

2015

Annotation and Transcription Start Site Analysis of contig70 in *Drosophila biarmipes*

Robin Wolschendorf
Grand Valley State University

Follow this and additional works at: <http://scholarworks.gvsu.edu/honorsprojects>



Part of the [Molecular Biology Commons](#)

Recommended Citation

Wolschendorf, Robin, "Annotation and Transcription Start Site Analysis of contig70 in *Drosophila biarmipes*" (2015). *Honors Projects*. 385.
<http://scholarworks.gvsu.edu/honorsprojects/385>

This Open Access is brought to you for free and open access by the Undergraduate Research and Creative Practice at ScholarWorks@GVSU. It has been accepted for inclusion in Honors Projects by an authorized administrator of ScholarWorks@GVSU. For more information, please contact scholarworks@gvsu.edu.

HONORS SENIOR PROJECT

Annotation and Transcription
Start Site Analysis of contig70 in
Drosophila biarmipes

Robin Wolschendorf
Grand Valley State University
Frederik Meijer Honors College
Winter 2015

Table of Contents

Introduction.....	3
GEP Annotation Report.....	4
Project Details.....	4
dbia_wnd.....	5
TSS for dbia_wnd.....	12
dbia_lush.....	27
TSS for dbia_lush.....	32
dbia_CG9372.....	42
TSS for dbia_CG9372.....	46
dbia_CG9376.....	51
TSS for dbia_CG9376.....	55
dbia_Lon.....	61
TSS for dbia_Lon.....	68
dbia_asf1.....	73
TSS for dbia_asf1.....	76
dbia_ms(3)76Cc.....	81
TSS for dbia_ms(3)76Cc.....	85
dbia_l(3)76BDm.....	89
TSS for dbia_l(3)76BDm.....	93
Annotation Files Merger.....	98
Other predicted genes.....	99

Introduction

The following is my annotation and transcription start site analysis report for contig70 in *D. biarmipes* to be submitted to the Genomics Education Partnership (GEP) at Washington University in St. Louis. The report was completed in accordance with the guidelines and template set forth by the GEP. There are eight complete genes found in contig70, for which I found the protein coding exon boundaries as well as the transcription start sites for all isoforms. All conclusions and reasoning for them are outlined in the report.

The GEP does genomics research, specifically comparing the genomes of all species of *Drosophila*. Students completing projects for the GEP analyze a portion of a genome of one of the species. The GEP collects all of this information and data and uses it for evolutionary analysis. Comparing the genomes of distinct species illuminates important regions of DNA, which are well conserved between species.

Students in CMB 440: Research Applications of *Drosophila* Genomics are expected to complete GEP reports for their individual projects. I took CMB 440 in Winter 2014, so this is the second project and report I have done. I decided to complete another report for my Honors Senior Project to assist Dr. Martin Burg in advancing the scope of the course. Before this semester, students were not expected to complete transcription start site (TSS) analysis for their projects. To ensure a smoother integration of TSS analysis, Dr. Burg asked me to become well versed in the process so I could teach it to students. I studied GEP materials and instructions over the course of the semester to compile a comprehensive instructional presentation on TSS analysis to give to the class. Along with the presentation, I made myself available once a week in class for students to ask me questions concerning TSS analysis in their individual projects.

- Robin Wolschendorf

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: Robin Wolschendorf
Student email: wolscher@mail.gvsu.edu
Faculty Advisor: Dr. Martin Burg
College/University: Grand Valley State University

Project details

Project name: contig70
Project species: *Drosophila biarmipes*
Date of submission: 4/16/2015
Size of project in base pairs: 40,000
Number of genes in project: 8

Does this report cover all genes and all isoforms or is it a partial report? Yes, all genes and isoforms

If this is a partial report because different students are working on different regions of this sequence, please report the region of the project covered by this report:

from base _____ to base _____

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 conclusions:

1. Perform a BLASTX search of the entire contig sequence against the non-redundant (*nr*) protein 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 BLASTP search using the predicted amino acid sequence against the protein database (*nr*) using the strategy described above.
3. Examine the gene expression tracks (e.g. cDNA/EST/RNA-Seq) for evidence of transcribed regions that do not correspond to alignments to known *D. melanogaster* proteins. Perform a BLASTX search against the *nr* database using these genomic regions to determine if the region is similar to any 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. Be sure to create enough Isoform Report Forms within your Gene Report Form for all isoforms.

Gene report form

Gene name (i.e. *D. mojavensis eyeless*): *D. biarmipes wnd*

Gene symbol (i.e. *dmoj_ey*): dbia wnd

Approximate location in project (from 5' end to 3' end): 4693-14366

Number of isoforms in *D. melanogaster*: 4

Number of isoforms in this project: 4

Complete the following table for all the isoforms in this project:

Name(s) of unique isoform(s) based on coding sequence	List of isoforms with identical coding sequences
wnd-PC	wnd-PA, wnd-PB
wnd-PD	

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 your project consensus sequence:

All 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 which supports the hypothesis of errors in the consensus sequence include: CDS alignment with frame-shifts or in-frame stop codons, multiple RNA-seq reads with discrepant alignments compared to the project sequence, multiple high quality discrepancies in the *Consed* assembly.

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

Using the Sequencer Updater (available through the GEP web site under “Projects” -> “Annotation Resources”), create a VCF (Variant Call Format) file that describes the changes to the consensus sequence you have identified above. **Paste a screenshot with the list of sequence changes below:**

Isoform report form

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

Gene-isoform name (i.e. dmoj_ey-PA): dbia wnd-PC

Names of the isoforms with identical coding sequences as this isoform
dbia wnd-PA, dbia wnd-PB

Is the 5' end of this isoform missing from the end of project: No

If so, how many exons are missing from the 5' end: _____

Is the 3' end of this isoform missing from the end of the project: No

If so, how many exons are missing from the 3' end: _____

1. 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 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 revise the submitted exon coordinates automatically using this VCF file.

The screenshot shows the Gene Model Checker web interface. On the left, the 'Configure Gene Model' section includes fields for 'Fosmid Sequence File' (C:\fakepath\contig70.fasta), 'Errors in Consensus Sequence?' (No), 'Ortholog in D. melanogaster:' (wnd-PC), 'Coding Exon Coordinates:' (4693-4850, 9739-10227, 10680-10871, 10934-11587, 11650-11800, 13009-13910, 13970-14363), 'Annotated Untranslated Regions?' (No), 'Orientation of Gene Relative to Query Sequence:' (Plus), 'Completeness of Gene Model Translation:' (Complete), 'Stop Codon Coordinates:' (14364-14366), 'Project Group:' (D. biarmipes 3L Control), and 'Project Name:' (contig70). On the right, the 'Checklist' tab is active, displaying a table of criteria and their status.

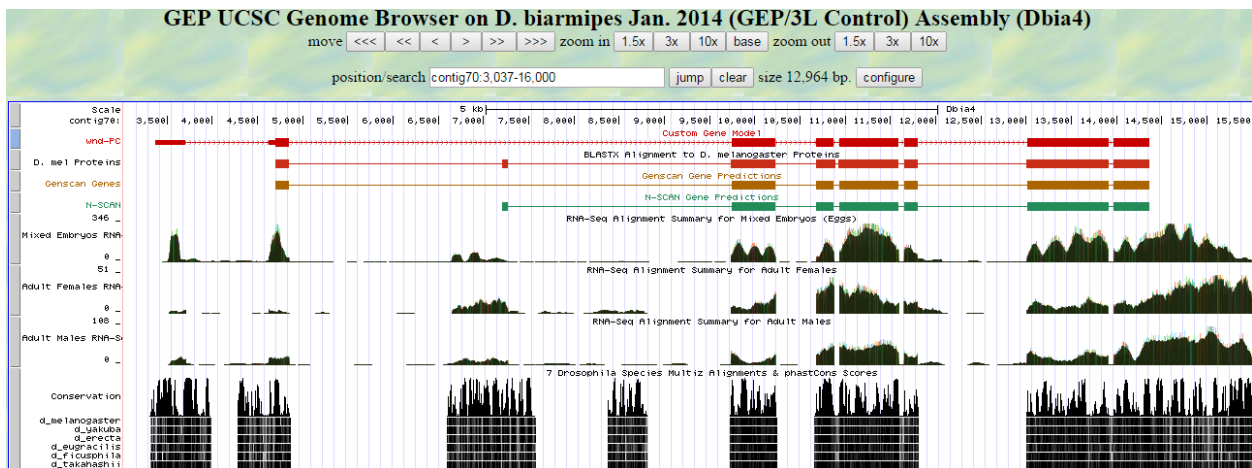
View	Criteria	Status	Message
	Check for Start Codon	Pass	
	Acceptor for CDS 1	Skip	Already checked for Start Codon
	Donor for CDS 1	Pass	
	Acceptor for CDS 2	Pass	
	Donor for CDS 2	Pass	
	Acceptor for CDS 3	Pass	
	Donor for CDS 3	Pass	
	Acceptor for CDS 4	Pass	
	Donor for CDS 4	Pass	
	Acceptor for CDS 5	Pass	
	Donor for CDS 5	Pass	
	Acceptor for CDS 6	Pass	
	Donor for CDS 6	Pass	
	Acceptor for CDS 7	Pass	
	Donor for CDS 7	Skip	Already checked for Stop Codon
	Check for Stop Codon	Pass	
	Additional Checks	Pass	
	Number of coding exons matched D. melanogaster or...	Pass	

2. View the gene model on the Genome Browser

Using the custom track feature from the Gene Model Checker (see page 10 of the Gene Model Checker user guide on how to do this; you can find the guide under “Help” -> “Documentations” -> “Web Framework” on the GEP website at <http://gеп.wustl.edu>). 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. Include the following evidence tracks in the screenshot if they are available.

1. A sequence alignment track (D. mel Protein 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, 3-way, 5-way or 7-way multiz)

Paste the screenshot of your gene model as shown on the Genome Browser 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 use the protein alignment generated by the Gene Model Checker or you can generate a new alignment using BLAST 2 Sequences (*bl2seq*). **Paste a screenshot of the protein alignment below:**

Alignment of wnd-PC vs. Submitted_Seq

[View plain text version](#)

Identity: 923/981 (94.1%), Similarity: 943/981 (96.1%), Gaps: 5/981 (0.5%)

```

wnd-PC      1  MQPFSDSLSKSRDLDLVT---AASRQQQLHGRRRHHGSPNLGLDQTONLRREMaCLQDE 57
Submitted_Seq 1  MQPFSDSLSKSRDLDLVNVAASRQQHQLYGRRRHHGSPNLGLDQTONLRREMaCLQDE 60

wnd-PC      58  FGHGTAAWDQAFSSDSDSPRLQHNNYAEIWDSSAENCCQAPFVGGAGAGGHE 117
Submitted_Seq 61  FGHGTASADAPFKSPDLSSPRLQHNNYAEIWDSSAENCCQAPFVGGAGAGGHE 120

wnd-PC      118  DKPIGMVYGLCCMKPVLSFSGKTVIEVKSQRSEDDWIPFSEITELEWLGSGAGAVFE 177
Submitted_Seq 121  DKPIGMVYGLCCMKPVLSFSGKTVIEVKSQRSEDDWIPFSEITELEWLGSGAGAVFE 180

wnd-PC      178  SRLKNSFVAVKVKELKEVDIHKLRKLDHENITKFGVCTQSPVFCIIMEPCPYQLONI 237
Submitted_Seq 181  SRLKNSFVAVKVKELKEVDIHKLRKLDHENITKFGVCTQSPVFCIIMEPCPYQLONI 240

wnd-PC      238  LKKEQVMLPSRLVSWSKQIALGMOYLHSHKIHRDLKSPNLLISTNEVVKISDFQTSREH 297
Submitted_Seq 241  LKKEQVMLPSRLVSWSKQIALGMOYLHSHKIHRDLKSPNLLISTNEVVKISDFQTSREH 300

wnd-PC      298  NEISPKMSFAGVAMADAEVIRNEPCEKVDINSYGVVLEMLTCEIPYKDVDSALITWG 357
Submitted_Seq 301  NEISPKMSFAGVAMADAEVIRNEPCEKVDINSYGVVLEMLTCEIPYKDVDSALITWG 360

wnd-PC      358  VGNNSIKLVPSCTCPGFKLLVYLVKCFKRNRPFRQILSHDLAGPELAKTKYQVFE 417
Submitted_Seq 361  VGNNSIKLVPSCTCPGFKLLVYLVKCFKRNRPFRQILSHDLAGPELAKTKYQVFE 420

wnd-PC      418  FQKSWKEVRSHEIKETIQNGNIIHKVQDLIKRRTAEWRHQAQIRMVYEDKIAKTKQVFE 477
Submitted_Seq 421  FQKSWKEVRSHEIKETIQNGNIIHKVQDLIKRRTAEWRHQAQIRMVYEDKIAKTKQVFE 480

wnd-PC      478  FLSQCHSQIQKKEKETAEERKLPQSGVYKPNRPFONTIRKMOHVRRLNPAALQQQSS 537
Submitted_Seq 481  FLSQCHSQIQKKEKETAEERKLPQSGVYKPNRPFONTIRKMOHVRRLNPAALQQQSS 539

wnd-PC      538  FPDRETTPEFVVKCMIAQLDSNCPKSYLANITPSSGLGAMPNKNKVFRRHRNAGGS 597
Submitted_Seq 540  FPDRETTPEFVVKCMIAQLDSNCPKSYLANITPSSGLGAMPNKNKVFRRHRNAGGS 599

wnd-PC      598  FGAPPKYSFPRDRBYSEPEPKRVQVLEBQVQDAMDVSEDTISPSAEAPRSQPIDVPE 657
Submitted_Seq 600  FGAPPKYSFPRDRBYSEPEPKRVQVLEBQVQDAMDVSEDTISPSAEAPRSQPIDVPE 659

wnd-PC      658  NHRQDFLQARVQITAAQARARSGSFSALGAVNPACPSNGNSISSELYQDACSSPD 717
Submitted_Seq 660  NHRQDFLQARVQITAAQARARSGSFSALGAVNPACPSNGNSISSELYQDACSSPD 719

wnd-PC      718  QYIDDVMNSNERLDMTECCSDNENLERLGRKVIETINENRSLIQSNTNSVSNADNGGCG 776
Submitted_Seq 720  QYIDDVMNSNERLDMTECCSDNENLERLGRKVIETINENRSLIQSNTNSVSNADNGGCG 779

wnd-PC      777  ASPLRLRSGNSPGLSRGSSSHKRRKHYLGNPNCGSSIGHANGSHEDQSWSDREGEV 836
Submitted_Seq 780  ASPLRLRSGNSPGLSRGSSSHKRRKHYLGNPNCGSSIGHANGSHEDQSWSDREGEV 839

wnd-PC      837  DYKVALRRRSICRQPIARGMRPRRSYKAPLSQKIAIHKRNVIIVSDEBENTSEYSHSPSS 896
Submitted_Seq 840  DYKVALRRRSICRQPIARGMRPRRSYKAPLSQKIAIHKRNVIIVSDEBENTSEYSHSPSS 899

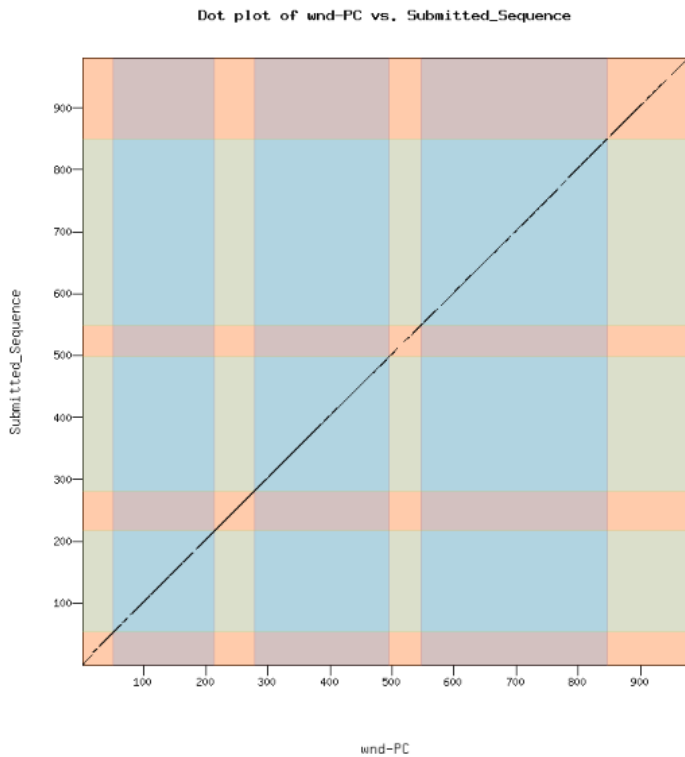
wnd-PC      897  QHSTLESNDIADMKKTQATSTGSSYSEPEDDSSSDDEQGNRPTEAKAEGPALPMGA 956
Submitted_Seq 900  QHSTLESNDIADMKKTQATSTGSSYSEPEDDSSSDDEQGNRPTEAKAEGPALPMGA 959

wnd-PC      957  VRSSDIISIPTEADGAVNMV 977
Submitted_Seq 960  VRSSDIISIPTEADGAVNMV 980
    
```

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). **Provide an explanation for any anomalies** on the dot plot (e.g. large gaps, regions with no sequence similarity).

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



Isoform report form

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

Gene-isoform name (i.e. dmoj_ey-PA): dbia wnd-PD

Names of the isoforms with identical coding sequences as this isoform

Is the 5' end of this isoform missing from the end of project: No

If so, how many exons are missing from the 5' end: _____

Is the 3' end of this isoform missing from the end of the project: No

If so, how many exons are missing from the 3' end: _____

1. 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 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 revise the submitted exon coordinates automatically using this VCF file.

Gene Model Checker

Configure Gene Model

Model Details

Fosmid Sequence File: C:\fakepath\contig70.fasta

Errors in Consensus Sequence? Yes No

Ortholog in *D. melanogaster*: wnd-PD

Coding Exon Coordinates: 7202-7269, 9739-10227, 10680-10871, 10934-11587, 11650-11800, 13009-13910, 13970-14363

Annotated Untranslated Regions? Yes No

Orientation of Gene Relative to Query Sequence: Plus Minus

Completeness of Gene Model Translation: Complete Partial

Stop Codon Coordinates: 14364-14366

Project Details

Project Group: *D. biarmipes* 3L Control

Project Name: contig70

Checklist

View	Criteria	Status	Message
<input type="checkbox"/>	Check for Start Codon	Pass	
<input type="checkbox"/>	Acceptor for CDS 1	Skip	Already checked for Start Codon
<input type="checkbox"/>	Donor for CDS 1	Pass	
<input type="checkbox"/>	Acceptor for CDS 2	Pass	
<input type="checkbox"/>	Donor for CDS 2	Pass	
<input type="checkbox"/>	Acceptor for CDS 3	Pass	
<input type="checkbox"/>	Donor for CDS 3	Pass	
<input type="checkbox"/>	Acceptor for CDS 4	Pass	
<input type="checkbox"/>	Donor for CDS 4	Pass	
<input type="checkbox"/>	Acceptor for CDS 5	Pass	
<input type="checkbox"/>	Donor for CDS 5	Pass	
<input type="checkbox"/>	Acceptor for CDS 6	Pass	
<input type="checkbox"/>	Donor for CDS 6	Pass	
<input type="checkbox"/>	Acceptor for CDS 7	Pass	
<input type="checkbox"/>	Donor for CDS 7	Skip	Already checked for Stop Codon
<input type="checkbox"/>	Check for Stop Codon	Pass	
<input type="checkbox"/>	Additional Checks	Pass	
<input type="checkbox"/>	Number of coding exons matched <i>D. melanogaster</i> or...	Pass	

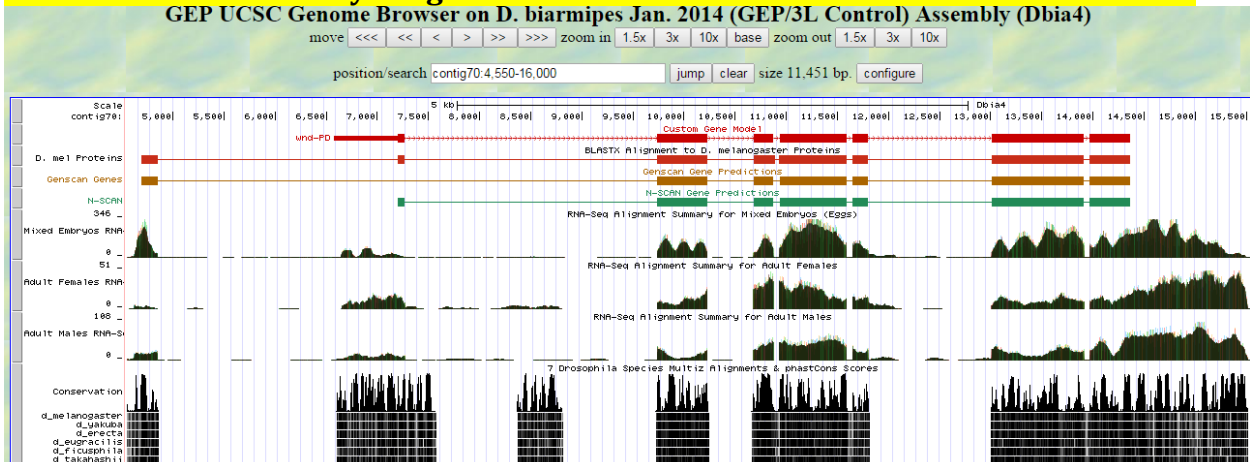
2. View the gene model on the Genome Browser

Using the custom track feature from the Gene Model Checker (see page 10 of the Gene Model Checker user guide on how to do this; you can find the guide under “Help” -> “Documentations” -> “Web Framework” on the GEP website at <http://gep.wustl.edu>).

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

1. A sequence alignment track (*D. mel* Protein 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, 3-way, 5-way or 7-way multiz)

Paste the screenshot of your gene model as shown on the Genome Browser below:
GEP UCSC Genome Browser on *D. biarmipes* Jan. 2014 (GEP/3L Control) Assembly (Dbia4)



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 use the protein alignment generated by the Gene Model Checker or you can generate a new alignment using BLAST 2 Sequences (*bl2seq*). **Paste a screenshot of the protein alignment below:**

Alignment of wnd-PD vs. Submitted_Seq

[View plain text version](#)

Identity: 896/951 (94.2%), Similarity: 916/951 (96.3%), Gaps: 2/951 (0.2%)

```
wnd-PD      1  WYVITPFFSHNDPDMILITKERSMAGIQDFPCHLTAATDLPFKSSDLDSPPLQHNHY 60
Submitted_Seq 1  WYVITPFFSHNDPDMILITKERSMAGIQDFPCHLTAATDLPFKSSDLDSPPLQHNHY 60

wnd-PD      61  ASITDSSAENFCQQRPPHVGCGAAGPGRDQPIGWYGLZCCHKPVSFIGKTVIEVK 120
Submitted_Seq 61  ASITDSSAENFCQQRPPHVGCGAAGPGRDQPIGWYGLZCCHKPVSFIGKTVIEVK 120

wnd-PD      121  SQRSRDWQIPFESITLWLSGAGQAVFSGRLKNETVAVKVKELKFTDIKLRKLOHE 180
Submitted_Seq 121  SQRSRDWQIPFESITLWLSGAGQAVFSGRLKNETVAVKVKELKFTDIKLRKLOHE 180

wnd-PD      181  NIIKFKGVCTQSPVFCIIIEFCPYGFLQNILKEBQVMLPSRLVSNKQIALGMQYLHSK 240
Submitted_Seq 181  NIIKFKGVCTQSPVFCIIIEFCPYGFLQNILKEBQVMLPSRLVSNKQIALGMQYLHSK 240

wnd-PD      241  IHRDLKSNLILISWVVKISDFGSRFNEISTKMSFACTVAMMADVIRNEPCEKV 300
Submitted_Seq 241  IHRDLKSNLILISWVVKISDFGSRFNEISTKMSFACTVAMMADVIRNEPCEKV 300

wnd-PD      301  DINSYQVWLEMDGCEFTYQVDSNATLWGVNNSLKHVYSPQCEGKELAVTQVTSK 360
Submitted_Seq 301  DINSYQVWLEMDGCEFTYQVDSNATLWGVNNSLKHVYSPQCEGKELAVTQVTSK 360

wnd-PD      361  RNRSPFQILSHDIAQPELLRKRKEQYFEPQKSWKEVRSKHEITONGTIRHKVEQDI 420
Submitted_Seq 361  RNRSPFQILSHDIAQPELLRKRKEQYFEPQKSWKEVRSKHEITONGTIRHKVEQDI 420

wnd-PD      421  IKRRAWRHHAQDIRMVFEEKIQKIQVFFLSECMQIQEKKEIARERKRLPGQVKE 480
Submitted_Seq 421  IKRRAWRHHAQDIRMVFEEKIQKIQVFFLSECMQIQEKKEIARERKRLPGQVKE 480

wnd-PD      481  NRREGNTIRKMQHYRRLNAPAAIQQSTTPDPETTPSPVKCMLYAQLDSCQPKSVY 540
Submitted_Seq 481  NRREGNTIRKMQHYRRLNAPAAIQQSTTPDPETTPSPVKCMLYAQLDSCQPKSVY 540

wnd-PD      541  ANITPSSGLGAMPNKNKVFRRHRNASGSPAPPKYSPTDRRYQSEPNRKVQLVERQ 600
Submitted_Seq 541  ANITPSSGLGAMPNKNKVFRRHRNASGSPAPPKYSPTDRRYQSEPNRKVQLVERQ 600

wnd-PD      601  QYDAMDVSEFDISPAEAPRSQPIDVDFVHRHQLPLQLRVOKIAQAQARASGTSSEA 660
Submitted_Seq 601  QYDAMDVSEFDISPAEAPRSQPIDVDFVHRHQLPLQLRVOKIAQAQARASGTSSEA 660

wnd-PD      661  AGAVNPACPSNGNSLSTSELFYQDACSSFPQIIDVMNSNERLDITECCSDNENLERLGR 720
Submitted_Seq 661  AGAVNPACPSNGNSLSTSELFYQDACSSFPQIIDVMNSNERLDITECCSDNENLERLGR 720

wnd-PD      721  XVIEPIENRSLTOSNTNSVSNADNGGG-ASPIELRESGNSPCLSRCSSTHSKRKHPI 779
Submitted_Seq 721  XVIEPIENRSLTOSNTNSVSNADNGGG-ASPIELRESGNSPCLSRCSSTHSKRKHPI 779

wnd-PD      780  SDNPNCGSSIGNANGSHEQDSWDEEGETTDYKYALRRBSIGROPIARGMRPRRSYKAPL 839
Submitted_Seq 780  SDNPNCGSSIGNANGSHEQDSWDEEGETTDYKYALRRBSIGROPIARGMRPRRSYKAPL 839

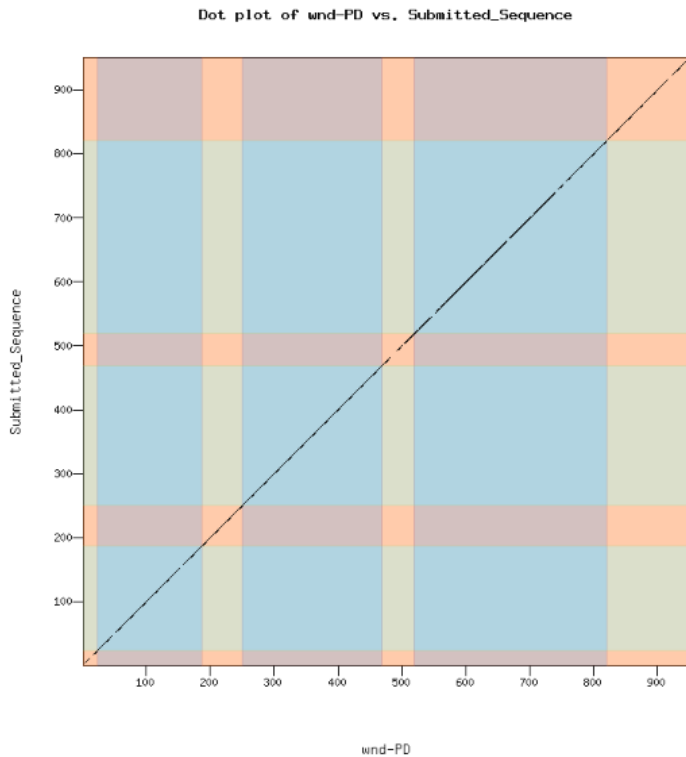
wnd-PD      840  SQKIAIHKRNVIIVSDEEENTSEYSHSPSQHSTLESNPDADMKKTQATSTSNVSEPE 899
Submitted_Seq 840  SQKIAIHKRNVIIVSDEEENTSEYSHSPSQHSTLESNPDADMKKTQATSTSNVSEPE 899

wnd-PD      900  EDDSDSDSDEEQNRPTAKAEGPALPMGAVRSDIISIPTEADGAVNMV 950
Submitted_Seq 900  EDDSDSDSDEEQNRPKATVEQALPMGAVRSDIISIPTEADGAVNMV 950
```

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). **Provide an explanation for any anomalies** on the dot plot (e.g. large gaps, regions with no sequence similarity).

Note: Large vertical and horizontal gap near exon boundaries in the dot plot often indicates that an incorrect splice site might have been picked. Please re-examine these regions and provide a detail 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 specified above. This section is OPTIONAL and you do not need to complete this section to submit the project.

Name(s) of isoform(s) with unique TSS	List of isoforms with identical TSS
wnd-PD	

Complete this report form for each unique TSS listed in the table above (copy and paste to create as many copies of TSS report form as needed):

Gene-isoform name (i.e. dbia_ey-RA): dbia_wnd-PD

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

Type of core promoter: (Peaked or Broad): Broad

Coordinates of the first transcribed exon: 6,576 – 7,269

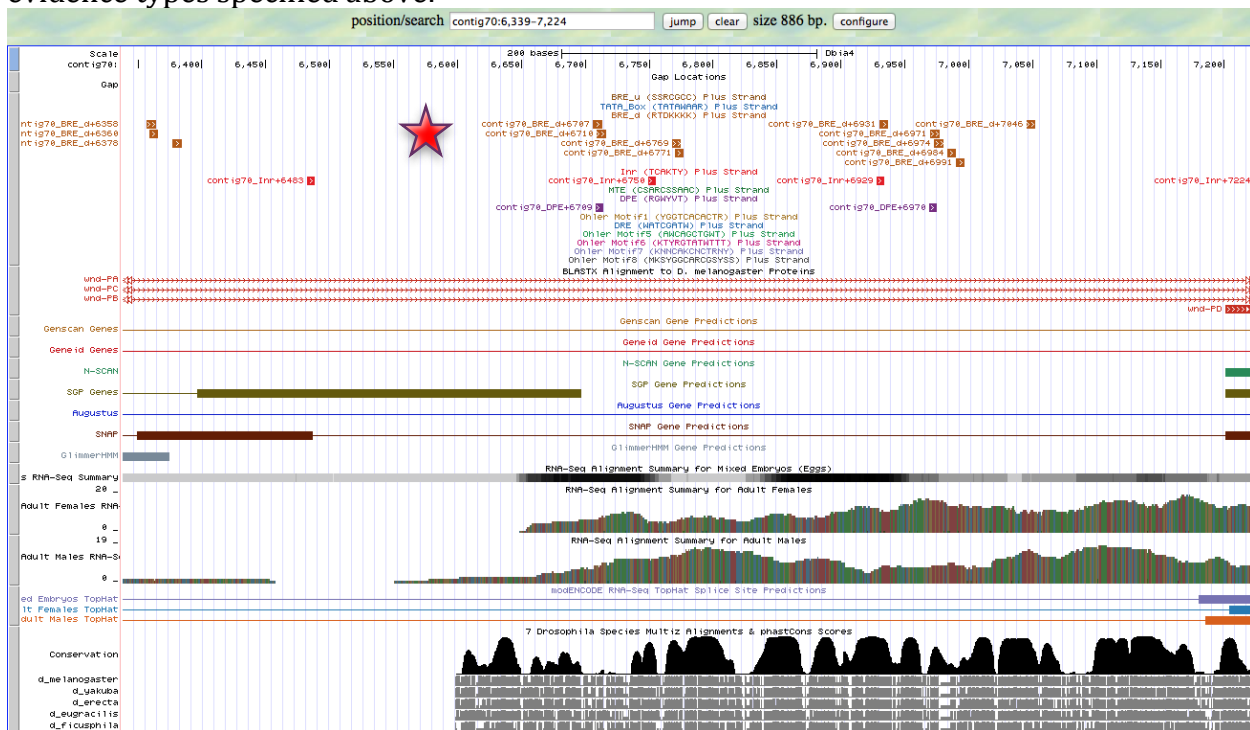
Coordinate(s) of TSS position(s): 6,576
 Coordinate(s) of TSS search region(s): 5,000 – 7,201

1. Evidence that supports the TSS annotation postulated above

Specify the type of evidence used to support the TSS annotation:

Evidence type	Support TSS annotation?	Refute TSS annotation?
blastn alignment of the initial exon from <i>D. melanogaster</i>	X	
RNA-Seq coverage and TopHat splice junctions	X	
Core promoter motifs		X
Sequence conservation with other <i>Drosophila</i> species	X	
Other (please specify)		

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



The core promoters in the search region do not help in predicting a TSS. They neither support my prediction nor strongly support another location. As seen in the figure above, none of the surrounding motifs support my predicted TSS, indicated by the red star.

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 below:**

Dbia4_dna range=contig70:1-40000 5'pad=0 3'pad=0 strand=+ repeatMasking=none
 Sequence ID: lcl|Query_8681 Length: 40000 Number of Matches: 12

Range 1: 6576 to 7269 [Graphics](#)

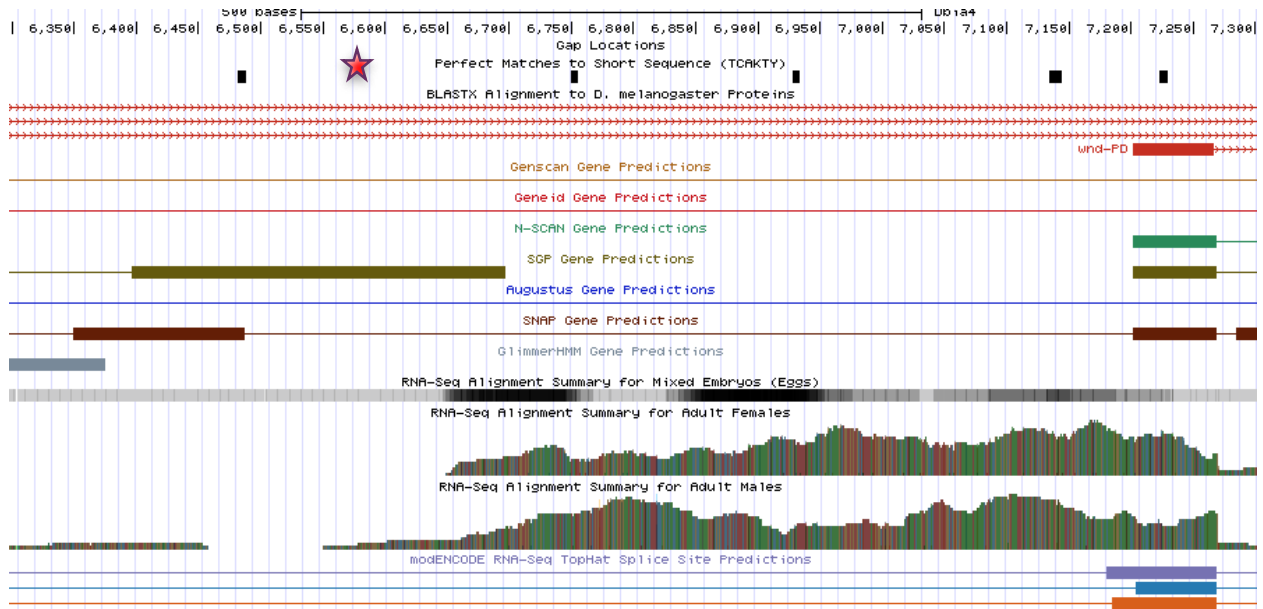
▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
633 bits(442)	0.0	598/724(83%)	41/724(5%)	Plus/Plus
Query 18		CTGATACTGATAATGTGGAGCTAGACCGTTATGCTCACGCT-GCGAGTCGCGATTAGTT		76
Sbjct 6576		CTGATAGTGATAATGGGGAGCTTTACCGCTATGCTCCCGCTCGCGAGTCGCGGTTAGTT		6635
Query 77		GGCGAAAATCAATCGCTTCTGGTGGTGCCTCTTTGCGCTTGAAAATTCGCAAGCCTGCCTT		136
Sbjct 6636		GGCGAAAATCAATCGTATCTGACGGTGCCTCTTCGCGCGTGGAATTCGCGAGCCTGCCTT		6695
Query 137		AATTTCGCTCCGTGTGTTGTGCGACAACAAAAACAAAAGCGAGCTTATCGCCAAATCAG		196
Sbjct 6696		AAAGTGCCTCGTGG--TTGTGGCACAGCAAGAACAACACGGCTTATCGCCAAATCAG		6753
Query 197		TTTGACGCCGTTGTCTGTGTGTTGGTGTGACGCAGTTGTGCGAAAATACGTGAAATTATTG		256
Sbjct 6754		TTTGACAGCCT-GTCTGTGTGTTGGAGTGACGCAGTGG--CGAAAATACGTGAAATCATT		6810
Query 257		CATTTTTCACCCCAACCCAACCAATAAGCAACAGTCTTGAAAAACCGCAACGAAAGT		316
Sbjct 6811		CACTTTTCACCCAGACAGC-----AACCAACTGT-AAAAACCGCAACGAAAGT		6861
Query 317		GTTACGCGTCGCGTTCGTCTCTCTTTTCTGTAATTAATAGCAACAAAAAGCGATACCAA		376
Sbjct 6862		GTTACGCGCTTCCCTCGTCTCT-TTTTCTGTAATTGA---CAACAAAAAGC-ATACCAA		6916
Query 377		CTTACCTTGACGTCACTTTTTTTTATTATATACATCGGACGACAAAAGGCGATTGTTGT		436
Sbjct 6917		CTCACCTTGACGTCACTTTTTTTTATAATAC-ACACATCGGACGACAAAAGGCGTGGTTGT		6975
Query 437		TGTT-----TTTTCGGTGGCTGCTGGTGCTTGGCACGAACTCACAATCAAAAGTTA		487
Sbjct 6976		TGTTGCTCGTTTTTTTATGGTGGCTGCTGGTGCTTGGCACGAACTTAAAATCAAAAGTTA		7035
Query 488		GCCAACTTTTGTGTTGGCCAGTGCCTAAAAACAAAAAAGCAGAACTGCTGGC-GA		546
Sbjct 7036		GCCAACTTTTGTGTTGGCTCGGTGCCTAAACAAAAATAAAGAA---AACTGTCGACCGA		7092
Query 547		AAAGCGAGCGTAAAAATTTGCATTAATGCAGTCTCGTCGCATGAATGAATGAAATCTTA		606
Sbjct 7093		AAGGCGAGCGTAAAAATTTGCATCAATGCAGTCTCGTCGCATGAATGAATGAAATCTTA		7152
Query 607		GTGCGCGACAACCTAATTACGAAAGGGAACCAAAAAAGTGGCAATAACAAATAAACCATGG		666
Sbjct 7153		GTGCGCGACAACCTAATTACGAAAGCAAAC---AAAAGAGG---GACAAATAAACCATGG		7205
Query 667		TCTACATCATACCGTCTTTAGCCACAATGACCCGCCGACATGATCATCACCAAAGAAA		726
Sbjct 7206		TCTACACCATACCGTCTTCAGTCACAATGAGCCGCCGACCCGATCATCACCAAAGAAA		7265
Query 727		AAAG 730		
Sbjct 7266		AAAG 7269		

If the TSS annotation is supported by RNA-Seq or RNA polymerase II data, **paste a Genome Browser screenshot of the region around the TSS (± 2 kb) with the evidence tracks listed below:**

1. Short Match results for the Inr motif (TCAKTY)
2. RNA-Seq Alignment Summary

3. RNA-Seq TopHat



The RNA-seq tracks show good support for my TSS prediction. This is predicted as the 5' UTR is relatively short. My predicted TSS is indicated by the red star in the figure above.

If the TSS annotation is supported by sequence conservation with other *Drosophila* species, **paste a screenshot of the pairwise alignment (e.g. from blastn, matcher) or the multiple sequence alignment (e.g. from clustalw, EvoPrinter, Multiz) below:**

Alignment block 1 of 1 in window, 6600 - 7446, 847 bps

```

B D D. biarmipes accgctatgctcccgcctcgcgagtcgcggttagttggcgaaaatcaatcgatctgacggctcgtctttc
B D D. melanogaster accggtatgctcacgc-tgcgagtcgcgatttagttggcgaaaatcaatcgcttctggtggtcgtcttt
B D D. yakuba accggtatgctcccgcctcgcgagtcgcgatttagttggcgaaaatcaatcgatctgacggctcgtcttt
B D D. erecta accggtatgctcccgcctcgcgagtcgcgatttagttggcgaaaatcaatcgcttctgacggctcgtctct
B D D. eugracilis gccactatgctcccgcctcgcgagttgagatttagttggcgaaaatcaatcgcttctgacggctcgtcttt
B D D. ficusphila atgcatatgctcccgc-----tcgcaattcagttggcgaaaatcgaaagattctgacggctg-tctgt
B D D. takahashii accgctatgctcccgcctcgcgagtcgcgatttagttggcgaaaatcaatcgcttccaacggctcgtcttt

D. biarmipes gcgctggaatttcgagcctgccttaaagtgcgctcgtg--gttgtg----gcac-----agcaag
D. melanogaster gcgcttgaaatttcgaagcctgccttaaattgcgctcgtggttgtg----cgac-----aacaaa
D. yakuba gcgctggaatttcgaagcctgccttaaattgcgctcgtggtggtg----cgacaaaaaaaaaacaaa
D. erecta gcgctggaatttcgaagcctgccttaaatttcgctcagtggttgtg----cgac-----aacaaa
D. eugracilis gcgctggaatttcgagcctatcttaaatttcgctcgtggttgtgagatgcac-----aataac
D. ficusphila gcgctggaatttcgagcagccttaaatacgcctcagtggttgcgtg----agcc-----aacaaa
D. takahashii gcgctggaatttcgagcctgccttaaattacgcgctgatttgtg----tgac-----aacaac

D. biarmipes aacaa-aacacggcttatcgcaaatcagtttgagcc-ctgtctggtggttgagtgacgcagt--ggc
D. melanogaster aacaaaagcg-agcttatcgcaaatcagtttgagccggttgtctggtggttggtggtgacgcagttgtgc
D. yakuba aacaaaagcg-ggcttatcgcaaatcagtttctggtggttctggtggttggtggtgacgcagttgtgc
D. erecta aacaaaagcg-ggcttatcgtcaaatcagtttgagccggttgtctggtggttggtggtgacgcagttgtgc
D. eugracilis aaaa--agtgtgcttatcgcaaatcagtttgcaattggttctggtggttggtggtgacgcagttgtgc
D. ficusphila aacaacagcacagcttatcgcaaatcagtttctactgctgtctggtggttgatggtgacgcagttgtgc
D. takahashii aaaaacaagactgcttatcgcaaatcagtttgagctcgtgtctggtggttggtggtgacgcagttgtgc

D. biarmipes gaaatacgtgaaatcattacactttt-t--cccc-----agacagcaaccaactgtt-aaaaacc
D. melanogaster gaaatacgtgaaattattgacttttt-c--cccccaaacccaaccaataagcaacagctcttgaaaaacc
D. yakuba gaaatacgtgaaattattacactttttac--cccccaaacccaaccaataagcaacagctcttgaaaaacc
D. erecta gaaatacgtgaaattattacacttttt-t--tcccccaaacccaaccaataagcaacagctcttgaaaaacc
D. eugracilis gaaatacgtgaaattattacactttt-c--cccctta-----aaccaataaccaactcttgaaaaacc
D. ficusphila gaaatacgtgaaattattaaactttt-t--tccc-----agccataaccaactcttgaaaaacc
D. takahashii gaaatacgtgaaattattacactttt-ttccacc-----aaccaataatcaactcttgaaaaacc

D. biarmipes gccaacgaaagtgttacgcgcttccctcgtctc-tttttcttgtaattgacaac----aaaagcataacc
D. melanogaster gccaacgaaagtgttacgcgctcgcgttcgtctctcttttcttgtaattaatagcaacaaaagcgataacc
D. yakuba gccaacgaaagtgttacgcgctcgcgttcgtctctctttttcttgtaattaatagcaacaaaagcgataacc
D. erecta accaacgaaagtgttacgcgctcgcgttcgtctctctttttcttgtaattaatagcaacaaaagcgataacc
D. eugracilis gccaacgaaagtgttacgcgctcgcgttcgtcgtctctttttcttgtaattaacgacgacaaaaaccataacc
D. ficusphila gccaacgaaagtgttacgcgctcgcgttcgtctctctttttcttgtaattaa-aacaacaaaaaac-atacc
D. takahashii gccaacgaaagtgttacgcgctcgcgttctgtctctttttcttgtaattaaacaac---aaaaaacacacc

D. biarmipes aactcaccttgacgtca--ttttttataat-acacacatcggacgacaaaaagcgctggttgttgttg
D. melanogaster aacttaccttgacgtcactttttttattat--tatacatcgaagcacaacaaaagcattattatttga----

```

The Multiz alignments somewhat support my TSS prediction as their conservation starts at 6,600.

2. Search for core promoter motifs

Note: The consensus sequences for the Drosophila core promoter motifs are available at: http://gander.wustl.edu/~wilson/core_promoter_motifs.html

Use the "Short Match" functionality in the GEP UCSC Genome Browser to search for each of the core promoter motifs listed below in the region surrounding the TSS (± 300 bp) in your project and in the TSS of the *D. melanogaster* ortholog. (For TSS annotations where you can only define a TSS search region, you should search for the core promoter motifs in the entire TSS search region).

Record the **orientation and the start coordinate** (e.g. +10000) of each motif match below. (Enter "NA" if the motif is not present.)

Core promoter motif	Your project	<i>D. melanogaster</i>
BRE ^u	NA	NA
TATA Box	NA	NA
BRE ^d	+6358, +6360, +6378, +6707, +6710, +6769, +6771	-19632874, -19632876, -19632878, -19632939, -19632941, -19632943, -19633182, -19633269, -19633271, -19633281, -19633283
Inr	+6483, +6750	-19633146, -19632899
MTE	NA	NA
DPE	+6709	-19632862, -19633193, -19633274
Ohler_motif1	NA	NA
DRE	NA	NA
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	NA	NA
Ohler_motif8	NA	NA

Transcription start sites (TSS) report form (optional)

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

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

Complete this report form for each unique TSS listed in the table above (copy and paste to create as many copies of TSS report form as needed):

Gene-isoform name (i.e. dbia_ey-RA): dbia_wnd-PC

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

Type of core promoter: (Peaked or Broad): Broad

Coordinates of the first transcribed exon: 3371-3697

Coordinate(s) of TSS position(s): 3371
 Coordinate(s) of TSS search region(s): 3,071 – 3,671

1. Evidence that supports the TSS annotation postulated above

Specify the type of evidence used to support the TSS annotation:

Evidence type	Support TSS annotation?	Refute TSS annotation?
blastn alignment of the initial exon from <i>D. melanogaster</i>	X	
RNA-Seq coverage and TopHat splice junctions	X	
Core promoter motifs	X	
Sequence conservation with other Drosophila species	X	
Other (please specify)		

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

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 below:**

Dbia4_dna range=contig70:1-40000 5'pad=0 3'pad=0 strand=+ repeatMasking=none
 Sequence ID: lcl|Query_45753 Length: 40000 Number of Matches: 6

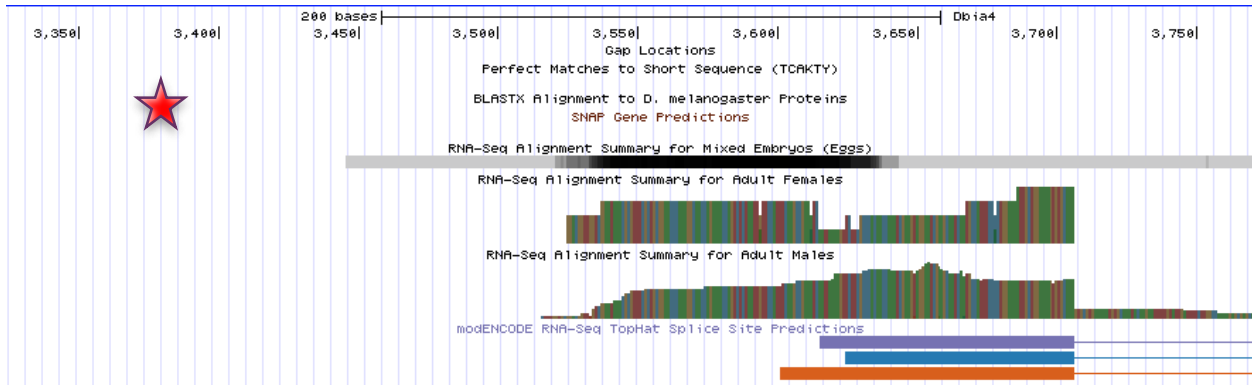
Range 1: 3380 to 3697 [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
278 bits(193)	2e-77	278/351(79%)	39/351(11%)	Plus/Plus
Query 9	TTCTCCCGCACACTTGC	CGCAGTGAACACAACCTGC	GCTCCAAACTGGTTTTGGTGGAGGA	68
Sbjct 3380	TTCTGCCGCACACATGC	CGCAGTG----CAACCTGC	GCTCCAAACTGGTTTCGGGGGAGGA	3435
Query 69	AATCCGCAAGAGAGAGCGCTAGAGAGCGGATTGGA	ACTCGGTTTCGGCCAAAGCCAAAGC	128	
Sbjct 3436	AATCCGCCGAGAGAG-----CGGATTGG-----	TTTCGGCCCAA-----	3468	
Query 129	AGAGCCAGCAGCCAGTTTTT-----GCTTTTTAGTCGATTGTATCTACCTTTGGTGC	CGG	183	
Sbjct 3469	--AGCCAGCAGCCAGTTTTTTTCTGCTTTTTTAGTCGACTTGTATCTACCTTTGGTGC	CGG	3526	
Query 184	ACGGTTGGTCGGAATAAACGCGTTTCGCGAGCGGAACTCCAAGAAGAGCAGAAAAC	CAGTC	243	
Sbjct 3527	ACGGTTGGTCGGAATAAACGCGTTTCGCTGCGGAAACCAAAGGCAACCAGAAAAC	CAGTC	3586	
Query 244	TTTAATTGTTCCATTTCAAGCTGAAATCAAATAAAAGTTCTTCGCCAATGGCTGAAT-AA	302		
Sbjct 3587	TCTAATTGTTCCATTTCAAGCCGAAATCAATTAAGGTTCTTCGCCCAACCAATCGA	3646		
Query 303	AGTACGAATCAATGCAATCACTACGATTTCGCAGTGGAAAAATTGCTGGAAAA	353		
Sbjct 3647	AGTAGGAATCGAAGCAATCAATGCCATTTCCCTGGAAAAATTGCTGGAAAA	3697		

BLASTn predicts the first transcribed exon on wnd-PC to be from about 3380-3697. If the first 8 nucleotides are included, BLAST supports a TSS at 3371.

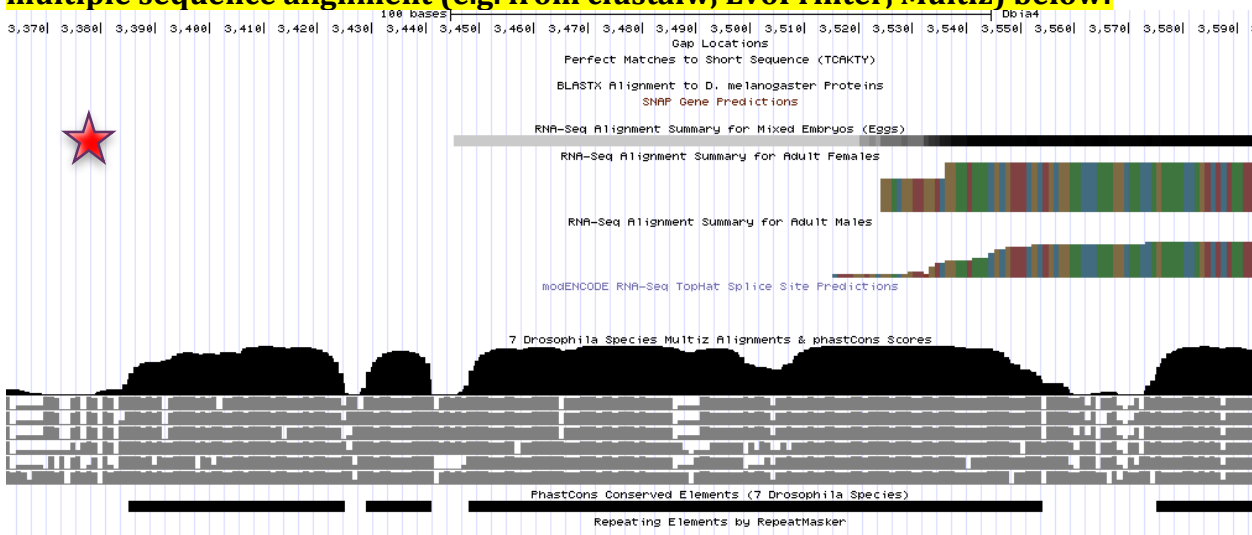
If the TSS annotation is supported by RNA-Seq or RNA polymerase II data, **paste a Genome Browser screenshot of the region around the TSS ($\pm 2\text{kb}$) with the evidence tracks listed below:**

1. Short Match results for the Inr motif (TCAKTY)
2. RNA-Seq Alignment Summary
3. RNA-Seq TopHat



The RNA-Seq alignments do not explicitly support my TSS prediction, but make it plausible. The RNA-Seq data just does not carry far enough into the 5' UTR. The red star in the figure above approximately indicates my TSS prediction.

If the TSS annotation is supported by sequence conservation with other *Drosophila* species, **paste a screenshot of the pairwise alignment (e.g. from blastn, matcher) or the multiple sequence alignment (e.g. from clustalw, EvoPrinter, Multiz) below:**



Alignment block 1 of 1 in window, 3360 - 3630, 271 bps

```

B D D. biarmipes ggctctctcgcgcacactctc-----ttctgccgcacacatgcgagcagtgcaac----ctgcgctcca
B D D. melanogaster cccgc-----cag-tcgc-----gtt-ctcccgcacacttgcgagtg-aacacaacctgcgctcca
B D D. yakuba cccgc-----caa-tcgc-----t-ctctgcacacttgcgagtgcaacgcaacctgcgctcca
B D D. erecta cccgc-----caa-tcgc-----t-ctcccgcacacttgcgagtgcaacgcaacctgcgctcga
B D D. eugracilis ggctc-----tttc-----t-ctcccgcacacatgcgagcagtgcaac----ctgcgctcca
B D D. ficusphila ggctc-----tctc-tcgtgctccccct-ccccgcacgcatgcgccgtacaac----ctgcgctcca
B D D. takahashii ggctctcttgccac-tctc-----t-ctgcccgcacacatgcgagcagtgcaac----ctgcgctcca

D. biarmipes aactggtttcgggggaggaaatccg-----cgagagagcggattgg-----tttcgg
D. melanogaster aactggttttggtggaggaaatccg---caagagagagcgcctagagagcggattggaactcggtttcgg
D. yakuba aactggttttgggggaggaaatccg---caggagagagcgcgagagagcggattggaactcggtttcgg
D. erecta aactggtttaaggggaggaaatccg---caggagagagcgcgagagagcggattggaactcggtttcgg
D. eugracilis aactggtttcaggggaggaaatccg--aaaaagagagagaacgagagagcggattgg-----gattcgg
D. ficusphila aactggtttcaggggaggaaatccgctctcgttttctcgtctctgcccggattgg-----gtttcgg
D. takahashii aactggtttcaggggaggaaatccg-----ccagagagcggattgg-----gtttcgg

D. biarmipes ccc-aa-----agccagcagccagtttttttctgcttttttagtcg-acttgtatctacctttgg
D. melanogaster cca-aagccaaagcagagccagcagccagttttt----gcttttttagtcg-atttgtatctacctttgg
D. yakuba cca-aagccaaagcagagccagcagccagttttt----gcttttttagtcg-atttgtatctacctttgg
D. erecta cca-aagccaaagcagagccagcagccagttttt----gcttttttagtcg-atttgtatctacctttgg
D. eugracilis cccaaa-----agccagcagccagtttttttc-----tttttattttagtctatctacctttgg
D. ficusphila ccc-aa-----agccagcagccagttttt---tgcttttta-----atttgtatctacctttgg
D. takahashii ccc-aa-----agccagcagccagttttttt-tgcttttttagtcg-atttgtatctacctttgg

D. biarmipes tgcggacggttggtcggaataaacgcggttctcgtcgcggaaccacaagcagaaaaccagtccttaa
D. melanogaster tgcggacggttggtcggaataaacgcggttctcgcagcggaaactccaagaagagcagaaaaccagtccttaa
D. yakuba tgcggacggttggtcggaataaacgcggttctcgcagcggaaactcc-agaagaccagaaaaccagtccttaa
D. erecta tgcggacggttggtcggaataaacgcggttctcgcagcggagctccaagaagagcagaaaaccagtccttaa
D. eugracilis tgcggacggttggtcggaataaacgcggttacgctgcggagcccaagaaaagcagaaaaccagtccttaa
D. ficusphila tgcggacggttggtcggaataaacgcggttctcgtcggggtcccaagaaaagcagaaaaccagtccttaa
D. takahashii tgcggacggttggtcggaataaacgcgattctcgcagcggaaaccaca-gacaacctgaaaaccagtccttaa

D. biarmipes ttgttccatttcaagccgaaatcaattaaagtctcttcg
D. melanogaster ttgttccatttcaagctgaaatcaattaaagtctcttcg
D. yakuba ttgttccatttcaagctaaaattaaattaaagtctcttcg
D. erecta ttgttccatttcaagctgaaatcaattaaagtctcttcg
D. eugracilis ttgttcgattcctagctaaaattaaaaacctctcttcg
D. ficusphila ttgttctgttttaagctgaaatcaattaaagtctcttcg
D. takahashii ttgttccatttcaagctgaaatcaattaaagtctcttcg

```

The red line and red star indicate my TSS prediction in the figures above. Sequence conservation shows a significant increase immediately downstream of the predicted TSS. Thereby, the conservation from Multiz alignments supports my prediction.

2. Search for core promoter motifs

Note: The consensus sequences for the *Drosophila* core promoter motifs are available at: http://gander.wustl.edu/~wilson/core_promoter_motifs.html

Use the "Short Match" functionality in the GEP UCSC Genome Browser to search for each of the core promoter motifs listed below in the region surrounding the TSS (± 300 bp) in your project and in the TSS of the *D. melanogaster* ortholog. (For TSS annotations where you can only define a TSS search region, you should search for the core promoter motifs in the entire TSS search region).

Record the **orientation and the start coordinate** (e.g. +10000) of each motif match below. (Enter "NA" if the motif is not present.)

Core promoter motif	Your project	<i>D. melanogaster</i>
---------------------	--------------	------------------------

BRE ^u	NA	NA
TATA Box	NA	NA
BRE ^d	+3179, +3342, +3348, +3453, +3481	-19635635, -19635723, -19635795, -19635799, -19635881, -1963640
Inr	+3129	NA
MTE	NA	NA
DPE	+3622	-19635500
Ohler_motif1	NA	NA
DRE	NA	NA
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	NA	NA
Ohler_motif8	NA	NA

The BRE^d motif at 3348 supports my predicted TSS at 3371.

Transcription start sites (TSS) report form (optional)

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

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

Complete this report form for each unique TSS listed in the table above (copy and paste to create as many copies of TSS report form as needed):

Gene-isoform name (i.e. dbia_ey-RA): dbia_wnd-PA

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

Type of core promoter: (Peaked or Broad): Broad

Coordinates of the first transcribed exon: 3,476 – 3,706

Coordinate(s) of TSS position(s): 3,476

Coordinate(s) of TSS search region(s):

1. Evidence that supports the TSS annotation postulated above

Specify the type of evidence used to support the TSS annotation:

Evidence type	Support TSS annotation?	Refute TSS annotation?
blastn alignment of the initial exon from <i>D. melanogaster</i>	X	
RNA-Seq coverage and TopHat splice junctions	X	
Core promoter motifs	X	
Sequence conservation with other <i>Drosophila</i> species	X	
Other (please specify)		

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

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 below:**

Dbia4_dna range=contig70:1-40000 5'pad=0 3'pad=0 strand=+ repeatMasking=none
Sequence ID: lcl|Query_38867 Length: 40000 Number of Matches: 13

Range 1: 3476 to 3706 [Graphics](#)

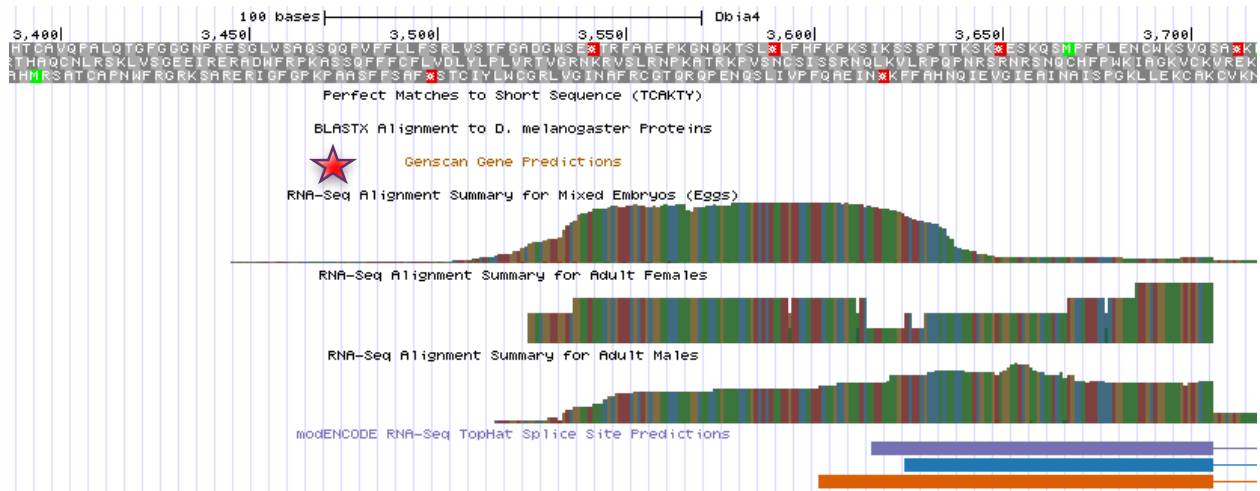
▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
226 bits(157)	4e-62	196/231(85%)	6/231(2%)	Plus/Plus
Query 1	AGCCAGTTTTT-----GCTTTT	TAGTCGATTTGTATCTACCTTTGGTGCGGACGGTTGGT		55
Sbjct 3476	AGCCAGTTTTTTTTCTGCTTTT	TAGTCGACTTGTATCTACCTTTGGTGCGGACGGTTGGT		3535
Query 56	CGGAATAAACGCGTTTCGCAGCGGA	ACTCCAAGAAGAGCAGAAAACCAGTCTTTAATTGT		115
Sbjct 3536	CGGAATAAACGCGTTTCGCTGCGGA	ACCCAAAGGCAACCAGAAAACCAGTCTCTAATTGT		3595
Query 116	TCCATTTCAAGCTGAAATCAAATA	AAAAGTTCTTCGCCAATGGCTGAAT-AAAGTACGAAT		174
Sbjct 3596	TCCATTTCAAGCCGAAATCAATTA	AAAAGTTCTTCGCCCAACAACCAATCGAAGTAGGAAT		3655
Query 175	CAATGCAATCACTACGATTCGCAG	TGGAAAATTGCTGGAAAAGTGAGCAA		225
Sbjct 3656	CGAAGCAATCAATGCCATTTCC	CCTGGAAAATTGCTGGAAAAGTGAGCAA		3706

The BLASTn alignment strongly supports a TSS prediction on 3,476.

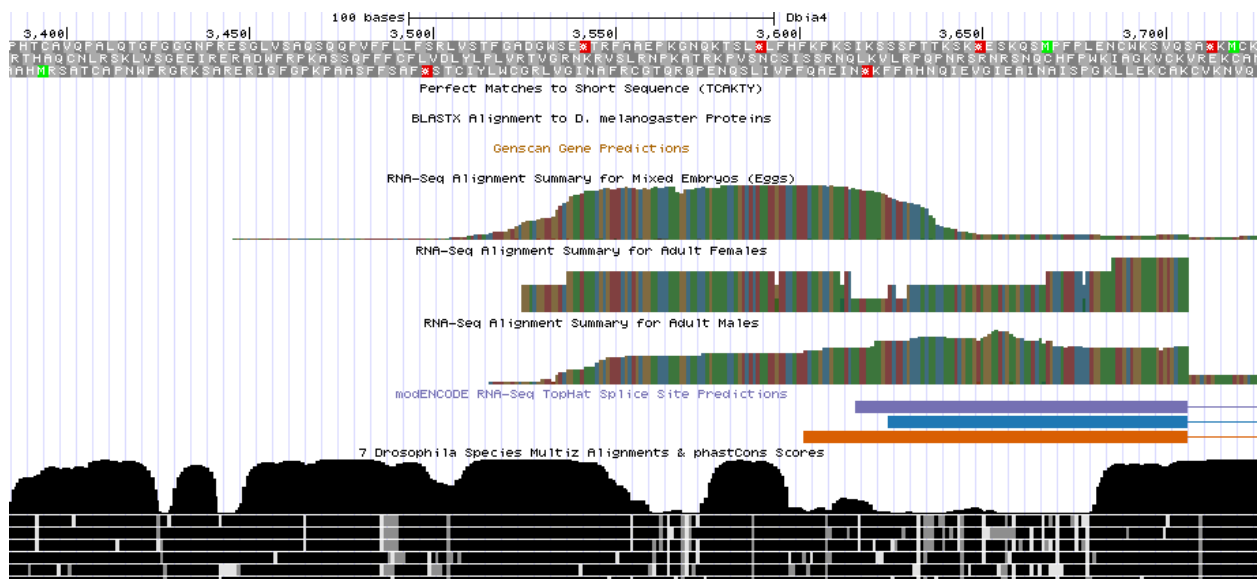
If the TSS annotation is supported by RNA-Seq or RNA polymerase II data, **paste a Genome Browser screenshot of the region around the TSS (± 2 kb) with the evidence tracks listed below:**

1. Short Match results for the Inr motif (TCAKTY)
2. RNA-Seq Alignment Summary
3. RNA-Seq TopHat



The RNA-Seq tracks support my TSS prediction, indicated by the red star in the figure above. The RNA-Seq data seems to pick up soon after the TSS.

If the TSS annotation is supported by sequence conservation with other *Drosophila* species, **paste a screenshot of the pairwise alignment (e.g. from blastn, matcher) or the multiple sequence alignment (e.g. from clustalw, EvoPrinter, Multiz) below:**



The conservation is just overall very good for the entire search region. This is likely due to the fact that the TSS for wnd-PC is just upstream and the TSS for wnd-PB is just downstream.

2. Search for core promoter motifs

Note: The consensus sequences for the *Drosophila* core promoter motifs are available at: http://gander.wustl.edu/~wilson/core_promoter_motifs.html

Use the "Short Match" functionality in the GEP UCSC Genome Browser to search for each of the core promoter motifs listed below in the region surrounding the TSS (± 300 bp) in your

project and in the TSS of the *D. melanogaster* ortholog. (For TSS annotations where you can only define a TSS search region, you should search for the core promoter motifs in the entire TSS search region).

Record the **orientation and the start coordinate** (e.g. +10000) of each motif match below. (Enter "NA" if the motif is not present.)

Core promoter motif	Your project	<i>D. melanogaster</i>
BRE ^u	NA	NA
TATA Box	NA	NA
BRE ^d	+3179, +3342, +3348, +3453, +3481, +3734	-19635415, -19635635, -19635723, -19635795, -19635799, -19635881
Inr	NA	NA
MTE	NA	NA
DPE	+3622	-19635500
Ohler_motif1	NA	NA
DRE	NA	NA
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	NA	NA
Ohler_motif8	NA	NA

The BRD^d motif at +3453 supports my TSS prediction at 3,476.

Transcription start sites (TSS) report form (optional)

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

Name(s) of isoform(s) with unique TSS	List of isoforms with identical TSS
wnd-PB	

Complete this report form for each unique TSS listed in the table above (copy and paste to create as many copies of TSS report form as needed):

Gene-isoform name (i.e. dbia_ey-RA): dbia_wnd-PB
 Names of the isoforms with the same TSS as this isoform:

Type of core promoter: (Peaked or Broad): Broad
 Coordinates of the first transcribed exon: 3,499-3,647
 Coordinate(s) of TSS position(s): 3,499
 Coordinate(s) of TSS search region(s): 3,450-3,600

1. Evidence that supports the TSS annotation postulated above

Specify the type of evidence used to support the TSS annotation:

Evidence type	Support TSS annotation?	Refute TSS annotation?
blastn alignment of the initial exon from <i>D. melanogaster</i>	X	
RNA-Seq coverage and TopHat splice junctions	X	
Core promoter motifs		X
Sequence conservation with other Drosophila species	X	
Other (please specify)		

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

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 below:**

Dbia4_dna range=contig70:1-40000 5'pad=0 3'pad=0 strand=+ repeatMasking=none
 Sequence ID: lcl|Query_37661 Length: 40000 Number of Matches: 6

Range 1: 3499 to 3632 [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
162 bits(112)	5e-43	123/134(92%)	0/134(0%)	Plus/Plus

```

Query 1      AGTCGATTTGTATCTACCTTTGGTGCGGACGTTGGTCGGAATAAACGCGTTTCGCAGCG 60
             ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
Sbjct 3499   AGTCGACTTGTATCTACCTTTGGTGCGGACGTTGGTCGGAATAAACGCGTTTCGCTGCG 3558

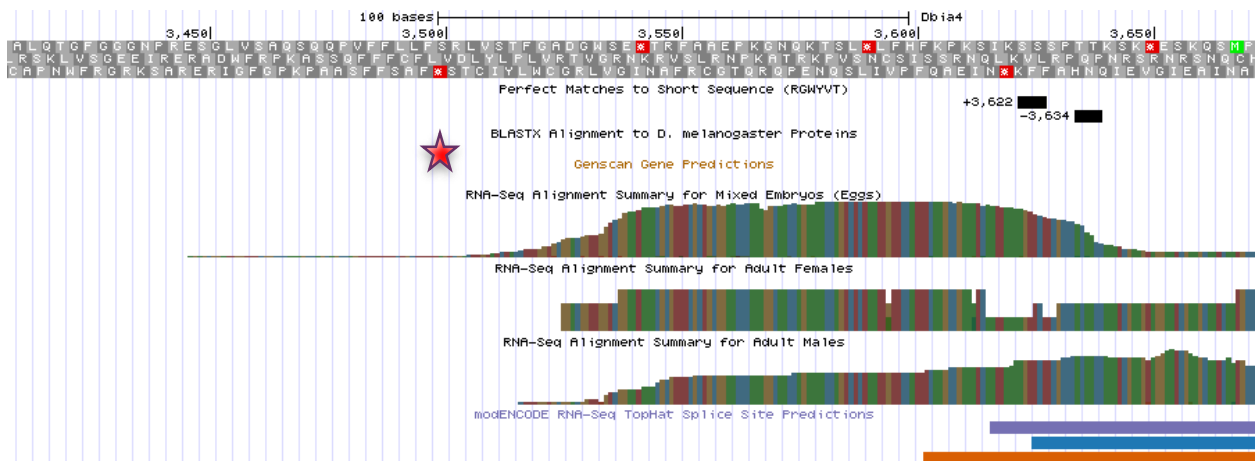
Query 61     GAACTCCAAGAAGAGCAGAAAACCGTCTTTAATTGTTCCATTTCAAGCTGAAATCAAAT 120
             ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
Sbjct 3559   GAACCCAAAGGCAACCAGAAAACCGTCTCTAATTGTTCCATTTCAAGCCGAAATCAATT 3618

Query 121    AAAAGTTCTTCGCC 134
             ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
Sbjct 3619    AAAAGTTCTTCGCC 3632
  
```

The BLASTn alignment shows good support for a TSS at 3,499.

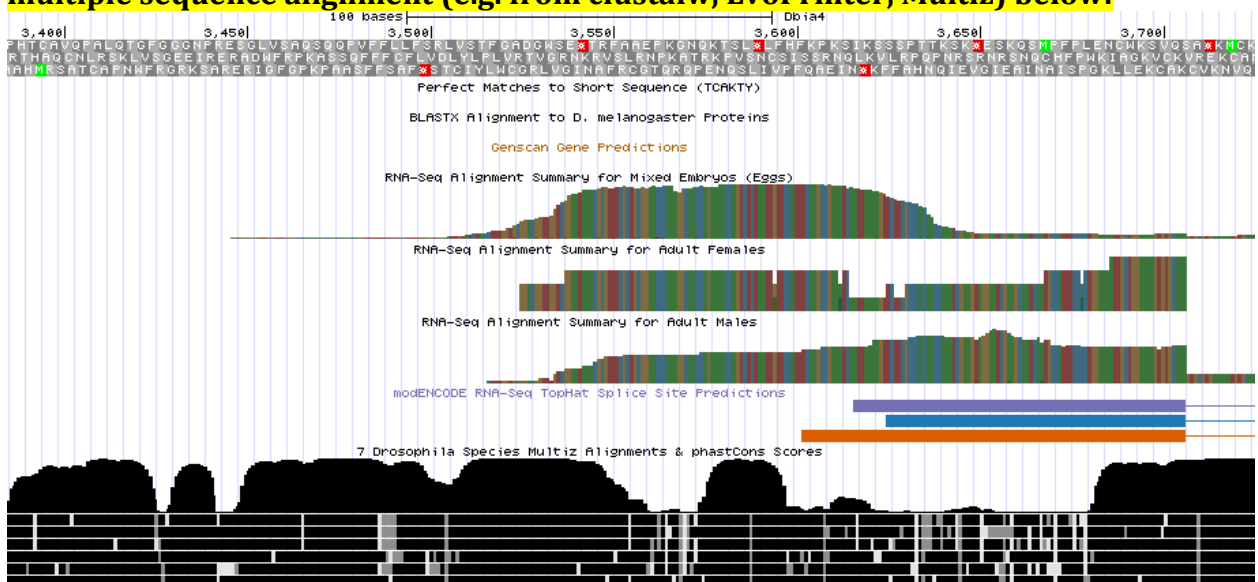
If the TSS annotation is supported by RNA-Seq or RNA polymerase II data, **paste a Genome Browser screenshot of the region around the TSS (± 2 kb) with the evidence tracks listed below:**

1. Short Match results for the Inr motif (TCAKTY)
2. RNA-Seq Alignment Summary
3. RNA-Seq TopHat



The RNA-Seq data supports the TSS prediction, indicated by the red star in the figure above. When comparing to the RNA-Seq of the TSS predictions for wnd-PC and wnd-PA, wnd-PB may be more expressed in adult males than in mixed embryos.

If the TSS annotation is supported by sequence conservation with other *Drosophila* species, **paste a screenshot of the pairwise alignment (e.g. from blastn, matcher) or the multiple sequence alignment (e.g. from clustalw, EvoPrinter, Multiz) below:**



The conservation is just overall very good for the entire search region. This is likely due to the TSS locations for other isoforms in the area.

2. Search for core promoter motifs

Note: The consensus sequences for the *Drosophila* core promoter motifs are available at: http://gander.wustl.edu/~wilson/core_promoter_motifs.html

Use the "Short Match" functionality in the GEP UCSC Genome Browser to search for each of the core promoter motifs listed below in the region surrounding the TSS (± 300 bp) in your project and in the TSS of the *D. melanogaster* ortholog. (For TSS annotations where you can only define a TSS search region, you should search for the core promoter motifs in the entire TSS search region).

Record the **orientation and the start coordinate** (e.g. +10000) of each motif match below. (Enter "NA" if the motif is not present.)

Core promoter motif	Your project	<i>D. melanogaster</i>
BRE ^u	NA	NA
TATA Box	NA	NA
BRE ^d	+3179, +3342, +3348, +3453, +3481, +3734	-19635415, -19635635, -19635723, -19635795, -19635799, -19635881
Inr	NA	NA
MTE	NA	NA
DPE	+3622	-19635500
Ohler_motif1	NA	NA
DRE	NA	NA
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	NA	NA
Ohler_motif8	NA	NA

None of the core promoters in the area support the predicted TSS.

Gene report form

Gene name (i.e. *D. mojavensis eyeless*): *D. biarmipes lush*

Gene symbol (i.e. dmoj_ey): dbia lush

Approximate location in project (from 5' end to 3' end): 38021-38629

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:

Name(s) of unique isoform(s) based on coding sequence	List of isoforms with identical coding sequences
lush-PA	lush-PB

--	--

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 your project consensus sequence:

All 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 which supports the hypothesis of errors in the consensus sequence include: CDS alignment with frame-shifts or in-frame stop codons, multiple RNA-seq reads with discrepant alignments compared to the project sequence, multiple high quality discrepancies in the *Consed* assembly.

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

Using the Sequencer Updater (available through the GEP web site under “Projects” -> “Annotation Resources”), create a VCF (Variant Call Format) file that describes the changes to the consensus sequence you have identified above. **Paste a screenshot with the list of sequence changes below:**

Isoform report form

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

Gene-isoform name (i.e. dmoj_ey-PA): dbia lush-PA

Names of the isoforms with identical coding sequences as this isoform
dbia lush-PB

Is the 5’ end of this isoform missing from the end of project: No

If so, how many exons are missing from the 5’ end: _____

Is the 3’ end of this isoform missing from the end of the project: No

If so, how many exons are missing from the 3’ end: _____

1. 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 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 revise the submitted exon coordinates automatically using this VCF file.

View	Criteria	Status	Message
	Check for Start Codon	Pass	
	Acceptor for CDS 1	Skip	Already checked for Start Codon
	Donor for CDS 1	Pass	
	Acceptor for CDS 2	Pass	
	Donor for CDS 2	Pass	
	Acceptor for CDS 3	Pass	
	Donor for CDS 3	Skip	Already checked for Stop Codon
	Check for Stop Codon	Pass	
	Additional Checks	Pass	
	Number of coding exons matched D. melanogaster or...	Pass	

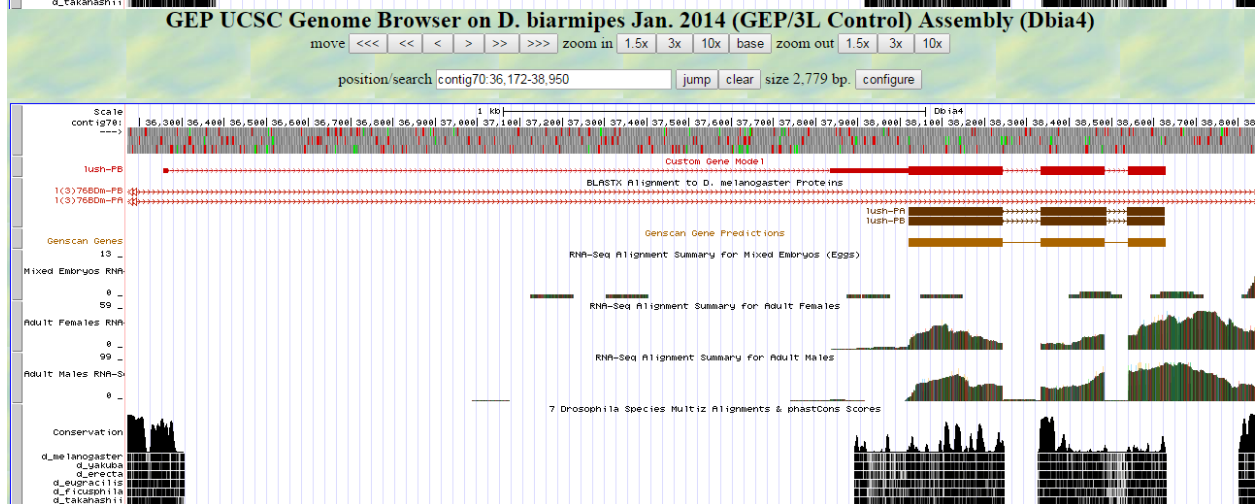
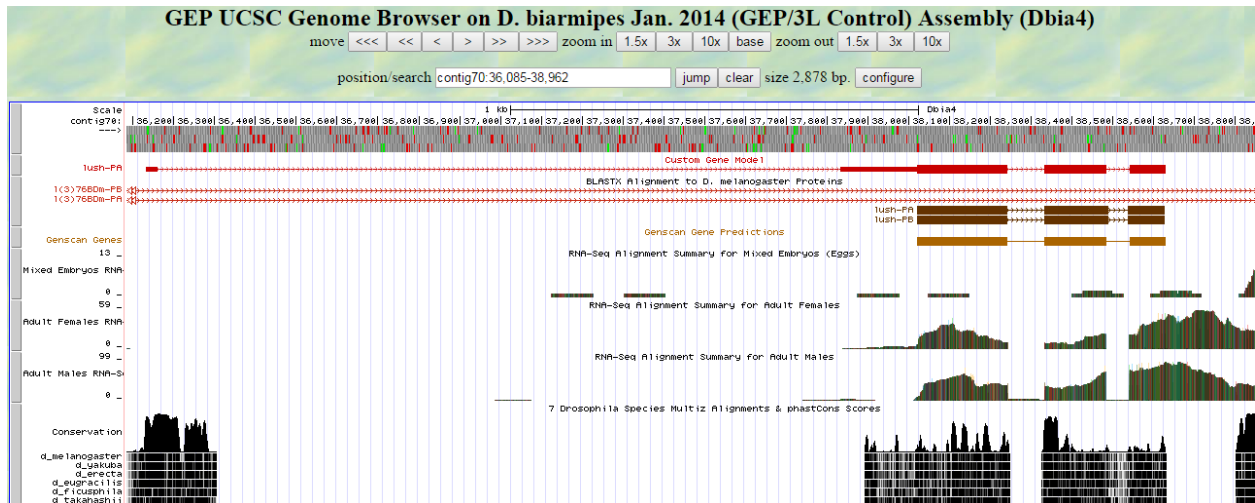
2. View the gene model on the Genome Browser

Using the custom track feature from the Gene Model Checker (see page 10 of the Gene Model Checker user guide on how to do this; you can find the guide under “Help” -> “Documentations” -> “Web Framework” on the GEP website at <http://gep.wustl.edu>).

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

5. A sequence alignment track (D. mel Protein or Other RefSeq)
6. At least one gene prediction track (e.g. Genscan)
7. At least one RNA-Seq track (e.g. RNA-Seq Alignment Summary)
8. A comparative genomics track
(e.g. Conservation, D. mel. Net Alignment, 3-way, 5-way or 7-way multiz)

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



Even though the coding exons are the same for lush-PA and PB, they have unique first transcribed exons. That is why I have included custom models for both above.

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 use the protein alignment generated by the Gene Model Checker or you can generate a new alignment using BLAST 2 Sequences (*bl2seq*). **Paste a screenshot of the protein alignment below:**

Alignment of lush-PA vs. Submitted_Seq

[View plain text version](#)

Identity: 128/153 (83.7%), Similarity: 139/153 (90.8%), Gaps: 0/153 (0.0%)

```

lush-PA      1 MKHWKRRSSAVFAIVLQVLVLLLPDPVAMTMEQFLTSLDMIRSGCAPKFKLKTEDLDRL 60
Submitted_Seq 1 MRHWQRSSSVLTIVLAVLGLLLPDPGTAMTMDQFLASLDMIRNGCAPKFKLNIEDLDRL 60

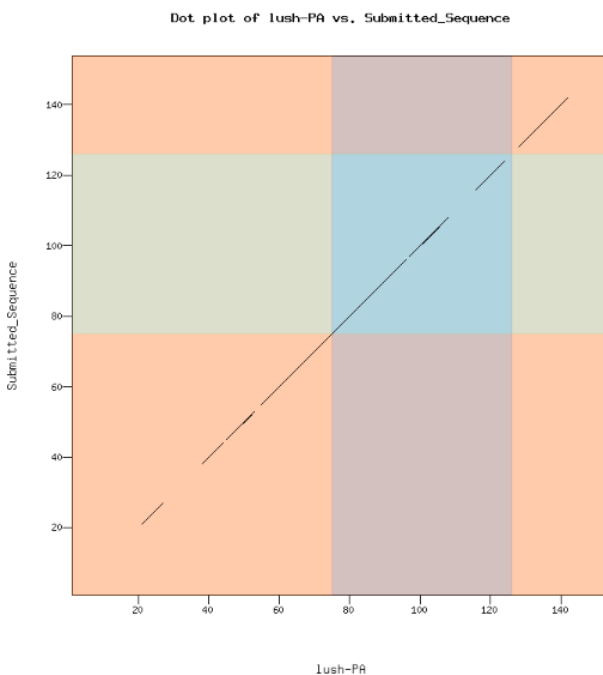
lush-PA      61 RVGDFNFPSPQDLMCYTKCVSLMAGTVNKKGFENAPKALAQPLVPPMEMMEMSRKSVEA 120
Submitted_Seq 61 RVGDFNFPSPQDLMCYTKCVSLMAGTVNKKGFENAPKALAQPLVPTMIEMSKKSVEA 120

lush-PA      121 CRDTEKQPKESCERVYQTAKCFSENADGQFMWP 153
Submitted_Seq 121 CRDAEKQPKESCERVYQTAKCLAENAECKFMWP 153
  
```

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). Provide an explanation for any anomalies on the dot plot (e.g. large gaps, regions with no sequence similarity).

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



The first and last exons have some mismatches with the *D. melanogaster* sequence. The middle exons in genes are usually the best conserved between species, and this is one of those cases. Relatedness can still easily be concluded based on the 83.7% identity and same length of both sequences.

Transcription start sites (TSS) report form (optional)

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

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

Complete this report form for each unique TSS listed in the table above (copy and paste to create as many copies of TSS report form as needed):

Gene-isoform name (i.e. dbia_ey-RA): dbia lush-PA

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

Type of core promoter: (Peaked or Broad): Broad

Coordinates of the first transcribed exon: 36133-36161

Coordinate(s) of TSS position(s): 36133

Coordinate(s) of TSS search region(s): 36133-38021

1. Evidence that supports the TSS annotation postulated above

Specify the type of evidence used to support the TSS annotation:

Evidence type	Support TSS annotation?	Refute TSS annotation?
blastn alignment of the initial exon from <i>D. melanogaster</i>	X	
RNA-Seq coverage and TopHat splice junctions	X	
Core promoter motifs		X
Sequence conservation with other <i>Drosophila</i> species	X	
Other (please specify)		

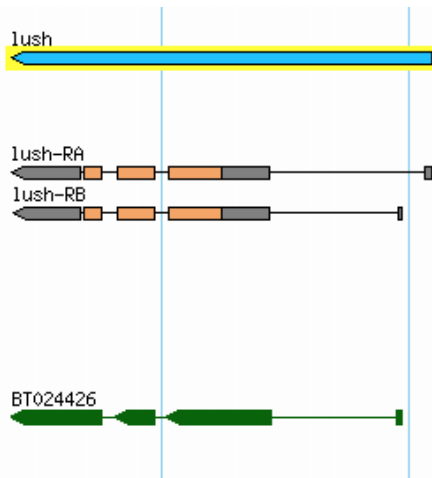
Provide an explanation if the TSS annotation is inconsistent with at least one of the evidence types specified above:



Searching the area just upstream of the first coding exon, which starts at 38021, there is a relatively dense collection of core promoter motifs. The Inr motif at 37874 and the BRE_d motif upstream of it give the indication of a possible TSS. The RNA-seq and conservation data also seem to somewhat support this TSS. However, a TSS at this location would remove a transcribed exon in the 5' UTR from the *D. melanogaster* model. The BLASTn alignment also refutes this TSS. My annotation of the TSS is not supported by any surrounding core promoter motifs.

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 below:**

The melanogaster model has 2 transcribed exons in the 5' UTR.



Dbia4_dna range=contig70:1-40000 5'pad=0 3'pad=0 strand=+ repeatMasking=none
 Sequence ID: lcl|207095 Length: 40000 Number of Matches: 84

Range 1: 36133 to 36161 [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
58.0 bits(29)	3e-12	29/29(100%)	0/29(0%)	Plus/Plus
Query 1	GTGATGTGCATCGCAAATGATCCGGTTCG	29		
Sbjct 36133	GTGATGTGCATCGCAAATGATCCGGTTCG	36161		

This is the BLASTn alignment for the first transcribed exon of lush-PA. It predicts a TSS at 36133.

Dbia4_dna range=contig70:1-40000 5'pad=0 3'pad=0 strand=+ repeatMasking=none
 Sequence ID: lcl|14135 Length: 40000 Number of Matches: 7

Range 1: 37893 to 38242 [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
251 bits(174)	4e-69	266/352(76%)	3/352(0%)	Plus/Plus
Query 67	AAATGGCGTGCGAGTGTGTTACTTAATTATTAATCCCTGCGATTTAATTACCACCTTTC	126		
Sbjct 37893	AAATGGCGTGCAAGTGCCTTGTAAATTACCAGAGCTACAAACGTGATAGCCAACTGTCC	37952		
Query 127	CTGCCACCATGATCA-CCTATAAAACTCTCCTACGACATGGTACTCAACGTATTTAGCT	185		
Sbjct 37953	AGCCAC-ATGGGCAGCCTATAAAGGCCTCCGCCCTCGTTGGCTAGTCGTCGTATTCG-T	38010		
Query 186	TCCGCCACCATGAAGCATTGGAACGACGCTCTCCGCTGTTTCGCGATCGTCCGTGCA	245		
Sbjct 38011	TCCGCCACCATGAGGCATTGGAGGCAACGCTCTTCCTCCGTTCTGACCATCGTCCGTGGC	38070		
Query 246	AGTGTGGTACTCCTGCTACCCGATCCTGCAGTTGCCATGACGATGGAGCAGTTCTTGAC	305		
Sbjct 38071	AGTTTTGGGCCCTCTCTTGCCAGATCCTGGGACAGCCATGACGATGGACCAGTTTTTGGC	38130		
Query 306	CTCGCTAGACATGATCCGCGAGTGCTGTGCGCCGAAGTTTAAGCTCAAAACAGAAGATCT	365		
Sbjct 38131	CTCGCTGGATATGATCCGGAATGTTGTGCGCCGAAGTTTAAGCTTAACATAGAAGATCT	38190		
Query 366	CGATCGGCTTCGCGTGGGTGATTTCAACTTTCGCCATCGCAGGATCTTATG	417		
Sbjct 38191	CGATCGGCTTCGCGTGGGGGATTTAATTTTCGCCGTGCGCAGGATCTCATG	38242		

This is the BLASTn alignment for the second transcribed exon of lush-PA. It does not align for the first 66 bases, but it gives a search region between about 37793-37893 to look for more evidence.

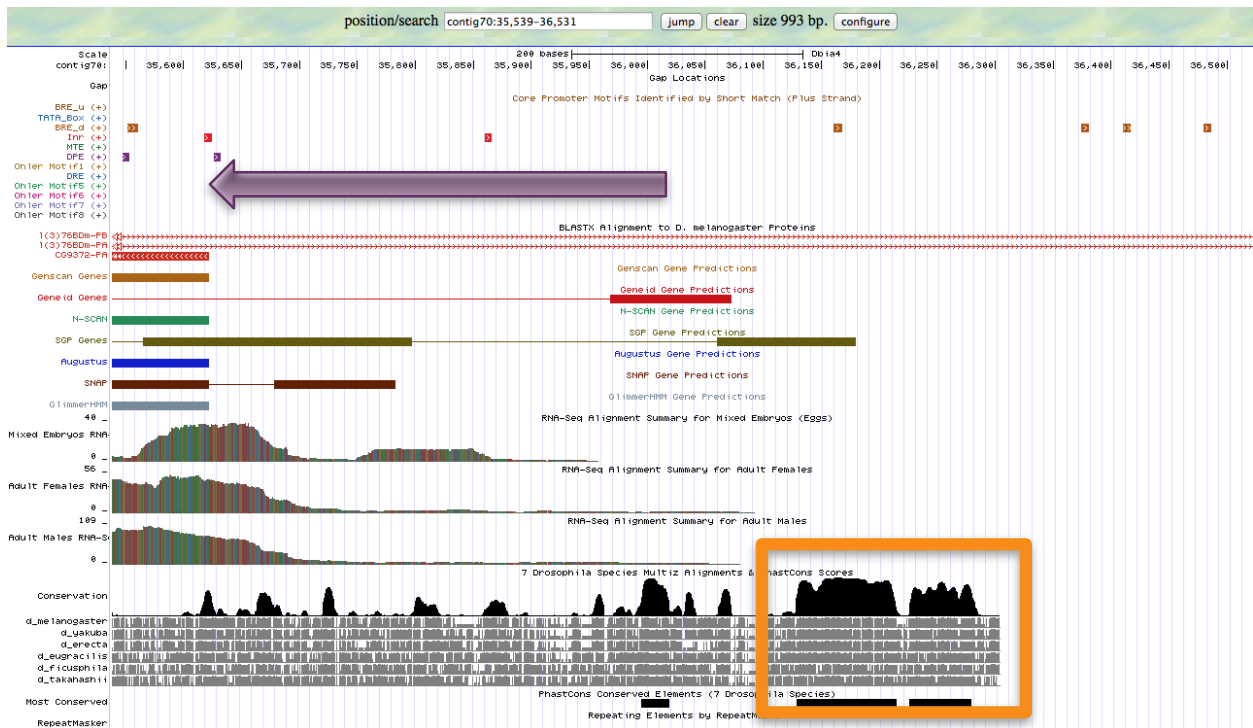
If the TSS annotation is supported by RNA-Seq or RNA polymerase II data, **paste a Genome Browser screenshot of the region around the TSS ($\pm 2\text{kb}$) with the evidence tracks listed below:**

4. Short Match results for the Inr motif (TCAKTY)
5. RNA-Seq Alignment Summary
6. RNA-Seq TopHat

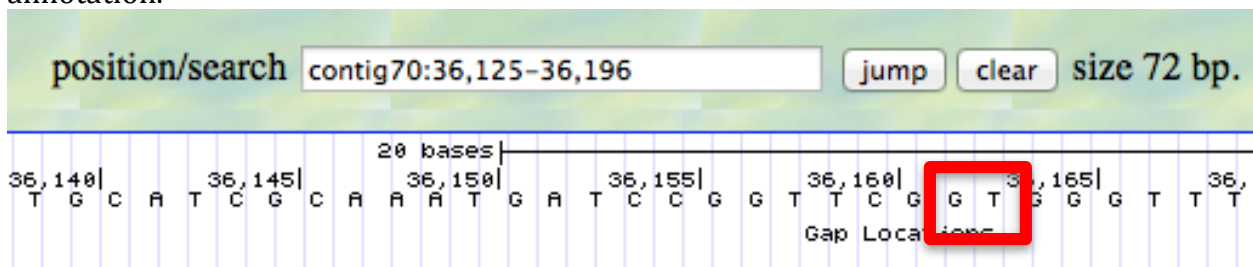


The second transcribed exon needs to have an acceptor site. TopHat indicates an acceptor site AG at 37832-37833. This prediction is supported by the BLASTn alignment, if the first 66 nucleotides aligned. It is also supported by RNA-seq data, which begins to improve coverage after this site.

If the TSS annotation is supported by sequence conservation with other Drosophila species, **paste a screenshot of the pairwise alignment (e.g. from blastn, matcher) or the multiple sequence alignment (e.g. from clustalw, EvoPrinter, Multiz) below:**



The purple arrow going to the left indicates the 5' UTR of the adjacent gene, CG9372. Thereby, it seems that the conservation shown by the Multiz alignments in the orange box is from the first transcribed exon of the lush gene. The BLASTn alignment indeed puts the exon in this region. This figure also shows the lack of core promoter motifs to support this annotation.



The BLASTn alignment of the first exon ends at 36,161. There is a GT donor site immediately after this alignment. Thereby, I annotated the first transcribed exon to the same start and end as the BLASTn alignment.

2. Search for core promoter motifs

Note: The consensus sequences for the *Drosophila* core promoter motifs are available at: http://gander.wustl.edu/~wilson/core_promoter_motifs.html

Use the "Short Match" functionality in the GEP UCSC Genome Browser to search for each of the core promoter motifs listed below in the region surrounding the TSS (± 300 bp) in your project and in the TSS of the *D. melanogaster* ortholog. (For TSS annotations where you can only define a TSS search region, you should search for the core promoter motifs in the entire TSS search region).

Record the **orientation and the start coordinate** (e.g. +10000) of each motif match below. (Enter "NA" if the motif is not present.)

Core promoter motif	Your project	<i>D. melanogaster</i>
BRE ^u	NA	NA
TATA Box	+37589	NA
BRE ^d	+36162, +36375, +36411, +36481, +36561, +36572, +36577, +36730, +36750, +36770, +36899, +37146, +37274, +37312, +37352, +37359, +37501, +37605, +37808, +37861, +37910,	-19606055, -19606184
Inr	+36957, +37047, +37350, +37874	-19606000, -19606110
MTE	NA	NA
DPE	+36601, +36986, +37199, +37730, +37880, +37995	-19605998
Ohler_motif1	NA	NA
DRE	NA	NA
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	NA	NA
Ohler_motif8	NA	NA

Transcription start sites (TSS) report form (optional)

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

Name(s) of isoform(s) with unique TSS	List of isoforms with identical TSS
lush-PB	

Complete this report form for each unique TSS listed in the table above (copy and paste to create as many copies of TSS report form as needed):

Gene-isoform name (i.e. dbia_ey-RA): dbia lush-PB

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

Type of core promoter: (Peaked or Broad): Broad

Coordinates of the first transcribed exon: 36256-36268

Coordinate(s) of TSS position(s): 36256

Coordinate(s) of TSS search region(s): 36256-38021

1. Evidence that supports the TSS annotation postulated above

Specify the type of evidence used to support the TSS annotation:

Evidence type	Support TSS annotation?	Refute TSS annotation?
blastn alignment of the initial exon from <i>D. melanogaster</i>	X	
RNA-Seq coverage and TopHat splice junctions	X	
Core promoter motifs		X
Sequence conservation with other Drosophila species	X	
Other (please specify)		

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

The core promoter motifs refute the TSS of lush-PB the same way they do lush-PA. There is another location supported by two core promoter motifs, but it is refuted by all other pieces of evidence and requires removing a transcribed exon from the *D. melanogaster* model. The annotated TSS at 36,256 is not supported by any core promoter motifs in the area.

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 below:**

Dbia4_dna range=contig70:1-40000 5'pad=0 3'pad=0 strand=+ repeatMasking=none
Sequence ID: lcl|97225 Length: 40000 Number of Matches: 46

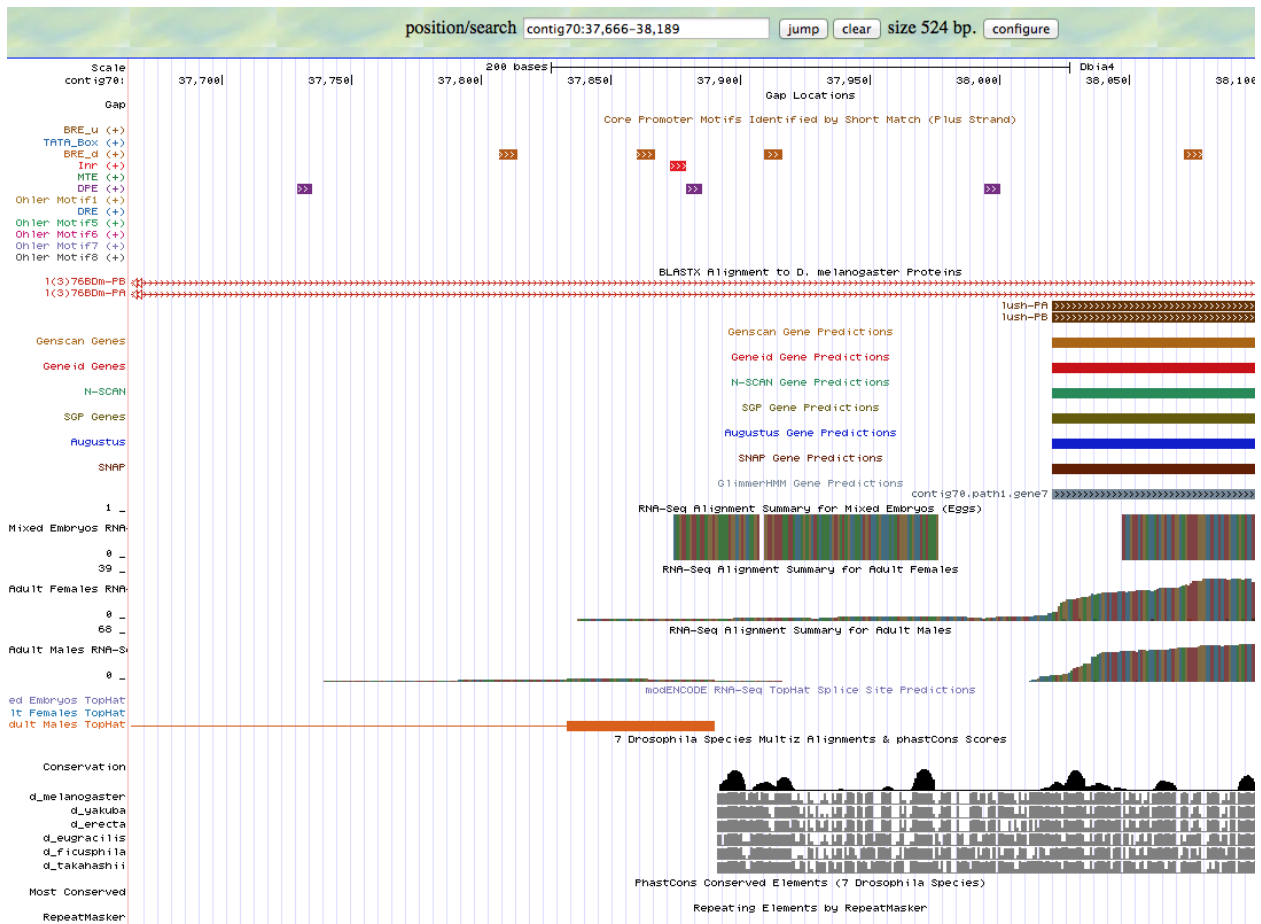
Range 1: 36256 to 36268 [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
26.3 bits(13)	0.003	13/13(100%)	0/13(0%)	Plus/Plus

```
Query 1      CTTGCACATCAA 13
           |||
Sbjct 36256  CTTGCACATCAA 36268
```

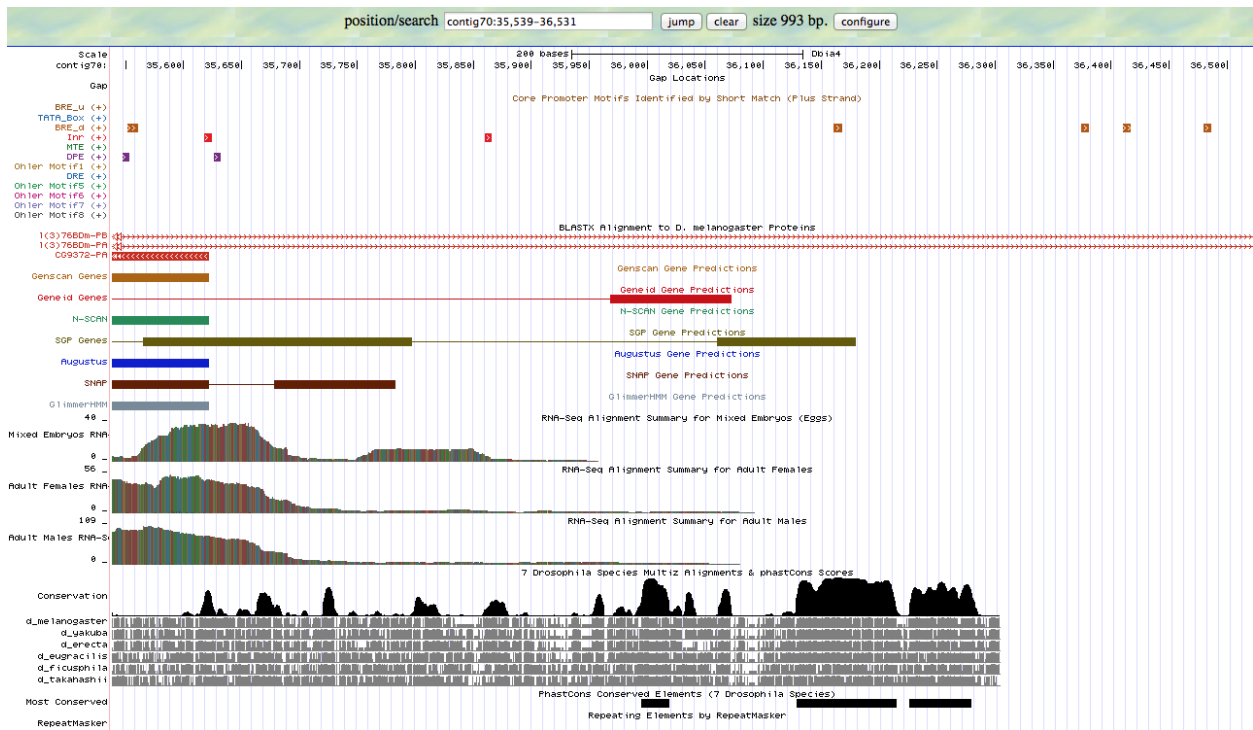
If the TSS annotation is supported by RNA-Seq or RNA polymerase II data, **paste a Genome Browser screenshot of the region around the TSS (± 2 kb) with the evidence tracks listed below:**

1. Short Match results for the Inr motif (TCAKTY)
2. RNA-Seq Alignment Summary
3. RNA-Seq TopHat



The second transcribed exon needs to have an acceptor site. TopHat indicates an acceptor site AG at 37832-37833. This prediction is supported by the BLASTn alignment, if the first 66 nucleotides aligned. It is also supported by RNA-seq data, which begins to improve coverage after this site.

If the TSS annotation is supported by sequence conservation with other *Drosophila* species, **paste a screenshot of the pairwise alignment (e.g. from blastn, matcher) or the multiple sequence alignment (e.g. from clustalw, EvoPrinter, Multiz) below:**



7 Drosophila Species Multiz Alignments & phastCons Scores

Conservation score statistics

Capitalize exons based on show bases

Place cursor over species for alignment detail. Click on 'B' to link to browser for aligned species, click on 'D' to get 1

Alignment block 1 of 1 in window, 36226 - 36304, 79 bps

```

B D   D. biarmipes   ttatattctctatttaattgccttctgacgcttgacatcaaaac---aataataaatcaactgcatgag
B D   D. melanogaster ttatattctctatttaattgccttttgacgcttgacatcaaaagttagtaataataaatcaattgcacggc
B D   D. yakuba     atattattctctatttaattgcattttgacgcttgacatcaaaagcagtaataataaatcaattgcattgac
B D   D. erecta     ctatattctctatttaattgccttttgacgcttgcaaatcaaaagcagtaataataaatcaattgcattgac
B D   D. eugracilis  ttatattcactatcttattgccttttgacgcttgacatcaaaacaataataataaatcaattacatgca
B D   D. ficusphila  ttat-ttctctacttaattgccttttgacgcttatacatcaaaac---agtaataaataaattacatgcc
B D   D. takahashii ttatattctctatttaattgcctcttgacgcttgacatcaaaac---aataataaatcaattgcaataa

```

```

D. biarmipes   aaaaccggaatc
D. melanogaster acaaccggaatc
D. yakuba     agaaccggaatc
D. erecta     agaaccggaatc
D. eugracilis  aaaaccggaatc
D. ficusphila  aaaaccggaatc
D. takahashii aaaccggaataa

```

BLASTn predicts the end of the first transcribed exon to be at 36,268. However, there is no donor sequence in this area for splicing. Looking at the Multiz conservation track in this area, there is inconsistency between the species. *D. biarmipes*, *D. ficusphila*, and *D. takahashii* have three gaps introduced. *D. melanogaster*, *D. yakuba*, and *D. erecta* all have a donating GT site, which could be used to end the exon. This needs to be investigated further to see if there is supposed to be a GT in the *D. biarmipes* sequence as well.

2. Search for core promoter motifs

Note: The consensus sequences for the *Drosophila* core promoter motifs are available at: http://gander.wustl.edu/~wilson/core_promoter_motifs.html

Use the "Short Match" functionality in the GEP UCSC Genome Browser to search for each of the core promoter motifs listed below in the region surrounding the TSS (± 300 bp) in your project and in the TSS of the *D. melanogaster* ortholog. (For TSS annotations where you can only define a TSS search region, you should search for the core promoter motifs in the entire TSS search region).

Record the **orientation and the start coordinate** (e.g. +10000) of each motif match below. (Enter "NA" if the motif is not present.)

Core promoter motif	Your project	<i>D. melanogaster</i>
BRE ^u	NA	NA
TATA Box	+37589	NA
BRE ^d	+36162, +36375, +36411, +36481, +36561, +36572, +36577, +36730, +36750, +36770, +36899, +37146, +37274, +37312, +37352, +37359, +37501, +37605, +37808, +37861, +37910,	-19606055, -19606184
Inr	+36957, +37047, +37350, +37874	-19606000, -19606110
MTE	NA	NA
DPE	+36601, +36986, +37199, +37730, +37880, +37995	-19605998
Ohler_motif1	NA	NA
DRE	NA	NA
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	NA	NA
Ohler_motif8	NA	NA

Gene report form

Gene name (i.e. *D. mojavensis eyeless*): *D. biarmipes CG9372*

Gene symbol (i.e. *dmoj_ey*): *dbia CG9372*

Approximate location in project (from 5' end to 3' end): 35622-33089

Number of isoforms in *D. melanogaster*: 1

Number of isoforms in this project: 1

Complete the following table for all the isoforms in this project:

Name(s) of unique isoform(s) based on coding sequence	List of isoforms with identical coding sequences
CG9372-PA	

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 your project consensus sequence:

All 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: No

1. Evidence that supports the consensus errors postulated above

Note: Evidence which supports the hypothesis of errors in the consensus sequence include: CDS alignment with frame-shifts or in-frame stop codons, multiple RNA-seq reads with discrepant alignments compared to the project sequence, multiple high quality discrepancies in the *Consed* assembly.

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

Using the Sequencer Updater (available through the GEP web site under "Projects" -> "Annotation Resources"), create a VCF (Variant Call Format) file that describes the changes

to the consensus sequence you have identified above. **Paste a screenshot with the list of sequence changes below:**

Isoform report form

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

Gene-isoform name (i.e. dmoj_ey-PA): dbia CG9372-PA

Names of the isoforms with identical coding sequences as this isoform

Is the 5' end of this isoform missing from the end of project: No

If so, how many exons are missing from the 5' end: _____

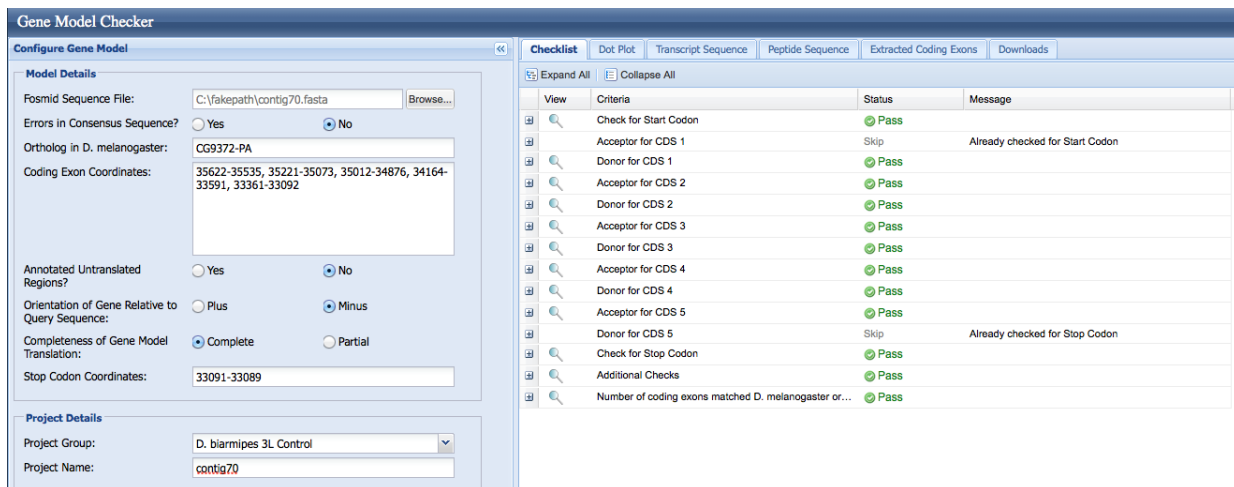
Is the 3' end of this isoform missing from the end of the project: No

If so, how many exons are missing from the 3' end: _____

1. 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 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 revise the submitted exon coordinates automatically using this VCF file.



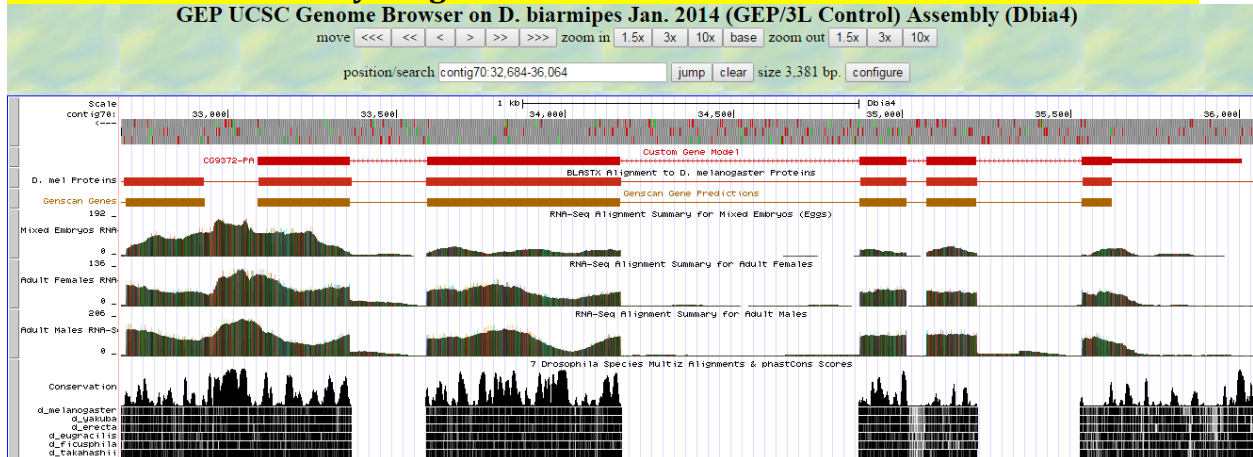
2. View the gene model on the Genome Browser

Using the custom track feature from the Gene Model Checker (see page 10 of the Gene Model Checker user guide on how to do this; you can find the guide under "Help" -> "Documentations" -> "Web Framework" on the GEP website at <http://gep.wustl.edu>). 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. Include the following evidence tracks in the screenshot if they are available.

9. A sequence alignment track (D. mel Protein or Other RefSeq)
10. At least one gene prediction track (e.g. Genscan)
11. At least one RNA-Seq track (e.g. RNA-Seq Alignment Summary)
12. A comparative genomics track (e.g. Conservation, D. mel. Net Alignment, 3-way, 5-way or 7-way multiz)

Paste the screenshot of your gene model as shown on the Genome Browser 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 use the protein alignment generated by the Gene Model Checker or you can generate a new alignment using BLAST 2 Sequences (*bl2seq*). **Paste a screenshot of the protein alignment below:**

Alignment of CG9372-PA vs. Submitted_Seq

[View plain text version](#)

Identity: 373/408 (91.4%), Similarity: 397/408 (97.3%), Gaps: 2/408 (0.5%)

```

CG9372-PA      1  MKAFLWALVILLGYIPOSTIAKTVSTDFLDLLDFSDNDEFQWGESENQVYENRTGENRVV  60
Submitted_Seq  1  MKAFLWAFVLLGFSVORILAKTITSYIDLDLLDFSDNDEFQWGESENQVFNRTGENQAV  60

CG9372-PA     61  SFLSOHRLNKRQAPTSOLLENKDYGACSTPLGEGRCRHIIYCRMPELKNDVWRLVSQLC  120
Submitted_Seq  61  NPLTRHRLNKRQAPNSQLLENKDYGPCSTPLGEGRCRHIIYCRMPELKNDVWRLVSQLC  120

CG9372-PA     121  IIEKSSIGICCTDOSTSNRFSFQVVTSADGDEPRIVNKPEQRCGGITTRQFPRLTGGRA  180
Submitted_Seq  121  IIEKSSIGICCTDOSTSNRFSFQIVTNPD--EPRIVNKPEQRCGGITTRQFPRLTGGRA  178

CG9372-PA     181  EPDEWPMAALLQEGLPFVWCGGVLITDRHVLTAAHCIYKKNKEDIFVRLGEYNTHMLNE  240
Submitted_Seq  179  EPDEWPMAALLQEGLPFVWCGGVLITDRHVLTAAHCIHKKNKEDIFVRLGEYNTHMLNE  238

CG9372-PA     241  TRARDFRIANMVLHIDYNPQNYDNDIAIVRIDRATIFNTYIWPVCMPPVNEDWSDRRAIV  300
Submitted_Seq  239  TRARDFRIANMVLHIDYNPQNYDNDIAIVRIDRATLFNTYIWPVCMPPINEDWAERNAIV  298

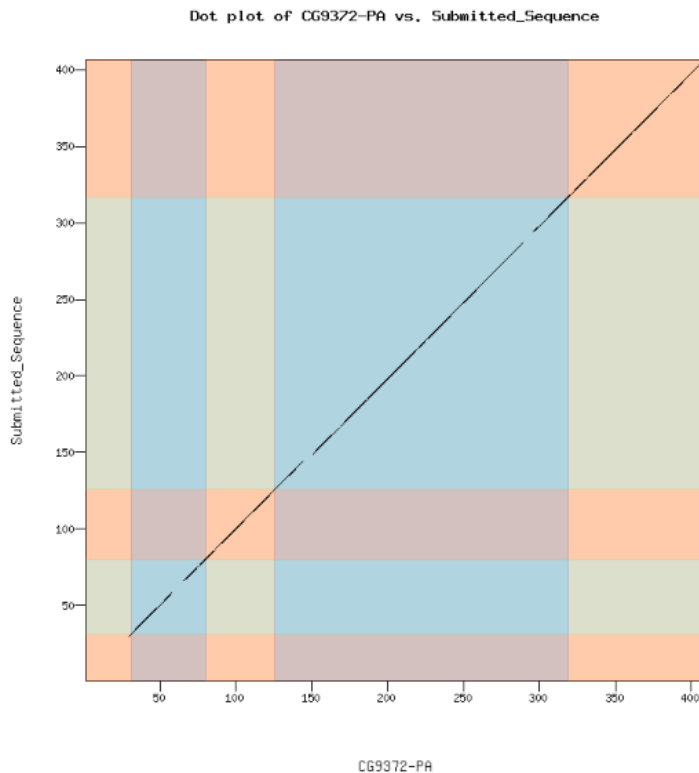
CG9372-PA     301  TGWGTQKFGGPHSNILMEVNLPVWQSDCRSFVQHVPDTAMCAGFPEGGQDSCQGDSGG  360
Submitted_Seq  299  TGWGTQKFGGPHSNILMEVNLPVWQSDCRASFVQHVPDTAMCAGFPEGGQDSCQGDSGG  358

CG9372-PA     361  PLLVQLPNQRWVTIGIVSWGVCGCGRPGIYTRVDRYLDWILANADV  408
Submitted_Seq  359  PLLVQLPNQRWVTIGIVSWGVCGCGRPGIYTRVDRYLDWILANADV  406
  
```

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). Provide an explanation for any anomalies on the dot plot (e.g. large gaps, regions with no sequence similarity).

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



The first exon is less conserved than the rest of the exons. However, it does have enough similarity, based on the protein alignment, to be annotated as it is.

Transcription start sites (TSS) report form (optional)

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

Name(s) of isoform(s) with unique TSS	List of isoforms with identical TSS
CG9372-PA	

Complete this report form for each unique TSS listed in the table above (copy and paste to create as many copies of TSS report form as needed):

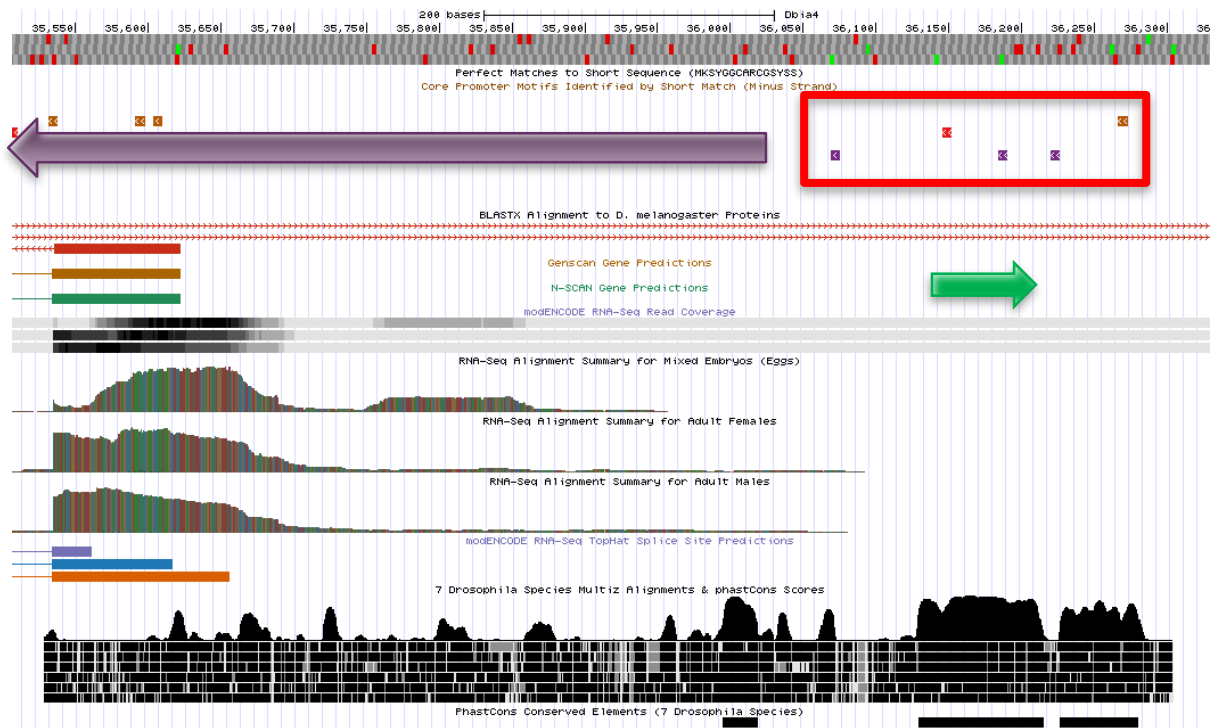
Gene-isoform name (i.e. dbia_ey-RA): dbia_CG9372-PA
 Names of the isoforms with the same TSS as this isoform:
 Type of core promoter: (Peaked or Broad): Broad
 Coordinates of the first transcribed exon: 36,009-35,535
 Coordinate(s) of TSS position(s): 36,009
 Coordinate(s) of TSS search region(s): 36,000-36,300

1. Evidence that supports the TSS annotation postulated above

Specify the type of evidence used to support the TSS annotation:

Evidence type	Support TSS annotation?	Refute TSS annotation?
blastn alignment of the initial exon from <i>D. melanogaster</i>	X	
RNA-Seq coverage and TopHat splice junctions	X	
Core promoter motifs		X
Sequence conservation with other <i>Drosophila</i> species	X	
Other (please specify)		

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



The core promoter motifs in the area refute the predicted TSS of 36,009. There are no core promoters supporting the predicted TSS. Upstream of the TSS, there is a greater concentration of core promoters, as shown by the red box in the figure above. The purple arrow shows the predicted first transcribed exon. The green arrow shows the first transcribed exon of the adjacent gene, lush-PA. Thus, the core promoters are not used to predict a TSS because it is unlikely that the two transcribed exons overlap.

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 below:**

Dbia4_dna range=contig70:1-40000 5'pad=0 3'pad=0 strand=+ repeatMasking=none
 Sequence ID: lc|26537 Length: 40000 Number of Matches: 9

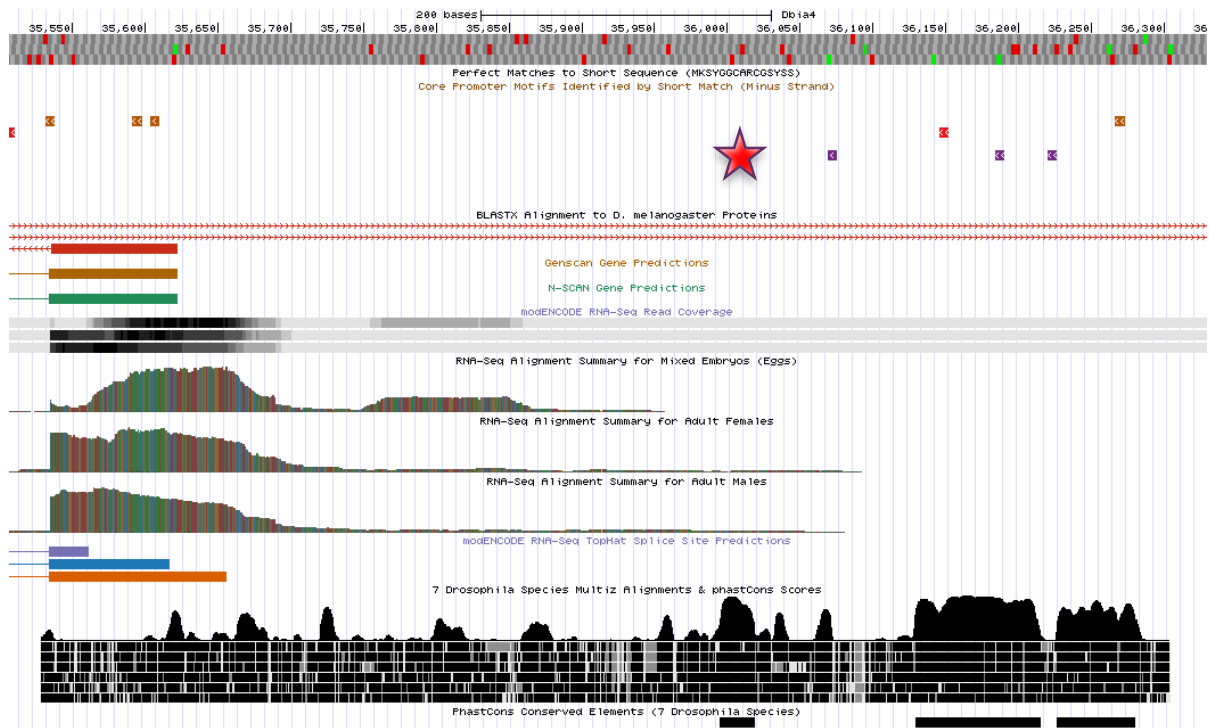
Range 1: 35535 to 36009 [Graphics](#) ▼ Next Match ▲ Previous

Score	Expect	Identities	Gaps	Strand
298 bits(207)	3e-83	361/491(74%)	46/491(9%)	Plus/Minus
Query 1	GCATTAAGGGGAC -GTCATTTCAAGTGGCTCGCTTTCCGCCCTACAGATAAGCCAAGTGC	59		
Sbjct 36009	GCATTAAGGGGACCGGCATCCAGTGCCTCGCTTTCCGCCCTACAGATAAGCCAAGAGC	35950		
Query 60	-----TCGGATTAAC--TCAAGTATTATATAATAGTTACTATATCGCTAGCGGTTTC	108		
Sbjct 35949	AGAAAGTACTCGGATTAATATTCGGGCGTTTCG-TAAAATTCACTATTTAAGT-GCGGTGC	35892		
Query 109	CAATCCCATTAAGCTTTTCGCCTGGGAAGTGCAT-GGAGTA-----AGA	152		
Sbjct 35891	CCATCCCAAAAGCTTTGTCTGGGAAGTGCATAGGAGTACGCGACTCGCCTGAGAGC	35832		
Query 153	CGCTTTAATAGCCAGGAAAACCCCATAGTGTGACGCATCGAACAGGTTCAACAGCAGTA	212		
Sbjct 35831	CGCTTCATTAGCCAGGAAAACCCCAAAAGTGTACCGCATTGAACAAGTCAAAAGCAGTG	35772		
Query 213	TCCAAGGCAAGACGTATGTAATTGAGCCGATCTGAGCTCAGC-TCCATATCGGAGAAGT	271		
Sbjct 35771	TCCAAGGCAAAATCGTAAGTC----GAGCCGGTCTCAGCTCGTCGTCCATATCGGAGCAGT	35716		
Query 272	AGCTTTAGCTGGTTGGAGAAGTGT-CAAACAGAAGTTCAGCGGGCAACGGAAGCGA	330		
Sbjct 35715	-----GCTCGTCGGAGAAGTGCACAGAGCAGAAGCGATTAGCTGGCAACGGAAGCGG	35663		
Query 331	CACGCGATTAATAATCCCACTGCACTCGCGAAAATAACTCAAAATGAAGGCATTTCTTTG	390		
Sbjct 35662	CACACGATAAAA---CCCAGTGCCTCGCGAAAATAACTCAAAATGAAGGCCTTTATTTG	35606		
Query 391	GGCATTGGTGATTTTACTGGGCTATATCCCGCAGAGTACAATCGCCAAGACTGTATCCAC	450		
Sbjct 35605	GGCTTTCGTGGTTTACTGGGCTTTCAGCGTGCAGCGTATACTCGCCAAAACAATAACCAG	35546		
Query 451	TGATTTTTTAG	461		
Sbjct 35545	TGATTATATAG	35535		

The BLASTn alignment supports the predicted TSS location at 36,009.

If the TSS annotation is supported by RNA-Seq or RNA polymerase II data, **paste a Genome Browser screenshot of the region around the TSS (± 2 kb) with the evidence tracks listed below:**

4. Short Match results for the Inr motif (TCAKTY)
5. RNA-Seq Alignment Summary
6. RNA-Seq TopHat



My predicted TSS, indicated by the red star, is supported by RNA-Seq data. The RNA-Seq does not pick up in read depth until downstream of the TSS.

If the TSS annotation is supported by sequence conservation with other *Drosophila* species, **paste a screenshot of the pairwise alignment (e.g. from blastn, matcher) or the multiple sequence alignment (e.g. from clustalw, EvoPrinter, Multiz) below:**

Alignment block 1 of 1 in window, 35971 - 36291, 321 bps

```

B D D. biarmipes      tgaaaagcgaggcactgggatgccgggtccccctttaatg ttatg-----agcttgggcccgaatttcgt--
B D D. melanogaster   cgaaaagcgaggcactgaaatg-acgtccccctttaatg ctctg-----agcttagaccgaatttggt--
B D D. yakuba         cgaaaagcgaggcactgagacg-cgggtccccctttaatg ttatgagctcagcttaggcccaatttatattt
B D D. erecta        cgaaaagcgaggcactgagatg-cgggtccccctttaatg ttatg-----agcttaggcccaattt-----
B D D. eugracilis    tggaaaagcgaggcactgggatgccgggtccccctttaatg ttatg-----agcttagaccgaattttgt--
B D D. ficusphila    tggaaaagcgagacactgatatgccgggtccccctttaatg ttatg-----agcttggaccgaatttcgt--
B D D. takahashii    tgaaaaacgagccactgagatgccgggtccccctttaatg ttatg-----agcttggaccgaattttgt--

D. biarmipes      -gtattatcttaatgttgctggca-acagtttggcaacatg-actccagggcgacctaccgcccatgatc
D. melanogaster   -gtattatctcaatgttgctggca-acagcttggcaacatggttgcaggggcga-gt-----agtgcctc
D. yakuba         cgtattatctcaatgttgctggca-atagcttggcaacatggttgcaggggtaagt-----agtgcctc
D. erecta        -----ttttcgtattatctca-atagcttggcaacatggttgcaggggcgaagt-----attgcctc
D. eugracilis    -gtattatctcaatgttgctggcatatagcttggcaacatgtttccaagggcaaaat-----attgcctc
D. ficusphila    -gtattatctcaatgttgctgact-tttgtttggcaacatgtttccaggcgctgaaa-----tttgcctc
D. takahashii    -gtattatctcaaagttgctgaca-aaagtttagcaaca-----ccaaaggcaaaat-----tgttgctc

D. biarmipes      acgatgatccactcggattctccgagga---- aaacgtgatgtgcatcgcaaatgatccgggtcgggtg
D. melanogaster   atgatgatcagtttggattctccgaggc---- gaacgtgatgtgcatcgcaaatgatccgggtcgggtg
D. yakuba         atgatgatcaatttgtattatccgagga---- gaacgtgatgtgcatcgcaaatgatccgggtcgggtg
D. erecta        atgacgatcagtttggattctccgagga---- gaacgtgatgtgcatcgcaaatgatccgggtcgggtg
D. eugracilis    acgatgatcaacttggattctccgagga---- aaacgtgatgtgcatcgcaaatgatccgggtcgggtg
D. ficusphila    atgataatccactcggattctctgagaa---- aaacgtgatgtgcatcggttaacgatccgggtcgggtg
D. takahashii    acgatgatcgactcggcttctccgatgaaat aaacgtgatggcattgcaaatgatccgggtcgggtg

D. biarmipes      gttttgattccgtag-agtaacatctccgaattactaagaaagattttataagtatacaacttatattct
D. melanogaster   gttttgattccatagaagtatcatctccgaaataactaaggaagattttata----ctcagttatattct
D. yakuba         gttttgattccatagaagtatcatctccgaaataactaaggaagattttataagtaactcaagatatattct
D. erecta        gttttgattccatagaagtatcatctccgaaataactaaggaagattttataagtgctcaagctatattct
D. eugracilis    gttttgattccatagaagtaacatctccgaaataactaaggaagattttataagtaacacagcttatattca
D. ficusphila    gttttgattccatagaagtaacatctccgaaataactgagggaagattttataactacacagcttat-ttct
D. takahashii    gttttaattccatagaagtaacatctccgaaataactaaggaagattttgtaggtacacagcttatattct

D. biarmipes      ctatttaattgccttctgacgcttgacacatcaaaaac---aataataaatcaactgcatga
D. melanogaster   ctatttaattgccttttgacgcttgacacatcaaaagtagtaataataaatcaattgcacgg
D. yakuba         ctatttaattgcattttgacgcttgacacatcaaaagcagtaataataaatcaattgcatgg
D. erecta        ctatttaattgccttttgacgcttgcaaatcaaaagcagtaataataaatcaattgcatgg
D. eugracilis    ctacttattgccttttgacgcttgacacatcaaaaacataataataaatcaattacatgc
D. ficusphila    ctacttaattgccttttgacgcttatacatcaaaaac---agtaataataataattacatgc
D. takahashii    ctatttaattgcctcttgacgcttgacacatcaaaaac---aataataaatcaattgcaata

```

The red box in the figure shows the first bases that are included in the first transcribed exon, according to my TSS prediction. The purple boxes show the conservation that is part of the adjacent lush gene.

2. Search for core promoter motifs

Note: The consensus sequences for the Drosophila core promoter motifs are available at: http://gander.wustl.edu/~wilson/core_promoter_motifs.html

Use the "Short Match" functionality in the GEP UCSC Genome Browser to search for each of the core promoter motifs listed below in the region surrounding the TSS (± 300 bp) in your project and in the TSS of the *D. melanogaster* ortholog. (For TSS annotations where you can only define a TSS search region, you should search for the core promoter motifs in the entire TSS search region).

Record the **orientation and the start coordinate** (e.g. +10000) of each motif match below. (Enter "NA" if the motif is not present.)

Core promoter motif	Your project	<i>D. melanogaster</i>
BRE ^u	NA	NA
TATA Box	NA	NA
BRE ^d	-36267	+19605909, +19606189, +19606384
Inr	-36147	+19606071, +19606223
MTE	NA	NA
DPE	-36070, -36185, -36221	+19606218, 19606429
Ohler_motif1	NA	NA
DRE	NA	NA
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	NA	NA
Ohler_motif8	NA	NA

Gene report form

Gene name (i.e. *D. mojavensis eyeless*): *D. biarmipes CG9376*

Gene symbol (i.e. dmoj_ey): *dbia CG9376*

Approximate location in project (from 5' end to 3' end): 27919-27209

Number of isoforms in *D. melanogaster*: 1

Number of isoforms in this project: 1

Complete the following table for all the isoforms in this project:

Name(s) of unique isoform(s) based on coding sequence	List of isoforms with identical coding sequences
CG9376-PA	

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 your project consensus sequence:

All 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 which supports the hypothesis of errors in the consensus sequence include: CDS alignment with frame-shifts or in-frame stop codons, multiple RNA-seq reads with discrepant alignments compared to the project sequence, multiple high quality discrepancies in the *Consed* assembly.

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

Using the Sequencer Updater (available through the GEP web site under “Projects” -> “Annotation Resources”), create a VCF (Variant Call Format) file that describes the changes to the consensus sequence you have identified above. **Paste a screenshot with the list of sequence changes below:**

Isoform report form

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

Gene-isoform name (i.e. dmoj_ey-PA): dbia 9376-PA

Names of the isoforms with identical coding sequences as this isoform

Is the 5’ end of this isoform missing from the end of project: No

If so, how many exons are missing from the 5’ end: _____

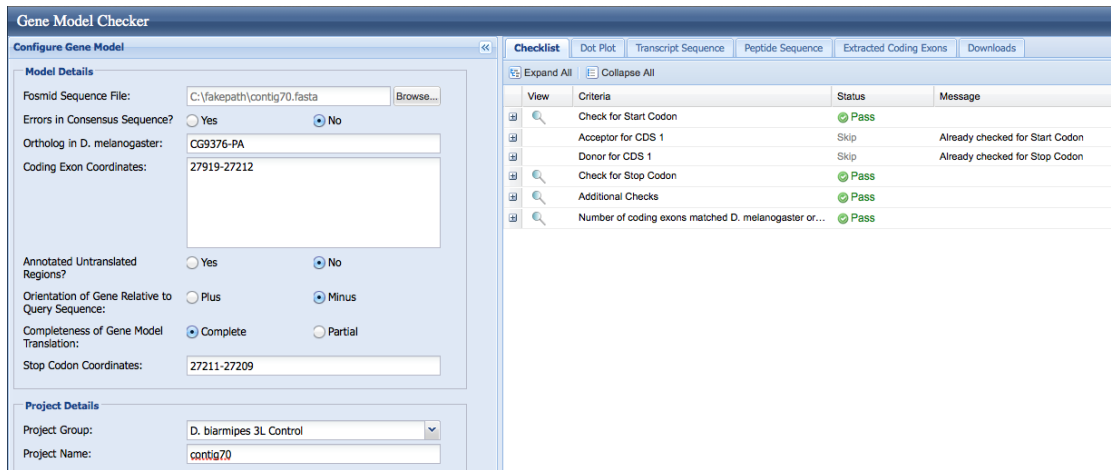
Is the 3’ end of this isoform missing from the end of the project: No

If so, how many exons are missing from the 3’ end: _____

1. 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 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 revise the submitted exon coordinates automatically using this VCF file.

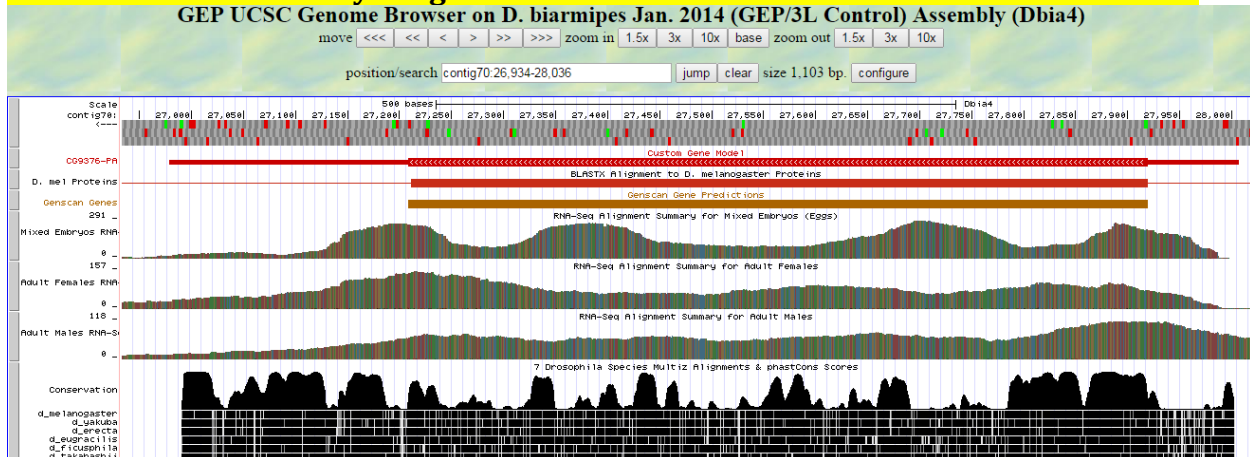


2. View the gene model on the Genome Browser

Using the custom track feature from the Gene Model Checker (see page 10 of the Gene Model Checker user guide on how to do this; you can find the guide under “Help” -> “Documentations” -> “Web Framework” on the GEP website at <http://gep.wustl.edu>). 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. Include the following evidence tracks in the screenshot if they are available.

13. A sequence alignment track (D. mel Protein or Other RefSeq)
14. At least one gene prediction track (e.g. Genscan)
15. At least one RNA-Seq track (e.g. RNA-Seq Alignment Summary)
16. A comparative genomics track (e.g. Conservation, D. mel. Net Alignment, 3-way, 5-way or 7-way multiz)

Paste the screenshot of your gene model as shown on the Genome Browser 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 use the protein alignment generated by the Gene Model Checker or you can generate a new alignment using BLAST 2 Sequences (*bl2seq*). **Paste a screenshot of the protein alignment below:**

Alignment of CG9376-PA vs. Submitted_Seq

[View plain text version](#)

Identity: 232/236 (98.3%), Similarity: 234/236 (99.2%), Gaps: 0/236 (0.0%)

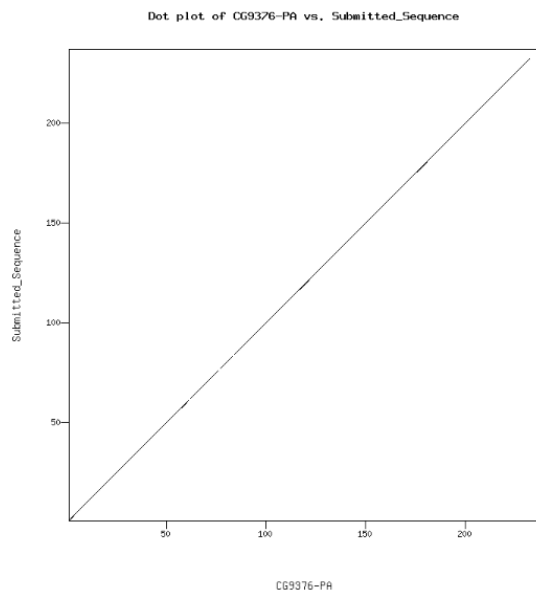
```

CG9376-PA      1  MWSRFDQOKTKHQLFHVSLSLATILICMAYMQYRRNWAHLGSFWDLSLVVPIVFLGELLKV  60
Submitted_Seq  1  MWSRFDQOKTKHQLFHVSLSLATILICMAYMQYRRNWAHLGSFWDLSLVVPIVFLGELLKV  60
CG9376-PA     61  VLARFYGKVEDGVLTAQRQKNSYFTPRELLGGFTLQFLCTLLYAFICILGAPVLGNY  120
Submitted_Seq  61  VLARFYGKVEDGVLTVQRQKASVFTPRELLGGFTLQFLCTLLYAFICILGAPVLGNY  120
CG9376-PA     121  BQTFVLALLMTLLTVSPTVFLGGGGALQVCPCEKPDFVTKCEDTALNLPKYNALGGILG  180
Submitted_Seq  121  BQTFVLALLMTLLTVSPTVFLGGGGALQVCPCEKPDFVTKCEDTALNLPKYNALGGILG  180
CG9376-PA     181  AWAGSVVAPLDWGRDWAQYPIPNVIGALLGSALGNIYACTHVLVYATARVYMTKRR  236
Submitted_Seq  181  AWAGSVVAPLDWGRDWAQYPIPNVIGALLGSALGNIYACTHVLVYATARVYMSKRR  236
    
```

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). **Provide an explanation for any anomalies** on the dot plot (e.g. large gaps, regions with no sequence similarity).

Note: Large vertical and horizontal gap near exon boundaries in the dot plot often indicates that an incorrect splice site might have been picked. Please re-examine these regions and provide a detail 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 specified above. This section is OPTIONAL and you do not need to complete this section to submit the project.

Name(s) of isoform(s) with unique TSS	List of isoforms with identical TSS
CG9376-PA	

Complete this report form for each unique TSS listed in the table above (copy and paste to create as many copies of TSS report form as needed):

Gene-isoform name (i.e. dbia_ey-RA): dbia CG9376-PA
 Names of the isoforms with the same TSS as this isoform:

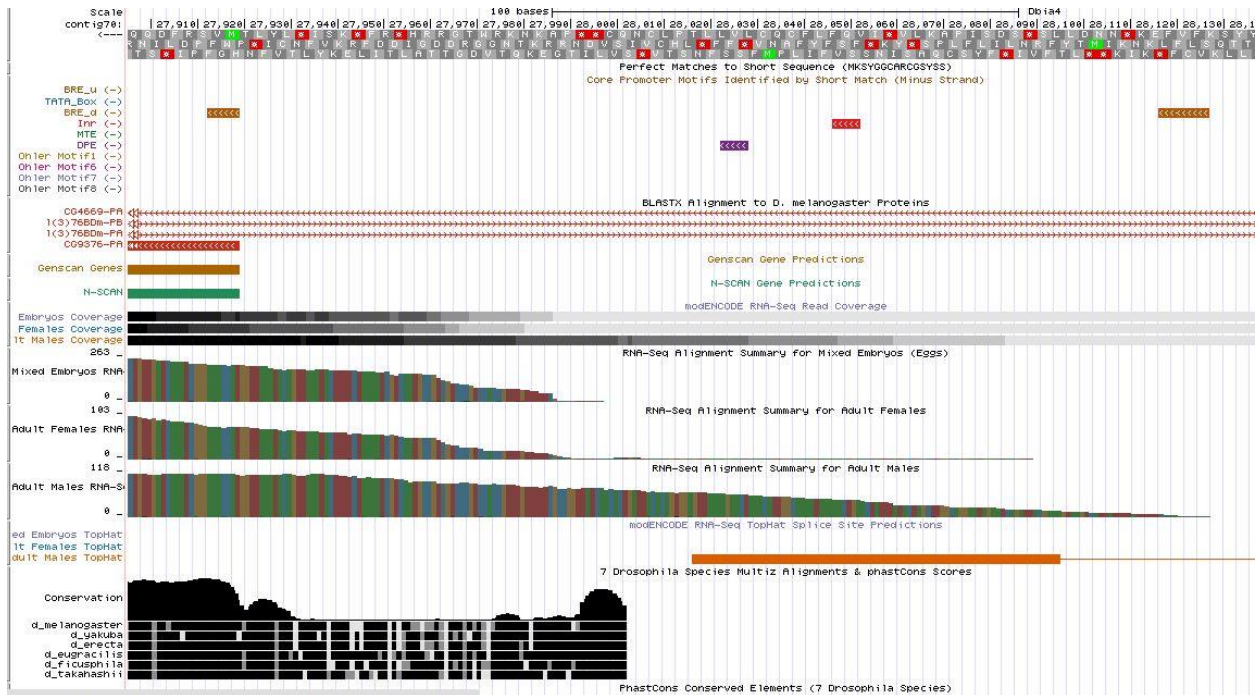
Type of core promoter: (Peaked or Broad): Peaked
 Coordinates of the first transcribed exon: 28007-26980
 Coordinate(s) of TSS position(s): 28008
 Coordinate(s) of TSS search region(s): 27970-28140

1. Evidence that supports the TSS annotation postulated above

Specify the type of evidence used to support the TSS annotation:

Evidence type	Support TSS annotation?	Refute TSS annotation?
blastn alignment of the initial exon from <i>D. melanogaster</i>	X	
RNA-Seq coverage and TopHat splice junctions	X	
Core promoter motifs		X
Sequence conservation with other <i>Drosophila</i> species	X	
Other (please specify)		

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



The core promoters in the area do not support the predicted TSS of 28,008. There are BRE_d motifs at 28142, 28121, and 28117. There is an InR motif at 28047 and a DPE motif at 28023. The proximity relationship of the DPE and InR motifs gives a possible indication of another TSS. However, there is not enough evidence to add this as a prediction, and more analysis may need to be done.

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 below:**

Dbia4_dna range=contig70:1-40000 5'pad=0 3'pad=0 strand=+ repeatMasking=none
 Sequence ID: lcl|13631 Length: 40000 Number of Matches: 17

Range 1: 26992 to 28004 [Graphics](#)

▼ Next Match ▲ Previous

Score	Expect	Identities	Gaps	Strand
1016 bits(710)	0.0	876/1026(85%)	19/1026(1%)	Plus/Minus
Query 4	TCAGTGTGATAGGCGGCGAAATACAAACGAAAGGCAGGAAAATATTGTCCAAGTTAAG			63
Sbjct 28004	TCAGTGTGATAACCGGCGAAAAACAAACGGTGGACGGGACGACGG-----CA--TTGACG			27951
Query 64	ATTTACCAAAATTGTATTTATCTGCATTTAATCATGGTTTCCCGATTGACCAACAAAAG			123
Sbjct 27950	ATTTTAGAAAA--GTATTTAACTGTATTTAACCATGGTTTCCCGATTGATCAACAGAAG			27893
Query 124	ACCAAGCACCAGCTCTCCACGTGTCCCTCAGCCTCGCCACCATCCTGATCTGCATGGCG			183
Sbjct 27892	ACCAAGCACCAGCTCTCCACGTGTCCCTCAGCCTGGCCACCATCCTGATCTGCATGGCG			27833
Query 184	TACATGCAGTATCGCCGGAATTGGGCACATCTCGGCAGCTTCTGGGATTCTCTTGTCTGTC			243
Sbjct 27832	TACATGCAGTATCGCCGGAACCTGGGCACATCTCGGCAGCTTCTGGGATTCTTTGGTGGTC			27773
Query 244	CCTATTGTTTTCTCGGAGAGCTGCTAAAAGTTGTTTTGGCTCGCTTACGGGAAAGGTT			303
Sbjct 27772	CCGATTGTTTTCTCGGTGAGCTTCTGAAAAGTTATACTGGCCGCTTCTATGGGAAAGTT			27713
Query 304	GAGGATGGCGTCCTAACCGCCAAACAGCGCCAGAAAAAGAACTCGTACTTTACGCCGCGG			363
Sbjct 27712	GAGGATGGCGTCCTGACTGTTAAACAGCGCCAGAAAAAGGCCTCGTACTTTACACCGCGG			27653
Query 364	GAGCTCCTCGGAGGATTACCCCTGCAGTTCCTGTGTACACTCCTCTACGCCTTTATCTGC			423
Sbjct 27652	GAGCTCCTCGGGGGATTCACTCTGCAGTTTCTTTGCACGCTGCTTTACGCCTTTATCTGC			27593
Query 424	ATAATTTTGGGAGCCCCGGTGTGGGCAACTATGAGCAGACCTTCGTCTTGGCCTTACTG			483
Sbjct 27592	ATCATTTTGGGAGCCCCGGTGTGGGCAACTACGAGCAGACCTTTGTCTTGGCCTTACTT			27533
Query 484	ATGACCTTGTGACGGTGTACCCACGGTTTTCTCCTCGGTGGCGGAGGAGCACTACAA			543
Sbjct 27532	ATGACACTGCTGACGGTGTGCGCAACTGTTTTCTGCTCGGGGGCGGCGGAGCCCTCCAG			27473

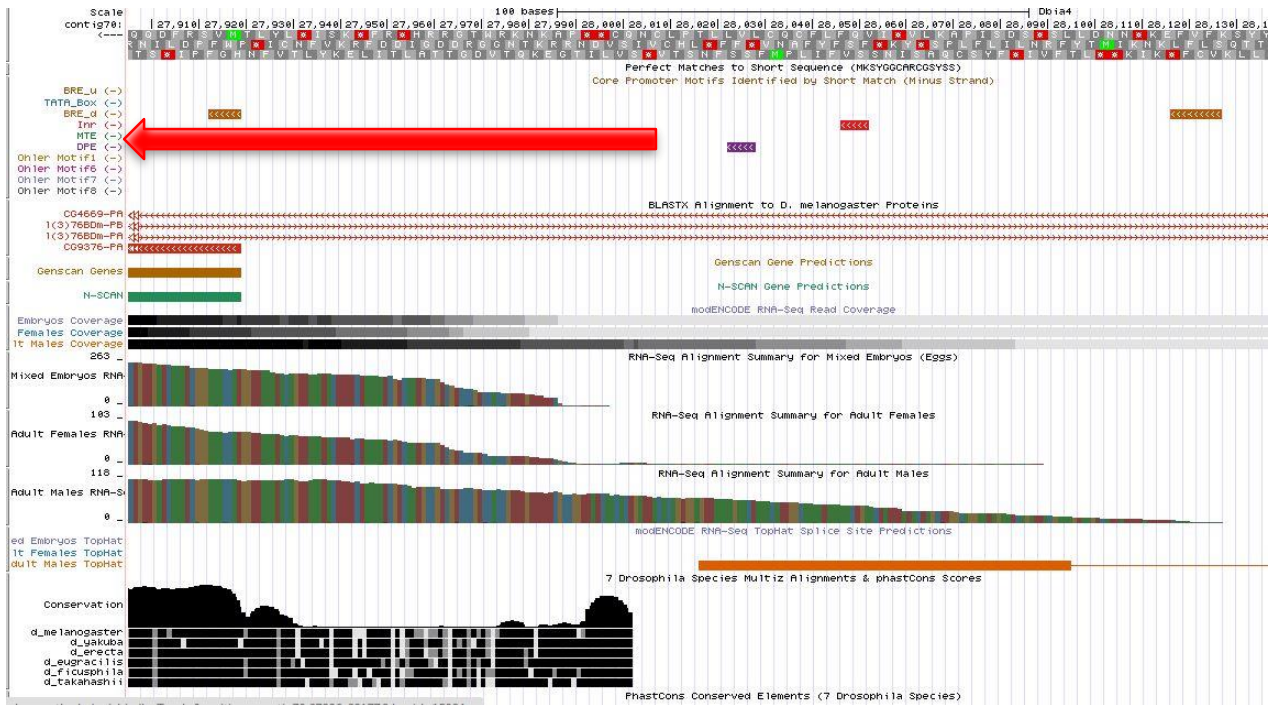
```

Query 544 GTGTGCTTCTGCGAGAAACCGGACTTTGTGACCAAGTGCGAGGACACGGCTCTGAATCTG 603
      |||
Sbjct 27472 GTTTGCTTCTGTGAGAAAGCCGGACTTTGTGACCAAGTGCGAGGATACGGCGCTGAACCTG 27413
      |||
Query 604 TTCAAGTACAATGCACTGGGCGGTATTCTGGGAGCTTGGGCCGGGAGCGTGGTCGCTCCA 663
      |||
Sbjct 27412 TTTAAGTACAATGCGCTGGGCGGGATTCTGGGCGCCTGGGCCGGAAGTGTGGTAGCTCCT 27353
      |||
Query 664 CTAGACTGGGAGCGTACTGGCAGGCTTATCCCATCCCAACGTGATCGGAGCACTGCTG 723
      |||
Sbjct 27352 CTAGACTGGGAGCGGACTGGCAGGCATACCCTATTCCGAATGTGATTGGAGCACTATTG 27293
      |||
Query 724 GGAAGCGCTCTGGGCAATATATACGCTTGTACGCATGTCCTCTACGCCACAGCTCGAGTT 783
      |||
Sbjct 27292 GGCAGCGCTCTGGGTAACATATACGCCTGTACACACGTCCTTTATGCCACGGCCCGTGT 27233
      |||
Query 784 TACATGACCAAGAAACGCACTTAAATCTAATAATAAAA-CTCTTCATTCTGCCACCAAGA 842
      |||
Sbjct 27232 TATATGAGCAAGAAACGCACTTAAACCTTGCAATAAATGCTATTAATCCTGCCACCAAGA 27173
      |||
Query 843 TTATTACGTTGCGCTGCCACTGAAATCCAC----TCATCCACTAAGAACCACGTCATCAT 898
      |||
Sbjct 27172 TTATTACGTTGCGCTGCCTCTGAATCCACACACTCATCCACTAAGAACCACGTCATCAT 27113
      |||
Query 899 CCCGCTAGTCAGCAGCTTAGGAACTACCAAAAAGCAATTACCACCTTAAATGTTACATGT 958
      |||
Sbjct 27112 CTCGACTAGTCAGCAGCTTAGGAACTACCAAAAAGCAATTAACAATTTAAAGTTACATAT 27053
      |||
Query 959 TTAATCATATTACTAGCGTGGATTATTAATGAATCTTTTTAAATCAATCAA-TAATAAA 1017
      |||
Sbjct 27052 TTAAGCATTTTACTAGTGTAG--TATTA---ATTCTTTTAAATCAATCAATTAATAAA 26998
      |||
Query 1018 CTAACA 1023
      |||
Sbjct 26997 CTAACA 26992

```

If the TSS annotation is supported by RNA-Seq or RNA polymerase II data, **paste a Genome Browser screenshot of the region around the TSS (± 2 kb) with the evidence tracks listed below:**

7. Short Match results for the Inr motif (TCAKTY)
8. RNA-Seq Alignment Summary
9. RNA-Seq TopHat



The predicted TSS of 28,008 is supported by the RNA-seq data from the mixed embryos and the adult females. The base of the red arrow in the figure above indicates the start of the 5' UTR. The adult males RNA-seq data seems to continue further upstream. This gene may actually be a broad promoter that have a different TSS for the sexes.

If the TSS annotation is supported by sequence conservation with other *Drosophila* species, **paste a screenshot of the pairwise alignment (e.g. from blastn, matcher) or the multiple sequence alignment (e.g. from clustalw, EvoPrinter, Multiz) below:**

7 Drosophila Species Multiz Alignments & phastCons Scores

Conservation score statistics

Capitalize exons based on show bases

Place cursor over species for alignment detail. Click on 'B' to link to browser for aligned species, click on

Alignment block 1 of 1 in window, 27896 - 28002, 107 bps

B D	D. biarmipes	ctgttgatcaaatcgggaaaccatggttaaatacagttaaat---acttttctaaaatcgtcaa-----
B D	D. melanogaster	ttgttggtcgaatcgggaaaccatgattaaatgcagataaatac-aattttggtaaatcttaaaccttgga
B D	D. yakuba	ttgttggtcaaaacgggaaaccattgttaaatacagataaaaac-aattttggtaaatcttaga-----
B D	D. erecta	ttgttggtcgaatcgggaaaccatggttaaatacagataaatac-aattttggtaaatcttagacttggt
B D	D. eugracilis	ctgttggtcgaatcgggaaaccatgattaaatgcaataaaaatat-agttttctaaaacctccga-----
B D	D. ficusphila	ttgttggtcgaatcgggaaaccatgattaaatgcagttaaatct-aaattt-tacaactctt-----
B D	D. takahashii	ctgttggtcgaatcgggaaaccatggttaaatacagttcaatagaattttctgaagttccttaa-----
	D. biarmipes	tgccg-----tcgtcccgtccaccggtttgtttttcgcggttatcaaacat
	D. melanogaster	caata-----ttttcctgcctttcgtttgtatttcgcccgttatcaaacat
	D. yakuba	tgctcgttatatTTTTTctctcttccggtttgtatttcgcccgttatcaaacat
	D. erecta	tgata-----ttttcctgtcttccggttatatttcgcccgttatcaaacat
	D. eugracilis	tgct-----tcttccctgcctccggtttatttttcgcccgttatcaaacat
	D. ficusphila	tgccg-----tcttctgactgcggtttgatttttgccgatttatcaaacat
	D. takahashii	tgccg-----ttttcctgacatccggtttgttttacgcccgttatcaaacat

The Multiz alignments support the TSS as there is a high level of conservation near the start of the 5' UTR. The conservation is good starting at 28,002.

2. Search for core promoter motifs

Note: The consensus sequences for the *Drosophila* core promoter motifs are available at: http://gander.wustl.edu/~wilson/core_promoter_motifs.html

Use the "Short Match" functionality in the GEP UCSC Genome Browser to search for each of the core promoter motifs listed below in the region surrounding the TSS (± 300 bp) in your project and in the TSS of the *D. melanogaster* ortholog. (For TSS annotations where you can only define a TSS search region, you should search for the core promoter motifs in the entire TSS search region).

Record the **orientation and the start coordinate** (e.g. +10000) of each motif match below. (Enter "NA" if the motif is not present.)

Core promoter motif	Your project	<i>D. melanogaster</i>
BRE ^u	NA	NA
TATA Box	NA	NA
BRE ^d	-28142, -28121, -28117, -27913, -27763	+19612155, +19612157, +19612227, +19612252, +19612317, +19612326, +19612338, +19612347, +19612387, +19612436, +19612500, +19612551, +19612701, +19612728, +19612731
Inr	-28047	NA
MTE	NA	NA
DPE	-28023, -27738	+19612727
Ohler_motif1	NA	NA
DRE	NA	NA
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	NA	NA
Ohler_motif8	NA	NA

Gene report form

Gene name (i.e. *D. mojavensis eyeless*): *D. biarmipes Lon*

Gene symbol (i.e. dmoj_ey): dbia Lon

Approximate location in project (from 5' end to 3' end): 28806-32932

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:

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

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 your project consensus sequence:

All 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 which supports the hypothesis of errors in the consensus sequence include: CDS alignment with frame-shifts or in-frame stop codons, multiple RNA-seq reads with discrepant alignments compared to the project sequence, multiple high quality discrepancies in the *Consed* assembly.

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

Using the Sequencer Updater (available through the GEP web site under "Projects" -> "Annotation Resources"), create a VCF (Variant Call Format) file that describes the changes to the consensus sequence you have identified above. **Paste a screenshot with the list of sequence changes below:**

Isoform report form

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

Gene-isoform name (i.e. dmoj_ey-PA): dbia Lon-PA

Names of the isoforms with identical coding sequences as this isoform

Is the 5' end of this isoform missing from the end of project: No

If so, how many exons are missing from the 5' end: _____

Is the 3' end of this isoform missing from the end of the project: No

If so, how many exons are missing from the 3' end: _____

1. 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 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 revise the submitted exon coordinates automatically using this VCF file.

View	Criteria	Status	Message
<input type="checkbox"/>	Check for Start Codon	Pass	
<input type="checkbox"/>	Acceptor for CDS 1	Skip	Already checked for Start Codon
<input type="checkbox"/>	Donor for CDS 1	Pass	
<input type="checkbox"/>	Acceptor for CDS 2	Pass	
<input type="checkbox"/>	Donor for CDS 2	Pass	
<input type="checkbox"/>	Acceptor for CDS 3	Pass	
<input type="checkbox"/>	Donor for CDS 3	Pass	
<input type="checkbox"/>	Acceptor for CDS 4	Pass	
<input type="checkbox"/>	Donor for CDS 4	Pass	
<input type="checkbox"/>	Acceptor for CDS 5	Pass	
<input type="checkbox"/>	Donor for CDS 5	Pass	
<input type="checkbox"/>	Acceptor for CDS 6	Pass	
<input type="checkbox"/>	Donor for CDS 6	Pass	
<input type="checkbox"/>	Acceptor for CDS 7	Pass	
<input type="checkbox"/>	Donor for CDS 7	Pass	
<input type="checkbox"/>	Acceptor for CDS 8	Pass	
<input type="checkbox"/>	Donor for CDS 8	Pass	
<input type="checkbox"/>	Acceptor for CDS 9	Pass	
<input type="checkbox"/>	Donor for CDS 9	Skip	Already checked for Stop Codon
<input type="checkbox"/>	Check for Stop Codon	Pass	
<input type="checkbox"/>	Additional Checks	Pass	
<input type="checkbox"/>	Number of coding exons matched D. melanogaster or...	Pass	

2. View the gene model on the Genome Browser

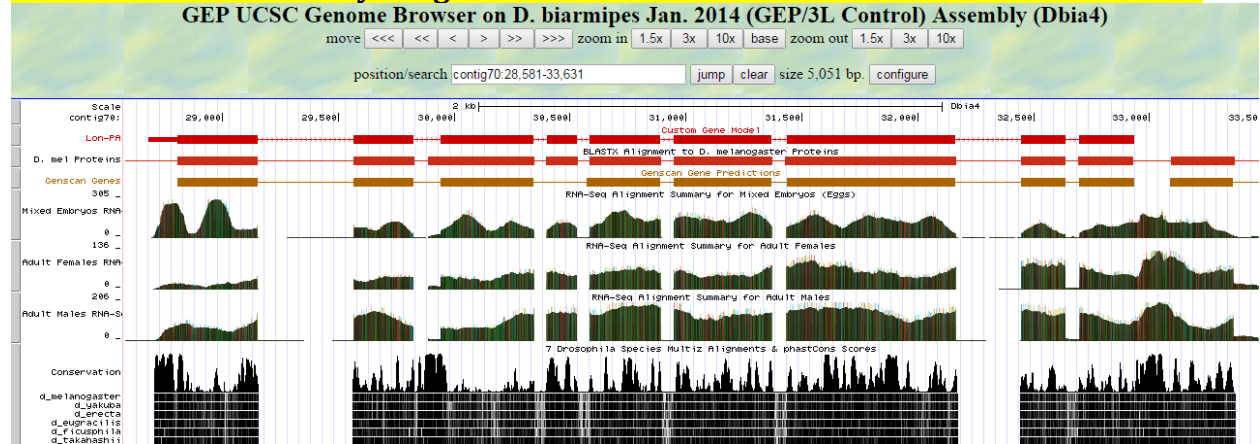
Using the custom track feature from the Gene Model Checker (see page 10 of the Gene Model Checker user guide on how to do this; you can find the guide under "Help" -> "Documentations" -> "Web Framework" on the GEP website at <http://gep.wustl.edu>).

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

17. A sequence alignment track (D. mel Protein or Other RefSeq)

- 18. At least one gene prediction track (e.g. Genscan)
- 19. At least one RNA-Seq track (e.g. RNA-Seq Alignment Summary)
- 20. A comparative genomics track (e.g. Conservation, D. mel. Net Alignment, 3-way, 5-way or 7-way multiz)

Paste the screenshot of your gene model as shown on the Genome Browser 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 use the protein alignment generated by the Gene Model Checker or you can generate a new alignment using BLAST 2 Sequences (*bl2seq*). **Paste a screenshot of the protein alignment below:**

Alignment of Lon-PA vs. Submitted_Seq

[View plain text version](#)

Identity: 914/1008 (90.7%), Similarity: 953/1008 (94.5%), Gaps: 6/1008 (0.6%)

```

Lon-PA      1  MLARATIRVPMWRGIASSVWVTRNPAQSSLVVCCRVVRAHLQRFHGANMNVVQVYRKRK  60
Submitted_Seq 1  MLARATIRVPMWRGIASSVWVTRNPAQSSLVVCCRVVRAHLQRFHGANMNVVQVYRKRK  60

Lon-PA      61  DDSMDLMDGHPNLSMDREAQLPATVAVDVPVHLLMRKRPFRPFMKIVVNSNDP  118
Submitted_Seq 61  DDSMDLMDGHPNLSMDREAQLPATVAVDVPVHLLMRKRPFRPFMKIVVNSNDP  120

Lon-PA      119  IMDLLRKRVKLNQPVGVVFKKSDQSEELTNDVYVNLGQFAQIQELGDLGDKRMVVV  178
Submitted_Seq 121  IMDLLRKRVKLNQPVGVVFKKSDQSEELTNDVYVNLGQFAQIQELGDLGDKRMVVV  180

Lon-PA      179  VRRIRVTVGVVVDVPPKPAQQSTQDAADAPIKRSDFARKPRGRIFRSRTKRSKSA  238
Submitted_Seq 181  VRRIRVTVGVVVDVPPKPAQQSTQDAADAPIKRSDFARKPRGRIFRSRTKRSKSA  239

Lon-PA      239  AARELIQNTLPEPLKSGKVSSESLPKPPTKEKIVEPETGAKENVQSAQVPLVIVE  298
Submitted_Seq 240  ATEENWVQWLEPPLKSGQVCEVSPKPPTEKRVQSGAGGATQASAPVPLVIVE  299

Lon-PA      299  ENVKQPIYKQTEVIALFOIILITLADLITNPNLYESELQQLBQMRVVDNPITFLCDL  358
Submitted_Seq 300  ENVKQPIYKQTEVIALFOIILITLADLITNPNLYESELQQLBQMRVVDNPITFLCDL  359

Lon-PA      359  VSLSAGIEPQKTLERNDPFRRLQLALTLKKELELSRLQOKIGREVEEKVQOHRKYY  418
Submitted_Seq 360  VSLSAGIEPQKTLERNDPFRRLQLALTLKKELELSRLQOKIGREVEEKVQOHRKYY  419

Lon-PA      419  IQGQLVKIKELGILTKDDKDAIETEYRKLKQKVVPEAMVIVDELTNLESHSSE  478
Submitted_Seq 420  IQGQLVKIKELGILTKDDKDAIETEYRKLKQKVVPEAMVIVDELTNLESHSSE  479

Lon-PA      479  NVFNYFDWTFELVGVISFENLCEKATLINDHGYMEDIKRIHETVIVSFKGQVY  538
Submitted_Seq 480  NVFNYFDWTFELVGVISFENLCEKATLINDHGYMEDIKRIHETVIVSFKGQVY  539

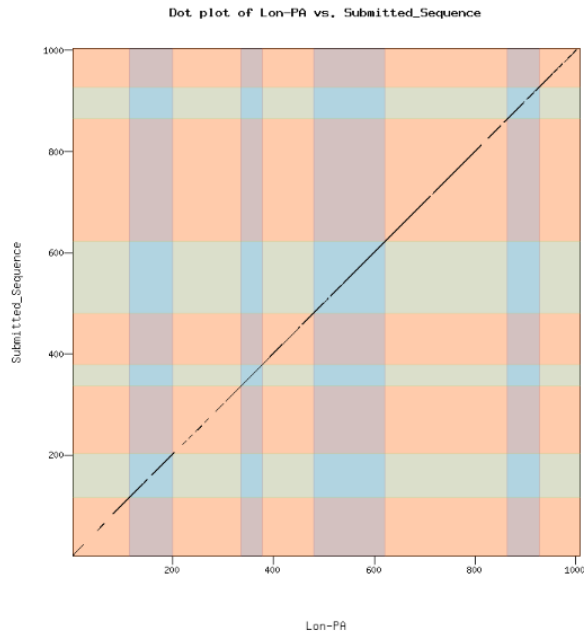
Lon-PA      539  KTLVCPGPGVGRKSTAKSIALANLREYFRFVGGMTDVAETKHRRYVGMAMPKRL  598
Submitted_Seq 540  KTLVCPGPGVGRKSTAKSIALANLREYFRFVGGMTDVAETKHRRYVGMAMPKRL  599

Lon-PA      599  ELKATVEMPLMLIDVVKIKGQVGGDSSALLHLDPGQANFLDHLIDVVDLSVLE  658
Submitted_Seq 600  ELKATVEMPLMLIDVVKIKGQVGGDSSALLHLDPGQANFLDHLIDVVDLSVLE  659
  
```

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). **Provide an explanation for any anomalies** on the dot plot (e.g. large gaps, regions with no sequence similarity).

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



Isoform report form

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

Gene-isoform name (i.e. dmoj_ey-PA): dbia_Lon-PC

Names of the isoforms with identical coding sequences as this isoform

Is the 5' end of this isoform missing from the end of project: No

If so, how many exons are missing from the 5' end: _____

Is the 3' end of this isoform missing from the end of the project: No

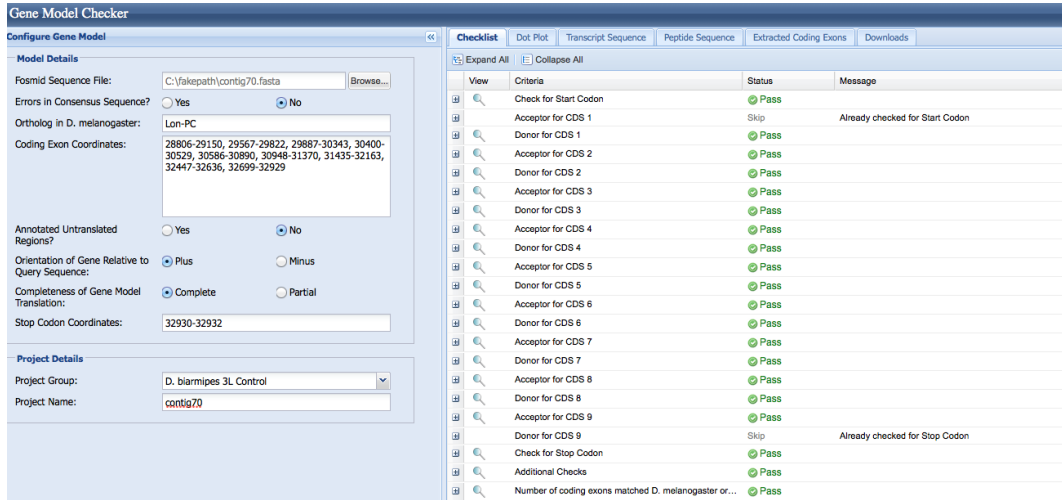
If so, how many exons are missing from the 3' end: _____

1. 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 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 revise the submitted exon coordinates automatically using this VCF file.



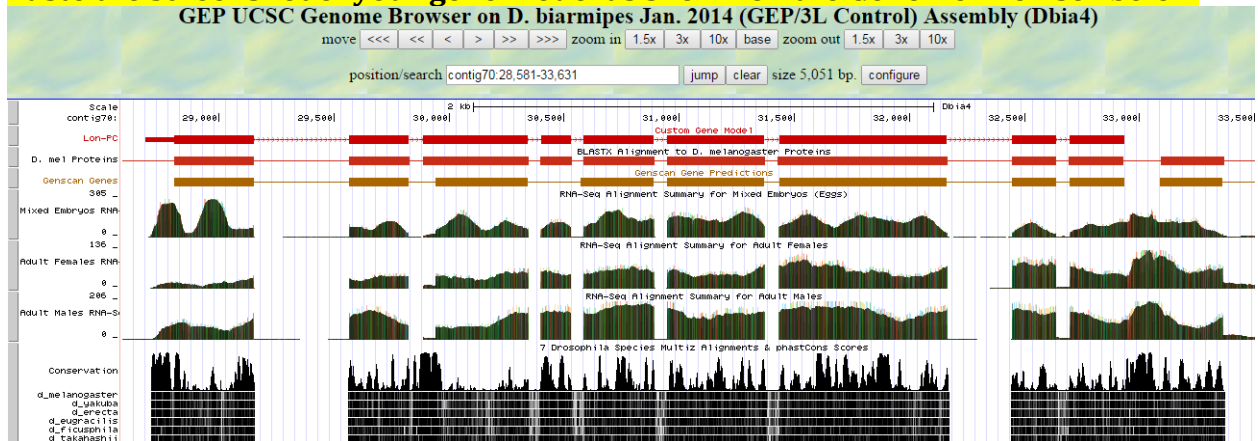
2. View the gene model on the Genome Browser

Using the custom track feature from the Gene Model Checker (see page 10 of the Gene Model Checker user guide on how to do this; you can find the guide under “Help” -> “Documentations” -> “Web Framework” on the GEP website at <http://gep.wustl.edu>).

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

5. A sequence alignment track (D. mel Protein or Other RefSeq)
6. At least one gene prediction track (e.g. Genscan)
7. At least one RNA-Seq track (e.g. RNA-Seq Alignment Summary)
8. A comparative genomics track (e.g. Conservation, D. mel. Net Alignment, 3-way, 5-way or 7-way multiz)

Paste the screenshot of your gene model as shown on the Genome Browser 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 use the protein alignment generated by the Gene Model Checker or you can generate a new alignment using BLAST 2 Sequences (*bl2seq*). **Paste a screenshot of the protein alignment below:**

Alignment of Lon-PC vs. Submitted_Seq

[View plain text version](#)

Identity: 932/1026 (90.8%), Similarity: 971/1026 (94.6%), Gaps: 6/1026 (0.6%)

```
Lon-PC      1 MLARAIRVPMRMTAASSVNNRNPQSSIMQNCRDPSLQLRFGANLAWQRFYSRKR 60
Submitted_Seq 1 MLARAIRVPMRMTAASSVNNRNPQSSIVQCCVRATLQRFGANLAWQRFYSRKR 60

Lon-PC      61 DSSNGDIING--PDLMSDQDTHLPATVAVPDPVWPHVPLLAMRKNPLFRPKIIVEVSNPI 118
Submitted_Seq 61 DSDSDLDMDGNPELMSDREAQLPATVAVPDPVWPHVPLLAMRKNPLFRPKIIVEVSNPI 120

Lon-PC      119 LMDLELRKVKIANQVYGVVFAKTCGSEFTIINTADWVNLGPFVQIQDFQDQKLSHWVV 178
Submitted_Seq 121 LMDLELRKVKIANQVYGVVFAKTCGSEFTIINTADWVNLGPFVQIQDFQDQKLSHWVV 180

Lon-PC      179 AHRRIYTCQVVEVPPPKVPMKTLHYPLFNKIQTPAEDQSTQQAADAPIKSRSDPAR 238
Submitted_Seq 181 AHRRIYTCQVVEVPPPKVPMKTLHYPLFNKIQTPAEDQSTDRAGEAP-KSRADPSR 239

Lon-PC      239 KPRGRIPRSRTGKSRSSAAAEELIQNOTLEPPLKSGKVESSELKPKPTEEKIVPEPTGAK 298
Submitted_Seq 240 KPRGRIPRSRAGKSRSSVATEEMVQNTLEPPLKSGQVECEKSPKPTTEGKRVQSQAGAE 299

Lon-PC      299 RNVNQSAPSAQPVLLIVEVENKQPIYKQTEEVKALQEIINKLRLDLYTMNPLVRSLSLQQM 358
Submitted_Seq 300 GEATQSAFSAFPVLLIVEVENKQPAYKQTEEVKALQEIINKLRLDLYTMNPLVRSLSLQQM 359

Lon-PC      359 HQNQRVNDNPIYLCDLGASISAGEPAELQKILSETDIPERLQLALVTLKKELELSRLQQ 418
Submitted_Seq 360 HQNQRVNDNPIYLCDLGASISAGEPAELQKILSETDIPERLQLALVTLKKELELSRLQQ 419

Lon-PC      419 KIGREVEEKVQQRKRYLQEQLVKIKKELGIEKDDKDAIGEKVREKIKDKVPEALMIV 478
Submitted_Seq 420 KIGREVEEKVQQRKRYLQEQLVKIKKELGIEKDDKDAIGEKVREKIKDKVPEALMIV 479

Lon-PC      479 IDBELTKLNFLESHSSEFVFRNYLDMGLSIPACVISTENICLEKATETIANDHYCMEDY 538
Submitted_Seq 480 IDBELTKLNFLESHSSEFVFRNYLDMGLSIPACVISTENICLEKATETIANDHYCMEDY 539

Lon-PC      539 KGRLEPIAVSLSKQVQKILCPGPPGVGKTSIAKSIANALNRVYRFVQGMQDVAE 598
Submitted_Seq 540 KGRLEPIAVSLSKQVQKILCPGPPGVGKTSIAKSIANALNRVYRFVQGMQDVAE 599

Lon-PC      599 LKGRHRVYVGMPEKTHQGLKXTIIEINPVDIVDQVTKGCTQGDPSALLELDPEQNA 658
Submitted_Seq 600 LKGRHRVYVGMPEKTHQGLKXTIIEINPVDIVDQVTKGCTQGDPSALLELDPEQNA 659

Lon-PC      659 NFLDHYLDVFPVLSRVLEICTANVIDTIPPELRDRMELTEMSGVABEKIATARQVLMQP 718
Submitted_Seq 660 NFLDHYLDVFPVLSRVLEICTANVIDTIPPELRDRMELTEMSGVABEKIATARQVLMQP 719

Lon-PC      719 AMKDCGLTDKHLINISEDALNMLIRSYCRESGVRNLQKHLKVIKRVAFRVVVKEGEHPV 778
Submitted_Seq 720 AMKDCGLTDKHLINISEDALNMLIRSYCRESGVRNLQKHLKVIKRVAFRVVVKEGEHPV 779

Lon-PC      779 NADNLTFGLKQIFSSDRMYATTPVGVVGLAWTANGGSSLYIETSRRHIRHGAKTDPNI 838
Submitted_Seq 780 NADNLTFGLKQIFSSDRMYATTPVGVVGLAWTANGGSSLYIETSRRHIRHGAKTDPNI 839

Lon-PC      839 VAGSLHITGNLGDVMKSAQIALTVARNFLVSLKPNLPLEQDIIHLVYDCAVTPDQDPS 898
Submitted_Seq 840 AGGZLHITGNLGDVMKSAQIALTVARNFLVSLKPNLPLEQDIIHLVYDCAVTPDQDPS 899

Lon-PC      899 AGIYIIFALVSLANGKFWRODIALMGEVSLKGVLEWGGIIEKMTIARRSGVNCILIPVD 958
Submitted_Seq 900 AGIYIIFALVSLANGKFWRODIALMGEVSLKGVLEWGGIIEKMTIARRSGVNCILIPVD 959

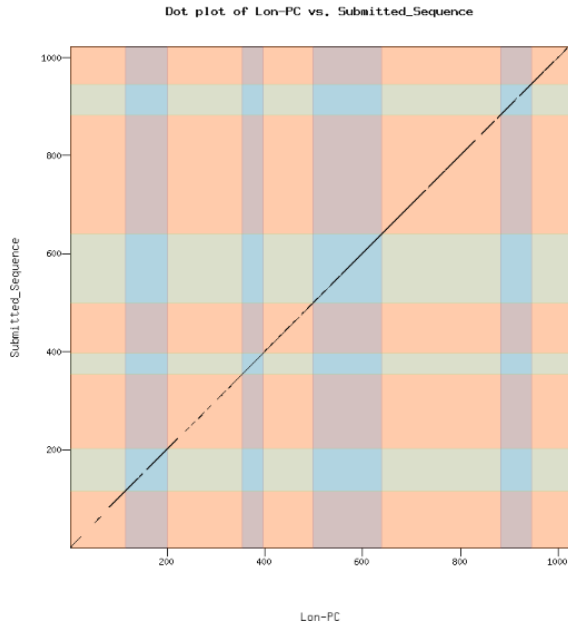
Lon-PC      959 NKKDFEELPTYITDGLVHFATTYEDVYKIAFDVTETTNNVEEQEPLQKLSAAAAAKS 1018
Submitted_Seq 960 NKKDFEELPTYITDGLVHFATTYEDVYKIAFDVSDA-ETTTNNVVEEQEPLQKLSAAAAAKS 1018

Lon-PC      1019 STMPYS 1024
Submitted_Seq 1019 ATWP--- 1022
```

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). **Provide an explanation for any anomalies** on the dot plot (e.g. large gaps, regions with no sequence similarity).

Note: Large vertical and horizontal gap near exon boundaries in the dot plot often indicates that an incorrect splice site might have been picked. Please re-examine these regions and provide a detail 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 specified above. This section is OPTIONAL and you do not need to complete this section to submit the project.

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

Complete this report form for each unique TSS listed in the table above (copy and paste to create as many copies of TSS report form as needed):

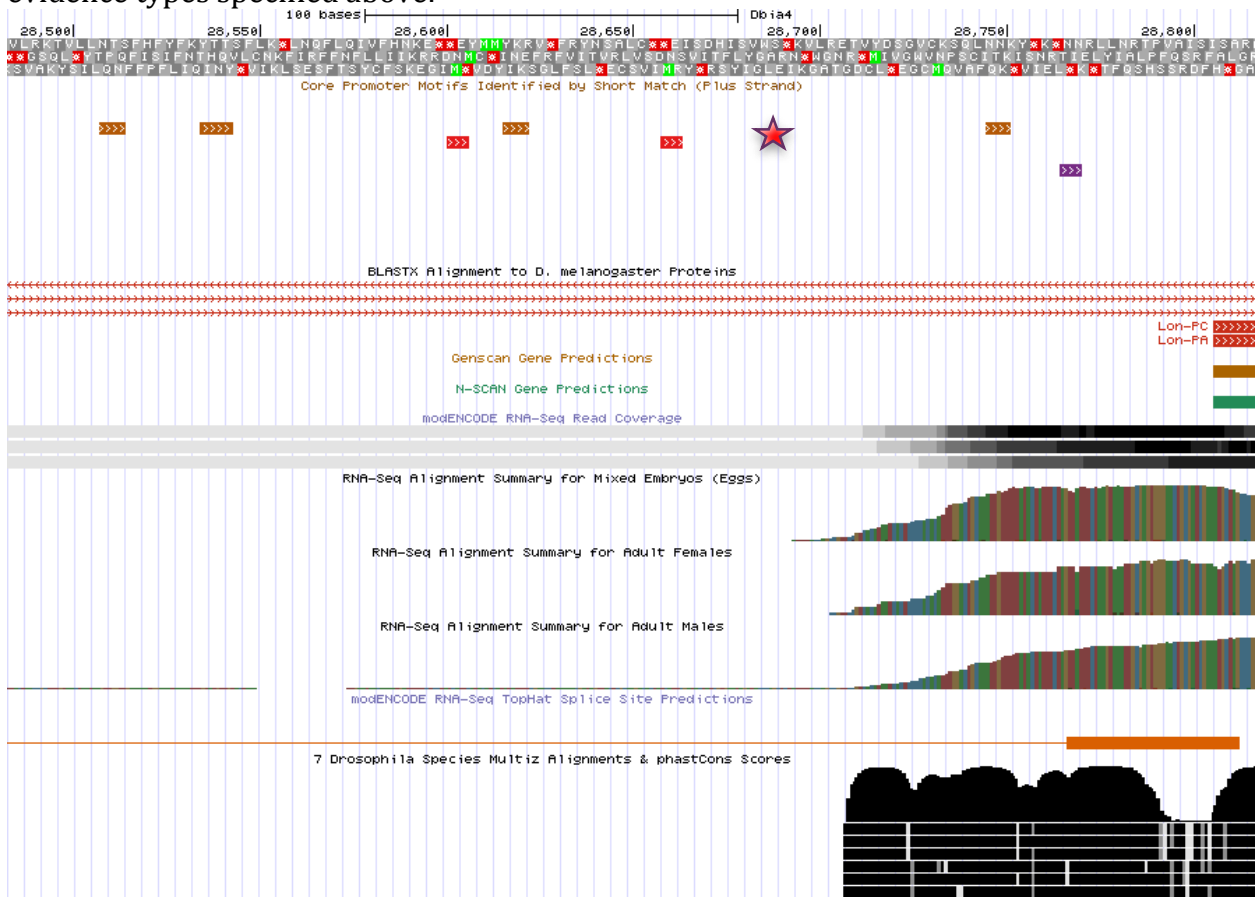
Gene-isoform name (i.e. dbia_ey-RA): dbia_Lon-PA
 Names of the isoforms with the same TSS as this isoform: dbia_Lon-PC
 Type of core promoter: (Peaked or Broad): Broad
 Coordinates of the first transcribed exon: 28,683-29,150
 Coordinate(s) of TSS position(s): 28,683
 Coordinate(s) of TSS search region(s): 28,600-28,805

1. Evidence that supports the TSS annotation postulated above

Specify the type of evidence used to support the TSS annotation:

Evidence type	Support TSS annotation?	Refute TSS annotation?
blastn alignment of the initial exon from <i>D. melanogaster</i>	X	
RNA-Seq coverage and TopHat splice junctions	X	
Core promoter motifs		X
Sequence conservation with other Drosophila species	X	
Other (please specify)		

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



The core promoters in the search region do not support my TSS prediction, indicated by the red star. Also, none of the core promoters would support the same TSS.

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 below:**

Dbia4_dna range=contig70:1-40000 5'pad=0 3'pad=0 strand=+ repeatMasking=none
 Sequence ID: lc||Query_39683 Length: 40000 Number of Matches: 11

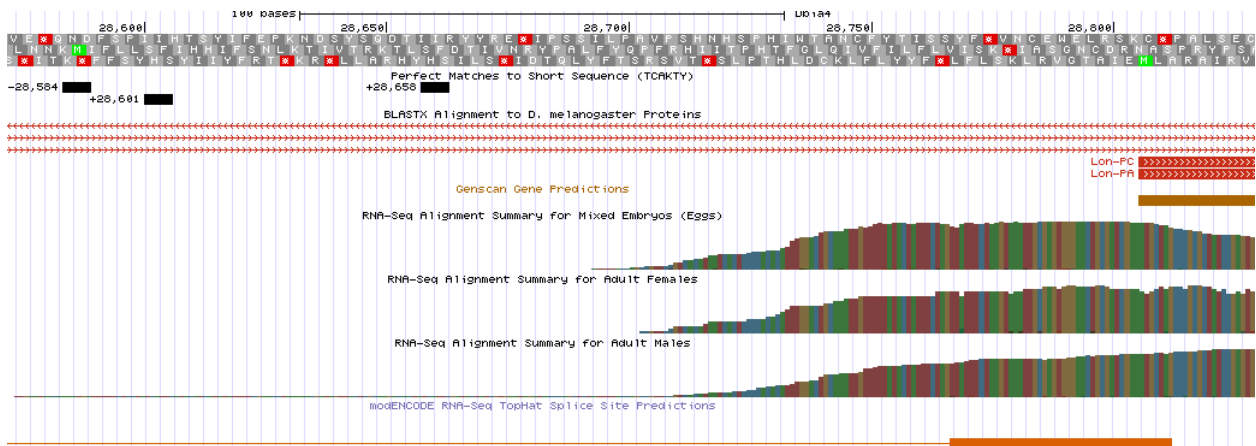
Range 1: 28693 to 29150 [Graphics](#)

▼ Next Match ▲ Prev

Score	Expect	Identities	Gaps	Strand
416 bits(290)	6e-119	377/458(82%)	11/458(2%)	Plus/Plus
Query 10	ATTTACCAGCCGT-----CACATAACTACTCACCACACATTTGGACTGCAAATGTTTT	64		
Sbjct 28693	ATTTACCAGCCGTTCCGTCACATAACTCCCCACACATTTGGACTGCAAATGTTTT	28752		
Query 65	TCTACCATTTCTAGTTATTTCTAAGTAAATTGCGAGTGGATATTGCTTTC AATATGTTAG	124		
Sbjct 28753	TATACTATTTCTAGTTATTTCTAAGTAAATTGCGAGTGGGAACTGCGATCGAAATGCTAG	28812		
Query 125	CCC GCGTATCCGAGTGC GCCCATGATGCGTGGCATCGCCTCGTCGT CAGTGTGGAACC	184		
Sbjct 28813	CCC GCGTATCCGAGTGC CCTATGATGCGGAGCATCGCCTCGTCGT CCGTGTGGACCC	28872		
Query 185	GGAATCGTCCCATT CAGAGTCCCTGATGCAATACTGCCGGGATCGGTCTTTCGCGCTCC	244		
Sbjct 28873	GCAACCGTCCC GCCCAGAGTCCCTGGTGAATGCTGCCGGGTTGGGGCAGCACCTCC	28932		
Query 245	AGCGGCTCCACGGAGCCAATTTGATGGTGCAGCGCTTCTACAGCCGCAAGCGGGATGATT	304		
Sbjct 28933	AGCGATTCATGGAGCAAACATGATGGTCCAACGTTTCTACAGCCGCAAGCGGGACGACT	28992		
Query 305	CCAACGGGGATATTAT -TATGG-----GACCCGATCTTATGTCCGATCAAGATACCCATC	358		
Sbjct 28993	CCGACGAGGATCTCATGGACGGTCATAATCCCGAGCTAATGTCCGATCGGGAAGCCAGT	29052		
Query 359	TTCCGGCAACTGTGGCGGTGCCGGACGTGTGGCCACATGTTCCGTTGTTGGCCATGCGCA	418		
Sbjct 29053	TGCCGGCCACTGTTGCGGTGCCGGATGTGTGGCCACATGTTCCGCTGTTGGCCATGCGAA	29112		
Query 419	AGAATCCTCTCTTTCCCGCTTATGAAGATAGTGGAG	456		
Sbjct 29113	AGAATCCCCTCTTTCCCGCTTATGAAGATTGTAGAG	29150		

If the TSS annotation is supported by RNA-Seq or RNA polymerase II data, **paste a Genome Browser screenshot of the region around the TSS (± 2 kb) with the evidence tracks listed below:**

10. Short Match results for the Inr motif (TCAKTY)
11. RNA-Seq Alignment Summary
12. RNA-Seq TopHat



There does not seem to be much difference in RNA expression in the different cell lines. All of the RNA-Seq alignments support my TSS prediction.

If the TSS annotation is supported by sequence conservation with other *Drosophila* species, **paste a screenshot of the pairwise alignment (e.g. from blastn, matcher) or the multiple sequence alignment (e.g. from clustalw, EvoPrinter, Multiz) below:**

```
Alignment block 1 of 1 in window, 28708 - 28832, 125 bps
B D D. biarmipes ccgtcacataatcactccccacacatttggactgcaaattg-tttttatactatttctagttatttctaa
B D D. melanogaster ccgtcacataatcactcaccacacatttggactgcaaattg-tttttctaccatttctagttatttctaa
B D D. yakuba ccgtcacataatcactcaccacacatttggactgcaaattg-tttttctactatttctagttatttctaa
B D D. erecta ccgtcacataatcactcaccacacatttggactgcaaattg-tttttctaccatttctagttatttctaa
B D D. eugracilis ccgtcacataatcactcctcacacactaggactgcaaattg-tttttataccatttctagatatttctaa
B D D. ficusphila ccgtcacataatcactcctcacacatttggactgcaaattgtttttttaccatttctagatatttctaa
B D D. takahashii ccgtcacataatcactcctcacacatttggg-tgcaaattg-tttttataccatttctagttatttctaa

D. biarmipes gtaaattgcgagtgggaactgcatcgaaatgctagcccgcgctatccgagtgcgt
D. melanogaster gtaaattgcgagtggatattgctttcaatatgtagcccgcgctatccgagtgcgc
D. yakuba gtaaattgcgagtggatattgctttcaatatgtagcccgcgctatccgagtgcgc
D. erecta gtaaattgcgagtggatactgctttcaatatgtagcccgcgctatccgagtgcgc
D. eugracilis gtaaattgcgattgggaattgtcatcaagatgctagcccgcgctatccgagtgcgt
D. ficusphila gtaaattgcgagtgggaattgtcatcgaaatgctagcccgcgctatccgagtgcgt
D. takahashii gtaaattgcgagtgggaattgccatcgagatgctagcccgcgctatccgagtgcgt
```

The conservation from Multiz only weakly supports my TSS prediction as the conservation does not start until 28,708.

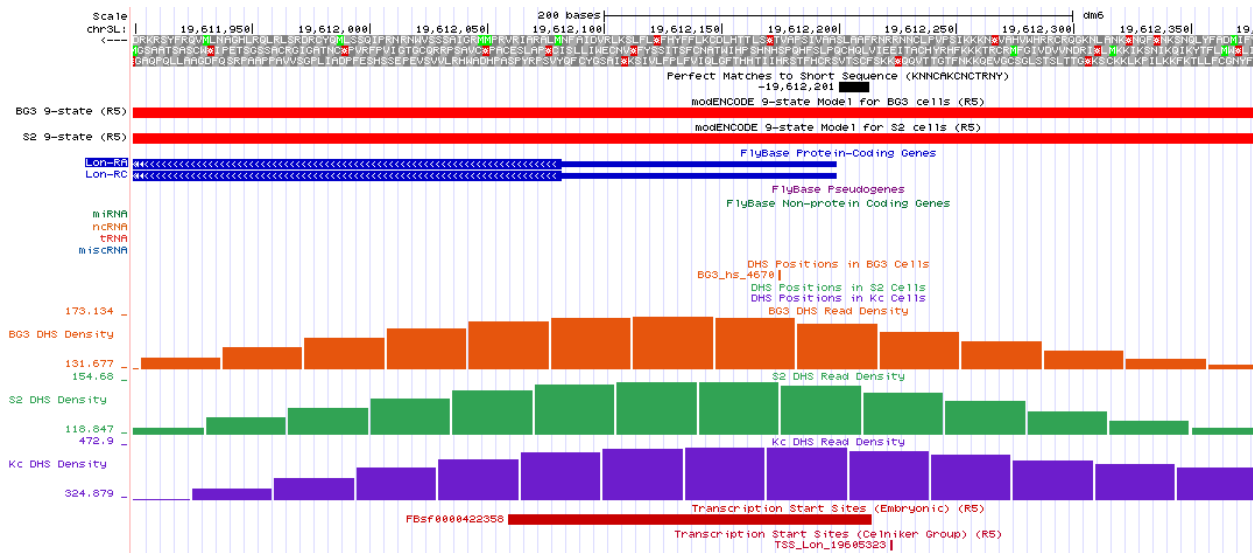
2. Search for core promoter motifs

Note: The consensus sequences for the *Drosophila* core promoter motifs are available at: http://gander.wustl.edu/~wilson/core_promoter_motifs.html

Use the "Short Match" functionality in the GEP UCSC Genome Browser to search for each of the core promoter motifs listed below in the region surrounding the TSS (± 300 bp) in your project and in the TSS of the *D. melanogaster* ortholog. (For TSS annotations where you can only define a TSS search region, you should search for the core promoter motifs in the entire TSS search region).

Record the **orientation and the start coordinate** (e.g. +10000) of each motif match below. (Enter "NA" if the motif is not present.)

Core promoter motif	Your project	<i>D. melanogaster</i>
BRE ^u	NA	NA
TATA Box	NA	NA
BRE ^d	+28403, +28435, +28463, +28508, +28535, +28537, +28616, +28745	-19612078, -19612137, - 19612279
Inr	+28601, +28658	-19612187
MTE	NA	NA
DPE	+28765	-19611965, -19612093, - 19612118
Ohler_motif1	NA	NA
DRE	NA	NA
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	NA	-19612201
Ohler_motif8	NA	NA



The *D. melanogaster* sequence has an Ohler_motif7 just upstream of the TSS. This could be significant and something to investigate more, as this motif is relatively rare to occur by chance.

Gene report form

Gene name (i.e. *D. mojavensis eyeless*): *D. biarmipes asf1*

Gene symbol (i.e. dmoj_ey): dbia asf1

Approximate location in project (from 5' end to 3' end): 21025-20375

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:

Name(s) of unique isoform(s) based on coding sequence	List of isoforms with identical coding sequences
asf1-PA	asf1-PB

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 your project consensus sequence:

All 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 which supports the hypothesis of errors in the consensus sequence include: CDS alignment with frame-shifts or in-frame stop codons, multiple RNA-seq reads with discrepant alignments compared to the project sequence, multiple high quality discrepancies in the *Consed* assembly.

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

Using the Sequencer Updater (available through the GEP web site under "Projects" -> "Annotation Resources"), create a VCF (Variant Call Format) file that describes the changes

to the consensus sequence you have identified above. **Paste a screenshot with the list of sequence changes below:**

Isoform report form

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

Gene-isoform name (i.e. dmoj_ey-PA): dbia asf1-PA

Names of the isoforms with identical coding sequences as this isoform
dbia asf1-PB

Is the 5' end of this isoform missing from the end of project: No

If so, how many exons are missing from the 5' end: _____

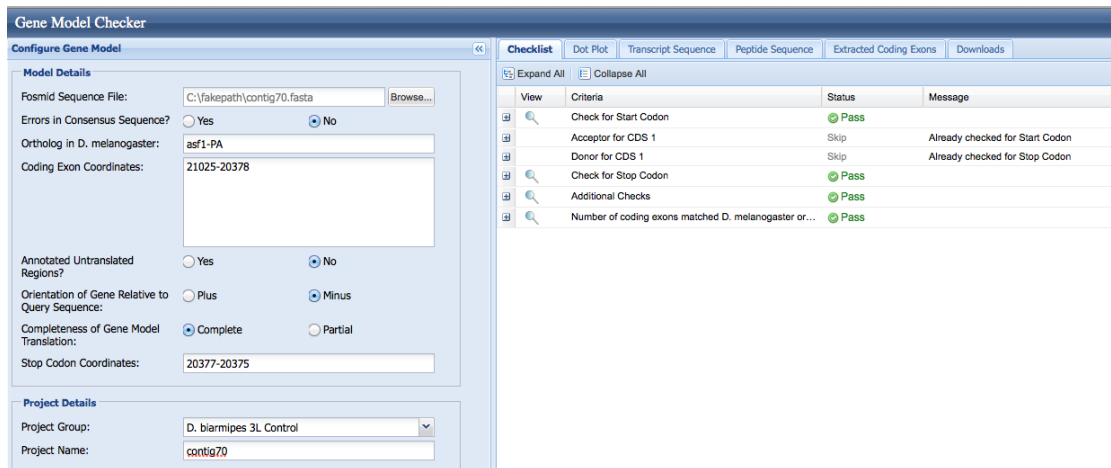
Is the 3' end of this isoform missing from the end of the project: No

If so, how many exons are missing from the 3' end: _____

1. 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 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 revise the submitted exon coordinates automatically using this VCF file.

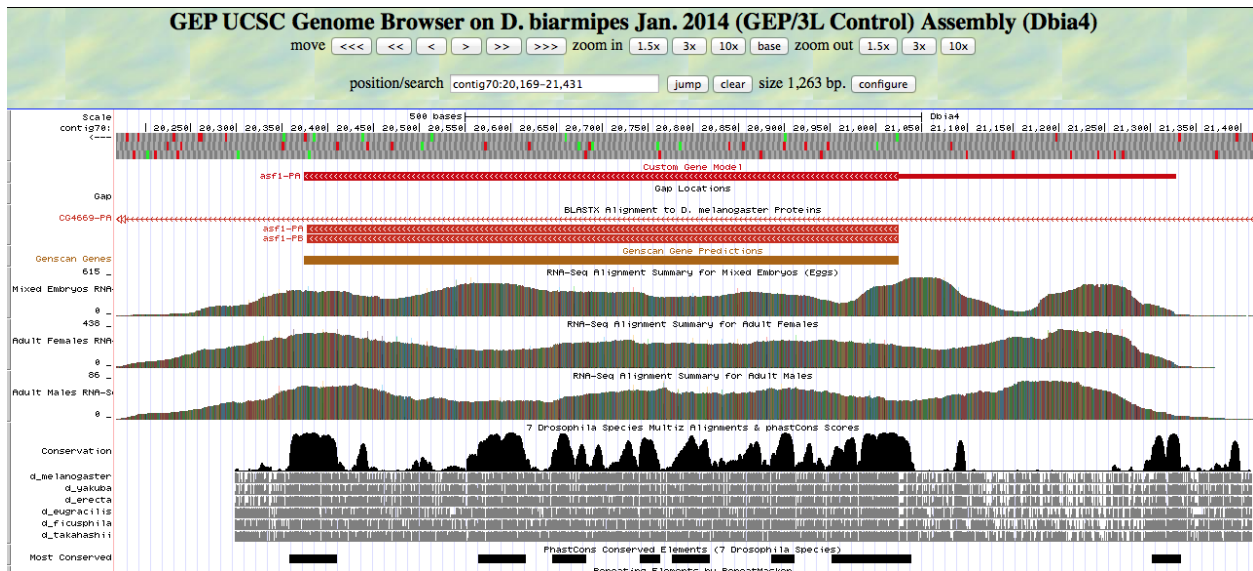


2. View the gene model on the Genome Browser

Using the custom track feature from the Gene Model Checker (see page 10 of the Gene Model Checker user guide on how to do this; you can find the guide under "Help" -> "Documentations" -> "Web Framework" on the GEP website at <http://gep.wustl.edu>).

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

21. A sequence alignment track (D. mel Protein or Other RefSeq)
22. At least one gene prediction track (e.g. Genscan)
23. At least one RNA-Seq track (e.g. RNA-Seq Alignment Summary)
24. A comparative genomics track (e.g. Conservation, D. mel. Net Alignment, 3-way, 5-way or 7-way multiz)



Paste the screenshot of your gene model as shown on the Genome Browser 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 use the protein alignment generated by the Gene Model Checker or you can generate a new alignment using BLAST 2 Sequences (*bl2seq*). Paste a screenshot of the protein alignment below:

Alignment of asf1-PA vs. Submitted_Seq

[View plain text version](#)

Identity: 206/218 (94.5%), Similarity: 210/218 (96.3%), Gaps: 2/218 (0.9%)

```

asf1-PA      1 MAKVHITNVVLDNPSFFNPFQPELTFECIEELKEDLEWKMIYVGSASEEHQVLDTI 60
Submitted_Seq 1 MAKVHITNVVLDNPSFFNPFQPELTFECIEELKEDLEWKMIYVGSASEEHQVLDTI 60
asf1-PA     61 YVGPVPEGRHIFVQADPPDVSKIPEPDVAVGTIVLLTCSYRQGFVRVGYVNDYADF 120
Submitted_Seq 61 YVGPVPEGRHIFVQADPPDVSKIPEPDVAVGTIVLLTCSYRQGFVRVGYVNDYADF 120
asf1-PA     121 EMRENPPKPLPEKLTRNILASKPRVTRFKINWDYGHINGNGVNGHODEMATDGFST 180
Submitted_Seq 121 EMRENPPKPLPEKLTRNILASKPRVTRFKINWDYGHINGN--GVNGHDEEMVTDGFST 178
asf1-PA     181 SEAASAVIHPEDDNSLAMPENGIKALNENSNSLAMEC 218
Submitted_Seq 179 SEVASVVVQPEDDNSLAMPENGIKALNENSNSLAMEC 216

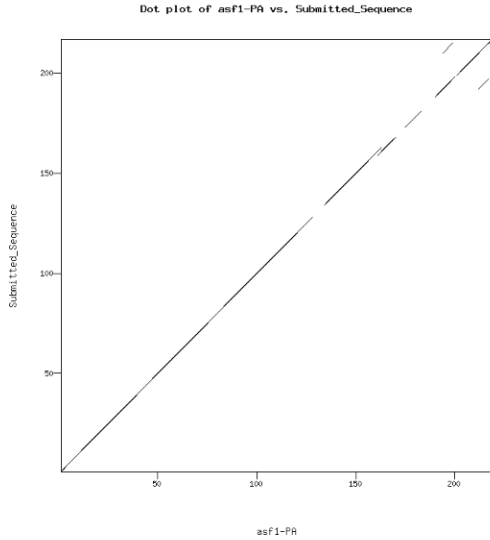
```

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). Provide an explanation for any anomalies on the dot plot (e.g. large gaps, regions with no sequence similarity).



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



The two lines near the end of the plot arise from the alignment with another part of the sequence due to a repetitive segment. The sequence that is repeated is “NSLAM”. It is seen twice in the last line of the protein alignment figure.

Transcription start sites (TSS) report form (optional)

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

Name(s) of isoform(s) with unique TSS	List of isoforms with identical TSS
asf1-PA	asf1-PB

Complete this report form for each unique TSS listed in the table above (copy and paste to create as many copies of TSS report form as needed):

Gene-isoform name (i.e. dbia_ey-RA): dbia_asf1-PA

Names of the isoforms with the same TSS as this isoform: dbia_asf1-PB

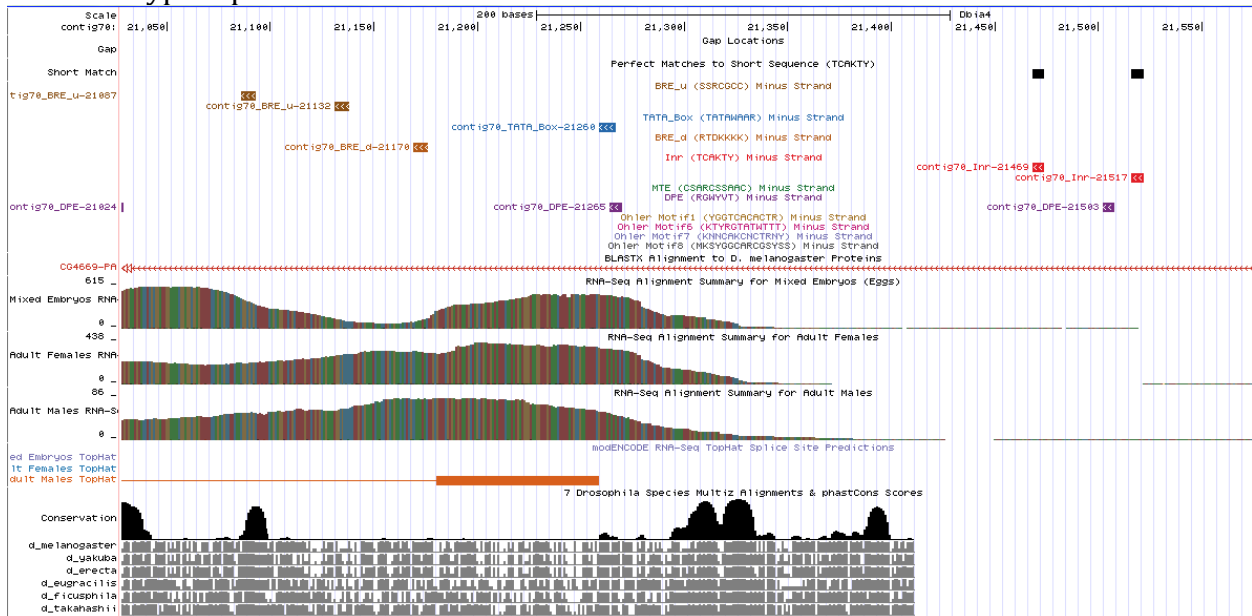
Type of core promoter: (Peaked or Broad): Peaked
 Coordinates of the first transcribed exon: 21,329-19,824(asf1-PA); 21,329-21,200(asf1-PB)
 Coordinate(s) of TSS position(s): 21,329
 Coordinate(s) of TSS search region(s): 21,200-21,500

1. Evidence that supports the TSS annotation postulated above

Specify the type of evidence used to support the TSS annotation:

Evidence type	Support TSS annotation?	Refute TSS annotation?
blastn alignment of the initial exon from <i>D. melanogaster</i>	X	
RNA-Seq coverage and TopHat splice junctions	X	
Core promoter motifs		X
Sequence conservation with other <i>Drosophila</i> species	X	
Other (please specify)		

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



None of the surrounding core promoters seem to support my TSS prediction or strongly support another location. It is interesting to note the variety of motifs in the search region. This variety was not seen in any of the other genes in the project.

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 below:**

Dbia4_dna range=contig70:1-40000 5'pad=0 3'pad=0 strand=+ repeatMasking=none
 Sequence ID: lcl|Query_32003 Length: 40000 Number of Matches: 22

Range 1: 20147 to 21329 [Graphics](#)

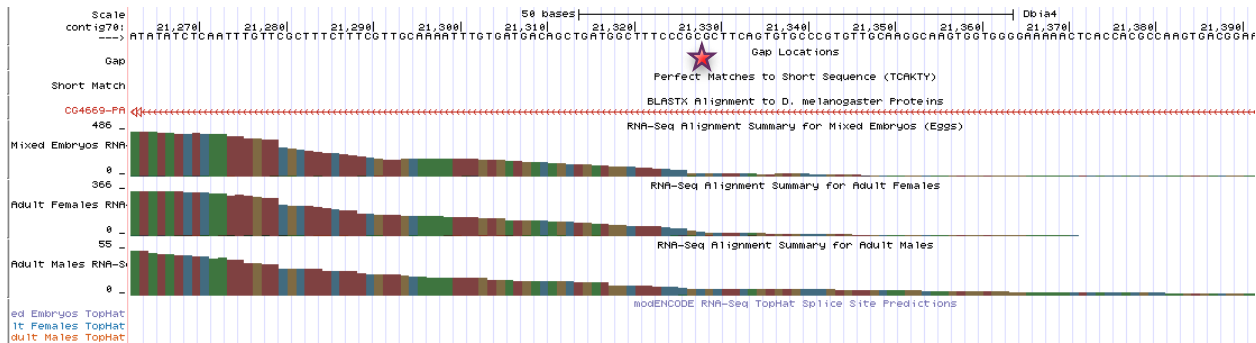
▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
900 bits(629)	0.0	929/1191(78%)	60/1191(5%)	Plus/Minus
Query 1	CGCGGGAAAGAAGTCAGCTGTCCACAT---TTGTGCGAAGACA---AGCGGAATAATTGC	54		
Sbjct 21329	CGCGGGAAAGCCATCAGCTGTATCACAAATTTTGAACGAAAGAAAGCGAACAAATTGA	21270		
Query 55	GATAAATAAAAGCAGAATAGCGACAGACAAT-TGCACAAACCTGATTCATGTTGTAACCA	113		
Sbjct 21269	GATATATAAACGAAAAATT-CGGCATAAAAAGTTTACAAACCTTACCCATCATATTGAAA	21211		
Query 114	AAACAACTACTAGCATTACAGTACATATATCAAGTATTTGCTGAGCGTTTTGCAGGTGA	173		
Sbjct 21210	TATCAATCAGCTAGCATTACATCACGAAACGCAAGTATTTGGCGAGTGTTCTGCAGGTTG	21151		
Query 174	ATTTTCGCAAAG-----CAATTCCTAGTCGGACAAAAGCAGAACCT-CTAAAAAAGCG	226		
Sbjct 21150	ATTTTCGACAAGGGGGCGCCATTTTGCACAGAACAAAAGCAGAACCTGCTGAAAAAAGCG	21091		
Query 227	CGCAGTCGAAGGAGTCTTTAAAATATCGGTCTACGCTCTGCAATCTCCGAGGTACA----	282		
Sbjct 21090	CGCCGTTGAAGGCGTCTGGAAAACTCCGCCTCCGTTTTCCAATCCGCGAGGTACATTCA	21031		
Query 283	--GTCATGGCCAAGGTGCACATCACCAACGTGGTGGTGGTGGACAACCCGAGCAGCTTCT	340		
Sbjct 21030	CAGTCATGGCCAAGGTGCACATCACCAATGTGGTGGTGGTGGACAACCCGAGCAGCTTTT	20971		
Query 341	TCAACCCCTTTTCAGTTCGAACTCACGTTTCGAGTGCATTGAGGAGCTAAAAGAGGATCTAG	400		
Sbjct 20970	TCAACCCCTTCCAGTTCGAACTGACCTTTGAGTGCATCGAGGAGCTGAAGGAAGACCTCG	20911		

The entire BLAST alignment would not fit in one screen-shot since it included both the UTR's and the coding exon. The alignment predicts a TSS at 21,329.

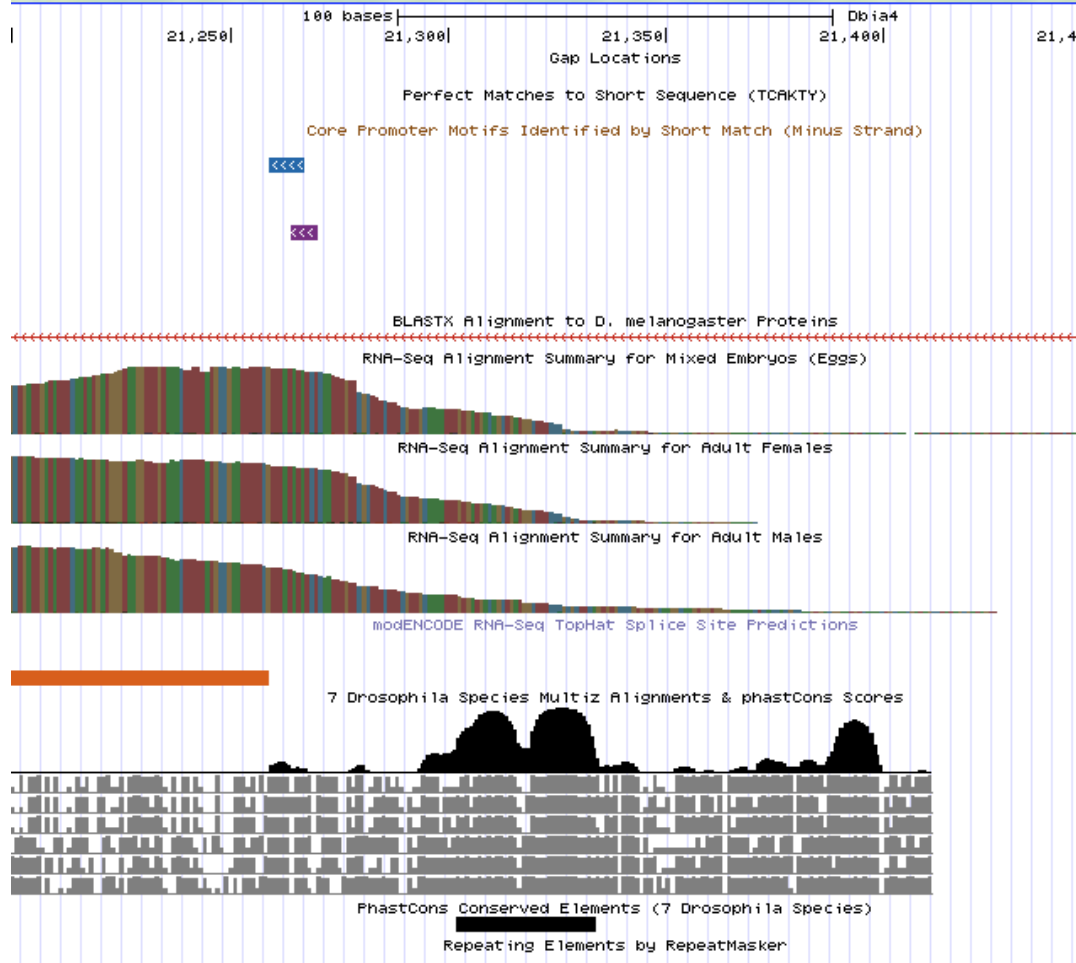
If the TSS annotation is supported by RNA-Seq or RNA polymerase II data, **paste a Genome Browser screenshot of the region around the TSS (± 2 kb) with the evidence tracks listed below:**

- 13. Short Match results for the Inr motif (TCAKTY)
- 14. RNA-Seq Alignment Summary
- 15. RNA-Seq TopHat



The RNA-seq alignments seem to support my TSS prediction, indicated by the red star, very well. The data depth increases soon after the TSS.

If the TSS annotation is supported by sequence conservation with other *Drosophila* species, **paste a screenshot of the pairwise alignment (e.g. from blastn, matcher) or the multiple sequence alignment (e.g. from clustalw, EvoPrinter, Multiz) below:**



Alignment block 1 of 1 in window. 21110 - 21411. 302 bps

```

B D D. biarmipes      tgcttttgttctgtgcagaaatggcgcccccttgcgaaatcaacctgcagaacctcgccaaacttg
B D D. melanogaster   tgcttttgcgactaggggaattg-----ct-ttgcgaaatcacctgcaaacgctcagcaaatacttg
B D D. yakuba         tgggttagcccgctctaggggaattgcgtgtct-ttgcgaaatgcacctgacaaacgctcgacaacacttg
B D D. erecta         tgcttttgcgocgctggggaattgcgtgtct-ttgcaaaatccacctgttaaacgctcggcaaacacttg
B D D. eugracilis     tacttctgttgcgagcaggacttacgtgctt-tgatgaaatacacctgctaaacactctactaacagttg
B D D. ficusphila     tgcttttgttccgtgcaggaaaaaacacct-tgaaaaagtgcgactgctgaaaccggatcaactcttg
B D D. takahashii    tgcttttgttccggcaggaatttcgttccc-tgctgaaatttgctgctaaacactcgacaactcacttg

D. biarmipes      cgtttcgtgatgtaatgctagctgattgatattcaatatgatgggtaaggtttgtaaacttttatgcc
D. melanogaster   atatatgtactgtaatgctagtagttgttttggttacaacatgaatcagggtttgtgcaattgtctgtcg
D. yakuba         atattcgtagtgggaatgctagtgattgttttggttataaatattaatcaggttcgtgcaattgtctgtcg
D. erecta         atatttgtagtgggaatgctagtaattgttttggttataacataaatcagggtttgtgcaattgtctgtcg
D. eugracilis     tagttcgttgtaacaatataaatgattaggtttgtaataatgactaagattcgtctactttcaaggct
D. ficusphila     cgttttgtgatgcaataatttctgtattgtttagtgaactttcttttccgggttcgtgactctctgcag
D. takahashii    ctgttcgcgatgcaatgctagctgattgtttgcgtaaatatttggatcagggtttgttcttttccactg

D. biarmipes      gaa-tttttcgtttatataatcctcaatttgttcg-----ctttctttcgttgcaaaatttggatgaca
D. melanogaster   cta-ttctgcttttattatcgcaattattccg-----cttgctcttcgcacaaa-----tgggtgaca
D. yakuba         cta-ttctgagttttatataatcgcaattattccg-----cttcttttctcacaattttgtggtgaca
D. erecta         cta-ttttgagttttattatcacaattattccg-----cttcttttctcacaatttctggtgaca
D. eugracilis     tta-ttcttcttttattatcggaatttcttcg-----ctttctttgctcgcgcaaaatctgtgatgaca
D. ficusphila     cgacttttaggtttatttattgtaattctttcg-----cctagttgctcgcgcaaaatttggatgaca
D. takahashii    gta-atttccgtttatttattgcaagaagtctcgtttcttcttttctttgctcgcgcaaaatttggatgaca

D. biarmipes      gctgatggctttcccgcgcttcagtggtgcccgtgttgcaaggcaagtgggtgggaaaaactcaccacgcc
D. melanogaster   gctgacttctttcccgcgcttcagagggaccgttcggtgtcgcaagtgggtggcgaaaaactcaccacgac
D. yakuba         gctgactgctttcccgcgcttcagtggtgaccgttctgtgacgcaagaggttgcggaaaaactcaccacgac
D. erecta         gctgacttctttcccgcgcttcaaagtgaccgttttgggtggcgcaagtgggtggcgaaaaactcaccacgac
D. eugracilis     gctgacggctttcccgcgcttcagtggtgaccgtt-----tagtgggtaaaatttcacaacgaa
D. ficusphila     gctgacggctttcccgcgcttcagtggtgaccgttttgaggagcaagcgggtgggtaaaactcaccacgac
D. takahashii    gctgacggctttcccgcgcttcagtggtgaccgttttccgacgcaagtgggtggcgaaaaactcaccacaac

D. biarmipes      aagtgacggaatgtcaaaactgcaacgatag
D. melanogaster   aagtggcggaatgtcaaaacgcgacaatag
D. yakuba         aagtttcggaatgtcaaaatacaatgatag
D. erecta         aagtggcggaatgtcaaaatacaacgatag
D. eugracilis     aagtgacggaatgtcaaaactgacgcgtag
D. ficusphila     aagtgacggaatgtcaaaacgcgcgatag
D. takahashii    aagtgacggaatgtcaaaactgcaacgatag

```

The Multiz alignments show a large increase in depth, at 21,234, just prior to my predicted TSS. The red box in the above figure shows what bases the peak refers to when looking at the UCSC Genome Browser overall.

2. Search for core promoter motifs

Note: The consensus sequences for the *Drosophila* core promoter motifs are available at: http://gander.wustl.edu/~wilson/core_promoter_motifs.html

Use the "Short Match" functionality in the GEP UCSC Genome Browser to search for each of the core promoter motifs listed below in the region surrounding the TSS (± 300 bp) in your project and in the TSS of the *D. melanogaster* ortholog. (For TSS annotations where you can only define a TSS search region, you should search for the core promoter motifs in the entire TSS search region).

Record the **orientation and the start coordinate** (e.g. +10000) of each motif match below. (Enter "NA" if the motif is not present.)

Core promoter motif	Your project	<i>D. melanogaster</i>
BRE ^u	-21087, -21132	NA
TATA Box	-21260	NA
BRE ^d	-21170	+19618608, +19618652, +19618660, +19618666, +19618668, +19618715, +19618774, +19618776, +19618919, +19618995, +19619041
Inr	-21469, -21517	NA
MTE	NA	NA
DPE	-21024, -21265, -21503	+19618747, +19618783, +19619175
Ohler_motif1	NA	NA
DRE	NA	NA
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	NA	NA
Ohler_motif8	NA	NA

None of the nearby core promoters support my TSS prediction

Gene report form

Gene name (i.e. *D. mojavensis eyeless*): *D. biarmipes ms(3)76Cc*

Gene symbol (i.e. dmoj_ey): *dbia ms(3)76Cc*

Approximate location in project (from 5' end to 3' end): 19540-15863

Number of isoforms in *D. melanogaster*: 1

Number of isoforms in this project: 1

Complete the following table for all the isoforms in this project:

Name(s) of unique isoform(s) based on coding sequence	List of isoforms with identical coding sequences
ms(3)76Cc-PA	

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 your project consensus sequence:

All 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 which supports the hypothesis of errors in the consensus sequence include: CDS alignment with frame-shifts or in-frame stop codons, multiple RNA-seq reads with discrepant alignments compared to the project sequence, multiple high quality discrepancies in the *Consed* assembly.

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

Using the Sequencer Updater (available through the GEP web site under “Projects” -> “Annotation Resources”), create a VCF (Variant Call Format) file that describes the changes to the consensus sequence you have identified above. **Paste a screenshot with the list of sequence changes below:**

Isoform report form

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

Gene-isoform name (i.e. dmoj_ey-PA): dbia ms(3)76Cc-PA

Names of the isoforms with identical coding sequences as this isoform

Is the 5' end of this isoform missing from the end of project: No

If so, how many exons are missing from the 5' end: _____

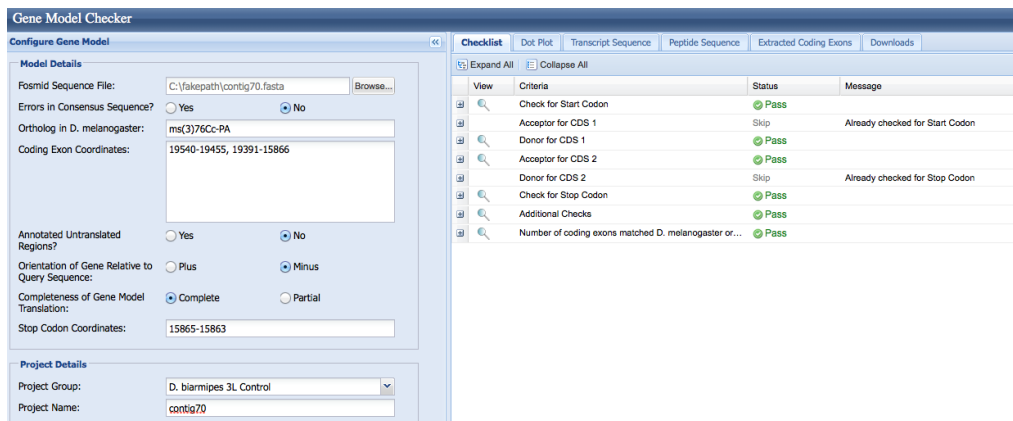
Is the 3' end of this isoform missing from the end of the project: No

If so, how many exons are missing from the 3' end: _____

1. 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 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 revise the submitted exon coordinates automatically using this VCF file.

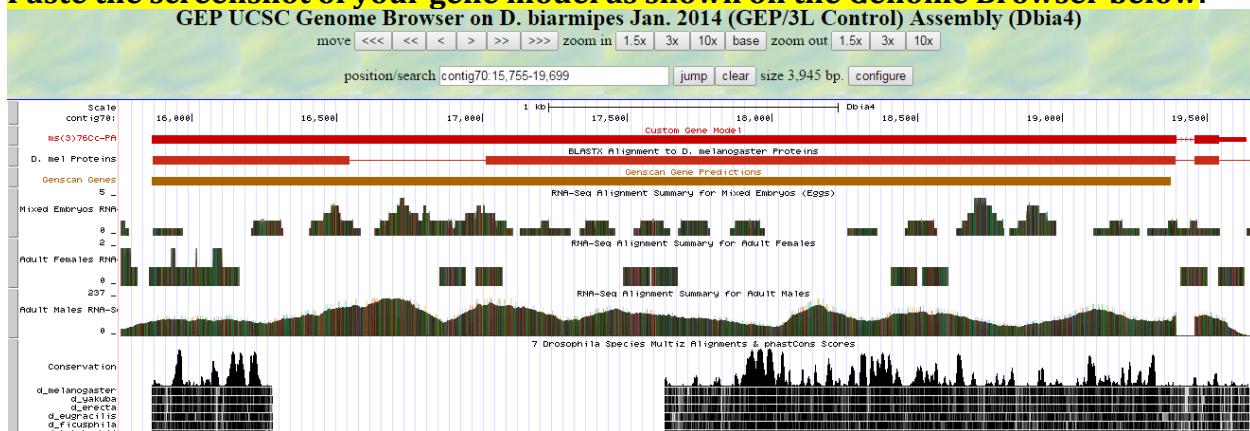


2. View the gene model on the Genome Browser

Using the custom track feature from the Gene Model Checker (see page 10 of the Gene Model Checker user guide on how to do this; you can find the guide under “Help” -> “Documentations” -> “Web Framework” on the GEP website at <http://gep.wustl.edu>). 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. Include the following evidence tracks in the screenshot if they are available.

25. A sequence alignment track (D. mel Protein or Other RefSeq)
26. At least one gene prediction track (e.g. Genscan)
27. At least one RNA-Seq track (e.g. RNA-Seq Alignment Summary)
28. A comparative genomics track (e.g. Conservation, D. mel. Net Alignment, 3-way, 5-way or 7-way multiz)

Paste the screenshot of your gene model as shown on the Genome Browser 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 use the protein alignment generated by the Gene Model Checker or you can generate a new alignment using BLAST 2 Sequences (*bl2seq*). **Paste a screenshot of the protein alignment below:**

Alignment of ms(3)76Cc-PA vs. Submitted_Seq

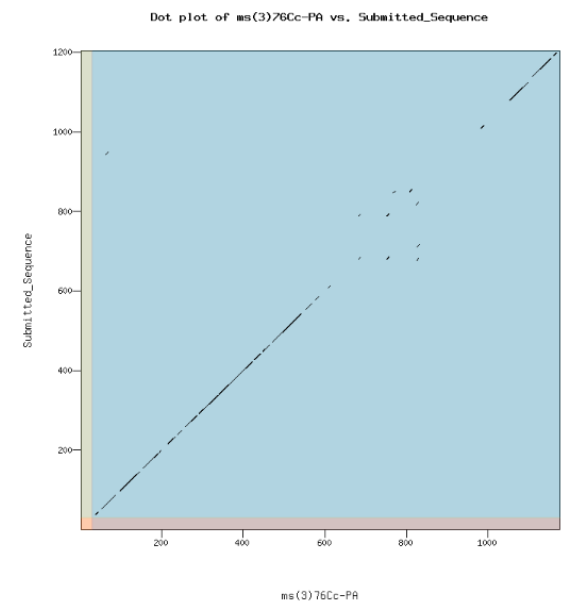
[View plain text version](#)

Identity: 765/1225 (62.4%), Similarity: 892/1225 (72.8%), Gaps: 72/1225 (5.9%)

ms(3)76Cc-PA	1	MPISQHDLARFASADGSPRISMSNSNWA LSYSDLDLQSGFCGGSSRTPEPCDHRKFC	60
Submitted_Seq	1	MPISQDLRLRLASTPSCDPLRTKQTSN YISYTDMLGSGKFCGSGSPRGASCDHRKFC	60
ms(3)76Cc-PA	61	SKVNRSDVDADSGCPVGLVCSNATVNPQCGSPKQGANVVFATYTKAFATYVLPPE	120
Submitted_Seq	61	SKVNRSEVZADSLGSPVGLVCSNATVNPQCGSSPQGANVVFATYTKAFATYVLPPE	120
ms(3)76Cc-PA	121	DDVFAKIDMLASGQDFAASDVAQCHDSDPDRHLLAEPPIYVNRSEKTKRNFNTEGHR	180
Submitted_Seq	121	DDVFAKIDMLASGQDFAASDVAQCHDSDPDRHLLAEPPIYVNRSEKTKRNFNTEGHR	180
ms(3)76Cc-PA	181	PTLAEPRVYGVGSPKFDQDPPIHNTGTHLQSPFKSRVYCMALTRVGHLLVRRRNVV	240
Submitted_Seq	181	PTLAEPRVYGVGSPKFDQDPPIHNTGTHLQSPFKSRVYCMALTRVGHLLVRRRNVV	240
ms(3)76Cc-PA	241	AVLDVKGKRDDEAVVDGCVMLVGLSNDVWVHLSNTEGHSFDEFSFRELAVVWRLE	300
Submitted_Seq	241	AVLDVKGKRDDEAVVDGCVMLVGLSNDVWVHLSNTEGHSFDEFSFRELAVVWRLE	300
ms(3)76Cc-PA	301	FPDCHYVDRFDSHRSFTEFVWNSYAVYKGVLELSLNDPVRNRSSIAVAVGSSVLAHQ	360
Submitted_Seq	301	FPDCHYVDRFDSHRSFTEFVWNSYAVYKGVLELSLNDPVRNRSSIAVAVGSSVLAHQ	360
ms(3)76Cc-PA	361	DHPANDNMLDRLISYGVLCBSCSDCIRDRPDIIDTFPQIRMGQVYLSAKIATAV	420
Submitted_Seq	361	DHPANDNMLDRLISYGVLCBSCSDCIRDRPDIIDTFPQIRMGQVYLSAKIATAV	420
ms(3)76Cc-PA	421	AVGHRGCVRIITQDFEARVSOALSELGNVWFQINQVAVIWKDGFVYVDPVREITVC	480
Submitted_Seq	421	AVGHRGCVRIITQDFEARVSOALSELGNVWFQINQVAVIWKDGFVYVDPVREITVC	480
ms(3)76Cc-PA	481	THVANDREGAKAVNRFDDQVLSVDFOLDEKSNQSAIVVVRIRNIAECDEGVA	540
Submitted_Seq	481	THVANDREGAKAVNRFDDQVLSVDFOLDEKSNQSAIVVVRIRNIAECDEGVA	540
ms(3)76Cc-PA	541	LPMDGDKDCEVSLNTEFFFDQCVAVCNKSAIETSDYEDVYSILADKSDIDDFQV	600
Submitted_Seq	541	LPMDGDKDCEVSLNTEFFFDQCVAVCNKSAIETSDYEDVYSILADKSDIDDFQV	600
ms(3)76Cc-PA	601	LDQNGGACDFCFE-EFATF-----PKR-RP-----KKNVNSKRVG-----LKKRFQ	645
Submitted_Seq	601	LDQNGGACDFCFE-EFATF-----PKR-RP-----KKNVNSKRVG-----LKKRFQ	645
ms(3)76Cc-PA	646	DKDPS-----GKMLP-----FASRFPND-----KDSPLDKKRSAN-----RQAGGRS---	686
Submitted_Seq	646	DLAFTSPITGQLVKSIVRQGGRRSNQSPAKSSVRFPTKSDVTTGRMKTFRQGGGTGTCI	720
ms(3)76Cc-PA	687	STKSPLTKSREN-FSNRIAKRSVLPKQV-----ELQTHFPKAPVPTPVVAPPGRIVR	743
Submitted_Seq	721	SAKSPERARFKEGLTPGRIMKNNITAFQTCNRDSSPSPERFKISTMPEVLSP-GKTLK	779
ms(3)76Cc-PA	744	STTKTEVGVGRSSNRSONNTLGRSRSREPPFERITTKKMQPKROAGMCS-----RDD	797
Submitted_Seq	780	KGIKTKRIQGRSSNHSPT-----KPSRSQDGSSTIKTK-BTSRQTCMGSSPTTKTKRKC	834
ms(3)76Cc-PA	798	STSPERC-----KPREGSPDRTIKKLVKSSRQGGVYCSPPVSVSLVYSKEVDVYPR	853
Submitted_Seq	835	CAVTTAMCSAHSRSREGSPGRIMKKNIKSPRESGGRSFNHPSPCSV-CESRHGSPFR	893
ms(3)76Cc-PA	854	IMIKSENTKDSGDTKSHLRKSRNDRSPGKTPGTPSRKMAKVDKSKGANPALDLSG	913
Submitted_Seq	894	TVLRSS-----KGNDSNGRGVFNKAP-----ITSDAASKLLPHSAISK--GPYTVNVR	945
ms(3)76Cc-PA	914	SEAIRTVQANKEASDRKGLASGVMLLRKVNKTDIALEPSKKAQPHILQGVSKS	973
Submitted_Seq	946	YRSNDTPTATNATPHKAE-----MYQBEPRERATVNSAKMVAQTHMLKQLSKR	999
ms(3)76Cc-PA	974	DLGVPAALAAHLNFPASNFWELDQKPEFYPKMNVRMAEQSEHFKNCCORDAERDS	1033
Submitted_Seq	1000	DLGVVPMANHLNFPVSDVFRQIDQKTTSLPASALIAREPHBEPGCCCGADNATV	1059
ms(3)76Cc-PA	1034	IGTSRMPVKFCPSRAPHILAVAGSBSQTVSSINRVLSGAPKVANRVLMTMPGNVYV	1093
Submitted_Seq	1060	STLQVSLPKYGFNRVQLLAVAGSBSQTVSSINRVLSGAPKVANRVLMTMPGNVYV	1119
ms(3)76Cc-PA	1094	HHDFVRS-----AAVYVYDFDCDCIDRFRHLDLSTQAGLIAFRKQGVNPKHIDSR	1149
Submitted_Seq	1120	AHPNYSRSDSAASFWYDFDCDCIDRFRHLDLSTQAGLIAFRKQGVNPKHIDSR	1179
ms(3)76Cc-PA	1150	SKAKMLDQSDHDPKIKQLLALR	1174
Submitted_Seq	1180	SKAATLRLNQLDDPPKRMOKLLVR	1204

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). **Provide an explanation for any anomalies** on the dot plot (e.g. large gaps, regions with no sequence similarity).

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



Relative to many other orthologs, these sequences do not have a large amount of identity. They are 62.4% identical. The first exon is very short and also does not have the best alignment. When compared to the entire sequence in the dot plot, it appears to not align at all. This is not the case, as seen in the protein alignment figure. The middle of the second exon does not align very well and many gaps are forced in both sequences.

Transcription start sites (TSS) report form (optional)

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

Name(s) of isoform(s) with unique TSS	List of isoforms with identical TSS
ms(3)76Cc-PA	

Complete this report form for each unique TSS listed in the table above (copy and paste to create as many copies of TSS report form as needed):

Gene-isoform name (i.e. dbia_ey-RA): dbia_ms(3)76Cc-PA
 Names of the isoforms with the same TSS as this isoform: _____
 Type of core promoter: (Peaked or Broad): Peaked
 Coordinates of the first transcribed exon: 19,634-19455
 Coordinate(s) of TSS position(s): 19,636
 Coordinate(s) of TSS search region(s): 19,541-19,683

1. Evidence that supports the TSS annotation postulated above

Specify the type of evidence used to support the TSS annotation:

Evidence type	Support TSS annotation?	Refute TSS annotation?
blastn alignment of the initial exon from <i>D. melanogaster</i>	X	
RNA-Seq coverage and TopHat splice junctions	X	
Core promoter motifs	X	
Sequence conservation with other Drosophila species	X	
Other (please specify)		

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

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 below:**

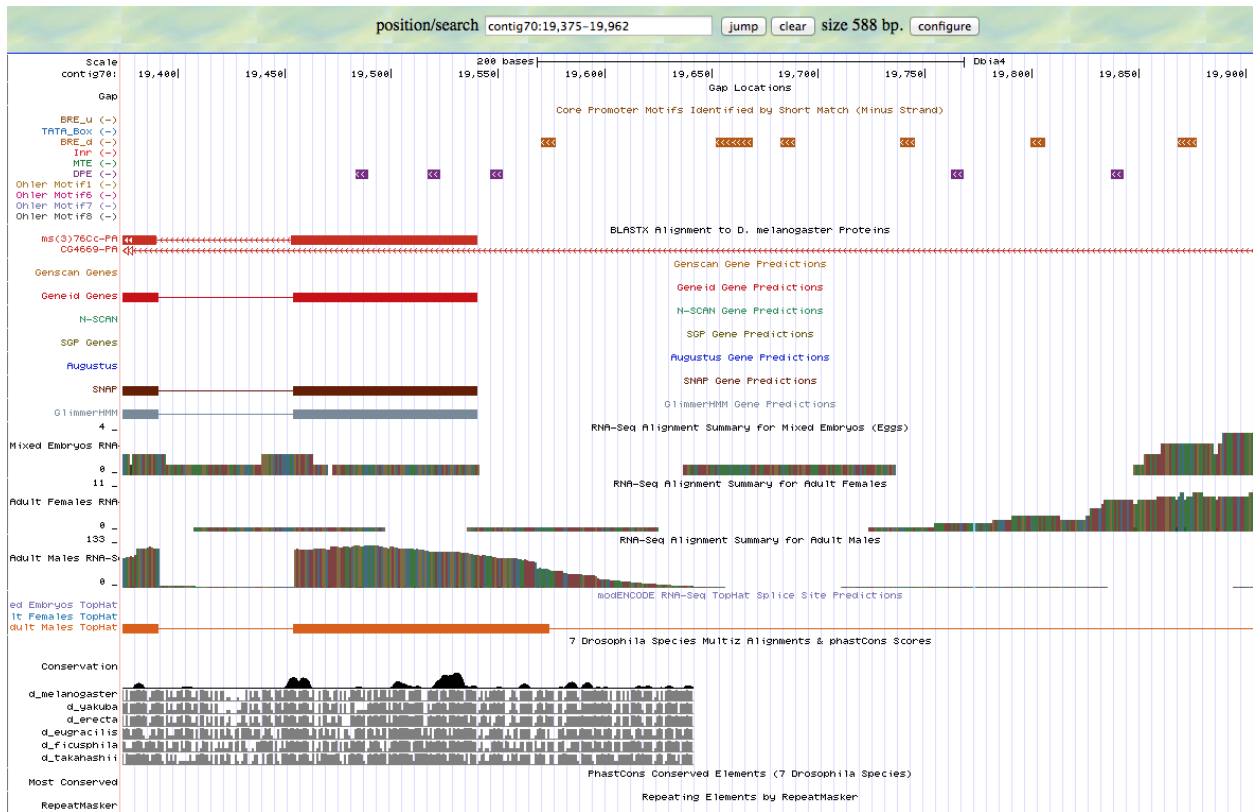
Dbia4_dna range=contig70:1-40000 5'pad=0 3'pad=0 strand=+ repeatMasking=none
 Sequence ID: lcl|1031 Length: 40000 Number of Matches: 31

Range 1: 19455 to 19634 [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
92.5 bits(63)	9e-22	129/191(68%)	11/191(5%)	Plus/Minus
Query 1	CTTCCTTTGTGCAATATTCAGCAACAATTACATTAGAACAAAAACAGAAAAAACTCAAA	60		
Sbjct 19634	CTTCCTATCTAAAATACTCAGCTAAATTTAAGTAGAACAAAACT--AAAAAAGGAAAA	19577		
Query 61	TCTCAAATCTTTGCTGTACAAAATTTAAGTAACAGACATCTTAGGATGCCTATCAGCCAA	120		
Sbjct 19576	T-----TTGTTGTGAAAAATTGGAGCTCCAGACATTTGAACATGCCGATCAGCCGA	19526		
Query 121	CACGATCTTGCCCGTTTTGGCCAGCGCCCGGGCAGTTGCCCTCGGATCAGCATGAGCAAC	180		
Sbjct 19525	CAGGATCTGCGCCGTTTTGGCCAGCACTCCGAGCTGTGGTCCTCGATTAACCAAAAAGCCAG	19466		
Query 181	TCGAGCAATAC	191		
Sbjct 19465	ACAAGCAACAC	19455		

If the TSS annotation is supported by RNA-Seq or RNA polymerase II data, **paste a Genome Browser screenshot of the region around the TSS (± 2 kb) with the evidence tracks listed below:**

1. Short Match results for the Inr motif (TCAKTY)
2. RNA-Seq Alignment Summary
3. RNA-Seq TopHat



There are no Inr motifs, as shown by the core promoter motifs track. The TSS prediction is supported by a BRE_d motif found at 19659-19653. The RNA-seq also supports the TSS prediction well, specifically the adult males track. This makes sense, as the protein coded for is known to be needed for sperm differentiation.

If the TSS annotation is supported by sequence conservation with other *Drosophila* species, **paste a screenshot of the pairwise alignment (e.g. from blastn, matcher) or the multiple sequence alignment (e.g. from clustalw, EvoPrinter, Multiz) below:**

Alignment block 1 of 1 in window, 19375 - 19641, 267 bps

```

B D D. biarmipes cagtatagctgatgtacctaagaataatcgggaaaaaggattattatttgaatacggcccagtttgggag
B D D. melanogaster ccgaatagctgagatacctaaggacgatcaggaggatggatcagt-ttaagcatttcacaaattctggag
B D D. yakuba cagaatagctgagatacctaaggacgatcaggagaacggatcagt-aaaagcattagccaaattctggag
B D D. erecta ccgaatagctgagatacctaaggacgatcaggaggacggttcagt-ttaaacattagctaaattccggag
B D D. eugracilis cagtgtagctatgatatactaggggcaatcaggaacatggatcagtttctgagaaccgccaatttcagag
B D D. ficusphila caacgtagctaattgatctaaagggtaaacagga-cacagataagcactcaaaaaccgctgcaattgggag
B D D. takahashii ccgtatagctgatatacctgaggaca-----aggaaggatcagtttctgacaatcaccgcatttgcag

D. biarmipes ctctccccacgtgttgcctgtctggcttttggttaatcgaggaccacagctcggagtgctggcacaacgg
D. melanogaster tactctccacgtattgctcgagtgctcatgctgatccgagggcaactgccggggcgctggcacaacgg
D. yakuba tactctccacgtattgctcgaattgctcttggtaaatgcg-----acagctcggggcgctgaccaaacgg
D. erecta tactctccacgtattgctcgagtgcccttagtactgagggacaacagctcggggcgctggcacaacgg
D. eugracilis taatacccacgtgttgcctgtgactcttcgagaacacaggaacaacaacttggcgtgctggcacaacgg
D. ficusphila cacatctcacgtgttgcctagattgctattgctgaaccgaggacagcagcttgcctgctggcacaacgg
D. takahashii cgcaccccacgtattgctcgaccgcttttggcaactcaggataaacagctcggcgctgctggcacaacgg

D. biarmipes cgcagatcctgtcggctgatcggcatgttcaaatgtctggagctccaattttcacacaaatttt----
D. melanogaster gcaagatcgtgttggctgataggcatcctaagatgtctgttacttaaattttgtacagcaaaagatttgag
D. yakuba ccaaaatcctgttggctgatcggcatcctcagatgtctgttacttaaattttgtacagcaaaaattt---
D. erecta ccaagatcctgttggctgatcgcacatcctcagaaatctgttactttaaattttgtacagcaaaaattt---
D. eugracilis tgcggttctcgttgcctgatcggcatcctgggatgtctttaaatacgaatttagtacaaaactaatt----
D. ficusphila cacagatcctgttggctgatcggcatcct-ggctgtctagaattttaa-tttgcgtaaaaaaattc---
D. takahashii cgcagatcctgttggctgattgccatcctcggatgtctaaaactcctaatttgacacaaaacaaattt---

D. biarmipes -----c--cttttttagtttttctactttaaatttagctgagtttttagataggaagaattgat
D. melanogaster atttgag-ttttttctgtttttgttctaatgtaattgttgcctgaatattgcacaaaggaagaattgat
D. yakuba -tccgagtttttttccggtttttgttctactggaa-ttttgcgtgaaaattgcagaaaggaagattgat
D. erecta -tccgag--tttttttagtttttctactggaatttttgcgtgaaaattgcagaaaggaagaattgat
D. eugracilis -ttccac--ttttttcagttttcgttttactagaatttttaaatgaatattccagaaaggaagattgat
D. ficusphila -ttacac-----tttttctgttttctactaaaatttttagctgaatattgcagaaaggaagattgat
D. takahashii -ttacac--cttttccgttttctacttggaaatttagttgagttatgcagaaaggaagattgat

```

The Multiz alignments show conservation with other species back to 19,641.

2. Search for core promoter motifs

Note: The consensus sequences for the *Drosophila* core promoter motifs are available at: http://gander.wustl.edu/~wilson/core_promoter_motifs.html

Use the "Short Match" functionality in the GEP UCSC Genome Browser to search for each of the core promoter motifs listed below in the region surrounding the TSS (± 300 bp) in your project and in the TSS of the *D. melanogaster* ortholog. (For TSS annotations where you can only define a TSS search region, you should search for the core promoter motifs in the entire TSS search region).

Record the **orientation and the start coordinate** (e.g. +10000) of each motif match below. (Enter "NA" if the motif is not present.)

Core promoter motif	Your project	<i>D. melanogaster</i>
BRE ^u	NA	NA
TATA Box	NA	NA
BRE ^d	-19577, -19659, -19664, -19667, -19669, -19689	+19620429, 19620417
Inr	NA	NA
MTE	NA	NA
DPE	-19552	+19620595

Ohler_motif1	NA	NA
DRE	NA	NA
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	NA	NA
Ohler_motif8	NA	NA

Gene report form

Gene name (i.e. *D. mojavensis eyeless*): *D. biarmipes l(3)76BDm*

Gene symbol (i.e. *dmoj_ey*): *dbia l(3)76BDm*

Approximate location in project (from 5' end to 3' end): 21856-26514

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:

Name(s) of unique isoform(s) based on coding sequence	List of isoforms with identical coding sequences
l(3)76BDm-PB	l(3)76BDm-PA

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 your project consensus sequence:

All 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 which supports the hypothesis of errors in the consensus sequence include: CDS alignment with frame-shifts or in-frame stop codons, multiple RNA-seq reads with discrepant alignments compared to the project sequence, multiple high quality discrepancies in the *Consed* assembly.

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

Using the Sequencer Updater (available through the GEP web site under “Projects” -> “Annotation Resources”), create a VCF (Variant Call Format) file that describes the changes to the consensus sequence you have identified above. **Paste a screenshot with the list of sequence changes below:**

Isoform report form

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

Gene-isoform name (i.e. dmoj_ey-PA): dbia l(3)76BDm-PB

Names of the isoforms with identical coding sequences as this isoform

dbia l(3)76BDm-PA

Is the 5' end of this isoform missing from the end of project: No

If so, how many exons are missing from the 5' end: _____

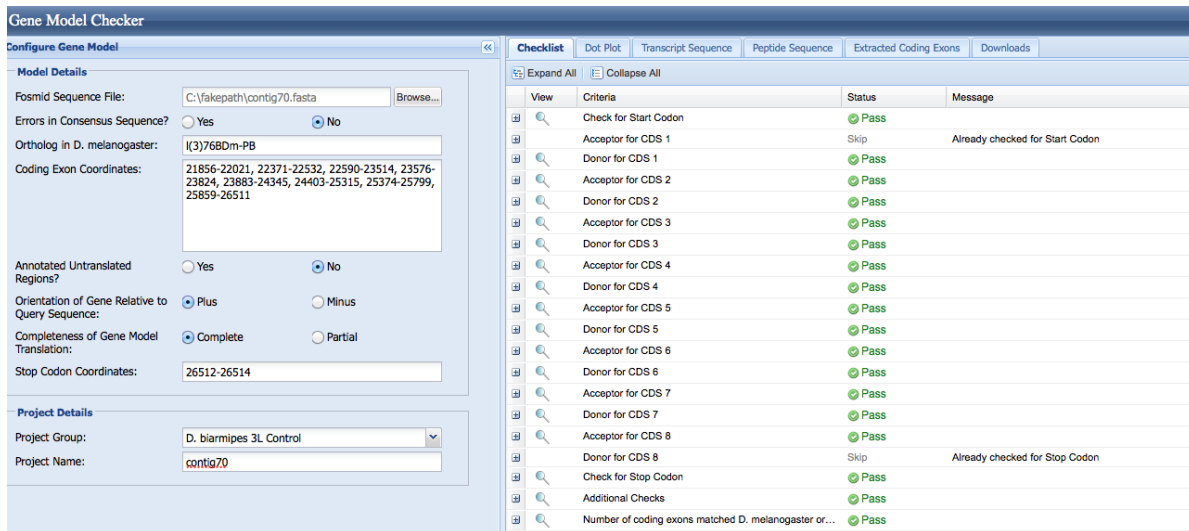
Is the 3' end of this isoform missing from the end of the project: No

If so, how many exons are missing from the 3' end: _____

1. 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 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 revise the submitted exon coordinates automatically using this VCF file.



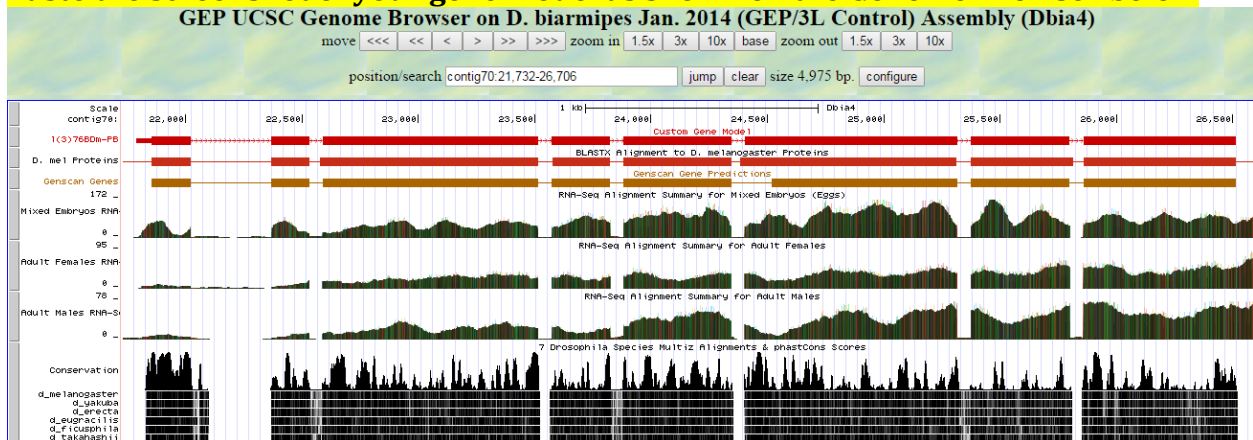
2. View the gene model on the Genome Browser

Using the custom track feature from the Gene Model Checker (see page 10 of the Gene Model Checker user guide on how to do this; you can find the guide under “Help” -> “Documentations” -> “Web Framework” on the GEP website at <http://gep.wustl.edu>).

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

29. A sequence alignment track (D. mel Protein or Other RefSeq)
30. At least one gene prediction track (e.g. Genscan)
31. At least one RNA-Seq track (e.g. RNA-Seq Alignment Summary)
32. A comparative genomics track (e.g. Conservation, D. mel. Net Alignment, 3-way, 5-way or 7-way multiz)

Paste the screenshot of your gene model as shown on the Genome Browser 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 use the protein alignment

generated by the Gene Model Checker or you can generate a new alignment using BLAST 2 Sequences (*bl2seq*). **Paste a screenshot of the protein alignment below:**

Alignment of 1(3)76BDm-PB vs. Submitted_Seq

[View plain text version](#)

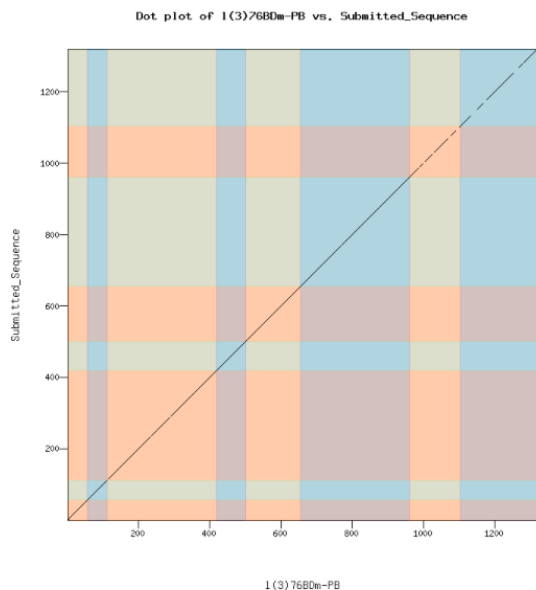
Identity: 1224/1319 (92.8%), Similarity: 1274/1319 (96.6%), Gaps: 0/1319 (0.0%)

<pre> 1(3)76BDm-PB 1 MHLHLCQGNVARDIIRNFSPFLGLVLSPPQADEICHRNLSFVBLLOPPAKLPNDLFFR 60 Submitted_Seq 1 MHLHLCQGNVARDIIRNFSPFLGLVLSPPQADEICHRNLSFVBLLOPPAKLPNDLFFR 60 1(3)76BDm-PB 61 VSGSTYSVRGLRLNFCQVDWHPFQTLARLKGMLNESVTNAHNDKRLRLATLCHDLELPTSE 120 Submitted_Seq 61 VSGSTYSVRGLRLNFCQVDWHPFQTLARLKGMLNESVTNAHNDKRLRLATLCHDLELPTSE 120 1(3)76BDm-PB 121 PWFQWRRTFTLVQFPADHEFTRHLLSCLLVLSADPQIVETAHKLQKRVQMQSITPQR 180 Submitted_Seq 121 PWFQWRRTFTLVQFPADHEFTRHLLSCLLVLSADPQIVETAHKLQKRVQMQSITPQR 180 1(3)76BDm-PB 181 LPKWFHPTFVLNGLVYLHPASQGLLKAQQGFELLKSTFDGSKPLVLSINSLDQGSAAEV 240 Submitted_Seq 181 LPKWFHPTFVLNGLVYLHPASQGLLKAQQGFELLKSTFDGSKPLVLSINSLDQGSAAEV 240 1(3)76BDm-PB 241 PDHNTFTIKRQPKSDANLPSITDLSAPKSPQAVSVISMPAMQMSQLLDGAHAQDSDLAME 300 Submitted_Seq 241 PDHNTFTIKRQPKSDANLPSITDLSAPKSPQAVSVISMPAMQMSQLLDGAHAQDSDLAME 300 1(3)76BDm-PB 301 HPLSPMQSATEANVSKFSISSESIASQTLNPNVWANELADAPHGQCLTRDIDNLRHF 360 Submitted_Seq 301 HPLSPMQSATEANVSKFSISSESIASQTLNPNVWANELADAPHGQCLTRDIDNLRHF 360 1(3)76BDm-PB 361 VQDYAVRALIPYIEHLVAILEAGVINKKGVSKSLLSATKRFVYSKPGAGANNQNAVITK 420 Submitted_Seq 361 VQDYAVRALIPYIEHLVAILEAGVINKKGVSKSLLSATKRFVYSKPGAGANNQNAVITK 420 1(3)76BDm-PB 421 NBSAHLQPRLLGDIYPMFGHYNTAFQVHQAQRDPNADSAWQYVAGALSMALSAFMLGI 480 Submitted_Seq 421 NBSAHLQPRLLGDIYPMFGHYNTAFQVHQAQRDPNADSAWQYVAGALSMALSAFMLGI 480 1(3)76BDm-PB 481 AQKTYDYHMDAIVCVLIVCKLQQPATRATLLSMECLKTARLYSEVAKQLIRMTNEESDI 540 Submitted_Seq 481 AQKTYDYHMDAIVCVLIVCKLQQPATRATLLSMECLKTARLYSEVAKQLIRMTNEESDI 540 1(3)76BDm-PB 541 RSALLLEQAAYCVLIVQPMHRRYAFHIVLGNRYSRAGQRKHAYRCYQAYQVFPQRRE 600 Submitted_Seq 541 RSALLLEQAAYCVLIVQPMHRRYAFHIVLGNRYSRAGQRKHAYRCYQAYQVFPQRRE 600 1(3)76BDm-PB 601 SLAEDHIQTVYAKQAYMLKQLREASRSFAHLRLRPSLQSAQQQTSFLKEYIQTONSLVYK 660 Submitted_Seq 601 SLAEDHIQTVYAKQAYMLKQLREASRSFAHLRLRPSLQSAQQQTSFLKEYIQTONSLVYK 660 </pre>	<pre> 1(3)76BDm-PB 661 SPSLGLLPHALFQVQSSVRVLTAVQSPSAVADRVPAINTDINSLTADDFINNKIEEMH 720 Submitted_Seq 661 SPSLGLLPHALFQVQSSVRVLTAVQSPSAVADRVPAINTDINSLTADDFINNKIEEMH 720 1(3)76BDm-PB 721 VITAAANNKPFVFKPSRYLYTKQPPALETPVAVQGGPPIELAVTLSSNSVQCRIALSIDLLM 780 Submitted_Seq 721 VITAAANNKPFVFKPSRYLYTKQPPALETPVAVQGGPPIELAVTLSSNSVQCRIALSIDLLM 780 1(3)76BDm-PB 781 KTLQNDNDVLSNACTYESSSDSANKIAVGAATKTSMASTKLDAQREFTTHFKLAPKLAG 840 Submitted_Seq 781 KTLQNDNDVLSNACTYESSSDSANKIAVGAATKTSMASTKLDAQREFTTHFKLAPKLAG 840 1(3)76BDm-PB 841 RLNLGVVCRVAAGADPAASLLQTLQFTQKTRPHNAKQSSQTMVNDRLTILKLVQLPAM 900 Submitted_Seq 841 RLNLGVVCRVAAGADPAASLLQTLQFTQKTRPHNAKQSSQTMVNDRLTILKLVQLPAM 900 1(3)76BDm-PB 901 SVSFTPVNRLLAGETIPVHVTRNMLGAPTEIYLGCDNPRCVSLLDHHSQMPALMSS 960 Submitted_Seq 901 SVSFTPVNRLLAGETIPVHVTRNMLGAPTEIYLGCDNPRCVSLLDHHSQMPALMSS 960 1(3)76BDm-PB 961 LRNLNDKLVKDKETRQQRVYRLNRPGLAALDAQOQTTLGLAVQAPHQGPFTRLLIFY 1020 Submitted_Seq 961 LRNLNDKLVKDKETRQQRVYRLNRPGLAALDAQOQTTLGLAVQAPHQGPFTRLLIFY 1020 1(3)76BDm-PB 1021 YSLPACTTAPIKYRLVRIHWQLQVENCLOADATCVVSNVAVNDELGLDINVRKHQAEGTE 1080 Submitted_Seq 1021 YSLPACTTAPIKYRLVRIHWQLQVENCLOADATCVVSNVAVNDELGLDINVRKHQAEGTE 1080 1(3)76BDm-PB 1081 VYNSISLYSEFKLPPRLHINSMVSPQAGACGLKSTICNFCQRLIKESQPKVY 1140 Submitted_Seq 1081 VYNSISLYSEFKLPPRLHINSMVSPQAGACGLKSTICNFCQRLIKESQPKVY 1140 1(3)76BDm-PB 1141 DITPTGTLAHSRSHLTLFVAQLQVLDQDPDIDPHSPTLNEVYVFNYSWHEFN 1200 Submitted_Seq 1141 DITPTGTLAHSRSHLTLFVAQLQVLDQDPDIDPHSPTLNEVYVFNYSWHEFN 1200 1(3)76BDm-PB 1201 SVADYVPRHATVWSAIVAREEQRLACCLHETIYTLFPTGCPDPAVASELRRRBS 1260 Submitted_Seq 1201 SVADYVPRHATVWSAIVAREEQRLACCLHETIYTLFPTGCPDPAVASELRRRBS 1260 1(3)76BDm-PB 1261 DDTLVNRRFFPLPEGALNEDAQPLADADESDYWFPHENYVQRCLLEDSFVAVKA 1319 Submitted_Seq 1261 DDTLVNRRFFPLPEGALNEDAQPLADADESDYWFPHENYVQRCLLEDSFVAVKA 1319 </pre>
---	---

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). **Provide an explanation for any anomalies** on the dot plot (e.g. large gaps, regions with no sequence similarity).

Note: Large vertical and horizontal gap near exon boundaries in the dot plot often indicates that an incorrect splice site might have been picked. Please re-examine these regions and provide a detail 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 specified above. This section is OPTIONAL and you do not need to complete this section to submit the project.

Name(s) of isoform(s) with unique TSS	List of isoforms with identical TSS
l(3)76BDm-PB	l(3)76BDm-PA

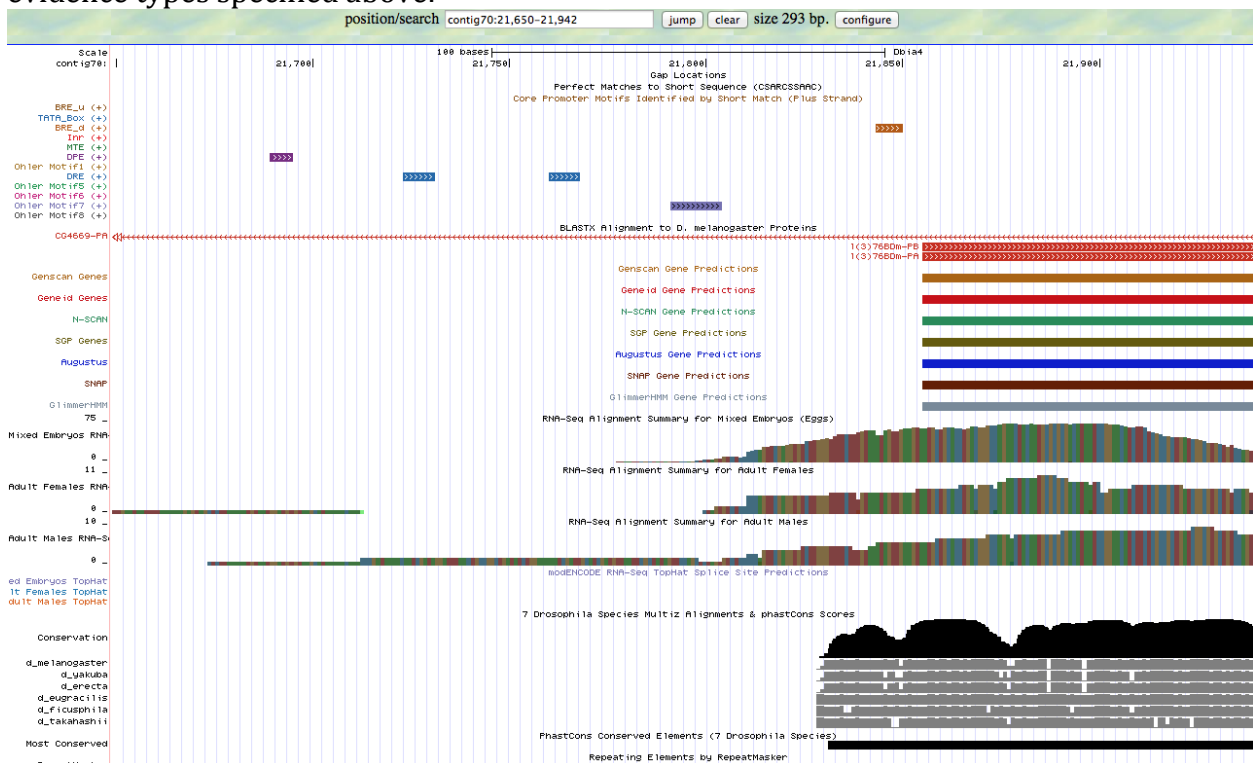
Complete this report form for each unique TSS listed in the table above (copy and paste to create as many copies of TSS report form as needed):

Gene-isoform name (i.e. dbia_ey-RA): dbia_l(3)76BDm-PB
 Names of the isoforms with the same TSS as this isoform: dbia_l(3)76BDm-PA
 Type of core promoter: (Peaked or Broad): Broad
 Coordinates of the first transcribed exon: 21,790-22,021
 Coordinate(s) of TSS position(s): 21,785
 Coordinate(s) of TSS search region(s): 21,690-21,855

1. Evidence that supports the TSS annotation postulated above
 Specify the type of evidence used to support the TSS annotation:

Evidence type	Support TSS annotation?	Refute TSS annotation?
blastn alignment of the initial exon from <i>D. melanogaster</i>	X	
RNA-Seq coverage and TopHat splice junctions	X	
Core promoter motifs	X	
Sequence conservation with other Drosophila species		X
Other (please specify)		

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



Alignment block 1 of 1 in window, 21830 - 21942, 113 bps

```

B D D. biarmipes ggttttcttggtaaatagttgaagtaatgctgcacctgacgggccagaactacaatgcgcgcgacataat
B D D. melanogaster tattttcttggtaaatagttggagtaatgctgcacctgacgggccagagttacaatgctcgcgacatcat
B D D. yakuba tattttcttggtaaataattagagtaatgctgcacctgacgggccaaagctacaatgcccgcgacatcat
B D D. erecta tattttcttggtaaataattggagtaatgctgcacctgacgggccagagctacaatgcccgcgacatcat
B D D. eugracilis ggttttcttggtaaatagttgaagtaatgctgcacctgacgggccagaactacaatgcgcgcgacataat
B D D. ficusphila ggttttccctgtaaatagttgaagtaatgctgcacctgacgggtcagaactacaatgcgcgcgatataat
B D D. takahashii ggttttcttggtaaatagta-aagtaatgctgcacctgacgggccagagctacaatgcgcgcgacataat

D. biarmipes ccggaacatcttctcgccgctgatcggcgtgctgtccagtccg
D. melanogaster ccggaacatcttctcgccgctgatcggcgtgctgtccagtcca
D. yakuba ccggaacatattctcgccgctgatcggcgtgctgtccagtcca
D. erecta ccggaacatcttctcgccgctgatcggcgtgctgtccagtccg
D. eugracilis ccggaacatcttctcgccgctgatcggcgtgctgtccagtcct
D. ficusphila ccggaacatcttctcgccgctgatcggcgtgctgtccagtccg
D. takahashii ccggaacatcttctcccaactgataggcgtgctgtccagtcctg

```

The predicted TSS is not supported by the conservation data given by the Multiz alignments. The predicted TSS is at 21,784 and the conservation begins at 21,830.

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 below:**

Dbia4_dna range=contig70:1-40000 5'pad=0 3'pad=0 strand=+ repeatMasking=none
 Sequence ID: lcl|19409 Length: 40000 Number of Matches: 10

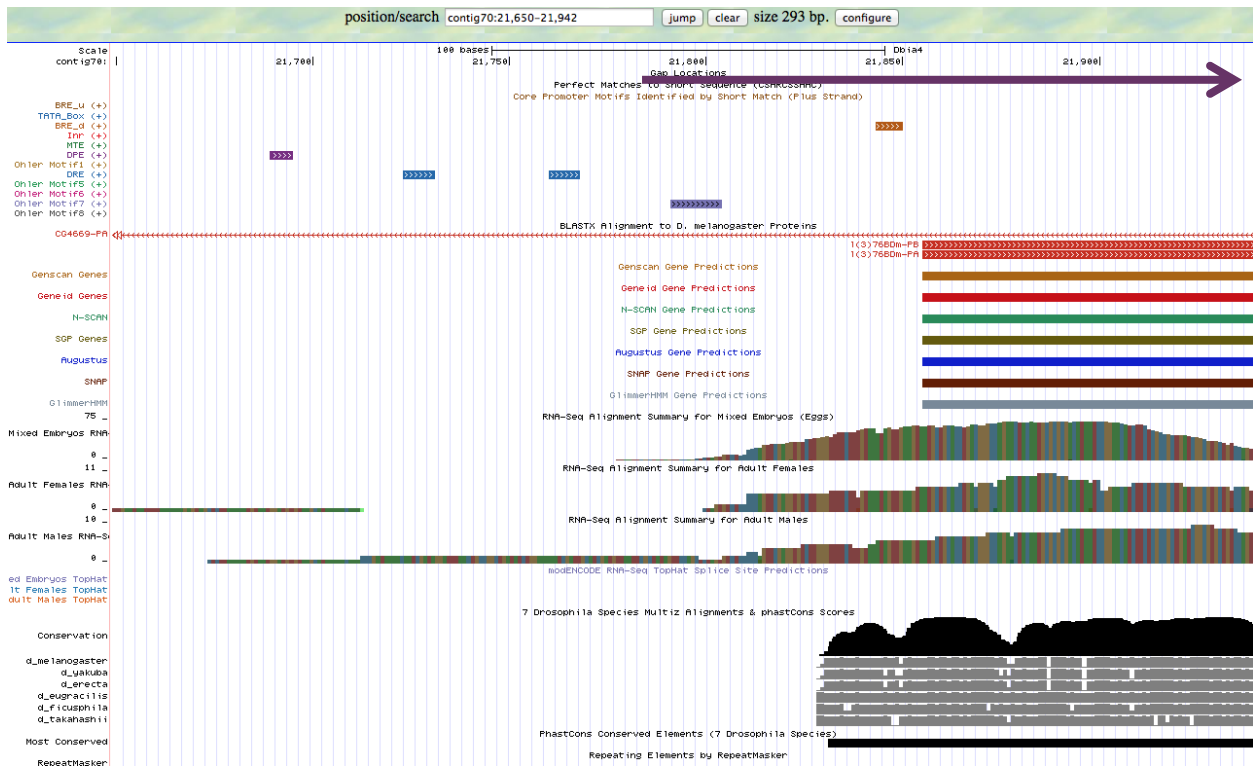
Range 1: 21790 to 22021 [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
255 bits(177)	1e-70	209/237(88%)	5/237(2%)	Plus/Plus
Query 6	AGTGCCATCTCCAGCTGCTCCCTGCAATTGTTGTGTGGTGTCTGTATTTTCTTTGTAAA	65		
Sbjct 21790	AGTTCCAGCTCTAGTGTCTCCCC-CAGTTC----GAGGTGTCTTGGGTTTTCTTTGTAAA	21844		
Query 66	TAGTTGGAGTAATGCTGCACCTGACGGGCCAGAGTTACAATGCTCGCGACATCATCCGGA	125		
Sbjct 21845	TAGTTGAAGTAATGCTGCACCTGACGGGCCAGAACTACAATGCGCGACATAATCCGGA	21904		
Query 126	ACATCTTCTCGCCGCTGATCGGCGTGCTGTCCAGTCCACAGGCGGACGAGATTTGCCACC	185		
Sbjct 21905	ACATCTTCTCGCCGCTGATCGGCGTGCTGTCCAGTCCGCGAGGCGGACGAGATCTGCCACC	21964		
Query 186	GGAACAACCTCTCCTTCGTGGAACCTCCTGCAGCCGTTTCGCAAAGTTGCCAAATGATG	242		
Sbjct 21965	GCAACAACCTCTCCTTCGTGGAGCTCCTGCAGCCGTTTCGCAAAGTTGCCAAATGATG	22021		

The BLAST alignment, even with more sensitive parameters, could not align the first 5 bases of the melanogaster ortholog. However, to preserve length, I am placing the TSS prediction at 21,785.

If the TSS annotation is supported by RNA-Seq or RNA polymerase II data, **paste a Genome Browser screenshot of the region around the TSS (± 2 kb) with the evidence tracks listed below:**

1. Short Match results for the Inr motif (TCAKTY)
2. RNA-Seq Alignment Summary
3. RNA-Seq TopHat



There are no Inr motifs in the search region, as indicated by the core promoter motifs track. There is a DRE, as well as two Ohler_motif7's, that may or may not support the TSS prediction. These core promoter motifs can indicate the presence of a TSS, but do not support a specific location. The RNA-Seq tracks seem to support the TSS prediction well. The purple arrow on the figure is used to indicate the approximate location of the TSS prediction.

If the TSS annotation is supported by sequence conservation with other *Drosophila* species, **paste a screenshot of the pairwise alignment (e.g. from blastn, matcher) or the multiple sequence alignment (e.g. from clustalw, EvoPrinter, Multiz) below:**

2. Search for core promoter motifs

Note: The consensus sequences for the *Drosophila* core promoter motifs are available at: http://gander.wustl.edu/~wilson/core_promoter_motifs.html

Use the "Short Match" functionality in the GEP UCSC Genome Browser to search for each of the core promoter motifs listed below in the region surrounding the TSS (± 300 bp) in your project and in the TSS of the *D. melanogaster* ortholog. (For TSS annotations where you can only define a TSS search region, you should search for the core promoter motifs in the entire TSS search region).

Record the **orientation and the start coordinate** (e.g. +10000) of each motif match below. (Enter "NA" if the motif is not present.)

Core promoter motif	Your project	<i>D. melanogaster</i>
---------------------	--------------	------------------------

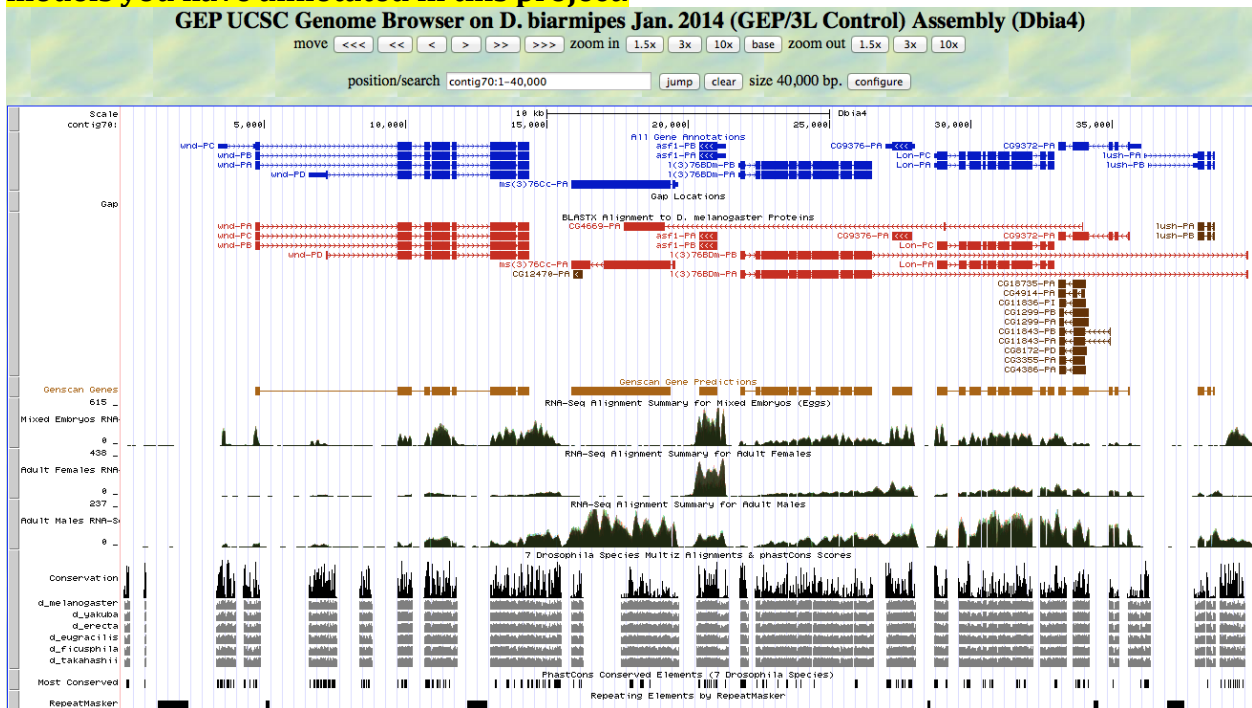
BRE ^u	NA	NA
TATA Box	NA	NA
BRE ^d	+21844	-19618677, -19618603, -19618535, -19618484, -19618481, -19618478, -19618476, -19618466, -19618451
Inr	NA	NA
MTE	NA	NA
DPE	+21690	NA
Ohler_motif1	NA	NA
DRE	+21724, +21761	NA
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	+21792	NA
Ohler_motif8	NA	NA

Preparing the project for submission

For each project, you should prepare the project GFF, transcripts 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.

The Annotation Files Merger also allows you to view all the gene models in the combined GFF file within the Genome Browser. Please refer to the Annotation Files Merger User Guide for detail instructions on how to view the combined GFF file on the Genome Browser (you can find the user guide under “Help” -> “Documentations” -> “Web Framework” on the GEP website at <http://gep.wustl.edu>).

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

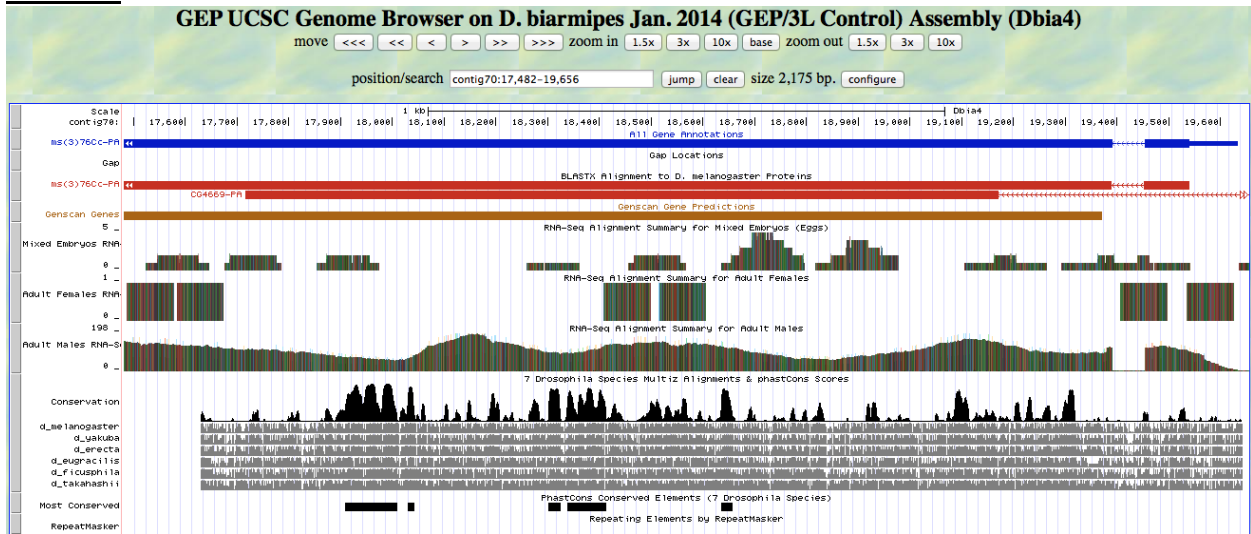


For projects with multiple errors in the consensus sequence, you should combine all the VCF files into a single project VCF file using the Annotation Files Merger (see the Annotation Files Merger User Guide for details). **Paste a screenshot (generated by the Annotation Files Merger) with all the consensus sequence errors you have identified in your project.**

Have you annotated all the genes that are in your project?

For each region of the project with gene predictions that do not overlap with putative orthologs identified in the BLASTX track, perform a BLASTP search using the predicted amino acid sequence against the non-redundant protein database (*nr*). **Provide a screenshot of the search results.** Provide an explanation for any significant (E-value < 1e-5) hits to known genes in the *nr* database and why you believe these hits do not correspond to real genes in your project.

CG4366



The Genome Browser included the gene CG4366 as a blastx prediction within the ms(3)76Cc-PA gene. Investigating this further, I took the region in which CG4366 was predicted and performed a BLAST search myself

CG4669 [Drosophila melanogaster]

Sequence ID: [ref|NP_647956.2|](#) Length: 611 Number of Matches: 1

[▶ See 2 more title\(s\)](#)

Range 1: 106 to 574 [GenPept](#) [Graphics](#)

▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
396 bits(1018)	1e-127	Compositional matrix adjust.	210/489(43%)	300/489(61%)	24/489(4%)	-1
Query 1457	VNFATYLKAFALIYVLPEDWTAEKIDMLLEEGTDLFRASSEVDQKDDQEEHKLEPEI					1278
Sbjct 106	V F A+ L K A F A Y + W + I D M++ E G L F S + D P+I					159
Query 1277	YTNEEKRIKRNFNLEGHFTFLALEPRYQGAGSKPLEEQPPHTIDNLRPVLLNFFKSSRYC					1098
Sbjct 160	Y E+++ R+F + F + L E ++ G + + I N L R + L F F K ++ Y					212
Query 1097	LLLTRVGHLLVWRRRKVFFVLDVKGRRIEDLETQDNGVAMLVCLKTIDNVVHLASNLGS					918
Sbjct 213	+ T +L L+W+ + V+ V L D+ G R ++ + G +L+ L K+ D N V V L S					272
Query 917	ISPQDEFITIRELVVVRLETPD--GRIYMRDTSHRISIEFRVVKNSYAYLKATLHLSLNEHD					744
Sbjct 273	+ +F+IRE++VVRL TP G+ + R+ R +F V+ YAY+K+ LHL+LN D					332
Query 743	PVRNRSSLMVAVGSILASKIDHPANWDTNMLDRLICYGVELCRNCWSDCLDRRRPIDLDT					564
Sbjct 333	+RNRS+L VAV + LASKIDHPA WD M D+++CYGV +C+NCW C+ +P+DLD					392
Query 563	FPTQLRMGQYVLELKLIPNVRAGHWKCGVRIIGTDFEAHVSEALNEFGNVVVFQINNQMYS					384
Sbjct 393	FP Q+R+GQ+V E+ L P N G W K C D F + +A L++ ++F Q I N N Q M Y+					452
Query 383	IWVKDEFYLLDPYRHTIVGTHVAEDKGEKAWATVRMFRDQLTMLSVMHMLKESNRQS					204
Sbjct 453	+W K +F YL+DPYRH IVG + E G E K A T V R M F + +L S V F H Q + L E S N R +					510
Query 203	AYYLVVVRIRNLAECPEGY--ALVPLSEVDGNADVKSLNEPILFNEQQGVRVCDKSLADI					30
Sbjct 511	+++H +RIRN+ ECP G AL+P EDV +V+SLNE I F+E C + L +I					565
Query 29	SDYEEDVIS 3					
Sbjct 566	SD+EED++S SDFEEDLVS 574					

CG4366 did appear as a match and the E-score is good. However, there are some concerns with the identity being only 43% and the 611 length protein only matching from 106-574.

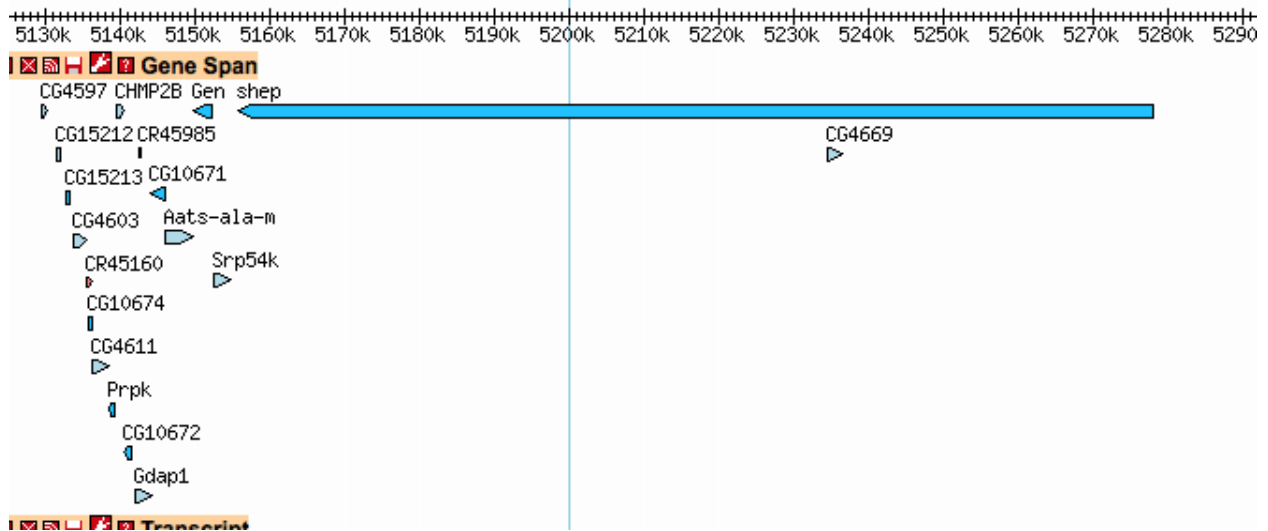
contig70 (40000 letters)

<p>RID JS89A40P113 (Expires on 04-15 04:03 am)</p> <p>Query ID Icl Query_20275</p> <p>Description contig70</p> <p>Molecule type nucleic acid</p> <p>Query Length 40000</p>	<p>Subject ID Icl Query_20277</p> <p>Description CG4669:1_6708_0</p> <p>Molecule type amino acid</p> <p>Subject Length 31</p> <p>Program BLASTX 2.2.31+ ▶ Citation</p>
---	---

ⓘ No significant similarity found. For reasons why, [click here](#)

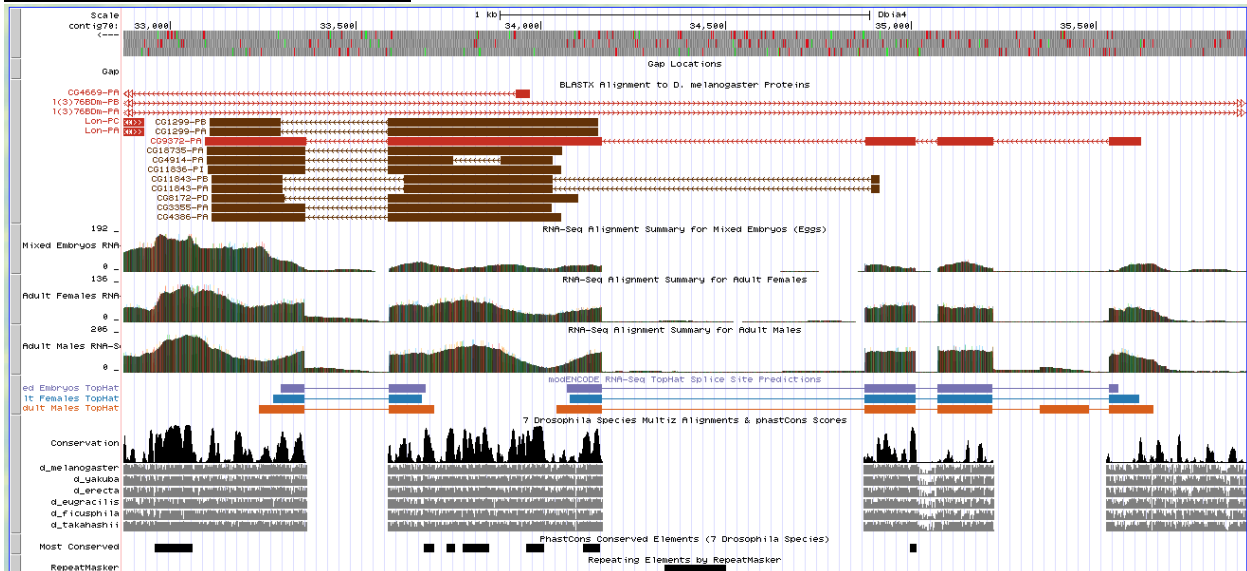
Other reports: [▶ Search Summary](#)

The first exon of CG4366 could not be found in my contig.



I looked at the *Drosophila melanogaster* model to see where and how CG4366 fit in relation to ms(3)76Cc-PA. I found that CG4366 is not even found in the same general location and around the same genes. This further raised the question of why it appeared on the initial blast search with a good e-value. It turns out that ms(3)76Cc and CG4366 are very similar genes as they are both expressed in male testes.

Predictions around CG9372



The gene CG9372, in red, is what I annotated and the model fit very well. The Blast tracks predicted additional genes in brown around and overlapping CG9372. They are brown to indicate that the E-value was not as high as normally expected. It turns out that all of these genes are similar to CG9372 to the point where they align, all be it poorly, on Blast. There was no other support for these genes being located here.