David Wang

Bio 434W

4/27/15

Annotation of contig27 in the Muller F Element of *D. elegans*

Abstract

Contig27 is a 60,000 bp region located in the Muller F element of the *D. elegans*. Genscan predicted six features in the region (Figure 1), while the BLASTx alignment track in the UCSC Browser showed two genes with *D. melanogaster* orthologs, with two of the Genscan features overlapping exons of a single ortholog (Figure 2). Two genes were annotated in this region: *myo* and *ey*. The *myo* gene was determined to have one coding isoform with three coding exons, while *ey* had four isoforms with different coding regions and between 5 and 9 coding exons. For the *myo* gene, a broad transcription start site was annotated within 10 bp. An analysis of the repeats found by Repeat Masker showed that there were 7 repeats in contig27 with lengths longer than 500 bp. The orthologous region in *D. melanogaster* contained only one such repeat. A ClustalW2 multiple sequence alignment of the *myo* gene from several species of *Drosophila* and other similar organisms showed that conserved domains were maintained in certain regions throughout all of these species, possibly indicating active sites or other important regions of the protein. Through the orientation of the two genes in the orthologous region in *D. melanogaster*, synteny is preserved in this region.

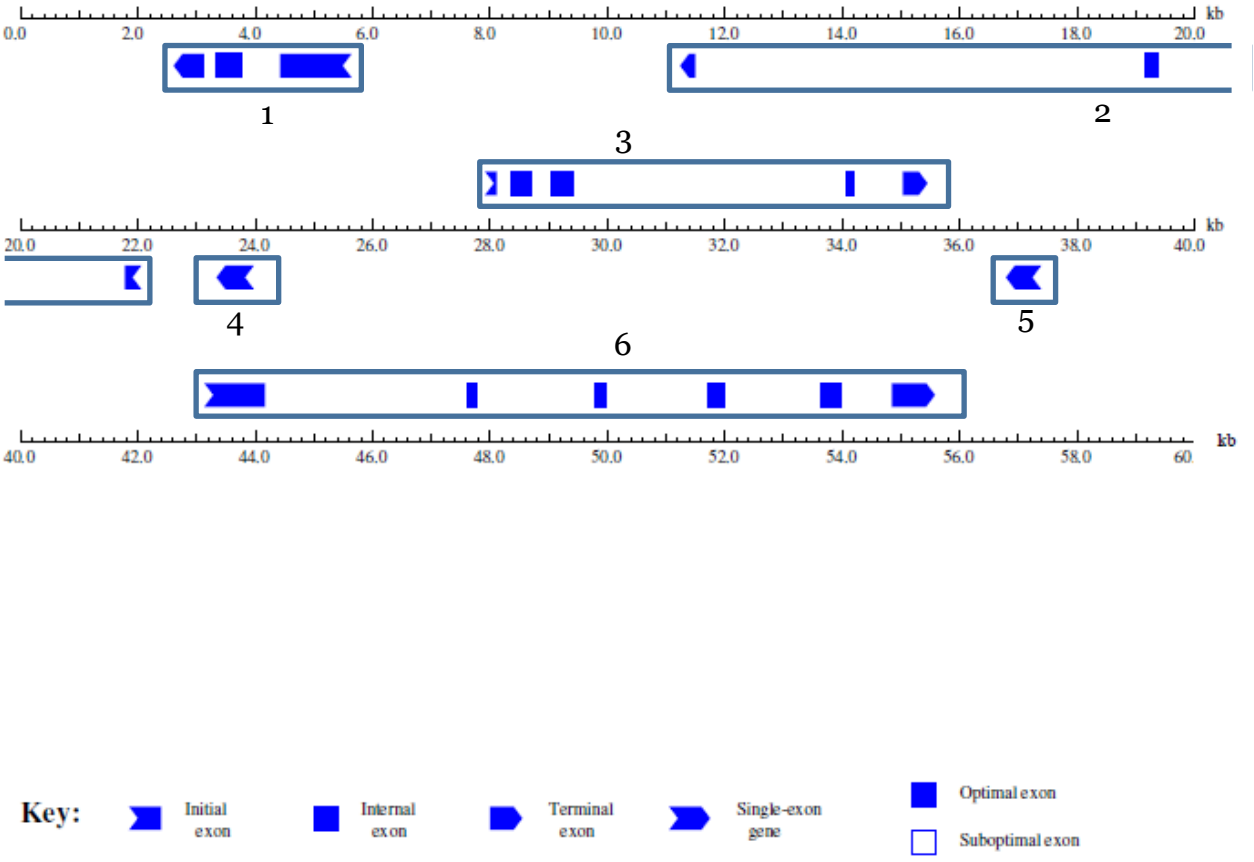


Figure 1: Genscan predicted genes in contig27. Predictions are numbered

Introduction

In *Drosophila*, the Muller F element or "Dot Chromosome" is a unique region. It shares many characteristics of heterochromatic regions, including low recombination rate and late replication. However, it also has a high gene density that is similar to known euchromatic regions. Understanding of the features of this element could be important in fully understanding the characteristics of heterochromatin and euchromatin.

The goal of this report is to annotate significant features of contig27, a 60,000 bp region of the Muller F element in *D. elegans* (Figure 2). In collaboration with other annotators, the entire Muller F element will be annotated. Since several other species of *Drosophila* have also had this region annotated, a comparative approach can be used to study the evolution of these genes across the different species. This could uncover further information about the function of these genes which is significant to understanding the characteristics of heterochromatic and euchromatic regions.

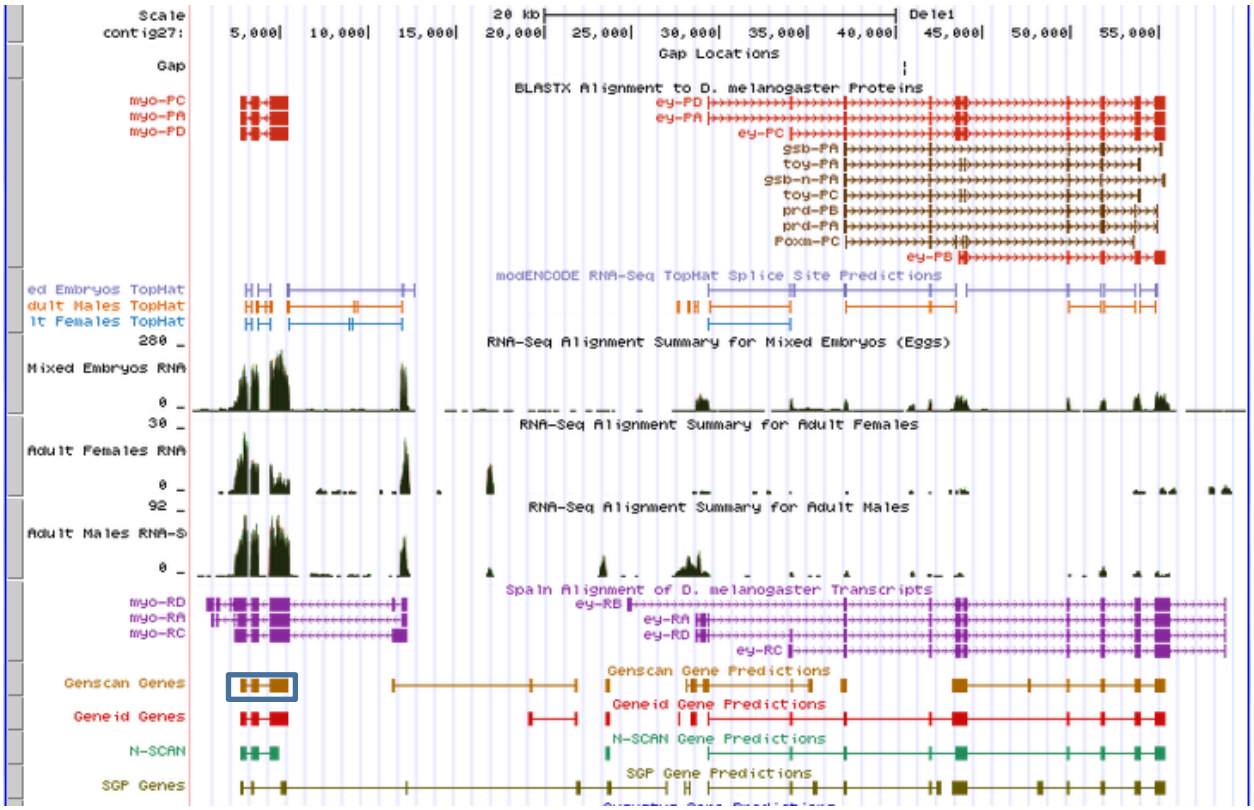


Figure 2: contig27 viewed in the UCSC Genome Browser. Genscan feature 1 is indicated

Genes

First Gene

The first gene annotated in this region corresponds to the first Genscan prediction in the contig (Figure 2). The general plan for annotating this feature is as follows. First, the *D. melanogaster* ortholog will be identified using a BLASTp alignment of the predicted protein from the Genscan prediction against the Annotated Proteins database (Figure 3). The starting point for this annotation was the Genscan predictions, as Genscan tends to overpredict genes more than the other gene predictors. Annotation involved using conservation with the BLASTx alignment of this ortholog, RNA-seq data, Tophat Junctions, and other tracks from the Jan. 2015 assembly of the *D. elegans* Dot Chromosome viewed in the GEP UCSC Browser. These tracks provided the evidence to determine splice sites and create a model for the gene. The model was checked using the Gene Model Checker on the GEP website. The *myo* gene ortholog was determined to be a 3 exon gene on the negative strand of this contig. The features of the final model for this gene can be seen in Table 1.

Send BLAST Hits to HitList

BLAST Hit Summary

	Description	Species	Score	E value
☑	myo-PD	Dmel	865.914	0
☑	myo-PC	Dmel	865.914	0
☑	myo-PA	Dmel	865.914	0
☑	dpp-PA	Dmel	77.411	4.53356e-14
☑	dpp-PE	Dmel	77.411	4.53356e-14
☑	dpp-PC	Dmel	77.411	4.53356e-14

Figure 3: BLASTp output shows *myo* is the clear ortholog. Query: Genscan predicted protein, Subject: Annotated Proteins

Exon Number	Start to Stop	Reference Frame	SS Acceptor Phase	SS Donor Phase
1	5490-4425	-1	N/A	1

2	3762-3344	-3	2	0
3	3127-2762	-3	0	N/A

Table 1: Characteristics of the three exons in the *myo* gene model are shown

In order to hypothesize a gene model for this Genscan prediction, I first searched for an appropriate ortholog in *D. melanogaster*. As shown in Figure 2, the BLASTx alignment track for *D. melanogaster* shows three isoforms for an ortholog in *D. melanogaster* that matches closely with the Genscan prediction. To confirm this assignment, a BLASTp search with the predicted protein from the Genscan prediction against the Annotated Proteins (AA) database in Flybase was run. The results can be seen in Figure 3. With such differences in scores and E-values, the *myo* gene was the clear ortholog in *D. melanogaster*. Since the three isoforms, myo-PA, myo-PC, and myo-PD only differ in untranslated regions, any of the three could be selected as the isoform for comparison for coding regions (Figure 4).

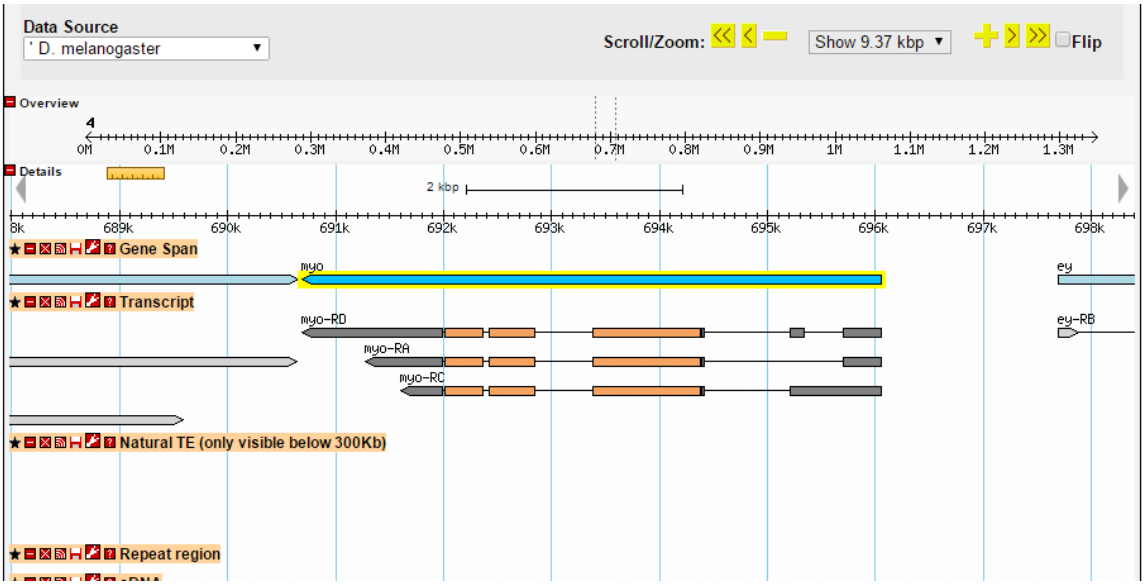


Figure 4: GBrowse view for the *myo* gene in *D. melanogaster:* all coding regions of the isoforms are identical

Using the *D. melanogaster* ortholog for comparison, determining the boundaries of the largest exon was a natural starting point to anchor a model of the entire gene. The largest exon is also the first exon translated for the *myo* gene. Close inspection of the beginning of this exon at the amino acid level reveals only one plausible starting point. This gene is on the negative strand. Using reading frames -2 or -3 results in premature stop codons very early in the gene (Figure 5). Since there is a stop codon almost immediately upstream from the methionine indicated, this must be the predicted start site. The beginning of a gene should be well conserved, so the conservation track result is expected. Further, the amino acid sequence in the first exon of this gene is well conserved in *D. melanogaster* if this reference frame is used (Figure 5, Figure 6). Thus, without excluding many amino acids, the proper start site for our gene model is as indicated in Figure 6. Inspection at the nucleotide level shows that this start point for this exon is at position 5490.

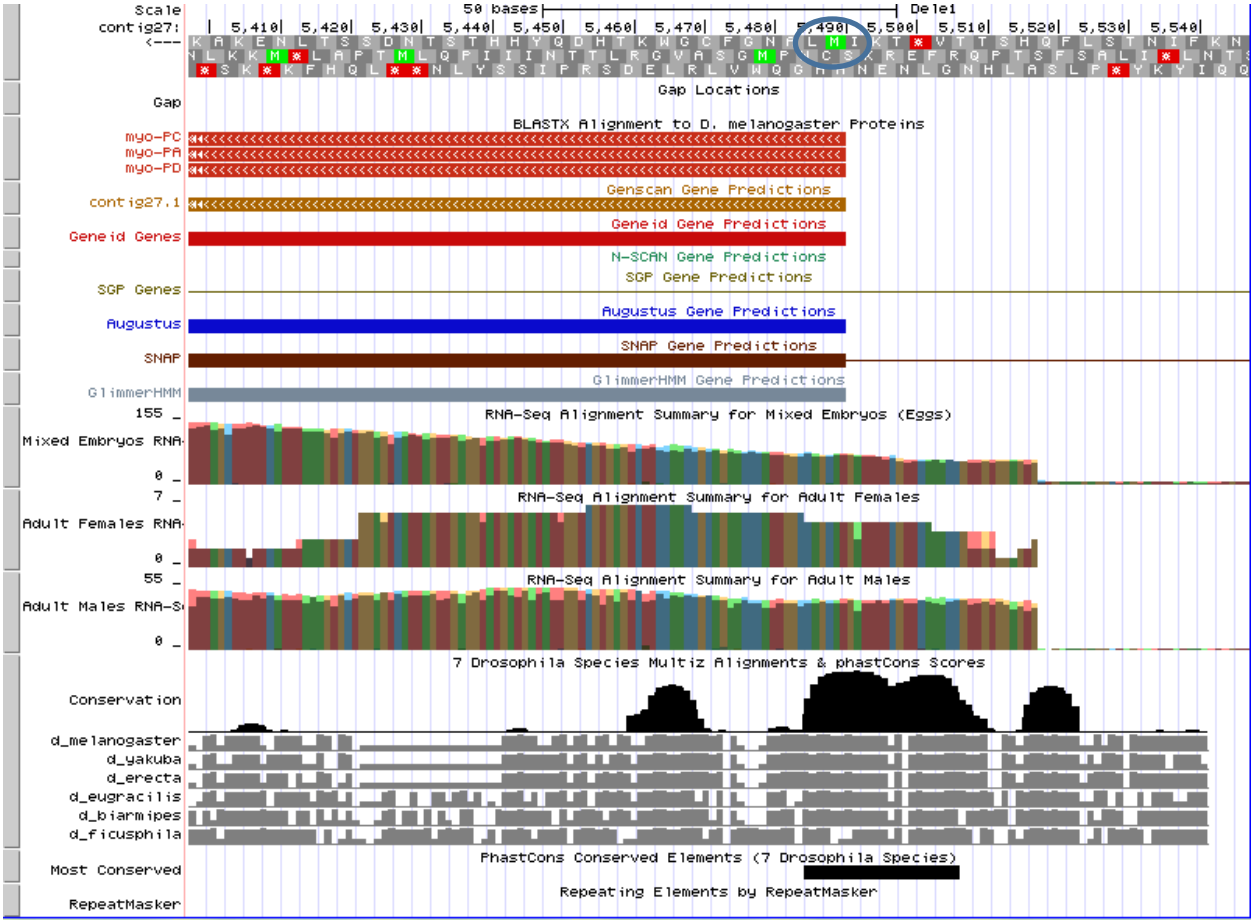


Figure 5: The predicted start of the first exon in *myo*. The start site inferred by location of stop codons is indicated. Use of this start site would result in a first exon that began with the amino acid sequence: MLANGFCGWK, mostly conserved in comparison with *D. melanogaster*

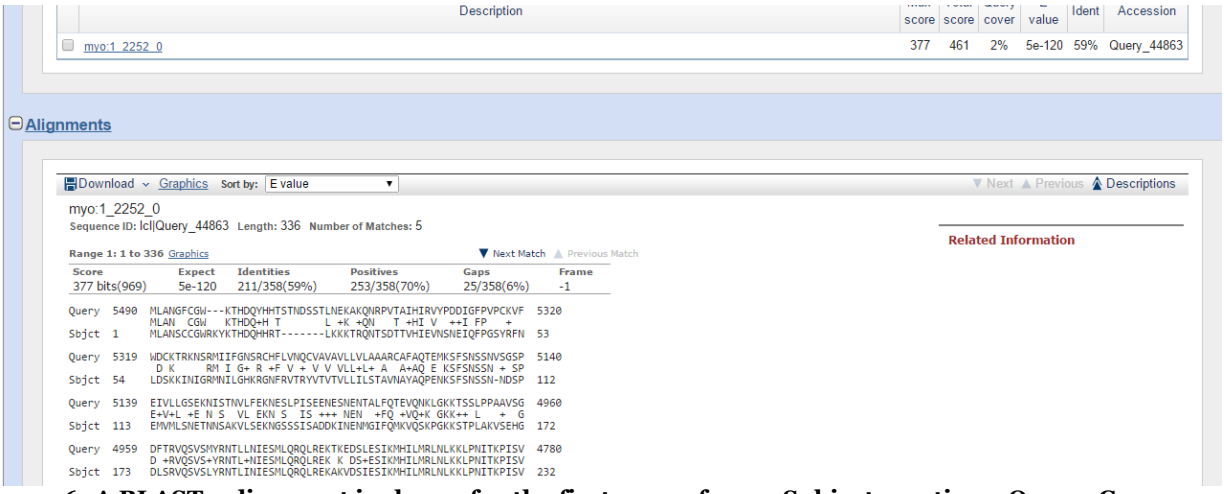


Figure 6: A BLASTx alignment is shown for the first exon of *myo*. Subject: contig27, Query: Genscan Prediction contig27.1. Seven of the first ten bases are conserved if this frame is used with even better conservation in other areas of the exon.

In order to determine the boundaries of introns and thus their adjacent exons, Genscan and other gene predictors search for the canonical 5' splice site, usually a GT, and the 3' splice acceptor, an AG. Most genes follow this model, so these sites with these two base combinations near the *D. melanogaster* splice sites were inspected first. For splicing to work, the number of bases directly adjacent to the intron that are not incorporated in codons on either side must add up to 3 or 0. Otherwise, there will be a frameshift in the model. Thus, the phase of a splice site refers to the number of bases not included in an exon directly adjacent to the splice. Further, the reading frame of exons can be inferred by conservation with *D. melanogaster*.

Consideration of these restrictions, the RNA-seq data, and the close inspection of the 5' splice donor site of the first intron at the nucleotide level yields only one logical site in the previously determined -1 reading frame. In the near vicinity of the ortholog's splice site, two GT's are present (Figure 7). However, RNA-seq data strongly support the first GT further upstream as the actual splice site. There is no evidence to suggest that the other GT is an alternative site. Thus, the boundary for the first exon of the *myo* gene model in *D. elegans* was determined to be at 4425.

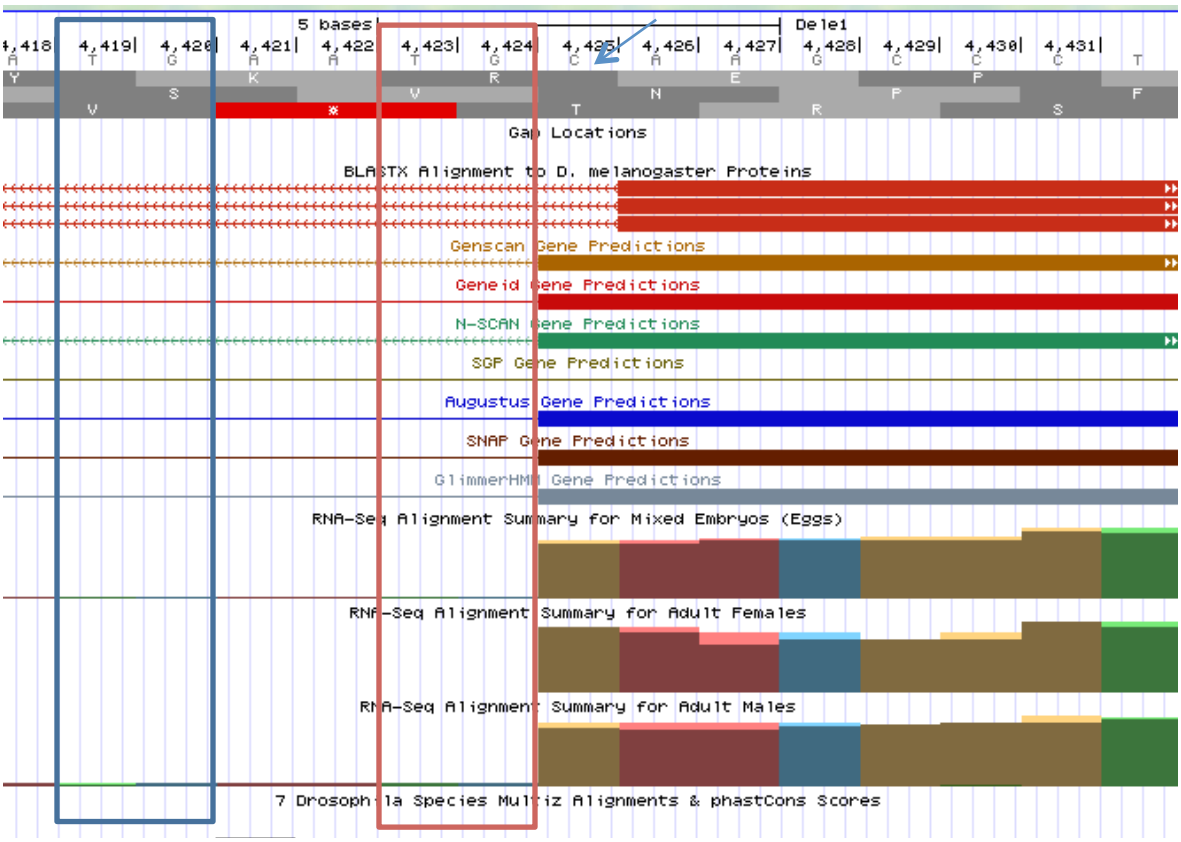


Figure 7: The 5' splice donor site for the first intron; the more likely of the two splice sites is indicated in red. Exon 1 is to the right of the splice and the top frame (-1) is the correct reading frame. A phase of one can be seen by the blue arrow.

Another tool that can be utilized to determine the boundaries of the intron is the TopHat Junctions tracks. Tophat is a tool derived from RNA-seq reads that span an intron entirely and are used to determine the location of splice sites. Each junction is given a score based on how much RNA-seq data support the junction. The TopHat junctions for this intron are shown in Figure 8. Only one of the junctions does not agree with the model for the 5' splice donor site hypothesized. However, this junction has a very low score of 1, so this is likely just a spurious result. Thus, TopHat Junctions, RNA-seq data, multiple gene predictors and the rules for splice sites all agree on the same 5'

splice donor site. Since the first exon is in the -1 frame, this splice site for the first exon has a phase of one.

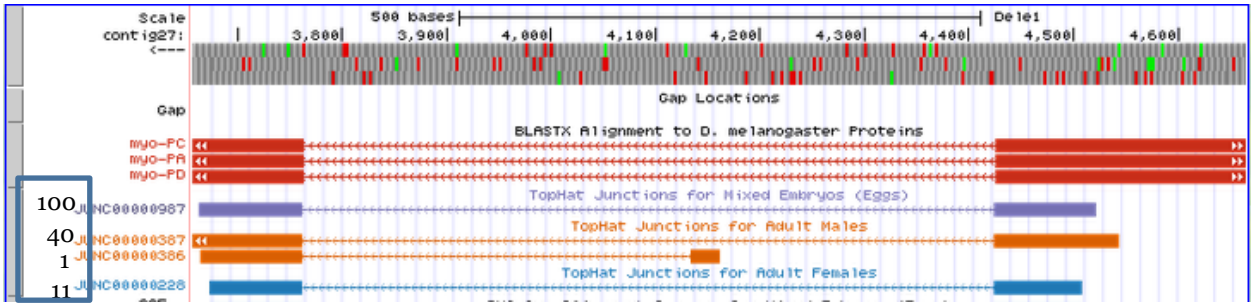


Figure 8: The TopHat junctions for the first intron; Scores for each junction are indicated

Since TopHat Junctions span predicted introns entirely, these tracks can be used to examine the splice acceptor site for this intron as well. Starting near the acceptor site of the *D. melanogaster* ortholog, most pieces of available data point to the same acceptor site. TopHat, RNA-seq, and all of the gene predictors indicate that 3762 is the splice acceptor site for the 2nd exon (Figure 9). This results in choosing the -3 reading frame for the 2nd exon. This is well supported by conservation with *D. melanogaster* (Figure 10).

Closer inspection of this region at the nucleotide level supports this hypothesized splice site. Acceptor splice sites are nearly always AG. There is only one AG within close vicinity of this region. This AG entirely corresponds to the prediction from RNA-seq and TopHat reads (Figure 9). Further, since the 5' splice donor site has a phase of one, this acceptor must have a phase of two. That points to the -3 translation frame as the proper frame. The -1 and -2 frame both contain many premature stop codons throughout the exon while the -3 frame does not, which further supports the 3' splice acceptor site at 3762.

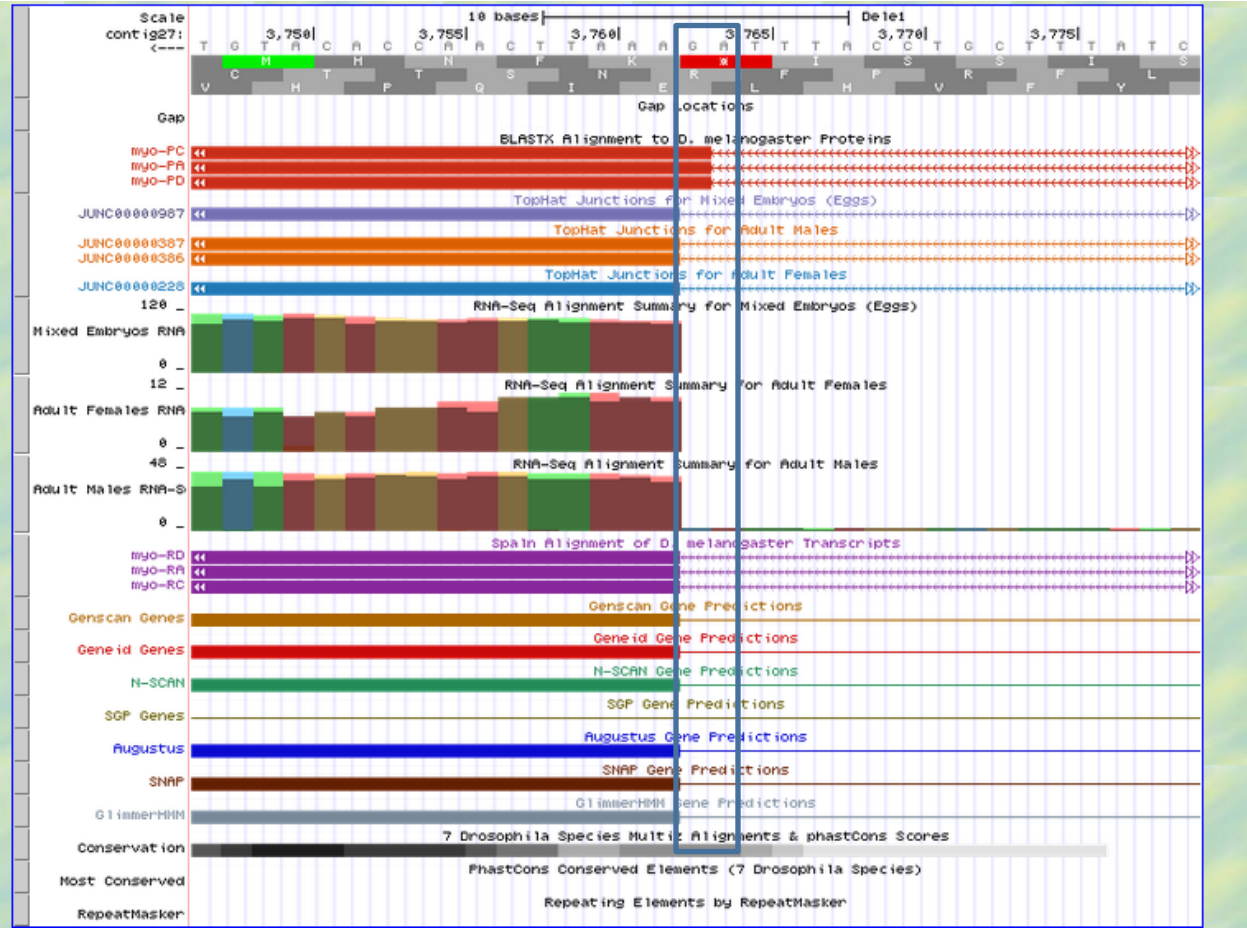


Figure 9: The 3' splice acceptor site between the junction of the first intron and second exon; the annotated splice site is indicated. The -3 reading frame results in a 2nd exon that begins with IQPHV

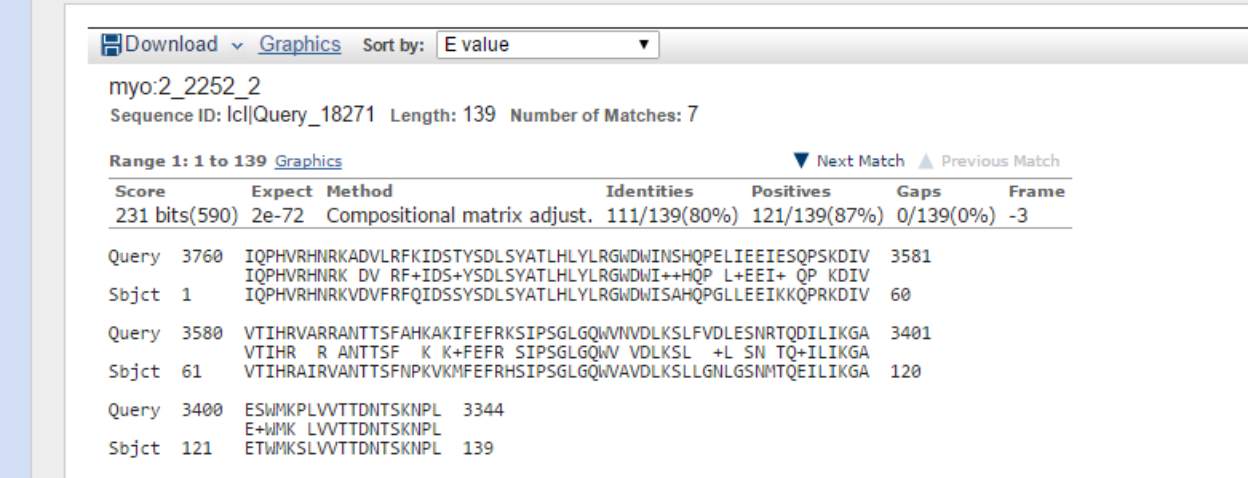


Figure 10: BLASTx alignment of the 2nd exon. Subject: 2nd exon of the *myo* gene in *D. melanogaster*, Query: contig27.fasta. The beginning of this exon is well conserved in the -3 reading frame

Since there is no premature stop codon and the RNA-seq data is continuous throughout the exon, the middle of this exon can be annotated in congruence with the ortholog. Annotating the 5' splice donor site at the junction between the second exon and the second intron follows a process similar to annotation of the first 5' donor site explained earlier in this report. The beginning of the exon in the *D. melanogaster* ortholog is used as a starting point. A GT combination that is supported by the other tracks and is near this location is a strong candidate for a 5' splice donor site.

A nucleotide level examination of the 5' splice donor region can be seen in Figure 11. There is only one GT in this region and its role as a 5' splice donor site is supported by data from TopHat, RNA-seq and several gene predictors. Thus, this is a likely candidate for the 5' splice donor site for the second intron for this gene. The gene model for *myo* in *D. elegans* has a second exon that ends at 3344. Since the second exon is in the -3 reading frame, this splice donor site at the exon end has phase 0.

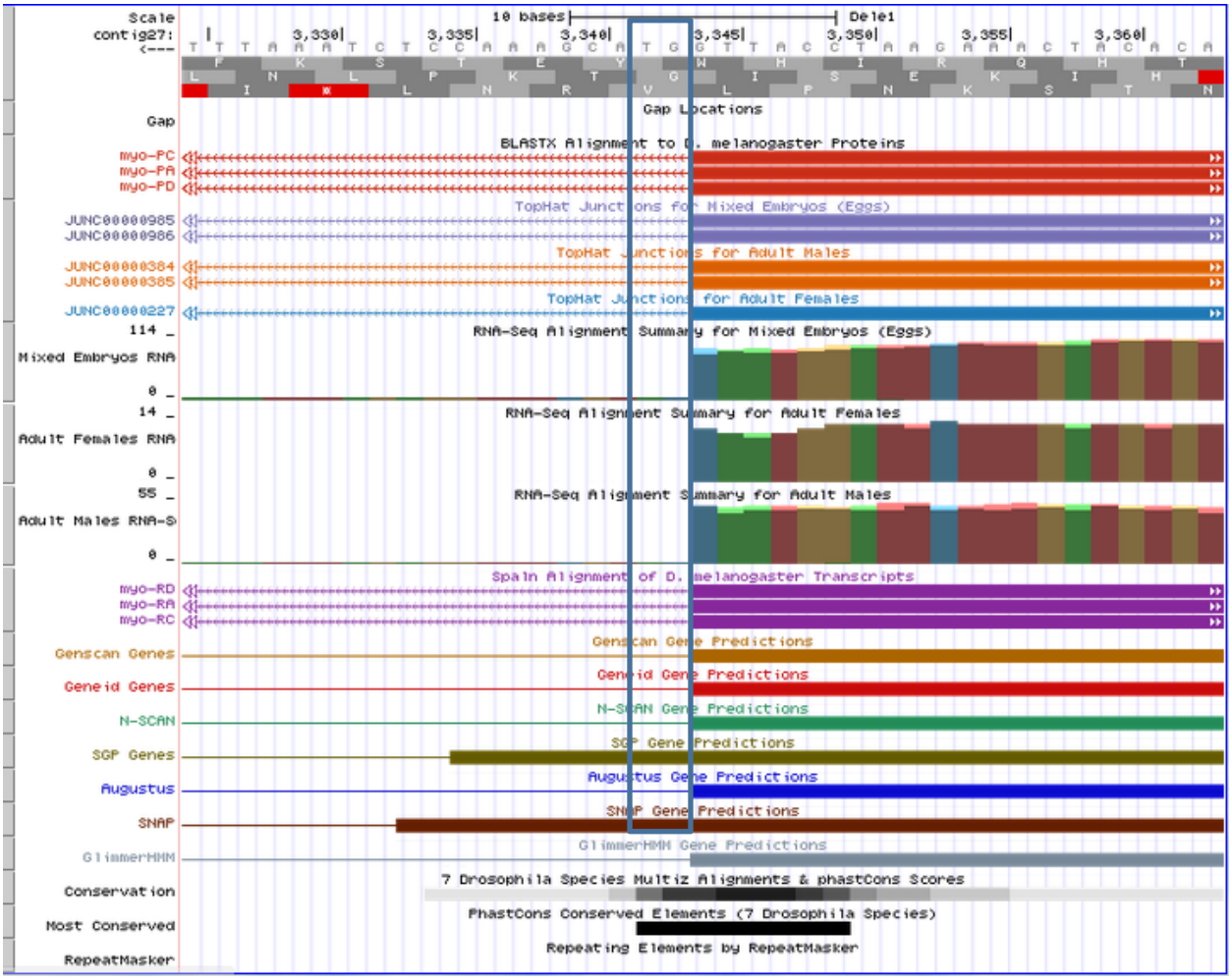


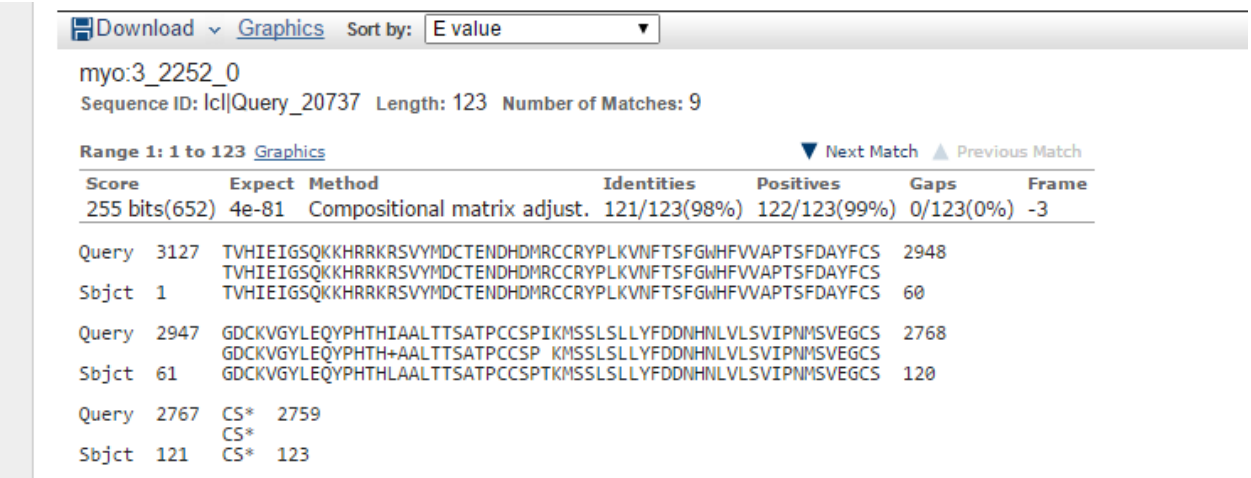
Figure 11: The junction between the 2nd exon and 2nd intron. The annotated splice donor site is indicated

The 3' splice acceptor site was revealed using the same method as above. The splice acceptor site in the BLASTx alignment of the *D. melanogaster* exon 3 was again used as a starting point. In this region, an AG combination that is supported as a splice site by RNA-seq and TopHat would be a likely 3' splice acceptor site candidate. Once a splice site was determined, the reference frame for the third and last coding exon was determined by matching its phase. Reference back to the BLASTx alignment shows that this results in an amino acid sequence that is conserved (Figure 12).

A nucleotide-level examination of the 3' splice acceptor region reveals two AG combinations near the *D. melanogaster* splice site (Figure 13). However, one of these

combinations is better supported as a splice acceptor site than the other. Both sites seem to have TopHat junctions as support, but observation of the scores for these junctions supports the upstream site. Further, RNA-seq data supports the upstream site strongly. The 0 phase of the 5' splice donor site indicates that this splice acceptor site must also have a phase of 0. Given the data solely at the junction between the 2nd intron and 3rd exon, the upstream AG is a more likely splice site, but both are possible. The more likely splice site would leave the third exon in the -3 reading frame due to phase. The downstream, less likely site would put the third exon in the -1 reading frame since the phase needs to be 0.

A broader inspection of the third exon shows that only the upstream splice site is possible for our model. If the downstream AG is the actual splice site, then the third exon would be in the -1 reading frame. However, this is not possible, as that would leave the exon with many premature stop codons (Figure 14). In contrast, the -3 reading frame has no premature stop codons, supporting the upstream AG as the more likely 3' splice acceptor site candidate. Thus, the starting point for the third exon is annotated at position 3127. This leads to the closest conservation of amino acids in the gene.


Download Graphics Sort by: E value

myo:3_2252_0

Sequence ID: lcl|Query_20737 Length: 123 Number of Matches: 9

Range 1: 1 to 123 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
255 bits(652)	4e-81	Compositional matrix adjust.	121/123(98%)	122/123(99%)	0/123(0%)	-3

```
Query  3127  TVHIEIGSQKKHRRKRSVYMDCTENDHDMRCCRYPLKVNFTSFGWHFVVAPTSFDAYFCS  2948
             TVHIEIGSQKKHRRKRSVYMDCTENDHDMRCCRYPLKVNFTSFGWHFVVAPTSFDAYFCS
Sbjct  1     TVHIEIGSQKKHRRKRSVYMDCTENDHDMRCCRYPLKVNFTSFGWHFVVAPTSFDAYFCS  60

Query  2947  GDCKVGYLEQYPHTHIAALTTSATPCCSPIKMSSLSLLYFDDNHNLVLSVIPNMSVEGCS  2768
             GDCKVGYLEQYPHTH+AALTTSATPCCSP KMSSLSLLYFDDNHNLVLSVIPNMSVEGCS
Sbjct  61    GDCKVGYLEQYPHTHLAALTTSATPCCSPTKMSSLSLLYFDDNHNLVLSVIPNMSVEGCS  120

Query  2767  CS*  2759
             CS*
Sbjct  121   CS*  123
```

Figure 12: BLASTx alignment of the third exon. Subject: 3rd exon of *myo* in *D. melanogaster*, Query: contig 27.fasta. This exon is very well conserved if the -3 reading frame is used.

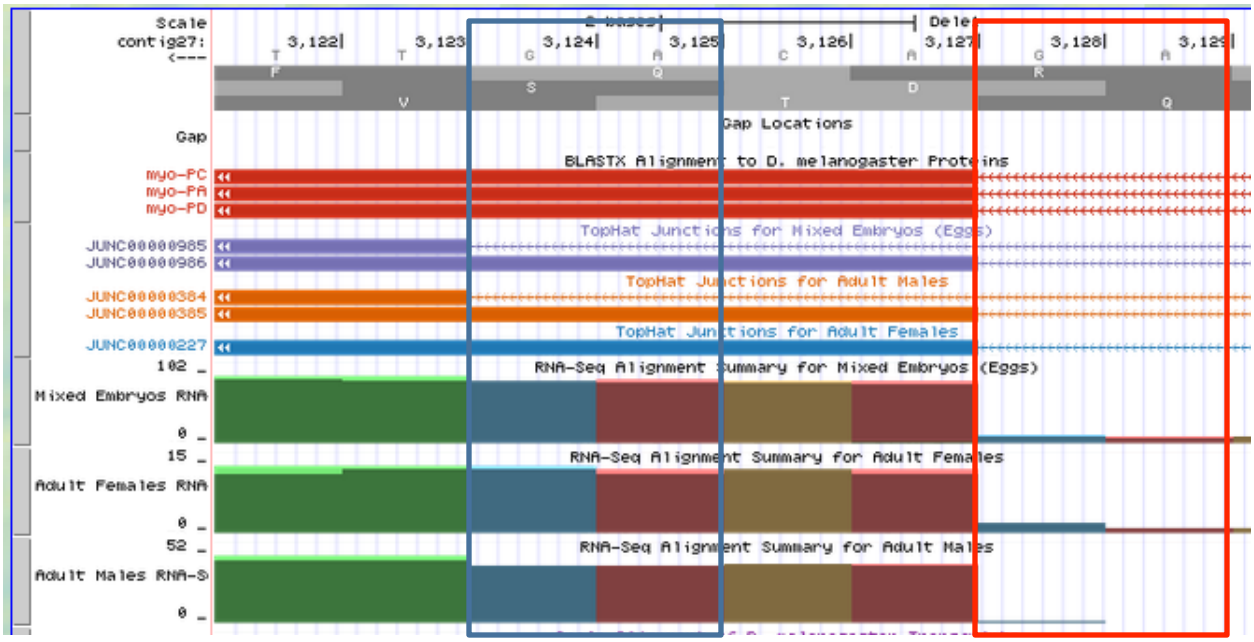


Figure 13: The junction between the 2nd intron and the 3rd exon. Both possible splice sites are indicated with the more likely site in red

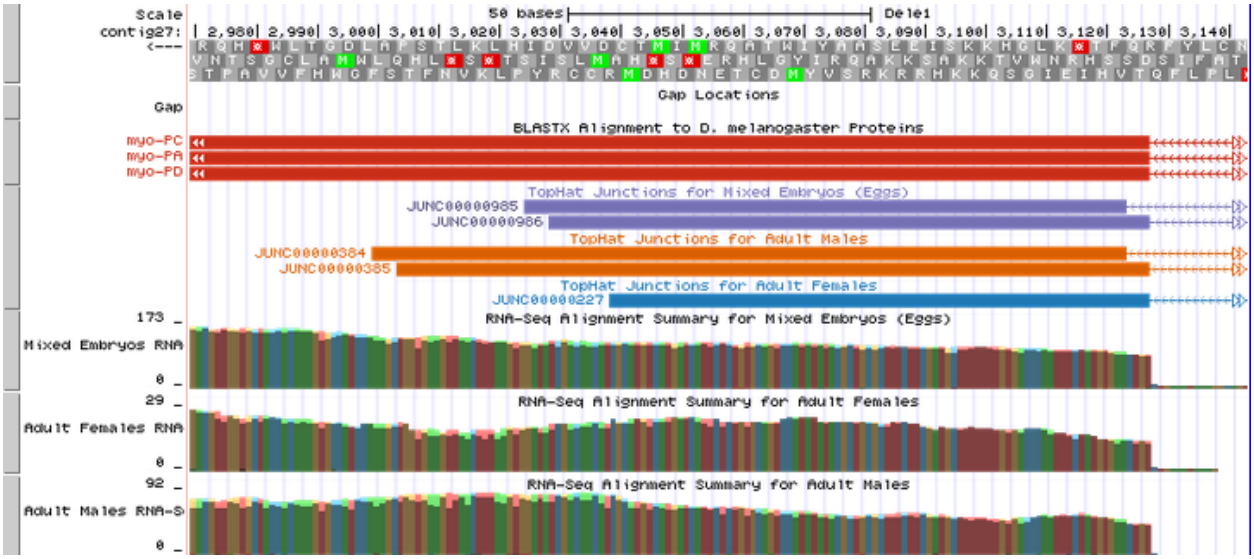


Figure 14: A broader view of the last exon; the -1 and -2 reading frames contain premature stop codons while the -3 frame does not

The last portion of the gene model to be annotated is the stop codon. An amino acid level examination reveals only one possible stop codon (Figure 15). The first stop codon in the -3 reading frame occurs at the site where a stop codon occurs in the *D. melanogaster* ortholog. This further supports the selection of the previous splice

acceptor site due to the functional reading frame. A 3' untranslated region causes RNA-seq data at the 3' end of genes to be uninformative; that the RNA-seq extends past the protein stop codon is expected. The stop codon is annotated at 2759-2761.

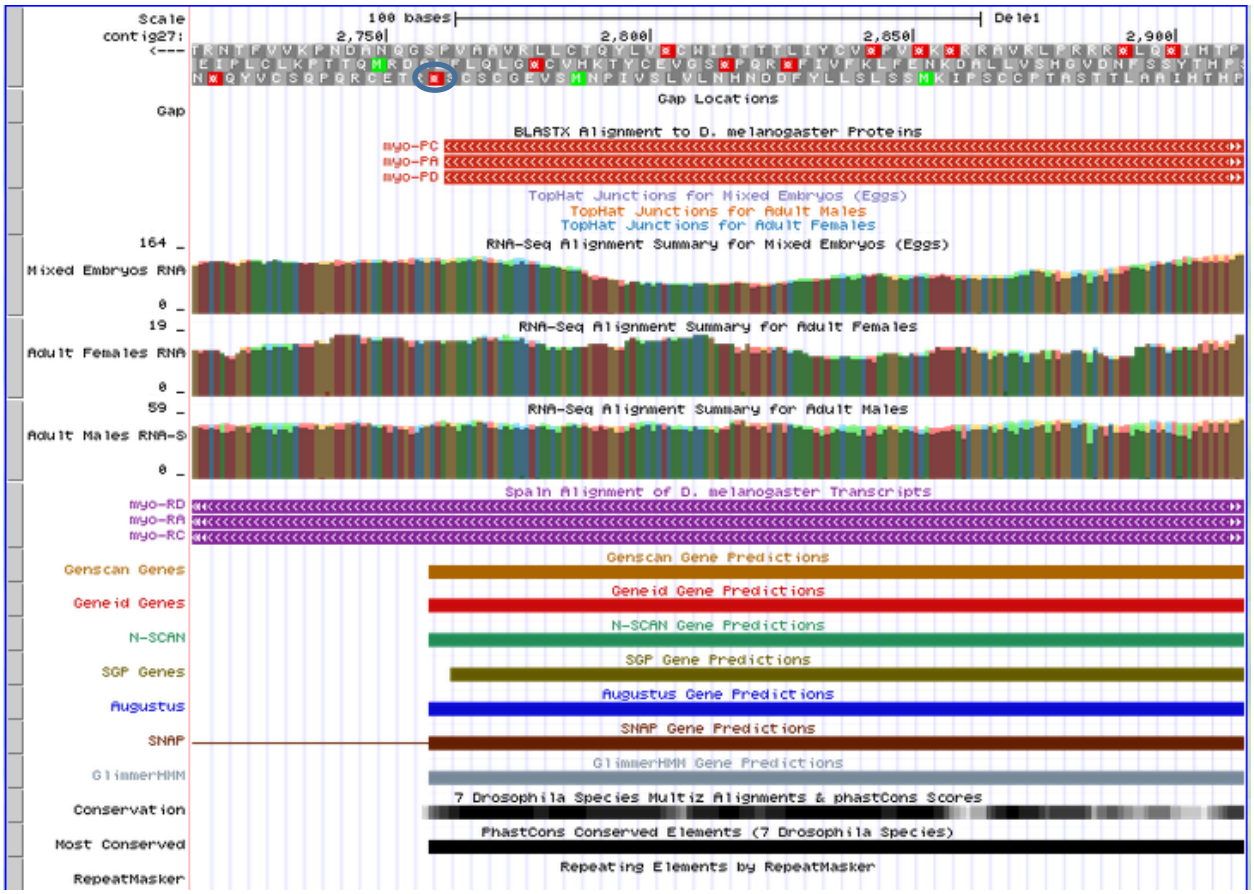


Figure 15: The stop codon for the *myo* gene is indicated

Confirming the First Gene

The characteristics of the gene model can be seen in Table 1. In order to check this model, the Gene Model Checker from the GEP website was utilized. The data from Table 1 was entered into the Gene Model Checker, and this model for the *myo* gene in *D. elegans* passed all of the tests in the checker. The output for the Gene Model Checker provides other tools that can be used to further support the gene model.

A dot plot of this gene model at the amino acid sequence level can be seen in Figure 16. This plot shows that the third exon of this gene is very well conserved between *D. elegans* and *D. melanogaster*. The second is less well conserved, and the first exon has relatively poor conservation. Since *D. elegans* and *D. melanogaster* usually have more conservation, this region is worth inspecting further. An amino acid sequence alignment shows that in particular, the first half of the first exon is poorly conserved (Figure 17).

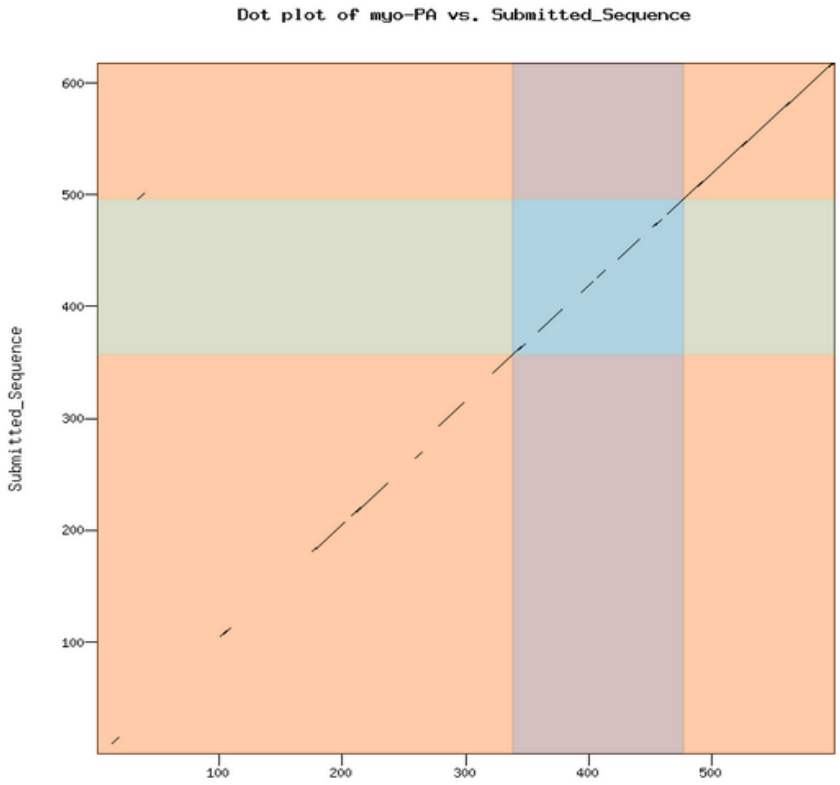


Figure 16: A dot plot of this gene *D. elegans* and *D. melanogaster;* the first exon has poor conservation as indicated by the gaps in the line

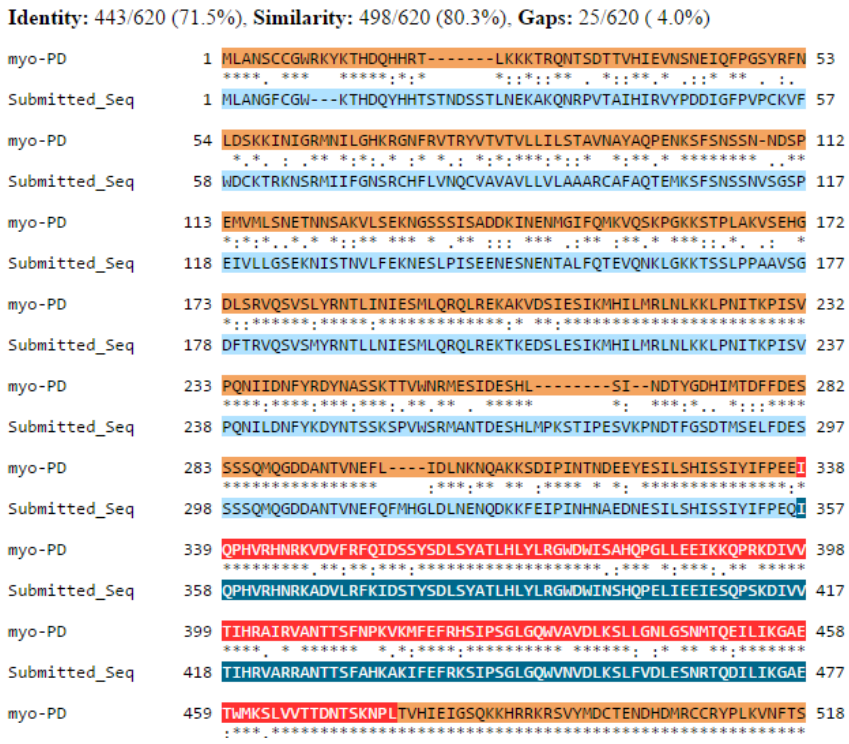


Figure 17: A protein alignment of the gene model against the *D. melanogaster* model. The first two exons and the beginning of the third exon are shown.

There is a chance that this first exon could be misplaced in the *D. elegans* model. If it is misplaced, the actual first exon would be upstream from the rest of the gene but still relatively near. Since there are about 55,000 bases upstream from this gene in contig27, the actual first exon is definitely contained within this contig. Thus, a BLASTx search of the contig27 .fasta file against the protein sequence of the first exon would uncover the actual location of the first exon if it is somewhere else. The output for this search can be seen in Figure 6. Since the protein homolog was only found in one location, the location of the first exon is as given by the model described in this report.

Region between the First and Second Genes

The second gene in this orthologous region in *D. melanogaster* as shown by the *D. melanogaster* BLASTx track in the genome browser is a multi-exon gene located

roughly from 30,000 and 55,000. In the region between this feature and the first gene though, Genscan predicted two features. These two features do not return significant matches anywhere in the Annotated Protein Database when a BLASTp search is run, so they are judged to be false positives and not considered further in annotation.

Second Gene

The rest of the Genscan predictions in contig27 overlap with a *D. melanogaster* ortholog shown by the *D. melanogaster* BLASTx Track. Annotation of this gene followed a procedure very similar to that of *myo* gene described above. The *D. melanogaster* ortholog determined was the *ey* gene. The *D. melanogaster ey* gene has four different isoforms with different coding regions (Figure 19). A summary of the boundaries of each exon can be seen in Tables 2-5.

Exon #	start	stop	Reference Frame	Acceptor Phase	Donor Phase
1	43629	44163	3	N/A	1
2	49788	49961	2	2	1
3	51712	51987	3	2	1
4	53634	53994	2	2	2
5	54840	55428	1	1	N/A

Table 2: Gene model summary for *ey-PB*

Exon #	start	stop	Reference Frame	Acceptor Phase	Donor Phase
1	34045	34210	1	N/A	1
2	37132	37297	3	2	2
3	41917	42097	2	1	0
4	43416	44163	3	0	1
5	49788	49961	2	2	1
6	51712	51987	3	2	1
7	53634	53994	2	2	2
8	54840	55428	1	1	N/A

Table 3: Gene model summary for *ey-PD*

Exon #	start	stop	Reference Frame	Acceptor Phase	Donor Phase
1	29313	29421	1	N/A	1
2	37132	37297	3	2	2
3	41917	42097	2	1	0
4	43416	44163	3	0	1
5	49788	49961	2	2	1
6	51712	51987	3	2	1
7	53634	53994	2	2	2
8	54840	55428	1	1	N/A

Table 4: Gene model summary for *ey-PA*

Exon #	start	stop	Reference Frame	Acceptor Phase	Donor Phase
1	29313	29421	3	N/A	1
2	34031	34210	1	2	1
3	37132	37297	3	2	2
4	41917	42097	2	1	0
5	43416	44163	3	0	1
6	49788	49961	2	2	1
7	51712	51987	3	2	1
8	53634	53994	2	2	2
9	54840	55428	1	1	N/A

Table 5: Gene Model summary for *ey-PC*

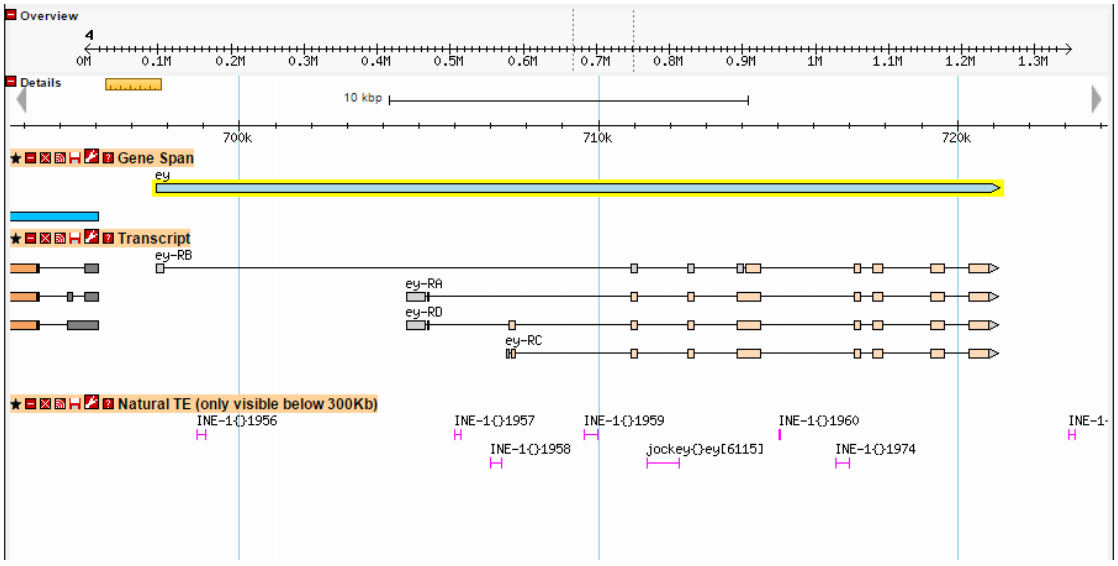


Figure 18: The four different coding isoforms of *ey* in *D. melanogaster* are shown in GBrowse view.

To determine the ortholog in *D. melanogaster*, the first region searched was the feature identified by Genscan Prediction contig27.6. This predicted protein was used in BLASTp and searched against the database of Annotated Proteins in *D. melanogaster* (Figure 19). There is a significant jump in the E-Value and score between the top four isoforms of the *ey* gene and the next most significant matches. To further support the claim that the appropriate *D. melanogaster* ortholog is the *ey* gene, Genscan prediction contig27.4 was also searched against the Annotated Proteins database. The result is another close match to *ey* isoforms without any other significant matches (Figure 20). Four isoforms were found in the *D. melanogaster* database, and four coding isoforms were annotated within this last gene in *D. elegans*.

BLAST Hit Summary				
	Description	Species	Score	E value
☑	ey-PC	Dmel	679.093	0
☑	ey-PD	Dmel	678.707	0
☑	ey-PA	Dmel	678.707	0
☑	ey-PB	Dmel	675.241	0
☑	toy-PA	Dmel	174.096	4.74429e-43
☑	toy-PC	Dmel	159.458	1.17938e-38
☑	al-PA	Dmel	110.923	5.27287e-24

Figure 19: BLASTp search to find the orthologous gene. Subject Annotated Proteins of *D. melanogaster*, Query: Predicted Genscan Protein contig 27.6. The top four matches are all isoforms of the same gene.

BLAST Hit Summary				
☑	Description	Species	Score	E value
☑	ey-PC	Dmel	68.1662	1.33863e-11
☑	ey-PD	Dmel	68.1662	1.38406e-11
☑	CG5347-PA	Dmel	29.6462	6.23443

>gnl|dmel|FBpp0099809 type=protein; loc=4:join(707506..707671, 710917..711082, 712493..712673, 713858..714512, 717127..717300, 717654..717929, 719274..719628, 720322..720922); ID=FBpp0099809; name=ey-PC; parent=FBgn0005558, FBtr0100395; dbxref=FlyBase_Annotation_IDs:CG1464-PC, FlyBase:FBpp0099809, GB_protein:AAX52513.1, REFSEQ:NP_001014694, GB_protein:AAX52513, FlyMine:FBpp0099809, modMine:FBpp0099809; MD5=545eb997734633237c4c96532b46af3f; length=857; release=r6.04; species=Dmel;
Length = 857

HSP # = 1 , **Score** = 68.1662 bits (165) , **Expect** = 1.33863e-11
Identities = 32 / 35 (91.4%) , **Positives** = 32 / 35 (91.4%)

Subject FASTA

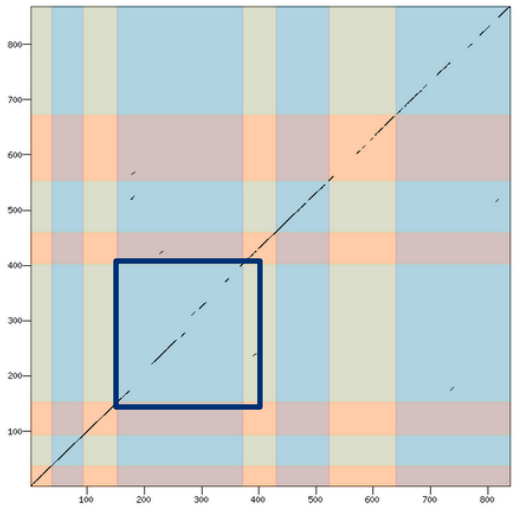
```
Query:   274          KPSPTMEAVEAGPASQPHSTSSYFATTYYHLTDDE   308
                      KPSPTMEAVEA  AS PHSTSSYFATTYYHLTDDE
Subject: 21           KPSPTMEAVEASTASHPHSTSSYFATTYYHLTDDE   55
```

Figure 20: BLASTp search indicating upstream features of the gene. Subject: Annotated Proteins of *D. melanogaster*. Query: Genscan Predicted Protein contig27.4.

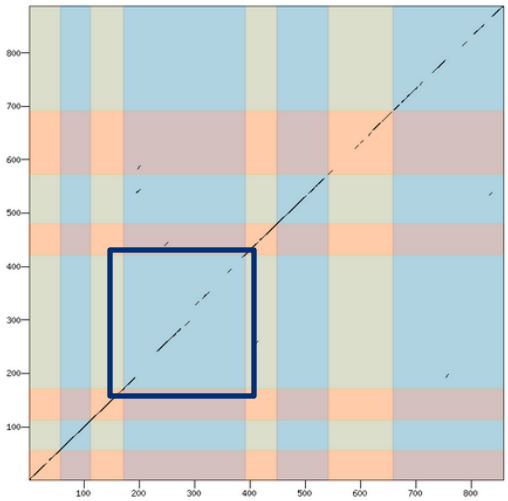
Once the proper *D. melanogaster* orthologs were identified, a model for each isoform could be developed in a manner similar to that described for the *myo* gene. For all four isoforms, splice sites were determined by examining canonical splice sites using the BLASTx alignment track as a guide. The exact location and reference frame of each splice was determined first by conservation with *D. melanogaster*. Finally, matching phases, location of premature stop codons, and the RNA-seq and TopHat Junction tracks were all considered when deciding on splice sites. Then, the gene models were all checked using the Gene Model Checker and examined using other outputs from the

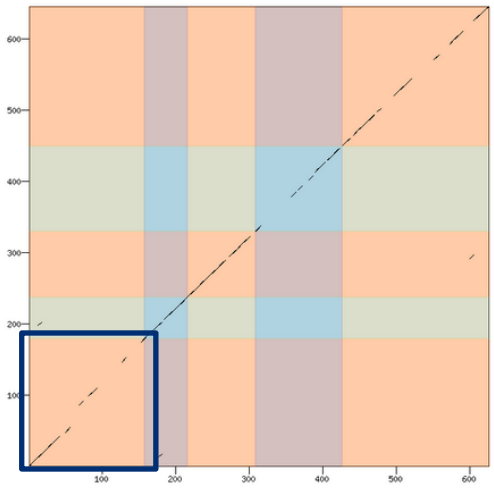
Gene Model Checker such as the Dot Plot and Protein Alignment outputs. The dot plots from each of the four isoforms can be seen in Figure 21.

a)

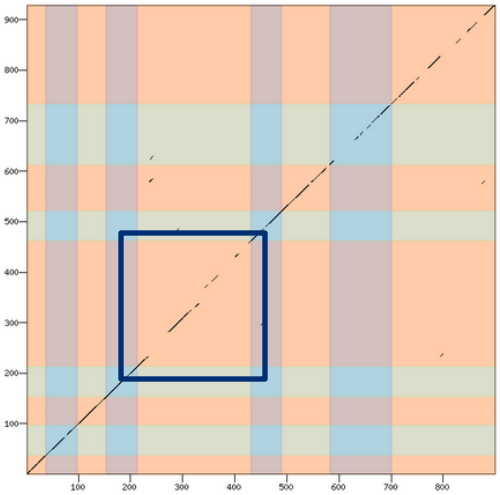


c)





b)



d)

Figure 21: The dot plots for the four isoforms of *ey*. a) *ey-PA*, b) *ey-PB*, c) *ey-PC*, d) *ey-PD*. The vertical axes represent the sequences of each isoform in contig27 and the horizontal axis denotes the *D. melanogaster* ortholog. The boxed region is an exon present in all four orthologs. The shift in the comparison line within this exon suggests an insertion or expansion

One region that seemed to have conflicted with some of the data given by the UCSC Genome Browser was located within the largest exon of each of the four isoforms, 6_931_0 or 7_931_0. The BLASTx Alignment, the Spaln Alignment of *D. melanogaster* Transcripts track, and the conservation in Drosophila tracks all seemed to suggest that there is a small intron in the middle of this gene (Figure 22). However, a review of each isoform in the Gene Record Finder indicated that this region was in fact one exon in *D. melanogaster* (Figure 23).

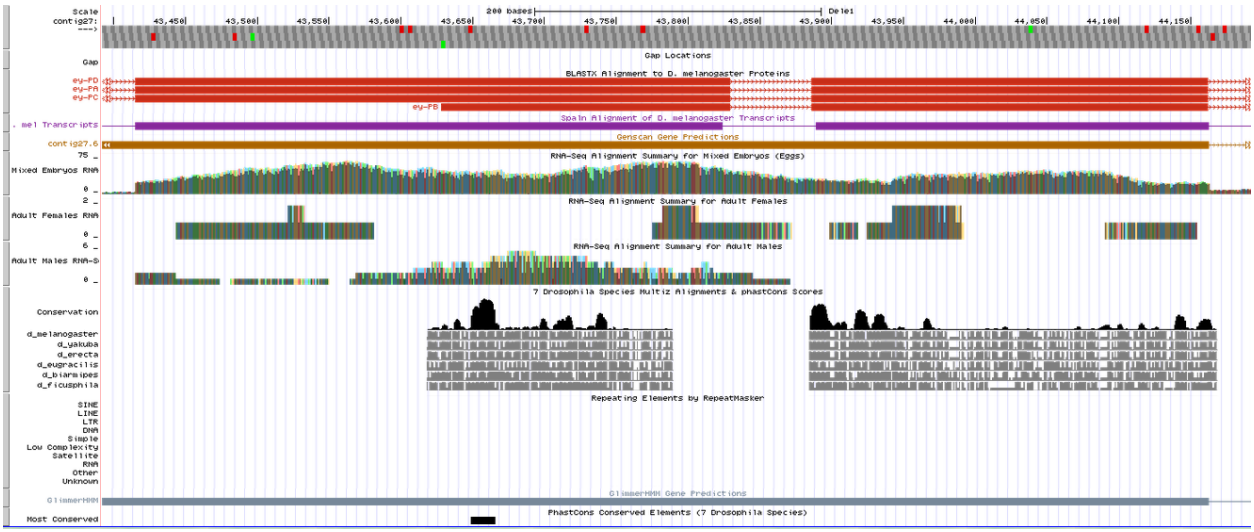
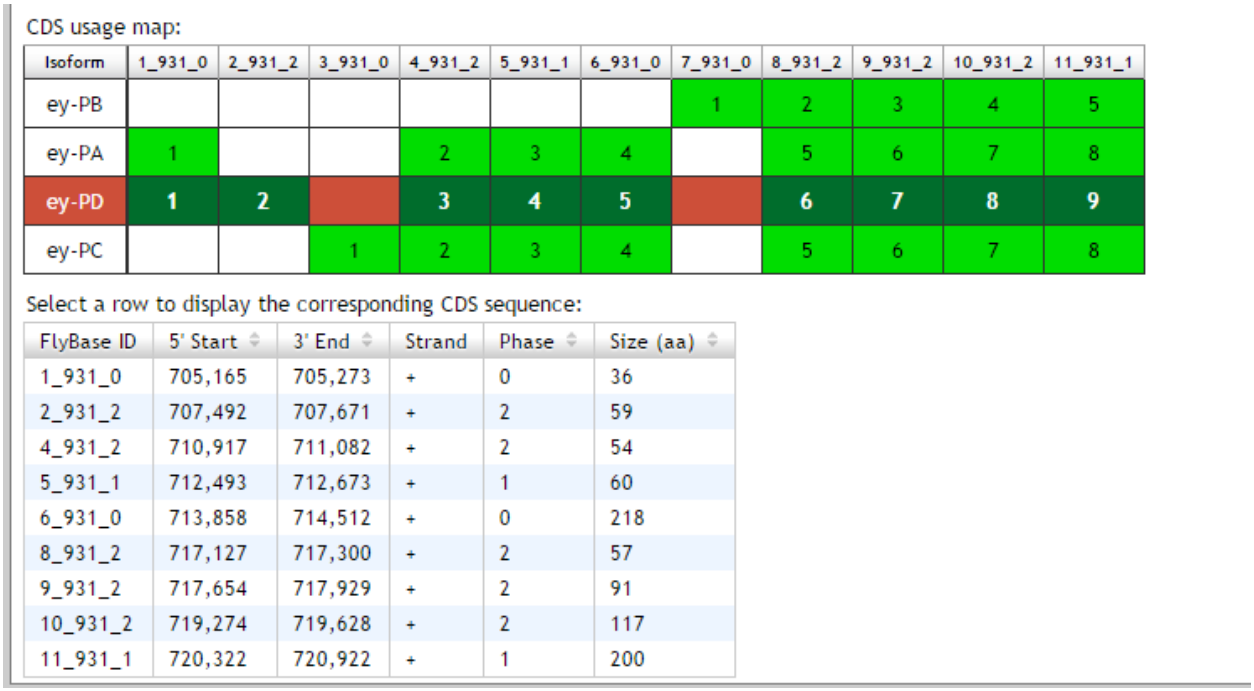


Figure 22: The possible intron in the middle of an exon for this gene.

This region also corresponds to the least well conserved region according to the dot plots with a shift in exon size. Closer inspection of this region at the amino acid level reveals that the potential small intron is actually a relatively simple sequence. This sequence is not within the orthologous region of *D. melanogaster,* and an inspection of the orthologous region of other species of *Drosophila* shows that this simple sequence is unique to *D. elegans* (Figure 24). Since this region is only present in *D. elegans,* further inspection was necessary to explore its characteristics. If this region is a small intron, then there must be compatible splice sites on either side of the intron.

Prior evidence showed that the beginning and end of exon 6_931_0 and 7_931_0 must be in the +3 reading frame (Tables 2-5). If this region is indeed a small intron, then the exons it separates must both be in the +3 reading frame. A nucleotide level inspection of this region shows six potential canonical splice donors and two potential splice acceptors (Figure 25). Introns must be at least 40bp in length, so only the two most upstream splice donor sites are viable. Relative to the +3 reading frame, both donors have phase 1, and the potential acceptor sites have phases of 0 and 1. Thus, none of these splice pairs are viable. Additionally, continuous RNA-seq reads through the potential intron support the conclusion that this is one continuous exon, and the region was annotated as such. It is possible that this simple sequence expansion was an evolutionary event that occurred recently in the evolutionary history of *D. elegans*, although the resulting protein is still functional.



CDS usage map:

Isoform	1_931_0	2_931_2	3_931_0	4_931_2	5_931_1	6_931_0	7_931_0	8_931_2	9_931_2	10_931_2	11_931_1
ey-PB							1	2	3	4	5
ey-PA	1			2	3	4		5	6	7	8
ey-PD	1	2		3	4	5		6	7	8	9
ey-PC			1	2	3	4		5	6	7	8

Select a row to display the corresponding CDS sequence:

FlyBase ID	5' Start	3' End	Strand	Phase	Size (aa)
1_931_0	705,165	705,273	+	0	36
2_931_2	707,492	707,671	+	2	59
4_931_2	710,917	711,082	+	2	54
5_931_1	712,493	712,673	+	1	60
6_931_0	713,858	714,512	+	0	218
8_931_2	717,127	717,300	+	2	57
9_931_2	717,654	717,929	+	2	91
10_931_2	719,274	719,628	+	2	117
11_931_1	720,322	720,922	+	1	200

Figure 23: Exons in Gene Record Finder. The intact exon is labeled as 6_931_0 and 7_931_0

```
sechelia_1          ARAAPLVGQSPNHLGTRS------------------SHP--QLVHGNHQALQQHQQQSWP 149
melanogaster_1      ARAAPLVGQSPNHLGTRS------------------SHP--QLVHGNHQALQQHQQQSWP 149
yakuba_1            ALAVP-LGQAPNHLVTHS------------------SHPLTQLVHGNHQALQQQQQQSWP 148
takahashi           AIAAPLVGQAPGHFEAHS------------------SHPLNQLVHGNQQALQQQQQ-SWP 148
biarmipes_1         AIAVPLVGQSPSHFGPHS------------------SHPLNQLVHGNQQALQQQQ--SWP 151
elegans             SIAAPLVGQAPGHFGSHSSHSHSHSHGLGHGHGHSHGHPLNQLVHGGQQVLQQQQQ-SWP 175
rhopaloa_1          ATAAPLVGQAPSHFGPHS------------------GHPLNQLVHGNQQVLQQQQQ-SWP 150
                    : *.* :**:*.*: .:*                  .**  *****.:*.***:*  ***
```

Figure 24: A multiple sequence alignment of exon 6_931_0 and orthologous regions of other *Drosophlila* species. The region unique to the *D. elegans* could be a simple sequence expansion

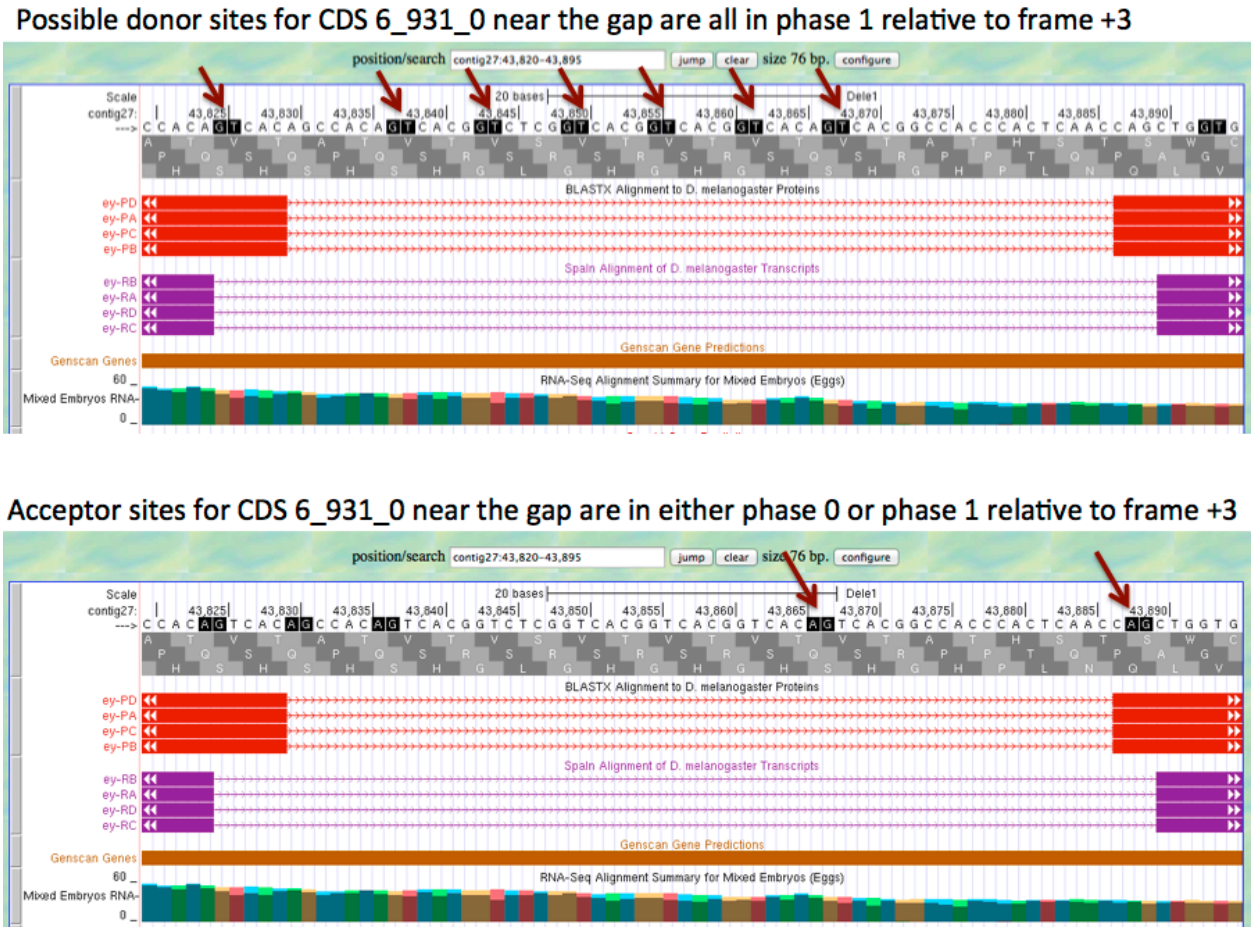


Figure 25 (Leung W., Personal Communication): The six potential splice donor sites and two potential splice acceptors. No donor-acceptor pairs have matching phases, so this region is very unlikely to be an intron. The region is annotated as one continuous reading frame.

Transcription Start Site (TSS) Estimates

The gene that was a strong candidate to find a TSS was *myo* because its RNA-seq data was so pronounced. Since each isoform had 5' non-translated exons that started at the same place, only one TSS was annotated for this gene. A more sensitive BLAST

search would be used to map the TSS and TSS search region on the *D. elegans* genome. Then a comparison with the first transcribed exon in *D. melanogaster* would be utilized to examine the TSS further. Regulatory motifs in *D. elegans* provided further evidence of the TSS.

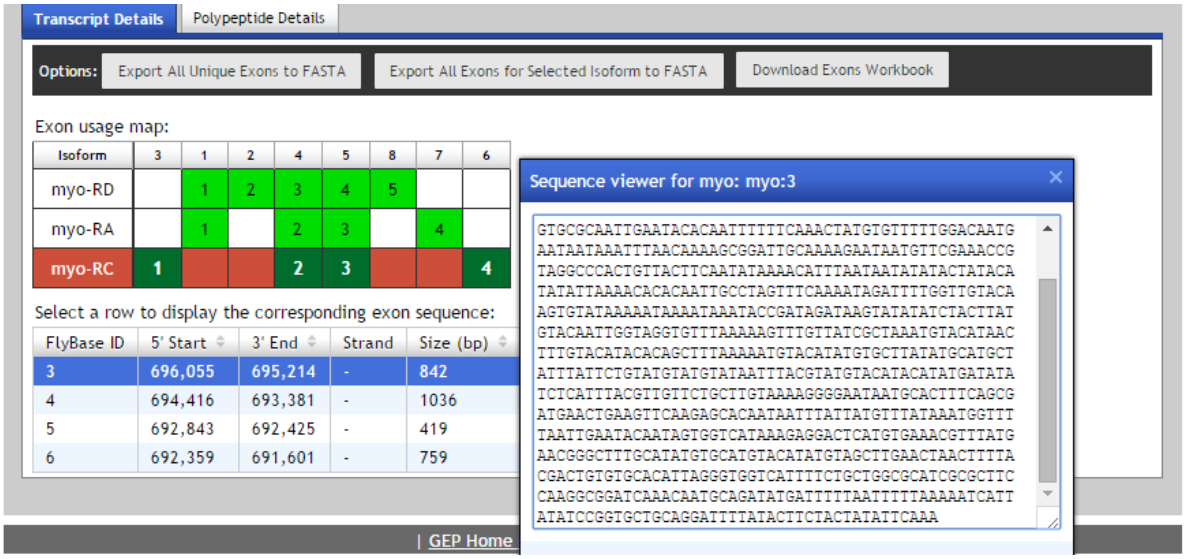


Figure 26: Gene Record Finder was used to find the sequence for the first transcribed exon in *myo-RC* for *D. melanogaster*

A BLASTn search using the first transcribed exon in *D. melanogaster myo-RC* against contig27 returns a close match that has five bases missing from the beginning of the match (Figure 27). The sixth base, which is the first in the BLAST output, is located at 12,213 in *D. elegans*. Since the TSS is located at the very beginning of the first transcribed exon, the TSS in *D. elegans* is located at 12,213 or upstream from that location.

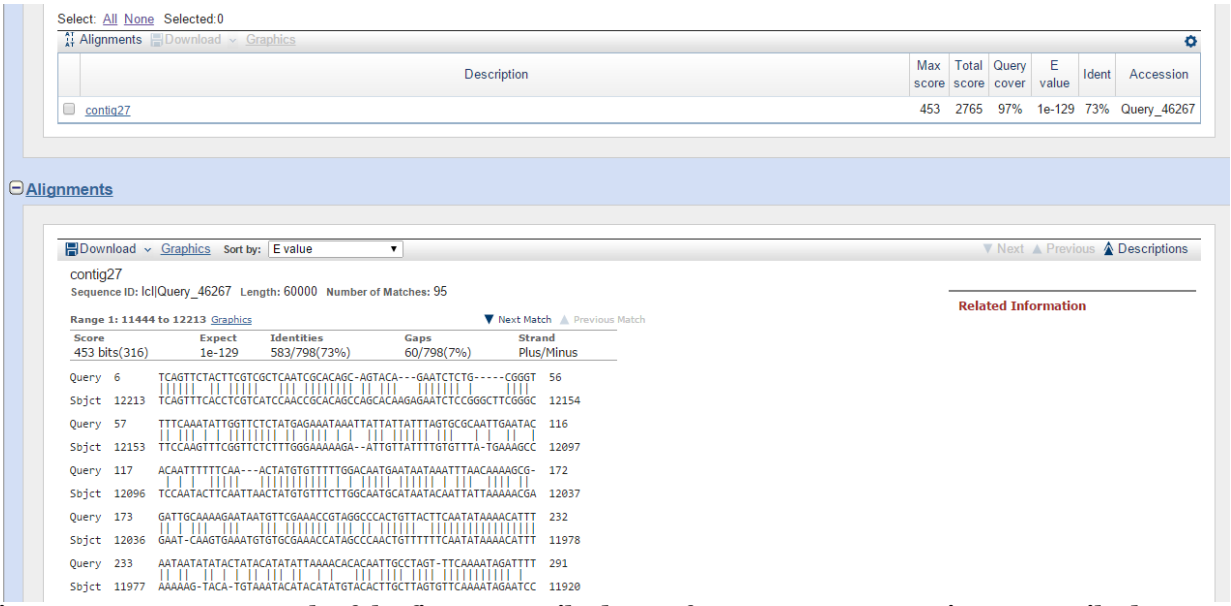


Figure 27: A BLASTn search of the first transcribed exon for *myo-RC*. Query: First transcribed exon from Gene Record Finder, Subject: contig27.fasta. Each isoform has the same start site for transcribed exons. Five bases are missing from the beginning of this exon so the TSS in *D. elegans* is located at 12,213 or earlier

Inspection of the orthologous TSS in *D. melanogaster* reveals that this TSS is a broad TSS without data that defines a singular TSS with great resolution (Figure 28). Similarly, in *D. elegans*, the RNA-seq data fails to resolve a singular TSS (Figure 29). The BLASTn result provides a starting point to search for a TSS, but the TSS search region has to span the RNA-seq reads that extend past the 5' end of the first transcribed exon.

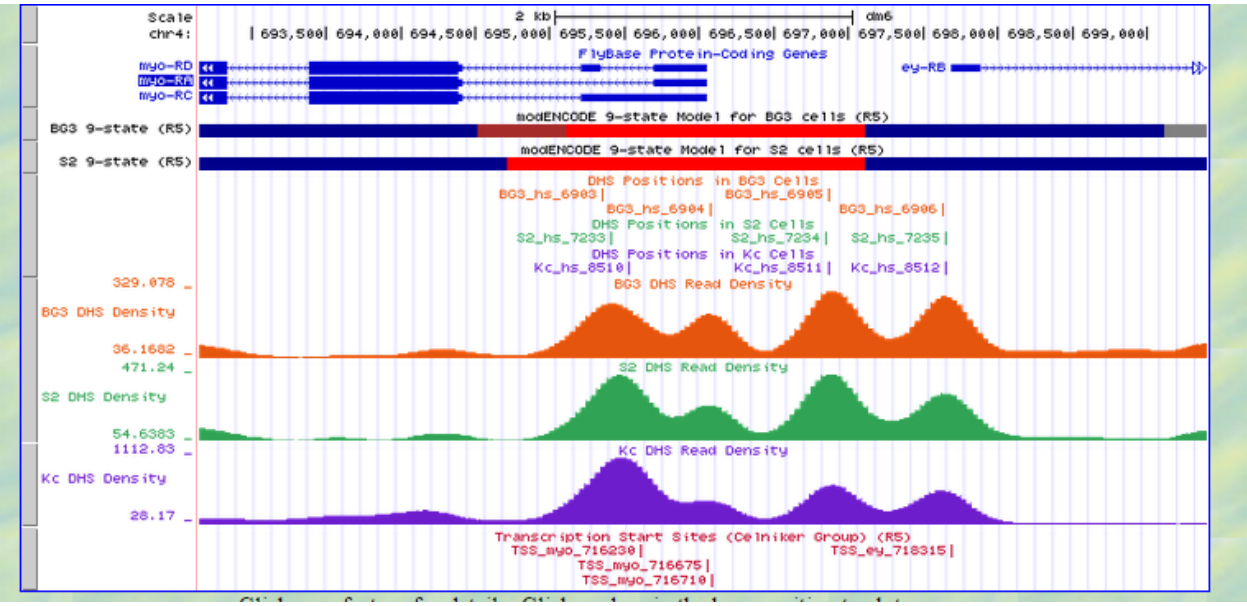


Figure 28: Evidence Tracks at the broad TSS in *D. melanogaster*. All of the evidence tracks define an area for the TSS at the beginning of the first transcribed exon

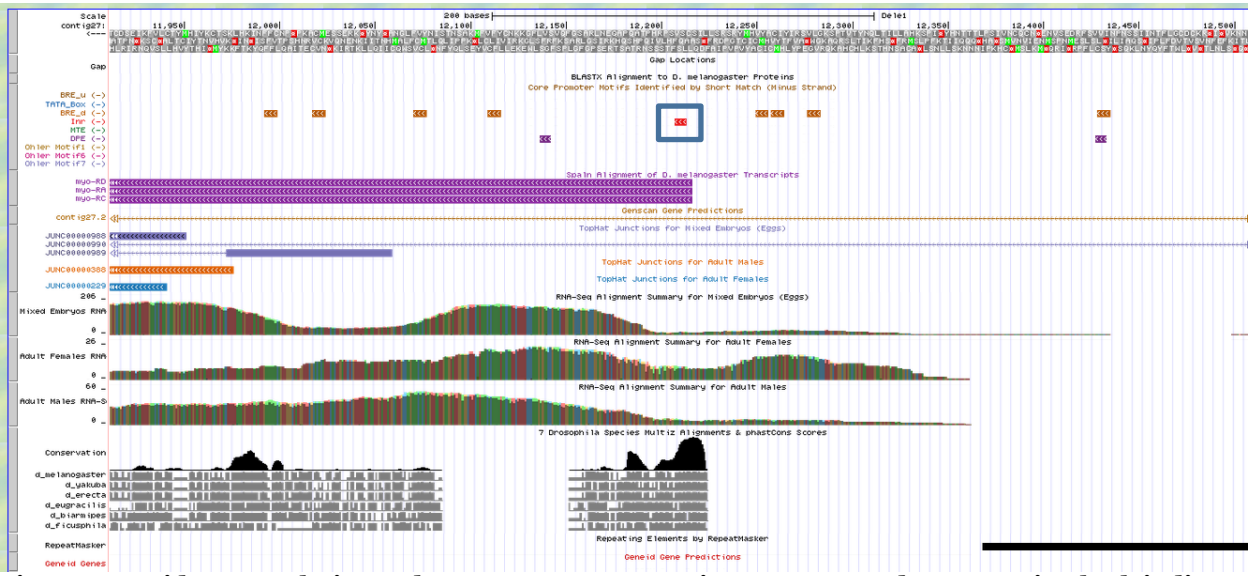


Figure 29: Evidence tracks in *D. elegans* near 12,215. Again, RNA-seq, and conservation both indicate a region for a potential TSS but not a single point. An initiator motif is indicated

Inspection of regulatory motifs in *D. elegans* near the TSS search region provides further evidence of the presence of a TSS in the region (Figure 29). In particular, an initiator motif with the sequence TCAKTY is present near position 12,213. As a core promoter motif, this sequence is usually found 2 bp upstream from the TSS. In this instance, this motif is located at position 12,213 which would indicate a TSS at 12,211.

Although no single point for a TSS was defined by all of the tracks together, a small region was defined to likely contain a TSS. However, the continuity of the RNA-seq tracks also requires an extension of the TSS search region. This report concludes that the TSS is likely located from 12,216-12,213. However, the search region is from 12,213 to 12,368.

Gene Evolution

Inspection of the BLASTp output of the *myo* gene in *D. melanogaster* shows that this protein has two conserved regions with a Transcription Growth Factor-beta (TGFB) domain located from 506 to 597 (Figure 30). The less conserved domain is also associated with TGF. To further analyze the evolution of these conserved domains, a ClustalW2 analysis was run on seven different species of *Drosophila* and four other similar species.

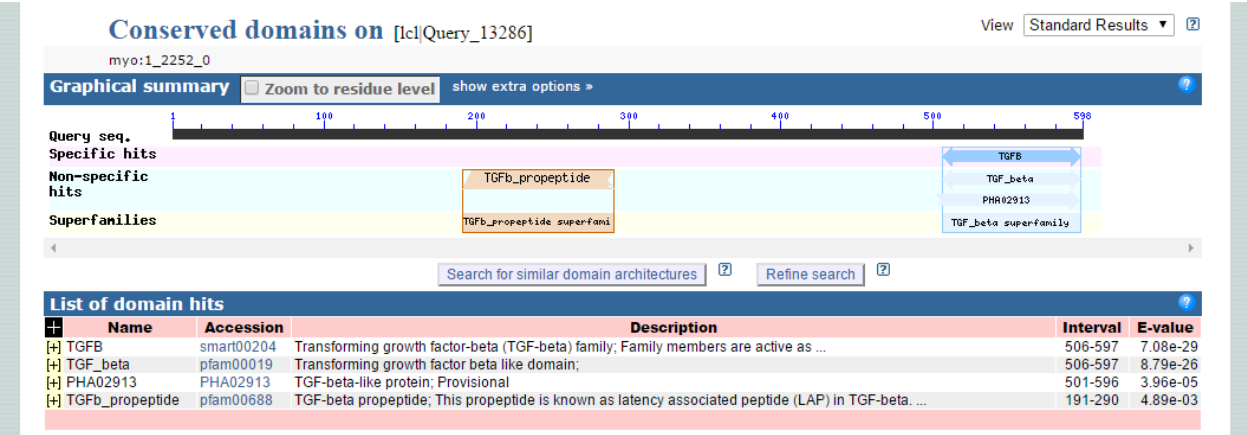


Figure 30: Summary of the conserved domains in the region

The output for this multiple sequence alignment can be seen in Figure 31. There were no matches for the first 500 bp outside of *Drosophila*, so while the two presumed conserved domains are both shown, only *Drosophila* species are shown for the first region.

A)


persimilis GVSMNRN-LINIDSILQRQLREKAKQDSLESIKMHILMRLNLKKLPNITK 233
pseudoobscura_pseudoobscura GVSMNRN-LINIDSILQRQLREKAKQDSLESIKMHILMRLNLKKLPNITK 233
sechellia SVSLYRNTLINIESMLQRQLREKAKVDSIESIKMHILMRLNLKKLPNITK 228
erecta SVSLYRNTLINIESMLQRQLREKAKVDSIESIKMHILMRLNLKKLPNITK 228
yakuba SVSMYRNTLINIESMLQRQLREKAKVDSIESIKMHILMRLNLKKLPNITK 213
grimshawi AVSMSRNSLISGDSVLQRQLREKAKQDSLESIKIHILMRLNLKRLPNITK 212
virilis GFSMIRNSLIGGDTVLQRQLREKAKQDSLESIKMHILMRLNLKKLPNITK 238
persimilis PISVPQNILEDFYKGYNASVTNTIWR---RRESTGESLSESPSVPKKHEQ 280
pseudoobscura_pseudoobscura PISVPQNILEDFYKGYNASVTNTIWR---RRESTGESLSESPSVPKKHEQ 280
sechellia PISVPQNIIDNFYRDYNASSKNTVWN---RMENIDESHLS---------I 266
erecta PISVPQNIIDNFYRDYNASSKNNVWN---RMENNEESHLS---------I 266
yakuba PISVPQNIIDNFYKNYNGSSKNKVWN---RMENNDESHLS---------I 251
grimshawi PIAVPQNILENFYKTYNASILS-----------TTKKRT--------IDQ 243
virilis PVSVPQTILEKFYKNYNASLSSSFRDHNNKPDSASTSRTGLYVPEEIVSD 288
persimilis TTNTSSDLVAGESD----------MSLSSQMQGDDANALNEFK--- 313
pseudoobscura_pseudoobscura TTNTSSDLVAGESD----------MSLSSQMQGDDANALNEFK--- 313
sechellia NDTYGGHIMTDFSD----------ESSSSQMQGDDANTVNEFL--- 299
erecta NDTYGGNIMTDFFE----------DSSSSQMQGDDANTVNEFL--- 299
yakuba NDTYGGHIMADFFD----------ESSSAQMQGDDANTVNEFL--- 284
grimshawi TLPNTENPTASLE----------NSSSEMQADDTNAFKDYLPFI 277
virilis SSETTSNVTKNIFNSQYTSAMQYTSNQNQNSSPEMQSDDPNNFKEFQ-FI 337
persimilis --LIYNTEIGIHEYQNIKQNFFMHNDDE----YESILSMISSIYVFPEQ- 356
pseudoobscura_pseudoobscura --LIYNTEIGIHEYQHIKQNFFMHNDDE----YESILSMISSIYVFPEQ- 356
sechellia --VDLNKNQAKKSDIPINT-----NDEE----YESILSMISSIYIFPEE- 337
erecta --IDLNKNQAKESDIPINT-----NDEE----YESILSMISSIYIFPEQ- 337
yakuba --IDLNKNQAKKSDIPIKT-----NDED----YESILSMISSIYIFPEK- 322
grimshawi QNFNFDEHRKFDSDYTLNGAVDAYGDNG---EQESILSMISSIYVFPEQH 324
virilis YNLGFDNHQNHKSTDSRSNDDDDNSDDNDDVEYESILSMISSIYVFPEQ- 386
persimilis LQPHVRHNRKTDVLRFKFDNSYSDISYATLHLYLRGWDWISTHQPELIEE 406
pseudoobscura_pseudoobscura LQPHVRHNRKTDVLRFKFDNSYSDISYATLHLYLRGWDWISTHQPELIEE 406
sechellia IQPHVRHNRKVDVFRFQIDSSYSDLSYATLHLYLRGWDWISAHQPGLLEE 387
erecta IQPHVRHNRKVDVFRFQIDSSYSDLSYATLHLYLRGWDWISTHQPGLIEE 387
yakuba IQPHVRHNRKVDVFRFQIDSSYTDLSYATLHLYLRGWDWISTHQPGLIEE 372
grimshawi KEPHLRHNRKSDVLRFKIDTGYSDISYVTLHLYLRGFDWIRSHQPKILEE 374
virilis --PHVRHNRKSDLLRFKFSTGYSDISHVTLHLYLRGLEWISAHQPKIIEE 434
persimilis IENQ-QSKDIVVAIHRAVRRANNTSFTHKAKMFEFRQKIPSGQGQWVNVD 455
pseudoobscura_pseudoobscura IENQ-QSKDIVVAIHRAVRRANNTSFTHKAKMFEFRQKIPSGQGQWVNVD 455
sechellia IKKQ-PRKDIVVTIHRAIRVANTTSFTPKVKMFEFRHSIPSGLGRWVAVD 436
erecta IKKQ-PCKDIVVTIHRAIRLANTTSFTPKVKMFEFRQSIPSGLGQWVTVD 436
yakuba IKKQ-PCKDIVVTIHRAIKLANSTSFTSKVKMFEFRQSIPSGLGQWVTVD 421
grimshawi FSDAKQNQDIIVALHRPIRRTNSSNYTHKAKMFEFRHKIPSGQGQWVNVD 424
virilis IADY-HNKDIVVALHRAIRRSNSTSYTHKAKIFEFRHKIPSGLGQWVNVD 483

B)

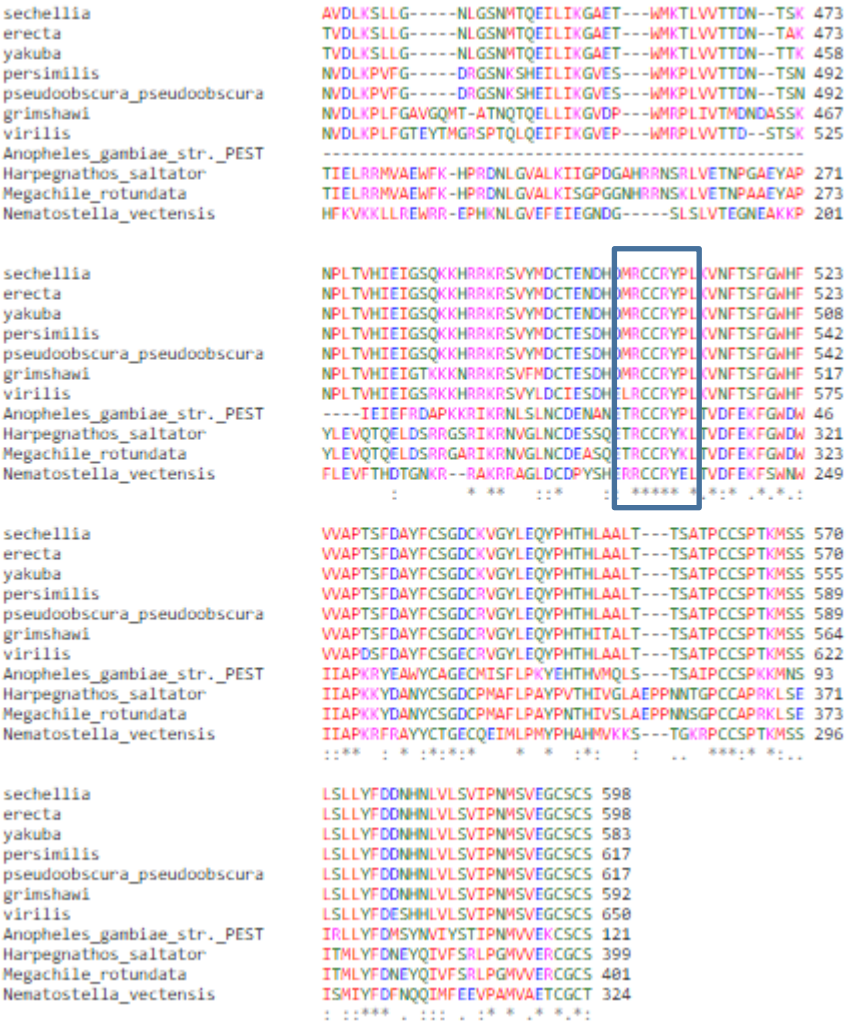


Figure 31: ClustalW2 multiple sequence alignments for *myo*. Amino acid numbers are indicated at each line of the alignment. A) An alignment involving species of *Drosophila* including the Transforming Growth Factor-beta (TGFb) propeptide region. The approximate boundaries of this region are indicated by arrows. B) An alignment involving *Drosophila* and other nearby species including the TGFb domain from 506-597. The earlier domain was not found in species outside of *Drosophila*. A perfect 5bp conserved region is indicated by the box.

About half of the first conserved domain is actually well conserved throughout *Drosophila* species. While the identified conserved domain was assessed by BLAST to span from 191 to 290, the conserved domain shown in the multiple sequence alignment is only about 62 amino acids long, from approximately 191-252 in *D. persimilis* (Figure 31A). However, this region is much more closely conserved than other non-conserved

domains of this protein, so it is possible that this region has significance, such as an active site, for the protein. However, this region is not present in other species outside of *Drosophila,* so it may only be significant in the *Drosophila* version of this protein.

The characteristics of the other conserved domain located from 506-597 in *D. elegans myo,* elucidate some aspects of this gene's evolution. First, some regions of this protein are more closely conserved than others. A particular 5 amino acid stretch is perfectly conserved even outside of *Drosophila*. This region could be important for the active site of the final protein or play another significant role in the protein's function so evolution has conserved it throughout different species. Additionally, in many regions there seem to be two distinct forms of the protein separated between *Drosophila* and other species. There were a series of evolutionary events in the dissociation of *Drosophila* into its own group of species that result in a different form of the protein. However, some of these regions are well conserved within either *Drosophila* or the other species, indicating that these regions may also play a significant role in protein function.

Repeats

In general, contig27 contains more repetitive elements than the corresponding orthologous region in *D. melanogaster*. In contig27, 32.5% of the total bases in the region are contained in repetitive regions. In contrast, only 13.28% of the bases in the orthologous region in *D. melanogaster* are contained in repeats. Further, there are 7 repeats that are longer than 500 bp in contig27, while the orthologous region only has one such repeat. A summary of the long repeats in contig27 can be seen in Table 6 and an overall summary of the repeats in contig27 can be seen in Table 7.

Start	Stop	Length	Repeat

16369	17154	785	rnd-5_family-2544
39702	41626	1924	rnd-5_family-237
44676	45248	572	rnd-5_family-2363
45823	46563	740	rnd-5_family-2544
45832	46638	806	rnd-2_family-33
58926	59702	776	rnd-5_family-237
58970	59842	872	rnd-5_family-2544

Table 6: All repeats longer than 500 bp in contig27

	number of elements*	length occupied	percentage of sequence
SINEs:	0	0 bp	0.00 %
ALUs	0	0 bp	0.00 %
MIRs	0	0 bp	0.00 %
LINEs:	8	1450 bp	2.42 %
LINE1	0	0 bp	0.00 %
LINE2	1	89 bp	0.15 %
L3/CR1	0	0 bp	0.00 %
LTR elements:	0	0 bp	0.00 %
ERVL	0	0 bp	0.00 %
ERVL-MaLRs	0	0 bp	0.00 %
ERV_classI	0	0 bp	0.00 %
ERV_classII	0	0 bp	0.00 %
DNA elements:	11	2045 bp	3.41 %
hAT-Charlie	0	0 bp	0.00 %
TcMar-Tigger	0	0 bp	0.00 %
Unclassified:	103	16191 bp	26.98 %
Total interspersed repeats:		19686 bp	32.81 %
Small RNA:	0	0 bp	0.00 %
Satellites:	0	0 bp	0.00 %
Simple repeats:	0	0 bp	0.00 %
Low complexity:	0	0 bp	0.00 %

Table 7: Summary of all repeats in contig27. 33% of all bases in contig27 are contained in repeats

Synteny

To examine synteny between contig 27 and the orthologous region of *D. melanogaster,* the orientation and order of the two genes in this region were observed. In the orthologous region in *D. melanogaster*, *myo* is a negative strand gene and upstream from the start site of *myo*, *ey* is a positive strand gene. These genes are closer

together in *D. melanogaster,* likely due to the difference in repeat density between contig27 and the orthologous region. A syntenic comparison of these two regions can be seen in Figure 32.

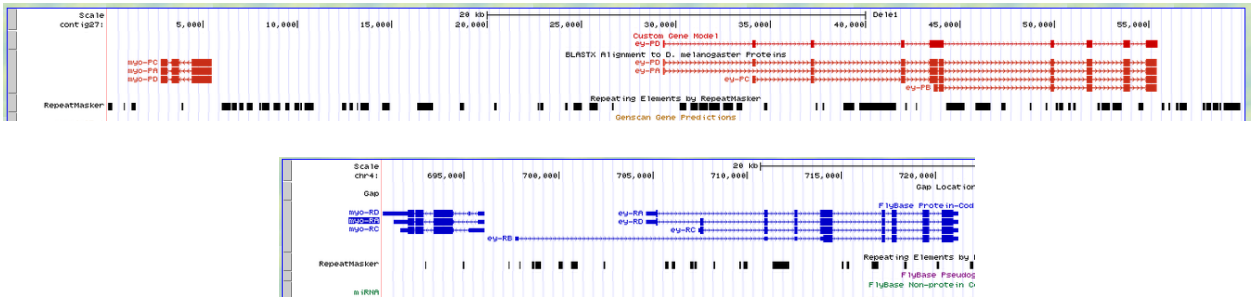


Figure 32: Syntenic comparison between contig27 of *D. elegans* (top) and the orthologous region of *D. melanogaster* (bottom).The bottom track of both regions is the RepeatMasker track. The same scale is used for each region.

Discussion

This 60,000 bp region in *D. elegans* contained two genes: *myo* and *ey*. These two genes were annotated by integrating information in the UCSC Genome Browser, BLAST, and ClustalW2. Further, the transcription start site and search region were annotated for *myo* and the evolution of *myo* was observed by multiple sequence alignment. This report discusses the methods of annotation and also examines the characteristics of repeats throughout the region and synteny with the orthologous region in *D. melanogaster*.

The findings from this report were all gathered using the protocol given by the Genome Institute Partnership. In this region in particular, annotation was often done with much contention from conflicting data, but unique conclusions were all well supported. Thus, the findings in this report are thought to be all appropriate conclusions that are not likely to be disputed.