**Overview**

Fosmid XAAA112 consists of 34,783 nucleotides. Blat results indicate that this fosmid has significant identity to the 2R chromosome of *D.melanogaster*. Evidence suggests that fosmid XAAA112 contains a gene with homology to unnamed gene CG-10440 in *D.melanogaster*. The entire coding region of the CG-10440 gene has high identity match to fosmid XAAA112. In addition, evidence suggests that fosmid XAAA112 also contains a portion of a gene with homology to CG-10079, an Epidermal Growth Factor Receptor (Egfr) gene in *D.melanogaster*. Exons 2-4 of the Egfr gene show high identity matches to this fosmid, however it appears that exon 1 is located off the end of the contig. Furthermore, XAAA112 contains a large tandem repeat region covering over 6kb of the fosmid and consisting of a 175bp cassette. Analysis suggests that this large tandem repeat is not present in the *D.melanogaster* genome. In addition to the tandem repeat, four repeat regions were identified by RepeatMasker software, none of which were located within gene regions. Figure 1 below illustrates a diagram of feature location on fosmid XAAA112.

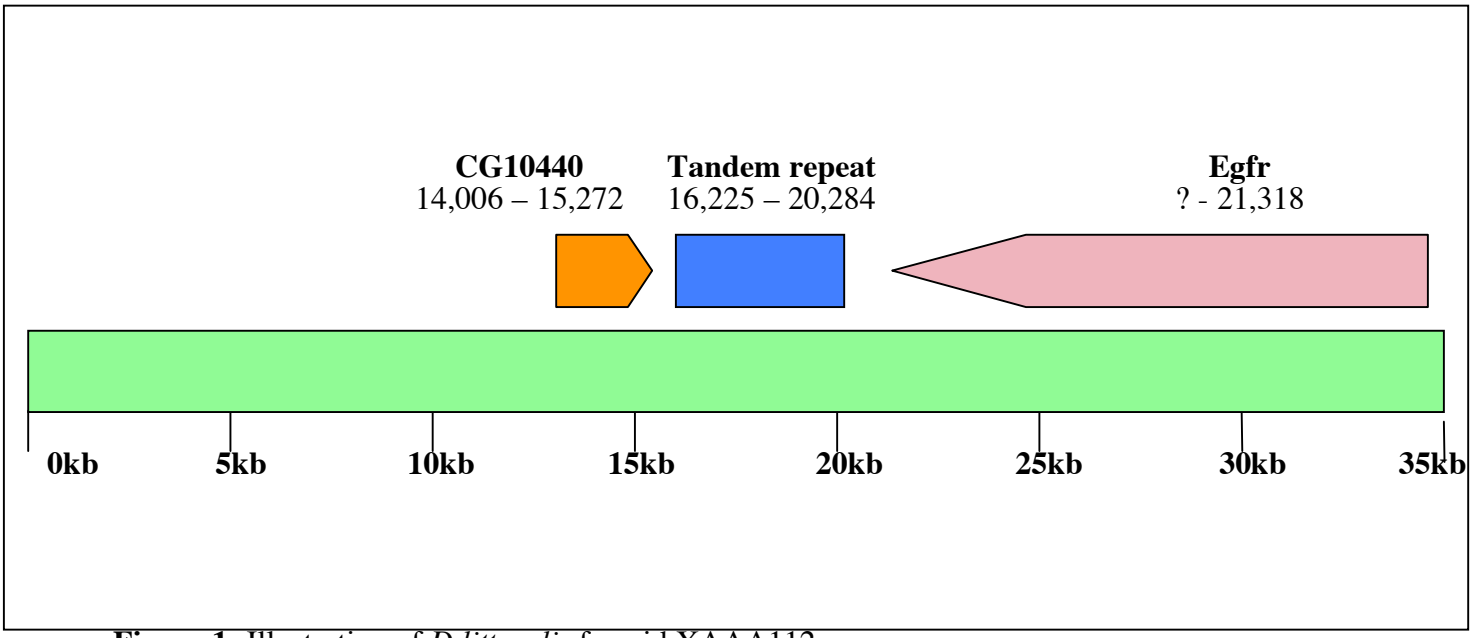


**Figure 1:** Illustration of *D.littoralis* fosmid XAAA112

| Feature | Location | Comments |
|---|---|---|
| Probable Gene | ?-21,318 | Homology to Egfr in D.melanogaster. First exon is off the end of fosmid **CG-10079** |
| Gene | 14,006 – 15, 272 | Homology to unnamed gene in D.melanogaster. **CG-10440** |
| Tandem Repeat | 16,225 – 20,284 | Not seen in *D.melanogaster* |

**Genes**

Fosmid XAAA112 appears to contain a gene with homology to an unnamed gene CG-10440 of *D.melanogaster*. This gene in *D.melanogaster* consists of 5 exons. Exons 1 and 2 are completely comprised of UTRs and do not show high identity to fosmid XAAA112. However the entire coding sequence appears to be represented on the fosmid as supported by Swisprot hits across the entire coding sequence and EST matches from the *D.melanogaster* database. CG-10440 has been sequenced in *D.melanogaster* and its amino acid sequence contains a BTB/POZ domain and a K+ channel tetramerisation domain[1]. The transcript length in D.Melanogaster has been reported to be 2348 base pairs in length, while the translation length is 338 residues[2]. Only one recorded allele has been identified in D.melanogaster and acknowledged as wild-type.

Table 1 below lists the predicted exon boundaries, as estimated by Ensembl software and blast2 evidence. Figure 2 below depicts an illustration of coding and noncoding regions of CG-10440 as well as the documented protein domains.

| Exon | Start | Stop |
|---|---|---|
| 1 | All UTR | |
| 2 | All UTR | |
| 3 | UTR….14006 | 14100 |
| 4 | 14184 | 14701 |
| 5 | 14830 | 15272…UTR |

* No significant similarity found to the UTR sequences

**Table 1:** Exon Boundaries for CG-10440 homolog

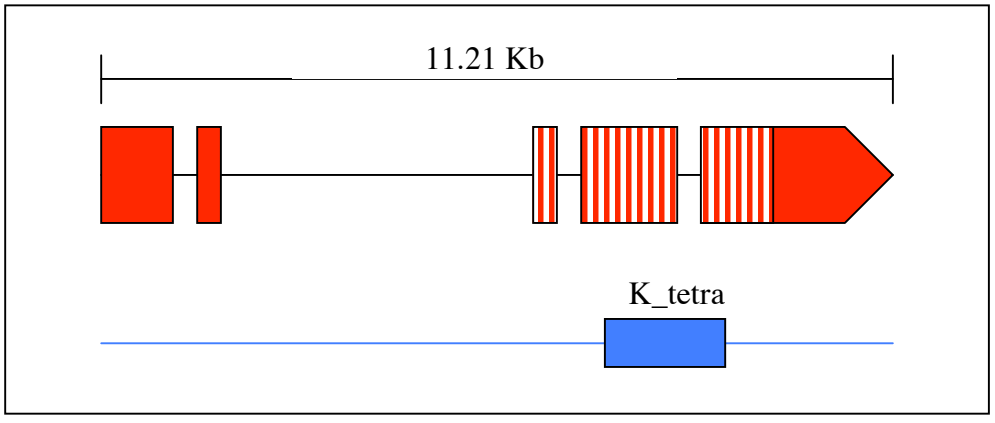


**Figure 2.** Depiction of CG-10440 in *D.Melanogaster*. Striped regions represent coding domains, solid regions represent UTRs.

Fosmid XAAA112 also appears to contain a gene with homology to an Epidermal Growth Factor Receptor gene CG-10079 of *D.melanogaster*. This gene in *D.melanogaster* also consists of 5 exons. Flybase has reported that this gene has two different splicing patterns in *D.melanogaster*, depicted Egfr-RA and Egfr-RB. Both of these splicing patterns show that exon 1 is considerably further upstream relative to the other four exons (see Figure 3). Due to this large inter-exon space, exon 1 does not appear to fall within fosmid XAAA112. However, there is substantial evidence that the other four exons do show identity to this fosmid. Since exon 1 does consist of some coding sequence, only a portion of the translated region appears to match the fosmid. CG10079-RA of *D.melanogaster* has a transcript length of 4563 bp and a translation length of 1377 residues[2]. Residues 95-1377 of the RA pattern show high match to my fosmid. CG10079-RB of *D.melanogaster* has a transcript length of 4648 bp and a translation length of 1426 residues[2]. Residues 54-1426 show high identity match to fosmid XAAA112 as reported by Swissprot and EST matches.

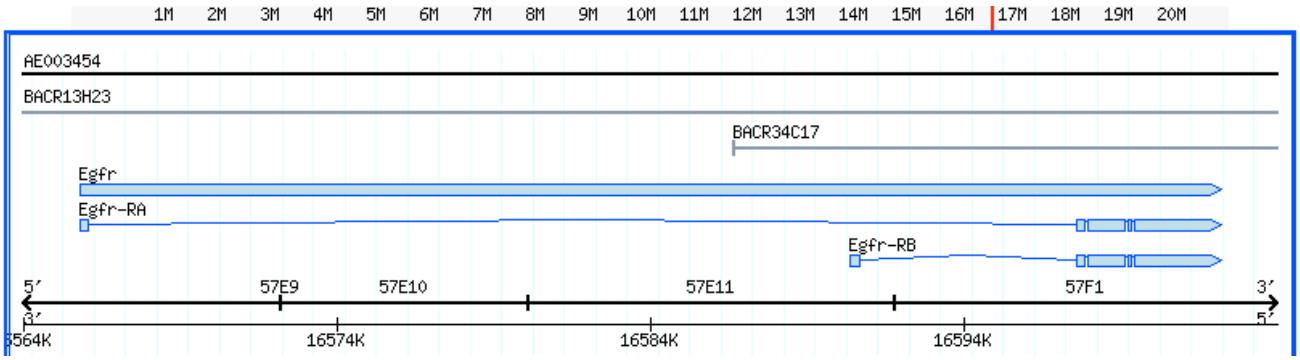


**Figure 3.** Splicing patterns of Egfr of *D.melanogaster*

CG-10079 has been sequenced in *D.melanogaster* and its protein product contains many common motifs such as an eukaryotic protein kinase, a tyrosine kinase catalytic domain, and a furin-like cysteine rich region[1]. The Egfr gene in *D.melanogaster* encodes a protein product with activity involved in eye morphogenesis and is expressed in the adult, the embryo, the larva, the ovary, and prepupa and pupa[1]. Flybase reports that there

are 125 recorded alleles: 20 in vitro constructs, 104 classical mutants, and 1 wild-type allele[1]. Research done on Egfr has reported that the gene is necessary for proper wing venation, development of the arista, legs, and female genital disc. Additionally research suggests that mutations in Egfr alter the distribution of macrochaetae[1].

Table 2 below lists the predicted exon boundaries, as predicted using Ensembl software and blast2 evidence. Figure 4 below depicts an illustration of coding and noncoding regions of Egfr as well as the protein domains.

| | | |
|---|---|---|
| Exon 1 | Off the end of contig | |
| Exon 2 | 25587 | 25370 |
| Exon 3 | 25284 | 24104 |
| Exon 4 | 23907 | 24039 |
| Exon 5 | 23841 | 21318 |

* No significant similarity found to the UTR sequences
**Table 2:** Exon Boundaries for CG-10079 (Egfr) homolog

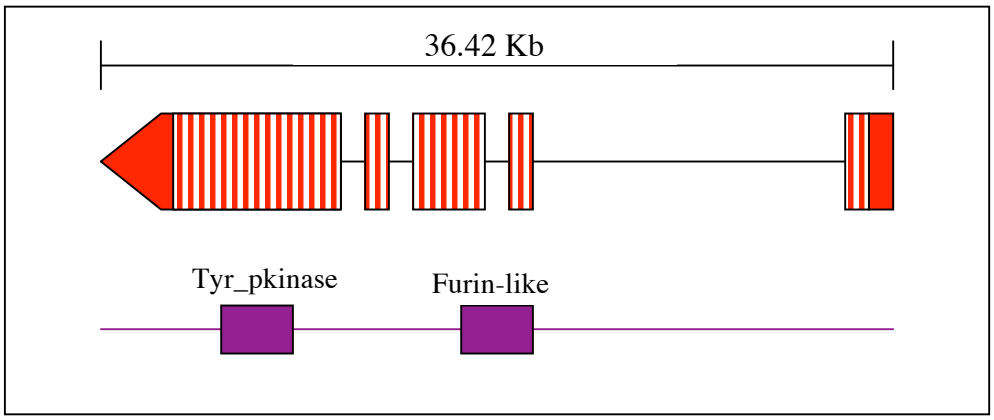


**Figure 4.** Depiction of CG-10079 in *D.Melanogaster*. Striped regions represent coding domains, solid regions represent UTRs.

In addition to the two genes that were documented above, GenScan software predicted the existence of two additional genes located upstream from the CG-10440 homolog. A Blast search using the entire XAAA112 fosmid against the *D.melanogaster* database and the Swissprot database showed to support for the existence of these two genes. Since GenScan predicted that these genes were located near 3kb and 8kb, I

extracted the first 14Kb of the fosmid and conducted a Blast search with this region against both the *D.melanogaster* database as well as the Swissprot database. No significant matches were reported. After examining the RepeatMasker output, I noticed that two small repeat regions were located near or within the predicted GenScan genes, and concluded that perhaps using the masked contig was to blame for the missing genes. However, even after extracting the first 14Kb of the unmasked fosmid and Blasting against *D.melanogaster* and Swissprot resulted in no significant hits. Therefore, I concluded that the two additional genes predicted by GenScan do not appear within fosmid XAAA112.

**Clustal Analysis**

Clustal Analysis was preformed using the Egfr gene. The first analysis was conducted comparing homologs from *D.littoralis, D.melanogaster,* and mosquito. The first analysis showed very high conservation between the species, therefore, it was concluded that the Egfr gene product is a very slowly evolving protein. In order to address the pattern, a second analysis was completed that included homologs from *C.elgans* and mouse. Again high levels of conservation were noted, suggesting that Egfr has significant importance in many species and has a relatively old evolutionary past.

Two regions exhibited very high levels of conservation across all the species; they occurred in the areas spanning from 270-470 and 915-1320. Interestingly these sections correspond to the Furin-like domain and the tyrosine-kinase domain respectively. An Ensembl protein report illustrates that the Furin-like region is located from 206-355 and the tyrosine-kinase domain spans from 967-1138 in *D.melanogaster*[2]. This evidence lends support to the idea that these two domains are necessary for proper functioning of the protein and thus show extremely high levels of conservation (Figure 5).

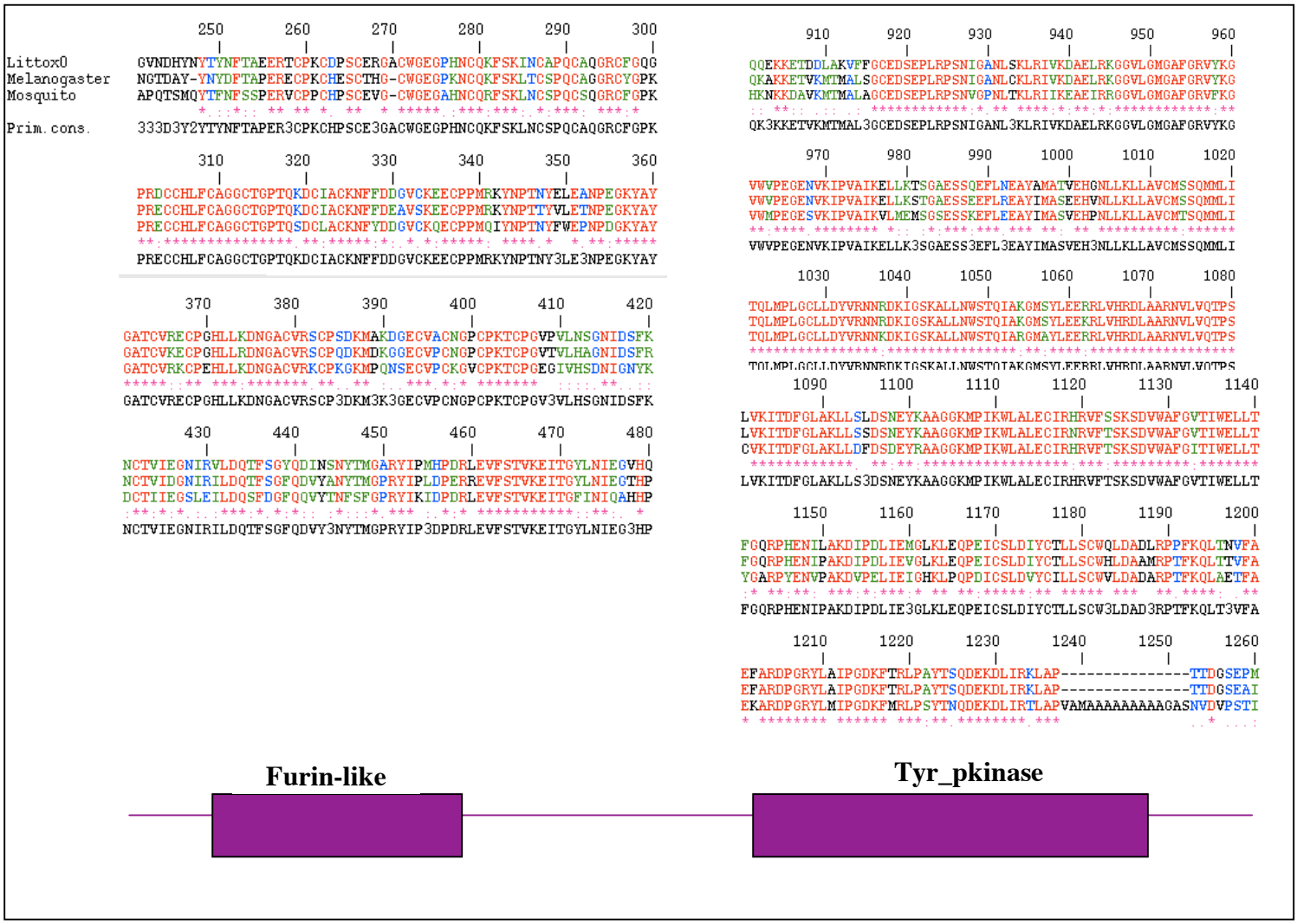


**Figure 5.** Clustal Analysis output comparing Egfr homologs from *D.littoralis, D.melanogaster*, and mosquito. Areas of particularly high levels of conservation correspond to Furin-like protein domain and a tyrosine kinase domain.

## Repeats

RepeatMasker software identified four repeat regions within fosmid XAAA112, these areas are listed in Table 3. Since this fosmid is most likely located within the 2R chromosome, this relatively low level of repeat DNA is not unusual. RepeatMasker did not identify the large tandem repeat located near 16-20kb. After extracting this region from the fosmid, I blasted the tandem repeat section against the *D.melanogaster* database.

No significant similarity was found, therefore I concluded that the tandem repeat is not found within the *D.melanogaster* genome. Furthermore, since two identified genes, CG10440 and CG10079, flank both ends of the repeat I compared the size of the region between these genes in both the *D.littoralis* and *D.melanogaster* genome. Interestingly, this region in *D.melanogaster* is only 819 nucleotides in length, while the homologous region in *D.littoralis* is over 6 kb and contains the tandem repeat. This finding lends further support to the conclusion that the tandem repeat region is unique to *D.littoralis.*

The percentage of repetitive DNA as calculated using both the RepeatMasker identified repeats and the unique tandem repeat was found to be 12.73% (calculation: 4431 / 34,783). Note that this number estimated the tandem repeat to be 4059 nucleotides in length. This size may need to be adjusted when finishing has been verified by the Genome Sequencing Center.

| Repeat | Repeat class/family | Position begin | Position End |
|---|---|---|---|
| ROO_I | LTR/Pao | 3262 | 3328 |
| ROO_I | LTR/Pao | 7337 | 7473 |
| GYPSY3_I | LTR/Gypsy | 30775 | 30854 |
| BS | LINE/Jockey | 33513 | 33604 |

**Table 3.** Repeats identified by RepeatMasker

**Synteny**

Fosmid XAAA112 shows high synteny to the *D.melanogaster* 2R chromosome. Both genes identified on the *D.littoralis* fosmid appear on the *D.melanogaster* chromosome in the same orientation. In addition, the blast searches using the first 14kb of the fosmid and the tandem repeat region showed highest identity to the a location near 16M of *D.melanogaster*. An illustration comparing sequence from the two species is shown in Figure 6.

A few additional genes: CG30222, CG33225, and CG10080, are located just downstream from the unnamed gene CG10440 on *D.melanogaster*. Due to the proximity of these additional genes to CG10440, it was reasonable to suggest that these genes might have homologs located in XAAA112. However, running a Blast2 search using this

fosmid and the protein sequence for these genes as reported by Ensembl revealed no evidence of the presence of these genes on the *D.littoralis* fosmid. Perhaps the distance between CG10440 and these additional genes is too vast, or perhaps the sequence of these genes has drifted too much and is undetectable by current Blast parameters.

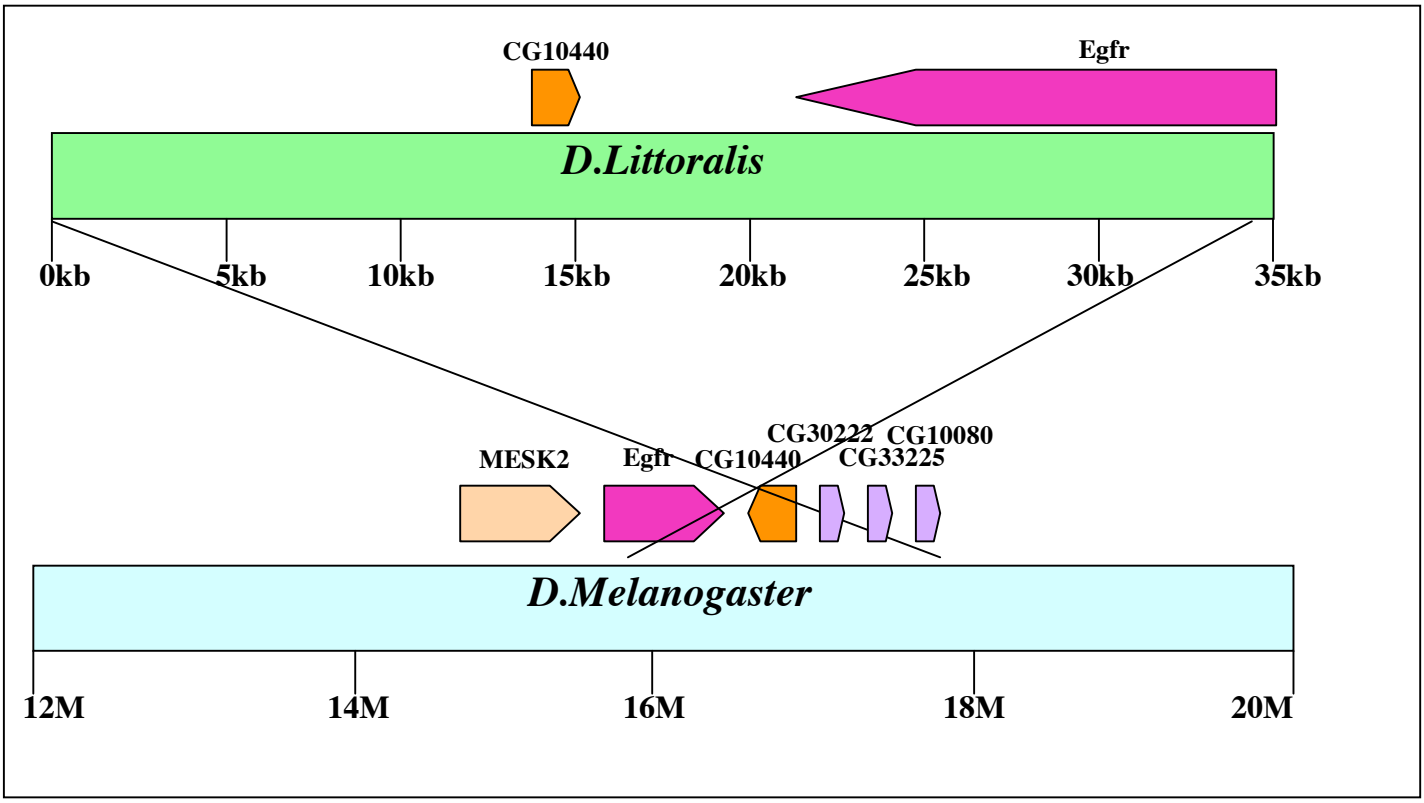


**Figure 6.** Illustration of *D.littoralis* fosmid XAAA112 in comparison to *D.melanogaster* chromosome 2R

## Appendix

Protein sequence for Egfr

>MMIISMWMSISRGLWDSSSIWSVLLILACMASITTSSSVSNAGYVDNGNMKVCIGTKSRLSVPSN
KEHHYRNLRDRYTNCTYVDGNLELTWLPNENLDLSFLDNIREVTGYILISHVDVKKVVFPKLQIIR
GRTLFSLSVEEEKYALFVTYSKMYTLEIPDLRDVLNGQVGFHNNYNLCHMRTIQWSEIVSNGTDA
YYNYDFTAPERECPKCHESCTHGCWGEGPKNCQKFSKLTCSPQCAGGRCYGPKPRECCHLFCAG
GCTGPTQKDCIACKNFFDEGVCKEECPPMRKYNPTTYVLETNPEGKYAYGATCVKECPGHLLRD
NGACVRSCPQDKMDKGGECVPCNGPCPKTCPGVTVLHAGNIDSFRNCTVIDGNIRILDQTFSGFQD
VYANYTMGPRYIPLDPERLEVFSTVKEITGYLNIEGTHPQFRNLSYFRNLETIHGRQLMESMFAAL
AIVKSSLYSLEMRNLKQISSGSVVIQHNRDLCYVSNIRWPAIQKEPEQKVWVNENLRADLCEKNGT
ICSDQCNEDGCWGAGTDQCLTCKNFNFNGTCIADCGYISNAYKFDNRTCKICHPECRTCNGAGAD
HCQECVHVRDGQHCVSECPKNKYNDRGVCRECHATCDGCTGPKDTIGIGACTTCNLAIINNDATV
KRCLLKDDKCPDGYFWEYVHPQEQGSLKPLAGRAVCRKCHPLCELCTNYGYHEQVCSKCTHYK
RREQCETECPADHYTDEEQRECFQCHPECNGCTGPGADDCKSCRNFKLFDANETGPYVNSTMFNC
TSKCPLEMRHVNYQYTAIGPYCAASPPRSSKITANLDVNMIFIITGAVLVPTICILCVVTYICRQKQK
AKKETVKMTMALSGCEDSEPLRPSNIGANLCKLRIVKDAELRKGGVLGMGAFGRVYKGVWVPE
GENVKIPVAIKELLKSTGAESSEEFLREAYIMASVEHVNLLKLLAVCMSSQMMLITQLMPLGCLLD
YVRNNRDKIGSKALLNWSTQIAKGMSYLEEKRLVHRDLAARNVLVQTPSLVKITDFGLAKLLSSD
SNEYKAAGGKMPIKWLALECIRNRVFTSKSDVWAFGVTIWELLTFGQRPHENIPAKDIPDLIEVGL
KLEQPEICSLDIYCTLLSCWHLDAAMRPTFKQLTTVFAEFARDPGRYLAIPGDKFTRLPAYTSQDE
KDLIRKLAPTTDGSEAIAEPDDYLQPKAAPGPSHRTDCTDEIPKLNRYCKDPSNKNSSTGDDETDSS
AREVGVGNLRLDLPVDEDDYLMPTCQPGPNNNNNINNPNQNNMAAVGVAAGYMDLIGVPVSVD
NPEYLLNAQTLGVGESPIPTQTIGIPVMGVPGTMEVKVPMPGSEPTSSDHEYYNDTQRELQPLHRN
RNTETRV

Nucleotide sequence for Egfr CG10079: (lowercase=up/downstream sequence; purple=noncoding regions; black=coding sequence) arranged by exon

gactaggcaccacaacgccacccccccccccctcaactcaggcaaagca

1
TCTCGAAGTAAACAAACAATTCTGCCGAGATTCGCGTGTTTGTACAGTTGGTGAATATAAGCA
AATCGACCGCGCATCGATATCATGATGATTATCAGCATGTGGATGAGCATATCGCGAGGATTG
TGGGACAGCAGCTCCATCTGGTCGGTGCTGCTGATCCTCGCCTGCATGGCATCCATCACCACA
AGCTCATCGGTCAGCAATGCCGGCTATGTGGATAATGGCAATATGAAAG

2
TCTGCATCGGCACTAAATCTCGGCTCTCCGTGCCCTCCAACAAGGAACATCATTACCGGAACC
TCAGAGATCGGTACACGAACTGTACGTATGTGGATGGCAACCTGGAGCTGACCTGGCTGCCCA
ACGAGAATTTGGACCTCAGCTTCCTGGACAACATACGGGAGGTCACCGGCTATATTCTGATCA
GTCATGTGGACGTTAAGAAAGTGGTATTTCCCAA

3
ACTACAAATCATTCGCGGACGCACGCTGTTCAGCTTATCCGTGGAGGAGGAGAAGTATGCCTT
GTTCGTCACTTATTCCAAAATGTACACGCTGGAGATTCCCGATCTACGCGATGTCTTAAATGG
CCAAGTGGGCTTCCACAACAACTACAATCTCTGCCACATGCGAACGATCCAGTGGTCGGAGAT
TGTATCCAACGGCACGGATGCATACTACAACTACGACTTTACTGCTCCGGAGCGCGAGTGTCC
CAAGTGCCACGAGAGCTGCACGCACGGATGTTGGGGCGAGGGTCCCAAGAATTGCCAGAAGT
TCAGCAAGCTCACCTGCTCGCCACAGTGTGCCGGAGGTCGTTGCTATGGACCAAAGCCGCGGG
AGTGTTGTCACCTCTTCTGCGCCGGAGGATGCACTGGTCCCACGCAAAAGGATTGCATCGCCT
GCAAGAACTTCTTCGACGAGGGCGTATGCAAGGAGGAATGCCCGCCCATGCGCAAGTACAAT
CCCACCACCTATGTTCTTGAAACGAATCCTGAGGGAAAGTATGCCTATGGTGCCACCTGCGTC

AAGGAGTGTCCCGGTCATCTGTTGCGTGATAATGGCGCCTGCGTGCGCAGCTGTCCCCAGGAC
AAGATGGACAAGGGGGGCGAGTGTGTGCCCTGCAATGGACCGTGCCCCAAAACCTGCCCGGG
CGTTACTGTCCTGCATGCCGGCAACATTGACTCGTTCCGGAATTGTACGGTGATCGATGGCAA
CATTCGCATTTTGGATCAGACCTTCTCGGGCTTCCAGGATGTCTATGCCAACTACACGATGGG
ACCACGATACATACCGCTGGATCCCGAGCGACTGGAGGTGTTCTCCACGGTGAAGGAGATCA
CCGGGTATCTGAATATCGAGGGAACCCACCCGCAGTTCCGGAATCTGTCGTACTTCCGCAATC
TGGAAACAATTCATGGCCGCCAGCTGATGGAGAGCATGTTTGCCGCTTTGGCGATCGTTAAGT
CATCCCTGTACAGCCTGGAGATGCGCAATCTGAAGCAGATTAGTTCCGGCAGTGTGGTCATCC
AGCATAATAGAGACCTCTGCTACGTAAGCAATATCCGTTGGCCGGCCATTCAGAAGGAGCCC
GAACAGAAGGTGTGGGTCAACGAGAATCTCAGGGCGGATCTATGCG

4
AGAAAAATGGAACCATTTGCTCGGATCAGTGCAACGAGGACGGCTGCTGGGGAGCTGGCACG
GATCAGTGCCTTACCTGCAAGAACTTCAATTTCAATGGCACCTGCATCGCCGACTGTGGTTAT
ATATCCAA

5
TGCCTACAAGTTTGACAATAGAACGTGCAAGATATGCCATCCAGAGTGCCGGACTTGCAATGG
AGCTGGAGCAGATCACTGCCAGGAGTGCGTCCATGTGAGGGACGGTCAGCACTGTGTGTCCG
AGTGCCCGAAGAACAAGTACAACGATCGTGGTGTCTGCCGAGAGTGCCACGCCACCTGCGAT
GGATGCACTGGGCCCAAGGACACCATCGGCATTGGAGCGTGTACAACGTGCAATTTGGCCATT
ATCAACAATGACGCCACAGTAAAACGCTGCCTGCTGAAGGAGACAAGTGCCCCGATGGGTAC
TTCTGGGAGTATGTGCATCCACAAGAGCAGGGATCGCTAAAGCCATTGGCCGGCAGAGCAGT
TTGCCGAAAGTGCCATCCCCTTTGCGAGCTGTGCACCAACTACGGATACCATGAACAGGTGTG
CTCCAAGTGCACCCACTACAAGCGACGAGAGCAGTGCGAGACCGAGTGTCCGGCCGATCACT
ACACGGATGAGGAGCAGCGCGAGTGCTTCCAGTGCCACCCAGAATGCAACGGTTGCACTGGT
CCGGGTGCCGACGATTGCAAGTCTTGTCGCAACTTCAAGTTGTTCGACGCGAATGAGACGGGT
CCCTATGTGAACTCCACGATGTTCAATTGCACCTCGAAGTGTCCCTTGGAGATGCGACATGTG
AACTATCAGTACACGGCCATTGGACCCTACTGTGCAGCTAGTCCGCCGAGGAGCAGCAAGAT
AACTGCCAATCTGGATGTGAACATGATCTTCATTATCACTGGTGCTGTTCTGGTGCCGACGATC
TGCATCCTCTGCGTGGTCACATACATTTGTCGGCAAAAGCAAAAGGCCAAGAAAGAAACAGT
GAAGATGACCATGGCTCTGTCCGGCTGTGAGGATTCCGAGCCGCTGCGTCCCTCGAACATTGG
AGCCAATCTATGCAAGTTGCGCATTGTCAAGGACGCCGAGTTGCGCAAGGGCGGAGTCCTCG
GAATGGGAGCCTTTGGACGAGTGTACAAGGGCGTTTGGGTGCCGGAGGGTGAGAACGTCAAG
ATTCCAGTGGCCATTAAGGAGCTGCTCAAGTCCACAGGCGCCGAGTCAAGCGAAGAGTTCCTC
CGCGAAGCCTACATCATGGCCTCTGTGGAGCACGTTAATCTGCTGAAGCTCCTGGCCGTCTGC
ATGTCCTCACAAATGATGCTAATCACGCAACTGATGCCGCTTGGCTGCCTGTTGGACTATGTG
CGAAATAACCGGGACAAGATCGGCTCTAAGGCTCTGCTCAACTGGAGCACGCAAATCGCCAA
GGGCATGTCGTATCTGGAGGAGAAGCGACTGGTCCACAGAGACTTGGCTGCCCGCAATGTCCT
GGTGCAGACTCCCTCGCTGGTGAAGATCACCGACTTTGGGCTGGCCAAGTTGCTGAGCAGCGA
TTCCAATGAGTACAAGGCTGCTGGCGGCAAGATGCCCATCAAGTGGTTGGCACTGGAGTGCAT
TCGCAATCGTGTATTCACCAGCAAGTCCGATGTCTGGGCCTTTGGTGTGACAATTTGGGAACT
GCTGACCTTTGGCCAGCGTCCACACGAGAACATCCCCGCTAAGGATATTCCCGATCTTATTGA
AGTCGGTCTGAAGCTGGAGCAGCCGGAGATTTGTTCGCTGGACATTTACTGCACACTTCTCTC
GTGCTGGCACTTGGATGCCGCCATGCGTCCAACCTTCAAGCAGCTGACTACGGTCTTTGCTGA
GTTCGCCAGAGATCCGGGTCGCTATCTGGCCATTCCCGGGGATAAGTTCACCCGGCTGCCGGC
CTACACGAGTCAGGATGAGAAGGATCTCATCCGAAAATTGGCTCCCACCACCGATGGGTCCG
AAGCCATTGCGGAACCCGATGACTACCTGCAACCCAAGGCAGCACCTGGTCCTAGTCACAGA
ACCGACTGCACGGATGAGATACCCAAGCTGAACCGCTACTGCAAGGATCCTAGCAACAAGAA
TTCGAGTACCGGAGACGATGAGACGGATTCGAGTGCCCGGGAAGTGGGCGTGGGTAATCTGC
GCCTCGATCTACCAGTCGATGAGGATGATTACCTGATGCCCACATGCCAACCGGGGCCCAACA
ACAACAACAACATAAATAATCCCAATCAAAACAATATGGCAGCTGTGGGCGTGGCTGCCGGC
TACATGGATCTCATCGGAGTGCCCGTTAGTGTGGACAATCCGGAGTATCTGCTAAACGCGCAG
ACACTGGGTGTTGGGGAGTCGCCGATACCCACCCAGACCATCGGGATACCGGTGATGGGAGT
CCCGGGCACCATGGAGGTCAAGGTGCCAATGCCAGGCAGTGAGCCAACGAGCTCCGATCACG
AGTACTACAATGATACCCAACGGGAGTTGCAGCCACTGCATCGAAACCGCAACACGGAGACG

AGGGTGTAGGCTCCAGTCGAGTAGGAGCAATTGCCAATGAAGAAGGAGAATCTTGCCAAGTG
CCTTTGGAAGCCATGCGGATATGCCTTTGCTGGCTGTTATCTTAGATAGGAGCCCGTGCGCCT
GTACAGAGCCTAGAACGACCCATAGATTTGTAAATTACTCTCTAGTGTACTTAGCCAGTCCCC
CTGCATCTTTGTGTATACTTCTCTATCCTAGCCCGTAAACCAGTAGTCAGCAGGGATCGTCGTC
GCGGTCCTGCGCTTGTAGAATCCATGTAATTACGAGAACAAAACACCGATAGTGCATTTCTTA
GACGCCTCCGCCATGCCAAATTCGAAGAGATCCCGGATC

cggatcggctccgccggagggcacacactaagctaagccgatgtccactt

Protein sequence from CG10440

>MDRERERDAKGLEPRDLSSTGRIYARSDIKISASPTVSPTISNSSSPTPTPPASSSVTPLGLPGAAAA
AAVAAAAAAAAGAGPGGGGASAGASSYLHNHKPISGIPCVAAASRYTAPVHIDVGGTIYTSSLET
LTKYPESKLAKLFNGQIPIVLDSLKQHYFIDRDGGMFRHILNFMRNSRLLIADDFPDLELLLEEARY
YEVEPMVKQLESMRKERVRNGNYLVAPPTPPARHIKTSPRTSGAPECNYEVVALHISPDLGERIML
SAERALLDELFPEANQATQSSRSGVSWNQGDWRQIIRFPLNGYCKLNSVQVLTRLLNAGFTIEAST
GGGVEGQQFSEYLLVRRVPM

Nucleotide sequence for CG10440 (lowercase=up/downstream sequence; purple=noncoding regions; black=coding sequence) arranged by exon

>tggccattcaccattcgccattcgccatccatcagccagcgagctaaatc

1
AGTATTCGAGCAACTCTCGACGCGAATGCAAGTGAGGTAAATTGCCAGCGCTCAAAAGTAGG
CAACACTTAGGTCTAACGGTGAACGGTGAACAGCGAACGGTTAACGTTTAACGGTGTGCGTTG
TGTACGGCAAGCGGAAACGGAAGTAAACACTCGTTGGCGCGCGCTCATTTGATTGAATTTGAA
AGGGTTCAACCGAAAGCGACTAGCGATCAGCAGATCGAAGCACACTGAGAATATCGAAGGTC
ACTGAAGATCGGACAAGACGAATATCGTTTATCTACGCGCGCGTTGTATTAAAAAAGAGTTAA
AGTGCATTTCCTAAATATACCAACCTTTTT

2
AACTCCGTTTGTTGCATTGGATCAACGATCACCGTTGACGGTGGATTCCTCCTGCCCCAAACG

3
GAAAAAATGCGATATGGACCGAGAGCGCGAAAGAGATGTCAAGGCCCTGGAGCCACGCGAC
TTATCCTCCACGGGCCGCATTTACGCCAGAAGCGATATTAAAATCAG

4
CTCCTCGCCAACCGTCTCGCCAACGATCTCAAACTCCTCTTCGCCCACACCGACGCCACCCGC
CAGTTCCTCAGTGACGCCGCTGGGCTTGCCCGGAGCGGTGGCAGCTGCTGCCGCCGCCGTCGG
AGGAGCCTCATCGGCGGGGGCCAGTTCCTATTTGCACGGCAACCACAAGCCCATCACCGGGA
TTCCGTGTGTGGCCGCCGCCTCCCGCTATACGGCGCCCGTTCACATCGACGTGGGCGGGACCA
TCTACACCAGCTCACTGGAGACGCTCACCAAGTACCCGGATCGAAGCTGGCCAAGCTCTTCAA
TGGCCAGATACCCATCGTGCTGGACTCGCTGAAGCAGCACTATTTCATCGATCGCGACGGGGG
CATGTTCCGCCACATTCTGAACTTTATGAGAAACTCAAGACTTCTGATTGCCGAGGATTTTCCC
GACCTGGAGCTGCTGCTGGAGGAGGCTCGCTACTACGAAGTGGAGC

5
CGATGATTAAGCAGCTGGAAAGCATGCGCAAGGATCGCGTTCGCAATGGCAACTATTTGGTG
GCGCCACCCACTCCACCGGCTCGCCACATCAAGACGAGCCCGAGGACCAGCGCATCGCCGGA
GTGCAATTACGAGGTAGTGGCGCTGCACATCTCGCCAGACCTTGGCGAACGCATAATGCTGTC
GGCAGAACGGGCTCTGCTCGACGAGCTCTTCCCGGAGGCCAGCCAAGCGACACAGTCGAGTC
GCAGCGGAGTGTCCTGGAACCAGGGCGACTGGGGGCAAATCATTCGCTTCCCCCTCAACGGCT

ACTGCAAGCTGAACTCGGTGCAGGTGCTCACCCGACTCCTCAACGCCG
GCTTCACCATCGAGGCCAGCGTTGGGGGCCAGCAGTTCTCCGAGTATCTGCTGGCGCGACGCG
TTCCAATGTGATAGGCTCTGGGTAGGGATTCCCCAGCTCCCCAAGAAGCTGCAGATGTGGCCC
TGATTCCGTCGGCTTCTGTGGCTTCTGTTTCCGGCTCCGGTTCGGACTCGGGTTCGGGTTCGGC
ATCGGGCTCTTTGCGACGCGGGCGTCTGATCCATGTGCGTTGCAGGCGCTAATTGATCTGGGT
TAGGTGACCCAGAATCGCAGGTGGGGAGAGTGGTCTATGAACAATTCCAAGTGATACAATGT
ACTGAGAAGGGCAAATGAAAAATCAAGTGTAACGCCGAATTGTTTATATATCTAACACACTA
AAACTCGATCCAACACCGTTTTGGGGACTGCGCATTGCTGCTC
TGAGCAGTAAAATACATAATTTAGAAATACTTGGACACCGTTCGGTGAGGATGCAAGTTCTGA
ATCTGAATCCAAATGTTATTTAATGTTTATCAATTAAGTTTGCTGGCACGGACACTGAACGCA
AACACGCTCGTGTGTATTATTAAGCATTTACAGAGCCAAGAGCCCAGCACTGATTTTGTTTTTT
AGCCTTGATTTTAAGTCGATATTGTTTTCTCTAAGTGTACCACCTAATGCTAAAGACCCTAAGT
TGTTTCATTAAATCAAGTTGGCCAGGGATATTAGCTGGCTTAAATTATTGTAACACACACGGA
AAACTATGGTAGTATCAACTCACAGCGAAAATAGTTTTGTAGTACTTAATTGATGTAATTGAA
CCTTTATGAATGTAAACTATGGGATAACTTTTTCAATATA
TACGATCCGACGTCTAAATAGTCTACACCAATTCAAATACAATTACATGCAAATATAGATATA
TTTAACGAAATACAGTTTGCGACCACGCCCTCTATAGAACAACCACTCAATTGAATATTTAAA
TAAAGAAAAAACATACGTGT

aaactgttgaatttcctctttggatgcttaatttgaaatttgtattcttg

Clustal: Search for Egfr in multiple species

>*D.Littoralis.*
XSSAAATVEDKNKGQEFVRKFLRFDLPAAALLINVRRVGTQLQLCLSTHGTLLRPTRRAA
YHDPYRDPLCIGTKSRLSVPSNRDHHYRNLRDRYTNCTYVDGNLELTWLQDPNLDLSFLA
NIREVTGYILISHVDVTKVIFPKLQIIRGRTLFSLTVDEAKYALFAAYSKMNTLEMPELR
DILQGWVGFFSNYNLCHTRSIVWREILSGVNDHYNYTYNFTAEERTCPKCDPSCERGACW
GEGPHNCQKFSKINCAPQCAQGRCFGQGPRDCCHLFCAGGCTGPTQKDCIACKNFFDDGV
CKEECPPMRKYNPTNYELEANPEGKYAYGATCVRECPGHLLKDNGACVRSCPSDKMAKDG
ECVACNGPCPKTCPGVPVLNSGNIDSFKNCTVIEGNIRVLDQTFSGYQDINSNYTMGARY
IPMHPDRLEVFSTVKEITGYLNIEGVHQHFKNLSYFRNLEVIHGRQLMESFFAALSIVKS
SLSSLELRNLKRINSGSIVIQHNEKLCYVSNIRWSTVQKSNDQMQYINENLNTSECRKAR
QVCSDQCNSDGCWGAGPDQCLNCKSFNYNGTCIADCRNVTKTYQFDAKTCKKCHPECRTC
SGPGAEHCDECVHVRDDQHCVTVCPENKYSDFGICRRCHETCEGCTGPNNTIGHGACNTC
NLAIINADATVERCLRKDDKCPDGYYWEYVHPQEQGSLKPLAGKAICRKCHPRCELCTNY
GFHEQVCSKCAGYKRREQCEDECPADHYADEEKRECFECHPECKGCTGPGSEDCLACRNF
KLFAESEVYDNSTLFNCTSKCPPELPHVNYQSHFIGPHCASSPPRGSKLTAGLNVNMIII
IFVAVFIPVICVLCVIIYICRQKQQEKKETDDLAKVFFGCEDSEPLRPSNIGANLSKLRI
VKDAELRKGGVLGMGAFGRVYKGVWVPEGENVKIPVAIKELLKTSGAESSQEFLNEAYAM
ATVEHGNLLKLLAVCMSSQMMLITQLMPLGCLLDYVRNNRDKIGSKALLNWSTQIAKGMS
YLEERRLVHRDLAARNVLVQTPSLVKITDFGLAKLLSLDSNEYKAAGGKMPIKWLALECI
RHRVFSSKSDVWAFGVTIWELLTFGQRPHENILAKDIPDLIEMGLKLEQPEICSLDIYCT
LLSCWQLDADLRPPFKQLTNVFAEFARDPGRYLAIPGDKFTRLPAYTSQDEKDLIRKLAP
TTDGSEPMVEADDYLQPKAAPGPSHRTDGTDEIPKLNRYCKDPSNKNSNSGGIGGDDETD
SNAREVGVGNLRLDLPVDEDDYLMPTGQGNPNGNNNNNNNNSNNPNNNMATAAATGYIDV
IGVSKRWEYIKQANTNPILALSHNQVPVSVDNPEYLLNALSAGEAPMPTQTIGIPVAGVP
GTMEVKMPGSESTSSDHEYYNDTQRELQPLQLLRSRSTETRV

>*D.Melanogaster*
MLLRRRNGPCPFPLLLLLLAHCICIWPASAARDRYARQNNRQRHQDIDRDRDRDRFLYRS
SSAQNRQRGGANFALGLGANGVTIPTSLEDKNKNEFVKGKICIGTKSRLSVPSNKEHHYR

NLRDRYTNCTYVDGNLKLTWLPNENLDLSFLDNIREVTGYILISHVDVKKVVFPKLQIIR
GRTLFSLSVEEEKYALFVTYSKMYTLEIPDLRDVLNGQVGFHNNYNLCHMRTIQWSEIVS
NGTDAYYNYDFTAPERECPKCHESCTHGCWGEGPKNCQKFSKLTCSPQCAGGRCYGPKPR
ECCHLFCAGGCTGPTQKDCIACKNFFDEAVSKEECPPMRKYNPTTYVLETNPEGKYAYGA
TCVKECPGHLLRDNGACVRSCPQDKMDKGGECVPCNGPCPKTCPGVTVLHAGNIDSFRNC
TVIDGNIRILDQTFSGFQDVYANYTMGPRYIPLDPERREVFSTVKEITGYLNIEGTHPQF
RNLSYFRNLETIHGRQLMESMFAALAIVKSSLYSLEMRNLKQISSGSVVIQHNRDLCYVS
NIRWPAIQKEPEQKVWVNENLRADLCEKNGTICSDQCNEDGCWGAGTDQCLTCKNFNFNG
TCIADCGYISNAYKFDNRTCKICHPECRTCNGAGADHCQECVHVRDGQHCVSECPKNKYN
DRGVCRECHATCDGCTGPKDTIGIGACTTCNLAIINNDATVKRCLLKDDKCPDGYFWEYV
HPQEQGSLKPLAGRAVCRKCHPLCELCTNYGYHEQVCSKCTHYKRREQCETECPADHYTD
EEQRECFQRHPECNGCTGPGADDCKSCRNFKLFDANETGPYVNSTMFNCTSKCPLEMRHV
NYQYTAIGPYCAASPPRSSKITANLDVNMIFIITGAVLVPTICILCVVTYICRQKQKAKK
ETVKMTMALSGCEDSEPLRPSNIGANLCKLRIVKDAELRKGGVLGMGAFGRVYKGVWVPE
GENVKIPVAIKELLKSTGAESSEEFLREAYIMASEEHVNLLKLLAVCMSSQMMLITQLMP
LGCLLDYVRNNRDKIGSKALLNWSTQIAKGMSYLEEKRLVHRDLAARNVLVQTPSLVKIT
DFGLAKLLSSDSNEYKAAGGKMPIKWLALECIRNRVFTSKSDVWAFGVTIWELLTFGQRP
HENIPAKDIPDLIEVGLKLEQPEICSLDIYCTLLSCWHLDAAMRPTFKQLTTVFAEFARD
PGRYLAIPGDKFTRLPAYTSQDEKDLIRKLAPTTDGSEAIAKPDDYLQPKAAPGPSHRTD
CTDEMPKLNRYCKDPSNKNSSTGDDERDSSAREVGVGNLRLDLPVDEDDYLMPTCQPGPN
NNNNMNNPNQNNMAAVGVAAGYMDLIGVPVSVDNPEYLLNAQTLGVGESPIPTQTIGIPV
MGGPGTMEVKVPMPGSEPTSSDHEYYNDTQRELQPLHRNRNTETRV

>Mosquito
CIGTNGRMSVPANREYHYKNLRDRYTNCTYVDGNLEITWIQNITDLNFLQHIREVTGYVLISHIDLP
QVILPRLQIIRGRTTFKLNKWEEAYGLFVSFSHMNTLELPALRDILGGSVGFFNNYNLCHMKSINW
EEILSAPQTSMQYTFNFSSPERVCPPCHPSCEVGCWGEGAHNCQRFSKLNCSPQCSQGRCFGPKPR
ECCHLFCAGGCTGPTQSDCLACKNFYDDGVCKQECPPMQIYNPTNYFWEPNPDGKYAYGATCVR
KCPEHLLKDNGACVRKCPKGKMPQNSECVPCKGVCPKTCPGEGIVHSDNIGNYKDCTIIEGSLEIL
DQSFDGFQQVYTNFSFGPRYIKIDPDRLEVFSTVKEITGFINIQAHHPNFTTLNYFRNLEVVGGRQL
KENLFASVYIVKTSLKSLELKSLKRVNSGSIVILENSDLCFVEDIDWSEIKKSSDHEVMVQKNRNAT
ECHEEGMECSEQCSKAGCWGKGPEQCLECKNVKYKGKCLDSCKSLPRLYSVDSKTCGDCHQECK
DFCYGPNEDNCGSCMNVKDGRFCVACPTTKHAMNGTCINCHKTCVGCRGPRDTIAPDGCISCDK
AIIGSDAKIERCLMKDESCPDGYYSDYVLQEEGPLKQLSGKAVCRKCHPRCKKCTGYGFHEQFCQ
ECTGYKKGEQCEDECPQDFYANEETRICLPCHQECRGCHGLGDDHCDECRNLKLFEGDPYDNATT
FTCVSNCPASHPYKRFPQEAGKIGPYCSADSMQSGLRIEPQTQVKIVMGSVMALILLCVVFGIAFVL
FSRHKNKKDAVKMTMALAGCEDSEPLRPSNVGPNLTKLRIIKEAEIRRGGVLGMGAFGRVFKGV
WMPEGESVKIPVAIKVLMEMSGSESSKEFLEEAYIMASVEHPNLLKLLAVCMTSQMMLITQLMPL
GCLLDYVRNNKDKIGSKALLNWSTQIARGMAYLEERRLVHRDLAARNVLVQTPSCVKITDFGLA
KLLDFDSDEYRAAGGKMPIKWLALECIRHRVFTSKSDVWAFGITIWELLTYGARPYENVPAKDVP
ELIEIGHKLPQPDICSLDVYCILLSCWVLDADARPTFKQLAETFAEKARDPGRYLMIPGDKFMRLPS
YTNQDEKDLIRTLAPVAMAAAAAAAAAGASNVDVPSTIAETDEYLQPKTRPSIMLPGPSAVEPSDE
MPKSLRYCKDPLKPDDETDGHGKEVGVGGIRLNLPLDEDDYLMPTCQSQNQSTPGYMDLIGVPAS
VDNPEYLMGSTQAIAGLAQGHTPSLSSASGSIGPKSADQPGAAAGLHHQQPSSPPTQTIGIPLSPTET
EATSSEHEYYNDLQREL
IPLHRNETTV

>*C.elegans*
MRYPPSIGSILLIIPIFLTFFGNSNAQLWKRCVSPQDCLCSGTTNGISRYGTGNILEDLETMYRGCRR
VYGNLEITWIEANEIKKWRESTNSTVDPKNEDSPLKSINFFDNLEEIRGSLIIYRANIQKISFPRLRVIY
GDEVFHDNALYIHKNDKVHEVVMRELRVIRNGSVTIQDNPKMCYIGDKIDWKELLYDPDVQKVE
TTNSHQHCYQNGKSMAKCHESCNDKCWGSGDNDCQRVYRSVCPKSCSQCFYSNSTSSYECCDSA
CLGGCTGHGPKNCIACSKYELDGICIETCPSRKIFNHKTGRLVFNPDGRYQNGNHCVKECPPELLIE
NDVCVRHCSDGHHYDATKDVRECEKCRSSSCPKICTVDGHLTNETLKNLEGCEQIDGHLIIEHAFT
YEQLKVLETVKIVSEYITIVQQNFYDLKFLKNLQIIEGRKLHNVRWALAIYQCDDLEELSLNSLKLI

KTGAVLIMKNHRLCYVSKIDWSSIITSKGKDNKPSLAIAENRDSKLCETEQRVCDKNCNKRGCWG
KEPEDCLECKTWKSVGTCVEKCDTKGFLRNQTSMKCERCSPECETCNGLGELDCLTCRHKTLYNS
DFGNRMECVHDCPVSHFPTQKNVCEKCHPTCYDNGCTGPDSNLGYGGCKQCKYAVKYENDTIFC
LQSSGMNNVCVENDLPNYYISTYDTEGVIETHCEKCSISCKTCSSAGRNVVQNKCVCKHVEYQPN
PSERICMDQCPVNSFMVPDTNNTVCKKCHHECDQNYHCANGQSTGCQKCKNFTVFKGDIAQCVS
ECPKNLPFSNPANGECLDYDIASRQRKTRMVIIGSVLFGFAVMFLFILLVYWRCQRIGKKLKIAEM
VDMPELTPIDASVRPNMSRICLIPSSELQTKLDKKLGAGAFGTVFAGIYYPKRAKNVKIPVAIKVFQ
TDQSQTDEMLEEATNMFRLRHDNLLKIIGFCMHDDGLKIVTIYRPLGNLQNFLKLHKENLGAREQ
VLYCYQIASGMQYLEKQRVVHRDLATRNVLVKKFNHVEITDFGLSKILKHDADSITIKSGKVA
IKWLAIEIFSKHCYTHASDVWAFGVTCWEIITFGQSPYQGMSTDSIHNFLKDGNRLSQPPNCSQDY
QELLRCWMADPKSRPGFEILYERFKEFCKVPQLFLENSNKISESDLSAEERFQTERIREMFDGNIDP
QMYFDQGSLPSMPSSPTSMATFTIPHGDLMNRMQSVNSSRYKTEPFDYGSTAQEDNSYLIPKTKEV
QQSAVLYTAVTNEDGQTELSPSNGDYYNQPNTPSSSSGYYNEPHLKTKKPETSEEAEAVQYENEE
VSQKETCL

>Human
MRPSGTAGAALLALLAALCPASRALEEKKVCQGTSNKLTQLGTFEDHFLSLQRMFNNCEV
VLGNLEITYVQRNYDLSFLKTIQEVAGYVLIALNTVERIPLENLQIIRGNMYYENSYALA
VLSNYDANKTGLKELPMRNLQEILHGAVRFSNNPALCNVESIQWRDIVSSDFLSNMSMDF
QNHLGSCQKCDPSCPNGSCWGAGEENCQKLTKIICAQQCSGRCRGKSPSDCCHNQCAAGC
TGPRESDCLVCRKFRDEATCKDTCPPLMLYNPTTYQMDVNPEGKYSFGATCVKKCPRNYV
VTDHGSCVRACGADSYEMEEDGVRKCKKCEGPCRKVCNGIGIGEFKDSLSINATNIKHFK
NCTSISGDLHILPVAFRGDSFTHTPPLDPQELDILKTVKEITGFLLIQAWPENRTDLHAF
ENLEIIRGRTKQHGQFSLAVVSLNITSLGLRSLKEISDGDVIISGNKNLCYANTINWKKL
FGTSGQKTKIISNRGENSCKATGQVCHALCSPEGCWGPEPRDCVSCRNVSRGRECVDKCN
LLEGEPREFVENSECIQCHPECLPQAMNITCTGRGPDNCIQCAHYIDGPHCVKTCPAGVM
GENNTLVWKYADAGHVCHLCHPNCTYGCTGPGLEGCPTNGPKIPSIATGMVGALLLLLVV
ALGIGLFMRRRHIVRKRTLRRLLQERELVEPLTPSGEAPNQALLRILKETEFKKIKVLGS
GAFGTVYKGLWIPEGEKVKIPVAIKELREATSPKANKEILDEAYVMASVDNPHVCRLLGI
CLTSTVQLITQLMPFGCLLDYVREHKDNIGSQYLLNWCVQIAKGMNYLEDRRLVHRDLAA
RNVLVKTPQHVKITDFGLAKLLGAEEKEYHAEGGKVPIKWMALESILHRIYTHQSDVWSY
GVTVWELMTFGSKPYDGIPASEISSILEKGERLPQPPICTIDVYMIMVKCWMIDADSRPK
FRELIIEFSKMARDPQRYLVIQGDERMHLPSPTDSNFYRALMDEEDMDDVVDADEYLIPQ
QGFFSSPSTSRTPLLSSLSATSNNSTVACIDRNGLQSCPIKEDSFLQRYSSDPTGALTED
SIDDTFLPVPEYINQSVPKRPAGSVQNPVYHNQPLNPAPSRDPHYQDPHSTAVGNPEYLN
TVQPTCVNSTFDSPAHWAQKGSHQISLDNPDYQQDFFPKEAKPNGIFKGSTAENAEYLRV
APQSSEFIGA

>Mouse
MRPSGTARTTLLVLLTALCAAGGALEEKKVCQGTSNRLTQLGTFEDHFLSLQRMYNNCEV
VLGNLEITYVQRNYDLSFLKTIQEVAGYVLIALNTVERIPLENLQIIRGNALYENTYALA
ILSNYGTNRTGLRELPMRNLQEILIGAVRFSNNPILCNMDTIQWRDIVQNVFMSNMSMDL
QSHPSSCPKCDPSCPNGSCWGGGEENCQKLTKIICAQQCSHRCRGRSPSDCCHNQCAAGC
TGPRESDCLVCQKFQDEATCKDTCPPLMLYNPTTYQMDVNPEGKYSFGATCVKKCPRNYV
VTDHGSCVRACGPDYYEVEEDGIRKCKKCDGPCRKVCNGIGIGEFKDTLSINATNIKHFK
YCTAISGDLHILPVAFKGDSFTRTPPLDPRELEILKTVKEITGFLLIQAWPDNWTDLHAF
ENLEIIRGRTKQHGQFSLAVVGLNITSLGLRSLKEISDGDVIISGNRNLCYANTINWKKL
FGTPNQKTKIMNNRAEKDCKAVNHVCNPLCSSEGCWGPEPRDCVSCQNVSRGRECVEKCN
ILEGEPREFVENSECIQCHPECLPQAMNITCTGRGPDNCIQCAHYIDGPHCVKTCPAGIM
GENNTLVWKYADANNVCHLCHANCTYGCAGPGLQGCEVWPSGPKIPSIATGIVGGLLFIV
VVALGIGLFMRRRHIVRKRTLRRLLQERELVEPLTPSGEAPNQAHLRILKETEFKKIKVL
GSGAFGTVYKGLWIPEGEKVKIPVAIKELREATSPKANKEILDEAYVMASVDNPHVCRLL
GICLTSTVQLITQLMPYGCLLDYVREHKDNIGSQYLLNWCVQIAKGMNYLEDRRLVHRDL
AARNVLVKTPQHVKITDFGLAKLLGAEEKEYHAEGGKVPIKWMALESILHRIYTHQSDVW

SYGVTVWELMTFGSKPYDGIPASDISSILEKGERLPQPPICTIDVYMIMVKCWMIDADSR
PKFRELILEFSKMARDPQRYLVIQGDERMHLPSPTDSNFYRALMDEEDMEDVVDADEYLI
PQQGFFNSPSTSRTPLLSSLSATSNNSTVACINRNGSCRVKEDAFLQRYSSDPTGAVTED
NIDDAFLPVPEYVNQSVPKRPAGSVQNPVYHNQPLHPAPGRDLHYQNPHSNAVGNPEYLN
TAQPTCLSSGFNSPALWIQKGSHQMSLDNPDYQQDFFPKETKPNGIFKGPTAENAEYLRV
APPSSEFIGA

## References

1. Flybase [1993-1997]. The Genetics Society of America [May 1, 2004]. Retrieved from: http://flybase.bio.indiana.edu

**2.** Project Ensembl. [2002-2004] EMBL-EBI, The Wellcome Trust, and Sanger Institute. [May 1, 2004]. Retrieved from http://ensemble.org