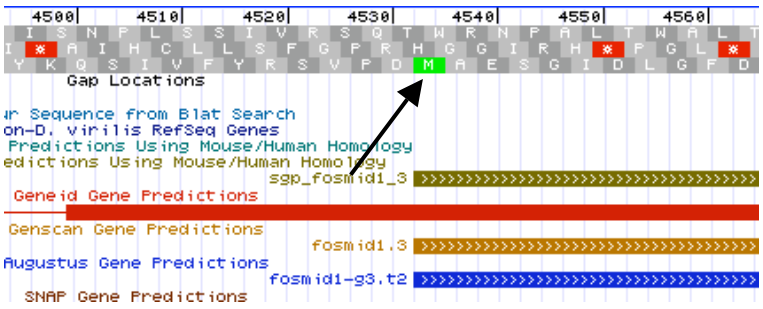


Figure 8: Genome Browser showing a start codon for Mitf in the same position (4533) predicted by the BlastX search (in frame +3)

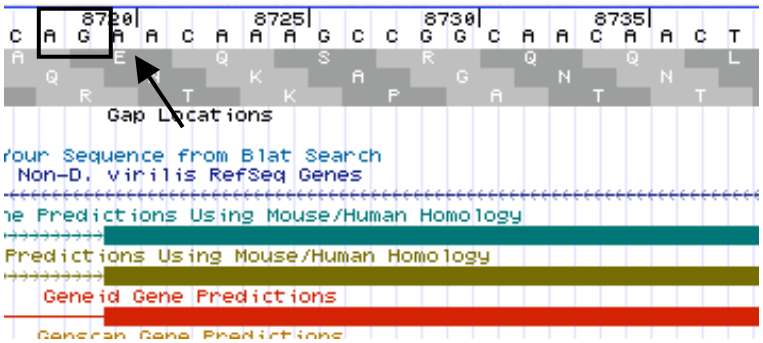


Figure 9: Genome Browser showing the splice site (in box outline) and first coding base (arrow) of exon 4. Note that this does not correspond to the site that was predicted to be the start of the exon from the BlastX output.

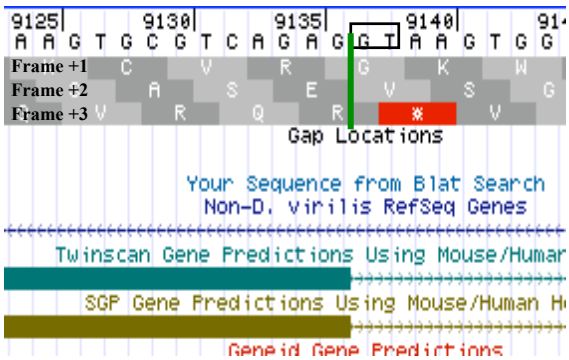


Figure 10: End of exon 4 with splice site donor in box. Note that exon 4 is in frame +2, so the splice site is in phase 0 (refer to green line indicating the end of the spliced exon).

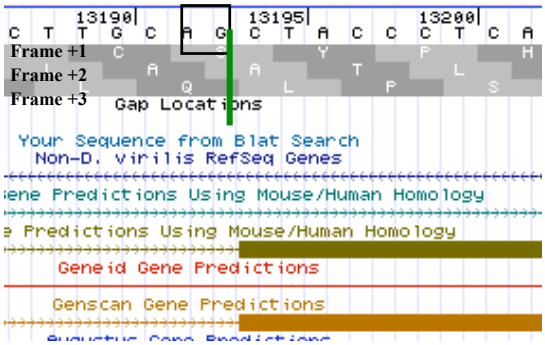


Figure 11: Beginning of exon 5 with splice site acceptor in box. Note that exon 5 is in frame +3, so the splice site is in phase 0 (refer to green line, indicating the start of the spliced exon).

Following the procedure outlined above, the boundaries of all the exons of the Mitf gene were located. The gene checker confirmed the proposed model. This same procedure was used to annotate the other seven features in fosmid 1. Although some of the features had more than one isoform, as recorded in the table above, only CG33521 had isoforms with different peptide sequences. However, the exons of this gene that differed in amino acid sequence were not included in the fosmid sequence (note on the

figure above that the end of the fosmid is in the middle of one of the CG33521 exons). Therefore, each isoform had an identical annotation and only differed in accession number. When the gene models of each feature were confirmed, the translated amino acid sequence was obtained from the gene checker program. These sequences were then each aligned with the amino acid sequence of the *D. melanogaster* protein using the Blast2 program. Refer to Appendix A for the Blast2 alignment of each gene. The one for Mitf is shown below.

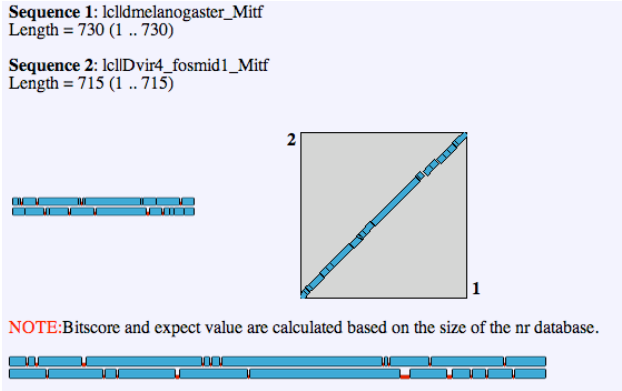


Figure 12: Blast2 alignment of *D. melanogaster* Mitf with *D. virilis* Mitf

Gene Function Summary

CG33521 has not been extensively characterized, but is known to contain a zinc-binding domain. CG17471 is likely a Phosphatidylinositol-4-phosphate 5-kinase. Mitf is a microphthalmia-associated transcription factor. A description of the function of CG41130 could not be found. Arf102F is an ADP ribosylation factor. No information was available about the function of CG31997. Ci encodes the protein cubitus interruptus, which contains a zinc finger domain and is a nuclear protein that plays a role in development. PlexB is a transmembrane receptor protein with tyrosine kinase activity.

III. Clustal Analysis

A Clustal Analysis was performed to investigate evolutionarily conserved regions of two of the genes found in fosmid 1. To begin this analysis, the BlastP program was used to search the non-redundant protein database for amino acid sequences that aligned with the sequences from Arf102F or CG17471. The amino acid sequence from numerous organisms of different evolutionary distance from *Drosophila virilis* were chosen for the Clustal analysis to facilitate investigation of the extent with which different regions of the protein had been conserved. These sequences were all compared to each other using a Clustal W multiple sequence alignment.

The Arf102F Clustal analysis showed extremely high conservation throughout all organisms from *D. virilis* to *Homo sapiens* and many species in between. Because of difficulties using the Clustal program, the amino acid sequence from the last species in the analysis (zebra fish) was not incorporated in the alignment. However, it was clear from the alignment that this protein has been conserved to an impressive extent. Very few amino acids are different among the species. Furthermore, those locations with differing amino acids still have amino acids with similar properties. Therefore, it was

concluded that the Arf102F gene product is very important for the function of all organisms included in the Clustal analysis and has consequently been under heavy selective pressure throughout evolution, thereby preventing much change in its sequence.

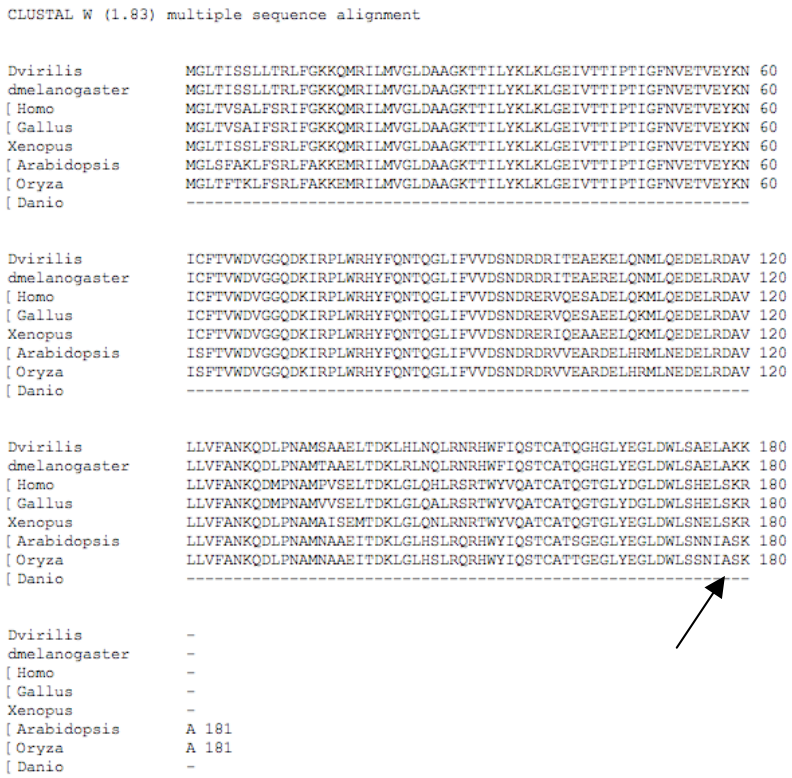


Figure 13: Clustal with Arf102F. Note the high conservation. The arrow points to a region where the amino acids seem to differ among species. However, leucine (L) and isoleucine (I) for example, are very similar.

Another Clustal analysis was performed using the sequence of the translated CG17471. This analysis also showed a high level of conservation, but not as extreme as that observed with the Arf102F analysis. In the alignment shown below, * beneath an amino acid signifies that this amino acid has been conserved in all of the sequences used for the alignment. Two vertical dots signify a high level of conservation although not all amino acids are the same in that location. One dot (.) indicates conservation, but not as high as that observed with two dots (the properties of the amino acids differ more from one another). CG17471 has undergone significant changes throughout evolution, but a number of important residues have been conserved allowing the protein to retain its function as it serves the needs of different organisms.

Figure 14: Clustal analysis of the CG17471 gene

```
clustalw.aln  (CG17471)
CLUSTAL W (1.83) multiple sequence alignment

Dvir                          --------------MDKKISSTSQPRILKKKHFRVKHQKVKLFRANEPI
Drosophila_melanogaster       --------------MEKKISSSSQPRILKKKHFRVKHQKVKLFRANEPI
Homo_sapiens                  ---MASSSVPPATVSAATAGPGPGFGFASKTKKKHFVQQKVKVFRAADPL
Mus_musculus                  ---MASSSVPPATAPAAAGGPGPGFGFASKTKKKHFVQQKVKVFRAADPL
Xenopus_tropicalis            MSSSGAMPAVSSASSSAAVGILSATTAKTKTKKKHFVQQKVKVFRASDPL
Danio_rerio                   --------MASLGNSGSASSPMVMLAPKTKTKKRHFMQQKVKVFRASDPM
rattus_norvegicus             --------MSSNCTSTTAVAVAPLSASKTKTKKKHFVCQKVKLFRASEPI
Gallus_gallus                 --------------MAAPGTVASVMASKTKTKKKHFVVQKVKLFRASDPL
Caenorhabditis_elegans        --------------------MSTKKKTKVLSKKKGKILVPKWKLFRAKEPV
                                                    ..*: :.   * *:*** :*:
```

```
Dvir                        LSVFMWGINHTINELSHVNIPVMLLPDDFRAYSKIKVDNHLFNKENMPSH
Drosophila_melanogaster     LSVFMWGINHTINELSHVNIPVMLLPDDFRAYSKIKVDNHLFNKENMPSH
Homo_sapiens                VGVFLWGVAHSINELSQVPPPVMLLPDDFKASSKIKVNNHLFHRENLPSH
Mus_musculus                VGVFLWGVAHSINELSQVPPPVMLLPDDFKASSKIKVNNHFFHRENLPSH
Xenopus_tropicalis          ISVFMWGVNHSVNELIQVPVPVMLLPDDFKANSKIKVTNHLFNRENLPSH
Danio_rerio                 LSVFMWGVNHSINDLNQVPVPVMLLPDDFKANTKIKVNNHLFNKENLPSH
rattus_norvegicus           LSVLMWGVNHTINELSNVPVPVMLMPDDFKAYSKIKVDNHLFNKENLPSR
Gallus_gallus               LSVLMWGVNHSINELSHVQIPVMLMPDDFKAYSKIKVDNHLFNKENMPSH
Caenorhabditis_elegans      LSVFMWGINHTVDQLLHVPPPGLLMPDDFKAYSKVKIDNHNFNKDIMPSH
                            :.*::**: *::::* :*  * :*:****:* :*:*: ** *::: :**:

Dvir                        FKVKEYCPLVFRNLRERFGVDDVDYRESLTRSQPLGIDSS---GKSGAQF
Drosophila_melanogaster     FKVKEYCPLVFRNLRERFGVDDVDYRESLTRSQPIQIDSS---GKSGAQF
Homo_sapiens                FKFKEYCPQVFRNLRDRFGIDDQDYLVSLTRN-PPSESEG---SDG--RF
Mus_musculus                FKFKEYCPQVFRNLRDRFAIDDHDYLVSLTRS-PPSETEG---SDG--RF
Xenopus_tropicalis          FKFKDYCPQVFRNLRERFGIDDQDFQASLTRSSPYCESEG---HDG--RF
Danio_rerio                 FEFKEYCPQVFRNLRERFGIEDLDYQASLARSAPMKGDGQ---GEG--LL
rattus_norvegicus           FKFKEYCPMVFRNLRERFGIDDQDYQNSVTRSAPINSDSQ---GRCGTRF
Gallus_gallus               FKFKEYCPMVFRNLRERFGIDDQDFQNSLTRSAPLANDSQ---ARSGARF
Caenorhabditis_elegans      YKVKEYCPNVFRNLREQFGVDNFEYLRSLTSYEPEPDLLDGSAKDSTPRF
                            ::.*:*** ******::*.::: ::  *::   *              :

Dvir                        YQSYDKFFIIKSLTSEEIERMHAFLKHYHPYVVERHGKTLLPQYLGMYRI
Drosophila_melanogaster     YQSYDKFFIIKSLTSEEIERMHAFLKQYHPYVVERHGKTLLPQYLGMYRI
Homo_sapiens                LISYDRTLVIKEVSSEDIADMHSNLSNYHQYIVKCHGNTLLPQFLGMYRV
Mus_musculus                LISYDRTLVIKEVSSEDIADMHSNLSNYHQYIVKCHGNTLLPQFLGMYRV
Xenopus_tropicalis          LLSYDKTLVIKEISSEDVADMHNILSHYHQHIVKCHGNTLLPQFLGMYRL
Danio_rerio                 FTSYDRTLIVKQISSEEVADMHNILSEYHQHIVKCHGSTLLPQFLGMYRI
rattus_norvegicus           LTTYDRRFVIKTVSSEDVAEMHNILKKYHQFIVECHGNTLLPQFLGMYRL
Gallus_gallus               HTSYDKRYIIKTITSEDVAEMHNILKKYHQFIVECHGNTLLPQFLGMYRL
Caenorhabditis_elegans      FISYDKKFVIKSMDSEAVAELHSVLRNYHQYVVEKQGKTLLPQYLGLYRL
                              :**:  ::* : ** :  :*  * .** .:*: :*.*****:**:**:

Dvir                        TVESVQYYFVVMRNVFSSQLTIHKKFDLKGSTVDREASEKELEKLLPTYK
Drosophila_melanogaster     TVESVQYYFVVMRNVFSSHLTIHKKFDLKGSTVDREASEKELEKNLPTFK
Homo_sapiens                SVDNEDSYMLVMRNMFSHRLPVHRKYDLKGSLVSREASDKEKVKELPTLR
Mus_musculus                SVENEDSYMLVMRNMFSHRLPVHRKYDLKGSLVSREASDKEKVKELPTLK
Xenopus_tropicalis          SVDNEDSYIMVMRNMFSHRLTVHRKYDLKGSLVSREASDKEKIKELPTLK
Danio_rerio                 TVESEDTYLIVMRNMFSHRLLVHRKYDLKGSLVDREASDKEKVKELPTFK
rattus_norvegicus           TVDGVETYMVVTRNVFSHRLTVHRKYDLKGSTVAREASDKEKAKDLPTFK
Gallus_gallus               TVDGVEIYMIVTRNVFSHRLSVYRKYDLKGSTVAREASDKEKAKELPTFK
Caenorhabditis_elegans      TIEGSETYLIVMRNVFGRKYGVHTKFDLKGSTVSRAASDKEKAKDLPTLK
                            :::. : *::* **:*. :  :: *:***** * * **:**  * *** :

Dvir                        DNDFIKQKVKLEIGEEPKKKLLDTLNNDVDLLTKLHIMDYSLLVGIHDCV
Drosophila_melanogaster     DNDFIKQKVKLDIGKEAKDKLMDTLSNDVDLLTKLHIMDYSLLVGVHDCV
Homo_sapiens                DMDFLNKNQKVYIGEEEKKIFLEKLKRDVEFLVQLKIMDYSLLLGIHDII
Mus_musculus                DMDFLNKNQKVYIGEEEKKVFLEKLKRDVEFLVQLKIMDYSLLLGIHDII
Xenopus_tropicalis          DMDFLNKSQKVYVDEEQKKNFMEKLKRDVDFLVQLKLMDYSLLLGIHEVF
Danio_rerio                 DMDFRNNMQKVYVTEEQKEKMMEKLNRDVEFLVKLKIMDYSLLLGIHDVA
rattus_norvegicus           DNDFLNEGQKLRVGEESKKNFLEKLKRDVEFLAQLKIMDYSLLVGIHDVD
Gallus_gallus               DNDFINDGQKIHIDENNKRMFLEKLKKDVEFLAQLKLMDYSLLVGIHDVE
Caenorhabditis_elegans      DNDFLEQNWKLNLPPEAGKLLIEMLTSDTEWLTRMHLMDYSLLVGIHDCE
                            * ** :.  *: :  :    ::: *. *.: *.::::******:*:*:
```

```
Dvir                          RAEEEALQGDSAQATGRSENSESEECDSGERWTYNTPPDS---------P
Drosophila_melanogaster       RAEEEALQGDNILTVGRSENSESEECDSGERWTYNTPPDS---------P
Homo_sapiens                  RGSEPEEEAPVREDESEVDGDCSLTGPPALVGSYGTSPEGIGGYIHSHRP
Mus_musculus                  RGSEPEEEGPVREEESEWDGDCNLAGPPALVGSYGTSPEGIGGYIHSHRP
Xenopus_tropicalis            RAEQEEEE------DLEEDHTENESSPHMNVGSYGTSPEGIAGYLNSHKP
Danio_rerio                   RGEREEEEAEEPCYEDDADPE-NGLAPALQVGSYGTSPEGIAGYMNSIKP
rattus_norvegicus             RAEQEETEVEDRAEEEECENDGVGGG---LLCSYGTPPDSPGNLLSFPRF
Gallus_gallus                 RAEQEEVECEENDGEDEGESDGTHP--------IGTPPDSPGNTLNSSLP
Caenorhabditis_elegans        RAAQEAANRPVEQNSEESGDE-----------LAPTPPDSPIPSTGGAFP
                              *. .   :                          *.*:.

Dvir                          RGAQYKEVVYEVDIYAIPSIEERREIYFIAIIDVLTQYGVKKQAAKAAKT
Drosophila_melanogaster       RGAQYKEVVYEVDIYDIPSIEEKREIYFIAIIDVLTQYGVKKQAAKAAKT
Homo_sapiens                  LGPGEFESFIDVYAIRSAEGAPQKEVYFMGLIDILTQYDAKKKAAHAAKT
Mus_musculus                  LGPGEFESFIDVYAIRSAEGAPQKEVYFMGLIDILTQYDAKKKAAHAAKT
Xenopus_tropicalis            LGPGEFEPIIDVYAIKSSDNAPQKEVYFMGLIDILTHYDAKKKAAHAAKT
Danio_rerio                   LGPGEFDPYIDVYAVKSAPGAPQREVYFMGLIDVLTQYDTKKKAAHAAKT
rattus_norvegicus             FGPGEFDPSVDVYAMKSHESAPKKEVYFMAIIDILTPYDAKKKAAHAAKT
Gallus_gallus                 LAPGEFDPAIDVYGIKSHESAPRKEVYFMAIIDILTHYDAKKKAAHAAKT
Caenorhabditis_elegans        GVSGGPDLDDEFYAIASPADFEKNLIYFIGLVDILTYYGIKKRSATAAKT
                                .   :   :.          :. :**:.::*:** *. **::* ****

Dvir                          VKYGSNVDGISTCDPEQYAKRFLEFMDKAIE
Drosophila_melanogaster       VKYGSNVDGISTCDPEQYAKRFLDFMDKAIE
Homo_sapiens                  VKHG-AGAEISTVHPEQYAKRFLDFITNIFA
Mus_musculus                  VKHG-AGAEISTVHPEQYAKRFLDFIANIFA
Xenopus_tropicalis            VKHG-AGAEISTVHPDQYGKRFLEFVTNIFA
Danio_rerio                   VKHG-AGAEISTVHPEQYAKRFREFISNIFA
rattus_norvegicus             VKHG-AGAEISTVNPEQYSKRFNEFMSNILT
Gallus_gallus                 VKHG-AGAEISTVNPEQYSKRFLDFIANILT
Caenorhabditis_elegans        VKYGSDAENISTVKPEQYAKRLVEFVSRALN
                              **:*     *** .*:**.**: :*: . :
```

One last Clustal analysis was performed with the region upstream of the Arf102F gene in multiple different *Drosophila* species. To carry out this analysis, the 1kb directly upstream of the Arf102F gene in 6 different *Drosophila* species including *D. yakuba, and D. ananassae,* was extracted using the UCSC genome browser. This 1kb of sequence from each of the 6 selected *Drosophila* species was aligned using Clustal W. If loci of significant conservation were observed, it could have been hypothesized that the region contains gene regulatory elements because these would have been under selective pressure and therefore would not have evolved as quickly as non-functional sequence. The results of this alignment showed no highly conserved regions among all of the species, but two regions of interest showing conservation are marked with red boxes in figure 16 below. Further investigation of these regions would require wet lab experiments.

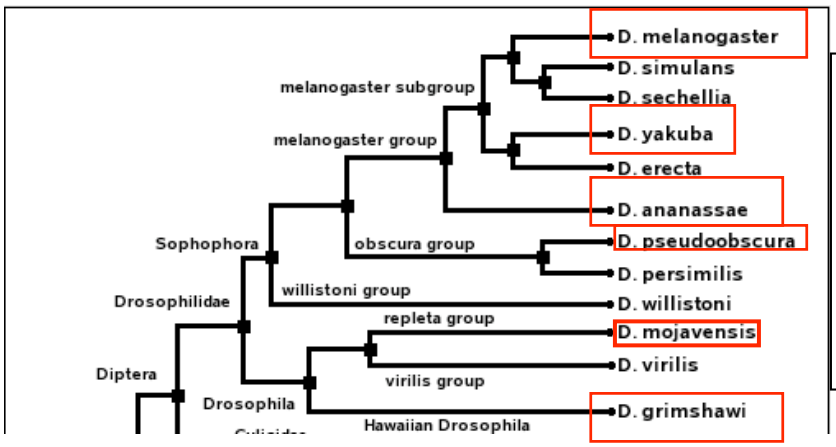


Figure 15: *Drosophila* species (in red boxes) from different subgroups were chosen for the Arf102F upstream Clustal analysis.

Figure 16: Clustal Analysis if the 1kb upstream from the Arf102F gene. Conserved regions of interest are in red boxes.

```
clustalw.aln

CLUSTAL W (1.83) multiple sequence alignment


dyakuba             AATTAACTGTGCACTTTAATGAAAGGACCGAAAAAATTTAATTCTTTTATTTGTTTACTT
dpseudoobscura      ----------ACTCTTTCGTTTGTTTAGACATTGGCTTTCGTCGTTTATTGTTATATCAC
dgrimshawi          ----------TATATAAATGCTAACTACACACACACATACACTATATG-TATGTCTGCTC
dananassae          ------------ATAAATTTCGCTCTTAACACACACTCAAGCCCTTCACCATGAAAAACA
dmelanogaster       ---------ACATTTTTTTTTTTATTTTTGATCAGAGGCAGAATATTTATGTTGTTAACTA
dmojavensis         ------------AATGAATACGAAATTCAGTTCAGATCGACGATTCCGAACGCCGACCAA
                                 : : :      :  :   :   .         *         : .

dyakuba             AGTTTCTTACTTCTGTCAGCAAAATGATTACTTATGGCGTTTGAGAAACCAG-TTGTAAG
dpseudoobscura      GATTTCTGGGTACT------------TATAAATCAGCAGTTTAAGTGGCTGGGTTCTGTG
dgrimshawi          ATTCATATATAAATG-----------CATTATGCATATATTGAATTTGATTAATTCTACA
dananassae          TCCAATAAGGTAACG----------ATGTAAAATCTCAAACTGATACTAAAGGTACGCAG
dmelanogaster       TTGCAAATCGCAACC-------------AAAATATAGTGATTGTCATTTATCTTGGCCTG
dmojavensis         TTCCGAAACGAGTCCG-----GTTCAAGCGAAAGTAAACATCGAATTGAAACTAACCGAC
                          :                        .:       :  .: :       :

dyakuba             TTTAAATACAGTAGTCAATACCCTGGAAATGTTTATCCGACATTTTTAGATTCTT--GTG
dpseudoobscura      TTTCTCTCCGTTGGTGTATAAGTG--------TGCGCCGGTATCCGTG------C--GTG
dgrimshawi          CTTTATACATGCAGGAAGAGCATT-----AATTAAGCCGGTGCAGTTG---------GCA
dananassae          TCTCTACTACAACTTGAATTCCCG--------ATCTCCCGTGCGCTTGGACTCCCGTGCA
dmelanogaster       ATATCGACATATTGGCAATAGCAG---------ACCTTTGATTAATAAAAATGTCTTAAT
dmojavensis         CAGTATCACAACTATTTATAGGCG----------AAATCGTGCTTGTG-----------G
                           .       :.:               .    .      :.

dyakuba             TAGTAGAGATTAATACTTGAAATGCGATAAAAAGAAACACCAT-ATAAGGCCCTGTG-CA
dpseudoobscura      ACGAAATGGTCGAGACCTCCTCTGCGG----AATTATCACCTT-TCAACAGGCTGCCGCC
dgrimshawi          TTTTTTTCCCCCAAATTTGGTGCGCGGCAACATTTGTCGCCATTTCAAGATGTAGTTTTC
dananassae          AGGTCCTTCTCGACAACACTAATATGGACACAAGGCTCTCAGCTGCGGGAACATGCG---
dmelanogaster       TTTTGGTTAATAGAGAGGAAGAACCTAAGTTTTTTGTTTTCAGGCAGCTTTATAGTGTTT
dmojavensis         ATTTTCGAGTTTCTTTAGTTGTTGCTGCTTTTGTTGTTATTGTTGATGTAGACATGAGTC
                    :  :                        .    :    :                :

dyakuba             GACTACGCTCCACCGATTTAAGCGACATTAAGTCCAGACCATATTTTATT---AACGTAA
dpseudoobscura      AGACACTATCCGCGCACATACACGTTTCAAGATGCATGCGTTGGCTCATGGAGTAGGCGC
dgrimshawi          AATTTCATAACAAGCTGG-----ATATTACGAAACAAATCTTGAAATAAG---CACTATA
dananassae          -GTTACATGGTATATTCAAAACGATAACATAAAAAAGATCTTGGCATCCCG-AAATTTCA
dmelanogaster       GAAAAATTTGTAACATCGTTTTTAAATTTGATTTTTAAATTTGATTTGAAAAAGTTGTTT
dmojavensis         GGTAACTGACAAGCTCTGCTTTGTCAACTAGGTGTTAGTCGTTGATATTAG-CGTATGTA
                     .  :.     .              : : . :  : .   *   :        :
```

```
dyakuba             TGAGTTAGGGAATGCTTAAACTGTTTTTTATGCTTACCATAGGTCCTGGGGTTGTTTATG
dpseudoobscura      TGAGCTAGTTGGTGGCTGGGTGGCTCGTACGGCCTTACGGGGAACGGTGGGATTTATTTA
dgrimshawi          TAAATTAAATTTAACATATGTATTTGTATTT---ATACTTATTAGTATTTTATATATGCA
dananassae          AGGCTTAAATCGACCTACCGTAGCCCCAAACCATACACAGTCACCATACACATAAGCTAG
dmelanogaster       TGGCAATTAGAATTCGCGAATAATACCATAG-TCCTACAGAGCACATTCG---TAAAAAA
dmojavensis         TGCATGTATATATATATATATATATATGTGCGTGTGTGTTTCTTTGTGGGTATAATGCAG
                    :.    :     :      .       :                        :    .

dyakuba             GGTGGGTTTTTGTTATCTATAGGTATTCAAATTTCTGGATATTCCACAACACCTTGACTG
dpseudoobscura      CAGAAGTGAAACTCCGATTCACACTGGCACCATCCGTCAAAAACAAGTCCATTAAATTTT
dgrimshawi          CATTAATTATTAAAACATAATTACACACCTCGGACCACTGTGCTAAAAACATTTAAATCA
dananassae          CTCGCTCTCAAGCAATTTTATTCATTGCAATCATATATAGAAATACTGTCCCAAACTTTT
dmelanogaster       CAGTTGTGATTTTTAAAAATTGTGTAACTCCG--AGTCTAAGTTATATGCTCACTAACCC
dmojavensis         CTATGAGTTTTCTTAAAACCTTGCACACATACATGCACATTGACAAAGAAAGAGCATTAA
                            ::   .  :  :   :  *          : :   .    .      :

dyakuba             TACTATTTATTTTCTTATTGGACTCAACATATATGGTTTATTTAGAAAAATATTAAAAAT
dpseudoobscura      CGCGGAAAATTTTTTAATTTTTTTCTTATAATGG-----ATGATGAAGGATGTAGCAATT
dgrimshawi          AATGGCGGATTTAGTGCAAG--CACAGCCAGAAGTATCGATATATCGGTGTTTAAATGTA
dananassae          CTTATTAAATTATCATCATTTTTCCAAAAATCTGGCCGCATTTCAATCAACCTAATAAGA
dmelanogaster       CTTGGGGGCAGCCTTTCTAGGTCAATTTAAATATTTCACTTAAAAAATGTACACGAAAAT
dmojavensis         TTAAGCCGG-TCCTTCAGTCTTACCACCAAATTTGGTACCTCCACATATTTCAACTACAA
                                 : . :     .:   :          *    .      :   :  :

dyakuba             CAAATATTTATAG-TCAATACCACCCTAAATAAATGGTCTGTAATTTACTTAG---CCTA
dpseudoobscura      ATTTGGGCTGTGGTTCATCAGCAAAATATGCGAACAGCCCGATAGCAGCAATAGGTCTGA
dgrimshawi          GTTTTGTCGTAATAAACTGCAAGTCAGTGGCAGTAATTTAGTTATCTTACTTATATTACT
dananassae          TTGATG-----ATTTCCATATCATCAAACAAAAGTGGCATTTCTACAATTTCCCTCCATC
dmelanogaster       GGTTTTATCGACATTCGACTTAATGGCGCAGTCAAAACATTGTATTATCGATGTAGTTTT
dmojavensis         TATATTACAGACTTAACACACCTACAATACTCAAGCACTATAATAAGACATTTTGGGGAA
                       :          :. :   .                      :      :

dyakuba             ACTTTAGTTAGACAATAACATAAAATAGCTATTATATTGGCAATAAAACATAGTGTGGAT
dpseudoobscura      TGACTAATCAGTCAGCAGAACAGATACGATCTGCAATCG----CAAAATATTGTGTAATT
dgrimshawi          AGAAATGTCAG-CTGTATCAATTACGAAATATATTCATATTTACTTAGCGTTCTAAAACA
dananassae          AGGTAAAAATTTGCCGTGTGAAATTAAACTTAGCTAGGGTCTTTTAAACTGTATG-GGCT
dmelanogaster       GTGTTTCTTTTTATTGTTTTCATCAGACCATGCTTAGGGTATAGTCAACATACTGTTATT
dmojavensis         AAATTGGATGAATAAAAAACTGTTCAGACAATACATGTG--TGCATCGATTTTTCAAGAT
                        :  :        :           .:    :   .     : ..   : *   . :

dyakuba             GAAGCCGTAGTAGCAATGATG--TAAGCATCCGGAAAACAATTTTATGGCATCAGGGTT-
dpseudoobscura      AGAAACTTATTTATTATTATTTTTCAGCACTTCCAATAAATCATTCTGGAATTTTTTTG-
dgrimshawi          AAATACAACAAAGTATCCATCTTTTGCTGCCCAGCATCGGTTTTTGCTGCCACCTAGCGG
dananassae          GGAATCGCCCTGGTAAATTTCTGGTATTTTTTTTAGATGAGAGTACTTGGTGGTGAACAG
dmelanogaster       AGTATTTTATAG--AAAACAGTATTTTTAACCATCGTTTAACTTTTCTGAAAAGTACAG-
dmojavensis         GATTTTGTAAAATCAATTTTTGTTTAAGTAATCGATAGCATATAAACTGAAATTATATGT
                    ..:     . :   ::   :              . :  .   ::   *

dyakuba             -TATGAAACTAAAAGCTCTAAGAAAAAGCCCAGTGT-AAACACCCCTTTGCTTCAAAAAA
dpseudoobscura      -CGTGTTGATGAGATCCTTGACATCGTTTAAAACGTGTAACCAAATTATGCATAGCAAAT
dgrimshawi          CCGTTCAAAGTTGGGGCAAATAGGCTTGCCTAGCAT-CGGTTGTGGTAGCATTCACACAG
dananassae          ACTCTGTCAAAAACAGTTTGTTTTATCTATTTGTTACTCATTCTCATAGCTTAAAAACAG
dmelanogaster       ---CATTATTAATCGAATGACTAGTTTACACGTCAATTTACTGAAGTTACTGAAGACAAG
dmojavensis         GCCGTTAGATAAATATATTTGAAATAGGAAAAATAAAATCCTCAAAACAATTTCCCTCTA
                          :    :                        :          :     :. . .:

dyakuba             ATGTAATTGCTTATCCCTTT-CCTACTCTTAG---CAAAAAG--CATTCGTCTAATGGCG
dpseudoobscura      CTTGACCCGATCAGATCTCT-CCTTCGAATACCTCCAAAAAA--AAAAACTACATAGGTG
dgrimshawi          GCGCCACCTAGCAGATTTCC-ACAGTTGCGAAAAATAAATCG--CAATGCATTTATAAAT
dananassae          TGTGTATGGATCGGAGTACTGGCTTTAACAATATGCTCATCG--GAGTCGCAATAGGACC
dmelanogaster       CTTTTTCCTATACGAAATAAAATAGTTTCTTATTAAACATGGGACTAACAATATCTAGTC
dmojavensis         GGTTGACTTTTCGATAATTAGGTTGGCAGGTTTTAATTTTTG--GTGACCGATTATACAG
                                    :     :      :     : :: .   : :     :  .

dyakuba             --CATACAAAGACAGAACACATTAAATGTCGGTGCAG--TTTTGTGTTTCTTTTCAGTGG
dpseudoobscura      --TTTGGGATGAAAAATAGATGTTGGCCGATGCGCAGACCTAAGCGATAAAATCCCCCGG
dgrimshawi          --TAAAAATTGTTAATTAAAAACAAATAAATAAAATTTGTTTAACAATATTTTATATGCA
dananassae          GTAAGACTTTGACTTAAAAAAATTGGTCCACAATAGTAAGATAAATTATTTTGATCTGAG
dmelanogaster       TTTTGACACGTTTGTTTGGAAAAAAACAAATGCGTATTCTTATGGGTATGTTCAAAATAC
dmojavensis         -----TATGTTTTCGATTGCATTCGGTCAATTTTGGTGCCAAATAAAATCTAAAATAAAT
                               :   :: ..:   ..   .          :::   ::: ::
```

III. Repeats

Before annotation began, RepeatMasker had identified the repeats within this fosmid. Only two repeats larger than 500 bp were located (refer to top two repeats in the repeat table in Appendix C), one of which is located within the ci gene. 13.31% of the fosmid is repetitive. The largest class of repeats present in this fosmid is simple repeats (5.55%). Of great interest in *Drosophila* repeat analyses are the transposable elements. These are the LINEs, LTR, and DNA elements. In total, transposable elements compose 3.16% of the fosmid. Refer to the table below for a more detailed repeat summary.

```
==================================================
file name: fosmid1.fasta
sequences:            1
total length:     48491 bp  (48491 bp excl N/X-runs)
GC level:         38.57 %
bases masked:      6456 bp ( 13.31 %)
==================================================
               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:               5          777 bp     1.60 %
      LINE1          0            0 bp     0.00 %
      LINE2          0            0 bp     0.00 %
      L3/CR1         0            0 bp     0.00 %

LTR elements:        0            0 bp     0.00 %
      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:        2          756 bp     1.56 %
      MER1_type      0            0 bp     0.00 %
      MER2_type      0            0 bp     0.00 %

Unclassified:        0            0 bp     0.00 %

Total interspersed repeats:     1533 bp     3.16 %


Small RNA:           0            0 bp     0.00 %

Satellites:          5          442 bp     0.91 %
Simple repeats:     35         2689 bp     5.55 %
Low complexity:     23          771 bp     1.59 %
--------------------------------------------------
```

Figure 17: Summary of repeats

IV. Syteny

A gene-by-gene comparison between *D.melanogaster* and *D. virilis* shows that synteny is maintained between three sets of genes: ci-plexB, CG17471-Mitf/CG41130, and CG33521-CG17471. However, the distance between these genes is not conserved between the species. The distance between ci and plexB is 3,931 bp in *D. melanogaster* compared to 1,541 in *D. virilis.* Similarily, there are 2,018 bp between CG17471 in *D. melanogaster,* but only 1,124 bp between them in *D. virilis.* CG17471 and CG33521 overlap by 285 bp in *D. melanogaster,* but are separated by 714 bp in *D. virilis.* The greater distance between genes in one species compared to another may be due to the presence of repeats in that locus. However, the table of repeats for the *D. virilis* fosmid does not indicate the presence of repetitive elements between CG17471 and CG33521. However, it is also significant to note that the UTR's of the *D. melanogaster* genes have been annotated along with their exons. UTR's were not characterized in the *D. virilis* annotation described above. Therefore, the coordinates of *D. melanogaster* genes in the Genome Browser include UTR's, but those of *D. virilis* do not. Consequently, the apparent overlap of CG17471 and CG33521 in *D. melanogaster* may be an overlap of their UTR's.

Since the lack of conserved synteny between other genes of the fosmid is characterized by the presence of other genes in *D. melanogaster* that are between genes located next to each other in *D. virilis.* Thus, CG31997 and Arf102F were translocated

from their neighboring genes in *D.melanogaster* to their location in *D. virilis.* Another difference between the *D. virilis* and *D. melanogaster* synteny is that the region of the chromosome containing CG33521, CG17141, and Mitf is flipped with respect to the rest of the chromosome. Therefore, in addition to the translocation events that moved CG31997 and Arf102F, an inversion also occurred.

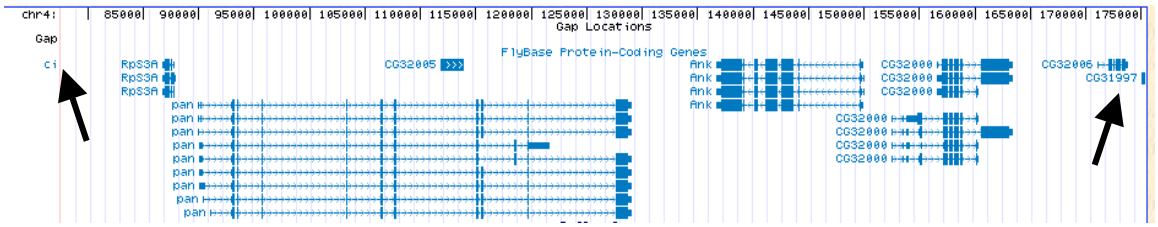


Figure 18: Genome browser showing the genes located between ci and CG31997 in *D. melanogaster.*

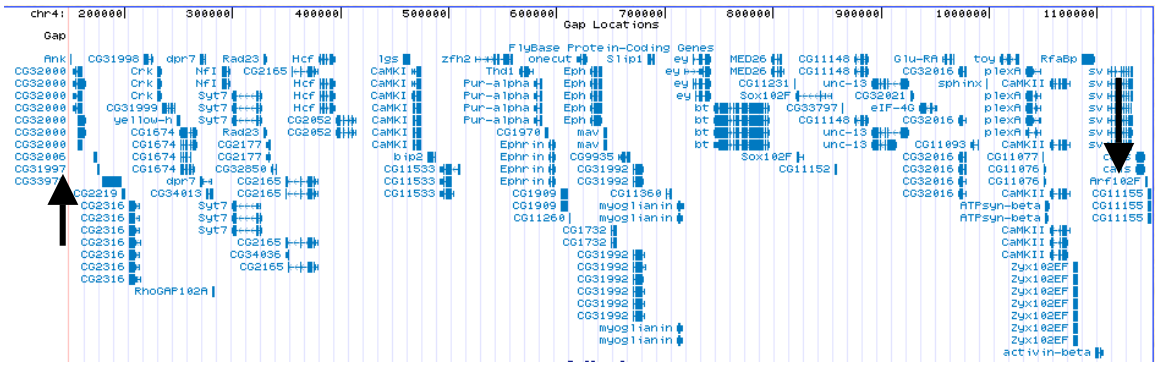


Figure 19: Genome browser showing the genes located between CG31997 and Arf102F in *D. melanogaster*

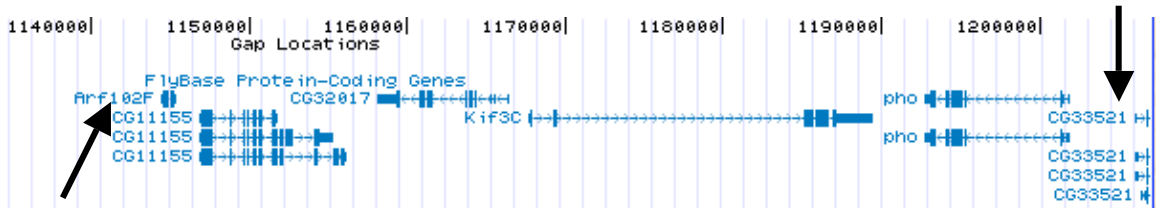


Figure 20: Genome browser showing the genes located between Arf102F and CG33521 in *D. melanogaster*

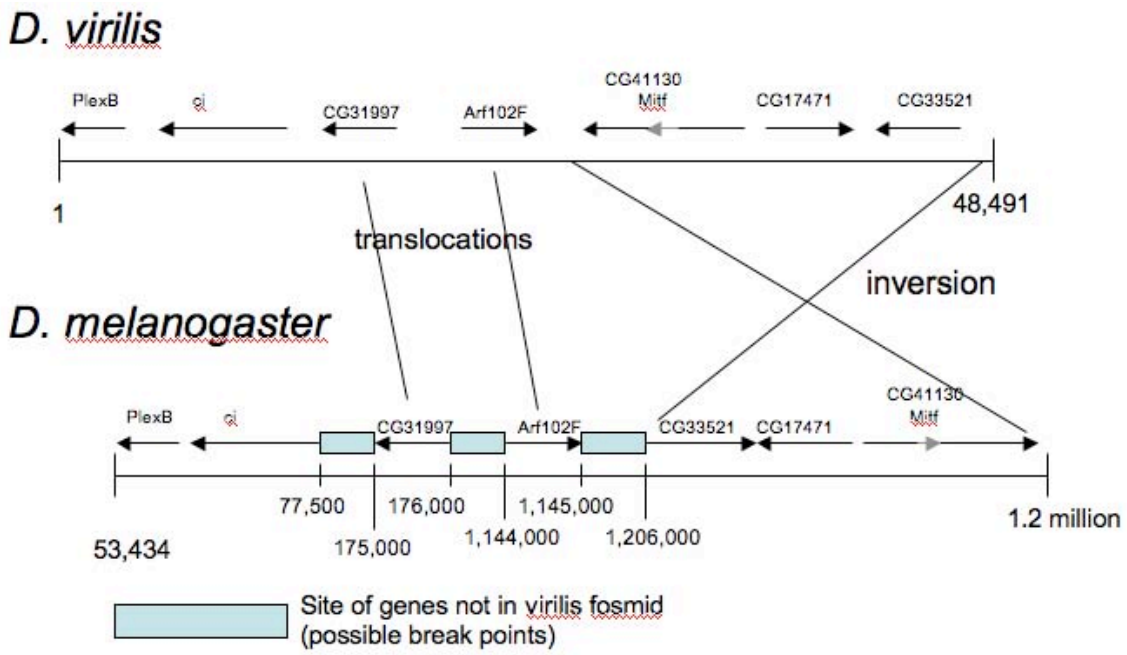


Figure 21: Synteny overview

V. Conclusion

After completing the annotation of fosmid 1, the results were compared to the features predicted by GENSCAN. GENSCAN had predicted 9 features, 8 of which corresponded to the annotated genes described above. To investigate the feature that did not correspond to an annotated gene, the DNA in the region it was located was extracted. Using BlastX, this sequence was compared to the annotated *D. melanogaster* proteins. One alignment was produced, but it was of very low quality and had no RefSeq data (see figure below). Therefore, it was determined that this feature was a mispredicted gene, and fosmid 1 has only 8 genes.

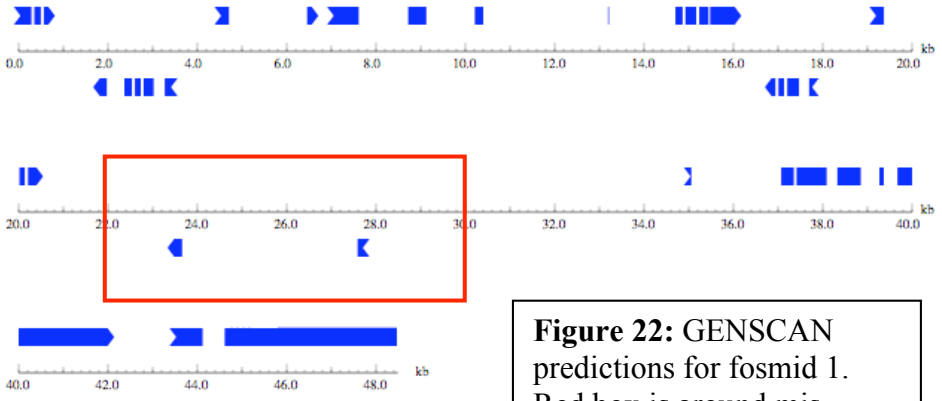


Figure 22: GENSCAN predictions for fosmid 1. Red box is around mis-predicted gene.

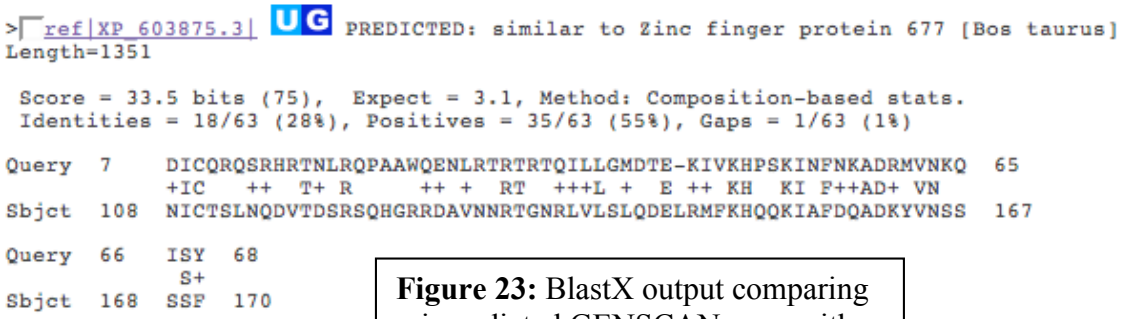


Figure 23: BlastX output comparing mispredicted GENSCAN gene with *D. melanogaster* annotated proteins.