Introduction

With rapid advances in sequencing technology, particularly with the development of second and third generation sequencing, genomes for organisms from all kingdoms and many phyla have been sequenced. The choke point for novel biological information, therefore, is not sequence data but rather the computational power to process the data as well as proper annotation of the data.

There are various degrees of annotation, much of which is assessed using computer algorithms. For example, simple gene prediction algorithms like GENSCAN have been developed for gene predictions while programs like RepeatMasker have been useful in identifying various repetitious elements. However, several problems emerge when relying on prediction algorithms. For example, gene prediction tools tend to fail in identifying short exons, particularly at the 5' and 3' ends of genes, as well as in identifying non-canonical splice sites. Often gene prediction algorithms are discrepant with one another, and many completely fail to take into consideration evidence from homology. Furthermore, computer algorithms can miss important features such as 5' and 3' untranslated regions (UTR) of genes, containing regulatory elements. Detailed annotation of a gene involves synthesizing information from homology, gene prediction programs, and knowledge of various features in the genome.

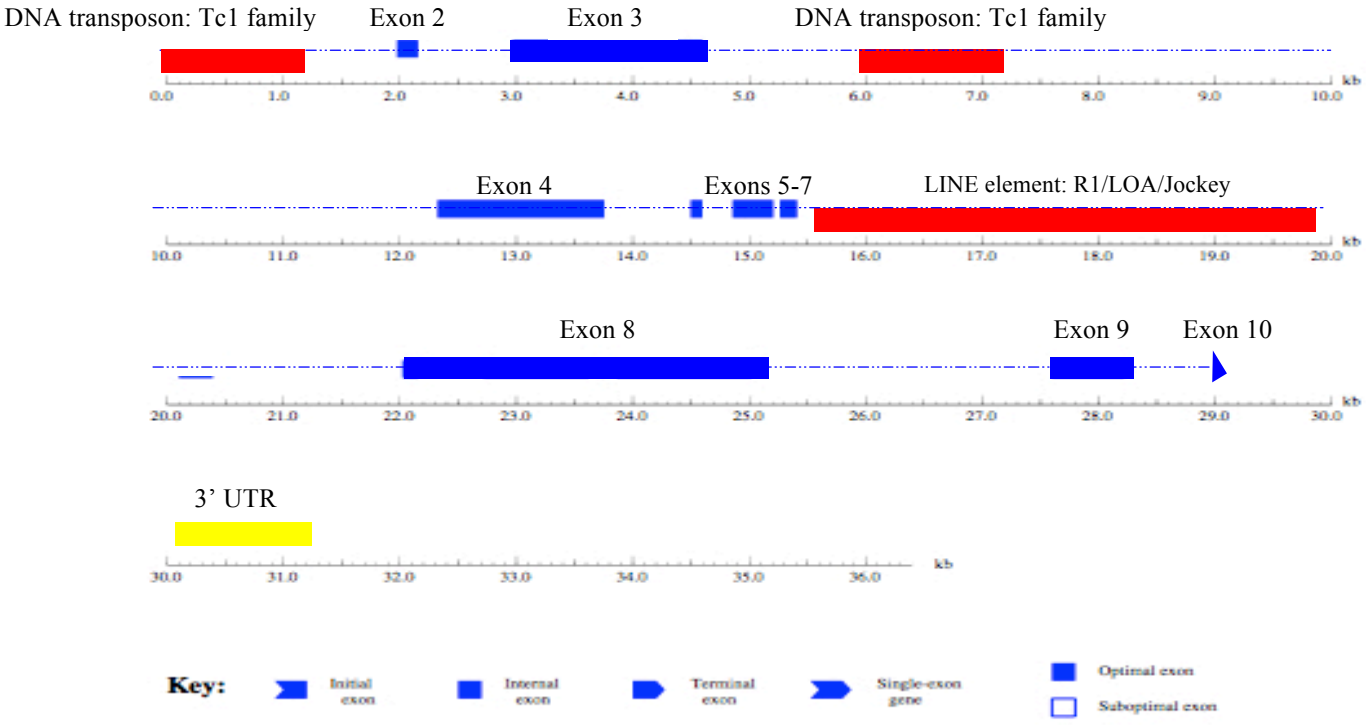
The goal in "Biology 4342: Research Exploration in Genomics" has been the finishing and annotating of the dot chromosome in a number of *Drosophila* species. The dot chromosome is characterized by large regions of heterochromatic packaging. Heterochromatin formation has been suggested as an ancient mechanism in eukaryotes to combat the viruses and transposable elements that insert into the genome. Heterochromatin formation has been associated with a high number of repeat sequences, particularly multiple copies of transposable elements. It is also known to silence nearby genes. What makes the dot chromosome so interesting is that there is evidence for significant gene transcription even though the majority of the chromosome contains heterochromatic packaging. Sequencing and annotating fosmids from the dot chromosome will reveal the presence of genes that may serve as targets of future studies of gene transcription in regions of heterochromatic packaging.

During the Spring 2009 semester, we focused on finishing and annotating a portion of the dot chromosome in *Drosophila grimshawi*, the Hawaiian fruitfly. In annotating a species such as *D. grimshawi*, it is important to use evidence from homology with a corresponding experimentally validated gene in the well-annotated *D. melanogaster* genome.

Overview

This project involved finishing and annotating Contig 9, a 37 kb region located at approximately positions 940 kb to 980 kb on the dot chromosome of *D. grimshawi*. Annotation revealed that this region consists of only a single gene, the ortholog of *Zfh2* in *D. melanogaster*. *Zfh2* is a ten exon gene in *D. melanogaster*, and there is only one isoform currently identified. Exons 2 to 10 seem to be located in Contig 9, and all of these exons show significant conservation in length, splice site, and amino acid sequence. Contig 9 also contains the 3' UTR for both *Zfh2* and *Thd1*, a gene transcribed in the reverse direction just downstream of Contig 9. Finally, Contig 9 also contains one particularly interesting repeat feature, a 5 kb LINE element located between exons 7 and 8 of the ortholog of *Zfh2*.

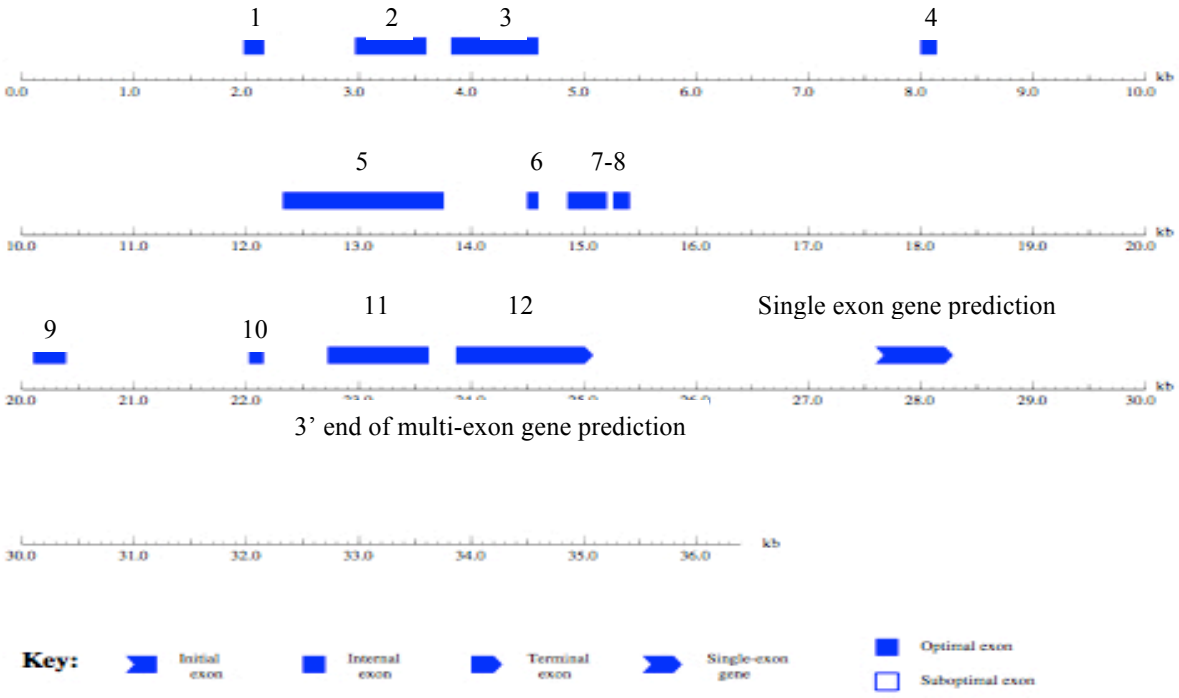
Figure 1: Final annotation map of Contig 9



Overview

The initial GENSCAN prediction in Figure 2 shows two predicted features, a 12 exon gene prediction missing the 5' end and one single exon gene prediction.

Figure 2: GENSCAN predicted genes in Contig 9 showing two predicted genes, one single exon gene and another multi-exon gene missing the 3' end.



Annotation of *Zfh2*

The first step to the annotation process was to identify a potential ortholog of Feature 1 in *D. melanogaster*. A BLASTx search of the nucleotide sequence of the first exon in Feature 1 (positions 1900 to 2200) to the *D. melanogaster* proteome reveals a match with 97% identity to Exon 2 in *Zfh2* (Figure 3).

Figure 3: BLASTx alignment of first exon in Feature 1 (Query) to Zfh2 protein (Subject)

>gnl|dmel|FBpp0088139 type=protein; loc=4:join(542130..543730, 544957..545138, 545530..547146, 550319..551719, 551782..552005, 552062..552409, 553409..553552, 553609..556506, 557801..558386, 559164..559180); ID=FBpp0088139; name=zfh2-PA; parent=FBgn0004607, FBtr0089070; dbxref=FlyBase_Annotation_IDs:CG1449-PA, FlyBase:FBpp0088139, GB_protein:AAF59339.2, REFSEQ:NP_524623, GB_protein:AAF59339; MD5=301fa19e73459ee0625b835b8fcc86b9; length=3005; release=r5.16; species=Dmel;
Length = 3005

HSP # = 1 , **Score =** 137.502 bits (345) , **Expect =** 1.42645e-33
Identities = 61 / 63 (96.8%) , **Positives =** 62 / 63 (98.4%)
Frame = +2

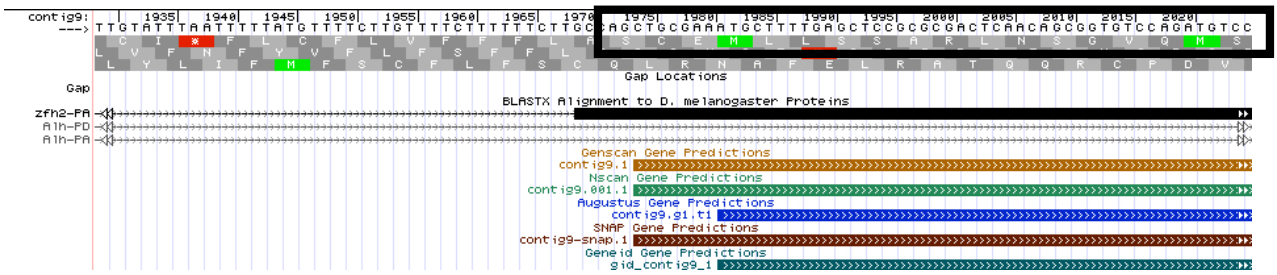
Subject FASTA

```
Query:   20         ASCEMLLSSARLNSGVQMSTRNSCKTLKCPQCNWHYKYQETLEIHMREKHPDGESACGYC   199
                    A CEMLL+SARLNSGVQMSTRNSCKTLKCPQCNWHYKYQETLEIHMREKHPDGESACGYC
Subject: 533        ARCEMLLNSARLNSGVQMSTRNSCKTLKCPQCNWHYKYQETLEIHMREKHPDGESACGYC   592

Query:   200        LAG   208
                    LAG
Subject: 593        LAG   595
```

Having identified Feature 1 as the ortholog of *Zfh2*, I proceeded to find the nucleotide and amino acid sequences of the experimentally validated protein on Gene Record Viewer. Gene Record Viewer indicates that there is only one known isoform of *Zfh2* in *D. melanogaster*. A comparison to the alignment shown in Figure 3 revealed that the alignment includes the entire Exon 2 from the arginine residue at position 534 to the alanine residue at position 594. The UCSC Genome Browser view shown in Figure 4 shows an AG splice acceptor site (at the start of the box in Figure 4) immediately before almost perfectly conserved sequence in Frame 1. The AG motif is found in 100% of splice acceptor sites. There is also no other AG motif in Frame 1 prior to the stop codon shown at positions 1936 to 1938. Finally, the AG splice acceptor site in phase 1 complements the phase 2 splice donor site on Exon 1, annotated by Noor Tazudeen, from the contig immediately upstream from Contig 9. This offers strong support that the ortholog of Exon 2 in *Zfh2* starts at position 1974.

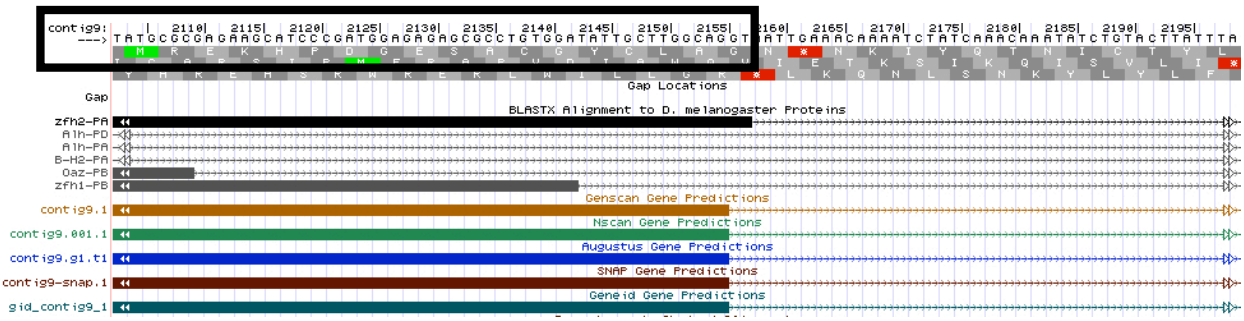
Figure 4: Start of the ortholog of Exon 2 in Zfh2 showing sequence in Frame 1 with AG splice acceptor site



The end of the ortholog of Exon 2 in *Zfh2* corresponds to a GT splice donor site located immediately after the last conserved amino acid to the second exon in *D. melanogaster*, an alanine (Figure 5). The GT motif is present in 95% of all splice donor sites. In this case, it is in Phase 1 of Frame 1 with a guanine base also conserved, and it is the only GT splice donor site located after the last highly conserved amino acid at position 2155 and before the stop codon at positions 2161 to 2163. In conclusion, the nucleotide sequence for Exon 2 can be identified as the sequence from position 1974 to 2155.

.

Figure 5: End of the ortholog of Exon 2 in Zfh2 showing the sequence in Frame 2 with the GT splice donor site



Ortholog of Exon 3

Having identified Feature 1 as the ortholog of *Zfh2* in *D. melanogaster*, the remaining exons were annotated using BLAST searches and ClustalW alignments of *Zfh2* amino acid sequences from Gene Record Viewer to nucleotide sequences downstream of predicted Exon 2. The following methodology was used to identify Exon 3: since Exon 2 ends at position 2155, a BLASTx search was performed on sequence in Contig 9 from position 2156 to the next exon of the GENSCAN prediction shown from positions 3000 to 3600 in Figure 2. This sequence aligned to Exon 3 of *Zfh2* in *D. melanogaster* (Figure 6). The translated nucleotide sequence for the second exon predicted by GENSCAN starts from residue 594 and ends at residue 824 of Exon 3 in *D. melanogaster* as indicated by the alignment in Figure 6. This is only a 230 amino acid sequence whereas the actual Exon 3 consists of 539 amino acids. This suggested that second and third exons as predicted by GENSCAN are actually a single exon, the ortholog of Exon 3 in *D. melanogaster*.

Figure 6: BLASTx search for the nucleotide sequence downstream of Exon 2 in Contig 9 (Query) to the protein database in D. melanogaster, showing alignment to a part of Exon 3 of Zfh2

>gnl|dmel|FBpp0088139 type=protein; loc=4:join(542130..543730, 544957..545138, 545530..547146, 550319..551719, 551782..552005, 552062..552409, 553409..553552, 553609..556506, 557801..558386, 559164..559180); ID=FBpp0088139; name=zfh2-PA; parent=FBgn0004607, FBtr0089070; dbxref=FlyBase_Annotation_IDs:CG1449-PA, FlyBase:FBpp0088139, GB_protein:AAF59339.2, REFSEQ:NP_524623, GB_protein:AAF59339; MD5=301fa19e73459ee0625b835b8fcc86b9; length=3005; release=r5.16; species=Dmel;
Length = 3005

HSP # = 1 , Score = 299.286 bits (765) , Expect = 3.81184e-81
Identities = 169 / 248 (68.1%) , Positives = 187 / 248 (75.4%) , Gaps = 4 / 248 (1.6%)
Frame = +1

Subject FASTA

```
Query:   808       AGQQHPRLARGESYSCGYKPYRCEICNYSTTTKGNLSIHMQSDKHLNNMQELNSSQNIVA   987
                   AGQQHPRLARGESYSCGYKPYRCEICNYSTTTKGNLSIHMQSDKHLNNMQELNSSQN+VA
Subject: 594       AGQQHPRLARGESYSCGYKPYRCEICNYSTTTKGNLSIHMQSDKHLNNMQELNSSQNMVA   653

Query:   988       VAAAA----KLMLPNPSPQGSVSGCSNAVVSANHQPVGGVCSSSAGSAGTVSNASNASNS   1155
                    AAAA    KL+L + SPQ + +  SN+   A     G   S+  G  GT S + NA+ S
Subject: 654       AAAAAAVTGKLLLSSSSPQVTAACPSNSGSGA-----GSGSSNIVG--GTASLSGNATPS   706

Query:   1156      STMGSGTSSGAGTGSSTGVSSLKPKPSFRCDICSYETSVARNLRIHMTSEKHTHNMAVLQ   1335
                    T     SS A  GS+T  +  KPKPSFRCDICSY+TSVARNLRIHMTSEKHTHNMAVLQ
Subject: 707       VT--GANSSNANAGSNTNNAGTKPKPSFRCDICSYDTSVARNLRIHMTSEKHTHNMAVLQ   764

Query:   1336      NNIKHIQAFNFLQQQQQQQSAAAVPGTASNSFLPVPEVALADLAYNQALMIQLIQHNAAN   1515
                   NNIKHIQAFNFL QQQQQ     +   +S SF+  PEVALADLAYNQALMIQL+     +
Subject: 765       NNIKHIQAFNFL-QQQQQSGTGNIASHSSGSFM--PEVALADLAYNQALMIQLL-----H   816

Query:   1516      QQQQQQQN   1539
                   QQQQ QQ+
Subject: 817       QQQQHQQS   824
```

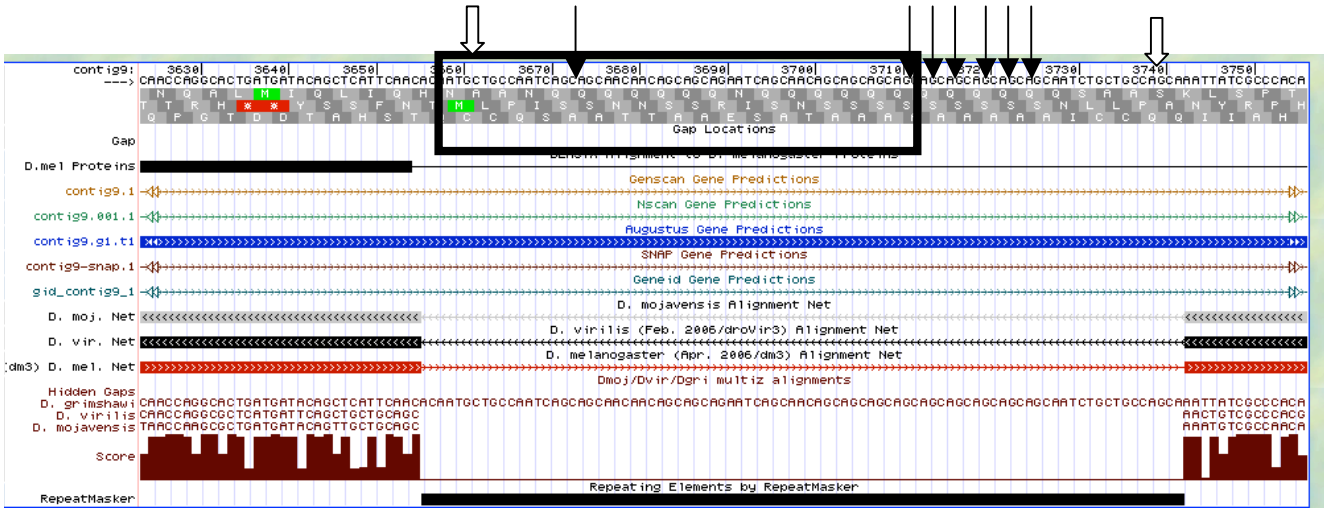
To determine whether the ortholog of Exon 3 is actually a combination of the second and third exons as predicted by GENSCAN, I conducted a ClustalW alignment between the amino acid sequence spanning the two exons predicted by GENSCAN to the amino acid sequence of the known Exon 3 in *D. melanogaster.* The results are shown in Figure 7 on the next page.

Figure 7: ClustalW alignment of the second and third exons of Contig 9 as predicted by GENSCAN to the amino acid sequence of Exon 3 in D. melanogaster. The arrows indicate the splice sites that preserve the most number of conserved bases while the potential intron is indicated by the box.

The ClustalW alignment shows an insertion in Exon 3 dominated by a poly-glutamine sequence. The insertion consists of 18 amino acids or approximately 50 nucleotides, raising the question whether there is a change in exon number. Although towards the shorter end of a possible intron length, the 50 nucleotide sequence cannot be completely ruled out as an intron.

A closer look at this region on the UCSC Genome Browser viewer reveals a simple repeat as indicated by Repeat Masker (Figure 8). In the viewer, the boxed region indicates the insertion beginning after the histidine residue and ending before the sixth glutamine residue in Frame 1 as indicated in Figure 7. A quick scan of the boxed region shows no GT splice donor sites. The closest upstream GT motif is located at position 3600 and marks the end of the second exon predicted by GENSCAN, NSCAN, and SNAP gene predictions. The closest complementary AG receptor site is found at 3740. The problem with these splice acceptor and donor sites is that they eliminate 23 perfectly conserved amino acid residues. Therefore, these are not likely to be actual splice acceptor and donor sites.

Figure 8: UCSC Genome Browser View showing the region between exons 2 and 3 predicted by GENSCAN in Frame 1. Notice the poly-Q insertion and arrows corresponding to potential splice donor-acceptor pairs; the two arrows on the left correspond to the GC splice donors while the seven arrows on the left correspond to the AG splice acceptors. The types of arrows correspond to splice pairs in the same phase.



Before ruling out the possibility of an intron, it is important to consider the presence of a non-canonical GC splice donor site, present in 5% of all splice sites. There are two possible GC splice donors as indicated by the arrows in Figure 8 and several potential receptor sites. The first pair (indicated by the block arrows) yields an approximately 80 base intron. However, this also eliminates ten highly conserved amino acid residues. For the second set, the only pair that preserves all conserved residues yields an approximately 40 base intron. This is unlikely because it is too small for a potential intron. Several AG acceptor sites that subsequently increase the size of the intron eliminate conserved residues outside the black box in Figure 8. While it is still possible that there is an intron from one of these pairs, it seems unlikely.

In conclusion, I decided to annotate a single exon spanning from base 2966 to 4627 in Contig 9. It is very unusual to find a change in exon number from one *Drosophila* species to the next. However, GC splice donor sites are only present in about 5% of all splice donor sites. Finally, the potential splice donor-acceptor pairs yield introns that are either too small or that eliminate highly conserved residues.

Ortholog of Exon 4

A ClustalW alignment of Exon 4 in *D. melanogaster* to the region downstream of the identified ortholog of Exon 3 in Contig 9 revealed alignment with a high conservation of amino acids from base positions 12300 to 13800 (Figure 9).

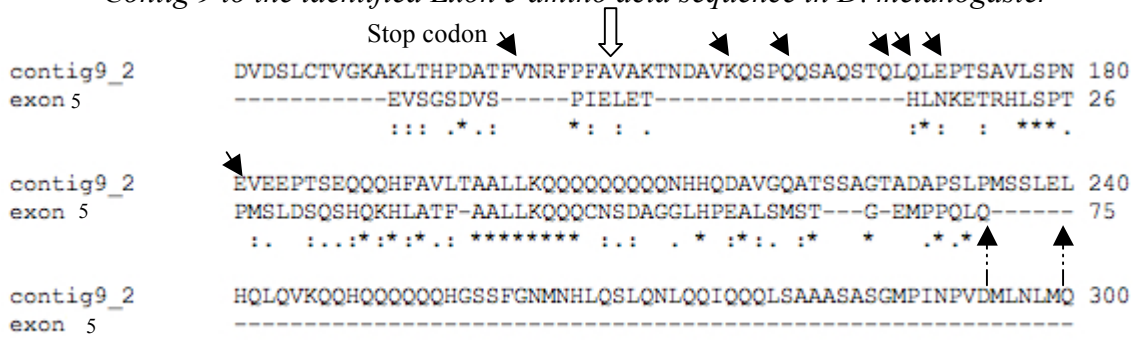
Figure 9: ClustalW alignment between amino acids from identified Zfh2 exon 4 in D. melanogaster to Contig 9, positions 12322 to 13755.

The start of the exon at position 12322 corresponds to the first conserved amino acid (aspartic acid) as well as an AG splice acceptor site of phase 2 complimentary to the GT splice donor site of phase 1 from the previous exon. More interesting in this exon was identifying the end of the ortholog of exon 4. The GENSCAN prediction shown as "contig9_2" in Figure 9 shows an extension of the exon after the last conserved amino acid. There were two possible splice donor sites, a GT site and a GC site. Since the *D. melanogaster* exon terminates with a GT site, it is unlikely that the GC site is the appropriate splice donor site in *D. grimshawi*. Furthermore, an AG splice acceptor site was found in phase 2 that conserved the most number of amino acid residues. This is indicated by the block arrow shown in Figure 10 on the next page. From this evidence, the ortholog of exon 4 was annotated as a region on Contig 9 spanning from base 12322 to 13755.

Ortholog of Exon 5

Alignment to the identified exon 5 in *D. melanogaster* revealed relatively poor conservation (Figure 10). A ClustalW alignment was much more useful than the alignment from a BLAST search since the BLAST search eliminated many residues at the start and end of the exon.

Figure 10: ClustalW alignment of the amino acid sequence downstream of Exon 4 in Frame 3 of Contig 9 to the identified Exon 5 amino acid sequence in D. melanogaster



With such low percent identity, it was difficult to determine a reliable splice acceptor site. Interestingly, a stop codon indicated in Figure 10 shows that eight amino acids from the Exon 5 in *D. melanogaster* must be absent in the ortholog in *D. grimshawi*. All AG sites in phase 2 have been indicated by block arrows while all other potential AG sites have been indicated by line arrows in Figure 10. The identified donor site in phase 1 from the annotation of the previous exon suggested that the AG site indicated by the block arrow was the best candidate for a splice acceptor site.

The only AG splice acceptor site in phase 2 to compliment the GT donor in phase 1 is also the splice site that conserves the largest number of amino acids. Furthermore, the TT nucleotides from the previous exon when added to the guanine give the amino acid, valine (block arrow in Figure 10). This residue maintains the same properties as the aligning isoleucine residue in the identified exon 5 of *D. melanogaster.* Using these lines of evidence, the start site for the exon was annotated at position 14496.

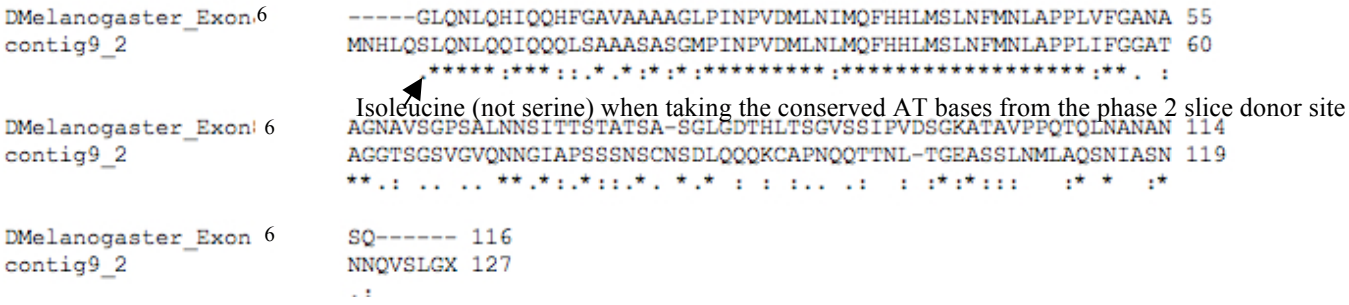
The low conservation at the 3' end also made identifying a GT donor site difficult. There are two potential GT sites after the last conserved base, both of which are in phase 2. These have been marked by two dashed arrows at the 3' end of Exon 5 in Figure 10. The first yields an AT pair of nucleotides while the second yields a GT pair of nucleotides (left over from the last complete amino acid). All other splice sites downstream are too close to the start of the next exon. The strongest evidence to identify the correct splice site was from the high conservation of the next exon. The two bases AT when combined with the thymine base yields the amino acid, isoleucine, which would be a perfectly conserved amino acid. This indicated that the first splice site at position 14758 (the first dashed arrow in Figure 10) was most likely the correct splice donor site. Furthermore, this site also does not add any extra amino acid residues as indicated by the arrow pointing to the last amino acid (glutamine) in Figure 10.

Based on the two splice sites, the ortholog of exon 5 can be annotated as the region on Contig 9 from base 14497 to 14758.

Ortholog of Exon 6

The same methods were used to identify exon 6. A ClustalW alignment reveals a much higher degree of conservation, particularly at the start of the exon (Figure 11).

Figure 11: ClustalW alignment of the amino acid sequence downstream of Exon 5 in Frame 3 of Contig 9 to the identified Exon 6 amino acid sequence in D. melanogaster



Since both potential splice donor sites from exon 5 are in phase 2, the splice acceptor site from exon 6 must be in phase 1. Conveniently, this is located at the start of the first conserved base to give an isoleucine with properties similar to the glycine in exon 6 in *D. melanogaster.* This AG splice site is located at position 14858.

While there is less conservation at the end of the ortholog of exon 6 than at the beginning, there is only one possible splice after the last conserved residue (glutamine) and before the stop codon in Frame 3. This is located at position 15206.

Based on the splice donor and acceptor sites, the most likely location of the ortholog of exon 6 in Contig 9 is from positions 14858 to 15206.

Ortholog of Exon 7

The same methods were used to identify exon 7. A ClustalW alignment reveals almost 100% identity between the ortholog of exon 7 in Contig 9 and exon 7 in *D. melanogaster* (Figure 12).

Figure 12: ClustalW alignment of the amino acid sequence downstream of Exon 6 in Frame 3 of Contig 9 to the identified Exon 7 amino acid sequence in D. melanogaster

```
                          Position 15264                                  Position 15410
Contig9                   QQLTSNQKRARTRITDDQLKILRAHFDINNSPSEESIMEMSQKANLPMK 49
Dmelanogasterexon7        -QLASNQKRARTRITDDQLKILRAHFDINNSPSEESIMEMSQKANLPMK 48
                           **:**********************************************
```

There is only one potential splice site before the conserved residues and after the stop codon at the 5' end of the exon. The AG acceptor site is located at position 15264. This is also in the appropriate phase (phase 0), complimenting the phase 0 donor site from the previous exon.

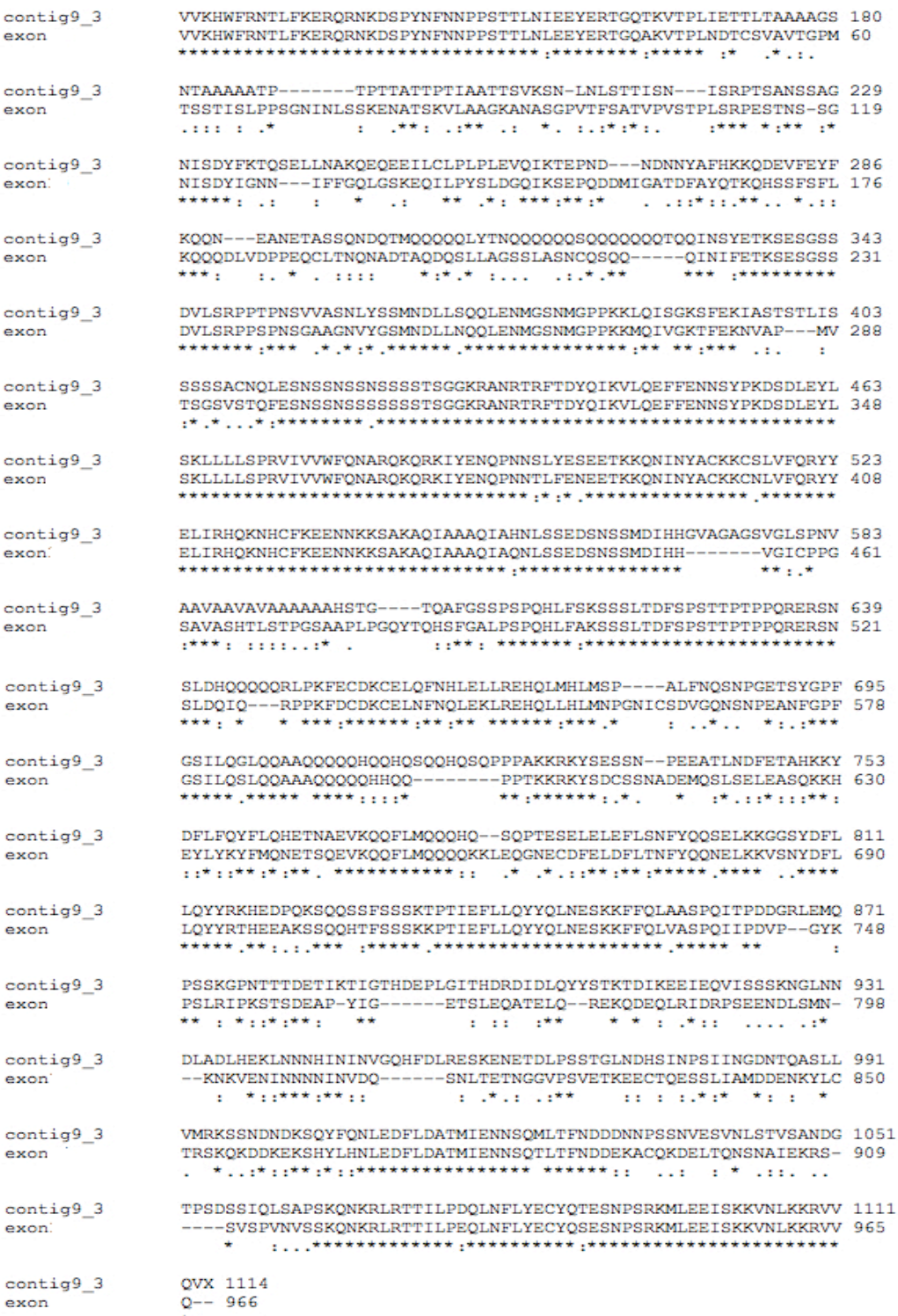
There is also only one potential splice site after the last conserved residue and before the stop codon at the 3' end of the exon. The GT splice donor site is located at position 15410 and is in phase 0.

From this annotation, the position of the ortholog of exon 7 has been identified as positions 15264 to 15410 on Contig 9.

Ortholog of Exon 8

The same methods were used to identify exon 8. Whereas all the gene prediction algorithms indicated two to four different exons, a ClustalW alignment reveals a single exon with high conservation along the whole length of a single open reading frame in Frame 3 (Figure 13).

Figure 13: ClustalW alignment of the amino acid sequence downstream of Exon 7 in Frame 3 of Contig 9, to the identified Exon 8 amino acid sequence in D. melanogaster



There is only one splice acceptor site before the first conserved amino acid and after the stop codon at the 5' end. The AG site is in phase 0 and is complimentary to the GT donor site that is also in phase 0. This indicates that the start of the exon is most likely at position 22020.

At this point, the GENSCAN algorithm includes a stop codon, suggesting that this exon is the last exon of Feature 1. However, Gene Record Viewer indicates that the actual gene in *D. melanogaster* consists of 10 exons. A reduction of exon number is rare in *Drosophila* species; it is more likely that GENSCAN incorrectly predicted the end of a feature. A search for the ortholog of exon 9 confirmed this.

There are two potential GT splice donor sites at the 3' end of the ortholog of exon 8. The best splice site occurs immediately after the last conserved base and adds no additional residues. This is in phase 0 and compliments the best splice acceptor site at the 5' end of exon 9. The other splice donor site adds one residue and is in phase 1.

From this analysis, the ortholog of exon 8 in Contig 9 was identified as the region from positions 22020 to 24995.

Orthologs of Exon 9 and 10

While GENSCAN predicted that Features 1 and 2 were actually separate genes, Gene Record Viewer indicates that *Zfh2* in *D. melanogaster* actually contains two additional exons. The last exon, Exon 10, is a four residue exon. A search for Exon 10 is difficult since it contains only four residues. Therefore, I performed a ClustalW alignment between the amino acid residues of Feature 2 to the amino acid sequence of both exons 9 and 10 in *D. melanogaster* (Figure 15).

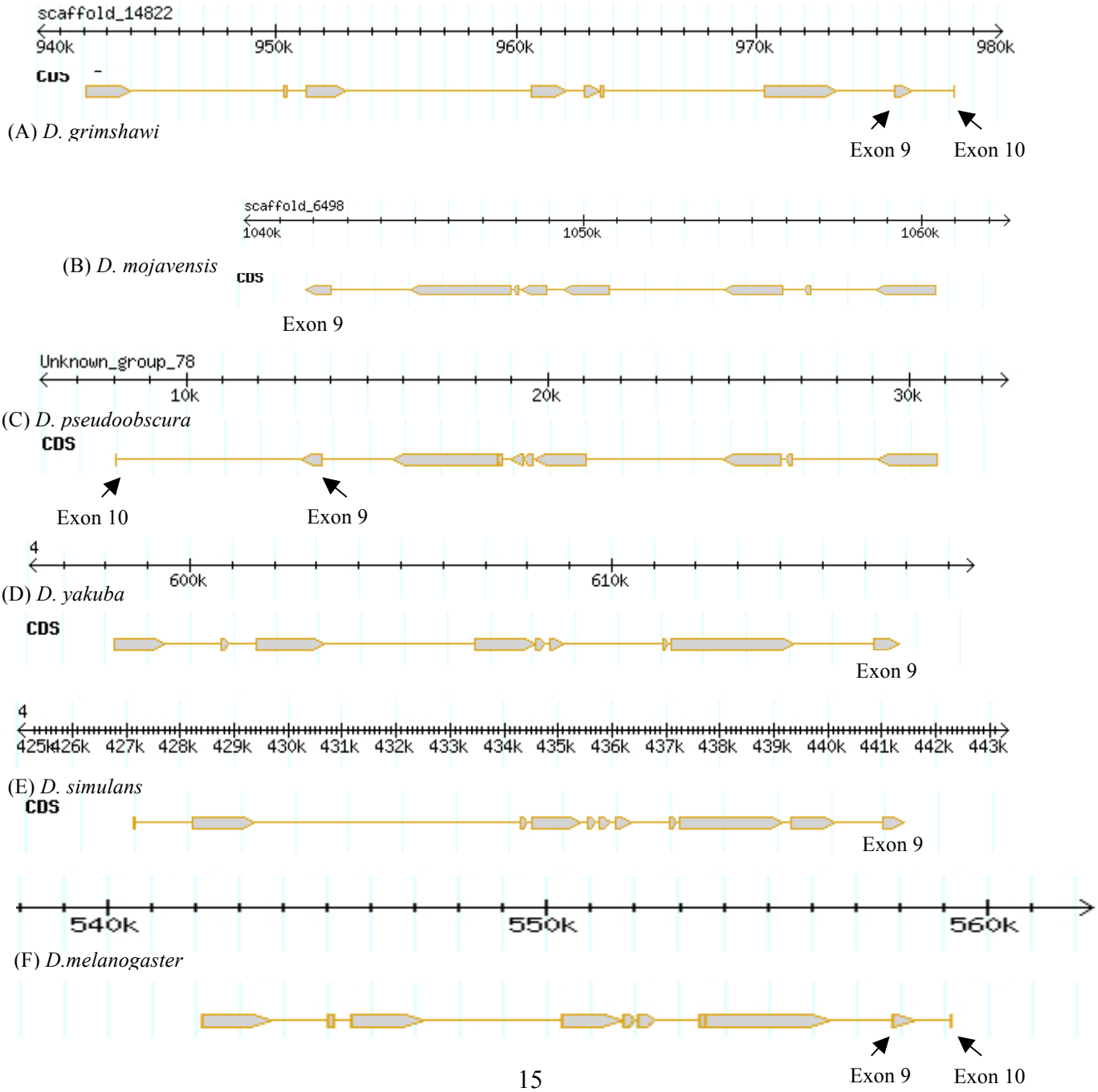
Figure 15: ClustalW alignment of the amino acid sequence from Feature 2 of the GENSCAN prediction in Frame 3 of Contig 9 to the identified Exon 9 and 10 amino acid sequences in D. melanogaster (exon 10 has been boxed)

The UCSC Genome Browser view of Contig 9 shows that none of the gene prediction algorithms include the ortholog of exon 10. This could mean that the exon (1) is no longer present in *D. grimshawi*, (2) has been incorporated into exon 9, or (3) is simply not detected by the gene prediction programs.

The second possibility is interesting because the ClustalW alignment in Figure 15 shows that a region within the GENSCAN and Augustus gene predictions aligns well to exon 10 just downstream of the highly conserved 3' end of exon 9. However, this evidence is insufficient to identify a reduction in exon number. Even though *D. grimshawi* is a distant relative of *D. melanogaster*, a change in exon number is still very rare.

Another source of information involves looking at synteny between different *Drosophila* species. To do so, I proceeded to conduct a BLASTx search on the FlyBase browser for the alignment of Contig 9 to *D. melanogaster, D. mojavensis, D. pseudoobscura, D. yakuba,* and *D. simulans*. It is important to recognize that synteny to any fruitfly species other than *D. melanogaster* is extremely weak evidence, particularly since annotation of the other species relies primarily on gene prediction algorithms. With this in mind, it still might give a general sense regarding the putative exon number and the presence of the ortholog of exon 10. The BLASTx search on Flybase provided a Mapviewer function to show the position of *Zfh2* in all species (Figure 16).

Figure 16: Map of predicted ortholog of Zfh2 in D. grimshawi (A), D. mojavensis (B), D. pseudoobscura (C), D. yakuba (D), D. simulans (E), and D. melanogaster (F) all arranged according to increasing evolutionary distance from D. grimshawi



The gene prediction algorithms suggest that the ortholog of exon 10 is missing from most of the *Drosophila* species. This might simply be a result of the small size of exon 10. However, exon 10 is included in the gene prediction for *D. grimshawi*. A ClustalW alignment of the nucleotide sequence of the known exon 10 in *D. melanogaster* to the nucleotide sequence downstream of exon 9 in Contig 9 as well as in two other *Drosophila* species, *D. yakuba* and *D. simulans*, reveals high conservation (Figure 17). This provides evidence that the gene prediction algorithms likely missed exon 10 in all *Drosophila* species due to its short length.

Figure 17: ClustalW alignment of known exon 10 in D. melanogaster to nucleotide sequence downstream of exon 9 in D. yakuba, D. simulans, and Contig 9

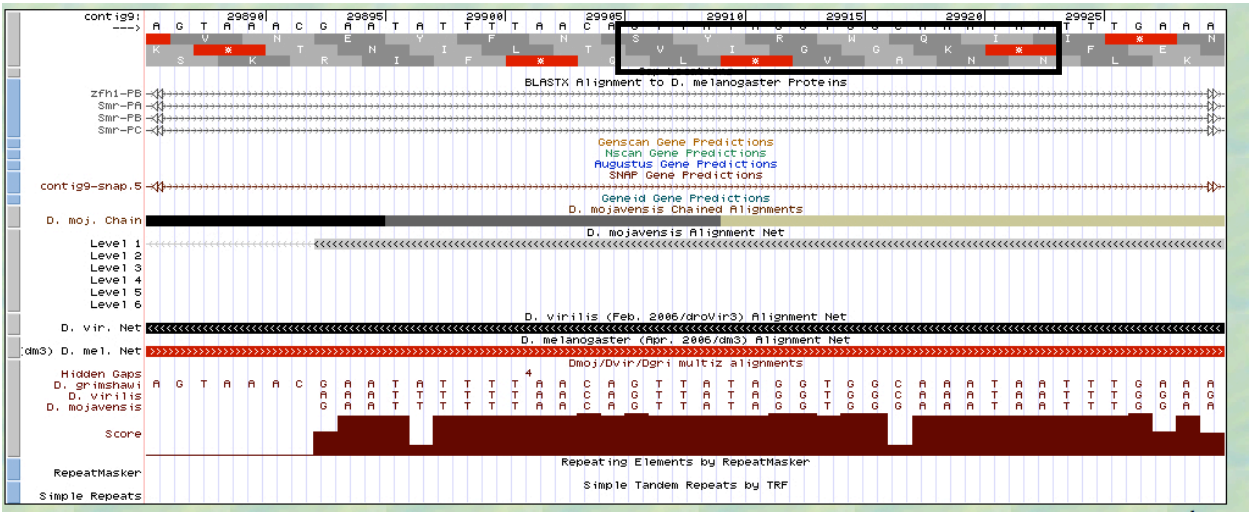
TAA stop codon

```
Dmelanogasterexon10     ------------------------------------TTATAGGTGGAAAGTAATCT 21
Dyakubaexon10           ------------------------------------TTATAGGTGGAAAGTAATCT 21
Dsimulansexon10         ------------------------------------TTATAGGTGGAAAGTAATCT 21
contig9                 TATACCTAAACTGCAATAAGTAAACGAATATTTTAACATTATAGGTGGCAAATAATTTG 1800
                                                           ********** ** *** *

Dmelanogasterexon10     GAGTGTTTATATGGCCTTATAACCGTGGATAG---TGGTTAAAATTTAG---------- 67
Dyakubaexon10           GAGTGTTTATATGGCCTTCTAACCGTGGATAC---TGGGTAAAAT------------- 63
Dsimulansexon10         GAGTGTTTATATGGCCTTATAACCGTGGATAG---TGGTTAAAAT------------- 63
contig9                 AAATAATTTG--GGCGCGATAATGCTGGATATCAGTAGGTAAGTTTCTCTATAATATATA 1858
                         * *  **    ***    ***   ******    * * ***  *
```

This sequence coding for VIGGK was identified on the UCSC Genome Browser window of Contig 9 at positions 29907 to 29923 with the TAA stop codon at positions 29921 to 29923 (Figure 18).

Figure 18: UCSC Genome Browser view of exon 10 in Contig 9 in Frame 2



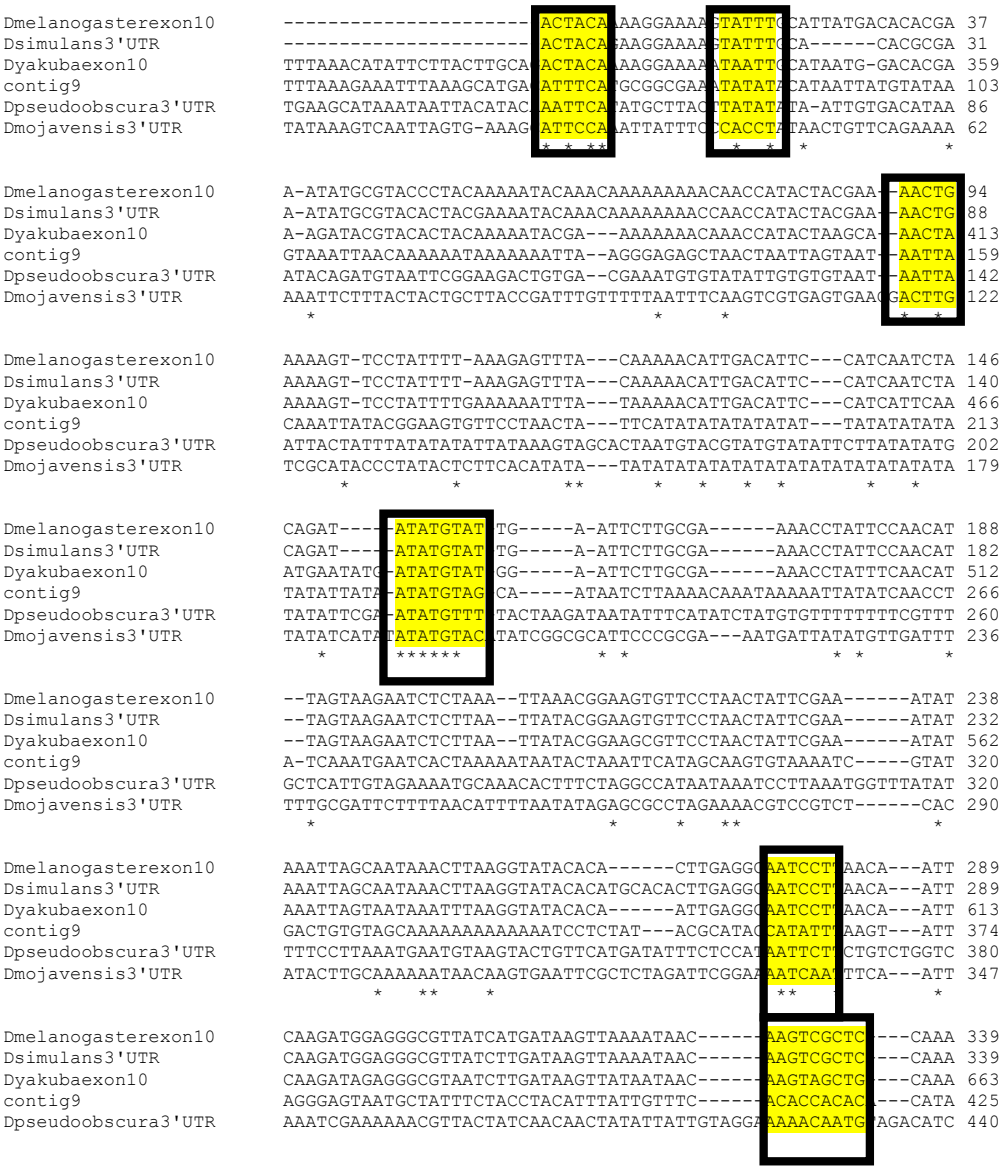
Finally, the only possible GT splice donor site at the 3' end of exon 9 is in phase 1, which is complimentary to the AG acceptor site at the 5' end of exon 10 in phase 2. This is in accordance with the annotation that exon 10 is present in Contig 9 and is important since there are no other possible splice donor-acceptor pairs.

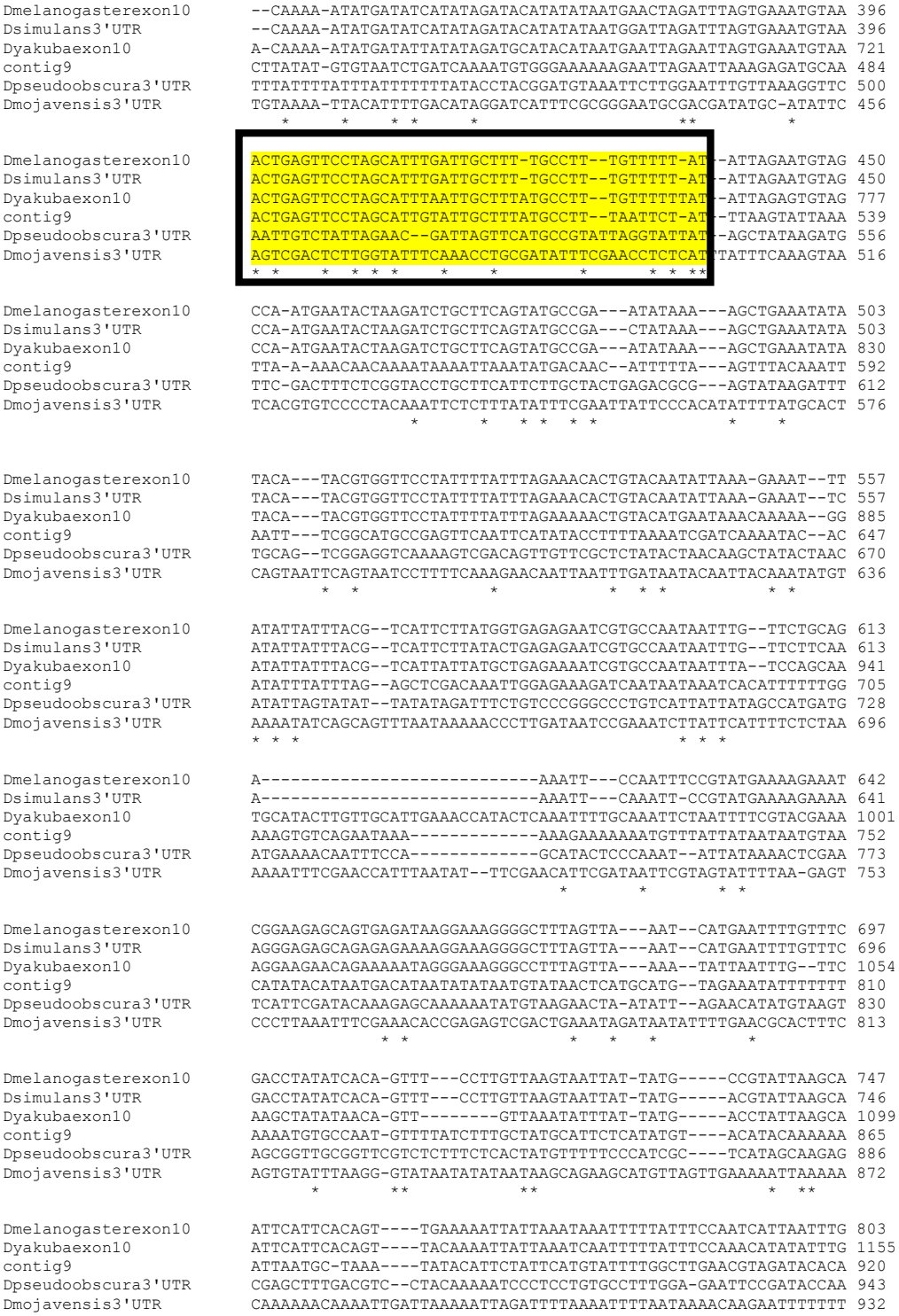
3’ UTR for *Zfh2*

The 3’ UTR for *Zfh2* has been identified in *D. melanogaster*. This region has been particularly interesting for its regulatory functions on the dot chromosome in the *Drosophila* species. A ClustalW alignment was performed on five of the *Drosophila* species to determine the location and conservation of potential 3’ UTR sequence (Figure 19). The location was identified as positions 30193 to 31050.

There seem to a few conserved motifs indicated by boxes in Figure 19. However, the 3’ UTR is poorly conserved across the *Drosophila* species. Interestingly, conservation seems to decrease with increasing distance from the 3’ end of the transcribed Exon 10.

Figure 19: ClustalW alignment of identified 3’ UTR in D. melanogaster from Gene Record Viewer to D. simulans, D. yakuba, D. grimshawi (Contig 9), and D. mojavensis. Regions of high conservation are highlighted.

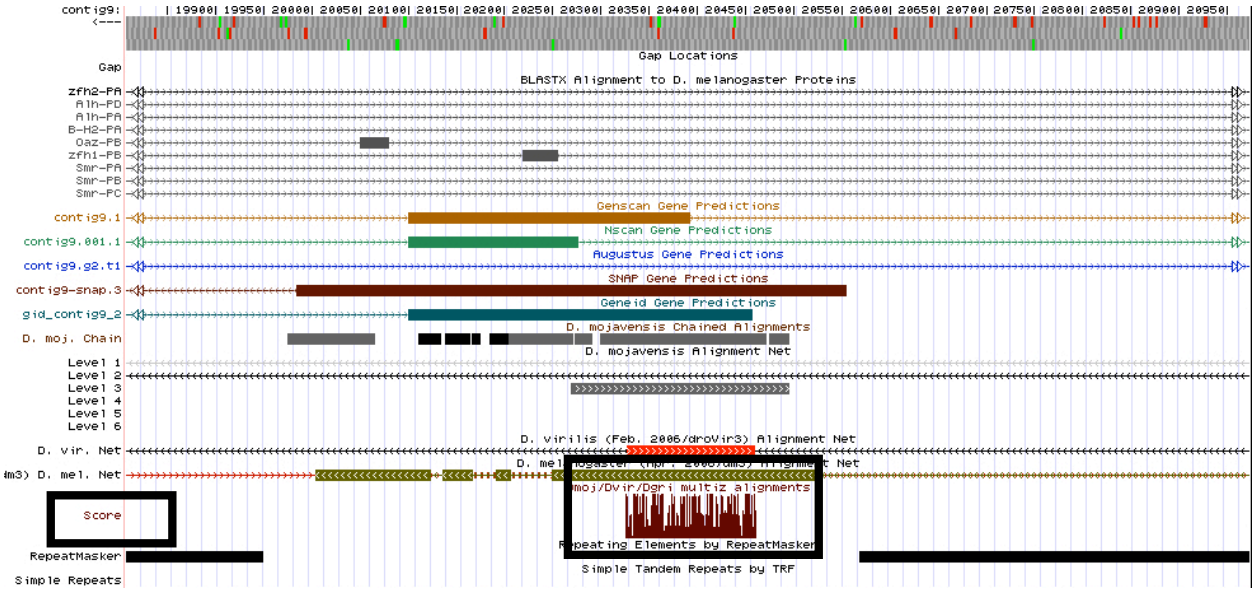



Dmelanogasterexon10 --CAAAA-ATATGATATCATATAGATACATATATAATGAACTAGATTTAGTGAAATGTAA 396
Dsimulans3'UTR --CAAAA-ATATGATATCATATAGATACATATATAATGGATTAGATTTAGTGAAATGTAA 396
Dyakubaexon10 A-CAAAA-ATATGATATTATATAGATGCATACATAATGAATTAGAATTAGTGAAATGTAA 721
contig9 CTTATAT-GTGTAATCTGATCAAAATGTGGGAAAAAAGAATTAGAATTAAAGAGATGCAA 484
Dpseudoobscura3'UTR TTTATTTTATTTATTTTTTTATACCTACGGATGTAAATTCTTGGAATTTGTTAAAGGTTC 500
Dmojavensis3'UTR TGTAAAA-TTACATTTTGACATAGGATCATTTCGCGGGAATGCGACGATATGC-ATATTC 456
Dmelanogasterexon10 ACTGAGTTCCTAGCATTTGATTGCTTT-TGCCTT--TGTTTTT-AT--ATTAGAATGTAG 450
Dsimulans3'UTR ACTGAGTTCCTAGCATTTGATTGCTTT-TGCCTT--TGTTTTT-AT--ATTAGAATGTAG 450
Dyakubaexon10 ACTGAGTTCCTAGCATTTAATTGCTTTATGCCTT--TGTTTTTTAT--ATTAGAGTGTAG 777
contig9 ACTGAGTTCCTAGCATTGTATTGCTTTATGCCTT--TAATTCT-AT--TTAAGTATTAAA 539
Dpseudoobscura3'UTR AATTGTCTATTAGAAC--GATTAGTTCATGCCGTATTAGGTATTAT--AGCTATAAGATG 556
Dmojavensis3'UTR AGTCGACTCTTGGTATTTCAAACCTGCGATATTTCGAACCTCTCATTTATTTCAAAGTAA 516
Dmelanogasterexon10 CCA-ATGAATACTAAGATCTGCTTCAGTATGCCGA---ATATAAA---AGCTGAAATATA 503
Dsimulans3'UTR CCA-ATGAATACTAAGATCTGCTTCAGTATGCCGA---CTATAAA---AGCTGAAATATA 503
Dyakubaexon10 CCA-ATGAATACTAAGATCTGCTTCAGTATGCCGA---ATATAAA---AGCTGAAATATA 830
contig9 TTA-A-AAACAACAAAATAAAATTAAATATGACAAC--ATTTTTA---AGTTTACAAATT 592
Dpseudoobscura3'UTR TTC-GACTTTCTCGGTACCTGCTTCATTCTTGCTACTGAGACGCG---AGTATAAGATTT 612
Dmojavensis3'UTR TCACGTGTCCCCTACAAATTCTCTTTATATTTCGAATTATTCCCACATATTTTATGCACT 576
Dmelanogasterexon10 TACA---TACGTGGTTCCTATTTTATTTAGAAACACTGTACAATATTAAA-GAAAT--TT 557
Dsimulans3'UTR TACA---TACGTGGTTCCTATTTTATTTAGAAACACTGTACAATATTAAA-GAAAT--TC 557
Dyakubaexon10 TACA---TACGTGGTTCCTATTTTATTTAGAAAAACTGTACATGAATAAACAAAAA--GG 885
contig9 AATT---TCGGCATGCCGAGTTCAATTCATATACCTTTTAAAATCGATCAAAATAC--AC 647
Dpseudoobscura3'UTR TGCAG--TCGGAGGTCAAAAGTCGACAGTTGTTCGCTCTATACTAACAAGCTATACTAAC 670
Dmojavensis3'UTR CAGTAATTCAGTAATCCTTTTCAAAGAACAATTAATTTGATAATACAATTACAAATATGT 636
Dmelanogasterexon10 ATATTATTTACG--TCATTCTTATGGTGAGAGAATCGTGCCAATAATTTG--TTCTGCAG 613
Dsimulans3'UTR ATATTATTTACG--TCATTCTTATACTGAGAGAATCGTGCCAATAATTTG--TTCTTCAA 613
Dyakubaexon10 ATATTATTTACG--TCATTATTATGCTGAGAAAATCGTGCCAATAATTTA--TCCAGCAA 941
contig9 ATATTTATTTAG--AGCTCGACAAATTGGAGAAAGATCAATAATAAATCACATTTTTTGG 705
Dpseudoobscura3'UTR ATATTAGTATAT--TATATAGATTTCTGTCCCGGGCCCTGTCATTATTATAGCCATGATG 728
Dmojavensis3'UTR AAAATATCAGCAGTTTAATAAAAACCCTTGATAATCCGAAATCTTATTCATTTTCTCTAA 696
Dmelanogasterexon10 A----------AAATT---CCAATTTCCGTATGAAAAGAAAT 642
Dsimulans3'UTR A----------AAATT---CAAATT-CCGTATGAAAAGAAAA 641
Dyakubaexon10 TGCATACTTGTTGCATTGAAACCATACTCAAATTTTGCAAATTCTAATTTTCGTACGAAA 1001
contig9 AAAGTGTCAGAATAAA----------AAAGAAAAAAATGTTTATTATAATAATGTAA 752
Dpseudoobscura3'UTR ATGAAAACAATTTCCA----------GCATACTCCCAAAT--ATTATAAAACTCGAA 773
Dmojavensis3'UTR AAAATTTCGAACCATTTAATAT--TTCGAACATTCGATAATTCGTAGTATTTTAA-GAGT 753
Dmelanogasterexon10 CGGAAGAGCAGTGAGATAAGGAAAGGGGCTTTAGTTA---AAT--CATGAATTTTGTTTC 697
Dsimulans3'UTR AGGGAGAGCAGAGAGAAAAGGAAAGGGGCTTTAGTTA---AAT--CATGAATTTTGTTTC 696
Dyakubaexon10 AGGAAGAACAGAAAAATAGGGAAAGGGCCTTTAGTTA---AAA--TATTAATTTG--TTC 1054
contig9 CATATACATAATGACATAATATATAATGTATAACTCATGCATG--TAGAAATATTTTTTT 810
Dpseudoobscura3'UTR TCATTCGATACAAAGAGCAAAAAATATGTAAGAACTA-ATATT--AGAACATATGTAAGT 830
Dmojavensis3'UTR CCCTTAAATTTCGAAACACCGAGAGTCGACTGAAATAGATAATATTTTGAACGCACTTTC 813
Dmelanogasterexon10 GACCTATATCACA-GTTT---CCTTGTTAAGTAATTAT-TATG-----CCGTATTAAGCA 747
Dsimulans3'UTR GACCTATATCACA-GTTT---CCTTGTTAAGTAATTAT-TATG-----ACGTATTAAGCA 746
Dyakubaexon10 AAGCTATATAACA-GTT--------GTTAAATATTTAT-TATG-----ACCTATTAAGCA 1099
contig9 AAAATGTGCCAAT-GTTTTATCTTTGCTATGCATTCTCATATGT----ACATACAAAAAA 865
Dpseudoobscura3'UTR AGCGGTTGCGGTTCGTCTCTTTCTCACTATGTTTTTCCCATCGC----TCATAGCAAGAG 886
Dmojavensis3'UTR AGTGTATTTAAGG-GTATAATATATAATAAGCAGAAGCATGTTAGTTGAAAAATTAAAAA 872
Dmelanogasterexon10 ATTCATTCACAGT----TGAAAAATTATTAAATAAATTTTTATTTCCAATCATTAATTTG 803
Dyakubaexon10 ATTCATTCACAGT----TACAAAATTATTAAATCAATTTTTATTTCCAAACATATATTTG 1155
contig9 ATTAATGC-TAAA----TATACATTCTATTCATGTATTTTGGCTTGAACGTAGATACACA 920
Dpseudoobscura3'UTR CGAGCTTTGACGTC--CTACAAAAATCCCTCCTGTGCCTTTGGA-GAATTCCGATACCAA 943
Dmojavensis3'UTR CAAAAAACAAAATTGATTAAAAATTAGATTTTAAAATTTTAATAAAACAAGAATTTTTTT 932

Other Features

GENSCAN included one exon in the *Zfh2* gene prediction that was not part of the ortholog of *Zfh2*. A closer look at this region reveals an open-reading frame several hundred bases in length, exon predictions from four different gene prediction tools, and relatively high conservation.

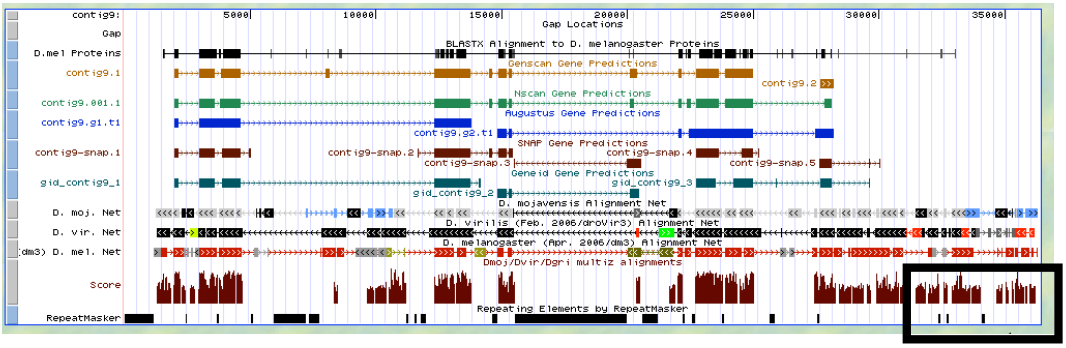
Figure 20: Unidentified feature in Contig 9 from bases 19900 to 20900. The high conservation is evident from the boxed score.



However, a BLASTx search for this region revealed no significant matches to any identified genes in *D. melanogaster*. Furthermore, there is no corresponding EST evidence in *D. grimshawi* in this region. In general, a look at HERNE window for the alignment of a BLASTn search on the "Other EST" database revealed no significant matches to any region in Contig 9 that was not part of *Zfh2* or a repeat region. It is possible that this corresponds to a transposon remnant missing from library of repetitious elements in the database.

Finally, a region of high conservation downstream of the 3' UTR of *Zfh2* seems to be the 3' UTR of the ortholog of *Thd1* which runs in the reverse direction and ends downstream of Contig 9.

Figure 21: UCSC Genome Browser view of Contig 9 (most regions of high conservation correspond to exons, but the end of Contig 9 does not)



Repeat Regions

The table output for Repeat Masker provides a detailed account of identifiable repeats in Contig 9 (Table 1). Contig 9 consists of 21.94% repetitive sequence. There are four LINE elements, one of which is 4437 bases long, two LTR elements, and six DNA transposons in Contig 9. The orthologous region in *D. melanogaster* was also analyzed using Repeat Masker. In *D. melanogaster*, only 17.39% of sequence consists of repetitive sequence. There are no LINEs or LTR elements.

Table 1: Repeat Masker output table showing identifiable repetitive elements present in Contig 9

file name: Contig 9
sequences: 1
total length: 36397 bp (36397 bp excl N/X-runs)
GC level: 36.65 %
bases masked: 9968 bp (27.39 %)

	number of elements*	length occupied	percentage of sequence
Retroelements	6	5249 bp	14.42 %
SINEs:	0	0 bp	0.00 %
Penelope	0	0 bp	0.00 %
LINEs:	4	4973 bp	13.66 %
CRE/SLACS	0	0 bp	0.00 %
L2/CR1/Rex	1	242 bp	0.66 %
R1/LOA/Jockey	1	4437 bp	12.19 %
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:	2	276 bp	0.76 %
BEL/Pao	0	0 bp	0.00 %
Ty1/Copia	1	81 bp	0.22 %
Gypsy/DIRS1	1	195 bp	0.54 %
Retroviral	0	0 bp	0.00 %
DNA transposons	6	2699 bp	7.42 %
hobo-Activator	0	0 bp	0.00 %
Tc1-IS630-Pogo	4	2375 bp	6.53 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	0	0 bp	0.00 %
Tourist/Harbinger	0	0 bp	0.00 %
Other (Mirage, P-element, Transib)	0	0 bp	0.00 %
Rolling-circles	0	0 bp	0.00 %
Unclassified:	0	0 bp	0.00 %
Total interspersed repeats:		7948 bp	21.84 %
Small RNA:	0	0 bp	0.00 %
Satellites:	0	0 bp	0.00 %
Simple repeats:	25	1337 bp	3.67 %
Low complexity:	9	683 bp	1.88 %

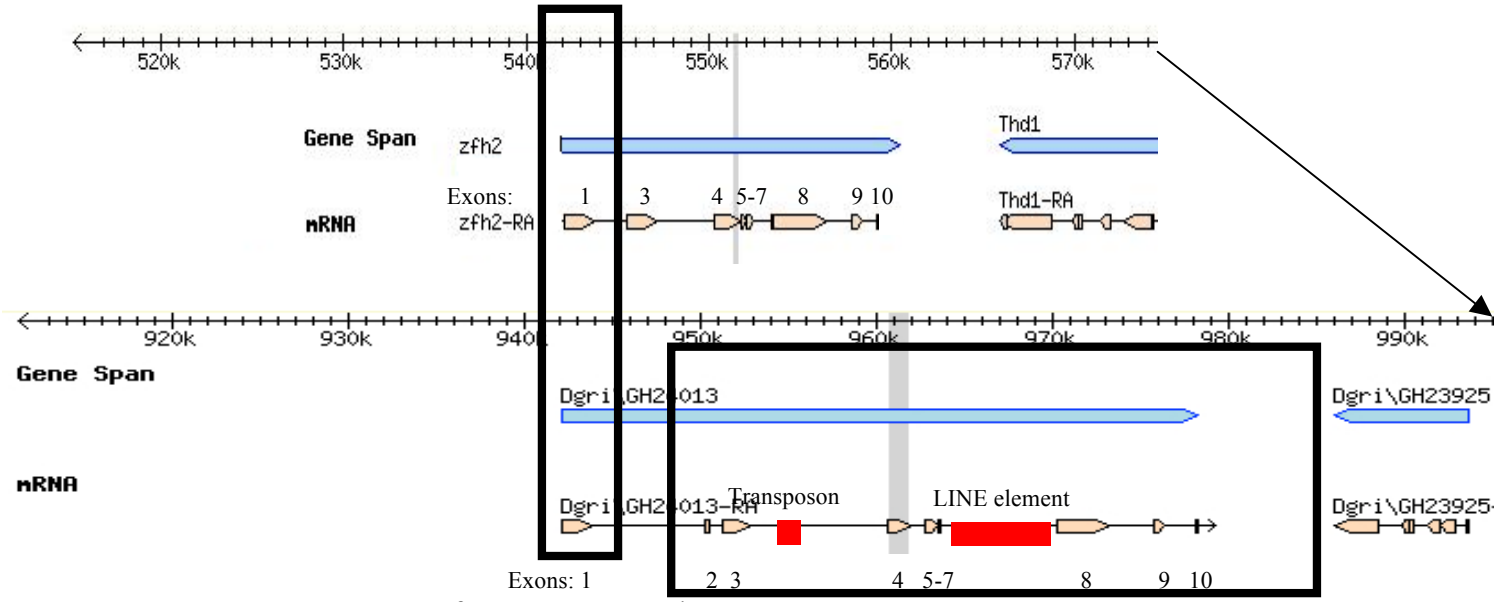
file name: *D. melanogaster repetitive elements*
sequences: 1
total length: 20650 bp (20650 bp excl N/X-runs)
GC level: 36.91 %
bases masked: 3592 bp (17.39 %)

	number of elements*	length occupied	percentage of sequence
Retroelements	0	0 bp	0.00 %
SINEs:	0	0 bp	0.00 %
Penelope	0	0 bp	0.00 %
LINEs:	0	0 bp	0.00 %
CRE/SLACS	0	0 bp	0.00 %
L2/CR1/Rex	0	0 bp	0.00 %
R1/LOA/Jockey	0	0 bp	0.00 %
R2/R4/NeSL	0	0 bp	0.00 %
RTE/Bov-B	0	0 bp	0.00 %
L1/CIN4	0	0 bp	0.00 %
LTR elements:	0	0 bp	0.00 %
BEL/Pao	0	0 bp	0.00 %
Ty1/Copia	0	0 bp	0.00 %
Gypsy/DIRS1	0	0 bp	0.00 %
Retroviral	0	0 bp	0.00 %
DNA transposons	13	3112 bp	15.07 %
hobo-Activator	0	0 bp	0.00 %
Tc1-IS630-Pogo	0	0 bp	0.00 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	0	0 bp	0.00 %
Tourist/Harbinger	0	0 bp	0.00 %
Other (Mirage, P-element, Transib)	1	41 bp	0.20 %
Rolling-circles	0	0 bp	0.00 %
Unclassified:	0	0 bp	0.00 %
Total interspersed repeats:		3112 bp	15.07 %
Small RNA:	0	0 bp	0.00 %
Satellites:	0	0 bp	0.00 %
Simple repeats:	1	173 bp	0.84 %
Low complexity:	9	307 bp	1.49 %

Synteny

Figure 22 shows a map of synteny centered around exon 1 of *Zfh2*. The top sequence shows the dot chromosome (chromosome 4) of *D. melanogaster* with *Zfh2* spanning over a 20 kb region. The bottom sequence shows the dot chromosome of *D. grimshawi* showing the ortholog of *Zfh2* spanning over a 36 kb region. The exon-by-exon analysis revealed that the size of each exon has not changed significantly. Rather, there are large insertions in the introns. Most significant is the insertion of a 4.4 kb LINE element between exons 7 and 8 in *D. grimshawi* not found in *D. melanogaster*. This is expected from the comparison of the Repeat Masker outputs in Table 1. The orthologous region in *D. grimshawi* consists of almost twice as much repetitive sequence as *D. melanogaster*. Finally, the dot plot in the Appendix comparing the ortholog of *Zfh2* in Contig 9 to the *Zfh2* in *D. melanogaster* also shows several insertions in the intronic sequence. The complete annotation map of Contig 9 is presented in Figure 23 on the next page.

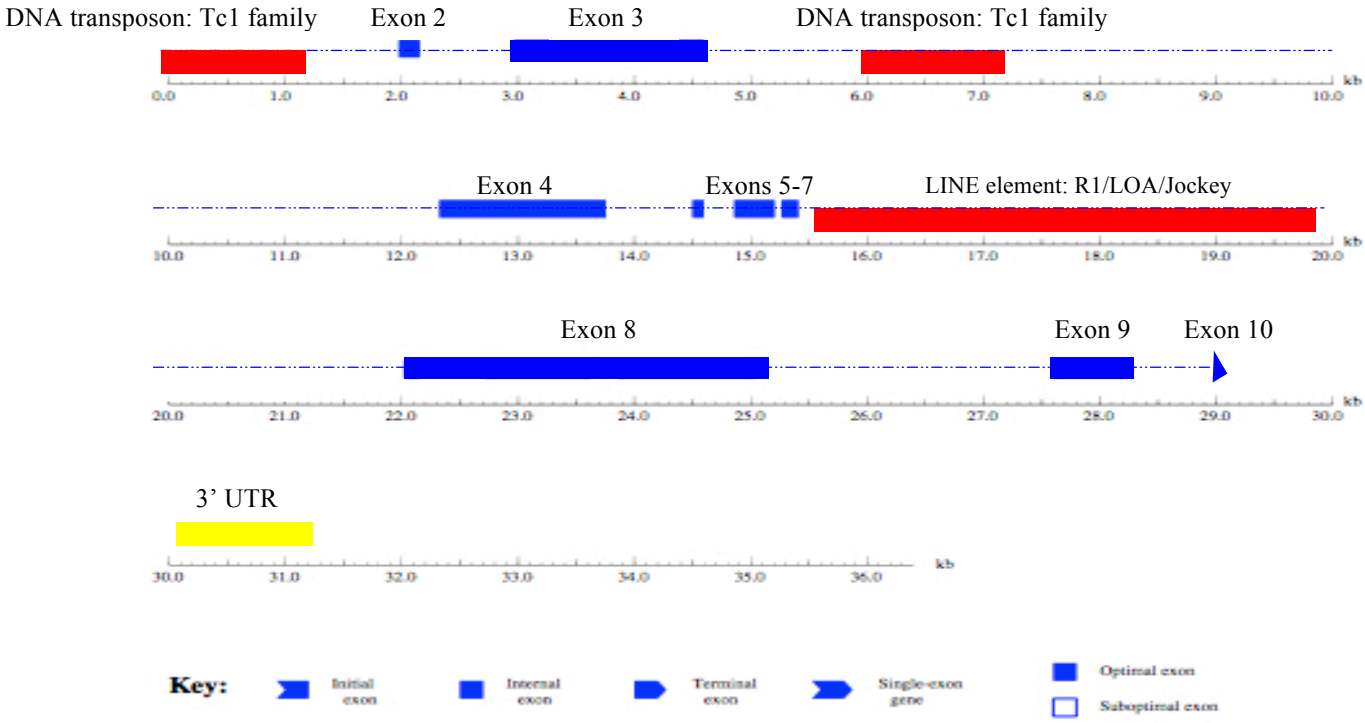
Figure 22: Synteny map of Zfh2 in D. melanogaster (on top) and D. grimshawi (on bottom). The map is centered around exon 1 and the scales of each have been adjusted so that 10 kb marks are exactly the same distance apart in both species. Exons have been labeled by number with significant transposable elements drawn in. The location of Contig 9 in the map is indicated by the large box from positions 949 kb to 985 kb on chromosome 4 in D. grimshawi.



Map of synteny centered on Exon 1 of *Zfh2*

36 kb nucleotide sequence of Contig 9 (boxed region)

Figure 23: Final Annotation Map of Contig 9



Appendix

Dot Plot of ortholog of *Zfh2* in Contig 9 to *Zfh2* in *D. melanogaster*

(Contig 9 is missing the 530 amino acid sequence in exon 1 of *Zfh2*)