Ryan Friedman
Bio 434W
Dr. Elgin
8 April 2016

Final Annotation Report of *Drosophila ficusphila* contig18

Abstract

Contig18, a 40 kb region of the Muller F element (dot chromosome) in *Drosophila ficusphila*, was annotated using a variety of bioinformatics tools, including *BLAST*, the Genomics Education Partnership mirror of the *UCSC Genome Browser*, FlyBase, *ab initio* gene predictors, *RepeatMasker*, and RNA-seq data. The *D. melanogaster* genome, which was used as a reference, contains four full genes and a non-coding RNA (ncRNA) in its orthologous region. All four genes, *NfI*, *Syt7*, *Rad23*, and *Zip102B*, were found in contig18. Every isoform and every exon in *D. melanogaster* was found and annotated in *D. ficusphila*. The ncRNA, *CR44023*, was not found in contig18 based on a *BLAST* search. However, *CR44023* contains a DINE transposable element in *D. melanogaster* and thus could be found elsewhere in *D. ficusphila*. The transcription start sites (TSS) for *NfI* and *Rad23* were found to be peaked promoters, as was one TSS for *Zip102B*. *D. melanogaster* has two TSS for *Zip102B*, but further analysis is necessary to determine if a second *Zip102B* TSS is found in contig18. *RepeatMasker* identified 34.72% of contig18 as repetitious elements based on a library for *D. ficusphila*. Nine repeats are over 500 bp long and are considered remnants of transposable elements. These long repeats are 98.52% of all repeats found by *RepeatMasker*. The pattern of genes in this region is syntenic with *D. melanogaster* and overall shows significant conservation.

Introduction

In eukaryotes, chromatin structure and packaging is a vital component of gene regulation, expression, and transcription. Chromatin can be broadly classified into two categories: euchromatin and heterochromatin. Active gene transcription is associated with euchromatin, where the chromatin is loosely packaged and therefore easily accessible to RNA polymerase and transcription factors (TFs). In contrast, gene silencing is associated with heterochromatin, apparently due to its dense packaging. The *Drosophila* genus Muller F element (dot chromosome) is almost entirely heterochromatic, yet contains approximately 80 genes which are actively expressed. Therefore, an intra-genus comparative analysis of this unique genomic region can help further our understanding of the relationship between chromatin structure and gene regulation.

The advancement of computational techniques and next-generation genome sequencing technology has allowed for the sequencing and assembly of many *Drosophila* species, which can now be annotated and analyzed for conserved sequence motifs. In turn, this may reveal how F element genes are regulated and expressed in such a unique environment. *D. melanogaster* has been used as the reference species for gene this annotation. This annotation of a 40 kb region of the F element in *D. ficusphila* is a part of a larger project to analyze the assembly of the chromosome as a whole. *D. ficusphila* is separated from *D. melanogaster* by approximately 10-15 million years; this is thought to be enough evolutionary time for the species to have diverged but still share recognizable conserved motifs.

The 40 kb region, hereby referred to as contig18, contains three gene-like features as predicted by *Genscan* and *N-SCAN*, two *ab initio* gene predictors. However, conservation

suggests there are four genes (Figure 1). Feature 3 was annotated first due to a general agreement between gene predictors and conservation. All four features were annotated following the methodology described for Feature 3.

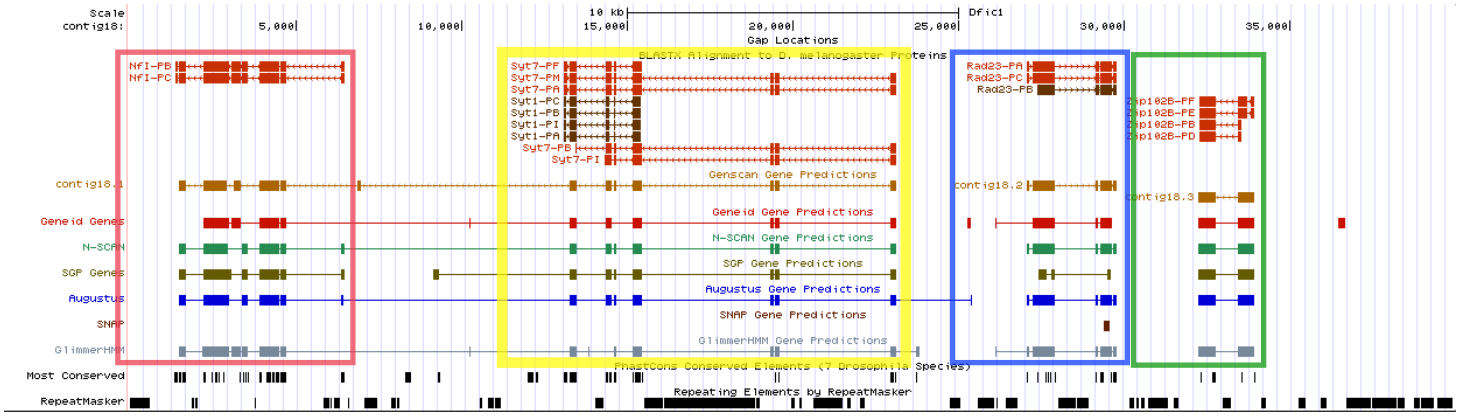


Figure 1: Contig18 in the *UCSC Genome Browser* GEP mirror *D. ficusphila* Jan. 2016 GEP/Dot assembly. *Genscan* and *N-SCAN* predict three gene-like features, yet conservation suggests four features exist. Features 1, 2, 3, and 4 are boxed in red, yellow, blue, and green, respectively.

Feature 3

An initial analysis of the properties of Feature 3 can be done using the GEP mirror of the *UCSC Genome Browser* (Figure 2). *Genscan* and *N-SCAN* both predict five exons on the forward strand spanning approximately 2.5 kb, which agrees with the RNA-seq data. Feature 3 appears to be highly transcribed and is surrounded by repetitious elements identified by *RepeatMasker*; it is an excellent example of a highly expressed gene in a heterochromatic environment.

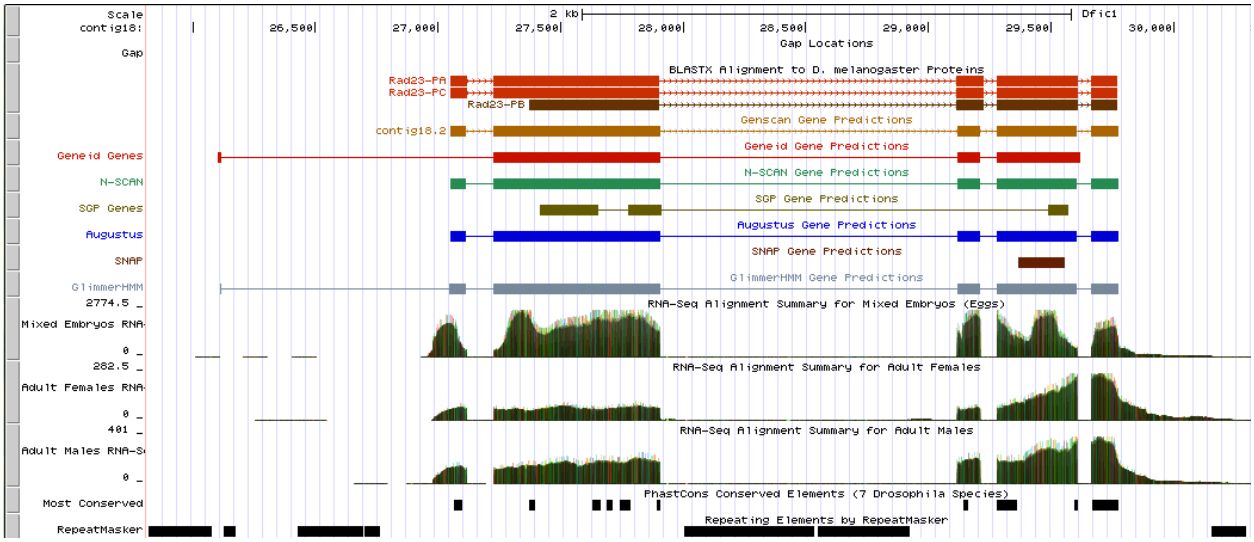
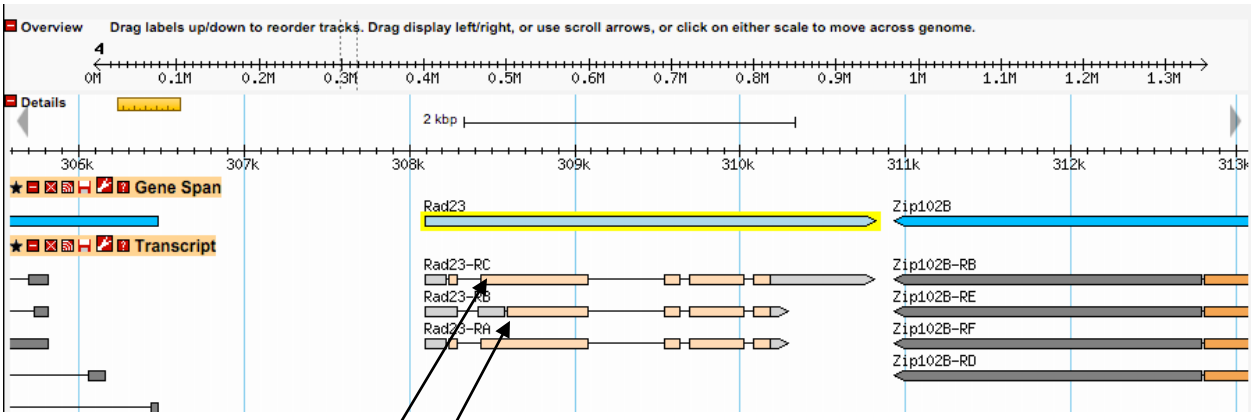


Figure 2: Closer view of Feature 3. *Genscan* and *N-SCAN* predictions closely agree with RNA-seq data.

BLAST Hit Summary				
Description	Species	Score	E value
Rad23-PC	Dmel	598.971	2.64624e-171
Rad23-PA	Dmel	598.971	2.64624e-171
Rad23-PB	Dmel	457.603	8.4015e-129
CG10694-PA	Dmel	110.153	3.82349e-24
RpL40-PB	Dmel	48.1358	1.74023e-05
RpL40-PA	Dmel	48.1358	1.74023e-05
RpS27A-PA	Dmel	45.4394	0.000102051

Figure 3: *BLASTP* results of searching the *Genscan* predicted peptide sequence of Feature 3 (query) to the *D. melanogaster* AA database (subject). *Rad23* isoforms have the highest scores and significantly smaller E-values; *Rad23* is therefore the likely ortholog of Feature 3.

A.



B.

CDS usage map:

Isoform	1_2346_0	2_2346_0	3_2346_0	4_2346_2	5_2346_0	6_2346_0
Rad23-PC	1	2		3	4	5
Rad23-PA	1	2		3	4	5
Rad23-PB			1	2	3	4

Isoforms with unique coding exons:

Unique isoform(s) based on coding sequence	Other isoforms with identical coding sequences
Rad23-PC	Rad23-PA
Rad23-PB	

Select a row to display the corresponding CDS sequence:

FlyBase ID	5' Start	3' End	Strand	Phase	Size (aa)
1_2346_0	308,224	308,289	+	0	22
2_2346_0	308,431	309,079	+	0	216
4_2346_2	309,539	309,633	+	2	31
5_2346_0	309,694	310,017	+	0	108
6_2346_0	310,079	310,189	+	0	37

Figure 4: FlyBase Annotation of *Rad23* in *D. melanogaster*. (A) Browser view of *Rad23*'s location on the F element. Light orange corresponds to coding regions and grey indicates UTRs. Arrows show 3_2346_0 occurs within 2_2346_0. **(B)** FlyBase annotation of coding exon usage. Isoforms A and C have identical coding sequences. Annotation of the isoform C coding exons is shown in the bottom table of the figure.

The predicted peptide sequence of Feature 3 was used as a query in a *BLASTP* search of the annotated proteins (AA) database of *D. melanogaster* in FlyBase (Figure 3). The A and C isoforms of *Rad23* had the lowest E-value, followed by the B isoform. All subsequent matches have much larger E-values, so the ortholog of Feature 3 is very likely *Rad23*. Preliminary *BLASTX* alignments of contig18 (query) to *D. melanogaster* proteins (subject) shown in the *UCSC Genome Browser* indicate that the A and C isoforms are larger than the B isoform, which would explain why the B isoform has a larger E-value (Figure 2).

According to the GEP *Gene Record Finder V1.3* and FlyBase Release 6.08, the *Rad23* gene is located on the plus strand of the fourth chromosome (F element) in *D. melanogaster*, which is further evidence that Feature 3 is an ortholog of *Rad23* (Figure 4). The A and C isoforms of *Rad23* have identical five-exon coding sequences and only differ in the downstream untranslated region (UTR). The B isoform is smaller and only contains four exons. The first exon of the B isoform (coding exon 3_2346_0) begins within the second exon of the A and C isoforms (coding exon 2_2346_0) and ends at the same location. The last three exons are identical for all three isoforms.

To determine the approximate exon boundaries of *Rad23* in *D. ficusphila*, a series of *BLASTX* searches was conducted using contig18 as the query and the protein sequence of each *D. melanogaster* coding exon as the subject (Table 1, Figure 5). Although Exon 3 is a smaller version of Exon 2 (Figure 4A), a *BLASTX* search was still necessary to determine if the ortholog to Exon 3 begins with a methionine, as anticipated. The best alignment places Exon 3 in the same region. There are five gaps totaling 14 bp within the same region of contig18 that align to both Exons 2 and 3 (Figure 5B, C); contig18 also does not align as well to the middle of Exon 5

(Figure 5E). Overall, the coding exons alignments are collinear. The entirety of each coding exon sequence aligns to contig18. Thus, *Rad23* is generally well conserved between *D. melanogaster* and *D. ficusphila*.

FlyBase ID	Exon ID	Coding Exon Size	Query Start	Query End	Reading Frame	Subject Start	Subject End	Isoforms Present
1_2346_0	Exon 1	22	27051	27116	+3	1	22	A,C
2_2346_0	Exon 2	216	27227	27904	+2	1	216	A,C
3_2346_0	Exon 3	167	27374	27904	+2	1	167	B
4_2346_2	Exon 4	31	29120	29212	+2	1	31	A,B,C
5_2346_0	Exon 5	108	29282	29608	+2	1	108	A,B,C
6_2346_0	Exon 6	37	29665	29775	+1	1	37	A,B,C

Table 1: Summary of *BLASTX* searches of contig18 (query) to each *D. melanogaster* coding exon of *Rad23* (subject).

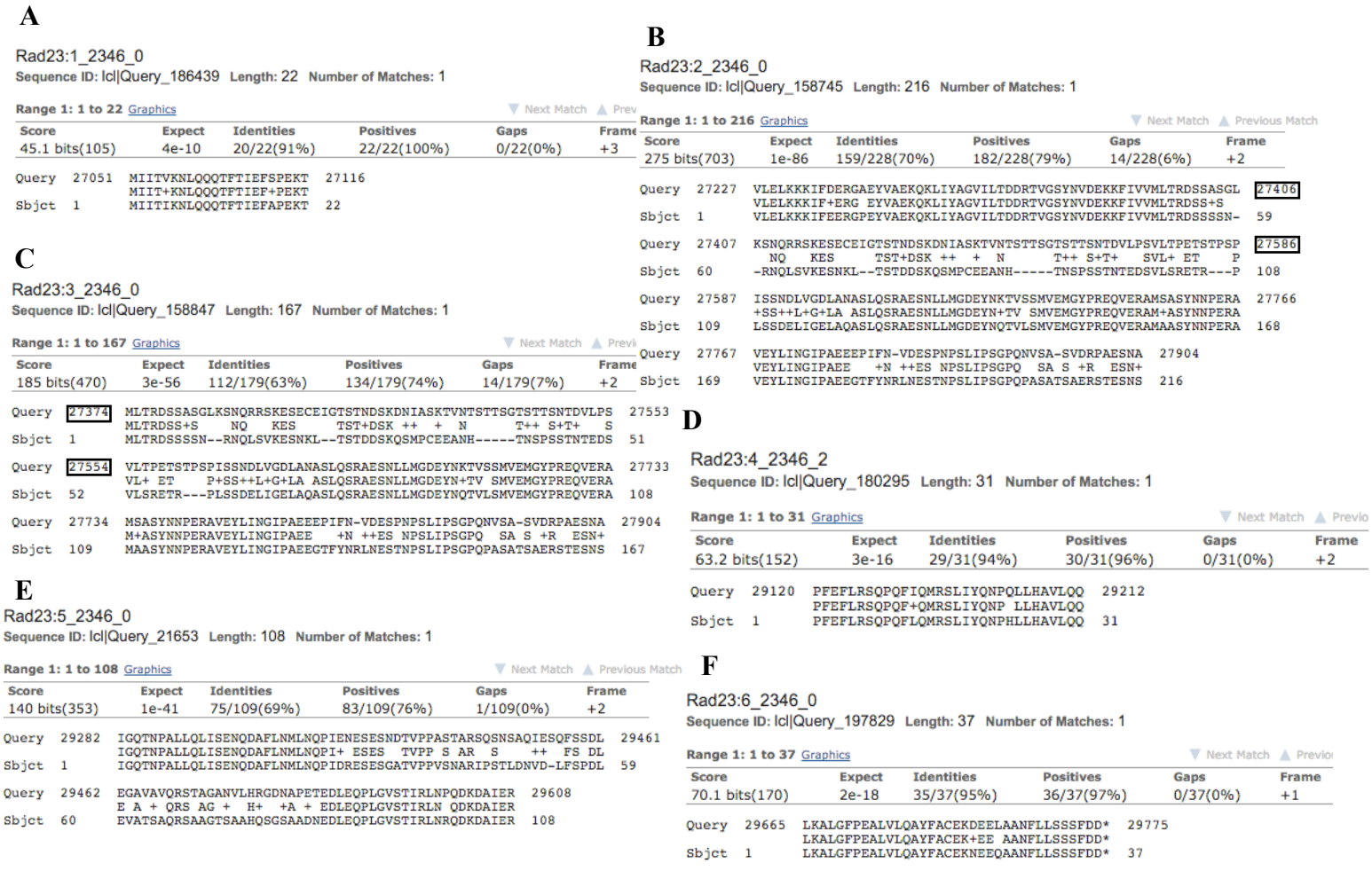


Figure 5: *BLASTX* alignments of contig18 (query) to each individual coding exon of *D. melanogaster Rad23* (subject). (A) 1_2346_0 **(B)** 2_2346_0 **(C)** 3_2346_0 **(D)** 4_2346_2 **(E)** 5_2346_0 **(F)** 6_2346_0. Boxes show that 14 gaps occur between bases 27,404-27,583 in both coding exons 2_2346_0 and 3_2346_0.

The approximate exon coordinates and reading frames as determined by *BLAST* were used with RNA-seq data and splice junctions predicted by *TopHat* to find the exact boundaries of each exon. The first coding exon of the A and C isoforms, Exon 1, is in the +3 reading frame. The only methionine nearby is located where contig18 aligned with Exon 1 (note amino acid sequence), indicating that the exon begins at 27051 bp (Figure 6).

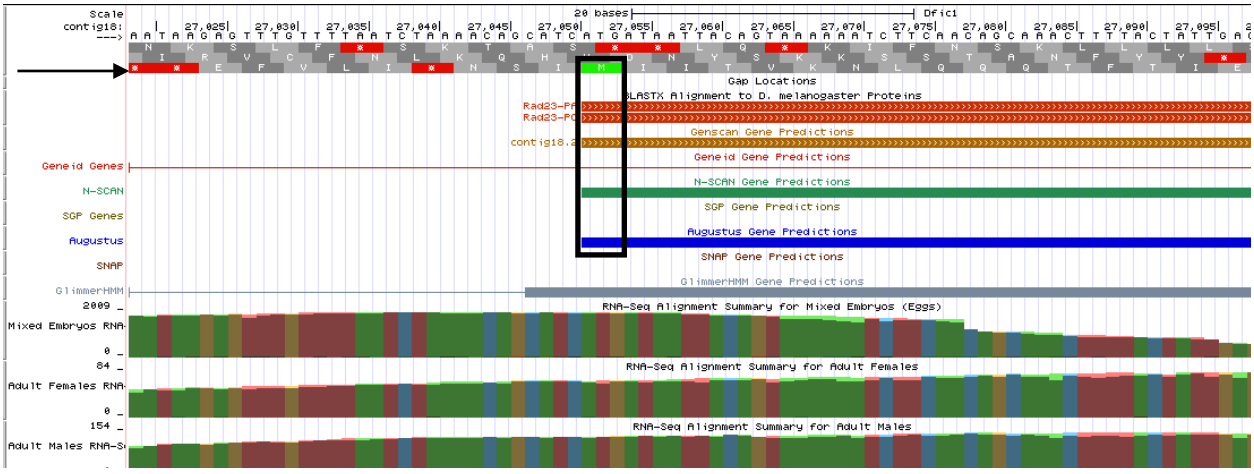


Figure 6: Identification of the methionine in Exon 1. This exon is in the +3 reading frame (arrow). There is only one methionine nearby and the amino acid sequence is conserved, matching *D. melanogaster*, so the exon begins at 27051 bp. Since this is the first exon, the upstream RNA-seq data corresponds to the 5' UTR.

Splice Sites

A splice site donor is identified by the bases GT (or in rare cases, GC) and an acceptor by bases AG. A splice site's phase is determined by how many nucleotides are between the splice site and the nearest full codon. The proper splice site donor/acceptor pair must have phases adding to 0 or 3 so that no incomplete codons exist; this prevents a frameshift. Ambiguous cases can be resolved by investigating RNA-seq data and *TopHat* junction predictions, while maintaining conservation of the amino acid sequence. Exon 1 has three candidate splice donors at bases 27117-8, 27121-2, and 27124-5, which are in phases 0, 1, and 1 for the +3 reading frame, respectively (Figure 7A). Only one potential splice acceptor exists for Exon 2 at 27225-6

bp (Figure 7B). This exon is in the +2 reading frame and the acceptor site is in phase 0, indicating the correct donor is at 27117-8 bp. This prediction is supported by RNA-seq data and a *TopHat* prediction with 321 supporting reads (Figure 7C). *TopHat* predicts two additional splice junctions, but one prediction (JUNC00000536) is not in phase and the other prediction (JUNC00000538) has only 11 supporting reads and is less conserved among the seven *Drosophila* species in the Multiz Alignments. No additional evidence supports JUNC00000538, so it is unlikely to be an actual splice junction and may be the result of a splicing error.

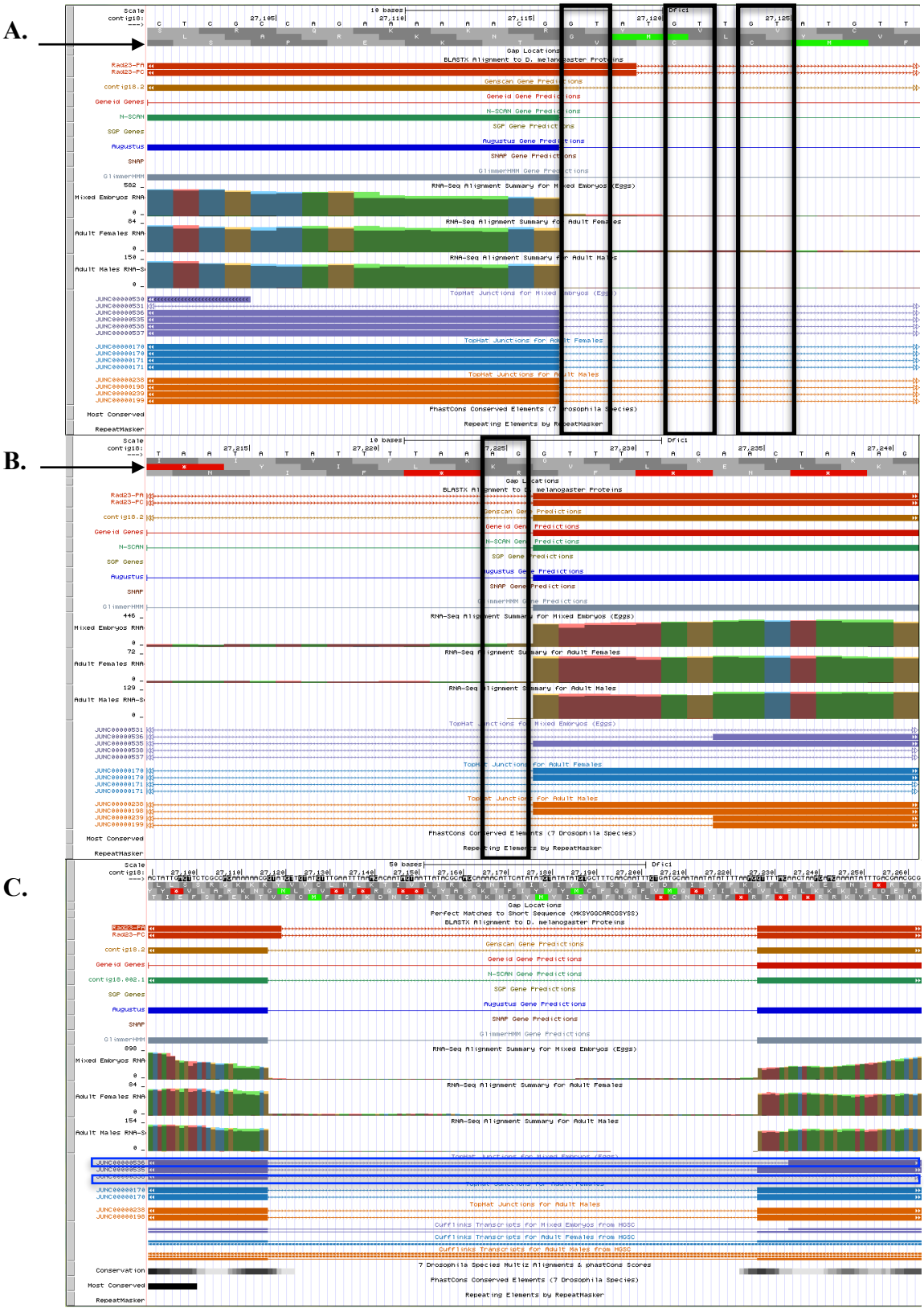


Figure 7: Splice junction donor on Exon 1 (A) and acceptor on Exon 2 (B). Arrows point to reading frames and black boxes correspond to candidate donors and acceptors, respectively. Reading frames are based on *BLASTX* findings shown in Table 1. This notation will be used for all splice site figures. Only one acceptor exists and it is in phase 0 in the +2 reading frame. The only donor candidate in phase 0 on the +3 reading frame is at 27117-8 bp. **(C)** Supporting RNA-seq data and *TopHat* predictions. Blue boxes correspond to *TopHat* predictions with a low number of supporting reads or that are out of phase.

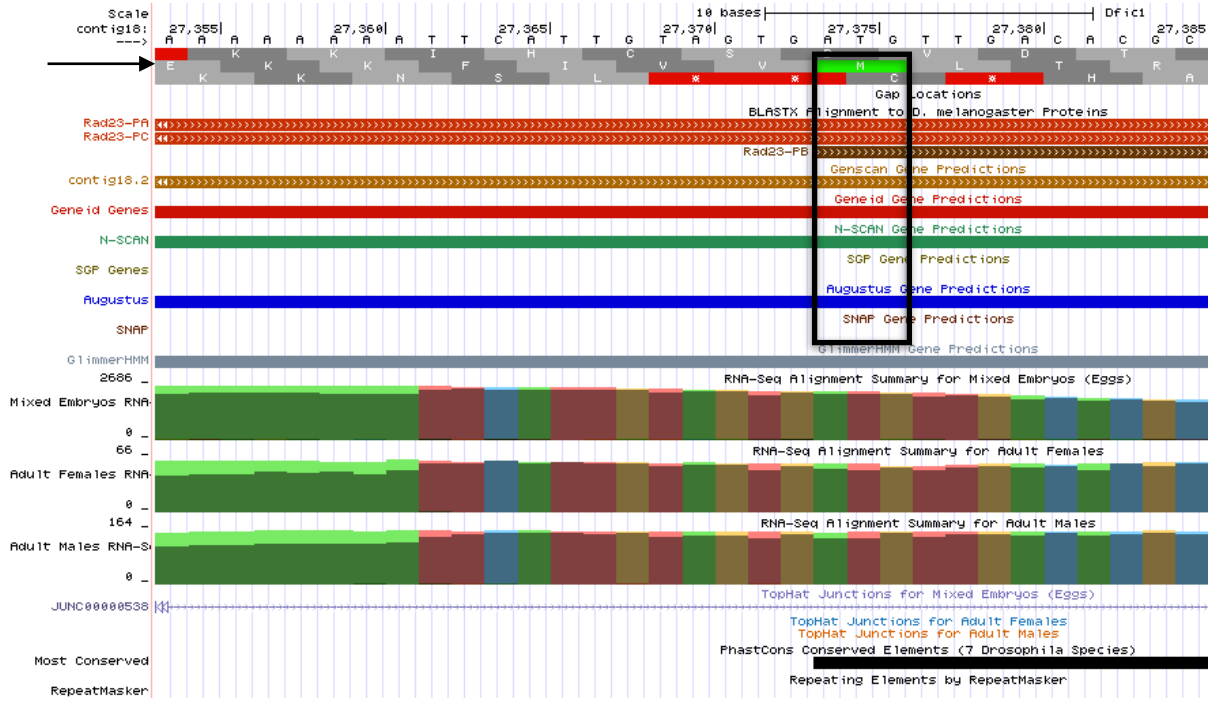


Figure 8: Identification of the methionine in Exon 3. This exon is in reading frame +2 (arrow) and only has one methionine nearby, so the exon begins at 27374 bp.

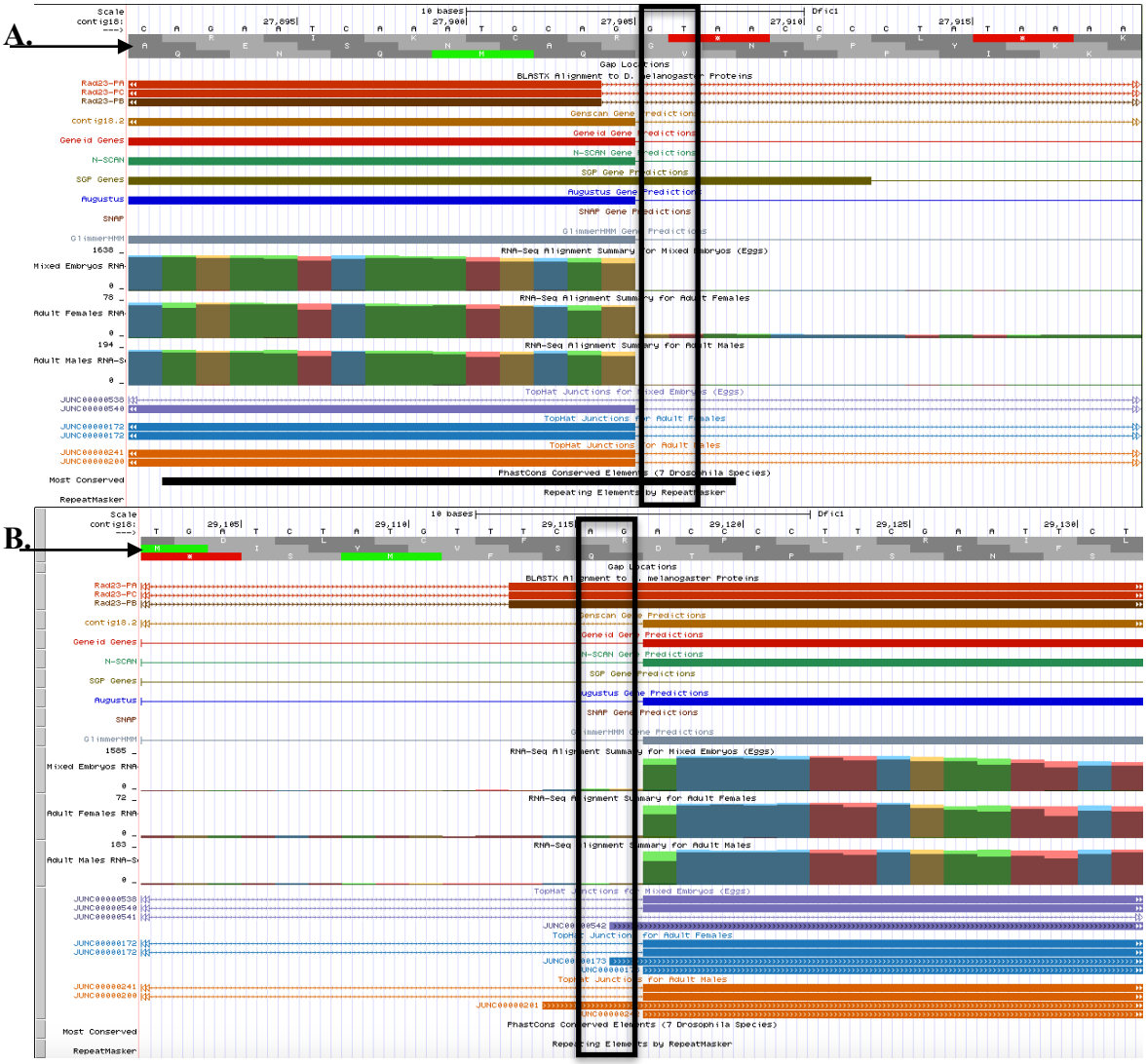


Figure 9: Splice donor on Exons 2 & 3 (A) and acceptor on Exon 4 (B). The only splice donor candidate is in phase 1 for the +2 frame and the only acceptor candidate is in phase 2 for the +2 frame.

The first coding exon for *Rad23-PB* (Exon 3, see Table 1) begins in the middle of Exon 2, which is a coding exon of isoforms A/C (see Figure 4). There is only one methionine in the +2 reading frame near the start of Exon 3, indicating that the ortholog of *Rad23-PB* begins at 27,374 bp (Figure 8), which yields the predicted amino acid sequence. Since both Exons 2 and 3 end at the same site (Table 1), they must have the same splice donor. Only one candidate is found near the end of these exons, which is in phase 1 for the +2 reading frame (Figure 9A). Exon 4, which is in the +2 reading frame, also only has one possible splice acceptor site near the end of the alignment; since the acceptor is in phase 2, this is the likely splice donor/acceptor pair (Figure 9B). While there are *TopHat* predictions for this splice junction with 1420-30 supporting reads in embryos, adult males, and adult females, there is an additional *TopHat* prediction in embryos with a donor at 28,748 supported by 48 reads (Figure 10). There are enough supporting reads for the prediction to be potentially correct, but there is little to no supporting RNA-seq evidence and the Multiz Alignments does not indicate conservation. *RepeatMasker* also indicated that this donor region is repetitious, so this splice junction was considered spurious.

Exon 4, which is in the +2 frame (Table 1), has two possible splice donors in phases 0 and 1. Due to the presence of a nearby upstream stop codon in the +2 frame, Exon 5 only has one candidate splice acceptor (Table 1, Figure 11). The acceptor is in phase 0, so the splice donor in Exon 4 must be the candidate in phase 0. This choice is in agreement with the RNA-seq evidence. Similarly, a stop codon is found shortly downstream of the 3' edge of Exon 5, leaving only one candidate splice donor in phase 0 (Figure 12). Exon 6 is in the +1 frame (Table 1) and begins immediately downstream of a stop codon. The only candidate splice acceptor is in phase 0, which is in agreement with the donor from Exon 5 (Figure 12) and RNA-seq data. Exon 6

ends with a stop codon at bases 29773-5. This is the end of the coding region for all three isoforms of *Rad23* (Figure 13) and is confirmed by amino acid sequence conservation (Figure 5F).

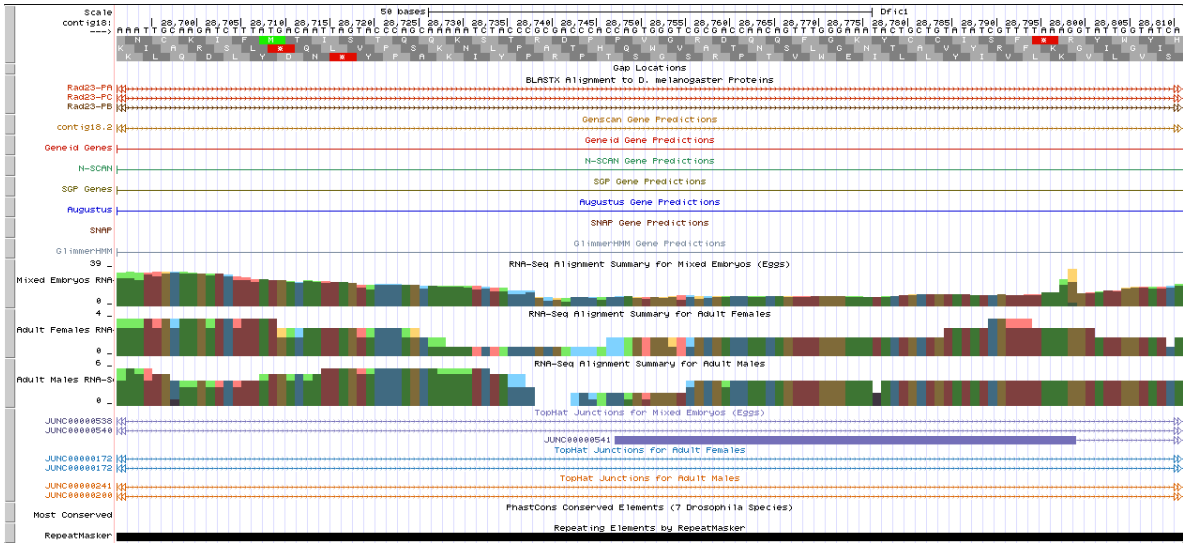


Figure 10: A spurious *TopHat* prediction. Despite 48 supporting reads, JUNC00000541 lacks supporting evidence from gene predictors, RNA-seq, and Multiz Alignment conservation. It also occurs within a repetitious element, indicating it is likely a remnant of a transposable element.

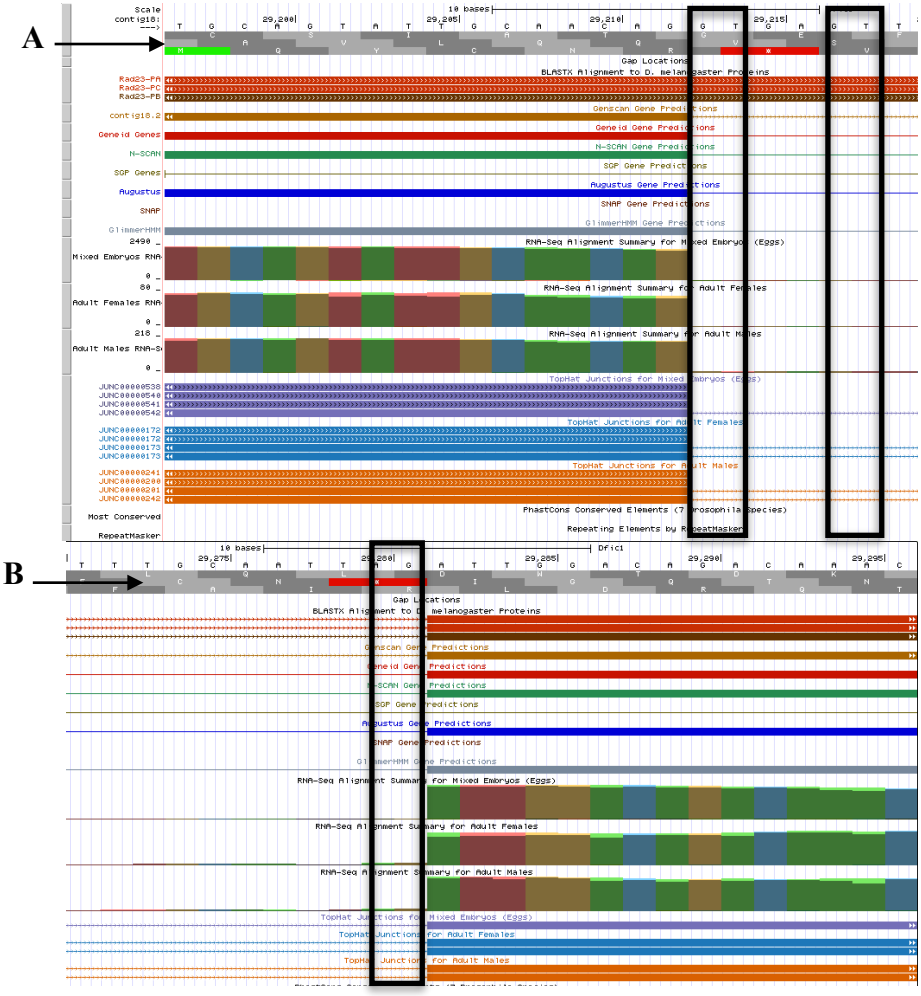


Figure 11: Candidate splice donors on Exon 4 (A) and acceptor on Exon 5 (B). The only splice acceptor is in phase 0 in the +2 frame, so the donor must be in phase 0 in the +2 frame.

Figure 12: Candidate splice donor on Exon 5 (A) and acceptor on Exon 6 (B). The only splice donor candidate is in phase 0 for the +2 frame. The only splice acceptor candidate is in phase 0 for the +1 frame.

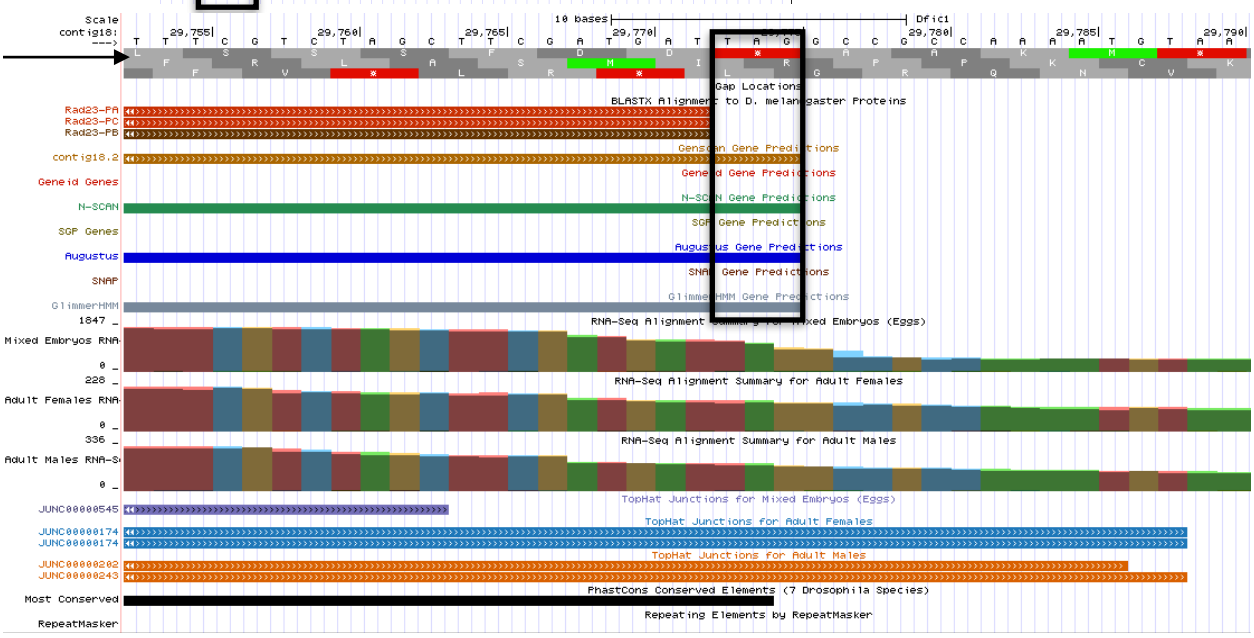


Figure 13: Identification of the stop codon in Exon 6. The exon is in the +1 frame (arrow, Table 1). The stop codon is correctly predicted by four *ab initio* gene predictors and confirmed by sequence homology.

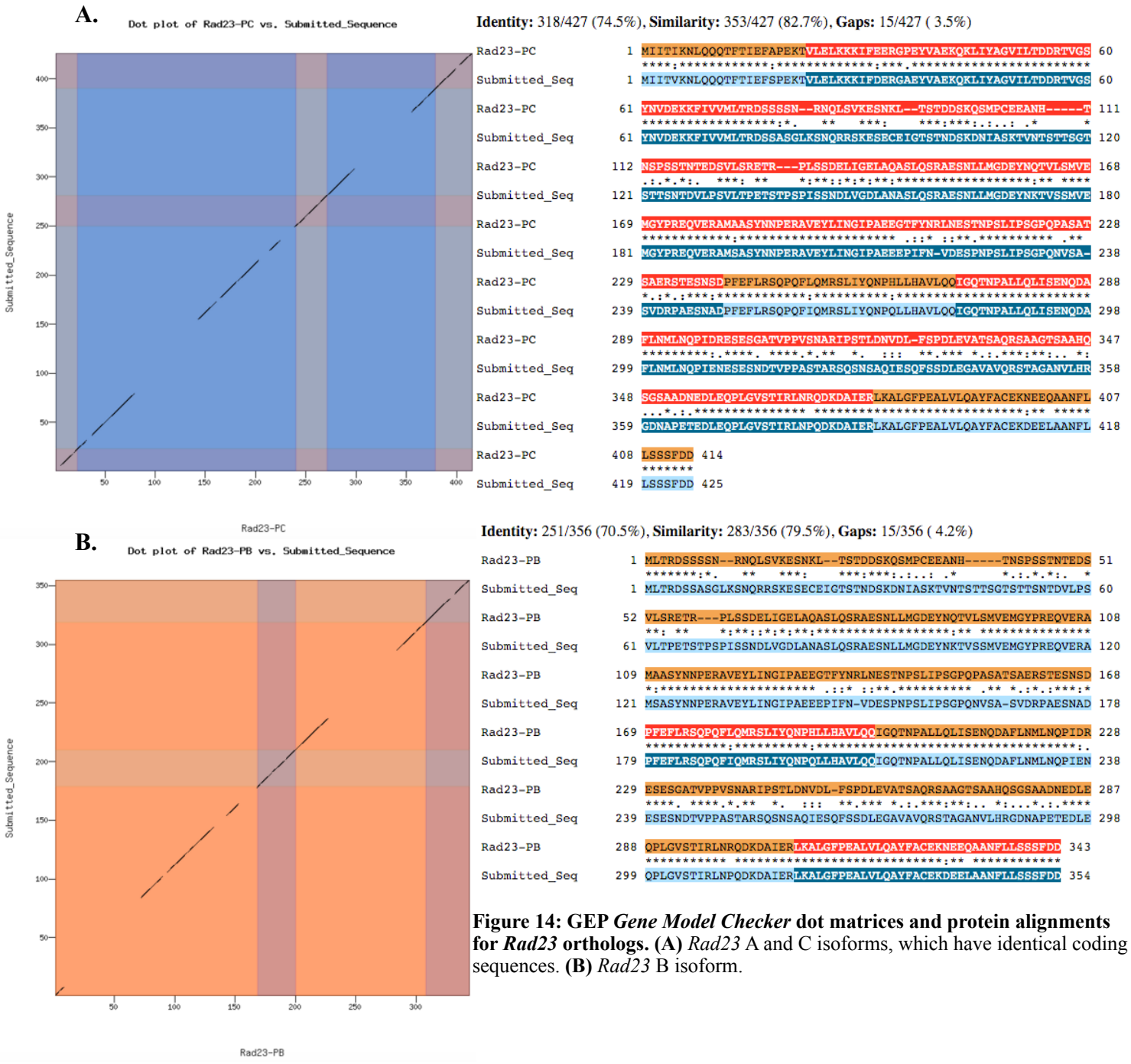


Figure 14: GEP *Gene Model Checker* dot matrices and protein alignments for *Rad23* orthologs. (A) *Rad23* A and C isoforms, which have identical coding sequences. **(B)** *Rad23* B isoform.

Exon ID	Coding Exon Size	Beginning	End	Reading Frame	Splice Acceptor Phase	Splice Donor Phase	Isoforms
Exon 1	22	27051	27116	+3	NA	0	A,C
Exon 2	216	27227	27905	+2	0	1	A,C
Exon 3	167	27374	27905	+2	NA	1	B
Exon 4	31	29118	29212	+2	2	0	A,B,C
Exon 5	108	29282	29608	+2	0	0	A,B,C
Exon 6	37	29665	29775	+1	0	NA	A,B,C

Table 2: Final exon annotation of the *D. ficusphila* ortholog of *Rad23*.

Verification of Gene Model

The exon annotations for the *D. ficusphila* ortholog of *Rad23* were checked using the GEP *Gene Model Checker* (Figure 14). Since the A and C isoforms have identical coding sequences, the resulting dot matrices and protein alignments are identical; C is shown. Furthermore, the B isoform begins with Exon 3 (which starts in the middle of Exon 2 of the A/C isoforms) and is identical to the A and C isoforms thereafter, so its dot matrix is a sub-matrix of that corresponding to the A and C isoforms. The same is true for the protein alignment. The middle portions of Exon 2 (and consequently, Exon 3) and Exon 5 show significant divergence from the *D. melanogaster* ortholog. The beginnings and ends of all coding exons are well-aligned, anchoring their positions and indicating *Rad23* is generally well conserved. A summary of the final exon annotation is shown in Table 2. The FlyBase annotation for *D. melanogaster Rad23* indicates it is a putative DNA repair protein involved in nucleotide-excision repair.

Transcription Start Site

One of the motivations for annotating the *D. ficusphila* F element is to identify the transcription start site(s) (TSS) of genes orthologous to *D. melanogaster*. The regions surrounding TSS orthologs, along with those of many other species within the genus, will be analyzed in an attempt to identify conserved sequence motifs. In *D. melanogaster*, the first transcribed exon of *Rad23* begins at the same location for all three isoforms (Figure 4A). The 9-state epigenomic landscape in both BG3 and S2 cell lines show the gene is actively transcribed in a heterochromatic environment (Figure 15). The DNase hypersensitive sites (DHS) for all three cell lines have the same, single peak and only a single TSS has been annotated by the modENCODE project (Hoskins, et. al. 2011). A core promoter region can be classified as peaked

(one TSS, one DHS), intermediate (one TSS, multiple DHS), or broad (multiple TSS, multiple DHS). Thus, *Rad23* has a peaked promoter in *D. melanogaster*, meaning the *Rad23* ortholog in *D. ficusphila* likely has one TSS.

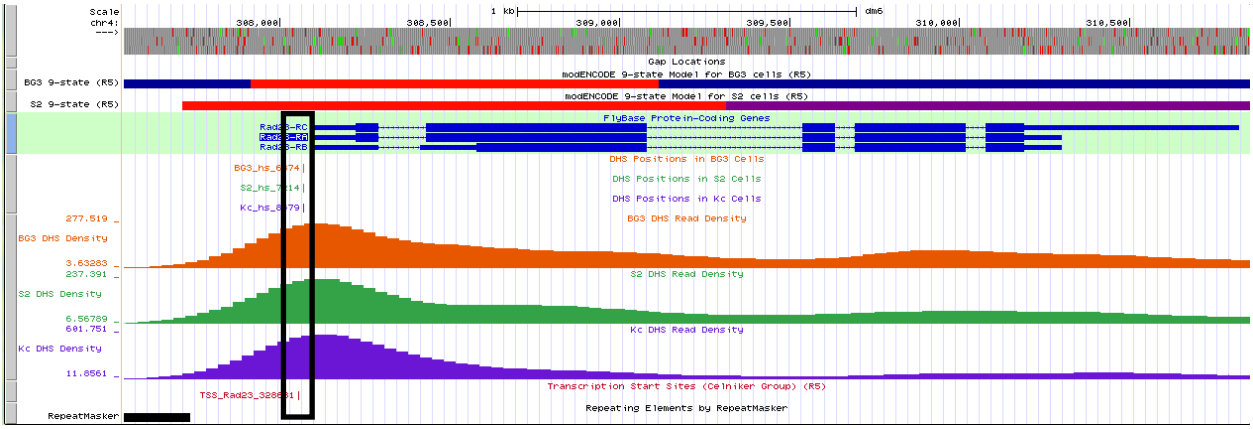


Figure 15: Core promoter region and 5' UTR of *Rad23* in *D. melanogaster*. The 9-state epigenomic tracks indicates the 5' end of *Rad23* is an "Active promoter / TSS Region" (red) surrounded by heterochromatin (dark blue). A single TSS has been previously annotated and there is only one DHS at the same site for all three cell lines (box). Thus, *Rad23* has a peaked promoter.

To estimate the location of the orthologous TSS in *D. ficusphila*, a *BLASTN* alignment was conducted using the first transcribed exon of *D. melanogaster Rad23* as the query and contig18 as the subject (Figure 16). Because UTRs are usually less conserved, sensitive *BLASTN* parameters were used (word size 7, match/mismatch scores 1/-1, gap cost 2 existence, 1 extension, allowing low complexity alignments). The best alignment, which is 75% identical and missing the first 15 bases of the query, was extrapolated to estimate a TSS at approximately 26923 bp.

contig18
Sequence ID: lcl|Query_78605 Length: 40000 Number of Matches: 45

Range 1: 26938 to 27116 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
122 bits(84)	8e-31	138/184(75%)	7/184(3%)	Plus/Plus

```
Query  16     GGTCACACTGATGACAAATCGTTTTTATCAAGCGATATTGGGAACTAACTTTCGAGTTTG  75
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  26938  GGTCACACTGATGACAATTCATTTTTTATATG----ATTGGTAT-TAAAAATCCAGTTTA  26992

Query  76     CA-TATGTGTATGCATATGTATATTTAATTAAGTTTTGGTTTTGTCTGATACA-CAAGAT  133
              || |||||||||||||||||||||||||||||||||||||||||||||||| ||||||
Sbjct  26993  TAGTTTACATATGCATATTTACACTTAATAAGAGTTTGTTTTAATCTAAAACAGCATCAT  27052

Query  134    GATTATTACAATTAAAAATCTTCAACAGCAAACTTTTACTATTGAGTTTGCCCCGGAAAA  193
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  27053  GATAATTACAGTAAAAAATCTTCAACAGCAAACTTTTACTATTGAGTTCTCGCCAGAAAA  27112

Query  194    AACG  197
              ||||
Sbjct  27113  AACG  27116
```

Figure 16: *BLASTN* alignment of the first transcribed exon of *Rad23* in *D. melanogaster* (query) and contig18 (subject).

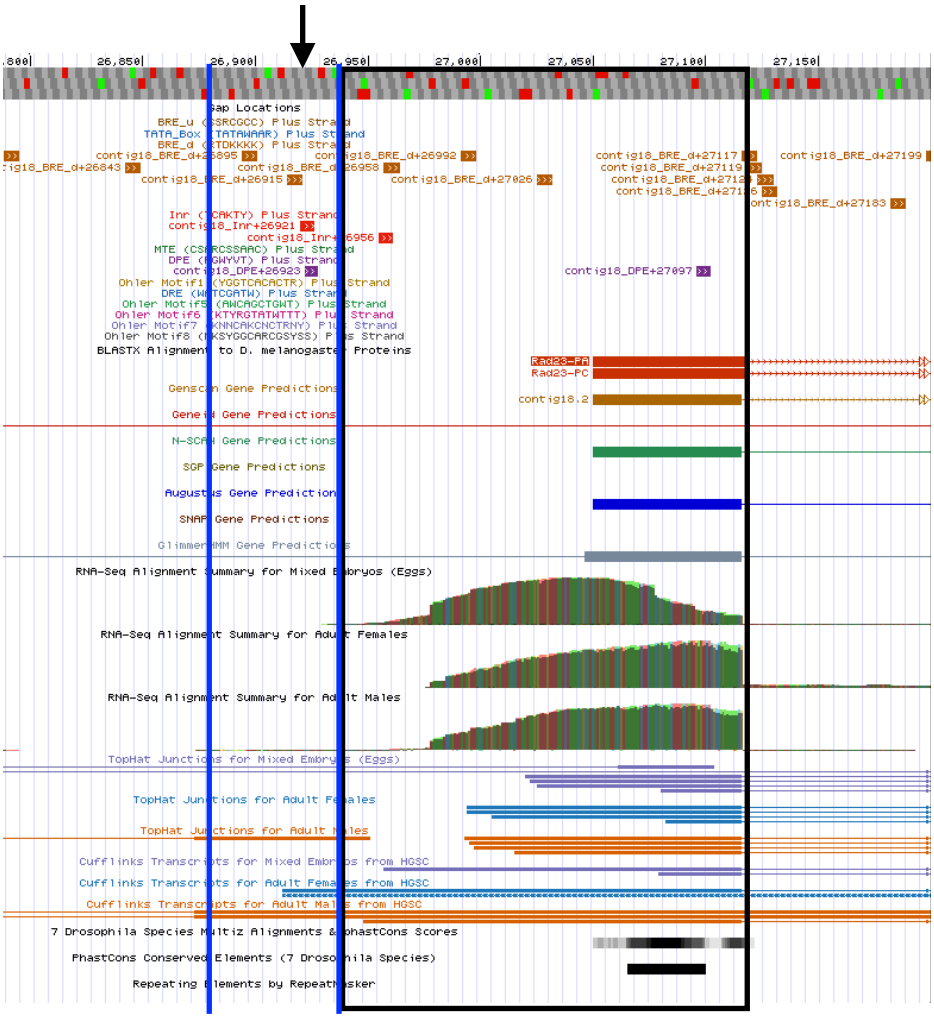


Figure 17: Core promoter motifs within 300 bp of the TSS search region (26874-26938 bp, between blue lines). The predicted TSS is at 26923 bp (arrow). The black box indicates the *BLASTN* alignment in Figure 16. *Cufflinks* evidence supports the putative TSS. Some *Cufflinks* transcripts and an Inr starting at 26956 bp suggest a TSS at 26958 bp, but there is insufficient evidence to annotate this as an actual TSS.

A small number of RNA-seq reads extended slightly upstream of the predicted TSS position (Figure 17). The TSS search region, which spans from the beginning of RNA-seq reads to where the *BLASTN* alignment begins, was defined as 26874-26938 bp. Any conserved core promoter motifs within 300 bp of both the TSS search region in *D. ficusphila* and the TSS in *D. melanogaster* are listed in Table 3. The presence of an Inr motif starting at 26921 bp and a BRE^{d} motif starting at 26915 bp support the prediction of a TSS at 26923 bp. Cufflinks transcripts and an Inr motif starting at 26956 bp suggest there may be an additional TSS at 26958 bp (Figure 17). However, the *BLASTN* alignment and lack of a second annotated TSS in *D. melanogaster* suggest that a TSS at 26958 bp is likely a false positive, but could be a feature in this species.

Although a DPE motif beginning at 26923 bp is within the search region (which would indicate a TSS at 26895 bp), there is no other evidence of a TSS to support this observation.

Motif	***D. ficusphila* position**	***D. melanogaster* position**
BRE[d]	26601, 26703, 26715, 26726, 26789, 26843, 36895, **26915**, 26958, 26992, 27026, 27117, 27119, 27124, 27126, 27183, 27199	307748, 307929, 307936, 308012, 308049, 308061, 308142, 308169, 308199, 308205, 308290
Inr	**26921**, 26956	
DPE	26630, 26724, 26923, 27097	
Ohler_motif1		308107

Table 3: Core promoter motifs within 300 bp of the *Rad23* TSS search region in *D. ficusphila* and the annotated TSS in *D. melanogaster*. Bold/underlined positions indicate those that support the hypothesis of a TSS at 26923 bp. All positions are on the plus strand. No evidence was found for BRE[u], TATA Box, MTE, DRE or Ohler motifs 5-8, despite a search.

In an attempt to further refine the TSS search region, the RNA-PolII ChIP-Seq data for *Rad23* was obtained from the August 2013 (GEP/Dot) assembly of *D. biarmipes*. The peak binding site was used as the query in a *BLASTN* alignment to the subject, contig18 (Figure 18). The aligned region of contig18, 26887-27061 bp, begins at essentially the same location as the TSS search region (26874-26938 bp). Although the TSS search region was not refined, it is supported by these results. The alignment is 68% identical, indicating moderate conservation of the 5' UTR in *Rad23* between *D. ficusphila, D. biarmipes,* and *D. melanogaster*.

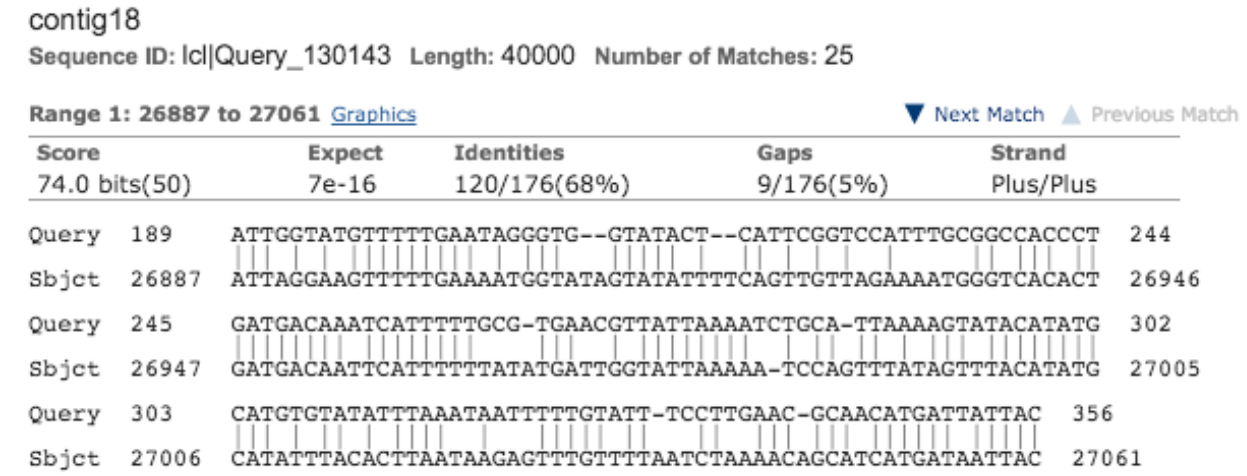


Figure 18: *BLASTN* alignment of the RNA-PolII ChIP-seq peak for *Rad23* in *D. biarmipes* (query) and contig18 (subject). The beginning of the alignment, at 26887 bp, is very close to the beginning of the TSS search region at 26874 bp. The 68% identity indicates moderate conservation of the 5' UTR.

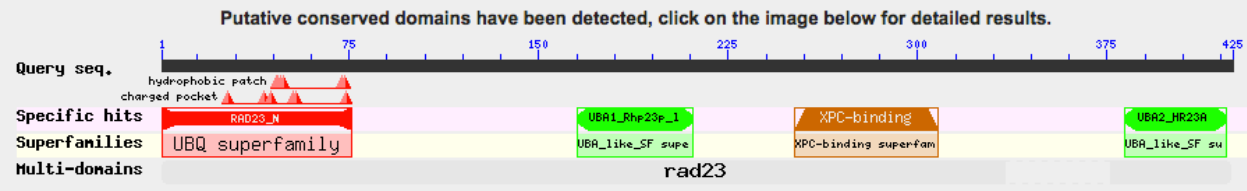


Figure 19: *BLASTP* search of the putative protein sequence from the *Gene Model Checker* for *Rad23* in *D. ficusphila* (query) to the non-redundant database (subject). Four conserved domains are identified, approximately from residues 1-75, 165-210, 255-308, and 382-425.

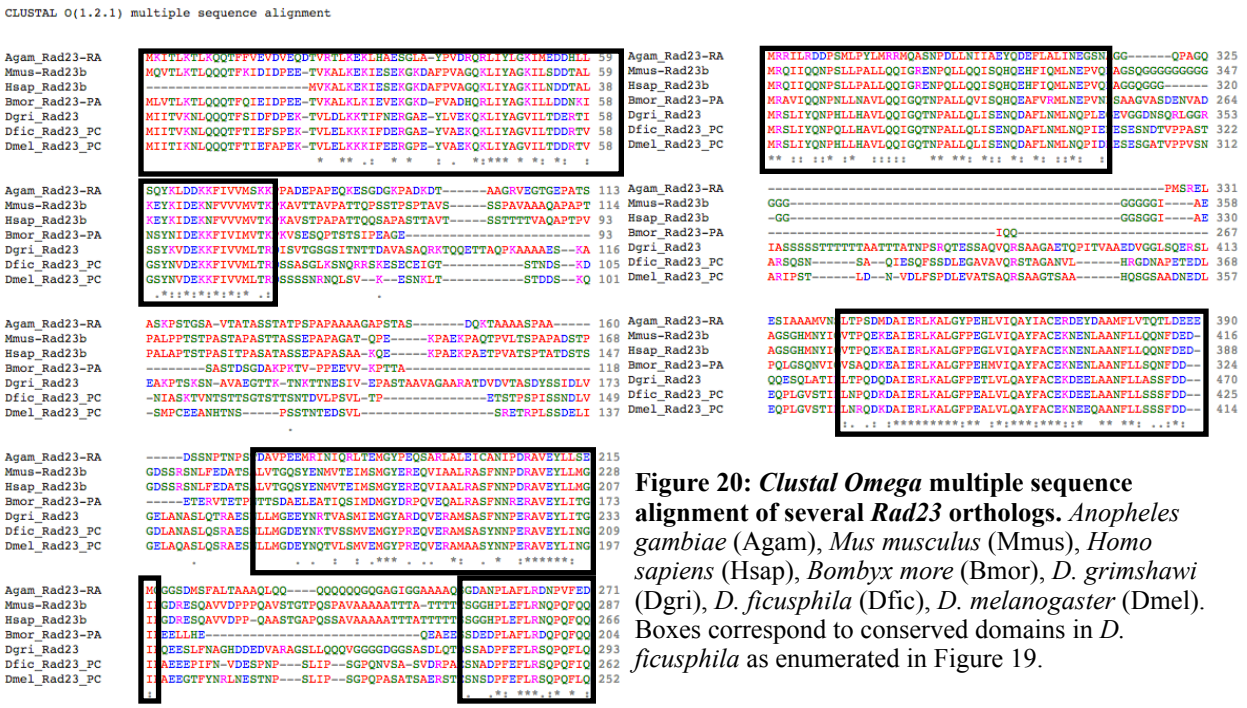


Figure 20: *Clustal Omega* multiple sequence alignment of several *Rad23* orthologs. *Anopheles gambiae* (Agam), *Mus musculus* (Mmus), *Homo sapiens* (Hsap), *Bombyx more* (Bmor), *D. grimshawi* (Dgri), *D. ficusphila* (Dfic), *D. melanogaster* (Dmel). Boxes correspond to conserved domains in *D. ficusphila* as enumerated in Figure 19.

Gene Evolution

Rad23 is very highly expressed, with over 2700 supporting RNA-seq reads in some parts of the gene (Figure 2). The predicted peptide sequence generated by the *Gene Model Checker* was used as the query for a *BLASTP* search of the non-redundant database (subject). Four conserved domains were identified (Figure 19). Since *Rad23* is a putative DNA-repair gene, a *Clustal Omega* multiple sequence alignment was carried out to determine to what extent the protein domains are conserved across the evolutionary tree. Orthologous protein sequences from *D. melanogaster*, *D. grimshawi*, *Anopheles gambiae* (Malaria mosquito), *Bombyx more* (Silkmoth), *Mus musculus* (house mouse), and *Homo sapiens* were used. The resulting alignment shows that although some divergence occurs in *Rad23* orthologs, the conserved domains are indeed very well conserved from flies to humans (Figure 20).

Feature 1

Four of the six *ab initio* gene predictors suggest Features 1 and 2 are one contiguous feature, yet conservation and preliminary *BLASTX* alignments suggest two different features (Figure 21). The predicted amino acid sequence from *Genscan* contig18.1 was used as the query in a *BLASTP* search to the FlyBase AA database for *D. melanogaster*. The first ~400 aa (Feature 2) matched to members of the *Synaptotagmin* (*Syt*) family, while the last ~600 aa (Feature 1) matched to *NfI*. The E-values indicated that Feature 1 is orthologous to *Syt7* and Feature 2, to *Nuclear factor I* (*NfI*) (Figure 22). Although the match to *Syt1* also has a small E-value, this gene is located on the 2L chromosome (B element) in *D. melanogaster; Syt7* is located on the F element, the orthologous position for contig18.

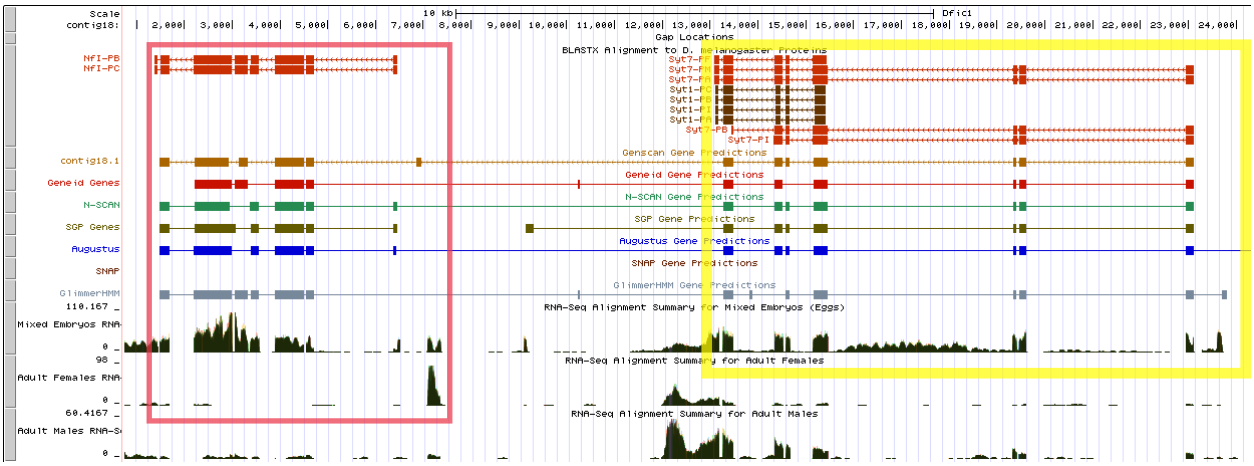


Figure 21: Closer view of Features 1 (red box) and 2 (yellow box). *Genscan*, *Geneid*, *N-SCAN*, and *Augustus* all suggest Features 1 and 2 are one continuous feature, yet preliminary *BLASTX* alignments suggest two distinct features.

A

BLAST Hit Summary

Description	Species	Score	E value
Nfl-PC	Dmel	944.88	0
Nfl-PB	Dmel	940.643	0
Syt7-PM	Dmel	772.311	0
Syt7-PA	Dmel	772.311	0
Syt7-PB	Dmel	649.432	0
Syt7-PI	Dmel	618.616	1.02242e-176
Syt7-PF	Dmel	530.791	2.55788e-150
Syt1-PI	Dmel	231.876	2.27829e-60
Syt1-PC	Dmel	231.876	2.27829e-60
Syt1-PB	Dmel	231.876	2.27829e-60

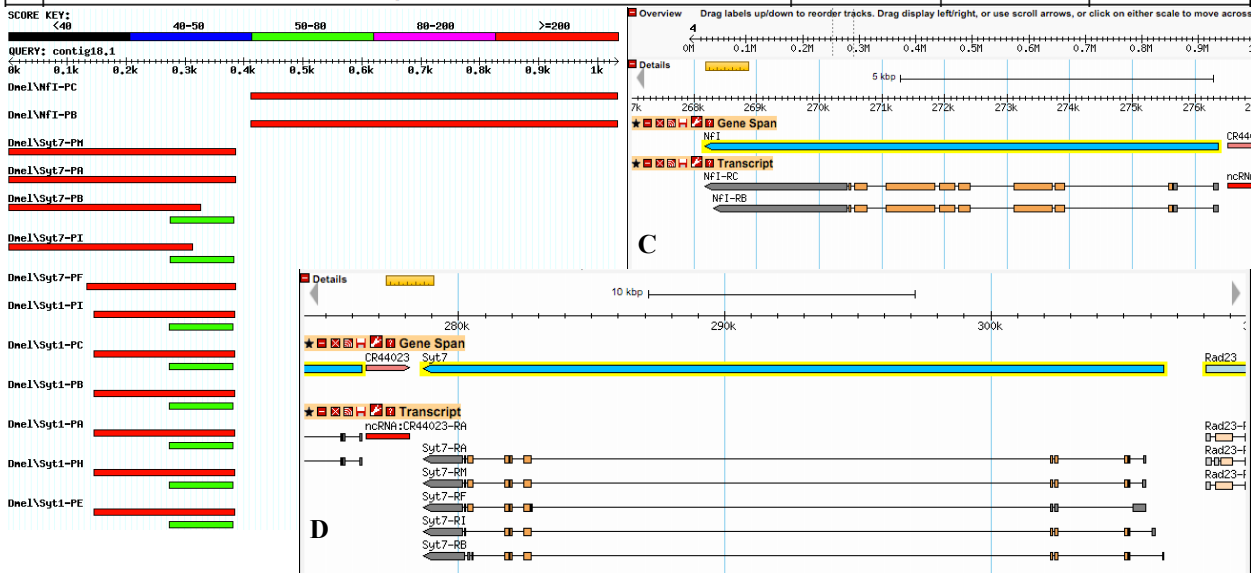


Figure 22: *BLASTP* search results of *Genscan* predicted peptide sequence contig18.1 (query) to the FlyBase AA database for *D. melanogaster* (subject). (A) *Nfl* (Feature 1) and *Syt7* (Feature 2) have substantially lower E-values than the next best result, *Syt1*. **(B)** The first ~400 aa of contig18.1 align to *Syt7* and *Syt1*, while the last ~600 aa align to *Nfl*, indicating contig18.1 is in fact two distinct features. **(C)** Feature 1 is orthologous to *Nfl*, which is found on the F element. Light orange corresponds to coding regions and grey indicates UTRs. **(D)** Feature 2 is very likely orthologous to *Syt7*, not *Syt1*. While *Syt1* is found on the B element in *D. melanogaster*, *Syt7* is found on the F element. *Syt7* is also adjacent to *CR44023*, which is also adjacent to *Nfl* (C), and *Rad23*.

FlyBase ID	Exon ID	Coding Exon Size	Query Start	Query End	Reading Frame	Subject Start	Subject End	Isoforms Present	Percent Identity
1_10065_0	Exon 1	27	6456	6376	-2	1	27	B,C	93%
2_10065_0	Exon 2	52	4705	4550	-1	1	52	B,C	100%
3_10065_0	Exon 3	199	4497	3904	-2	1	199	B,C	91%
4_10065_2	Exon 4	60	3563	3384	-3	1	60	B,C	73%
5_10065_2	Exon 5	81	3315	3061	-2	1	81	C	65%
6_10065_2	Exon 6	85	3327	3061	-2	1	85	B	66%
7_10065_2	Exon 7	259	2987	2208	-3	3	259	B,C	76%
8_10065_0	Exon 8	65	1687	1494	-1	1	65	B,C	86%
9_10065_2	Exon 9	16	1429	1381	-2	1	16	B,C	88%

Table 4: Summary of *BLASTX* searches of contig18 (query) to each *D. melanogaster* coding exon of *NfI* (subject). Exons 5 and 6 begin at different positions but cover the same region; these exons are the only difference between the B and C isoforms.

Contig18 was used as the query in a series of pairwise *BLASTX* alignments with each coding exon of *NfI* (Table 4), which has two isoforms. Except for the first two bases of *D. melanogaster NfI* coding exon 7_10065_2 (Exon 7), the entirety of every coding exon aligns with contig18, indicating that *NfI* is highly conserved. The FlyBase annotation of *D. melanogaster NfI* indicates it is a putative transcription factor involved in sequence-specific DNA binding, so high conservation is expected. This conserves the amino acid sequence in comparison to *D. melanogaster*. However, we cannot rule out the presence of a novel isoform using the 1443-4 splice acceptor in *D. ficusphila*.

The identification of splice donor/acceptor pairs was straightforward in most cases; only cases that were more unusual are discussed here. The only splice site donor for Exon 8, which is in the -1 reading frame, that has supporting RNA-seq and *TopHat* evidence is located at 1490-1 bp and is in phase 1 (Figure 23A). However, two splice site acceptor candidates for Exon 9, which is in the -2 reading frame, both had nearly equal levels of supporting evidence (Figure 23B). Candidate acceptor sites in phase 2 are located at 1431-2 and 1443-4 bp. The former,

which has 42 supporting *TopHat* splice junctions, is located where the majority of RNA-seq alignments end. The latter, where the final RNA-seq alignment ends, has 43 supporting *TopHat* splice junctions. Since the *BLASTX* alignment of Exon 9 begins at 1429 bp and the candidate splice acceptor at 1431-2 bp is more conserved in the Multiz alignment, only the splice acceptor at 1431-2 bp was annotated.

Figure 23: Splice site donor/acceptor pair for Exons 8 and 9. (A) Exon 8, in the -1 reading frame (Table 4), has a candidate splice site donor at 1490-1 bp supported by RNA-seq alignments, *TopHat* splice junctions, and conservation. **(B)** Exon 9, in the -2 reading frame (Table 4), has candidate splice sites at 1431-2 bp and 1443-4 bp. Both are supported by RNA-seq alignments and equally many *TopHat* splice junctions. However, the Multiz conservation and *BLASTX* alignment (Table 4) indicate the splice site at 1431-2 bp is more likely. The splice site at 1443-4 bp was not annotated.

Verification of Gene Model

The exon annotation for the *D. ficusphila* ortholog of *NfI* was checked using the GEP *Gene Model Checker* (Figure 24). Both the B and C isoforms of *NfI* were annotated. The dot matrices show that *NfI* is very well-conserved between *D. ficusphila* and *D. melanogaster*. A summary of the final exon-by-exon annotation is shown in Table 5.

Exon ID	Coding Exon Size	Beginning	End	Reading Frame	Splice Acceptor Phase	Splice Donor Phase	Isoforms Present
Exon 1	27	6456	6376	-2		0	B,C
Exon 2	52	4705	4550	-1	0	0	B,C
Exon 3	199	4497	3903	-2	0	1	B,C
Exon 4	60	3565	3383	-3	2	1	B,C
Exon 5	81	3317	3060	-2	2	1	C
Exon 6	85	3329	3060	-2	2	1	B
Exon 7	259	2995	2208	-3	2	0	B,C
Exon 8	65	1687	1492	-1	0	2	B,C
Exon 9	16	1430	1381	-2	1		B,C

Table 5: Final exon annotation of the *D. ficusphila* ortholog to *NfI*.

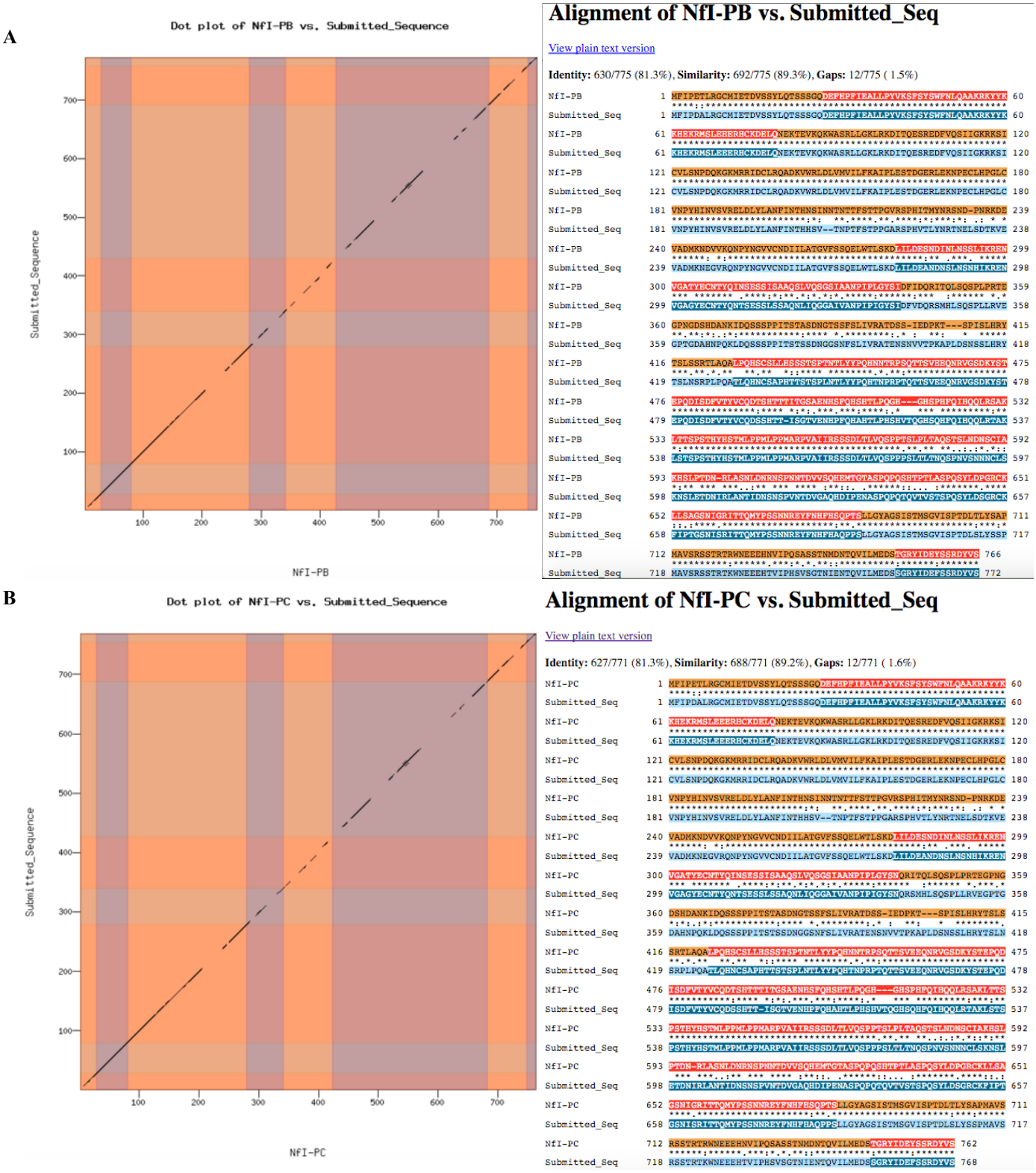


Figure 24: GEP *Gene Model Checker* dot matrices (left) and protein alignment (right) for *NfI-PB* (A) and *NfI-PC* (B). Both the dot matrices and protein alignments show that *NfI* is very well conserved. The beginning fifth exon for isoform C (Exon 5) is slightly less conserved than that of isoform B (Exon 6). However, Exon 5 begins just 12 bp after Exon 6 and both exons end at the same position (Table 5), so the exons are nearly identical.

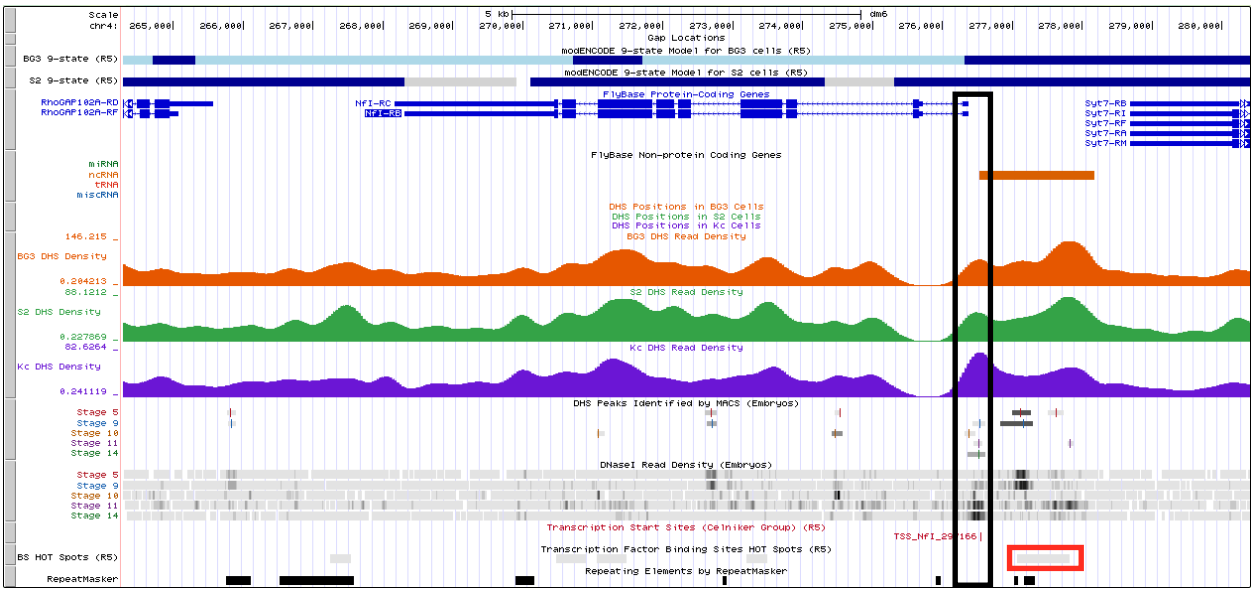


Figure 25: Core promoter region and 5' UTR of *Nfl* in *D. melanogaster*. The 9-state epigenomic tracks indicates *Nfl* is in a "heterochromatin-like region embedded in euchromatin" (light blue) surrounded by heterochromatin (dark blue). A single TSS has been previously annotated. There are no DHS annotated for BG3, S2, or Kc cell lines, but there is a DHS at the same site as the TSS in multiple embryonic stages (black box). Other embryonic DHS are likely due to the ncRNA. Thus, *Nfl* has a peaked promoter, which is supported by the weak TF binding hotspot signal in the red box.

Transcription Start Site

The 9-state epigenomic landscape of *D. melanogaster* in both S2 and BG3 cell lines indicates *Nfl* is located in a "heterochromatin-like region embedded in euchromatin" surrounded by heterochromatin (Figure 25). Although no DHS are annotated for BG3, S2, or Kc cell lines, there are DHS peaks annotated in multiple embryonic stages. Due to the presence of a noncoding RNA (ncRNA) gene immediately upstream, only the DHS peak nearest *Nfl* was considered to be a DHS for *Nfl*. Furthermore, only one TSS has been annotated for *Nfl* in *D. melanogaster* at 276,540 bp. A weak signal for a TF hotspot overlaps with the ncRNA and is in a logical position for upstream TFs of *Nfl*. Thus, *Nfl* is considered a peaked promoter in *D. melanogaster*. RNA-PolII ChIP-seq enrichment peaks in *D. biarmipes* indicate *Nfl* is also a peaked promoter in *D. biarmipes*.

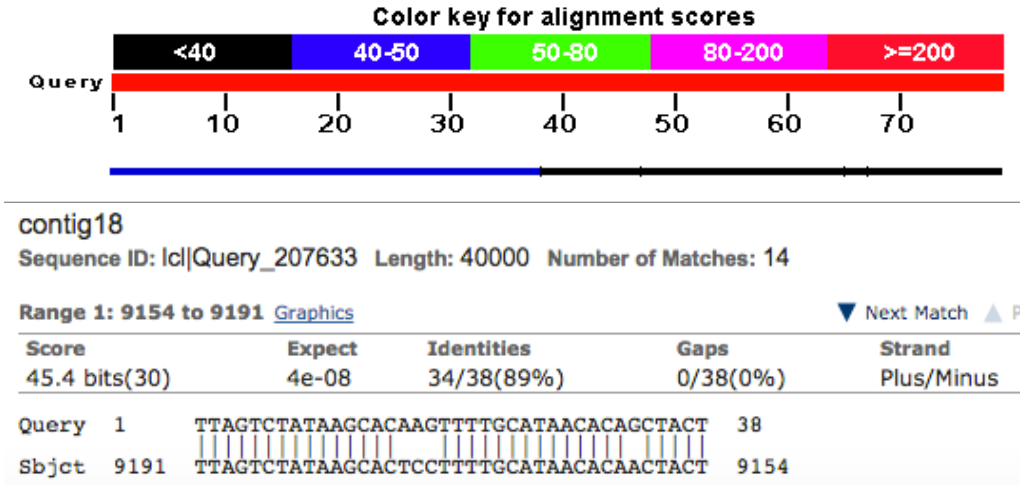


Figure 26: *BLASTN* alignment of the first transcribed exon of *D. melanogaster NfI* (query) and contig18 (subject). Top panel: Only the first 38 of 79 bp of the query align to contig18 with a meaningful score. Bottom panel: The first 38 bp of the query is 89% identical to the corresponding range of contig18 and places the putative TSS at 9191 bp.

Motif	*D. ficusphila* position	*D. melanogaster* position
TATA Box	9087	276690
BRE[d]	8887, 8898, 8959, 9050, 9053, 9057, 9104, 9106, 9259, 9261, 9263	276255, 276258, 276471, 276473, 276506, 276596, 276653, 276700, 276702, 276704, 276718, 276721, 276750, 276764, 276766
Inr	8963	276327
DPE	9328, 9462	276325, 276392, 276626

Table 6: Core promoter motifs within 300 bp of the *NfI* TSS search region in *D. ficusphila* and the annotated TSS in *D. melanogaster*. No core promoter motifs support the hypothesis of a TSS between 9191-9224 bp. All positions are on the minus strand. No evidence was found for BRE[u], MTE, DRE or Ohler motifs 1 and 5-8, despite a search.

The two *NfI* isoforms in *D. melanogaster* share the same first transcribed exon (Figure 22C). The only meaningful *BLASTN* alignment between the first transcribed exon of *NfI* (query) and contig18 (subject) placed the first 38 bp of the 79 bp exon at 9154-91 bp of contig18, which is approximately 2.7 kb upstream from the first coding exon (Figure 26, Table 5). Although the putative TSS at 9191 bp is far from the first coding exon in *D. ficusphila*, RNA-seq alignments and *RepeatMasker* results indicate that an uncategorized repetitious element, perhaps a Helitron,

is approximately located between 7000-7500 bp (Figure 27). RNA-seq alignments, *TopHat* splice junctions, and *Cufflinks* transcripts further support a putative TSS at 9191 bp. No core promoter motifs are located within 300 bp of the TSS search region (Table 6), which is 9191-9224 bp.

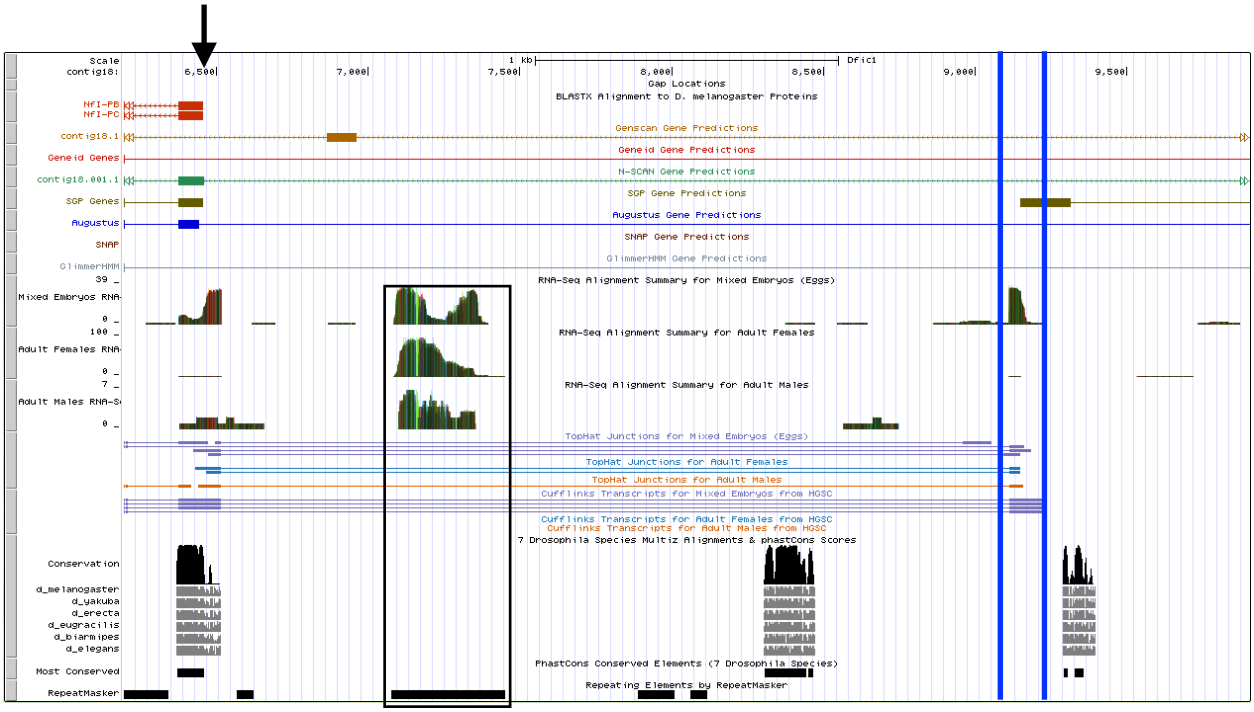


Figure 27: The 5' (upstream) UTR of *Nfl* in *D. ficusphila*. The TSS search region (between blue lines) is approximately 2.7 kb from the start codon (arrow), yet is supported by the *BLASTN* alignment in Figure 26, RNA-seq data, *TopHat* splice junctions, and *Cufflinks* transcripts. The black box highlights a ~500 bp region of high RNA-seq levels, identified as a repetitious element by *RepeatMasker*.

Feature 2

As indicated above, Feature 2 is likely orthologous to *D. melanogaster Syt7* (Figures 21, 22D). *Syt7* has five isoforms, but the A and M isoforms have identical coding sequences. The FlyBase annotation of *D. melanogaster Syt7* indicates that it is involved in calcium-dependent phospholipid binding and vesicle exocytosis of neurotransmitters. The results of the *BLASTX* alignments of contig18 (query) and *D. melanogaster* coding exons of *Syt7* (subject) are shown in Table 7. Coding exon 9_9584_0 (Exon 10), a terminal exon, is only two aa in *D. melanogaster*, so the GEP *Small Exon Finder* was used to find the orthologous exon in contig18 (Figure 28).

The bounds of the search region were defined by the flanking exons in *D. melanogaster*, i.e. the end of Exon 9 and the beginning of Exon 11. Since Exon 10 is translated to F* in *D. melanogaster*, the orthologous position in *D. ficusphila* is very likely at 13373-8 bp. Splice junction analysis revealed that the splice donor of Exon 9 is in phase 0 (see Table 8), further indicating this position for Exon 11. Contig18 aligns to the entirety of every exon, with 100% identity for nine of the eleven coding exons and at least 90% identity for all but one coding exon, indicating that *Syt7* is very highly conserved.

FlyBase ID	Exon ID	Coding Exon Size	Query Start	Query End	Query Frame	Subject Start	Subject End	Isoforms Present	Percent Identity
1_9584_0	Exon 1	53	23112	22954	-2	1	53	A,B,I,M	100%
2_9584_1	Exon 2	46	19605	19468	-2	1	46	A,B,I,M	93%
3_9584_1	Exon 3	27	19411	19331	-1	1	27	A,B,I,M	63%
5_9584_0	Exon 4	101	15458	15156	-3	1	101	A,B,I,M	100%
4_9584_0	Exon 5	96	15443	15156	-3	1	96	F	100%
6_9584_0	Exon 6	27	14667	14587	-2	1	27	A,B,F,I,M	100%
7_9584_0	Exon 7	56	14521	14354	-1	1	56	A,B,F,I,M	100%
10_9584_2	Exon 8	70	13482	13273	-2	1	70	A,F,M	100%
8_9584_2	Exon 9	13	13482	13444	-2	1	13	B	100%
9_9584_0	Exon 10	2	13378	13373	-1	1	2	B	100%
11_9584_2	Exon 11	33	13187	13089	-3	1	33	A,F,I,M	100%

Table 7: Summary of *BLASTX* searches of contig18 (query) to each *D. melanogaster* coding exon of *Syt7* (subject). Exon 10 is too small to be aligned with a *BLASTX* search and had to be identified using the GEP *Small Exon Finder* (Figure 28).

Search results

List of CDS that matched the search criteria:

Start	End	Translation	Acceptor Phase	Donor Phase	Sequence
13373	13378	F*	0	NA	TTTTAA
13268	13274	W*	1	NA	CTGGTAA

Figure 28: Finding the *Syt7* Exon 10 ortholog in contig18 using the GEP *Small Exon Finder*. The bounds of the search region were determined by the ending position of Exon 9 and the beginning position of Exon 11, which flank Exon 10 in *D. melanogaster*. The coding sequence is F* and the splice donor of Exon 9 is in phase 0, so the match at 13373-8 bp was chosen.

Although Exon 1 is the first coding exon for *Syt7*, two *TopHat* splice junctions extend upstream of it (Figure 29). These two splice junctions, JUNC00000528 and 529, have eleven and 50 supporting reads, respectively. According to the gene map of *Syt7* in *D. melanogaster* (Figure 22D), all five *Syt7* isoforms have 5' UTRs that begin within 1 kb of each other. Thus, these *TopHat* splice junctions likely correspond to these UTRs and indicate potential locations of TSSs. Further analysis would be required, as *Syt7* was not annotated for a TSS here due to lack of time.

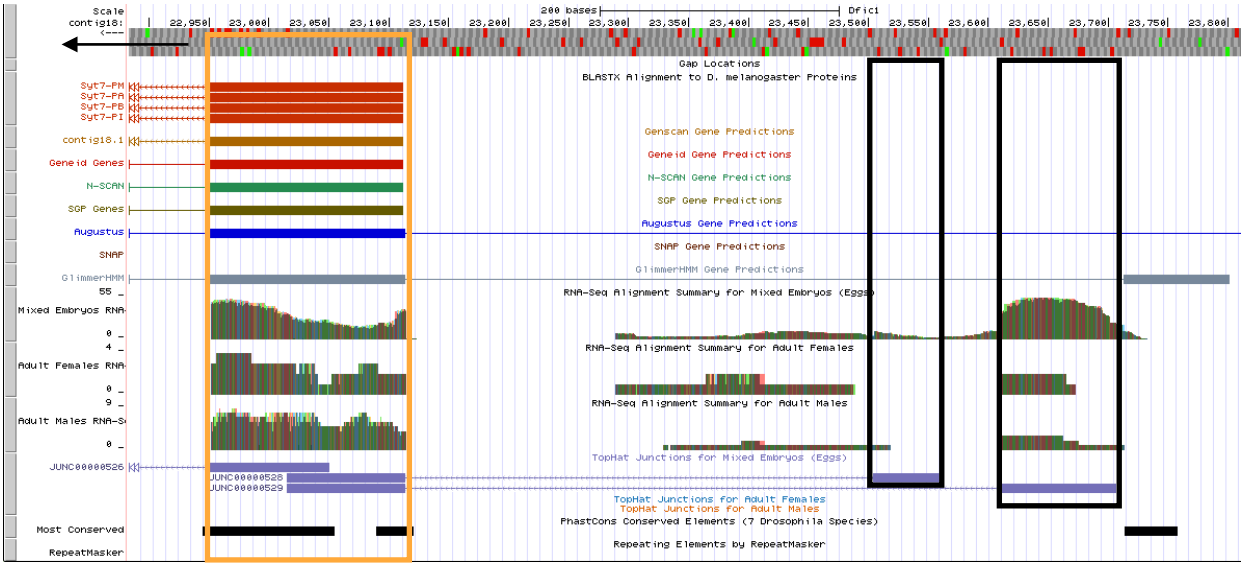


Figure 29: *TopHat* splice junctions extending upstream of Exon 1 for *D. ficusphila Syt7* (black boxes). Exon 1 (gold box) is the first coding exon for *Syt7*, yet JUNC00000528 and 529 extend upstream into regions with RNA-seq data, particularly the latter junction. These splice junctions likely correspond to one or more UTRs based on the gene map of *D. melanogaster Syt7* (Figure 22D).

Verification of Gene Model

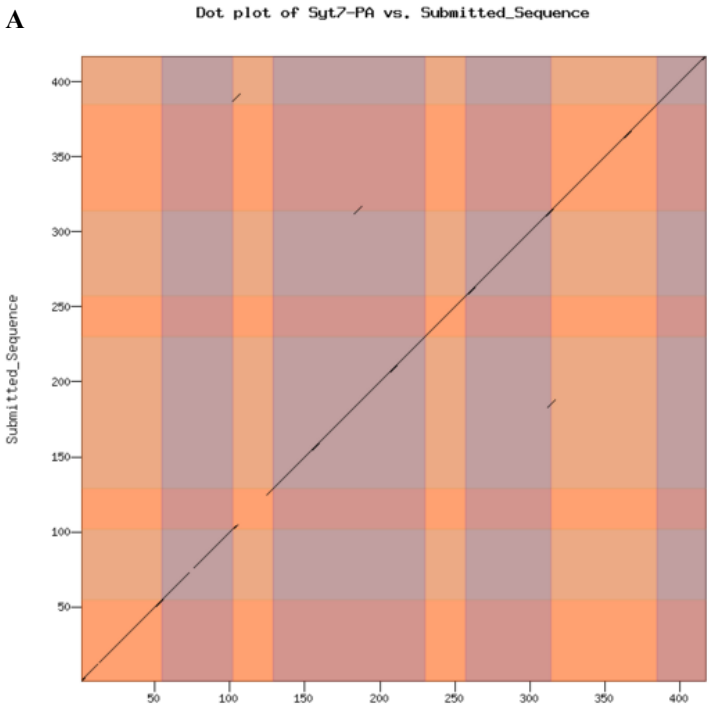
The *D. ficusphila* ortholog of *Syt7* was checked using the GEP *Gene Model Checker* (Figure 30). All five isoforms (A, B, F, I, and M) were annotated. Since the A and M isoforms have identical coding sequences, the dot matrices and protein alignments are identical; A is shown. Exon 3 shows significant divergence from *D. melanogaster*, but the beginning and end of the exon is conserved, providing anchors for the exon's location. All other exons are very highly

conserved: every putative isoform is over 95% similar to the *D. melanogaster* ortholog and the F isoform is completely identical. A summary of the final exon-by-exon annotation is shown in Table 8.

Exon ID	Coding Exon Size	Beginning	End	Reading Frame	Splice Acceptor Phase	Splice Donor Phase	Isoforms Present
Exon 1	53	23112	22952	-2	NA	2	A,B,I,M
Exon 2	46	19606	19466	-2	1	2	A,B,I,M
Exon 3	27	19412	19331	-1	1	0	A,B,I,M
Exon 4	101	15458	15156	-3	0	0	A,B,I,M
Exon 5	96	15443	15156	-3	NA	0	F
Exon 6	27	14667	14587	-2	0	0	A,B,F,I,M
Exon 7	56	14521	14353	-1	0	1	A,B,F,I,M
Exon 8	70	13484	13272	-2	2	1	A,F,M
Exon 9	13	13484	13444	-2	2	0	B
Exon 10	2	13378	13373	-1	0	NA	B
Exon 11	33	13189	13089	-3	2	NA	A,F,I,M

Table 8: Final exon annotation of the *D. ficusphila* ortholog to *Syt7*. Splice donor/acceptor pairs for the terminal exons are color-coded based on the corresponding isoform(s).

A



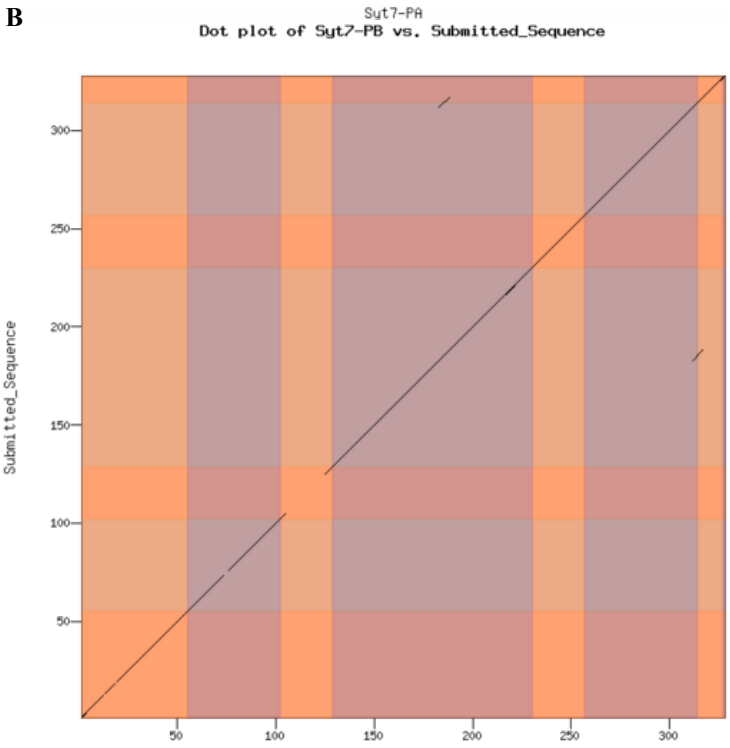
Alignment of Syt7-PA vs. Submitted_Seq

View plain text version

Identity: 401/416 (96.4%), **Similarity:** 405/416 (97.4%), **Gaps:** 0/416 (0.0%)

```
Syt7-PA           1 MASIVLIACLAILGLIITIALFLAGGYLWWRHKRSQLQFIEPNEDEESSSYSLRAAQDIV 60
                    **********:*****:*********************************************
Submitted_Seq     1 MASIVLIACLAVLGLIIAIALFLAGGYLWWRHKRSQLQFIEPNEDEESSSYSLRAAQDIV 60

Syt7-PA          61 DSGNPPTKPQVPVAHAITTPLQNNINRKLNGFLSLRTPLIGGSGASQTKPQIISSVGNPG 120
                    ***********..:*****************************:: :* * :**.** *
Submitted_Seq    61 DSGNPPTKPQVPAPQAITTPLQNNINRKLNGFLSLRTPLIGGSGTTPAKQQNVSSAGNVG 120

Syt7-PA         121 DGTTKDSANKSISMTDMYLDSTDPSENVGQIHFSLEYDFQNTTLILKVLQGKELPAKDLS 180
                    ***:************************************************************
Submitted_Seq   121 DGTSKDSANKSISMTDMYLDSTDPSENVGQIHFSLEYDFQNTTLILKVLQGKELPAKDLS 180

Syt7-PA         181 GTSDPYVRVTLLPDKKHRLETKIKRRTLNPRWNETFYFEGFPIQKLQSRVLHLHVFDYDR 240
                    ************************************************************
Submitted_Seq   181 GTSDPYVRVTLLPDKKHRLETKIKRRTLNPRWNETFYFEGFPIQKLQSRVLHLHVFDYDR 240

Syt7-PA         241 FSRDDSIGEVFLPLCQVDFAGKQSFWKALKPPAKDKCGELLSSLCYHPSNSILTLTLIKA 300
                    ************************************************************
Submitted_Seq   241 FSRDDSIGEVFLPLCQVDFAGKQSFWKALKPPAKDKCGELLSSLCYHPSNSILTLTLIKA 300

Syt7-PA         301 RNLKAKDINGKSDPYVKVWLQFGDKRVEKRKTPIFTCTLNPVFNESFSFNVPWEKIRECS 360
                    ************************************************************
Submitted_Seq   301 RNLKAKDINGKSDPYVKVWLQFGDKRVEKRKTPIFTCTLNPVFNESFSFNVPWEKIRECS 360

Syt7-PA         361 LDVMVMDFDNIGRNELIGRILLACKNGSGASETKHWQDMISKPRQTVVQWHRLKPE 416
                    ************************************************************
Submitted_Seq   361 LDVMVMDFDNIGRNELIGRILLACKNGSGASETKHWQDMISKPRQTVVQWHRLKPE 416
```

B



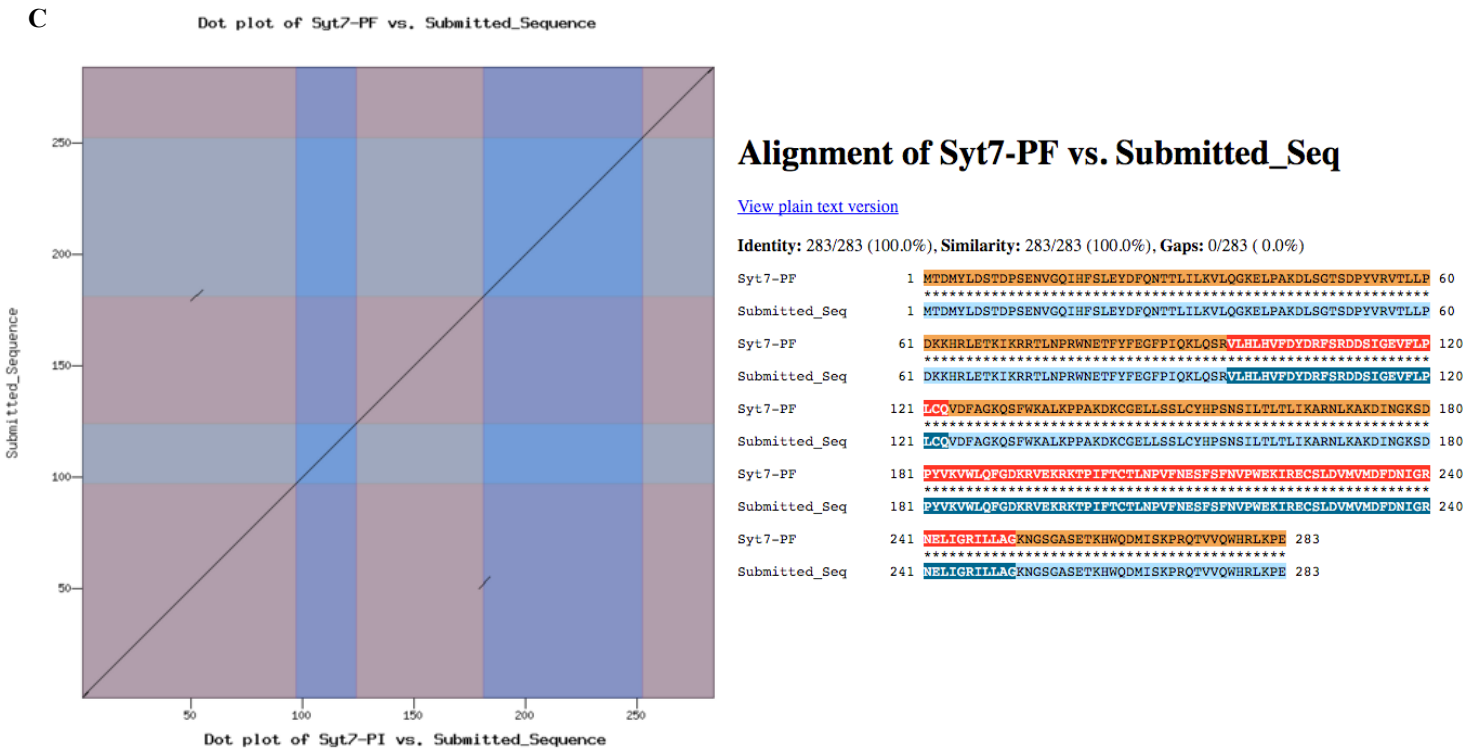
Alignment of Syt7-PB vs. Submitted_Seq

View plain text version

Identity: 312/327 (95.4%), **Similarity:** 316/327 (96.6%), **Gaps:** 0/327 (0.0%)

```
Syt7-PB           1 MASIVLIACLAILGLIITIALFLAGGYLWWRHKRSQLQFIEPNEDEESSSYSLRAAQDIV 60
                    **********:*****:*********************************************
Submitted_Seq     1 MASIVLIACLAVLGLIIAIALFLAGGYLWWRHKRSQLQFIEPNEDEESSSYSLRAAQDIV 60

Syt7-PB          61 DSGNPPTKPQVPVAHAITTPLQNNINRKLNGFLSLRTPLICGSGASQTKPQIISSVGNPG 120
                    ***********..:*****************************:: :* * :**.** *
Submitted_Seq    61 DSGNPPTKPQVPAPQAITTPLQNNINRKLNGFLSLRTPLICGSGTTPAKQQNVSSAGNVG 120

Syt7-PB         121 DGTTKDSANKSISMTDMYLDSTDPSENVGQIHFSLEYDFQNTTLILKVLQGKELPAKDLS 180
                    ***:************************************************************
Submitted_Seq   121 DGTSKDSANKSISMTDMYLDSTDPSENVGQIHFSLEYDFQNTTLILKVLQGKELPAKDLS 180

Syt7-PB         181 GTSDPYVRVTLLPDKKHRLETKIKRRTLNPRWNETFYFEGFPIQKLQSRVLHLHVFDYDR 240
                    ************************************************************
Submitted_Seq   181 GTSDPYVRVTLLPDKKHRLETKIKRRTLNPRWNETFYFEGFPIQKLQSRVLHLHVFDYDR 240

Syt7-PB         241 FSRDDSIGEVFLPLCQVDFAGKQSFWKALKPPAKDKCGELLSSLCYHPSNSILTLTLIKA 300
                    ************************************************************
Submitted_Seq   241 FSRDDSIGEVFLPLCQVDFAGKQSFWKALKPPAKDKCGELLSSLCYHPSNSILTLTLIKA 300

Syt7-PB         301 RNLKAKDINGKSDPYVKVWLQFGDKRF 327
                    ***************************
Submitted_Seq   301 RNLKAKDINGKSDPYVKVWLQFGDKRF 327
```

C



D

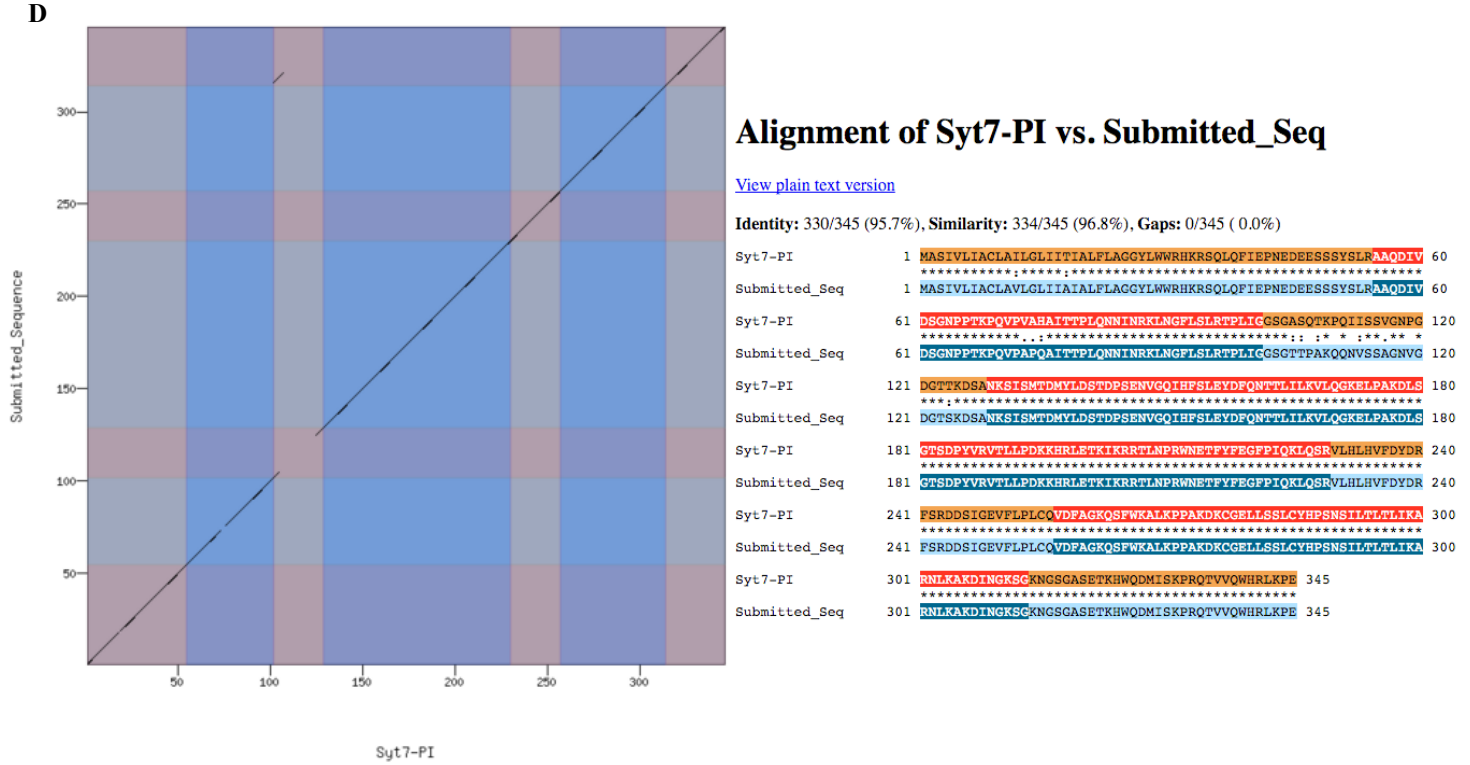


Figure 30: GEP *Gene Model Checker* dot matrices (left) and protein alignment (right) for *Syt7-PA* (A), *Syt7-PB* (B), *Syt7-PF* (C), and *Syt7-PI* (D). *Syt7-PM* is an identical coding sequence as *Syt7-PA*, so its dot matrix and protein alignment is identical to (A). Both the dot matrices and protein alignments show that *Syt7* is very well conserved. Most of Exon 3 has diverged significantly from the *D. melanogaster* ortholog, but the ends of the exon align well, indicating the exon likely exists.

Feature 4

All six *ab initio* gene predictors identified Feature 4, which is in agreement with preliminary *BLASTX* alignments, RNA-seq data, and *TopHat* splice junctions (Figure 31). The predicted peptide sequence from *Genscan* was used as the query in a *BLASTP* search to the FlyBase AA database for *D. melanogaster* (subject) (Figure 32). The best match, *Zinc/iron regulated transporter-related protein 102B* (*Zip102B*), has isoforms with E-values 98-133 orders of magnitude smaller than any other match and is on the fourth chromosome, indicating that Feature 4 is very likely orthologous to *Zip102B*. The FlyBase annotation of *D. melanogaster Zip102B* indicates it is a putative transmembrane transport protein for metal ions. The gene map and isoforms of *D. melanogaster Zip102B* is shown in Figure 33.

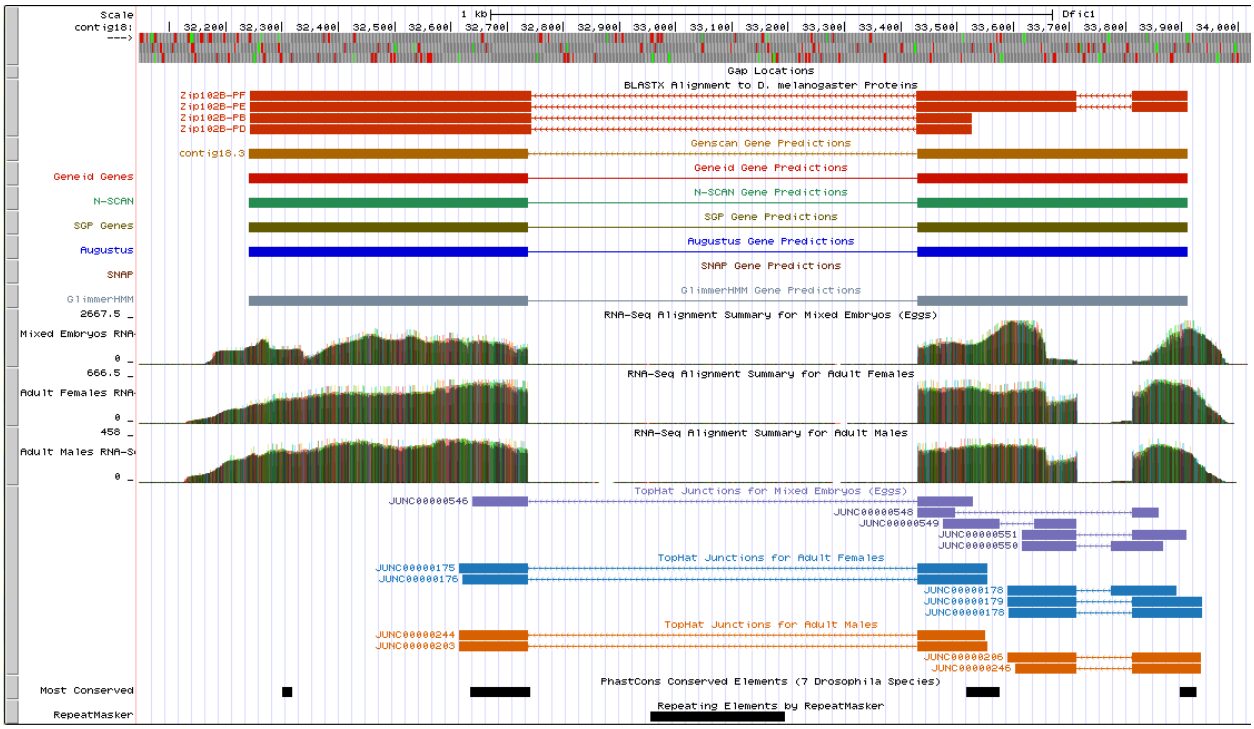


Figure 31: Closer view of Feature 4. All six *ab initio* gene predictors identify Feature 4 and agree with RNA-seq data and preliminary *BLASTX* alignments.

BLAST Hit Summary				
Description	Species	Score	E value
Zip102B-PF	Dmel	477.248	7.87444e-135
Zip102B-PE	Dmel	477.248	7.87444e-135
Zip102B-PD	Dmel	360.918	9.47813e-100
Zip102B-PB	Dmel	360.918	9.47813e-100
Catsup-PA	Dmel	37.3502	0.0230235
Zip99C-PJ	Dmel	34.6538	0.15821
Zip99C-PI	Dmel	34.6538	0.15821

Figure 32: *BLASTP* results of searching the *Genscan* predicted peptide sequence of Feature 4 (query) to the *D. melanogaster* FlyBase AA database (subject). *Zip102B* isoforms have the highest scores and smallest E-values. Thus, it is the likely ortholog of Feature 4.

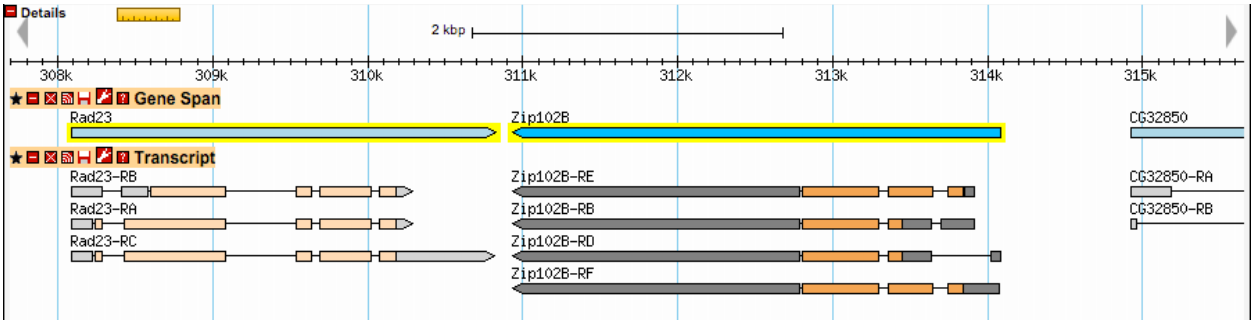


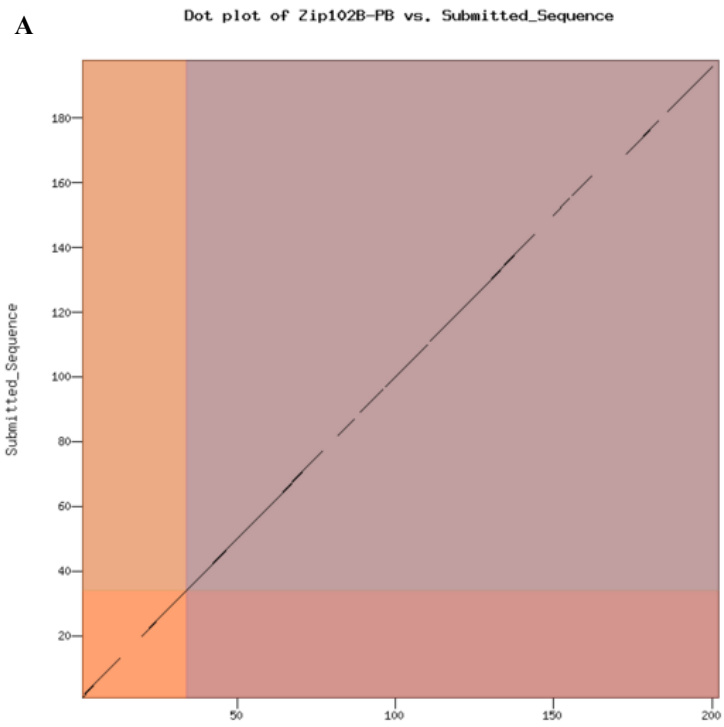
Figure 33: FlyBase gene map of *Zip102B* on the fourth chromosome in *D. melanogaster.* Light orange corresponds to coding regions and grey indicates UTRs. The B and D isoforms have identical coding sequences, as do the E and F isoforms. *Zip102B* is also syntenic with *Rad23*. In contig18, *Rad23* is syntenic with Feature 4 (Figure 1).

To determine the approximate coordinates of the exons orthologous to those of *Zip102B*, contig18 was used as a query in pairwise *BLASTX* alignments with coding exons of *D. melanogaster Zip102B* as the subject (Table 9). The entire subject was covered in every pairwise alignment. Exact coordinates were determined following the splice site analysis protocol used for Feature 3. This analysis was straightforward and did not identify any unusual features of the gene assembly.

FlyBase ID	Exon ID	Coding Exon Size	Query Start	Query End	Reading Frame	Subject Start	Subject End	Isoforms Present	Percent Identity
1_9585_0	Exon 1	33	33911	33813	-3	1	33	E,F	91%
3_9585_0	Exon 2	95	33713	33432	-3	1	95	E,F	86%
2_9585_0	Exon 3	32	33527	33432	-3	1	32	B,D	91%
4_9585_2	Exon 4	169	32735	32241	-3	1	169	B,D,E,F	86%

Table 9: Summary of *BLASTX* searches of contig18 (query) to each *D. melanogaster* coding exon of *Zip102B* (subject). All four coding exons are completely covered in the alignments.

A



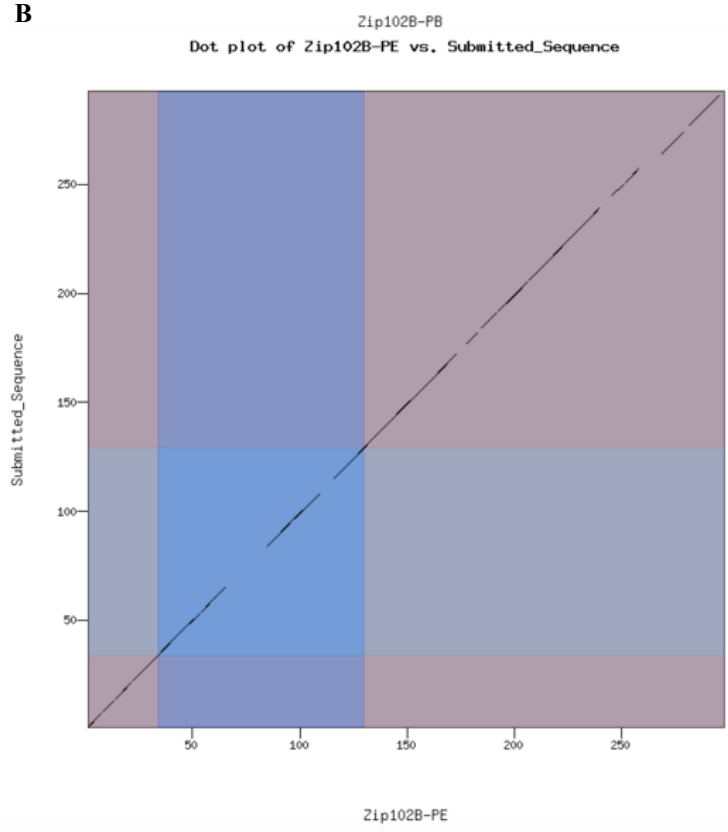
Alignment of Zip102B-PB vs. Submitted_Seq

View plain text version

Identity: 176/201 (87.6%), **Similarity:** 192/201 (95.5%), **Gaps:** 4/201 (2.0%)

```
Zip102B-PB          1 MMLVDISQRKSNVGSNKKNNATLTLGLVVHAAADGVALGAAATTSHQDVEIIVFLAIMLH 60
                      ***********:**:**.****************************************
Submitted_Seq       1 MMLVDISQRKSNIGSDKKSNATLTLGLVVHAAADGVALGAAATTSHQDVEIIVFLAIMLH 60

Zip102B-PB         61 KAPAAFGLVTFLLHEKVDRHQIRRHLVLFSLSAPLMTILTYFGIGQEQKDTLNSVNATGI 120
                      ****************:**::******.:*******:************:**********
Submitted_Seq      61 KAPAAFGLVTFLLHEKLDRKKIRRHLAVFSLSAPLLTILTYFGIGQEQKETLNSVNATGI 120

Zip102B-PB        121 AMLFSAGTFLYVATVHVLPELTQGGFTKSDQHDYCLLEESRGVATNDINGSNSIQALKYS 180
                      ***********************.**::::***** ******    ::****:********
Submitted_Seq     121 AMLFSAGTFLYVATVHVLPELTQNGFSQTDQHDYRLLEESR----DEINGSHSIQALKYS 176

Zip102B-PB        181 ELVILICGALLPLIITFGHNE 201
                      **:*:***************:*
Submitted_Seq     177 ELLIMICGALLPLIITFGHKE 197
```

B



Alignment of Zip102B-PE vs. Submitted_Seq

View plain text version

Identity: 259/297 (87.2%), **Similarity:** 278/297 (93.6%), **Gaps:** 5/297 (1.7%)

```
Zip102B-PE          1 MAEETIILILLVLVMLVGSYLAGSIPMLMKLSEEKLKCVTVLGAGLLVGTALVVIIPEGI 60
                      ******:*****:******* ****************************.*******
Submitted_Seq       1 MAEETIVLILLVFVMLVGSYAAGSIPMLMKLSEEKLKCVTVLGAGLLVGTALAVIIPEGI 60

Zip102B-PE         61 RSLYVGSGQSQKRTSVPEQPDYSQTIGLSLVLGFVFMMLVDISQRKSNVGSNKKNNATLT 120
                      ****:****.*    * **:  ***:**********************:**:**.*****
Submitted_Seq      61 RSLYMGSGQPQL-ISDPEHQDYSKTIGLSLVLGFVFMMLVDISQRKSNIGSDKKSNATLT 119

Zip102B-PE        121 LGLVVHAAADGVALGAAATTSHQDVEIIVFLAIMLHKAPAAFGLVTFLLHEKVDRHQIRR 180
                      ************************************************:**:.***
Submitted_Seq     120 LGLVVHAAADGVALGAAATTSHQDVEIIVFLAIMLHKAPAAFGLVTFLLHEKLDRKKIRR 179

Zip102B-PE        181 HLVLFSLSAPLMTILTYFGIGQEQKDTLNSVNATGIAMLFSAGTFLYVATVHVLPELTQG 240
                      **.:*******:**************:*********************************.
Submitted_Seq     180 HLAVFSLSAPLLTILTYFGIGQEQKETLNSVNATGIAMLFSAGTFLYVATVHVLPELTQN 239

Zip102B-PE        241 GFTKSDQHDYCLLEESRGVATNDINGSNSIQALKYSELVILICGALLPLIITFGHNH 297
                      **:::***** ******    ::*****:**********:*:**************:*
Submitted_Seq     240 GFSQTDQHDYRLLEESR----DEINGSHSIQALKYSELLIMICGALLPLIITFGHKH 292
```

Figure 34: GEP *Gene Model Checker* dot matrices (left) and protein alignment (right) for *Zip102B-PB* (A) and *Zip102B-PE* (B). *Zip102B-PD* has an identical coding sequence as *Zip102B-PB*, so its dot matrix and protein alignment are identical to (A). The same is true regarding *Zip102B-PE* and *Zip102-PF*. Both the dot matrices and protein alignments show that *Zip102B* is very well conserved.

Verification of Gene Model

The *D. ficusphila* ortholog of *Zip102B* was checked using the GEP *Gene Model Checker* (Figure 34). All four isoforms (B, D, E, and F) were annotated. Since the B and D isoforms have identical coding sequences, the dot matrices and protein alignments are identical; B is shown. The same is true for the E and F isoforms; E is shown. Furthermore, the B and D isoforms begin with Exon 3 (which starts in the middle of Exon 2) and is identical to the E and F isoforms thereafter, so the dot matrix for the B and D isoforms is a sub-matrix of that corresponding to the E and F isoforms. The exons show little divergence from those in *D. melanogaster*, indicating *Zip102B* is generally well-conserved. A summary of the final exon-by-exon annotation is shown in Table 10.

Exon ID	Coding exon size	Beginning	End	Reading Frame	Splice Acceptor Phase	Splice Donor Phase	Isoforms Present
Exon 1	33	33911	33813	-3	NA	0	E,F
Exon 2	95	33713	33431	-3	0	1	E,F
Exon 3	32	33527	33431	-3	NA	1	B,D
Exon 4	169	32737	32241	-3	2	NA	B,D,E,F

Table 10: Final exon annotation of the *D. ficusphila* ortholog to *Zip102B*. Splice donor/acceptor pairs for the terminal exons are color-coded based on the corresponding isoform(s).

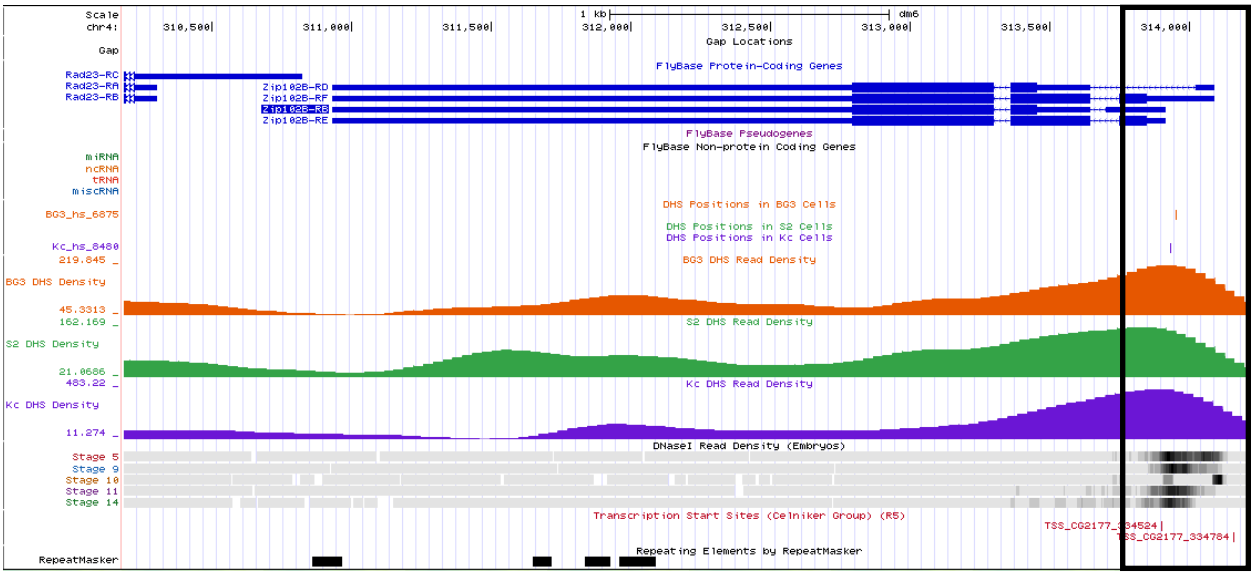


Figure 35: The 5' UTR and promoter region of *Zip102B* in *D. melanogaster*. The 9-state epigenomic tracks indicates the 5' end of *Zip102B* is an "Active promoter / TSS Region" (red) surrounded by heterochromatin (dark blue). Two TSS have bene previously annotated, one for the B and E isoforms and one for the D and F isoforms. Only one DHS exists in both the BG3 and Kc cell lines. Thus, *Zip102B* has two peaked promoters.

Transcription Start Site

The 9-state epigenomic landscape of *D. melanogaster Zip102B* shows the gene is actively transcribed in a heterochromatic environment for both BG3 and S2 cell lines (Figure 35). The D and F isoforms begin at the same location, as do the B and E isoforms (Figure 33). There are two TSS annotated by the modENCODE project at 314158 bp and 313898 bp, the former corresponding to the B and E isoforms, the latter to the D and F isoforms. BG3 and Kc cell lines both have one DHS, while no DHS exists in S2 cell lines. Thus, *Zip102B* has two peaked promoters, one for the B and E isoforms and one for the D and F isoforms.

The first transcribed exons in the E and F isoforms also contain the first coding exon for the respective isoforms (Figure 33). Both of these exons were used as the query in pairwise *BLASTN* alignments to contig18, the subject. While all but the first four bp of the first transcribed exon for *Zip102B-PE* aligns to contig18, the first 178 bp (52%) of *Zip102B-PF* fails to align (Figure 36). Neither alignment extends beyond 33980 bp; the first coding exon begins at 33911 bp (Table 10).

Figure 36: Pairwise *BLASTN* alignments of contig18 (subject), where the query is the first coding exon for *Zip102B-PE* (top) and *Zip102B-PF* (bottom). Both alignments span from the beginning of the *Zip102B* coding region (Table 10) to 33980 bp, although the first 178 bp do not align for *Zip102-PF*.

The TSS search region was determined to be between 33911 bp—the beginning of the coding region for *Zip102B*—and 34020 bp, where RNA-seq data begins (Figure 37). Although there is a *TopHat* splice junction extending beyond this region, it has less than ten supporting reads and spans across regions identified as repetitious by *RepeatMasker*. Any RNA-seq, *TopHat*, and *Cufflinks* data extending into this region was considered spurious. Very few core promoter motifs were found within the TSS search region (Table 11), although a DPE motif beginning at 33910 bp may suggest a TSS at 33938 bp. However, a second TSS could not be identified. RNA-PolII ChIP-seq data from *D. biarmipes* only indicates one peak for its *Zip102B* ortholog, yet extends approximately 500 bp upstream (Figure 38). Further analysis is necessary to determine whether there are one or two TSSs in the *D. ficusphila* ortholog of *Zip102B*.

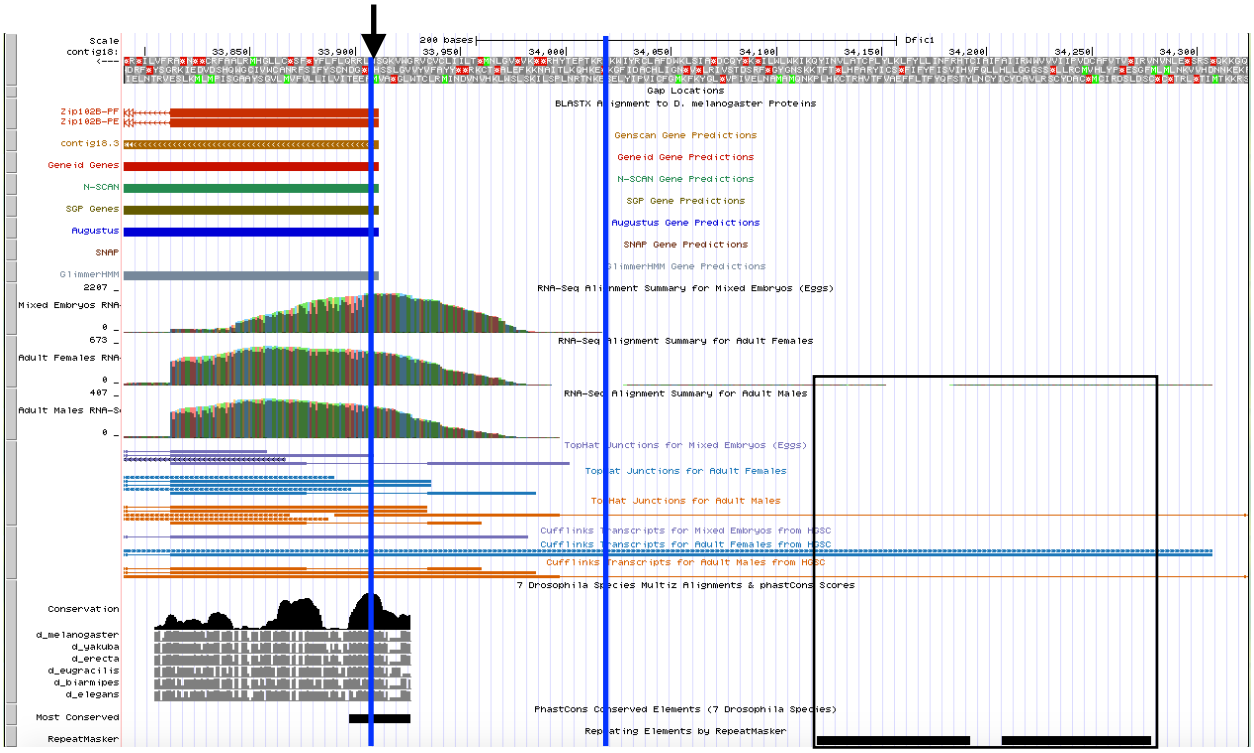


Figure 37: TSS search region for the *Zip102B* ortholog in *D. ficusphila* (blue lines). The TSS search region is bound by the beginning of the coding region (arrow) to where RNA-seq data ends. Although there are low levels of RNA-seq data, *TopHat* splice junction predictions, and *Cufflinks* transcripts beyond these regions (black box), they overlap with repetitious elements identified by *RepeatMasker* and thus was considered spurious.

Motif	*D. ficusphila* position	*D. melanogaster* position
TATA Box		314214, 314346
BRE[d]	33921, 33924, 33934	313666, 313692, 313792, 313805, 313807, 313957, 313976, 314000, 314019, 314105, 314232, 314234, 314441
Inr		313819
DPE	**33910**	313803, 314009, 314026, 314131

Table 11: Core promoter motifs within 300 bp of the *Zip102B* TSS search region in *D. ficusphila* and the two annotated TSS in *D. melanogaster*. A DPE found at 33910 bp (bold/underlined) suggests a TSS may exist at 33938. All positions are on the minus strand. No evidence was found for BRE[u], MTE, DRE or Ohler motifs 1 and 5-8, despite a search.

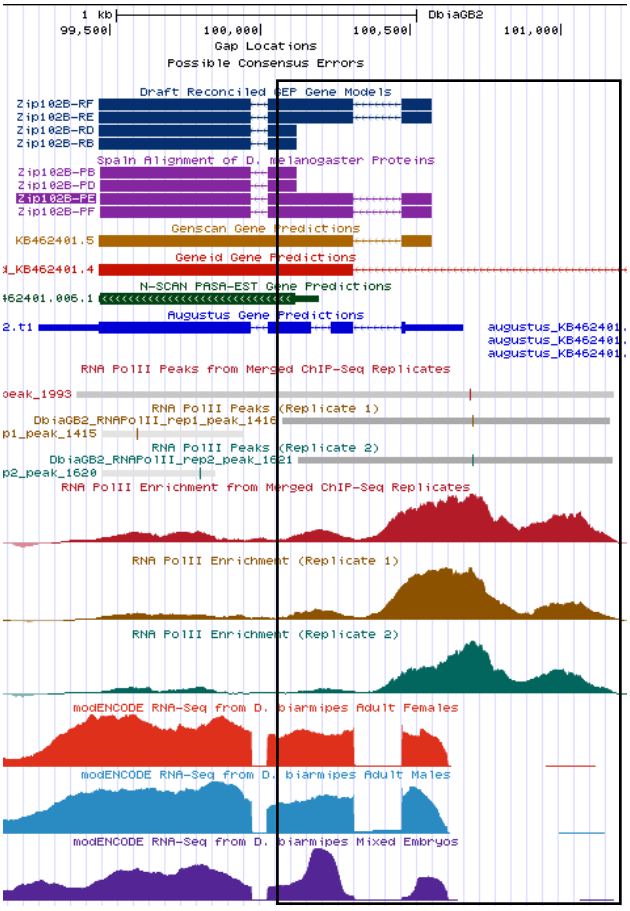


Figure 38: RNA-PolII ChIP-seq data for the *Zip102B* ortholog in *D. biarmipes*. Although there is only one peak in the 5' UTR region, the enriched region extends approximately 500 bp upstream of the peak (box). Thus, this evidence neither supports nor refutes the hypothesis of a single TSS in *D. ficusphila Zip102B*. Further analysis is necessary.

Repetitious Elements

Identification of repetitious elements allows for the classification of regions that may be remnants of transposable elements (TEs). *RepeatMasker* was used to identify repetitious elements of contig18 using a *D. ficusphila* repeat library. The results, shown in Figure 39,

identify 34.72% (13889/40000 bp) of contig18 as repetitious elements. Any repeat larger than 500 bp was considered a remnant of a TE and is listed in Table 12. These nine repeats account for 34.21% (13684/40000 bp) of contig18 and 98.52% (13684/13889 bp) of all repeats identified by *RepeatMasker*. To ensure none of these repetitious elements are due to *Wolbachia*, an endosymbiotic organism for some *Drosophila* species, the *Wolbachia* Riverside strain from *D. simulans* (GenBank accession CP001391.1) was used as the query in a *BLASTN* search of contig18 (subject). Only low complexity AT repeats matched (not shown), so no fragments of the *Wolbachia* genome are found in contig18.

```
==================================================
file name: contig18.fasta
sequences:            1
total length:      40000 bp  (40000 bp excl N/X-runs)
GC level:         35.15 %
bases masked:      13889 bp ( 34.72 %)
==================================================
               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:                7         3552 bp    8.88 %
      LINE1           0            0 bp    0.00 %
      LINE2           0            0 bp    0.00 %
      L3/CR1          0            0 bp    0.00 %

LTR elements:         6          940 bp    2.35 %
      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:        16         3830 bp    9.57 %
     hAT-Charlie      0            0 bp    0.00 %
     TcMar-Tigger     0            0 bp    0.00 %

Unclassified:        31         5567 bp   13.92 %

Total interspersed repeats:    13889 bp   34.72 %


Small RNA:            0            0 bp    0.00 %

Satellites:           0            0 bp    0.00 %
Simple repeats:       0            0 bp    0.00 %
Low complexity:       0            0 bp    0.00 %
==================================================
```

Figure 39: Repetitious elements in contig18 identified by *RepeatMasker* based on a library of *D. ficusphila* repeats. Of the 40000 bp in contig18, 13889 bp (34.72%) were identified as repeats.

Pos in contig18		Repeat Size	Matching repeat	Repeat class/family
Begin	End			
7	591	585	rnd-1_family-87	DNA/MITE
15524	16092	569	rnd-1_family-281	LINE
16122	17140	1019	rnd-1_family-406	LINE
16886	17739	854	rnd-1_family-282	LINE
17394	18434	1041	rnd-1_family-134	LINE
20606	21493	888	rnd-5_family-34	RC/Helitron
30711	31318	608	rnd-1_family-107	DNA/MITE
36679	37420	742	rnd-1_family-243	DNA/MITE
37458	38200	743	rnd-1_family-79	Unknown

Table 12: Repetitious elements in contig18 identified by *RepeatMasker* larger than 500 bp. These repeats are considered to be remnants of TEs and account for 98.52% of all repetitious elements identified by *RepeatMasker*.

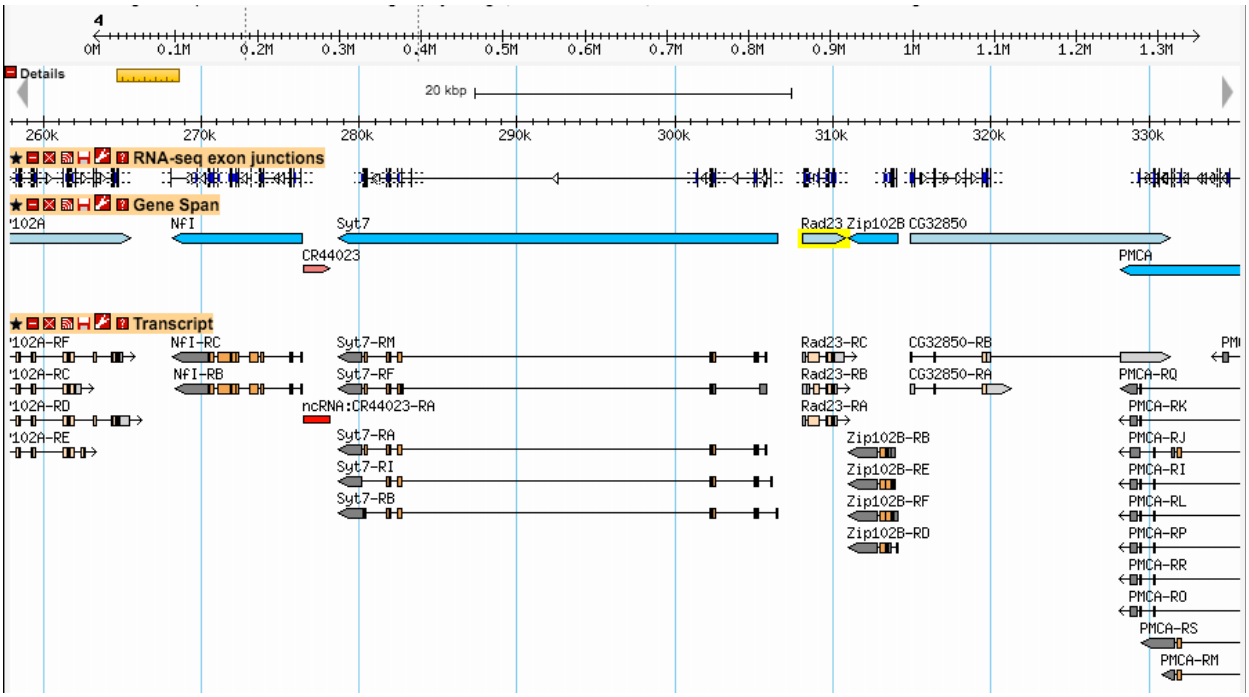


Figure 40: Gene map of the contig18 orthologous region in *D. melanogaster*. *NfI*, *Syt7*, *Rad23*, and *Zip102B* are in synteny with the orthologs on contig18. No *ab initio* gene predictor identified features orthologous to *CG32850*. These features required further inspection.

Synteny

To analyze the synteny of annotated genes in contig18 and *D. melanogaster*, the gene map of *D. melanogaster* had to be analyzed for the entire region orthologous to contig18 (Figure 40). Two features in *D. melanogaster* were not identified by any *ab initio* gene predictors and required further investigation. *CG32850* flanks *Zip102B*, which ends at 33911 bp (Table 10). The region of contig18 following *Zip102B* (i.e. 33,911-40,000 bp) was used as a query in a *BLASTX* search to both the NCBI *D. melanogaster* RefSeq protein database and the FlyBase *D. melanogaster* AA database (subjects). No significant similarity was found, indicating that *CG32850* is no located in contig18, but might be in the adjacent contig.

Since *CR44023* is a ncRNA, it cannot be recognized by the gene predictors a *BLASTX* search is not possible. Instead, *CR44023* was used as the query in a *BLASTN* alignment with

contig18 (subject). Only the first 103 bp out of 1652 bp (6.23%) aligned to contig18. In *D. melanogaster*, a DINE repeat is found within *CR44023* (Figure 41). DINE repeats are TEs, so it is possible the *CR44023* ortholog is found elsewhere in *D. ficusphila*. Indeed, there is little to no conservation in the Multiz alignments aside from the very beginning of the ncRNA. All conservation is within 250 bp of the *NfI* TSS annotated by modENCODE, so this can likely be attributed to regulatory regions necessary for TF binding. Thus, if the *CR44023* ortholog exists, it is likely elsewhere in the *D. ficusphila* genome, if present at all, and it was not annotated in contig18. Thus, the *D. ficusphila* orthologs of *NfI*, *Syt7*, *Rad23*, and *Zip102B* are in a syntenic configuration relative to their *D. melanogaster* orthologs (Figures 40, 42).

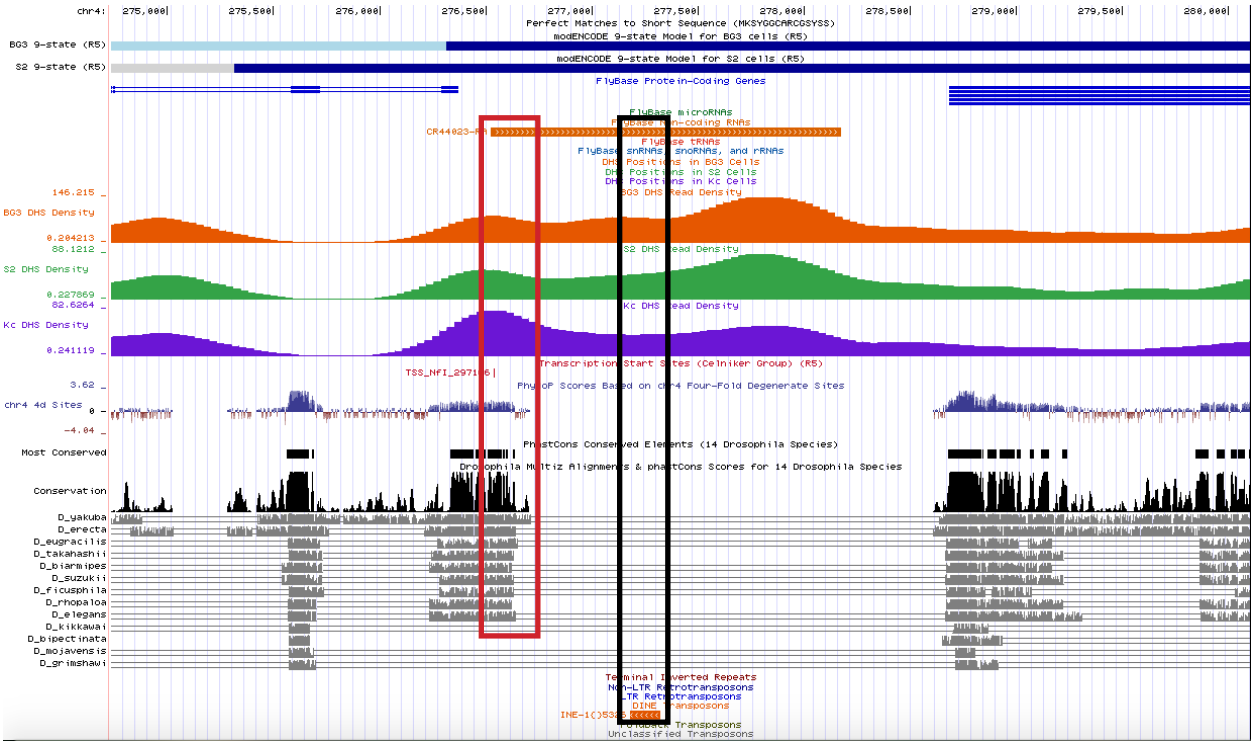


Figure 41: *CR44023* in *D. melanogaster*. A DINE TE within the ncRNA has been annotated (black box), indicating *CR44023* may be located elsewhere in *D. ficusphila*. The only Multiz alignment conservation occurs within 250 bp of the *NfI* TSS (red box) and is more likely due to TF activity near the TSS, not *CR44023*.

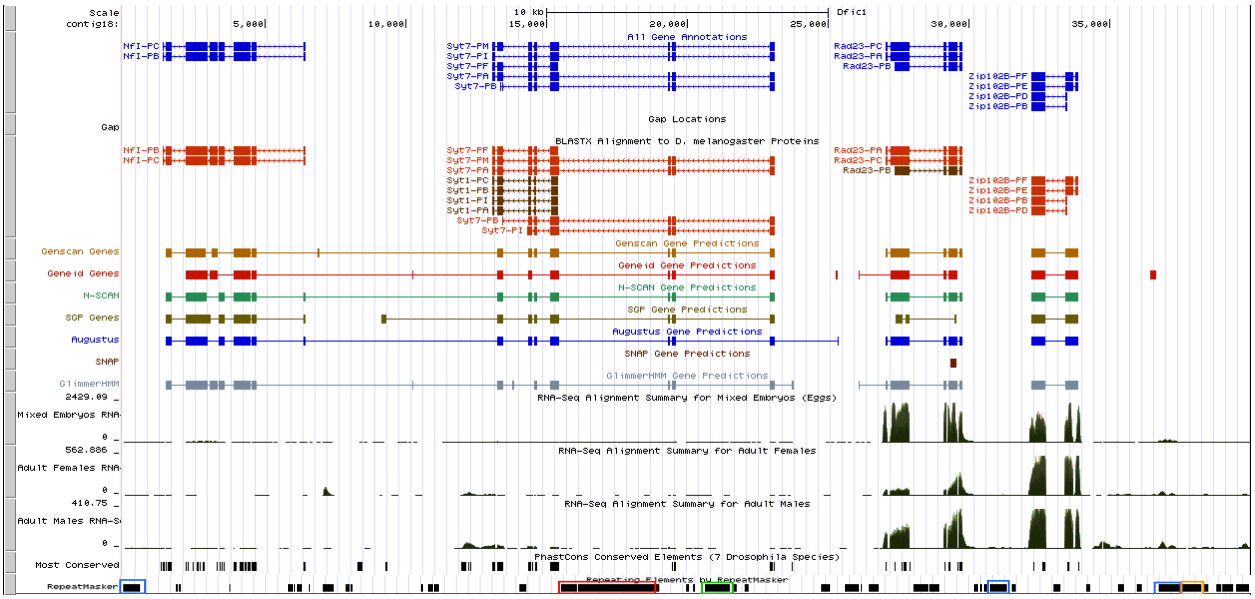


Figure 42: Final annotation of contig18, including TE remnants. Annotated genes and isoforms are in the top track in blue. TE remnants identified by *RepeatMasker*, in the bottom track, are boxed by the repeat class/family: DNA/MITE (blue), LINE (red), RC/Helitron (green), and unknown (orange).

Discussion

The final annotation for contig18, including TE remnants, is shown in Figure 42. Contig18 contains four genes, *Nfl*, *Syt7*, *Rad23*, and *Zip102B*. All exons and isoforms found in *D. melanogaster* are also found in contig18 of *D. ficusphila*. The TSS for *Rad23* was annotated at 26923 bp with strong support from a *BLASTN* alignment, core promoter motifs, and RNA-seq data. The TSS for *Nfl* was annotated at 9191 bp, far from the beginning of the coding region. However, a repetitious element was identified in this range, explaining the distance. Only one putative TSS, at 33938 bp, for *Zip102B* was annotated, but this is only due to a DPE core promoter motif located within the TSS search region. Further analysis is required to determine if a second TSS (for *Zip102B-PD* and *Zip102B-PF*) is found in *D. ficusphila*, which was expected based on the modENCODE annotation of the *D. melanogaster* gene. A *Clustal Omega* alignment of Rad23 protein orthologs revealed that while some divergence exists, the protein motifs identified by a *BLASTP* search of *D. ficusphila* Rad23 are well conserved from flies to humans.

CG32850, which is in the adjacent sequence *D. melanogaster*, was not found in contig18 and should be in the adjacent contig. The ncRNA *CR44023* also was not found, but because it is not well conserved and contains a TE, it could be found elsewhere in the *D. ficusphila* genome. This completes the annotation of contig18 in *D. ficusphila*, which contributes to the full annotation of the genome. This also allows for an intra-genus analysis of the Muller F element in an attempt to identify novel sequence motifs and provides useful data for future comparative genomics studies.

References

Graveley, Brenton R., et al. "The developmental transcriptome of Drosophila melanogaster." *Nature* 471.7339 (2011): 473-479.

Appendix

FASTA, PEP, and GFF files for each isoform of each gene are submitted electronically.

Addendum

The *Ekar* ortholog in contig33 was also annotated. For further details, see the report of Ron Nwumeh, since the contig was his project.

Acknowledgements

I would like to thank the Bio 434W teaching team: Dr. Sarah Elgin, Dr. Christopher Shaffer, Dr. Wilson Leung, Yu He, and Daniel Cui Zhou for their guidance and expertise, and Dr. April Bednarski for help in improving writing. I would also like to thank Washington University in St. Louis and the Genomics Education Partnership for providing opportunities for guided research.