Annotating *D. virilis* Fosmid 37A19

David Desruisseau
May 4, 2006

BIO434W Genomics

Summary

Drosophila virilis fosmid 37A19 (also Fosmid 14) contains six GenScan gene predictions (Figure 1.1). The matches to *Drosophila melanogaster* genes in the same region were very good, indicating high synteny in the region of this fosmid, though with some examples of gene inversion and also intron insertion through duplication events. The genes found on either end of the fosmid run off the end and would therefore require more sequence to fully annotate. A Blast search of the repeats in the fosmid showed no novel repeats. A near-scale final map of the fosmid is given below, where gene directionality is given by the series of arrows above each gene (Figure 1.2).

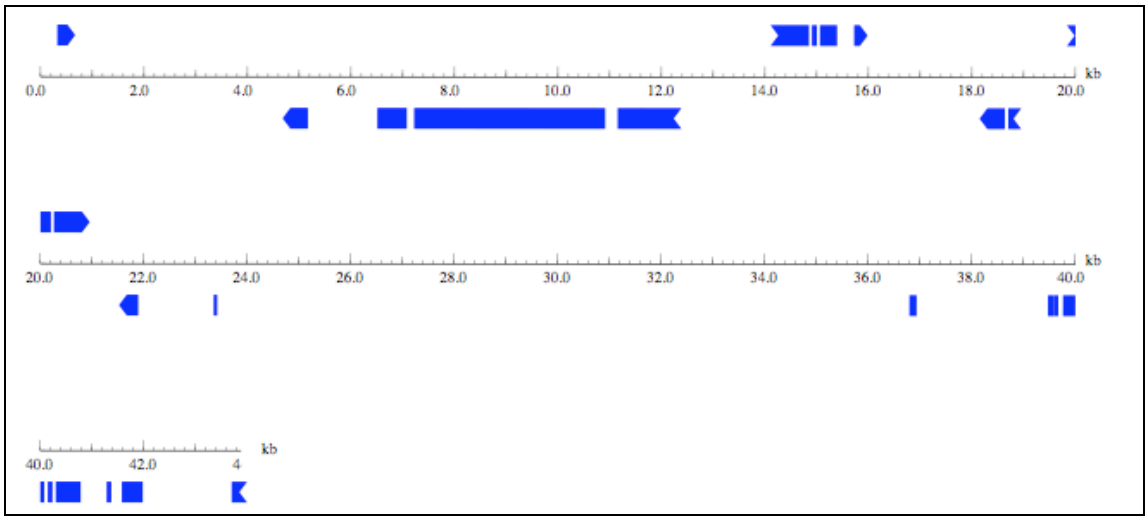


Figure 1.1: GenScan output for Fosmid 14 (see Appendix for full version)

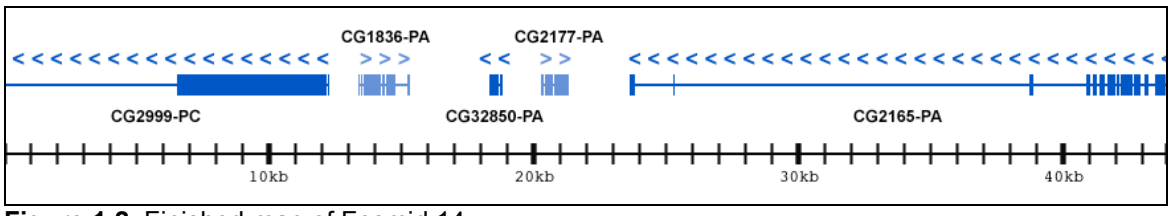


Figure 1.2: Finished map of Fosmid 14

Feature 14.3

I began my annotation work with the Washington University version of the UCSC Genome Browser (WU UCSC Browser herein), in which I first viewed the April 2006 assembly of my entire 44 kb fosmid. This initial view gave me a general impression of the genetic features I would be annotating, indicated by the six GenScan gene predictions in my fosmid (GenScan predictions in brown, Figure 2).

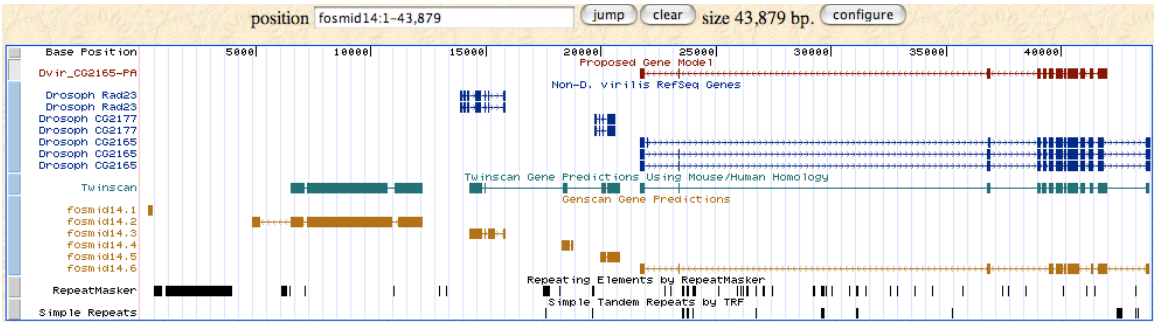


Figure 2

I chose to being my annotation with feature 14.3, so I obtained its translated protein from predicted mRNA by clicking the fosmid14.3 link in the initial WU UCSC Browser view. The predicted amino acid sequence is shown below:

```
>fosmid14.3_prot
MLTRDISGTSSGSSNNTNTEAVSSQQARKQAKETTERSTQDEPLVESKPA
VQVKESSSSKKGAKTNKITSEAGEEVGSTGAGSPAPASTTGSTTDYSSID
LVGELANTSLQTRAESNLLMGEEFNRTVASMVEMGYPREQVERAMAASFN
NPERAVEYLINGIPQEENLFTPGDDEESSRASNIHQGAASDLPAESAADP
FEFLRSQPQFLQMRSLIYQNPHLLHAVLQQIGQTNPALLQLISENQDAFL
NMLNQPLEDEVATNAQRLGRTQSNSSRTENLTSSASQAATTEGQRSAAGS
ENQPISVALEGDGTVSAERNVPTESLATIRLTPQDQDAIERLKALGFPEA
LVLQAYFACEKDEELAANFLLSSSFDD
```

Using this peptide sequence, I ran a FlyBase *tblastn* BLAST search of the annotated proteins (AA) database for *Drosophila melanogaster*. This returned good matches to the two isoforms (A and B) of Rad23, both with very low expect-values, on the order of 10^{-88}. These results indicated that feature 14.3 was very likely an orthologue of *D. melanogaster* Rad23. Focusing on this gene, I located the Ensembl Fruitfly entries for both previously annotated Rad23 isoforms in *D. melanogaster*. In order to obtain the most extensive annotation results possible for *D. virilis*, I chose the to use the Rad23A isoform in my search since it contains more exonic sequence than the B isoform, as indicated by the greater amount of filled black region in Figure 3.

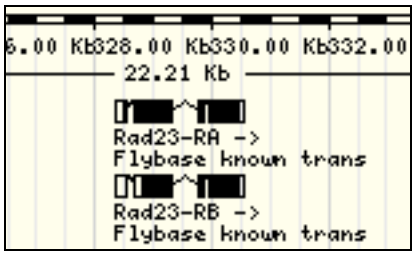


Figure 3

Through this non-specific Rad23 Ensembl entry, I then accessed the peptide information entry for Rad23A. This displayed annotation information on *D. melanogaster* Rad23A, providing the amino acid sequence for each of the gene's five exons, as well as a graphical representation of exon sizes (exons shown in purple, Figure 4). The other tracks (Pfam, Prosite profiles, and PRINTS) represent entries from other databases characterizing the listed sequence motifs.

The Pfam database is a database of protein sequence alignments provided by the Sanger Institute. Prosite is a database of protein families and domains provided by Expasy proteomics analysis servers identifying possible functions of newly-discovered proteins and analysis of known proteins for previously undetermined activity. The PRINTS database identifies protein families by functional attributes by means of a database of protein fingerprints. In this case, the matches to ubiquitin indicate the presence of this small regulatory protein that is 'ubiquitous' in eukaryotes, hence its name. UBA binds ubiquitin and is a commonly occurring sequence motif found in proteins involved in the ubiquitin/proteasome pathway. Of interest here is this pathway's involvement in excision repair by cell signaling with protein kinases. This explains the presence of these motifs in Rad23, a DNA excision-repair protein.

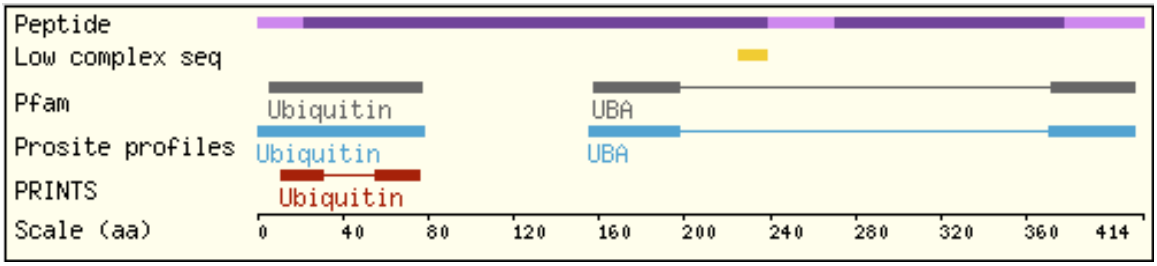


Figure 4

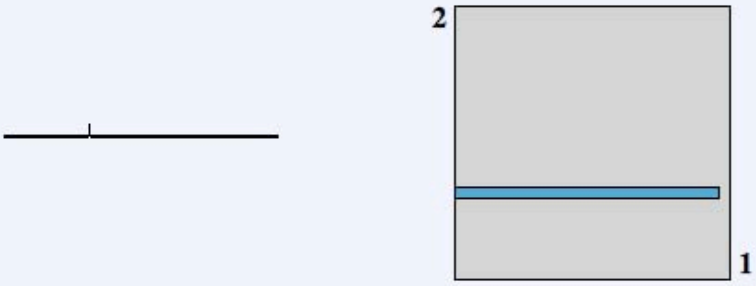
Using this amino acid sequence, I ran a Blast2 *tblastn* comparison between each exonic region and the entire unmasked Fosmid 14 sequence (fos14.fasta). Each search was performed with the filter off and the expect value set to 1,000 from the default of 10, in order to insure that all relevant results would be displayed. The Blast2 results for the first *D. melanogaster* exon resulted in a high-quality match to my fosmid from 13,888 – 13,953 bp (Figure 5). The next step would be to verify these bounds using the WU UCSC Browser to confirm exact base pair exon boundaries. The entire gene itself must be initiated by a Met (methionine) codon M and terminated by a stop codon *, easily distinguished by their coloration in the Browser. Aside from these gene end-specific exceptions, each individual exon must be preceded by the sequence AG (acceptor site) and followed by the sequence GT (donor site). The Browser helps in this regard, by providing the option of displaying predicted splice sites, in which it will locate these acceptor/donor sequences at likely locations.

BLAST 2 SEQUENCES RESULTS VERSION TBLASTN 2.2.13 [Nov-27-2005]

Matrix BLOSUM62 gap open: 11 gap extension: 1
x_dropoff: 50 expect: 1000.(wordsize: 3 Filter View option Standard
Masking character option X for protein, n for nucleotide Masking color option Black
Show CDS translation Align

Sequence 1: lcllseq_1
Length = 22 (1 .. 22)

Sequence 2: lcllfosmid14
Length = 43879 (1 .. 43879)



NOTE:Bitscore and expect value are calculated based on the size of the nr database.

```
 Score = 42.7 bits (99),  Expect = 0.019
 Identities = 19/22 (86%), Positives = 21/22 (95%), Gaps = 0/22 (0%)
 Frame = +1

Query  1      MIITIKNLQQQTFTIEFAPEKT  22
              MIIT+KNLQQQTFTI+F PEKT
Sbjct  13888  MIITVKNLQQQTFTIDFDPEKT  13953
```

Figure 5

The WU UCSC Browser confirmed these findings and displayed a Met codon flanking the 5' end and a GT intron donor site at the 3' end (Figures 6.1 and 6.2, respectively).

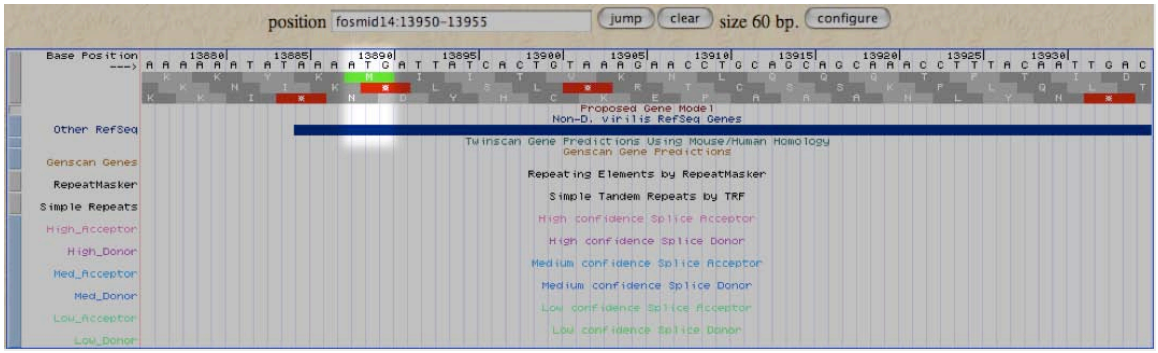


Figure 6.1

Figure 6.2

Results for the second *D. melanogaster* exon were slightly more complicated, since the Blast2 results were grouped into two highly matching regions, suggesting the insertion of an intron in the equivalent *D. virilis* region.

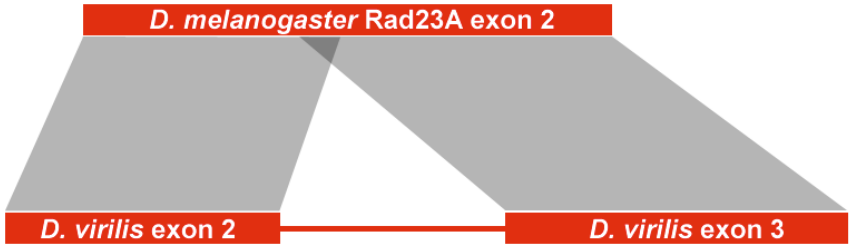


Figure 7.1: Initially-predicted duplication event

The *D. melanogaster* sequence matched the *D. virilis* sequence at 14,036 – 14,128 bp and 14,176 – 14,766 bp, for *D. virilis* exons 2 and 3 as illustrated in Figure 7.1. Checking for the appropriate GT/AG intron donor/acceptor sites confirmed the exact location of these matches, moving the initial matches slightly in order to not only match align with splice sites, but also to keep the exons in frame. However, in locating the precise exon ends and examining the resulting peptide sequence of the translated *D. virilis* product, I realized that there was no true duplication (Figure 7.2).

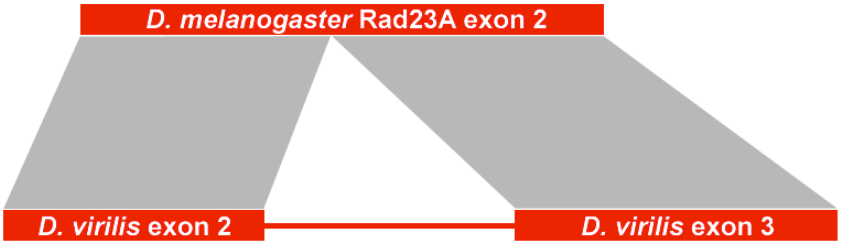


Figure 7.2: Actual exon reorganization consisting solely of intron insertion

Looking at the sequence matches below illustrates that the region previously believed to be duplicated (highlighted), matches with identity only at either end (red letters), meaning that there is no duplication. In this case, what I had originally thought to be a notable genetic anomaly turned out to be simply a misinterpretation of Blast2 results. This example underscores the importance of remaining skeptical of Blast2 matches so as to avoid jumping to conclusions regarding genetic importance.

Blast2 sequence alignment for *D. virilis* Exon 2:

```
Score = 46.2 bits (108),  Expect(2) = 5e-47
 Identities = 22/31 (70%), Positives = 25/31 (80%), Gaps = 0/31 (0%)
 Frame = +2

Query  1      VLELKKKIFEERGPEYVAEKQKLIYAGVILT  31
              VLELK++IF ERG EY  EKQKLIYAG  L+
Sbjct  14036  VLELKRQIFNERGAEYFVEKQKLIYAGTQLS  14128
```

Blast2 sequence alignment for *D. virilis* Exon 3:

```
Score =  167 bits (424),  Expect(2) = 5e-47
 Identities = 102/197 (51%), Positives = 125/197 (63%), Gaps = 36/197 (18%)
 Frame = +1

Query  23     LIYAGVILTDDRTVGSYNVDEKKFIVVMLTRD------SSSSNRNQLSV---------KE  67
              L + GVILTDDRT+ SY VDEKKFIVVMLTRD       SS+N N  +V         KE
Sbjct  14176  LFF*GVILTDDRTINSYKVDEKKFIVVMLTRDISGTSSGSSNNTNTEAVSSQQARKQAKE  14355

Query  68     SNKLTSTDD----SKQSMPCEEANHTNSPSSTNTEDSVLSRE-----------------T  106
              + + ++ D+    SK ++  +E++ +   + TN   S    E                  T
Sbjct  14356  TTERSTQDEPLVESKPAVQVKESSSSKKGAKTNKITSEAGEEVGSTGAGSPAPASTTGST  14535

Query  107    RPLSSDELIGELAQASLQSRAESNLLMGDEYNQTVLSMVEMGYPREQVERAMAASYNNPE  166
                 SS +L+GELA  SLQ+RAESNLLMG+E+N+TV SMVEMGYPREQVERAMAAS+NNPE
Sbjct  14536  TDYSSIDLVGELANTSLQTRAESNLLMGEEFNRTVASMVEMGYPREQVERAMAASFNNPE  14715

Query  167    RAVEYLINGIPAEEGTF  183
              RAVEYLINGIP EE  F
Sbjct  14716  RAVEYLINGIPQEENLF  14766
```

Following this methodology, I located the remainder of the exons for this feature, compensating for frame differences between exons where necessary in order to achieve a consistent reading frame in the final amino acid product. For example, in the case of locating the 3' end of *D. virilis* Rad23A exon 3, the exon is in reading frame 1 (the first row of gray peptide sequence in Browser view), but ends at 14,851 bp due to the presence of a GT donor site. This effectively leaves the base pair G as an incomplete codon at the end of this exon (Figure 8.1). However, moving to the 5' end of the next exon reveals that the AG acceptor site leaves two base pairs (AC) in the third frame, complementing the single base pair from exon 3 (Figure 8.2). These base pairs combine during transcriptional intron-removal to form a single codon, producing the desired peptide sequence for the gene.

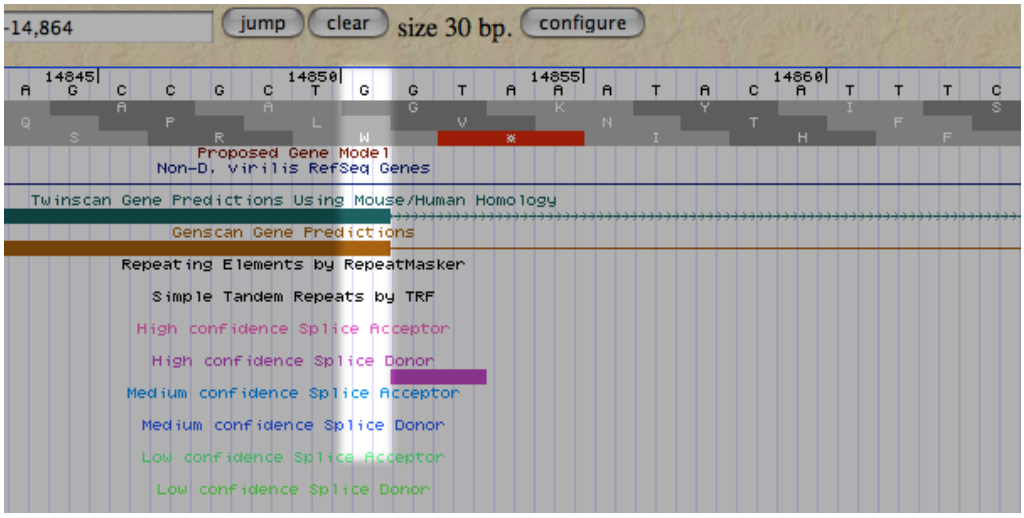


Figure 8.1

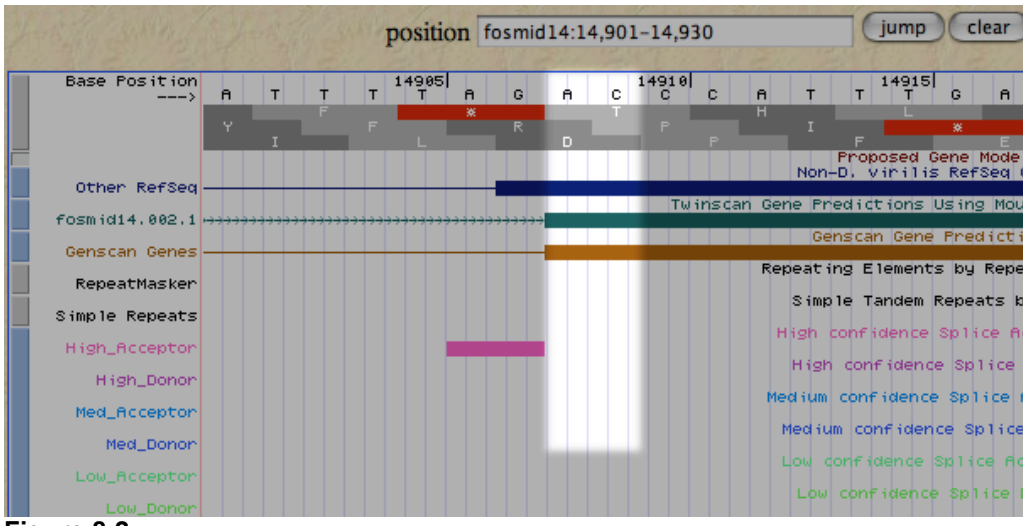


Figure 8.2

In order to confirm and finalize exon boundaries, I performed a predicted protein translation of the concatenated exonic sequences to identify the presence of any stop codons introduced by erroneously located exon splice sites. Translation of the nucleotide sequence using the ExPASy tools indicated several stop codons within my translated sequence in the desired directionality and frame (5'-3' Frame 1), indicating some exon boundaries were likely off by a small number of base pairs. This was easily remedied, by double-checking boundaries in the WU UCSC Browser, one at a time, and correcting any apparent errors. Simply rechecking my splice sites revealed several exon bounds that I had simply misreported as being one or two base pairs from the correct position. Noting these changes and translating the re-extracted exonic regions confirmed these changes were correct as my second ExPASy translation product contained no stop codons. This confirmation strongly supports my conclusion that *D. virilis* Feature 14.3 is an orthologueue of *D. melanogaster* Rad23A. The final peptide sequence of this annotated feature is provided below:

```
>D virilis - Putative orthologue of D melanogaster CG1836-PA
MIITVKNLQQQTFTIDFDPEKTVLELKRQIFNERGAEYFVEKQKLIYAGVI
LTDDRTINSYKVDEKKFIVVMLTRDISGTSSGSSNNTNTEAVSSQQARKQA
KETTERSTQDEPLVESKPAVQVKESSSSKKGAKTNKITSEAGEEVGSTGAG
SPAPASTTGSTTDYSSIDLVGELANTSLQTRAESNLLMGEEFNRTVASMVE
MGYPREQVERAMAASFNNPERAVEYLINGIPQEENLFTPGDDEESSRASNI
HQGAASDLPAESAADPFEFLRSQPQFLQMRSLIYQNPHLLHAVLQQIGQTN
PALLQLISENQDAFLNMLNQPLEDEVATNAQRLGRTQSNSSRTENLTSSAS
QAATTEGQRSAAGSENQPISVALEGDGTVSAERNVPTESLATIRLTPQDQD
AIERLKALGFPEALVLQAYFACEKDEELAANFLLSSSFDD
```

Feature 14.4

In annotating feature 14.4, I followed nearly the same procedure as I did for annotating feature 14.3. This is only a two-exon feature, so annotation was relatively straightforward. After obtaining its GenScan translated protein sequence from the WU UCSC Browser, I ran a FlyBASE *tblastn* BLAST search against the AA database for *D. melanogaster*, resulting in a single good match with an expect-value on the order of 10^{-53}, sufficiently small for good confidence. Using the transcript identification number CG32850 from the BLAST result, I searched the Ensembl site to obtain the protein entry for the gene.

I then performed a Blast2 search (Blast2, *tblastn*, filter off, expect 1000) with the first exon, which returned a good match. The Blast2 results also indicated that the match was in the *D. virilis* minus (–) strand. Initially, I did not know to reverse the reading frame directionality in the WU UCSC Browser, but exon identification was straightforward after clicking the arrow in the corner of the Browser view to reverse the reading frames (Figure 9).

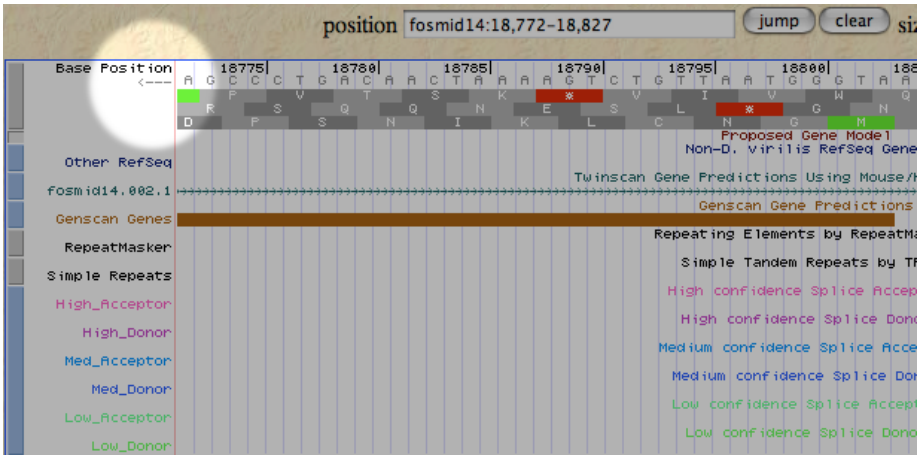


Figure 9

Again, after following the same methodology as for my first annotated feature, I located and verified the exon boundaries for the two exons of the gene, obtaining the final translated peptide sequence:

```
>D virilis - Putative orthologue of D melanogaster CG32850-PA
MGNCLKINSPDDISLLRGSESISGQDSGPMPIYQHEPLPQMFYPASSASHS
AASATNMSEEDQIKIAKRIGLVQHLPIGTYDGNLKKARECVICMVEFCVDE
AVRYLPCMHIYHVHCIDDWLMRSLTCPSCLEPVDAALLTSYETT
```

Feature 14.5

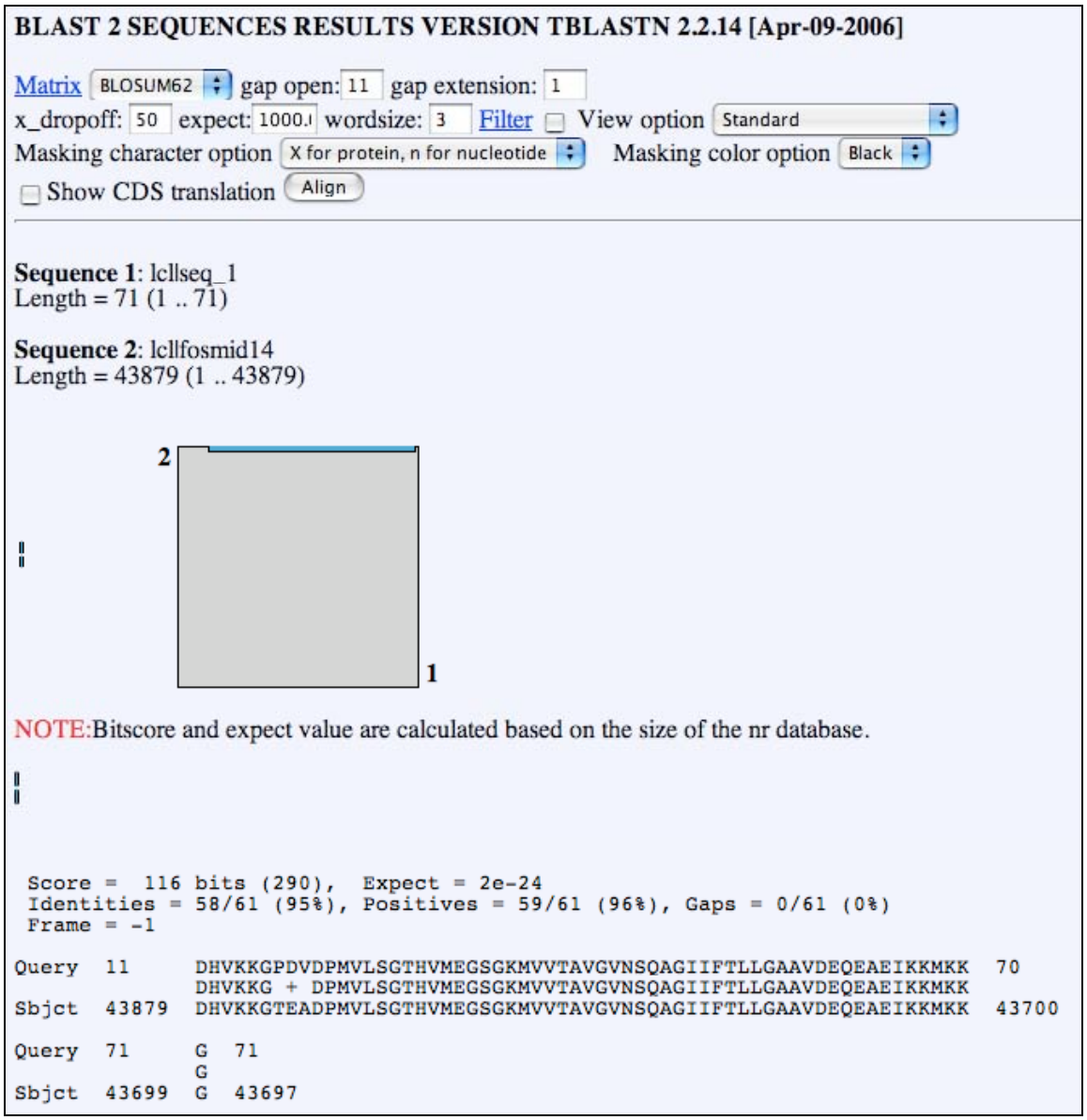
Annotating this feature presented nothing unordinary and proved to be straightforward after having successfully annotated the previous two features. As with the other features, I obtained the predicted peptide sequence, performed an identical BLAST search, and obtained high-quality matches to the A and B isoforms of CG2177. As the A-isoform contained more protein-coding region, I selected it for my Blast2 comparison. The Blast2 results (*tblastn*, filter off, expect 1000) provided exon boundaries that I quickly confirmed after referring back to the WU UCSC Browser. The final translated product is provided below.

```
>D virilis - Putative orthologue of D melanogaster CG2177-PA
MEETIILILLVIVMLVGSYLSGSIPLVMKLSEEKLKFVTVLGAGLLVGTAL
TVIIPEGIRSLYMDTQRPQTDANTSQLSAGLIKPHDYSRTIGLSLVLGFVF
MMLVDQVSQRKTNKGNENDKNITATLGLVVHAAADGVALGAAATTSHQDVE
IIVFLAIMLHKAPAAFGLVSFLLHEKVERQQIRKHLGIFSLSAPLLTLLTY
FGIGQEQRETLNSVNATGIAMLFSAGTFLYVATVHVLPELTQGGCSHPNRK
GGPGKYDYHAIEESREAATNDSSANGSAQSYNVQGLRYSELIIMICGALLP
LVITFGHQH
```

Feature 14.6

I began annotation of feature 14.6 just as I had the other features, obtaining the predicted peptide sequence to establish its counterpart in *D. melanogaster*. The feature was matched to CG2165-PA. Next, I performed Blast2 comparisons for each exon, using the same parameters as previously. Initially, I was concerned when I could not locate the first exons of the gene. There was no match in the Blast2 comparison between my fosmid and the first three exons of the *D.*

melanogaster gene. The fourth exon matched only partially to my fosmid sequence, though identity was high for the matching region (Figure 10).



BLAST 2 SEQUENCES RESULTS VERSION TBLASTN 2.2.14 [Apr-09-2006]

Matrix BLOSUM62 gap open: 11 gap extension: 1
x_dropoff: 50 expect: 1000.0 wordsize: 3 Filter View option Standard
Masking character option X for protein, n for nucleotide Masking color option Black
Show CDS translation Align

Sequence 1: lcllseq_1
Length = 71 (1 .. 71)

Sequence 2: lcllfosmid14
Length = 43879 (1 .. 43879)

2

1

NOTE:Bitscore and expect value are calculated based on the size of the nr database.

```
 Score =  116 bits (290),  Expect = 2e-24
 Identities = 58/61 (95%), Positives = 59/61 (96%), Gaps = 0/61 (0%)
 Frame = -1

Query  11     DHVKKGPDVDPMVLSGTHVMEGSGKMVVTAVGVNSQAGIIFTLLGAAVDEQEAEIKKMKK  70
              DHVKKG + DPMVLSGTHVMEGSGKMVVTAVGVNSQAGIIFTLLGAAVDEQEAEIKKMKK
Sbjct  43879  DHVKKGTEADPMVLSGTHVMEGSGKMVVTAVGVNSQAGIIFTLLGAAVDEQEAEIKKMKK  43700

Query  71     G  71
              G
Sbjct  43699  G  43697
```

Figure 10

Blast2 searches for *D. melanogaster* exons 5-16 produced very good results and convinced me that a gene orthologue was indeed present in my fosmid and that it simply ran off the right end of the fosmid. Matching to the minus strand, the first several exons could not be matched simply because there was no sequence available for that region. After discussing this with a BIO4342 TA, it was decided that I would annotate only complete exons, meaning that I would ignore the first four exons of this gene. The Blast2 matches for exons 5-16 were all of very high quality and made it easy to locate all boundaries using the WU UCSC Browser. The translated product for these particular exons (5-16) is provided on the following page.

```
>D virilis - Putative orthologue of D melanogaster CG2165-PA
ENDGRIPIKETSHVTQPRSASEAVKSESDGNHVQQSSTPAPETGHKKEKSV
LQAKLTKLAIQIGYAGSTIAVLTVFILIIQFCIKTFVIDEKPWKNTYANNL
VKHLIIGVTVLVVAVPEGLPLAVTLSLAYSVKKMMKDNNLVRHLDACETMG
NATAICSDKTGTLTTNRMTVVQSYICEKLCKPGPKPGDIPIQVGNKTECAL
LGFVQGLGVKYQSIRDEIPEDRFTRVYTFNSVRKSMGTVIPRPNGGYRLYT
KGASEIIMKKCAFIYGHEGTLEKFTRDMQERLIREVIEPMACDGLRTISVA
YRDFVPGKAAINEVHIDGEPNWDDEENIMSNLTCLCVVGIEDPVRPEVPDA
IRKCQRAGITVRMVTGDNINTARSIASKCGILRPNDDFLILEGKEFNRRIR
DSNGDIQQHLLDKVWPKLRVLARSSPTDKYTLVKGMIDSAVTDNREVVAVT
GDGTNDGPALKKADVGFAMGIAGTDVAKEASDIILTDDNFSSIVKAVMWGR
NVYDSIAKFLQFQLTVNVVAVIVAFIGACAVQDSPLKAVQMLWVNLIMDTL
ASLALATEVPTPDLLLRKPYGRTKPLISRTMMKNILGQALYQLVIIFGLLF
VGDLILDIESGRGQDLNAGPTQHFTIIFNTFVMMTLFNEINARKIHGQRNV
VIIQYGKMAFSTKALSLDQWLWCVFFGIGTLVWGQLITSVPTRKLPKILSW
GRGHPEEYTDAMNLGEERFDSIDSDKKPRAGQILWIRGLTRLQTQVIGGEL
QERLIPVPYSKSNTDQAIRVVNAFRQGLDARYGEHTNTSLAEVLRKQTSLS
KRLSETSSIDYADNIPDELTIPEIDVERLSSHSHTETAV
```

Features 14.1 and 14.2

Left with only two un-annotated features, I first focused my attention on GenScan prediction 14.1, because of its small size (Figure 11). Not surprisingly, the FlyBLAST search (*blastp*, AA, *D. melanogaster*) resulted in no viable matches; the only hits had extremely high e-values of 40. With this in mind, I moved on to feature 14.2, with the understanding that initial and final exons often have weak matches, and the only route in annotating them is to find neighboring exons in flanking fosmid sequence.

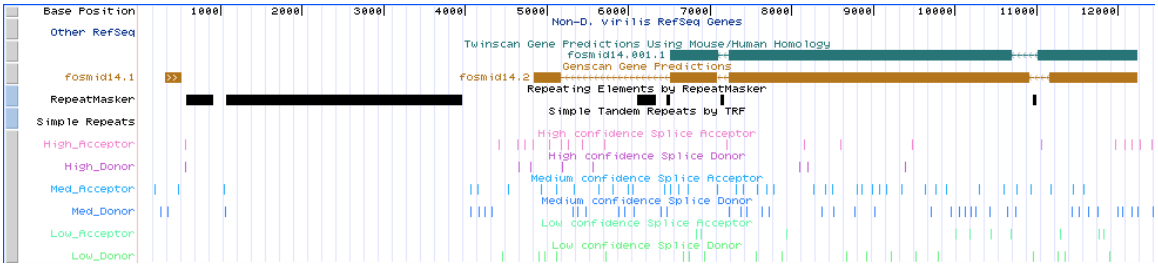


Figure 11

FlyBLAST searches (*blastp*, AA, *D. melanogaster*) produced very good matches to the A and C isoforms of the gene unc-13. In fact, the expect values were of the magnitude of 10^{-121} for both of these matches. Therefore, I was very confident that I had found an orthologue of the *D. melanogaster* gene in my fosmid. After accessing the Ensembl peptide entry for the C-isoform of this gene, I proceeded to begin matching these to my fosmid with Blast2 searches. However, the first exon of this gene was comprised of only two matching peptides, making it impossible for me to locate in a standard Blast2 search. Knowing this, I decided it would be more effective for me to simply locate the very large exon 2 first, and then look upstream of this region in the WU UCSC Browser to locate the first exon. This technique worked perfectly, as exon 2 matched extremely well, allowing me to use this as a landmark is searching for exon 1. Moving upstream from the second exon in the Browser led me to the MT amino acid sequence I was looking for, flanked on its 3' end by the desired intron donor site. This confirmed the location of exon 1 with reasonable confidence.

Moving forward, however, I could not locate exons 3-23 of the gene, so I suspected that the gene ran off the end of my fosmid, just as 14.6 had run off the other end. This seems the most logical conclusion to me, as this feature lies relatively close to the left end of my fosmid and runs in the negative direction. This is also a likely explanation for the presence of feature 14.1. GenScan likely erroneously predicted this gene due to the presence of gene-like or ORF sequence within the actual unc-13-PC gene present here. The translated peptide product for the two fully-present exons of the unc-13-PC gene is provided below.

```
>D virilis - Putative orthologue of D melanogaster CG2999-PC
MTHYARDECYHNNKNDALSTDSRPINNSTIYNYDEEHSGAYVYDEYGVSEE
YSSDVNWPRGDDHIFGDRSFNQNYAHYTDEYVPEEQNVAGLPTAEYWDENI
GAPIDYGGAGYGYGSTAPMKHNGCGSSTKYSDDVPIEGTYTRRGNRIGVCL
PTAPASSTTIVMDNSKYRMLPHAAAYQNVHAENNLNGHSGYASNAGLIDER
IGSFISLPAAPMRTLPEPQQRPSSSHRLFEDNNLKADAFGNQNVRDSVSAT
LGSARGPTSYDVFYQMARPYTSMLPLDYTDCSEFGGNSDNLSAFSDTPPVS
NAQHKLQQQRKISLMMAMTTASVIASGETRVPVQHSNRRIDNYSGSATTVV
PCCDENISICASSSSSVAFTNATTIALARWSTTDAQTTTTMTTTTTTTRKL
PKRLPSPAIQYKSSHTTGTSVPEVISEITFPSQAQFEKNHRAKQLPKLPLP
KAQANLNSVYIGSDISINRTASMVLSSGISVSSSPERLPAIPDSERMTTHM
APSSPMMSTTAVAFVDGANNFFRLEESQAFTKMCSRTQLQSNQSEVPRSVD
SFRSWPPTSIATESLSLYQSELSSDKAKLNFTVPAKLTQKTSTELAVEQES
SSAVPKSNHYHSAFTATETTVTIPALCESKTSSFDISDYLKPYTFDSLKFS
TENDHTMIPTVARTTFDTYVSCRDSLDSFKLTSLESQMPNSITSWTTSPIT
IESSLLSGDVAAVLTAPKEEMVTTPPSESSSTLSYLPNKPILPIPTTMEGK
KSDQFAATSDSHDAVTTLSHLSNVFTTNVMASSSSTIDIINTSPASVSTPA
SIIAHLPTLSYIDYMKKFELPELPPILDICSDETTAVFPETSSADNVTATT
DAGFRKFNTSLFAETETFKSTATMPSHTHPTESSMHCVHECNEFGVQSYQP
NQYQSEIPTCSPNLAPINTVSTHSVMNKYQDIMSTEAKLDSGISLQLEPQI
SETVPDYLSSHIDQTLVSTVSAFDDSFYDSFNVDLTALTATIAHIESETCL
TPALKPNSNDTLNYLPCDPSSNAYMDVRETGSSNELNESFGFRQGGYYKPS
QKQQQQPCVQDEAPHTPQKSKVVASAATSVLGGLSKGIKGGLDGVLSGVSS
TMEATKQYPANNNSKKSFGFGLASKLVPNVGGLLSSSNKQDQQQEAAPTIS
TVSVIISTADSAQHMLNCQSTSKGYEYVNENVIDTNTNTDVSVSHTPTTGY
LEELRSVQNNSSVLETKSNLAYDYLGNYEPDSSAIEVIGLNDDYRYGEYSA
PYQTTDEGHIDSLYSYTGPYKSIVGSEMELGSTMAVDTIPFTQYQSSMSSS
NLPAAEQTSMLETGIISVTETGTSKSQSVAPKVGIKKDRSSTVGSTGGMLG
SIFGKAAAAVQSATQAVNQGASSVASAVAQKASTPSAVVTTVGQTKPRDVP
AISSSNQSATPVSSVSFVSNIEQEPTQAANTISTSNASGAIGFHRVIDNDY
NELVTNPDSVSSQYSTAGDDYENSNTHMLGDDMSVHMNINDNKAYSIYYSG
GNQIQQIDTKDTPAVHSSMLGKVIAGQKLLPAVPSAVSTGKKLPTINGKSG
LLIKQQPTEIFDDESDDDLTEYFMGHPDKVDHEPNYCIDSEQDDYYMDPQQ
TTPSSRGANDYYEQVSAGYDYREDFFNEEDEYKYLEQQTQQQQHQYQHQHQ
SQKLLLSHKQSSLDYVDDEQNDDFLNETYQSDEDCGNYLDESSSGSVGIAE
GRKQQAEDNGFITGTTSIATMGMNTTPAIANTAERSSLRTSKAAGICLQLG
EGHQQIKKQDSIIIEEEESSPIDLNNDCRVAAHMSPEGDADDDDQLADLLP
TGGQPQKKKILMRGETEEVVSGHMQMIRKPEITAKQRWHWAYNKIIMQLN
```

Repeats

In order to locate any novel repeats in my fosmid, I ran a BLAST search on the goose.wustl.edu server through SSH using my unmasked fosmid sequence and searching against the sequence database of all fosmids, which had been pre-compiled on the server. In order to characterize any repeats, I viewed the resulting output file in the Herne browser, using the option to hide any already-characterized repeats. Viewing the output led me to only a single region of interest between 31.5 kb and 32 kb (Figure 12).

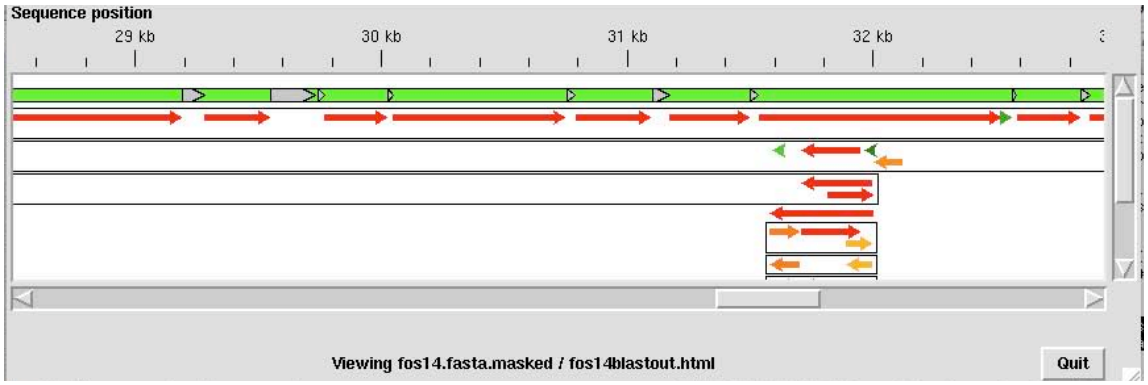


Figure 12

After establishing the exact ends to this region by clicking on each arrow to view a detailed entry for each result, I extracted the region (including 20 bps up- and downstream) and ran a Blast search with this sequence against the Drosophila repeats database that had been prepared on the WU Goose server. With the e-value set to 10^{-5}, no results were found. Even after repeating the Blast search with increased e-values, there were still no matches. This leads me to conclude that my fosmid contains no novel repeats, likely due at least in part to the fact that it contains so many transcribed genes.

Synteny with D. melanogaster

The genetic features comprising *D. virilis* Fosmid 14 are highly similar in relative organization and function to the *D. melanogaster* counterparts. As shown in the Fosmid 14 summary table (Appendix 2), all features correspond to chromosome 4. While there are examples of intron insertion (as described in depth with Feature 14.3), the relative arrangement of all genetic features shows high synteny between *D. virilis* and *D. melanogaster* in this region.

ClustalW analysis

For my ClustalW multiple-sequence analysis, I used the Rad23-PA gene homologue present in the fosmid, since I assumed this gene would be relatively highly conserved due to its radiation damage repair function. Using the concatenated exonic sequence from *D. virilis*, I ran a BLAST search on the NCBI site (*blastn*, *nr*). The results of this Blast search matched to several species; I chose to compare sequence from humans, zebrafish, mice, and of course *D. melanogaster* and *D. virilis*. After obtaining the RefSeq protein entries in the Genomes section of the NCBI Blast site, I used ClustalW to perform a multiple alignment between all these

sequences. The resulting alignment showed surprisingly little conservation overall, suggesting that this gene is not actually highly-conserved between these species. The results are shown below:

```
Hsapiens_NM_002874              MQVTLKTLQQQTFKIDIDPEETVKALKEKIESEKGKDAFPVAGQKLIYAG 50
Mmusculus_NP_033037             MQVTLKTLQQQTFKIDIDPEETVKALKEKIESEKGKDAFPVAGQKLIYAG 50
Dreriozebrafish_NP_956858       MQITLKTLQQQTFKIDIDAEETVKALKEKIENEKGKDGFPVAGQKLIYAG 50
Dmelanogaster_Rad23-PA          MIITIKNLQQQTFTIEFAPEKTVLELKKKIFEERGPE-YVAEKQKLIYAG 49
Dvir_CG1836-PA                  MIITVKNLQQQTFTIDFDPEKTVLELKRQIFNERGAE-YFVEKQKLIYAG 49
                                * :*:*.******.*:: .*:**  **.:* .*:* : : .  *******

Hsapiens_NM_002874              KILNDDTALKEYKIDEKNFVVVMVTKPKA-----------VSTPAPATTQ 89
Mmusculus_NP_033037             KILSDDTALKEYKIDEKNFVVVMVTKPKA-----------VTTAVPATTQ 89
Dreriozebrafish_NP_956858       KILSDDTALKEYKIDEKNFVVVMVTKPKS-----------ASAPAPPSSS 89
Dmelanogaster_Rad23-PA          VILTDDRTVGSYNVDEKKFIVVMLTRDSS---------------SSNRNQ 84
Dvir_CG1836-PA                  VILTDDRTINSYKVDEKKFIVVMLTRDISGTSSGSSNNTNTEAVSSQQAR 99
                                 **.** :: .*::***:*:***:*:  :                .

Hsapiens_NM_002874              QSAPASTTAVTSSTTTTVAQAPTPVPALAPTSTPASITPASATASSEPAP 139
Mmusculus_NP_033037             PSSTPSPTTVSSSPAVAAAQAPAPTPALAPTSTPASTTPASTTASSEPAP 139
Dreriozebrafish_NP_956858       SSSSSSSTTASAS--------------AAPSAAPVSESPS----EEEKKP 121
Dmelanogaster_Rad23-PA          LSVKESNKLTSTDD----SKQSMPCEEANHTNSPSSTNTEDSVLSRET-- 128
Dvir_CG1836-PA                  KQAKETTERSTQDEPLVESKPAVQVKESSSSKKGAKTNKITSEAGEEVGS 149
                                 .   :    : .                :    . .       *

Hsapiens_NM_002874              ASAAKQEKPAEKPAETPVATSPTATDSTSGDSSRSNLFEDATSALVTGQS 189
Mmusculus_NP_033037             AGATQPEKPAEKPAQTPVLTSPAPADSTPGDSSRSNLFEDATSALVTGQS 189
Dreriozebrafish_NP_956858       S----EEKPSSDPAP---ATTPVSSGSLPN----ANIFEEATSALVTGQS 160
Dmelanogaster_Rad23-PA          -----------RPLSSD---------ELIGELAQASLQSRAESNLLMGDE 158
Dvir_CG1836-PA                  T-----GAGSPAPASTTGSTTDYSSIDLVGELANTSLQTRAESNLLMGEE 194
                                            *             .  .    :.:   * * *: *:.

Hsapiens_NM_002874              YENMVTEIMSMGYEREQVIAALRASFNNPDRAVEYLLMGIPGDRESQAVV 239
Mmusculus_NP_033037             YENMVTEIMSMGYEREQVIAALRASFNNPDRAVEYLLMGIPGDRESQAVV 239
Dreriozebrafish_NP_956858       YENMVTEIMLMGYERDRVVAALRASFNNPDRAVEYLLTGIPAEGEG-SVV 209
Dmelanogaster_Rad23-PA          YNQTVLSMVEMGYPREQVERAMAASYNNPERAVEYLINGIPAEEGTFYNR 208
Dvir_CG1836-PA                  FNRTVASMVEMGYPREQVERAMAASFNNPERAVEYLINGIPQEENLFTPG 244
                                ::. * .:: *** *::*  *: **:***:******: *** :

Hsapiens_NM_002874              DPP-QAASTGAPQSSAVAAAAAT-------TTATTTTTSSGGHPLEFLRN 281
Mmusculus_NP_033037             DPPPQAVSTGTPQSPAVAAAAAT-------TTATTTTTS-GGHPLEFLRN 281
Dreriozebrafish_NP_956858       GAVDAVSPSGSTPASAPAPAISTGLSSPSSTAPAQPSSASGANPLEFLRN 259
Dmelanogaster_Rad23-PA          LN-ESTNPSLIPSGPQPAS-----------ATSAERSTESNSDPFEFLRS 246
Dvir_CG1836-PA                  DDEESSRASNIHQG----------------AASDLPAESAADPFEFLRS 277
                                       .:    .                 :.:  .:   ..*:****.

Hsapiens_NM_002874              QPQFQQMRQIIQQNPSLLPALLQQIGRENPQLLQQISQHQEHFIQMLNEP 331
Mmusculus_NP_033037             QPQFQQMRQIIQQNPSLLPALLQQIGRENPQLLQQISQHQEHFIQMLNEP 331
Dreriozebrafish_NP_956858       QPQFLQMRQIIQQNPSLLPALLQQIGRENPQLLQQISSHQEQFIQMLNEP 309
Dmelanogaster_Rad23-PA          QPQFLQMRSLIYQNPHLLHAVLQQIGQTNPALLQLISENQDAFLNMLNQP 296
Dvir_CG1836-PA                  QPQFLQMRSLIYQNPHLLHAVLQQIGQTNPALLQLISENQDAFLNMLNQP 327
                                **** ***.:* *** ** *:*****: ** *** **.:*: *::***:*

Hsapiens_NM_002874              VQEAGGQGGGGGGGSGG------------------------IAEAGSG-- 355
Mmusculus_NP_033037             VQEAGGQGGGGGGGGGGG-----------------GGGGGGIAEAGSG-- 362
Dreriozebrafish_NP_956858       VQEAG-QGGGAGG----------------------------VAEAGGG-- 328
Dmelanogaster_Rad23-PA          IDRESESGATVPPVSNARIPSTLDNVDLFSPDLEVATS-AQRSAAGT--- 342
Dvir_CG1836-PA                  LEDEVATNAQRLGRTQSNSSRTEN---LTSSASQAATTEGQRSAAGSENQ 374
                                ::     ..                                  : **

Hsapiens_NM_002874              ---------------------HMNYIQVTPQEKEAIERLKALGFPEGLVI 384
Mmusculus_NP_033037             ---------------------HMNYIQVTPQEKEAIERLKALGFPEGLVI 391
Dreriozebrafish_NP_956858       ---------------------HMNYIQVTPQEKEAIERLKALGFPEGLVI 357
Dmelanogaster_Rad23-PA          --SAAHQSGSAADNEDLEQPLGVSTIRLNRQDKDAIERLKALGFPEALVL 390
Dvir_CG1836-PA                  PISVALEGDGTVSAERNVPTESLATIRLTPQDQDAIERLKALGFPEALVL 424
                                                    :  *::. *:::************.**:

Hsapiens_NM_002874              QAYFACEKNENLAANFLLQQNFDED 409
Mmusculus_NP_033037             QAYFACEKNENLAANFLLQQNFDED 416
```

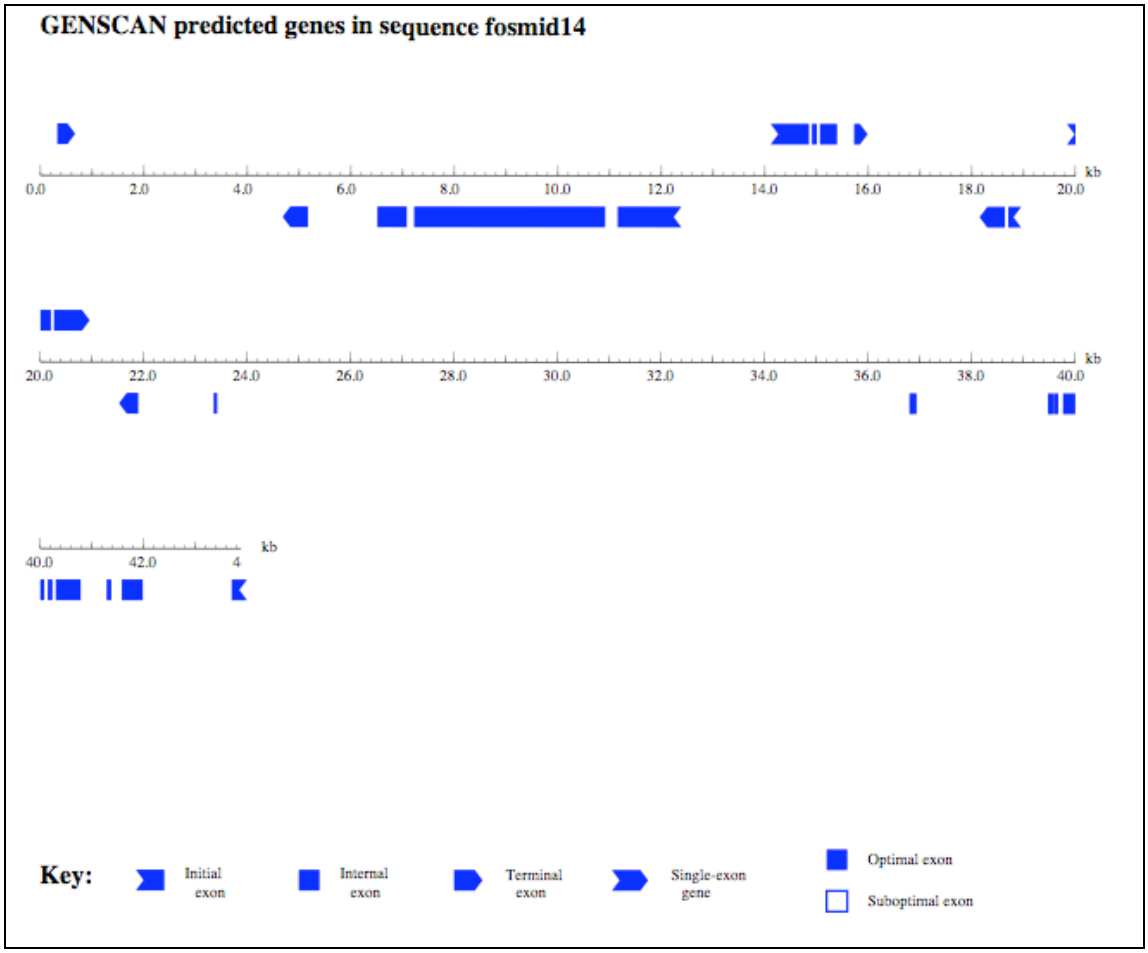
```
Dreriozebrafish_NP_956858      QAYFACEKNENLAANFLLQQNFDDD 382
Dmelanogaster_Rad23-PA         QAYFACEKNEEQAANFLLSSSFDD- 414
Dvir_CG1836-PA                 QAYFACEKDEELAANFLLSSSFDD- 448
                               ********:*: ******...**:
```

Next, I would analyze the conservation of the upstream promoter region of this gene. However, because promoter regions tend to be significantly less conserved than genes themselves, I would use only fly species for this ClustalW alignment, for evolutionary closeness. After obtaining the DNA sequences for the 1000 bp block of DNA upstream of Rad23 in *D. sechellia*, *D. simulans*, *D. mojavensis*, and *D. virilis*, I ran a the ClustalW alignment to find that there was significant alignment between these four fly species, especially in TATA box-like regions, which is understandable, considering genetic function. The ClustalW alignment follows.

```
CLUSTAL W (1.83) multiple sequence alignment


Dsechellia       ----------------------------------------TCTTGGCAATTCACGATATG 20
Dsimulans        ATGCAGCCATATTATCTCTACCTTGACAAGTGCATCGTTTTCTTGGCAATTCACGATATG 60
Dmojavensis      ------------------------------------------------------TTAA 4
Dvirilis         -----------AAACAAATAAATACATATGTTTTACATATCCTTATGTACTGTAAATATT 49
                                                                         *

Dsechellia       TCACGTGAAAAAGAAAGCTATGACCAGGATAGGCCAGGA--AAACACGAAAATCTGATG- 77
Dsimulans        TCACGCGAAAAAGAAAGCTATGACCAGGATAGGCCAGGA--AGACACGAAAATCTGATG- 117
Dmojavensis      TTGTTAGCATTGTAAATTTA-GCTCAAGATTAG---------AAATTGGATTGCTTAATT- 54
Dvirilis         TCATTAAGGCT-TATTTTTATTTCTATTGTTATGTATGTTTTATTTCTATTATTTTATTC 108
                 *            *    **     *   *                  *     * **

Dsechellia       CTA-ACACAAGAATAGAAAATTAATTCACACTAACAGA--CTTATTA----TACTAAG-- 128
Dsimulans        CTA-ACACAAGAATAGAAAATTAATTCACACTAACAGA--CTTATTA----TACTAAG-- 168
Dmojavensis      GTA-A---AAAAAAAGAATATTAGTTAACGTTACCGTATCCTTATTAAAAATATTTGA-- 108
Dvirilis         ATATATACATACAAACAACTTTTGTTCTTTTTAAGAAAGCCATATTTTCACTTTTAAATT 168
                  ** *   *   * * **  **  **     **    *  * ****     *  *

Dsechellia       --TGGATTGCTTAGTGGAAAAAGAATGAGTTCGTGT---ATTCTTTTATTATTCGTT--- 180
Dsimulans        --TGGATTGCTTAGTGGAAAAGGAATGAGTTCGTGT---ATTCTTTTATTATTCGTT--- 220
Dmojavensis      --AGGCAAATATTGTAGATATATAAACGTAATATAC---ATATTTACAGTACGCGTT--- 160
Dvirilis         TGACACCTATTTTATCGACCCGCTTGCAAAGTGTACCAAACACGTGTAACACACAATGGG 228
                            *  * **             *     *    *  *  *  *  *

Dsechellia       ACAAAATTA---TTTAGAATTAAT--ATATTTGTCATGTTTGACAAAAAA--TTCTCGAT 233
Dsimulans        ACAAAATTA---TTTAGAATTAAC--ATATTTGTCATGTTTGACAAAAAA--TTCTCGAT 273
Dmojavensis      CCTATCTCAAATTTTGTAAATCAC--TTGTTCGGTGCGGTTGA-AAAACT--TATTTTGT 215
Dvirilis         TTCACCGTAA-TGTTGTATTTTGCCATTGATTAGTTAACTTGACGGACTTACTACCCTGG 287
                    *    *    **  *  *      *  *        ****   *     *

Dsechellia       ACTTAC---TTCTTTTAAAAACCTCGTTGATGATGCAGCAATAGCTATCTTAATTTTTTA 290
Dsimulans        ACTTAC---TTCTTTTAAAAACC------------------------------------- 293
Dmojavensis      ATTGTA---TAATTCTGCCAGTA------------------------------------- 235
Dvirilis         ATTGAAATTTAGATTTCTTAATTACAGTTAGCCGTAACCGTCACC--------------- 332
                 * *      *   * *   *

Dsechellia       AAAACCCCGTTGATGATGCAGCAATAGCTATCTTAATTTCGAATTATTTGTGAATCTCAT 350
Dsimulans        ------CCGTTGATGATGAAGCAATAGTTATCTTAATTTCGAATTATTTGTGAATCTCAT 347
Dmojavensis      -------TGAGTATGATAATGGCGCATGCGCGCCAAAAAAAAATTGCTT-TAACAGTTA- 286
Dvirilis         ---GTCTTGTCCACATAGACAATTAAAATGCGCTCTTATCAATTTAATT-TAATAAAGAA 388
                        *   *           *              * **  ** * *     *

Dsechellia       ATTAACTCAAACGGAAAGGATTATAATAATCTCATTGCTGTATATCTTTTGAATTACGAT 410
Dsimulans        ATTAACTCAAACGGAAAGGATTATAATAATCTCATTGCTGTATATCTTTTGAATTACGAT 407
Dmojavensis      ATTAGCATAACC--AACTGATAGTTACTTTGGCGTTGCCAGACTGCAACT-ATTTATATT 343
Dvirilis         AT--ATTTAATCAACTATTTTTGTTATCCGCAATTTAAACAATTGTAAATGAGGCTTGTC 446
                 **      ** *        *  * *        **     *       * *

Dsechellia       GTTAGTCATTCAACGCCTATGTTGGACTAAA-ATATTTATAAGGAAATACCGAAAATATT 469
Dsimulans        GTTAGTCATTCAACGCCTATGTTGGACTAAA-ATATTTATAAGGAAATACCGAAAATATT 466
```

```
Dmojavensis      GTAATATTTATAATAAATATACCAAGTCATATAAAATTATATGGAAAAAC---AATTACT 400
Dvirilis         TTCACTTGTTCAAAACGGATGAAAAAACTTA---TTTTGTATTGTATAATTCTAACTGTG 503
                  * *    *  **     **          *     ** **  * *  *    ** *

Dsechellia       AAA--AATTACTACAATAGG-ACAAAAGTTTATTTGTAAC---GTTCATTATGACTCTTA 523
Dsimulans        AAA--AATTACTACAATAGG-ATAGAAGTTTATTTGTAAC---GTTCATTATGACTCTTA 520
Dmojavensis      CAGTTAATCAACATATTA---ATTAAATTTTATTTTATTT---ATTTATTTT--TTTTTA 452
Dvirilis         TGCGTACGAAACACGGTGCATGCGCAAACAAAATCGCTTCGGAAGTTAATAGGATTTGCA 563
                      *   *  *   *        **    * *          * * *     *   *

Dsechellia       TTACAAATCTTTACATCTTTATCTTAC-GTGTAAACGTGTTATAAGGCTTATGCTGCGAT 582
Dsimulans        TTACAAATCTTTACATCTTTATCTTATTGTGTGAACGTGTTATAAGGCTTATGCTGCGAT 580
Dmojavensis      TTTTA-----TCACGTTTTATTAATAT-GT-TTAACATATCAGTCAGTATATATAAAAGT 505
Dvirilis         ACTGACAGTTACAAGTTTTTATACGATACTCTAAAGATTGCTACTA--ACATCCTACATT 621
                     *       *  * **  *   *   * * **  *           **       *

Dsechellia       TTGGTGTCTTTTAAGCGAGAAGGTATTATTTGCCAGTCGATGTGCGGTCACCCTGATG-A 641
Dsimulans        TTGGTGTCTTTTAAACGAGAAGGTATTATTTGCCAGTCGATGTGCGGTCACCCTGATG-A 639
Dmojavensis      ACAGTACATATTCAGGCACATTTTGGTATTTAAAAA-CATTGAGCCGG-ACTCTGTTATA 563
Dvirilis         GTAGTGGATCATGAA--------TATTATTCCTCTG-CATTTCGAATGTCTCCTGTAA-A 671
                    **   *  * *         *  ****       *  *  *        ***    *

Dsechellia       CAAATCGTTTTTATCAAGCGATTTTTGGAACTAACTTTGAGTTTGCATATGCGTATGCAT 701
Dsimulans        CAAATCGTTTTTATCAAGCGATTTTGGGAACTAACTTTGAGTTTGCATATGCGTATGCAT 699
Dmojavensis      TAGGTGCTGCCGGTGGAGTCAATCAAAAGAATGATATGTTTTTTGTAGATTTATATAGAT 623
Dvirilis         AAATTTGTAATTGTGTAATTATAAGTGTGCTTATTGTTTATTATATACATTTCTACGTCT 731
                  *  *  *     *  *   *          *    *    * *  * **   **    *

Dsechellia       AGTATATTTAATTAAGTTTGGTTTTGTCTAAT--ACACAAGATGATTATTACAATTAAAA 759
Dsimulans        AGTATATTTAATTAAGTTTGGTTTTGTCTAAT--ACACACGATGATTATTACAATTAAAA 757
Dmojavensis      GCTATATT------AGGACCGTTATAATAAAC--ACATAAATTAACTAAAAAATCTTCTA 675
Dvirilis         GTTATT---------GCCTTGCTGCAAAAAATCAATTCAATTTGTACACAAGCACAACTG 782
                   ***          *    * *      **   *   *   *    *  *

Dsechellia       ATCTTCA-ACAGCAAACTTTTACTATAGAGTTTGCCCCGGAAAAAACGGTATGTGTATAA 818
Dsimulans        ATCTTCA-ACAGCAAACTTTTACTATAGAGTTTGCCCCGGAAAAAACGGTATGTGTATAA 816
Dmojavensis      TTATTAAGACTACGCACATTCAGTATTTTTTCATATCTTAAGGACATCCTTAAAATCTAT 735
Dvirilis         CAATTTA-ACTGTACAGAACCCATGTATTACCACA-CGTAAATACA-----AATACATCG 835
                    ** * **     *       * *          *   *  * *          *

Dsechellia       ATTGGCATATACAAGTATAATCACATATATGTGTATACAATTATTGCAAATGCATATGTG 878
Dsimulans        ATTGGCATATACAAATATAATCACATATATGTGTATACAATTATTGCAAATGCATATGTG 876
Dmojavensis      ATTACTA-ATAT---TTTAATTACAAACAATAATATCGCTTTATTCGATATTCG-----A 786
Dvirilis         ATAAG-AGAAATAGAATTTACTCTTCACGGAAATATCGTTTTCTTC--CATTCG-----A 887
                 **    * * *      * *      *      ***    ** **    ** *

Dsechellia       TGCTTATACAGGTATTTCTTACAATTAATTTCTGATACCACAATAAGCAGTGATTCTT-A 937
Dsimulans        TGCTTATACAGGTATTTCTTACCATTAATTTCTGATACCATAATAAGCATTGCTTCTTTA 936
Dmojavensis      TATTAATA----TATTTTGGACCAT----CTCTAGTGTC--GATGACAAACAATTTTGTG 836
Dvirilis         TATTTATA------TTTTAGGCCAC----CCCTAGTGTC--GATGACAAACATTTTCGTG 935
                 *  * ***      ***    * *       **  *  *   ** *  *    **

Dsechellia       TTACAAAAGG--TTTTGGAACTGA-AGAAGAAAATATTTGACGAA-CGCGGTTCAGAGTA 993
Dsimulans        TTACAAAAGG--TTTTGGAACTGA-AAAAAAAGATATTTGACGAA-CGCGGTCCCGAGTA 992
Dmojavensis      AGAAAGAAGA--GACCAAGGCAAA-AGTTAACAATACCAATTTAGTCCCAAATAAAAATT 893
Dvirilis         AGAAAAAAAAAAGAGCAAGGAAAATAATTAAAAATAACGAATTAT-TGCTAACAAAAAAA 994
                   * * **               * *    *  ***       *    *       *

Dsechellia       CGTCGCC- 1000
Dsimulans        CGTCTCCA 1000
Dmojavensis      TGTAAAA- 900
Dvirilis         TATAAA-- 1000
                   *
```

Appendix 1 – Full version of GenScan output for Fosmid 14



Appendix 2 – Exon boundaries summary table

	start	end	phase	STRAND	dvir chrom
Feature 14.2.1*	12231	12236	0	-	4
Feature 14.2.2	6510	12161	0	-	4
Feature 14.3.1	13888	13953	0	+	4
Feature 14.3.2	14036	14114	0	+	4
Feature 14.3.3	14189	14851	2	+	4
Feature 14.3.4	14908	15002	1	+	4
Feature 14.3.5	15061	15393	0	+	4
Feature 14.3.6	15723	15830	0	+	4
Feature 14.4.1	18702	18803	0	-	4
Feature 14.4.2	18302	18637	0	-	4
Feature 14.5.1	19725	19820	0	+	4
Feature 14.5.2	19892	20201	0	+	4
Feature 14.5.3	20260	20798	0	+	4
Feature 14.6.1	41577	41978	2	-	4
Feature 14.6.2	41221	41370	0	-	4
Feature 14.6.3	40812	41053	0	-	4
Feature 14.6.4	40299	40743	1	-	4
Feature 14.6.5	40147	40243	0	-	4
Feature 14.6.6	39763	40076	2	-	4
Feature 14.6.7	39472	39670	0	-	4
Feature 14.6.8	39219	39364	2	-	4
Feature 14.6.9	38946	39097	0	-	4
Feature 14.6.10	36795	36933	1	-	4
Feature 14.6.11	23349	23417	0	-	4
Feature 14.6.12	21672	21890	0	-	4

* Feature 14.2.1 denotes Fosmid 14, Predicted Feature 2, Exon 1