Sonal Singhal
3 May 2006
Bio 4342W

Annotating the *D. virilis* Fourth Chromosome: Fosmid 99M21

Abstract

In this project, I annotated a chunk of the *D. virilis* fourth chromosome (fosmid 99M21) by considering genes, repeat structure, synteny, and conserved coding and non-coding regions. Using multiple tools and databases, I was able to complete this project. Two partial but likely functional genes are described, one that shows similarity to *toy* and the other that shows similarity to cathepsin-L in comparison to *D. melanogaster*. Analysis of repeats increased the overall percentage of repeats by nearly 10% and found four possible novel repeats. Studying the synteny of this fosmid suggests that the *D. virilis* dot chromosome might share genetic material with *D. melanogaster* chromosome 2. Finally, ClustalW analysis helped identify a putative promoter for *toy* and showed the remarkable conservation of cathepsin-L. My final annotation is shown in Figure 1.

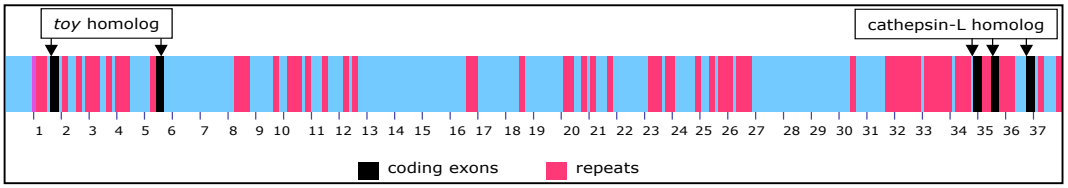


Figure 1: Final annotation of Fosmid 99M21.

Introduction

Sequencing a genome merely provides the order of base pairs on a DNA strand—an appropriate analogy might be that it provides us with a code that we must decipher before it makes sense. Annotation can be seen as the process of decoding a genome, because it extracts important biological information by analyzing the functional elements in a genome. In this class, we are annotating the fourth and largely euchromatic chromosome of *Drosophila virilis* so that we can compare it to the already-annotated fourth and largely heterochromatic chromosome of *Drosophila melanogaster*. By specifically considering chromosomal-wide changes in repeat density and distribution, synteny, and gene organization, we hope to better understand how heterochromatin forms. In this paper, I discuss my contribution to this project: the annotation of a fosmid (99M21), containing sequence from the fourth chromosome of *Drosophila virilis*. This fosmid is approximately 37 Kb in size and has a G/C content of 38%. With respect to my fosmid, I will discuss (1) identified genes, (2) repeat structure, (3) synteny with *D. melanogaster*, and (4) conservation of genic and non-genic regions.

Gene Finding

Method

I used the same basic procedure to annotate all genes in the fosmid 99M21. Following application of RepeatMasker, the *ab initio* gene finder Genscan was used to identify all possible coding features in the fosmid. As shown in Figure 2, Genscan predicted three features in my fosmid. Each feature was handled separately. To determine possible homology for the prospective gene in *D. melanogaster*, I used Blat to search the *D. melanogaster* genome and

blastx to serach the refseq database, looking for matches to my predicted coding sequence (cds). Through these searches, I was able to find the putative homolog for the feature in question, and I also determined what portion of the putative homolog was encoded by my fosmid. I then found the putative homolog in Ensembl and used the transcript information available through this website to build a gene model. A gene model describes the spacing, number, and length of exons for the gene in question. For this project, when multiple transcripts are available, we used the transcript that contains the most genetic information (i.e., that has the most amino acids).

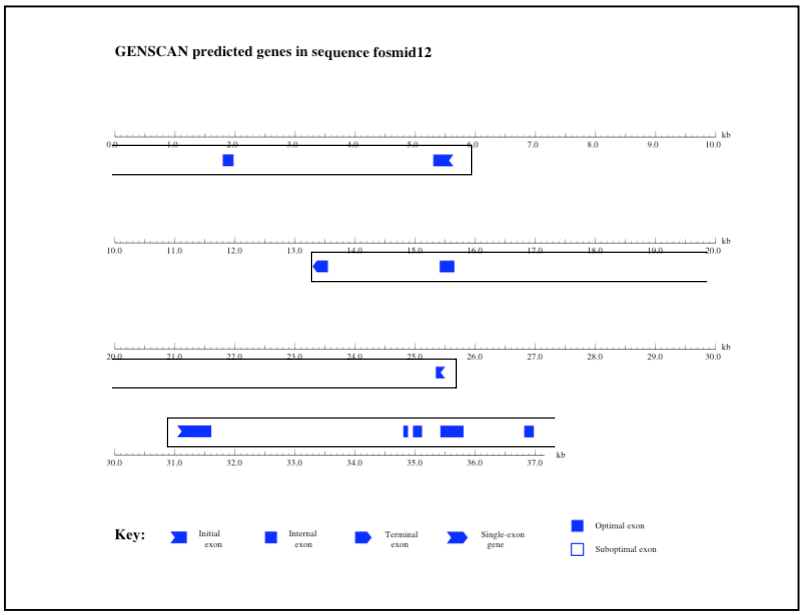


Figure 2: Genscan output for Fosmid 99M21.

After determining the gene model, I then attempted to characterize each exon in the model separately. I used bl2seq (tblastn), using the amino acid sequence from an exon to search the fosmid sequence to determine exon boundaries. In order to maximize matches, all searches were run without the "low complexity filter" and with an expect value of 1000. Generally, the results from these searches described the exon boundaries well, providing me with the coordinates for both the start and stop sites for the exon. I then did a first-pass check to determine if the exons were described (as appropriate) by start and stop codons and splice acceptor and donor sites. I modified my descriptions as necessary to conform to these rules without changing significantly the peptide that would result. Once my initial characterization of the exons had been confirmed, I used Wilson Leung's program "Annotation Check" to do a more thorough and reliable check of my annotation. If my annotation passed the check, I used bl2seq (blastp) to compare the polypeptide predicted by the concantenated exons to the homologous polypeptide from *D. melanogaster*. If I saw any drastic deviations, I re-evaluated my annotation as necessary—particularly, ensuring that exon boundaries and exon phases had been accurately defined. Here, exon phase refers to how the reading frame of each exon compares to the gene as a whole. Finally, because students in previous classes had already annotated both of my genes, I compared my annotation to the earlier class annotation.

Feature 1

Feature 1 consists of two exons and is on the minus strand of the fosmid. Genscan predicted an initial exon, but it did not predict a terminal exon. Thus, my initial suspicion was that the fosmid only contains a partial 5' region of the gene. Using blastp, the predicted Genscan peptide was used to search the complete coding sequence (CDS) database. Blastp predicted a conserved homeodomain in the peptide, suggesting that this peptide encodes a transcription factor. Indeed, the search shows that the predicted peptide has high homology to the *toy* gene from *D. melanogaster*, which encodes a transcription factor that is similar to *eyeless*. *toy* is located on the fourth chromosome of *D. melanogaster*. Developed with *D. melanogaster* transcript information from Ensembl, the gene model for *toy* is unambigious—there is only one characterized splicing of *toy* mRNA which consists of seven exons (Figure 3). Using bl2seq, I used the peptide predicted from each individual exon to search my masked fosmid sequence using tblastx. Doing so confirmed that my fosmid contains the first two exons of the *toy* gene (Figure 4, Table 1). As determined by the Annotation Check program, these predicted exons are corroborated by the presence of appropriate splice acceptor and donor sites. Further, the final peptide shares very high homology with the *toy* peptide from *D. melanogaster*.

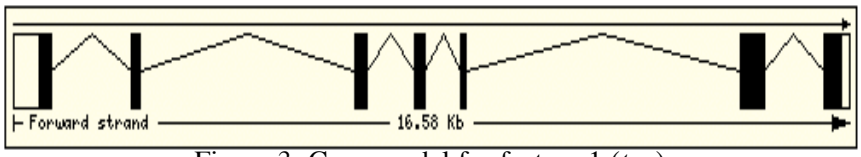


Figure 3: Gene model for feature 1 (*toy*).

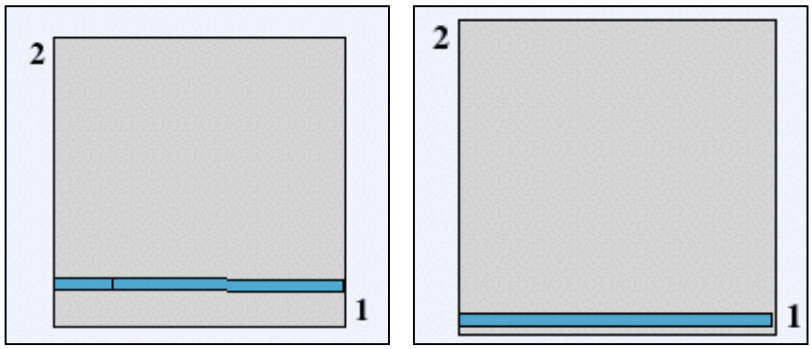


Figure 4: bl2seq matches for feature 1; exon 1 (L) and exon 2 (R). The X axis is the exon sequence from *D. melanogaster*, the Y axis is the fosmid sequence.

feature 1	% similarity	% positive	boundaries
exon 1	94%	94%	5559-5297
exon 2	100%	100%	1973-1793

Table 1: Exon predictions for feature 1.

As can be seen in Table 1, the *D. virilis* and *D. melanogaster* genes are very similar, despite the long evolutionary history that separates these two species. This similarity is not surprising, considering the importance of the genes. *toy*, or *twin of eyeless*, is a homeodomain-containing transcription factor that is key to proper imaginal disk development in insects.

Mutations in *toy* can lead to homeotic mutants, which often have abnormal or misplaced eyes. *toy* is the paralog of *eyeless* and likely arose due to a gene duplication. In other species, *toy* shows similarity to Pax-6 proteins (Gehring and Ikeo, 1999).

Feature 2

Genscan predicted that Feature 2 consisted of three exons on the minus strand. A blastp search of the predicted amino acid sequence against the nr database shows that this feature is likely a misprediction. The predicted protein has no good matches in the database (Figure 5). As a further check, I extracted the region in which this gene is found (15 Kb to 26 Kb) and used tblastx to search for any open reading frames matching the refseq database. As no significant matches were found, this confirmed my initial conclusion that this feature is a misprediction.

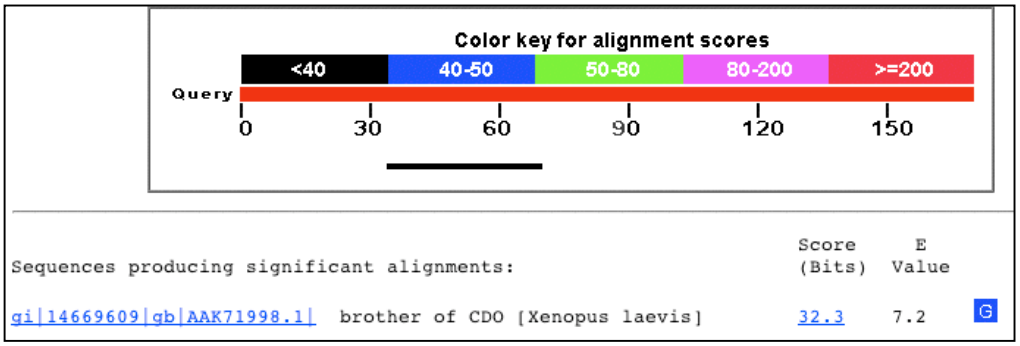


Figure 5: Blastp result for Genscan predicted peptide.

Feature 3

Feature 3, as predicted by Genscan, consists of five exons and is on the plus strand of the fosmid. Genscan predicted an initial exon, but because it did not predict a terminal exon, my initial suspicion was that the fosmid only contains the partial 5' region of the gene. Using blastp, the predicted Genscan peptide was used to search the complete coding sequence (CDS) database. Blastp predicted a conserved peptidase domain in the peptide, suggesting that this peptide encodes a protease. Indeed, this search shows that the predicted peptide has high homology to the CG5367 gene from *D. melanogaster*, which is a putative cathepsin-L gene. Developed with *D. melanogaster* transcript information from Ensembl, the gene model for CG5367 is unambigious—there is only one characterized splicing of CG5367 mRNA which consists of five exons (Figure 6). Using bl2seq, I searched with the peptide predicted from each individual exon against my masked fosmid sequence using tblastx. Doing so confirmed that my fosmid contains exons 2-4 of CG5367 (Figure 7, Table 2). However, I was unable to find a significant match to exon 1, which problematically left the protein without a start codon. Finally, I noted that CG5367 is on chromosome arm 2L in *D. melanogaster*.

Figure 6: Gene model for feature 3.

feature 3	% similarity	% positive	boundaries
exon 2	57%	75%	35001-35114
exon 3	75%	86%	35418-35804
exon 4	83%	92%	36814-36972

Table 2: Exon match for feature 3.

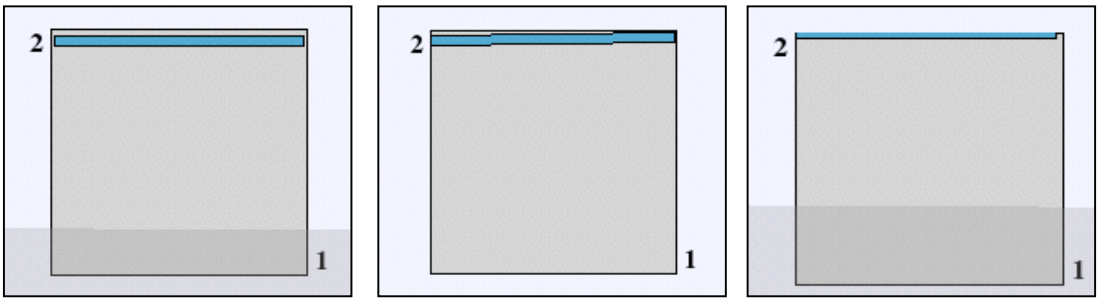


Figure 7: bl2seq matches for feature 3; exons 2-4 from L→R. The X axis is the exon sequence from *D. melanogaster*, the Y axis is the fosmid sequence.

Determining the location of exon 1 was a challenge. First, I used ClustalW, which is a more sensitive alignment tool than BLAST, to find whether exon 1 from *D. melanogaster* had a significant match to *D. virilis*. This method was unsuccessful. Second, I decided to explore whether the fact that CG5367 is on two different chromosomes in *D. virilis* and *D. melanogaster* is the basis for the missing exon 1. Using FlyBase, I located the position of a few genes and exons 1 and 3 of CG5367 in a number of *Drosophila* species (Table 3). Because few species' genomes outside that of *D. melanogaster* have been completely annotated, many of these locations are given with respect to a scaffold number rather than a chromosome. Further, these matches are putative, and thus, should be considered cautiously. Despite these limitations, a clear pattern arises: in *D. melanogaster* and the closely related species *D. simulans*, *D. yakuba*, and *D. erecta*, exons 1 and 3 of CG5367 are uniformly on chromosome 2. In *D. virilis* and the closely related species *D. grimshawi*, this search suggests that exon 1 has remained on chromosome 2 while exon 3 is now on chromosome 4. *D. mojavenesis* remains an ambiguous case. This suggests that there was a translocation of part of chromosome 2 onto chromosome 4 that led to the exons of CG5367 being split apart. Other fosmids that our class is studying also show evidence of a translocation.

				two coding genes on D.melanogaster 2L near feature 3		D. melanogaster chr4 gene
species	exon 1	exon 3	*toy*	CG5369-PA	CG5366-PA	PlexA-RB
	scaffold or chromosome on which the following elements are found					
D. grimshawi	15126	14822	14822	15126	15126	14822
D. virilis	12963	13052	13052	12963	12963	13052
D. mojavenesis	6496	6498	6498	6500	6540	6498
D. melanogaster	chr2L	chr2L	chr4	chr2L	chr2L	chr4
D. simulans	chr2L	chr2L	chr4	chr2L	chr2L	chr4
D. yakuba	chr2L	chr2L	chr4	chr2L	chr2L	chr4
D. erecta	4929	4929	4512	4929	4929	4512

Table 3: Location of exons 1 and 3 of cathepsin-L and other genes.

Because this gene has no clearly defined start codon, it is tempting to characterize it as a pseudogene. Yet, there is remarkable conservation of the gene between *D. virilis* and *D. melanogaster*, and there is no evidence of a pseudogene (i.e., frameshift mutation or premature stop codon). As such, I propose that CG5367 in *D. virilis* has a novel start codon, either (1) upstream of exon 2 or (2) within the exons already characterized. To explore the first possibility, I found all possible open reading frames (ORFs) up to 6 Kb away from the start of exon 1. Here, I chose to define putative ORFs as any region of amino acids at least 40 amino acids in length that begins with a start codon and that does not contain either repetitive elements or stop codons. I was able to identify three ORFs that met these criteria. The first ORF found is not conserved across *D. virilis*, *D. grimshawi*, and *D. mojavenesis*—in particular, in the same ORF in the other species there are numerous stop codons. As the ORF is postulated to code for a functional product, we would expect to see conservation in these three closely related organisms. The second ORF, when searched against the refseq database using blastx, is found to contain part of a putative retroviral element. Similarly, the third ORF, while not identified as repetitive DNA, is not unique. As such, these three ORFs are not good candidates for the first exon of CG5367.

To explore the second possibility, I searched for possible start codons within the exons already annotated. The first start codon I found was 11 amino acids into exon 2 (shown in red in Figure 8)—however, this start codon is not conserved in *D. grimshawi* and *D. mojavenesis*. The second possible start codon found is conserved in the three species (shown in blue in Figure 8), but it is the last amino acid of exon 2. I doubt that a gene would begin with an exon that is only 1 amino acid long. Although this can happen, the conservation upstream of the start codon is atypical for a UTR. Typically, these regions diverge quickly. As such, I choose to annotate the gene with the assumption that the coding sequence begins with the amino acid shown in red in Figure 5. The final annotation is summarized in Table 2. As determined by the Annotation Check program, these predicted exons are corroborated by the presence of appropriate splice acceptor and donor sites. Further, the final peptide shares high homology with the CG5367 peptide from *D. melanogaster*. Although this annotation is not ideal, it is the most satisfying option. Unfortunately, there is no expressed sequence tag (EST) data currently available to help evaluate this annotation. As we get expression data for *D. virilis* and closely related species, I believe we will be better able to annotate this gene.

```
dvir            NNRSYARSHDEMRSYEAYEENQIIVNEHNTYYETGKSSFRLATNTMADMNTDSYLKGYLR 60
dgri            --KSYMRLHDEIRSYKSFEENIRIVNEHNKFYETGKSSFRLSTNTMADMNTDSYLQGFLR 58
dmoj            --RSYIRAHDELRSYKAYEENHQIVEAHNKDYDSGRSSFRLAANTLADMSTDSYLKGFLR 58
                  :** * ***:***:::***  **: **. *::*:*****::**:***.*****:*:**

dvir            LLRSPEISDSDNIADIVGSPLMNNVPESFDWRKKGFITPLYNQQSCGSCYAFSIAQSIEG 120
dgri            LLRSPPNSTTDNIADIVGSPLMNNVPESFDWRKKGFNTPPYNQQSCGSCYAFSVAQSIEG 118
dmoj            LLRSPPISASDNMVDIVGATLMDNVPDSLDWRKKGFATPSYNQQSCGSCYAFSIAQSIEG 118
                ***** * :**:.****:.**:***:*:******* ** **************:******

dvir            QVFKRTGKIVALSEQQIVDCSVSHGNQGCIGGSLRNTLRYLQATGGLMR---------- 169
dgri            QVFKRTGKLLALSEQQIVDCSVSHGNHGCIGGSLRNTLTYLQATGGLMR---------- 167
dmoj            QVFKRTGRILSLSEQQIVDCSISHGNQGCTGGSLRNTLRYLQATGGLMRYAFSIAQSIEG 178
                *******::::**********:****:** ******** **********
```

Figure 8: Second exon alignment of cathepsin-L.

CG5367, as a putative cathepsin-L, encodes a cysteine protease. Cysteine proteases are key to proteolysis, a common method of degrading cellular products. Although little is known about CG5367 in flies, mutations in the gene can be lethal. In humans, cathepsin-L is the most-

active isoform of cathepsin (OMIM 2006). Localized in lysozymes, cathepsin-L helps affect the development and degradation of epithelial cells.

Repetitive Elements

Method

The RepeatMasker library commonly used to mask repetitive elements is designed for use with *D. melanogaster*, however, *D. virilis* contains some repetitive elements that are unique to its species and its relatives. Wilson Leung had updated the database in 2005 to reflect these unique repeats, but some repeats still remain unidentified. Further, some repeats go unidentified because they are too short or are interrupted by another repeat. To remedy this, I looked for unidentified repeats in my masked fosmid sequence by using blastn to search the *D. virilis* fosmid database. By viewing the BLAST results in Herne, I was able to delineate what regions contained multiple hits to different fosmids. These regions are identified as repeats, which I further characterized as either putative novel repeats or continuations of existing repeats. To determine the identity of putative novel repeats, I used blastn to look for the extracted repeat in the repeat database. Further, I used ClustalW to determine if the repeats contained any internal repeats. I also used blastx to look for the extracted repeat in the refseq database to ensure that it did not contain coding material. These steps ensured that the repeat was truly novel.

Results

An initial scan of my fosmid with RepeatMasker identifies 23.3% of the sequence as repeats: specifically, LINEs are 9.0%, DNA elements are 8.6%, and simple and low complexity repeats are 4.0% (Appendix B). Through the BLAST search, I identified an additional 27 repeats, 5 of which I identified as continuations of existing repeats due to their proximity to existing repeats (Appendix C). An additional 18 repeats (which represent 383 base pairs of sequence) will not be discussed because they are less than 30 base pairs long and because they do not seem to be continuations of existing repeats. The remaining 4 repeats are putative novel repeats. In total, the repeats represent 2723 base pairs of sequence, bringing the revised percentage of repetitive sequence to 30.6%.

The four putative novel repeats (repeats 23, 71, 82, 87; shown in Figure 9 and listed in Appendix C) do not match significantly to any currently defined repeats nor do they show significant similarity to any coding region. Extracting these regions and using BLAT to find possible matches in *D. virilis* confirms that all four are true repeats as they match to multiple regions in the genome. Repeats 23 and 87 are the same repeats, although one of these repeats is inverted with respect to the other. Further, repeat 87 occurs between exons 3 and 4 of CG5387. Repeat 71 contains a tandem inverted repeat. I was unable to further characterize repeat 82. These repeats should be characterized further to determine if they are truly novel as this initial evidence suggests they are. In the future, it would be of interest to search for remnants of *gag*, *pol* or DNA transposase in these putative novel repeats. These three elements are common markers of retroviral or DNA elements, and it is common to find them at the end of repeats.

Synteny

Method

To determine synteny, or conservation of gene order, I considered location and orientation of the gene features on my fosmid of *D. virilis* and how they compare to *D. melanogaster*. Location and orientation of gene features in *D. melanogaster* was determined by

Figure 9: Repeats 23, 71, 82, and 87 (Left → Right) as seen in Herne. These regions are characterized as repeats because they match to many regions in other *D. virilis* fosmids

using Ensembl. To better define changes in chromosomal evolution, I also determined where random, non-repetitive elements of *D. virilis* sequence matched to *D. melanogaster*. I used Blat and blastn to do so, searching the *D. melanogaster* genome in both cases. I only considered matches that had a Blat score higher than 75 or E-value smaller than 10^{-15}.

Results

Feature 1 (*toy)* is on the minus strand of the fourth chromosome of *D. virilis* and on the plus strand of the fourth chromosome of *D. melanogaster*. Feature 3 (putative cathepsin-L) is on the positive strand of the fourth chromosome of *D. virilis* and on the negative strand of the second chromosome (long arm) of *D. melanogaster*. This result suggests that there was a possible translocation or transposition between chromosome 2 and chromosome 4 after *D. virilis* and *D. melanogaster* diverged from each other. Further, the first ~10 Kb of the *D. virilis* fosmid matched to the minus strand of *D. melanogaster*, whereas the next ~14 Kb of the *D. virilis* fosmid matched to the plus strand of *D. melanogaster* chromosome 4. This pattern could possibly have been caused by an inversion of part of chromosome 4 in either *D. virilis* or *D. melanogaster*. Finally, in terms of spacing, it does not seem that similar regions of sequence have remained equidistant between the two species. This could be due to expansions of repeats between the two species: quantifying the amount of repetitive sequence in syntenic regions of *D. virilis* and *D. melanogaster* might help us better understand this result. Results are summarized in Figure 10 and Table 4.

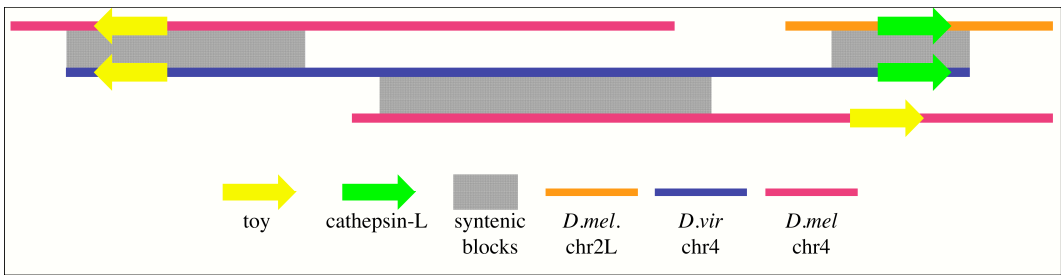


Figure 10: Synteny map of 99M21.

D. virilis region	percent identity	match to *D. melanogaster*	strand
1787-1979	88	chr4:1012505-1012310	minus
5292-5475	83	chr4:1010757-1010574	minus
5507-5559	90	chr4:1010554-1010502	minus
9400-9452	94	chr4:1008932-1008880	minus
13498-13668	86	chr4:1001914-1002065	plus
15381-15407	96	chr4:1002965-1002991	plus
15629-15684	92	chr4:1003356-1003410	plus
15782-15874	93	chr4:1003464-1003556	plus
18223-18318	94	chr4:1004998-1005093	plus
18478-18507	93	chr4:1005246-1005275	plus
24216-24407	92	chr4:1006567-1006760	plus
35521-35607	82	chr2L:10355766-10355680	minus

Table 4: Synteny summary table.

ClustalW Genic Analysis

Method

To consider conservation of one of my genes over time, I did a ClustalW analysis of the amino acid sequence for the predicted peptide and its orthologs in other species. ClustalW is a global alignment tool that finds regions of maximum conservation. Here, I considered feature 3 (the putative cathepsin-L). I chose to analyze feature 3, because unlike *toy* (feature 1), cathepsin is a less critical protein and thus I expected to find more divergence. I thought this would make this analysis more interesting, and further, I hoped that this ClustalW analysis might help me understand if the missing exon 1 of my putative cathepsin-L is crucial. By using the predicted peptide sequence for feature 3 (as found in Appendix D) and blastp, I searched for orthologs of this feature in the refseq database. This search identified numerous cathepsin-L genes in other species, such as mouse, rat, cow, pig, dog, human, flesh fly, and *D. melanogaster*. Through the NCBI database, I was able to find the amino acid sequences for this protein for these species. I then used ClustalW to align these sequences.

Results

The alignment showed more similarity across species than I expected. Much of this similarity was associated with a conserved protease domain, as shown highlighted in blue in Figure 11. This domain, as outlined by blastp, is peptidase C1A and is common to all cytosine proteases. Thus, the high conservation seen in this domain is not surprising. It is interesting to note, however, that this conservation is largely seen as conservation of amino acid similarity (shown by a semi-colon under given residue) rather than identity (shown by an asterisk under given residue). Further, in the species that do have an exon 1, there is substantially less conservation than in the rest of the protein (Fig. 12). I hypothesize that the lower level of conservation in exon 1 is because the N-terminus of cathepsins is the propetide for the protein. This propeptide is cleaved to activate the enzyme, and thus, it does not have enzymatic activity. I suspect this might explain some of my difficulty in locating exon 1 of the protein in *D. virilis* (as discussed under Feature 3).

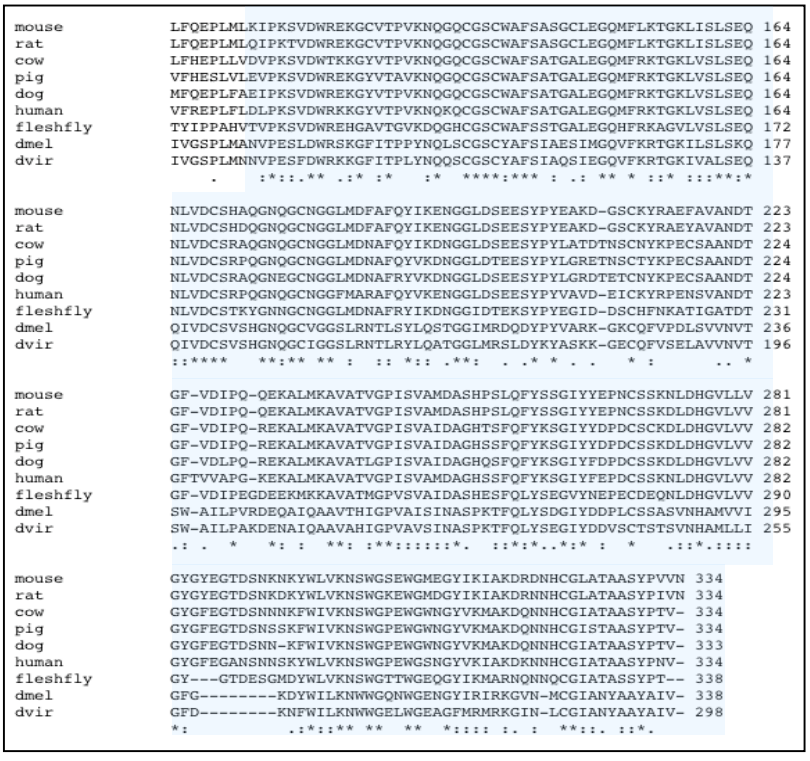
ClustalW Analysis of UTR and Promoter Regions

Method

To define functional conserved non-genic (CNGs) elements, I used ClustalW to align nucleotide sequences between both closely related species and more distant species. Here, I considered the upstream region of *toy* because it is a highly conserved gene across *Drosophila* and it has a well-characterized 5' untranslated region. These factors made it more likely that I would be able to identify any functional CNGs. To do this, I used Blat to identify the location of exon 1 of *toy* in *D. simulans*, *D. sechellia*, *D. melanogaster*, *D. erecta*, *D. yakuba*, *D. ananassae*, *D. virilis*, *D. grimshawi*, and *D. mojavensis*. We can loosely group the first six species as being distant relatives of *D. virilis* and the latter two species as close relatives of *D. virilis*. I then extracted the 1000 base pairs of DNA sequence that precede exon 1 in each of these species, reverse-complementing sequence as necessary. Because the UTR is nearly 500 base pairs long, extracting 1 Kb ensured that I would also be able to find any CNGs upstream of the UTR.

Results

There is very little conservation in the UTR in region. However, the ClustalW alignment shows a putative promoter region, which occurs 40 base pairs from the start of the UTR as defined in *D. melanogaster* (outlined in red in Fig. 13). This promoter region is about 30 base pairs long, which is fairly typical for *Drosophila* promoters. Also, comparing the promoter sequence with its reverse-complement shows that the sequence is palindromic (highlighted in green in Fig. 14). Palindromic sequence is commonly found in protein-binding DNA regions, thus providing additional evidence suggesting that this region is indeed a promoter. One would need to splice this element to a reporter gene for an *in vivo* test to confirm this prediction.



```
mouse          LFQEPLMLKIPKSVDWREKGCVTPVKNQGQCGSCWAFSASGCLEGQMFLKTGKLISLSEQ 164
rat            LFQEPLMLQIPKTVDWREKGCVTPVKNQGQCGSCWAFSASGCLEGQMFLKTGKLISLSEQ 164
cow            LFHEPLLVDVPKSVDWTKKGYVTPVKNQGQCGSCWAFSATGALEGQMFRKTGKLVSLSEQ 164
pig            VFHESLVLEVPKSVDWREKGYVTAVKNQGQCGSCWAFSATGALEGQMFRKTGKLVSLSEQ 164
dog            MFQEPLFAEIPKSVDWREKGYVTPVKNQGQCGSCWAFSATGALEGQMFRKTGKLVSLSEQ 164
human          VFREPLFLDLPKSVDWRKKGYVTPVKNQKQCGSCWAFSATGALEGQMFRKTGKLVSLSEQ 164
fleshfly       TYIPPAHVTVPKSVDWREHGAVTGVKDQGHCGSCWAFSSTGALEGQHFRKAGVLVSLSEQ 172
dmel           IVGSPLMANVPESLDWRSKGFITPPYNQLSCGSCYAFSIAESIMGQVFKRTGKILSLSKQ 177
dvir           IVGSPLMNNVPESFDWRKKGFITPLYNQQSCGSCYAFSIAQSIEGQVFKRTGKIVALSEQ 137
                   .   :*::.** .:* :*   :*  ****:*** : .: ** * ::* :::**:*

mouse          NLVDCSHAQGNQGCNGGLMDFAFQYIKENGGLDSEESYPYEAKD-GSCKYRAEFAVANDT 223
rat            NLVDCSHDQGNQGCNGGLMDFAFQYIKENGGLDSEESYPYEAKD-GSCKYRAEYAVANDT 223
cow            NLVDCSRAQGNQGCNGGLMDNAFQYIKDNGGLDSEESYPYLATDTNSCNYKPECSAANDT 224
pig            NLVDCSRPQGNQGCNGGLMDNAFQYVKDNGGLDTEESYPYLGRETNSCTYKPECSAANDT 224
dog            NLVDCSRAQGNEGCNGGLMDNAFRYVKDNGGLDSEESYPYLGRDTETCNYKPECSAANDT 224
human          NLVDCSRPQGNQGCNGGFMARAFQYVKENGGLDSEESYPYVAVD-EICKYRPENSVANDT 223
fleshfly       NLVDCSTKYGNNGCNGGLMDNAFRYIKDNGGIDTEKSYPYEGID-DSCHFNKATIGATDT 231
dmel           QIVDCSVSHGNQGCVGGSLRNTLSYLQSTGGIMRDQDYPYVARK-GKCQFVPDLSVVNVT 236
dvir           QIVDCSVSHGNQGCIGGSLRNTLRYLQATGGLMRSLDYKYASKK-GECQFVSELAVVNVT 196
               ::****   **:** ** :  :: *:: .**:  . .* * . .   * :      .. *

mouse          GF-VDIPQ-QEKALMKAVATVGPISVAMDASHPSLQFYSSGIYYEPNCSSKNLDHGVLLV 281
rat            GF-VDIPQ-QEKALMKAVATVGPISVAMDASHPSLQFYSSGIYYEPNCSSKDLDHGVLVV 281
cow            GF-VDIPQ-REKALMKAVATVGPISVAIDAGHTSFQFYKSGIYYDPDCSCKDLDHGVLVV 282
pig            GF-VDIPQ-REKALMKAVATVGPISVAIDAGHSSFQFYKSGIYYDPDCSSKDLDHGVLVV 282
dog            GF-VDLPQ-REKALMKAVATLGPISVAIDAGHQSFQFYKSGIYFDPDCSSKDLDHGVLVV 282
human          GFTVVAPG-KEKALMKAVATVGPISVAMDAGHSSFQFYKSGIYFEPDCSSKNLDHGVLVV 282
fleshfly       GF-VDIPEGDEEKMKKAVATMGPVSVAIDASHESFQLYSEGVYNEPECDEQNLDHGVLVV 290
dmel           SW-AILPVRDEQAIQAAVTHIGPVAISINASPKTFQLYSDGIYDDPLCSSASVNHAMVVI 295
dvir           SW-AILPAKDENAIQAAVAHIGPVAVSINASPKTFQLYSEGIYDDVSCTSTSVNHAMLLI 255
               .: .  *   *: :  **: :**:::::::*.  ::*:*..*:* :  *   .::*.:::::

mouse          GYGYEGTDSNKNKYWLVKNSWGSEWGMEGYIKIAKDRDNHCGLATAASYPVVN 334
rat            GYGYEGTDSNKDKYWLVKNSWGKEWGMDGYIKIAKDRNNHCGLATAASYPIVN 334
cow            GYGFEGTDSNNNKFWIVKNSWGPEWGWNGYVKMAKDQNNHCGIATAASYPTV- 334
pig            GYGFEGTDSNSSKFWIVKNSWGPEWGWNGYVKMAKDQNNHCGISTAASYPTV- 334
dog            GYGFEGTDSNN-KFWIVKNSWGPEWGWNGYVKMAKDQNNHCGIATAASYPTV- 333
human          GYGFEGANSNNSKYWLVKNSWGPEWGSNGYVKIAKDKNNHCGIATAASYPNV- 334
fleshfly       GY---GTDESGMDYWLVKNSWGTTWGEQGYIKMARNQNNQCGIATASSYPT-- 338
dmel           GFG--------KDYWILKNWWGQNWGENGYIRIRKGVN-MCGIANYAAYAIV- 338
dvir           GFD--------KNFWILKNWWGELWGEAGFMRMRKGIN-LCGIANYAAYAIV- 298
               *:          .:*::** **  **  *:::: :. :  **::. ::*.
```

Figure 11: ClustalW analysis of putative cathepsin-L.

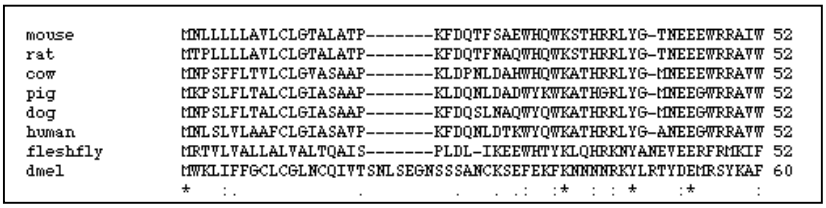


Figure 12: Exon 1 - ClustalW analysis of cathepsin-L.

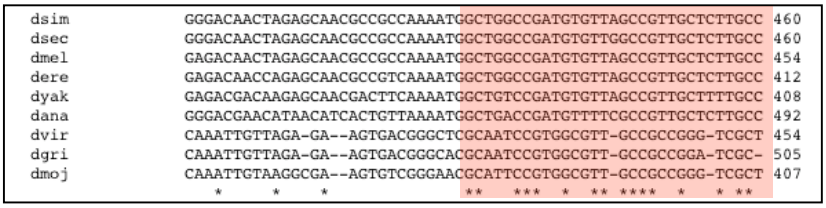


Figure 13: Putative promoter of *toy* gene.

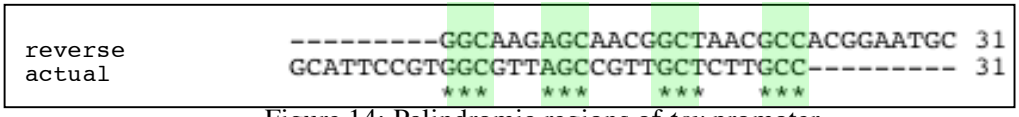


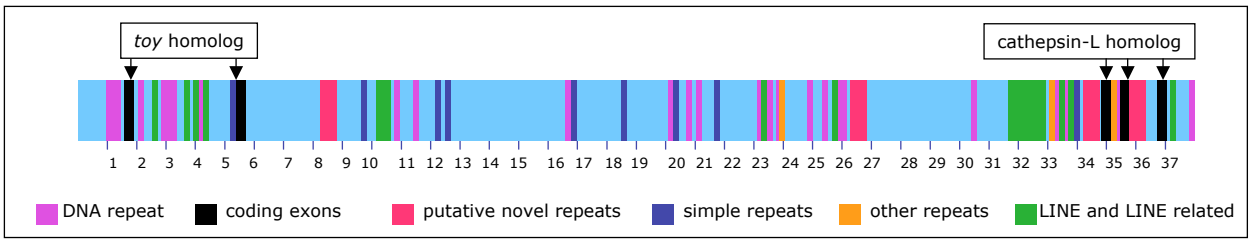
Figure 14: Palindromic regions of *toy* promoter.

Conclusion

Annotation is a difficult process, but, as we have more sequence information from the Drosophila family, the process will be facilitated. As can be seen in the example of Feature 3, having comparative resources can better allow one to consider a problem—even if the solution does not become immediately more obvious. Increased sequence information can also help us consider processes of chromosomal evolution as outlined in our discussion of synteny, and it can also help us find important CNGs as discussed in our ClustalW analysis of the upstream region of *toy*.

Additionally, this project shows the importance of the increased synergy between numerous genomic browsers as we move towards the genomic era. No browser or database—whether it is NCBI, Ensembl, FlyBase, or the UCSC genome browser—offers all the tools necessary to complete a successful annotation project. Thus, moving seamlessly between browsers is central to annotating effectively. While the browsers are somewhat compatible, this can be improved. Doing so will make annotation of future genomes less costly and more efficient.

Final Fosmid



References

"Cathpesin L: CTSL." 2006. Online Mendelian Inheritance in Man. 30 April 2006. <http://ncbi.nlm.nih.gov/entrez/dispmim.cgi?id=116880>.

Gehring WJ and K Ikeo. 1999. Pax 6: Mastering eye morphogenesis and eye evolution. Trends in Genetics 15: 371-344.

Appendix A

```
Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------

 1.02 Intr -   1973   1793  181  1  1   76   94   106 0.909   8.95
 1.01 Init -   5559   5297  263  0  2   58   94   177 0.391  12.00
 1.00 Prom -   8444   8405   40                             -5.75

 2.04 PlyA -   8629   8624    6                              1.05
 2.03 Term -  13545  13363  183  0  0   21   44   140 0.382  -0.54
 2.02 Intr -  15650  15406  245  0  2    4   40   408 0.245  23.89
 2.01 Init -  25423  25342   82  1  1   41   97    72 0.427   4.58
 2.00 Prom -  28700  28661   40                              0.05

 3.00 Prom +  29259  29298   40                             -8.85
 3.01 Init +  31113  31606  494  2  2   47   38   367 0.716  22.27
 3.02 Intr +  34802  34880   79  2  1   69   86    42 0.004   0.73
 3.03 Intr +  34962  35114  153  2  0    8   31   205 0.923   6.15
 3.04 Intr +  35418  35804  387  2  0   25   85   228 0.863  10.26
 3.05 Intr +  36814  36974  161  0  2  114   82    73 0.877   7.16

Predicted peptide sequence(s):

Predicted coding sequence(s):

>fosmid12|GENSCAN_predicted_peptide_1|148_aa
MMLTTEHIMHGHPHSSVGVGVGQSALFGCSTAGHSGINQLGGVYVNGRPLPDSTRQKIVE
LAHSGARPCDISRILQVSNGCVSKILGRYYETGSIKPRAIGGSKPRVATTPVVQKIADYK
RECPSIFAWEIRDRLLSEQVCNSDNIPS

>fosmid12|GENSCAN_predicted_CDS_1|444_bp
atgatgctaacaacggaacacattatgcatggacatccccattcgtccgtcggcgtgggg
gtgggccaaagtgcactgttcggctgctcgacagcgggacacagcggaattaaccagctg
ggaggcgtctatgtgaacggcagacccctgcccgattccacacgccagaagattgtcgag
ctggcacattccggcgctcggccatgtgatatttctcgaatacttcaagtttccaatggc
tgcgtaagcaaaatcttgggcagatattatgaaactggatctatcaaaccccgagcgata
ggtggttcaaagccgcgagtggctacgacgcccgtcgttcaaaaaattgccgattacaaa
agagaatgtccgagcatatttgcgtgggaaattcgtgatcggctgctatcggagcaagta
tgcaatagtgataatattccaagc

>fosmid12|GENSCAN_predicted_peptide_2|169_aa
MRKPDEDPTAAQEIHLSGNKGNLHLYVTVCQRHVAERLFFLVRWQQQAVNSSCEWAAERR
PSSDDRRPTTVDRLCAGDVKAQAQEEQPHHQPAEQRHRATPSEKSEANELFKIIFPYPAS
NTQRVRSDLLHQACQDLAATTPPPPHQLWTTVHVHHYHRSEQLTYYARA
```

>fosmid12|GENSCAN_predicted_CDS_2|510_bp
atgcgcaaacctgatgaggacccaactgcagcacaagaaattcatttatcagggaacaaa
ggtaatttgcatttgtatgtaacagtttgtcagcgacacgtggcggaacgactgttcttt
ctagttcgttggcaacagcaggcggttaattctagctgcgagtgggctgcggagcgtcga
ccatcgagcgacgaccgtcgaccaacgaccgtcgaccgactgtgtgccggtgacgtaaag
gctcaggctcaggaggaacagcctcatcatcagccggcagagcaacggcacagggcaacc
ccctcagagaagagtgaagccaacgagttatttaaaataatatttccgtatccggcaagt
aatacgcaacgcgtgcggagcgacctacttcaccaggcctgccaagacctggcagctaca
acccctccgccgccccaccaactgtggacgacggtccatgtgcatcattatcatcgctcg
gaacagctcacttactacgcgagagcctga

>fosmid12|GENSCAN_predicted_peptide_3|425_aa
MKESCDGFAKSSRMARSPLGVPTAAPIETNSGREPKSANYSRIRKRINITRCFRCLEFGH
LARHCKSGLDRSNLCRRCGGKDHLAKDCKQEPQCMLCKERNTDCKHIAGSGRCLCLGAPW
PREIEIDTAEPQPLRGCSGPTGVKYGRRNNLRAIQAYRWKWLGGNLQNGNMVICHKNESG
QMYKERFEIFKKINNRSYARSHDEMRSYEAYEENQIIVNEHNTYYETGKSSFRLATNTMA
DMNTDSYLKGYLRLLRSPEISDSDNIADIVGSPLMNNVPESFDWRKKGFITPLYNQQSCG
SCYAFSIAQSIEGQVFKRTGKIVALSEQQIVDCSVSHGNQGCIGGSLRNTLRYLQATGGL
MRSLDYKYASKKGECQFVSELAVVNVTSWAILPAKDENAIQAAVAHIGPVAVSINASPKT
FQLYX

>fosmid12|GENSCAN_predicted_CDS_3|1275_bp
atgaaggagagctgtgatggttttgcaaagagcagcagaatggcgagatcaccactaggt
gttccaacagccgccccgatagagacgaactcaggacgggagccaaagtcggcaaactac
tcgagaatccgcaaacggattaacatcaccaggtgtttcagatgcctcgagtttggacat
ctagcaagacactgtaaaagtggtctagacagatcgaatctgtgccgacgatgcggagga
aaagaccatctggctaaggactgcaaacaagagccacagtgcatgctctgtaaggaaagg
aacaccgactgcaaacacattgccggcagcggcagatgcctgtgtttaggagcgccctgg
ccaagagaaatagaaatagacacagctgaacctcaaccactgcgaggctgctcaggccct
acgggagtgaaatacggacgtcgcaataatctgcgagccatacaagcctatcgctggaag
tggctgggtggcaacttgcaaaatggtaatatggttatatgtcataaaaatgaatcaggt
cagatgtataaggaacgatttgaaattttcaagaaaataaacaatagaagctatgcacgt
tcccatgatgaaatgcgcagctatgaagcctatgaggaaaatcaaataattgtcaacgaa
cataatacgtattacgaaactggaaaaagcagctttcgattagcaacaaacacaatggct
gacatgaataccgattcatacctcaagggatatttacgtttattacggagtccagagatt
tctgattcggacaatattgccgacattgttggatcaccgctgatgaataatgttcctgaa
agttttgattggcgtaaaaagggatttattacaccactgtataatcaacaaagctgcggc
tcctgctatgccttcagtatagctcaaagtatagaagggcaggtgttcaagcgcaccggt
aagattgtggccctaagtgaacaacaaattgtggactgtagtgtctcccatggcaatcaa
ggctgtatcgggggctcactacgaaacactctaagatatctacaggctaccgggggtcta
atgagatcccttgattacaaatatgcctcaaagaaaggagaatgccaattcgttagcgaa
ctcgctgtagtcaatgtgacatcgtgggccattttaccggcaaaggatgaaaacgcaatt
caagcagctgtggcacatattggtccagttgcagtctccattaacgcaagtcccaaaact
tttcaactttatagn

Appendix B

	number of elements	length occupied	percentage of sequence
SINEs:	0	0 bp	0.00%
ALUs	0	0 bp	0.00%
MIRs	0	0 bp	0.00%
LINEs:	9	3328 bp	8.96%
LINE1	0	0 bp	0.00%
LINE2	0	0 bp	0.00%
L3/CR1	0	0 bp	0.00%
LTR elements:	1	56 bp	0.15%
MaLRs	0	0 bp	0.00%
ERVL	0	0 bp	0.00%
ERV_class I	0	0 bp	0.00%
ERV_class II	0	0 bp	0.00%
DNA elements:	20	3186 bp	8.58%
MER1_type	0	0 bp	0.00%
MER2_type	0	0 bp	0.00%
Unclassified:	1	142 bp	0.38%
Total interspersed repeats		6712 bp	18.07%
Small RNA:	0	0 bp	0.00%
Satellites:	0	0 bp	0.00%
Simple repeats	22	1410 bp	3.80%
Low complexity	15	443 bp	1.19%

Appendix C: All Repeats and Including Found Repeats

repeat number	start	end	matching repeat	class/family
1	1158	1230	dvir.16.2.centroid	DNA
2	1197	1243	PENELOPE	LINE
3	1241	1557	dvir.16.2.centroid	DNA
4	1626	1705	dvir.16.2.centroid	DNA
5	2414	2493	dvir.16.2.centroid	DNA
6	2460	2642	PENELOPE	LINE
7	2642	3043	dvir.16.2.centroid	DNA
8	3049	3121	dvir.16.2.centroid	DNA
9	3116	3185	dvir.16.17.centroid	DNA
10	3407	3654	G4_DM	LINE/Jockey

Appendix D: Final Annotation Files

CDS:

```
>Dvir_CG11186-RA [gene=Dvir\CG11186-RA] [note=Putative ortholog of Drosophila
```

melanogaster gene CG11186-RA, partial CDS]
ATGATGCTAACAACGGAACACATTATGCATGGACATCCCCATTCGTCCGTCGGCGTGGGGGTGGGCCAAAGTGCACT
GTTCGGCTGCTCGACAGCGGGACACAGCGGAATTAACCAGCTGGGAGGCGTCTATGTGAACGGCAGACCCCTGCCCG
ATTCCACACGCCAGAAGATTGTCGAGCTGGCACATTCCGGCGCTCGGCCATGTGATATTTCTCGAATACTTCAAGTT
TCCAATGGCTGCGTAAGCAAAATCTTGGGCAGATATTATGAAACTGGATCTATCAAACCCCGAGCGATAGGTGGTTC
AAAGCCGCGAGTGGCTACGACGCCCGTCGTTCAAAAAATTGCCGATTACAAAAGAGAATGTCCGAGCATATTTGCGT
GGGAAATTCGTGATCGGCTGCTATCGGAGCAAGTATGCAATAGTGATAATATTCCAA

>Dvir_CG5367-RA [gene=Dvir\CG5367-RA] [note=Putative ortholog of Drosophila
melanogaster gene CG5367-RA, partial CDS]
ATGCGCAGCTATGAAGCCTATGAGGAAAATCAAATAATTGTCAACGAACATAATACGTATTACGAAACTGGAAAAAG
CAGCTTTCGATTAGCAACAAACACAATGGCTGACATGAATACCGATTCATACCTCAAGGGATATTTACGTTTATTAC
GGAGTCCAGAGATTTCTGATTCGGACAATATTGCCGACATTGTTGGATCACCGCTGATGAATAATGTTCCTGAAAGT
TTTGATTGGCGTAAAAAGGGATTTATTACACCACTGTATAATCAACAAAGCTGCGGCTCCTGCTATGCCTTCAGTAT
AGCTCAAAGTATAGAAGGGCAGGTGTTCAAGCGCACCGGTAAGATTGTGGCCCTAAGTGAACAACAAATTGTGGACT
GTAGTGTCTCCCATGGCAATCAAGGCTGTATCGGGGGCTCACTACGAAACACTCTAAGATATCTACAGGCTACCGGG
GGTCTAATGAGATCCCTTGATTACAAATATGCCTCAAAGAAAGGAGAATGCCAATTCGTTAGCGAACTCGCTGTAGT
CAATGTGACATCGTGGGCCATTTTACCGGCAAAGGATGAAAACGCAATTCAAGCAGCTGTGGCACATATTGGTCCAG
TTGCAGTCTCCATTAACGCAAGTCCCAAAACTTTTCAACTTTATAG

Translation:

>Dvir_CG11186-RA [gene=Dvir\CG11186-RA] [protein=Dvir_CG11186-RA]
[note=Putative ortholog of Drosophila melanogaster gene CG11186-RA]
MMLTTEHIMHGHPHSSVGVGVGQSALFGCSTAGHSGINQLGGVYVNGRPLPDSTRQKIVELAHSGARPCDISRILQV
SNGCVSKILGRYYETGSIKPRAIGGSKPRVATTPVVQKIADYKRECPSIFAWEIRDRLLSEQVCNSDNIPS

>Dvir_CG5367-RA [gene=Dvir\CG5367-RA] [protein=Dvir_CG5367-RA] [note=Putative
ortholog of Drosophila melanogaster gene CG5367-RA]
MRSYEAYEENQIIVNEHNTYYETGKSSFRLATNTMADMNTDSYLKGYLRLLRSPEISDSDNIADIVGSPLMNNVPES
FDWRKKGFITPLYNQQSCGSCYAFSIAQSIEGQVFKRTGKIVALSEQQIVDCSVSHGNQGCIGGSLRNTLRYLQAT
GGLMRSLDYKYASKKGECQFVSELAVVNVTSWAILPAKDENAIQAAVAHIGPVAVSINASPKTFQLY