Biology 191 Annotation Report

Student names: _____Key_____ **Lab Section:** _____

Gene/Isoform report form - fill in the requested information:

Gene name: _____D. biarmipes yellow-h_____
Isoform symbol: _____Dbia yellow-h-PA_____
Size of project in base pairs (from beginning of start codon through end of stop codon):
_____9,866 base pairs (bp)_____
Is the 5' end of this isoform missing from the end of project: _____no_____
If so, how many exons are missing from the 5' end: _____
Is the 3' end of this isoform missing from the end of the project: _____no_____
If so, how many exons are missing from the 3' end: _____

1. Individual CDS BLASTX Results

Paste the best alignment for each CDS below:

CDS 1:

```
>lcl|37865 yellow-h:1_1556_0
Length=78

Score = 77.8 bits (190),  Expect = 2e-20, Method: Compositional matrix adjust.
Identities = 46/78 (59%), Positives = 52/78 (67%), Gaps = 6/78 (8%)
Frame = -1

Query  41412  MISLPRFCRFTMFIFFLKT--GNFIVQPVSQKLEDNEYAQKS----LQSESKLEVVNEWK  41251
              M S+  F   T  +F  KT  GN  VQPV Q L+  EY  +S    LQSES+LE+V EWK
Sbjct  1      MQSMTIFNIITQLVFLSKTLNGNLSVQPVFQTLDGYEYTSQSFSQNLQSESQLEIVYEWK  60

Query  41250  YLDFEYSTFVQRQQSILN  41197
              YLDF YSTFVQRQQSILN
Sbjct  61     YLDFLYSTFVQRQQSILN  78      Full-length
41,195, phase 2
```

CDS 2:

```
>lcl|37181 yellow-h:2_1556_1
Length=222

Score =   422 bits (1084),  Expect = 2e-137, Method: Compositional matrix adjust.
Identities = 198/222 (89%), Positives = 211/222 (95%), Gaps = 0/222 (0%)
Frame = -1
```

33,589 phase 1
Two AG's in a row: second does not have correct frame and will eliminate one aa

```
Query  33588  DFVPKNNLPLGIDVHQNRLFVTTPRWKDGVPASLGTIPFPPTESSPAIRPYPNWEAHGNP  33409
              DFVPKNNLPLGIDVH NRLFVTTPRWK+GVPASLGT+PFPP ESSPAI+PYPNWEAHGNP
Sbjct  1      DFVPKNNLPLGIDVHNNRLFVTTPRWKNGVPASLGTLPFPPKESSPAIKPYPNWEAHGNP  60
```

```
Query  33408  KNPDCLKLMSVYRTAVDRCQRIWIIDSGIVNATVNLNQICPPKIVVYDLKKDELIIRYNL  33229
               NPDC KLMSVYRTAVDRC RIW+IDSGIVNAT+NLNQICPPKIVVYDLK DELI+RYNL
Sbjct  61     NNPDCSKLMSVYRTAVDRCDRIWLIDSGIVNATINLNQICPPKIVVYDLKSDELIVRYNL  120

Query  33228  EASQVKQDSLHSNIVVDIGDHCDDAHAIVSDVWRFGLVVYSLSKNRSWRVTNYNFYPDPV  33049
               EAS VKQDSLHSNIVVDIG+ CDDAHAIVSDVWRFGL+VYSLSKNRSWRVTNYNFYPDP
Sbjct  121    EASHVKQDSLHSNIVVDIGEDCDDAHAIVSDVWRFGLLVYSLSKNRSWRVTNYNFYPDPF  180

Query  33048  ASDFNIYGLNFQWLDGVFGMTISYNENMMQRVLYFHPMASFK  32923
               ASDFN+YGLNFQWLDGVFGM+I YN+ +M+RVLYFHPMASFK
Sbjct  181    ASDFNVYGLNFQWLDGVFGMSIYYNKKIMERVLYFHPMASFK  222   Full-length
```

32,923 phase 0

CDS 3:

```
>lcl|59669 yellow-h:3_1556_0
Length=163

Score =   261 bits (668),  Expect = 1e-82, Method: Compositional matrix adjust.
Identities = 123/164 (75%), Positives = 143/164 (87%), Gaps = 3/164 (2%)
Frame = -1
```

32,031 phase 0

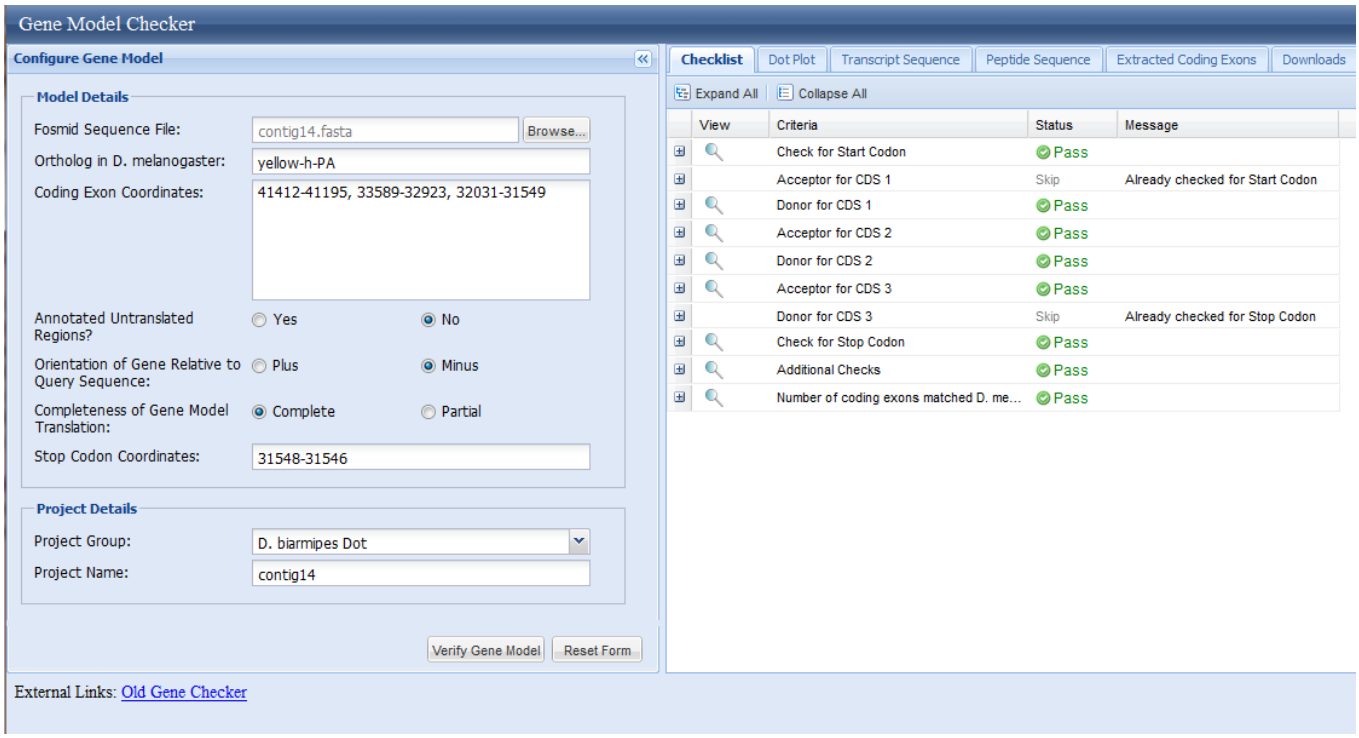
Two AG's separated by one bp: second HAS correct frame but will eliminate one amino acid

```
Query  32031  EFMVPMDLLLNESLWKSNNQDNAKYFFSIGDRGYNSQSSTSAITRSGVMFFTQVHQDNIG  31852
              EFMVPM++LLNES+W++N Q+ AKYF  IGDRGYNSQSSTS +TR+G+MFFTQVHQD+IG
Sbjct  1      EFMVPMNILLNESVWQTNTQEYAKYFIPIGDRGYNSQSSTSGVTRNGIMFFTQVHQDDIG  60

Query  31851  CWDTSKPYTRAHIERF--LENGPNLIQFPNDLKVDNEDDQSIWIISNRLPIFLYSNLDYG  31678
              CWDTSKPYTRAH+ +F  +EN  NLIQFPNDLKVD E DQ++W+ISNRLPIFLYSNLDYG
Sbjct  61     CWDTSKPYTRAHLGKFHNMENS-NLIQFPNDLKVDKEKDQNVWLISNRLPIFLYSNLDYG  119

Query  31677  EINFRILKVKVKTAISNSICNPENRYINGSKSTFVLIEEGQCY*  31546
              E+NFRILK  V   I NS+CNP+N YIN SKS FVLIEEGQC+*
Sbjct  120    EVNFRILKANVNKIIRNSVCNPDNSYINTSKSAFVLIEEGQCF*  163   Full-length
```

31,549 Stop at 31,548-31,546

2. Gene Model Checker checklist

Enter the coordinates of your final gene model for this isoform into the Gene Model Checker. Take a screenshot, including the Checklist results, by simultaneously pressing the Ctrl, Alt and Print Screen buttons. **Paste the screenshot below (and then use the Crop tool under Format to re-size the image showing just the pertinent parts):**



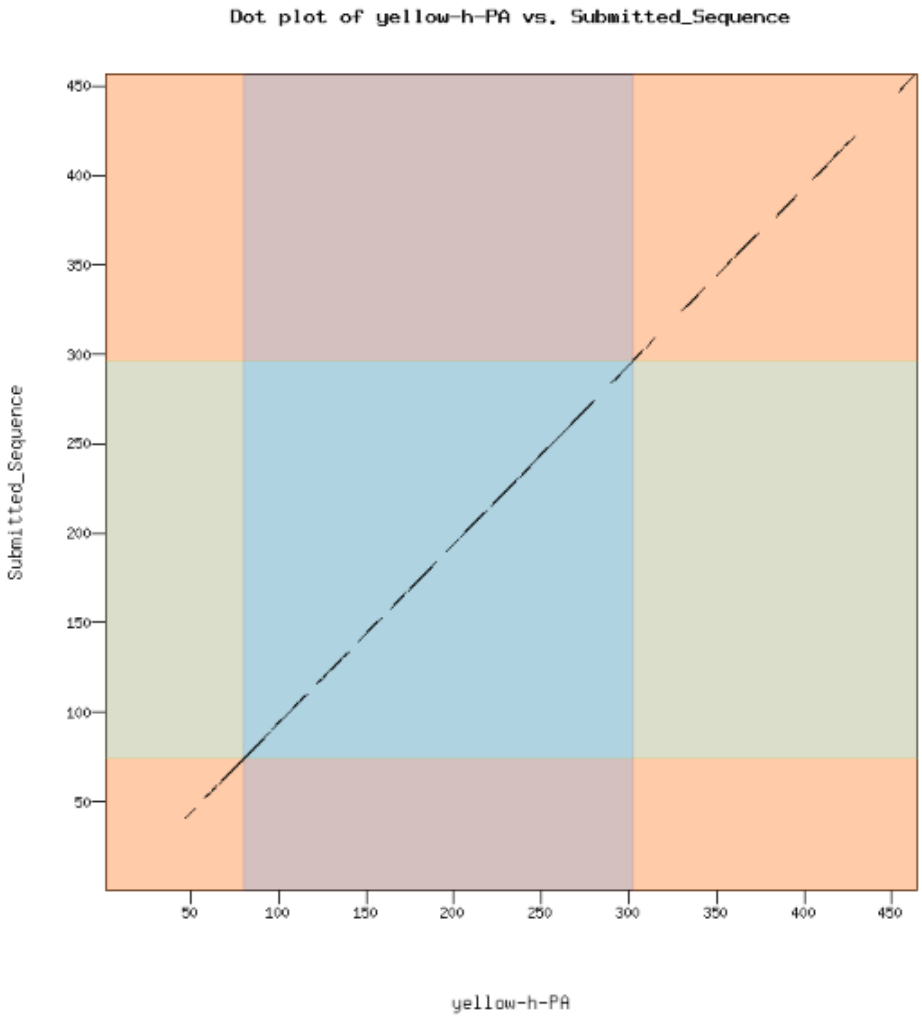
3. View the predicted polypeptide sequence

Click on the 'Peptide Sequence' tab in the right window of the Gene Model Checker to see the amino acid sequence of the polypeptide encoded by your gene model coordinates. **Paste the amino acid sequence below:**

```
>yellow-h-PA_peptide
1     MISLPRFCRFTMFIFFLKTGNFIVQPVSQKLEDNEYAQKSLQSESKLEVVNEWKYLDFEY
61    STFVQRQQSILNGDFVPKNNLPLGIDVHQNRLFVTTPRWKDGVPASLGTIPFPPTESSPA
121   IRPYPNWEAHGNPKNPDCLKLMSVYRTAVDRCQRIWIIDSGIVNATVNLNQICPPKIVVY
181   DLKKDELIIRYNLEASQVKQDSLHSNIVVDIGDHCDDAHAIVSDVWRFGLVVYSLSKNRS
241   WRVTNYNFYPDPVASDFNIYGLNFQWLDGVFGMTISYNENMMQRVLYFHPMASFKEFMVP
301   MDLLLNESLWKSNNQDNAKYFFSIGDRGYNSQSSTSAITRSGVMFFTQVHQDNIGCWDTS
361   KPYTRAHIERFLENGPNLIQFPNDLKVDNEDDQSIWIISNRLPIFLYSNLDYGEINFRIL
421   KVKVKTAISNSICNPENRYINGSKSTFVLIEEGQCY
```

4. Dot plot between the submitted model and the *D. melanogaster ortholog*

Click on the 'Dot Plot' tab in the right window of the Gene Model Checker to view a schematic alignment of the amino acid sequence encoded by your submitted gene model against the sequence of the putative *D. melanogaster* yellow-h ortholog (generated by the Gene Model Checker). **Paste a copy of the dot plot below**:



Provide an explanation for any anomalies in the dot plot (e.g. small or large gaps, such as at the lower-left end of the line):

The line is quite straight, indicating a very good alignment with no major structural differences between the two genes (e.g., the three coding blocks are about the same length in both species). The small gaps within the line can be attributed to short gaps (---) within the coding blocks or individual amino acid mismatches (see the alignment in the next section for a detailed view of these). The left end of the line is missing because there is much less sequence similarity of the first 45 amino acids of the polypeptide chain than in the rest of the alignment (see next section). Since these gaps or mismatches are present in the alignments of the individual CDS (see results of individual CDS-by-CDS alignments), evolution has brought about these differences in the time since the two species diverged. In other words, if the annotation has been done correctly for this gene, all differences in the two amino acid sequences can be accounted for in the initial CDS-by-CDS alignments (and NOT by student errors).

Note: Large <u>vertical and horizontal gaps</u> near exon boundaries in the dot plot often indicate that an incorrect splice site might have been picked. Please re-examine these regions (if any) and provide a detailed justification as to why you have selected this particular set of donor and acceptor sites.

This only applies if a student selects incorrect exon/intron boundaries.

5. Alignment between the *D. melanogaster ortholog* and the submitted model

Click on the 'View protein alignment' link in the Dot Plot window to see a detailed alignment of each amino acid from the putative *D. melanogaster* yellow-h ortholog versus the *D. biarmipes* amino acid sequence predicted from your model. **Copy and paste the colored version of the alignment below:**

Alignment of yellow-h-PA vs. Submitted_Seq

View plain text version

Identity: 367/464 (79.1%), **Similarity:** 406/464 (87.5%), **Gaps:** 9/464 (1.9%)

```
yellow-h-PA         1 MQSMTIFNIITQLVFLSKTLNGNLSVQPVFQTLDGYEYTSQSFSQNLQSESQLEIVYEWK 60
                      * *:. *  :* ::*: **  **: **** *.*:. **:.:*    *****:**:* ***
Submitted_Seq       1 MISLPRFCRFTMFIFFLKT--GNFIVQPVSQKLEDNEYAQKS----LQSESKLEVVNEWK 54

yellow-h-PA        61 YLDFLYSTFVQRQQSILNGDFVPKNNLPLGIDVHNNRLFVTTPRWKNGVPASLGTLPFPP 120
                      **** ******************************:***********:********:****
Submitted_Seq      55 YLDFEYSTFVQRQQSILNGDFVPKNNLPLGIDVHQNRLFVTTPRWKDGVPASLGTIPFPP 114

yellow-h-PA       121 KESSPAIKPYPNWEAHGNPNNPDCSKLMSVYRTAVDRCDRIWLIDSGIVNATINLNQICP 180
                      .******:***********:**** *************:***:*********:*******
Submitted_Seq     115 TESSPAIRPYPNWEAHGNPKNPDCLKLMSVYRTAVDRCQRIWIIDSGIVNATVNLNQICP 174

yellow-h-PA       181 PKIVVYDLKSDELIVRYNLEASHVKQDSLHSNIVVDIGEDCDDAHAIVSDVWRFGLLVYS 240
                      *********.****:*******:***************:.*****************:***
Submitted_Seq     175 PKIVVYDLKKDELIIRYNLEASQVKQDSLHSNIVVDIGDHCDDAHAIVSDVWRFGLVVYS 234

yellow-h-PA       241 LSKNRSWRVTNYNFYPDPFASDFNVYGLNFQWLDGVFGMSIYYNKKIMERVLYFHPMASF 300
                      ******************.*****:**************:* **:::*:***********
Submitted_Seq     235 LSKNRSWRVTNYNFYPDPVASDFNIYGLNFQWLDGVFGMTISYNENMMQRVLYFHPMASF 294

yellow-h-PA       301 KEFMVPMNILLNESVWQTNTQEYAKYFIPIGDRGYNSQSSTSGVTRNGIMFFTQVHQDDI 360
                      *******::*****:*::*.*: ****:.*************.:**.*:*********:*
Submitted_Seq     295 KEFMVPMDLLLNESLWKSNNQDNAKYFFSIGDRGYNSQSSTSAITRSGVMFFTQVHQDNI 354

yellow-h-PA       361 GCWDTSKPYTRAHLGKFHNMENS-NLIQFPNDLKVDKEKDQNVWLISNRLPIFLYSNLDY 419
                      *************: :*  :**. ************:*.**.:*:***************
Submitted_Seq     355 GCWDTSKPYTRAHIERF--LENGPNLIQFPNDLKVDNEDDQSIWIISNRLPIFLYSNLDY 412

yellow-h-PA       420 GEVNFRILKANVNKIIRNSVCNPDNSYINTSKSAFVLIEEGQCF 463
                      **:******.:*:. * **:***:* *** ***:*********:
Submitted_Seq     413 GEINFRILKVKVKTAISNSICNPENRYINGSKSTFVLIEEGQCY 456
```

6. Analysis of the Alignment

- What does a * (star) in the middle row indicate?

  An identical amino acid at that position of the polypeptide chain in the two species

- What does it mean when there is no symbol at all in the middle row between two aligned amino acids?

  Spaces mark mismatched amino acids that are chemically and/or physically dissimilar, such as isoleucine [I, with a nonpolar side chain] and glutamine [Q, with a polar $-NH_2$ group].

  Note that instead of using a '+' for mismatched but chemically- or physically-similar amino acids (as was the case in the previous BLAST alignments), this program uses a : (for more similar) or a . (for less similar). See the handout for a Venn diagram that shows these relationships between the different amino acids.

- What do dashes (---) in one sequence or the other indicate?

  That those amino acids do not exist in the polypeptide encoded by the *D. melanogaster yellow-h* gene or the predicted protein of *D. biarmipes*, respectively. [Note the *D. biarmipes* sequence is the bottom Sbjct sequence and the *D. melanogaster* polypeptide is the top sequence in this alignment.] Students should understand that these same gaps are not in the polypeptide chains (otherwise they would be in pieces!), but have been introduced by the alignment software to maximize sequence alignments in the rest of the polypeptide chain.

- What do the different colored, paired blocks of amino acids represent (e.g., the top orange/light blue, the red/dark-blue and the bottom orange/light-blue blocks)?

  The first, second and third coding blocks, which were 78, 222 and 163 amino acids long in the *D. melanogaster yellow-h* orthologue, respectively. Note that the Gene Model Checker software spliced out the introns and translated the resultant mRNA to create one, contiguous polypeptide chain☺!

- If you carefully count the number of amino acids in the top orange/blue block, it is 79 amino acids long. This is one more than the number of amino acids in the first CDS of the *D. melanogaster* ortholog that you used for the initial blastx alignment. Where did this 79th amino acid come from?

  From the amino acid that is encoded after splicing by a codon that is split in two pieces in the genomic DNA by exon/intron boundaries at either end of the first intron. This results from exon/intron boundaries with phases other than zero. The phases at the ends of the first intron must add up to three, to create a codon that maintains the correct reading frame in the transition from CDS 1 to CDS 2 in the spliced mRNA.

- How is the above alignment with the protein predicted by your gene model different (and hopefully more precise) than the first one you analyzed on p. 23 of the lab?

  The BLAST alignment on p. 22 is shorter, as it is missing the first 21 and last 4 amino acids of the *Drosophila melanogaster* yellow-h protein. Since BLAST looks for local regions of

similarity, it likely left off these regions, since to include them probably lowered the overall score (due to how different these regions are in the two species). Yet more evidence in favor of human brains to create a better gene model than computers!

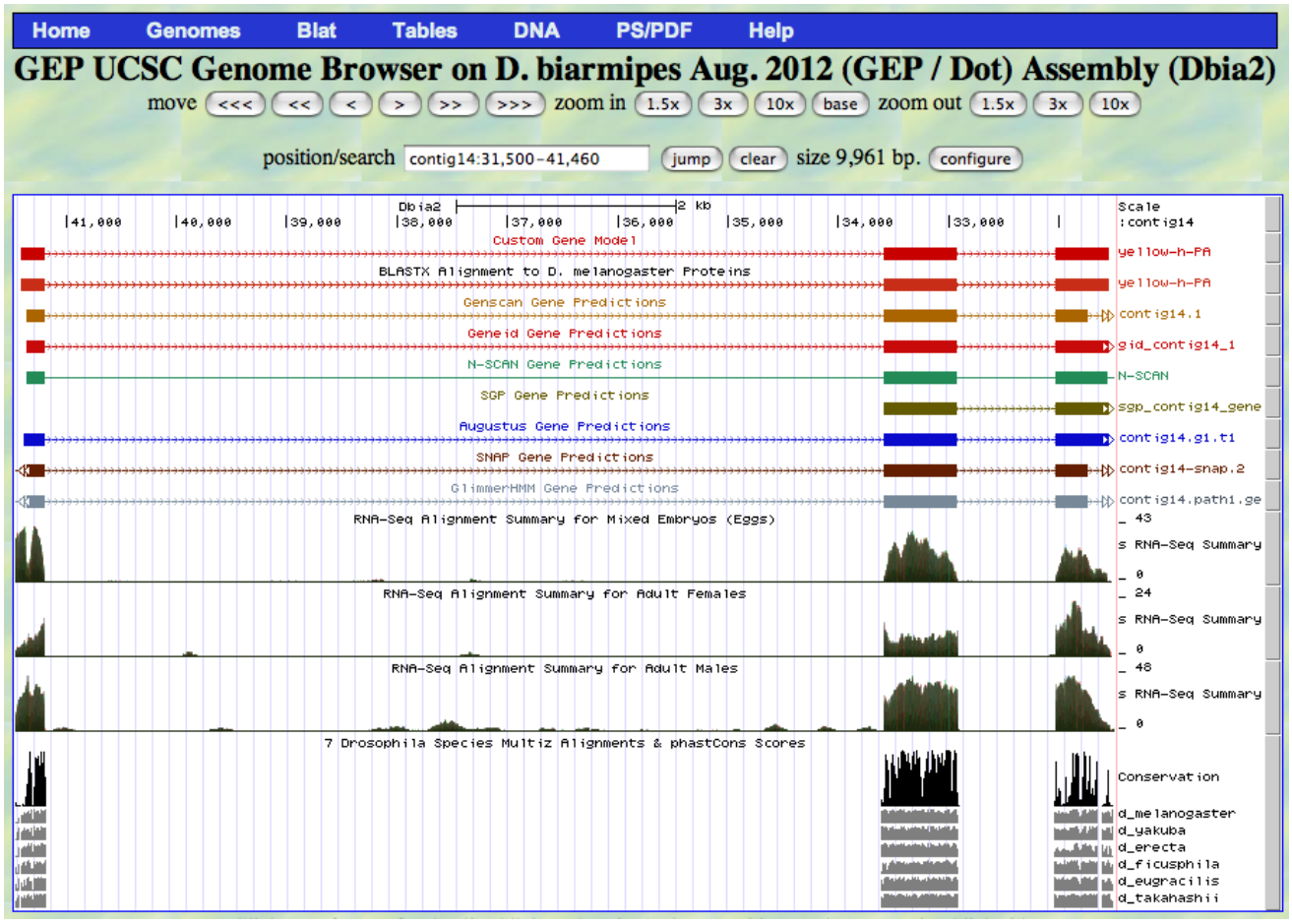
7. View the gene model on the Genome Browser

Using the custom track feature from the Gene Model Checker (see page 10 of the Gene Model Checker User Guide on how to do this, if need be). Capture a screenshot of your gene model shown on the Genome Browser for your project; zoom in (by increasing the overall length of your gene model by ~50 bp at both ends) so that your gene model fills the Browser window from left to right. Include the following evidence tracks in the screenshot:

1. A sequence alignment track (D. mel Proteins), set to pack
2. The three or four Gene Prediction Tracks that best support your model (e.g., Genscan, N-SCAN, etc.,) set to dense
3. At least one RNA-Seq Track (e.g., RNA-Seq Alignment Summary), set to show
4. A comparative genomics track [e.g., Conservation), set to full

Click the 'reverse' button below the Browser window so the gene model runs from left to right in the 5' to 3' direction.

Paste the screenshot of your gene model as shown on the Genome Browser below:



Comment on the quality of your predicted gene model.

If there are any discrepancies between the different Gene Prediction tracks, discuss which discrepancies you think support or do not support your particular gene model and whether or not you consider your model more accurate than some (or all!) of the gene predictors.

The gene model is a much better model than those in any of the gene predictor tracks, which have missing or truncated exons (as well as a fusion of this feature with the CG31999 one). The gene model is a very close match to the BLASTX alignment. However, we know that the gene model is even better than the BLASTX alignment, since when you zoomed in to the ends of the individual CDS on pp. 29-30, you saw that the exon length was off by one base at the end of the first exon and beginning of the second exon in the BLASTX track. Thus, human brains prevail once again!!!