Annotating Drosophila Erecta, fosmid 14

Ari Akerstein,
12.20.07
SFSU Dept. of Biology
Biol 871 (Bioinformatics), Dr. Chris Smith

Abstract,

Fosmid14 of Drosophila Erecta shows high similarity to Drosophila Melanogaster. Fosmid analysis reveals 3 nearly identical orthologous genes with similar intron / exon patterns, gene location and sizes. Furthermore, three putative genes were predicted with varying degrees of evidence.

However despite the general similarity in gene structure between the two species, there is some difference in the placement and density of transposable elements (TEs). Dere shows a higher density of TEs on both the forward and reverse strands. Of particular interest is a TE cluster upstream of Dere CG7140 in roughly the space one would expect an ortholog to Dmel CG7139 (absent on fosmid14). While perplexing, this may suggest the erosion of a gene (CG7139) between species with replacement by a cluster of TEs. Further, Clustal QT alignments and Ka/Ks analysis indicate strong purifying selection.

Global structure, Dere fosmid14

Dere fosmid 14 from GEP was uploaded in .xml format to Apollo, a gene curation program. The GEP project report indicated there were 6 predicted genes on Dere Fosmid14. The GEP project report estimated 3 of these to be real. Repeat masker predicted a repeat density of 2.42%.

Using Flybase D. melanogaster sequence as reference, several blasts were run (blastp, blastn, tblastn) against each gene - both total sequence blasts as well as exon by exon - to validate these predictions. It was concluded that the three orthologous genes were properly predicted while only one of the three genes predicted by gene predictor programs (Genscan, SNAP) were correct.

The overarching assumption in this work is that annotation of D. Erecta should be done in comparison to its better-understood relative Drosophila Melanogaster. Indeed, one expects the two genetic profiles to be nearly identical. This raises the somewhat philosophical question of how many / which (types) of genes ultimately distinguish one species from another. Perhaps it is not the gene structure, per se, but other features such as repeats or transposable elements that account for variety between close species. Regardless, one expects to find inevitable differences between genomes, but

that they should remain generally similar. Features considered most notable are deviations in genomic data as this indicates features that, at least in part, distinguish one species from another. How exactly such differences contribute to speciation remains a puzzle. Suffice it to say that the annotators alterations to the fosmid were noted and accounted for to the greatest extent possible. What follows is a discussion about the features considered the most interesting on the fosmid.

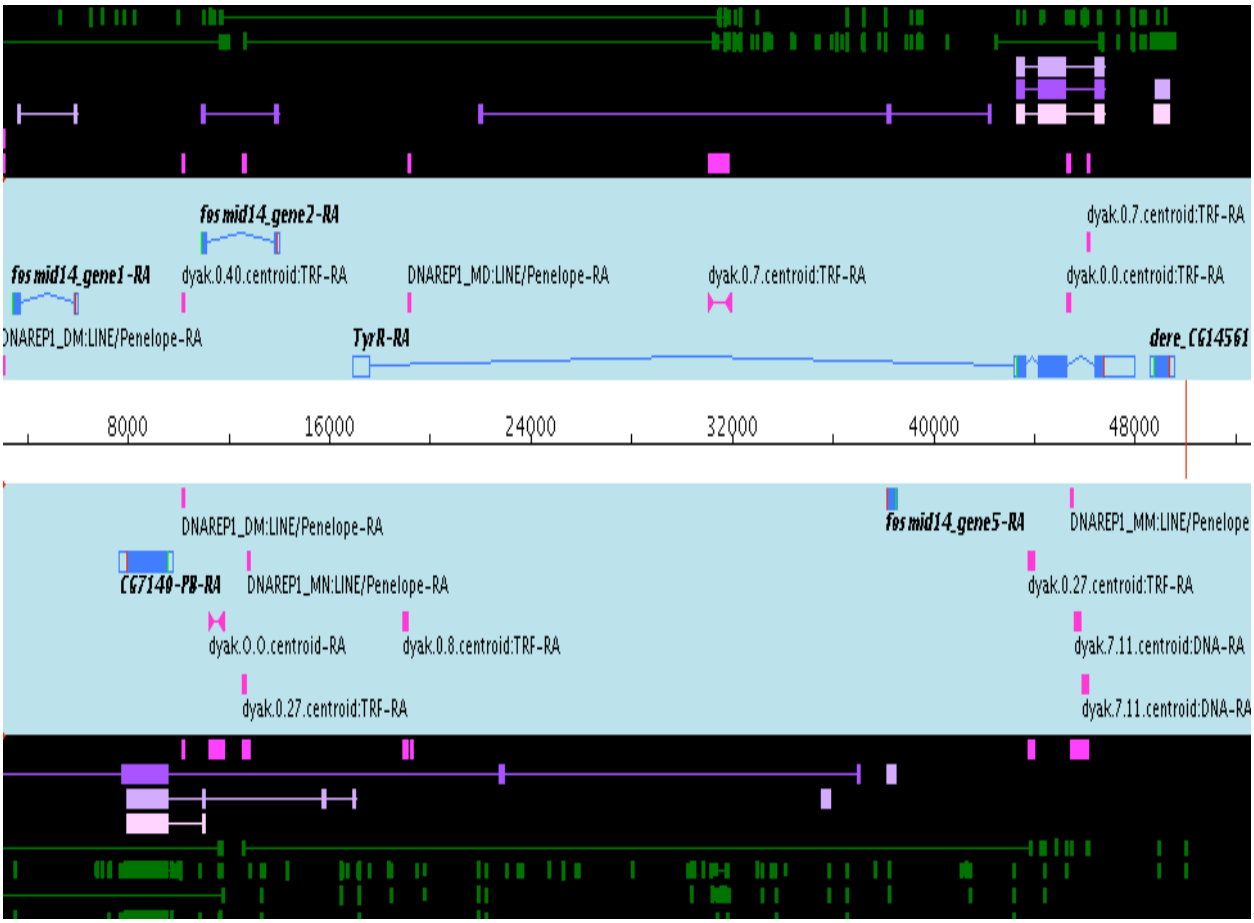


Fig 1. Global structure of Drosophila Erecta, fosmid 14, (including putative genes)

Gene by Gene analysis

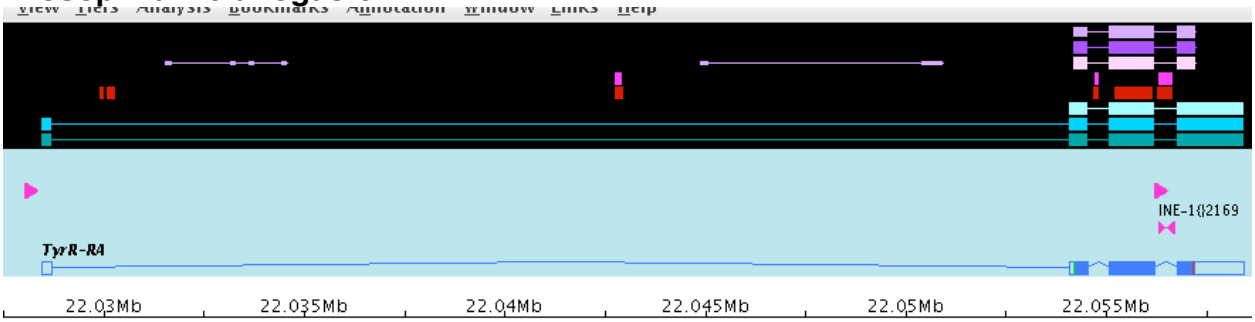
TyrR

The molecular function of TyrR is described as being involved in smell perception, muscle contraction, nerve-nerve synaptic transmission, G-protein coupled receptor activity. There is a possible phenotype association with neuro-muscular

junctions. Blast results reveal 97% similarity and 96% identity. The number of exons and introns is preserved between the genes, as is a long, distinctive 5' UTR. Splice sites are well-predicted between all the gene predictor programs (SNAP, Genscan etc.). There are two TEs within the Dere first intron that are also found in Dmel. However these TEs were unannotated in Dmel. The third intron in Dere is slightly longer.

GleanR comparisons reveal this gene is well-conserved between Drosophilidae.

Drosophila Melanogaster



Drosophila Erecta

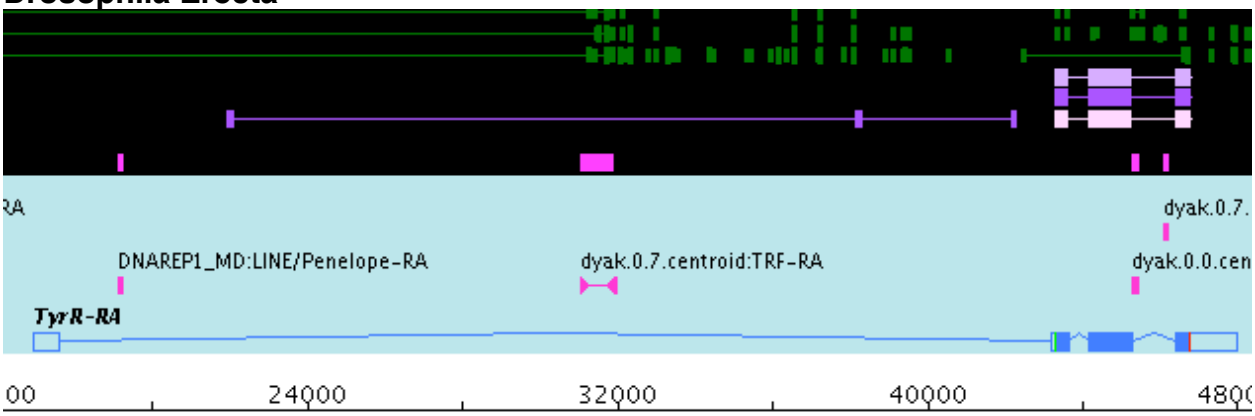
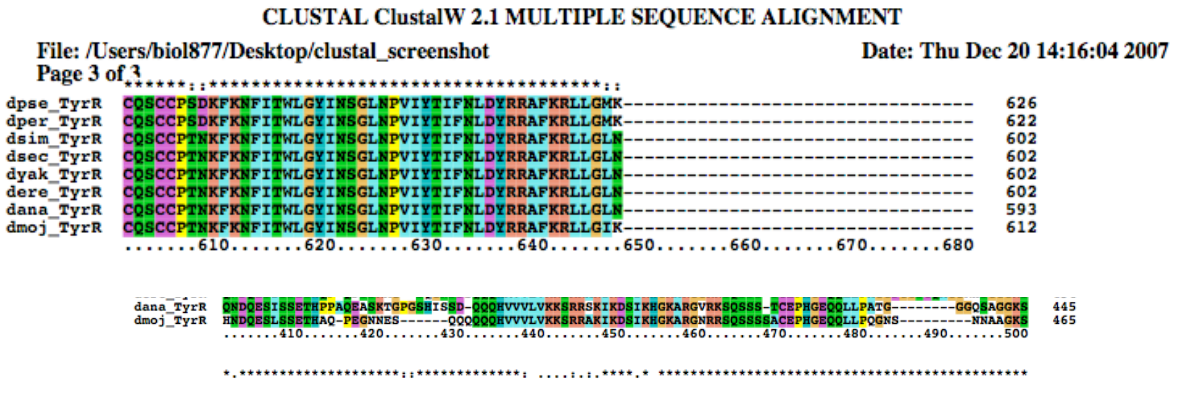


Fig 2. Side by side comparison between Dmel TyrR and Dere TyrR



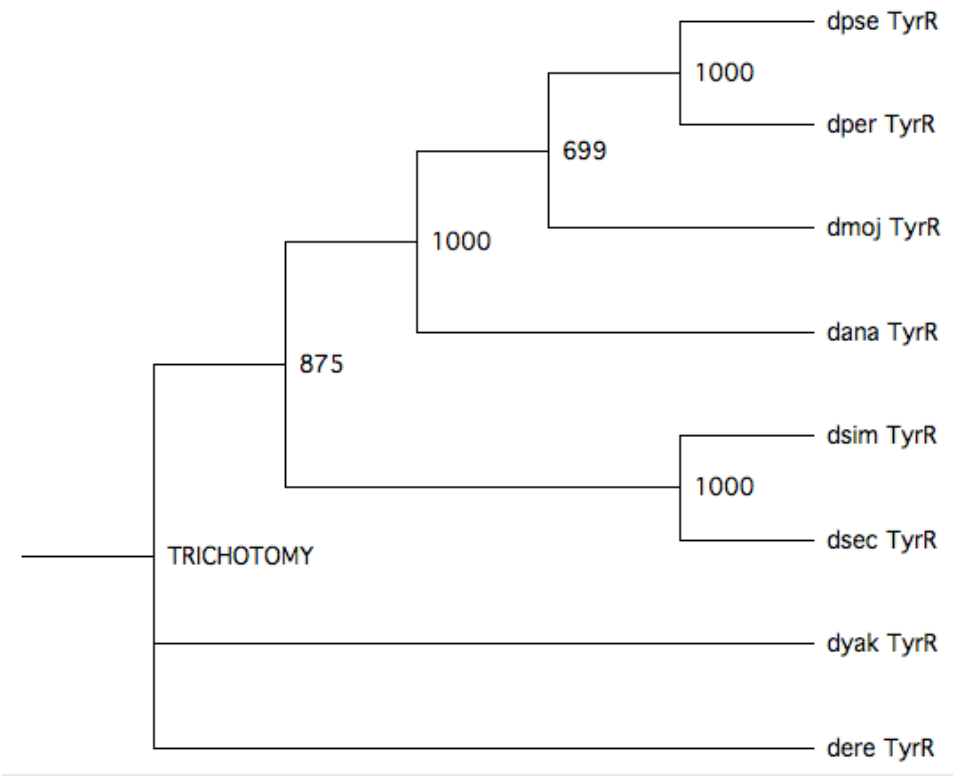


Fig.4. Tree view for Dere TyrR based on Clustal evidence.

Synonymous (K_S) and non-synonymous (K_A) substitution rates calcualted by codeml in the PAML package:

$K_S = 0.2348$

$K_A = 0.0120$

$K_A/K_S = 0.0511$

You can see the parameters used in codeml to calculate these values.

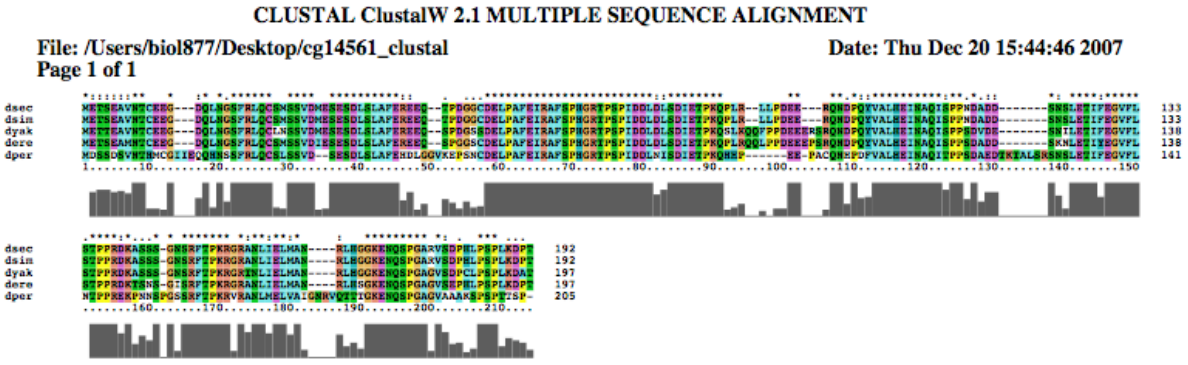
If you use these values, please cite the following peper:

- Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**:725-736.

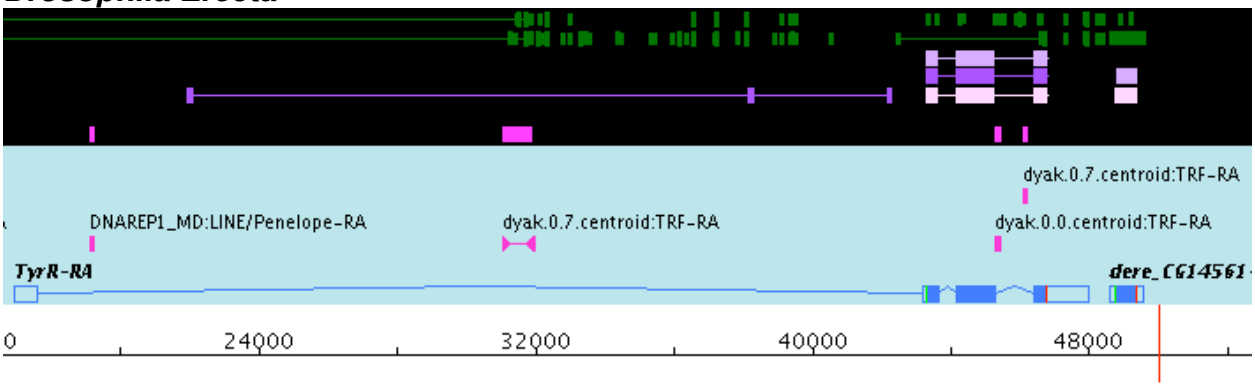
Fig 5. Ka/Ks analysis (Ka/Ks score = .0511) for TyrR indicates strong purifying selection

CG14561

Blast analysis shows these genes to be 88% identical and 92% similar. This is a one-exon gene with well-predicted splice sites and size. UTR structure is only slightly altered between species. Molecular function remains unknown. This seems a straight-forward example of a gene that is well-preserved and well-predicted.



Drosophila Erecta



Drosophila Melanogaster

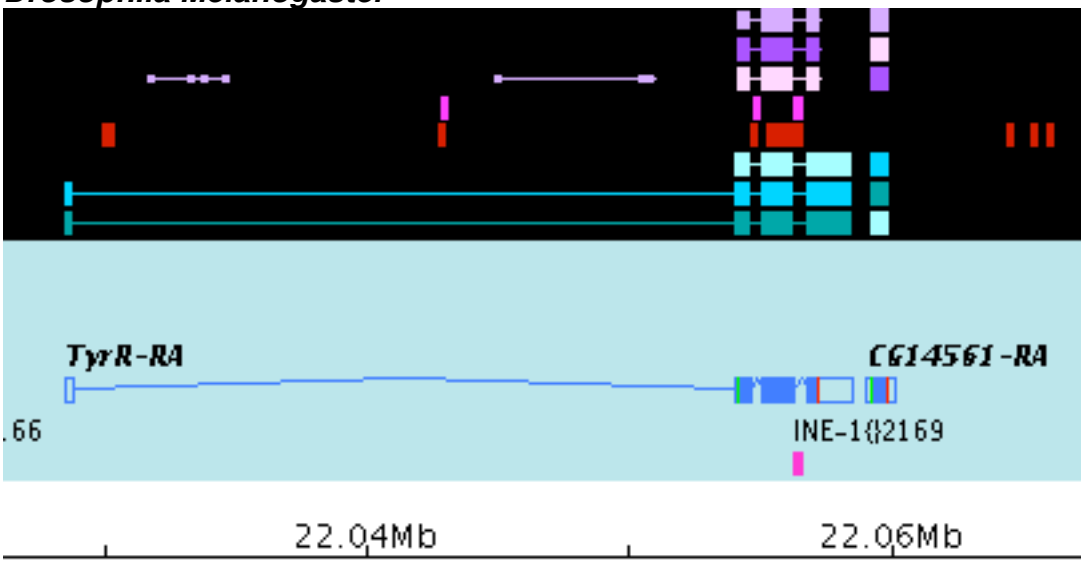


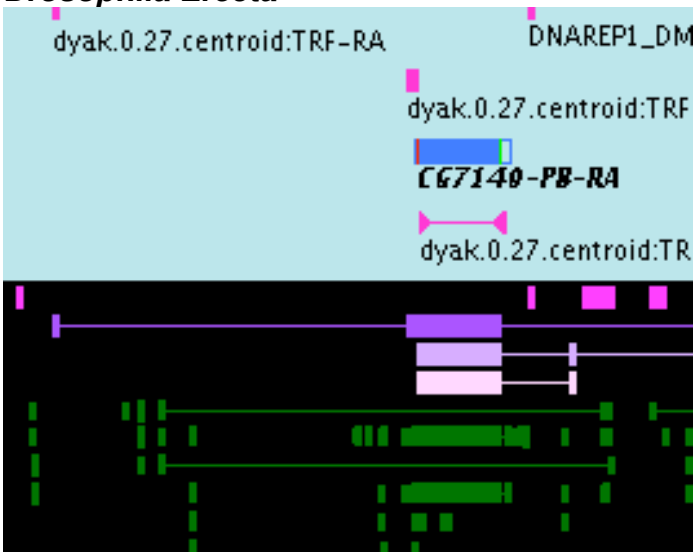
Fig. 7. Comparison between Dmel CG14561 and Dere CG14561

CG7140

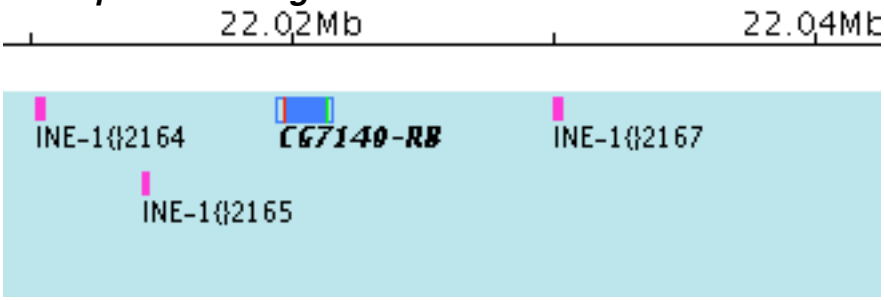
CG7140 is 96% similar and 93% identical between species. Molecular function is described as glucose-6-phosphate 1-dehydrogenase activity. The biological function is in glucose and monosaccharide metabolism. Gene predictors (Genscan) falsely predict multiple exons. Blast analysis indicates this is however a single exon gene. Interestingly

there is a high TE density upstream of CG7140 not seen in Dmel. cDNA evidence is present though splice sites aren't well defined. This gene was blasted against GleanR proteins and is well conserved between most Drosophilidae.

Drosophila Erecta



Drosophila Melanogaster



Flg 8. Comparison between CG7140 in Dere and Dmel, respectively

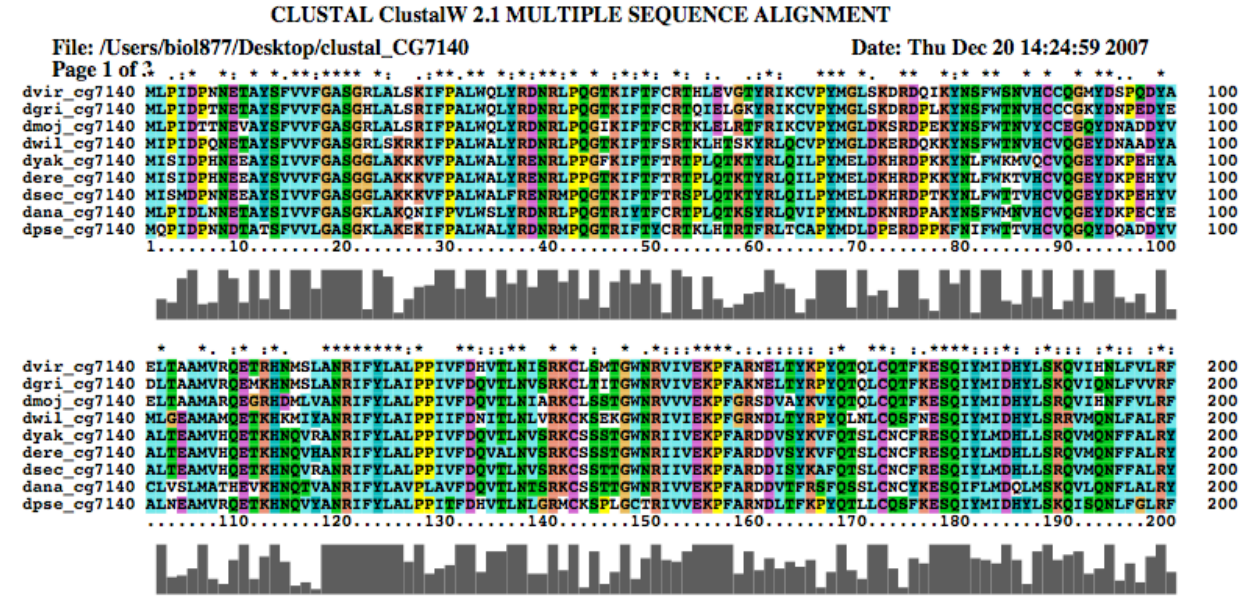


Fig 9. Clustal alignment for CG7140 between some members of drosophilidae

PAL2NAL output

```
   2     93
dmoj_cg7140
ATGCCCGACGTGGCTGCGATCCAGCTCGTAAATCTGCGTCTGCAGCTGGCACAGCGCAAG
GGTGTCGACCGCGTGGGCCCGATGGCCCGCGGC
dere_cg7140
ATGTCGGATGTGGCGGCACTGGAGCTGCACAATTTGCGGCTTCAACATGCTGCTCGCCGA
GCGCGTGATCGAGCTGGACCACGACATCGGGGA
```

Synonymous (K_S) and non-synonymous (K_A) substitution rates calcualted by codeml in the PAML package:

$K_S = 10.8878$

$K_A = 0.3395$

$K_A/K_S = 0.0312$

You can see the parameters used in codeml to calculate these values.

If you use these values, please cite the following peper:

- Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**:725-736.

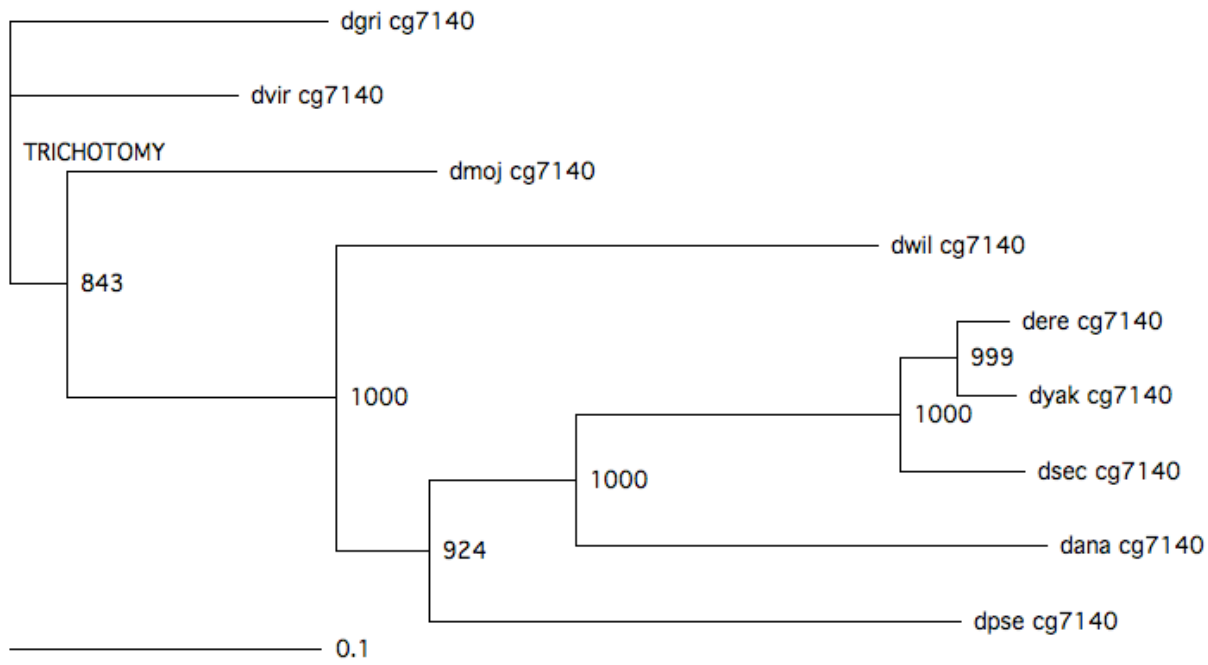
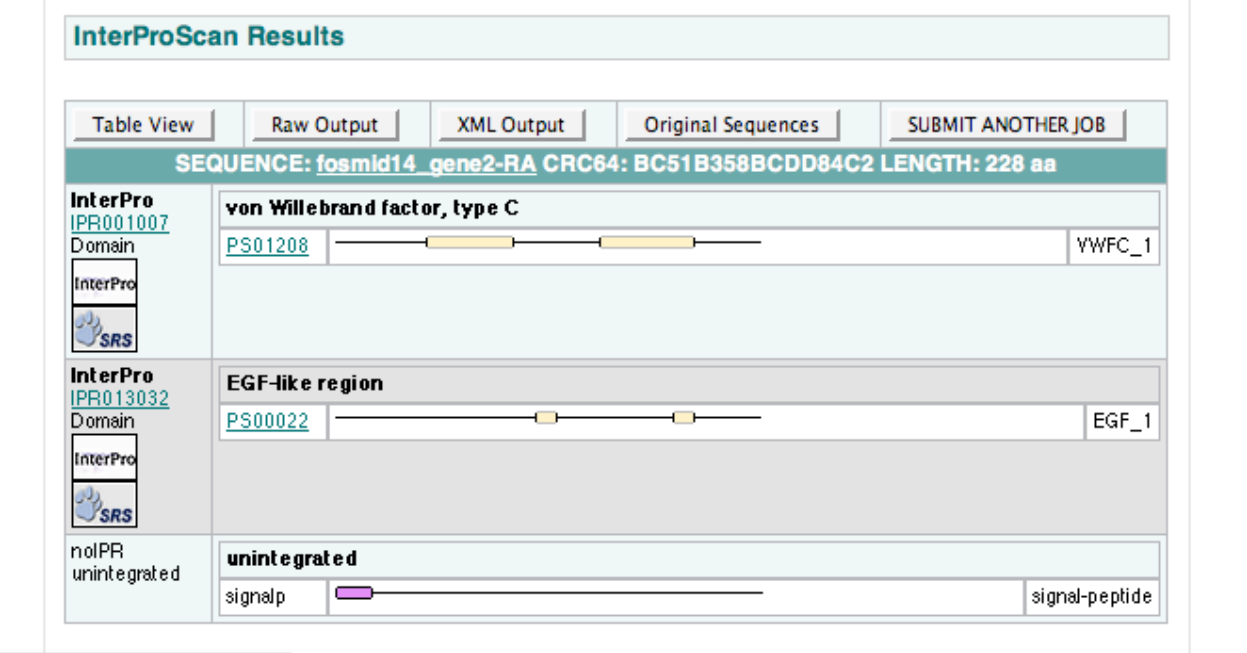


Fig 11. Tree view for CG7140 based on Clustal data.

Fosmid 14, Gene 2

This was the second predicted gene on fosmid 14 and the only one that was kept. Initial evidence came from *Genscan.* Blastp against all drosophilidae GleanR proteins were run on Flybase (www.flybase.org) with no hits found. Interproscan analysis however reveals some conserved functional domains (Fig. 5). Despite spotty cDNA evidence, the combination of gene prediction coupled with conserved protein domains seems enough to warrant keeping.



Flg 12. Interproscan results for fosmid14, gene2, illustrating conserved protein domains despite a lack of supporting Blast data.

What does it mean that there are no gleanR results against the near relatives and yet some conserved functional domains? This is a question without a good answer. In the absence of more information the gene is worth annotating.

Other predicted genes: Fosmid14 gene1 and gene3

While these predicted genes were eventually deleted, a brief discussion about the decision-making process is probably in order.

With respect to gene 3, if one compares the Dmel region just upstream of CG7140 and opposite TyrR it seems reasonable to think that fosmid14_gene3 might actually be dmel_CG7139. Working under this assumption several blasts were run, against dmel_CG7139, tblastn was run against dmel all translations and finally blastp against GleanR (Drosophilidae) was tested. All results indicated that indeed there is no correlation between the two genes. Fosmid14_gene3 nucleotide and peptide sequences were both run on interproscan to determine if any functional domains exist. Again no hits. This led to the conclusion that the region, for reasons unknown, was falsely predicted by Genscan. While a more detailed analysis of the Genscan algorithm might provide further insight it is beyond the scope of this paper.

Fosmid14_gene1 followed a similar treatment as gene3, though in this case it seemed less likely from the outset that this was a gene, as the Dmel region 5' of TyrR shows little activity (e.g. no repeats, TEs or genes). The results suggest that this too was a false-positive prediction.

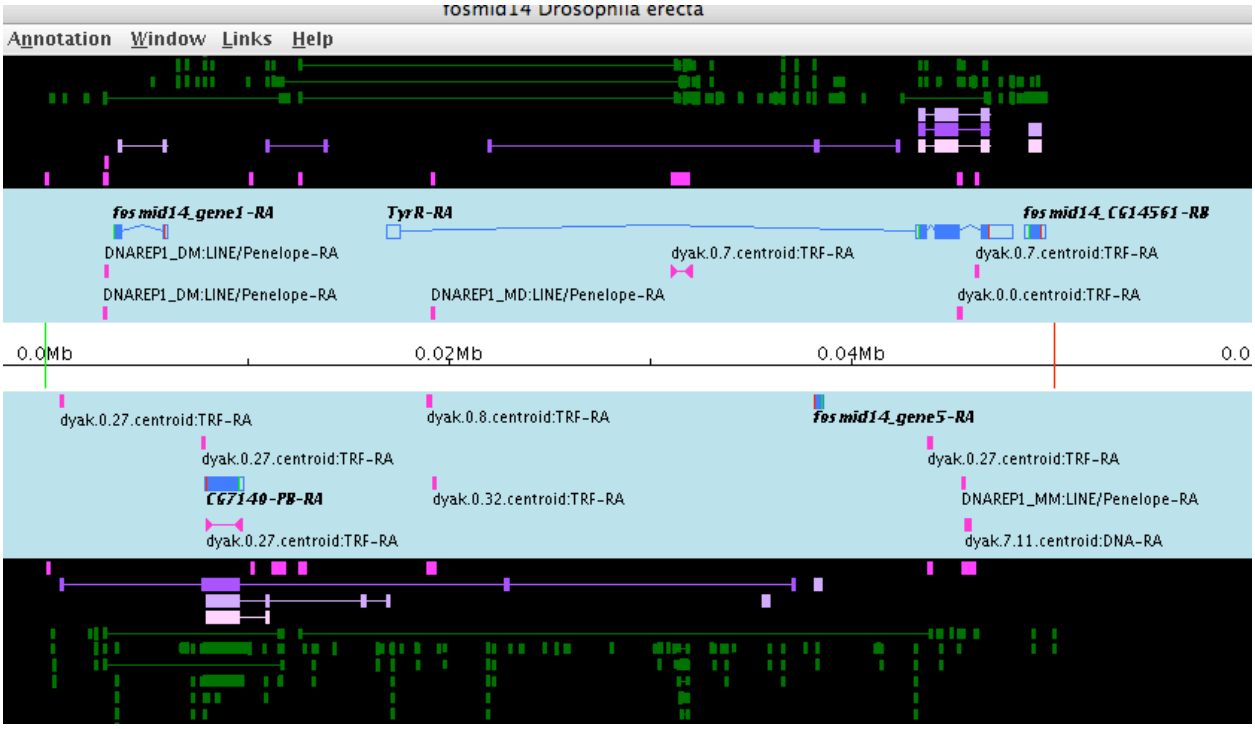


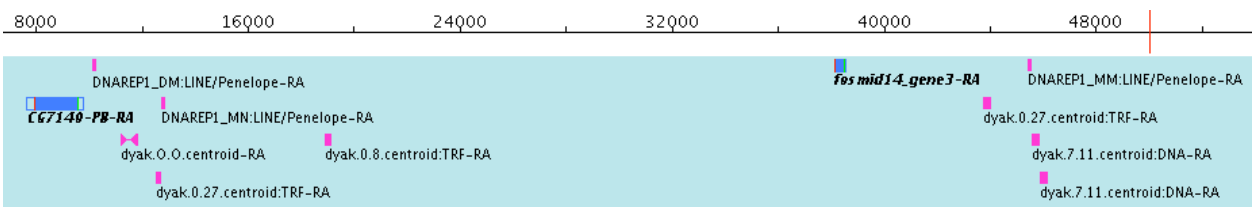
Fig 13. View of Predicted genes 1 and 3 (note: gene 5, bottom right is actually gene 3 misnamed)

Transposable Element structure

Analysis of TE density in and around the previously mentioned genes reveals some differences between Dmel and Dere. By examining the relative density of TE regions one observes a higher frequency of TEs in Dere on both the forward and reverse strands. Perhaps the most interesting finding with respect to TEs is a density just upstream of fosmid14, gene3 (since determined to be a false positive gene prediction).

One might hypothesize that this repeat density can be somehow correlated with Dmel CG7139, which is just upstream of CG7140. Might this be illustrative of the erosion of CG7139 as it is invaded by TEs? Might TEs simply manifest at the sites of gene erosion for other reasons? Such questions, while interesting, are beyond the scope of this paper.

Drosophila Erecta



```
                                                                    Smallest
                                                                      Sum
                                                                High  Probability
Sequences producing High-scoring Segment Pairs:                 Score  P(N)      N

fosmid14                                                          181  0.9997    2


>fosmid14
        Length = 50,000

  Plus Strand HSPs:

 Score = 181 (33.2 bits), Expect = 8.0, Sum P(2) = 0.9997
 Identities = 155/272 (56%), Positives = 155/272 (56%), Strand = Plus / Plus

Query: 73134 TTTCAAAGTAAATTCTGGTATT--TCCT-TATTTATATTTTTTAATACTACATCTGATAA 73190
             |||||||||||||| |||  |||  |  | || |||  |   |||| ||| ||||  |
Sbjct: 49737 TTTCAAAGTAAATCCTGAAATTGATTGTCTAATTAAGTAGGTTAA-ACTTCATCAAAAGT 49795

Query: 73191 ATCAATTTCATTCATGGAGTATAA-A-ACTTGATTTTTCTTTGAAATAAA-T-AATACTA 73246
              || | |  || | |  | || || | | |  ||  ||   || || ||  | || | ||
Sbjct: 49796 TTCCAATA-ATACGTTTACTAAAATAGAATATATAGTTTGATGTAACAATCTGAAAAATA 49854

Query: 73247 AGTCTATGTTGTTTTTTTCTGCTGTTTTTTTAGCAACTTCT-TTTATCTTGTTAAAAACA 73305
             | | || | |   ||  | ||  |  |||| |  ||| ||  |||||  |||    |  |
Sbjct: 49855 AATGTAAGATTAATTACTATGA-GCGTTTTCACGAACGTCAATTTATGATGTGCGTATTA 49913

Query: 73306 CTATCTAC-TGCGTCCCTATAACTTTCTTCCAGGTATGACATTTTTTTAGTGTAATTATC 73364
             ||| | |  | ||  |  | ||||||  |  |  ||| | ||  |||   | ||   ||
Sbjct: 49914 CTA-CGATGTTCGATCAAA-AACTTTTCTGAAAATATCA-ATAGTTTC--T-TACCAATA 49967

Query: 73365 TCTAGCATTTAAAGCTAGCTTGTTAAA-TTGC 73395
             |  | ||||||| | ||  || |  |  ||||
Sbjct: 49968 TG-A-CATTTAATGGTAATTTTTGGACCTTGC 49997

 Score = 127 (25.1 bits), Expect = 8.0, Sum P(2) = 0.9997
 Identities = 143/248 (57%), Positives = 143/248 (57%), Strand = Plus / Plus

Query: 61701 TGAAGAAGTATATA-TACTTTTTAAGTATAT-GTAACCATAATACATCTCAAAATACAAA 61758
             || ||  | ||| | || ||||||||||||| ||| |  |||   |||| |||   ||
Sbjct:  7526 TGTAGCCG-ATAAAATAGTTTTTAAGTATATAGTATCTTTAAGT-ATCTGAAAGAGCATT 7583

Query: 61759 GTTCAAC-AG--T-TA-ATT-GT-GTGAAAGATAAAATTGCTAATGGCT-ACTTAAATCC 61810
             || |||  |   | || ||  || | || ||  |  || || ||    | | || |
Sbjct:  7584 GTGCAATTAAACTCTACATCAGTTGCGACAGC-AGCAT-GCAAAACATTGAATTGACAAG 7641
```

D.mel	D.ere
forward strand:	
5' of TyrR:	
INE-1{}2161	*DNAREP1_DM:LINE/Penelope-RA*
INE-1{}2163	*dyak.0.40.centroid:TRF-RA*
INE-1{}2166	
TyrR bet. exon 1-2	
gene?:CG:temp1:3L_22000000_22100000-RA	*DNAREP1_MD:LINE/Penelope-RA*
	dyak.0.7.centroid:TRF-RA
TyrR bet. exon 3-4	
INE-1{}2169	dyak.0.0.centroid:TRF-RA
	dyak.0.7.centroid:TRF-RA
Reverse Strand:	
3' of (novel) fosmid14-gene3	
none	DNAREP1_MM:LINE/Penelope-RA
	dyak.0.27.centroid:TRF-RA, genomic range (45855-45567)
	dyak.7.11.centroid:DNA-RA, genomic range (46138-45899)
Between gene3 and CG7140	
INE-1{}2167	dyak.0.32.centroid:TRF-RA
	dyak.0.8.centroid:TRF-RA
	dyak.0.27.centroid:TRF-RA
	dyak.O.O.centroid-RA
	DNAREP1_MN:LINE/Penelope-RA, genomic range (12820-12757)
	DNAREP1_DM:LINE/Penelope-RA, genomic range (10194-10155)
5' of CG7140	
INE-1{}2165	dyak.0.27.centroid:TRF-RA
INE-1{}2164	
INE-1{}2155	

```
==================================================
file name: fosmid14.fasta
sequences:             1
total length:      50000 bp  (50000 bp excl N/X-runs)
GC level:         43.77 %
bases masked:       3821 bp (  7.64 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
LINEs:                 8         898 bp     1.80 %

DNA elements:          2         311 bp     0.62 %

Total interspersed repeats:     1209 bp     2.42 %

Simple repeats:        2          64 bp     0.13 %
Low complexity:        6         176 bp     0.35 %
==================================================

* most repeats fragmented by insertions or deletions
  have been counted as one element
```

```
==================================================
file name: fosmid14.fasta
sequences:             1
total length:      50000 bp  (50000 bp excl N/X-runs)
GC level:         43.77 %
bases masked:       3821 bp (  7.64 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
LINEs:                 8         898 bp     1.80 %

DNA elements:          2         311 bp     0.62 %

Total interspersed repeats:     1209 bp     2.42 %

Simple repeats:        2          64 bp     0.13 %
Low complexity:        6         176 bp     0.35 %
==================================================

* most repeats fragmented by insertions or deletions
  have been counted as one element
```

References:

Dr. Chris Smith lecture Notes.
Biol 871, Bioinformatics, Fall 2007
SFSU Biology Department