Annotation of *Drosophila erecta* Fosmid 14

Richard L. O'Neal

Wofford College
Dr. Robert Moss

Overview

This report covers the annotation of fosmid14, a 50kb fragment on the long arm of chromosome 3 in *Drosophila erecta*. This project is important to the scientific community because it allows for a better comparison between species, the ability to conduct any number of further analyses, and an overall expansion of the current genomic database. In this report, I will be annotating all important genes as well as verifying any major repeats that occur within this fosmid. Through annotation analysis, three genes were verified to be located within this fosmid: CG7140, TyrR, and CG14561. Figure 1 shows a final map of fosmid14, noting all genes annotated and showing synteny with *D. mel*.

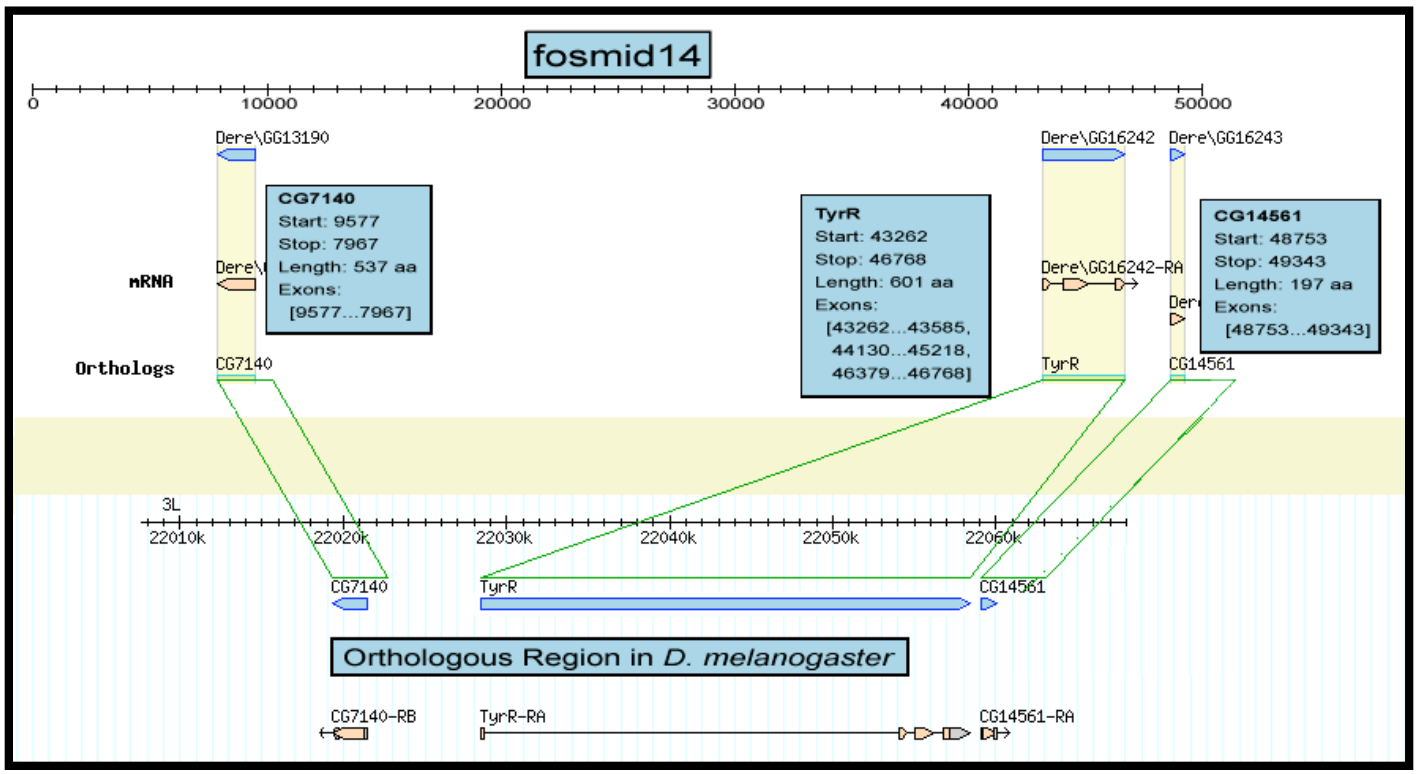

Figure 1. Map of fosmid14 showing all annotated genes as well as their ortholog in *D. mel*.

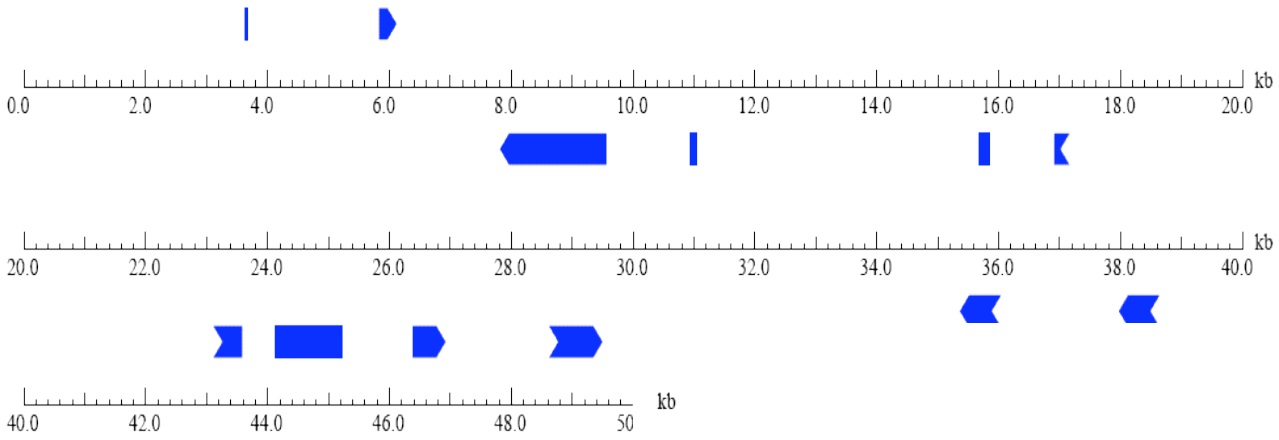

Figure 2. Genscan map output of fosmid14.

Genscan Predicted Genes		
Feature	Gene	Accession Number in *D.mel*
1	Mispredicted	Feature mispredicted by Genscan
2	CG4170	NP_649376.3
3	Mispredicted	Feature mispredicted by Genscan
4	Mispredicted	Feature mispredicted by Genscan
5	TyrR	Incorrect Genscan splicing
6	CG14561	NP_649377.1

Table 1.

N-Scan Predicted Genes		
Feature	Gene	Accession Number in *D.mel*
1	CG4170	NP_649376.3
2	TyrR	NP_524419.2
3	CG14561	NP_649377.1

Table 2.

Gene 1: CG4170 Ortholog

The first real gene that was predicted by both Genscan and N-Scan was verified to be an ortholog of *D. melanogaster* CG4170. The peptide predicted by N-Scan was found to be 537 amino acids long while Genscan predicted a protein with a length of 657. Genscan had begun predicting the amino acid sequence one start codon too early, accounting for a longer peptide. Later analysis recognized this error and the correct translation is indicated in this report. The transcribed protein of this predicted feature is 537 amino acids long and runs in the reverse direction of fosmid14. This predicted protein was analyzed using BlastP and found to be 93% identical to CG4170 located on chromosome 3L of *D. melanogaster*.

```
>ref|NP_649376.3| UG CG7140 [Drosophila melanogaster]
 gb|AAF51801.4|  CG7140 [Drosophila melanogaster]
Length=533

 GENE ID: 40445 CG7140 | CG7140 gene product from transcript CG7140-RB
[Drosophila melanogaster] (10 or fewer PubMed links)

 Score = 1056 bits (2730),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 496/533 (93%), Positives = 515/533 (96%), Gaps = 0/533 (0%)

Query  1    MISIDPHNEEAYSVVVFGASGGLAKKKVFPALWALYRENRLPPGTKIFTFTRTPLQTKTY  60
            MIS+DPHNEEAYS+VVFGASGGLAKKKVFPALWAL+RENRLP GTKIFTFTR+PLQTKTY
Sbjct  1    MISMDPHNEEAYSIVVFGASGGLAKKKVFPALWALFRENRLPQGTKIFTFTRSPLQTKTY  60

Query  61   RLQILPYMELDKHRDPKKYNLFWKTVHCVQGEYDKPEHYVALTEAMVHQETKHNQVHANR  120
            RLQILPYMELDKHRDPKKYNLFW TVHCVQGEYDKPE+YVALTEAMVHQETKHNQV ANR
Sbjct  61   RLQILPYMELDKHRDPKKYNLFWTTVHCVQGEYDKPENYVALTEAMVHQETKHNQVRANR  120

Query  121  IFYLALPPIVFDQVALNVSRKCSSSTGWNRIIVEKPFARDDVSYKVFQTSLCNCFRESQI  180
            IFYLALPPIVFDQV LNVSRKCSS+TGWNRIIVEKPFARDD+SYK FQTSLCNCFRESQI
Sbjct  121  IFYLALPPIVFDQVTLNVSRKCSSTTGWNRIIVEKPFARDDISYKAFQTSLCNCFRESQI  180

Query  181  YLMDHLLSRQVMQNFFALRYSNHLWGETLNNRHVAAVMISVKCELPVAASRADYFNQFGI  240
            YLMDHLLSRQVMQNFFALRYSNHLW ETLN+RHVAAVMIS+KCELPV+ +RADYFNQFGI
Sbjct  181  YLMDHLLSRQVMQNFFALRYSNHLWAETLNHRHVAAVMISIKCELPVSVNRADYFNQFGI  240

Query  241  IRDLMTNHMIQTLAMLAMDQPYANTADDLRVERLKVLRQVLTPHIGDVVLAQYRNNRRES  300
            IRDLMTNHMIQ LAMLAMDQPYANTADDLR ERLKVLRQVLTP+IGDVVLAQYRNNRRE+
Sbjct  241  IRDLMTNHMIQMLAMLAMDQPYANTADDLRAERLKVLRQVLTPNIGDVVLAQYRNNRRET  300

Query  301  EPAKCGYTEHTYIPKDSFTPTFALVVLQINNRRWSGVPFILRAGKALNDTKSEVRIQYKP  360
            +PAKCGYTEHTYIPKDSFTPTFALVVL INNRRW+GVPFILRAGKALNDTKSEVRIQYKP
Sbjct  301  DPAKCGYTEHTYIPKDSFTPTFALVVLHINNRRWTGVPFILRAGKALNDTKSEVRIQYKP  360

Query  361  VDCDAFDSDSADIRNELVLRSFPTEEVFMRMRLKRQGESICLRESEVNLRVDDRGPKGLQ  420
            VDCD F SDS DIRNELVLRSFPTEEVFMRMRLKRQGE ICLRESE+NLRVDDRGPKGLQ
Sbjct  361  VDCDTFHSDSTDIRNELVLRSFPTEEVFMRMRLKRQGEDICLRESEINLRVDDRGPKGLQ  420

Query  421  GLPGFLLNVFHGDQTLFMRSDEQCEIWRIFSPVLATIDSDRPRPLHYDFGSRGPLLAYRK  480
            GLPG+LLNVF GDQTLFMRSDEQCEIWRIFSPVLATIDSDRPRPLHYDFGSRGPLLAYRK
Sbjct  421  GLPGYLLNVFQGDQTLFMRSDEQCEIWRIFSPVLATIDSDRPRPLHYDFGSRGPLLAYRK  480

Query  481  AERAGFVFFASDEWHQSEETLEYTVKRSKQLIGPHTALKPVRDPRCKRLSSNP  533
            AERAGFVFFA+DEWHQSEETLEYTVK SKQLIGPHTALKPVRDPR KR +SNP
Sbjct  481  AERAGFVFFATDEWHQSEETLEYTVKNSKQLIGPHTALKPVRDPRSKRSNSNP  533
```

Figure 3. BlastP comparison of *D. erecta* fosmid14 feature against *D. mel.* CG7410.

The Blast results reveal that there has been an extremely high conservation of this gene between these two species of *Drosophila*. Not only are the peptide sequences nearly identical, but in both species, the translated region of this gene occurs on a single exon on chromosome 3L. The biggest difference between the two is that the ortholog in *D. erecta* has an additional 4 amino acids immediately preceding the stop codon. Presumably, this mutation does not affect the viability of the protein because the rest of the sequence has been faithfully conserved with that of *D. melanogaster.* This gene is involved in the glucose metabolic pathway and is structurally very similar to Zwischenferment, isoform A (NP_523411.1), a glucose-6-phosphate dehydrogenase located on chromosome X of *D. mel.* (E value = 4e-113).

To verify that Genscan and N-Scan had not made any errors in their predictions, the entire fosmid was run through the *sixpack* tool of the EMBOSS suite (http://sbcr.bii.a-star.edu.sg/cgi-bin/emboss/menu/sixpack). By doing this, the positions of the start and stop codons of this feature were rechecked and the coding sequence was verified to all be contained within the same reading frame. After completing all of these analyses, the proposed gene was verified using Washington University's Gene Model Checker (http://gander.wustl.edu/~wilson/genechecker/index.html), and is proposed in high confidence to be an ortholog of CG7140 found in *D. melanogaster*).

Gene 2: TyrR Ortholog

The next gene found in this fosmid was independently predicted by both Genscan and N-Scan with no discrepancies between their proposed peptide sequences. This feature is 601 amino acids long and is located in the positive direction on fosmid14. A BlastP analysis predicted that this feature is an ortholog of the *D. melanogaster* Tyramine Receptor located on chromosome 3L with an identity of 96% and an E-value of 0.0.

```
>ref|NP_524419.2| UG tyramine receptor [Drosophila melanogaster]
 sp|P22270.2|OAR_DROME  RecName: Full=Tyramine/octopamine receptor; AltName: Ful
Flags: Precursor
 emb|CAA38565.1|  Tyramine-dro receptor [Drosophila melanogaster]
 gb|AAF51802.1|  tyramine receptor [Drosophila melanogaster]
 gb|ABE73326.1|  IP02940p [Drosophila melanogaster]
 gb|ACL88716.1|  TyrR-PA [synthetic construct]
Length=601

 GENE ID: 42452 TyrR | Tyramine receptor [Drosophila melanogaster]
(Over 10 PubMed links)

 Score = 1104 bits (2856),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 579/601 (96%), Positives = 586/601 (97%), Gaps = 0/601 (0%)

Query  1    MPSADQILFVNVTTTVAAAALTAAAAVSTTNSGSDHAARGYADTDADAGMGTETVANISG  60
            MPSADQILFVNVTTTVAAAALTAAAAVSTT SGS +AARGY D+D DAGMGTE VANISG
Sbjct  1    MPSADQILFVNVTTTVAAAALTAAAAVSTTKSGSGNAARGYTDSDDDAGMGTEAVANISG  60

Query  61   SLVEGLTTVTAALSTAQADPDSAGECDGAVDELHASVLGLQLAVPEWEALLTALVLSVII  120
            SLVEGLTTVTAALSTAQAD DSAGEC+GAV+ELHAS+LGLQLAVPEWEALLTALVLSVII
Sbjct  61   SLVEGLTTVTAALSTAQADKDSAGECEGAVEELHASILGLQLAVPEWEALLTALVLSVII  120

Query  121  VLTIIGNILVILSVFTYKPLRIVQNFFIVSLAVADLTVALLVLPFNVAYSILGRWEFGIH  180
            VLTIIGNILVILSVFTYKPLRIVQNFFIVSLAVADLTVALLVLPFNVAYSILGRWEFGIH
Sbjct  121  VLTIIGNILVILSVFTYKPLRIVQNFFIVSLAVADLTVALLVLPFNVAYSILGRWEFGIH  180

Query  181  LCKLWLTCDVLCCTSSILNLCAIALDRYWAITDPINYAQKRTVGRVLLLISGVWLLSLLI  240
            LCKLWLTCDVLCCTSSILNLCAIALDRYWAITDPINYAQKRTVGRVLLLISGVWLLSLLI
Sbjct  181  LCKLWLTCDVLCCTSSILNLCAIALDRYWAITDPINYAQKRTVGRVLLLISGVWLLSLLI  240

Query  241  SSPPLIGWNDWPDEFTSATPCELTSQRGYVIYSSLGSFFIPLAIMTIVYIEIFVATRRRL  300
            SSPPLIGWNDWPDEFTSATPCELTSQRGYVIYSSLGSFFIPLAIMTIVYIEIFVATRRRL
Sbjct  241  SSPPLIGWNDWPDEFTSATPCELTSQRGYVIYSSLGSFFIPLAIMTIVYIEIFVATRRRL  300

Query  301  RERARANKLNTIALKSTELEPMANSSPVAASNSGSKSRLLASWLCCGRDRAQFATPMIQN  360
            RERARANKLNTIALKSTELEPMANSSPVAASNSGSKSRLLASWLCCGRDRAQFATPMIQN
Sbjct  301  RERARANKLNTIALKSTELEPMANSSPVAASNSGSKSRLLASWLCCGRDRAQFATPMIQN  360

Query  361  DQESISSETHQPQDSSKAGSQGNSDSQQQHMVVLVKKSRRAKIKDSIKHGKARGGRKSQS  420
            DQESISSETHQPQDSSKAG  GNSD QQQH+VVLVKKSRRAK KDSIKHGK RGGRKSQS
Sbjct  361  DQESISSETHQPQDSSKAGPHGNSDPQQQHVVVLVKKSRRAKTKDSIKHGKTRGGRKSQS  420

Query  421  SSTCEPHGEQQLLPAGGDGGSCQAGGRHSGGGKSDAEISTESGSDPKGCIQVCVTQADEQ  480
            SSTCEPHGEQQLLPAGGDGGSCQ GG HSGGGKSDAEISTESGSDPKGCIQVCVTQADEQ
Sbjct  421  SSTCEPHGEQQLLPAGGDGGSCQPGGGHSGGGKSDAEISTESGSDPKGCIQVCVTQADEQ  480



Query  481  TSLKLTPPQSSTGAAAVSATPLQKKTSGVNQFIEEKQKISLSKERRAARTLGIIMGVFVI  540
            TSLKLTPPQSSTG AAVS TPLQKKTSGVNQFIEEKQKISLSKERRAARTLGIIMGVFVI
Sbjct  481  TSLKLTPPQSSTGVAAVSVTPLQKKTSGVNQFIEEKQKISLSKERRAARTLGIIMGVFVI  540

Query  541  CWLPFFLMYVILPFCQSCCPTNKFKNFITWLGYINSGLNPVIYTIFNLDYRRAFKRLLGL  600
            CWLPFFLMYVILPFCQ+CCPTNKFKNFITWLGYINSGLNPVIYTIFNLDYRRAFKRLLGL
Sbjct  541  CWLPFFLMYVILPFCQTCCPTNKFKNFITWLGYINSGLNPVIYTIFNLDYRRAFKRLLGL  600

Query  601  N  601
            N
Sbjct  601  N  601
```

Figure 4. BlastP comparison of *D. erecta* fosmid14 feature against *D. mel.* TyrR.

This is the only multi-exonal translated feature in fosmid14. Both Genscan and N-Scan predicted two splice sites within the peptide for a total of three translated exons. These splice sites as well as the start and stop codons were again verified using tools in the Emboss suite. This feature in fosmid14 and its proposed ortholog have the exact same number of amino acids with no gaps, and all splice sites between exons correlate completely.

In *D. mel,* the Tyramine receptor is involved in the olfactory pathway. Due to high amount of conservation between species, the orthologous feature in fosmid14 is predicted to also be a receptor involved in the same pathways. Amongst other species of Drosophila, it also shares a high amount of conservation with G-protein coupled receptors (notably tyramine and octopamine).

Based on the independent predictions of Genscan and N-Scan, a high correlation proposed by BlastP, splice verification by Emboss, and a positive test result of Gene Model Checker, this second feature on fosmid14 is proposed in high confidence to be an ortholog of *D. melanogaster*'s TyrR gene.

Gene 3: CG14561 Ortholog

The final gene found in fosmid14 is proposed to be an ortholog of CG14561 in *D. melanogaster*. It is a single-exon gene that runs in the positive direction along this sequence. Genscan predicted a peptide sequence with a length of 191 amino acids, while N-Scan proposed one with a length of 197. This discrepancy was the result of N-Scan beginning translation one start codon before that one predicted by Genscan. The Blast comparison to the orthologous region in *D. melanogaster* suggests that the N-Scan prediction is indeed the correct one.

```
>ref|NP_649377.1| UG CG14561 [Drosophila melanogaster]
 gb|AAF51803.1|  CG14561 [Drosophila melanogaster]
 gb|AAS15647.1|  UT01525p [Drosophila melanogaster]
Length=194

 GENE ID: 40447 CG14561 | CG14561 gene product from transcript CG14561-RA
[Drosophila melanogaster] (10 or fewer PubMed links)

 Score =  345 bits (885),  Expect = 1e-93, Method: Compositional matrix adjust.
 Identities = 176/197 (89%), Positives = 183/197 (92%), Gaps = 3/197 (1%)

Query  1    METSEAMNTCEEGDQLNGSFRLQCSMSSVDIESESDLSLAFEREEQSPGGSCDELPAFEI  60
            METSEA+NTCEEGDQLNGSFRLQCSMSSVD+ESESDLSLAFEREEQ P G CDELPAFEI
Sbjct  1    METSEAVNTCEEGDQLNGSFRLQCSMSSVDMESESDLSLAFEREEQIPDGGCDELPAFEI  60

Query  61   RAFSPHGRTPSPIDDLDLSDIETPKQPLRQQLPPDEEEPSRQNDPQYVALHEINAQISPP  120
            RAFSPHGRTPSPIDDLDLSDIETPKQP RQ L P EEEP  QNDPQYVALHEINAQISPP
Sbjct  61   RAFSPHGRTPSPIDDLDLSDIETPKQPQRQHLLP-EEEP--QNDPQYVALHEINAQISPP  117

Query  121  SDADDSKNLETIYEGVFLSTPPRDKTSNSGISRFTPKRGRANLIELMANRLHSGKENQSP  180
            +DADDS +LETI+EGVFLSTPPRDK S+SG SRFTPKRGRANLIELMANRLH GKENQSP
Sbjct  118  NDADDSNSLETIFEGVFLSTPPRDKASSSGNSRFTPKRGRANLIELMANRLHGGKENQSP  177

Query  181  GAGVSEPHLPSPLKDPT  197
            GA VS+PHLPSPLKDPT
Sbjct  178  GARVSDPHLPSPLKDPT  194
```

Figure 4. BlastP comparison of *D. erecta* fosmid14 feature against *D. mel.* CG14561.

A BlastP analysis predicted that this feature was an ortholog of CG14561 in *D. melanogaster*, a single-exon gene located on chromosome 3L. There was a high degree of conservation between these two features (89% identity, e-value=1e-93) that suggests that they are indeed orthologs of one another. Like with other features in this fosmid, the start and stop codons were checked using the Emboss suite, and the final model was verified using Gene Model Checker.

This gene's function has not yet been well-researched in *D. melanogaster*. However it shares an 88 bp anticodon-binding domain with Arthrobacter phenylalanyl-tRNA synthetase, so it is possible that this gene could share a similar function.

Based on the evidence outlined above, this feature on fosmid14 of *D. erecta* is proposed in high confidence to be an ortholog of CG14561 on chromosome 3L of *D. melanogaster*.

Clustal Analysis

I chose to perform a ClustalW2 alignment on my second annotated gene in this fosmid: TyrR. A comparison of this gene between four species of *Drosophila* (*erecta, yakuba, melanogaster,* and *virilis*), a honeybee, and a moth revealed a very high degree of conservation in the peptide sequences. In fact, a span between 105 and 311 (on the *D. erecta* protein) showed an almost exact conservation between all six species with no gaps. There was also a 96 amino acid stretch on the end of this protein that was conserved in all species (See Figure 5).

```
Dyak_TyrR               KGCIQVCVTQADEQTSLKLTPPQSSTGVAAVSATPLQKKTSGVNQFIEEKQKISLSKERR 526
Dere_TyrR               KGCIQVCVTQADEQTSLKLTPPQSSTGAAAVSATPLQKKTSGVNQFIEEKQKISLSKERR 526
Dmel_TyrR               KGCIQVCVTQADEQTSLKLTPPQSSTGVAAVSVTPLQKKTSGVNQFIEEKQKISLSKERR 526
Dvir_TyrR               KGCIQVCVTQAEEQTSLKLTPPQSSSTGATAAAPALQKKPSTVNQFIEEKQKISLSKERR 533
Apis_mellifera_TyrR     ------------------TTSSRRTTGSRAAAT-----TTTVYQFIEERQRISLSKERR 324
Bombix_mori_TyrR        -----------KETHEDNMIEITEAAPVKIQKRPKQNQTNAVYQFIEEKQRISLTRERR 403
                                           :             .. * *****:*:***::***

Dyak_TyrR               AARTLGIIMGVFVICWLPFFLMYVILPFCQSCCPTNKFKNFITWLGYINSGLNPVIYTIF 586
Dere_TyrR               AARTLGIIMGVFVICWLPFFLMYVILPFCQSCCPTNKFKNFITWLGYINSGLNPVIYTIF 586
Dmel_TyrR               AARTLGIIMGVFVICWLPFFLMYVILPFCQTCCPTNKFKNFITWLGYINSGLNPVIYTIF 586
Dvir_TyrR               AARTLGIIMGVFVICWLPFFLMYVILPFCQSCCPTNKFKNFITWLGYINSGLNPVIYTIF 593
Apis_mellifera_TyrR     AARTLGVIMGVFVVCWLPFFLMYVIVPFCPDCCPSDRMVYFITWLGYVNSALNPLIYTIF 384
Bombix_mori_TyrR        AARTLGIIMGVFVVCWLPFFVIYLVIPFCVSCCLSNKFINFITWLGYVNSALNPLIYTIF 463
                        ******:******:******::*:::***  ** ::::  *******:**.***:*****

Dyak_TyrR               NLDYRRAFKRLLGLN- 601
Dere_TyrR               NLDYRRAFKRLLGLN- 601
Dmel_TyrR               NLDYRRAFKRLLGLN- 601
Dvir_TyrR               NLDYRRAFKRLLGIN- 608
Apis_mellifera_TyrR     NLDYRRAFRRLLRIR- 399
Bombix_mori_TyrR        NMDFRRAFKKLLFIKC 479
                        *:*:****::** :.
```

Figure 5. A ClustalW2 alignment of the 3' end of TyrR in six different insects.

Using Flybase, I obtained the extended gene sequences of TyrR in each of the four species of *Drosophila* above. From each, I extracted a sequence beginning 2000 bps in front of the start codon, and continuing 500 bp into the first coding sequence (this provided Clustal with a firm anchoring point for analysis). In doing this, I discovered that in *D. virilis*, there are two copies of this gene back-to-back, so for this species, I extracted sequences from both copies. After performing a ClustalW2 analysis on all 5 sequences, I discovered that there is indeed a highly conserved region preceding the start codon in each of these cases. In fact, these regions are conserved between these species well beyond 1500 bps before the start codon. There were several areas within this region that showed a particularly high degree of conservation. The first was a segment of about 100 bps preceding the start codon by 200 bps. The next was a sequence of around 500 bps spanning a region 400-900 bps before the start codon. The final highly conserved untranslated region that I noted was 200 bps long, centered around 1500 bases before the start codon. From this analysis, I would predict that there is indeed substantial conservation of the untranslated region preceding the coding sequence of the TyrR peptide. This is also supported by the fact that *D. melanogaster* is known to have a large intron in this region. Therefore, a point mutation in this region could potentially affect splicing or translation of the

TyrR gene, thus giving a possible explanation why this region has been so conserved between species of *Drosophila*. The UCSC genome browser provides a visualization of genetic conservation between 12 species of *Drosophila* as well as mosquito, honeybee, and beetle (See Figure 6 below). Here one can see that throughout fosmid14 there is a high degree of conservation.

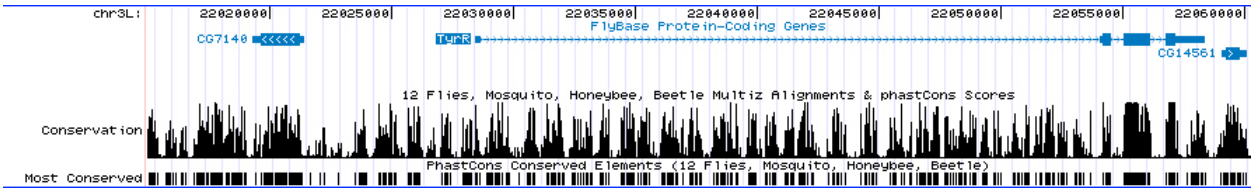

Figure 6. UCSC browser view of the conserved regions of *D. erecta* fosmid14 between 14 other insects.

Repeat analysis

Repeat Masker recognized 28 repeats in three different families. Table 3 below outlines and describes each of these features, while Table 4 summarizes this data by family. Two of the annotated repeats exceeded 500 bps in length. The first was found beginning at position 11197 of fosmid14 and continuing until 11732. The second was found beginning at the 31301 base and continuing until 31881. These repeats are all shown in relation to their position in *D. erecta* in Figures 7 and 8 below.

Begin	End	Repeat	Family	ID
42	141	DNAREP1_DM	DNA/Helitron	1
2864	2957	DNAREP1_DM	DNA/Helitron	2
2966	3097	DNAREP1_DM	DNA/Helitron	2
10046	10071	AT_rich	Low Complexity	3
10049	10074	AT_rich	Low Complexity	4
10155	10194	DNAREP1_DM	DNA/Helitron	5
11197	11732	*Helitron-1_DYak	DNA/Helitron	6
11742	11822	DNAREP1_DM	DNA/Helitron	7
12534	12694	Helitron-1_DYak	DNA/Helitron	8
12757	12820	DNAREP1_DM	DNA/Helitron	9
18760	18782	AT_rich	Low Complexity	10
18887	19208	Helitron-1_DYak	DNA/Helitron	11
19214	19282	Helitron-1_DYak	DNA/Helitron	6
22984	23019	AT_rich	Low Complexity	12
26209	26239	AT_rich	Low Complexity	13
26219	26244	AT_rich	Low Complexity	14
31036	31280	Helitron-1_DYak	DNA/Helitron	15
31301	31881	*DNAREP1_DM	DNA/Helitron	16
36573	36597	AT_rich	Low Complexity	17
38752	38781	(TCTG) n	Simple Repeat	18
39125	39159	AT_rich	Low Complexity	19
43733	43987	Helitron-1_DYak	DNA/Helitron	20
45250	45429	Helitron-1_DYak	DNA/Helitron	21
45429	45523	DNAREP1_DM	DNA/Helitron	22
45523	45777	Helitron-1_DYak	DNA/Helitron	23
45781	45855	DNAREP1_DM	DNA/Helitron	24
46002	46138	DNAREP1_DM	DNA/Helitron	25
48337	48370	(TAAA) n	Simple Repeat	26

Table 3. Repeats in fosmid14 masked by Repeat Masker. *2 repeats were longer than 500 bp

Table 4. Summary of repeats found in fosmid14.

Repeat	Number of Elements	Percentage of fosmid14
DNA transposons	16	6.74%
Simple Repeats	2	0.13%
Low Complexity	8	0.37%

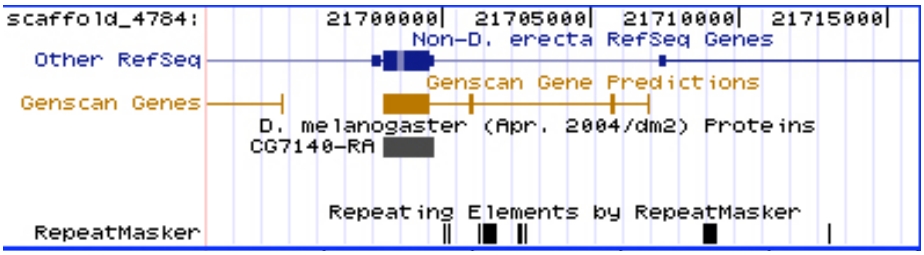

Figure 7. Bases 1-25000 of fosmid14 including CG7140 and 14 of 28 repeats. (See Table 3)

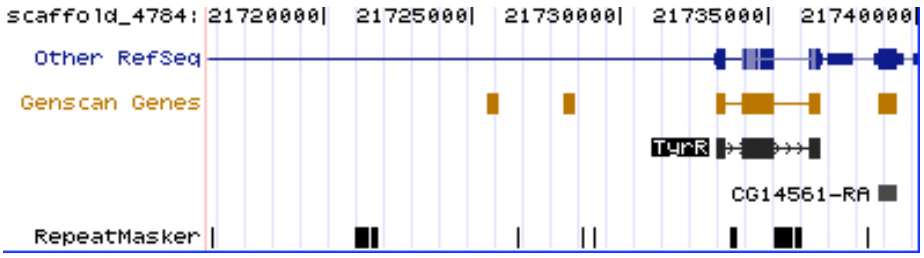

Figure 8. Bases 25001-50000 of fosmid14 including TyrR, CG14561 and 14 repeats. (See Table 3)

The large majority of these repeats are conserved with *D. melanogaster*, with the major exception being those that occur in the first 11000 bps of fosmid14. These, however, are not novel repeats as they all share a high correlation with known repeat families within *Drosophila.* There is also a large grouping of repeats within the introns of the TyrR gene.

Synteny

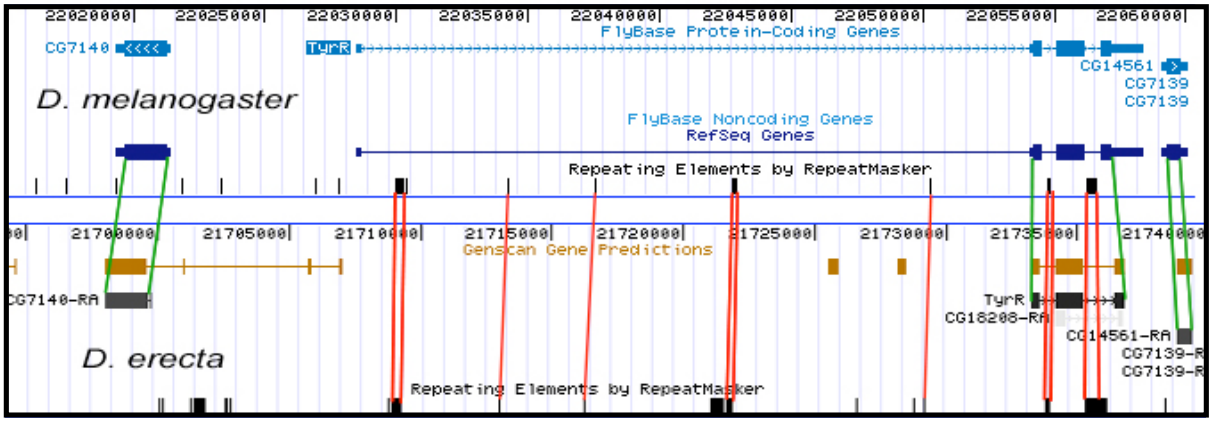

Figure 9. Synteny comparison of *D. erecta* with *D. melanogaster* (genes shown in green, repeats in red).

Fosmid14 has a high degree of synteny with *D. melanogaster*. All genes in this region occur in both species in the same order and the same direction. They even share a large number of repeats. The most significant divergence between these two species is the insertion of a 535 bp transposon between CG7140 and TyrR in *D. erecta* (See Table 3). Even the flanking regions of the genes have retained a relative amount of conservation between the two species, as referenced in my ClustalW alignment of TyrR. Based on the evidence of coding sequence conservation, repeat analysis, and the overall position of these features in relation to one another, there have been no major recombinatorial events in fosmid14 compared to *D. melanogaster*.