F Element Project: Annotation Report

Faculty instructor(s): Sarah C.R. Elgin
College/university: Washington University in St. Louis
Course number: Bio 4342
Course name: Research Explorations in Genomics

Project Details

Project name: contig10
Project species: *D. biarmipes*
Date of submission: 12/26/2023
Size of project in base pairs: 43,013
Number of genes in project: 3

Does this report cover all of the genes or is it a partial report? Partial report
If this is a partial report, please indicate the region of the project covered by this report:
From base 25,000 to base 28,000

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. 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 below).

Gene Report Form

Gene name (e.g., *D. ananassae eyeless*): *D. biarmipes CG31997*
Gene symbol (e.g., *dana_ey*): *dbia_CG31997*

Approximate location in project (from 5' end to 3' end): 25673-27471
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**
CG31997-PB	CG31997-PA

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

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.):

NA

Note: In addition to submitting your annotation report, you will also submit gene model files which describe your isoform(s) as a DNA sequence (FASTA), a peptide sequence (PEP), and as a collection of exon coordinates that can be visualized on the GEP UCSC Genome Browser (GFF). While we only require one Isoform Report form per unique coding sequence, **we also require a full set of gene model files (GFF, FASTA, and PEP) for *ALL* isoforms, even if their coding sequence is identical** to that of another isoform. See page 31 of the Gene Model Checker User Guide for details.

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:
NA

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 Sequence 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): dbia_CG31997-PB

Names of any additional isoforms with identical coding sequences:
dbia_CG31997-PA

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)

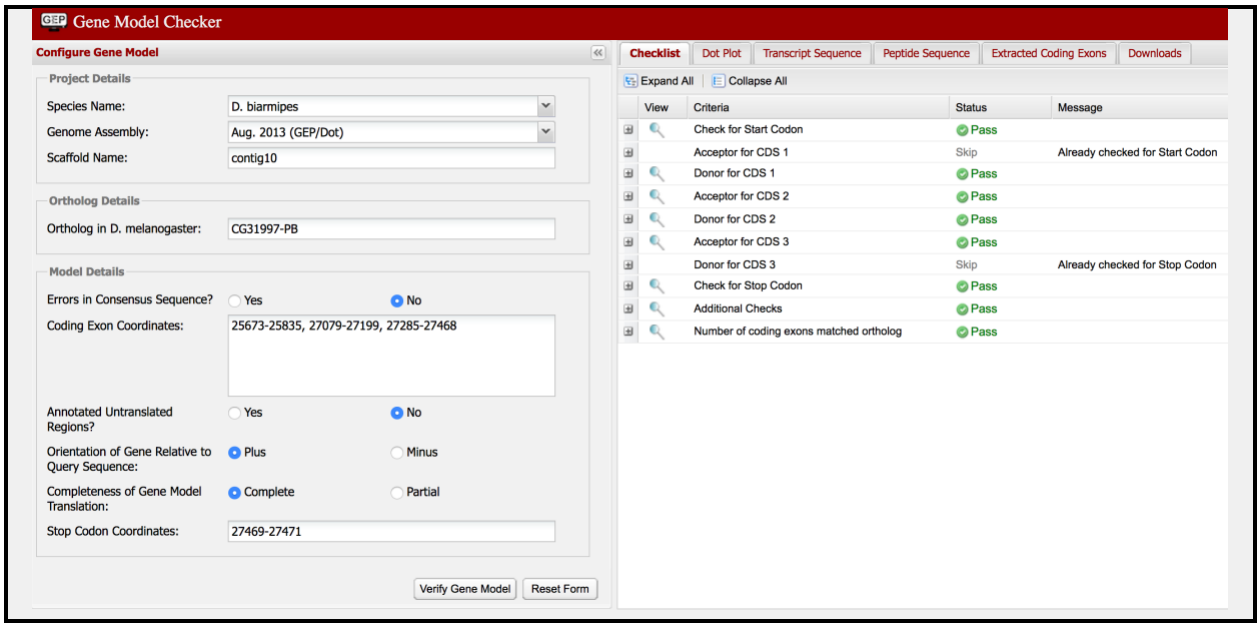
1. Gene Model Checker checklist

Coordinates of your final gene model for this isoform:
25673-25835, 27079-27199, 27285-27468

Stop codon coordinates: 27469-27471

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: This screenshot should show the "**Configure Gene Model**" panel with the exon coordinates and the "**Checklist**" panel with all the checklist items (i.e., from the criteria "Check for Start Codon" to "Number of coding exons matched ortholog"). If necessary, include multiple screenshots of the "Checklist" panel to capture all the checklist items.

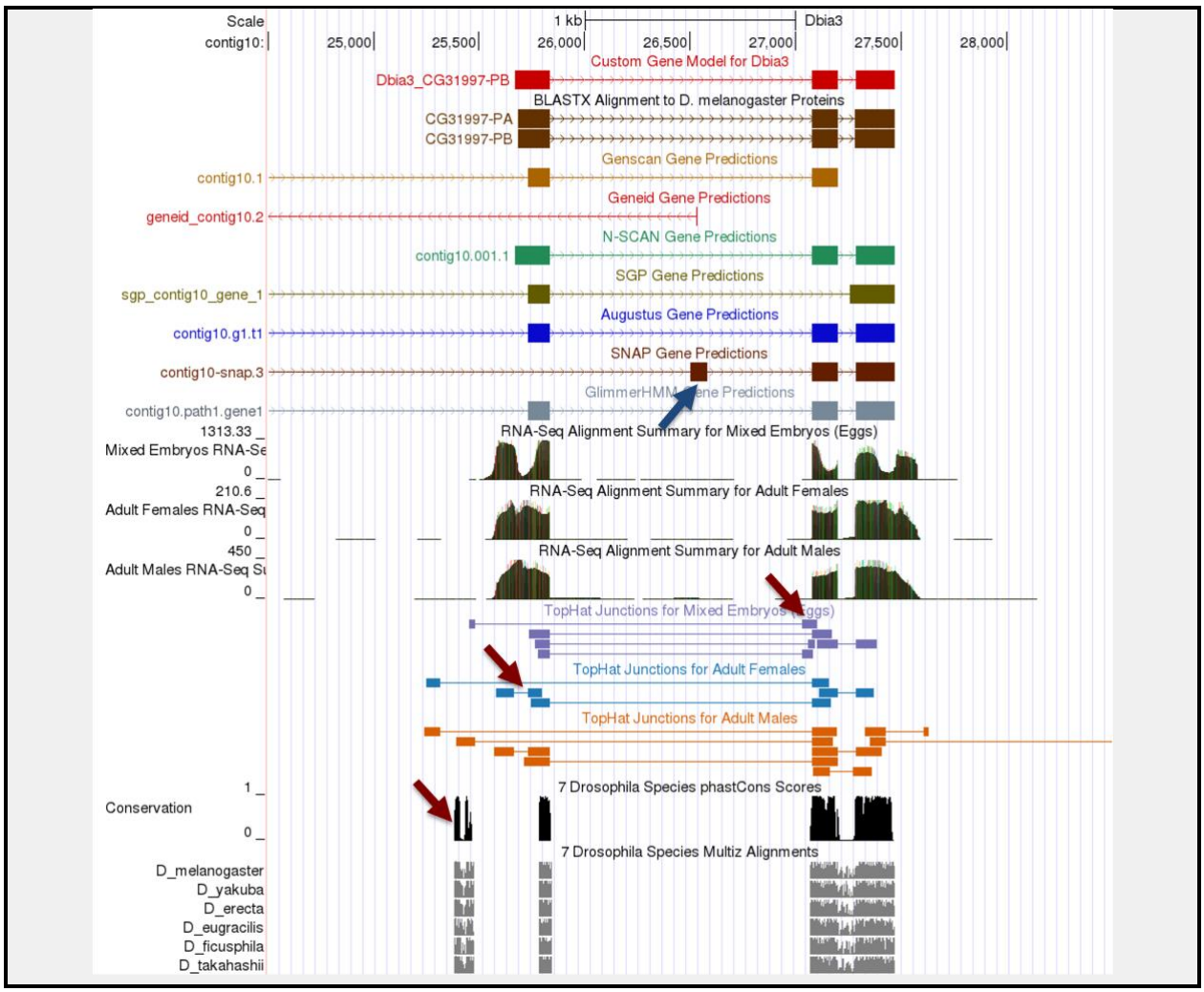


2. View the gene model on the Genome Browser

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:**

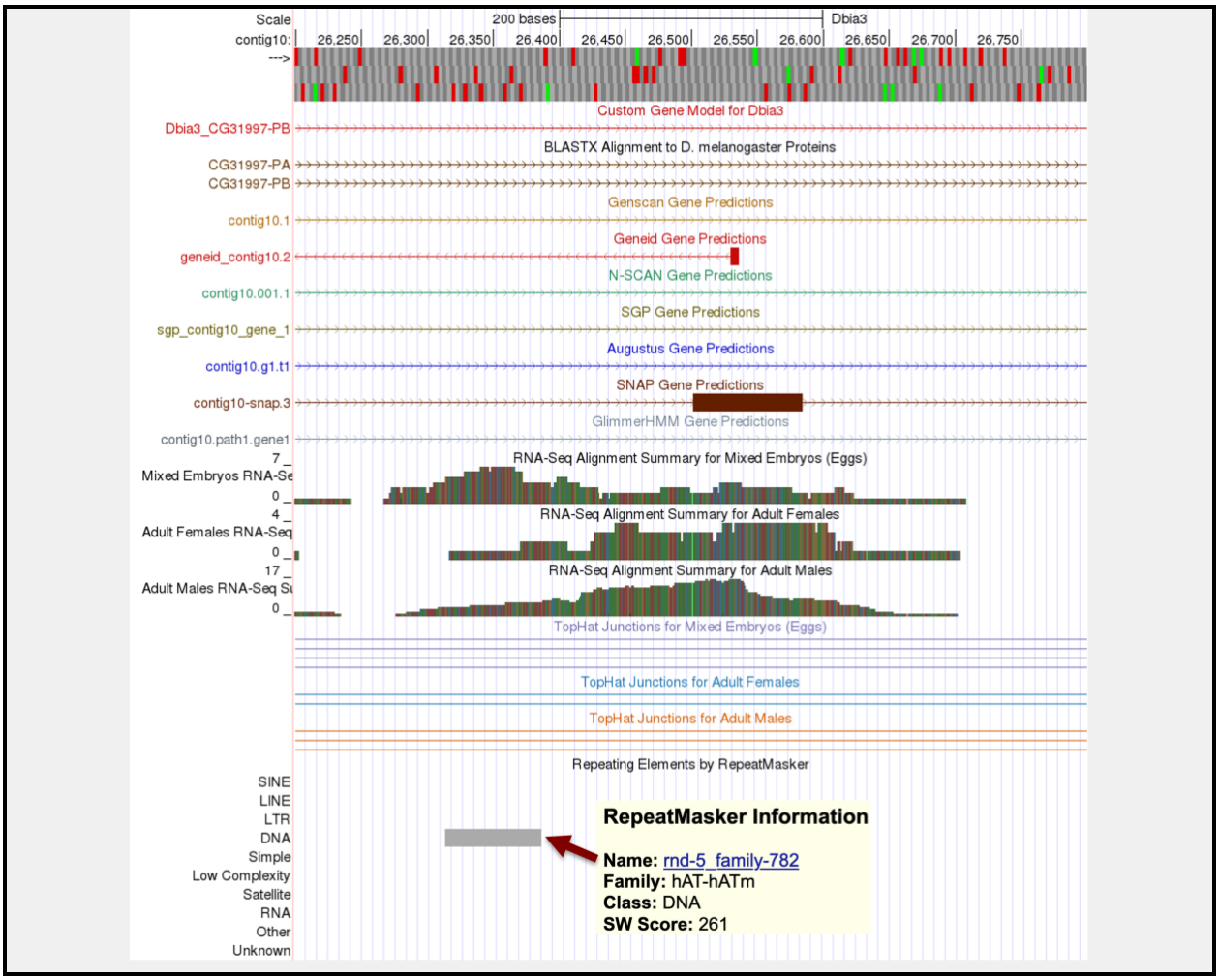
1. A sequence alignment track (e.g., D. mel Proteins)
2. At least one gene prediction track (e.g., Genscan)
3. At least one RNA-Seq track (e.g., RNA-Seq Coverage)
4. A comparative genomics track (e.g., D. mel. Net Alignment, Conservation)

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

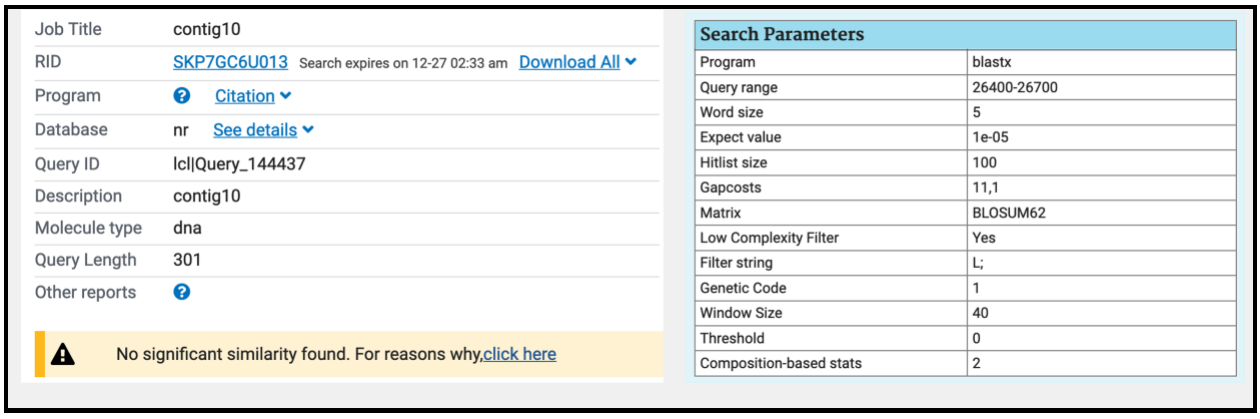


Low-frequency RNA-Seq exon junctions not annotated:
The evidence from the RNA-Seq TopHat evidence tracks and Multiz alignments suggest that there might be additional isoforms because of alternative splicing at the 5' end of this gene (red arrows in the screenshot above). However, because most of the TopHat junctions are supported by less than 10 reads, there is insufficient evidence to postulate the presence of multiple novel isoforms in *D. biarmipes* compared to *D. melanogaster*.

Extra CDS predicted by the SNAP gene predictor:
SNAP predicted a CDS at 26,502-26,584 (blue arrow in the screenshot above) between the first and second CDS's of *CG31997*. The RNA-Seq Alignment Summary track shows that the region surrounding this region has low (<20 reads) RNA-Seq read coverage and the region is adjacent to a hAT DNA transposon fragment (see screenshot below).



NCBI *blastx* search of the genomic region surrounding the SNAP CDS prediction (contig10:26400-26700) against the nr database did not detect any significant (E-value < 1e-5) sequence similarity to known proteins in the nr database (see screenshot below).

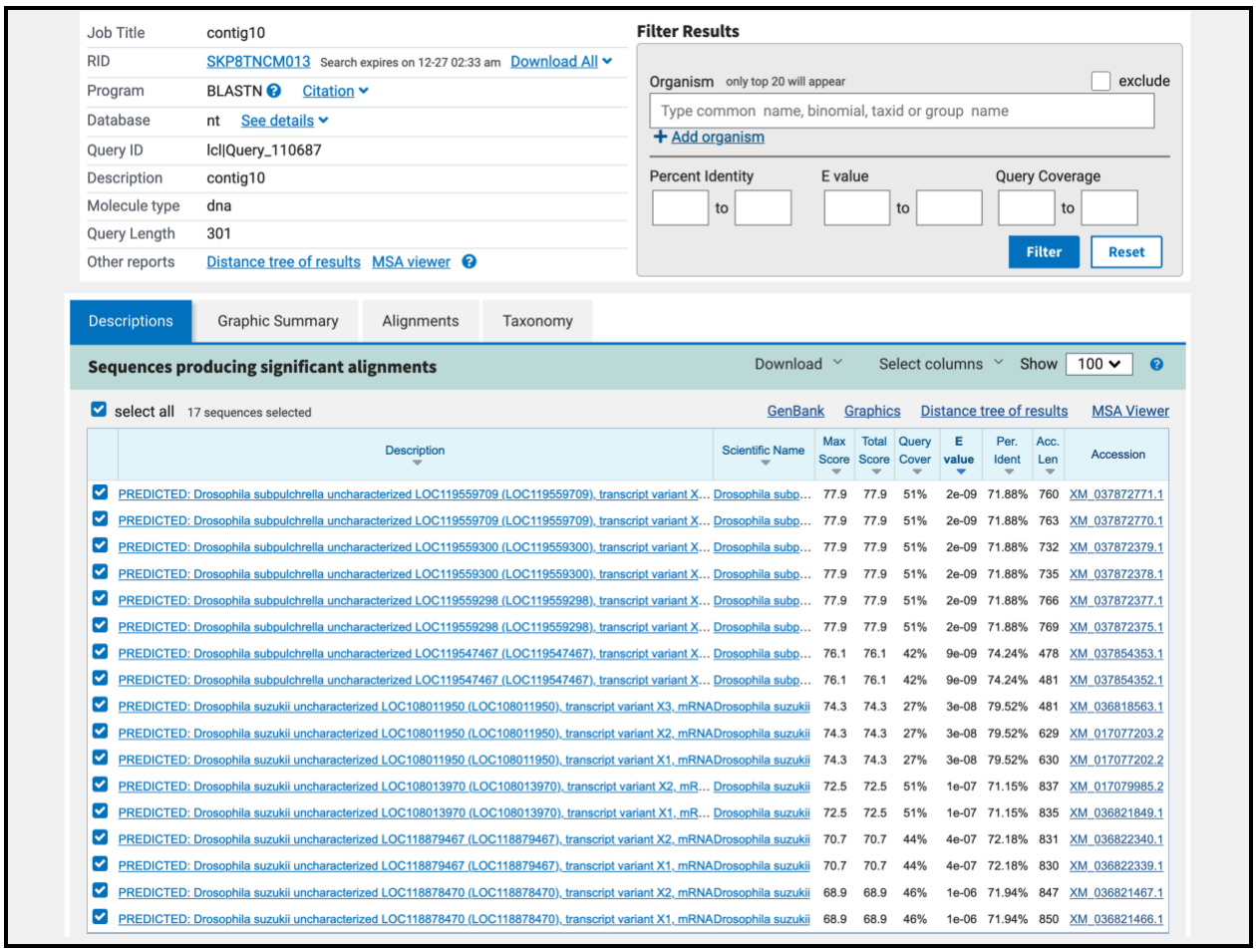


Job Title	contig10
RID	SKP7GC6U013 Search expires on 12-27 02:33 am Download All
Program	Citation
Database	nr See details
Query ID	lcl|Query_144437
Description	contig10
Molecule type	dna
Query Length	301
Other reports	

No significant similarity found. For reasons why,click here

Search Parameters	
Program	blastx
Query range	26400-26700
Word size	5
Expect value	1e-05
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Low Complexity Filter	Yes
Filter string	L;
Genetic Code	1
Window Size	40
Threshold	0
Composition-based stats	2

A NCBI *blastn* search of this region against the nt database detected 17 significant matches to predicted mRNAs in *Drosophila subpulchrella* and *Drosophila suzukii* (see screenshot below). Both *Drosophila* species are members of the suzukii subgroup.



Job Title	contig10
RID	SKP8TNCM013 Search expires on 12-27 02:33 am Download All
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_110687
Description	contig10
Molecule type	dna
Query Length	301
Other reports	Distance tree of results MSA viewer

Filter Results

Organism only top 20 will appear exclude

Type common name, binomial, taxid or group name

Add organism

Percent Identity to

E value to

Query Coverage to

Filter Reset

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments

Download Select columns Show 100

select all 17 sequences selected

GenBank Graphics Distance tree of results MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
PREDICTED: Drosophila subpulchrella uncharacterized LOC119559709 (LOC119559709), transcript variant X...	Drosophila subp...	77.9	77.9	51%	2e-09	71.88%	760	XM_037872771.1
PREDICTED: Drosophila subpulchrella uncharacterized LOC119559709 (LOC119559709), transcript variant X...	Drosophila subp...	77.9	77.9	51%	2e-09	71.88%	763	XM_037872770.1
PREDICTED: Drosophila subpulchrella uncharacterized LOC119559300 (LOC119559300), transcript variant X...	Drosophila subp...	77.9	77.9	51%	2e-09	71.88%	732	XM_037872379.1
PREDICTED: Drosophila subpulchrella uncharacterized LOC119559300 (LOC119559300), transcript variant X...	Drosophila subp...	77.9	77.9	51%	2e-09	71.88%	735	XM_037872378.1
PREDICTED: Drosophila subpulchrella uncharacterized LOC119559298 (LOC119559298), transcript variant X...	Drosophila subp...	77.9	77.9	51%	2e-09	71.88%	766	XM_037872377.1
PREDICTED: Drosophila subpulchrella uncharacterized LOC119559298 (LOC119559298), transcript variant X...	Drosophila subp...	77.9	77.9	51%	2e-09	71.88%	769	XM_037872375.1
PREDICTED: Drosophila subpulchrella uncharacterized LOC119547467 (LOC119547467), transcript variant X...	Drosophila subp...	76.1	76.1	42%	9e-09	74.24%	478	XM_037854353.1
PREDICTED: Drosophila subpulchrella uncharacterized LOC119547467 (LOC119547467), transcript variant X...	Drosophila subp...	76.1	76.1	42%	9e-09	74.24%	481	XM_037854352.1
PREDICTED: Drosophila suzukii uncharacterized LOC108011950 (LOC108011950), transcript variant X3, mRNA	Drosophila suzukii	74.3	74.3	27%	3e-08	79.52%	481	XM_036818563.1
PREDICTED: Drosophila suzukii uncharacterized LOC108011950 (LOC108011950), transcript variant X2, mRNA	Drosophila suzukii	74.3	74.3	27%	3e-08	79.52%	629	XM_017077203.2
PREDICTED: Drosophila suzukii uncharacterized LOC108011950 (LOC108011950), transcript variant X1, mRNA	Drosophila suzukii	74.3	74.3	27%	3e-08	79.52%	630	XM_017077202.2
PREDICTED: Drosophila suzukii uncharacterized LOC108013970 (LOC108013970), transcript variant X2, mR...	Drosophila suzukii	72.5	72.5	51%	1e-07	71.15%	837	XM_017079985.2
PREDICTED: Drosophila suzukii uncharacterized LOC108013970 (LOC108013970), transcript variant X1, mR...	Drosophila suzukii	72.5	72.5	51%	1e-07	71.15%	835	XM_036821849.1
PREDICTED: Drosophila suzukii uncharacterized LOC118879467 (LOC118879467), transcript variant X2, mRNA	Drosophila suzukii	70.7	70.7	44%	4e-07	72.18%	831	XM_036822340.1
PREDICTED: Drosophila suzukii uncharacterized LOC118879467 (LOC118879467), transcript variant X1, mRNA	Drosophila suzukii	70.7	70.7	44%	4e-07	72.18%	830	XM_036822339.1
PREDICTED: Drosophila suzukii uncharacterized LOC118878470 (LOC118878470), transcript variant X2, mRNA	Drosophila suzukii	68.9	68.9	46%	1e-06	71.94%	847	XM_036821467.1
PREDICTED: Drosophila suzukii uncharacterized LOC118878470 (LOC118878470), transcript variant X1, mRNA	Drosophila suzukii	68.9	68.9	46%	1e-06	71.94%	850	XM_036821466.1

The E-values for the *D. subpulchrella* matches range from 2e-09 to 9e-09, and they correspond to four different predicted genes (LOC119559709, LOC119559300, LOC119559298, and LOC119547467). The E-values for the *D. suzukii* matches range from 3e-08 to 1e-06, and they correspond to four different predicted genes (LOC108011950, LOC108013970, LOC118879467, and LOC118878470). All of these matches are RefSeq predictions that have not been confirmed experimentally. There are no significant matches to RefSeq records that are supported by experimental evidence and no significant matches to mRNAs in other species outside of the suzukii subgroup.

Collectively, while we could not reject the possibility that this region of contig10 contains an untranslated region of a nearby gene, there is insufficient evidence to postulate a novel isoform of *CG31997* in *D. biarmipes* compared to *D. melanogaster*. Given the proximity of this feature to the hAT DNA transposon and the multiple matches to predicted transcripts in *D. subpulchrella* and *D. suzukii*, an alternative explanation is that the feature is part of a transposon that is found in *D. biarmipes*, *D. subpulchrella*, and *D. suzukii*. Hence we have omitted this predicted CDS in our annotation of the *CG31997* ortholog in *D. biarmipes*.

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 website. **Paste a screenshot of the protein alignment into the box below:**

Alignment of Dmel_CG31997-PB vs. Dbia3_CG31997-PB

View plain text version

Download alignment image

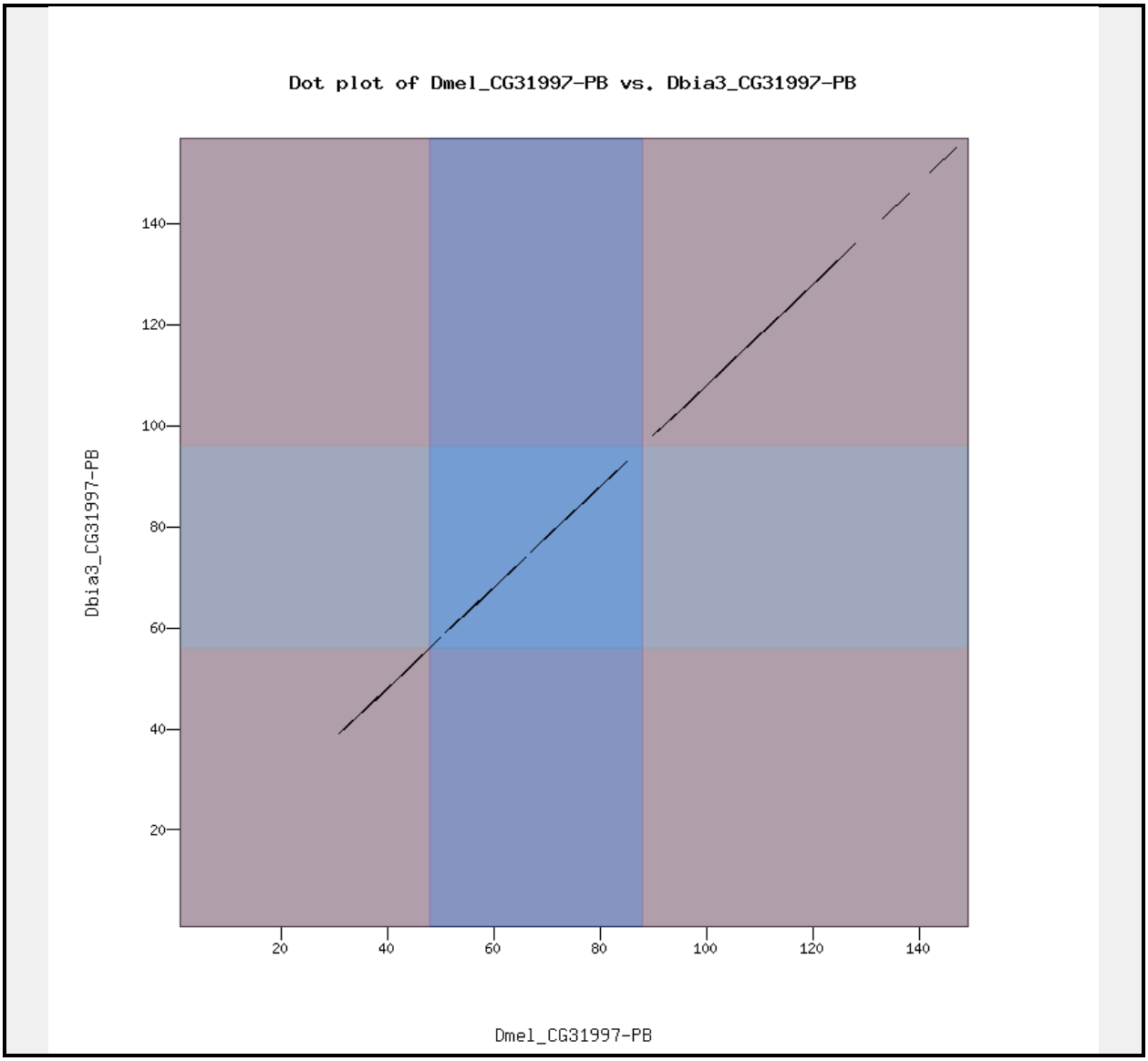
Identity: 121/156 (77.6%), **Similarity:** 132/156 (84.6%), **Gaps:** 8/156 (5.1%)

```
Dmel_CG31997-PB      1 MSFHFA--VLTLILTAFTVS---LCAEQKITKSDA---GEIRIFKRLIPADVLRDFPGMC 52
                       *.***   :* ***    ::   :.****: *.::   ******************.**
Dbia3_CG31997-PB     1 MGFHFHFDILLLILLLTLIAPFCIAAEQKVLKDETANVGEIRIFKRLIPADVLRDFPAMC 60

Dmel_CG31997-PB     53 FASTRCATVEPGKSWDLTPFCGRSTCVQNEENDAKLFELVEDCGPLPLANDKCKLDTEKT 112
                       ************:******************::**:************************
Dbia3_CG31997-PB    61 FASTRCATVEPGKTWDLTPFCGRSTCVQNEENETKLLELVEDCGPLPLANDKCKLDTEKT 120

Dmel_CG31997-PB    113 NKTASFPYCCPIFTCDPGVKLEYPEIGKDNDKKNSE 148
                       ***************:*** *****:**:***** *
Dbia3_CG31997-PB   121 NKTASFPYCCPIFTCEPGVALEYPEVGKENDKKNVE 156
```

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.



The dot plot shows that the last two CDS's of CG31997-PB are highly conserved between the proposed *D. biarmipes* gene model and the *D. melanogaster* ortholog. Examination of the protein alignment at the end of the second and third CDS's indicate that the amino acids have similar chemical properties even though they are not identical. In addition, the lengths of these two CDS's are the same between *D. biarmipes* and *D. melanogaster*.

The dot plot shows that the beginning of the first CDS of CG31997-PB is only weakly conserved between *D. biarmipes* and *D. melanogaster*. In addition, the dot plot shows that the first CDS of the *D. biarmipes* gene model is longer than the orthologous CDS in *D. melanogaster*. The protein alignment shows that there are 8 additional amino acids within the first CDS in the proposed *D. biarmipes* gene model compared to *D. melanogaster*.

Examination of this region in the GEP UCSC Genome Browser shows that there is only one methionine in frame +2 that could serve as the start codon for CG31997-PB (see screenshot below). The expansion of this CDS is consistent with the *blastx* alignment, the N-SCAN gene prediction, and the available RNA-Seq data. Consequently, our annotation has expanded the size of this CDS (1_10720_0) in order to retain this isoform in *D. biarmipes*.