GEP Annotation Report

**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.

Student name: Wilson Leung  
Student email: student@example.edu  
Faculty advisor: Sarah C.R. Elgin  
College/university: Washington University in St. Louis

**Project details**

Project name: contig10  
Project species: *D. biarmipes*  
Date of submission: 12/30/2018  
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

**Instructions for project with no genes**

If you believe that the project does not contain any genes, please provide the following evidence to support your conclusion:

1. Perform a NCBI BLASTX search of the entire contig sequence against the “non-redundant protein sequences (nr)” database. Provide an explanation for any significant (E-value < 1e-5) hits to known genes in the *nr* database as to why they do not correspond to real genes in the project.

2. For each Genscan prediction, perform a NCBI BLASTP search of the predicted amino acid sequence against the *nr* protein database using the strategy described above.

3. Examine the gene expression tracks (*e.g.*, RNA-Seq) for evidence of transcribed regions that do not correspond to alignments to known *D. melanogaster* proteins. Perform a NCBI BLASTX search against the *nr* protein database using these genomic regions to determine if they show sequence similarity to known or predicted proteins in the *nr* database.
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.

**Gene report form**

Gene name (e.g., *D. biarmipes* eyeless): *D. biarmipes* CG31997  
Gene symbol (e.g., dbia_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 for all the isoforms in this project:

<table>
<thead>
<tr>
<th>Name(s) of unique isoform(s) based on coding sequence</th>
<th>List of isoforms with identical coding sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG31997-PB</td>
<td>CG31997-PA</td>
</tr>
</tbody>
</table>

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

**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 they have identical coding sequences as other isoforms.

**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
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 name (e.g., dbia_ey-PA): dbia_CG31997-PB
Names of the isoforms with identical coding sequences as this isoform: dbia_CG31997-PA

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

1. 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
Use the custom track feature from the Gene Model Checker to capture a screenshot of your gene model shown on the Genome Browser for your project. Zoom in so that only this isoform is in the screenshot. (See page 12 of the Gene Model Checker user guide on how to do this; you can find the guide under “Help” ➔ “Documentations” ➔ “Web Framework” on the GEP website at http://gep.wustl.edu.)

Include the following evidence tracks in the screenshot if they are available:

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

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 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).
A NCBI BLASTN search of this region against the nt database detected five significant matches to predicted mRNAs in *Drosophila suzukii* (see screenshot below).

The E-values for these *D. suzukii* matches range from 3e-10 to 2e-06, and they correspond to three different predicted genes (LOC108013970, LOC108011950, and LOC108014610). 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 besides *D. suzukii*.

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. suzukii*, an alternative explanation is that the feature is part of a transposon that is found in both *D. biarmipes* 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 (*bl2seq*) at the NCBI BLAST web site. **Paste a screenshot of the protein alignment into the box below:**

![Alignment of CG31997-PB vs. Submitted_Seq](attachment:image.png)

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

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

![Dot plot of CG31997-PB vs. Submitted_Seq](attachment:image.png)
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_10777_0) in order to retain this isoform in *D. biarmipes*.

**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.

![RNA-Seq Alignment Summary](image_url)