Annotation of *Drosophila grimshawi* Contig 10

Matthew Kwong

4/20/2010

Overview:

Annotation projects involving multiple species of the same genus are important for understanding genomic evolution, adaptation, and conserved function. The *Drosophila* 12 Genomes Consortium paper titled, "Evolution of genes and genomes on the *Drosophila* phylogeny" demonstrates the wealth of knowledge that can be drawn from comparing annotated genomes.[1] This paper is a continuation of the *Drosophila* annotation project and discusses the work in annotating contig 10 of *Drosophila grimshawi*. The contig spans 28 kb in the *D. grimshawi* chromosome four. The initial *Genscan*[2] prediction calls for two features. The identity of Feature One is determined to be the beginning portion of the gene *eyeless* (with six unique exons) and contains the three isoforms *ey-PA, ey-PC,* and *ey-PD* (refer to Tables 2, 3, and 4 for specifics on exon number, location, and length). *ClustalW*[4] analysis shows that the isoform *ey-PD* shares close alignment between the species *D. grimshawi, D. melanogaster,* and *D. virilis*. An additional *ClustalW* alignment aligning *D. grimshawi* with the Body Louse, Red Flour Beetle, African Clawed Frog, and Human (species names listed on Table 5) shows evolutionary differentiation of *ey* orthologues. The *ClustalW* alignment on the 5' untranslated region (UTR) of *ey-PA* shows no conservation. A Bl2seq[6] analysis shows that all exons of *ey* for each isoform in contig 10 are accounted for. Further analysis of the *Genscan* predicted feature two and unaligned exon nets no results that indicate a gene or coding exon. There is only a 6.3% occurance of repeats in contig 10 according to *RepeatMasker*[5], and only one Line/LOA larger than 0.5 kb (0.892 kb in length). The *ey* gene is syntenic between the *D. grimshawi* and *D. melanogaster* species with the genes *myoglianin* and *bt* flanking the downstream and upstream regions respectively. Figure 1 shows the final gene map of contig 10 and includes the exons present in each of the three isoforms and the location of the 0.892 kb repeat. Table 1 shows information about the *eyeless* gene with respect to contig 10. The *eyeless* gene is the master regulator for eye formation in all *Drosophila* species and homologs exist in all animal species examined.

Feature # and Feature location	Relative orientation determined by *Genscan*	Number of unique exons in contig	Feature name and isoforms present	Flybase ID/ Gene accession #
1 (17645- N/A)	negative	6	*eyeless* *ey-PA* *ey-PC* *ey-PD*	FBpp0088300/ Hmm1732033

Table 1: Final annotation result of *eyeless* gene with respect to contig 10.

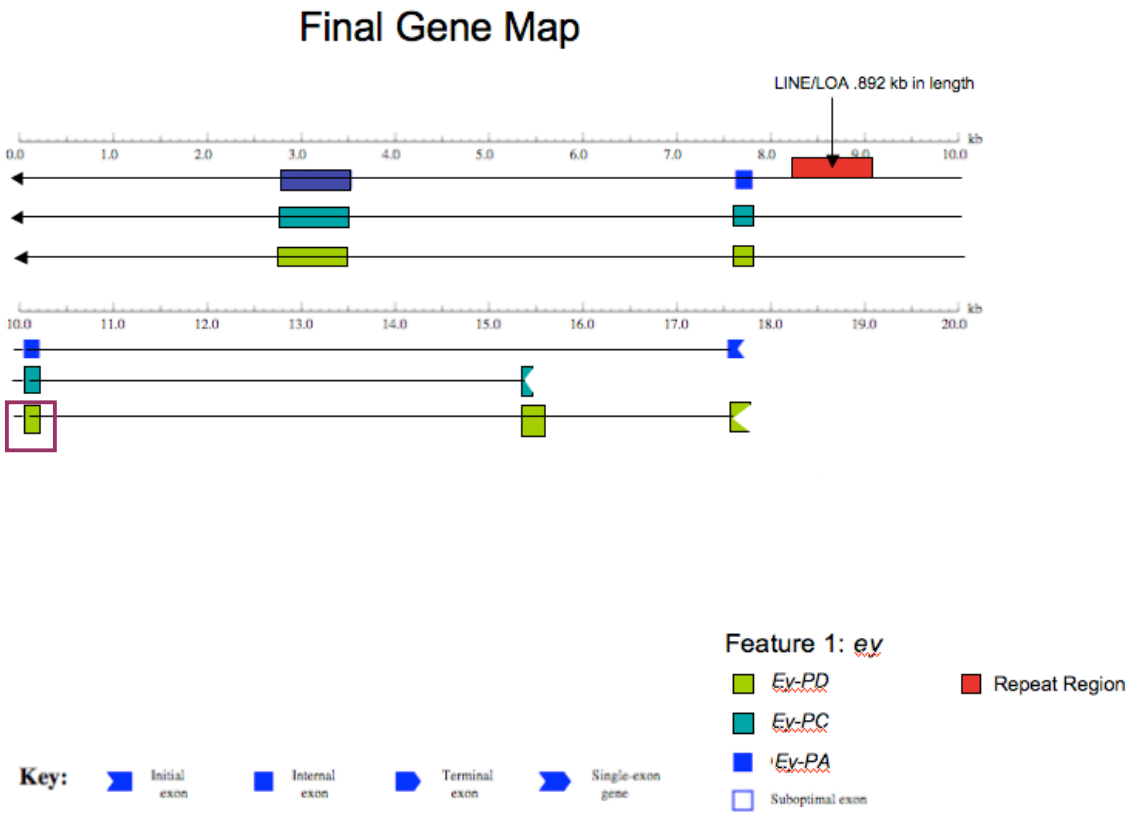


Figure 1: The Final Gene Map shows the locations of the exons for each isoform of the *ey* gene. The arrows indicate the relative orientation of the gene and connect the exons of each isoform that make up the gene. Isoform *ey-PA* is in blue, *ey-PC* is in teal, and *ey-PD* is in light green. The repeat region larger than 0.5kb is shown in red. The purple box demonstrates how the third exon for *ey-PD* is the second for the other two isoforms.

Gene:

Initial *Genscan* Prediction:

Genscan[7], a computational *ab initio* gene finder, predicts that contig 10 contains two features, both oriented in the negative direction (Figure 2). It predicts that feature one begins in the 17 kb region and spans beyond contig 10 in the negative direction. *Genscan* predicts this feature to have seven exons. It predicts that feature two ends in the 19 kb region and spans beyond contig 10 in the positive direction. *Genscan* predicts this feature to have two exons.

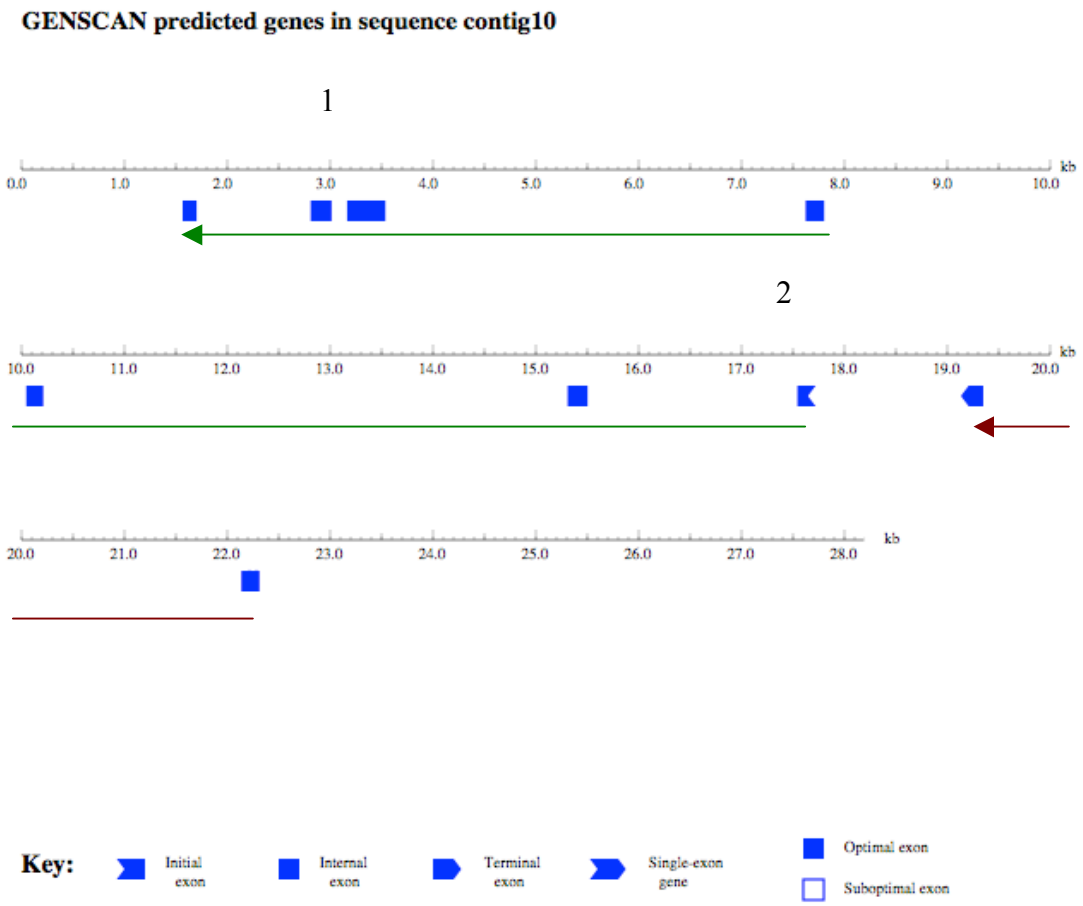


Fig. 2: Initial *Genscan* predictions predict two features as two independent genes oriented in the negative direction. Feature 1 spans the length and beyond the head of the green arrow in the negative direction. Feature 2 spans the length of the red arrow and extends beyond the back of the arrow in the positive direction (the arrowheads indicate orientation).

Initial Gene Browser:

The initial gene browser, which is available at gander.wustl.edu[2], shows a Blastx alignment between *D. gimshawi* and *D. melanogaster* for the gene *eyeless (ey)*, and a closer look reveals an alignment to three isoforms of the *ey* protein: *ey-PA*, *ey-PD*, *ey-PD*. Multiple gene finders also shown in the gene browser share many locations of exon agreement with the Blastx alignment. Unlike the *Genscan* prediction, the Blastx alignment only predicts one gene and thus one feature. One should note that the repeat masking program, *RepeatMasker*, did not mask much of the squence in contig 10. This indicates that there are few repeats in the contig.

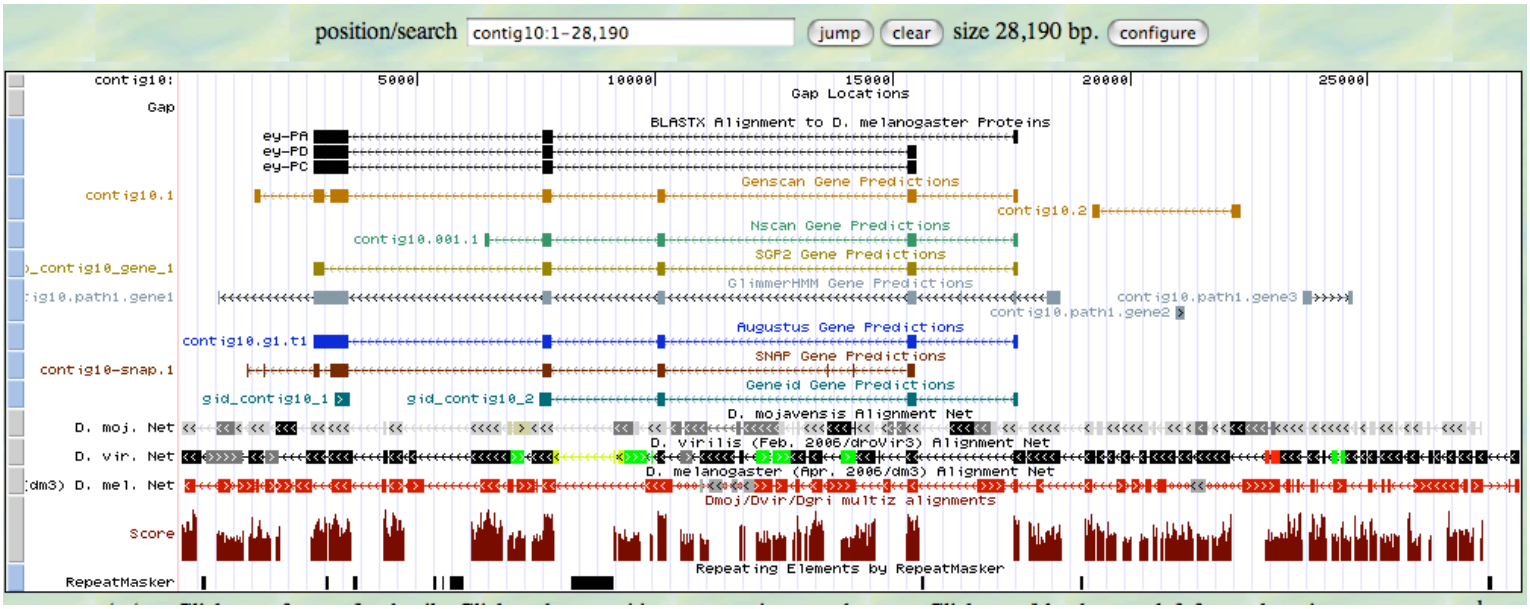


Fig. 2: Initial gene browser showing a Blastx alignment of contig 10 to *D. melanogaster*, results from various gene finders, and *RepeatMasker* results.

Feature 1- *eyeless*

Gene Record Finder, an online database detailing the transcript and polypeptide details of *D. melanogaster* proteins, shows that *ey* is on chromosome 4 in *D. melanogaster* and that there are four known isoforms of the gene: *ey-PA, ey-PB, ey-PC,* and *ey-PD* (Figure 3). As stated above, the Blastx results indicate that the isoforms *ey-PA, ey-PC, and ey-PD* exist in *D. grimshawi*. According to the Gene Record Finder transcript details (Figure 4), there are eleven possible exons for *ey* depending on the isoform. Even at first glance, it does not appear that contig 10 contains eleven unique exons. The Blastx[6] result indicates three for each isoform, and *Genscan* predicts seven exons for the feature that aligns with the Blastx result (feature 1). Thus, there is a high possibility that contig 10 contains only part of the *ey* gene. Further comparison between the genome browser and transcript details indicate that it likely contains the beginning of the *ey* gene based on orientation and because the isoforms present appear to have different exons from each other. The transcript details indicate that only the first four exons in *ey* are variable between isoforms, while the remaining seven are in all isoforms. Additional analysis descussed below allows me to conclude that only six unique exons are present in contig 10. The polypeptide details for these are shown in Figure 5.

Gene Details

FlyBase ID	FlyBase Name	Chr	Start	End	Strand
FBgn0005558	ey	4	718,315	741,787	+

mRNA Details

Select a row to display the corresponding transcript and peptide details:

FlyBase ID	FlyBase Name	Chr	Start	End	Strand
Unique	Unique	4	718,315	741,787	+
FBtr0089235	ey-RB	4	718,315	741,787	+
FBtr0089236	ey-RA	4	725,703	741,787	+
FBtr0100396	ey-RD	4	725,703	741,787	+
FBtr0100395	ey-RC	4	728,055	741,787	+

Fig 3: Gene record finder details demonstrating the location, direction and different isoform types of *ey* in *D. melanogaster*.

Transcript Details | Polypeptide Details

Options: Export Sequences for Selected Isoform to FASTA

Exon usage map:

Isoforms / ID	CG1464:1	CG1464:9	CG1464:17	CG1464:26	CG1464:2	CG1464:3	CG1464:4	CG1464:5	CG1464:6	CG1464:7	CG1464:8
Unique	X	X	X	X	X	X	X	X	X	X	X
ey-RB	X				X	X	X	X	X	X	X
ey-RA		X			X	X	X	X	X	X	X
ey-RD		X		X	X	X	X	X	X	X	X
ey-RC			X		X	X	X	X	X	X	X

Fig 4: Gene Record Finder transcript details showing eleven unique exons for *ey*.

Transcript Details | **Polypeptide Details**

Options: Export Sequences for Selected Isoform to FASTA

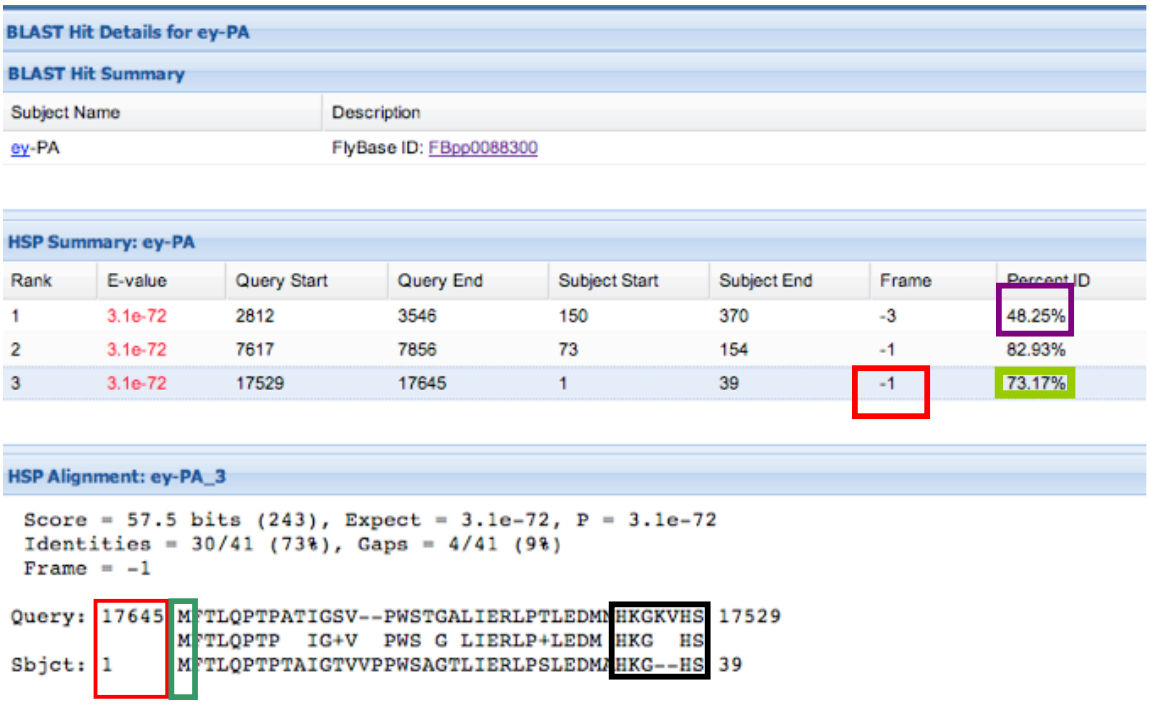
CDS usage map:

Isoforms / ID	CDS_CG1464:9_912	CDS_CG1464:26_912	CDS_CG1464:17_912	CDS_CG1464:2_912	CDS_CG1464:3_912	CDS_CG1464:4_912
Unique	X	X	X	X	X	X
ey-RB						
ey-RA	X			X	X	X
ey-RD	X	X		X	X	X
ey-RC			X	X	X	X

Figure 5: Gene Record Finder polypeptide details for the six unique exons for *ey* represented in Contig 10.

Finding the *ey* Start Site:

Looking at the Blast hit summary for isoform *ey-PA* (Figure 6), one can see that the methionine, which indicates the gene start site, exists in the negative frame. This methionine is found to align with the exon that has a 73.17% identity (in lime green) with the annotated *ey-PA* protein of *D. melanogaster*. The HSP alignment shows the methionine corresponding to the query end site, which is additional evidence that the protein is in a negative frame. The negative orientation also corresponds with the *Genscan* prediction in Figure 2, which also predicts the negative frame.



BLAST Hit Details for ey-PA

BLAST Hit Summary

Subject Name	Description
ey-PA	FlyBase ID: FBpp0088300

HSP Summary: ey-PA

Rank	E-value	Query Start	Query End	Subject Start	Subject End	Frame	Percent ID
1	3.1e-72	2812	3546	150	370	-3	48.25%
2	3.1e-72	7617	7856	73	154	-1	82.93%
3	3.1e-72	17529	17645	1	39	-1	73.17%

HSP Alignment: ey-PA_3

```
 Score = 57.5 bits (243), Expect = 3.1e-72, P = 3.1e-72
 Identities = 30/41 (73%), Gaps = 4/41 (9%)
 Frame = -1

Query: 17645 MFTLQPTPATIGSV--PWSTGALIERLPTLEDMHKGKVHS 17529
           MFTLQPTP  IG+V  PWS G LIERLP+LEDM HKG  HS
Sbjct: 1     MFTLQPTPTAIGTVVPPWSAGTLIERLPSLEDMAHKG--HS 39
```

Figure 6: The Blast hit summary shows the negative orientation (in red) of *ey-PA* and highlights the location of the subject start site (in green) with respect to the query. The query sequence represents the location in the contig a distance from the arbitrary start site and the subject represents the amino acid number for the *ey-PA* protein.

One feature in Figure 6 to note is the 48.25% percent ID of the exon with query sequence bp 2812-3546 (in purple). A similarly low percent ID was also the case in the Blast results for the first and last exons in the isoform *ey-PC* and *ey-PD*. The likely explanation for the low percent identity of select exons between *D. grimshawi* and *D. melanogaster* is the evolutionary divergence of the two species.

A close up view of the *ey-PA* starting region shows that the methionine is in the location predicted by both the Blastx alignment and the gene finders (Figure 7). A similar method of looking at the Blast hit summary and then comparing the Blastx alignment result with the gene finder predictions was used to find the start sites for the other isoforms *ey-PC* and *ey-PD*. These results are summarized in Tables 3 and 4, respectively. Since all three

isoforms have a negative frame, it is appropriate to reverse the gene browser in order to place the gene start sites in the left-hand side (Figure 8).

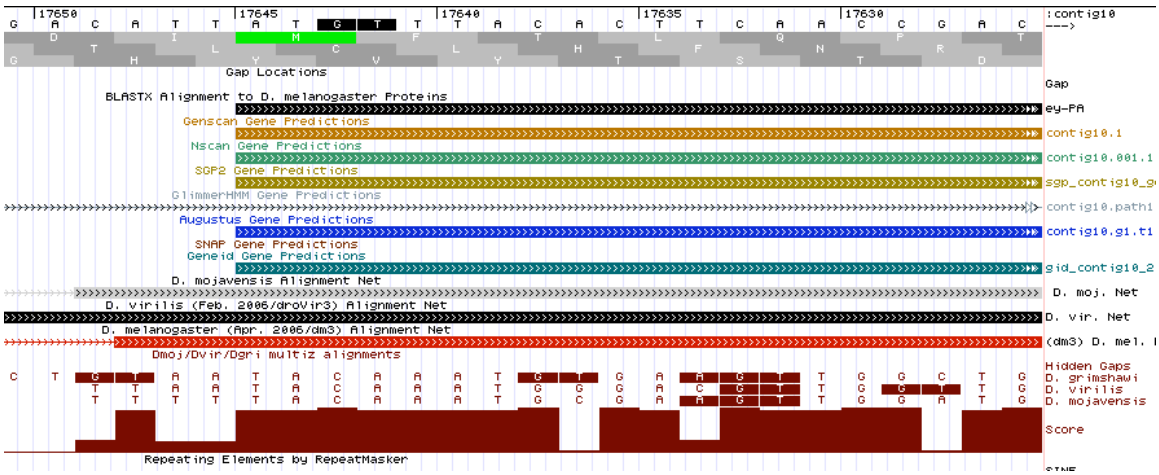


Figure 7: A close up on the genome browser showing the methionine start site for *ey-PA* and the agreement between the Blastx alignment and gene finders.

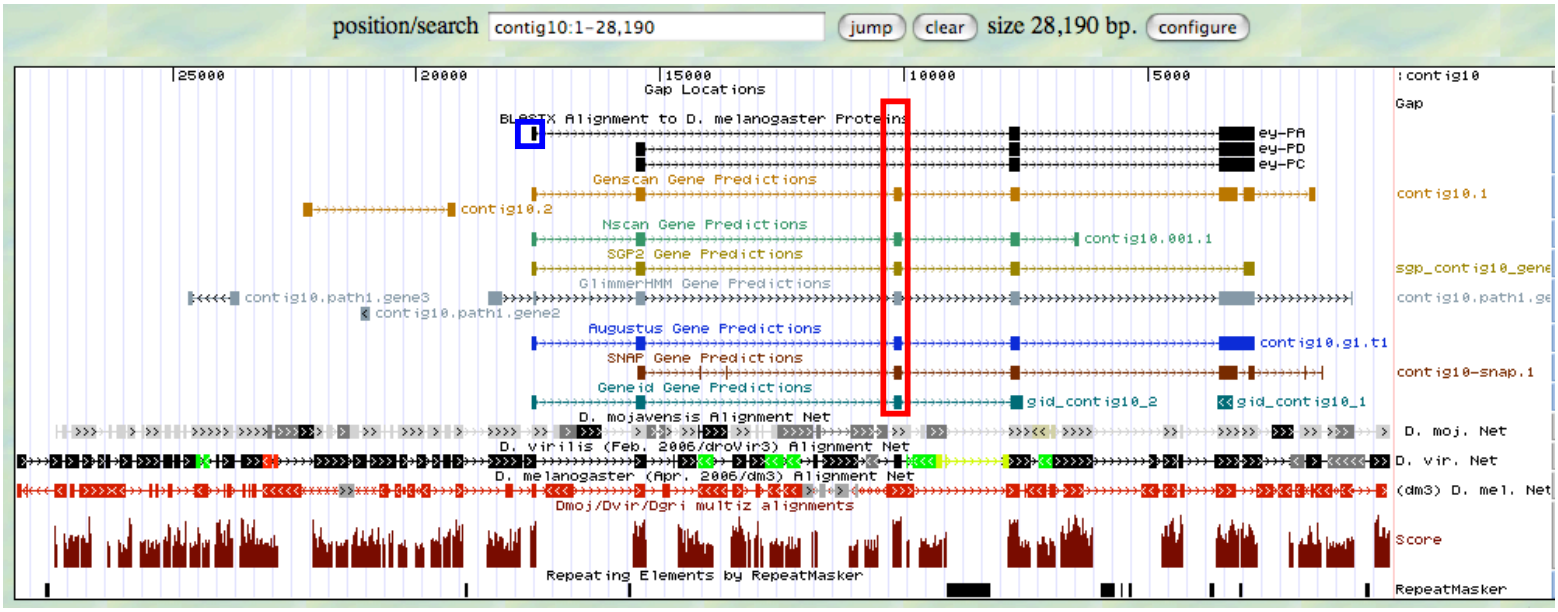


Fig 8: Reversed gene browser showing the *ey* start sites on the left-hand side and moving in a negative frame. The blue box represents the first exon for *ey-PA* and the red box indicates Exon 2. Both are referred to in subsequent sections.

Donor and Acceptor Sites:

The donor and acceptor sites of each exon known to be in the isoform of *ey* in question (from the Gene Record Finder shown in Figure 5) were identified to construct a gene

model of *ey* in contig 10. These sites indicate the beginning and end of the intronic regions that separate the exons. Donor sites have a characteristic GT nucleotide sequence and acceptor sites have a characteristic AG sequence. The criteria is that the corresonding donor and acceptor sites have a phase that add up to three or zero. This is the way that the amino acid sequences are conserved from one exon to the next. There are often many possible candidates for donor sites (and sometimes acceptor sites as well). One considers all donor sites until the nearest stop codon for that frame. The best candidate is the one that matches the acceptor site with the proper phase and is closest to the end of the exon boundary determined by homology.

Consider the exon 1 donating site for isoform *ey-PA* (blue box in Figure 8 shows this exon). Looking at Figure 6, one can see that this exon, whose query sequence is bp 17645-17529, has a final exon peptide sequence of HKGKVHS. Looking at a close-up window of the donating site (Figure 9), one can see the peptide sequence corresponding with the final exon (blue box). This observation demonstrates that the phase shift is 1. From Figure 6, one can see that the HKG and HS are conserved amino acids (black box). Thus, it is likely that either of these regions can signify the end of exon 1. Figure 9 shows that the BlastX alignment extends further, indicating that the alignment suggests the HS amino acids signify the end of exon 1. On the other hand, the gene finders predict the HKG amino acids to signify the end of exon 1 because their prediction of the exonic region ends with the K amino acid. In this case, it is likely that the gene finders predicted correctly because the only possible donating sites are within the region that the Blastx alignment predicted as an exon. The two "GT" nucleotides begin at the bp 17542 (red box) region and bp 17536 (green box) region. The red box GT has a phase of 1 and the green box GT has a phase of 0. Both these locations are candidates for being donor sites. One can identify the likely donor site for exon one when one has a strong candidate for the acceptor site of exon 2.

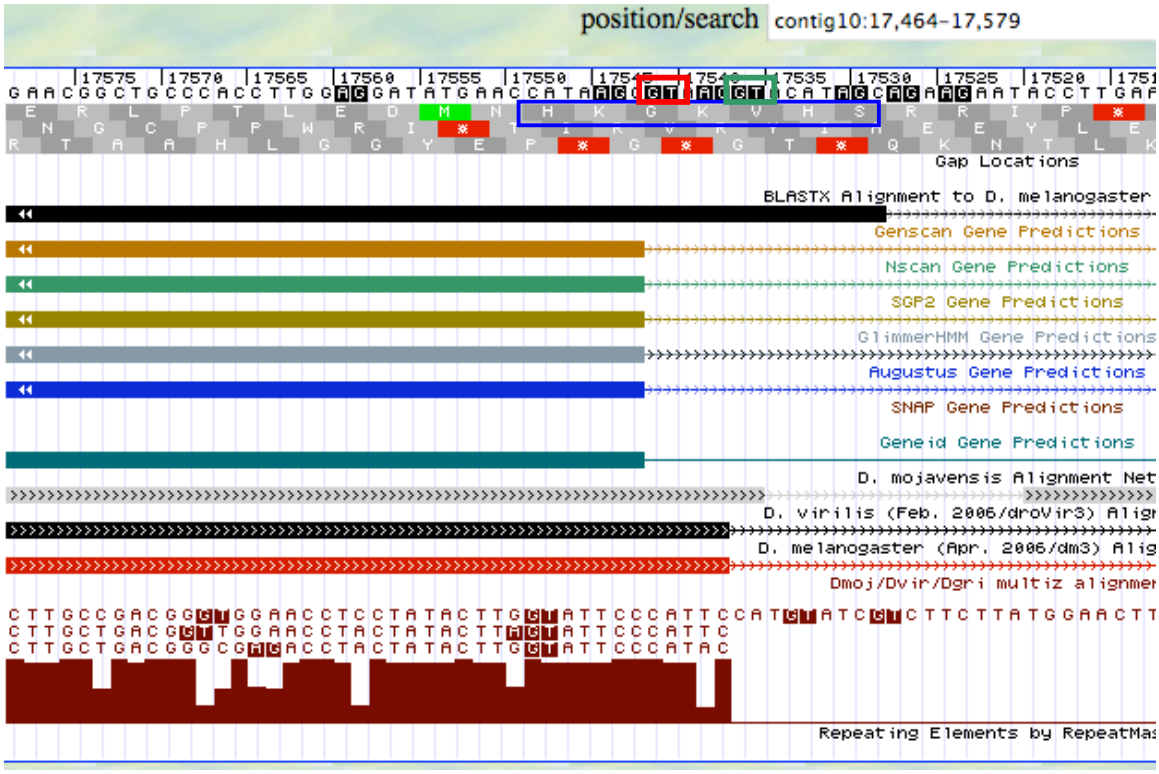


Figure 9: Exon 1 donating site that shows the possible donators at locations bp 17542 (red box) and bp 17537 (green box).

Exon 2:

Gene Record Finder indicates that the *ey-PA, ey-PC,* and *ey-PD* isoforms all have an exon (CDS_CG1464:2_912) not identified by the Blastx alignment (Figure 10 blue box). This exon is predicted by the gene finders and its location is shown by the red box on Figure 8. Figure 10 shows the peptide sequence for this exon and its location in *D. melanogaster*. A closer look at the tBlastn[6] text output shows that, based on percent identity and location in the contig, this peptide sequence is present in *D. grimshawi* (Figure 11). This is strong evidence that this exon exists in these isoforms of *ey*. This exon is the second exon in isoforms *ey-PA* and *ey-PC*, but is the third exon in *ey-PD* as shown by the purple box in Figure 1. This paper will view this exon from the isoform *ey-PA* viewpoint and be referred to as exon 2.

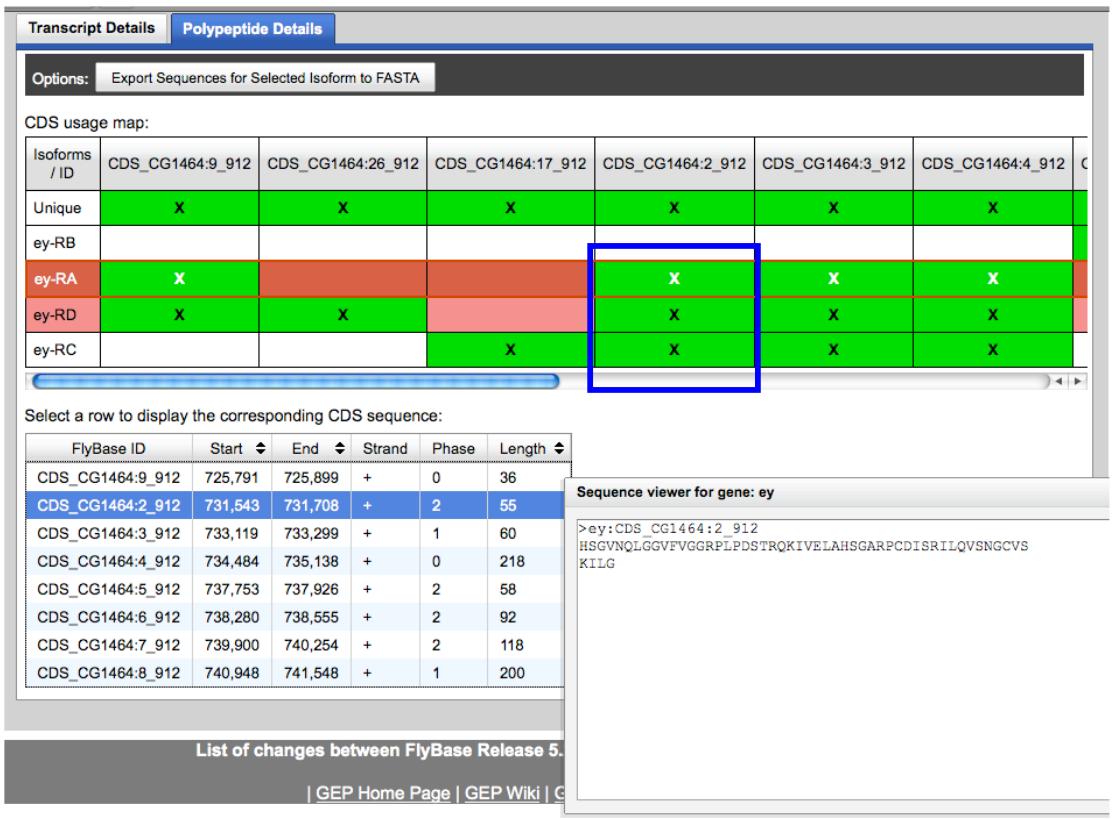


Figure 10: Exon 2 gene record finder information showing the exon identity (blue box) in various isoforms and the polypeptide sequence.

```
 Score =  109 bits (273),  Expect = 3e-28, Method: Compositional matrix adjust.
 Identities = 54/54 (100%), Positives = 54/54 (100%), Gaps = 0/54 (0%)
 Frame = -3

Query  1      HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILG  54
              HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILG
Sbjct  10209  HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILG  10048
```

Figure 11: tBlastn text output shows *D. grimshawi* (subject sequence) has a polypeptide coding sequence corresponding to exon 2 of *D. melanogaster* (query sequence).

The acceptor site in exon 2 is apparent as shown by the red box in Figure 12. The AG nucleotide begins at the bp 10213 location, which is just outside of the entronic region predicted by *Genscan*. It is the only strong candidate in the region because other AG sites are either within the predicted entronic region or a fair number of base pairs away. The bp 10213 AG acceptor site has a phase of 2, meaning that it matches with the exon 1 GT donor site at bp 17542 that has a phase of 1 (Figure 9, red box). Furthermore, as shown in Figure 11, there is reasonable conservation in the beginning of exon 2 (reading frame 3). Twelve of the first thirteen amino acid sequences are conserved between *D. melanogaster* and *D. grimshawi*. This conservation implies that the exon likely begins at an AG acceptor site near this area of conservation. Thus, one can conclude that the exon 1 donor site is at bp 17542 and the exon 2 acceptor site is at bp 10213 for *ey-PA*.

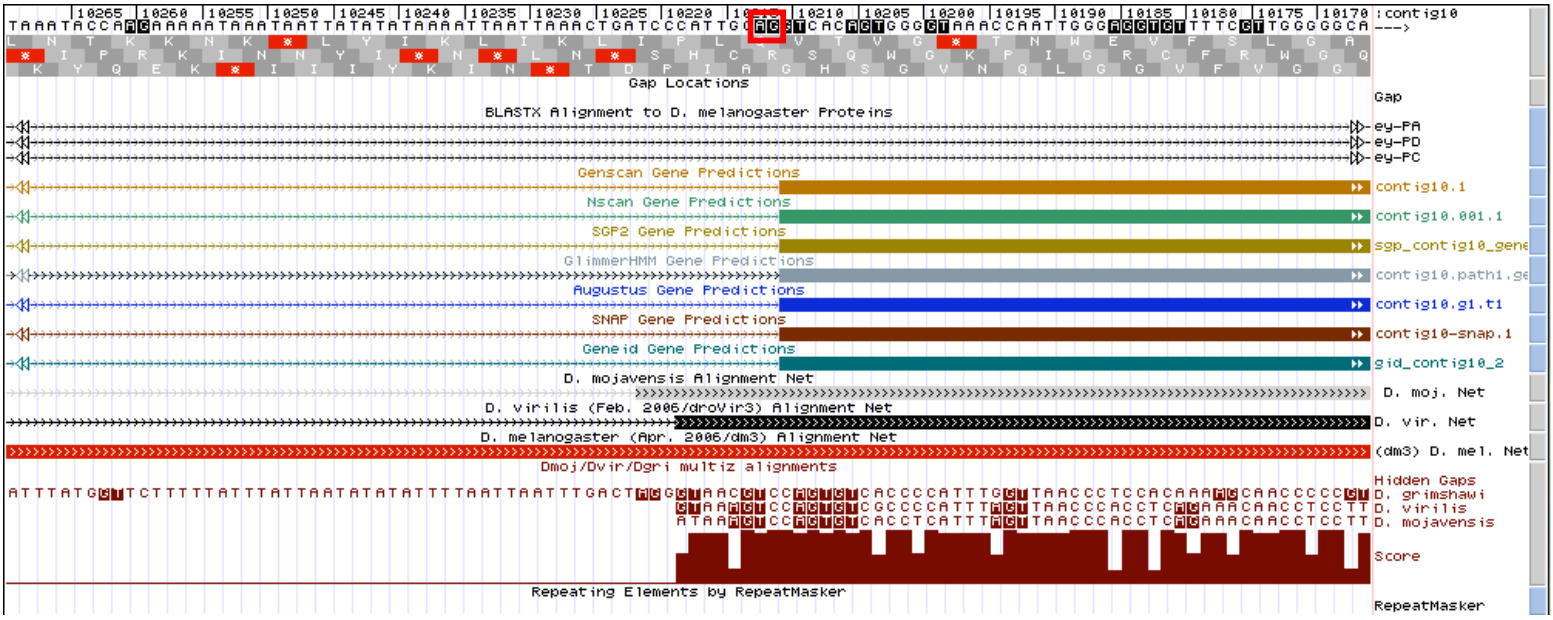


Figure 12: exon 2 acceptor +2 phase AG

After finding the donor site and the methionine for exon 1, one can hypothesize a query exon length for exon 1. The hypothesized length and query location for exon 1 would be bp 17645-17543 because the methionine location is bp 17645 and the entronic region ends at bp 17543 because the donor site occurs at bp 17542. The same technique for finding the donor and acceptor sites is used for every consecutive exon pair of each isoform. The gene model for an isoform is created when one can hypothesize the length and query location for each exon in the isoform in addition to the isoform start and end sites. The remaining information gathered to create the gene model is compiled in Table 2 (isoform *ey-PA*), Table 3 (isoform *ey-PD*), and Table 4 (isoform *ey-PC)*.

Gene Model Tables

Note:

-Tables 2,3,4 show the combined Blastx alignment and gene predictor data.

-All isoforms accepted by gene model checker.

-In these tables, strong acceptor and donator sites indicate the sites used in the gene model, while weak sites indicate the candidates not used in the gene model.

Isoform 1: *ey-PA* Fly-Base: FBpp0088300

Exon #	Exon ID	Query	Subject	Frame	Acceptor/strength/phase/location	Donor/strength/phase/location
1	CDS_CG1464:9_912	17645-17543	1-36	-1	Methionine/strong/17645	GT/strong/+1/17542 GT/weak/+2/17537
2	CDS_CG1464:2_912	10211-10046	37-92	-3	AG/strong/+2/10213	GT/strong/+2/10045
3	CDS_CG1464:3_912	7803-7623	93-153	-1	AG/strong/+1/7840 AG/weak/0/7835 AG/weak/0/7805	GT/strong/0/7622 GT/weak/+1/7618 GT/weak/+2/7587
4	CDS_CG1464:4_912	3537-2811	154-372	-3	AG/strong/0/3539	GT/strong/+1/2810 GT/strong/+2/2806

Table 2

Table 2 Notes:

-Exon 1 donor site misread by Blastx—gene model predictions net the logical result

-Exon 2 not shown by Blastx, but predicted by other gene predictors and present in *D. melanogaster*. The Blastx text output also shows that this exon should exist in this isoform.

-Exon 3 start site mispredicted by Blastx (at bp7855), predicted correctly by other gene predictors (at bp 7802)

-Exon 4 start site mispredicted by Blastx (at bp 3545), predicted correctly by other gene predictors (at bp 3536)

-Exon 4—gene model checker verifies donor site as bp 2811, but Blastx labels it as bp 2812

Isoform 2: ey-PD Fly-Base: FBpp0099810

Exon #	Exon ID	Query	Subject	Frame	Acceptor/strength /phase/location	Donor/strength/phase /location
1	CDS_CG 1464:9_9 12	17645-17543	1-36	-1	Methionine/stron g/ 17645	GT/strong/+1/17542 GT/weak/+2/17537
2	CDS_CG 1464:26_ 912	15503-15309	37- 97	-3	AG/strong/+2/155 05	GT/strong/+1/15308 GT/weak/+2/15304 GT/weak/+1/15295
3	CDS_CG 1464:2_9 12	10211-10046	98-153	-3	AG/strong/+2/102 13	GT/strong/+2/10045
4	CDS_CG 1464:3_9 12	7803-7623	154-214	-1	AG/strong/+1/784 0 AG/weak/0/7835 AG/weak/0/7805	GT/strong/0/7622 GT/weak/+1/7618 GT/weak/+2/7587
5	CDS_CG 1464:4_9 12	3537-2811	215-433	-3	AG/strong/0/3539	GT/strong/+1/2810 GT/strong/+2/2806

Table 3

Table 3 Notes:

-Blastx did not identify exon 1 (CDS_CG1464:9_912), but gene record finder states that the exon is present in this isoform of *D. melanogaster*. Furthermore, a methionine was not found at the start site of exon 2, which provides evidence that exon 1 indeed belongs in this isoform.

-Exon 3 not predicted by Blastx, but predicted by other gene predictors and present in *D. melanogaster*. The Blastx text output also shows that this exon should exist in this isoform.

-Exon 4 start site mispredicted by Blastx (at bp7855), predicted correctly by other gene predictors (at bp 7802)

-Exon 5 start site mispredicted by Blastx (at bp 3545), predicted correctly by other gene predictors (at bp 3536)

-Exon 5—gene model checker verifies donor site as bp 2811, but Blastx labels it as bp 2812

Isoform 3: ey-PC Fly-Base: FBpp0099809

Exon ID	Exon ID	Query	Subject	Frame	Acceptor/strength/phase/location	Donor/strength/phase/location
1	CDS_CG1464:17_912	15489-15309	1-55	-3	Methionine/strong/15489	GT/weak/+1/15306
2	CDS_CG1464:2_912	10211-10046	56-111	-3	AG/strong/+2/10213	GT/strong/+2/10045
3	CDS_CG1464:3_912	7803-7623	112-172	-1	AG/strong/+1/7840 AG/weak/0/7835 AG/weak/0/7805	GT/strong/0/7622 GT/weak/+1/7618 GT/weak/+2/7587
4	CDS_CG1464:4_912	3537-2811	173-391	-3	AG/strong/0/3539	GT/strong/+1/2810 GT/strong/+2/2806

Table 4

Table 4 notes:

-Exon 1- Blastx predicted methionine location correctly. Since gene finders overlap multiple isoforms in their prediction, cannot identify start site for isoform 3 from gene finder data.

-Exon 2 not shown by Blastx, but predicted by other gene predictors and present in *D. melanogaster*. The Blastx text output also shows that this exon should exist in this isoform.

-Exon 3 start site mispredicted by Blastx (at bp7855), predicted correctly by other gene predictors (at bp 7802)

-Exon 4 start site mispredicted by Blastx (at bp 3545), predicted correctly by other gene predictors (at bp 3536)

-Exon 4—gene model checker verifies donor site as bp 2811, but Blastx labels it as bp 2812

Final Exon in Contig 10

The final exon in contig 10 (exon 4 for isoforms *ey-PA, ey-PC* and exon 5 for isoform *ey-PD*) likely does not represent the end of the gene since *ey* has eleven unique exons with either eight or nine being expressed (depending on the isoform). Thus, the *ey* gene likely extends past contig 10. The final exon, as shown in Figure 13, contains no stop codons and two possible splice sites, meaning there is no evidence that translation ends with this exon. However, since there is no information past contig 10, this gene segment may or may not be a pseudogene. There is no evidence suggesting that this is a pseudogene, but the possiblity cannot be ruled out without a complete gene model and evidence of expression. Figure 14 shows this final exon's potential donor sites to the next exon. The two candidates are the GT nucleotide sites at bp 2810 with a phase of 1 (Figure 14, red box) and at bp 2806 with a phase of 2 (Figure 14, green box).

HSP Summary: ey-PA

Rank	E-value	Query Start	Query End	Subject Start	Subject End	Frame	Percent ID
1	3.1e-72	2812	3546	150	370	-3	48.25%
2	3.1e-72	7617	7856	73	154	-1	82.93%
3	3.1e-72	17529	17645	1	39	-1	73.17%

HSP Alignment: ey-PA_1

```
 Score = 149.1 bits (669), Expect = 3.1e-72, P = 3.1e-72
 Identities = 124/257 (48%), Gaps = 48/257 (18%)
 Frame = -3

Query: 3546 ILKVSSINRVLRNLAAQKEQQSSVGSASSSNSNPGANS-KATGSGTAAGTVTGTANGNNI 3370
            I  VSSINRVLRNLAAQKEQQS+ GS SSS S  G NS  A  S +  G V+  A+G
Sbjct: 150  IPSVSSINRVLRNLAAQKEQQST-GSGSSSTS-AG-NSISAKVSVSIGGNVSNVASG--- 203

Query: 3369 GSSNGTGLGNAGNELIQTATPLNSSESGGASNSGEGSEQESIYEKLRLLNTQHVAA---L 3199
              S GT L ++  +L+QTATPLNSSESGGASNSGEGSEQE+IYEKLRLLNTQH A    L
Sbjct: 204  --SRGT-LSSS-TDLMQTATPLNSSESGGASNSGEGSEQEAIYEKLRLLNTQHAAGPGPL 259

Query: 3198 DVA-TAPQTMTS-SHF---SPHP-L-HASHCXXXXXXXXXXXXXXXXXXXXXCWPPRH 3040
            + A  AP    S +H    S HP L H +H                       WPPRH
Sbjct: 260  EPARAAPLVGQSPNHLGTRSSHPQLVHGNH---------------QALQQHQQQSWPPRH 304

Query: 3039 YSTGSWYAAPLNGSSNDISASPGVLSVAGGYGNPGAALAPAPAHPLTPPTDLINIGGPSV 2860
            YS GSWY   L  S   IS++P + SV   Y + G    P+ AH L+PP D+ ++   S+
Sbjct: 305  YS-GSWYPTSL--SEIPISSAPNIASVTA-YAS-G----PSLAHSLSPPNDIESLA--SI 353

Query: 2859 -NLGNCTIAPDDVMLKK 2812
             +  NC +A +D+ LKK
Sbjct: 354  GHQRNCPVATEDIHLKK 370
```

External Links: BLAST Viewer for Dgri4_contig10

Done

Figure 13: Query sequence of final exon in Contig 10 shows no stop codons and equivalent conservation with other exons.

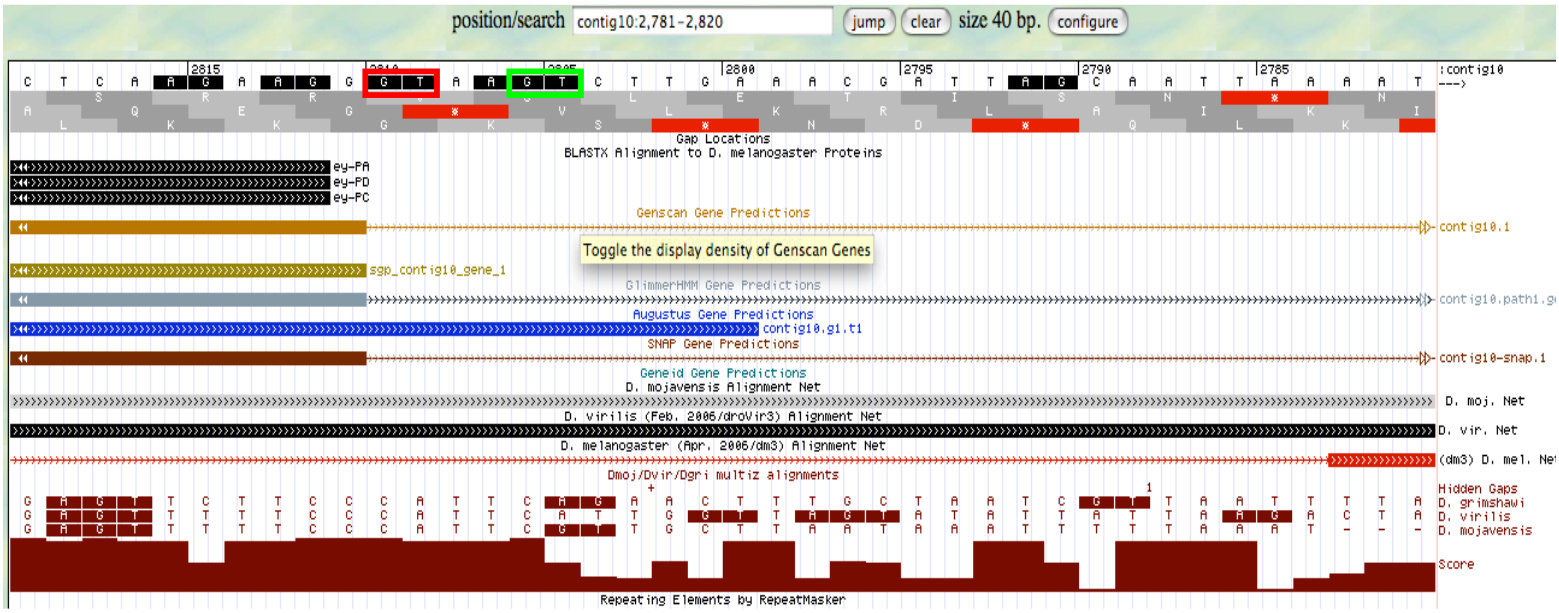


Figure 14: Final exon candidate donor sites include GT nucleotide regions at bp 2810 with a phase of 1 (red box) and at bp 2806 with a phase of 2 (green box).

***Gene Model Checker*:**

The *Gene Model Checker*[8] is a program used to ensure the proper donor and acceptor sites, in addition to the start and ending sites for each exon in an isoform. Since this contig is likely an incomplete gene, the option titled, "Missing 3' end of translated region" was chosen. Isoforms *ey-PA*, *ey-PC* and *ey-PD* underwent identical scrutiny through *Gene Model Checker*, and each met all the criteria. Figure 15a compares the submitted *D. grimshawi* gene model sequence for isoform *ey-PD* to the accepted *ey-PD* isoform sequence from *D. melanogaster* in a dot plot format. Figure 15b shows the amino acid Bl2seq alignment that the dot plot is based upon. The same colored boxed regions correspond to both figures. Isoform *ey-PD* is used as the example because it contains the greatest number of exons and serves as the best representation for all the isoforms. The dot plot shows the homology between *ey-PD* of the two species because it shows the relative amino acid positions of the isoform. The straight lines with a constant slope show areas of homology between the two species because they represent identical amino acids in identical locations. The red, brown, and pink boxes in Figures 15a and 15b indicate gaps between the lines and represent nonhomologous regions between the species. This idea is supported by the continuity of the homologous regions shown by the equal slope of each line. The equivalent slopes before and after the gaps indicate equivalent gene length in homologous regions and equivalent location and length in nonhomologous regions between each species. Based on the lengths of the solid lines and the locations of the gaps, it does not appear that the regions match exon/intron boundaries. Instead, as the Bl2seq of the coding portions of the gene alignment demonstrates, the gaps occur in regions of nonhomology. The blue circle marks the approximate end of *ey-PD* in contig 10 according to the dot plot. This makes sense because the *ey-PD* segment homologous between the two species in contig 10 is approximately 383 amino acids in length.

The green square designates a shift in the dot plot where the accepted *D. melanogaster ey-PD* sequence is aligned with an equivalent length *D. grimshawi* sequence, but the "poly-Q" nature of the *D. grimshawi* sequence causes it to match at multiple locations. Looking at the green box in Figure 15b, one will note that while the amino acid length is conserved, there is a "poly-Q" string of amino acids in the 356-375 region of the *D. grimshawi* sequence (the query). This corresponds to the length of the shift in the dot plot with respect to the y-axis (*D. grimshawi*). With respect to the dot plot x-axis (*D. melanogaster)*, the shift occurs in a short location near amino acid 350. In Figure 15b, this corresponds to the three Q alignment between the two species (353-355 with respect to *D. melanogaster* and 373-375 with respect to *D. grimshawi*). Thus, it is likely that the shift is due to the "poly-Q" string of the *D.grimshawi* amino acid sequence repeatidly matching with the three Q amino acids that align between the two species according to BLAST results. This explains the vertical nature of the shift. The multiple alignment idea is further supported by the observation that there is a gap in the *D. melanogaster* axis just before this shift. The gap represents the amino acids in *D. melanogaster* that are not aligned with *D. grimshawi* according to the dot plot. The amino acids 336-349 with respect to *D. melanogaster* are unaligned and is the gap location. Other isoforms also provided dot plots. Peptide sequences and transcript

sequences were derived from all three isoforms through *Gene Model Checker*. These dot plots and sequences are available in the appendix.

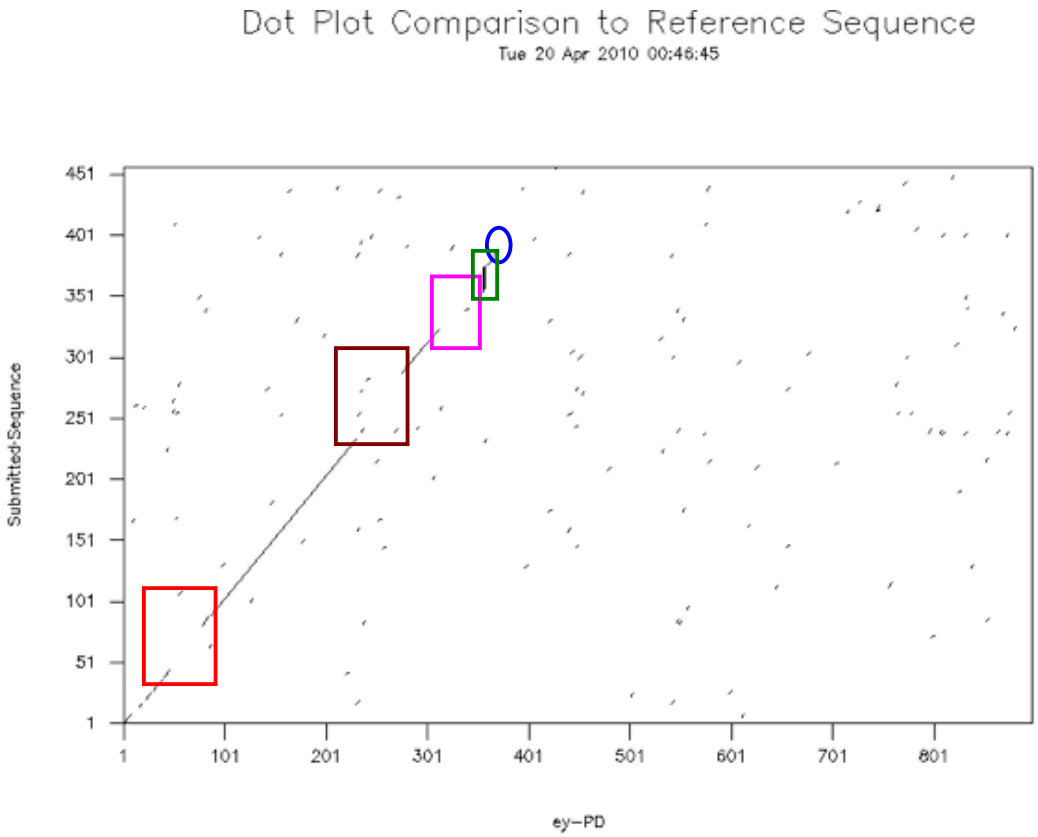


Figure 15a: Dot plot comparing the submitted *D. grimshawi* gene model sequence for isoform *ey-PD* to the accepted *ey-PD* isoform sequence from *D. melanogaster*. The red, brown, and pink rectangles show the gaps between homologous regions represented by the straight line. They represent the nonhomologous regions marked in Figure 15b. The blue circle indicates the end of *ey-PD* in contig 10 near the 383 amino acid mark (with respect to query). The green square indicates the shift due to the "poly-Q" string shown. All colors and shapes match the Bl2seq alignment in Figure 15b.

Figure 15b: : This Bl2seq alignment compares the submitted *D. grimshawi* gene model sequence (query) for isoform *ey-PD* to the accepted *ey-PD* isoform sequence from *D. melanogaster* (subject). The red, brown, and pink rectangles show the amino acid gaps between homologous regions. The blue circle indicates the end of *ey-PD* in contig 10 near the 383 amino acid mark (with respect to query). The green square indicates the "poly-Q" string that leads to the vertical shift observed in the dot plot. All colors and shapes match the dot plot in Figure 15a.

***ClustalW* Analysis:**

ey-PD ClustalW

ClustalW[4] is a multiple alignment program that gauges the amount of evolutionary divergence between sequences of multiple species. A *ClustalW* alignment was done for the polypeptide sequence of isoform *ey-PD* between the species *D. grimshawi, D. melanogaster,* and *D. virilis*. *D. mojavensis* was not included because the *ey-PD* isoform could not be obtained. The *ey-PD* isoform was chosen because it represents the greatest number of exons in the *ey* protein. Figure 17 shows a high and consistent alignment score between the three *Drosophila* species. This is good evidence that *ey* is a well-conserved gene. Figure 18 shows the alignment between the three *Drosophila* species

and supports the contention that *ey* is a well-conserved gene. The *D. virilis* sequence used is the orthologous isoform to *ey-PD* in *D. grimshawi* according to the GEP *D. virilis* annotations (http://gander.wustl.edu/~wilson/bio4342/dvirgenes/). A bl2seq confirms that the *D. mojavensis* sequence is orthologous to *ey-PD* in *D. virilis*. Lastly, Figure 19 shows a cladogram that indicates the predicted lineage of the species based on the alignment scores for the *ey-PD* isoform. Since there are only three closely aligned species derived from three *Drosophila* species, this *ClustalW* alignment and the resulting cladogram is not very informative.

Scores Table

Sort by | Sequence Number | View Output File

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	grimshawi	457	2	melanogaster	430	64
1	grimshawi	457	3	virilus	455	72
2	melanogaster	430	3	virilus	455	67

PLEASE NOTE: Some scores may be missing from the above table if the alignment was done using multiple CPU mode. Please check the output.

Sort by | Sequence Number | View Output File

Figure 17: *ClustalW* Shows a high and consistent alignment score between the three *Drosophila* species. This is good evidence that *ey* is a well-conserved gene.

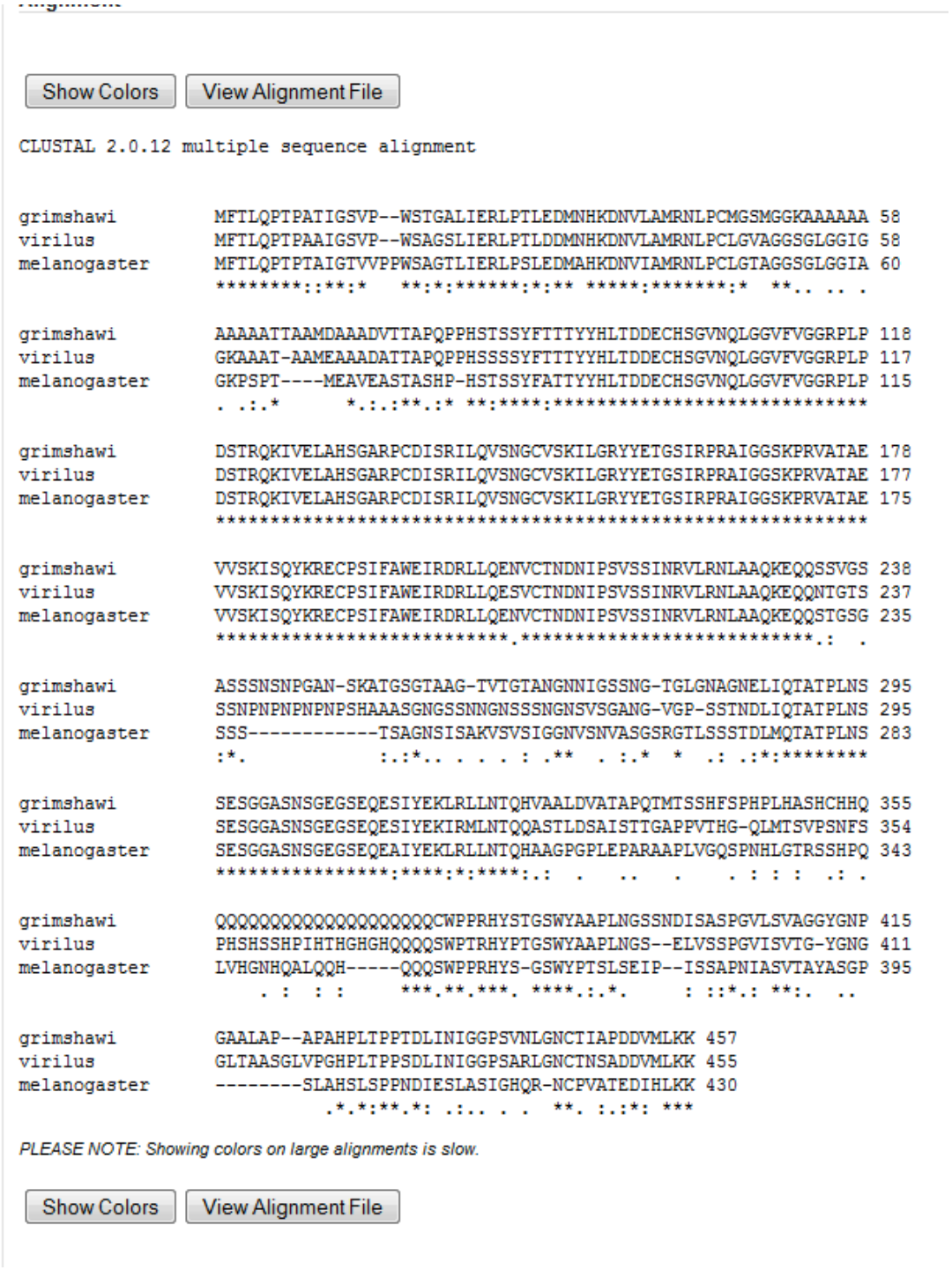


Figure 18: Alignment window between the three *Drosophila* species showing *ey* as a well-conserved gene.

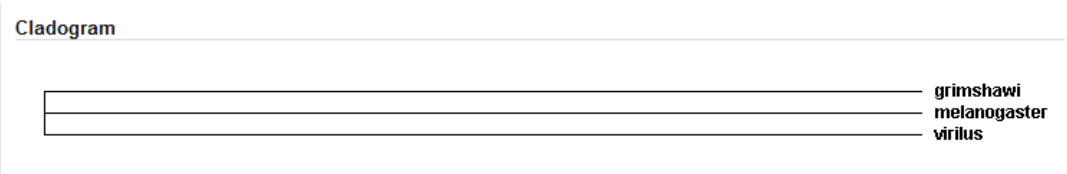


Figure 19: The cladogram indicates the predicted lineage of the species based on the alignment scores for the *ey-PD* isoform.

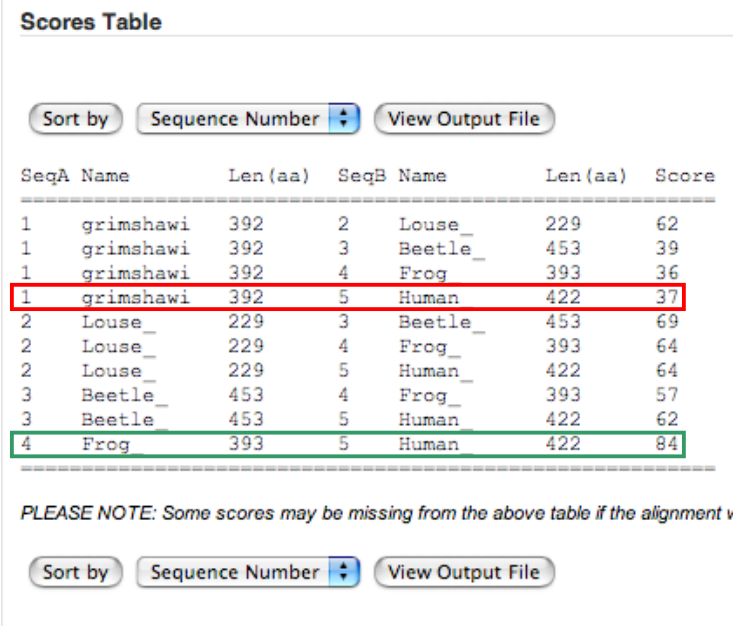
Broader *ClustalW* analysis on *ey-PA*

Since the *ClustalW* alignment comparing the three *Drosophila* species showed such a high degree of alignment that it is difficult to distinguish the evolutionary deviation between species for *ey-PD*, I chose four additional more distantly related species to preform a broader *ClustalW* alignment on against *ey-PA*. They are organized into Table 5 (does not include *D. grimshawi*). Three of these species have the ortholog to *ey* named "paired box protein" *Pax-6*. This was found to be the ortholog for these species through the high alignment scores produced in a Blastn search of *ey-PA*.

Common Name	Species	Blast ID	Ortholog to *ey*
Body Louse	*Pediculus humanus corporis*	gi|242009926|ref|XP_002425733.1|	Paired box protein *Pax-6*
Red Flour Beetle	*Tribolium castaneum*	gi|160333791|ref|NP_001103907.1|	*ey*
African Clawed Frog	*Xenopus laevis*	gi|288557282|ref|NP_001165666.1|	Paired box protein *Pax-6*
Human	*Homo Sapiens*	gi|4505615|ref|NP_000271.1|	Paired box protein *Pax-6*

Table 5: Species in addition to *D. grimshawi* used in the broader *ClustalW* analysis

Figure 20 depicts the *ClustalW* alignment scores table, which shows the quality of alignment between amino acid sequences of the five species. The alignment scores make sense with the expected divergence between evolutionarily disimilar versus more similar species. For example, the red box in Figure 20 indicates a poor alignment between the evolutionarily distant species *D. grimshawi* and Human. On the other hand, the green box indicates a high alignment score between the more closely related Frog and Human species. These scores demonstrate the evolutionary divergence of *ey* orthologues better than the three *Drosophila* alignment because there is a wider gradient of alignment scores. Thus, one can differentiate locations within the *ey* orthologues that have heavily diverged from those that have moderately divereged or have been conserved.



Scores Table

Sort by | Sequence Number | View Output File

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	grimshawi	392	2	Louse_	229	62
1	grimshawi	392	3	Beetle_	453	39
1	grimshawi	392	4	Frog_	393	36
1	grimshawi	392	5	Human_	422	37
2	Louse_	229	3	Beetle_	453	69
2	Louse_	229	4	Frog_	393	64
2	Louse_	229	5	Human_	422	64
3	Beetle_	453	4	Frog_	393	57
3	Beetle_	453	5	Human_	422	62
4	Frog_	393	5	Human_	422	84

PLEASE NOTE: Some scores may be missing from the above table if the alignment v

Sort by | Sequence Number | View Output File

Figure 20: *ClustalW* alignment scores table depicting *ey* ortholog amino acid length and scores. The red box indicates a poor alignment between the evolutionarily distant species *D. grimshawi* and Human. The green box indicates a high alignment score between the more closely related Frog and Human species.

Looking at the *ClustalW* multiple alignment map (Figure 21), one can observe the specific locations of amino acid divergence and conservation between species. One should note the region of complete alignment between the five species represented by the long consecutive rows of stars underneath the alignment. Considering the evolutionary distance between many of these species (ie., Human and *D. grimshawi*), such alignment shows the importance of the *ey* ortholog in organismal survival and fitness. Furthermore, one can argue that due to such conservation in the aligned portion, this specific amino acid sequence is essential to all species evolutionarily proximal to the five considered. The less conserved portions of the map shows the locations that some species deviate in sequence from others and gives clues to the evolutionary divergence of *ey* orthologues. For example, the black arrows indicate regions within or near the completely aligned region where the Frog and Human align separately from the Beetle, Louse, and *D. grimshawi*. These positions could indicate areas that create divergence in the *ey* ortholog. The green arrows indicate regions near the completely aligned area where *D. grimshawi* is the only nonaligning species. Like the black arrow marked regions, these regions have flexibility in amino acid identity and also show the divergence between *D. grimshawi* and the four other species. The red box shows a region where the Frog, Human, and Beetle are the only aligning species and the blue box shows regions where the Frog and Human are the only aligning species. These areas demonstrate divergence of a portion of the *ey* orthologue gene. It is likely that such regions are responsible for the differences in *ey* orthologue function between species. Lastly, Figure 22 shows the cladogram derived from the *ey* orthologue alignment.

Figure 21: *ClustalW* analysis depicting the alignment between the African Clawed Frog, the Human, the Red Flour Beetle, the Body Louse, and *D. grimshawi*. The stars indicate the region of alignment between the five species. The black arrows indicate regions

within or near the completely aligned region where the Frog and Human align separately from the Beetle, Louse, and *D. grimshawi*. The green arrows indicate regions near the completely aligned area where *D.* grimshawi is the only nonaligning species. The red box shows a region where the Frog, Human, and Beetle are the only aligning species. The blue box shows regions where the Frog and Human are the only aligning species.

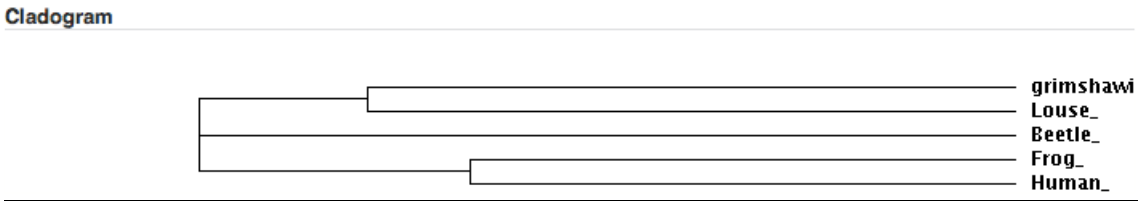


Fig. 22: The cladogram indicates the predicted lineage of five species based on the alignment scores for the *ey* orthologues.

UTR *ClustalW*

A second *ClustalW* alignment was preformed by aligning the nucleotide sequence of the 5' untranslated region of *ey-PA*. The four species *D. grimshawi, D. melanogaster, D. virilis,* and *D. mojavensis* were used in the alignment. Figure 23 demonstrates the low alignment score produced from the nucleotide sequence alignment. This is expected because the untranslated region is likely not functional and thus not evolutionarily conserved. The alignment map (Figure 24) demonstrates a similar conclusion since there are very sparse locations of universal alignment. Furthermore, since this a nucleotide rather than amino acid alignment, it is likely that the aligning areas are due to chance alone. One can further note the alignment of the last five bases that include the ATG nucleotide sequence and the beginning of exon 1.

Scores Table

Sort by | Sequence Number | View Output File

SeqA	Name	Len(nt)	SeqB	Name	Len(nt)	Score
1	grimshawi	106	2	melanogaster	94	12
1	grimshawi	106	3	virilis	106	15
1	grimshawi	106	4	mojavensis	106	28
2	melanogaster	94	3	virilis	106	9
2	melanogaster	94	4	mojavensis	106	10
3	virilis	106	4	mojavensis	106	11

PLEASE NOTE: Some scores may be missing from the above table if the alignment was done

Sort by | Sequence Number | View Output File

Figure 23:*ClustalW* Alignment score table for *ey-PA* 5' untranslated region. A low alignment score indicating weak conservation was observed.

```
CLUSTAL 2.0.12 multiple sequence alignment


grimshawi         -ATAAAAAAAAAAATACACAGTGACTCAA--GTGAAATCGAAAGTCTATATATAGTTTGT 57
mojavensis        CCAGCAAGAGAGTGAATCTAACTGCTGGA--ATAACTACAAAAAATA-ATAAGCACGTACA 57
virilis           --AAGAGCAACTTTTTTTGTGTTCCTTGACGGTTCGATCCAAGATCCA-AGATCGTTGGG 57
melanogaster      ----------TTCGCACGGCGTGCGTTTG-GCTGAACACAGCAGTCTCTTGGC--TAAAG 47
                                         *     *     *     *

grimshawi         CGGTGATCGATATCGGATAATTTGAGTACATATCGCGGACATTATGTTT 106
mojavensis        AAAAAAAAAAAAAAAAAAGAAAGAAAGAAATAAATAAATAAATAAAAAATGTTT 106
virilis           ATCGCAAGCATCTTGTTAACCACAAAAACATATCTTTTTAATTATGTTT 106
melanogaster      CTTTCATGAGCAGTGCATGTAATAAAAACTGA--GATCCAACTATGTTT 94
                       *                  *  *   *        *  ******
```

Figure 24: Nucleotide *ClustalW* alignment on the 5' untranslated region of *ey-PA* showing sparse alignment (the last five base pairs are the beginning of exon 1).

Bl2seq Alignment Analysis against *D. melanogaster*

A Bl2seq[6] analysis of each *D. grimshawi* isoform against *D. melanogaster* shows complete query coverage (Figure 25). This indicates that the gene models proposed incorporate all the exons in the portion of *ey* that is in contig 10. The alignment map in Figure 26 shows a 62% identity for isoform *ey-PD*. This mediocre identity is expected considering the evolutionary divergence between *D. grimshawi* and *D. melanogaster* contrasted with the common genus between the species.

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
3595	ey-PA melanogaster	420	420	100%	4e-122	

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
25441	ey-PC melanogaster	419	419	100%	1e-121	

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
31053	ey-PD melanogaster	475	475	100%	1e-138	

Figure 25: Bl2seq result showing 100% query coverage for all three isoforms between *D. grimshawi* and *D. melanogaster*. Amino acid sequences were used in the search.

```
>lcl|31053 ey-PD melanogaster
Length=427

 Score =  475 bits (1222),  Expect = 1e-138, Method: Compositional matrix adjust.
 Identities = 290/465 (62%), Positives = 332/465 (71%), Gaps = 46/465 (9%)

Query  1    MFTLQPTPATIGSV--PWSTGALIERLPTLEDMNHKDNVLAMRNLPCMGSMGGKAAAAAA  58
            MFTLQPTP  IG+V  PWS G LIERLP+LEDM HK NV+AMRNLPC+G+ GG      A
Sbjct  1    MFTLQPTPTAIGTVVPPWSAGTLIERLPSLEDMAHK-NVIAMRNLPCLGTAGGSGLGGIA  59

Query  59   AAAAATTAAMDAAADVTTAPQPPHSTSSYFTTTYYHLTDDECHSGVNQLGGVFVGGRPLP  118
               + T  A++A    +TA  P HSTSSYF TTYYHLTDDE HSGVNQLGGVFVGGRPLP
Sbjct  60   GKPSPTMEAVEA----STASHP-HSTSSYFATTYYHLTDDE-HSGVNQLGGVFVGGRPLP  113

Query  119  DSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATAE  178
            DSTRQKIVELAHSGARPCDISRILQVSNGCVSKILG YYETGSIRPRAIGGSKPRVATAE
Sbjct  114  DSTRQKIVELAHSGARPCDISRILQVSNGCVSKILG-YYETGSIRPRAIGGSKPRVATAE  172

Query  179  VVSKISQYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSINRVLRNLAAQKEQQSSVGS  238
            VVSKISQYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSINRVLRNLAAQKEQQS+
Sbjct  173  VVSKISQYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSINRVLRNLAAQKEQQSTGSG  232

Query  239  ASSSNSNPGANSKATGSGTAAGTVTGTANGNNIGSSNGTGLGNAGNELIQTATPLNSSES  298
            +SS+++  G +  A  S +  G V+  A+G+        G  ++  +L+QTATPLNSSES
Sbjct  233  SSSTSA--GNSISAKVSVSIGGNVSNVASGSR-------GTLSSSTDLMQTATPLNSSES  283

Query  299  GGASNSGEGSEQESIYEKLRLLNTQHVAA---LDVATAPQTMTSSHFSPHPLHASHCHHQ  355
            GGASNSGEGSEQE+IYEKLRLLNTQH A    L+ A A   +  S          +H   +
Sbjct  284  GGASNSGEGSEQEAIYEKLRLLNTQHAAGPGPLEPARAAPLVGQS--------PNHLGTR  335

Query  356  QQQQQQQQQQQQQQQQQQQQCWPPRHYSTGSWYAAPLNGSSNDISASPGVLSVAGGYGNP  415
                Q      Q  QQ QQQ WPPRHYS GSWY  P + S   IS++P + SV
Sbjct  336  SSHPQLVHGNHQALQQHQQQSWPPRHYS-GSWY--PTSLSEIPISSAPNIASVT------  386

Query  416  GAALAPAPAHPLTPPTDLINIGGPSVNLG---NCTIAPDDVMLKK  457
              A  P+ AH L+PP D+ ++     ++G   NC +A +D+ LKK
Sbjct  387  AYASGPSLAHSLSPPNDIESLA----SIGHQRNCPVATEDIHLKK  427
```

Figure 26: Bl2seq alignment of isoform *ey-PD* between *D. grimshawi* and *D. melanogaster*. The percent identity and percent positives are expected from such distantly related *Drosophila* species.

Feature 2 and Unaligned *Genscan* Exon

Feature 2, shown in the red box in Figure 27, was predicted by *Genscan* to be oriented in the negative direction, end in the 19 kb region, and originate from past the positive end of contig 10. However, feature 2's existance was immediately cast into doubt since neither the Blastx alignment or any of the gene predictors had exons in these locations. Further analysis of the feature included both a Blastn and Blastp search of the nucleotide and amino acid sequences used by *Genscan* (Figure 29). The Blastn of the nucleotide sequence with the NCBI database produced no hits and the Blastp search produced only the results shown in Figure 30. These hits have high e-values and are not proteins belonging to the *Drosophila* species. Thus, the second feature predicted by *Genscan* is a miscall. Furthermore, one exon predicted to be part of the *ey* assembly (shown by the blue box in Figure 27) does not align with the Blastx alignment or any of the other gene predictors. A Blastn search of the exported nucleotide sequence from this region produced no significant alignment results. Thus, this unaligned exon is likely a miscall by *Genscan*.

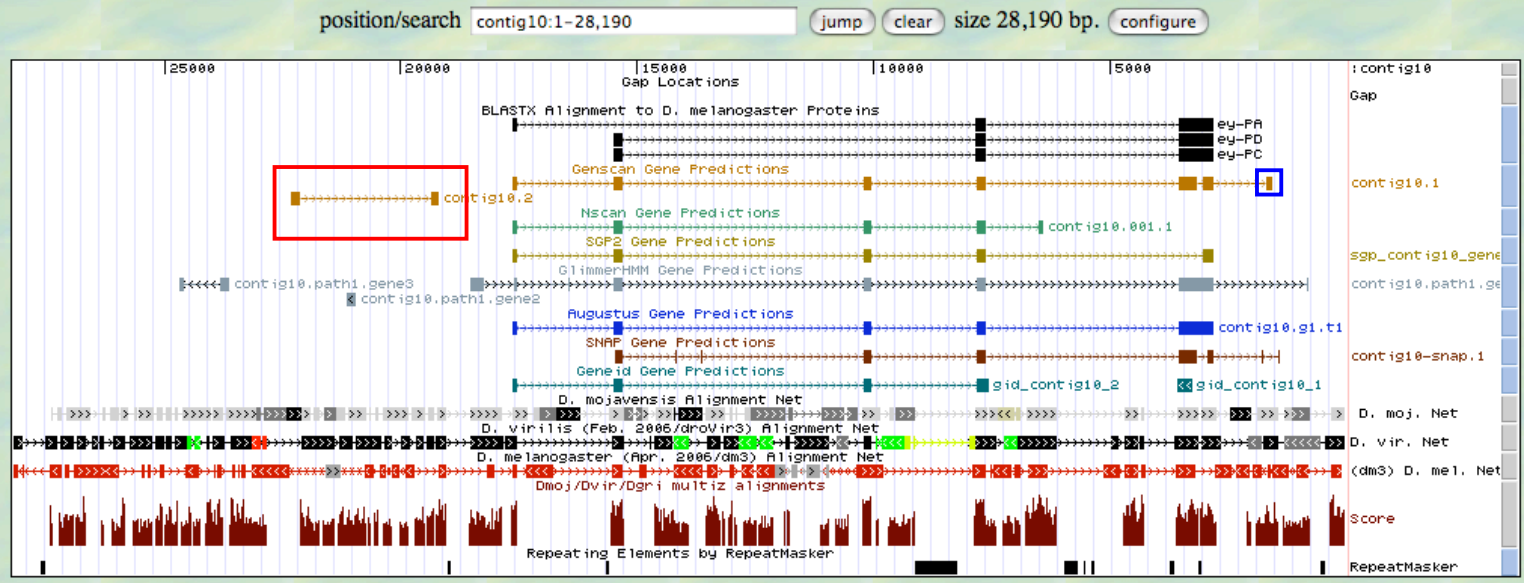


Figure 27: Browser view showing the *Genscan* predicted feature 2 (red box) and unaligned exon (blue box). Further analysis shows that both are miscalls by *Genscan*.

```
>contig10|GENSCAN_predicted_peptide_2|106_aa
XSLADLLTAADHTFIPLTLVFWDRIRSVVVFRRLSSKALNATGNALTCLFIRYAMFVIAL
CGCVLLRHFENKRKRPFNSYSLRVTRYPLTGYSVTEKREQEQRPQA

>contig10|GENSCAN_predicted_CDS_2|321_bp
nnctcgctcgctgatctgctgactgctgcagaccacacattcattcctctcactctcgtc
ttttgggatcgtatccgaagcgttgttgtttttcgtaggcttagcagtaaagctttaaat
gctaccggtaatgctttaacttgtctgtttatccgttacgctatgtttgtaattgctctt
tgtgggtgtgtcttgcttcgtcattttgaaaacaagcgcaaacgcccattcaacagttac
tcgttacgcgttacccgctacccgttaaccggttacagtgttacagagaaacgggaacag
gaacaacgcccacaagcataa
```

Figure 29: The peptide and nucleotide sequences called by *Genscan* for feature 2.

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
XP_800010.2	PREDICTED: similar to SJCHGC07951 protein [Strongylocentrotus purpura	34.3	34.3	68%	4.4	UG
CBL50022.1	Amino acid permease [Lactobacillus crispatus ST1]	33.5	33.5	39%	7.3	
ZP_05550122.1	amino acid permease [Lactobacillus crispatus 125-2-CHN] >gb|EEU18382.	33.5	33.5	39%	7.3	
ZP_03995853.1	APC family amino acid-polyamine-organocation transporter [Lactobacillus	33.5	33.5	39%	7.3	
AAU82844.1	serine/threonine protein kinase related protein [uncultured archaeon GZf	33.5	33.5	46%	8.6	

Figure 30: The Blastp alignment with the *Genscan* predicted peptide sequence. All matches have high e-values and are not proteins found in *Drosophila*.

RepeatMasker

RepeatMasker[5] is a program that identifies and masks nucleotide repeats. It masked only 6.30% of bases in contig 10, indicating a sparse population of repeats in this region (Figure: 31). Figure 32 shows the text output of the repeats masked by *RepeatMasker*. Most are short and less than 0.5 kb. Only the Line/LOA repeat that matches with Baggins1 (boxed in red) is greater than 0.5 kb (0.892kb).

```
╞=================================================
file name: contig10.fasta
sequences:            1
total length:     28190 bp  (28190 bp excl N/X-runs)
GC level:         32.13 %
bases masked:      1776 bp ( 6.30 %)
==================================================
               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:                2          962 bp    3.41 %
      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 %
      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:         5          423 bp    1.50 %
     hAT-Charlie      0            0 bp    0.00 %
     TcMar-Tigger     0            0 bp    0.00 %

Unclassified:         0            0 bp    0.00 %

Total interspersed repeats:     1385 bp    4.91 %


Small RNA:            0            0 bp    0.00 %

Satellites:           4          276 bp    0.98 %
Simple repeats:       1          115 bp    0.41 %
Low complexity:       0            0 bp    0.00 %
==================================================

* most repeats fragmented by insertions or deletions
  have been counted as one element
```

Figure 31: List of masked repeats by *RepeatMasker*. It masked 6.3% of bases in contig 10, which indicates sparse repeats in the contig.

contig10.fasta.out

SW score	perc div.	perc del.	perc ins.	query sequence	position in query begin	end	(left)		matching repeat	repeat class/family	position in repeat begin	end	(left)	ID	
370	23.5	1.7	0.0	contig10	436	550	(27640)	+	(CTG)n	Simple_repeat	2	118	(0)	1	
302	13.0	0.0	0.0	contig10	3058	3126	(25064)	C	TART_DV	LINE/telomeric	(3629)	11470	11402	2	
301	18.8	0.0	1.4	contig10	3640	3709	(24481)	C	dmoj.0.14.centroi	Satellite	(3)	942	874	3	
246	20.0	4.9	8.5	contig10	3649	3730	(24460)	C	dmoj.1.54.centroid	DNA	(6)	807	729	4	*
237	18.8	0.0	8.6	contig10	5317	5386	(22804)	C	dmoj.0.34.centroi	Satellite	(0)	927	864	5	*
254	23.7	0.0	0.0	contig10	5335	5393	(22797)	C	Helitron-1_DVir	DNA/Helitron	(0)	8816	8758	6	
231	12.8	0.0	0.0	contig10	5521	5559	(22631)	+	dvir.16.2.centroid	DNA	427	465	(608)	7	
1391	18.6	3.2	0.7	contig10	5678	5958	(22232)	C	dmoj.1.16.centroid	DNA	(170)	293	6	8	
2036	29.5	6.0	1.7	contig10	8216	9108	(19082)	C	Baggins1	LINE/LOA	(534)	4919	3988	9	
230	19.3	0.0	0.0	contig10	15599	15655	(12535)	+	dmoj.0.51.centroi	Satellite	697	753	(106)	10	
225	26.1	0.0	8.0	contig10	18936	19010	(9180)	+	dwil.28.21.centroid	DNA	183	251	(477)	11	
257	18.7	0.0	5.1	contig10	27525	27603	(587)	+	agam.228.353.centroi	Satellite	97	171	(1189)	12	

Figure 32: Text output of all repeats in Contig 10 masked by *RepeatMasker*. Only the Line/LOA repeat that matches with Baggins1 is greater than 0.5 kb (0.892kb).

Synteny

As shown in Figure 33, the *ey* gene resides in chromosome 4 of both *D. melanogaster* and *D. grimshawi*. It exists in essentially the same location, at a similar length, and at the same relative orientation in both species. The same genes that flank *ey* also exist between the two species. The gene *bt* flanks *ey* in the upstream location and the gene *myoglianin* flanks *ey* in the downstream location. The relative orientation and length of these genes are also largely conserved. Thus, *ey* is syntenic between the two species. This is further evidence that this gene is well-conserved. An interesting side-note is that two genes downstream of *ey*, *slip1* and *CG11360*, seem to be inverted between *D. melanogaster* and *D. grimshawi*. This indicates an inversion event just downstream of the *ey* gene.

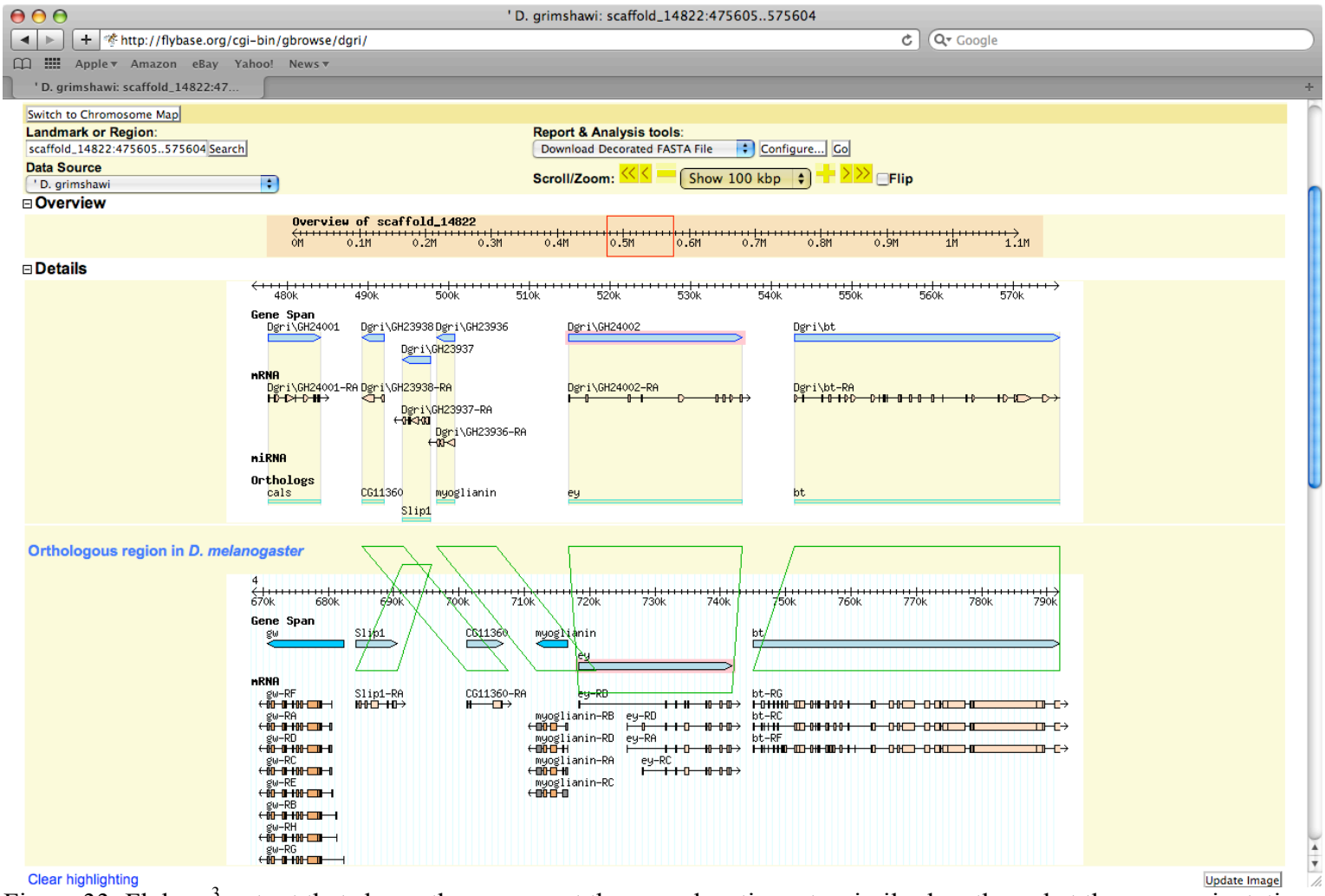


Figure 33: Flybase[3] output that shows the *ey* gene at the same location, at a similar length, and at the same orientation in both *D. melanogaster* and *D. grimshawi*. The flanking genes *myoglianin* and *bt* are also the same between the two species. This indicates that *ey* is syntenic between these two species.

Additional Work:

Contig 10 is largely annotated because the 3 isoforms of *eyeless* all have a verifiable gene model and most descripancies in the contig have been resolved. However, since contig 10 only contains the first six unique exons of *ey*, it would be wise to look at adjacent contigs for the remainder of the gene. This is especially true because there is no stop codon and two possible splice sites on the last exon of all three isoforms in this contig. It is likely that translation continues to the adjacent contig, but there is also no proof that this portion of *ey* is not a part of a pseudogene. Thus, looking at the adjacent contig in the negative direction (likely contig 9) would allow one to verify the entirety of the *eyeless* gene. One should note however, that analysis of neither contig 9 or contig 11 show any mention of the *eyeless* gene.[9]

Conclusion

Contig 10 of the *D. grimshawi* genome contains the beginning exons of isoforms *ey-PA*, *ey-PC,* and *ey-PD* of the *eyeless* gene (Refer to Tables 2, 3, and 4 for specifics on exon number, location, and length). A gene model was successfully constructed for all three isoforms. Isoform *ey-PA* of this gene shares close alignment between the species *D. grimshawi, D. melanogaster, D. virilis,* and *D. mojavensis*. Information about the evolution of the *eyeless* orthologues can be derived from a broader alignment. Untranslated regions of *ey* are much less conserved. Feature 2 and the unaligned exon (Figure 27) predicted by *Genscan* are likely mispredictions. There is only a 6.3% occurance of repeats in contig 10 according to *RepeatMasker* and one LINE/LOA element greater than 0.5 kb (0.892 kb). The *ey* gene was found to be syntenic between *D. grimshawi* and *D. melanogaster*, with the flanking genes being *bt* and *myoglianin* in both cases. As shown by Figure 34, the *eyeless* gene present in this region is largely annotated and contig 10 annotation can be considered complete except for the additional work described above.

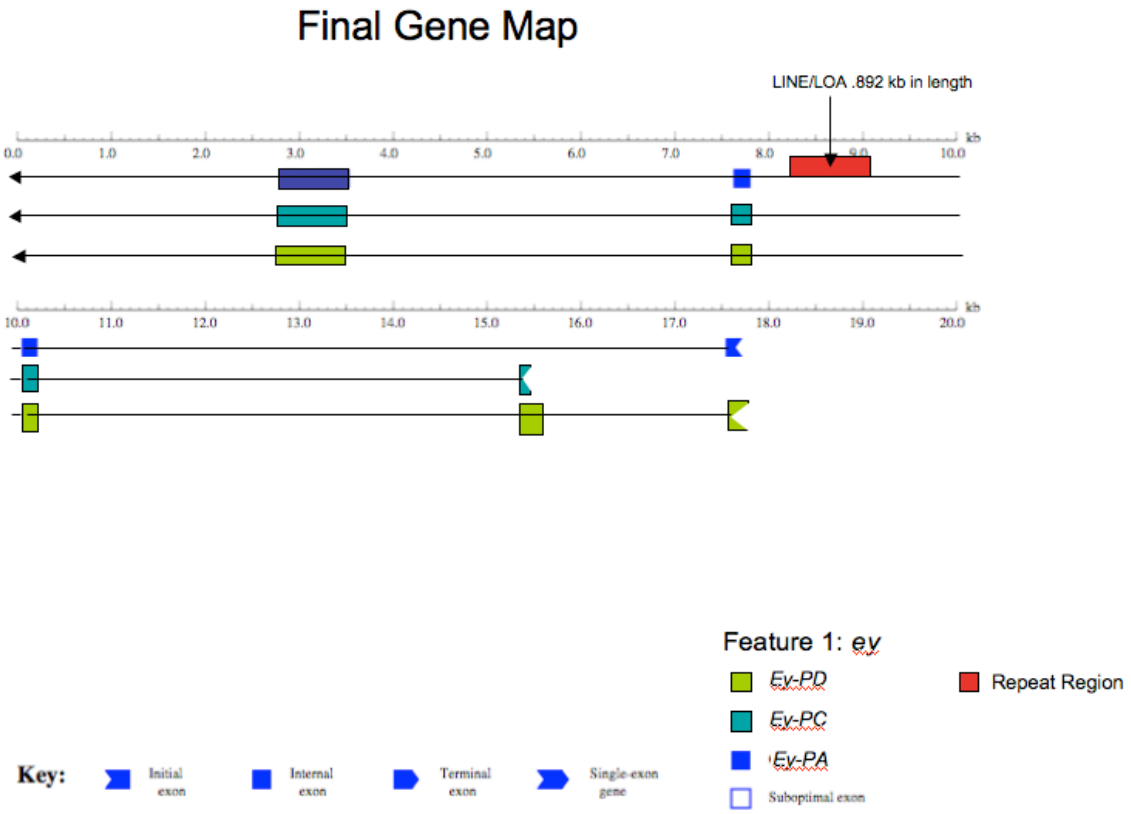


Figure 34: The Final Gene Map shows the locations of the exons for each isoform of the *ey* gene. The arrows indicate the relative orientation of the gene and connects the exons of each isoform that make up the gene. Isoform *ey-PA* is in blue, *ey-PC* is in teal, and *ey-PD* is in yellow. The repeat region larger than 0.5kb is shown in red.

For information on raw nucleotide data, *Gene Model Checker* outputs, *ClustalW* inputs, etc. contact Wilson Leung for the *D. grimshawi* Contig 10 Annotation Report Appendix.

Acknowledgements:

Special thanks to Wilson Leung, Aaditya Khatri, Gabe Haller, and Jeannette Wong for guiding me in annotating contig 10 and ensuring a successful outcome.

Special thanks to Professor Elgin, Professor Mardis and Professor Shaffer for making this course possible and ensuring that we have all the tools and knowledge necessary to partake in our endeavor.

Additional thanks to Professor Elgin, Dr. Chavez, and Jimmy Ma for editing this report.

Thanks to the Howard Hughes Medical Institute for their funding and support to make this course possible.

References

1. Drosophila 12 Genomes Consortium. "Evolution of genes and genomes on the *Drosophila* phylogeny." *Nature*. Nov. 8, 2007: 25-40.

2. Kent, Jim *et al.* "UCSC Genome Browser at Washington University on *D. grimshawi* Apr. 2010 Assembly." (http://gander/cgi-bin/hgTracks). 23 Apr. 2010

3. "Flybase." 19 March 2010. (http://flybase.org). 23 Apr. 2010

4. Larkin, MA *et al.* "Clustal W and Clustal X version 2.0" (http://www.ebi.ac.uk/Tools/clustalw2/index.html). 23 Apr. 2010

5. Smit, Arian *et al.* "*RepeatMasker*." (http://www.repeatmasker.org/). 23 Apr. 2010

6. "Basic Local Alignment Search Tool (BLAST)." (http://blast.ncbi.nlm.nih.gov/Blast.cgi). 23 Apr. 2010

7. Burge, Chris and Karlin, S. "Prediction of complete gene structures in human genomic DNA." Molecular Biology 268 (1997): 78-94.

8. Leung, Wilson *et al*. "Genomics Edcation Partnership Annotation Tools." (http://gep.wustl.edu/). 23 Apr. 2010

9. Personal communication, Ruth Howe and Lucy Liu, Washington University in St. Louis