# F Element Project: Annotation Report

Faculty instructor(s): James Price
College/university: Utah Valley University
Course number: BIOL 3515
Course name: Advanced Genetics Laboratory

## 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 (The student who completes this report)

| | |
|---|---|
| **First name** | Steven |
| **Middle initials** | C |
| **Last name** | Smith |
| **Author name** (name that will appear on the publication): | Steven Smith |
| **Permanent Email address** (one you will use five years from now): | Panzerfaust412@hotmail.com |
| **Alternative Email address** (optional): | |
| **Check this box to indicate that you have read and accept the co-authors responsibilities** | ☒ |

## Project Details

Project name: contig48
Project species: D. ananassae
Date of submission: _____
Size of project in base pairs: 1-60,000
Number of genes in project: 8

Does this report cover all of the genes or is it a partial report? All
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 CG14452-PA

Gene symbol (*e.g., dana_ey*): name of gene if not given in record finder

Although I was able to create a hypothetical gene model, reported below, I have doubts as to whether this gene is active in this species. Here are my reasons: ……

No RNA-seq to support gene expression. This may be a pseudogene or is expressed in different stages of the life cycle.

Approximate location in project (from 5' end to 3' end): _____
Number of isoforms in *D. melanogaster:* _____
Number of isoforms in this project: _____

**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** |
|---|---|
| | CG12546-PA |
| | |
| | |
| | |
| | |
| | |

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 <u>ALL</u> isoforms**, irrespective of whether their coding sequence is identical to that of another isoform.

## 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): _____

Names of any additional isoforms with identical coding sequences:

_____

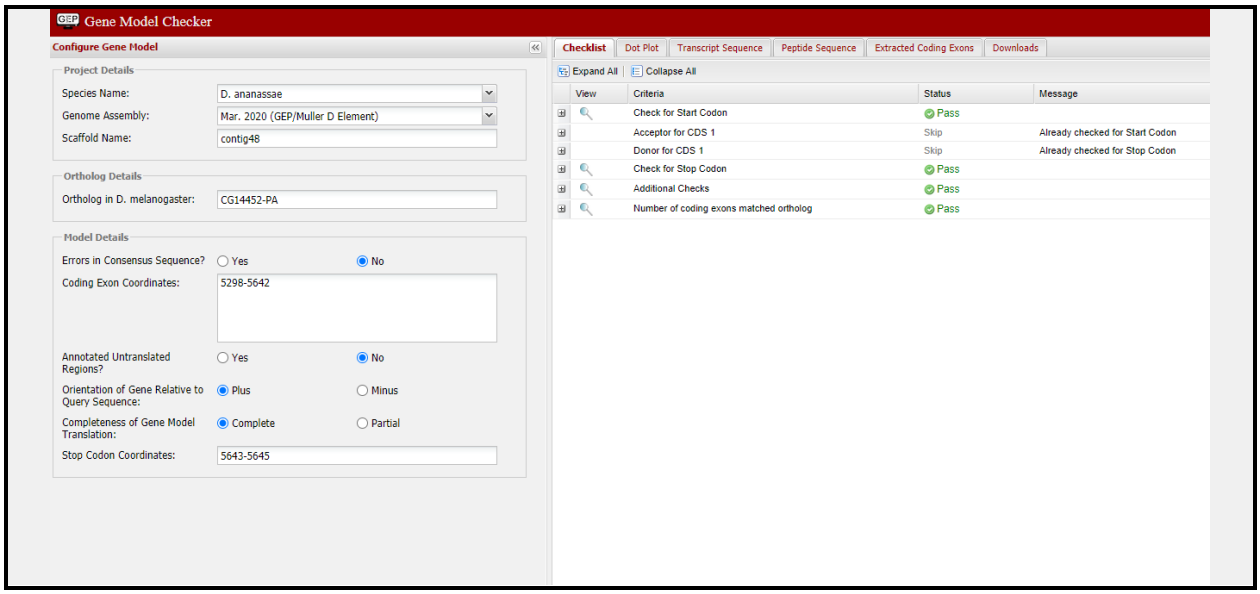
Is the 5' end of this isoform missing from the end of the project? _____
If so, how many putative exons are missing from the 5' end: _____
Is the 3' end of this isoform missing from the end of the project? _____
If so, how many putative exons are missing from the 3' end: _____

(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 **<u>original project sequence</u>**. 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.

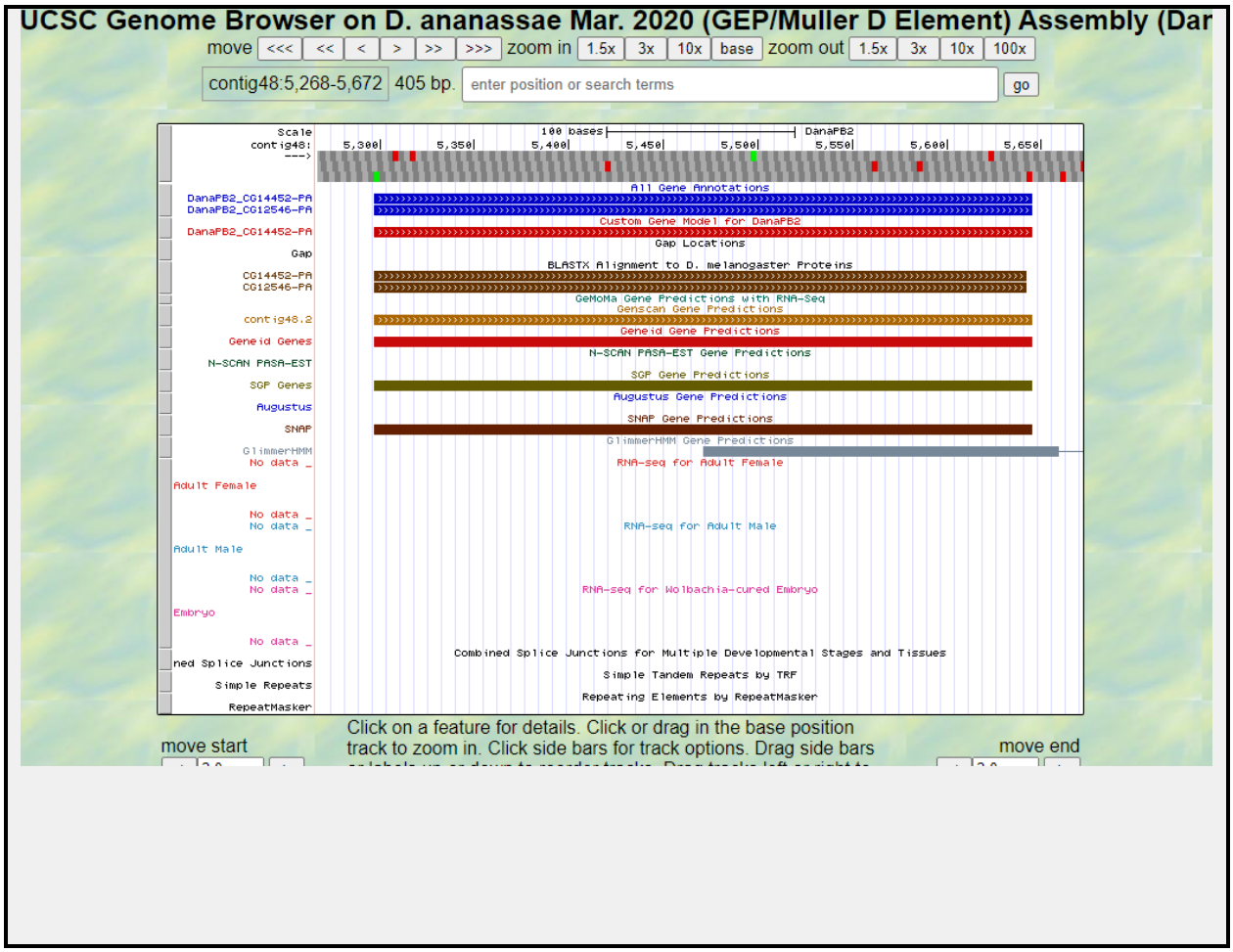



## 2. View the gene model on the Genome Browser

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:**

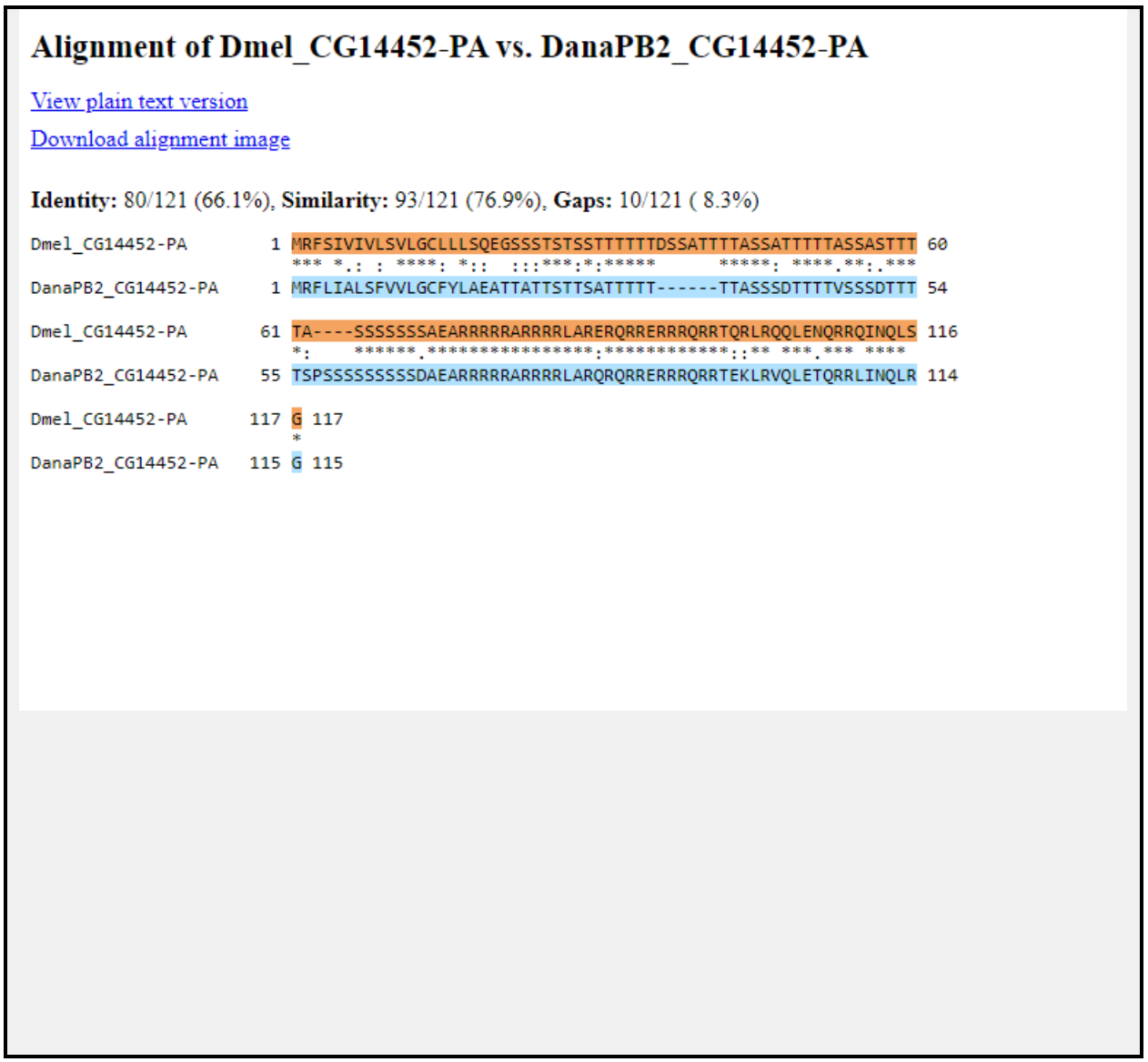
1. A sequence alignment track (*e.g.*, D. mel Proteins)
2. At least one gene prediction track (*e.g.*, Genscan)
3. At least one RNA-Seq track (*e.g.*, RNA-Seq Coverage)
4. A comparative genomics track (*e.g.*, D. mel. Net Alignment, Conservation)

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


UCSC Genome Browser on D. ananassae Mar. 2020 (GEP/Muller D Element) Assembly (Dar
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
contig48:5,268-5,672 405 bp. enter position or search terms go
Scale
contig48: 5,300 5,350 5,400 5,450 5,500 5,550 5,600 5,650
100 bases
DanaFB2
All Gene Annotations
DanaFB2_CG14452-PA
DanaFB2_CG12546-PA
Custom Gene Model for DanaFB2
DanaFB2_CG14452-PA
Gap Locations
Gap
BLASTX Alignment to D. melanogaster Proteins
CG14452-PA
CG12546-PA
GeMoMa Gene Predictions with RNA-Seq
Genscan Gene Predictions
contig48.2
Geneid Gene Predictions
Geneid Genes
N-SCAN PASA-EST Gene Predictions
N-SCAN PASA-EST
SGP Gene Predictions
SGP Genes
Augustus Gene Predictions
Augustus
SNAP Gene Predictions
SNAP
GlimmerHMM Gene Predictions
GlimmerHMM
RNA-seq for Adult Female
No data
Adult Female
RNA-seq for Adult Male
No data
No data
Adult Male
RNA-seq for Wolbachia-cured Embryo
No data
No data
Embryo
Combined Splice Junctions for Multiple Developmental Stages and Tissues
No data
ned Splice Junctions
Simple Tandem Repeats by TRF
Simple Repeats
Repeating Elements by RepeatMasker
RepeatMasker
move start
Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars
move end

## 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_CG14452-PA vs. DanaPB2_CG14452-PA**

View plain text version

Download alignment image

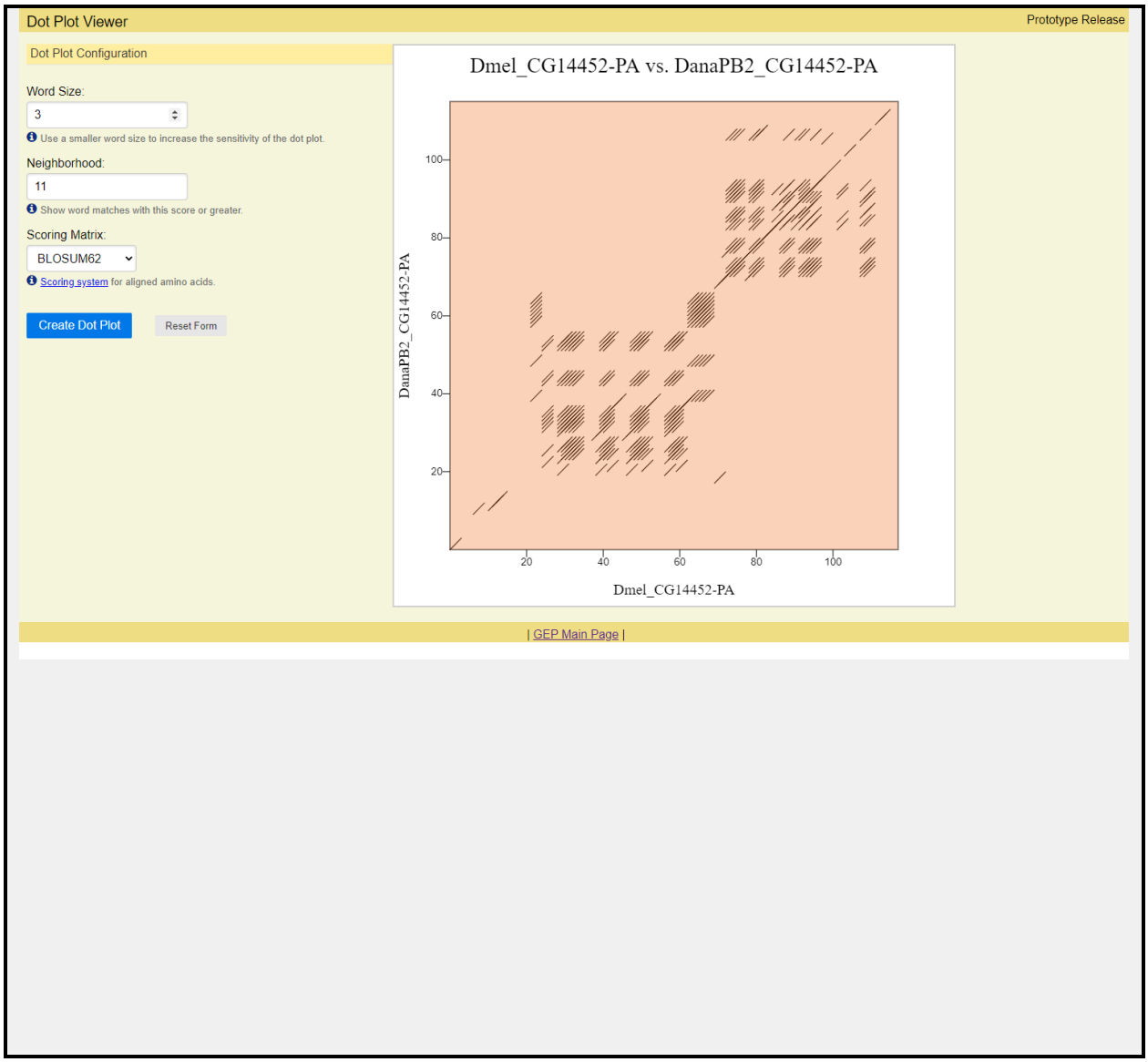
**Identity:** 80/121 (66.1%), **Similarity:** 93/121 (76.9%), **Gaps:** 10/121 ( 8.3%)

```
Dmel_CG14452-PA        1 MRFSIVIVLSVLGCLLLSQEGSSSTSTSSTTTTTTDSSATTTTASSATTTTTASSASTTT 60
                         *** *.: : ****: *::  :::***:*:*****       *****: ****.**:.***
DanaPB2_CG14452-PA     1 MRFLIALSFVVLGCFYLAEATTATTSTTSATTTTT------TTASSSDTTTTVSSSDTTT 54

Dmel_CG14452-PA       61 TA----SSSSSSSAEARRRRRARRRRLARERQRRERRRQRRTQRLRQQLENQRRQINQLS 116
                         *:    ******.***************:***********::** ***.*** ****
DanaPB2_CG14452-PA    55 TSPSSSSSSSSSDAEARRRRRARRRRLARQRQRRERRRQRRTEKLRVQLETQRRLINQLR 114

Dmel_CG14452-PA      117 G 117
                         *
DanaPB2_CG14452-PA   115 G 115
```

## 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.



## Gene Report Form

Gene name (*e.g.*, *D. ananassae eyeless*): D. ananassae CG32453-PB
Gene symbol (*e.g.*, *dana_ey*): DanaPB2_CG32453-PB name of gene if not given in record finder

Although I was able to create a hypothetical gene model, reported below, I have doubts as to whether this gene is active in this species. Here are my reasons: ……

No RNA-seq in adult females to support gene expression. Likely expressed in different stages of the life cycle.

Approximate location in project (from 5' end to 3' end): 7,427-7774
Number of isoforms in *D. melanogaster:*
Number of isoforms in this project: 3

**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** |
|---|---|
| | CG32453-PA |
| | CG14454-PA |
| | CG14454-PB |
| | |
| | |
| | |

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.):

## 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):

Names of any additional isoforms with identical coding sequences:

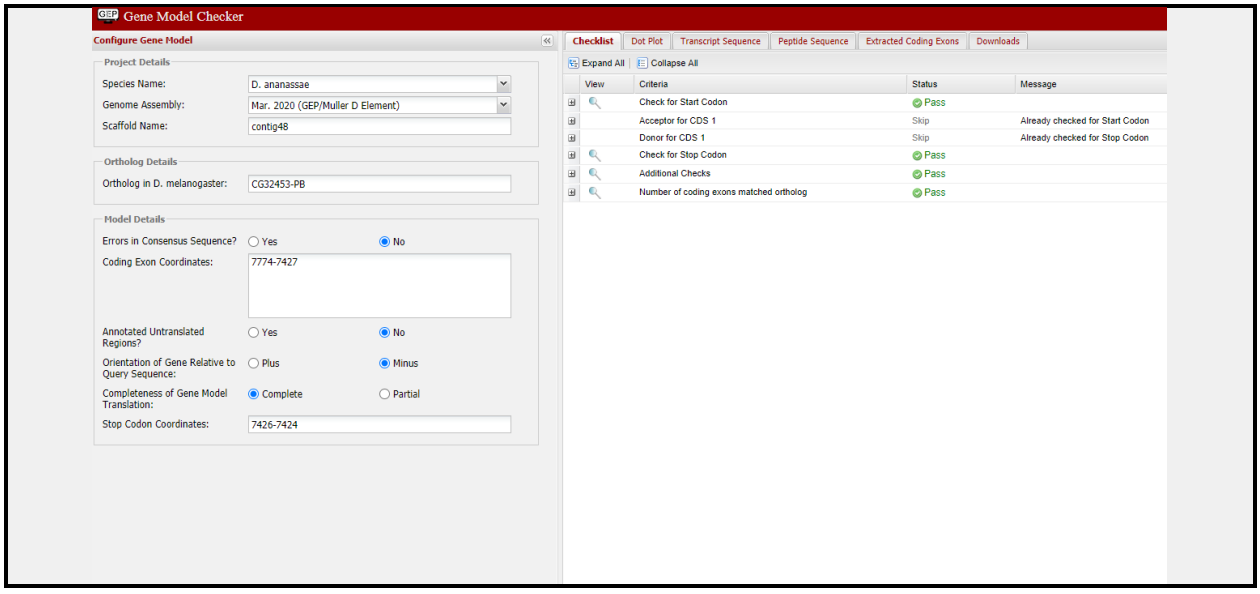
Is the 5' end of this isoform missing from the end of the project? _____
If so, how many putative exons are missing from the 5' end: _____
Is the 3' end of this isoform missing from the end of the project? _____
If so, how many putative exons are missing from the 3' end: _____

(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.



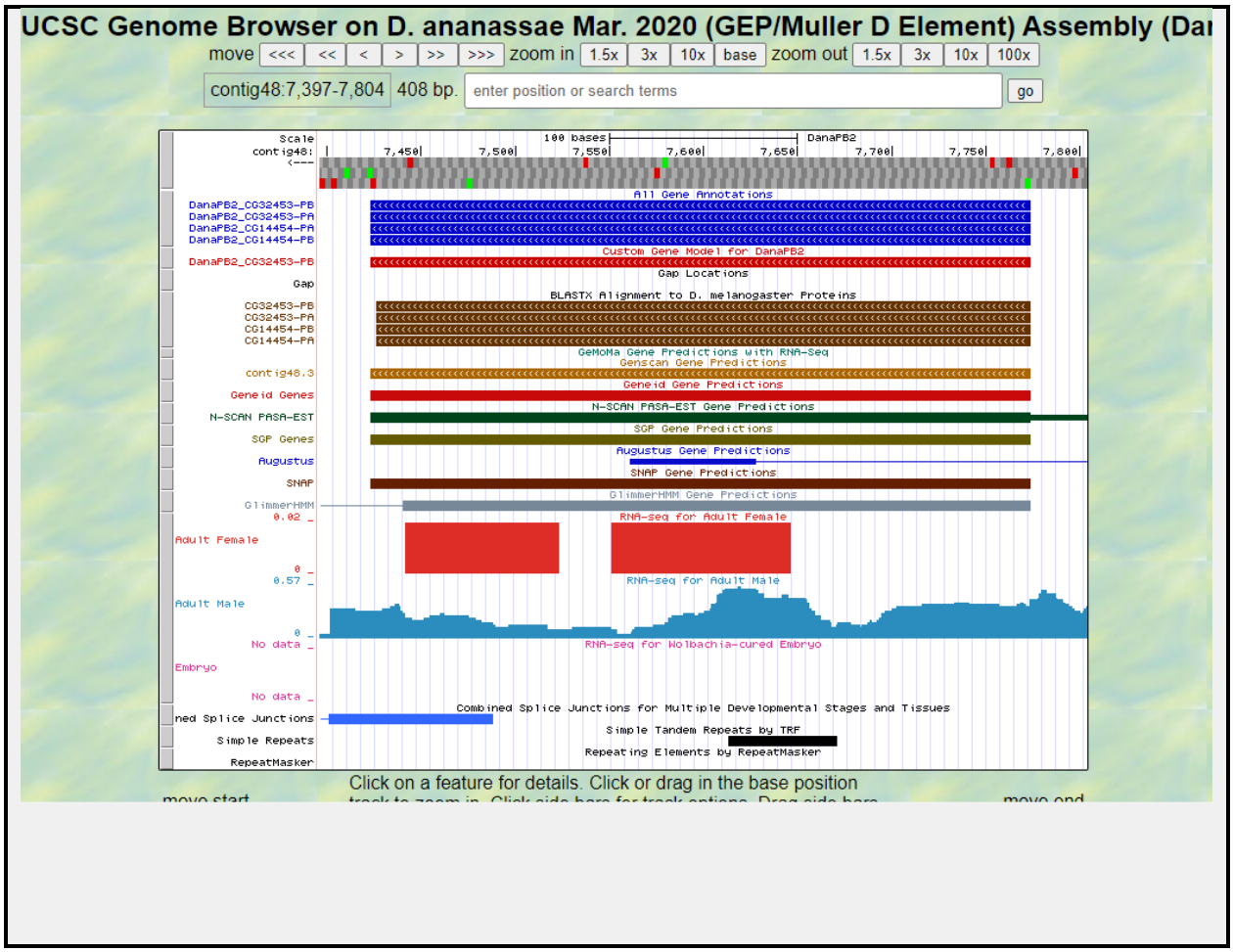

## 2. View the gene model on the Genome Browser

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)

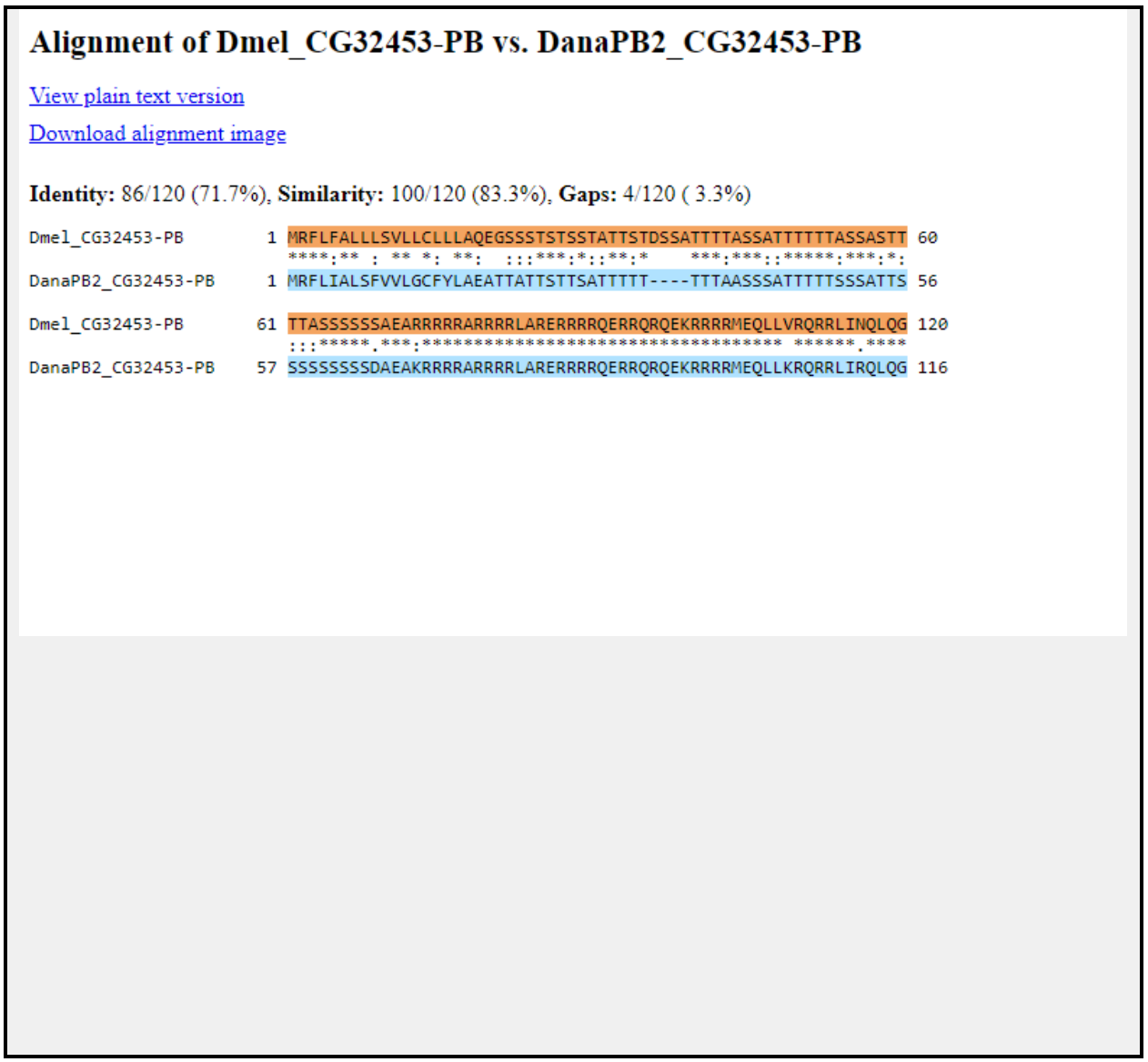
4. A comparative genomics track (*e.g.*, D. mel. Net Alignment, Conservation)

**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_CG32453-PB vs. DanaPB2_CG32453-PB**

View plain text version

Download alignment image

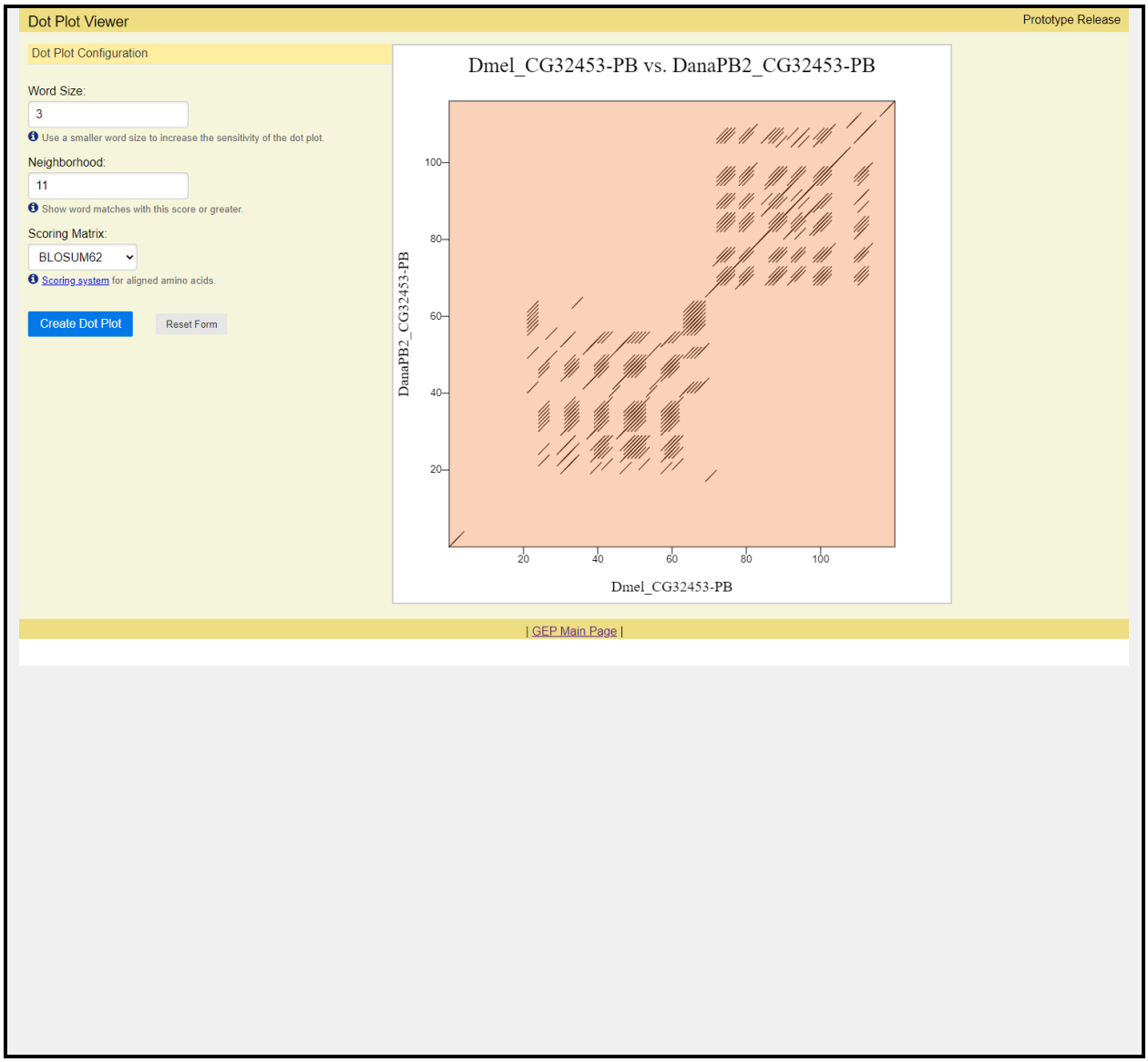
**Identity:** 86/120 (71.7%), **Similarity:** 100/120 (83.3%), **Gaps:** 4/120 ( 3.3%)

```
Dmel_CG32453-PB        1 MRFLFALLLSVLLCLLLAQEGSSSTSTSSTATTSTDSSATTTTASSATTTTTTASSASTT 60
                         ****:** : ** *: **:  :::***:*::**:*    ***:***::*****:***:*:
DanaPB2_CG32453-PB     1 MRFLIALSFVVLGCFYLAEATTATTSTTSATTTTT----TTTAASSSATTTTTSSSATTS 56

Dmel_CG32453-PB       61 TTASSSSSSAEARRRRRARRRRLARERRRRQERRQRQEKRRRRMEQLLVRQRRLINQLQG 120
                         :::*****.***:********************************** ******.****
DanaPB2_CG32453-PB    57 SSSSSSSSDAEAKRRRRARRRRLARERRRRQERRQRQEKRRRRMEQLLKRQRRLIRQLQG 116
```

## 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.



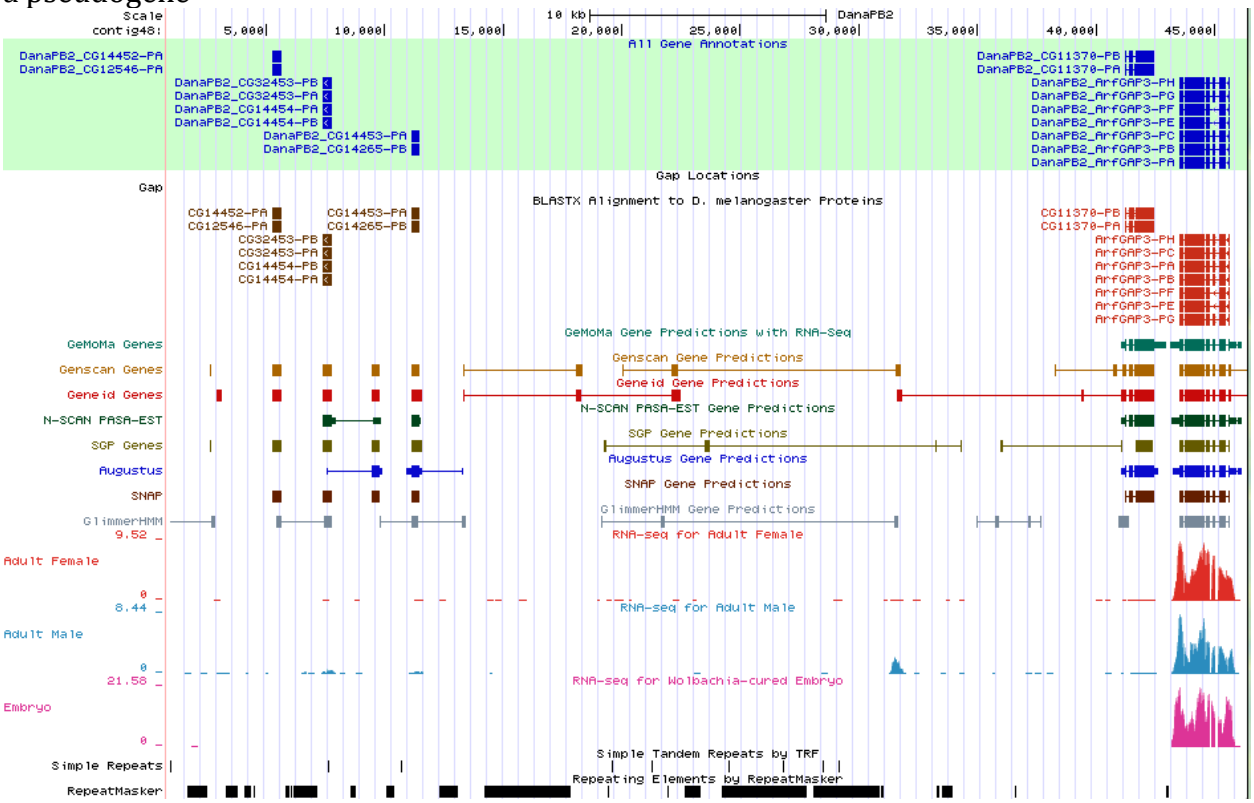
# Gene Report Form

Gene name (*e.g.*, *D. ananassae eyeless*): D. ananassae CG14453-PA

Gene symbol (*e.g.*, *dana_ey*): name of gene if not given in record finder

Although I was able to create a hypothetical gene model, reported below, I have doubts as to whether this gene is active in this species. Here are my reasons: ……

Little or no RNA-seq throughout entirety of exons to support gene expression. This may be a pseudogene




.

Approximate location in project (from 5' end to 3' end): 11,135-11,503
Number of isoforms in *D. melanogaster:*
Number of isoforms in this project:

**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 |
|---|---|
| | CG14265-PB |
| | |
| | |
| | |
| | |
| | |

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.):

## 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): _____

Names of any additional isoforms with identical coding sequences:

_____

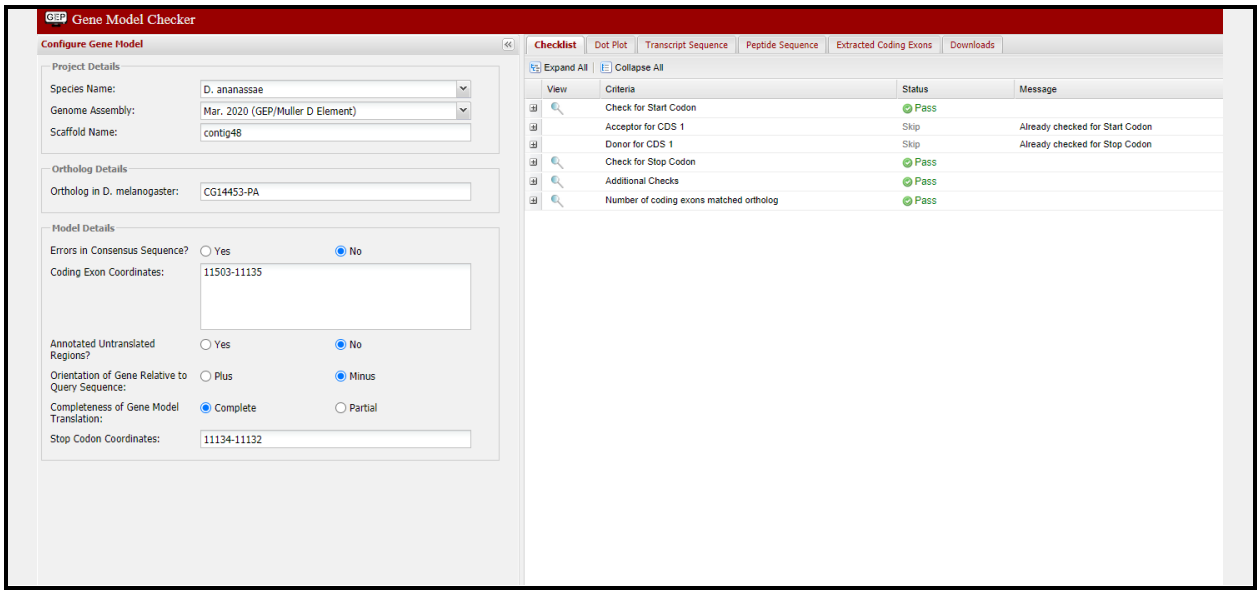
Is the 5' end of this isoform missing from the end of the project? _____
If so, how many putative exons are missing from the 5' end: _____
Is the 3' end of this isoform missing from the end of the project? _____
If so, how many putative exons are missing from the 3' end: _____

(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.

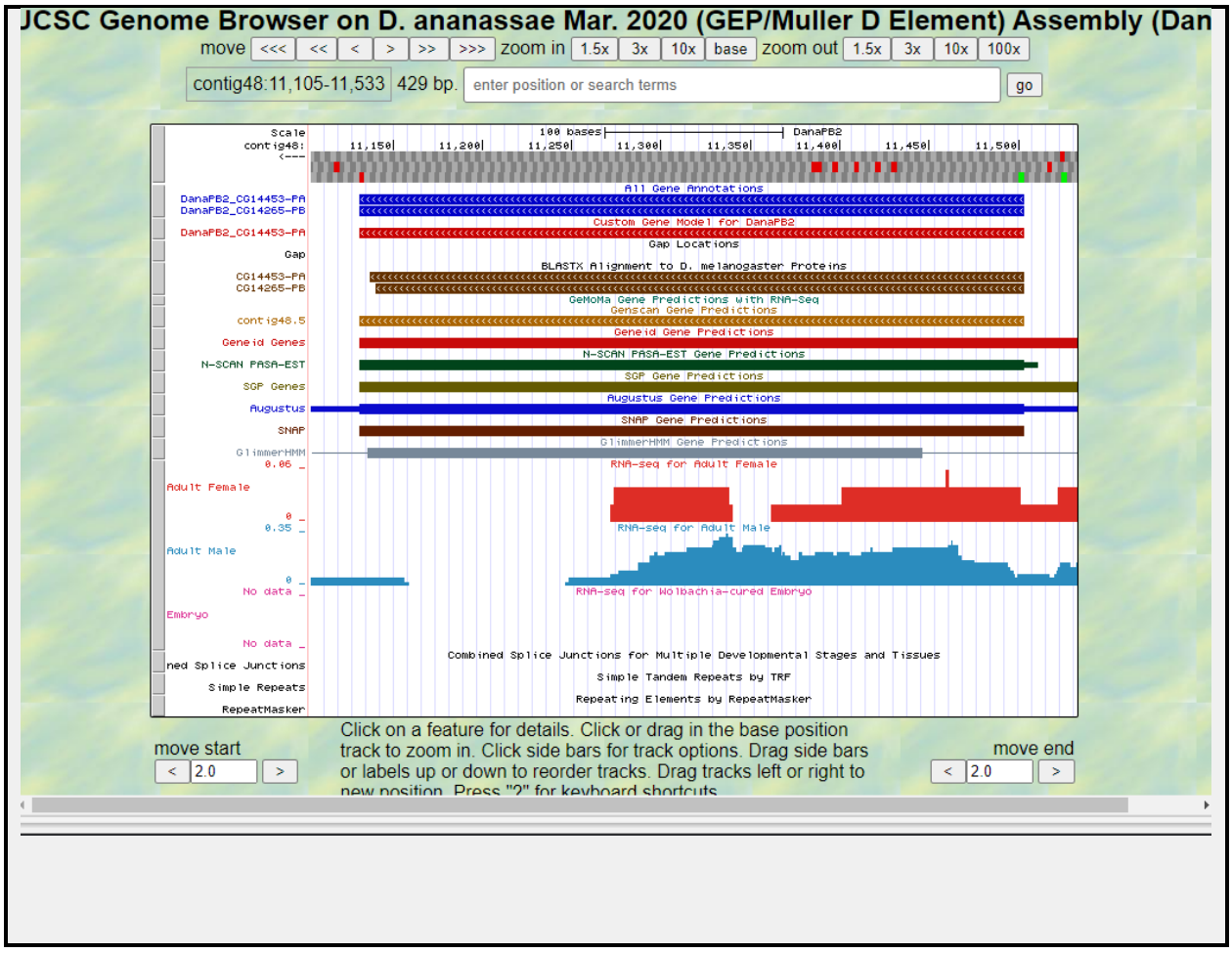



## 2. View the gene model on the Genome Browser

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:**

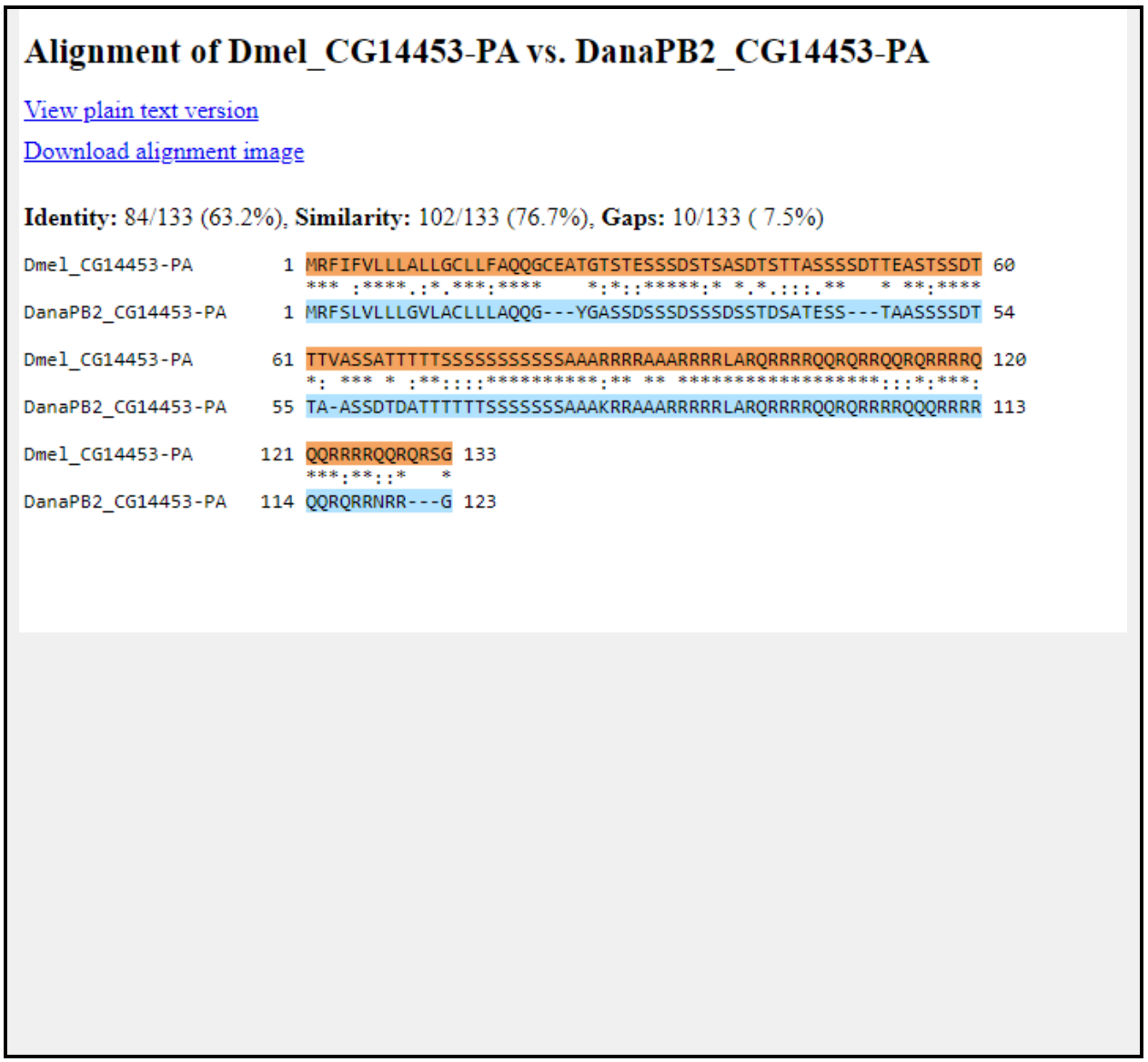
1. A sequence alignment track (*e.g.*, D. mel Proteins)
2. At least one gene prediction track (*e.g.*, Genscan)
3. At least one RNA-Seq track (*e.g.*, RNA-Seq Coverage)
4. A comparative genomics track (*e.g.*, D. mel. Net Alignment, Conservation)

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


JCSC Genome Browser on D. ananassae Mar. 2020 (GEP/Muller D Element) Assembly (Dan
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
contig48:11,105-11,533 429 bp. enter position or search terms go
Scale
contig48:
100 bases
DanaPB2
11,150 11,200 11,250 11,300 11,350 11,400 11,450 11,500
All Gene Annotations
DanaPB2_CG14453-PA
DanaPB2_CG14265-PB
Custom Gene Model for DanaPB2
DanaPB2_CG14453-PA
Gap Locations
Gap
BLASTX Alignment to D. melanogaster Proteins
CG14453-PA
CG14265-PB
GeMoMa Gene Predictions with RNA-Seq
Genscan Gene Predictions
contig48.5
Geneid Gene Predictions
Geneid Genes
N-SCAN PASA-EST Gene Predictions
N-SCAN PASA-EST
SGP Gene Predictions
SGP Genes
Augustus Gene Predictions
Augustus
SNAP Gene Predictions
SNAP
GlimmerHMM Gene Predictions
GlimmerHMM
0.05
RNA-seq for Adult Female
Adult Female
0
0.35
RNA-seq for Adult Male
Adult Male
0
No data
RNA-seq for Wolbachia-cured Embryo
Embryo
No data
Combined Splice Junctions for Multiple Developmental Stages and Tissues
ned Splice Junctions
Simple Tandem Repeats by TRF
Simple Repeats
Repeating Elements by RepeatMasker
RepeatMasker
move start
< 2.0 >
Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts
move end
< 2.0 >

## 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_CG14453-PA vs. DanaPB2_CG14453-PA**

View plain text version

Download alignment image

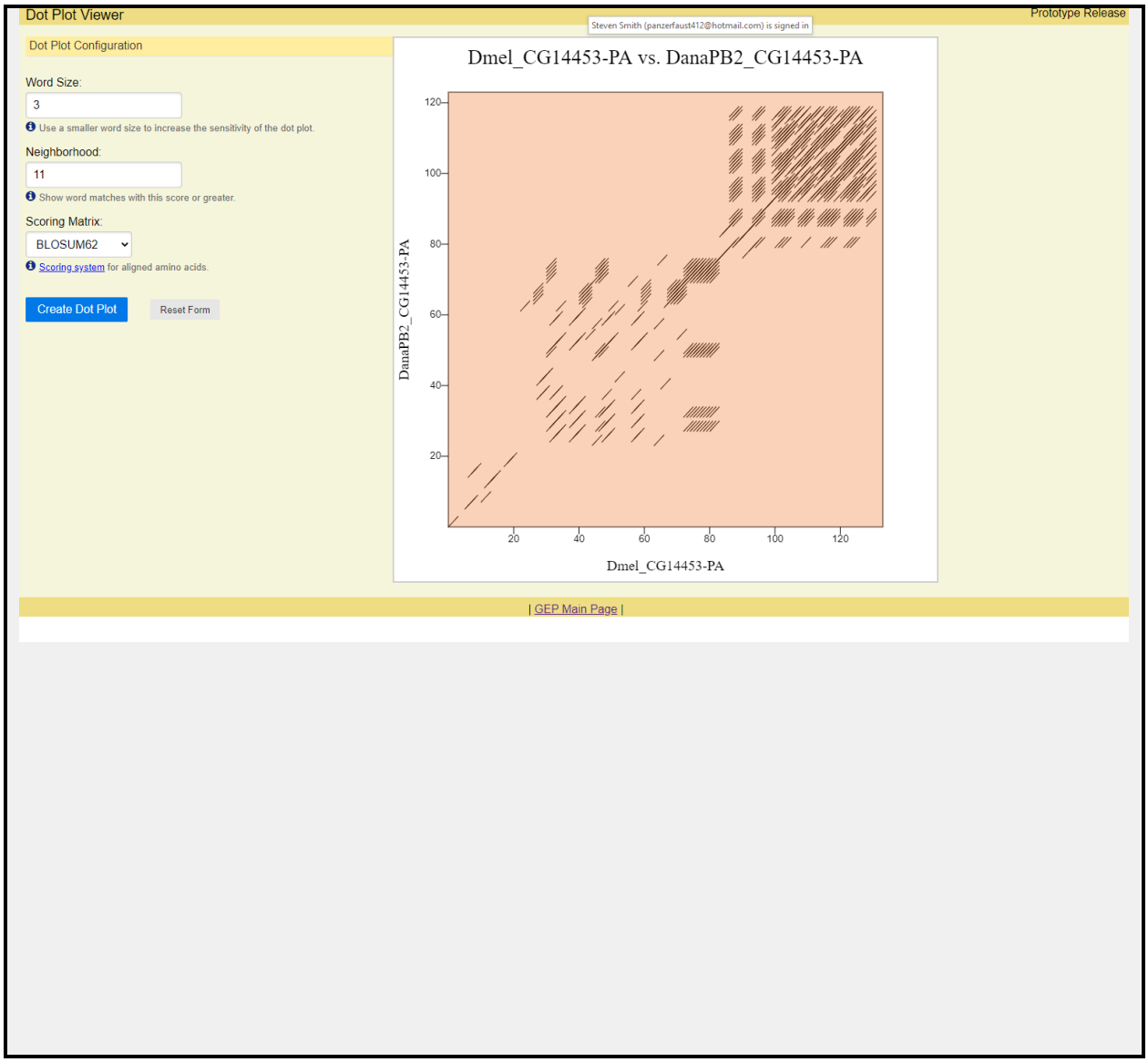
**Identity:** 84/133 (63.2%), **Similarity:** 102/133 (76.7%), **Gaps:** 10/133 ( 7.5%)

```
Dmel_CG14453-PA        1 MRFIFVLLLALLGCLLFAQQGCEATGTSTESSSDSTSASDTSTTASSSSDTTEASTSSDT 60
                         *** :****.:*.***:****    *:*::*****:* *.*.:::.**    * **:****
DanaPB2_CG14453-PA     1 MRFSLVLLLGVLACLLLAQQG---YGASSDSSSDSSSDSSTDSATESS---TAASSSSDT 54

Dmel_CG14453-PA       61 TTVASSATTTTTSSSSSSSSSSAAARRRRAAARRRRLARQRRRRQQRQRRQQRQRRRRQ 120
                         *: *** * :**::::**********:** ** *****************:::*:***:
DanaPB2_CG14453-PA    55 TA-ASSDTDATTTTTTSSSSSSSAAAKRRAAARRRRRLARQRRRRQQRQRRRRQQQRRRR 113

Dmel_CG14453-PA      121 QQRRRRQQRQRSG 133
                         ***:**::*   *
DanaPB2_CG14453-PA   114 QQRQRRNRR---G 123
```

## 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.



## Gene Report Form

Gene name (*e.g.*, *D. ananassae eyeless*): D. ananassae CG11370-PB 1

Gene symbol (*e.g.*, *dana_ey*): DanaPB2 CG11370-PB name of gene if not given in record finder

Although I was able to create a hypothetical gene model, reported below, I have doubts as to whether this gene is active in this species. Here are my reasons: ……

Approximate location in project (from 5' end to 3' end): 41,180-42,441
Number of isoforms in *D. melanogaster:*
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** |
|---|---|
| CG11370-PA | |
| | |
| | |
| | |
| | |
| | |

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.):

## 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): CG11370-PA

Names of any additional isoforms with identical coding sequences:

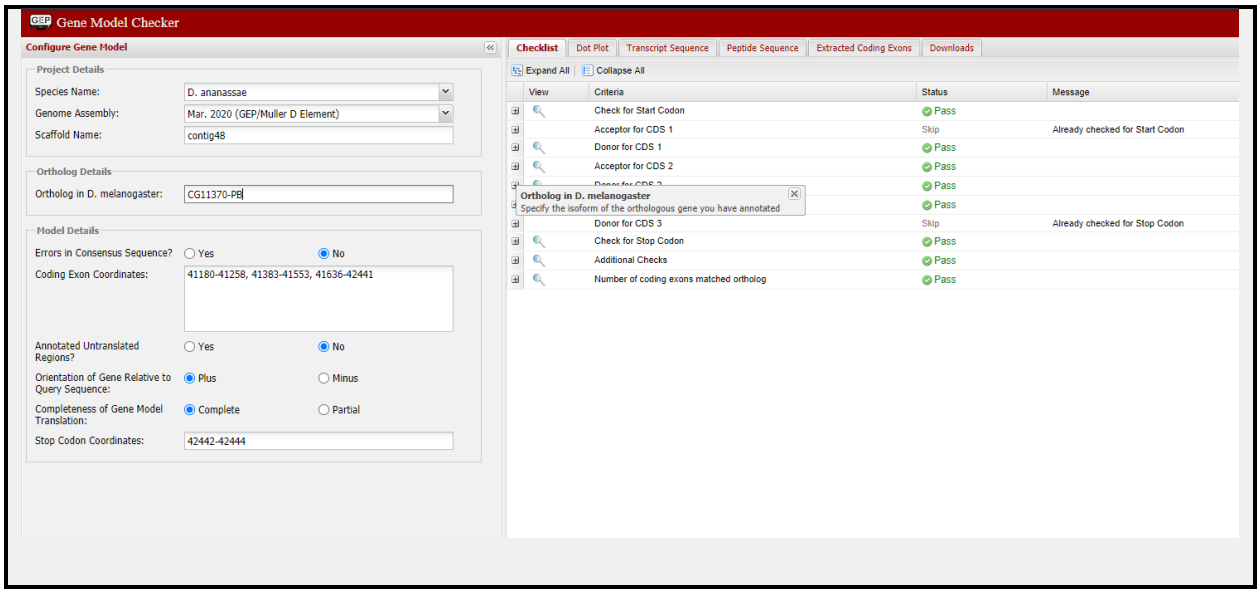
Is the 5' end of this isoform missing from the end of the project? No
If so, how many putative exons are missing from the 5' end: _____
Is the 3' end of this isoform missing from the end of the project? No
If so, how many putative exons are missing from the 3' end: _____

(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.



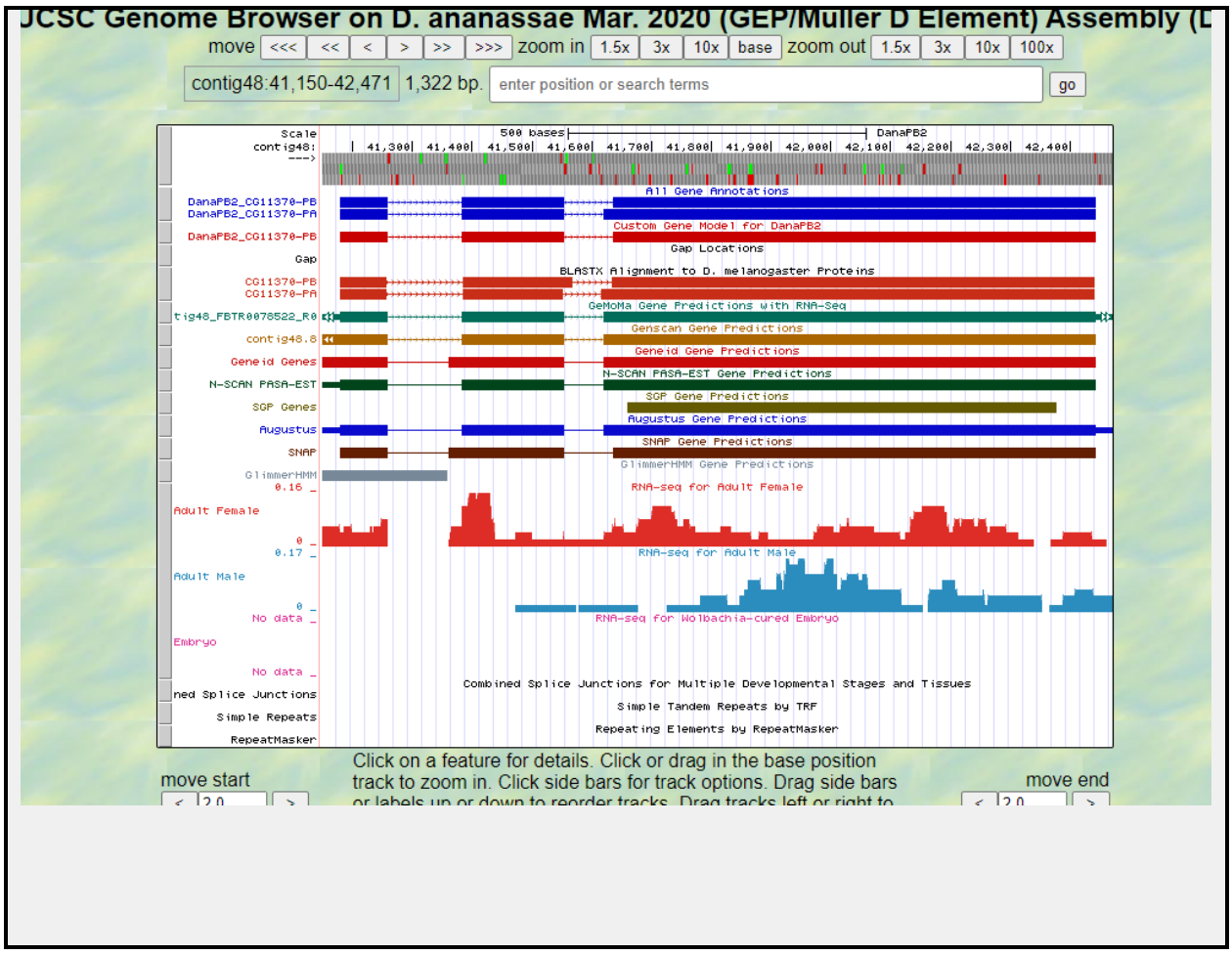

## 2. View the gene model on the Genome Browser

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:**

5. A sequence alignment track (*e.g.*, D. mel Proteins)
6. At least one gene prediction track (*e.g.*, Genscan)

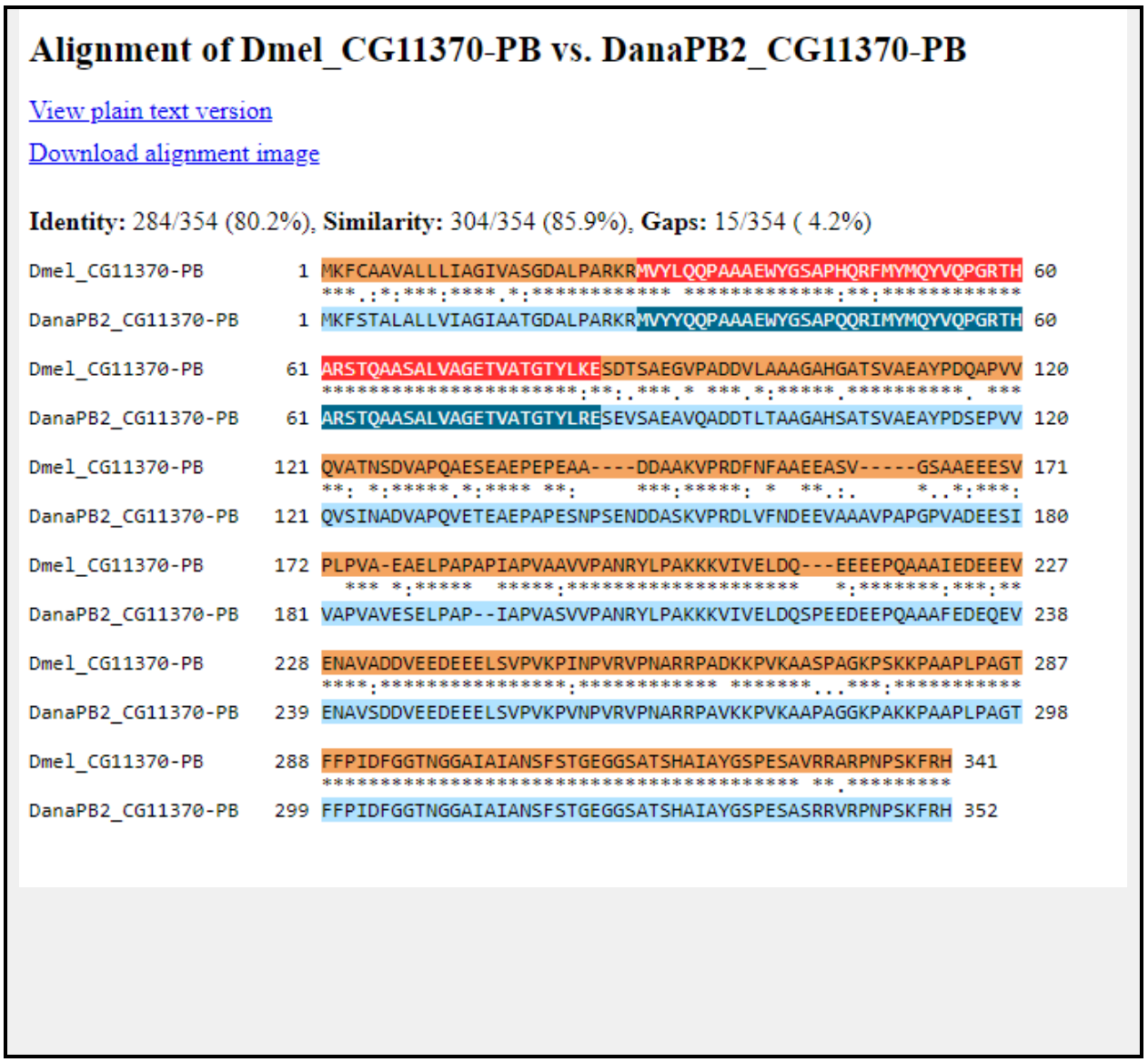
7. At least one RNA-Seq track (*e.g.*, RNA-Seq Coverage)
8. A comparative genomics track (*e.g.*, D. mel. Net Alignment, Conservation)

**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_CG11370-PB vs. DanaPB2_CG11370-PB**

View plain text version

Download alignment image

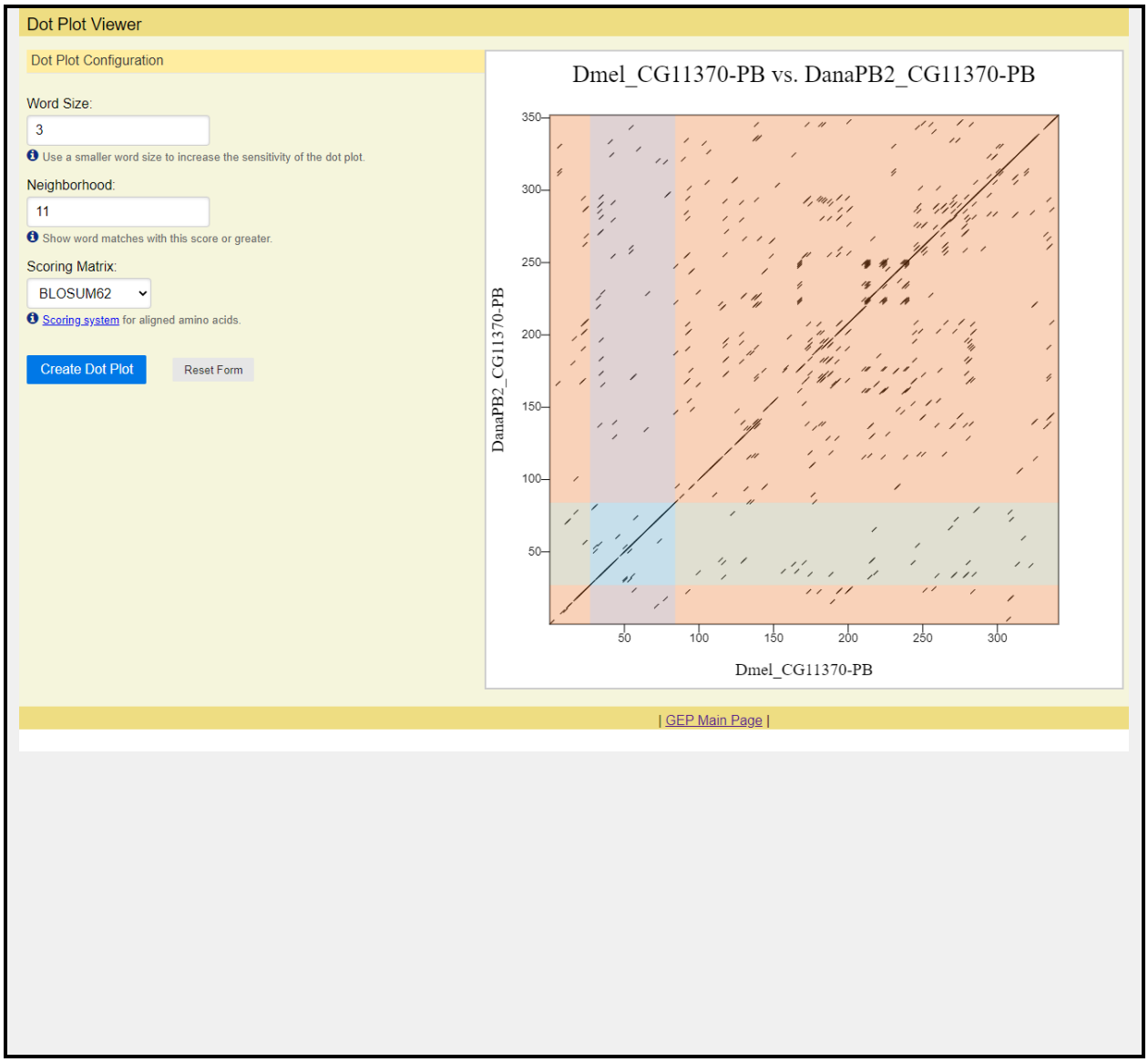
**Identity:** 284/354 (80.2%), **Similarity:** 304/354 (85.9%), **Gaps:** 15/354 ( 4.2%)

```
Dmel_CG11370-PB       1 MKFCAAVALLLIAGIVASGDALPARKRMVYLQQPAAAEWYGSAPHQRFMYMQYVQPGRTH 60
                        ***.:*:***:****.*:*********** ************:**:************
DanaPB2_CG11370-PB    1 MKFSTALALLVIAGIAATGDALPARKRMVYYQQPAAAEWYGSAPQQRIMYMQYVQPGRTH 60

Dmel_CG11370-PB      61 ARSTQAASALVAGETVATGTYLKESDTSAEGVPADDVLAAAGAHGATSVAEAYPDQAPVV 120
                        **********************:**:.***.* ***.*:*****.**********. ***
DanaPB2_CG11370-PB    61 ARSTQAASALVAGETVATGTYLRESEVSAEAVQADDTLTAAGAHSATSVAEAYPDSEPVV 120

Dmel_CG11370-PB      121 QVATNSDVAPQAESEAEPEPEAA----DDAAKVPRDFNFAAEEASV-----GSAAEEESV 171
                        **: *:*****.*:**** **:     ***:*****: *  **.:.     *..*:***:
DanaPB2_CG11370-PB    121 QVSINADVAPQVETEAEPAPESNPSENDDASKVPRDLVFNDEEVAAAVPAPGPVADEESI 180

Dmel_CG11370-PB      172 PLPVA-EAELPAPAPIAPVAAVVPANRYLPAKKKVIVELDQ---EEEEPQAAAIEDEEEV 227
                           *** *:*****  *****:********************   *:*******:***:**
DanaPB2_CG11370-PB    181 VAPVAVESELPAP--IAPVASVVPANRYLPAKKKVIVELDQSPEEDEEPQAAAFEDEQEV 238

Dmel_CG11370-PB      228 ENAVADDVEEDEEELSVPVKPINPVRVPNARRPADKKPVKAASPAGKPSKKPAAPLPAGT 287
                        ****:***************:*********** *******...***:***********
DanaPB2_CG11370-PB    239 ENAVSDDVEEDEEELSVPVKPVNPVRVPNARRPAVKKPVKAAPAGGKPAKKPAAPLPAGT 298

Dmel_CG11370-PB      288 FFPIDFGGTNGGAIAIANSFSTGEGGSATSHAIAYGSPESAVRRARPNPSKFRH 341
                        ************************************** **.*********
DanaPB2_CG11370-PB    299 FFPIDFGGTNGGAIAIANSFSTGEGGSATSHAIAYGSPESASRRVRPNPSKFRH 352
```

## 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.



## Gene Report Form

Gene name (*e.g.*, *D. ananassae eyeless*): D. ananassae ArfGAP3-PH
Gene symbol (*e.g.*, *dana_ey*): name of gene if not given in record finder

Although I was able to create a hypothetical gene model, reported below, I have doubts as to whether this gene is active in this species. Here are my reasons: ……

Approximate location in project (from 5' end to 3' end): 45,635-43,495
Number of isoforms in *D. melanogaster:*
Number of isoforms in this project: 7

**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** |
|---|---|
| ArfGAP3-PF | ArfGAP3-PC |
| ArfGAP3-PG | ArfGAP3-PA |
| | ArfGAP3-PB |
| | ArfGAP3-PG |
| | ArfGAP3-PE (identical to ArfGAP3-PF) |
| | |

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.):

## 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): ArfGAP3-PF

Names of any additional isoforms with identical coding sequences:
ArfGAP3-PE

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)

## 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): ArfGAP3-PG

Names of any additional isoforms with identical coding sequences:

_____

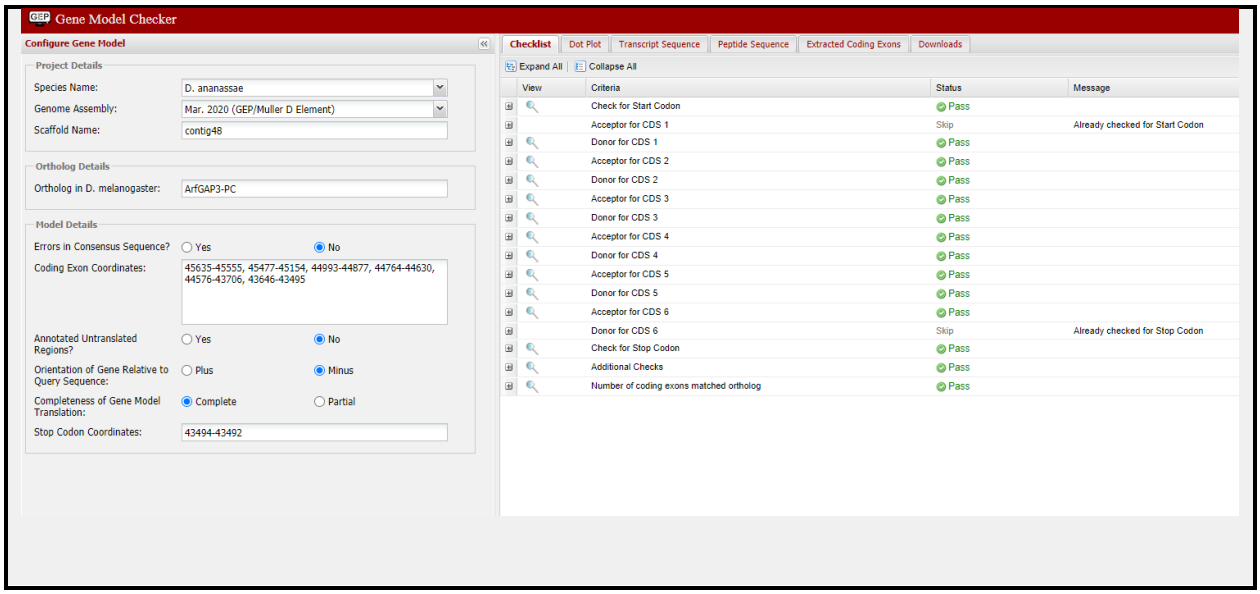
Is the 5' end of this isoform missing from the end of the project? No
If so, how many putative exons are missing from the 5' end: _____
Is the 3' end of this isoform missing from the end of the project? No
If so, how many putative exons are missing from the 3' end: _____

(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.

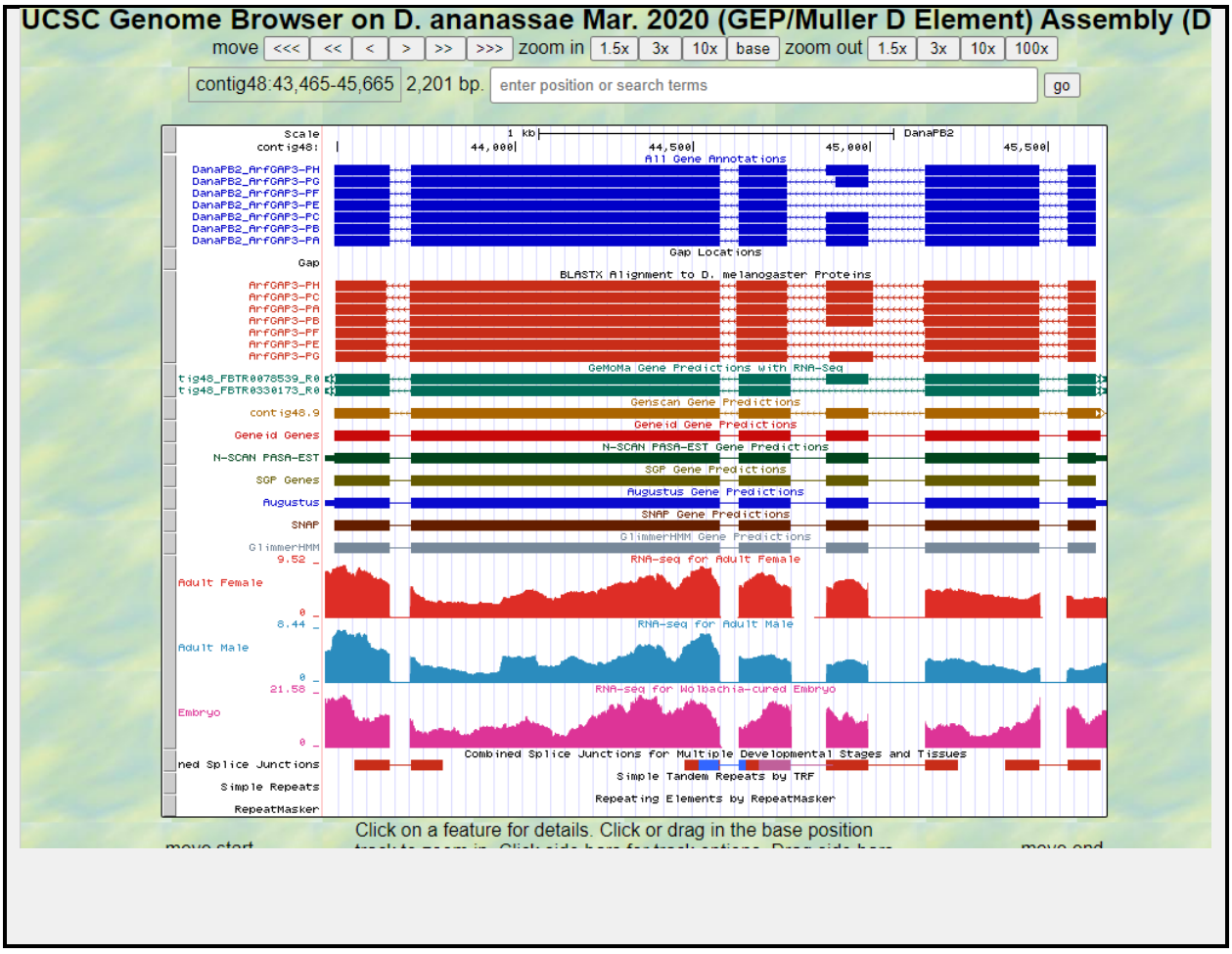



## 2. View the gene model on the Genome Browser

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:**


UCSC Genome Browser on D. ananassae Mar. 2020 (GEP/Muller D Element) Assembly (D
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
contig48:43,465-45,665 2,201 bp. enter position or search terms go
Scale
contig48:
1 kb
44,000
44,500
45,000
45,500
DanaPB2
All Gene Annotations
DanaPB2_ArfGAP3-PH
DanaPB2_ArfGAP3-PG
DanaPB2_ArfGAP3-PF
DanaPB2_ArfGAP3-PE
DanaPB2_ArfGAP3-PC
DanaPB2_ArfGAP3-PB
DanaPB2_ArfGAP3-PA
Gap
Gap Locations
BLASTX Alignment to D. melanogaster Proteins
ArfGAP3-PH
ArfGAP3-PC
ArfGAP3-PA
ArfGAP3-PB
ArfGAP3-PF
ArfGAP3-PE
ArfGAP3-PG
GeMoMa Gene Predictions with RNA-Seq
tig48_FBTR0078539_R0
tig48_FBTR0330173_R0
Genscan Gene Predictions
contig48.9
Geneid Gene Predictions
Geneid Genes
N-SCAN PASA-EST Gene Predictions
N-SCAN PASA-EST
SGP Gene Predictions
SGP Genes
Augustus Gene Predictions
Augustus
SNAP Gene Predictions
SNAP
GlimmerHMM Gene Predictions
GlimmerHMM
RNA-seq for Adult Female
9.52
Adult Female
0
RNA-seq for Adult Male
8.44
Adult Male
0
RNA-seq for Wolbachia-cured Embryo
21.58
Embryo
0
Combined Splice Junctions for Multiple Developmental Stages and Tissues
ned Splice Junctions
Simple Tandem Repeats by TRF
Simple Repeats
Repeating Elements by RepeatMasker
RepeatMasker
Click on a feature for details. Click or drag in the base position

## 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_ArfGAP3-PC vs. DanaPB2_ArfGAP3-PC**

View plain text version

Download alignment image

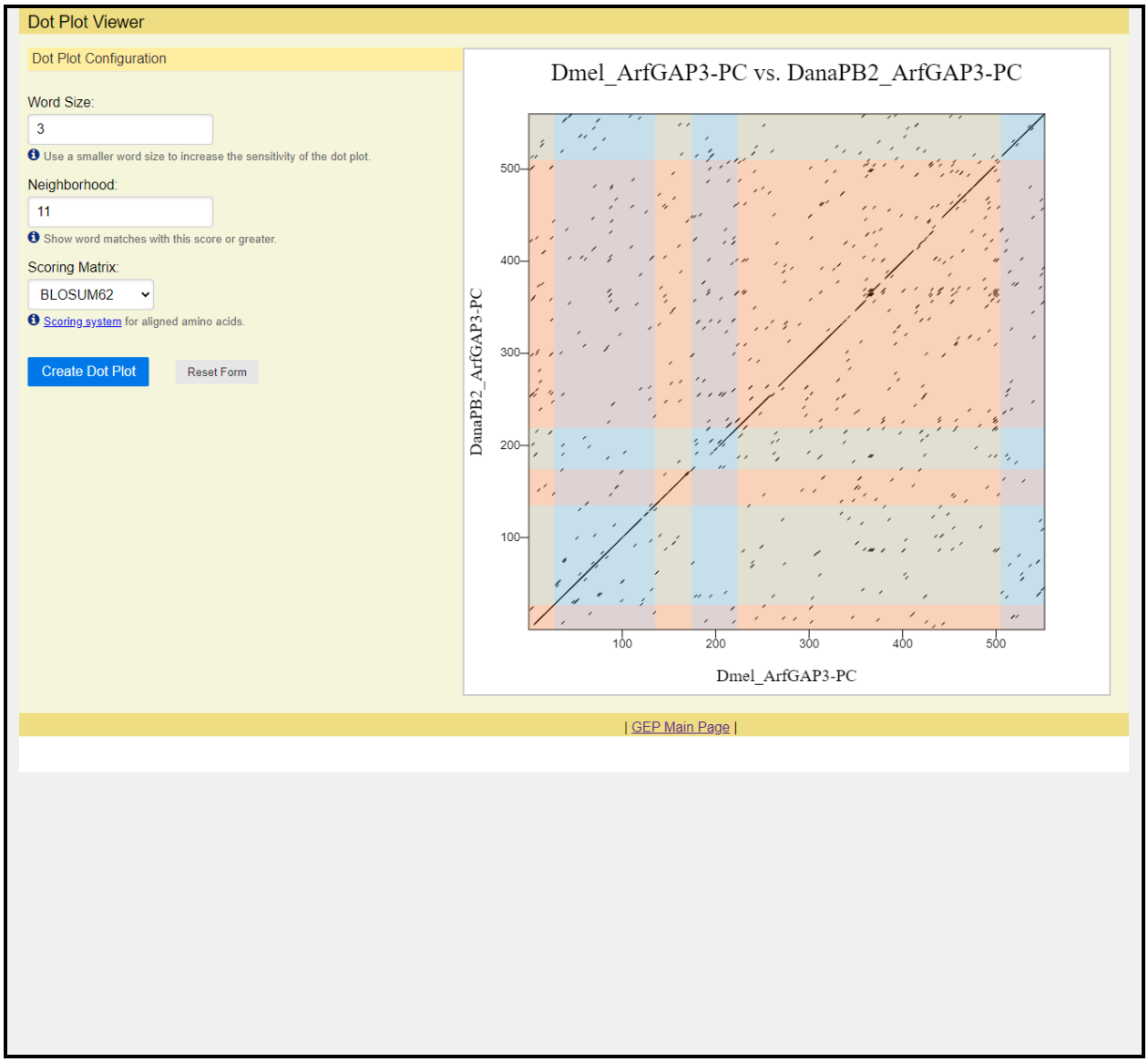
**Identity:** 451/564 (80.0%), **Similarity:** 498/564 (88.3%), **Gaps:** 16/564 ( 2.8%)

```
Dmel_ArfGAP3-PC       1 MASPAAGPSKQEIESVFSRLRAQPANKSCFDCAAKAPTWSSVTYGIFICIDCSAVHRNLG  60
                        **: ::**:********:******************************************
DanaPB2_ArfGAP3-PC    1 MATQTTGPTKQEIESVFTRLRAQPANKSCFDCAAKAPTWSSVTYGIFICIDCSAVHRNLG  60

Dmel_ArfGAP3-PC      61 VHLTFVRSTNLDTNWTWLQLRQMQLGGNANAAQFFRAHNCSTTDAQVKYNSRAAQLYRDK 120
                        ************************************:***:.******************
DanaPB2_ArfGAP3-PC    61 VHLTFVRSTNLDTNWTWLQLRQMQLGGNANAAQFFRSHNCTNTDAQVKYNSRAAQLYRDK 120

Dmel_ArfGAP3-PC      121 LCAQAQQAMKTHGTKLHLEQTDKSEGNEAAREEDFFAQCDNEVDFNVQNNNVSKDPNPPT 180
                        *.:******:*.*********.:******:*:*****:***:******:*** .*    *
DanaPB2_ArfGAP3-PC    121 LSSQAQQAIKVHGTKLHLEQSEKSEGNESAKEEDFFSQCDHEVDFNVENNNCKK----PI 176

Dmel_ArfGAP3-PC      181 VAPVISVETQQGGAPSVNLLNSVVPAAVPSSIGARKVQPKKGGLGARKVGGLGATKVKTN 240
                         * : . *.* *..*:*:****.**:****:**.**:*********************:*
DanaPB2_ArfGAP3-PC    177 PALIKDSEPQLGSGPTVDLLNSFVPTAVPSTIGVRKIQPKKGGLGARKVGGLGATKVKAN 236

Dmel_ArfGAP3-PC      241 FADIEARANAANEMKTSAA-AAPVVKPQTAEDELETVASMRLAYQELSMQKTREEAKLKT 299
                        ****************** *:   * :***:*:**************:**********:
DanaPB2_ArfGAP3-PC    237 FADIEARANAANEMKTSAAPASQPDKLKTAEEEVETVASMRLAYQELSLQKTREEAKLKS 296

Dmel_ArfGAP3-PC      300 MDPAKAKQMERLGMGFNLRGSDMAHSALGDMETIQQSAAPKAKLSLLESENFFTDFSLYG 359
                        ***************.*****:****:********:.*** *** **.:.****:****
DanaPB2_ArfGAP3-PC    297 MDPAKAKQMERLGMGFSLRGSDVAHSAIGDMETIQQTVAPKNKLSSLENDSFFTDYSLYG 356

Dmel_ArfGAP3-PC      360 NS--SSGGGG-GGSSEKRESSVGGTSKLDKFELDALGYETIEPIGGSHSNITSMFSRSND 416
                        **  :**.** ****:****   .:*********************:*.******* . *
DanaPB2_ArfGAP3-PC    357 NSPAASGNGGVGGSSDKRESIETSSSKLDKFELDALGYETIEPIGGTHGNITSMFSSNYD 416

Dmel_ArfGAP3-PC      417 YDKPKTSAPVKKNSGS----SQTHTKG-GTSTDPVIAQQKFGNSKGFGSDQYFASEQSSA 471
                         :***:*.*.*..***     ***  *. .:*.************************:*
DanaPB2_ArfGAP3-PC    417 SEKPKSSPPAKSSSGSASSVSQTTGKNKNSSNDPVIAQQKFGNSKGFGSDQYFASEQSAA 476

Dmel_ArfGAP3-PC      472 DVSASLNRFQGSRAISSSDYFGDGSPGGTGGNRAS---SVNFSAPDLDVESVKESVRQGV 528
                        *:**.*********************.*:.*.*..    ********************
DanaPB2_ArfGAP3-PC    477 DISANLNRFQGSRAISSSDYFGDGSPAGSSGSRGGYSPSVNFSAPDLDVESVKESVRQGV 536

Dmel_ArfGAP3-PC      529 HKVAGRLSNLANDVMTSWQDKYGY 552
                        ************************
DanaPB2_ArfGAP3-PC    537 HKVAGRLSNLANDVMTSWQDKYGY 560
```

## 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.



## 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

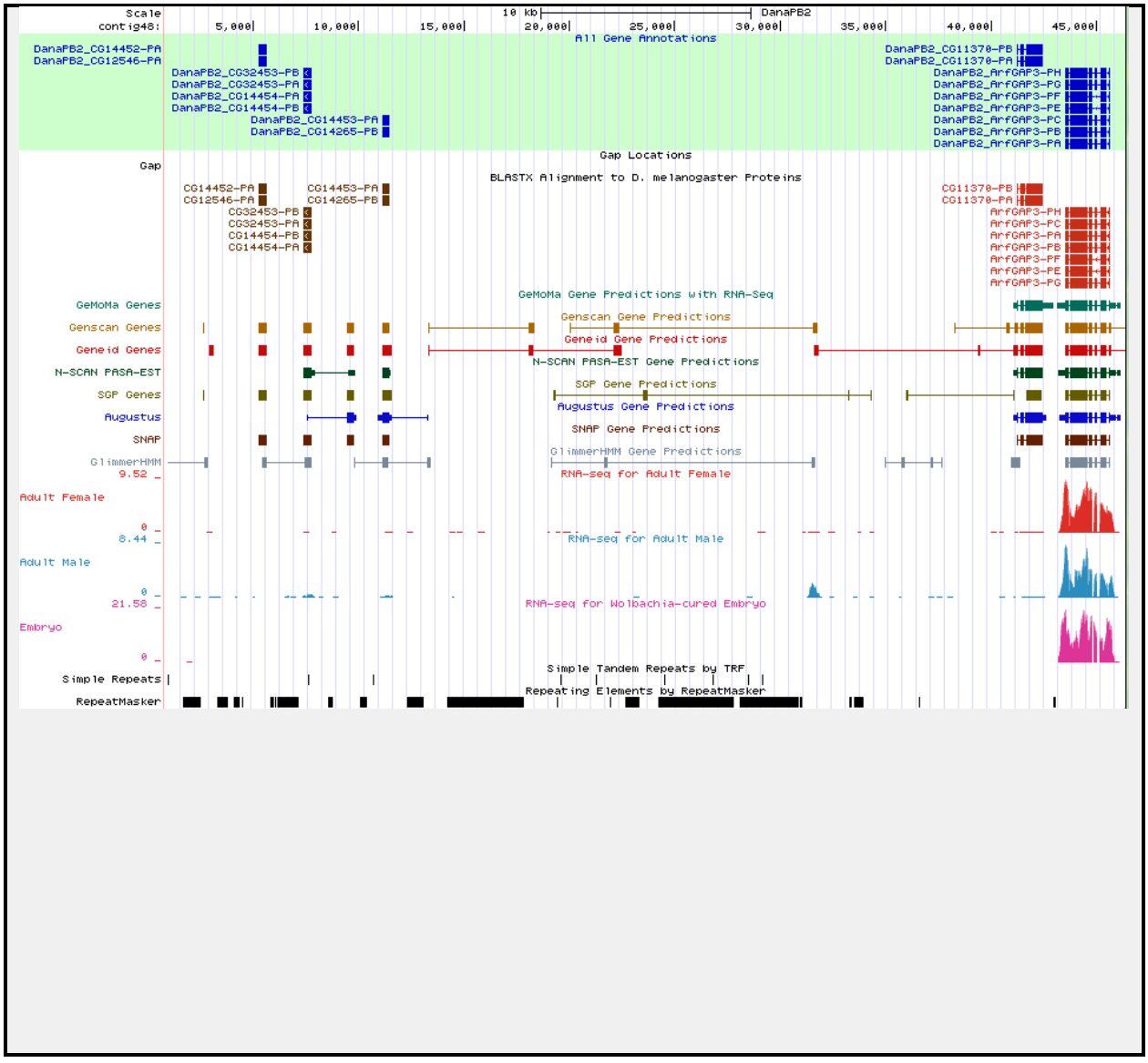
3. Perform a NCBI BLASTP search of the predicted protein sequence from Genscan against the "**Reference proteins (refseq_protein)**" database. Report the significant matches (E-value < 1e-5) to curated RefSeq gene models:
    - Protein records curated by the NCBI RefSeq database have the prefix "**NP_**"

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:

## 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.**



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.