12-2020

Genomics Education Partnership F Element Annotation Report

Amanda Moy
Grand Valley State University

Genomics Education Partnership F Element Annotation Report

Amanda Moy

Advisor: Dr. Martin Burg, Department Biomedical Sciences, Department of Cell and Molecular Biology

The Genomics Education Partnership (GEP), headquartered at the University of Alabama, is a collection of over 100 universities that provide training and resources in order to provide students experiential learning in bioinformatics and genomics (1). The GEP hosts numerous research projects, including the F element project. The F element project has the main focus of annotating the F element genes of the fruit fly species *D. ananassae*, *D. bipectinata*, *D. kikkawai*, and *D. takahashii* (2). The Muller F element is the smallest chromosome in *Drosophila* species. However, the four species listed above have a notably larger F element than other *Drosophila* species. Therefore, this project seeks to explore the expansion of the F element in these species in order to determine the evolutionary impacts of the changes in exon, intron, and chromosome size.

Genome annotation is the process of identifying the location of genes within the genome. My project focused on annotating a region of the Muller D element, which serves as a control and reference region for the F element research project. I examined contig26 of *D. ananassae*, with contig referring to a region of a whole genome. I located the intron/exon boundaries of the coding genes of contig26, as well as the transcription start sites of those genes. This data will be submitted to the Genomics Education Partnership to be used as part of the F element research project, and essentially allows the GEP to know the location of all coding genes within contig26 of the *D. ananassae* genome.

My process for genome annotation starts by identifying a noted feature within the contig using the GEP USCS Genome Browser (3). This means that a computer program believes that there is a coding gene in this region of the contig (though that isn't always the case). I then use Flybase Blast to determine if there is a coding gene in the feature, and the identity of any gene that might be present in that feature. Once I have identified a match to a gene, I can compare the sequence of the *D. ananassae* genome to the gene sequence in *D. melanogaster*, and determine the intron/exon boundaries of the gene. *D. melanogaster* is used because it serves as a model species of genetics work and the genome of the species has already been sequenced and analyzed. Once all intron/exon boundaries are found, the coordinates of the gene are checked in Gene Model Checker, which makes sure that the motifs needed on either end of an intron/exon boundary are present within the gene model I created. Then, I can search for the transcription start site of the gene using RAMPAGE, RNA-Seq, ATAC-Seq, and NCBI BLAST data (4,5). The following report is in the GEP format and contains the annotation data for contig26 of *D. ananassae*.

Literature Cited

1. "Genomics Education Partnership (GEP) Mission." *Genomics Education Partnership*, https://thegep.org/about/.
2. "Expansion of Drosophila Muller F Element." *Genomics Education Partnership*, https://thegep.org.
3. Leung, Wilson. *Annotation of a Drosophila Gene*. Genomics Education Partnership, 31 Aug. 2020.

4. *TSS Annotation Workflow*. Genomics Education Partnership, https://community.gep.wustl.edu/repository/course_materials_WU/annotation/Search_TSS_Workflow.pdf.
5. Leung, Wilson. *Annotation of Transcription Start Sites in Drosophila*. Genomics Education Partnership, 20 Aug. 2020.

Relevant Literature

1. Leung, Wilson, et al. “Drosophila Muller F Elements Maintain a Distinct Set of Genomic Properties Over 40 Million Years of Evolution.” *G3: Genes, Genomes, Genetics*, vol. 5, no. 5, G3: Genes, Genomes, Genetics, May 2015, pp. 719–40. *www.g3journal.org*, doi:10.1534/g3.114.015966.
2. Slawson, Elizabeth E., et al. “Comparison of Dot Chromosome Sequences from D. Melanogaster and D. Virilis reveals an Enrichment of DNA Transposon Sequences in Heterochromatic Domains.” *Genome Biology*, vol. 7, no. 2, Feb. 2006, p. R15. *BioMed Central*, doi:10.1186/gb-2006-7-2-r15.
3. Wangler, Michael F., et al. “Fruit Flies in Biomedical Research.” *Genetics*, vol. 199, no. 3, Genetics, Mar. 2015, pp. 639–53. *www.genetics.org*, doi:10.1534/genetics.114.171785.
4. Lopatto, David, et al. “A Central Support System Can Facilitate Implementation and Sustainability of a Classroom-Based Undergraduate Research Experience (CURE) in Genomics.” *CBE—*

Life Sciences Education, vol. 13, no. 4, American Society for Cell Biology (lse), Dec. 2014, pp. 711–23. *lifescied.org (Atypon)*, doi:10.1187/cbe.13-10-0200.

F Element Project: Annotation Report

Faculty instructor(s): Dr. Martin Burg
College/university: Grand Valley State University
Course number: HNR 499
Course name: Honors Senior Project

Authorship Information for GEP Scientific Publications

☒ By checking this box, I/we grant permission for the Genomics Education Partnership (GEP) to use the annotation data produced in this report in future scientific publications.

Note: Please skip the rest of this section if **more than three students** contribute to this annotation report. When more than three students contribute to an annotation project, the class as a whole will be acknowledged in future GEP scientific publications.

Co-authors Responsibilities

In order to be a co-author on a GEP publication, you must review, critique, and approve the final gene models and manuscript, responding promptly to requests to read and approve. As part of the preparations for the microPublication article, co-authors are required to validate specific data within the manuscript, supplemental materials, and GenBank submission (the specific details will depend on each annotation project). In most cases, the manuscript preparation process will take approximately 3–5 hours of your time.

The above requirements mean that we must be able to contact you when the GEP microPublication, and later, the scientific paper with meta-analysis, is ready for your review and approval. **If we cannot reach you at that time, you will not be a co-author on our GEP scientific publications,** as scientific journals require all co-authors to have read and approved the manuscript.

Please provide your contact information below. Note that your name and contact information will be publicly available through the scientific publication and the GenBank record (this is standard for all scientific publications.). Please list the authors in ascending alphabetical order by last name. (The actual order of the student co-authors in the scientific publication will be determined by a random number generator.)

Contact information for Author #1

First name	Amanda
Middle initials	L
Last name	Moy
Author name (name that will appear on the publication):	Amanda Moy
Permanent Email address (one you will use five years from now):	amandamoy7@gmail.com
Alternative Email address (optional):	
Check this box to indicate that you have read and accept the co-authors responsibilities	☒

Contact information for Author #2

First name	
Middle initials	
Last name	
Author name (name that will appear on the publication):	
Permanent Email address (one you will use five years from now):	
Alternative Email address (optional):	
Check this box to indicate that you have read and accept the co-authors responsibilities	☐

Contact information for Author #3

First name _____

Middle initials _____

Last name _____

Author name
(name that will appear on the publication): _____

Permanent Email address
(one you will use five years from now): _____

Alternative Email address (optional): _____

Check this box to indicate that you have read and accept the co-authors responsibilities ☐

Project Details

Project name: Contig26
Project species: D. ananassae
Date of submission: 12/15/2020
Size of project in base pairs: 66,525
Number of genes in project: 4

Does this report cover all of the genes or is it a partial report? Covers all genes
If this is a partial report, please indicate the region of the project covered by this report:
From base _____ to base _____

Note: For each gene described in this annotation report, you should also prepare the corresponding **GFF, transcript and peptide sequence files** as part of your submission.

Complete the following Gene Report Form for each gene in your project. Copy and paste the sections below to create as many copies as needed within this report. Be sure to create enough Isoform Report Forms within your Gene Report Form for all isoforms. If you cannot find evidence for any protein-coding genes in the project, jump to the "Check for additional features in your project" section.

Gene Report Form

Gene name (*e.g.*, *D. ananassae eyeless*): *D. ananassae CG18223*
Gene symbol (*e.g.*, *dana_ey*): *dana_CG18223*

Approximate location in project (from 5' end to 3' end): 14,673-16,483
Number of isoforms in *D. melanogaster:* 2
Number of isoforms in this project: 2

Complete the following table, including all of the isoforms in this project:

Name(s) of unique isoform(s) based on coding sequence	**List of isoforms with identical coding sequences**
CG18223-RA	
CG18223-RC	

Names of the isoforms with unique coding sequences in *D. melanogaster* that are absent in this species: _____

Provide the evidence (text and figures) which support the hypothesis that these isoforms are absent in this species (*e.g.*, changes in canonical splice sites, gene structure, etc.):

Note: For isoforms with identical coding sequence, you only need to complete the Isoform Report Form for one of these isoforms (i.e. using the name of the isoform listed in the left column of the table above). However, you should **generate GFF, transcript, and peptide sequence files for ALL isoforms**, irrespective of whether their coding sequence is identical to that of another isoform.

Consensus Sequence Errors Report Form

Complete this section if you have identified errors in the project consensus sequence that affect the annotation of the gene described above.

All of the coordinates reported in this section should be relative to the coordinates of the original project sequence.

Location(s) in the project sequence with consensus errors:

_____

1. Evidence that supports the consensus errors postulated above

Note: Evidence that could be used to support the hypothesis of errors within the consensus sequence includes a CDS alignment with frame shifts or in-frame stop codons, and RNA-Seq reads with discrepant alignments compared to the project sequence.

2. Generate a VCF file which describes the changes to the consensus sequence

Use the Sequencer Updater to create a Variant Call Format (VCF) file that describes the changes to the consensus sequence you have identified above. **Paste a screenshot with the list of sequence changes into the box below:**

Isoform Report Form

Complete this report form for each unique isoform listed in the table above. Copy and paste this form to create as many copies of this Isoform Report Form as needed.

Gene-isoform symbol (*e.g.*, dana_ey-PA): *dana_CG18223-RA*

Names of any additional isoforms with identical coding sequences:

_____

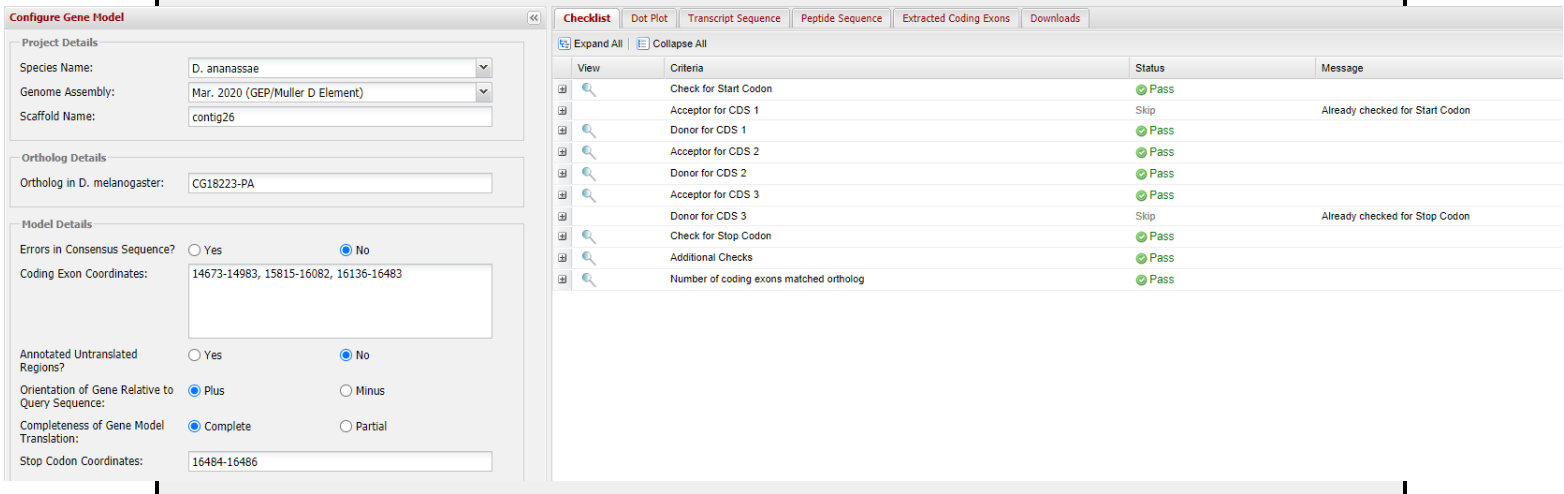
Is the 5' end of this isoform missing from the end of the project? No
If so, how many putative exons are missing from the 5' end: _____
Is the 3' end of this isoform missing from the end of the project? No
If so, how many putative exons are missing from the 3' end: _____

(Define "putative exons" based on the exons present in the *D. melanogaster* ortholog)

1. Gene Model Checker checklist

Enter the coordinates of your final gene model for this isoform into the Gene Model Checker and **paste a screenshot of the checklist results into the box below:**

Note: For projects with consensus sequence errors, report the exon coordinates relative to the **original project sequence**. Include the VCF file you have generated above when you submit the gene model to the Gene Model Checker. The Gene Model Checker will use this VCF file to automatically revise the submitted exon coordinates.

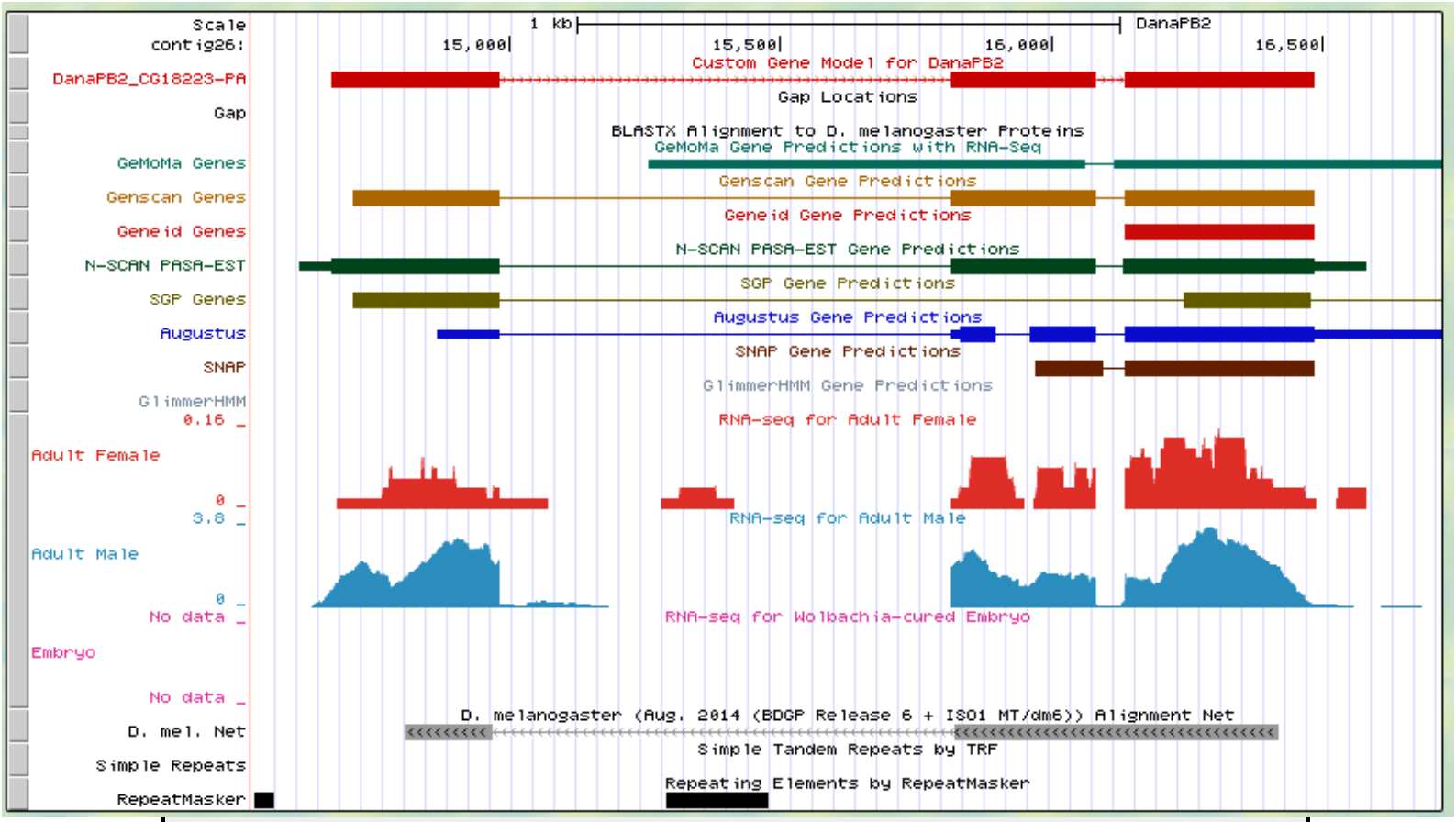


2. View the gene model on the Genome Browser

Click on the magnifying glass icon under the "Checklist" tab of the Gene Model Checker to view your gene model on the GEP UCSC Genome Browser. Zoom in so that **only this isoform is in the genome browser window, and capture a screenshot that includes the following evidence tracks if they are available:**

1. A sequence alignment track (*e.g.*, D. mel Proteins)
2. At least one gene prediction track (*e.g.*, Genscan)
3. At least one RNA-Seq track (*e.g.*, RNA-Seq Coverage)
4. A comparative genomics track (*e.g.*, D. mel. Net Alignment, Conservation)

Paste a screenshot of your gene model as shown on the GEP UCSC Genome Browser into the box below:



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

Show an alignment between the protein sequence for your gene model and the protein sequence from the putative *D. melanogaster* ortholog. You can either use the protein alignment generated by the Gene Model Checker (available through the "**View protein alignment**" link under the "Dot Plot" tab) or you can generate a new alignment using the "Align two or more sequences" feature at the NCBI BLAST web site. **Paste a screenshot of the protein alignment into the box below:**

Alignment of Dmel_CG18223-PA vs. DanaPB2_CG18223-PA

View plain text version

Download alignment image

Identity: 99/330 (30.0%), **Similarity:** 156/330 (47.3%), **Gaps:** 29/330 (8.8%)

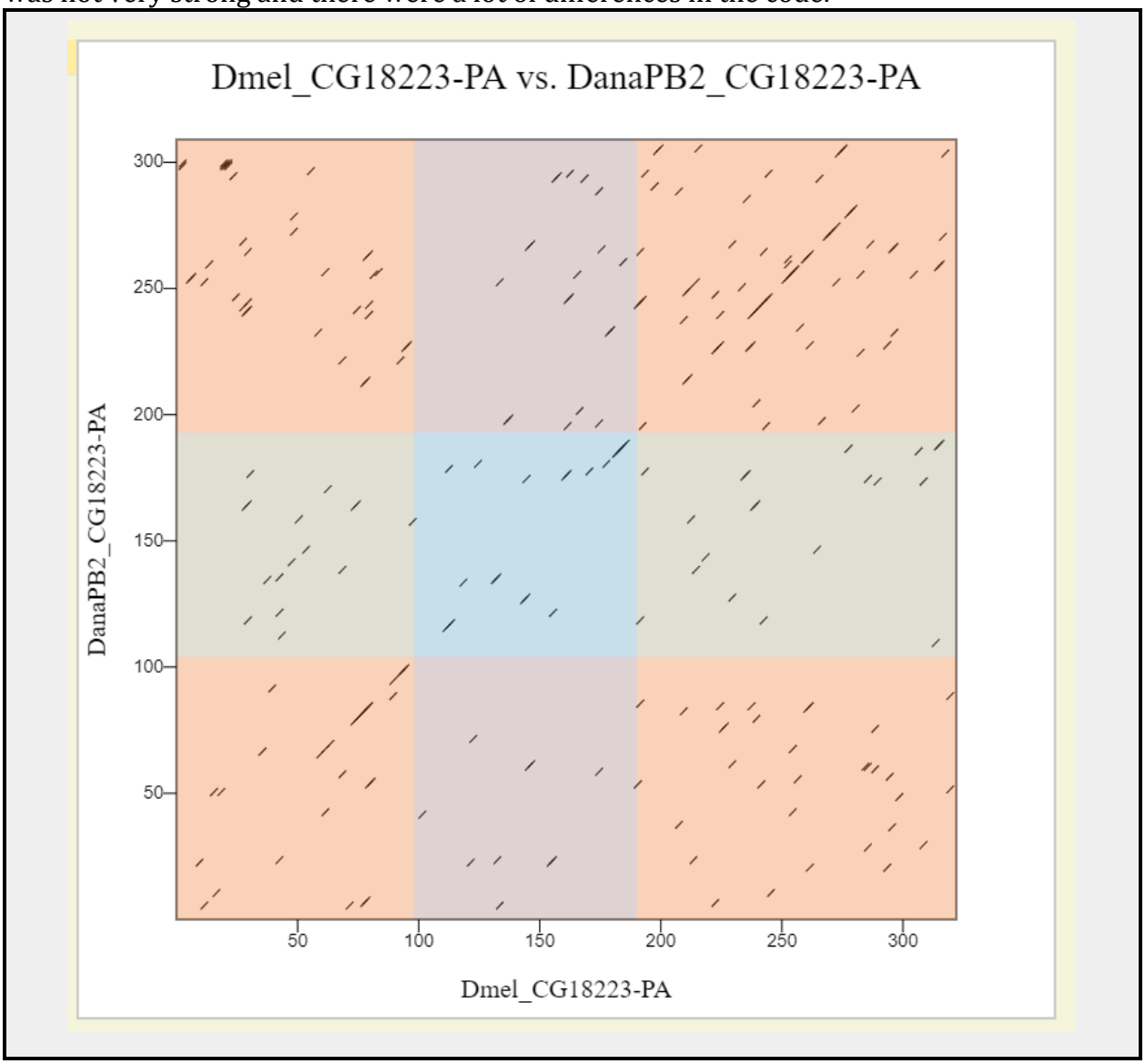
```
Dmel_CG18223-PA       1 MLLLKYGVSRKIFSFLLFLLLLPILDAGD------PIGSHFVRRRAKRLSSPYFDKEKT 54
                        * *:*      : *.         : : *:      * . :.* .   ::  .   : ::
DanaPB2_CG18223-PA    1 MSLVKLFCILLFASVAESQCTSKKLQQNGNSSAVEPPNDLKVVSNSDYQFLVTGGYRPES 60

Dmel_CG18223-PA      55 LVLAKYVVSIRSRRPHKLFGDNHFCGGVIISRTYILTSAHCAMDKRKIVHRSRVLVVVAG 114
                          *.*:***:*. :  : *****.*** **::  ***:***. : .: ::  * * ****
DanaPB2_CG18223-PA    61 NNLVKHVVSLRKGKEGN-FGDNHVCGGSIITKRAILTAAHCVYELNQKIKAWR-LSVVAG 118

Dmel_CG18223-PA     115 TTNRLKSRKGLSLNMEVKKIFVPDKFTVFNT-NNIALMMLAKKLPLDNPLVGVINLPTAD 173
                        *.:** . :  .: :.**:* *  :*.  .  *::*:::: *:*.:.:. ...*.:.. :
DanaPB2_CG18223-PA    119 TPKRLVKINTTQI-VRVKRIRVHPRFKKRTLRNDLAILIMEKELSMGDS-TDTISIANKN 176

Dmel_CG18223-PA     174 PEPGLNYTVLGWGRIFKGGPLASDILHIDVELLPRDICEKKVHIFKEEMMCAGNLNNTMD 233
                        * .*** **:*** ::: ***..:::: *::: *:.:***   :  . ****..  *  :
DanaPB2_CG18223-PA    177 PVAGLNCTVIGWGTLLELGPLPDEVVNGDMQIQPKSFCEKIEGFRAHGMMCASSSAN-YE 235

Dmel_CG18223-PA     234 ENPCAGDTGSPLIFNETVFGVVSYRVGCGSKTLPSIYTNVYMHMDWINGIMNNNEANRLC 293
                          :.* **:*.*** :*.*:*:**: .***. . ..***:** : ***:   :*. .: :
DanaPB2_CG18223-PA    236 VDSCQGDSGGPLICDEKVYGIVSFGFGCGEPNNAGIYTDVYHYHDWIE---KNSCPEDVM 292

Dmel_CG18223-PA     294 YSPNYLFTTIGIIIGNKILKSWGFYL-ITI 322
                          *  *:  *             :** : :
DanaPB2_CG18223-PA    293 SLPLVLLLLI------------IYLHLGL 309
```

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

Paste a screenshot of the dot plot (generated by the Gene Model Checker) of your submitted model against the putative *D. melanogaster* ortholog into the box below.
Provide an explanation for any anomalies on the dot plot (*e.g.*, large gaps, regions with no sequence similarity, indications of significant insertions or deletions).

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 and provide a justification as to why you have selected this particular set of donor and acceptor sites.

There are gaps in this alignment because the Blast between the gene in *D. ana* and *D. mel* was not very strong and there were a lot of differences in the code.



Isoform Report Form

Complete this report form for each unique isoform listed in the table above. Copy and paste this form to create as many copies of this Isoform Report Form as needed.

Gene-isoform symbol (*e.g.*, dana_ey-PA): *dana_CG18223-RC*

Names of any additional isoforms with identical coding sequences:

_____

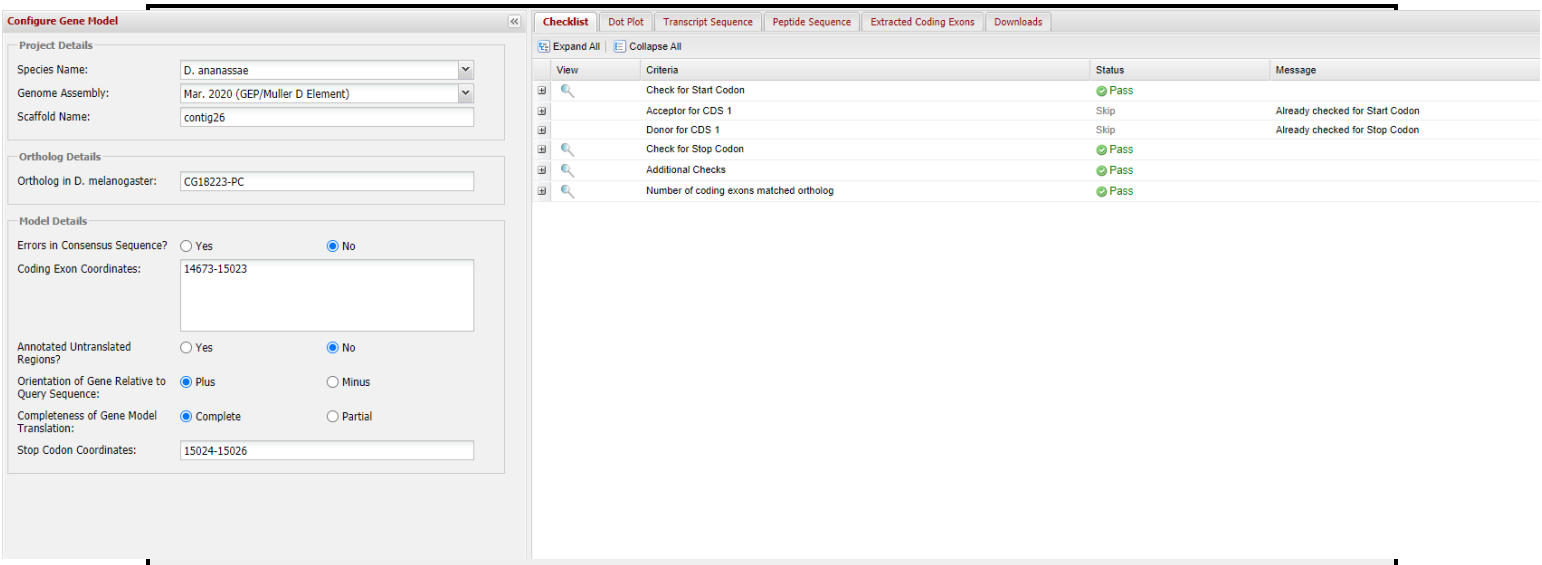
Is the 5' end of this isoform missing from the end of the project? No
If so, how many putative exons are missing from the 5' end: _____
Is the 3' end of this isoform missing from the end of the project? No
If so, how many putative exons are missing from the 3' end: _____

(Define "putative exons" based on the exons present in the *D. melanogaster* ortholog)

1. Gene Model Checker checklist

Enter the coordinates of your final gene model for this isoform into the Gene Model Checker and **paste a screenshot of the checklist results into the box below:**

Note: For projects with consensus sequence errors, report the exon coordinates relative to the **original project sequence**. Include the VCF file you have generated above when you submit the gene model to the Gene Model Checker. The Gene Model Checker will use this VCF file to automatically revise the submitted exon coordinates.

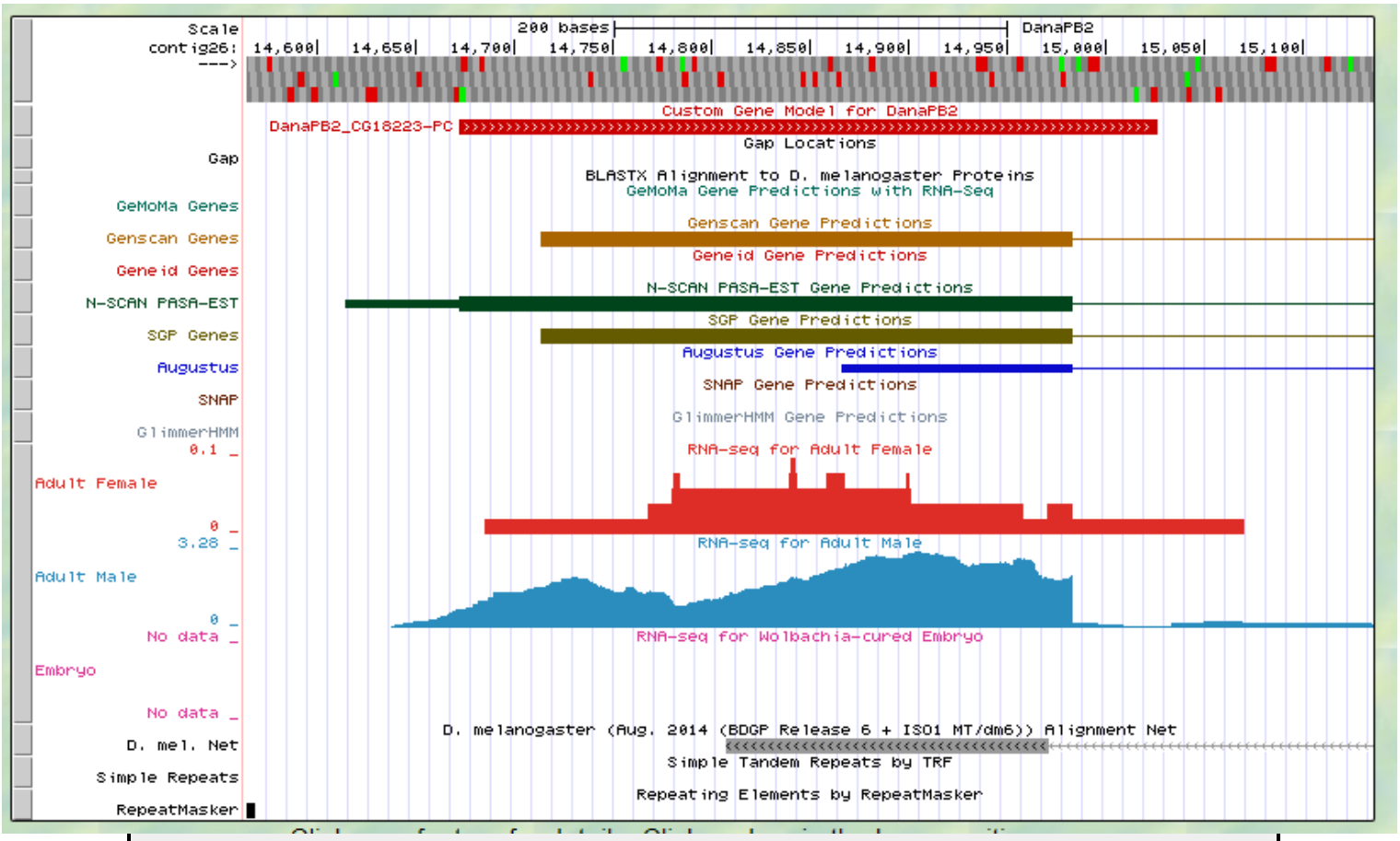


2. View the gene model on the Genome Browser

Click on the magnifying glass icon under the "Checklist" tab of the Gene Model Checker to view your gene model on the GEP UCSC Genome Browser. Zoom in so that **only this isoform is in the genome browser window, and capture a screenshot that includes the following evidence tracks if they are available:**

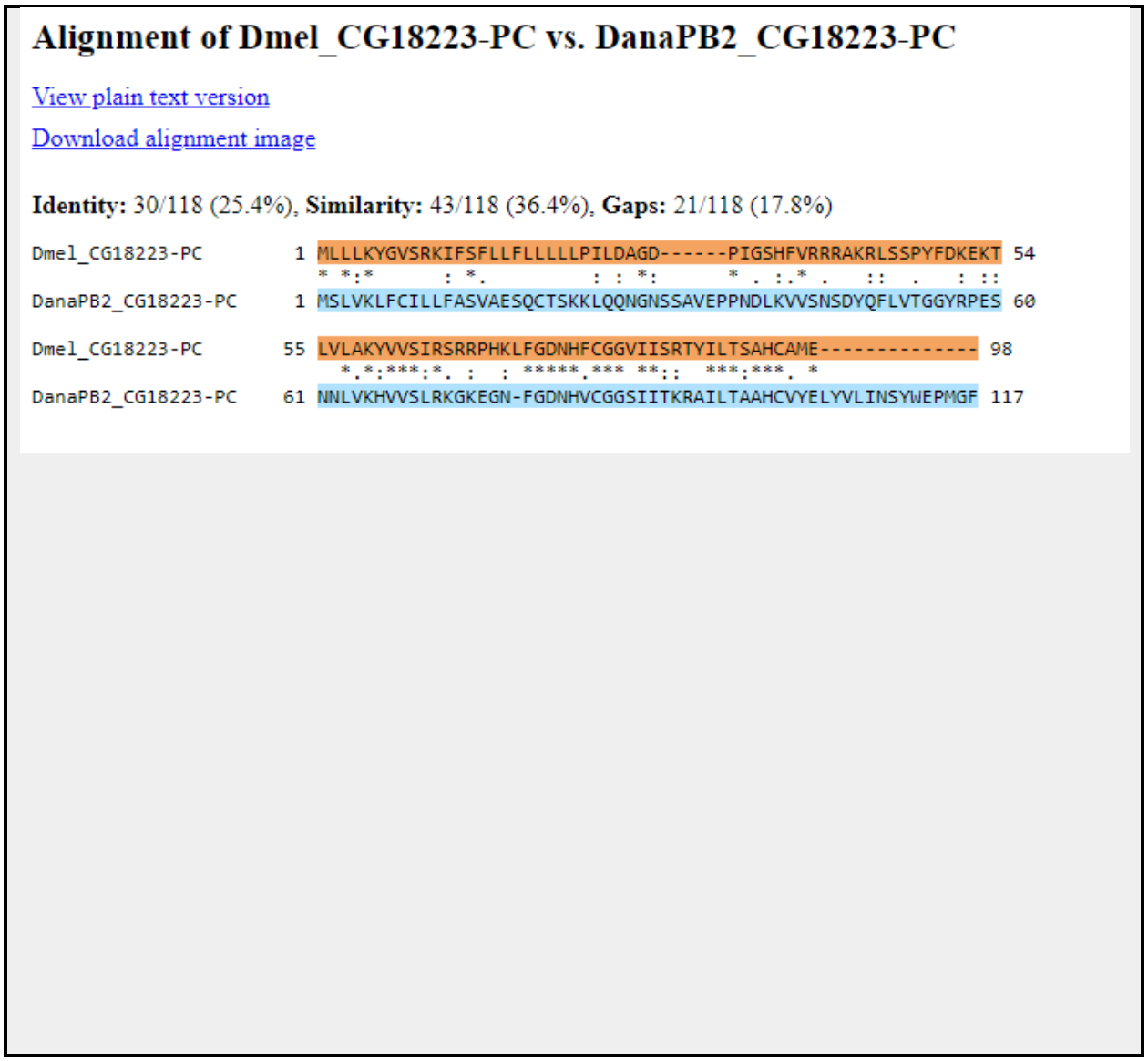
1. A sequence alignment track (*e.g.*, D. mel Proteins)
2. At least one gene prediction track (*e.g.*, Genscan)
3. At least one RNA-Seq track (*e.g.*, RNA-Seq Coverage)
4. A comparative genomics track (*e.g.*, D. mel. Net Alignment, Conservation)

Paste a screenshot of your gene model as shown on the GEP UCSC Genome Browser into the box below:



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

Show an alignment between the protein sequence for your gene model and the protein sequence from the putative *D. melanogaster* ortholog. You can either use the protein alignment generated by the Gene Model Checker (available through the "**View protein alignment**" link under the "Dot Plot" tab) or you can generate a new alignment using the "Align two or more sequences" feature at the NCBI BLAST web site. **Paste a screenshot of the protein alignment into the box below:**



Alignment of Dmel_CG18223-PC vs. DanaPB2_CG18223-PC

View plain text version

Download alignment image

Identity: 30/118 (25.4%), **Similarity:** 43/118 (36.4%), **Gaps:** 21/118 (17.8%)

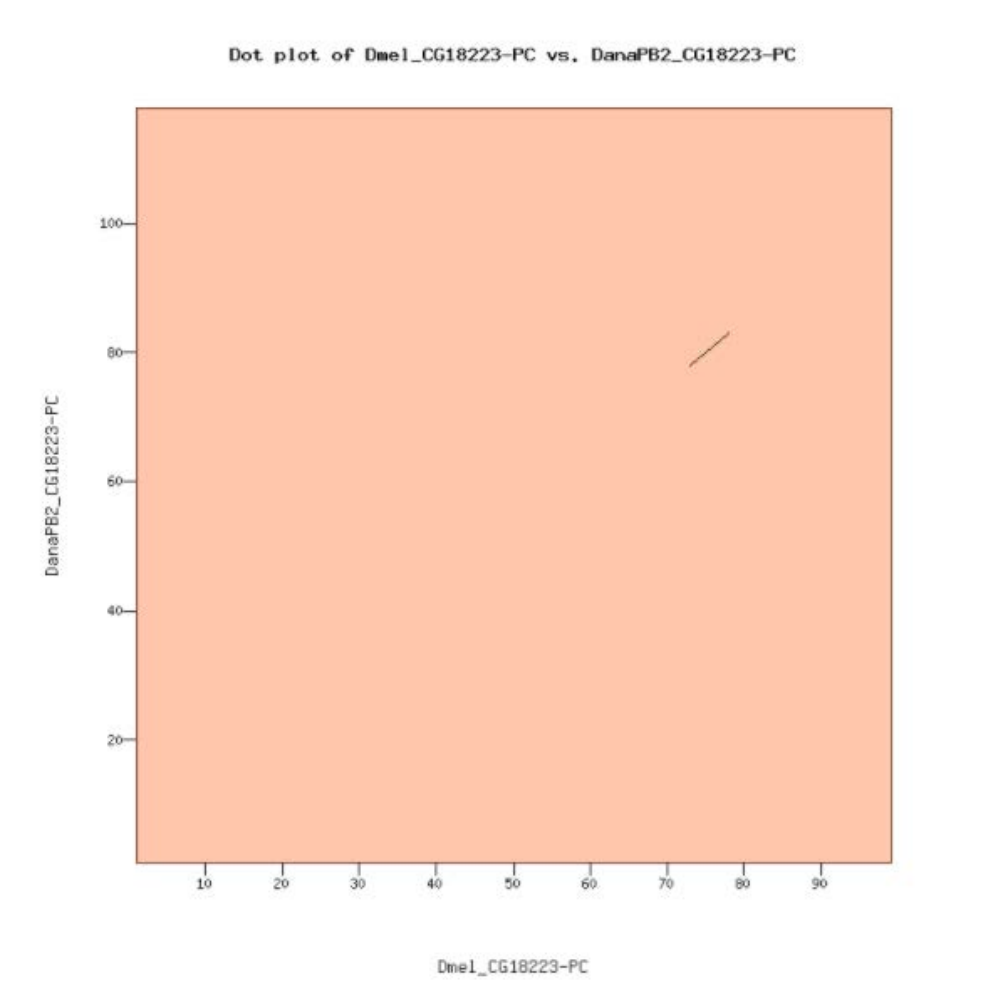
```
Dmel_CG18223-PC        1 MLLLKYGVSRKIFSFLLFLLLLLPILDAGD------PIGSHFVRRRAKRLSSPYFDKEKT 54
                         * *:*      : *.         : : *:      * . :.* .   ::  .   : ::
DanaPB2_CG18223-PC     1 MSLVKLFCILLFASVAESQCTSKKLQQNGNSSAVEPPNDLKVVSNSDYQFLVTGGYRPES 60

Dmel_CG18223-PC       55 LVLAKYVVSIRSRRPHKLFGDNHFCGGVIISRTYILTSAHCAME-------------- 98
                           *.*:***:*. :  : *****.*** **::  ***:***. *
DanaPB2_CG18223-PC    61 NNLVKHVVSLRKGKEGN-FGDNHVCGGSIITKRAILTAAHCVYELYVLINSYWEPMGF 117
```

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

Paste a screenshot of the dot plot (generated by the Gene Model Checker) of your submitted model against the putative *D. melanogaster* ortholog into the box below.
Provide an explanation for any anomalies on the dot plot (*e.g.*, large gaps, regions with no sequence similarity, indications of significant insertions or deletions).

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

There are gaps in this alignment because the Blast between the gene in *D. ana* and *D. mel* was not very strong and there were a lot of differences in the code. Additionally, this is a single exon gene so the differences in the code as indicated by the Blast search affect the dot plot more than a gene that has multiple exons.



Note: Complete this section if you have annotated the TSS for the gene above. This section is **optional** and you do not need to complete this section to submit the project.

Transcription Start Sites (TSS) Report Form (Optional)

Gene name (*e.g., D. ananassae eyeless*): *D. ananassae CG18223*
Gene symbol (*e.g., dana_ey*): *dana CG18223*

Name(s) of isoform(s) with unique TSS	List of isoforms with identical TSS
CG18223-RA	CG18223-RC

Names of the isoforms with unique TSS in *D. melanogaster* that are absent in this species:

Provide the evidence (text and figures) which support the hypothesis that these isoforms are absent in this species (*e.g.*, changes in canonical splice sites, gene structure, etc.):

Isoform TSS Report

Complete an Isoform TSS report (through page 71) for each unique TSS listed in the table above. Copy and paste this form to create as many copies as needed.

Gene-isoform name (*e.g.*, *dana_ey-RA*): *dana_CG18223-RA*

Names of the isoforms with the same TSS as this isoform: *dana_CG18223-RC*

Type of core promoter in *D. melanogaster* (see table below):
(**Peaked** / Intermediate / Broad / Insufficient Evidence)

The type of core promoter is defined by the number of TSS annotated by the Celniker group at modENCODE and the number of DHS positions:

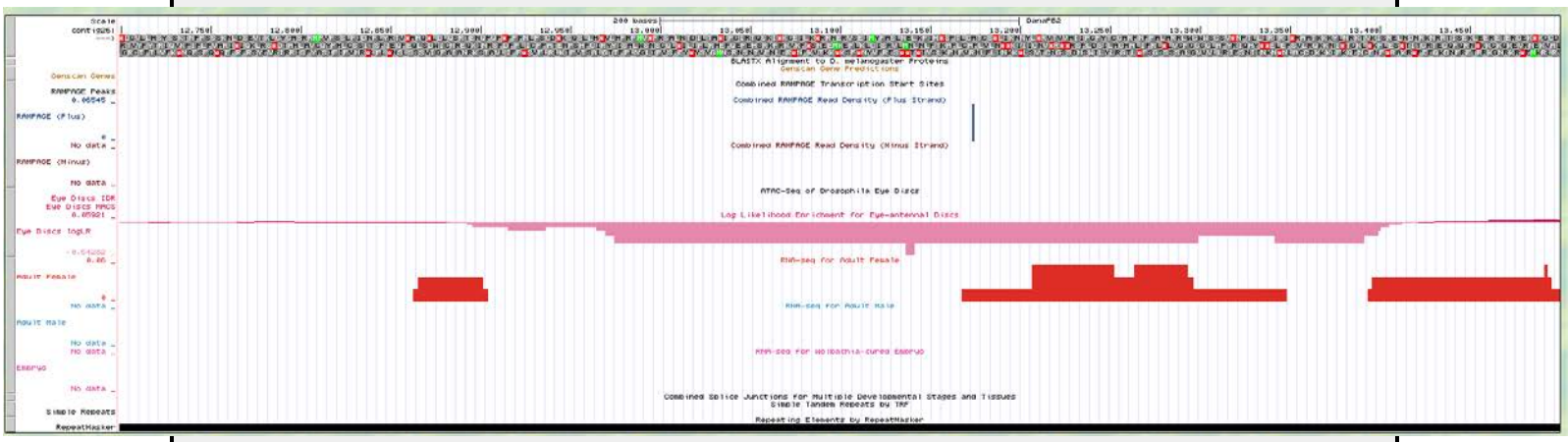
Type of core promoter	# annotated TSS	# DHS positions
Peaked	1	0
	0	1
	1	1
Intermediate	≤ 1	> 1
	> 1	≤ 1
Broad	> 1	> 1
Insufficient Evidence	0	0

1. Turn on RAMPAGE evidence tracks

Coordinates of the TSS position based on position with the highest RAMPAGE read density
13,173-13,174

Coordinates of the narrow TSS search region based on RAMPAGE peaks
13,000-13,200

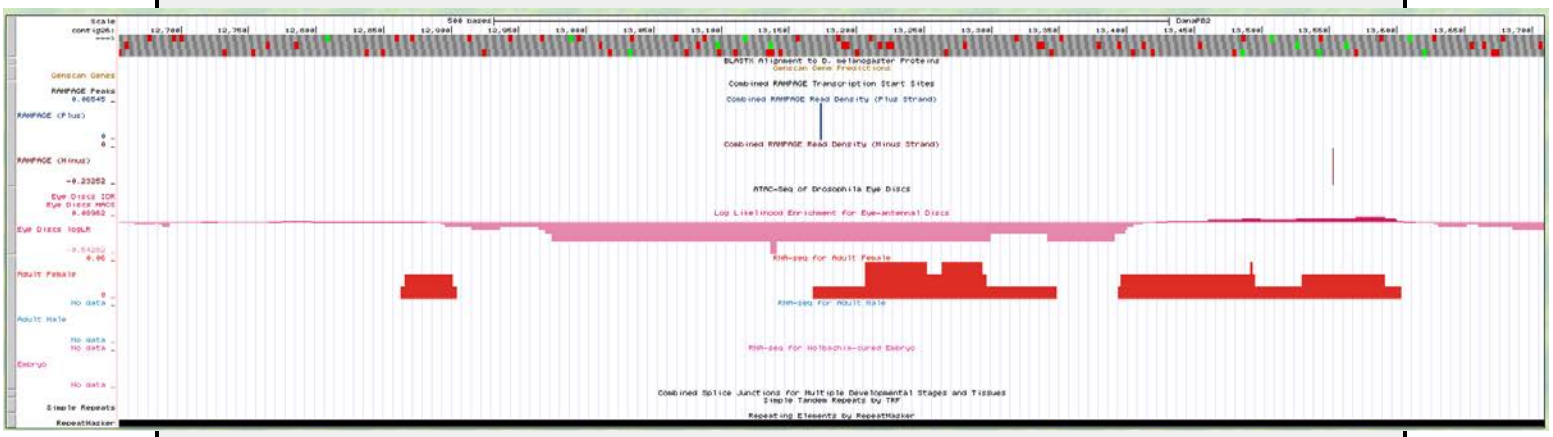
If the TSS position and narrow TSS search region are supported by RAMPAGE data, **paste a Genome Browser screenshot of the region surrounding the putative TSS (±300bp) showing the Combined RAMPAGE TSS evidence track:**



2. Turn on ATAC-Seq evidence track

If the wide TSS search region is supported by ATAC-Seq data, **paste a Genome Browser screenshot of the region surrounding the putative TSS (±300bp) showing the Eye Discs ATAC-Seq evidence track:**

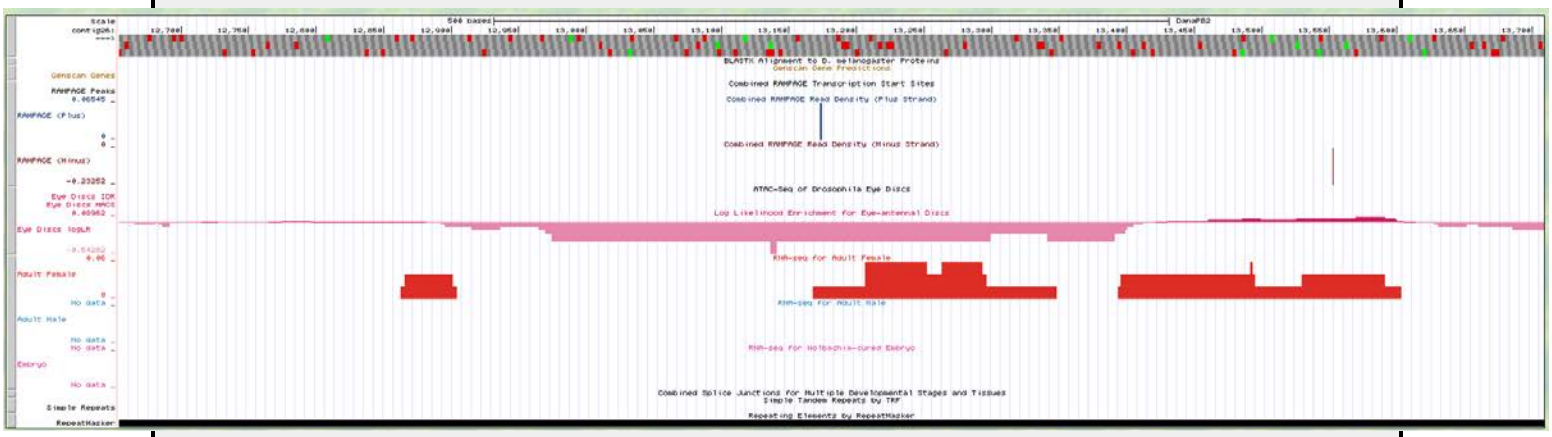
Coordinates of the wide TSS search region based on ATAC-Seq peaks
12,955-13,409



3. Turn on RNA-Seq evidence tracks

If the TSS annotation is supported by RNA-Seq read coverage or splice junction predictions (*e.g.*, regtools), **paste a Genome Browser screenshot of the region surrounding the putative TSS (±300bp) showing the following evidence tracks:**

1. RNA-Seq Coverage or RNA-Seq Alignment Summary
2. Combined Splice Junctions or RNA-Seq TopHat



If the RNA-Seq evidence tracks indicate the TSS position, list it here: 13,169-13,603

4. Annotate the first transcribed exon

Coordinates of the first transcribed exon based on blastn alignment:
21,101-21,217

Does the blastn alignment cover the entire *D. melanogaster* first transcribed exon?
No

If not, specify the parts of the *D. melanogaster* exon that are missing from the blastn alignment.
1-186, 304-316

If the TSS annotation is supported by blastn alignment of the initial transcribed exon against the contig sequence, **paste a screenshot of the blastn alignment into the box below:**

DanaPB2_dna range=contig26:1-66525 5'pad=0 3'pad=0 strand=+ repeatMasking=none

Sequence ID: Query_38847 Length: 66525 Number of Matches: 1

Range 1: 21101 to 21217 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
41.1 bits(27)	8e-06	79/121(65%)	8/121(6%)	Plus/Plus

```
Query  187    TCTGGCCAAATATGTAGTCTCGATTCGATCC--CGAAGACCCCACAAG-TTATTTGGCGA  243
              ||||  ||| || ||||| ||  | ||  |   | ||||    | ||| || ||||| ||
Sbjct  21101  TCTGATCAAGTACGTAGTGTCCTTGCGGGCGTACAAAGAG---AAAAGCTTCTTTGGAGA  21157

Query  244    CAATCACTTCTGCGGTGGCGTGATCATATCGAGAACCTACA-TCCTCACCTCCGCCCACT  302
               || |||   ||||| || | ||| || |||| | | | |  |||| ||  | |||||||
Sbjct  21158  TAACCACGCATGCGGAGGTGCGATAATTTCGA-AGCGTGCCGTCCTGACAGCAGCCCACT  21216

Query  303    G  303
              |
Sbjct  21217  G  21217
```

5. Turn on comparative genomics tracks

If the TSS annotation is supported by sequence conservation with other *Drosophila* species, **paste a screenshot of the multiple sequence alignment (*e.g.*, from Clustal Omega, ROAST) into the box below:**

6. Summarize the evidence that supports the TSS annotation postulated above

Coordinate(s) of the TSS position(s):

Based on RAMPAGE data: 13,000-13,200
Based on ATAC-Seq data: 12,955-13,409
Based on RNA-Seq data: 13,169-13,603
Based on blastn alignment: 20,195
Based on other evidence (please specify): _____

Note: If the blastn alignment for the initial transcribed exon is a partial alignment, you can **extrapolate the TSS position** based on the number of nucleotides that are missing from the beginning of the exon. (Enter "Insufficient evidence" if you cannot determine the TSS position based on the available evidence.)

Were you able to define a TSS position based on the available evidence? Yes
If so, indicate in the table below the evidence that supports this TSS position

If not, were you able to define a TSS search region? _____
If so, indicate in the table below the evidence that supports the TSS search region(s)

Evidence type	Support	Refute	Neither
RAMPAGE peaks and read density	☒	☐	☐
ATAC-Seq peaks and log likelihood enrichment profile	☒	☐	☐
RNA-Seq coverage and splice junctions	☒	☐	☐
blastn alignment of the initial exon from *D. melanogaster*	☐	☐	☒
Sequence conservation with other *Drosophila* species (*e.g.*, "Conservation" track on the Genome Browser)	☐	☐	☒
Other (please specify) [*e.g.*, Gnomon, N-SCAN, Augustus TSS predictions; histone modifications (ChIP-Seq data)].	☐	☐	☒

Note: The evidence type refutes the TSS annotation only if it **suggests an alternate TSS position**. For example, the presence of RNA-Seq read coverage upstream of the annotated TSS indicates that the TSS is located further upstream and it would be considered to be evidence against the annotated TSS; check "Refute." In contrast, the lack of RNA-Seq read coverage is a negative result that neither supports nor refutes the TSS annotation; check "Neither."

Gene Report Form

Gene name (*e.g.*, *D. ananassae eyeless*): *D. ananassae Tetraspanin 68C*
Gene symbol (*e.g.*, *dana_ey*): *dana_Tsp68C*

Approximate location in project (from 5' end to 3' end): 17,937-19,238
Number of isoforms in *D. melanogaster:* 2
Number of isoforms in this project: 2

Complete the following table, including all of the isoforms in this project:

Name(s) of unique isoform(s) based on coding sequence	**List of isoforms with identical coding sequences**
Tsp68C-PA	*Tsp68C-PC*

Names of the isoforms with unique coding sequences in *D. melanogaster* that are absent in this species: _____

Provide the evidence (text and figures) which support the hypothesis that these isoforms are absent in this species (*e.g.*, changes in canonical splice sites, gene structure, etc.):

Note: For isoforms with identical coding sequence, you only need to complete the Isoform Report Form for one of these isoforms (i.e. using the name of the isoform listed in the left column of the table above). However, you should **generate GFF, transcript, and peptide sequence files for ALL isoforms**, irrespective of whether their coding sequence is identical to that of another isoform.

Consensus Sequence Errors Report Form

Complete this section if you have identified errors in the project consensus sequence that affect the annotation of the gene described above.

All of the coordinates reported in this section should be relative to the coordinates of the original project sequence.

Location(s) in the project sequence with consensus errors:

_____

1. Evidence that supports the consensus errors postulated above

Note: Evidence that could be used to support the hypothesis of errors within the consensus sequence includes a CDS alignment with frame shifts or in-frame stop codons, and RNA-Seq reads with discrepant alignments compared to the project sequence.

2. Generate a VCF file which describes the changes to the consensus sequence

Use the Sequencer Updater to create a Variant Call Format (VCF) file that describes the changes to the consensus sequence you have identified above. **Paste a screenshot with the list of sequence changes into the box below:**

Isoform Report Form

Complete this report form for each unique isoform listed in the table above. Copy and paste this form to create as many copies of this Isoform Report Form as needed.

Gene-isoform symbol (*e.g.*, dana_ey-PA): *dana_Tsp68C-PA*

Names of any additional isoforms with identical coding sequences: *dana_Tsp68C-PC*

_____

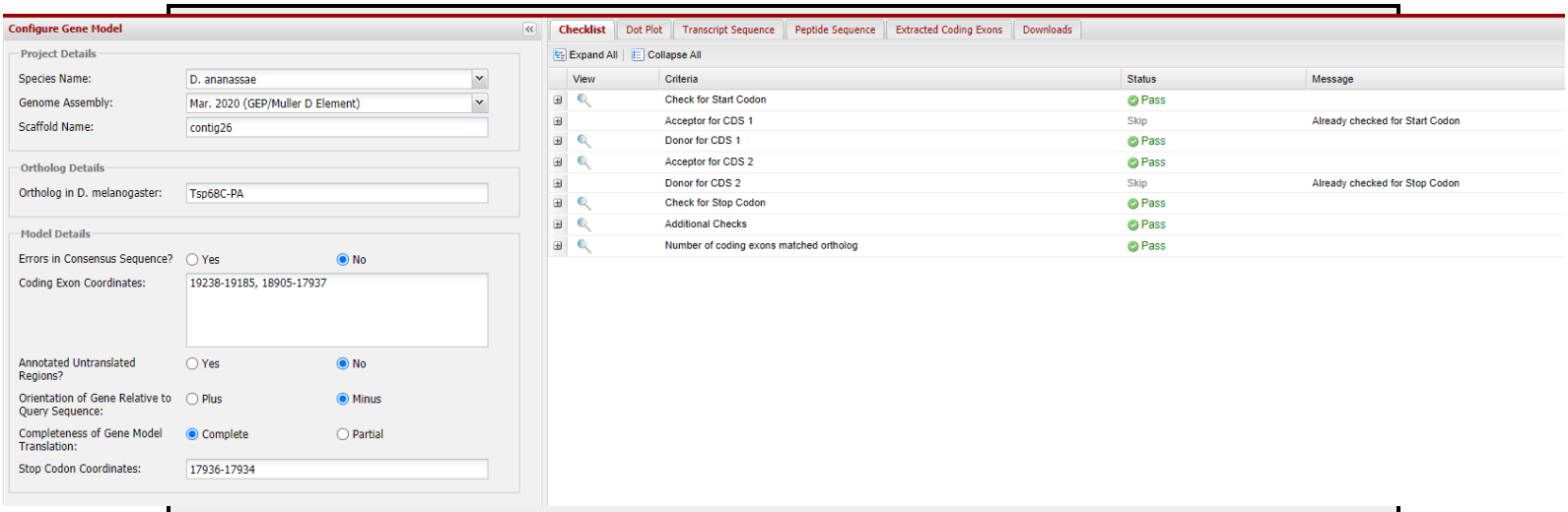
Is the 5' end of this isoform missing from the end of the project? No
If so, how many putative exons are missing from the 5' end: _____
Is the 3' end of this isoform missing from the end of the project? No
If so, how many putative exons are missing from the 3' end: _____

(Define "putative exons" based on the exons present in the *D. melanogaster* ortholog)

1. Gene Model Checker checklist

Enter the coordinates of your final gene model for this isoform into the Gene Model Checker and **paste a screenshot of the checklist results into the box below:**

Note: For projects with consensus sequence errors, report the exon coordinates relative to the **original project sequence**. Include the VCF file you have generated above when you submit the gene model to the Gene Model Checker. The Gene Model Checker will use this VCF file to automatically revise the submitted exon coordinates.

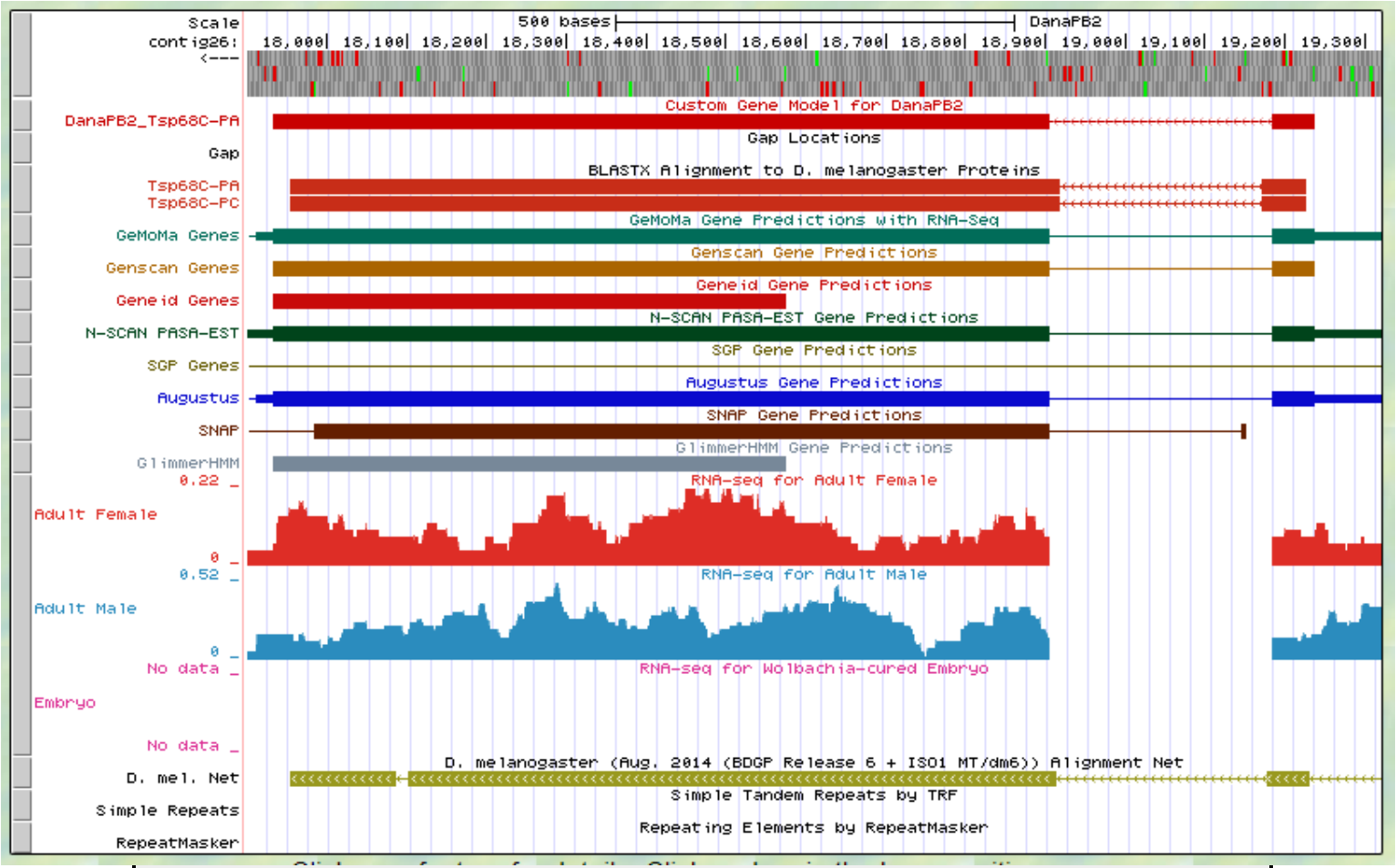


2. View the gene model on the Genome Browser

Click on the magnifying glass icon under the "Checklist" tab of the Gene Model Checker to view your gene model on the GEP UCSC Genome Browser. Zoom in so that **only this isoform is in the genome browser window, and capture a screenshot that includes the following evidence tracks if they are available:**

1. A sequence alignment track (*e.g.*, D. mel Proteins)
2. At least one gene prediction track (*e.g.*, Genscan)
3. At least one RNA-Seq track (*e.g.*, RNA-Seq Coverage)
4. A comparative genomics track (*e.g.*, D. mel. Net Alignment, Conservation)

Paste a screenshot of your gene model as shown on the GEP UCSC Genome Browser into the box below:



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

Show an alignment between the protein sequence for your gene model and the protein sequence from the putative *D. melanogaster* ortholog. You can either use the protein alignment generated by the Gene Model Checker (available through the "**View protein alignment**" link under the "Dot Plot" tab) or you can generate a new alignment using the "Align two or more sequences" feature at the NCBI BLAST web site. **Paste a screenshot of the protein alignment into the box below:**

Alignment of Dmel_Tsp68C-PA vs. DanaPB2_Tsp68C-PA

View plain text version

Download alignment image

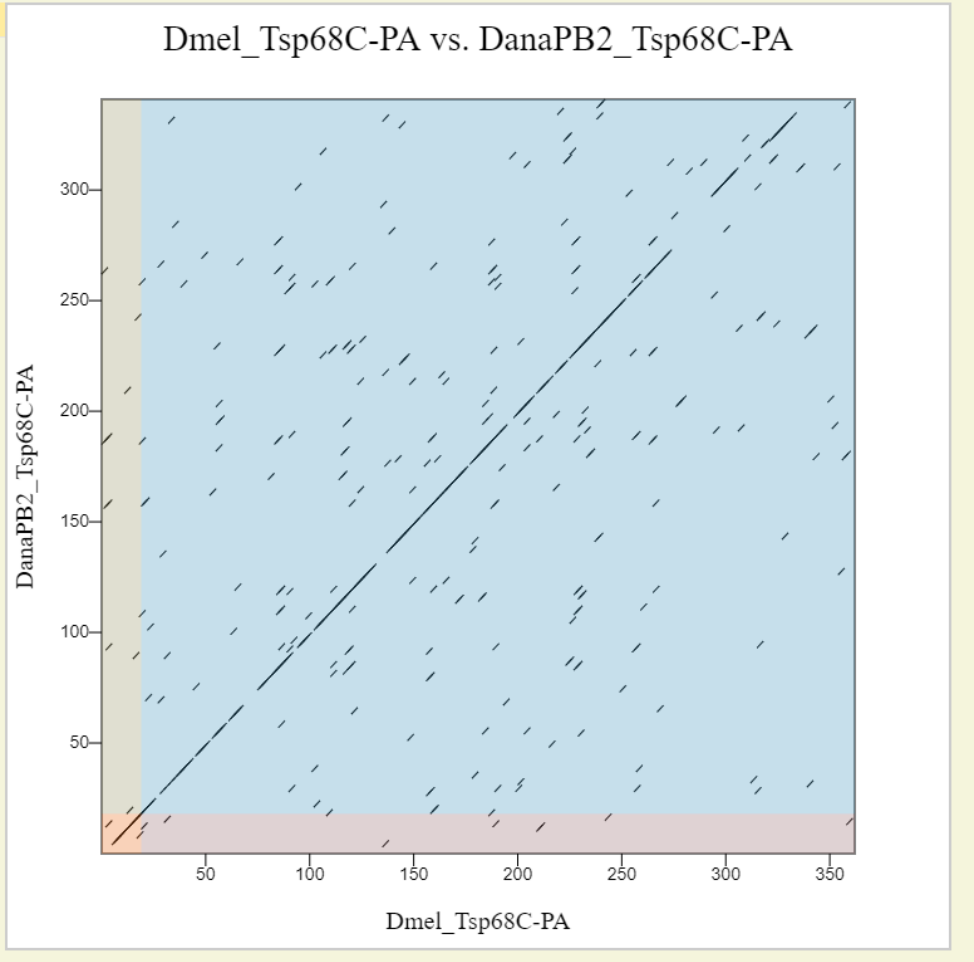
Identity: 254/367 (69.2%), **Similarity:** 287/367 (78.2%), **Gaps:** 31/367 (8.4%)

```
Dmel_Tsp68C-PA       1 MACCFNYKFVLNLCNFLFLICGLLLVVSGLYIFSDNKRILLSRLLAASSDRLSSLPQPLL 60
                       *. . :***:***************:.** *:.**.**:*****:*******.*******
DanaPB2_Tsp68C-PA    1 MGRS-HYKFLLNLCNFLFLICGLLLIASGCYLYSDGKRVLLSRLFAASSDRLNSLPQPLL 59

Dmel_Tsp68C-PA      61 FYIALGVAIAGFVATLAAVVGFWASCLHTYCFLTIYFLSVVVLLLTESVLCLAITLWPHC 120
                       .******* **.:*.****** **.***** *  *******.*******:**********
DanaPB2_Tsp68C-PA    60 HYIALGVASAGLIAVLAAVVGCWANCLHTYIFHCIYFLSVVALLLTESVVCLAITLWPHC 119

Dmel_Tsp68C-PA      121 LGISLDETQMVRSLQSNYGVPGQEQFTNALDLAQVRFGCCGMRSSLDYDTSLWRLQGYGQ 180
                       *******:****:**: ***********:****.*******:************.:**
DanaPB2_Tsp68C-PA   120 LGISLDESQMVRALQALYGVPGQEQFTNAMDLAQARFGCCGMKSSLDYDTSLWRLQNFGQ 179

Dmel_Tsp68C-PA      181 RNWPVPLSCCFLKNAGHSMAYLDPKPANESMCQSLERLSYERERHTESCLPHLDNWYREQ 240
                       **********.*:** *.:*******.* *:****:* ******.*********:****:
DanaPB2_Tsp68C-PA   180 RNWPVPLSCCVLENAPHPLAYLDPKPVNASVCQSLDRSSYERERFTESCLPHLDDWYREH 239

Dmel_Tsp68C-PA      241 YSIFLGASLILAMIEFCVLLAIIMSCTGLASQRARLKK-----PVQEMRTQKVKSRQTLI 295
                       *********:**.:******.*****:****:** **.     * :: * : *:*. :**
DanaPB2_Tsp68C-PA   240 YSIFLGASLVLALVEFCVLLCIIMSCAGLASERAVLKSCKEGMPDRKRRLKVVQSNLALI 299

Dmel_Tsp68C-PA      296 ENIYEPDVELRENSNHSGDGIYLGPASRHVSSEDFKELYIKPRDLYKQHNLRTSPANRPT 355
                       ******:**** ::.** :.   ** *********.* * :
DanaPB2_Tsp68C-PA   300 ENIYEPEVELRPKARHSTEAA--GP-SRHVSSEDFSEYYSR------------------- 337

Dmel_Tsp68C-PA      356 QMRNYLV 362
                         *:*
DanaPB2_Tsp68C-PA   338 --RQYQ- 341
```

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

Paste a screenshot of the dot plot (generated by the Gene Model Checker) of your submitted model against the putative *D. melanogaster* ortholog into the box below.
Provide an explanation for any anomalies on the dot plot (*e.g.*, large gaps, regions with no sequence similarity, indications of significant insertions or deletions).

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 and provide a justification as to why you have selected this particular set of donor and acceptor sites.



Transcription Start Sites (TSS) Report Form (Optional)

Note: Complete this section if you have annotated the TSS for the gene above. This section is **optional** and you do not need to complete this section to submit the project.

Gene name (*e.g.*, *D. ananassae eyeless*): *D. ananassae Tetraspanin 68C*
Gene symbol (*e.g.*, *dana_ey*): *dana_Tsp68C*

Name(s) of isoform(s) with unique TSS	List of isoforms with identical TSS
Tsp68C-PA	Tsp68C-PC

Names of the isoforms with unique TSS in *D. melanogaster* that are absent in this species:

_____

Provide the evidence (text and figures) which support the hypothesis that these isoforms are absent in this species (*e.g.*, changes in canonical splice sites, gene structure, etc.):

Isoform TSS Report

Complete an Isoform TSS report (through page 71) for each unique TSS listed in the table above. Copy and paste this form to create as many copies as needed.

Gene-isoform name (*e.g.*, *dana_ey-RA*): *dana_Tsp68C-PA*

Names of the isoforms with the same TSS as this isoform: *dana_Tsp68C-PC*

Type of core promoter in *D. melanogaster* (see table below):
(**Peaked** / Intermediate / Broad / Insufficient Evidence)

The type of core promoter is defined by the number of TSS annotated by the Celniker group at modENCODE and the number of DHS positions:

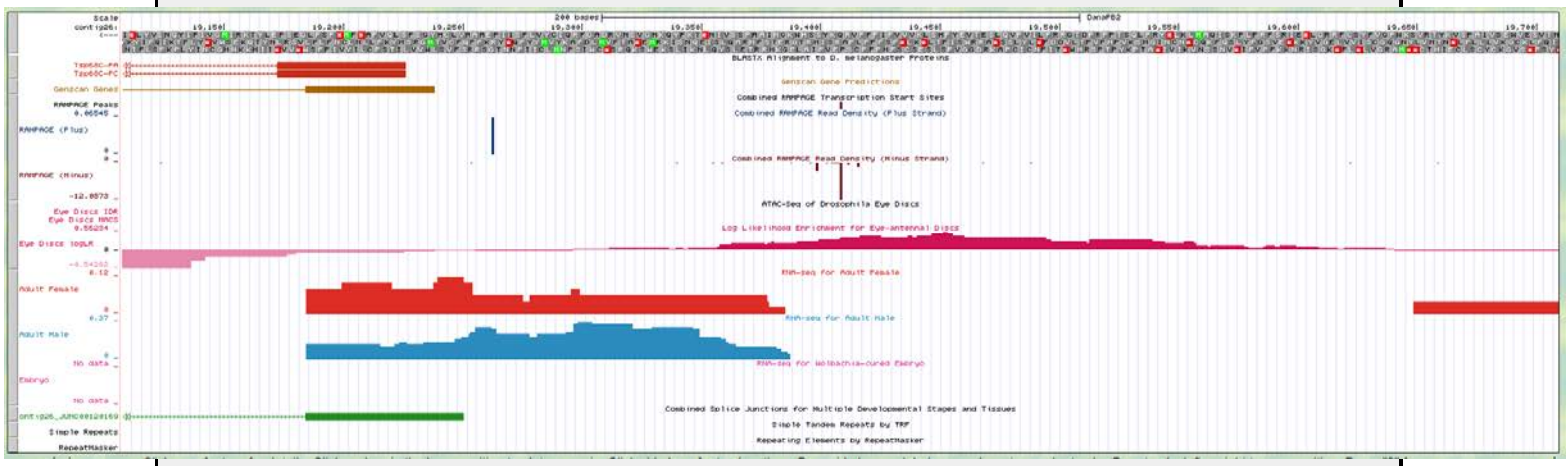
Type of core promoter	# annotated TSS	# DHS positions
Peaked	1	0
	0	1
	1	1
Intermediate	≤ 1	> 1
	> 1	≤ 1
Broad	> 1	> 1
Insufficient Evidence	0	0

1. Turn on RAMPAGE evidence tracks

Coordinates of the TSS position based on position with the highest RAMPAGE read density
19,408-19,409

Coordinates of the narrow TSS search region based on RAMPAGE peaks
19,108-19,709

If the TSS position and narrow TSS search region are supported by RAMPAGE data, **paste a Genome Browser screenshot of the region surrounding the putative TSS (±300bp) showing the Combined RAMPAGE TSS evidence track:**



2. Turn on ATAC-Seq evidence track

If the wide TSS search region is supported by ATAC-Seq data, **paste a Genome Browser screenshot of the region surrounding the putative TSS (±300bp) showing the Eye Discs ATAC-Seq evidence track:**

Coordinates of the wide TSS search region based on ATAC-Seq peaks
19,056-19,858

3. Turn on RNA-Seq evidence tracks

If the TSS annotation is supported by RNA-Seq read coverage or splice junction predictions (*e.g.*, regtools), **paste a Genome Browser screenshot of the region surrounding the putative TSS (±300bp) showing the following evidence tracks:**

1. RNA-Seq Coverage or RNA-Seq Alignment Summary
2. Combined Splice Junctions or RNA-Seq TopHat

If the RNA-Seq evidence tracks indicate the TSS position, list it here: 18,884-19,687

4. Annotate the first transcribed exon

Coordinates of the first transcribed exon based on blastn alignment: 19,231-19,185

Does the blastn alignment cover the entire *D. melanogaster* first transcribed exon?
No

If not, specify the parts of the *D. melanogaster* exon that are missing from the blastn alignment.
1-215

If the TSS annotation is supported by blastn alignment of the initial transcribed exon against the contig sequence, **paste a screenshot of the blastn alignment into the box below:**

DanaPB2_dna range=contig26:1-66525 5'pad=0 3'pad=0 strand=+ repeatMasking=none

Sequence ID: **Query_31481** Length: **66525** Number of Matches: **1**

Range 1: 19185 to 19231 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
41.1 bits(27)	6e-06	37/47(79%)	0/47(0%)	Plus/Minus

```
Query  216    GTTTTAATTACAAATTTGTGCTCAATCTATGTAACTTTTTGTTTTTG  262
              ||| | ||||||| ||| |||| ||| ||||||| ||| | || |||
Sbjct  19231  GTTCTCATTACAAGTTTTTGCTTAATTTATGTAATTTTCTTTTCTTG  19185
```

5. Turn on comparative genomics tracks

If the TSS annotation is supported by sequence conservation with other *Drosophila* species, **paste a screenshot of the multiple sequence alignment (*e.g.*, from Clustal Omega, ROAST) into the box below:**

6. Summarize the evidence that supports the TSS annotation postulated above

Coordinate(s) of the TSS position(s):

Based on RAMPAGE data: 19,108-19,709
Based on ATAC-Seq data: 19,056-19,858
Based on RNA-Seq data: 18,884-19,687
Based on blastn alignment: 19,061
Based on other evidence (please specify): _____

Note: If the blastn alignment for the initial transcribed exon is a partial alignment, you can **extrapolate the TSS position** based on the number of nucleotides that are missing from the beginning of the exon. (Enter "Insufficient evidence" if you cannot determine the TSS position based on the available evidence.)

Were you able to define a TSS position based on the available evidence? Yes
If so, indicate in the table below the evidence that supports this TSS position

If not, were you able to define a TSS search region? _____
If so, indicate in the table below the evidence that supports the TSS search region(s)

Evidence type	Support	Refute	Neither
RAMPAGE peaks and read density	☒	☐	☐
ATAC-Seq peaks and log likelihood enrichment profile	☒	☐	☐
RNA-Seq coverage and splice junctions	☒	☐	☐
blastn alignment of the initial exon from *D. melanogaster*	☒	☐	☐
Sequence conservation with other *Drosophila* species (*e.g.*, "Conservation" track on the Genome Browser)	☐	☐	☒
Other (please specify) [*e.g.*, Gnomon, N-SCAN, Augustus TSS predictions; histone modifications (ChIP-Seq data)].	☐	☐	☒

Note: The evidence type refutes the TSS annotation only if it **suggests an alternate TSS position**. For example, the presence of RNA-Seq read coverage upstream of the annotated TSS indicates that the TSS is located further upstream and it would be considered to be evidence against the annotated TSS; check "Refute." In contrast, the lack of RNA-Seq read coverage is a negative result that neither supports nor refutes the TSS annotation; check "Neither."

Provide an explanation if the TSS annotation is inconsistent with at least one of the evidence types specified above:

Gene Report Form

Gene name (*e.g., D. ananassae eyeless*): *D. ananassae CG8757*
Gene symbol (*e.g., dana_ey*): *dana_CG8757*

Approximate location in project (from 5' end to 3' end): 25,049-25,926
Number of isoforms in *D. melanogaster:* 1
Number of isoforms in this project: 1

Complete the following table, including all of the isoforms in this project:

Name(s) of unique isoform(s) based on coding sequence	**List of isoforms with identical coding sequences**
CG8757-RB	

Names of the isoforms with unique coding sequences in *D. melanogaster* that are absent in this species: _____

Provide the evidence (text and figures) which support the hypothesis that these isoforms are absent in this species (*e.g.*, changes in canonical splice sites, gene structure, etc.):

Note: For isoforms with identical coding sequence, you only need to complete the Isoform Report Form for one of these isoforms (i.e. using the name of the isoform listed in the left column of the table above). However, you should **generate GFF, transcript, and peptide sequence files for ALL isoforms**, irrespective of whether their coding sequence is identical to that of another isoform.

Consensus Sequence Errors Report Form

Complete this section if you have identified errors in the project consensus sequence that affect the annotation of the gene described above.

All of the coordinates reported in this section should be relative to the coordinates of the original project sequence.

Location(s) in the project sequence with consensus errors:

_____

1. Evidence that supports the consensus errors postulated above

Note: Evidence that could be used to support the hypothesis of errors within the consensus sequence includes a CDS alignment with frame shifts or in-frame stop codons, and RNA-Seq reads with discrepant alignments compared to the project sequence.

2. Generate a VCF file which describes the changes to the consensus sequence

Use the Sequencer Updater to create a Variant Call Format (VCF) file that describes the changes to the consensus sequence you have identified above. **Paste a screenshot with the list of sequence changes into the box below:**

Isoform Report Form

Complete this report form for each unique isoform listed in the table above. Copy and paste this form to create as many copies of this Isoform Report Form as needed.

Gene-isoform symbol (*e.g.*, dana_ey-PA): *dana_CG8757-RB*

Names of any additional isoforms with identical coding sequences:

_____

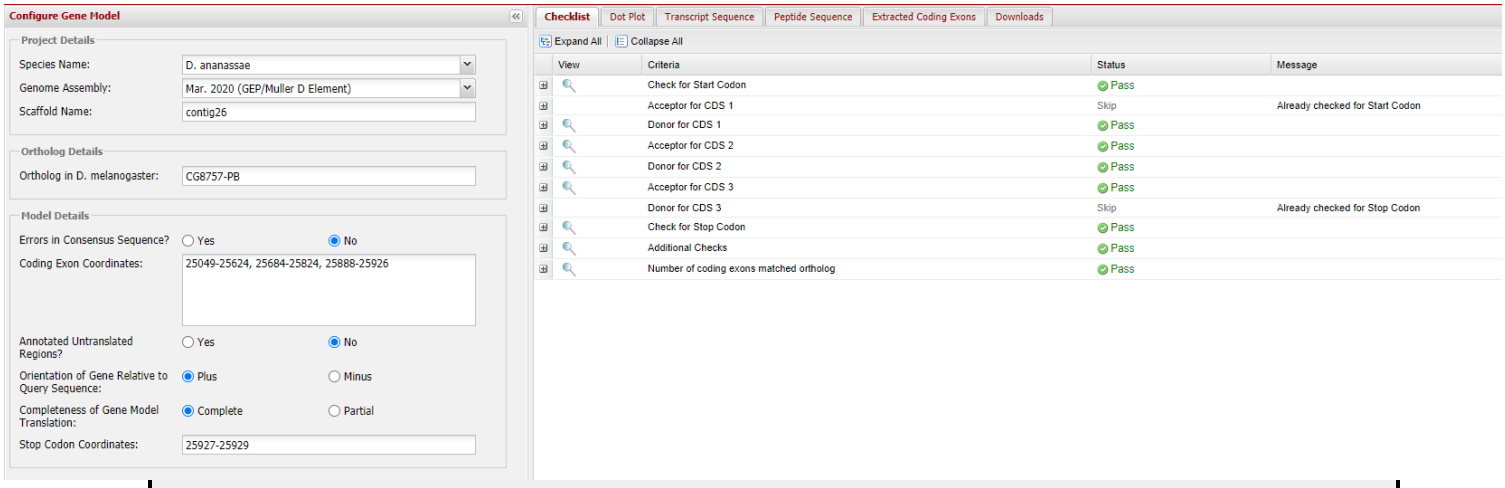
Is the 5' end of this isoform missing from the end of the project? No
If so, how many putative exons are missing from the 5' end: _____
Is the 3' end of this isoform missing from the end of the project? No
If so, how many putative exons are missing from the 3' end: _____

(Define "putative exons" based on the exons present in the *D. melanogaster* ortholog)

1. Gene Model Checker checklist

Enter the coordinates of your final gene model for this isoform into the Gene Model Checker and **paste a screenshot of the checklist results into the box below:**

Note: For projects with consensus sequence errors, report the exon coordinates relative to the **original project sequence**. Include the VCF file you have generated above when you submit the gene model to the Gene Model Checker. The Gene Model Checker will use this VCF file to automatically revise the submitted exon coordinates.



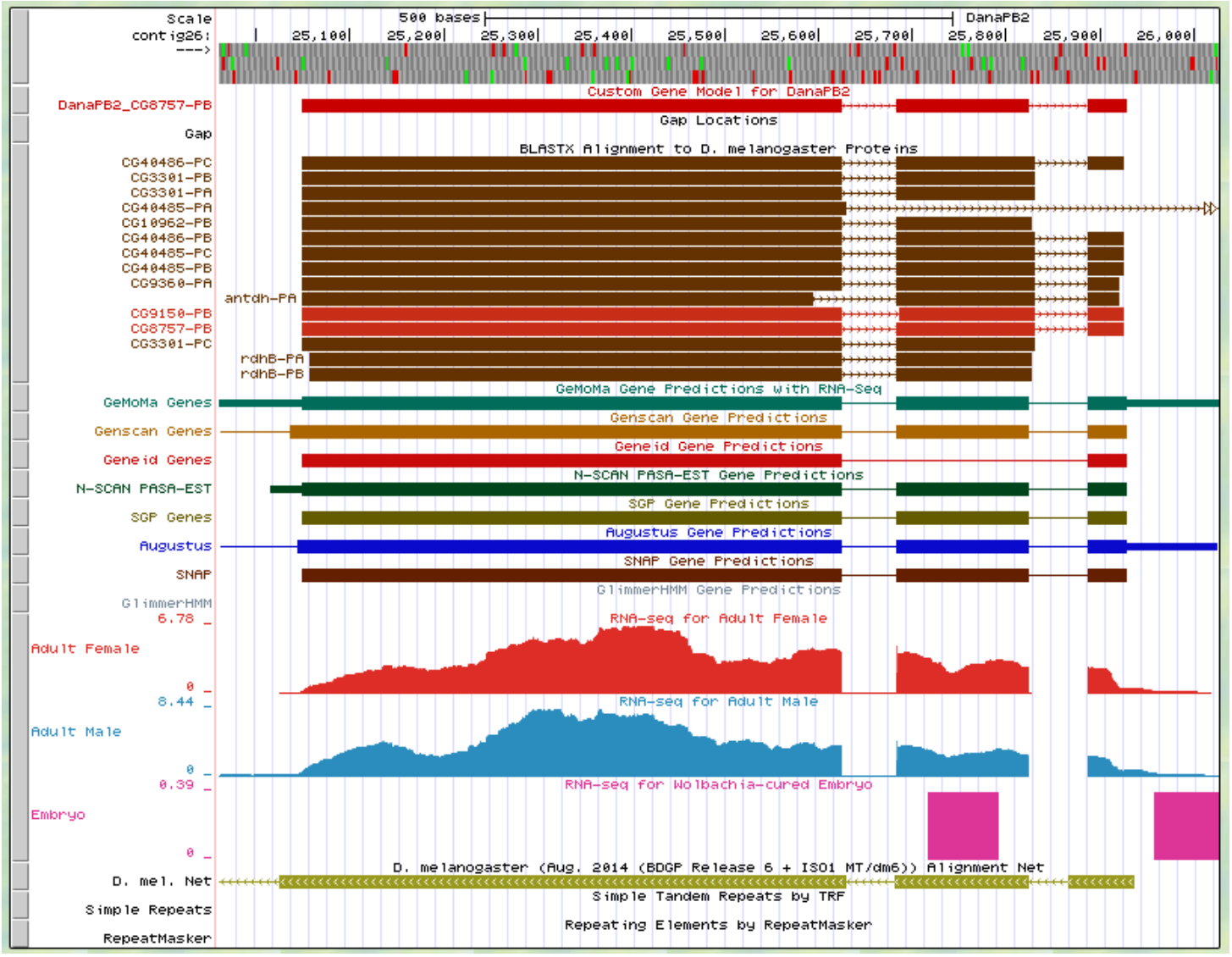
2. View the gene model on the Genome Browser

Click on the magnifying glass icon under the "Checklist" tab of the Gene Model Checker to view your gene model on the GEP UCSC Genome Browser. Zoom in so that **only this isoform is in the genome browser window, and capture a screenshot that includes the following evidence tracks if they are available:**

1. A sequence alignment track (*e.g.*, D. mel Proteins)

2. At least one gene prediction track (*e.g.*, Genscan)
3. At least one RNA-Seq track (*e.g.*, RNA-Seq Coverage)
4. A comparative genomics track (*e.g.*, D. mel. Net Alignment, Conservation)

Paste a screenshot of your gene model as shown on the GEP UCSC Genome Browser into the box below:



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

Show an alignment between the protein sequence for your gene model and the protein sequence from the putative *D. melanogaster* ortholog. You can either use the protein alignment generated by the Gene Model Checker (available through the "**View protein alignment**" link under the "Dot Plot" tab) or you can generate a new alignment using the "Align two or more sequences" feature at the NCBI BLAST web site. **Paste a screenshot of the protein alignment into the box below:**

Alignment of Dmel_CG8757-PB vs. DanaPB2_CG8757-PB

View plain text version

Download alignment image

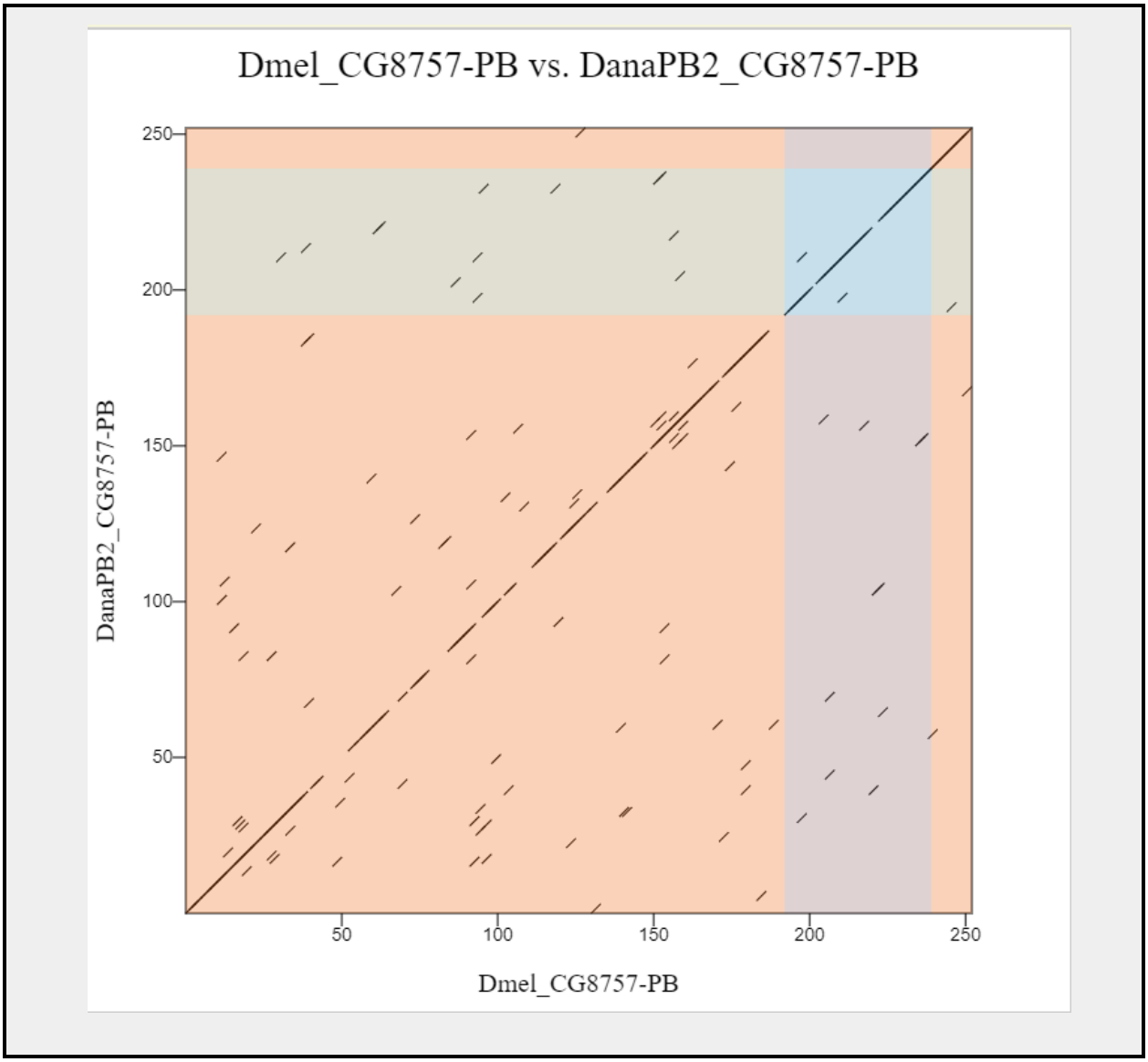
Identity: 192/252 (76.2%), **Similarity:** 223/252 (88.5%), **Gaps:** 0/252 (0.0%)

```
Dmel_CG8757-PB        1 MERWCNKVAVVSGASAGIGAACTRALIGAGMIVVGLARRHERVEKLRSGLSLEQQSRLHA 60
                        **** *****:****:******:***:****:*******::***:**. ** *:***:**
DanaPB2_CG8757-PB     1 MERWRNKVAVISGASSGIGAACARALVGAGMVVVGLARRQDRVEQLRQELSAEKQSRVHA 60

Dmel_CG8757-PB       61 IKCDITQEDQVLKAFDWTCRQLGGVDVLVSNAGIIGTGELSERDDGPAMRSTIETNIMGT 120
                        *:**:::::** ***:*   :**.****:**** :.***** :**. ** **:***:**
DanaPB2_CG8757-PB     61 IRCDVSRNEQVAKAFEWIKMNLGAVDVLISNAGTMATGELSGQDDAGAMLSTMETNVMGG 120

Dmel_CG8757-PB       121 VYCVRESFRSMKRRGTEGHVVIVNSVAGYQVPNLGPQLPSLNIYPATKFALRAMNEIYRQ 180
                        ***:**:*:**:.* :******:***** ****************:**** ********
DanaPB2_CG8757-PB     121 VYCIREAFQSMRERERAEGHVVIMNSVAGQQVPNLGPQLPSLNIYPASKFALTAMNEIYRQ 180

Dmel_CG8757-PB       181 EFQRHKTAVRVSTVSPGIVDTVILPEQIQGIIKQHMPMLRSDDVADAVLWAIGTPPNVQV 240
                        ******* *:*:*:******* ***:*************:. ******************
DanaPB2_CG8757-PB     181 EFQRHKTLVKVTTISPGIVDTDILPQQIQGIIKQHMPMLKCADVADAVLWAIGTPPNVQV 240

Dmel_CG8757-PB       241 HNITIKPQGEKF 252
                        **:*********
DanaPB2_CG8757-PB     241 HNVTIKPQGEKF 252
```

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

Paste a screenshot of the dot plot (generated by the Gene Model Checker) of your submitted model against the putative *D. melanogaster* ortholog into the box below. Provide an explanation for any anomalies on the dot plot (*e.g.*, large gaps, regions with no sequence similarity, indications of significant insertions or deletions).

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 and provide a justification as to why you have selected this particular set of donor and acceptor sites.



Transcription Start Sites (TSS) Report Form (Optional)

Note: Complete this section if you have annotated the TSS for the gene above. This section is **optional** and you do not need to complete this section to submit the project.

Gene name (*e.g.*, *D. ananassae eyeless*): *D. ananassae_CG8757*
Gene symbol (*e.g.*, *dana_ey*): *dana_CG8757*

Name(s) of isoform(s) with unique TSS	List of isoforms with identical TSS
CG8757-RB	

Names of the isoforms with unique TSS in *D. melanogaster* that are absent in this species:

_____

Provide the evidence (text and figures) which support the hypothesis that these isoforms are absent in this species (*e.g.*, changes in canonical splice sites, gene structure, etc.):

Isoform TSS Report

Complete an Isoform TSS report (through page 71) for each unique TSS listed in the table above. Copy and paste this form to create as many copies as needed.

Gene-isoform name (*e.g.*, *dana_ey-RA*): *dana_CG8757-RB*

Names of the isoforms with the same TSS as this isoform:

_____

Type of core promoter in *D. melanogaster* (see table below):
(**Peaked** / Intermediate / Broad / Insufficient Evidence)

_____

The type of core promoter is defined by the number of TSS annotated by the Celniker group at modENCODE and the number of DHS positions:

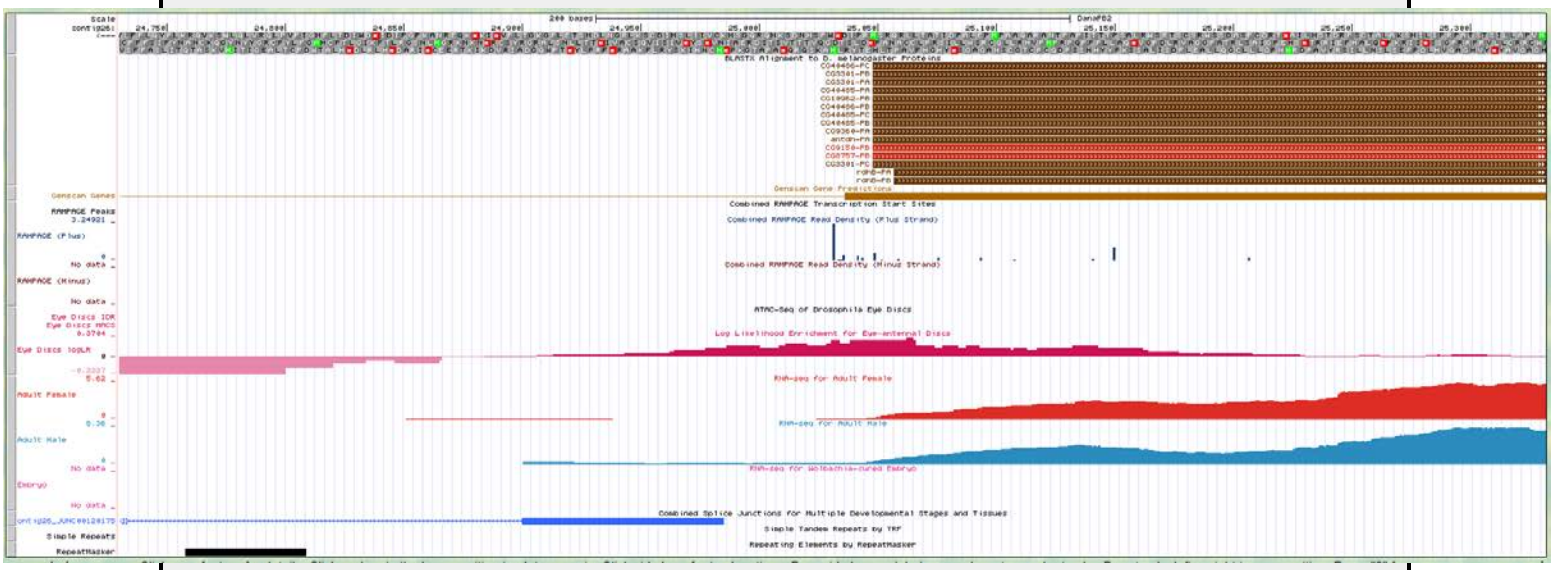
Type of core promoter	# annotated TSS	# DHS positions
Peaked	1	0
	0	1
	1	1
Intermediate	≤ 1	> 1
	> 1	≤ 1
Broad	> 1	> 1
Insufficient Evidence	0	0

1. Turn on RAMPAGE evidence tracks

Coordinates of the TSS position based on position with the highest RAMPAGE read density
25,031-25,032

Coordinates of the narrow TSS search region based on RAMPAGE peaks
24,731-25,332

If the TSS position and narrow TSS search region are supported by RAMPAGE data, **paste a Genome Browser screenshot of the region surrounding the putative TSS (±300bp) showing the Combined RAMPAGE TSS evidence track:**



2. Turn on ATAC-Seq evidence track

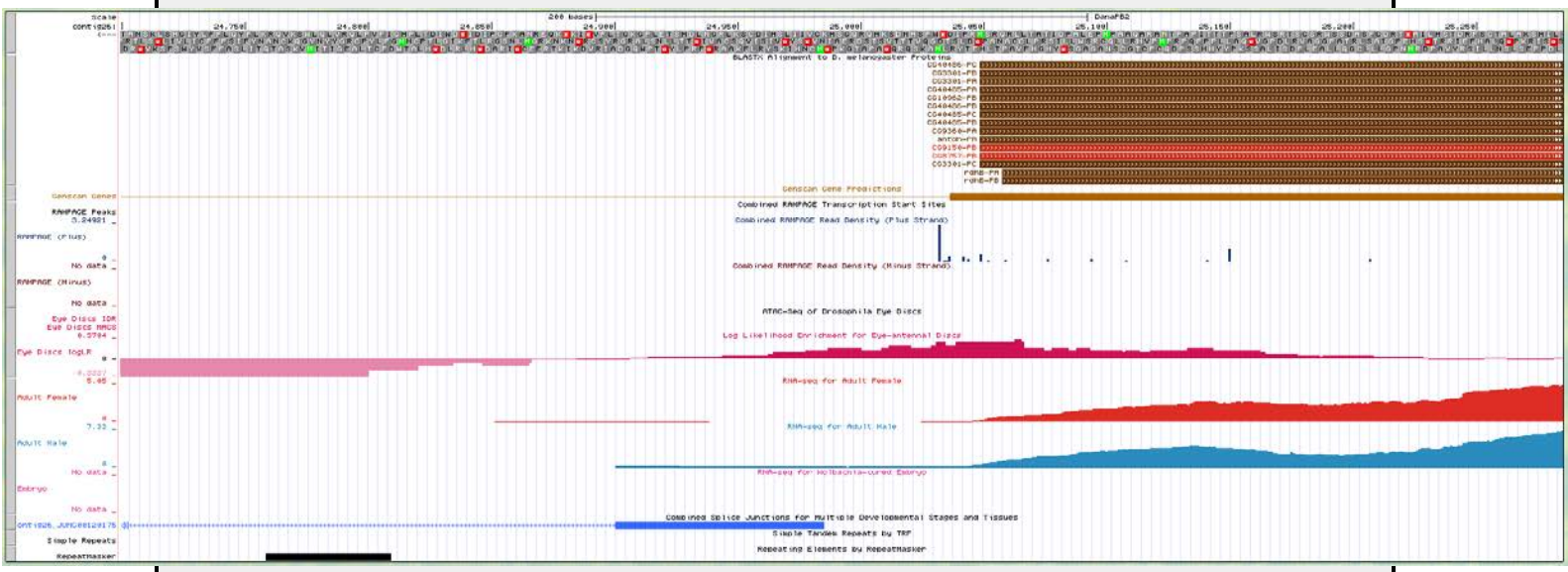
If the wide TSS search region is supported by ATAC-Seq data, **paste a Genome Browser screenshot of the region surrounding the putative TSS (±300bp) showing the Eye Discs ATAC-Seq evidence track:**

Coordinates of the wide TSS search region based on ATAC-Seq peaks
24,664-25,417

3. Turn on RNA-Seq evidence tracks

If the TSS annotation is supported by RNA-Seq read coverage or splice junction predictions (*e.g.*, regtools), **paste a Genome Browser screenshot of the region surrounding the putative TSS (±300bp) showing the following evidence tracks:**

1. RNA-Seq Coverage or RNA-Seq Alignment Summary
2. Combined Splice Junctions or RNA-Seq TopHat



If the RNA-Seq evidence tracks indicate the TSS position, list it here: 24,700-25,285

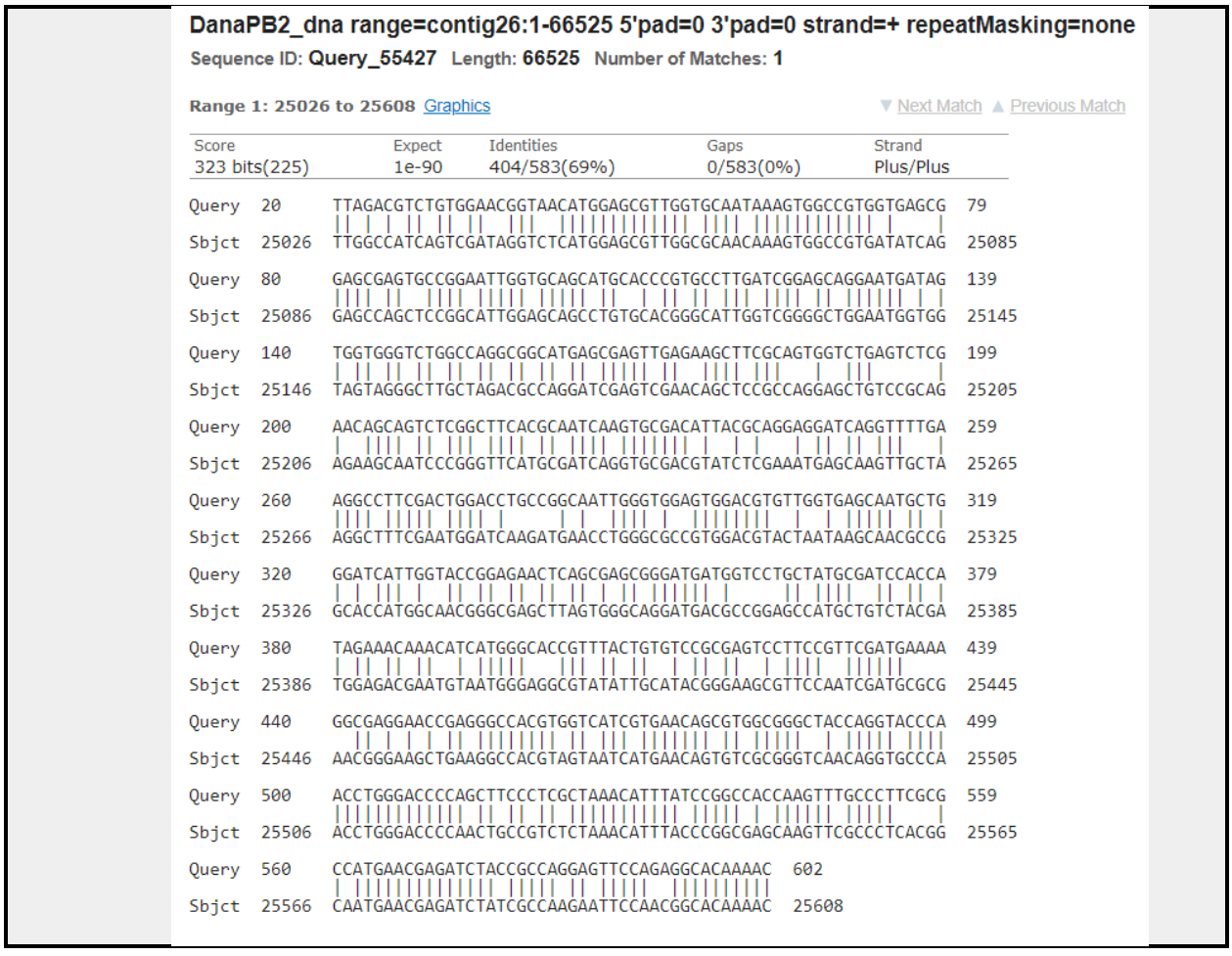
4. Annotate the first transcribed exon

Coordinates of the first transcribed exon based on blastn alignment:
25,006-25,623

Does the blastn alignment cover the entire *D. melanogaster* first transcribed exon?
No

If not, specify the parts of the *D. melanogaster* exon that are missing from the blastn alignment.
1-20, 603-618

If the TSS annotation is supported by blastn alignment of the initial transcribed exon against the contig sequence, **paste a screenshot of the blastn alignment into the box below:**



DanaPB2_dna range=contig26:1-66525 5'pad=0 3'pad=0 strand=+ repeatMasking=none

Sequence ID: Query_55427 Length: 66525 Number of Matches: 1

Range 1: 25026 to 25608 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
323 bits(225)	1e-90	404/583(69%)	0/583(0%)	Plus/Plus

```
Query  20     TTAGACGTCTGTGGAACGGTAACATGGAGCGTTGGTGCAATAAAGTGGCCGTGGTGAGCG  79
              || | | || || ||  |||  |||||||||||||| |||| ||||||||||||     |
Sbjct  25026  TTGGCCATCAGTCGATAGGTCTCATGGAGCGTTGGCGCAACAAAGTGGCCGTGATATCAG  25085

Query  80     GAGCGAGTGCCGGAATTGGTGCAGCATGCACCCGTGCCTTGATCGGAGCAGGAATGATAG  139
              |||| ||  |||| ||||| ||||| ||  | || || ||| |||| || |||||| | |
Sbjct  25086  GAGCCAGCTCCGGCATTGGAGCAGCCTGTGCACGGGCATTGGTCGGGGCTGGAATGGTGG  25145

Query  140    TGGTGGGTCTGGCCAGGCGGCATGAGCGAGTTGAGAAGCTTCGCAGTGGTCTGAGTCTCG  199
              | || || || || || || || ||||| ||  |||| |||   |  |||       |
Sbjct  25146  TAGTAGGGCTTGCTAGACGCCAGGATCGAGTCGAACAGCTCCGCCAGGAGCTGTCCGCAG  25205

Query  200    AACAGCAGTCTCGGCTTCACGCAATCAAGTGCGACATTACGCAGGAGGATCAGGTTTTGA  259
              |  |||| || ||| |||| || |||| |||||||| |  | |   | || || |||   |
Sbjct  25206  AGAAGCAATCCCGGGTTCATGCGATCAGGTGCGACGTATCTCGAAATGAGCAAGTTGCTA  25265

Query  260    AGGCCTTCGACTGGACCTGCCGGCAATTGGGTGGAGTGGACGTGTTGGTGAGCAATGCTG  319
              |||| ||||| |||| |      | |  |||| |  ||||||||  |  | ||||| || |
Sbjct  25266  AGGCTTTCGAATGGATCAAGATGAACCTGGGCGCCGTGGACGTACTAATAAGCAACGCCG  25325

Query  320    GGATCATTGGTACCGGAGAACTCAGCGAGCGGGATGATGGTCCTGCTATGCGATCCACCA  379
              | | ||| |  || || || || || | || |||||| |      || ||||  || || |
Sbjct  25326  GCACCATGGCAACGGGCGAGCTTAGTGGGCAGGATGACGCCGGAGCCATGCTGTCTACGA  25385

Query  380    TAGAAACAAACATCATGGGCACCGTTTACTGTGTCCGCGAGTCCTTCCGTTCGATGAAAA  439
              | || || ||  | |||||   ||| || ||  | || ||  | ||||  ||||||
Sbjct  25386  TGGAGACGAATGTAATGGGAGGCGTATATTGCATACGGGAAGCGTTCCAATCGATGCGCG  25445

Query  440    GGCGAGGAACCGAGGGCCACGTGGTCATCGTGAACAGCGTGGCGGGCTACCAGGTACCCA  499
              || | | | || |||||||| || ||| ||||||| || |||||  | |||||  ||||
Sbjct  25446  AACGGGAAGCTGAAGGCCACGTAGTAATCATGAACAGTGTCGCGGGTCAACAGGTGCCCA  25505

Query  500    ACCTGGGACCCCAGCTTCCCTCGCTAAACATTTATCCGGCCACCAAGTTTGCCCTTCGCG  559
              |||||||||||||| || || || |||||||||||| ||||| | |||||| |||||    |
Sbjct  25506  ACCTGGGACCCCAACTGCCGTCTCTAAACATTTACCCGGCGAGCAAGTTCGCCCTCACGG  25565

Query  560    CCATGAACGAGATCTACCGCCAGGAGTTCCAGAGGCACAAAAC  602
              | |||||||||||||| ||||| || |||||   ||||||||||
Sbjct  25566  CAATGAACGAGATCTATCGCCAAGAATTCCAACGGCACAAAAC  25608
```

5. Turn on comparative genomics tracks

If the TSS annotation is supported by sequence conservation with other *Drosophila* species, **paste a screenshot of the multiple sequence alignment (*e.g.*, from Clustal Omega, ROAST) into the box below:**

6. Summarize the evidence that supports the TSS annotation postulated above

Coordinate(s) of the TSS position(s):

Based on RAMPAGE data: 24,731-25,332
Based on ATAC-Seq data: 24,664-25,417
Based on RNA-Seq data: 24,700-25,285
Based on blastn alignment: 25,006-25,623
Based on other evidence (please specify): _____

Note: If the blastn alignment for the initial transcribed exon is a partial alignment, you can **extrapolate the TSS position** based on the number of nucleotides that are missing from the beginning of the exon. (Enter "Insufficient evidence" if you cannot determine the TSS position based on the available evidence.)

Were you able to define a TSS position based on the available evidence? Yes
If so, indicate in the table below the evidence that supports this TSS position

If not, were you able to define a TSS search region? _____
If so, indicate in the table below the evidence that supports the TSS search region(s)

Evidence type	Support	Refute	Neither
RAMPAGE peaks and read density	☒	☐	☐
ATAC-Seq peaks and log likelihood enrichment profile	☒	☐	☐
RNA-Seq coverage and splice junctions	☒	☐	☐
blastn alignment of the initial exon from *D. melanogaster*	☒	☐	☐
Sequence conservation with other *Drosophila* species (*e.g.*, "Conservation" track on the Genome Browser)	☐	☐	☒
Other (please specify) [*e.g.*, Gnomon, N-SCAN, Augustus TSS predictions; histone modifications (ChIP-Seq data)].	☐	☐	☒

Note: The evidence type refutes the TSS annotation only if it **suggests an alternate TSS position**. For example, the presence of RNA-Seq read coverage upstream of the annotated TSS indicates that the TSS is located further upstream and it would be considered to be evidence against the annotated TSS; check "Refute." In contrast, the lack of RNA-Seq read coverage is a negative result that neither supports nor refutes the TSS annotation; check "Neither."

Gene Report Form

Gene name (*e.g.*, *D. ananassae eyeless*): *D. ananassae Hemolectin*
Gene symbol (*e.g.*, *dana_ey*): *dana_Hml*

Approximate location in project (from 5' end to 3' end): 27,564-42,593
Number of isoforms in *D. melanogaster:* 2
Number of isoforms in this project: 2

Complete the following table, including all of the isoforms in this project:

Name(s) of unique isoform(s) based on coding sequence	**List of isoforms with identical coding sequences**
Hml-RA	*Hml-RB*

Names of the isoforms with unique coding sequences in *D. melanogaster* that are absent in this species: _____

Provide the evidence (text and figures) which support the hypothesis that these isoforms are absent in this species (*e.g.*, changes in canonical splice sites, gene structure, etc.):

Note: For isoforms with identical coding sequence, you only need to complete the Isoform Report Form for one of these isoforms (i.e. using the name of the isoform listed in the left column of the table above). However, you should **generate GFF, transcript, and peptide sequence files for ALL isoforms**, irrespective of whether their coding sequence is identical to that of another isoform.

Consensus Sequence Errors Report Form

Complete this section if you have identified errors in the project consensus sequence that affect the annotation of the gene described above.

All of the coordinates reported in this section should be relative to the coordinates of the original project sequence.

Location(s) in the project sequence with consensus errors:

_____

1. Evidence that supports the consensus errors postulated above

Note: Evidence that could be used to support the hypothesis of errors within the consensus sequence includes a CDS alignment with frame shifts or in-frame stop codons, and RNA-Seq reads with discrepant alignments compared to the project sequence.

2. Generate a VCF file which describes the changes to the consensus sequence

Use the Sequencer Updater to create a Variant Call Format (VCF) file that describes the changes to the consensus sequence you have identified above. **Paste a screenshot with the list of sequence changes into the box below:**

Isoform Report Form

Complete this report form for each unique isoform listed in the table above. Copy and paste this form to create as many copies of this Isoform Report Form as needed.

Gene-isoform symbol (*e.g.*, dana_ey-PA): *dana_Hml-RA*

Names of any additional isoforms with identical coding sequences: *dana_Hml-RB*

_____

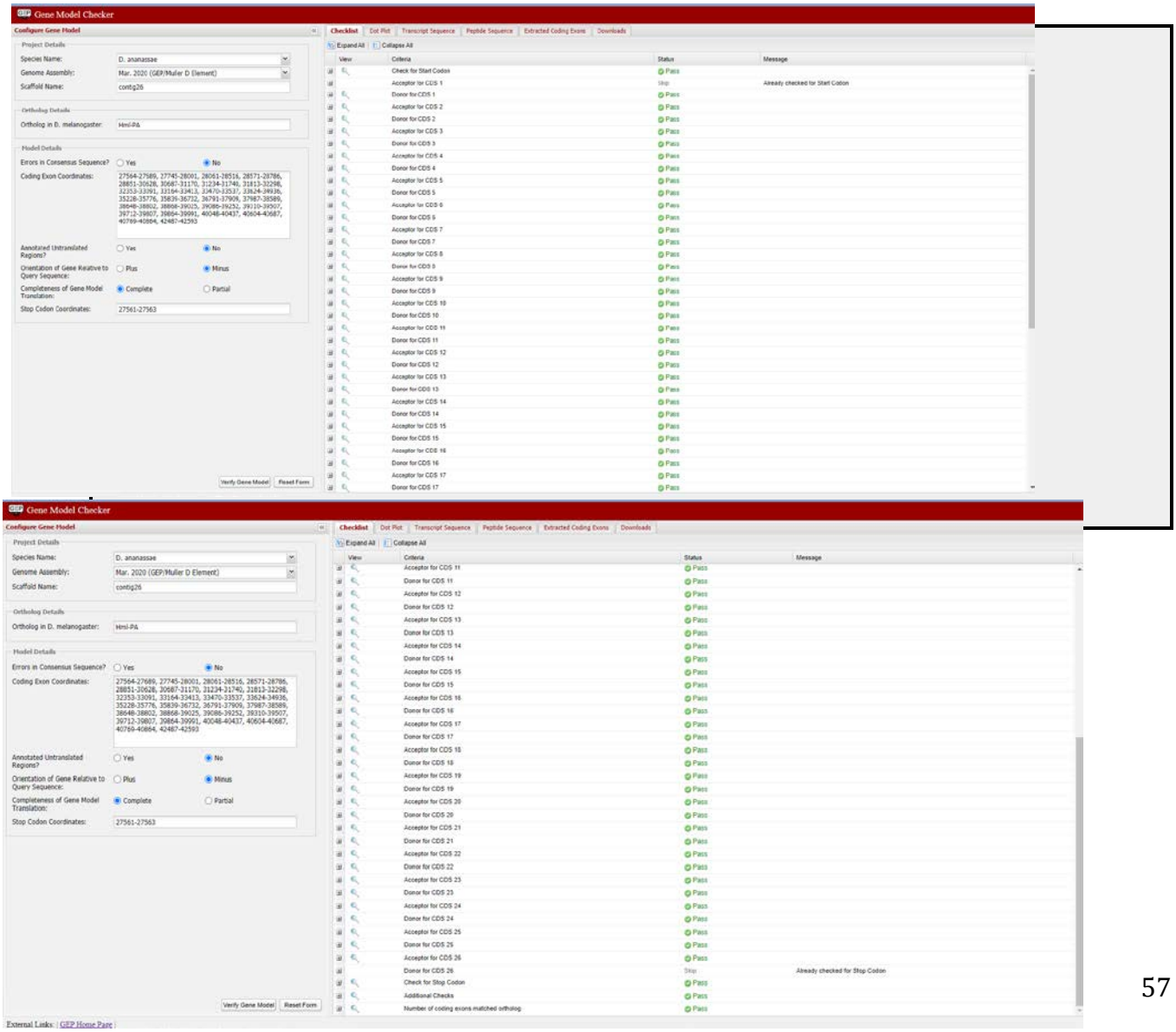
Is the 5' end of this isoform missing from the end of the project? No
If so, how many putative exons are missing from the 5' end: _____
Is the 3' end of this isoform missing from the end of the project? No
If so, how many putative exons are missing from the 3' end: _____

(Define "putative exons" based on the exons present in the *D. melanogaster* ortholog)

1. Gene Model Checker checklist

Enter the coordinates of your final gene model for this isoform into the Gene Model Checker and **paste a screenshot of the checklist results into the box below:**

Note: For projects with consensus sequence errors, report the exon coordinates relative to the **original project sequence**. Include the VCF file you have generated above when you submit the gene model to the Gene Model Checker. The Gene Model Checker will use this VCF file to automatically revise the submitted exon coordinates.

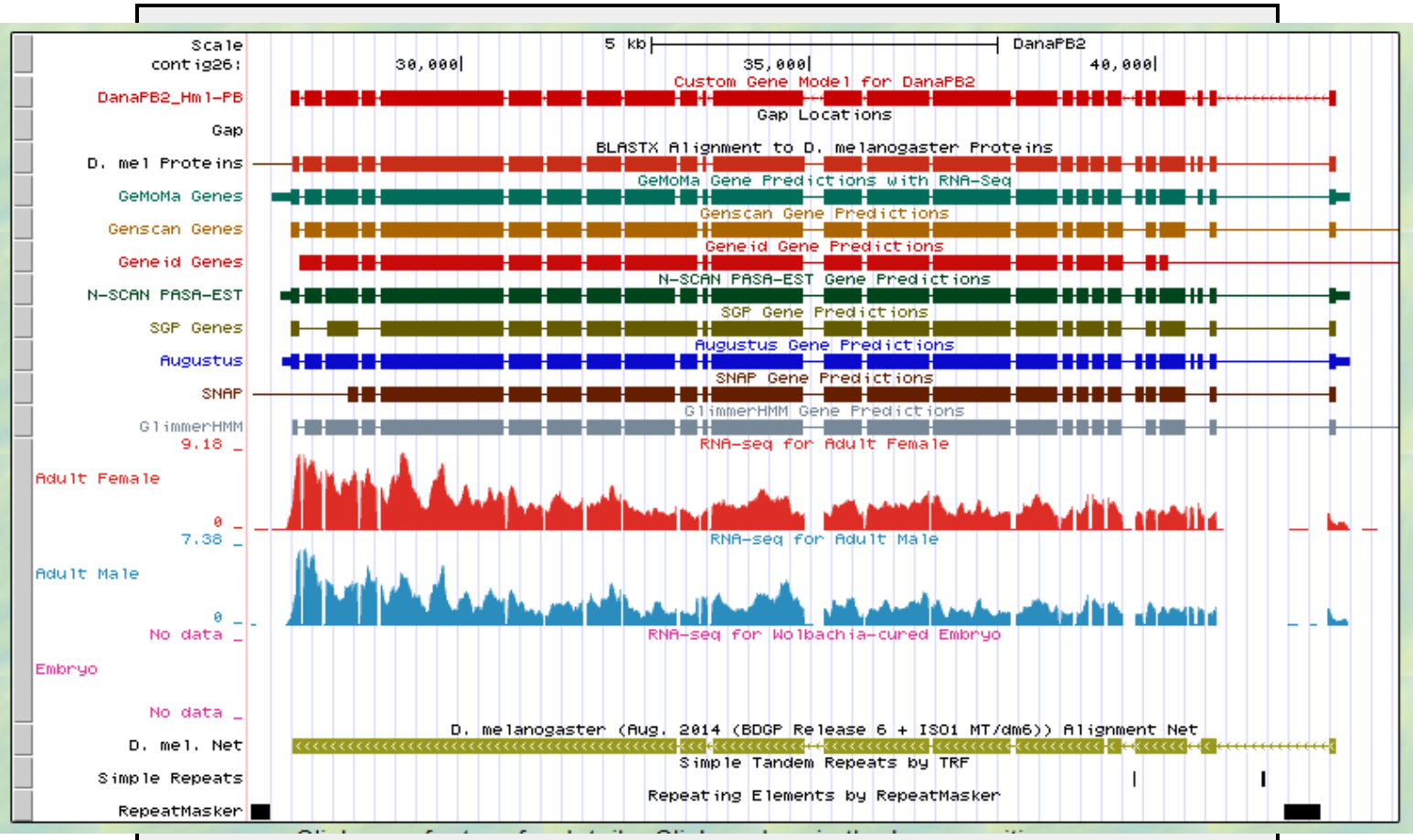


2. View the gene model on the Genome Browser

Click on the magnifying glass icon under the "Checklist" tab of the Gene Model Checker to view your gene model on the GEP UCSC Genome Browser. Zoom in so that **only this isoform is in the genome browser window, and capture a screenshot that includes the following evidence tracks if they are available:**

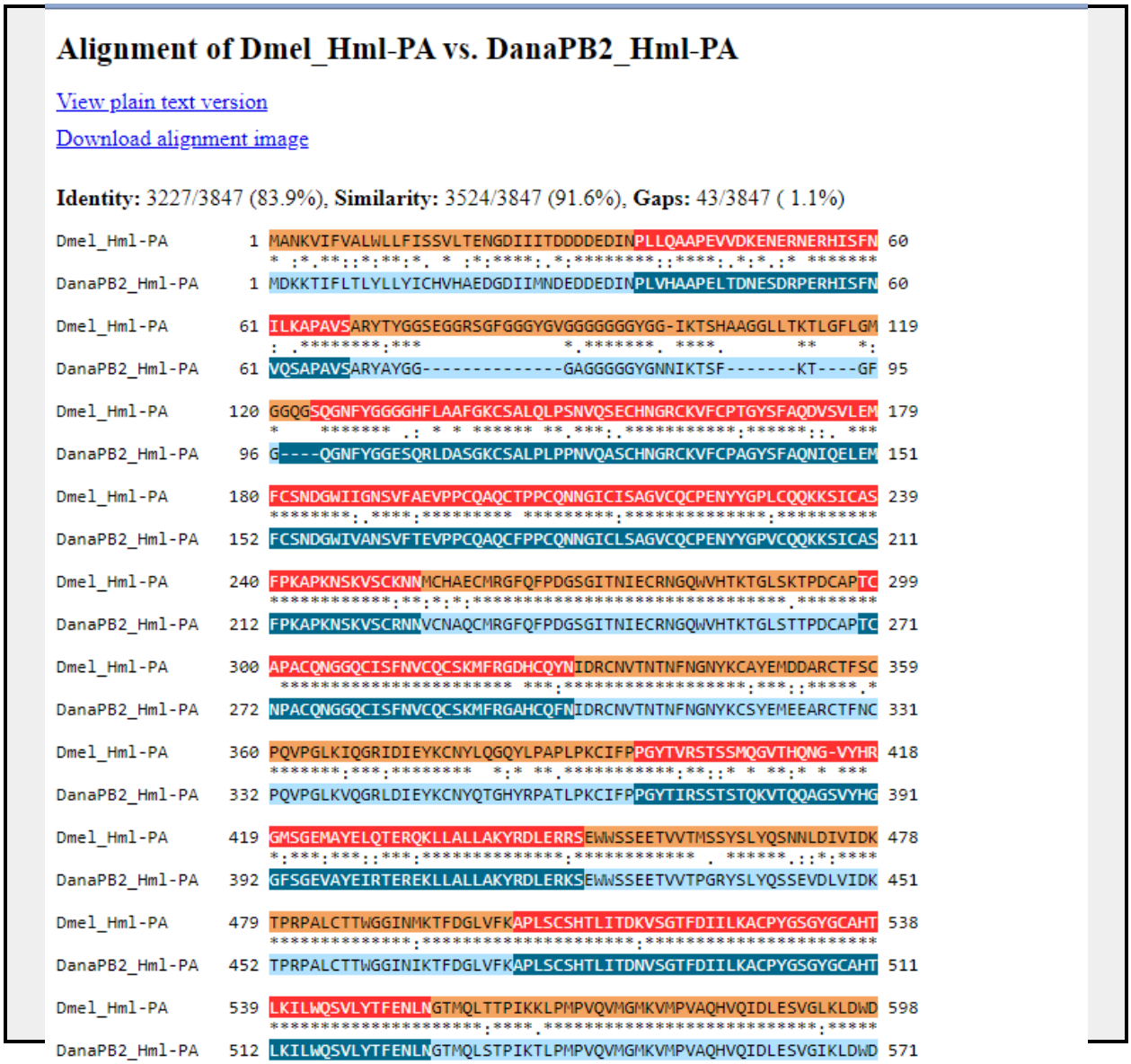
1. A sequence alignment track (*e.g.*, D. mel Proteins)
2. At least one gene prediction track (*e.g.*, Genscan)
3. At least one RNA-Seq track (*e.g.*, RNA-Seq Coverage)
4. A comparative genomics track (*e.g.*, D. mel. Net Alignment, Conservation)

Paste a screenshot of your gene model as shown on the GEP UCSC Genome Browser into the box below:



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

Show an alignment between the protein sequence for your gene model and the protein sequence from the putative *D. melanogaster* ortholog. You can either use the protein alignment generated by the Gene Model Checker (available through the "**View protein alignment**" link under the "Dot Plot" tab) or you can generate a new alignment using the "Align two or more sequences" feature at the NCBI BLAST web site. **Paste a screenshot of the protein alignment into the box below:**



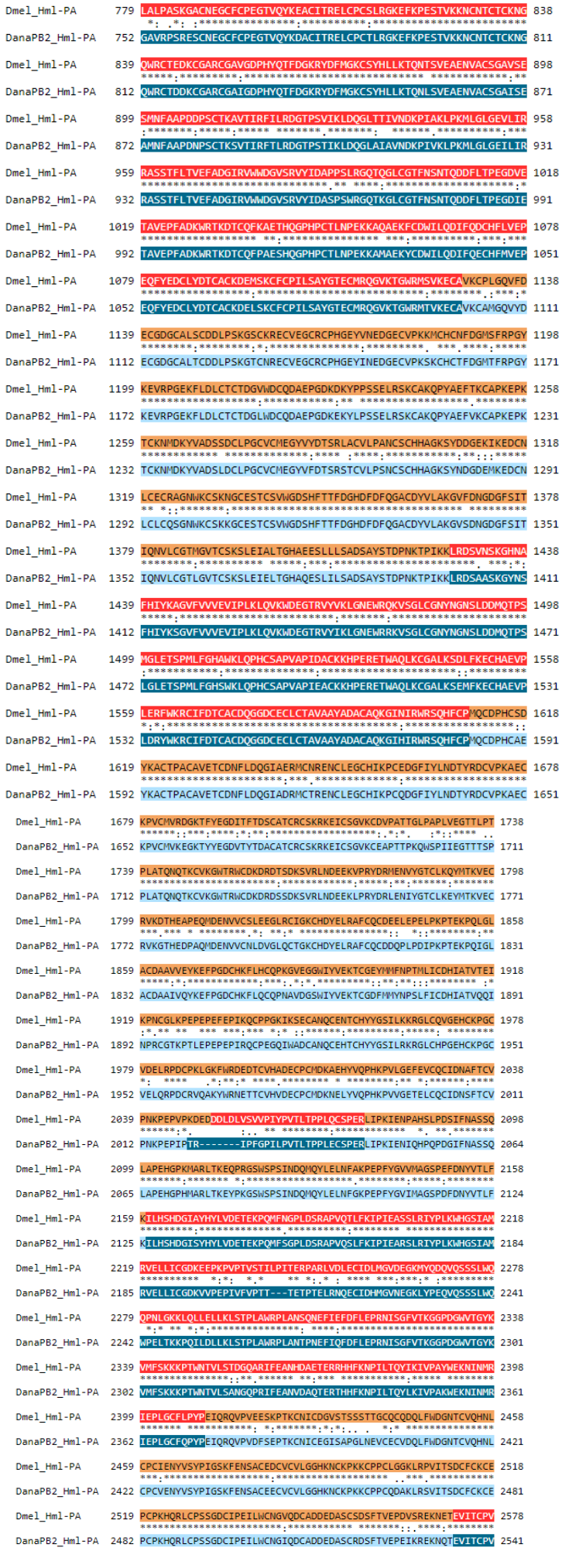
Alignment of Dmel_Hml-PA vs. DanaPB2_Hml-PA

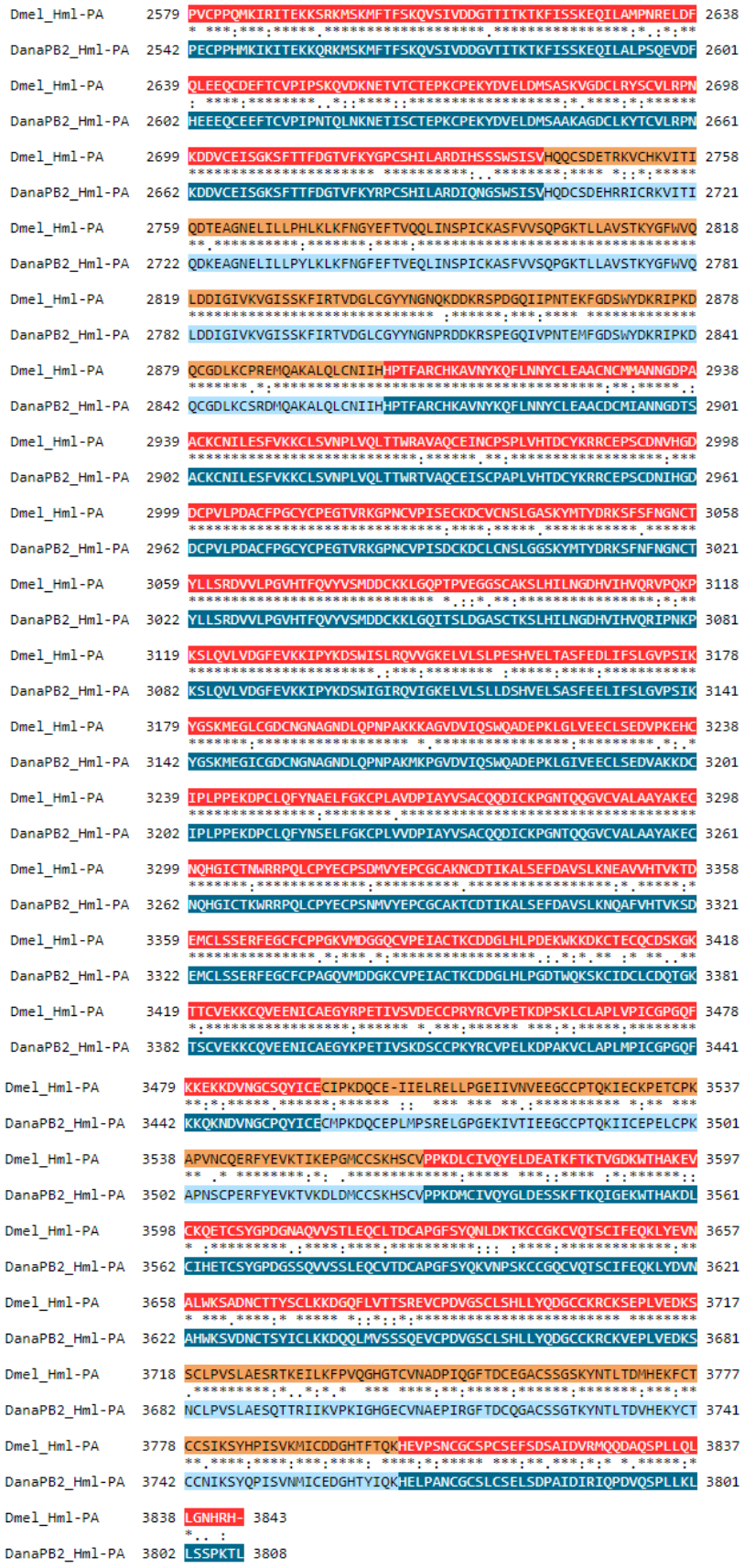
View plain text version

Download alignment image

Identity: 3227/3847 (83.9%), **Similarity:** 3524/3847 (91.6%), **Gaps:** 43/3847 (1.1%)

```
Dmel_Hml-PA       1 MANKVIFVALWLLFISSVLTENGDIIITDDDDEDINPLLQAAPEVVDKENERNERHISFN 60
                    * :*.**::*:**:*. * :*:****:.*:********::****:.*:*.:* *******
DanaPB2_Hml-PA    1 MDKKTIFLTLYLLYICHVHAEDGDIIMNDEDDEDINPLVHAAPELTDNESDRPERHISFN 60

Dmel_Hml-PA      61 ILKAPAVSARYTYGGSEGGRSGFGGGYGVGGGGGGGYGG-IKTSHAAGGLLTKTLGFLGM 119
                    : .********:***               *.*******. ****.       **    *:
DanaPB2_Hml-PA    61 VQSAPAVSARYAYGG--------------GAGGGGGYGNNIKTSF-------KT----GF 95

Dmel_Hml-PA      120 GGQGSQGNFYGGGGHFLAAFGKCSALQLPSNVQSECHNGRCKVFCPTGYSFAQDVSVLEM 179
                    *    ******* .: * * ****** **.***:.***********:******::. ***
DanaPB2_Hml-PA    96 G----QGNFYGGGESQRLDASGKCSALPLPPNVQASCHNGRCKVFCPAGYSFAQNIQELEM 151

Dmel_Hml-PA      180 FCSNDGWIIGNSVFAEVPPCQAQCTPPCQNNGICISAGVCQCPENYYGPLCQQKKSICAS 239
                    ********:.****:********* *********:**********:**********
DanaPB2_Hml-PA   152 FCSNDGWIVANSVFTEVPPCQAQCFPPCQNNGICLSAGVCQCPENYYGPVCQQKKSICAS 211

Dmel_Hml-PA      240 FPKAPKNSKVSCKNNMCHAECMRGFQFPDGSGITNIECRNGQWVHTKTGLSKTPDCAPTC 299
                    ************:**:*:*:*********************************.********
DanaPB2_Hml-PA   212 FPKAPKNSKVSCRNNVCNAQCMRGFQFPDGSGITNIECRNGQWVHTKTGLSTTPDCAPTC 271

Dmel_Hml-PA      300 APACQNGGQCISFNVCQCSKMFRGDHCQYNIDRCNVTNTNFNGNYKCAYEMDDARCTFSC 359
                      ********************** ***:*****************:***::*****.*
DanaPB2_Hml-PA   272 NPACQNGGQCISFNVCQCSKMFRGAHCQFNIDRCNVTNTNFNGNYKCSYEMEEARCTFNC 331

Dmel_Hml-PA      360 PQVPGLKIQGRIDIEYKCNYLQGQYLPAPLPKCIFPPGYTVRSTSSMQGVTHQNG-VYHR 418
                    *******:***:********  *:* **.***********:**::* * **:* * ***
DanaPB2_Hml-PA   332 PQVPGLKVQGRLDIEYKCNYQTGHYRPATLPKCIFPPGYTIRSSTSTQKVTQQAGSVYHG 391

Dmel_Hml-PA      419 GMSGEMAYELQTERQKLLALLAKYRDLERRSEWNSSEETVVTMSSYSLYQSNNLDIVIDK 478
                    *:***:***::***:*************:***********  . ******.::*:****
DanaPB2_Hml-PA   392 GFSGEVAYEIRTEREKLLALLAKYRDLERKSEWNSSEETVVTPGRYSLYQSSEVDLVIDK 451

Dmel_Hml-PA      479 TPRPALCTTWGGINMKTFDGLVFKAPLSCSHTLITDKVSGTFDIILKACPYGSGYGCAHT 538
                    **************:**********************:**********************
DanaPB2_Hml-PA   452 TPRPALCTTWGGINIKTFDGLVFKAPLSCSHTLITDNVSGTFDIILKACPYGSGYGCAHT 511

Dmel_Hml-PA      539 LKILWQSVLYTFENLNGTMQLTTPIKKLPMPVQVMGMKVMPVAQHVQIDLESVGLKLDWD 598
                    **********************:****.*************************:******
DanaPB2_Hml-PA   512 LKILWQSVLYTFENLNGTMQLSTPIKTLPMPVQVMGMKVMPVAQHVQIDLESVGIKLDWD 571

Dmel_Hml-PA      599 HRQYVSVQAGPQMWGKVGGLCGTLDGDPNTDLTSRTGKKLATVKAFADAWRVEDRSELCQ 658
                    *******:**************:****  ****.*:***** ***************:**
DanaPB2_Hml-PA   572 HRQYVSVHAGPQMWGKVGGLCGSLDGDFNTDLVSKTGKKLETVKAFADAWRVEDRSEMCQ 631

Dmel_Hml-PA      659 VENSAEMEFGMDSCEQSKLQKAVSVCERLLANEKLGDCIKPFNYDALIRTCMADYCNCAN 718
                    ******:**::** *:*:***.*********************:****:**********
DanaPB2_Hml-PA   632 VENSAELDFGVESCPQAKMQKAASVCERLLANEKLGDCIKPFNYEALIRSCMADYCNCAN 691

Dmel_Hml-PA      719 REHPESCNCDAIAMLAKECAFKGIKLEHGWRNLEICPISCGFGRVYQACGPNVEPTCDSD 778
                    :*********:**************************************************:
DanaPB2_Hml-PA   692 QEHPESCNCDSIAMLAKECAFKGIKLEHGWRNLEICPISCGFGRVYQACGPNVEPTCDSN 751
```

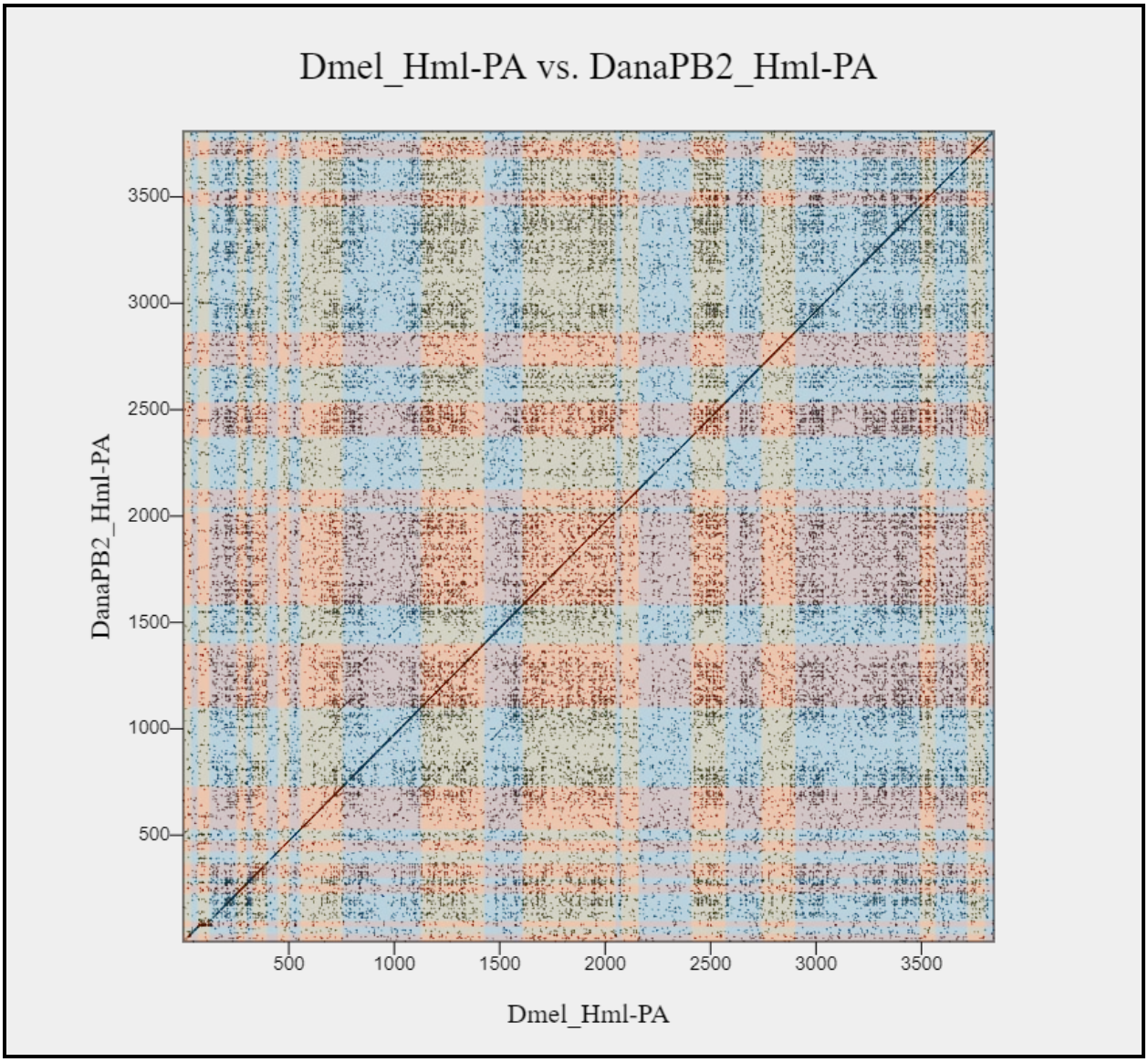

Dmel_Hml-PA 779 LALPASKGACNEGCFCPEGTVQYKEACITRELCPCSLRGKEFKPESTVKKNCNTCTCKNG 838
DanaPB2_Hml-PA 752 GAVRPSRESCNEGCFCPEGTVQYKDACITRELCPCTLRGKEFKPESTVKKNCNTCTCKNG 811
Dmel_Hml-PA 839 QWRCTEDKCGARCGAVGDPHYQTFDGKRYDFMGKCSYHLLKTQNTSVEAENVACSGAVSE 898
DanaPB2_Hml-PA 812 QWRCTDDKCGARCGAIGDPHYQTFDGKRYDFMGKCSYHLLKTQNLSVEAENVACSGAISE 871
Dmel_Hml-PA 899 SMNFAAPDDPSCTKAVTIRFILRDGTPSVIKLDQGLTTIVNDKPIAKLPKMLGLGEVLIR 958
DanaPB2_Hml-PA 872 AMNFAAPDNPSCTKSVTIRFTLRDGTPSTIKLDQGLAIAVNDKPIVKLPKMLGLGEILIR 931
Dmel_Hml-PA 959 RASSTFLTVEFADGIRVWWDGVSRVYIDAPPSLRGQTQGLCGTFNSNTQDDFLTPEGDVE 1018
DanaPB2_Hml-PA 932 RASSTFLTVEFADGIRVWWDGVSRVYIDASPSWRGQTKGLCGTFNSNTQDDFLTPEGDIE 991
Dmel_Hml-PA 1019 TAVEPFADKWRTKDTCQFKAETHQGPHPCTLNPEKKAQAEKFCDWILQDIFQDCHFLVEP 1078
DanaPB2_Hml-PA 992 TAVEPFADKWRTKDTCQFPAESHQGPHPCTLNPEKKAMAEKYCDWILQDIFQECHFMVEP 1051
Dmel_Hml-PA 1079 EQFYEDCLYDTCACKDEMSKCFCPILSAYGTECMRQGVKTGWRMSVKECAVKCPLGQVFD 1138
DanaPB2_Hml-PA 1052 EQFYEDCLYDTCACKDELSKCFCPILSAYGTECMRQGVKTGWRMTVKECAVKCAMGQVYD 1111
Dmel_Hml-PA 1139 ECGDGCALSCDDLPSKGSCKRECVEGCRCPHGEYVNEDGECVPKKMCHCNFDGMSFRPGY 1198
DanaPB2_Hml-PA 1112 ECGDGCALTCDDLPSKGTCNRECVEGCRCPHGEYINEDGECVPKSKCHCTFDGMTFRPGY 1171
Dmel_Hml-PA 1199 KEVRPGEKFLDLCTCTDGVWDCQDAEPGDKDKYPPSSELRSKCAKQPYAEFTKCAPKEPK 1258
DanaPB2_Hml-PA 1172 KEVRPGEKFLDLCTCTDGLWDCQDAEPGDKEKYLPSSELRSKCAKQPYAEFVKCAPKEPK 1231
Dmel_Hml-PA 1259 TCKNMDKYVADSSDCLPGCVCMEGYVYDTSRLACVLPANCSCHHAGKSYDDGEKIKEDCN 1318
DanaPB2_Hml-PA 1232 TCKNMDKYVADSLDCLPGCVCMEGYVFDTSRSTCVLPSNCSCHHAGKSYNDGDEMKEDCN 1291
Dmel_Hml-PA 1319 LCECRAGNWKCSKNGCESTCSVWGDSHFTTFDGHDFDFQGACDYVLAKGVFDNGDGFSIT 1378
DanaPB2_Hml-PA 1292 LCLCQSGNWKCSKKGCESTCSVWGDSHFTTFDGHDFDFQGACDYVLAKGVSDNGDGFSIT 1351
Dmel_Hml-PA 1379 IQNVLCGTMGVTCSKSLEIALTGHAEESLLLSADSAYSTDPNKTPIKKLRDSVNSKGHNA 1438
DanaPB2_Hml-PA 1352 IQNVLCGTLGVTCSKSLEIELTGHAQESLILSADSAYSTDPNKTPIKKLRDSAASKGYNS 1411
Dmel_Hml-PA 1439 FHIYKAGVFVVVEVIPLKLQVKWDEGTRVYVKLGNEWRQKVSGLCGNYNGNSLDDMQTPS 1498
DanaPB2_Hml-PA 1412 FHIYKSGVFVVVEVIPLKLQVKWDEGTRVYIKLGNEWRRKVSGLCGNYNGNSLDDMQTPS 1471
Dmel_Hml-PA 1499 MGLETSPMLFGHAWKLQPHCSAPVAPIDACKKHPERETWAQLKCGALKSDLFKECHAEVP 1558
DanaPB2_Hml-PA 1472 LGLETSPMLFGHSWKLQPHCSAPVAPIEACKKHPERETWAQLKCGALKSEMFKECHAEVP 1531
Dmel_Hml-PA 1559 LERFWKRCIFDTCACDQGGDCECLCTAVAAYADACAQKGINIRWRSQHFCPMQCDPHCSD 1618
DanaPB2_Hml-PA 1532 LDRYWKRCIFDTCACDQGGDCECLCTAVAAYADACAQKGIHIRWRSQHFCPMQCDPHCAE 1591
Dmel_Hml-PA 1619 YKACTPACAVETCDNFLDQGIAERMCNRENCLEGCHIKPCEDGFIYLNDTYRDCVPKAEC 1678
DanaPB2_Hml-PA 1592 YKACTPACAVETCDNFLDQGIADRMCTRENCLEGCHIKPCQDGFIYLNDTYRDCVPKAEC 1651
Dmel_Hml-PA 1679 KPVCMVRDGKTFYEGDITFTDSCATCRCSKRKEICSGVKCDVPATTGLPAPLVEGTTLPT 1738
DanaPB2_Hml-PA 1652 KPVCMVKEGKTYYEGDVTYTDACATCRCSKRKEICSGVKCEAPTTPKQWSPIIEGTTTSP 1711
Dmel_Hml-PA 1739 PLATQNQTKCVKGWTRWCDKDRDTSDKSVRLNDEEKVPRYDRMENVYGTCLKQYMTKVEC 1798
DanaPB2_Hml-PA 1712 PLATQNQTKCVKGWTRWCDKDRDSSDKSVRLNDEEKLPRYDRLENIYGTCLKEYMTKVEC 1771
Dmel_Hml-PA 1799 RVKDTHEAPEQMDENVVCSLEEGLRCIGKCHDYELRAFCQCDEELEPELPKPTEKPQLGL 1858
DanaPB2_Hml-PA 1772 RVKGTHEDPAQMDENVVCNLDVGLQCTGKCHDYELRAFCQCDDQPLPDIPKPTEKPQIGL 1831
Dmel_Hml-PA 1859 ACDAAVVEYKEFPGDCHKFLHCQPKGVEGGWIYVEKTCGEYMMFNPTMLICDHIATVTEI 1918
DanaPB2_Hml-PA 1832 ACDAAIVQYKEFPGDCHKFLQCQPNAVDGSWIYVEKTCGDFMMYNPSLFICDHIATVQQI 1891
Dmel_Hml-PA 1919 KPNCGLKPEPEPEFEPIKQCPPGKIKSECANQCENTCHYYGSILKKRGLCQVGEHCKPGC 1978
DanaPB2_Hml-PA 1892 NPRCGTKPTLEPEPEPIRQCPEGQIWADCANQCEHTCHYYGSILKRGLCHPGEHCKPGC 1951
Dmel_Hml-PA 1979 VDELRPDCPKLGKFWRDEDTCVHADECPCMDKAEHYVQPHKPVLGEFEVCQCIDNAFTCV 2038
DanaPB2_Hml-PA 1952 VELQRPDCRVQAKYWRNETTCVHVDECPCMDKNELYVQPHKPVVGETELCQCIDNSFTCV 2011
Dmel_Hml-PA 2039 PNKPEPVPKDEDDDLDLVSVVPIYPVTLTPPLQCSPERLIPKIENPAHSLPDSIFNASSQ 2098
DanaPB2_Hml-PA 2012 PNKPEPIPTR-------IPFGPILPVTLTPPLECSPERLIPKIENIQHPQPDGIFNASSQ 2064
Dmel_Hml-PA 2099 LAPEHGPKMARLTKEQPRGSWSPSINDQMQYLELNFAKPEPFYGVVMAGSPEFDNYVTLF 2158
DanaPB2_Hml-PA 2065 LAPEHGPHMARLTKEYPKGSWSPSINDQMQYLELNFGKPEPFYGVIMAGSPDFDNYVTLF 2124
Dmel_Hml-PA 2159 KILHSHDGIAYHYLVDETEKPQMFNGPLDSRAPVQTLFKIPIEASSLRIYPLKWHGSIAM 2218
DanaPB2_Hml-PA 2125 KILHSHDGISYHYLVDETEKPQMFSGPLDSRAPVQSLFKIPIEARSLRIYPLKWHGSIAM 2184
Dmel_Hml-PA 2219 RVELLICGDKEEPKPVPTVSTILPITERPARLVDLECIDLMGVDEGKMYQDQVQSSSLWQ 2278
DanaPB2_Hml-PA 2185 RVELLICGDKVVPEPIVFVPTT---TETPTELRNQECIDHMGVNEGKLYPEQVQSSSLWQ 2241
Dmel_Hml-PA 2279 QPNLGKKLQLLELLKLSTPLAWRPLANSQNEFIEFDFLEPRNISGFVTKGGPDGWVTGYK 2338
DanaPB2_Hml-PA 2242 WPELTKKPQILDLLKLSTPLAWRPLANTPNEFIQFDFLEPRNISGFVTKGGPDGWVTGYK 2301
Dmel_Hml-PA 2339 VMFSKKKPTWNTVLSTDGQARIFEANHDAETERRHHFKNPILTQYIKIVPAYWEKNINMR 2398
DanaPB2_Hml-PA 2302 VMFSKKKPTWNTVLSANGQPRIFEANVDAQTERTHHFKNPILTQYLKIVPAKWEKNINMR 2361
Dmel_Hml-PA 2399 IEPLGCFLPYPEIQRQVPVEESKPTKCNICDGVSTSSSTTGCQCQDQLFWDGNTCVQHNL 2458
DanaPB2_Hml-PA 2362 IEPLGCFQPYPEIQRQVPVDFSEPTKCNICEGISAPGLNEVCECVDQLFWDGNTCVQHNL 2421
Dmel_Hml-PA 2459 CPCIENYVSYPIGSKFENSACEDCVCVLGGHKNCKPKKCPPCLGGKLRPVITSDCFCKCE 2518
DanaPB2_Hml-PA 2422 CPCVENYVSYPIGSKFENSACEECVCVLGGHKNCKPKKCPPCQDAKLRSVITSDCFCKCE 2481
Dmel_Hml-PA 2519 PCPKHQRLCPSSGDCIPEILWCNGVQDCADDEDASCSDSFTVEPDVSREKNETEVITCPV 2578
DanaPB2_Hml-PA 2482 PCPKHQRLCPSSGDCIPEILWCNGIQDCADDEDASCRDSFTVEPEIKREKNQTEVITCPV 2541


Dmel_Hml-PA 2579 PVCPPQMKIRITEKKSRKMSKMFTFSKQVSIVDDGTTITKTKFISSKEQILAMPNRELDF 2638
DanaPB2_Hml-PA 2542 PECPPHMKIKITEKKQRKMSKMFTFSKQVSIVDDGVTITKTKFISSKEQILALPSQEVDF 2601
Dmel_Hml-PA 2639 QLEEQCDEFTCVPIPSKQVDKNETVTCTEPKCPEKYDVELDMSASKVGDCLRYSCVLRPN 2698
DanaPB2_Hml-PA 2602 HEEEQCEEFTCVPIPNTQLNKNETISCTEPKCPEKYDVELDMSAAKAGDCLKYTCVLRPN 2661
Dmel_Hml-PA 2699 KDDVCEISGKSFTTFDGTVFKYGPCSHILARDIHSSSWSISVHQQCSDETRKVCHKVITI 2758
DanaPB2_Hml-PA 2662 KDDVCEISGKSFTTFDGTVFKYRPCSHILARDIQNGSWSISVHQDCSDEHRRICRKVITI 2721
Dmel_Hml-PA 2759 QDTEAGNELILLPHLKLKFNGYEFTVQQLINSPICKASFVVSQPGKTLLAVSTKYGFWVQ 2818
DanaPB2_Hml-PA 2722 QDKEAGNELILLPYLKLKFNGFEFTVEQLINSPICKASFVVSQPGKTLLAVSTKYGFWVQ 2781
Dmel_Hml-PA 2819 LDDIGIVKVGISSKFIRTVDGLCGYYNGNQKDDKRSPDGQIIPNTEKFGDSWYDKRIPKD 2878
DanaPB2_Hml-PA 2782 LDDIGIVKVGISSKFIRTVDGLCGYYNGNPRDDKRSPEGQIVPNTEMFGDSWYDKRIPKD 2841
Dmel_Hml-PA 2879 QCGDLKCPREMQAKALQLCNIIHHPTFARCHKAVNYKQFLNNYCLEAACNCMMANNGDPA 2938
DanaPB2_Hml-PA 2842 QCGDLKCSRDMQAKALQLCNIIHHPTFARCHKAVNYKQFLNNYCLEAACDCMIANNGDTS 2901
Dmel_Hml-PA 2939 ACKCNILESFVKKCLSVNPLVQLTTWRAVAQCEINCPSPLVHTDCYKRRCEPSCDNVHGD 2998
DanaPB2_Hml-PA 2902 ACKCNILESFVKKCLSVNPLVQLTTWRTVAQCEISCPAPLVHTDCYKRRCEPSCDNIHGD 2961
Dmel_Hml-PA 2999 DCPVLPDACFPGCYCPEGTVRKGPNCVPISECKDCVCNSLGASKYMTYDRKSFSFNGNCT 3058
DanaPB2_Hml-PA 2962 DCPVLPDACFPGCYCPEGTVRKGPNCVPISDCKDCLCNSLGGSKYMTYDRKSFNFNGNCT 3021
Dmel_Hml-PA 3059 YLLSRDVVLPGVHTFQVYVSMDDCKKLGQPTPVEGGSCAKSLHILNGDHVIHVQRVPQKP 3118
DanaPB2_Hml-PA 3022 YLLSRDVVLPGVHTFQVYVSMDDCKKLGQITSLDGASCTKSLHILNGDHVIHVQRIPNKP 3081
Dmel_Hml-PA 3119 KSLQVLVDGFEVKKIPYKDSWISLRQVVGKELVLSLPESHVELTASFEDLIFSLGVPSIK 3178
DanaPB2_Hml-PA 3082 KSLQVLVDGFEVKKIPYKDSWIGIRQVIGKELVLSLLDSHVELSASFEELIFSLGVPSIK 3141
Dmel_Hml-PA 3179 YGSKMEGLCGDCNGNAGNDLQPNPAKKKAGVDVIQSWQADEPKLGLVEECLSEDVPKEHC 3238
DanaPB2_Hml-PA 3142 YGSKMEGICGDCNGNAGNDLQPNPAKMKPGVDVIQSWQADEPKLGIVEECLSEDVAKKDC 3201
Dmel_Hml-PA 3239 IPLPPEKDPCLQFYNAELFGKCPLAVDPIAYVSACQQDICKPGNTQQGVCVALAAYAKEC 3298
DanaPB2_Hml-PA 3202 IPLPPEKDPCLQFYNSELFGKCPLVVDPIAYVSACQQDICKPGNTQQGVCVALAAYAKEC 3261
Dmel_Hml-PA 3299 NQHGICTNWRRPQLCPYECPSDMVYEPCGCAKNCDTIKALSEFDAVSLKNEAVVHTVKTD 3358
DanaPB2_Hml-PA 3262 NQHGICTKWRRPQLCPYECPSNMVYEPCGCAKTCDTIKALSEFDAVSLKNQAFVHTVKSD 3321
Dmel_Hml-PA 3359 EMCLSSERFEGCFCPPGKVMDGGQCVPEIACTKCDDGLHLPDEKWKKDKCTECQCDSKGK 3418
DanaPB2_Hml-PA 3322 EMCLSSERFEGCFCPAGQVMDDGKCVPEIACTKCDDGLHLPGDTWQKSKCIDCLCDQTGK 3381
Dmel_Hml-PA 3419 TTCVEKKCQVEENICAEGYRPETIVSVDECCPRYRCVPETKDPSKLCLAPLVPICGPGQF 3478
DanaPB2_Hml-PA 3382 TSCVEKKCQVEENICAEGYKPETIVSKDSCCPKYRCVPELKDPAKVCLAPLMPICGPGQF 3441
Dmel_Hml-PA 3479 KKEKKDVNGCSQYICECIPKDQCE-IIELRELLPGEIIVNVEEGCCPTQKIECKPETCPK 3537
DanaPB2_Hml-PA 3442 KKQKNDVNGCPQYICECMPKDQCEPLMPSRELGPGEKIVTIEEGCCPTQKIICEPELCPK 3501
Dmel_Hml-PA 3538 APVNCQERFYEVKTIKEPGMCCSKHSCVPPKDLCIVQYELDEATKFTKTVGDKWTHAKEV 3597
DanaPB2_Hml-PA 3502 APNSCPERFYEVKTVKDLDMCCSKHSCVPPKDMCIVQYGLDESSKFTKQIGEKWTHAKDL 3561
Dmel_Hml-PA 3598 CKQETCSYGPDGNAQVVSTLEQCLTDCAPGFSYQNLDKTKCCGKCVQTSCIFEQKLYEVN 3657
DanaPB2_Hml-PA 3562 CIHETCSYGPDGSSQVVSSLEQCVTDCAPGFSYQKVNPSKCCGQCVQTSCIFEQKLYDVN 3621
Dmel_Hml-PA 3658 ALWKSADNCTTYSCLKKDGQFLVTTSREVCPDVGSCLSHLLYQDGCCKRCKSEPLVEDKS 3717
DanaPB2_Hml-PA 3622 AHWKSVDNCTSYICLKKDQQLMVSSSQEVCPDVGSCLSHLLYQDGCCKRCKVEPLVEDKS 3681
Dmel_Hml-PA 3718 SCLPVSLAESRTKEILKFPVQGHGTCVNADPIQGFTDCEGACSSGSKYNTLTDMHEKFCT 3777
DanaPB2_Hml-PA 3682 NCLPVSLAESQTTRIIKVPKIGHGECVNAEPIRGFTDCQGACSSGTKYNTLTDVHEKYCT 3741
Dmel_Hml-PA 3778 CCSIKSYHPISVKMICDDGHTFTQKHEVPSNCGCSPCSEFSDSAIDVRMQQDAQSPLLQL 3837
DanaPB2_Hml-PA 3742 CCNIKSYQPISVNMICEDGHTYIQKHELPANCGCSLCSELSDPAIDIRIQPDVQSPLLKL 3801
Dmel_Hml-PA 3838 LGNHRH- 3843
DanaPB2_Hml-PA 3802 LSSPKTL 3808

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

Paste a screenshot of the dot plot (generated by the Gene Model Checker) of your submitted model against the putative *D. melanogaster* ortholog into the box below.
Provide an explanation for any anomalies on the dot plot (*e.g.*, large gaps, regions with no sequence similarity, indications of significant insertions or deletions).

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 and provide a justification as to why you have selected this particular set of donor and acceptor sites.



Transcription Start Sites (TSS) Report Form (Optional)

Note: Complete this section if you have annotated the TSS for the gene above. This section is **optional** and you do not need to complete this section to submit the project.

Gene name (*e.g.*, *D. ananassae eyeless*): *D. ananassae Hemolectin*
Gene symbol (*e.g.*, *dana_ey*): *dana_Hml*

Name(s) of isoform(s) with unique TSS	List of isoforms with identical TSS
Hml-RA	Hml-RB

Names of the isoforms with unique TSS in *D. melanogaster* that are absent in this species:

_____

Provide the evidence (text and figures) which support the hypothesis that these isoforms are absent in this species (*e.g.*, changes in canonical splice sites, gene structure, etc.):

Isoform TSS Report

Complete an Isoform TSS report (through page 71) for each unique TSS listed in the table above. Copy and paste this form to create as many copies as needed.

Gene-isoform name (*e.g.*, *dana_ey-RA*): *dana_Hml-RA*

Names of the isoforms with the same TSS as this isoform: *dana_Hml-RB*

Type of core promoter in *D. melanogaster* (see table below):
(Peaked / **Intermediate** / Broad / Insufficient Evidence)

The type of core promoter is defined by the number of TSS annotated by the Celniker group at modENCODE and the number of DHS positions:

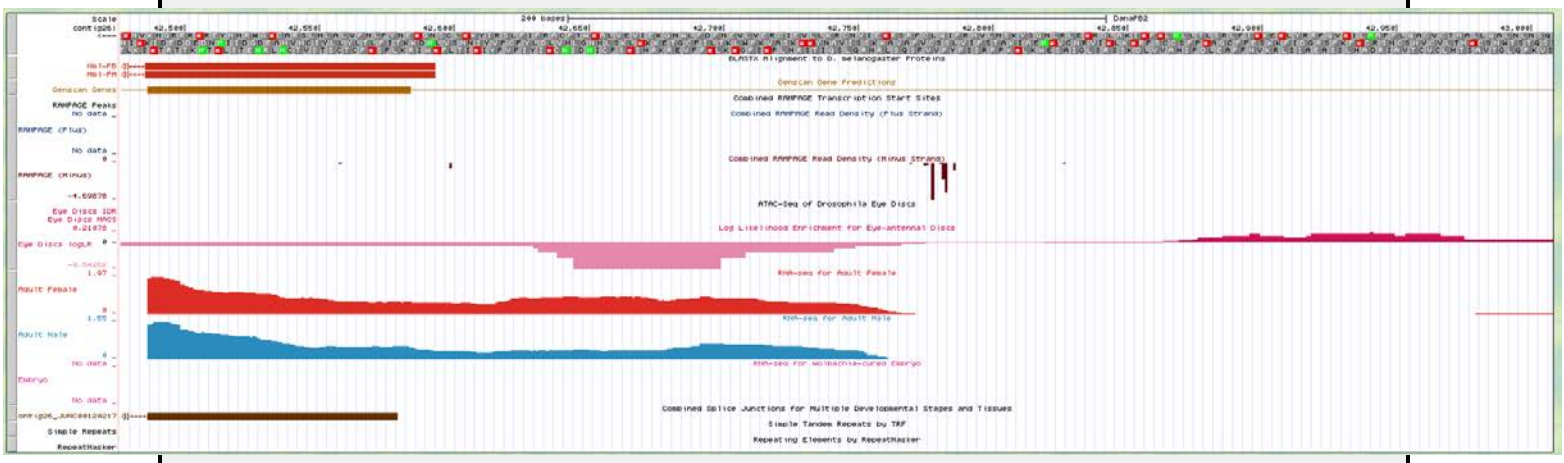
Type of core promoter	# annotated TSS	# DHS positions
Peaked	1	0
	0	1
	1	1
Intermediate	≤ 1	> 1
	> 1	≤ 1
Broad	> 1	> 1
Insufficient Evidence	0	0

1. Turn on RAMPAGE evidence tracks

Coordinates of the TSS position based on position with the highest RAMPAGE read density
42,777-42,778

Coordinates of the narrow TSS search region based on RAMPAGE peaks
42,477-43,008

If the TSS position and narrow TSS search region are supported by RAMPAGE data, **paste a Genome Browser screenshot of the region surrounding the putative TSS (±300bp) showing the Combined RAMPAGE TSS evidence track:**



2. Turn on ATAC-Seq evidence track

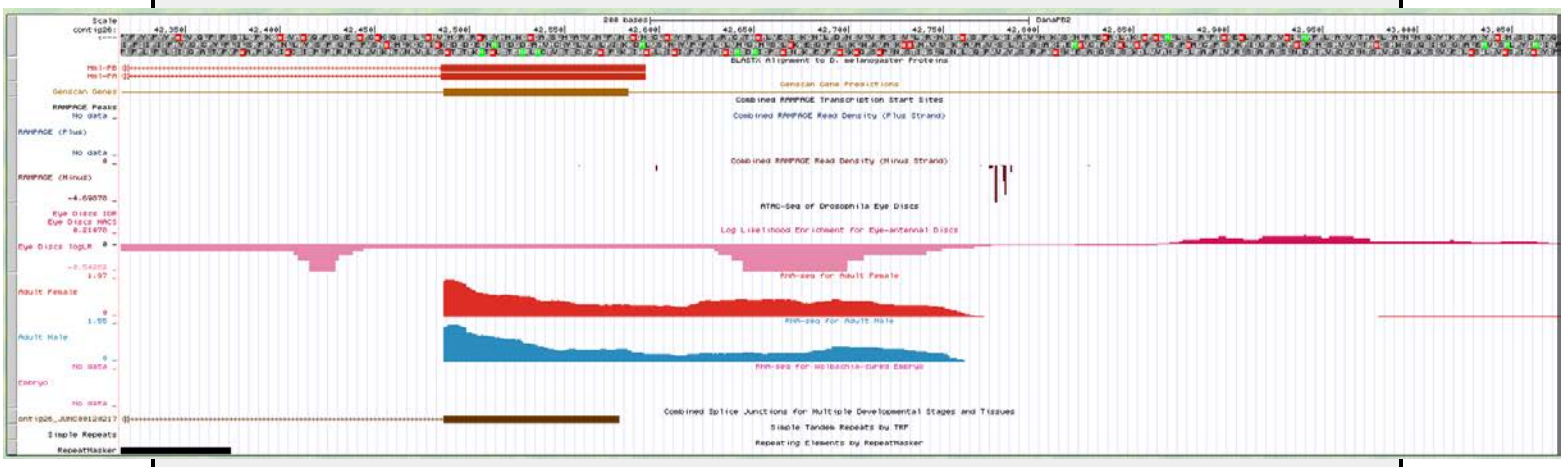
If the wide TSS search region is supported by ATAC-Seq data, **paste a Genome Browser screenshot of the region surrounding the putative TSS (±300bp) showing the Eye Discs ATAC-Seq evidence track:**

Coordinates of the wide TSS search region based on ATAC-Seq peaks

3. Turn on RNA-Seq evidence tracks

If the TSS annotation is supported by RNA-Seq read coverage or splice junction predictions (*e.g.*, regtools), **paste a Genome Browser screenshot of the region surrounding the putative TSS (±300bp) showing the following evidence tracks:**

1. RNA-Seq Coverage or RNA-Seq Alignment Summary
2. Combined Splice Junctions or RNA-Seq TopHat



If the RNA-Seq evidence tracks indicate the TSS position, list it here: 42,317-43,074

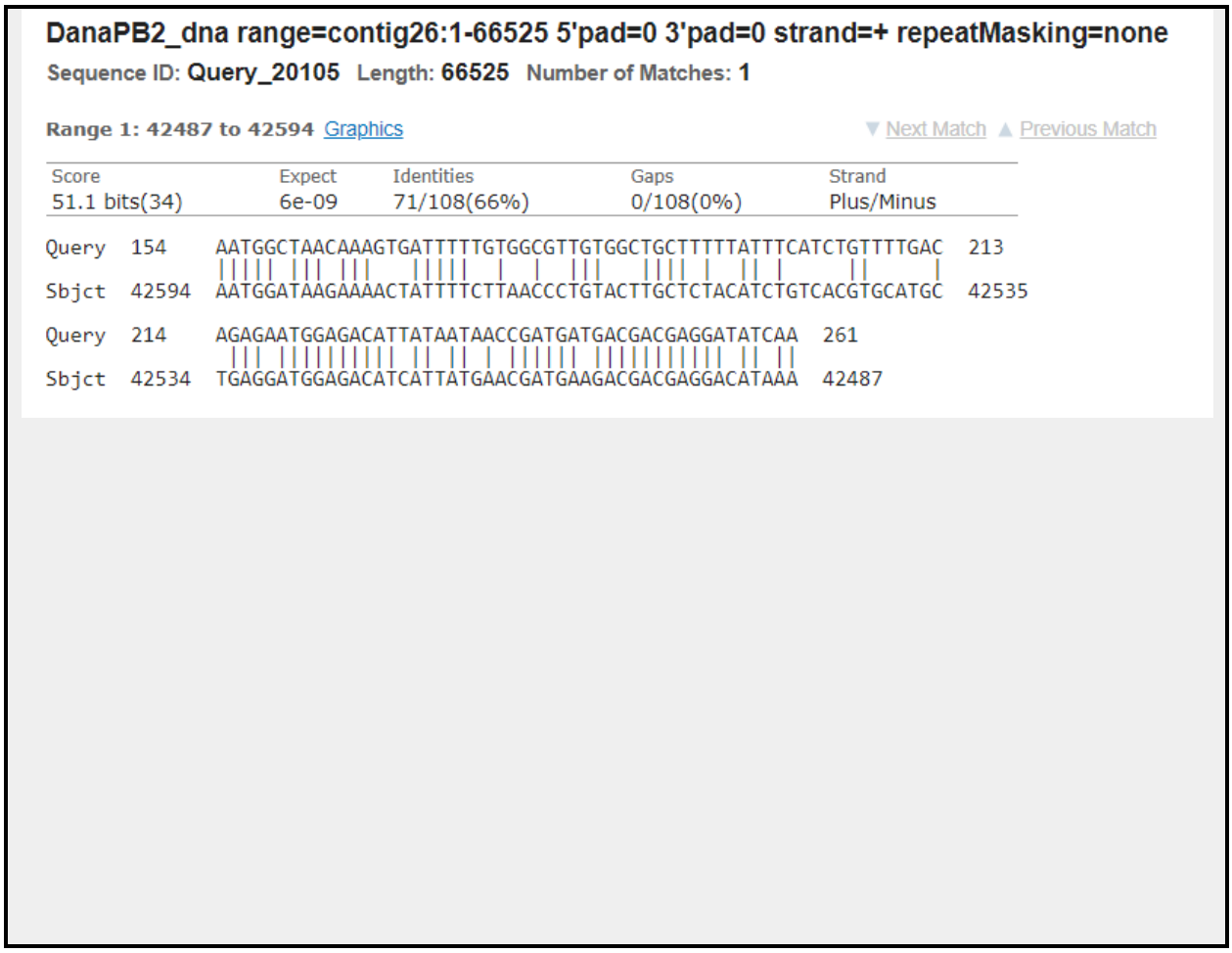
4. Annotate the first transcribed exon

Coordinates of the first transcribed exon based on blastn alignment:
42,441-42,487

Does the blastn alignment cover the entire *D. melanogaster* first transcribed exon?
No

If not, specify the parts of the *D. melanogaster* exon that are missing from the blastn alignment.
1-153

If the TSS annotation is supported by blastn alignment of the initial transcribed exon against the contig sequence, **paste a screenshot of the blastn alignment into the box below:**



5. Turn on comparative genomics tracks

If the TSS annotation is supported by sequence conservation with other *Drosophila* species, **paste a screenshot of the multiple sequence alignment (*e.g.*, from Clustal Omega, ROAST) into the box below:**

6. Summarize the evidence that supports the TSS annotation postulated above

Coordinate(s) of the TSS position(s):

Based on RAMPAGE data: 42,477-43,008
Based on ATAC-Seq data: _____
Based on RNA-Seq data: 42,317-43,074
Based on blastn alignment: 42,441-42,487
Based on other evidence (please specify): _____

Note: If the blastn alignment for the initial transcribed exon is a partial alignment, you can **extrapolate the TSS position** based on the number of nucleotides that are missing from the beginning of the exon. (Enter "Insufficient evidence" if you cannot determine the TSS position based on the available evidence.)

Were you able to define a TSS position based on the available evidence? Yes
If so, indicate in the table below the evidence that supports this TSS position

If not, were you able to define a TSS search region? _____
If so, indicate in the table below the evidence that supports the TSS search region(s)

Evidence type	Support	Refute	Neither
RAMPAGE peaks and read density	☒	☐	☐
ATAC-Seq peaks and log likelihood enrichment profile	☐	☐	☒
RNA-Seq coverage and splice junctions	☒	☐	☐
blastn alignment of the initial exon from *D. melanogaster*	☒	☐	☐
Sequence conservation with other *Drosophila* species (*e.g.*, "Conservation" track on the Genome Browser)	☐	☐	☒
Other (please specify) [*e.g.*, Gnomon, N-SCAN, Augustus TSS predictions; histone modifications (ChIP-Seq data)].	☐	☐	☒

Note: The evidence type refutes the TSS annotation only if it **suggests an alternate TSS position**. For example, the presence of RNA-Seq read coverage upstream of the annotated TSS indicates that the TSS is located further upstream and it would be considered to be evidence against the annotated TSS; check "Refute." In contrast, the lack of RNA-Seq read coverage is a negative result that neither supports nor refutes the TSS annotation; check "Neither."

Provide an explanation if the TSS annotation is inconsistent with at least one of the evidence types specified above:

Check for additional features in your project

For each Genscan gene prediction that does not overlap with the genes you have already annotated, perform the following analyses to determine if the feature corresponds to a protein-coding gene, pseudogene, or partial gene duplication.

1. Perform a FlyBase BLASTP search of the predicted protein sequence from Genscan against the *D. melanogaster* "**Annotated proteins**" database. Report the significant matches (E-value < 1e-5) to protein sequences in *D. melanogaster*:

2. If there are significant matches to *D. melanogaster* proteins, analyze the genomic region immediately surrounding the Genscan prediction using the exon-by-exon strategy. Report your findings:
    - If the feature is a functional protein-coding gene, construct the gene model in the target species and provide the supporting evidence for the gene model in a new Gene Report Form

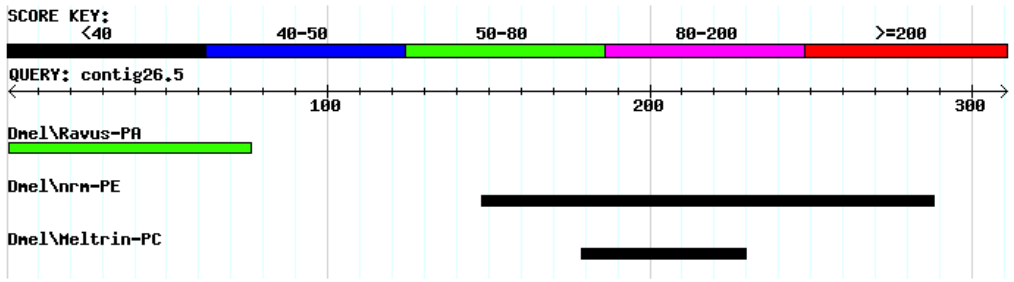
    - If the feature is a pseudogene or a partial gene duplication, provide the evidence (text and figures) which support these hypotheses:
        - Evidence for a pseudogene includes in-frame stop codons, and frame shifts within coding exons
        - Changes in gene structure from a multi-exon gene in *D. melanogaster* to a single exon gene in the target species could indicate a retrotransposed pseudogene

3. Perform a NCBI BLASTP search of the predicted protein sequence from Genscan against the "**Reference proteins (refseq_protein)**" database. Report the significant matches (E-value < 1e-5) to curated RefSeq gene models:
    - Protein records curated by the NCBI RefSeq database have the prefix "**NP_**"

4. Examine the gene expression tracks (*e.g.*, RNA-Seq data) for evidence of transcribed regions that do not correspond to the features you have already annotated or transposon remnants identified by RepeatMasker. Perform an NCBI BLASTX search of these genomic regions against the **refseq_protein** database to determine if they show significant similarity (E-value < 1e-5) to curated RefSeq gene models (i.e. protein records with the prefix "**NP_**"). Report as above:

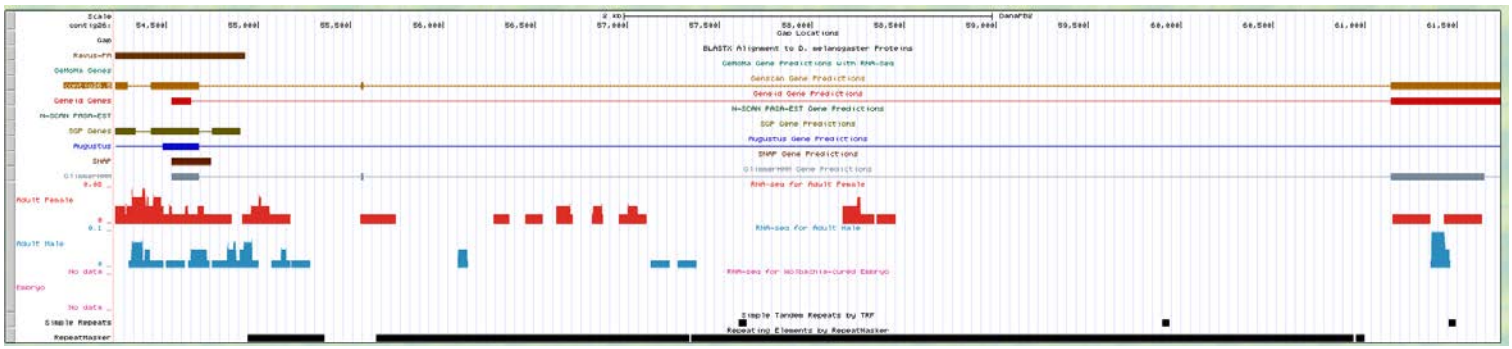
Contig 26.5:
FlyBase BlastP shows one significant match to a protein sequence in *D. melanogaster*, *Ravus-PA*. However, *Ravus-PA* is on the 3R chromosome instead of the 3L chromosome, and also the first exon is not present in the contig. Therefore, it is most likely a pseudogene.

Query	
Description	Length
contig26.5 length=310	310



Send BLAST Hits to HitList

BLAST Hit Summary				
☑	Description	Species	Score	E value
☑	Ravus-PA	Dmel	71.633	1.08942e-12
☑	nrm-PE	Dmel	32.7278	0.542398
☑	Meltrin-PC	Dmel	28.4906	8.44301



ravus [Drosophila melanogaster]

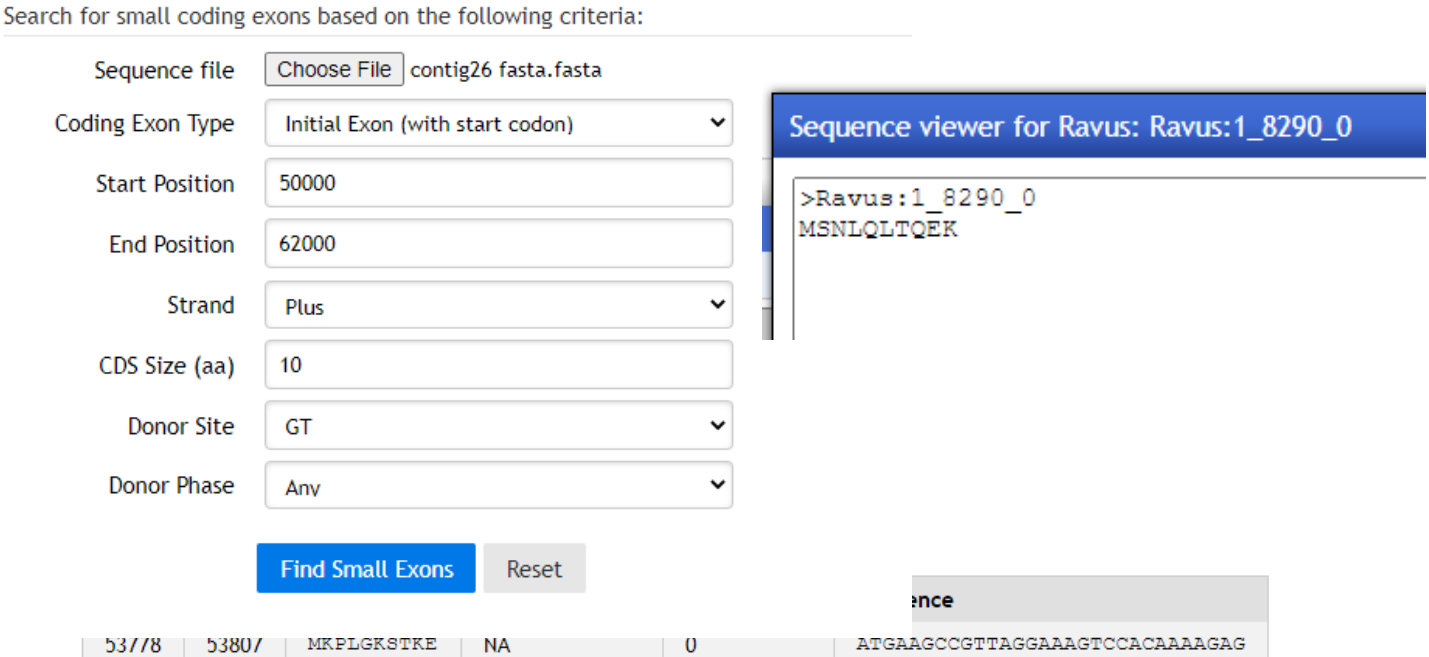
Sequence ID: NP_731779.1 Length: 359 Number of Matches: 1

Range 1: 1 to 122 GenPept Graphics ▼ Next Match ▲ Previous Match

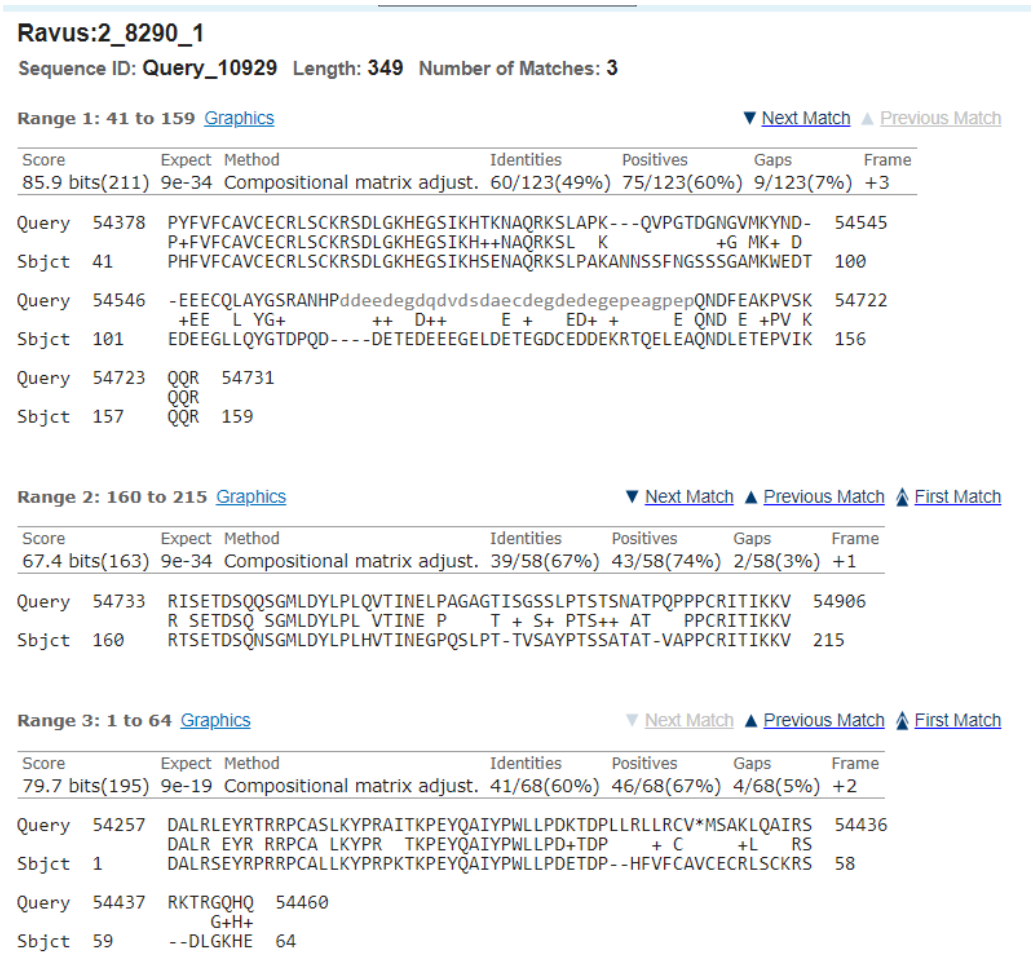
Score	Expect	Method	Identities	Positives	Gaps
71.6 bits(174)	3e-10	Compositional matrix adjust.	51/122(42%)	59/122(48%)	47/122(38%)

```
Query  1    MANVQLTQEKLDALRLEYRTRRPC-----------------------------------  24
            M+N+QLTQEKLDALR EYR RRPC
Sbjct  1    MSNLQLTQEKLDALRSEYRPRRPCALLKYPRPKTKPEYQAIYPWLLPDETDPHFVFCAVC  60

Query  25   ------KRSDLGKHEGSIKHTKNAQRKSL---APKQVPGTDGNGVMKYND--EEECQLAY  73
                 KRSDLGKHEGSIKH++NAQRKSL    A          +G MK+ D   +EE  L Y
Sbjct  61   ECRLSCKRSDLGKHEGSIKHSENAQRKSLPAKANNSSFNGSSSGAMKWEDTEDEEGLLQY  120

Query  74   GS  75
            G+
Sbjct  121  GT  122
```

As you can see from the screenshot above, *Ravus-PA* is a significant match both in FlyBase Blast and BlastP. However, the first exon was attempted to be found using small exon finder. However, despite using a very large search area, no matches or close matches to the first exon could be found. Through BlastP, the second exon could be found, but the match was very poor. There is not a significant amount of RNA-Seq data. The BlastX of the data shows 57 significant matches, most of them to various hypothetical proteins, and different transposases in various species. Given these facts plus the fact that *Ravus* should not be on this chromosome, I believe that this is most likely a pseudogene.



53778	53807	MKPLGKSTKE	NA	0	ATGAAGCCGTTAGGAAAGTCCACAAAAGAG
54134	54163	MGCKPEPKAA	NA	0	ATGGGCTGCAAACCAGAGCCAAAGGCAGCC
54472	54503	MHSESPWLPN	NA	2	ATGCACAGCGAAAGTCCTTGGCTCCCAAACAA
55539	55568	MFFLFSIGRK	NA	0	ATGTTCTTCTTATTTAGTATTGGAAGAAAG
56610	56639	MAEVMLRFGP	NA	0	ATGGCGGAGGTTATGTTAAGGTTTGGGCCT
57596	57625	MSFNLEPFEA	NA	0	ATGAGCTTTAATCTGGAGCCTTTTGAGGCT
58415	58446	MWTILFPTQQ	NA	2	ATGTGGACGATCTTATTCCCGACCCAACAACA
58648	58677	MDFQKTTSQN	NA	0	ATGGATTTCCAAAAAACCACAAGCCAAAAT
58648	58679	MDFQKTTSQN	NA	2	ATGGATTTCCAAAAAACCACAAGCCAAAATGT
59473	59502	MLSILSPVPQ	NA	0	ATGTTGTCGATCCTCAGCCCAGTTCCTCAA
59868	59898	MIFLCYYFII	NA	1	ATGATTTTTTTGTGTTATTATTTCATTATTT
61296	61326	MGGAMCHVGI	NA	1	ATGGGAGGGGCCATGTGCCATGTTGGCATAA



Ravus:2_8290_1

Sequence ID: Query_10929 Length: 349 Number of Matches: 3

Range 1: 41 to 159 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
85.9 bits(211)	9e-34	Compositional matrix adjust.	60/123(49%)	75/123(60%)	9/123(7%)	+3

```
Query  54378  PYFVFCAVCECRLSCKRSDLGKHEGSIKHTKNAQRKSLAPK---QVPGTDGNGVMKYND-  54545
              P+FVFCAVCECRLSCKRSDLGKHEGSIKH++NAQRKSL  K          +G MK+ D
Sbjct  41     PHFVFCAVCECRLSCKRSDLGKHEGSIKHSENAQRKSLPAKANNSSFNGSSSGAMKWEDT  100

Query  54546  -EEECQLAYGSRANHPddeedegdqdvdsdaecdegdedegepeagpepQNDFEAKPVSK  54722
               +EE  L YG+        ++  D++     E +   ED+ +     E QND E +PV K
Sbjct  101    EDEEGLLQYGTDPQD----DETEDEEEGELDETEGDCEDDEKRTQELEAQNDLETEPVIK  156

Query  54723  QQR  54731
              QQR
Sbjct  157    QQR  159
```

Range 2: 160 to 215 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
67.4 bits(163)	9e-34	Compositional matrix adjust.	39/58(67%)	43/58(74%)	2/58(3%)	+1

```
Query  54733  RISETDSQQSGMLDYLPLQVTINELPAGAGTISGSSLPTSTSNATPQPPPCRITIKKV  54906
              R SETDSQ SGMLDYLPL VTINE P    T + S+ PTS++ AT   PPCRITIKKV
Sbjct  160    RTSETDSQNSGMLDYLPLHVTINEGPQSLPT-TVSAYPTSSATAT-VAPPCRITIKKV  215
```

Range 3: 1 to 64 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
79.7 bits(195)	9e-19	Compositional matrix adjust.	41/68(60%)	46/68(67%)	4/68(5%)	+2

```
Query  54257  DALRLEYRTRRPCASLKYPRAITKPEYQAIYPWLLPDKTDPLLRLLRCV*MSAKLQAIRS  54436
              DALR EYR RRPCA LKYPR  TKPEYQAIYPWLLPD+TDP    + C     +L   RS
Sbjct  1      DALRSEYRPRRPCALLKYPRPKTKPEYQAIYPWLLPDETDP--HFVFCAVCECRLSCKRS  58

Query  54437  RKTRGQHQ  54460
                G+H+
Sbjct  59     --DLGKHE  64
```

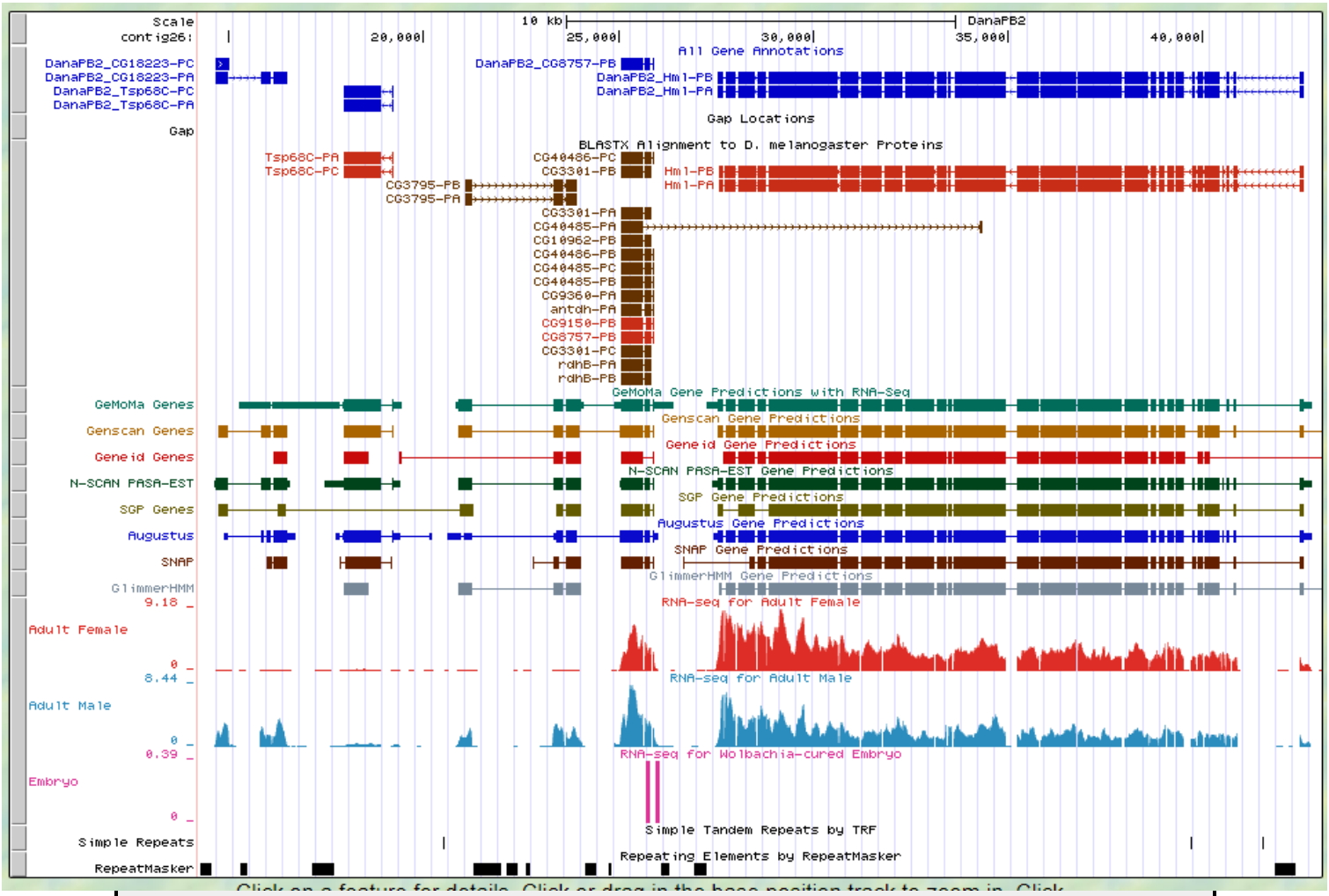
Contig 26.6:

There are no significant matches to protein sequences in *D. melanogaster*. There are no matches to any reference proteins in NCBI Blastp. RNA-Seq data does not provide evidence for there being transcription in this region and there are no matches to any reference proteins using NCBI Blastx.

Preparing the Project for Submission

For each project, you should prepare the project GFF, transcript, and peptide sequence files for **ALL** isoforms along with this report. You can combine the individual files generated by the Gene Model Checker into a single file using the Annotation Files Merger. Once you have combined the GFF files into a single file, click on the "**Show Track**" button to view all the gene models in the combined GFF file within the Genome Browser.

Paste a screenshot (generated by the Annotation Files Merger) with all the gene models you have annotated in this project into the box below.



For projects with multiple errors in the consensus sequence, you should combine all the VCF files into a single project VCF file using the Annotation Files Merger. **Paste a screenshot of the Genome Browser (generated by the Annotation Files Merger) showing the locations of all the consensus errors with respect to the original project sequence into the box below.**

Thank you for your submission, and congratulations on completing your analysis of this region of this genome. Our planned GEP meta-analysis of the genes and genomes in this study depends on the high quality annotations accomplished by GEP students.