Rachel Greenstein
Bio 434W – Annotation Report
May 2, 2012

Annotation of Fosmid24 from *Drosophila erecta* Chromosome 3L

Overview:

Using publicly available computational genomics tools including alignment programs such as the Basic Local Alignment Search Tool (BLAST) and the BLAST-Like Alignment Tool (BLAT), gene predictors like GENSCAN, and the UCSC Genome Browser, *Drosophila erecta* Fosmid24 was examined for relevant genomic elements – genes, pseudogenes, repetitious elements – and this information was compiled into a final annotated map of the region (partial map included here as Figure 1). The goal of this study is the annotation and identification of all the genes present within the fosmid sequence with the well-annotated *Drosophila melanogaster* sequence serving as reference for conservation-based analysis. Fosmid24 is located on the 3L-extended region of *D. erecta* and contains nine genes, all found in the same order and orientation as in the homologous region of *D. melanogaster*. Several of the genes are well conserved between *D. melanogaster* and *D. erecta*, while others show greater sequence divergence. The repetitious content of the region is low, consistent with its location in a euchromatic arm. A graphical representation of the annotated region follows (Figure 1).

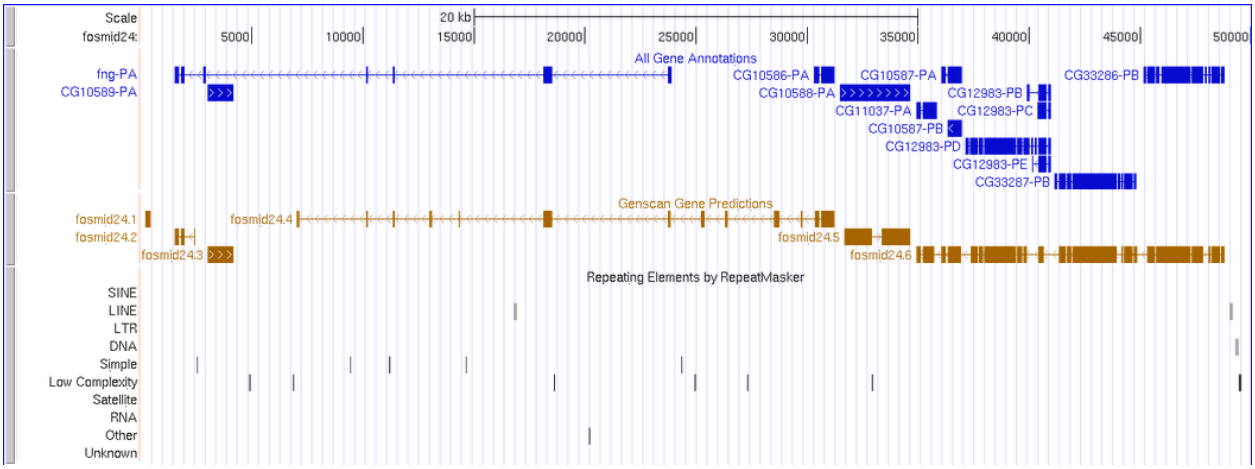


Figure 1: Graphical representation of final annotation from Genome Browser showing the custom track in blue, the GENSCAN predictions in brown, and the repeats in grey.

Introduction:

Given the recent innovations in sequencing technology and progress in computational capabilities, it has become a more feasible task to undertake massive comparative analyses of the genomic content of multiple species. *Drosophila melanogaster*, the well-characterized and iconic model organism for many genetic studies, serves as a good base for the creation of a study that looks to draw conclusions about the processes of evolution, adaptation, and conservation. In 2005, the *Drosophila* 12 Genomes Consortium published a paper in which they detailed the whole genome sequencing and assembly for 12 species of *Drosophila,* as well as drew some initial conclusions regarding important genomic phenomena such as chromosomal rearrangements, gene density and repetitive content. As a direct result of this study, other groups, including the Genomics Education Partnership (GEP) based at the Genome Institute at

Washington University in St. Louis, have begun efforts to improve the quality of the data by finishing genomic assemblies, as well as creating careful, sequence-based annotations of relevant features.

Well-curated genome assemblies are the ideal input for comparative genomics studies that seek to examine the processes of adaptation, conservation, and evolution via sequence-based analysis. Comparisons of homologous genomic regions between closely related species will hopefully yield information on the primary sequence changes that determine the differences between them. Additionally, these types of analyses can provide new data on phenomena such as repeat expansions, chromosomal translocations, and the evolution of new genes, and offer some perspective on the mechanisms that shape their occurrence. The data collected by the GEP will serve this purpose well in future comparative studies that endeavor to untangle the mystery of the evolution that has shaped the heterochromatic domains of the *Drosophila* genus.

Initial Findings:

One of the first steps in the process of genome annotation is the utilization of gene predictors to get a general sense of putative coding exons. Computer programs such as GENSCAN offer a good starting point for the identification of genes and gene-like regions even in genomes with complex structure, such as those of higher eukaryotes. While they do not often predict genes with complete accuracy, their ability to find long open reading frames offers researchers a general idea of where a gene might be located along a chromosome, providing them with a good base for further investigation. The initial GENSCAN output for Fosmid24 predicts six features, three on the forward strand and three on the reverse strand (Figure 2).

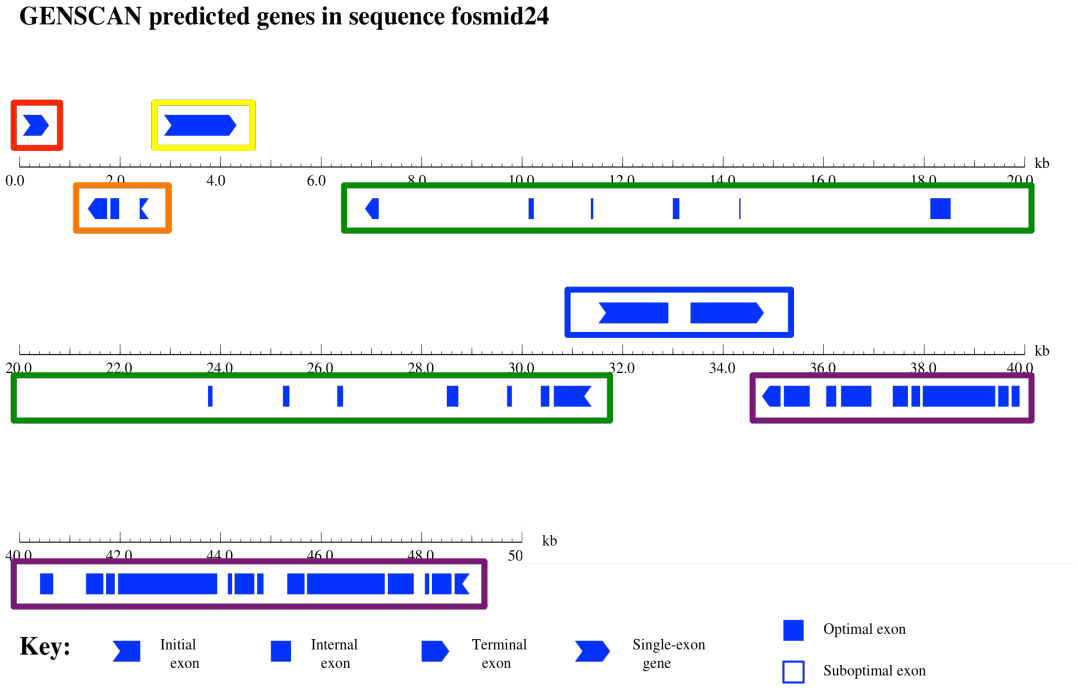


Figure 2: Initial GENSCAN output identifies six predicted coding features.

To begin identification of the features present in Fosmid24, I consulted the UCSC Genome Browser Mirror hosted by Washington University. The output track showing BLASTx hits from a search that used the nucleotide sequence from Fosmid24 as a query against the *D.*

melanogaster protein database is significantly different from the predicted GENSCAN output (Figure 3). The BLASTx search returned 13 hits to known *D. melanogaster* proteins, although several are completely overlapping, which suggested some of the hits are extraneous. Another point is that the three longer hits on the right-hand side of the browser window are visually similar to each other and to the largest feature predicted by GENSCAN (marked purple in Figure 2 above, Figure 3 below). Additionally, there is another set of three hits that are similar in exon organization (marked with green arrows in Figure 3).

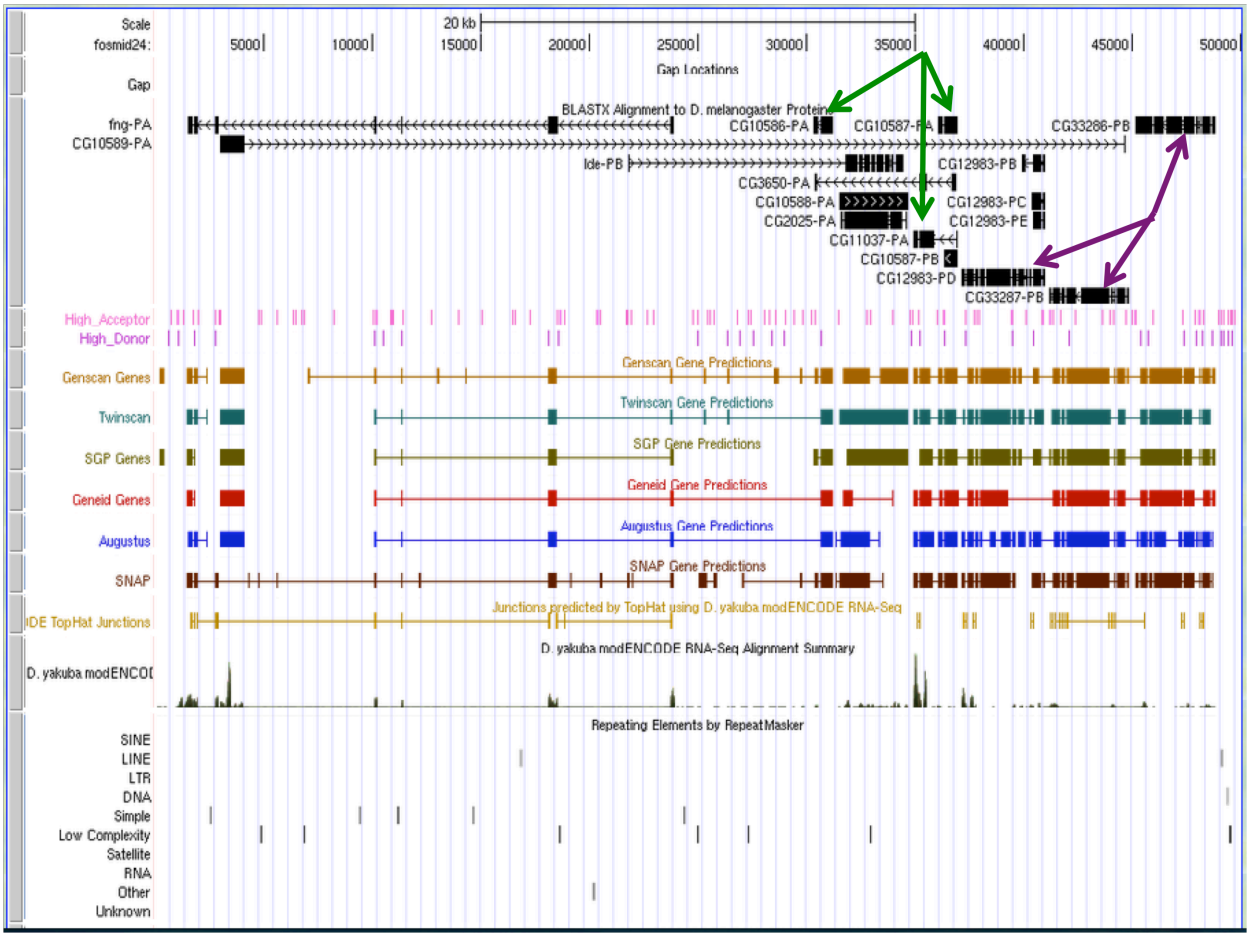


Figure 3: UCSC Genome Browser Mirror showing BLASTx hits to *D. melanogaster* proteins, gene prediction tracks, splice site predictions, *D. yakuba* modENCODE RNA-Seq data, and repetitious elements.

Discovery of a Gene Family:

Because these BLASTx hits differed significantly from the GENSCAN output, I decided to investigate them further. A tBLASTn search that queried the protein sequence from the *D. melanogaster* protein CG10586 or *Sems* (green arrows) against the unmasked Fosmid24 sequence returned 3 alignments within the fosmid, only one of which was full length, with an E-value of 7×10^{-138} (Figure 4).

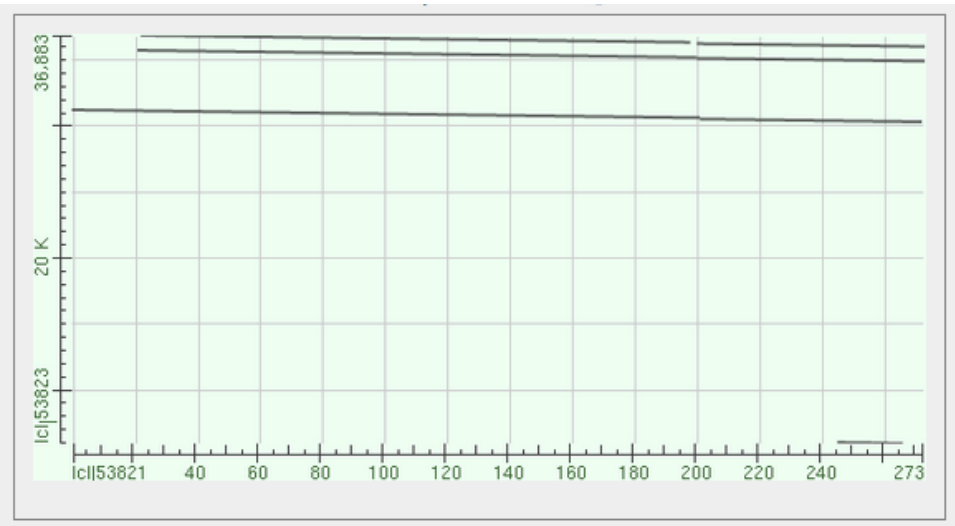
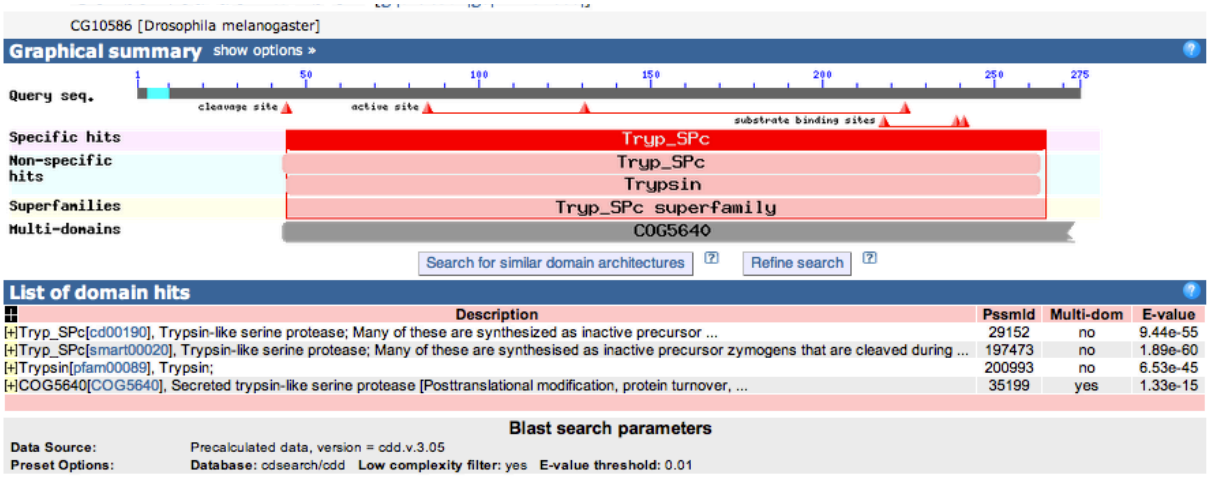


Figure 4: 2-D dot plot from the output of the tBLASTn search of *CG10586* protein sequence against Fosmid24.

To determine if CG10586 belongs to a protein family, a BLASTp search of its sequence against the *D. melanogaster* RefSeq protein database was run. This search revealed that *CG10586* contains a conserved Trypsin-like Serine Protease domain and its sequence has significant homology with several proteins that also had hits to Fosmid 24- CG10587 and CG11037 (Figure 5).



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NP_649270.1	CG10586 [Drosophila melanogaster]	570	570	100%	0.0	100%	U G M
NP_649272.1	CG11037 [Drosophila melanogaster]	348	348	91%	1e-121	67%	U G M
NP_649273.1	CG10587, isoform A [Drosophila melanogaster]	341	341	99%	3e-119	60%	U G M
NP_001137989.1	CG10587, isoform B [Drosophila melanogaster]	222	222	72%	2e-73	55%	G M
NP_611971.1	CG3650 [Drosophila melanogaster]	205	205	82%	2e-66	43%	U G M
NP_728833.1	CG32271 [Drosophila melanogaster]	177	177	82%	2e-55	40%	U G M
NP_788455.1	CG33159 [Drosophila melanogaster]	161	161	89%	2e-49	39%	U G M
NP_728837.1	CG32270 [Drosophila melanogaster]	150	150	81%	3e-45	38%	M
NP_647788.1	CG32269 [Drosophila melanogaster]	144	144	74%	2e-42	38%	U G M
NP_788456.1	CG33160 [Drosophila melanogaster]	141	141	96%	7e-42	34%	U G M
NP_523518.2	trypsin 29F [Drosophila melanogaster]	125	125	96%	8e-36	37%	U G M

Figure 5: Conserved Trypsin-like Serine Protease domain found in *D. melanogaster* protein *CG10586* (top); BLASTp hits to other potential genes found in Fosmid24, CG10587 and CG11037 (bottom).

Similar results are obtained with tBLASTn searches that use the protein sequences of CG10587 and CG11037 as a query against the Fosmid24 sequence. *CG10586* shows full-length alignment

to the first of these hits, CG11037 to the second, and CG10587 to the third. A BLAT alignment of all three protein sequences to the *D. erecta* 3L-extended assembly clearly identifies the orthology and order of the three features (Figure 6, top). Based on this information, it was concluded that all three of these genes represent a gene family and all are contained within the sequence of Fosmid24 in the order presented above. A comparison with the orthologous chromosomal region in *D. melanogaster* reveals that these three genes are present in the same order and orientation, thus conserving synteny (Figure 6, bottom). It should be noted that all three genes are located on the reverse strand.

D. erecta BLAT Results

BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	CG10586	493	19	271	275	82.8%	fosmid24	+-	30332	31180	849
browser details	CG10586	493	19	271	275	82.8%	fosmid23	+-	5332	6180	849
browser details	CG10586	249	23	273	275	67.0%	fosmid24	+-	34928	36883	1956
browser details	CG10586	249	23	273	275	67.0%	fosmid23	+-	9928	11883	1956
browser details	CG10586	148	109	271	275	65.7%	fosmid24	+-	36042	36625	584
browser details	CG10586	148	109	271	275	65.7%	fosmid23	+-	11042	11625	584
browser details	CG10587	593	1	276	276	85.9%	fosmid24	+-	36033	36955	923
browser details	CG10587	593	1	276	276	85.9%	fosmid23	+-	11033	11955	923
browser details	CG10587	262	25	276	276	67.7%	fosmid24	+-	34925	35746	822
browser details	CG10587	262	25	276	276	67.7%	fosmid23	+-	9925	10746	822
browser details	CG10587	233	25	274	276	71.7%	fosmid24	+-	30329	31168	840
browser details	CG10587	233	25	274	276	71.7%	fosmid23	+-	5329	6168	840
browser details	CG11037	629	5	292	292	86.5%	fosmid24	+-	34925	35854	930
browser details	CG11037	629	5	292	292	86.5%	fosmid23	+-	9925	10854	930
browser details	CG11037	266	61	292	292	77.4%	fosmid24	+-	36033	36823	791
browser details	CG11037	266	61	292	292	77.4%	fosmid23	+-	11033	11823	791
browser details	CG11037	244	41	290	292	66.9%	fosmid24	+-	30329	31168	840
browser details	CG11037	244	41	290	292	66.9%	fosmid23	+-	5329	6168	840

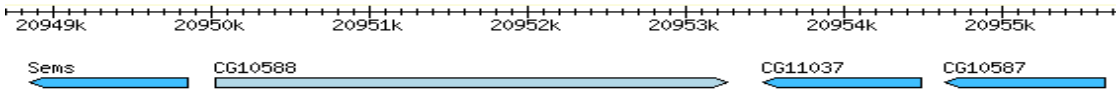


Figure 6: BLAT search of all three sequences identifies feature orthology (top); GBrowse output from Flybase showing the genes in the orthologous region of chromosome 3L in *D. melanogaster* (bottom)

Once the putative orthology of the paralogs had been assigned, the next step was to identify exon and intron coordinates and verify that the *D. erecta* sequences met the qualifications for true coding genes. I began with the protein predicted to be the ortholog to *D. melanogaster* CG10587. The workflow for annotating this feature will be described in full here and followed by results from a similar procedure for the features corresponding to CG10586 and CG11037. Any differences in method will be noted.

Annotation of CG10587 Ortholog:

Isoform A:

From the BLASTx track on the Genome Browser, it is clear that the best hit to *D. melanogaster* protein CG10587 is the third one in the gene family, so initial analysis was restricted to the region of this hit. Using the protein sequence from Flybase for CG10587-PA, a tBLASTn search of the A isoform's two exons queried against the Fosmid24 nucleotide sequence with region limited to 36,000 to 37,000bp showed nearly complete alignment with only the first 16 amino acids missing from the first exon. The regions at the 3′ end of the first exon and the 5′ end of the second exon that span the single intron are well conserved. The E-values for the exons are $2e^{-116}$ and $2e^{-48}$, respectively (Figure 7).

```
>lcl|45309 fosmid24
Length=50001

 Score =  326 bits (835),  Expect = 2e-116, Method: Compositional matrix adjust.
 Identities = 155/188 (82%), Positives = 171/188 (91%), Gaps = 0/188 (0%)
 Frame = -3

Query  16     SIEVLAQDLNQTIDVNKLAKIVQRPGFQTRVVGGDVTTNAQLGGYLIALRYEMNFVCGGT  75
              S  VL QDLNQTIDV KLA  VQ PGFQ+RVVGG+VTTNA+LGGYLIALRYE  F+CGG+
Sbjct  36910  STAVLGQDLNQTIDVRKLAAAVQSPGFQSRVVGGEVTTNARLGGYLIALRYEEAFICGGS  36731

Query  76     LLHDLIVLTAAHCFLGRVKISDWLAVGGASKLNDRGIQRQVKEVIKSAEFREDDMNMDVA  135
              LLH+LIVLTAAHCFLGRVKIS WLAVG ASKLNDRGIQR+VKEVIKSA+FREDDMNMDVA
Sbjct  36730  LLHELIVLTAAHCFLGRVKISSWLAVGSASKLNDRGIQREVKEVIKSAQFREDDMNMDVA  36551

Query  136    ILRLKKPMKGKSLGQLILCKKQLMPGTELRVSGWGLTENSEFGPQKLLRTVTVPVVDKKK  195
              IL LKKPM+ ++LG+L+LCKK L+PGTELRVSGWGLT  +EFGPQKLLRTVTVP+VDKK
Sbjct  36550  ILLLKKPMRHRTLGKLVLCKKHLVPGTELRVSGWGLTNPNEFGPQKLLRTVTVPIVDKKI  36371

Query  196    CRASYLPT  203
              CRASYLPT
Sbjct  36370  CRASYLPT  36347
```

```
Score =  146 bits (368),  Expect = 2e-48, Method: Compositional matrix adjust.
Identities = 72/73 (99%), Positives = 72/73 (99%), Gaps = 0/73 (0%)
Frame = -2

Query  1      HLTDSMFCAGVLGKKDACTFDSGGPLVYKNQVCGIVSFGIGCASKRYYGVYTDIMYVKPF  60
               LTDSMFCAGVLGKKDACTFDSGGPLVYKNQVCGIVSFGIGCASKRYYGVYTDIMYVKPF
Sbjct  36248  QLTDSMFCAGVLGKKDACTFDSGGPLVYKNQVCGIVSFGIGCASKRYYGVYTDIMYVKPF  36069

Query  61     IEQSIKVLLAKR*  73
              IEQSIKVLLAKR*
Sbjct  36068  IEQSIKVLLAKR*  36030
```

Figure 7: tBLASTn alignments of the two exons from *D. melanogaster* protein CG10587-PA to the Fosmid24 nucleotide sequence.

To identify the coordinates for the *D. erecta* ortholog, the sequences for the *D. melanogaster* exons were queried via BLAT against the 3L-extended assembly within the Genome Browser output. To identify the intron, I searched for a 5′ to 3′ GT nucleotide pair in the reverse frame. There is a GT pair in the +1 phase relative to the -3 frame where the first exon is located; this pair was also identified as a high quality donor site by the splice prediction track, which lends confidence to this choice (Figure 8). Because it is phase 1, there is an extra G base that is left over from the alignment, which must be considered in defining the intron acceptor site.

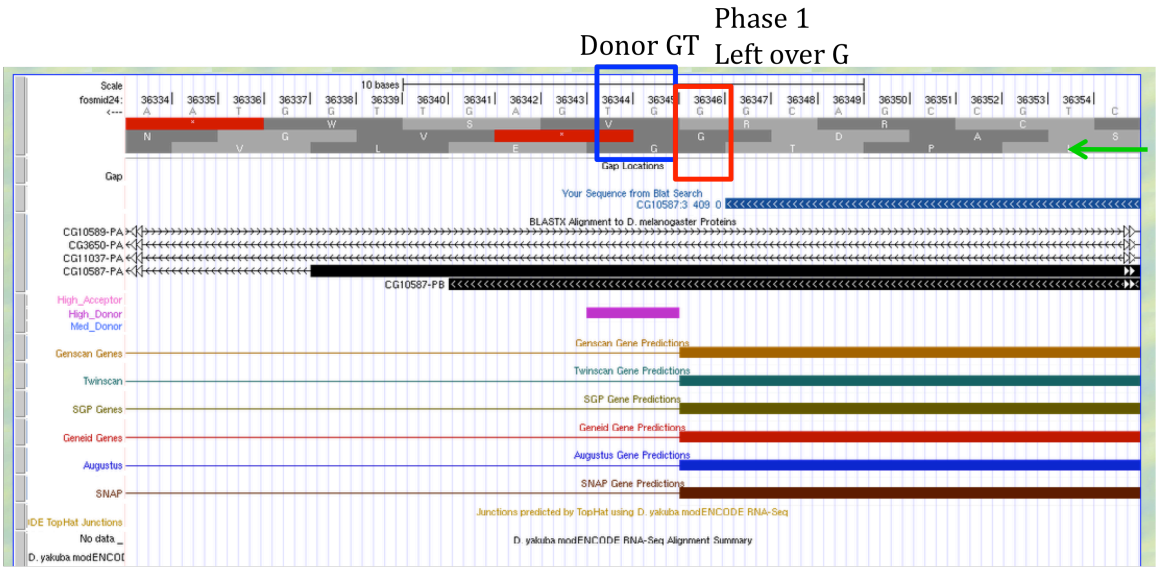


Figure 8: A close-up view of Fosmid24 showing the intron donor site for CG10587-PA ortholog.

An AG acceptor site in phase 2 relative to the second exon in frame -2 also agrees with the splice site predictor (Figure 9). The extra G from the donor splice fits into the position of the missing G from the acceptor splice site to yield a valine codon that is expressed in the corresponding protein, which is consistent with the splicing in *D. melanogaster*. It should be noted that in Figure 8 (below) the BLASTx alignment is overextended, while the BLAT alignment is truncated relative to the annotated protein and gene predictor outputs.

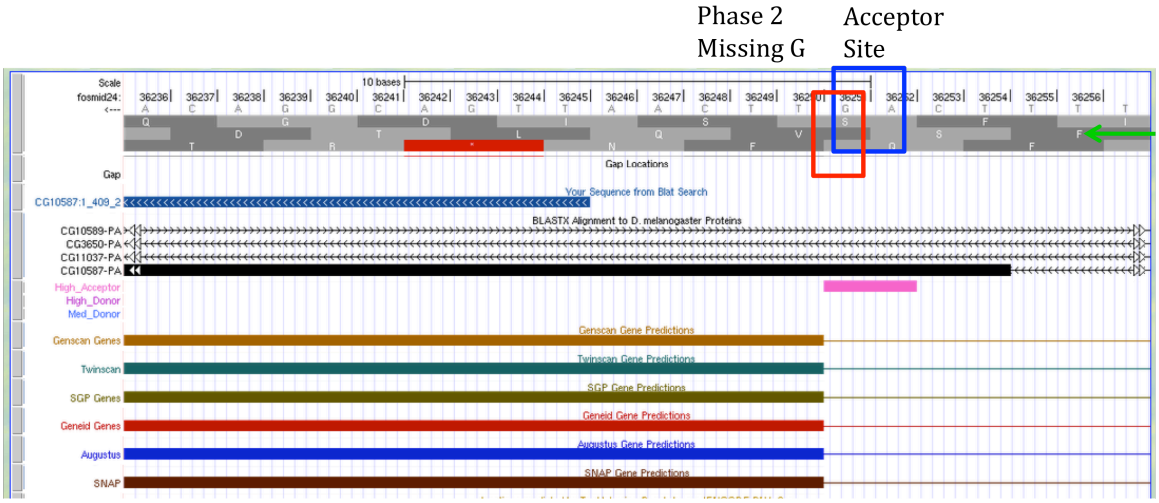


Figure 9: A close-up view of Fosmid24 showing the intron acceptor site for CG10587-PA ortholog.

Finally, the start codon was identified from the BLAT and BLASTx alignments, although it was not aligned by the initial tBLASTn search (Figure 10). A visual inspection of the sequence shows that the translated nucleotide sequence is similar to the beginning of the *D. melanogaster* ortholog. Additionally, the fact that this is the only in-frame AUG start codon in this region reinforces the annotation.

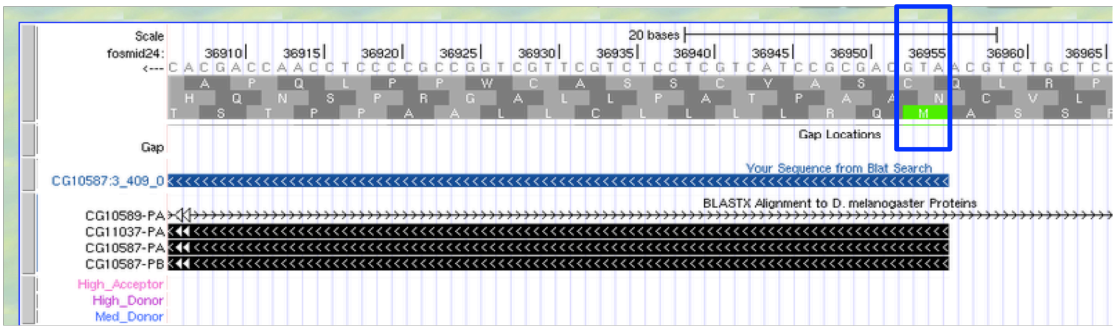


Figure 10: A close-up view of Fosmid24 showing the start codon for CG10587-PA ortholog.

Looking at the 3′ end of the alignment for the second exon, an in-frame stop codon can be found in the -2 frame (Figure 11). This aligns to the end of the *D. melanogaster* protein and is the logical choice for the *D. erecta* ortholog in this annotation.

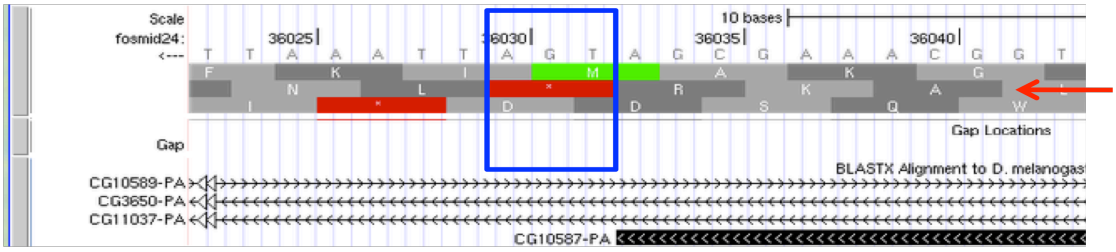


Figure 11: A close-up view of Fosmid24 showing the stop codon for CG10587-PA ortholog.

Isoform B:

The B isoform of *D. melanogaster* CG10587 has the same start codon as the A isoform; however it consists of only one coding exon that extends beyond where the A isoform first exon ends. A tBLASTn search of the isoform B protein sequence against the fosmid forms an alignment up to

residue 205, while the query contains 225 amino acids (Figure 12). The first 16 residues are unaligned as above.

```
>lcl|6071 fosmid24
Length=50001

 Score =  330 bits (845),  Expect = 3e-117, Method: Compositional matrix adjust.
 Identities = 157/190 (83%), Positives = 173/190 (91%), Gaps = 0/190 (0%)
 Frame = -3

Query  16     SIEVLAQDLNQTIDVNKLAKIVQRPGFQTRVVGGDVTTNAQLGGYLIALRYEMNFVCGGT  75
              S  VL QDLNQTIDV KLA  VQ PGFQ+RVVGG+VTTNA+LGGYLIALRYE  F+CGG+
Sbjct  36910  STAVLGQDLNQTIDVRKLAAAVQSPGFQSRVVGGEVTTNARLGGYLIALRYEEAFICGGS  36731

Query  76     LLHDLIVLTAAHCFLGRVKISDWLAVGGASKLNDRGIQRQVKEVIKSAEFREDDMNMDVA  135
              LLH+LIVLTAAHCFLGRVKIS WLAVG ASKLNDRGIQR+VKEVIKSA+FREDDMNMDVA
Sbjct  36730  LLHELIVLTAAHCFLGRVKISSWLAVGSASKLNDRGIQREVKEVIKSAQFREDDMNMDVA  36551

Query  136    ILRLKKPMKGKSLGQLILCKKQLMPGTELRVSGWGLTENSEFGPQKLLRTVTVPVVDKKK  195
              IL LKKPM+ ++LG+L+LCKK L+PGTELRVSGWGLT  +EFGPQKLLRTVTVP+VDKK
Sbjct  36550  ILLLKKPMRHRTLGKLVLCKKHLVPGTELRVSGWGLTNPNEFGPQKLLRTVTVPIVDKKI  36371

Query  196    CRASYLPTGE  205
              CRASYLPTGE
Sbjct  36370  CRASYLPTGE  36341
```

Figure 12: tBLASTn alignment of *D. melanogaster* protein CG10587-PB to the Fosmid24 nucleotide sequence.

To investigate a possible cause for the unaligned 3′ end, I took a closer look at the nucleotide sequence that corresponds to the end of the alignment at position 36341bp in the Genome Browser window. With this exon coding in the -3 frame, it was clear there was an in-frame stop codon that truncates this protein relative to the *D. melanogaster* ortholog (Figure 13). Of the 225 residues in the *D. melanogaster* protein, 209 can be accounted for in the *D. erecta* sequence from Fosmid24.

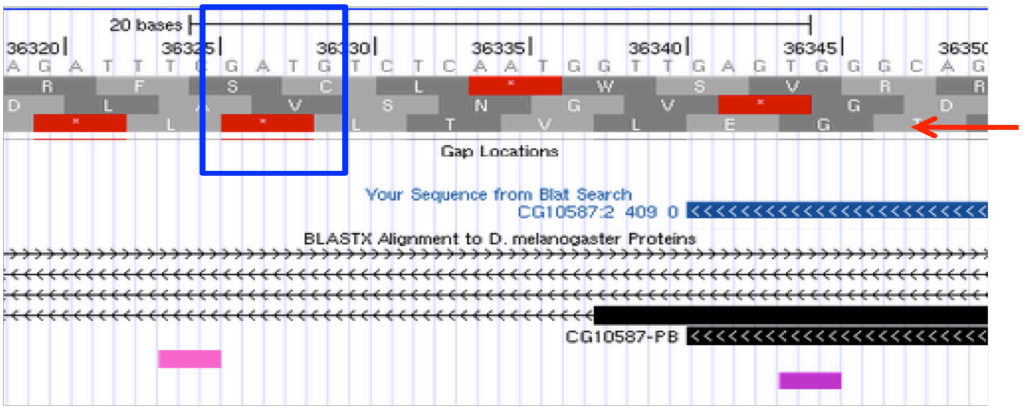


Figure 13: A close-up view of Fosmid24 showing the premature stop codon for CG10587-PB ortholog.

Because it is possible that the rest of the sequence could be present in a second exon downstream, I ran a tBLASTn search that used the protein sequence as a query against the fosmid sequence from the end of the aligned portion of the B isoform to the start of the second exon in the A isoform. This search found no alignments. I repeated the search with the nucleotide sequence of the exon using the BLASTn algorithm and found an alignment from fosmid position 36313-36336, which includes the codons for some of the truncated amino acids. Because this region extends from the end of the previous isoform B alignment, it cannot be a new exon and the multiple in-frame stop codons in this region only confirm the truncation.

In summary, the *D. erecta* ortholog to *D. melanogaster* protein CG10587 has two isoforms. Isoform A has a full length alignment of its two exons, while Isoform B contains only one exon as expected, though it is truncated 16 residues relative to the protein in *D. melanogaster*. A table summarizing the results follows (Table 1).

CG10587	Start	End	Frame	Phase Start/End	Isoform A	Isoform B	Comments
Exon 3_409_0	36955	36346	-3	Start/1	Exon 1	N/A	
Exon 1_409_2	36250	36033	-2	2/Stop	Exon 2	N/A	
Exon 2_409_0	36955	36329	-3	N/A	N/A	Exon 1	Truncated at residue 209/225

Table 1: Summary of exons from CG10587 ortholog

The coordinates of all the exons for each isoform were submitted to the Gene Model Checker with the outputs included below (Figure 14). Both isoforms shows alignments with no gaps other than the isoform B truncation and had varied levels of conservation to their *D. melanogaster* orthologs.

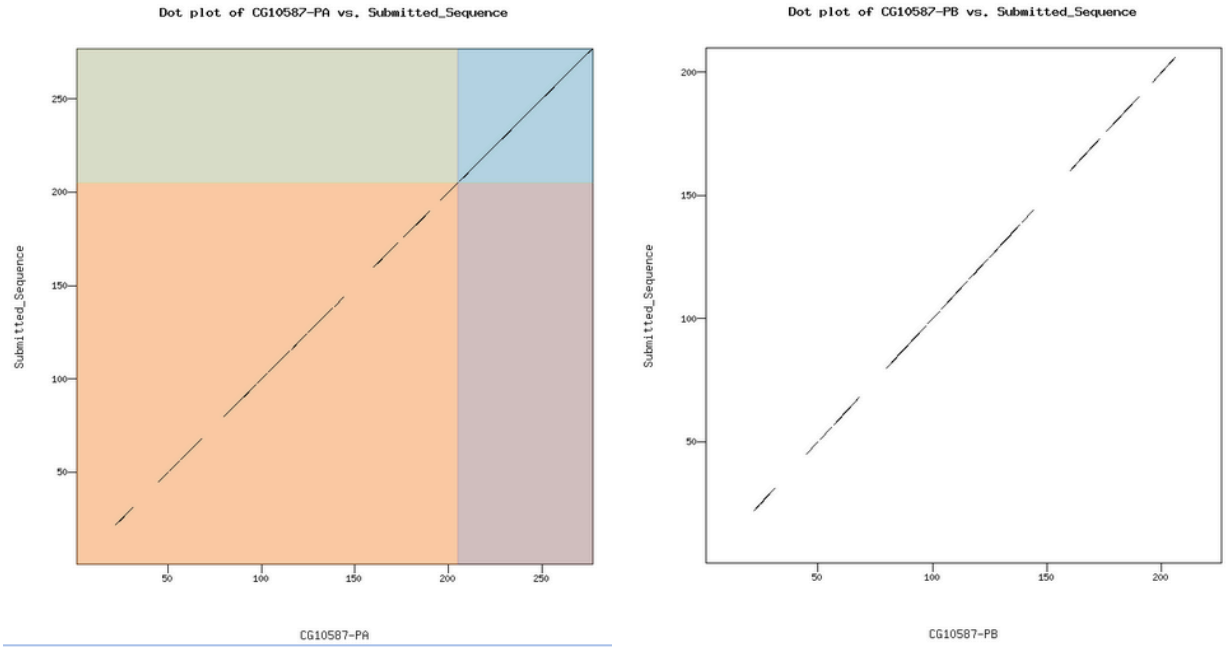


Figure 14: Gene Model Checker output for the A (left) and B (right) isoforms of the CG10587 ortholog displayed both as 2-D dot plots.

Annotation of CG11037 Ortholog:

The Flybase record for *D. melanogaster* protein CG11037 identifies only one isoform. A tBLASTn search with the protein sequence of CG11037 against the Fosmid24 sequence gave a hit with complete coverage of the query sequence and an E-value of $3e^{-166}$ along with two other incomplete hits with high E-values as expected. Next, a BLAT search of the *D. melanogaster* exon sequences against the *D. erecta* assembly confirmed that the best match to this protein was located in Fosmid24. The start codon was identified though a combination of visual inspection of the correct reading frame and alignments via BLASTx and BLAT to the CG11037 sequence. The site for the intron is supported by the modENCODE RNA-Seq data mapped from *D. yakuba*, modENCODE TopHat junctions, and the BLAT and BLASTx alignments (Figure 15). The GT donor site and AG acceptor sites are both in agreement with the high quality sites identified by

the splice site predictor algorithm and match up well having phases 1 and 2 respectively. The stop codon is located in the same position as the *D. melanogaster* ortholog.

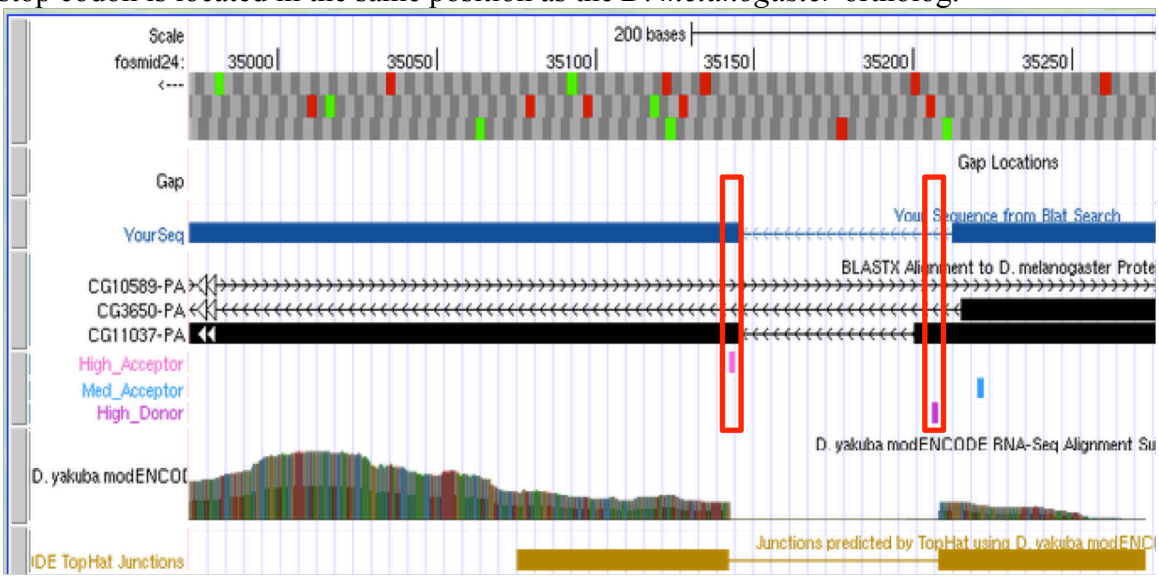


Figure 15: BLAT and BLASTx alignments to *D. melanogaster* protein CG11037, the GENSCAN predicted output, and the modENCODE RNA-Seq and TopHat Junctions tracks from *D. yakuba* support this intron splice.

To conclude, the CG11037 ortholog has a single isoform with two exons whose annotations have been supported by multiple lines of evidence. A summary table is included below (Table 2).

CG11037	Start	End	Frame	Phase at Start/End	Isoform A
Exon 2_432_0	35866	35209	-3	Start/1	Exon 1
Exon 1_432_2	35142	34925	-3	2/Stop	Exon 2

Table 2: Summary of exons from CG11037 ortholog

This set of coordinates was submitted to the Gene Model Checker and shows and alignment with no gaps relative to *D. melanogaster* as seen in the 2-D dot plot below (Figure 16). There is greater conservation of the second exon sequence relative to *D. melanogaster*, consistent with a more conserved second exon in the CG10587-PA.

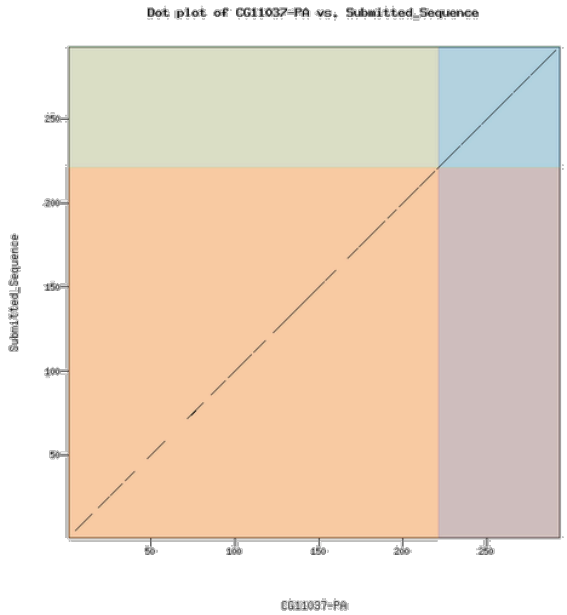


Figure 16: Gene Model Checker output for the CG11037 ortholog displayed as a 2-D dot plot.

Annotation of CG10586/Sems:
The Flybase record for *D. melanogaster* protein CG10586 (alternative name *Sems*) identifies only one isoform. A tBLASTn search with the protein sequence of CG10586 against the Fosmid24 sequence gave an alignment that contains 272 of the 275 residues from the query and an E-value of $7e^{-138}$ along with two other incomplete hits with high E-values, as expected. Next a BLAT search with the *D. melanogaster* exon sequences against the *D. erecta* assembly confirmed that the best match to this protein was located in Fosmid24. Again, the start codon was identified through a combination of visual inspection of the correct reading frame and alignments via BLASTx and BLAT to the CG10586 sequence. The site for the intron is supported by the modENCODE RNA-Seq data mapped from *D. yakuba*, BLAT and BLASTx alignments at least on the upstream side (Figure 17). More specifically, the GT donor site agrees with the high quality prediction from the splice site predictor program; this is the only GT pair close to the end of the BLAT and BLASTx alignments and it is in phase 1. The AG acceptor site agrees with the medium quality acceptor prediction from the splice predictor program; it is the only AG pair in the vicinity and is located in phase 2. The valine codon truncated by removal of the G from the acceptor in phase 2 is restored with extra G base from the phase 1 donor site. This valine is contained within the *D. melanogaster* protein sequence.

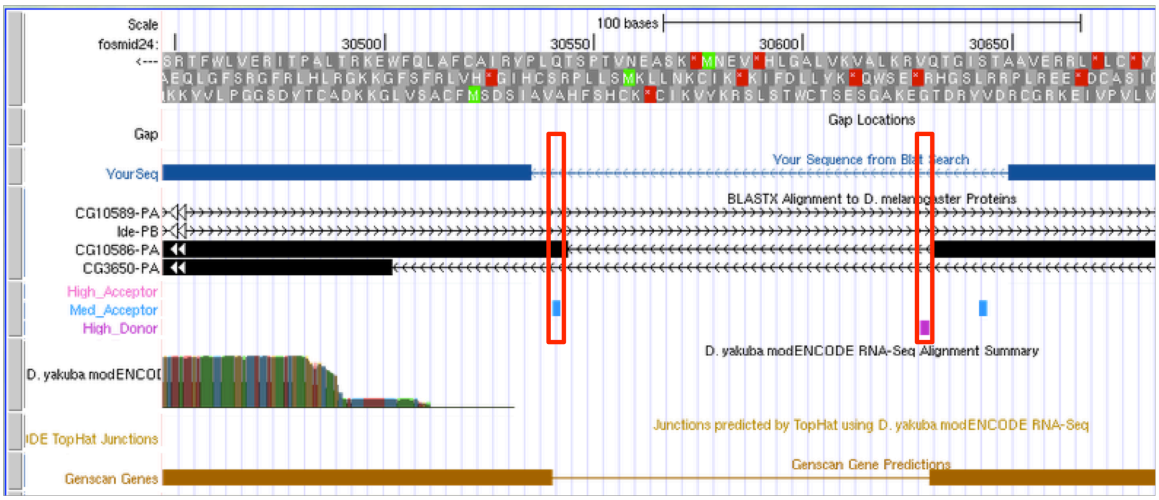


Figure 17: BLAT and BLASTx alignments to *D. melanogaster* protein CG10586, and the GENSCAN predicted output, tracks from *D. yakuba* support this intron splice.

Because the alignment via tBLASTn only contained 272 of 275 amino acids, it merited further investigation to determine the correct location of the stop codon. I took a closer look at the nucleotide sequence that corresponded to the end of the BLASTx alignment in the Genome Browser window. With this exon coding in the -3 frame, it is clear that there is an in-frame stop codon that truncates this protein relative to the *D. melanogaster* ortholog (Figure 18). Of the 275 residues in the *D. melanogaster* protein, 273 can be accounted for in the *D. erecta* sequence from Fosmid24.

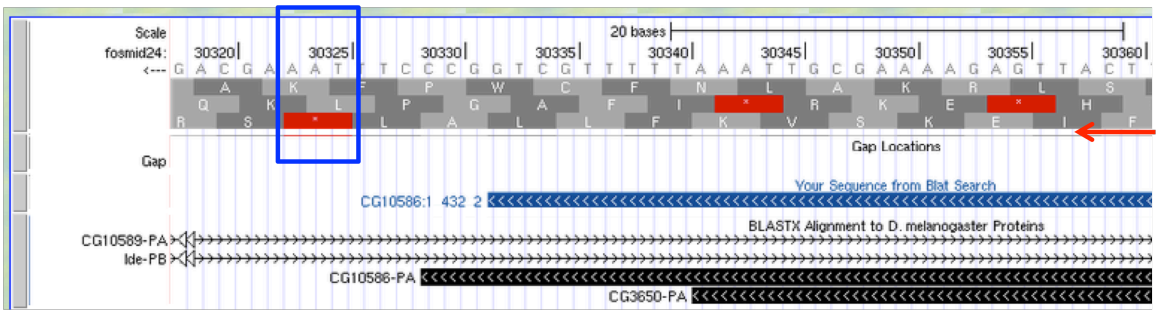


Figure 18: Premature stop codon in the 3' end of the CG10586 sequence

To conclude, the CG10586 ortholog has a single isoform with two exons whose annotations have been supported by multiple lines of evidence. A summary table is included below (Table 3).

CG10586	Start	End	Frame	Phase at Start/End	Isoform A	Comments
Exon 2_432_0	31234	30631	-3	Start/1	Exon 1	
Exon 1_432_2	30540	30326	-3	2/Stop	Exon 2	Premature stop codon, 273/275 residues present

Table 3: Summary of exons from CG10586 ortholog

The coordinates for the above gene model were submitted to the Gene Model Checker and have alignments with no gaps relative to *D. melanogaster* as seen in the 2-D dot plot below (Figure 19). There is greater conservation of the second exon sequence relative to *D. melanogaster*, consistent with a more conserved second exon in CG10587-PA and CG11037 annotated previously.

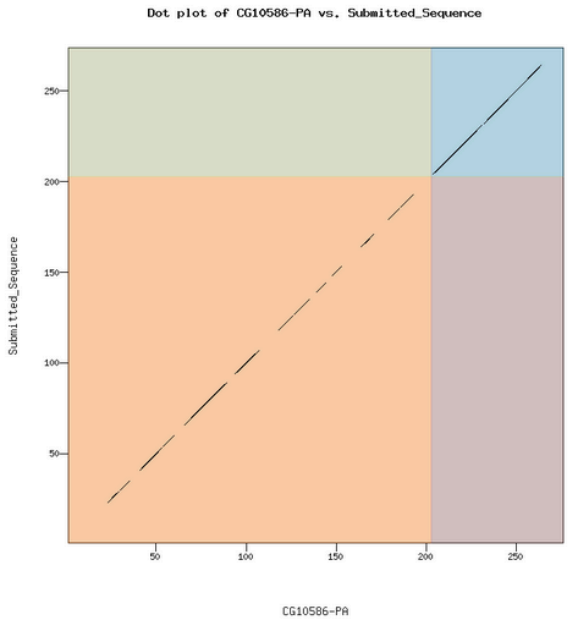


Figure 19: Gene Model Checker output for the CG10586 ortholog displayed as a 2-D dot plot.

Annotation of CG10588 Ortholog:

Based on the *D. melanogaster* BLASTx track from the Genome Browser interface, there is a large alignment block present from 31500 to 34500bp that shows alignment with multiple proteins. GENSCAN also identifies this region as a feature of interest consisting of two exons (Figure 2, blue box). Of the three protein hits, the alignment to *D. melanogaster* protein CG10588 is the most complete with an E-value of zero, complete coverage, and 88% identity. A BLASTp search of the GENSCAN feature output against the CG10588 protein sequence aligns the entirety of the predicted feature with an E-value of zero but reveals that the prediction program did not identify a region that corresponds to bases 1-62 and 480-623 of the CG10588 protein.

An external BLASTx search of the Fosmid24 sequence against the *D. melanogaster* CG10588 protein gives a complete alignment with an E-value of zero that includes the bases missed by GENSCAN. A tBLASTn search using the CG10588 protein as a query against the fosmid sequence shows that this protein aligns from 31486 to 34659bp in the Fosmid24 assembly. The coordinates for the exon were submitted to the Gene Model Checker and have alignments with no gaps (Figure 20). A summary table is included for reference (Table 4). Because this gene is well conserved and *D. erecta* is closely related to *D. melanogaster,* I attempted to identify the untranslated regions (UTRs) of the gene by searching the whole fosmid sequence with the mRNA transcript for the *D. melanogaster* CG10588. The BLASTn algorithm found an alignment to the fosmid region from 31450bp to 34675bp. Given that these points lie outside the annotated exon region, they likely correspond to this gene's UTRs. Unfortunately, there is no RNA-Seq data in this region to confirm this conclusion.

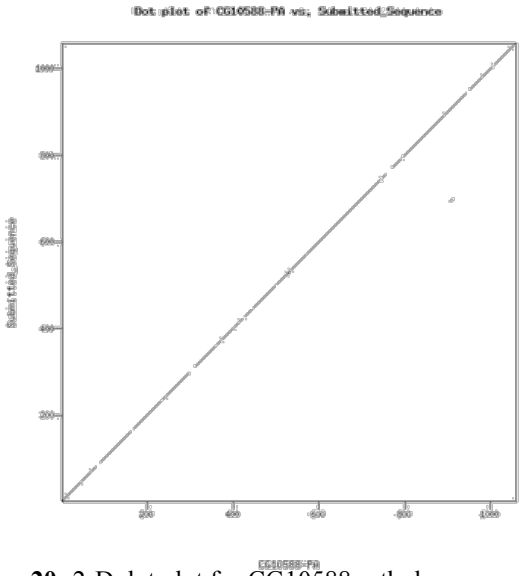


Figure 20: 2-D dot plot for CG10588 ortholog

CG10588	Start	End	Frame	Phase at Start/End
Exon	31486	34659	+1	Start/Stop
Transcript	31450	34675	N/A	N/A

Table 4: Summary of exons from CG10588 ortholog

Though I was able to identify a transcript region beyond the coordinates of the annotated gene by aligning the mRNA, I performed an additional ClustalW2 analysis to see if I could find any well-conserved promoter elements in the 5` UTR. Using the sequence anchored between the start of the CG10588 ortholog and the adjacent CG10586/Sems ortholog, I completed a multiple sequence alignment with the ClustalW2 program using sequences from *D. erecta, D. yakuba, D. ananassae,* and *D. melanogaster* (Figure 21). There are some regions that show a low level of conservation located near the end of the alignment, which corresponds to the region closest to the start codon of the CG10588 gene but no region shows enough conservation to conclude it is a definitive promoter element.

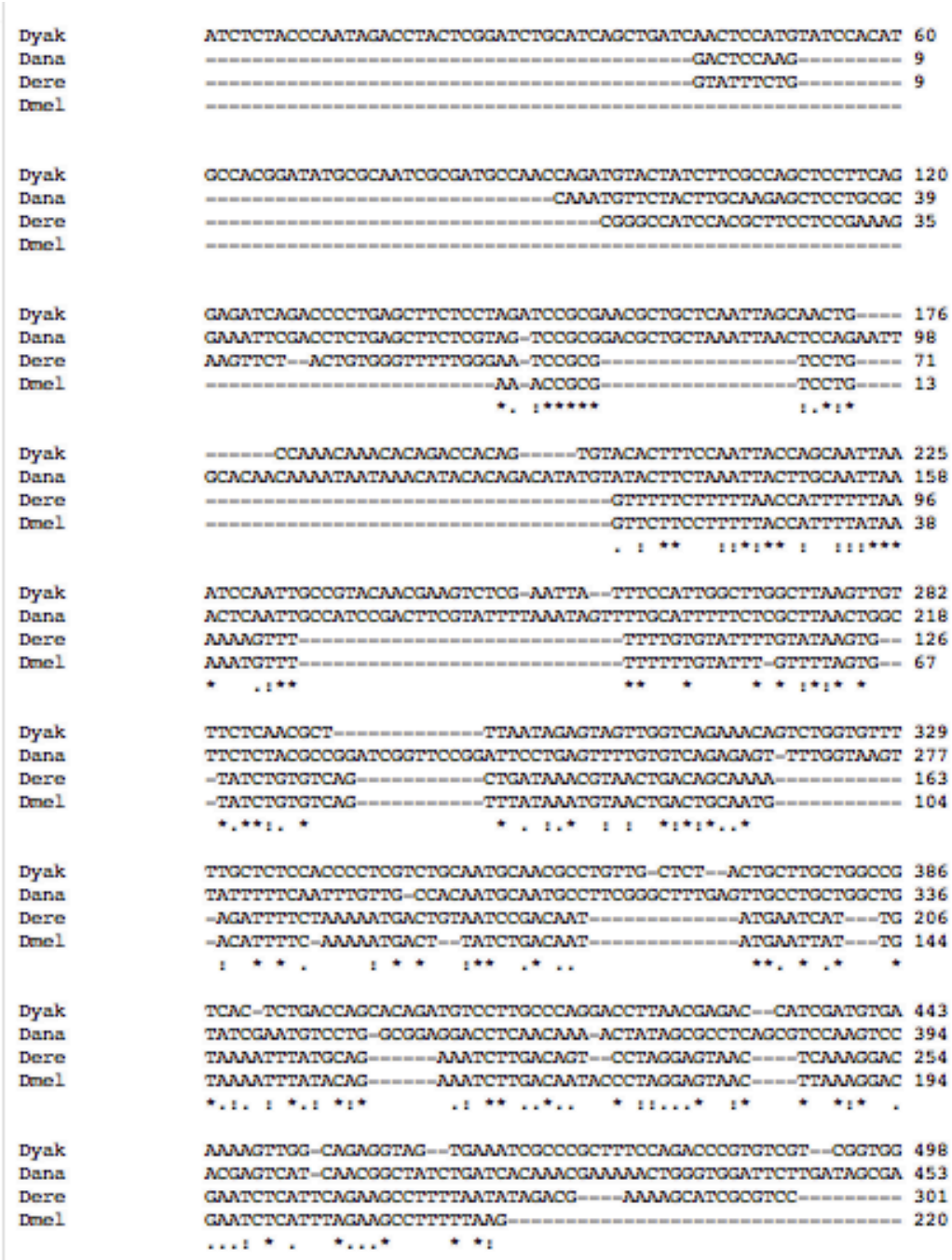


Figure 21: ClustalW2 alignment for the 5` UTR region of the CG10588 ortholog identifies no conclusive elements.

Annotation of CG10589 Ortholog:

A single exon gene codes for the *D. melanogaster* protein CG10589. A BLASTx search of the Fosmid24 sequence against *D. melanogaster* proteins gives a hit to this gene with one major alignment block, between position 3000 and 4200bp. When the CG10589 protein sequence is used as a query via BLAT back to the *D. erecta* assembly there is also one alignment block that is located in the same region as the large alignment block from the BLASTx search. In an attempt to locate the end of the protein, I took a closer look at the nucleotide sequence that corresponds to the end of the BLAT alignment in the Genome Browser window. With this exon

coding in the +1 frame, it was clear there was an in-frame stop codon that truncates this protein relative to the *D. melanogaster* ortholog (Figure 22). Of the 386 residues in the *D. melanogaster* protein, 380 can be accounted for in the *D. erecta* sequence from Fosmid24. A table summarizing the feature is included for reference (Table 5).

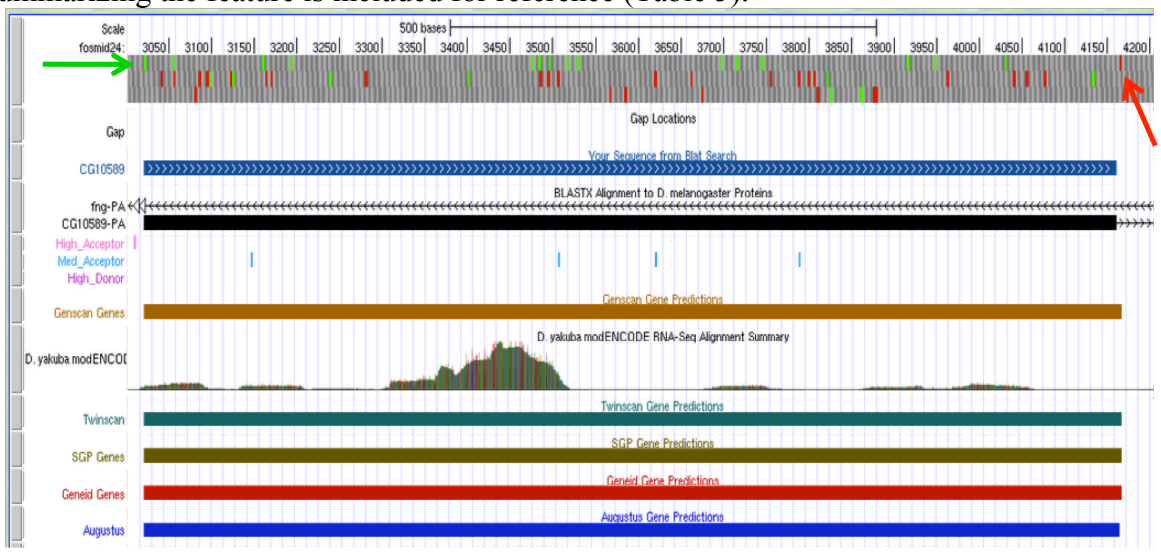


Figure 22: Genome Browser showing the alignment of CG10589 ortholog in the +1 frame with the start codon (green arrow) and stop codon (red arrow) identified.

CG10589	Start	End	Frame	Phase at Start/End	Comments
Exon 1_322_0	3022	4164	+1	Start/Stop	Truncated at residue 380/386

Table 5: Summary of exons from CG10589 ortholog

The coordinates for this gene were submitted to the Gene Model Checker and it has an alignment with no gaps as can be seen in the 2-D dot plot comparison with the *D. melanogaster* protein sequence (Figure 23). Aside from the truncation, this protein is well-conserved.

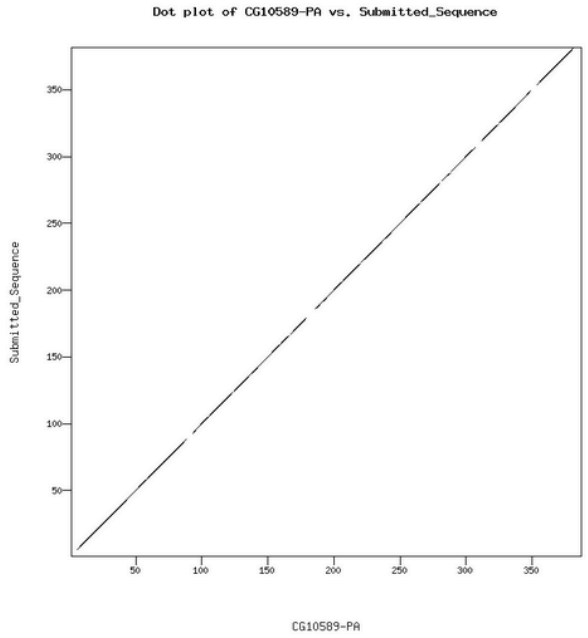


Figure 23: Gene Model Checker output for the CG10589 ortholog displayed as a 2-D dot plot.

Annotation of *fng* Ortholog:

The *D. melanogaster* gene *fng* contains 7 exons, 6 introns, and has only one isoform. The BLASTx track of hits to *D. melanogaster* proteins shows an alignment with blocks located between positions 1500 and 240000bp in the reverse direction, corresponding to the first alignments in the fosmid. GENSCAN predicts three features in this region, with two in the forward strand and one on the reverse strand. However, a tBLASTn search of the *fng* protein against Fosmid24's sequence gave an alignment with complete coverage and an E-value of 7e-70. An examination of the 2-D dot blot confirms that the complete query is contained within the fosmid, although some of the alignment blocks are clearly overextended (Figure 24).

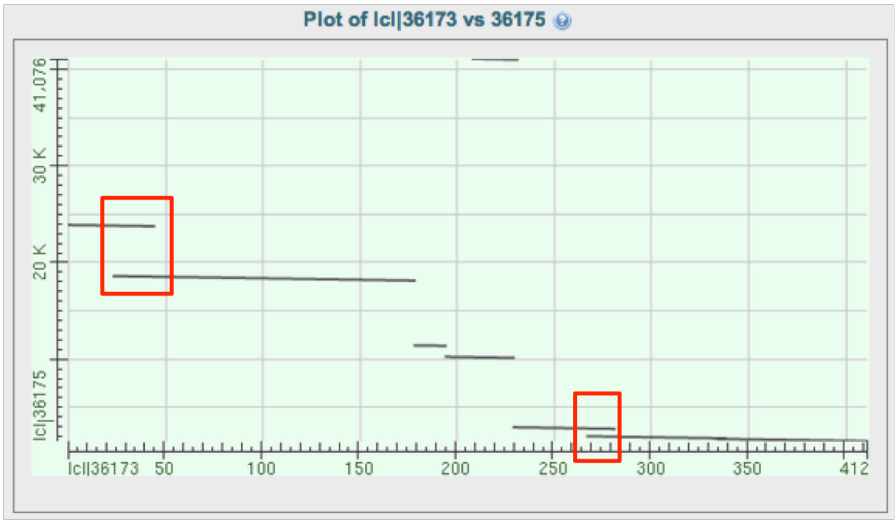


Figure 24: 2-D dot plot from the tBLASTn search of FNG protein against Fosmid24 with overextensions boxed

A BLAT search of the FNG protein sequence against the *D. erecta* 3L extended assembly gives a clear reason for discrepancy between BLAST/BLAT and GENSCAN. The previously annotated CG10589 ortholog is present within the fourth intron region of *fng*. This gene arrangement is homologous and syntenic to that found in the corresponding region in *D. melanogaster,* which supports the annotation. The in-frame start codon is consistent with BLAT and BLAST alignments, while the stop codon is supported by BLAST, BLAT, and the GENSCAN prediction. All intron donor and acceptor sites were determined in the manner demonstrated above. The results are summarized in the table below (Table 6).

fng	**Start**	**End**	**Frame**	**Phase at Start/End**
Exon 1_325_0	23880	23749	-1	Start/0
Exon 2_325_0	18527	18123	-2	0/0
Exon 3_325_0	11415	11367	-1	0/1
Exon 4_325_2	10231	10129	-2	2/2
Exon 5_325_1	2938	2800	-1	1/0
Exon 6_325_0	1980	1806	-1	0/1
Exon 7_325_2	1740	1508	-3	2/Stop

Table 6: Summary of exons from *fng* ortholog

The coordinates for the gene model were submitted to the Gene Model Checker and have an alignment with no gaps (Figure 25). This protein is extremely well-conserved between *D. erecta* and *D. melanogaster* with only two amino acid substitutions between them.

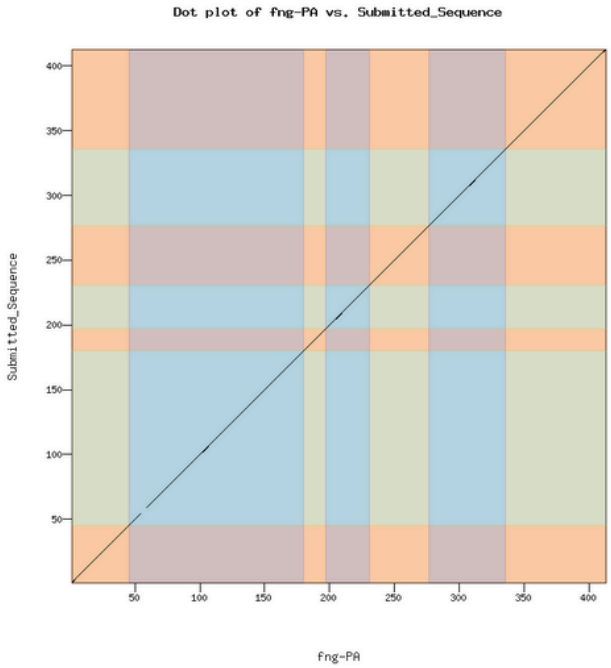


Figure 25: Gene Model Checker output for the *fng* ortholog displayed as a 2-D dot plot.

Identification of a Second Gene Family:

A tBLASTn search of the *D. melanogaster* CG12983 protein against Fosmid24 sequence gives three sets of alignment blocks with the first set being the most complete (Figure 26)

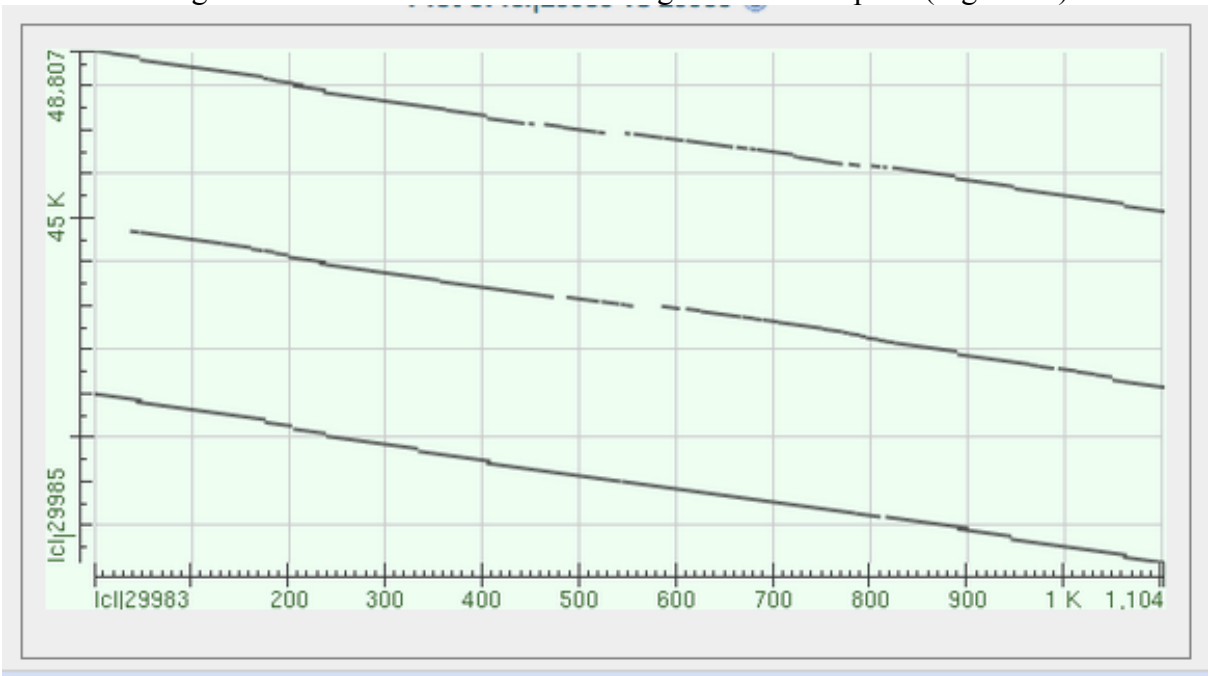


Figure 26: 2-D dot plot from the tBLASTn search of *CG12983* protein sequence against part of Fosmid24.

Two tBLASTn searches of the *D. melanogaster* CG33287 and CG33286 protein sequences against Fosmid24 each give 3 sets of alignment blocks with the second and third set being the most complete respectively. To determine if these features belong to a protein family, a BLASTp search of CG33287 protein against the *D. melanogaster* RefSeq protein database was run. This search revealed that CG33287 contains a conserved Casc1 Superfamily domain and its sequence has significant identity to several proteins that also showed BLASTx hits to Fosmid 24-CG33286 and CG12983 (Figure 27). The Casc1 (Cancer susceptibility candidate 1) domain is primarily found in eukaryotes and is known to have many SNPs that are associated with cancer susceptibility.

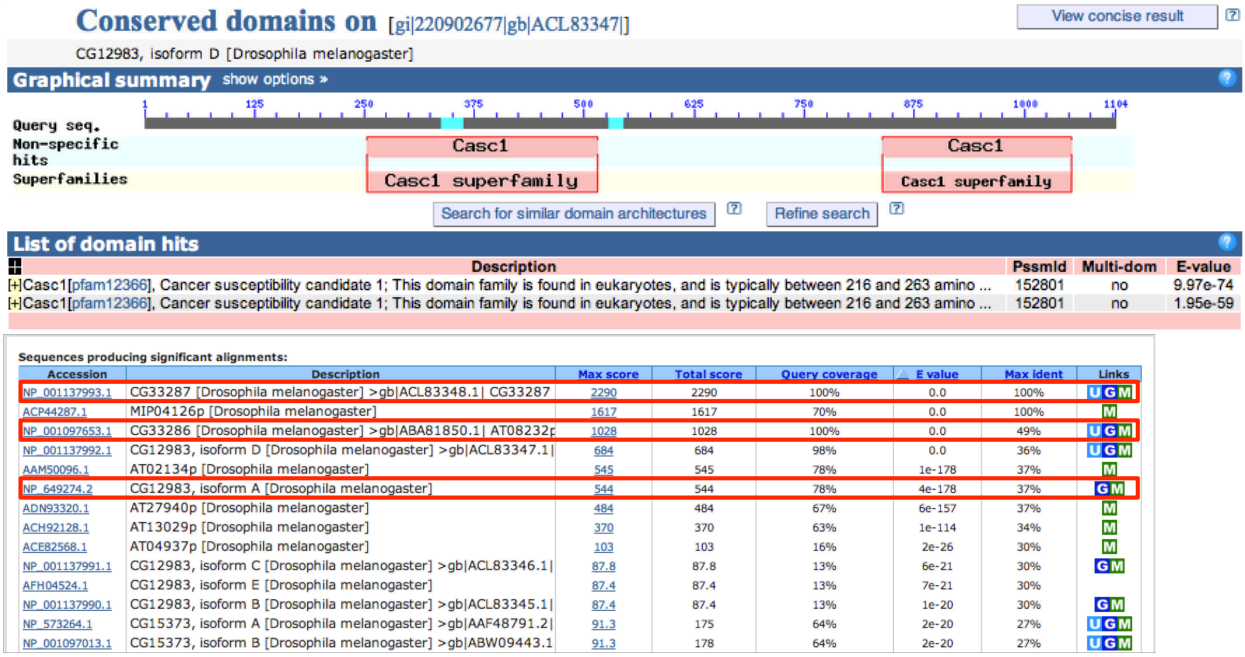
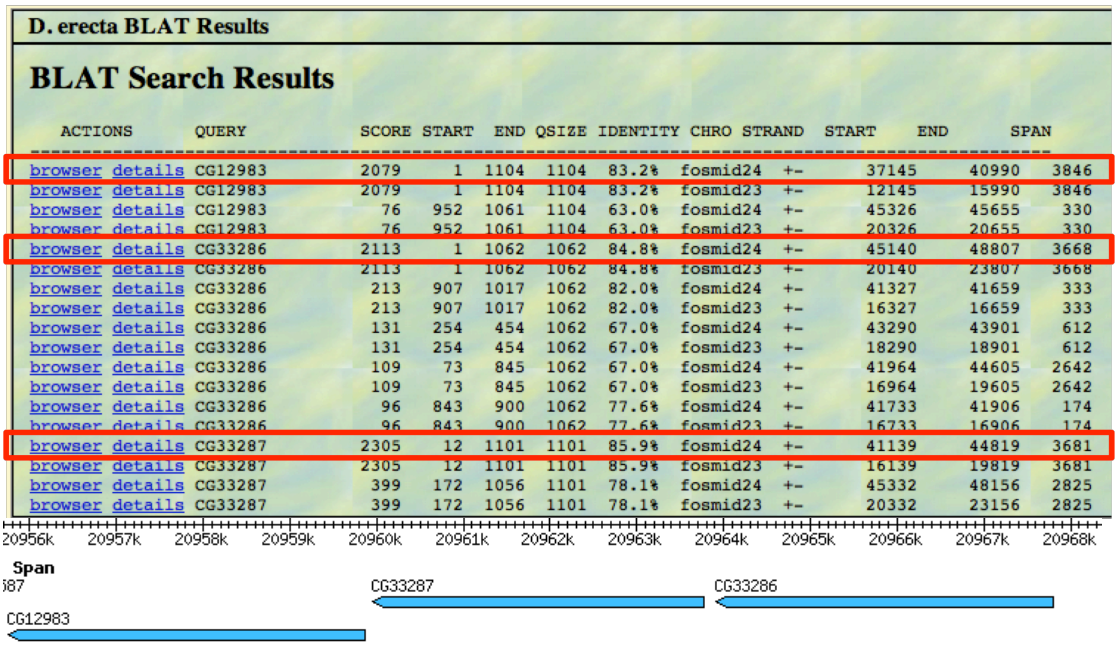


Figure 27: Conserved Cancer susceptibility candidate 1 domain found in *D. melanogaster* protein *CG33287* (top); BLASTp hits to other potential genes found in Fosmid24, CG33286 and CG12983 (bottom).

A BLAT search that used of all three of the protein sequences as queries against the *D. erecta* 3L extended assembly confirms the putative orthology with CG12983, CG33287, and CG33286 ordered from left to right on the assembly (Figure 28, top). This is syntenic with the orthologous region in *D. melanogaster,* as order and orientation of the genes has been conserved (Figure 28, bottom).



D. erecta BLAT Results

BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	CG12983	2079	1	1104	1104	83.2%	fosmid24	+-	37145	40990	3846
browser details	CG12983	2079	1	1104	1104	83.2%	fosmid23	+-	12145	15990	3846
browser details	CG12983	76	952	1061	1104	63.0%	fosmid24	+-	45326	45655	330
browser details	CG12983	76	952	1061	1104	63.0%	fosmid23	+-	20326	20655	330
browser details	CG33286	2113	1	1062	1062	84.8%	fosmid24	+-	45140	48807	3668
browser details	CG33286	2113	1	1062	1062	84.8%	fosmid23	+-	20140	23807	3668
browser details	CG33286	213	907	1017	1062	82.0%	fosmid24	+-	41327	41659	333
browser details	CG33286	213	907	1017	1062	82.0%	fosmid23	+-	16327	16659	333
browser details	CG33286	131	254	454	1062	67.0%	fosmid24	+-	43290	43901	612
browser details	CG33286	131	254	454	1062	67.0%	fosmid23	+-	18290	18901	612
browser details	CG33286	109	73	845	1062	67.0%	fosmid24	+-	41964	44605	2642
browser details	CG33286	109	73	845	1062	67.0%	fosmid23	+-	16964	19605	2642
browser details	CG33286	96	843	900	1062	77.6%	fosmid24	+-	41733	41906	174
browser details	CG33286	96	843	900	1062	77.6%	fosmid23	+-	16733	16906	174
browser details	CG33287	2305	12	1101	1101	85.9%	fosmid24	+-	41139	44819	3681
browser details	CG33287	2305	12	1101	1101	85.9%	fosmid23	+-	16139	19819	3681
browser details	CG33287	399	172	1056	1101	78.1%	fosmid24	+-	45332	48156	2825
browser details	CG33287	399	172	1056	1101	78.1%	fosmid23	+-	20332	23156	2825

Figure 28: BLAT search of all three sequences identifies feature orthology (top); GBrowse output from Flybase showing the genes in the orthologous region of chromosome 3L in *D. melanogaster* (bottom)

Annotation of CG12983 Ortholog:

The *D. melanogaster* gene CG12983 contains four isoforms (B, C, D, and E), the longest of which contains ten exons and nine introns. The BLASTx track of hits to *D. melanogaster* proteins shows an alignment with blocks located between positions 37000 and 41000bp in the reverse direction, while GENSCAN predicts this region is part of a much longer feature. A BLAT search of the individual exon sequences from *D. melanogaster* against the *D. erecta* 3L-Extended assembly was used to identify positions of the donor and acceptor splice sites as well as start and stop codons. All intron donor and acceptor sites were determined in the manner previously described. The results are summarized in the table below (Table 7).

CG12983	Start	End	Frame	Phase	PB	PC	PD	PE	Comments
16_535_0	40990	40852	-3	Start/1	X	X	X	X	
13_535_2	40795	40407	-2	2/0	X		X	X	
11_534_2	40795	40374	-2	2/Stop		X			
12_535_0	40341	40256	-1	0/2			X		
10_536_0	40190	40158	-2	0/Stop				X	8 AA extension
10_535_1	40190	40089	-3	1/2			X		
9_534_0	40028	39894	-2	0/Stop	X				
9_535_1	40028	39741	-3	1/2			X		
8_535_1	39680	39473	-3	1/0			X		
7_535_0	39413	37972	-2	0/1			X		
6_535_1	37919	37748	-3	1/0			X		
5_535_0	37677	37333	-1	0/0			X		GC donor
2_535_0	37273	37145	-3	0/Stop			X		

Table 7: Summary of exons from CG12983 ortholog

For the donor site in exon 5_535_0, I have annotated a non-canonical GC splice site that is not present in *D. melanogaster*. The *D. yakuba* modENCODE RNA-Seq and TopHat Junctions support this as the splice site and there is no evidence of splicing that deviates from these locations. The GC site is in the zero phase relative to the -1 frame, which is required to match up with the subsequent exon. There are two other GT pairs before the first stop codon in that frame but neither is in the zero phase and thus do not match up to the next exon whose acceptor site is in phase zero relative to the -3 frame (Figure 29, top). A canonical GT is found in *D. yakuba* (Figure 29, bottom).

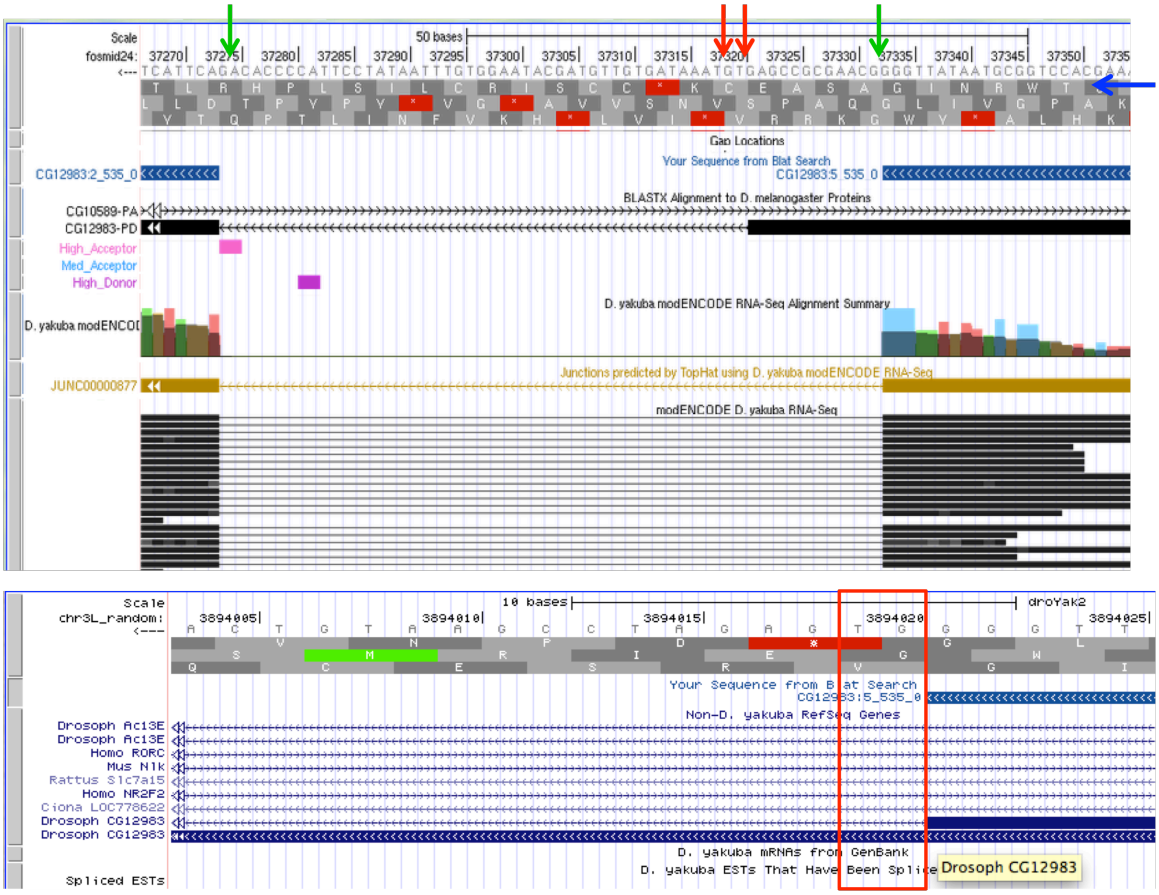


Figure 29: The GC donor and AG acceptor sites annotated for CG12983 ortholog marked in green have matching phases; other candidate GT donors do not match phase with the acceptor site (top); Canonical GC donor is found in *D. yakuba* ortholog (bottom)

The terminal exon of *D. erecta* isoform E (10_536_0) is 8 residues longer than the *D. melanogaster* sequence (Figure 30, top and middle). Because the exon in *D. melanogaster* contains only three amino acid residues, a BLAT search with the nucleotide sequence as a query against the *D. erecta* 3L – Extended assembly was necessary to determine its location. A similar BLAT search of the *D. yakuba* assembly shows that the extension is also contained within its sequence (Figure 30, bottom).

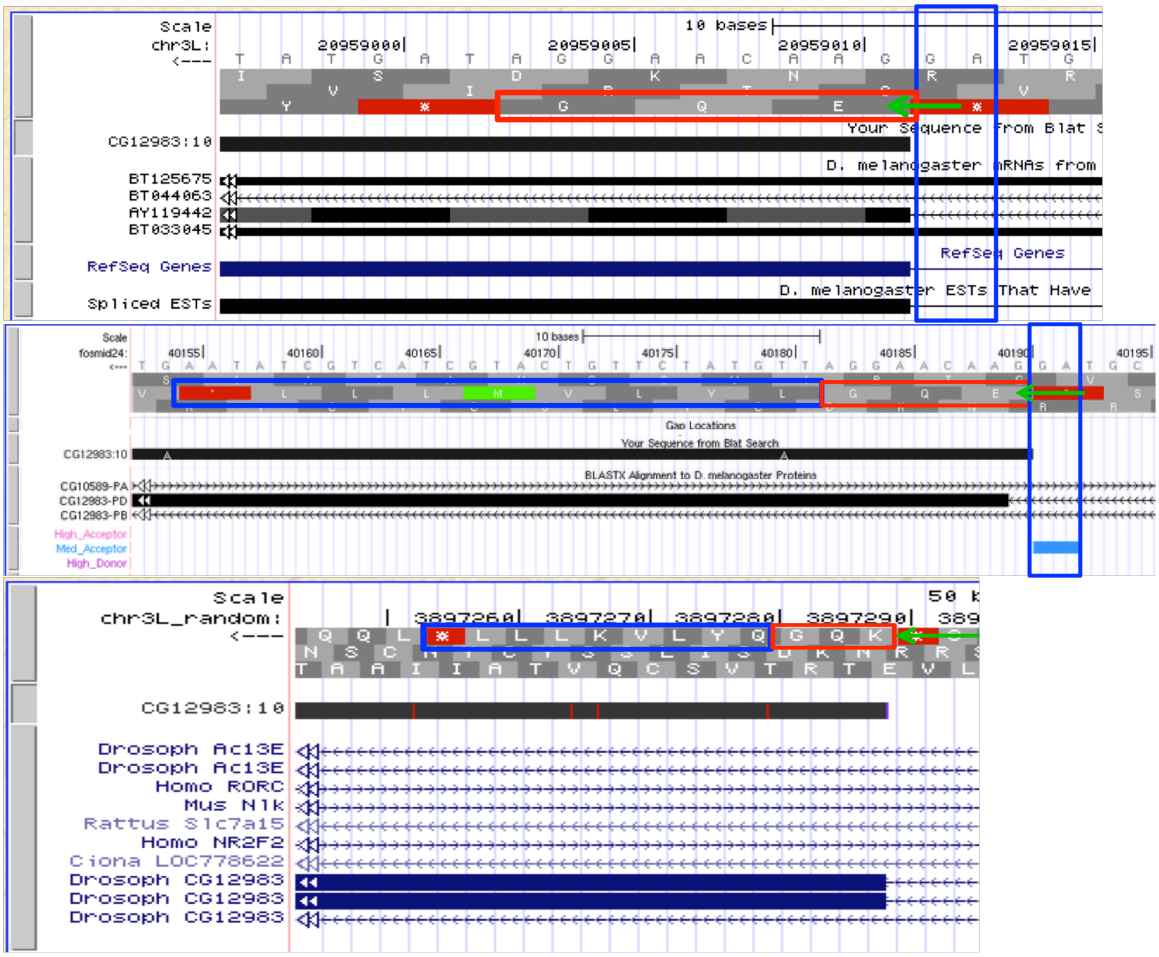
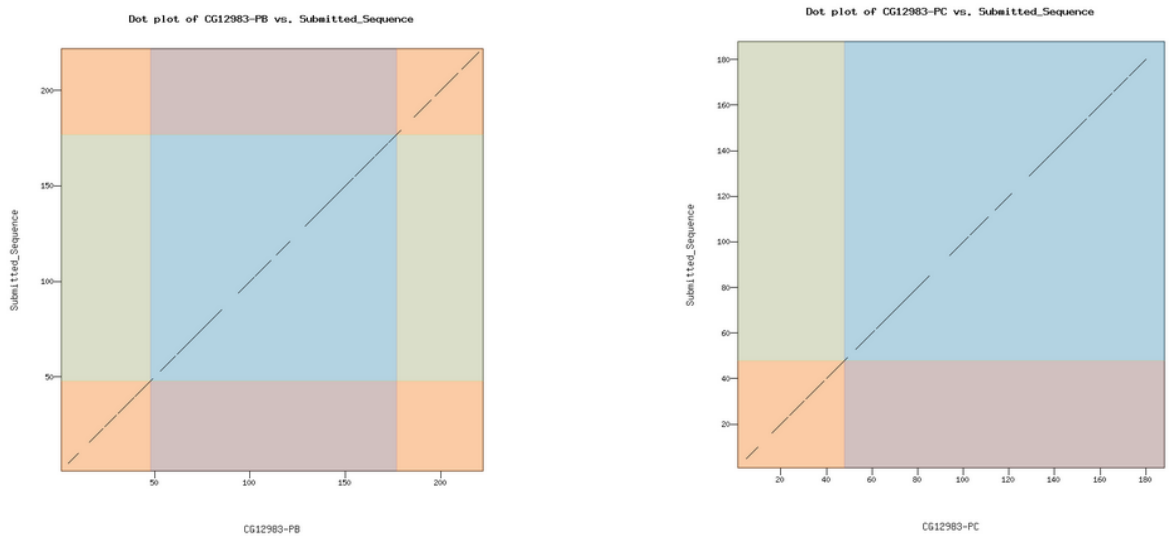


Figure 30: 8 residue extension is found in *D. erecta* (middle) and *D. yakuba* (bottom) relative to the *D. melanogaster* (top) sequence. *D. melanogaster* residues are marked in red and residues in the extension are blue.

The coordinates of all of the exons for each isoform were submitted to the Gene Model Checker with the 2-D dot plot outputs included below (Figure 31). The extension of isoform E is clearly identified in the plot by having a line with slope less than one.



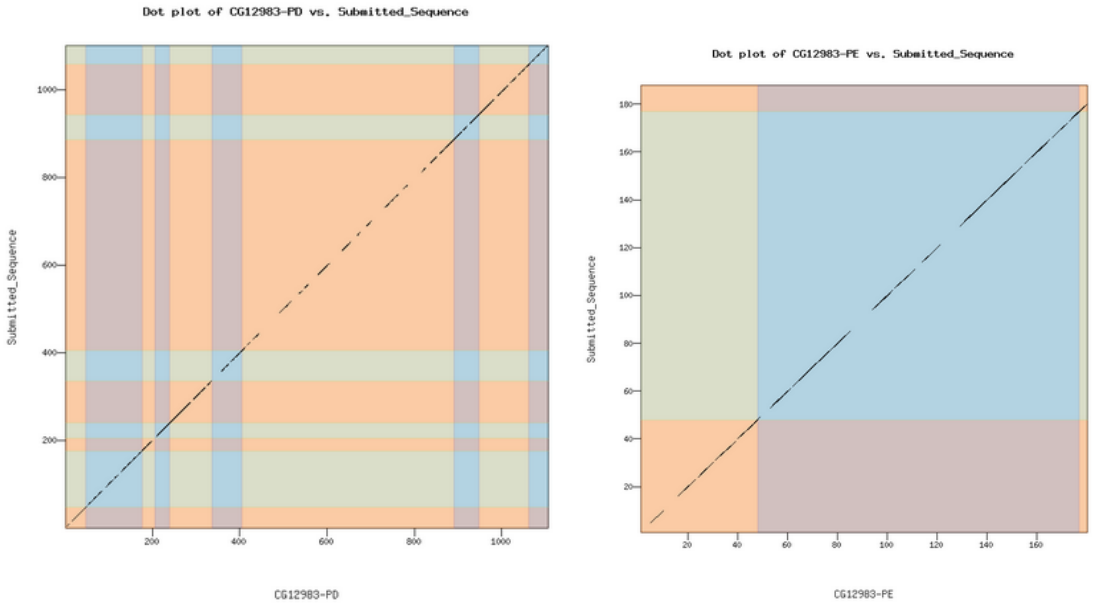


Figure 31: Gene Model Checker output for the CG12983 ortholog displayed as a 2-D dot plot.

Annotation of CG33286:

The *D. melanogaster* gene CG33286 contains one isoform with nine exons and eight introns. The BLASTx track of hits to *D. melanogaster* proteins shows an alignment with blocks located between positions 450000 and 49000bp in the reverse direction, while GENSCAN predicts this region is part of the same longer feature as CG12983. A BLAT search of the individual exon sequences from *D. melanogaster* against the *D. erecta* 3L-Extended assembly was used to identify positions of the donor and acceptor splice sites as well as start and stop codons (Figure 32). All intron donor and acceptor sites were determined in the manner previously described. The results are summarized in the table below (Table 8).

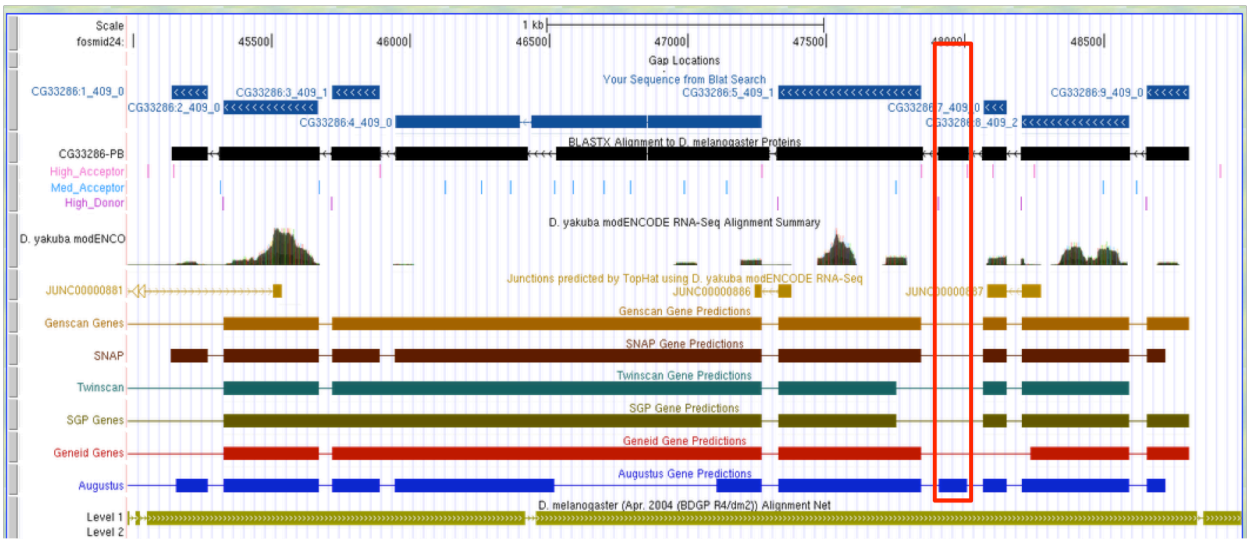


Figure 32: Genome Browser output showing BLAT alignment of individual exons of *D. melanogaster* CG33286 to the *D. erecta* 3L- extended assembly on Fosmid 24. One exon was unaligned by BLAT and is boxed in red.

CG33286	Start	End	Frame	Phase Start/End	Comments
9_409_0	48807	48657	-1	Start/1	
8_409_2	48593	48205	-1	2/0	
7_409_0	48150	48065	-1	0/2	
6_409_1	48006	47905	-2	1/2	Not found by BLAT/GENSCAN
5_409_1	47843	47327	-3	1/0	
4_409_0	47266	45945	-3	0/2	Insertion relative to *D. melanogaster*
3_409_1	45890	45719	-3	1/0	
2_409_0	45670	45326	-3	0/0	
1_409_0	45268	45140	-3	0/Stop	

Table 8: Summary of exons from CG33286 ortholog

Exons 6_409_1 was not identified by GENSCAN, nor was it aligned by BLAT. Both the BLASTx track on the Genome Browser and an external tBLASTn search of the protein sequence against Fosmid 24 identified its position successfully. Exon 4_409_0 has a 13 residue insertion relative to *D. melanogaster* which can be noted by the shift in the line of the 2-D dot plot output from the Gene Model Checker included below (Figure 33)

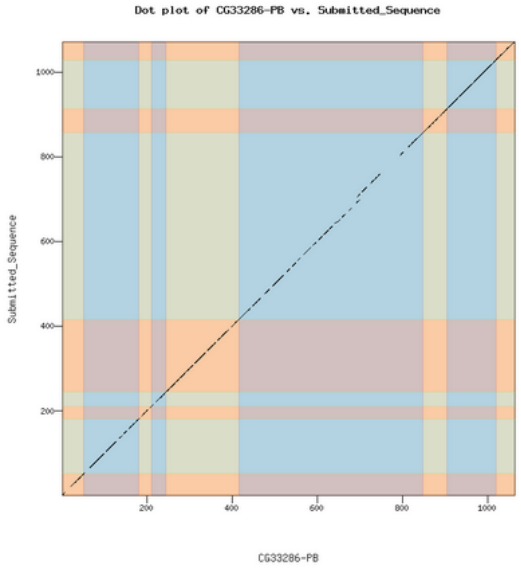


Figure 33: Gene Model Checker output for the CG33286 ortholog displayed as 2-D dot plot.

Annotation of CG33287:

The *D. melanogaster* gene CG33287 contains one isoform with nine exons and eight introns, all with high similarity to CG33286. The BLASTx track of hits to *D. melanogaster* proteins shows an alignment with blocks located between positions 410000 and 45000bp in the reverse direction, while GENSCAN predicts this region is part of the same longer feature as CG12983 and CG33286. A BLAT search of the individual exon sequences from *D. melanogaster* against the *D. erecta* 3L-Extended assembly was used to identify positions of the donor and acceptor splice sites as well as start and stop codons (Figure 34). All intron donor and acceptor sites were determined in the manner previously described. The results are summarized in the table below (Table 9).

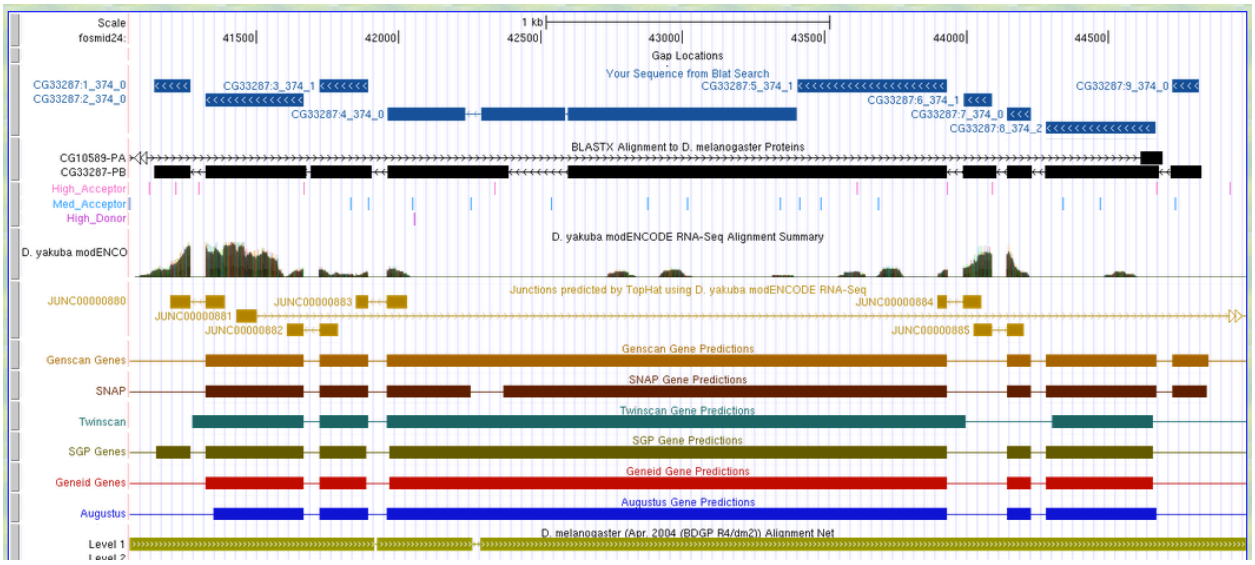


Figure 34: Genome Browser output showing BLAT alignment of individual exons of *D. melanogaster* CG33287 to the *D. erecta* 3L- extended assembly on Fosmid 24.

CG33287	Start	End	Frame	Phase Start/End	Comments
9_374_0	44846	44726	-2	Start/1	
8_374_2	44670	44279	-3	2/0	
7_374_0	44228	44143	-2	0/2	
6_374_1	44090	43989	-3	1/2	
5_374_1/4_374_0	43932	41959	-2	1/2	Merged two exons
3_374_1	41892	41721	-2	1/0	
2_374_0	41665	41321	-3	0/0	
1_374_0	41267	41139	-2	0/Stop	

Table 9: Summary of exons from CG33287 ortholog

It should be noted that the sequences corresponding to *D. melanogaster* exons 5_374_1 and 4_374_0 have been merged in this *D. erecta* ortholog. The individual exons sequences are located in the same frame and are separated by three base pairs to form one contiguous open reading frame. This merge is not present in *D. yakuba* as seen below, with the *D. melanogaster* region included for reference (Figure 35)

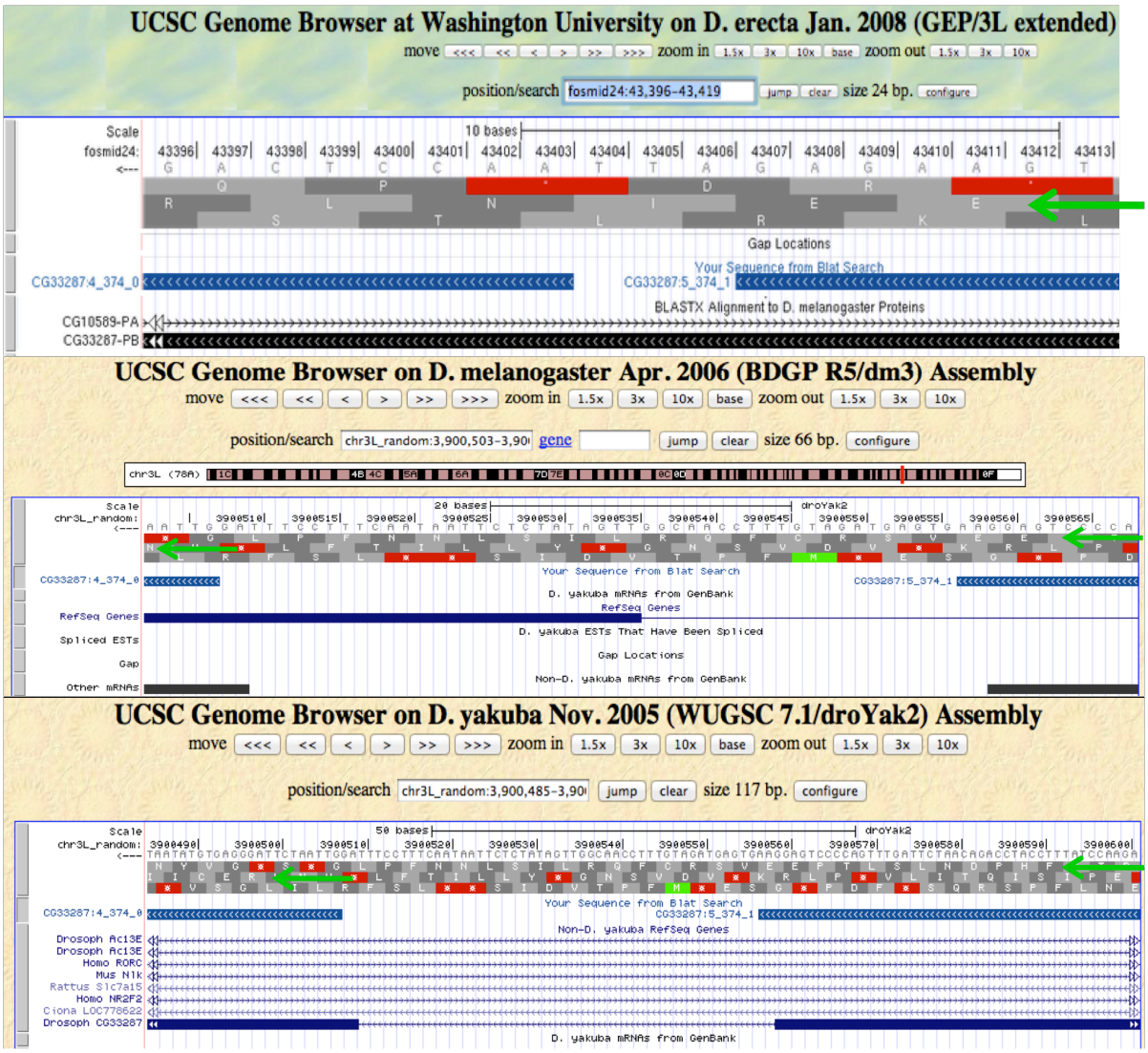


Figure 35: UCSC Genome Broswer and Washington University Mirror outputs showing the merged exon present in *D. erecta* (top) is two separated exons in both *D. melanogaster* (middle) and *D. yakuba* (bottom).

The coordinates for this gene model were submitted to the Gene Model Checker and the output shows a good alignment with no gaps. The 2-D dot plot output is included below (Figure 36).

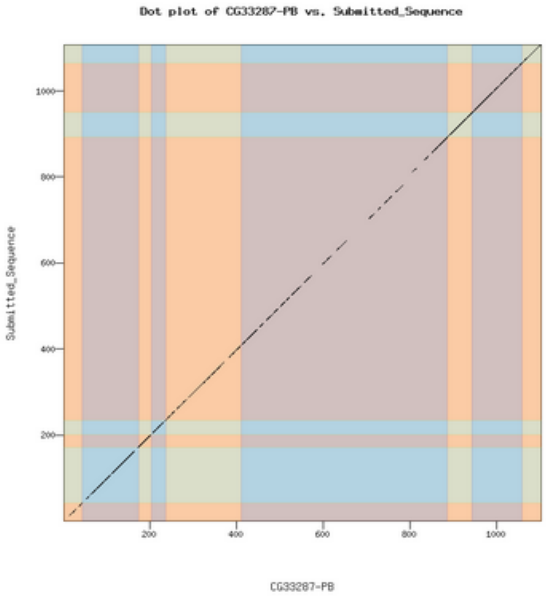


Figure 36: Gene Model Checker output for the CG33287 ortholog displayed as a 2-D dot plot.

Investigation of Other Possible Features:
External NCBI BLASTx searches that used exons called by GENSCAN but not identified by the original BLASTx output track as queries against all *Drosophila* proteins found no experimentally verified hits for any of them. Based on these results, I concluded that they were most likely spurious predictions made by GENSCAN, simply because it recognized them as potential open reading frames. Analysis of the RNA-Seq track mapped from the modENCODE *D. yakuba* data shows several high peaks of expression. All of these major peaks occur within coding genes. The highest peak located at about 35kb in the assembly is from a region in the second exon of the CG11037 gene, which has significant identity to the second exons of the other two proteins in its gene family in the fosmid. No other features or repeats were found by this logic. A figure showing the GENSCAN predictions compared to the final annotations and the modENCODE RNA-Seq Summary track is included for reference (Figure 37).

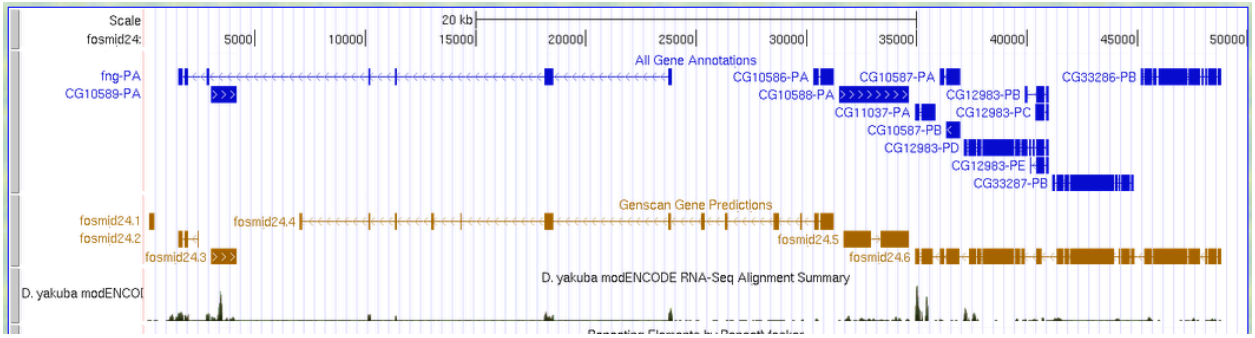


Figure 37: Output from Genome Browser that compares the GECSCAN predictions to the final gene annotations

Clustal Analysis:
The *fng* protein contains a conserved Galactosyl-transferase domain in its C-terminal region. In order to attempt to identify any other conserved regions of the protein, I performed a multiple sequence alignment of *fng* homologs from *D. erecta, D. melanogaster, D. mojavensis, D. virilis, Bombyx mori* and *Anopheles gambiae*. The protein sequences from these five species were submitted to the ClustalW2 alignment program with the graphical output included below (Figure 38).

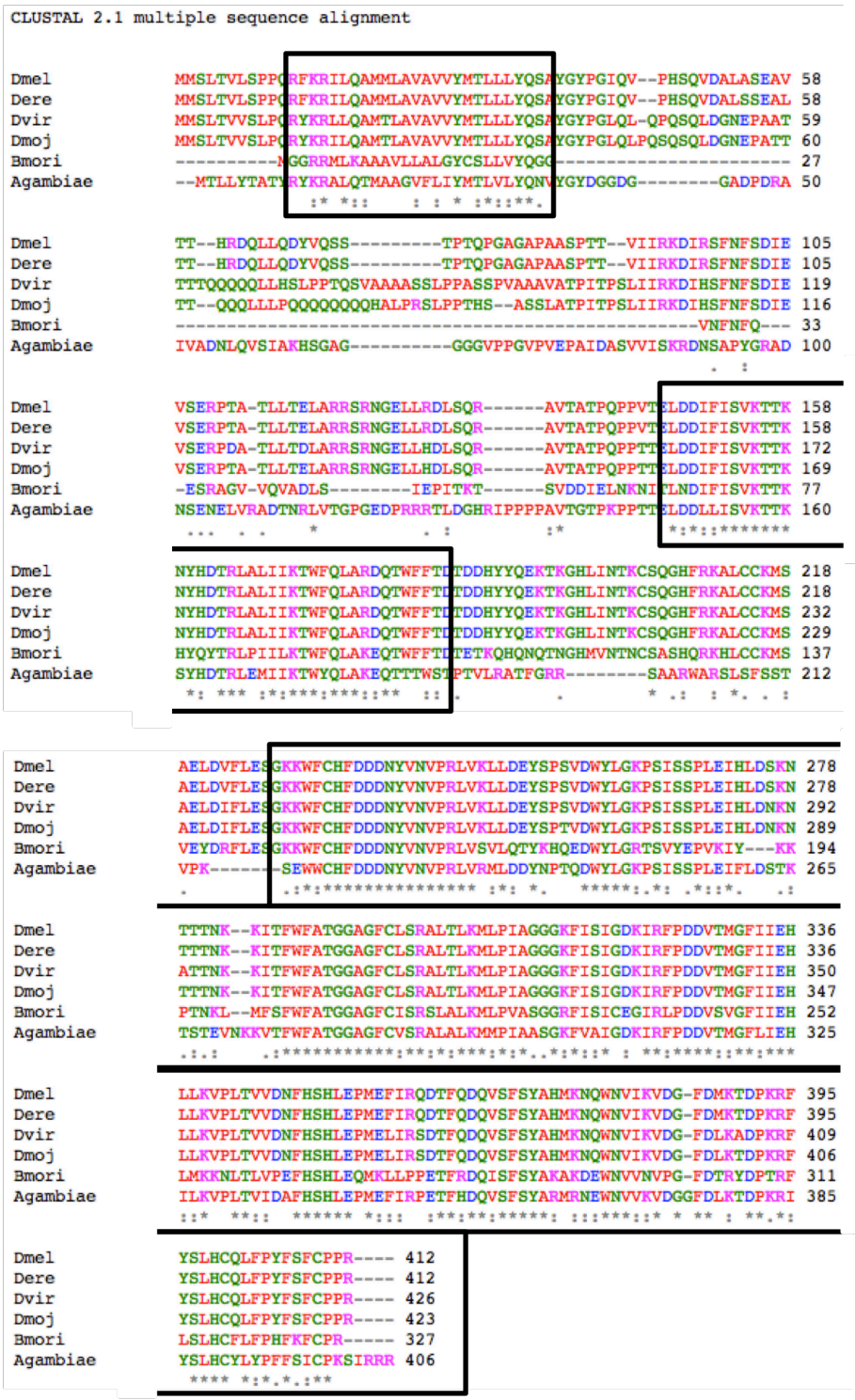
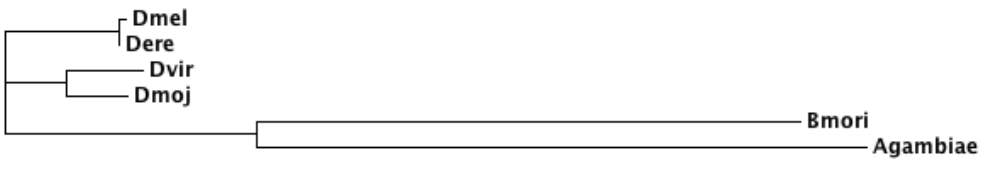


Figure 38a: Output from ClustalW2 analysis with three well-conserved domains indicated.

Figure 38b: Guide tree used for ClustalW2 analysis agrees with known phylogeny.



From the alignment above, it is clear that the galactosyl-transferase domain is well conserved as expected, but there is also a second somewhat conserved domain at the N-terminus of the protein. Although there are some variations in the outgroup sequences from *B. mori* and *A. gambiae* the region still maintains its overall hydrophobic character. It may warrant further investigation to see if this conserved region has some function, perhaps in substrate recognition or the maintenance of correct protein conformation.

Repeats:

Because this fosmid is located in the 3L extended region of *D. erecta,* there are not many repetitious sequences present, as is to be expected. The total repeat content identified by RepeatMasker is 1.92%. None of the identified repeats are over 500bp in length and the only one that falls within a coding exon is an AT-rich region contained within the CG10588 ortholog, whose corresponding amino acid sequence is present in the protein. Repeat content is summarized in the following output table from RepeatMasker (Figure 39/ Table 10)

```
==================================================
file name: RM2_fosmid24.fasta_1335626785
sequences:             1
total length:      50001 bp  (50001 bp excl N/X-runs)
GC level:         45.67 %
bases masked:        687 bp ( 1.37 %)
==================================================
                number of      length   percentage
                elements*    occupied  of sequence
--------------------------------------------------
Retroelements            2          230 bp    0.46 %
   SINEs:                0            0 bp    0.00 %
   Penelope              0            0 bp    0.00 %
   LINEs:                2          230 bp    0.46 %
    CRE/SLACS            0            0 bp    0.00 %
     L2/CR1/Rex          0            0 bp    0.00 %
     R1/LOA/Jockey       2          230 bp    0.46 %
     R2/R4/NeSL          0            0 bp    0.00 %
     RTE/Bov-B           0            0 bp    0.00 %
     L1/CIN4             0            0 bp    0.00 %
   LTR elements:         0            0 bp    0.00 %
     BEL/Pao             0            0 bp    0.00 %
     Ty1/Copia           0            0 bp    0.00 %
     Gypsy/DIRS1         0            0 bp    0.00 %
       Retroviral        0            0 bp    0.00 %

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

Rolling-circles          0            0 bp    0.00 %

Unclassified:            0            0 bp    0.00 %

Total interspersed repeats:          230 bp    0.46 %


Small RNA:               0            0 bp    0.00 %

Satellites:              0            0 bp    0.00 %
Simple repeats:          5          168 bp    0.34 %
Low complexity:          7          289 bp    0.58 %
==================================================
```

Figure 39: Summary of repetitious content of Fosmid24 as identified by RepeatMasker (above)

Table 10: Summary of individual repeat elements in Fosmid24 (below)

ID	size	SW score	perc div.	perc del.	perc ins.	start	end	matching repeat	repeat class/family
1	36	195	16.2	0	0	2545	2581	(CTG)n	Simple_repeat
2	20	21	0	0	0	4905	4925	AT_rich	Low_complexity
3	36	23	5.4	0	0	6849	6885	AT_rich	Low_complexity
4	35	207	16.7	0	0	9428	9463	(CTG)n	Simple_repeat
5	22	207	0	0	0	11203	11225	(CGGT)n	Simple_repeat
6	41	240	14.3	0	0	14625	14666	(TTGG)n	Simple_repeat

7	118	381	19	6.7	2.5	16795	16913	TART_DV	LINE/telomeric
8	54	55	0	0	0	18605	18659	AT_rich	Low_complexity
9	87	281	24.2	0	0	20166	20253	dvir.8.41.centroid	Other
10	29	195	10	0	0	24323	24352	(TC)n	Simple_repeat
11	26	27	0	0	0	24951	24977	AT_rich	Low_complexity
12	40	27	4.9	0	0	27325	27365	AT_rich	Low_complexity
13	27	28	0	0	0	32936	32963	AT_rich	Low_complexity
14	108	650	5.5	21.1	0	49054	49162	G4_DM	LINE/Jockey
15	152	266	23.6	3.9	8.5	49294	49446	dana.92.139.centroid	DNA
16	112	29	10.6	0	0	49457	49569	AT_rich	Low_complexity

RepeatMasker was run on the orthologous region in *D. melanogaster*; there are a similar number of retroelements, though it contains more low complexity regions as can be seen in the comparison of the two Genome Browser outputs below (Figure 40). The total repeat density is comparable, with 1.92% repeats in *D. erecta* and 1.45% in *D. melanogaster*. Both species have a gene density of 9 in 50kb. The *D. melanogaster* RepeatMasker output is included for reference (Figure 41).

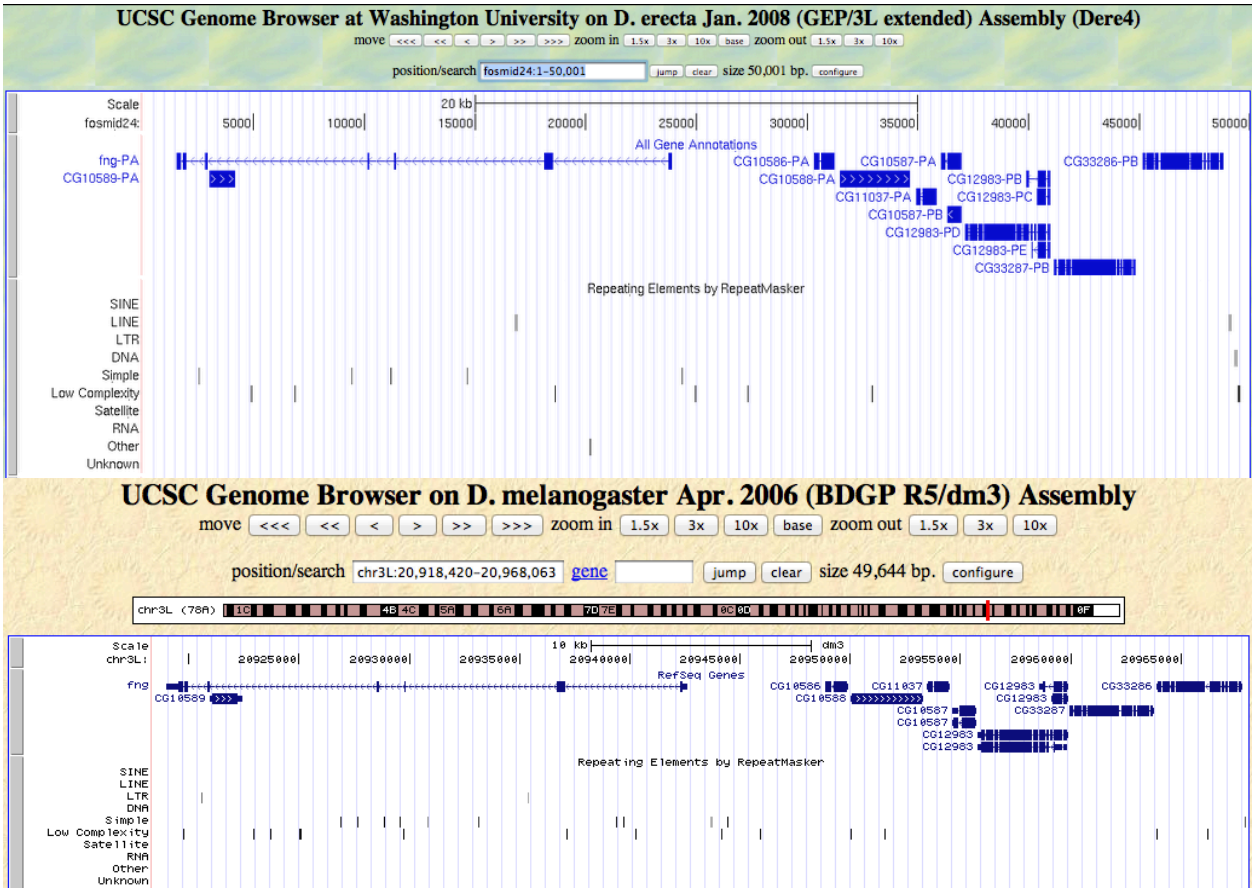


Figure 40: Genome Browser view comparing gene annotations and repeat elements between *D. erecta* (top) and *D. melanogaster* (bottom)

```
==================================================
file name: D.melanogaster
sequences:             1
total length:      55001 bp  (55001 bp excl N/X-runs)
GC level:         43.38 %
bases masked:        799 bp ( 1.45 %)
==================================================
                number of      length   percentage
                elements*    occupied  of sequence
--------------------------------------------------
Retroelements            3          244 bp    0.44 %
   SINEs:                0            0 bp    0.00 %
   Penelope              0            0 bp    0.00 %
   LINEs:                2          193 bp    0.35 %
    CRE/SLACS            0            0 bp    0.00 %
      L2/CR1/Rex         0            0 bp    0.00 %
      R1/LOA/Jockey      2          193 bp    0.35 %
      R2/R4/NeSL         0            0 bp    0.00 %
      RTE/Bov-B          0            0 bp    0.00 %
      L1/CIN4            0            0 bp    0.00 %
   LTR elements:         1           51 bp    0.09 %
      BEL/Pao            1           51 bp    0.09 %
      Ty1/Copia          0            0 bp    0.00 %
      Gypsy/DIRS1        0            0 bp    0.00 %
        Retroviral       0            0 bp    0.00 %

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

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

Rolling-circles          0            0 bp    0.00 %

Unclassified:            0            0 bp    0.00 %

Total interspersed repeats:          244 bp    0.44 %


Small RNA:               0            0 bp    0.00 %

Satellites:              0            0 bp    0.00 %
Simple repeats:         11          403 bp    0.73 %
Low complexity:          6          194 bp    0.35 %
--------------------------------------------------
```

Figure 41: Summary of repetitious content of the orthologous region in *D. melanogaster* as identified by RepeatMasker

Synteny:

The order and orientation of these genes is consistent with the *D. melanogaster* chromosome 3L assembly as seen in the comparison below that is adapted from the GBrowse output from Flybase (Figure 42).

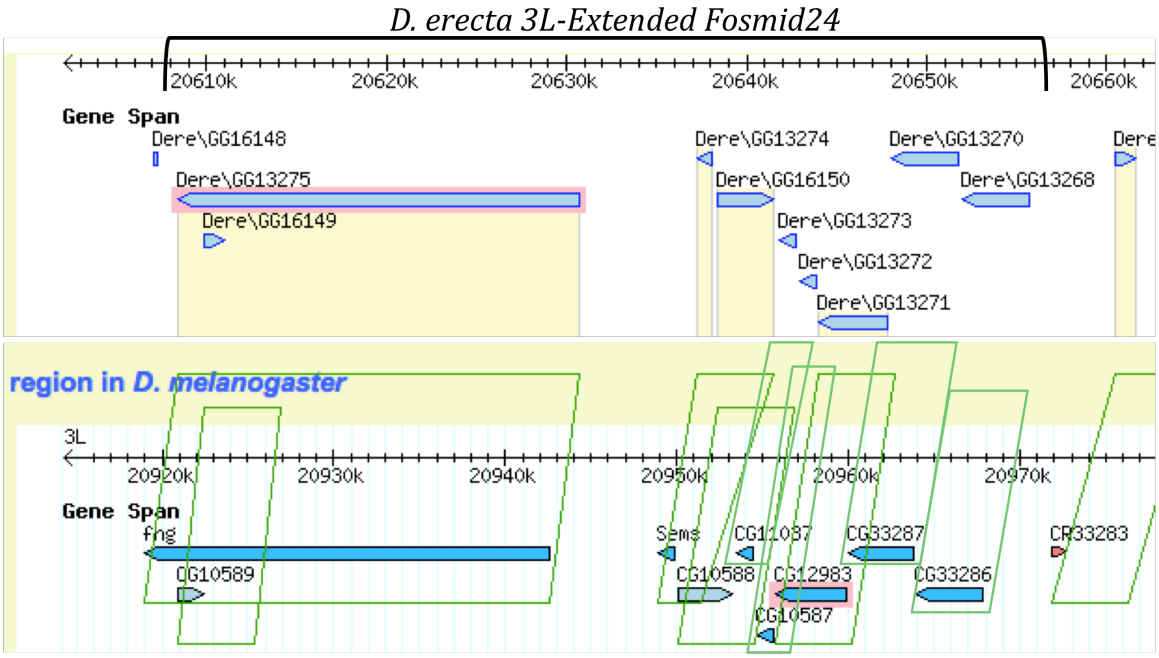


Figure 42: Synteny analysis based on GBrowse output from Flybase shows synteny is conserved between species

A more thorough analysis of synteny in this region was necessary to understand the evolution of the two gene families contained within the fosmid sequence. To find the region in other species that corresponds to the fosmid, I used BLAT on the official UCSC genome browser to search the genome assemblies of various *Drosophila* species with the peptide sequence from the *D. erecta* CG11037 ortholog I annotated. Once I identified the region with the best match, I zoomed out to get a view of the surrounding area. One of the display tracks showed hits to annotated *D. melanogaster* proteins and because these other assemblies have not been individually annotated, I used the order and orientation of these hits to assess synteny of the genes (Figure 43). It should be noted that because this is a gene family with highly similar adjacent sequences, there is an increased probability for miss-assembly by collapsed repeats. Because I do not know the quality of these publicly available assemblies, the following analysis assumes the published assemblies are correct in this region.

While the *D. erecta, D. yakuba* and *D. melanogaster* regions are completely syntenic, this region is slightly different in more distantly related species. In *D. ananassae*, the only difference is the absence of the CG10588 ortholog, which has a better match to a different scaffold assembly. The two gene families are present in full (Figure 43). In other species such as *D. persimilis, D. pseudoobscura, D.virilis,* and *D. mojavensis* not only is the CG10588 ortholog absent but the gene families have fewer members. There is only one copy of a *Sems* family protein, suggesting that there was a triplication in this region that likely occurred after the divergence of the melanogaster group from the obscura group in the *Drosophila* phylogeny. In the other gene family, there is only one copy of the CG33286/CG33287 paralogs in addition to CG12893 ortholog. From sequence comparisons, it is clear that CG33286 and CG33287 are more closely related to each other than either is to CG12983. This is consistent with a their creation via a duplication event that occurred in the same time frame as the Sems family triplication. An earlier duplication is likely responsible for differentiation of the CG33286/CG33287 ancestor and the CG12983 ortholog from a previous single gene progenitor.

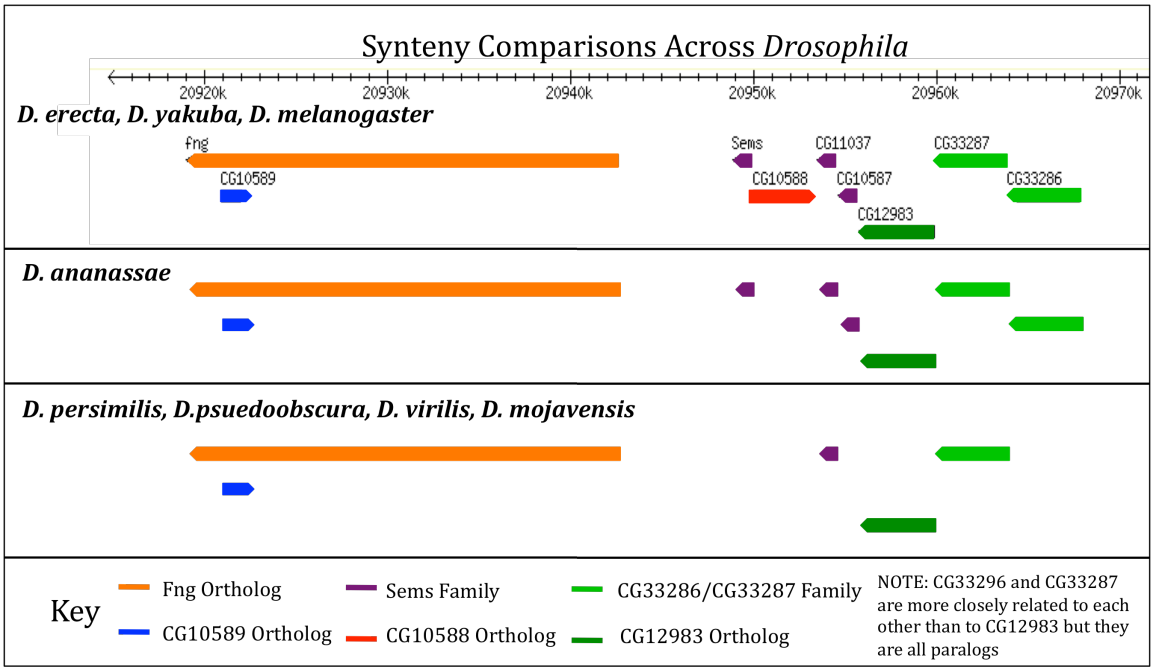


Figure 43: Figure comparing gene order and orientation across multiple species of *Drosophila*

Discussion:

Based on this annotation of Fosmid 24, the results support the high degree of similarity one expects to find between species as closely related as *D. melanogaster* and *D. erecta*. Because this fosmid is from a euchromatin control region located on chromosome three, there is a higher relative gene density and lower repetitious content as compared to the fourth chromosome, whose evolution is the GEP's main focus of study. Synteny has been conserved and there are no wanderer genes whose chromatin state warrants further study like those observed by Leung et el (2010) in a comparative study between *D. melanogaster* and *D. virilis*. The only significant differences from *D. melanogaster* presently identified are several truncated proteins that have lost residues from their terminal exons, one terminal exon extension, one non-canonical GC splice site, and one case of two merged exons. One other feature of interest is the high levels of conservation found in the Trypsin-like Serine Protease Domain containing gene family and the *fng* protein. In the case of the *fng* protein, Clustal analysis suggests there may be at least one conserved region apart from the putative galactosyl-transferase domain.

Based on the in-depth synteny analysis, a further investigation of the history of the two gene families in this fosmid could certainly yield interesting results. My preliminary analysis suggests the origin of the *Sems* family was after the divergence of the *melanogaster* and *obscura* subgroups, while the other group of paralogs likely occurred via two distinct events with enough time between them to account for the significant differences between the CG12983 and CG33297/CG33286 sequences. The insertion of the CG10588 likely occurred after the split of the *D. ananassae* lineage from *D. melanogaster, D. erecta*, and *D. yakuba* as its ortholog is found on a different scaffold. When the *D. ananasse* sequence is better assembled it would be interesting to investigate the chromosomal locus of this gene and perhaps investigate the mechanism of this translocation.

Ultimately, this region offers the GEP a base for comparing euchromatin and heterochromatin. Analysis of the *D. erecta* and other *Drosophila* dot chromosomes hopes to uncover new information about the maintenance and evolution of heterochromatin domains and to further characterize the genes expressed from what appears to be a heterochromatic state. To do this, well-annotated, euchromatic control regions are a necessary tool and for this reason the findings for this region in *D. erecta* are of interest.

Appendix: GFF, pep and fasta files for all genes to be submitted to Wilson Leung upon completion of the project.

Acknowledgements: Special thanks to William Barshop, Katherine Geist, Wilson Leung, Dr. Chavez, Prof. Shaffer, and Prof. Elgin for all their help with this project.

References:

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

Leung, Wilson *et al.* 2010. Evolution of a Distinct Genomic Domain in Drosophila: Comparative Analysis of the Dot Chromosome in *Drosophila melanogaster* and *Drosophila virilis*." Genetics. 185: 1519-1534.