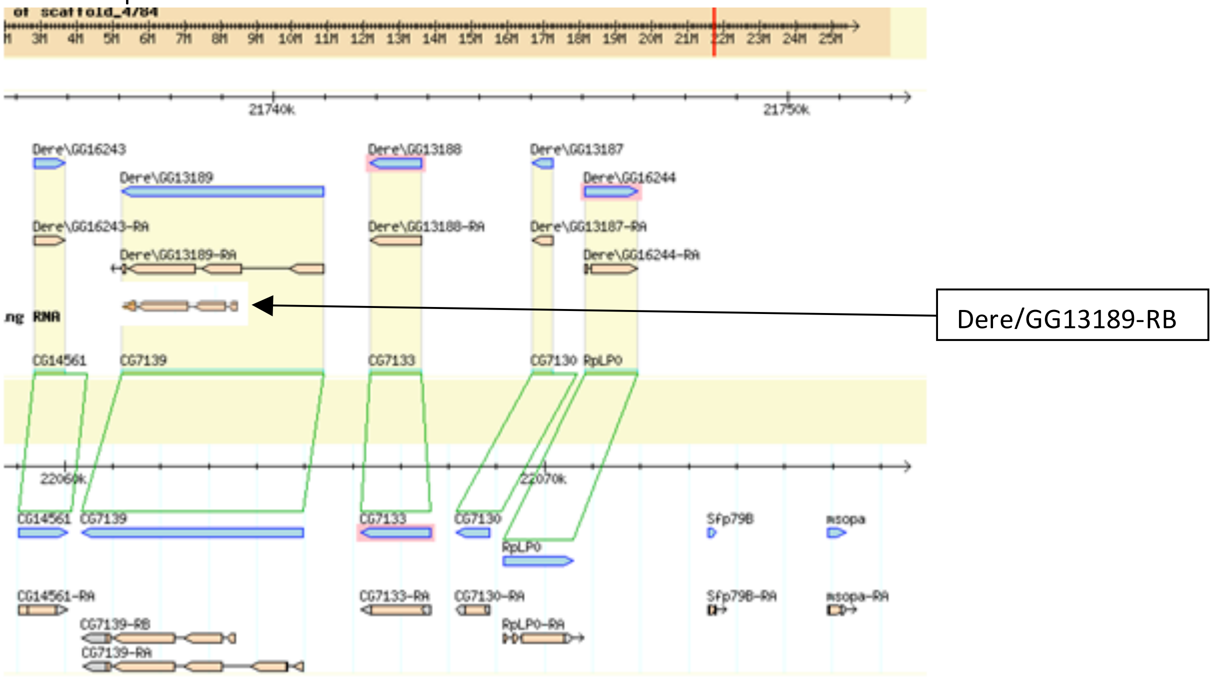
The Characterization of Orthologous CG14561, CG7139-PA, CG7139-PB, CG7133, CG7130 and the Ineffectual Exclusion of *Drosophila melanogaster* Exonic Sequence in RpLP0 in *Drosophila erecta*

Genomics Education Partnership
Fosmid 16
Harry Quedenfeld

Overview

The area represented by my fosmid starts at approximately base pair number 22,058,504 and proceeds 44,625 base pairs until about 22,103,128 in *Drosophila melanogaster*, though fosmid 16 is 40,000bp in *D. ere.*, which shows that either the latter has deletions or the former insertions. The figure below shows the gene-containing region, as the rest of DNA 3' of msopa contains no genes. The intention of this project was to determine genes within my fosmid of *Drosophila erecta* and probable function of those genes based on *D. mel.* as a reference genome. The transcripts of the genes CG14561, CG7139, CG7133, CG7130, RpLP0, Sfp79B and msopa were found to have homologous regions with my fosmid by use of Blast2, and all confirmed genes had the same amount of exons as their orthologous genes in *D. mel* (RpLP0 is an exception due to Genechecker's insufficiency to detect untranslated exons). Sfp79B and msopa are thought to be pseudogenes because of their short peptide length and high degree of mutation from *D. mel.* Evidence from *D. mel.* supports that similar genes with similar functional proteins are found in *D. ere.* for all genes except Sfp79B and msopa. RpLP0's first exon and a portion of the second are part of the 3' untranslated region in *D. mel*. RpLP0 in *D. mel*. does not have a start codon in its first exon, but in the middle of its second. In *D. ere.*, however, there is a mutation which causes for a premature start and reading of a premature stop codon. If this normally untranslated 1st exon was transcribed and translated, the protein would be nonfunctional and a pseudogene. If the 1st exon is not transcribed, then the RpLP0 protein in *D. ere.* is 98% identical to *D. mel.*'s, which is expected for a ribosomal protein. Genes CG14561, CG7139 and CG7130 in *D. ere.* have no significant differences from their *D. mel.* ortholog. CG7133 in *D. ere.* is similar length but shows a 55% identity with *D. mel.'*s, so it probably has a different function in *D. ere.* This protein is not highly conserved within close relatives of *D. ere.* and is still thought to be a gene, which evidences its probability to be a gene in *D. ere.* as well. In the species *Drosophila simulans*, *Drosophila sechellia* and *Drosophila yakuba*, CG7133 is not highly conserved but is still a putative gene, thus this is also a gene in *D. ere.* The gene Sfp79B is a pseudogene because it is only 4 amino acids in length due to a mutation that codes for an early start codon, which is followed by a stop codon. Msopa is also thought to be a pseudogene because it is only 64 amino acids long and has a large deletion when compared to *D. mel.*



Genes

[flybase.org, Genome Browser
scaffold_4784:21732372..21752371]

CG14561/NM_141120

CG14561-PA is the name of the ortholog in melanogaster to the gene found in fosmid 16 of *D. ere.* It has a single isoform, CG14561-PA. This gene codes from 753-1343 and the stop codon is present at 1344-1346 in fosmid 16. It is a single exon gene and codes in the plus direction. It is a gene in *D. mel.* and thus is protein-coding, yet its function and gene ontology is unknown. The BlastP between the protein for the gene I annotated in *D. ere.* and the known gene in *D. mel.* yields a significant e-value of 3e-100 and with 89% identity, showing that they are rather homologous. The function in *D. mel.* is likely the same as the function in *D. ere.*, so once the function is found, it is safe to assume its function in *D. ere.* is also found. This gene has orthologs in 10 closely related Drosophila species, so it originated from a common ancestor between these. Once there is a reference within one of these 10, its function can likely be applied to all because this sequence is moderately conserved. Orthologs are named (Dspecies/name): Dana\GF10942, Dgri\GH15124, Dmoj\GI13775, Dper\GL25424, Dpse\GA13080, Dsec\GM22428, Dsim\GD15018, Dvir\GJ13580, Dwil\GK10706 and Dyak\GE23023.

Exon
1

Start	End	Stop codon
753	1343	1344-1346

CG7139/NM_141121

This gene has two mRNA products, or two isoforms, named CG7139-PA and CG7139-PB in *D. mel.* Both isoforms are conserved rather highly in *D. ere.*, still having 4 exons in isoform A and 2 exons in isoform B, both running in the minus direction. Isoform B consists of the last two exons of isoform A, thus is missing the first two exons from isoform A. The function of this gene, in either isoform, is not known. CG7139 has orthologs in 10 other drosophilid species, in which they're named (Dspecies/name): Dana\GF23494, Dgri\GH16330, Dmoj\GI11522, Dper\GL25453, Dpse\GA20130, Dsec\GM22101, Dsim\GD12077, Dvir\GJ11777, Dwil\GK12132 and Dyak\GE22996. The coordinates in fosmid 16 for isoform A are 6696-6618, 6329-6277, 4789-3997 and 3876-2591, with a stop codon at 2590-2588. The coordinates for isoform B are 4789-3997 and 3876-2591 with a stop codon at 2590-2588. The last two exons in isoform A are the exact same present in isoform B; since the first codon in the third exon in isoform A is a Methionine (a start codon), isoform B begins at the same location. The first two exons of isoform A can be excluded to yield isoform B, thus this gene codes for two mRNA and protein products that may play differing or similar roles in a cell. This is the only gene in fosmid 16 that displays two mRNAs for one gene. The protein products yielded by my annotation of *D. ere.*'s gene are similar to those in *D. mel.* BlastP showed an e-value of 0.0 and an identity of 85% with both isoforms.

CG7139-PA

Exon	Start	End	Stop codon
1	6696	6618	-
2	6329	6277	-
3	4789	3997	-
4	3876	2591	2590-2588

CG7139-PB

Exon	Start	End	Stop codon
1	4789	3997	-
2	3876	2591	2590-2588

CG7133/NM_141122

This gene is a single exon gene transcribed in the minus direction. The coding region coordinates in fosmid 16 is 8274-7288 and the stop codon is from 7287-7285. It has a single isoform, CG7133-PA. This gene has its gene ontology defined. Its molecular function is described as a protein that binds to unfolded protein. It functions in heat shock protein (HSP) binding. Thus this protein is a heat shock protein that acts as a chaperone protein to fold unfolded proteins that are either recently translated or were denatured by heat. The e-value from a BlastP between the *D. mel.* CG7133 protein and the *D. ere.*'s orthologous protein is 3e-81, and identity sits at 55%. Sixty-five percent score as positives, meaning 10% of amino acids that differ are similar in both. Still, this lack of similarity is surprising for such an important protein. When the *D. mel.* protein is blasted against other closely related drosophilid species, none of them have high similarity, all having identities around or less than 50%. This either shows that this is a new mutation in *D. mel.* that causes for a functional HSP or it is a functional HSP in all species but does not need to conserve 50% of its peptide sequence to remain functional. If the former is the case, then that supports that this gene is a pseudogene in *D. ere.* There are no other copies of a similar gene in *D. ere.*, so I do not believe it to be a pseudogene; rather, I believe that this gene is functional in all of the related species because only conserving a certain area is necessary to retain its function. In fact, about the first 65 amino acids in CG7133's protein in *D. mel.* show moderate homology to many other related species, which suggests this region is more selected for than the other residues of the protein, which suggests it may be essential for the protein's interaction. This gene has 4 orthologs in closely related species, meaning it is a newer gene than CG14561, CG7139 and CG7130.

Exon	Start	End	Stop codon
1	8274	7288	7287-7285

CG7130/NM_141123

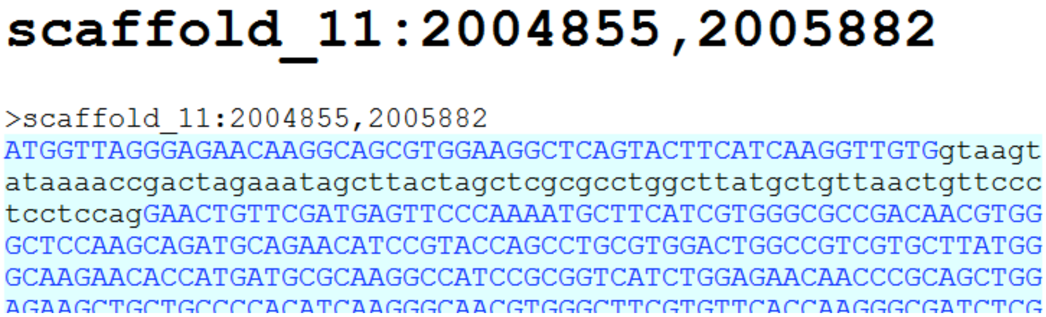
This gene is a single exon gene transcribed in the minus direction. The coding region coordinates in fosmid 16 is 10823-10443 and the stop codon is from 10442-10440. It has a single isoform, CG7130-PA. This gene has its gene ontology defined. Its molecular function is described as a protein that binds to unfolded protein. It functions in HSP binding. Thus this protein is a HSP that acts as a chaperone protein to fold unfolded proteins that are either recently translated or were denatured by heat. The BlastP between the *D. mel.* CG7130 protein and the protein I predicted in *D. ere.* yields an e-value of 1e-69 and a 92% amino acid identity. Since CG7130 and CG7133 are not highly paralogous, it is not likely that CG7130 can compensate for a lack of function in CG7133, which further supports that CG7133 is not a pseudogene (identity: 32%). This partial identity can be accounted for by the fact that they perform similar functions, but their structure would be different enough that one would most likely not be able to compensate for another. This gene has 10 orthologs, which means this HSP has been around longer than CG7133. Both CG7130 and CG7133 are in the same protein family, HSP40 (http://www.uniprot.org/uniprot/Q9VNW0,http://www.uniprot.org/uniprot/?query=CG7133&sort=score) , however their different structures have not yet been determined. At this time there is not enough information to determine whether they bind to the same substrates, but I hypothesize that they do not because of their high lack of amino acid identity.

CG7130

Exon	Start	End	Stop codon
1	10823	10443	10442-10440

RpLP0/NM_079487
Its gene ontology is defined as having molecular functions in repairing DNA, specifically as a DNA-(apurinic or apyrimidinic site) lyase activity, which is an enzyme that cuts out nucleotides to be replaced. Inside the cell it is thought to be a constituent of a ribosome, and its biological functions are translation, DNA repair, translational elongation and ribosome biogenesis. There were probably mistakes made in its GO definition because for it to repair DNA and be a part of a ribosome require different structures, for the same protein to perform such different functions is unlikely. It is a two-exon gene coding in the plus direction. The coordinates for the exons within my fosmid are: 11450-11503, 11579-12475, with the stop at 12476-12478; it codes in the plus direction. In *D. mel.* it is annotated to have 3 exons, however the first exon and a portion of the second are not translated into amino acids, and are thus a 3' untranslated region. The 3' untranslated region (and 1st exon) is at 11204-11302, with the second exon starting at 11410 but not coding until 11450. In *D. ere.* however, there is a missense mutation that causes for a premature start codon, which results in a premature stop codon being read and a highly truncated and dissimilar protein, thus this usually untranslated 1st exon must **not** be an exon in *D. ere.* because RpLP0 is highly conserved among drosophilids. *D. ere.*'s RpLP0 contains 2 exons, the first is the same as the coding region of the 2nd exon in *D. mel.*, and the 3rd has identical intron/exon borders. Since the 1st exon in *D. mel.* no longer exists as an exon in *D. ere.*, the proteins are the same length and their identity is 98%, with an e-value of 0.0. Thus this protein is highly conserved between species and performs the same function in *D. ere.* as in *D. mel.* RpLP0 has 10 orthologous genes in different *Drosophila* species including erecta. The orthologous genes are in (Species/Ortholog): Dana\GF10946, Dere\GG16244, Dgri\GH14667, Dmoj\GI13777\, Dpse\GA20389, Dsec\GM22429, Dsim\GD15019, Dvir\GJ13582, Dwil\GK20443 and Dyak\RpLP0. The two more closely related species, [R.1] Dsec and [R.2] Dana are shown below, both of which also exclude the 3' untranslated exon in order to conserve the protein sequence. Figure R.3 shows Dmel's RpLP0 nucleotide sequence, with a first exon non-coding and partial 2nd non-coding exon, indicated by black, capital letters. These untranslated regions may be a De novo mutation that changes the start codon in the first exon. Thus Dmel probably shows a new inclusion of this untranslated exon, instead of Dere showing a new exclusion because other species are similar to Dere, not Dmel. [Source: Decorated FASTA, flybase.org]

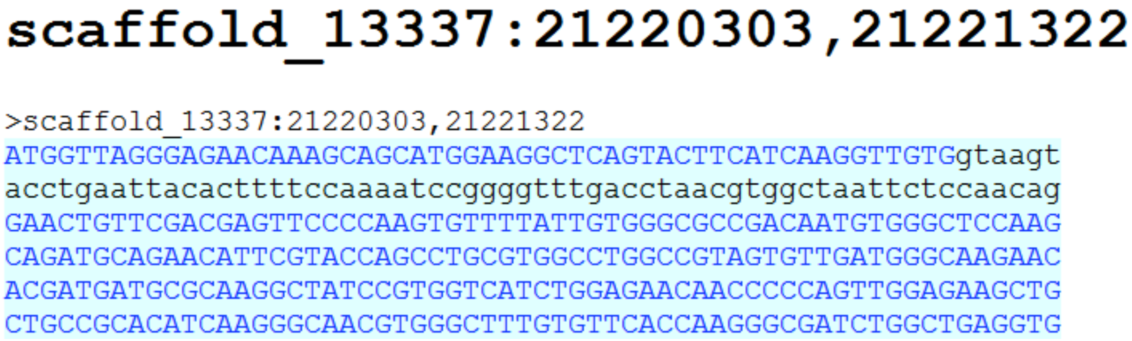
[R.1]



scaffold_11:2004855,2005882

```
>scaffold_11:2004855,2005882
ATGGTTAGGGAGAACAAGGCAGCGTGGAAGGCTCAGTACTTCATCAAGGTTGTGgtaagt
ataaaaccgactagaaatagcttactagctcgcgcctggcttatgctgttaactgttccc
tcctccagGAACTGTTCGATGAGTTCCCAAAATGCTTCATCGTGGGCGCCGACAACGTGG
GCTCCAAGCAGATGCAGAACATCCGTACCAGCCTGCGTGGACTGGCCGTCGTGCTTATGG
GCAAGAACACCATGATGCGCAAGGCCATCCGCGGTCATCTGGAGAACAACCCGCAGCTGG
AGAAGCTGCTGCCCCACATCAAGGGCAACGTGGGCTTCGTGTTCACCAAGGGCGATCTCG
```

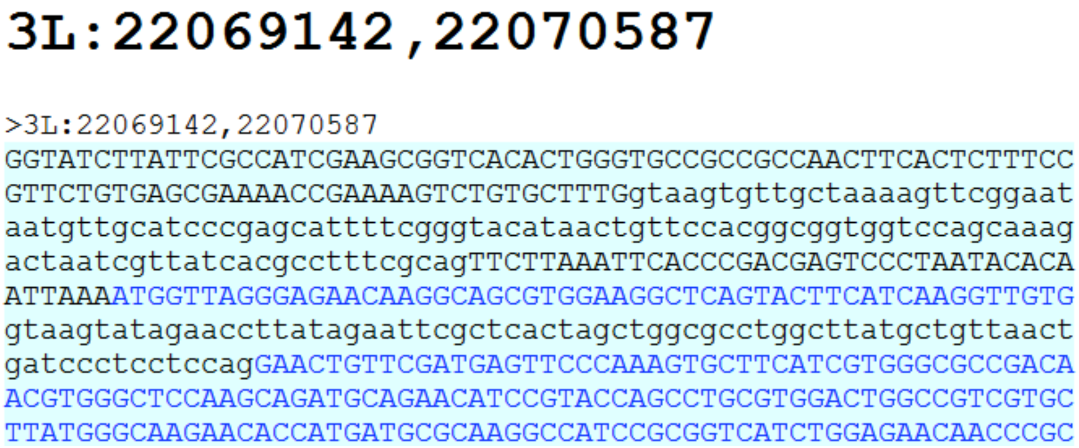
[R.2]



scaffold_13337:21220303,21221322

```
>scaffold_13337:21220303,21221322
ATGGTTAGGGAGAACAAAGCAGCATGGAAGGCTCAGTACTTCATCAAGGTTGTGgtaagt
acctgaattacacttttccaaaatccggggtttgacctaacgtggctaattctccaacag
GAACTGTTCGACGAGTTCCCCAAGTGTTTTATTGTGGGCGCCGACAATGTGGGCTCCAAG
CAGATGCAGAACATTCGTACCAGCCTGCGTGGCCTGGCCGTAGTGTTGATGGGCAAGAAC
ACGATGATGCGCAAGGCTATCCGTGGTCATCTGGAGAACAACCCCCAGTTGGAGAAGCTG
CTGCCGCACATCAAGGGCAACGTGGGCTTTGTGTTCACCAAGGGCGATCTGGCTGAGGTG
```

[R.3]



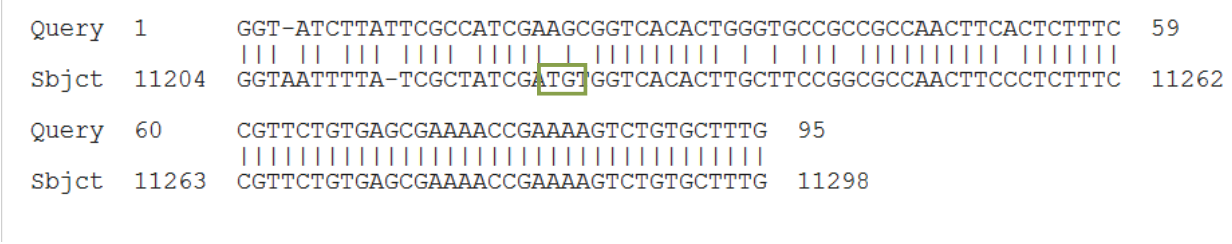
3L:22069142,22070587

```
>3L:22069142,22070587
GGTATCTTATTCGCCATCGAAGCGGTCACACTGGGTGCCGCCGCCAACTTCACTCTTTCC
GTTCTGTGAGCGAAAACCGAAAAGTCTGTGCTTTGgtaagtgttgctaaaagttcggaat
aatgttgcatcccgagcattttcgggtacataactgttccacggcggtggtccagcaaag
actaatcgttatcacgcctttcgcagTTCTTAAATTCACCCGACGAGTCCCTAATACACA
ATTAAAATGGTTAGGGAGAACAAGGCAGCGTGGAAGGCTCAGTACTTCATCAAGGTTGTG
gtaagtatagaaccttatagaattcgctcactagctggcgcctggcttatgctgttaact
gatccctcctccagGAACTGTTCGATGAGTTCCCAAAGTGCTTCATCGTGGGCGCCGACA
ACGTGGGCTCCAAGCAGATGCAGAACATCCGTACCAGCCTGCGTGGACTGGCCGTCGTGC
TTATGGGCAAGAACACCATGATGCGCAAGGCCATCCGCGGTCATCTGGAGAACAACCCGC
```

Exon	Start	End	Stop codon
1	11450	11503	-
2	11579	12475	12476-12478

Query: *D. mel.* RpLP0 mRNA, Sbjct: fosmid16

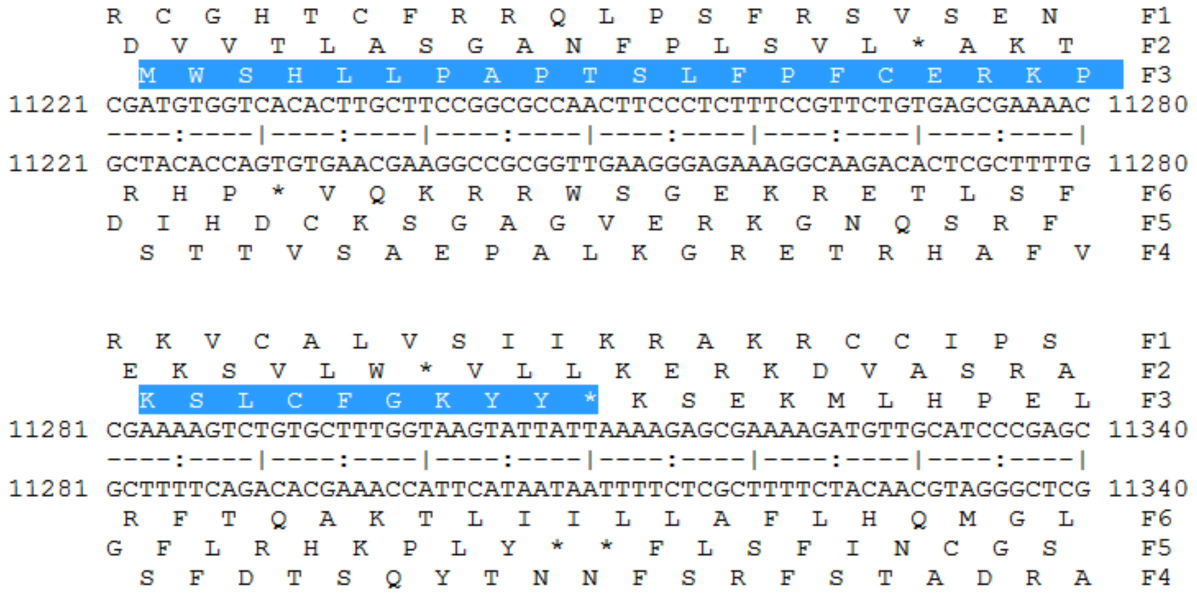
The green box below indicates the missense mutation and the premature start codon in the region homologous to the 1st exon in *D. mel.* If an exon, this normally untranslated exon would be translated in *D. ere.*, thus it must no longer be an exon in *D. ere.* because conservation of this protein is essential, with 89% amino acid identity or higher across *D. sec., D. ere., D. ana.* and *D. yak.*



```
Query  1      GGT-ATCTTATTCGCCATCGAAGCGGTCACACTGGGTGCCGCCGCCAACTTCACTCTTTC  59
              ||| || ||| |||| ||||| | ||||||||| | | ||| |||||||||| |||||||
Sbjct  11204  GGTAATTTTA-TCGCTATCGATGTGGTCACACTTGCTTCCGGCGCCAACTTCCCTCTTTC  11262

Query  60     CGTTCTGTGAGCGAAAACCGAAAAGTCTGTGCTTTG  95
              ||||||||||||||||||||||||||||||||||||
Sbjct  11263  CGTTCTGTGAGCGAAAACCGAAAAGTCTGTGCTTTG  11298
```

[NCBI Blast2P]

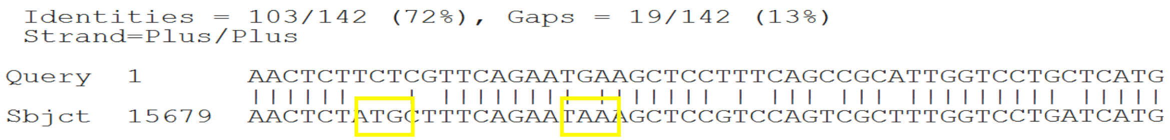
If this premature start codon is coded for, it will result in a protein that shares no homology with RpLP0, as it is only coded for in the first exon and the RpLP0 gene in *D. mel.* and other species has an untranslated first exon. This protein would only be 29 amino acids long and would most likely be nonfunctional. In order for *D. ere.* to survive, I hypothesize that this untranslated exon is excluded in transcription of mRNA in *D. ere.* There is no other protein that looks like RpLP0 in *D. ere.*, so this important gene is not a pseudogene copy.



```
      R  C  G  H  T  C  F  R  R  Q  L  P  S  F  R  S  V  S  E  N     F1
       D  V  V  T  L  A  S  G  A  N  F  P  L  S  V  L  *  A  K  T    F2
        M  W  S  H  L  L  P  A  P  T  S  L  F  P  F  C  E  R  K  P   F3
11221 CGATGTGGTCACACTTGCTTCCGGCGCCAACTTCCCTCTTTCCGTTCTGTGAGCGAAAAC 11280
      ----:----|----:----|----:----|----:----|----:----|----:----|
11221 GCTACACCAGTGTGAACGAAGGCCGCGGTTGAAGGGAGAAAGGCAAGACACTCGCTTTTG 11280
       R  H  P  *  V  Q  K  R  R  W  S  G  E  K  R  E  T  L  S  F    F6
      D  I  H  D  C  K  S  G  A  G  V  E  R  K  G  N  Q  S  R  F     F5
        S  T  T  V  S  A  E  P  A  L  K  G  R  E  T  R  H  A  F  V   F4


      R  K  V  C  A  L  V  S  I  I  K  R  A  K  R  C  C  I  P  S     F1
       E  K  S  V  L  W  *  V  L  L  K  E  R  K  D  V  A  S  R  A    F2
        K  S  L  C  F  G  K  Y  Y  *  K  S  E  K  M  L  H  P  E  L   F3
11281 CGAAAAGTCTGTGCTTTGGTAAGTATTATTAAAAGAGCGAAAAGATGTTGCATCCCGAGC 11340
      ----:----|----:----|----:----|----:----|----:----|----:----|
11281 GCTTTTCAGACACGAAACCATTCATAATAATTTTCTCGCTTTTCTACAACGTAGGGCTCG 11340
       R  F  T  Q  A  K  T  L  I  I  L  L  A  F  L  H  Q  M  G  L    F6
      G  F  L  R  H  K  P  L  Y  *  *  F  L  S  F  I  N  C  G  S     F5
        S  F  D  T  S  Q  Y  T  N  N  F  S  R  F  S  T  A  D  R  A   F4
```

Sfp79B/ACG69555

Not only is Sfp79B a very short protein in *D. mel.* of only 35 amino acids, it is truncated to 4 amino acids in *D. ere.*, which eliminates any possibility of the Sfp79B gene coding for a functional protein in *D. ere.* Based on its length, it must be a pseudogene. As is seen in the figure below, there is a missense mutation that causes for a premature start codon, which after 9 amino acids, is followed by a stop codon (TAA). It would have been a single exon gene in the plus direction, if functional. Sfp79B is a putative seminal fluid protein in *D. mel.* because it has a predicted structure similar to other seminal fluid proteins (thus is not confirmed). BlastP of the 4 amino acid sequence against *D. mel.* yielded no significant results. This gene is not anywhere else in *D. ere.*, but since there are many seminal fluid proteins that can compensate for its absence, it does not affect fertility. It could also be a pseudogene in *D. mel.* as well because it is short (<100 amino acids) and has not been curated.



```
 Identities = 103/142 (72%), Gaps = 19/142 (13%)
 Strand=Plus/Plus

Query  1      AACTCTTCTCGTTCAGAATGAAGCTCCTTTCAGCCGCATTGGTCCTGCTCATG
              ||||||   | |||||||| ||||||| | ||| ||| ||||||||| |||||
Sbjct  15679  AACTCTATGCTTTCAGAATAAAGCTCCGTCCAGTCGCTTTGGTCCTGATCATG
```

Query: *D. mel.*
Sbjct: fosmid16 (Tool: NCBI's Blast2n)

Msopa/NM_080120

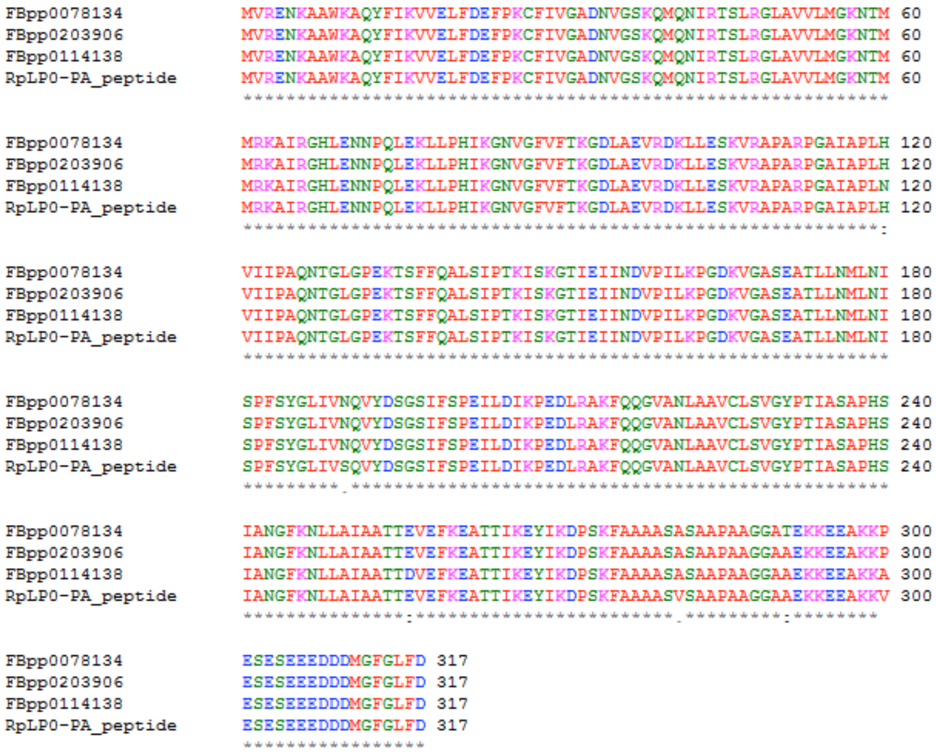
This feature is probably a pseudogene because it is only 63 amino acids long. The *D. mel.* msopa may also be a pseudogene because it was annotated by looking for ORFs that might code for proteins that look like a certain protein family, and has not been curated. *D. mel.*'s msopa protein is only 83 amino acids long. Regardless of its legitimacy as a gene in *D. mel.*, in *D. ere.*, msopa has a 20 amino acid deletion compared to *D. mel.*, is significantly shorter and only has a 46% identity with *D. mel.*'s. These three facts point to its being a pseudogene. This gene is thought to be involved in immunity or defense, however nothing is specified because this gene has not been researched. The protein is not known to exist, and at 83 amino acids it is unlikely to be a functional gene. For it to exist as 63 amino acids is even more unlikely. Msopa does not occur anywhere else in the *D. ere.* genome, the highest percent identity is the region in my fosmid with 46% shared amino acids. Furthermore, msopa is not conserved between other species (BlastP 68% identity with *Drosophila sechellia* 46% in *Drosophila yakuba,* NCBI BlastP). With such a lack of conservation and such a short length, it is probably a pseudogene. This evidence is stronger than its inferred existence due to predicted structure similarity because the researchers did not account for its lack of conservation in species. De novo mutations are highly unlikely. There are 4 orthologs to msopa in the other drosophilid species: Dere\GG16245, Dsec\GM22431, Dsim\GD15020 and Dyak\GE22603, however in Dsim it is 122 amino acids long, almost twice as long as Dere's supposed msopa ortholog. I do not agree that GG16245 is functional in *D. ere.* or that GM22431 is similar in *D. sec.* because its protein is only 67 amino acids long, given the median peptide length in all Drosophilid species is 373. [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1150220]

GENSCAN

Genscan predicted 6 features for my masked fosmid, three of which were single-exon. Genscan predicted single-exon genes rather well. The first example is CG14561, in which Genscan predicted a region in my fosmid that, when blasted against melanogaster, turned out to be the middle of CG14561; the start codon was late and the stop codon was early, probably in the wrong reading frame. It predicted the third exon of CG7139 to be a double-exon gene, which is understandable because it begins with a start codon (as does isoform B) and there is a splice junction (GT) right before the stop codon (TGA), the "T" of the stop codon is the "T" of the "GT" splice junction. For the third feature, Genscan also predicted a single exon gene in the middle of CG7133, but it only takes up about one third of the actual gene. This probably happened because Genscan read in the wrong frame, causing a late start codon and a premature stop. The fourth predicted feature was single-exon and turned out to be homologous to the center of CG7130, only slightly short on both ends probably due to reading the incorrect frame. The fifth predicted feature, when extracted from fosmid 16 and blasted against *D. mel.*, is highly homologous to the RpLP0 gene, only missing the first exon and predicting it has two. Genscan did predict this accurately by overlooking the first exon, which contains a mutation for a start codon. Genscan most likely skipped this because the resulting protein would be too short, thus it ended up predicting RpLP0 correctly. The sixth feature predicted a two-exon gene that spans from 18704-39859. When these bases are extracted from fosmid 16, they are homologous with regions after msopa but there are no genes in *D. melanogaster* in any of the homologous regions. The Genscan coordinates given matched up with splice junctions GT and AG, but there is a stop codon in the reading frame and cuts the first predicted exon short, with no splice junction nearby. Due to the fact that Genscan predicted the first exon incorrectly and the exons are very distant from each other, there is not strong evidence for a new gene here in *D. ere.* There are 4 introns predicted to be between the two exons, but there are 10kbp between the first exon and first intron, which is not likely to occur because exons and introns are usually continuous. Since none of the Genscan predicted genes matched up with the real genes perfectly, it cannot be solely relied on; however, it proves to be a good starting point. Blasting the region of the predicted feature against *D. mel.* will result in homologous regions, and Genome View will show if it is part of any gene, which is helpful.

CLUSTAL analysis - RpLP0 analysis

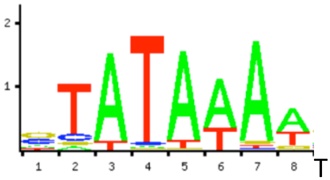
As the image from CLUSTALW2 illustrates, RpLP0 is a highly conserved protein because it has an essential function as a constituent of a ribosome. Ribosomes are necessary in every cell all the time, if they mutate the cell will die. If an organism has a mutation in RpLP0, it will likely die. There are 5 mutations throughout the 4 organisms, but they all result in a similar amino acid. This means they have the same charge and/or polarity, so the protein's function would not be altered significantly by such a mutation. KEY: FBpp0078134: *D. melanogaster* RpLP0, FBpp0203906: *D. sechellia* GM22429, FBpp0114138: *D. ananassae* GF10946, RpLP0-PA_peptide: *D. ere.*'s orthologous protein of RpLP0 from fosmid 16, GG16244.



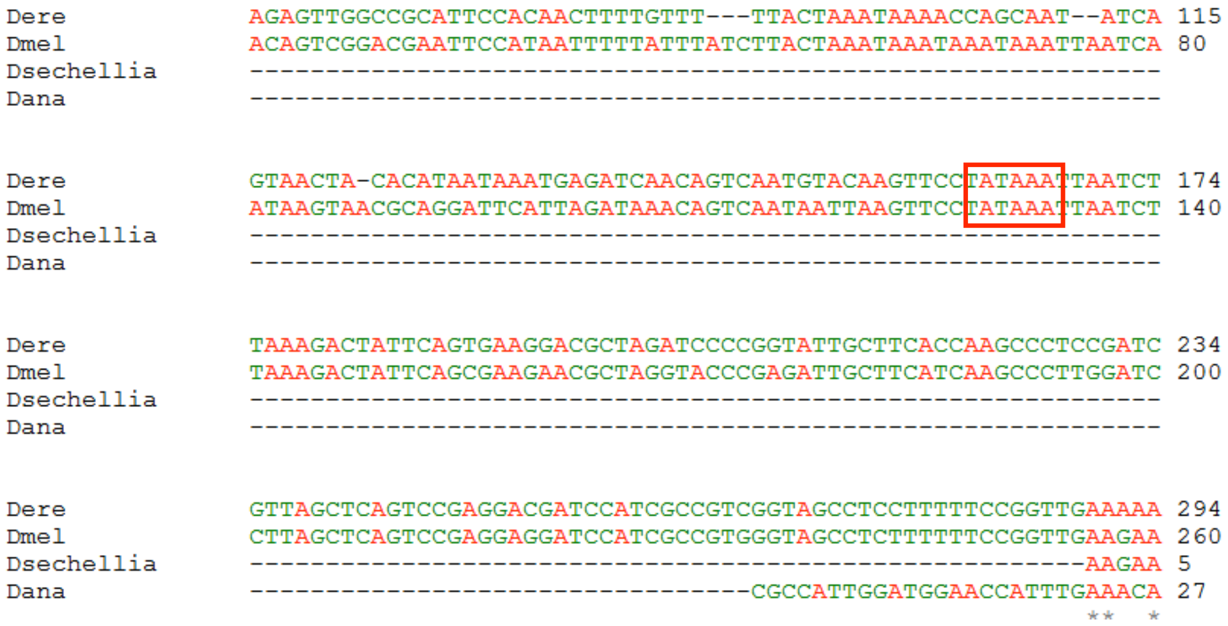
```
FBpp0078134         MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGLAVVLMGKNTM 60
FBpp0203906         MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGLAVVLMGKNTM 60
FBpp0114138         MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGLAVVLMGKNTM 60
RpLP0-PA_peptide    MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGLAVVLMGKNTM 60
                    ************************************************************

FBpp0078134         MRKAIRGHLENNPQLEKLLPHIKGNVGFVFTKGDLAEVRDKLLESKVRAPARPGAIAPLH 120
FBpp0203906         MRKAIRGHLENNPQLEKLLPHIKGNVGFVFTKGDLAEVRDKLLESKVRAPARPGAIAPLH 120
FBpp0114138         MRKAIRGHLENNPQLEKLLPHIKGNVGFVFTKGDLAEVRDKLLESKVRAPARPGAIAPLN 120
RpLP0-PA_peptide    MRKAIRGHLENNPQLEKLLPHIKGNVGFVFTKGDLAEVRDKLLESKVRAPARPGAIAPLH 120
                    ***********************************************************:

FBpp0078134         VIIPAQNTGLGPEKTSFFQALSIPTKISKGTIEIINDVPILKPGDKVGASEATLLNMLNI 180
FBpp0203906         VIIPAQNTGLGPEKTSFFQALSIPTKISKGTIEIINDVPILKPGDKVGASEATLLNMLNI 180
FBpp0114138         VIIPAQNTGLGPEKTSFFQALSIPTKISKGTIEIINDVPILKPGDKVGASEATLLNMLNI 180
RpLP0-PA_peptide    VIIPAQNTGLGPEKTSFFQALSIPTKISKGTIEIINDVPILKPGDKVGASEATLLNMLNI 180
                    ************************************************************

FBpp0078134         SPFSYGLIVNQVYDSGSIFSPEILDIKPEDLRAKFQQGVANLAAVCLSVGYPTIASAPHS 240
FBpp0203906         SPFSYGLIVNQVYDSGSIFSPEILDIKPEDLRAKFQQGVANLAAVCLSVGYPTIASAPHS 240
FBpp0114138         SPFSYGLIVNQVYDSGSIFSPEILDIKPEDLRAKFQQGVANLAAVCLSVGYPTIASAPHS 240
RpLP0-PA_peptide    SPFSYGLIVSQVYDSGSIFSPEILDIKPEDLRAKFQQGVANLAAVCLSVGYPTIASAPHS 240
                    *********.**************************************************

FBpp0078134         IANGFKNLLAIAATTEVEFKEATTIKEYIKDPSKFAAAASASAAPAAGGATEKKEEAKKP 300
FBpp0203906         IANGFKNLLAIAATTEVEFKEATTIKEYIKDPSKFAAAASASAAPAAGGAAEKKEEAKKP 300
FBpp0114138         IANGFKNLLAIAATTDVEFKEATTIKEYIKDPSKFAAAASASAAPAAGGAAEKKEEAKKA 300
RpLP0-PA_peptide    IANGFKNLLAIAATTEVEFKEATTIKEYIKDPSKFAAAASVSAAPAAGGAAEKKEEAKKV 300
                    ***************:************************.*********:******** 

FBpp0078134         ESESEEEDDDMGFGLFD 317
FBpp0203906         ESESEEEDDDMGFGLFD 317
FBpp0114138         ESESEEEDDDMGFGLFD 317
RpLP0-PA_peptide    ESESEEEDDDMGFGLFD 317
                    *****************
```

5' Upstream Regions

These alignments are performed using the 1,000 bp 5' upstream from RpLP0 in *D. mel.* and its orthologs in *D. ere, D. sec.* and *D. ana.* There is a conserved TATA box in both the *D. ere.* and *D. mel.* 5' upstream regions, shown by the red box below. The TATA box is at 162-167 in *D. ere.* and 127-132 in *D. mel.* TATA box used here is defined as any sequence upstream of the initiator with 5 of 6 nucleotides conforming to the consensus TATAAA. (Locations are based on t he 1000bp extract 5' of RpLP0 or its orthologs).

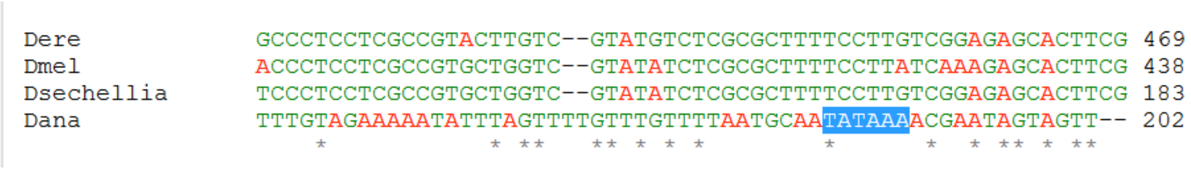


The figure to the left illustrates the frequency of bases within a TATA box. [Source: http://jaspar.genereg.net/cgi-bin/jaspar_db.pl]



```
Dere            AGAGTTGGCCGCATTCCACAACTTTTGTTT---TTACTAAATAAAACCAGCAAT--ATCA 115
Dmel            ACAGTCGGACGAATTCCATAATTTTTATTTATCTTACTAAATAAATAAATAAATTAATCA 80
Dsechellia      ------------------------------------------------------------
Dana            ------------------------------------------------------------

Dere            GTAACTA-CACATAATAAATGAGATCAACAGTCAATGTACAAGTTCCTATAAATTAATCT 174
Dmel            ATAAGTAACGCAGGATTCATTAGATAAACAGTCAATAATTAAGTTCCTATAAATTAATCT 140
Dsechellia      ------------------------------------------------------------
Dana            ------------------------------------------------------------

Dere            TAAAGACTATTCAGTGAAGGACGCTAGATCCCCGGTATTGCTTCACCAAGCCCTCCGATC 234
Dmel            TAAAGACTATTCAGCGAAGAACGCTAGGTACCCGAGATTGCTTCATCAAGCCCTTGGATC 200
Dsechellia      ------------------------------------------------------------
Dana            ------------------------------------------------------------

Dere            GTTAGCTCAGTCCGAGGACGATCCATCGCCGTCGGTAGCCTCCTTTTTCCGGTTGAAAAA 294
Dmel            CTTAGCTCAGTCCGAGGAGGATCCATCGCCGTGGGTAGCCTCTTTTTTCCGGTTGAAGAA 260
Dsechellia      -------------------------------------------------------AAGAA 5
Dana            ---------------------------------CGCCATTGGATGGAACCATTTGAAACA 27
                                                               **  *
```

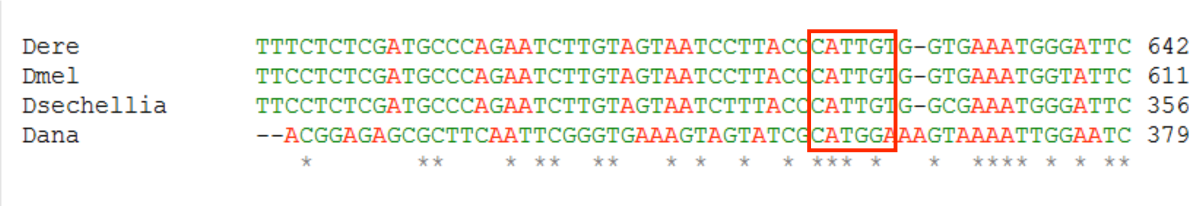
D. ana. also has a TATA box [shown below] within 20bp of *D. ere.*’s and 60 of *D. mel.*’s, however *D. sec.* does not possess a quality TATA box (It has more than one different base from TATAAA) within its equivalent 5’ region of DNA, thus it likely does not have one. The TATA box location is at 190-196.



```
Dere            GCCCTCCTCGCCGTACTTGTC--GTATGTCTCGCGCTTTTCCTTGTCGGAGAGCACTTCG 469
Dmel            ACCCTCCTCGCCGTGCTGGTC--GTATATCTCGCGCTTTTCCTTATCAAAGAGCACTTCG 438
Dsechellia      TCCCTCCTCGCCGTGCTGGTC--GTATATCTCGCGCTTTTCCTTGTCGGAGAGCACTTCG 183
Dana            TTTGTAGAAAAATATTTAGTTTTGTTTGTTTTAATGCAATATAAAACGAATAGTAGTT-- 202
                    *           * **   ** * * *        *      *  * ** * **
```

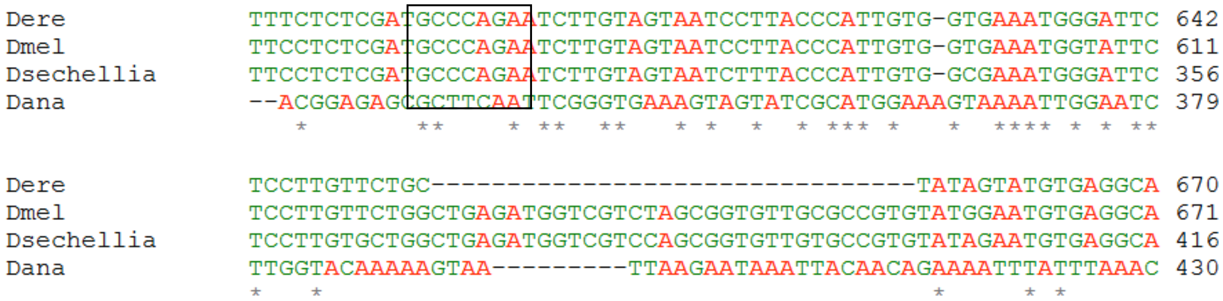
D. ana. Is clearly the least similar among the four; however, *D. mel* and *D. ere*. show a higher similarity in the 1,000 bp before RpLP0 than the others do to them. This is not in agreement with what is expected because *D. sec.* is more closely related to *D. mel.* than *D. ere.* is, however these alignments attest that this 5’ upstream region have higher homology between *D. mel.* and *D. ere.* according to TATA box locations. The first 200 base pairs of *D. sec.* and *D. ana.* not align with *D. mel.* and *D. ere.* because this region was not conserved enough to align between 4 species perfectly, There might have been an insertion in *D. sec* and *D. ana*. that does not align with the first 250 base pairs of the other two species. *D. ana.* is clearly the most distant relative of the other 3 species, as is visible with the ‘show colors’ option in CLUSTAL.

The initiator CCATTG was found in all four sequences in the alignment, indicated by the red box below. *D. ana.* has a missense mutation that leads the C to be replaced with a G, but this is still an initiation sequence.



```
Dere            TTTCTCTCGATGCCCAGAATCTTGTAGTAATCCTTACCCATTGTG-GTGAAATGGGATTC 642
Dmel            TTCCTCTCGATGCCCAGAATCTTGTAGTAATCCTTACCCATTGTG-GTGAAATGGTATTC 611
Dsechellia      TTCCTCTCGATGCCCAGAATCTTGTAGTAATCTTTACCCATTGTG-GCGAAATGGGATTC 356
Dana            --ACGGAGAGCGCTTCAATTCGGGTGAAAGTAGTATCGCATGGAAAGTAAAATTGGAATC 379
                   *       **    * **  **   * *  *  * *** *   *  **** * * **
```

A downstream promoter element (DPE) is any 6 nucleotide sequence at exactly +28 to +33 with 5 of 6 nucleotides conforming to the DPE functional range set A/G/T - C/G - A/T - C/T - A/C/G - C/T. As seen in the figure below, there are no DPE in any of the species because there is a C rich region here, which causes the third nucleotide to always be a C and eliminates the chance for a DPE region. The *D. ana.* sequence GCTTCA would work if the last A was a T, so *D. ana.* also probably does not have a DPE 28-35 3' of the initiator. [Black box below indicates 28-35 bp after the initiator, where DPEs are possible].



```
Dere            TTTCTCTCGATGCCCAGAATCTTGTAGTAATCCTTACCCATTGTG-GTGAAATGGGATTC 642
Dmel            TTCCTCTCGATGCCCAGAATCTTGTAGTAATCCTTACCCATTGTG-GTGAAATGGTATTC 611
Dsechellia      TTCCTCTCGATGCCCAGAATCTTGTAGTAATCTTTACCCATTGTG-GCGAAATGGGATTC 356
Dana            --ACGGAGAGCGCTTCAATTCGGGTGAAAGTAGTATCGCATGGAAAGTAAAATTGGAATC 379
                   *       **    * **  **   * *  *  * *** *   *  **** * * **

Dere            TCCTTGTTCTGC--------------------------------TATAGTATGTGAGGCA 670
Dmel            TCCTTGTTCTGGCTGAGATGGTCGTCTAGCGGTGTTGCGCCGTGTATGGAATGTGAGGCA 671
Dsechellia      TCCTTGTGCTGGCTGAGATGGTCGTCCAGCGGTGTTGTGCCGTGTATAGAATGTGAGGCA 416
Dana            TTGGTACAAAAAGTAA---------TTAAGAATAAATTACAACAGAAAATTTATTTAAAC 430
                *   *                                        *     * *
```

[Source for known sequences: http://www-biology.ucsd.edu/labs/Kadonaga/DCPD.html]

Repeats
[From Repeat Masker]
Most of fosmid 16 is high complexity, despite the last 20kbp being vacant of any genes in *D. mel.* and *D. ere.*, only 6.31% is repetitive, with nearly 5% of that being interspersed repeats, of which were mostly DNA transposons. Given that transposons account for a higher percentage of the entire *Drosophila* genome on average, the low percentage of DNA transposons evidences that it is a possible gene-rich area. There is only 0.15% of my fosmid composed of Retroelements, which are also typically much more common, which shows this area has been conserved. It is expected that, any organisms that had many transposons or retroelement would have mutations in this area and increase their chance of mutating fatally. The organisms that lived do not have mutations in these genes caused by transposon elements, so we expect less repeats in a gene-rich region. However, the first 20,000 bp of fosmid 16 actually contain nearly twice as many repeats [Fig2.2]. Thus, it is not reliable to judge whether a region is gene rich by the proportion of repeats because in this case, the gene-rich region within a larger area actually has almost all of the repeats; 2425 of 2526 repeats were found in the gene-rich region. This shows that transposons and other repeatable elements are present but do not cause any fatal mutations, and sheds light that the coding regions may not be exceptionally stable. [Areas in yellow refer to specifics mentioned]

[2.1] Summary:

```
==================================================
file name: RM2_Fosmid16.txt_1241404234
sequences:            1
```

```
total length:      40000 bp  (40000 bp excl N/X-runs)
GC level:         43.93 %
bases masked:       2526 bp ( 6.31 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
Retroelements            1           60 bp    0.15 %
   SINEs:                0            0 bp    0.00 %
   Penelope              0            0 bp    0.00 %
   LINEs:                1           60 bp    0.15 %
    CRE/SLACS            0            0 bp    0.00 %
     L2/CR1/Rex          0            0 bp    0.00 %
     R1/LOA/Jockey       0            0 bp    0.00 %
     R2/R4/NeSL          0            0 bp    0.00 %
     RTE/Bov-B           0            0 bp    0.00 %
     L1/CIN4             0            0 bp    0.00 %
   LTR elements:         0            0 bp    0.00 %
     BEL/Pao             0            0 bp    0.00 %
     Ty1/Copia           0            0 bp    0.00 %
     Gypsy/DIRS1         0            0 bp    0.00 %
       Retroviral        0            0 bp    0.00 %

DNA transposons         11         1899 bp    4.75 %
   hobo-Activator        0            0 bp    0.00 %
   Tc1-IS630-Pogo        0            0 bp    0.00 %
   En-Spm                0            0 bp    0.00 %
   MuDR-IS905            0            0 bp    0.00 %
   PiggyBac              0            0 bp    0.00 %
   Tourist/Harbinger     0            0 bp    0.00 %
   Other (Mirage,        3          692 bp    1.73 %
    P-element, Transib)

Rolling-circles          0            0 bp    0.00 %

Unclassified:            0            0 bp    0.00 %

Total interspersed repeats:        1959 bp    4.90 %


Small RNA:               0            0 bp    0.00 %

Satellites:              0            0 bp    0.00 %
Simple repeats:          5          246 bp    0.61 %
Low complexity:          6          321 bp    0.80 %
```

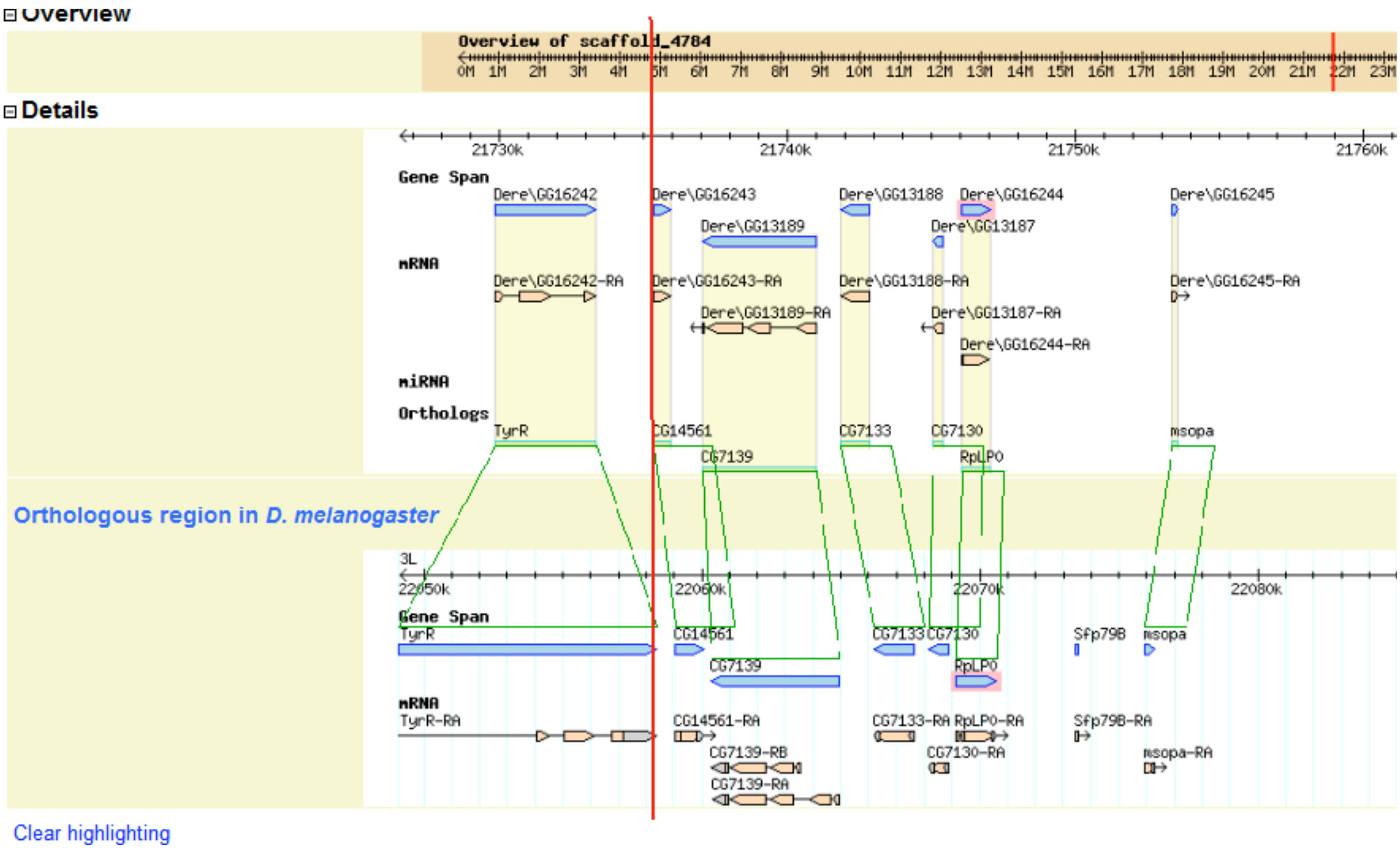
[2.2] Summary:

```
==================================================
file name: RM2sequpload_1242055364
sequences:             1
total length:      20000 bp  (20000 bp excl N/X-runs)
GC level:         43.84 %
bases masked:       2425 bp ( 12.12 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
Retroelements            1           60 bp    0.30 %
   SINEs:                0            0 bp    0.00 %
   Penelope              0            0 bp    0.00 %
   LINEs:                1           60 bp    0.30 %
    CRE/SLACS            0            0 bp    0.00 %
     L2/CR1/Rex          0            0 bp    0.00 %
     R1/LOA/Jockey       0            0 bp    0.00 %
     R2/R4/NeSL          0            0 bp    0.00 %
     RTE/Bov-B           0            0 bp    0.00 %
     L1/CIN4             0            0 bp    0.00 %
   LTR elements:         0            0 bp    0.00 %
     BEL/Pao             0            0 bp    0.00 %
     Ty1/Copia           0            0 bp    0.00 %
     Gypsy/DIRS1         0            0 bp    0.00 %
       Retroviral        0            0 bp    0.00 %

DNA transposons         10         1843 bp    9.21 %
   hobo-Activator        0            0 bp    0.00 %
   Tc1-IS630-Pogo        0            0 bp    0.00 %
   En-Spm                0            0 bp    0.00 %
   MuDR-IS905            0            0 bp    0.00 %
   PiggyBac              0            0 bp    0.00 %
   Tourist/Harbinger     0            0 bp    0.00 %
   Other (Mirage,        3          692 bp    3.46 %
    P-element, Transib)

Rolling-circles          0            0 bp    0.00 %

Unclassified:            0            0 bp    0.00 %

Total interspersed repeats:        1903 bp    9.52 %


Small RNA:               0            0 bp    0.00 %

Satellites:              0            0 bp    0.00 %
Simple repeats:          4          209 bp    1.04 %
Low complexity:          9          313 bp    1.56 %
==================================================
```

Synteny

Key: [Left most] vertical red line at approximately 21735kbp in *D. ere.* is the start of fosmid16. From the 3' is the entire coding region, all 3' after msopa is non-coding DNA in both *D. mel*. and *D. ere.*

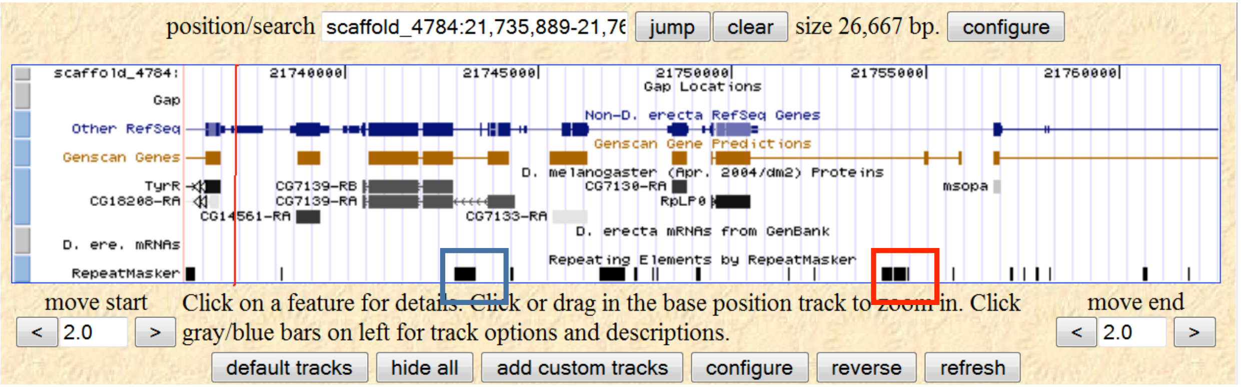
[1] Orthology between Dere Scaffold_4784 (top) and Dmel Chr3L (bottom)



[flybase.org GBrowser]

This shows high synteny between Dere and Dmel because all of the genes in Dere are in the same order as Dmel's, code in the same direction and are similar distances apart.

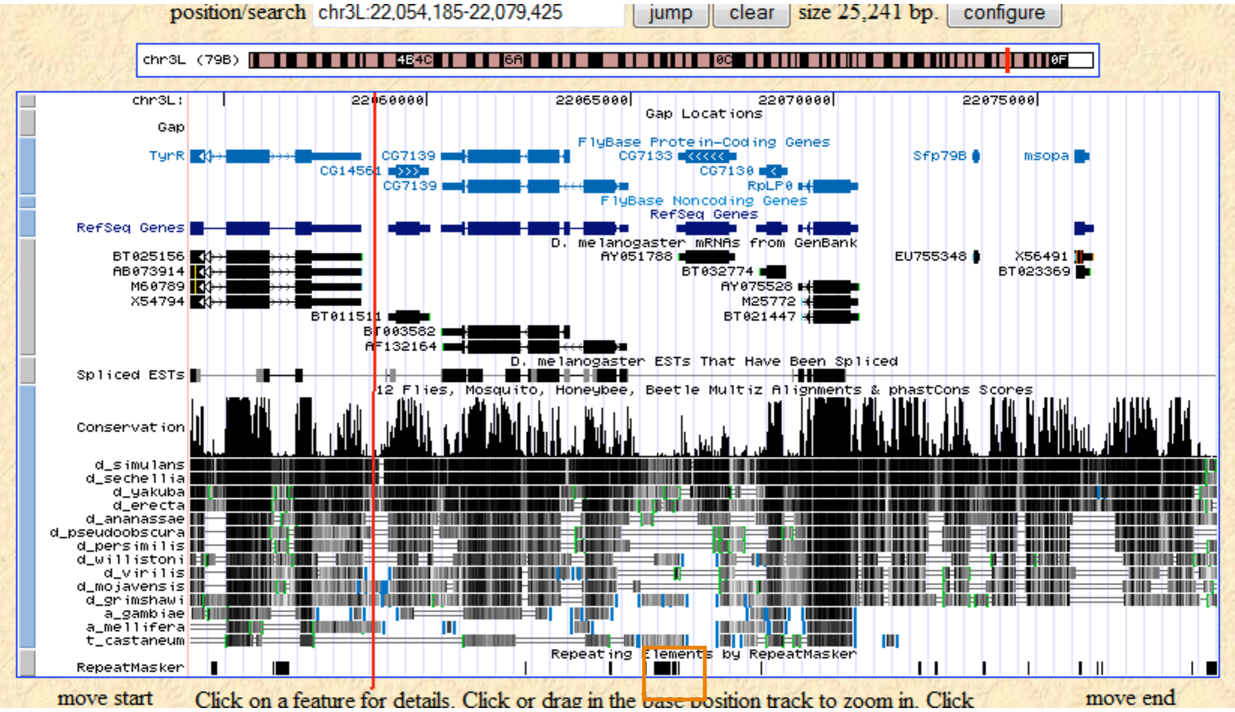
[2] Fosmid16 region in D. ere.


position/search scaffold_4784:21,735,889-21,76 jump clear size 26,667 bp. configure

Click on a feature for details. Click or drag in the base position track to zoom in. Click gray/blue bars on left for track options and descriptions.

[UCSC Genome Browser: http://genome.ucsc.edu]

[3] Fosmid 16 region in D. mel.



position/search chr3L:22,054,185-22,079,425 size 25,241 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click

[UCSC Genome Browser: http://genome.ucsc.edu]

The above images show homology [1] and repeat elements in Dere and Dmel [2,3]. Fosmid16 is on chromosome 3L in D. melanogaster and Scaffold_4784 in Dere. Synteny has been preserved because the genes on fosmid16 are all from the same region of the *D. melanogaster* genome; that is, they are in the same order, spaced similarly and on the same chromosome in both *D. ere.* and *D. mel.* Thus there is no evidence of any chromosomal mutations such as inversions or transpositions that have occurred since these two species split. There are repeats present in Dere [2] between the first and second exon (minus direction) that are not present in Dmel (blue box [2]), which suggests a transposable element inserted itself between those exons (the first two in CG7139-PA). There is also another insertion of repeats in Dere between RpLP0 and msop that is not present in Dmel (red box in [2]). The orange box in [3] shows repeats that intervene CG1739 and CG1733 in Dmel, but come after CG7133 in Dere, which evidences minor DNA rearrangement; however, this does not alter the synteny of genes in this region.