2015

Annotation and Transcription Start Site Analysis of contig70 in Drosophila biarmipes

Robin Wolschendorf
Grand Valley State University

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

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

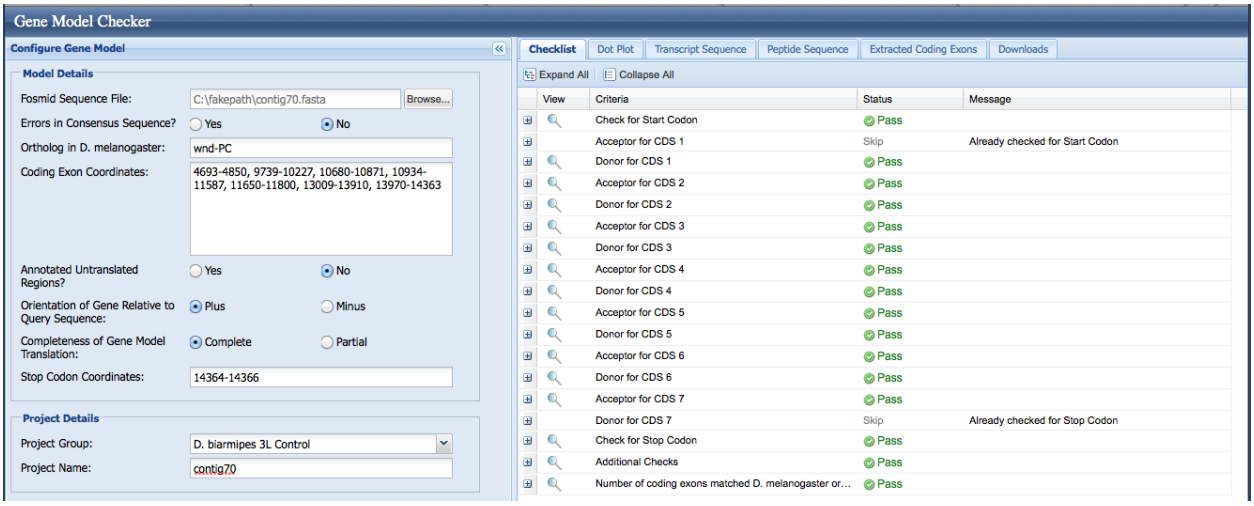
Complete this report form for each unique isoform listed in the table above (copy and paste to create as many copies of this Isoform Report Form as needed):

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.

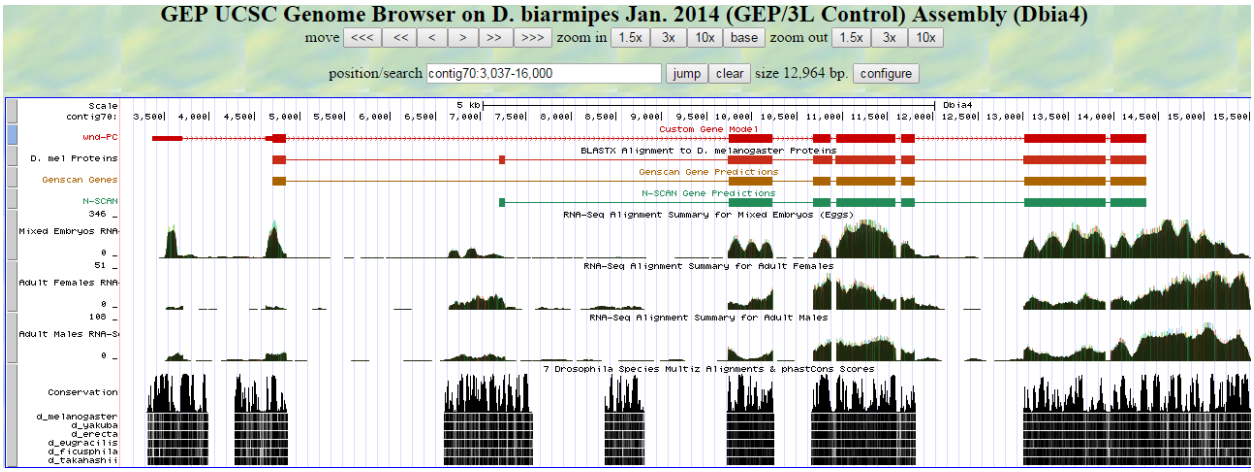


2. View the gene model on the Genome Browser

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:



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

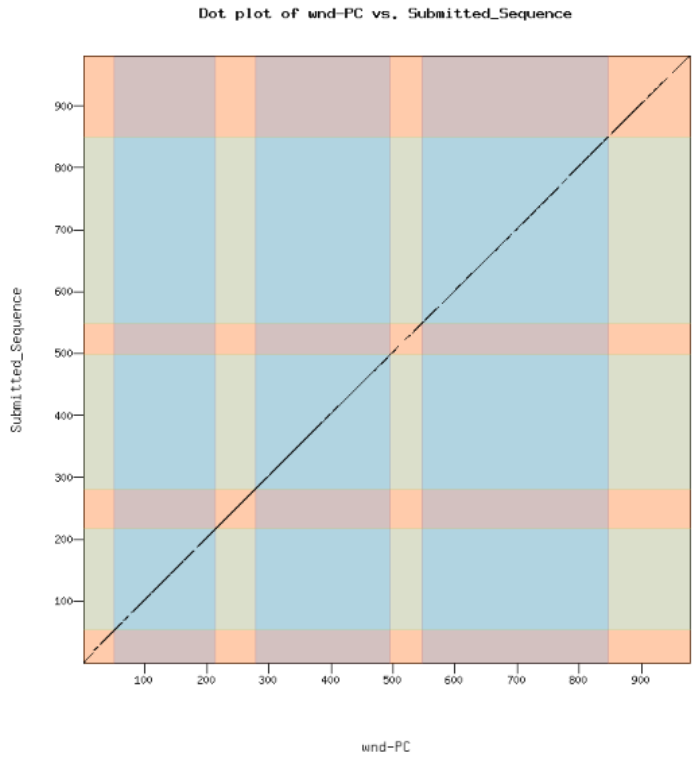
Identity: 923/981 (94.1%), **Similarity:** 943/981 (96.1%), **Gaps:** 5/981 (0.5%)

```
wnd-PC              1 MQPFSDSLSKSRDDLVT---AASRQQQQLHGRRRHHGSSPNLSLDQTDNLRRSMACLQDE 57
                      ****************.   ******:**:**********************************
Submitted_Seq       1 MQPFSDSLSKSRDDLVNVAAAASRQQHQLYGRRRHHGSSPNLSLDQTDNLRRSMACLQDE 60

wnd-PC             58 FGHLGIAATDLPFKSSDLDSPPRLQHHNNYAEITDSSAENTCQQRWPPHVGGAGAAFGHP 117
                      *****:**********.****.*************************************.******
Submitted_Seq      61 FGHLGLSATDLPFKSPDLDSSPRLQHHNNYAEITDSSAENTCQQRWPPHVGGAAAAFGHP 120

wnd-PC            118 DKPIGWMYGLLGCMKPVLSPIGKTGVIEVKSQRSEDWQIPFESITELEWLGSGAQGAVFS 177
                      ************************:**********:**********:***************
Submitted_Seq     121 DKPIGWMYGLLGCMKPVLSPIGKSGVIEVKSQRSEDWQIPFESITDLEWLGSGAQGAVFS 180

wnd-PC            178 GRLKNETVAVKKVKELKETDIKHLRKLDHENIIKFKGVCTQSPVFCIIMEFCPYGPLQNI 237
                      *:**** ************************************************************
Submitted_Seq     181 GKLKNEIVAVKKVKELKETDIKHLRKLDHENIIKFKGVCTQSPVFCIIMEFCPYGPLQNI 240

wnd-PC            238 LKEEQVMLPSRLVSWSKQIALGMQYLHSHKIIHRDLKSPNTLISTNEVVKISDFGTSREW 297
                      ************************************************************:
Submitted_Seq     241 LKEEQVMLPSRLVSWSKQIALGMQYLHSHKIIHRDLKSPNTLISTNEVVKISDFGTSREF 300

wnd-PC            298 NEISTKMSFAGTVAWMAPEVIRNEPCSEKVDIWSYGVVLWEMLTCEIPYKDVDSSAIIWG 357
                      ************************************************************
Submitted_Segq    301 NEISTKMSFAGTVAWMAPEVIRNEPCSEKVDIWSYGVVLWEMLTCEIPYKDVDSSAIIWG 360

wnd-PC            358 VGNNSLKLLVPSTCPEGFKLLVKLCWKSKPRNRPSFRQILSHLDIAGPELLRKTEKQYFE 417
                      ************************************************************
Submitted_Seq     361 VGNNSLKLLVPSTCPEGFKLLVKLCWKSKPRNRPSFRQILSHLDIAGPELLRKTEKQYFE 420

wnd-PC            418 TQKSWKEEVRSHLKEITQNGTNIHKYEQDLIKRRTAEWRHAQDIRMVYEDKLQKTNQLFF 477
                      ********************.************************:**********
Submitted_Seq     421 TQKSWKEEVRSHLKEITQNGTSIHKYEQDLIKRRTAEWRHAQDIRMVYEEKLQKTNQLFF 480

wnd-PC            478 ELSECMSQLQEKEKEIAERERKLPGSGYKPNRRFGNTIRKMQHYRRRLNPAPAAIQQQST 537
                      ******************************.*:*:***:***:*******.*** ******
Submitted_Seq     481 ELSECMSQLQEKEKEIAERERKLPGSGYKPTRHLGNTLRKMKHYRRRLNAAPA-IQQQST 539

wnd-PC            538 TPDPETTPESPVKCMLYAQLDSNCQPKSYLANIIPSSGLGAPMPNKNKVFRHRRNASGS 597
                      ************ ********************.**:*.*******  ******.***
Submitted_Seq     540 TPDPETTPESPVKCTLYAQLDSNCQPKSYLANIIPGSGMGGTMPNKNKNFRHRRNGSGS 599

wnd-PC            598 PGAPPKYSPTRDRRYQSEPENRKVQLVERQTQTDAMDVSETDISPSAEAPRSQPIDVPVP 657
                      ************************************************************.*
Submitted_Seq     600 PGAPPKYSPTRDRRYQSEPENRKVQLVERQTQTDAMDVSETDISPSAEAPRSQPIDVPAP 659

wnd-PC            658 NHRQLPLQLQRVQKIAQAQARARSGSTSSAAGAVNPACPSNGNSLSTSELTYQDACSSPD 717
                      ************************************:***.************************
Submitted_Seq     660 NHRQLPLQLQRVQKIAQAQARARSGSTSSAAGALNPASPSNGNSLSTSELTYQDACSSPD 719

wnd-PC            718 QLIDDVMNSNERLDMTECCSDNENLERLGRKVIEPINENRLSIQSNTNSVSNADNGNGG- 776
                      **************:**************************************** .*:.:*****
Submitted_Seq     720 QLIDDVMNSNERLDITECCSDNENLERLGRKVIEPINENRLSIQSNTNSNCNTENGNGGG 779

wnd-PC            777 ASPLELRESGNSPCLSRCSSTHSKRRKHPLGDNPNGSSIGNANGESHEQDSWSDEEGETT 836
                      ***:** ********************:**********.********************
Submitted_Seq     780 ASPVELLESGNSPCLSRCSSTHSKRRKQPLGDNPNGSSISNANGESHEQDSWSDEEGETT 839

wnd-PC            837 DYKYALRRRSIGRQPIARGMRPRRSYKAPLSQKIAIHKRNVVIVSDEEENTSEYSHSPSS 896
                      * ********:** ************************************************
Submitted_Seq     840 DCKYALRRRSFGRLPIARGMRPRRSYKAPLSQKIAIHKRNVVIVSDEEENTSEYSHSPSS 899

wnd-PC            897 QHSTLESNTDIADMKKTQATSTSSNSYSEPEDDSSDSSDEQGNRPTEAKAEGPALPMGA 956
                      ********:*:************:***************:***** .:*..** ******
Submitted_Seq     900 QHSTLESNADMADMKKTQATSTSTNSYSEPEDDSSDSSDDEQGNRVPKATVEGQALPMGA 959

wnd-PC            957 VRSSDIISIPTFEADGAVNMV 977
                      **********************
Submitted_Seq     960 VRSSDIISIPTFEADGAVNMV 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 <u>vertical and horizontal gap</u> 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

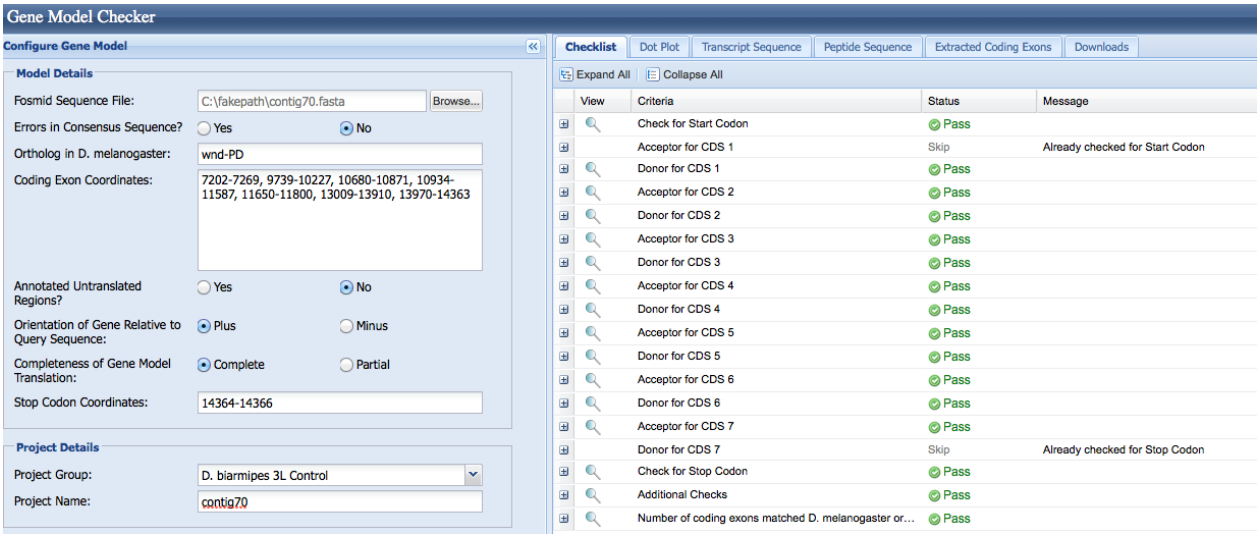
Complete this report form for each unique isoform listed in the table above (copy and paste to create as many copies of this Isoform Report Form as needed):

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.

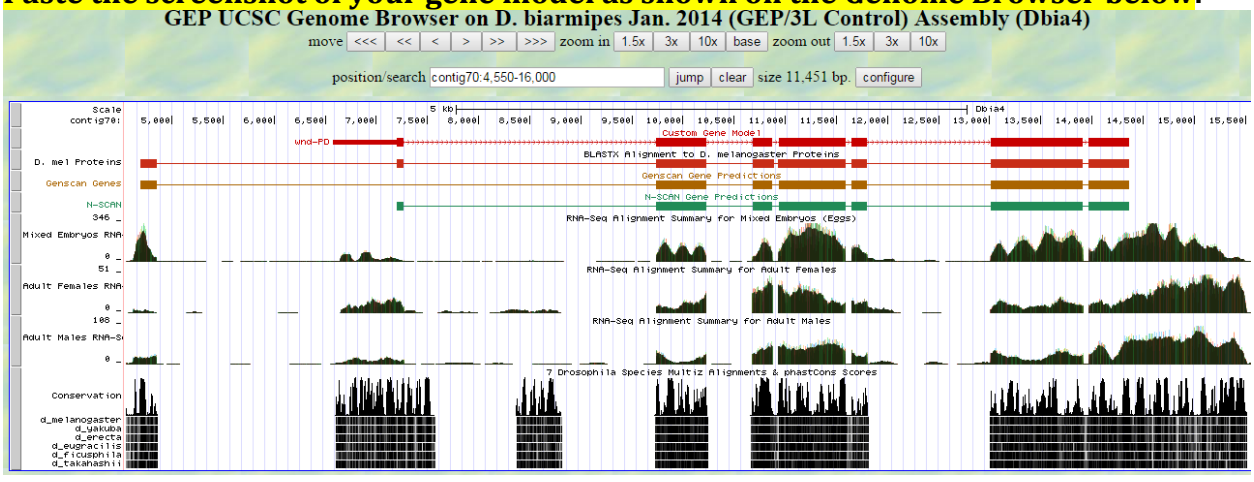


2. View the gene model on the Genome Browser

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:



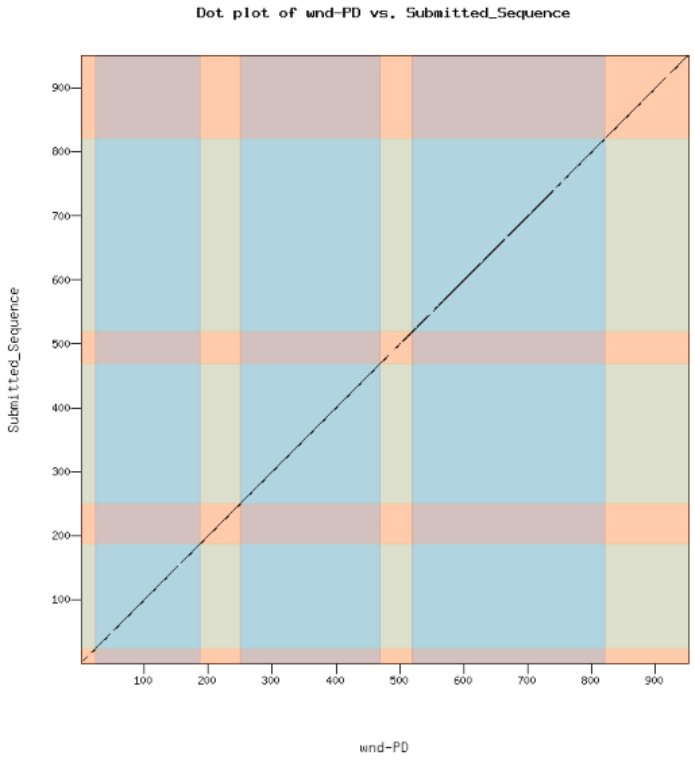
3. Alignment between the submitted model and the *D. melanogaster ortholog*

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:**

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 <u>vertical and horizontal gap</u> 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 <u>OPTIONAL</u> 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

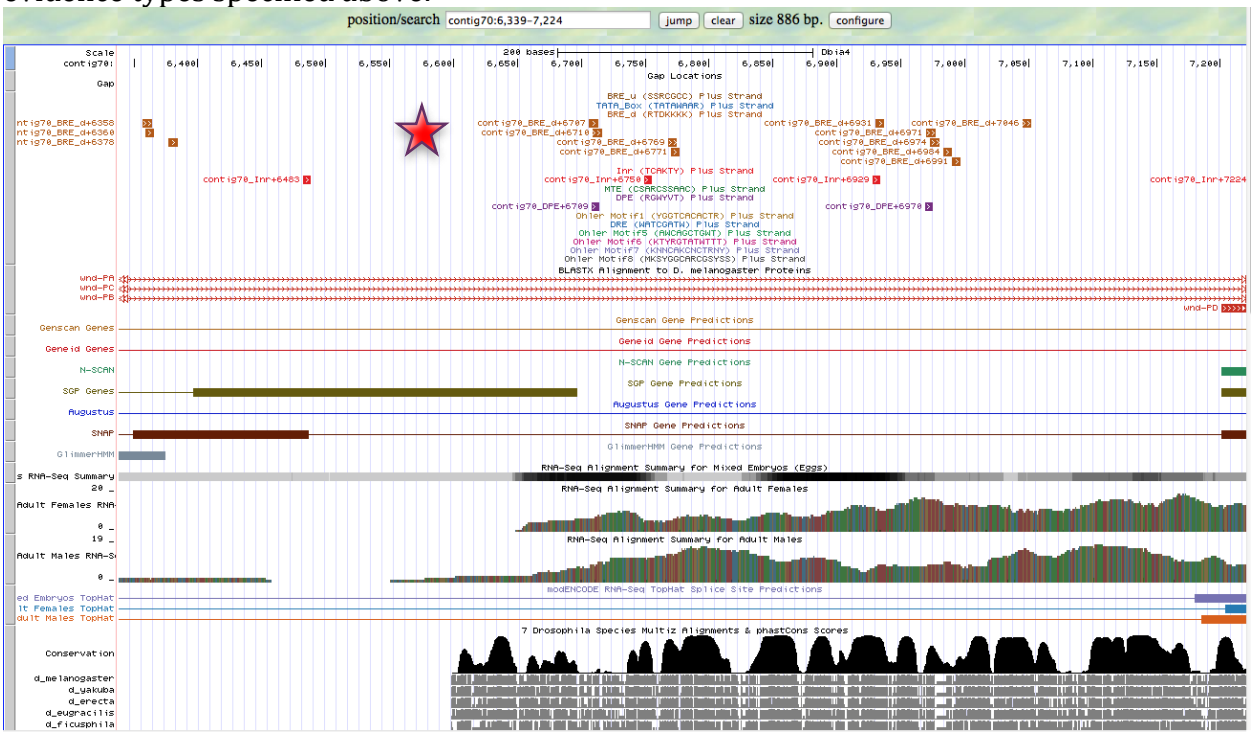
Coordinate(s) of TSS position(s): 6,576
Coordinate(s) of TSS search region(s): 5,000 – 7,201

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 *D. melanogaster*	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 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

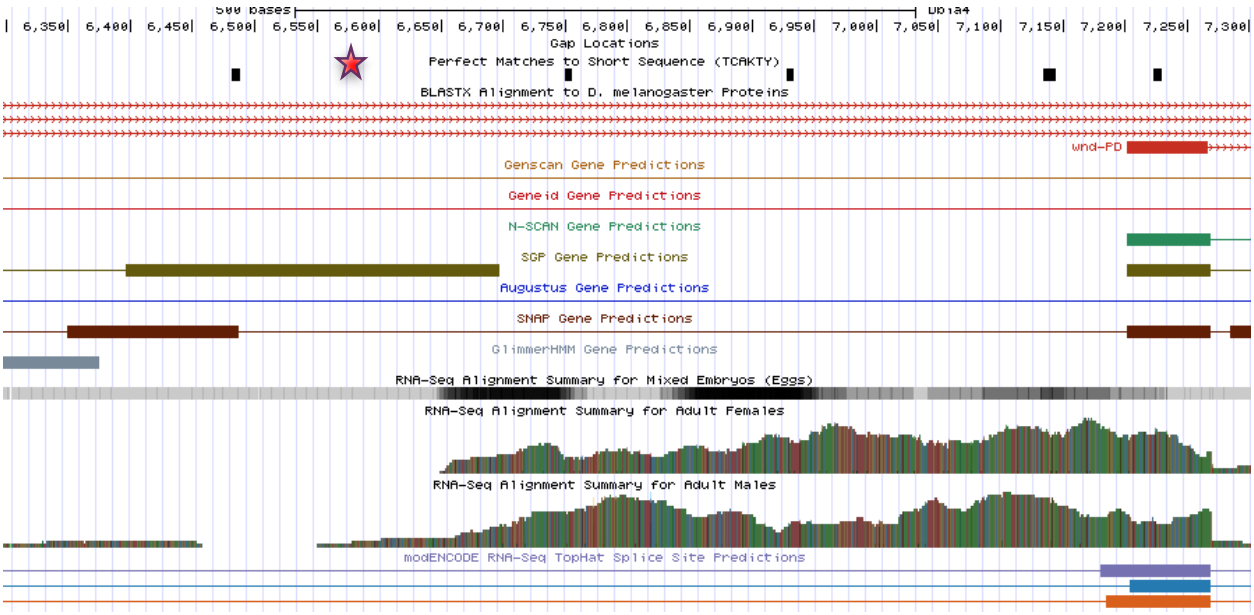
Score	Expect	Identities	Gaps	Strand
633 bits(442)	0.0	598/724(83%)	41/724(5%)	Plus/Plus

```
Query  18    CTGATACTGATAATGTGGAGCTAGACCGTTATGCTCACGCT-GCGAGTCGCGATTTAGTT  76
             |||||| |||||||| |||||| |||| ||||||| |||| |||||||||| ||||||||
Sbjct  6576  CTGATAGTGATAATGGGGAGCTTTACCGCTATGCTCCCGCTCGCGAGTCGCGGTTTAGTT  6635

Query  77    GGCGAAAATCAATCGCTTCTGGTGGTCGCTCTTTGCGCTTGAAATTCGCAAGCCTGCCTT  136
             ||||||||||||||   ||||  |||||||||| |||| || |||||||  ||||||||||
Sbjct  6636  GGCGAAAATCAATCGTATCTGACGGTCGCTCTTCGCGCGTGGAATTCGCGAGCCTGCCTT  6695

Query  137   AATTTGCGTCCGTGTGTTTGTGCGACAACAAAAACAAAAGCGAGCTTATCGCCAAATCAG  196
             ||  ||||  ||||   |||||  ||| ||| ||||||||     ||||||||||||||||
Sbjct  6696  AAAGTGCGCTCGTGG--TTGTGGCACAGCAAGAACAAAACACGGCTTATCGCCAAATCAG  6753

Query  197   TTTGCAGCCGTTGTCTGTGTGTTGGTGTGACGCAGTTGTGCGAAATACGTGAAATTATTG  256
             |||||||  | | ||||||||||||| |||||||||| |   |||||||||||||| |||
Sbjct  6754  TTTGCAGGCCT-GTCTGTGTGTTGGAGTGACGCAGTGG--CGAAATACGTGAAATCATTA  6810

Query  257   CATTTTTCCCCCCCAAACCCAACCAATAAGCAACAGTCTTGAAAAACCGCCAACGAAAGT  316
             ||  ||||  ||||   | | |          || |||||  | ||  ||||||||||||||||||
Sbjct  6811  CACTTTTTCCCCAGACAGC---------AACCAACACTGTT-AAAAACCGCCAACGAAAGT  6861

Query  317   GTTACGCGTCGCGTTCGTCTCTCTTTTCTTGTAATTAATAGCAACAAAAAGCGATACCAA  376
             ||||||||    |  |||||||| |||||||||||||  |    |||||||||| ||||||||
Sbjct  6862  GTTACGCGCTTCCCTCGTCTCT-TTTTCTTGTAATTGA---CAACAAAAAGC-ATACCAA  6916

Query  377   CTTACCTTGACGTCACTTTTTTTTATTATTATACATCGGACGACAAAAGGCGATTGTTGT  436
             ||  |||||||||||| |||||| ||  ||  | ||||||||||||||| |||  | |||||
Sbjct  6917  CTCACCTTGACGTCATTTTTTTTATAATAC-ACACATCGGACGACAAAAAGCGCTGGTTGT  6975

Query  437   TGTT---------TTTTCGGTGGCTGCTGGTGCTTGGCACGAACTCACAATCAAAAGTTA  487
             ||||         ||  | |||||||||||||||||||||||||  | ||||||||||||
Sbjct  6976  TGTTGCTCGTTTTTTTATTGGTGGCTGCTGGTGCTTGGCACGAACTTAAAATCAAAAGTTA  7035

Query  488   GCCAACTTTTGTTGTTGGCCCAGTGCGTAAAAACAAAAAAAAAACAGAACTGCTGGC-GA  546
             |||||||||||||||||| | |||||||| | ||| ||| ||   |||||  | | ||
Sbjct  7036  GCCAACTTTTGTTGTTGGCTCGGTGCGTAAACAAAAATAAAGAA---AACTGTCGACCGA  7092

Query  547   AAAGCGAGCGTAAAATTTTGCATTAATGCAGTCTCGTCGCATGAATGAATGAAATTCTTA  606
             || |||||||||||| |||||||| |||||||||| ||||||||||||||||||||||||
Sbjct  7093  AAGGCGAGCGTAAAAATTTGCATCAATGCAGTCGCGTCGCATGAATGAATGAAATTCTTA  7152

Query  607   GTGCGCGACAACTAATTACGAAAGGGAACCAAAAAAGTGGCAATAACAAATAAACCATGG  666
             ||||||||||||||||||||||||  |||   |||||  ||      ||||||||||||||
Sbjct  7153  GTGCGCGACAACTAATTACGAAAGCAAAC---AAAAGAGG----GACAAATAAACCATGG  7205

Query  667   TCTACATCATACCGTTCTTTAGCCACAATGACCCGCCCGACATGATCATCACCAAAGAAA  726
             ||||||  |||||||||||| || |||||||  ||||||||||   ||||||||||||||||
Sbjct  7206  TCTACACCATACCGTTCTTCAGTCACAATGAGCCGCCCGACCCGATCATCACCAAAGAAA  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 (±2kb) with the evidence tracks listed below:**

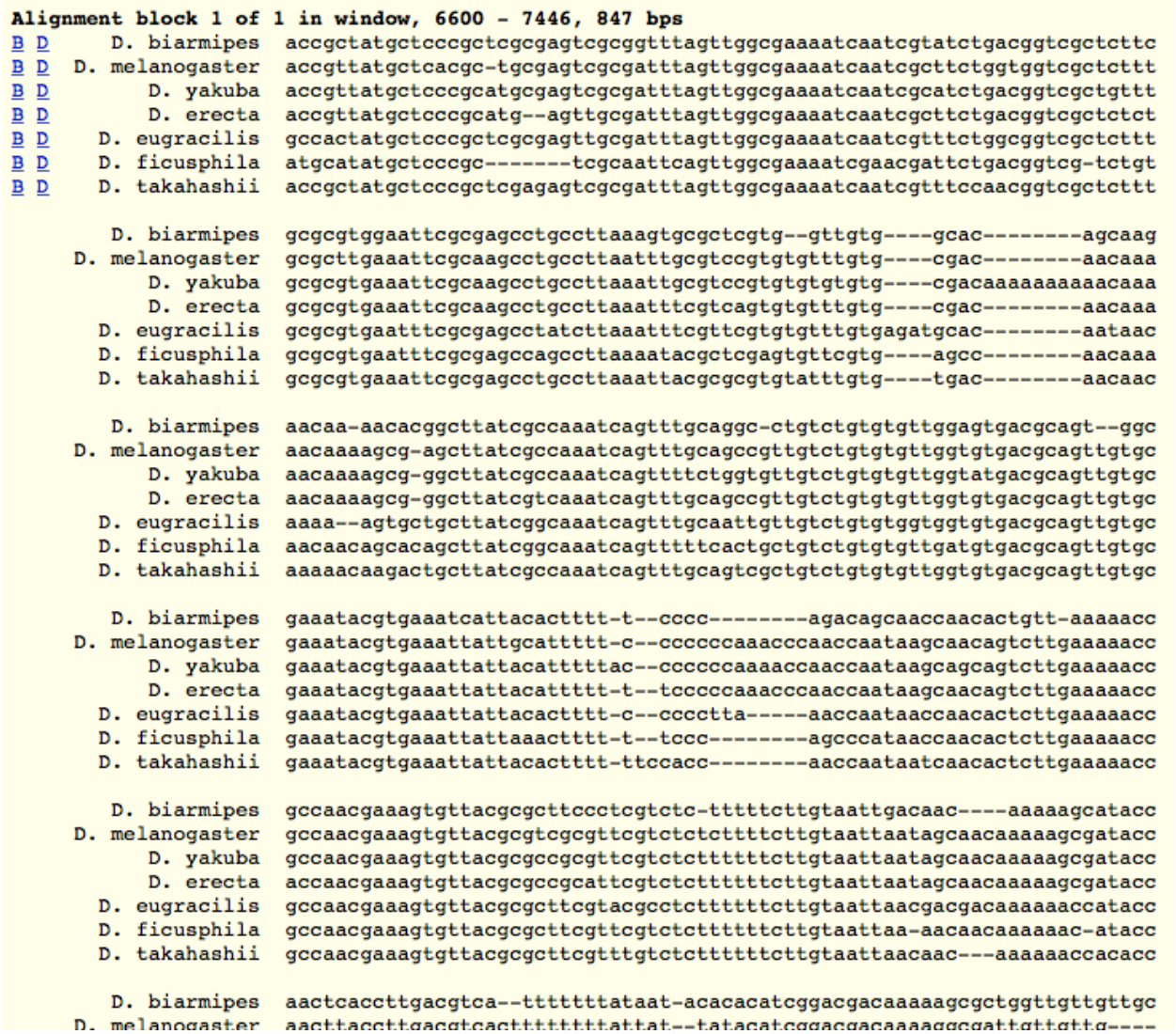
1. Short Match results for the Inr motif (TCAKTY)
2. RNA-Seq Alignment Summary

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:**



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 (±300bp) 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	*D. melanogaster*
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 <u>OPTIONAL</u> 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 *D. melanogaster*	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

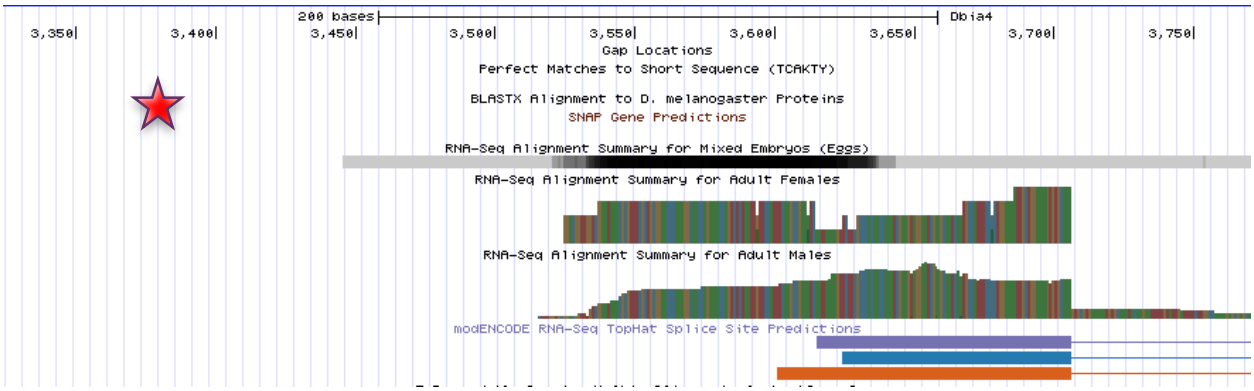
Score	Expect	Identities	Gaps	Strand
278 bits(193)	2e-77	278/351(79%)	39/351(11%)	Plus/Plus

```
Query  9     TTCTCCCGCACACTTGCGCAGTGAACACAACCTGCGCTCCAAACTGGTTTTGGTGGAGGA  68
Sbjct  3380  TTCTGCCGCACACATGCGCAGTG----CAACCTGCGCTCCAAACTGGTTTCGGGGGAGGA  3435

Query  69    AATCCGCAAGAGAGAGCGCTAGAGAGCGGATTGGAACTCGGTTTCGGCCAAAGCCAAAGC  128
Sbjct  3436  AATCCGCGAGAGAG------------CGGATTGG-------TTTCGGCCCAA--------  3468

Query  129   AGAGCCAGCAGCCAGTTTTT-----GCTTTTTAGTCGATTTGTATCTACCTTTGGTGCGG  183
Sbjct  3469  --AGCCAGCAGCCAGTTTTTTTTCTGCTTTTTAGTCGACTTGTATCTACCTTTGGTGCGG  3526

Query  184   ACGGTTGGTCGGAATAAACGCGTTTCGCAGCGGAACTCCAAGAAGAGCAGAAAACCAGTC  243
Sbjct  3527  ACGGTTGGTCGGAATAAACGCGTTTCGCTGCGGAACCCAAAGGCAACCAGAAAACCAGTC  3586

Query  244   TTTAATTGTTCCATTTCAAGCTGAAATCAAATAAAAGTTCTTCGCCAATGGCTGAAT-AA  302
Sbjct  3587  TCTAATTGTTCCATTTCAAGCCGAAATCAATTAAAAGTTCTTCGCCCACAACCAAATCGA  3646

Query  303   AGTACGAATCAATGCAATCACTACGATTCGCAGTGGAAAATTGCTGGAAAA  353
Sbjct  3647  AGTAGGAATCGAAGCAATCAATGCCATTTCCCCTGGAAAATTGCTGGAAAA  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.

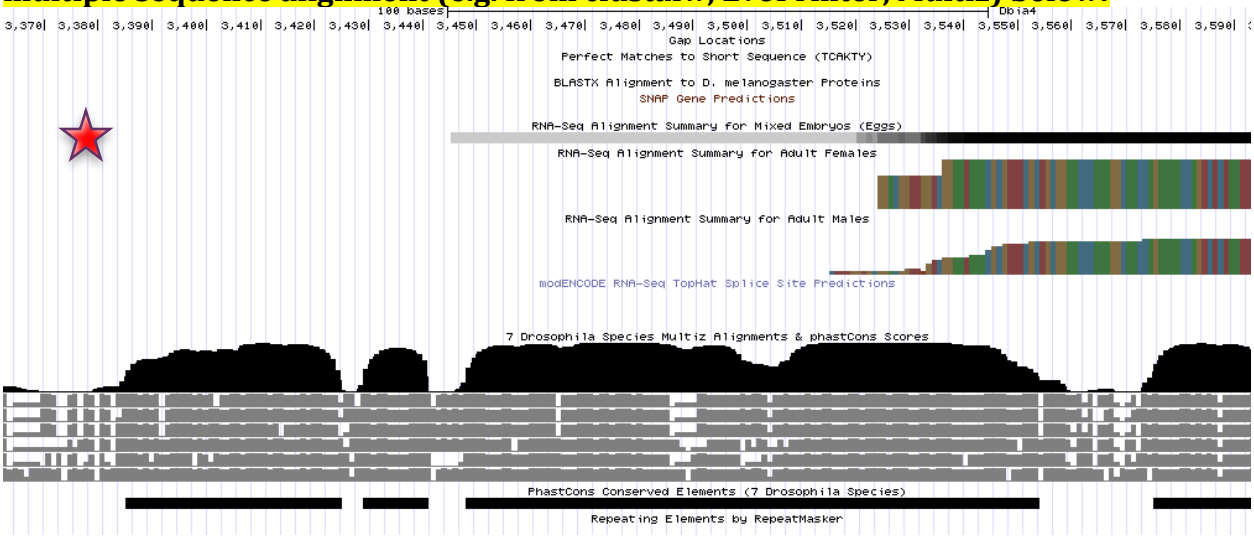
If the TSS annotation is supported by RNA-Seq or RNA polymerase II data, **paste a Genome Browser screenshot of the region around the TSS (±2kb) with the evidence tracks listed below:**

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:**



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 (±300bp) 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**	***D. melanogaster***

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 <u>OPTIONAL</u> 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 *D. melanogaster*	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_38867 Length: 40000 Number of Matches: 13

Range 1: 3476 to 3706 Graphics ▼ Next Match ▲ Previous Match

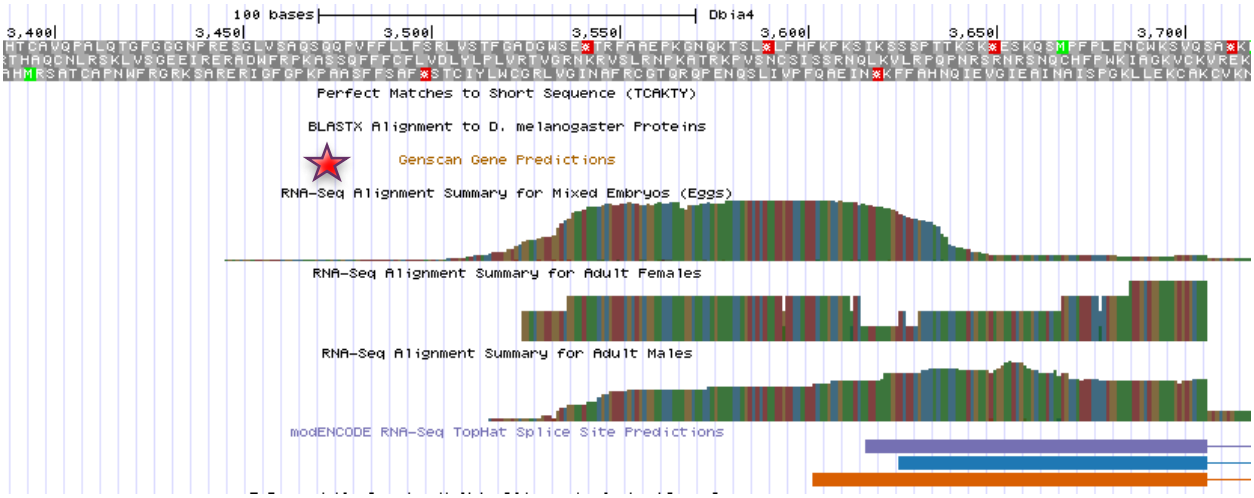
Score	Expect	Identities	Gaps	Strand
226 bits(157)	4e-62	196/231(85%)	6/231(2%)	Plus/Plus

```
Query  1     AGCCAGTTTTT-----GCTTTTTAGTCGATTTGTATCTACCTTTGGTGCGGACGGTTGGT  55
             ||||||||||      |||||||||||| ||||||||||||||||||||||||||||||
Sbjct  3476  AGCCAGTTTTTTTTCTGCTTTTTAGTCGACTTGTATCTACCTTTGGTGCGGACGGTTGGT  3535

Query  56    CGGAATAAACGCGTTTCGCAGCGGAACTCCAAGAAGAGCAGAAAACCAGTCTTTAATTGT  115
             |||||||||||||||||| ||||||| | |||   | ||||||||||||| |||||||
Sbjct  3536  CGGAATAAACGCGTTTCGCTGCGGAACCCAAAGGCAACCAGAAAACCAGTCTCTAATTGT  3595

Query  116   TCCATTTCAAGCTGAAATCAAATAAAAGTTCTTCGCCAATGGCTGAAT-AAAGTACGAAT  174
             |||||||||| |||||||| |||||||||||||| |   | |||  ||||| ||||
Sbjct  3596  TCCATTTCAAGCCGAAATCAATTAAAAGTTCTTCGCCCACAACCAAATCGAAGTAGGAAT  3655

Query  175   CAATGCAATCACTACGATTCGCAGTGGAAAATTGCTGGAAAAGTGAGCAAA  225
             | | ||||||| | | |||  | ||||||||||||||||||||| |||||
Sbjct  3656  CGAAGCAATCAATGCCATTTCCCCTGGAAAATTGCTGGAAAAGTGTGCAAA  3706
```

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

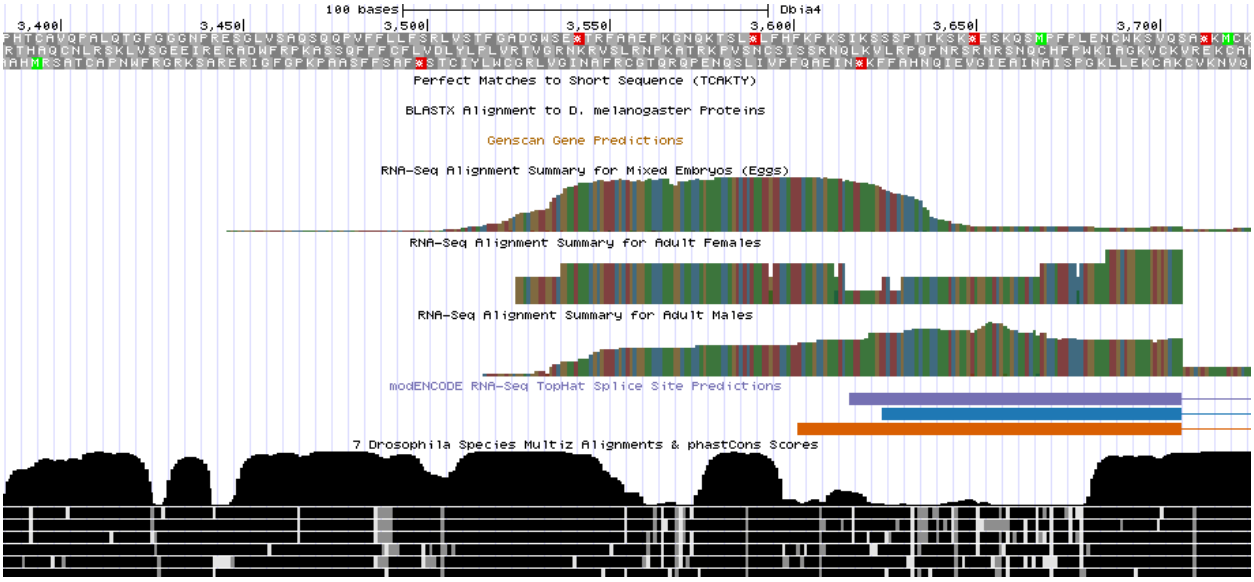
If the TSS annotation is supported by RNA-Seq or RNA polymerase II data, **paste a Genome Browser screenshot of the region around the TSS (±2kb) with the evidence tracks listed below:**

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 (±300bp) 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**	***D. melanogaster***
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 <u>OPTIONAL</u> 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 *D. melanogaster*	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

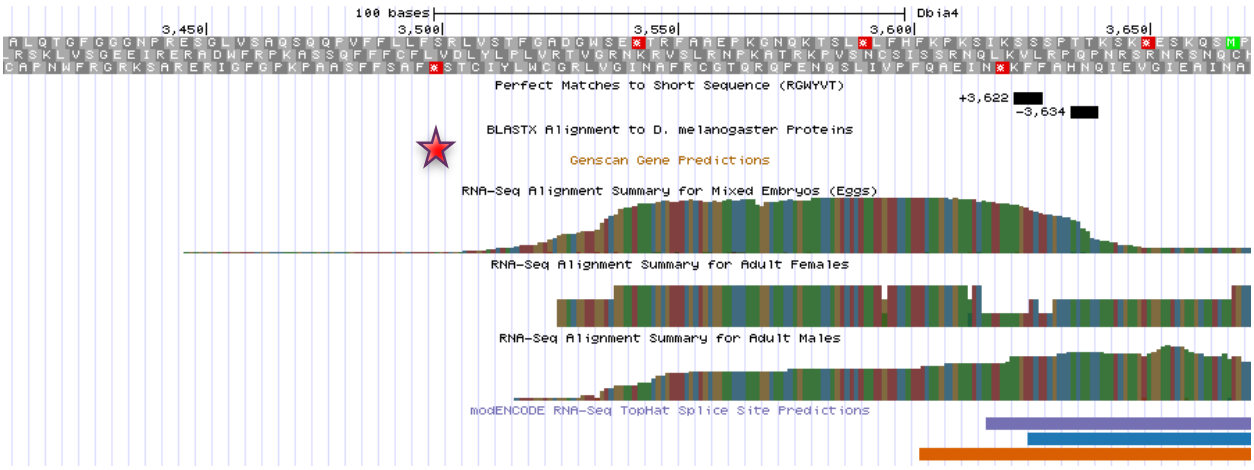
Score	Expect	Identities	Gaps	Strand
162 bits(112)	5e-43	123/134(92%)	0/134(0%)	Plus/Plus

```
Query  1     AGTCGATTTGTATCTACCTTTGGTGCGGACGGTTGGTCGGAATAAACGCGTTTCGCAGCG  60
             |||||| |||||||||||||||||||||||||||||||||||||||||||||||||| |||
Sbjct  3499  AGTCGACTTGTATCTACCTTTGGTGCGGACGGTTGGTCGGAATAAACGCGTTTCGCTGCG  3558

Query  61    GAACTCCAAGAAGAGCAGAAAACCAGTCTTTAATTGTTCCATTTCAAGCTGAAATCAAAT  120
             |||| | |||   | |||||||||||||| |||||||||||||||||| |||||||| |
Sbjct  3559  GAACCCAAAGGCAACCAGAAAACCAGTCTCTAATTGTTCCATTTCAAGCCGAAATCAATT  3618

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

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

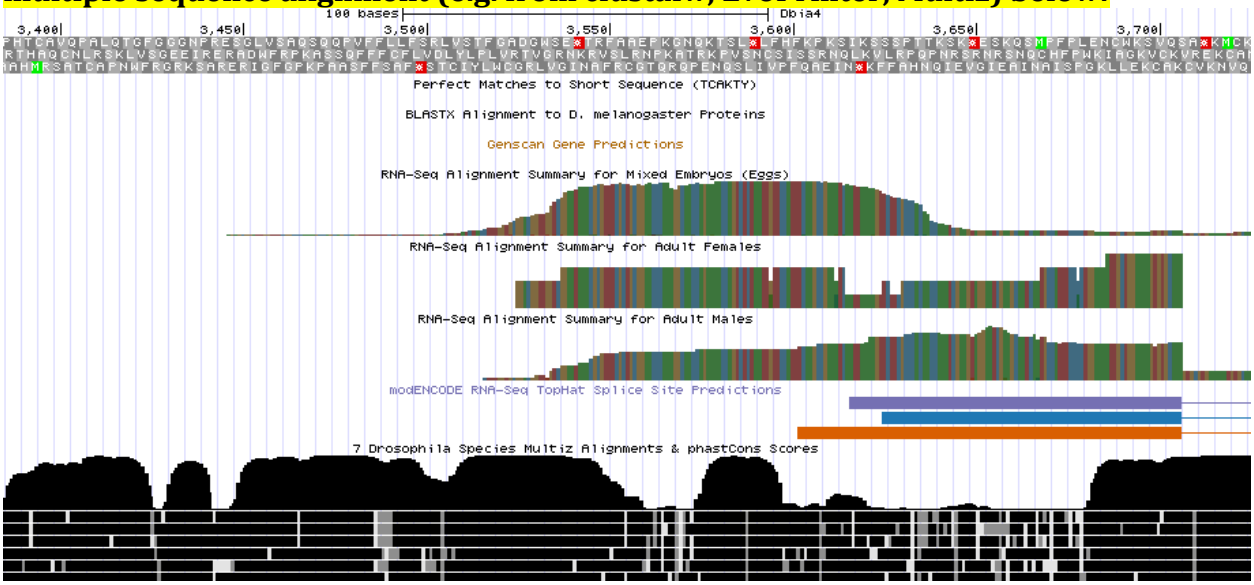
If the TSS annotation is supported by RNA-Seq or RNA polymerase II data, **paste a Genome Browser screenshot of the region around the TSS (±2kb) with the evidence tracks listed below:**

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 (±300bp) 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**	***D. melanogaster***
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

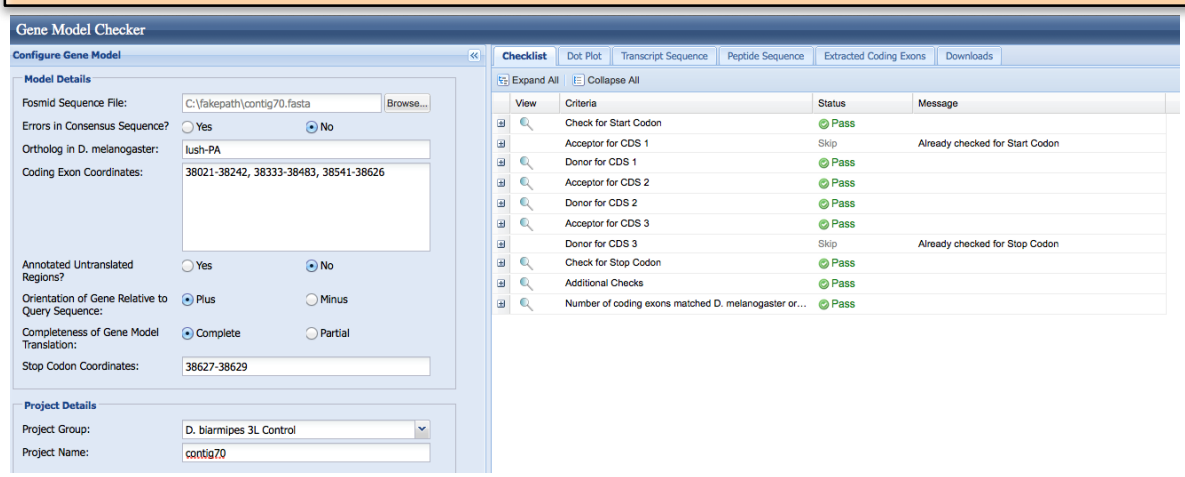
Complete this report form for each unique isoform listed in the table above (copy and paste to create as many copies of this Isoform Report Form as needed):

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.

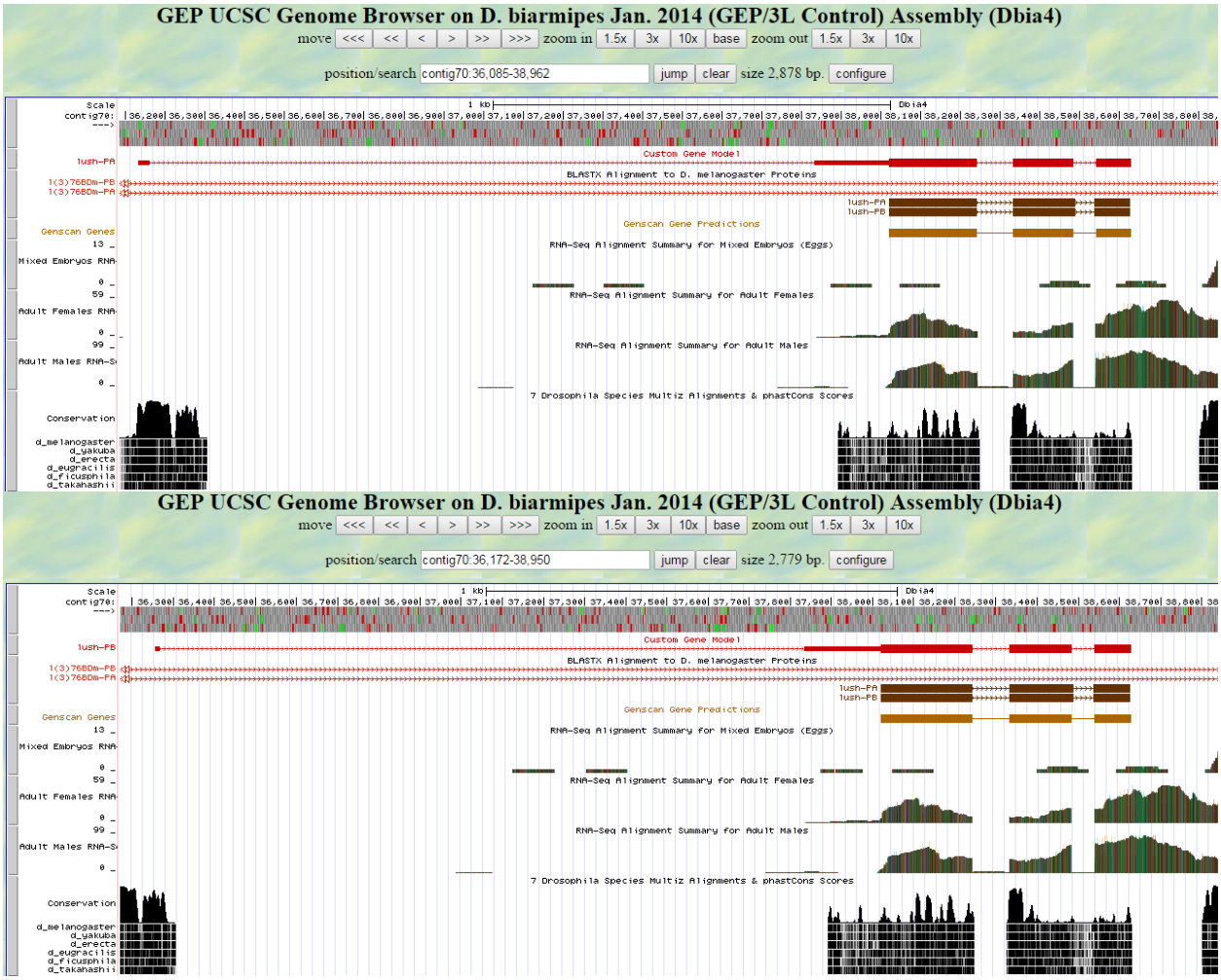


2. View the gene model on the Genome Browser

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

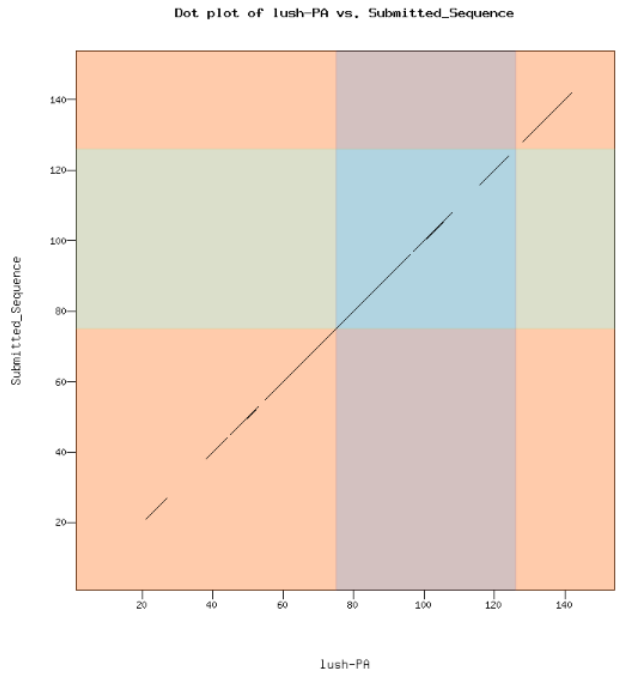
Identity: 128/153 (83.7%), **Similarity:** 139/153 (90.8%), **Gaps:** 0/153 (0.0%)

```
lush-PA             1 MKHWKRRSSAVFAIVLQVLVLLLPDPAVAMTMEQFLTSLDMIRSGCAPKFKLKTEDLDRL 60
                      *:**::***:*::*** ** ******..****:***:******.*******: ******
Submitted_Seq       1 MRHWRQRSSSVLTIVLAVLGLLLPDPGTAMTMDQFLASLDMIRNGCAPKFKLNIEDLDRL 60

lush-PA            61 RVGDFNFPPSQDLMCYTKCVSLMAGTVNKKGEFNAPKALAQLPHLVPPEMMEMSRKSVEA 120
                      ***********************************.***********.**:***:*****
Submitted_Seq      61 RVGDFNFPPSQDLMCYTKCVSLMAGTVNKKGEFNAAKALAQLPHLVPTEMIEMSKKSVEA 120

lush-PA           121 CRDTHKQFKESCERVYQTAKCFSENADGQFMWP 153
                      ***:** ***************::***:*:****
Submitted_Seq     121 CRDAHKAFKESCERVYQTAKCLAENAEGKFMWP 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:

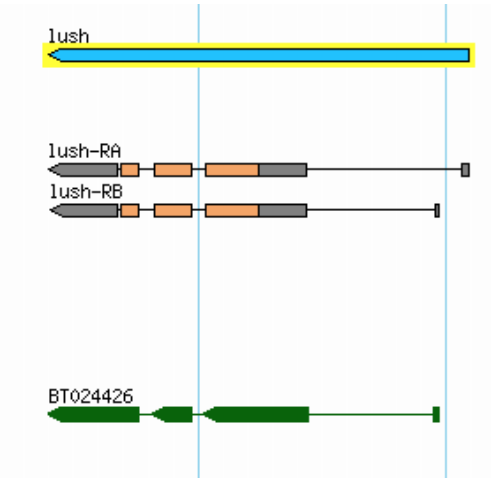
Evidence type	Support TSS annotation?	Refute TSS annotation?
blastn alignment of the initial exon from *D. melanogaster*	X	
RNA-Seq coverage and TopHat splice junctions	X	
Core promoter motifs		X
Sequence conservation with other Drosophila species	X	
Other (please specify)		

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

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     AAATGGCGTGCGAGTGTGTTACTTAATTATTAATCCCTGCGATTTAATTACCACCTTTCC  126
              |||||||||| |||| |||  |||||||  |   |      |  | ||  ||| || |||
Sbjct  37893  AAATGGCGTGCAAGTGCGTTGTTTAATTACCAGAGCTACAAACGTGATAGCCAACTGTCC  37952

Query  127    CTGCCACCATGATCA-CCTATAAAACTCTCCTACGACATGGTTACTCAACGTATTTAGCT  185
                ||||  |||   ||  |||||||    ||||  |    |||  ||  ||  ||||||  | |
Sbjct  37953  AGCCCAC-ATGGGCAGCCTATAAAGGCCTCCGCCTCGTTGGCTAGTCGTCGTATTCCG-T  38010

Query  186    TTCCGCCACCATGAAGCATTGGAAACGACGCTCTTCCGCTGTTTTCGCGATCGTCCTGCA  245
              | ||||||||||||| |||||||| | ||||||||||| | |||  |  | ||||||||||
Sbjct  38011  TCCCGCCACCATGAGGCATTGGAGGCAACGCTCTTCCTCCGTTCTGACCATCGTCCTGGC  38070

Query  246    AGTGCTGGTACTCCTGCTACCCGATCCTGCAGTTGCCATGACGATGGAGCAGTTCTTGAC  305
              |||  |||  |||||  | || |||||||      |||||||||||||| |||||  ||| |
Sbjct  38071  AGTTTTGGGCCTCCTCTTGCCAGATCCTGGGACAGCCATGACGATGGACCAGTTTTTGGC  38130

Query  306    CTCGCTAGACATGATCCGCAGTGGCTGTGCGCCGAAGTTTAAGCTCAAAACAGAAGATCT  365
              ||||||  ||  |||||||| | |||  ||||||||||||||||||||  ||  | |||||||||
Sbjct  38131  CTCGCTGGATATGATCCGGAATGGTTGTGCGCCGAAGTTTAAGCTTAACATAGAAGATCT  38190

Query  366    CGATCGGCTTCGCGTGGGTGATTTCAACTTTCCGCCATCGCAGGATCTTATG  417
              |||||||||||||||||  ||||| ||  |||||||| |||||||||| |||
Sbjct  38191  CGATCGGCTTCGCGTGGGGGATTTTAATTTTCCGCCGTCGCAGGATCTCATG  38242
```

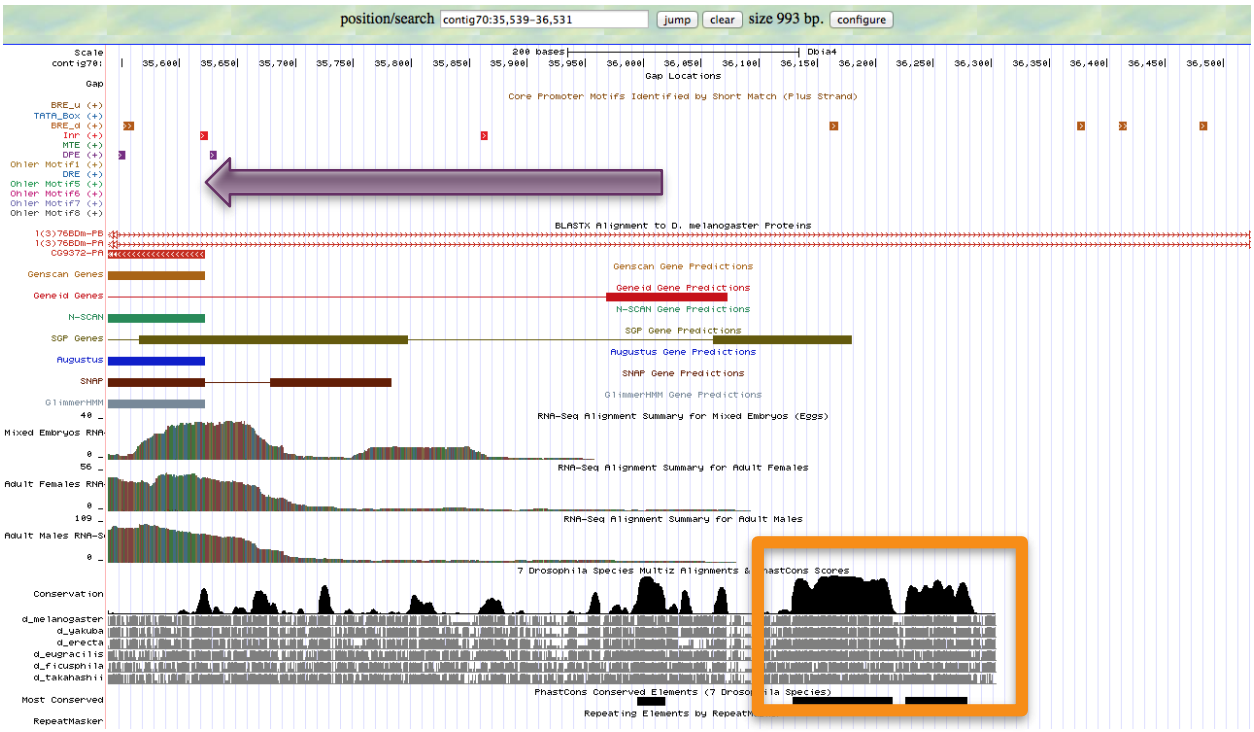
This is the BLASTn alignment for the second transcribed exon of lush-PA. It does not align for the first 66 bases, but it gives a search region between about 37793-37893 to look for more evidence.

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 (±2kb) with the evidence tracks listed below:**

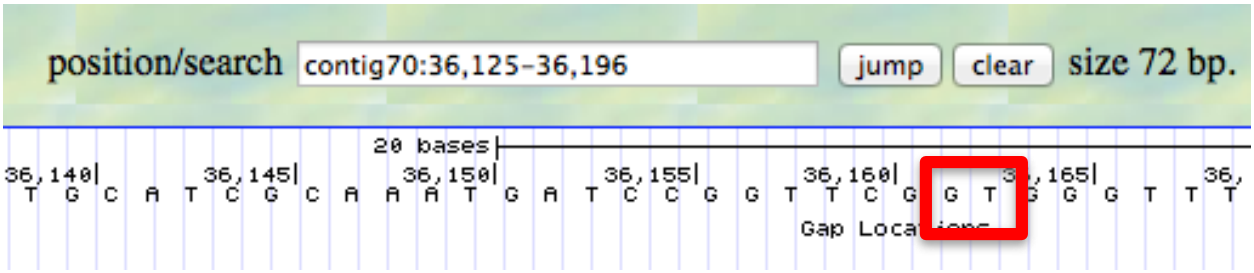
4. Short Match results for the Inr motif (TCAKTY)
5. RNA-Seq Alignment Summary
6. RNA-Seq TopHat

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 (±300bp) 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	*D. melanogaster*
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 <u>OPTIONAL</u> 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 *D. melanogaster*	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

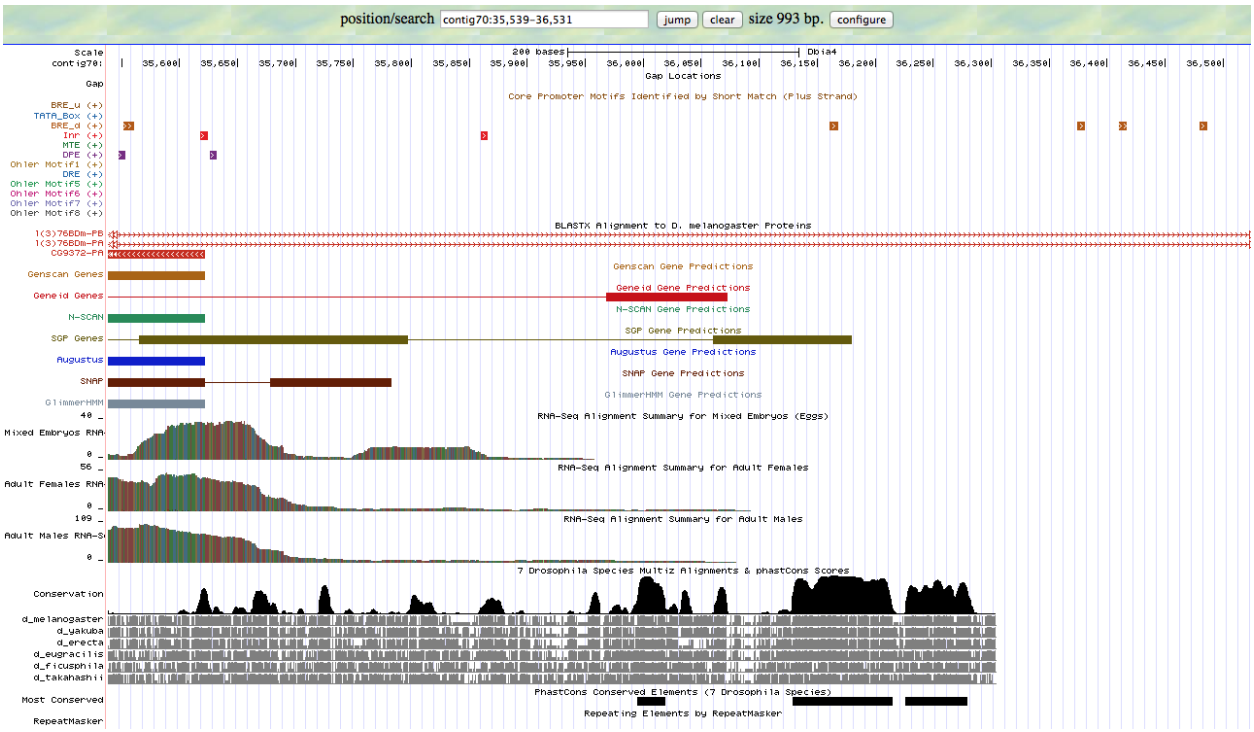
```
Query  1      CTTGCACATCAAA  13
              |||||||||||||
Sbjct  36256  CTTGCACATCAAA  36268
```

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 (±2kb) 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 coding exons based on Augustus show all bases

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

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

```
B D      D. biarmipes  ttatattctctatttaattgccttctgacgcttgcacatcaaaac---aataataaatcaactgcatgag
B D  D. melanogaster  ttatattctctatttaattgccttttgacgcttgcacatcaaagtagtaataataaatcaattgcacggc
B D         D. yakuba  atatattctctatttaattgcattttgacgcttgcacatcaaagcagtaataataaatcaattgcatggc
B D         D. erecta  ctatattctctatttaattgcctttgacgcttgcaaatcaaagcagtaataataaatcaattgcatggc
B D     D. eugracilis  ttatattcactatcttattgcctttgacgcttgcacatcaaaacaataataataaatcaattacatgca
B D     D. ficusphila  ttat-ttctctacttaattgcctttgacgcttatacatcaaaac---agtaataaataaattacatgcc
B D     D. takahashii  ttatattctctatttaattgcctcttgacgcttgcacatcaaaac---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 (±300bp) 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	*D. melanogaster*
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

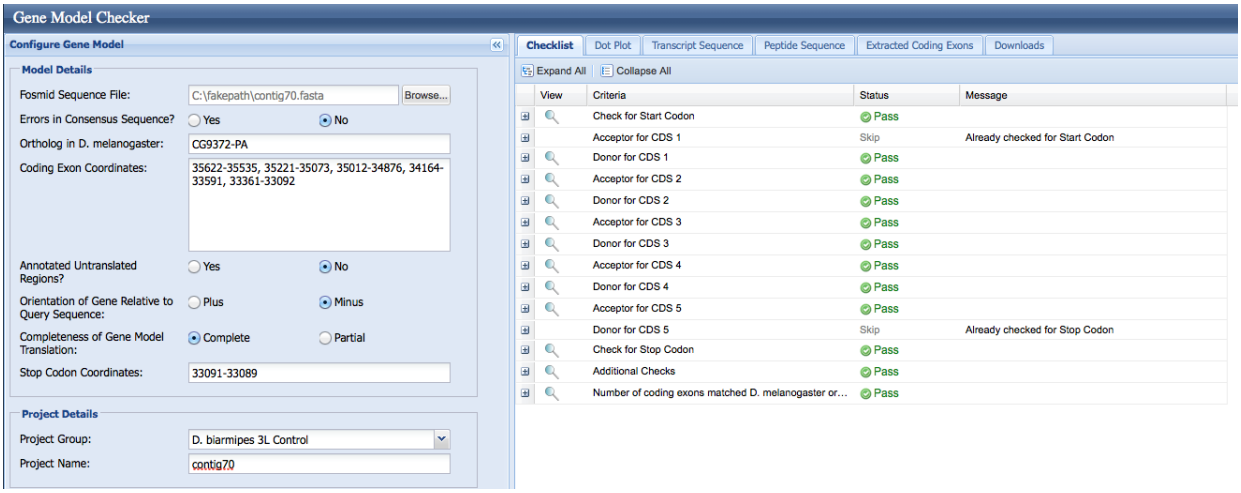
Complete this report form for each unique isoform listed in the table above (copy and paste to create as many copies of this Isoform Report Form as needed):

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.



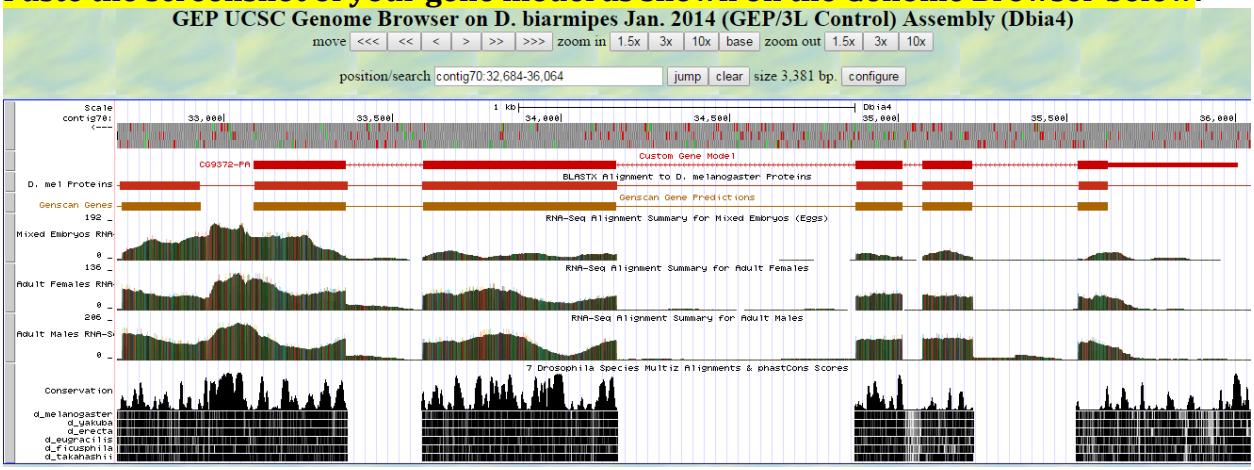
2. View the gene model on the Genome Browser

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

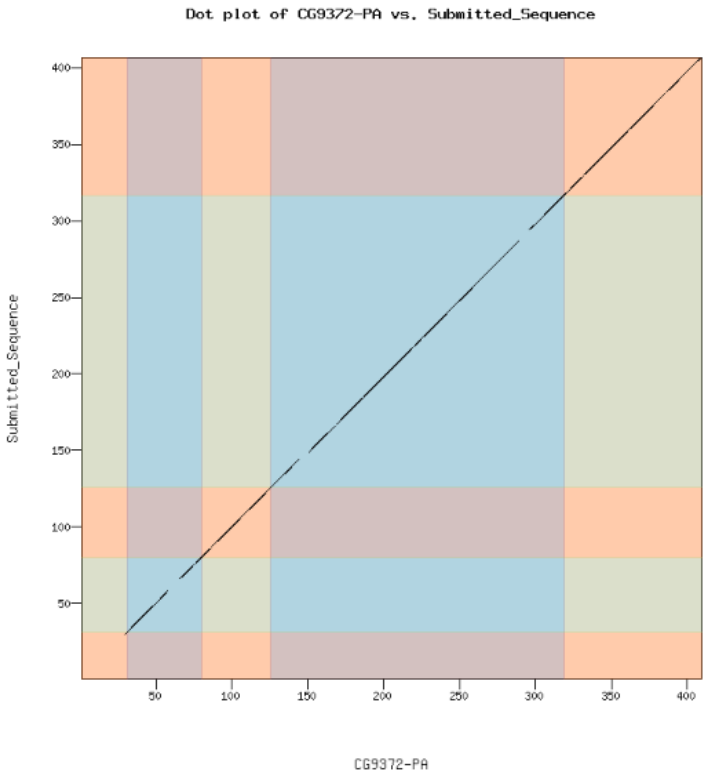
Identity: 373/408 (91.4%), **Similarity:** 397/408 (97.3%), **Gaps:** 2/408 (0.5%)

```
CG9372-PA           1 MKAFLWALVILLGYIPQSTIAKTVSTDPLDLLDFSDNDEFQWGESENQVYENRTGENRVV 60
                      ****:**:*:***:   *   :***:::*::********************:*******:.*
Submitted_Seq       1 MKAFIWAFVVLLGFSVQRILAKTITSDYIDLLDFSDNDEFQWGESENQVFENRTGENQAV 60

CG9372-PA          61 SFLSQHRLNKRQAPTSQLLENKDYGACSTPLGESGRCRHIIYCRMPELKNDVWRLVSQLC 120
                      .**::*********.**********.**********************************
Submitted_Seq      61 NFLTRHRLNKRQAPNSQLLENKDYGPCSTPLGESGRCRHIIYCRMPELKNDVWRLVSQLC 120

CG9372-PA         121 IIEKSSIGICCTDQSTSNRFSPQVVTSADGDEPRIVNKPEQRGCGITSRQFPRLTGGRPA 180
                      **********************:**..*  ***************:************
Submitted_Seq     121 IIEKSSIGICCTDQSTSNRFSPQIVTNPD--EPRIVNKPEQRGCGITTRQFPRLTGGRPA 178

CG9372-PA         181 EPDEWPWMAALLQEGLPFVWCGGVLITDRHVLTAAHCIYKKNKEDIFVRLGEYNTHMLNE 240
                      ************************************:********************
Submitted_Seq     179 EPDEWPWMAALLQEGLPFVWCGGVLITDRHVLTAAHGIHKKNKEDIFVRLGEYNTHMLNE 238

CG9372-PA         241 TRARDFRIANMVLHIDYNPQNYDNDIAIVRIDRATIFNTYIWPVCMPPVNEDWSDRNAIV 300
                      ************************************:************:****::*****
Submitted_Seq     239 TRARDFRIANMVLHIDYNPQNYDNDIAIVRIDRATLFNTYIWPVCMPPINEDWAERNAIV 298

CG9372-PA         301 TGWGTQKFGGPHSNILMEVNLPVWKQSDCRSSFVQHVPDTAMCAGFPEGGQDSCQGDSGG 360
                      ****************************:*************************
Submitted_Seq     299 TGWGTQKFGGPHSNILMEVNLPVWKQSDCRASFVQHVPDTAMCAGFPEGGQDSCQGDSGG 358

CG9372-PA         361 PLLVQLPNQRWVTIGIVSWGVGCGQRGRPGIYTRVDRYLDWILANADV 408
                      ************************:**********************
Submitted_Seq     359 PLLVQLPNQRWVTIGIVSWGVGCGERGRPGIYTRVDRYLDWILANADV 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 <u>vertical and horizontal gap</u> 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 <u>OPTIONAL</u> 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):

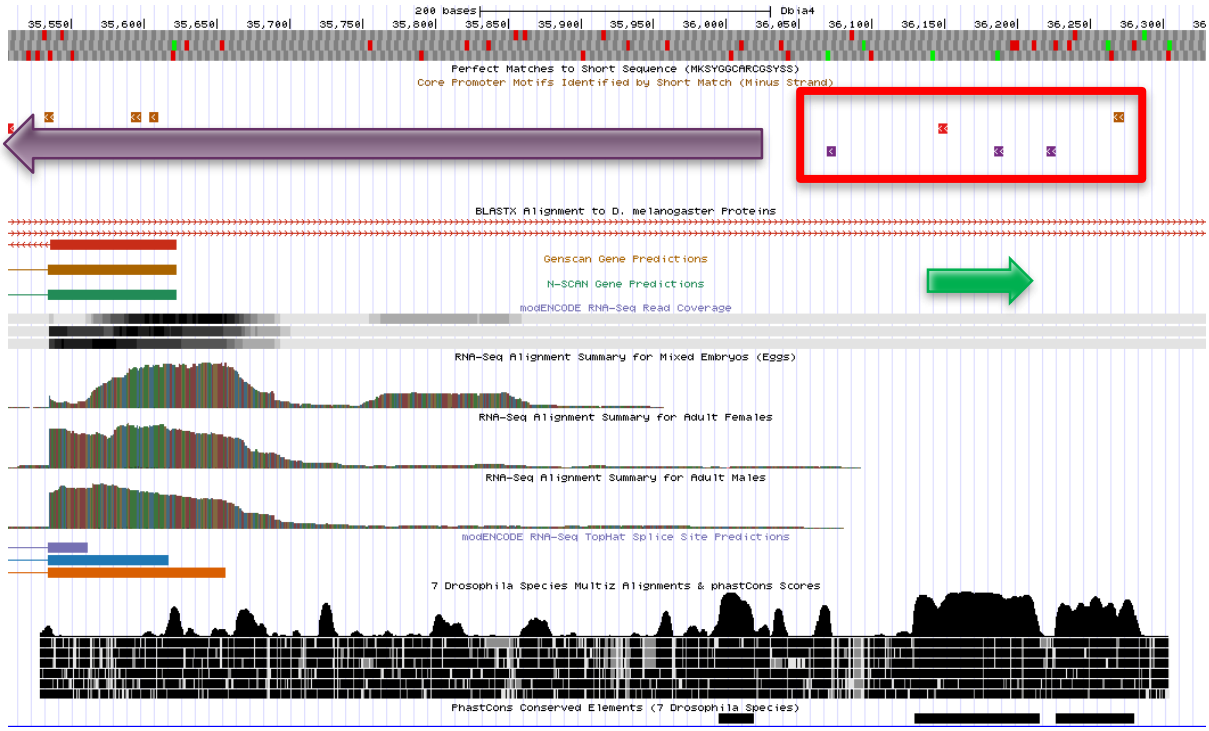
Gene-isoform name (i.e. dbia_ey-RA): dbia_CG9372-PA
Names of the isoforms with the same TSS as this isoform:
Type of core promoter: (Peaked or Broad): Broad
Coordinates of the first transcribed exon: 36,009-35,535
Coordinate(s) of TSS position(s): 36,009
Coordinate(s) of TSS search region(s): 36,000-36,300

1. Evidence that supports the TSS annotation postulated above

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

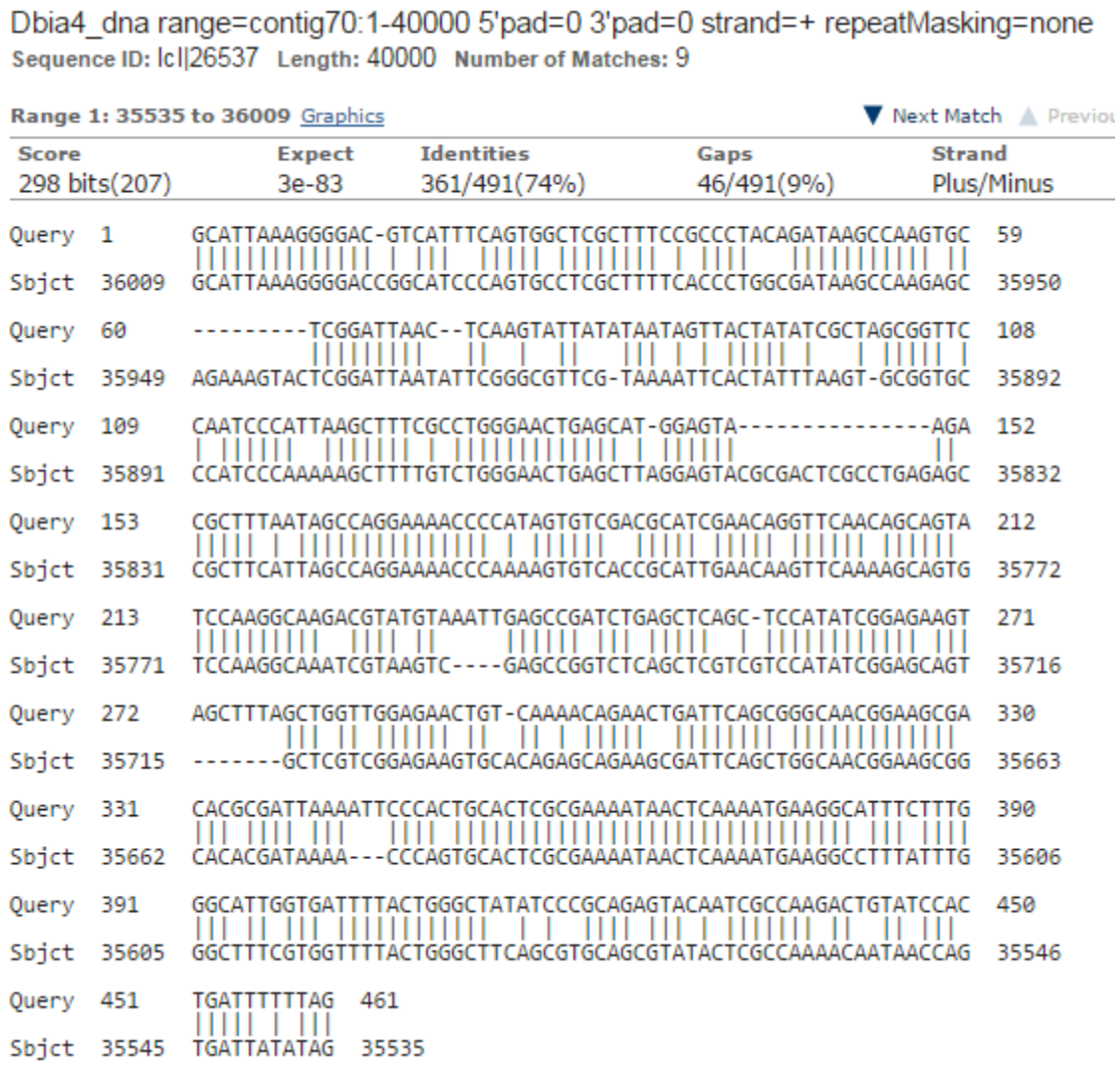
Evidence type	Support TSS annotation?	Refute TSS annotation?
blastn alignment of the initial exon from *D. melanogaster*	X	
RNA-Seq coverage and TopHat splice junctions	X	
Core promoter motifs		X
Sequence conservation with other Drosophila species	X	
Other (please specify)		

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



The core promoter motifs 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.

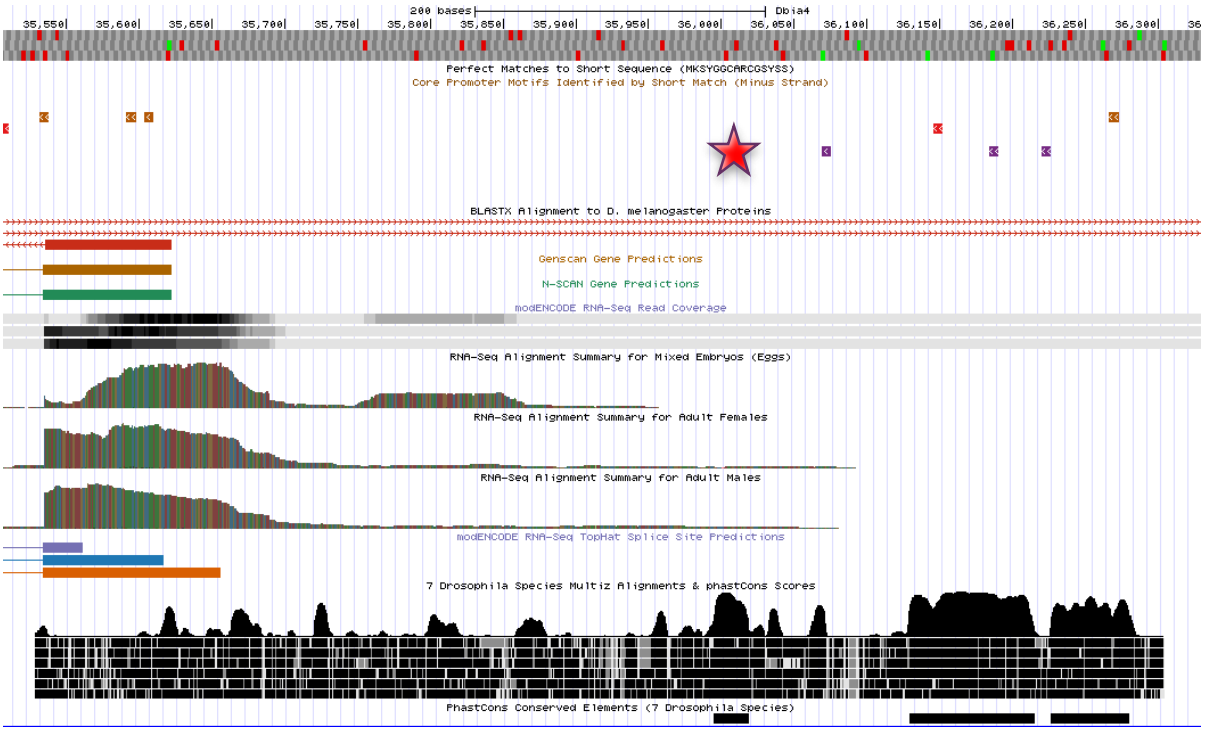
If the TSS annotation is supported by blastn alignment of the initial transcribed exon against the contig sequence, **paste a screenshot of the blastn alignment below**:



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

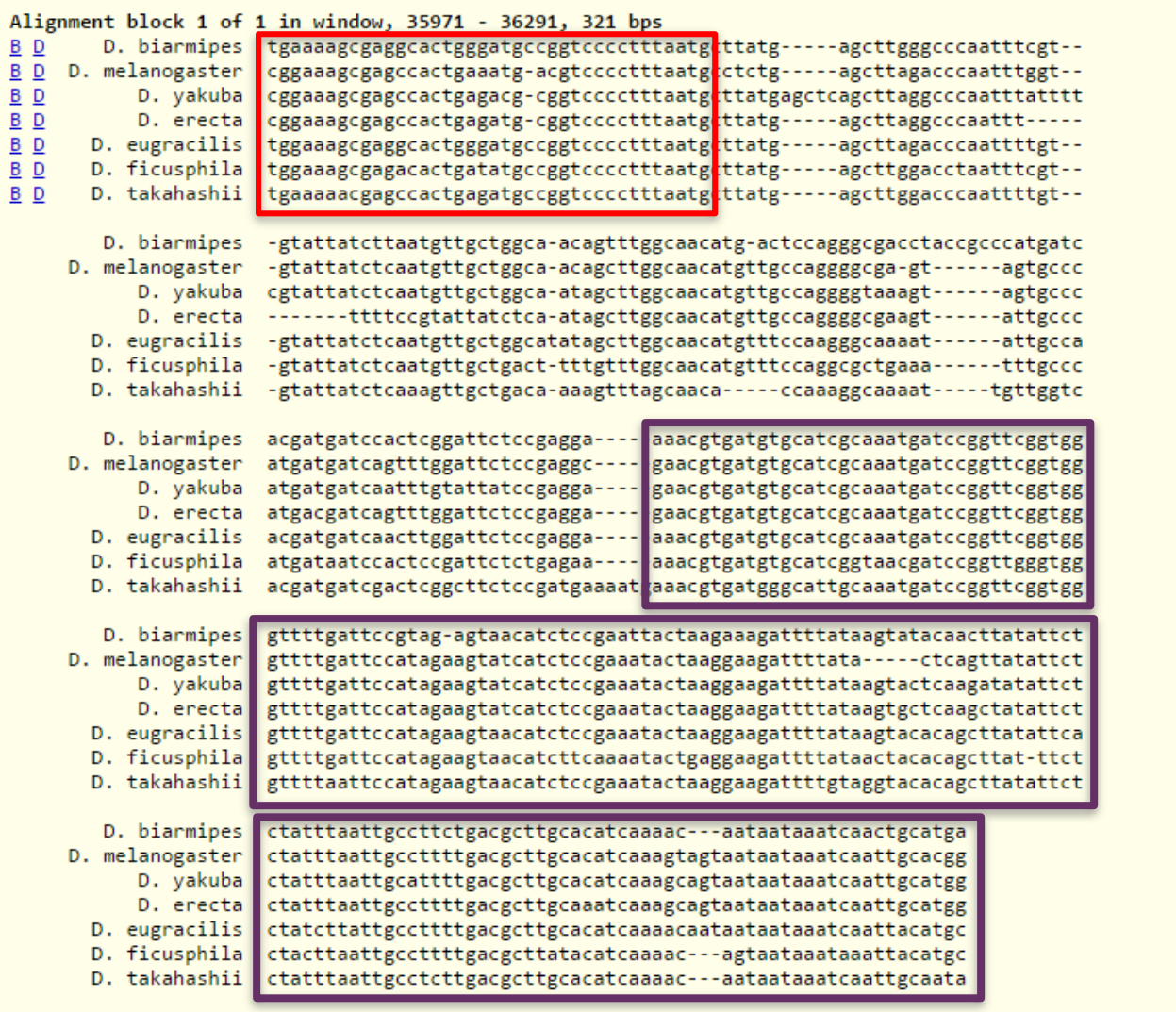
If the TSS annotation is supported by RNA-Seq or RNA polymerase II data, **paste a Genome Browser screenshot of the region around the TSS (±2kb) with the evidence tracks listed below:**

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:**



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 (±300bp) 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	*D. melanogaster*
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

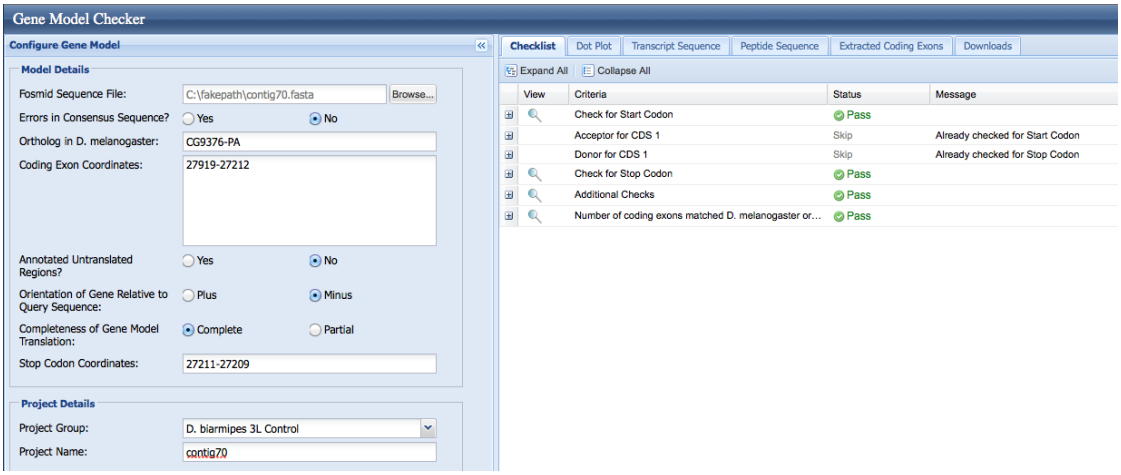
Complete this report form for each unique isoform listed in the table above (copy and paste to create as many copies of this Isoform Report Form as needed):

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.

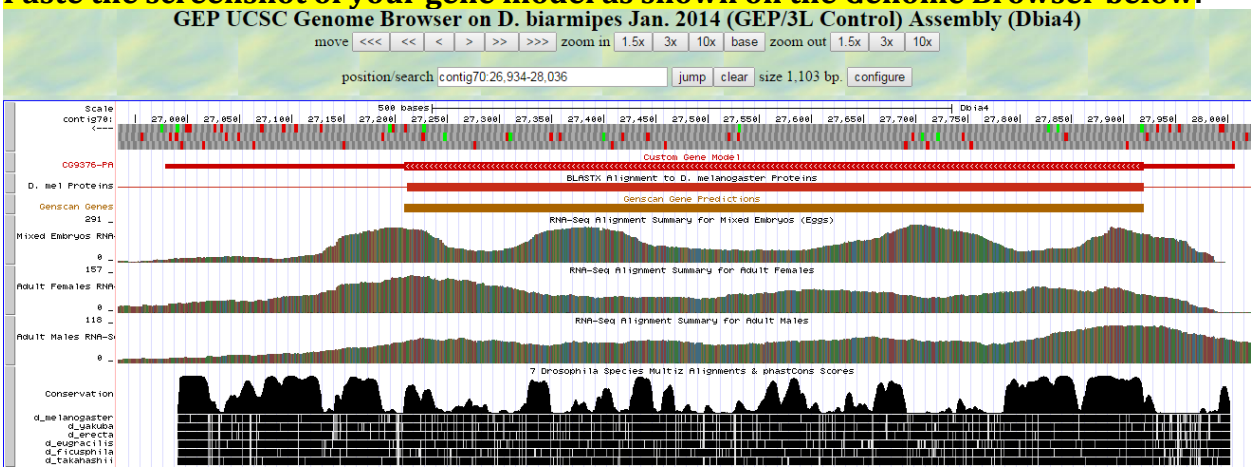


2. View the gene model on the Genome Browser

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

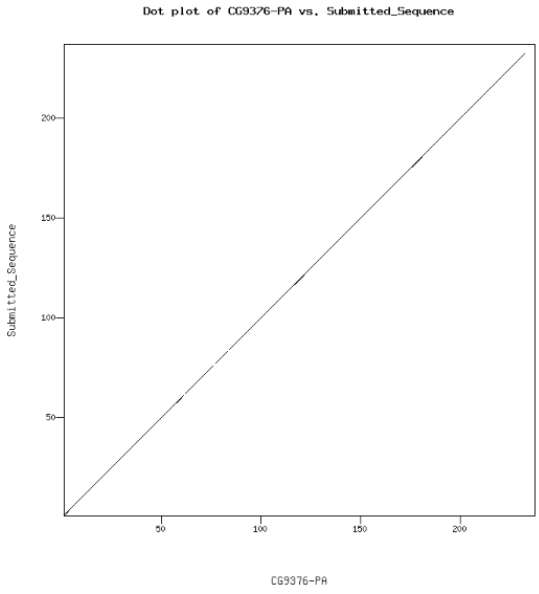
Identity: 232/236 (98.3%), **Similarity:** 234/236 (99.2%), **Gaps:** 0/236 (0.0%)

```
CG9376-PA           1 MVSRFDQQKTKHQLFHVSLSLATILICMAYMQYRRNWAHLGSFWDSLVVPIVFLGELLKV 60
                      ************************************************************
Submitted_Seq       1 MVSRFDQQKTKHQLFHVSLSLATILICMAYMQYRRNWAHLGSFWDSLVVPIVFLGELLKV 60

CG9376-PA          61 VLARFYGKVEDGVLTAKQRQKKNSYFTPRELLGGFTLQFLCTLLYAFICIILGAPVLGNY 120
                      :**************.****** ************************************
Submitted_Seq      61 ILARFYGKVEDGVLTVKQRQKKASYFTPRELLGGFTLQFLCTLLYAFICIILGAPVLGNY 120

CG9376-PA         121 EQTFVLALLMTLLTVSPTVFLLGGGGALQVCFCEKPDFVTKCEDTALNLFKYNALGGILG 180
                      ************************************************************
Submitted_Seq     121 EQTFVLALLMTLLTVSPTVFLLGGGGALQVCFCEKPDFVTKCEDTALNLFKYNALGGILG 180

CG9376-PA         181 AWAGSVVAPLDWGRDWQAYPIPNVIGALLGSALGNIYACTHVLYATARVYMTKKRT 236
                      ***************************************************:****
Submitted_Seq     181 AWAGSVVAPLDWGRDWQAYPIPNVIGALLGSALGNIYACTHVLYATARVYMSKKRT 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 <u>vertical and horizontal gap</u> 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 <u>OPTIONAL</u> 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):

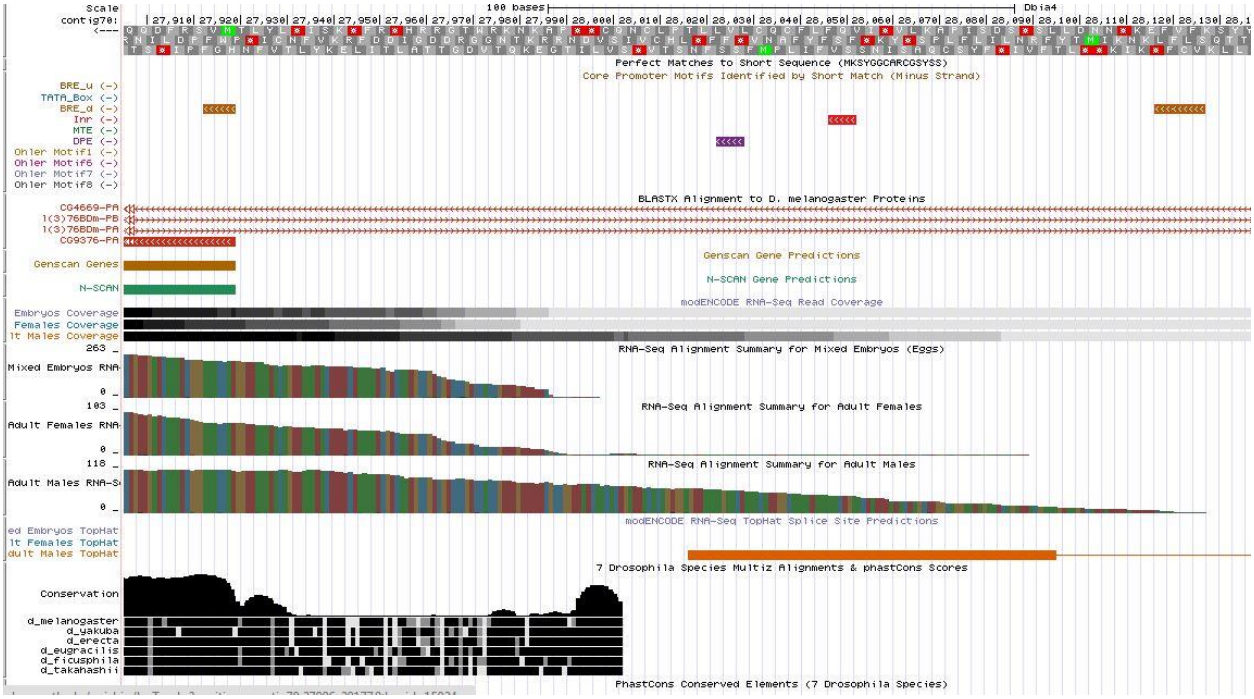
Gene-isoform name (i.e. dbia_ey-RA): <u>dbia_CG9376-PA</u>
Names of the isoforms with the same TSS as this isoform:
_____
Type of core promoter: (Peaked or Broad): <u>Peaked</u>
Coordinates of the first transcribed exon: <u>28007-26980</u>
Coordinate(s) of TSS position(s): <u>28008</u>
Coordinate(s) of TSS search region(s): <u>27970-28140</u>

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 *D. melanogaster*	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 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 ▲ Previo

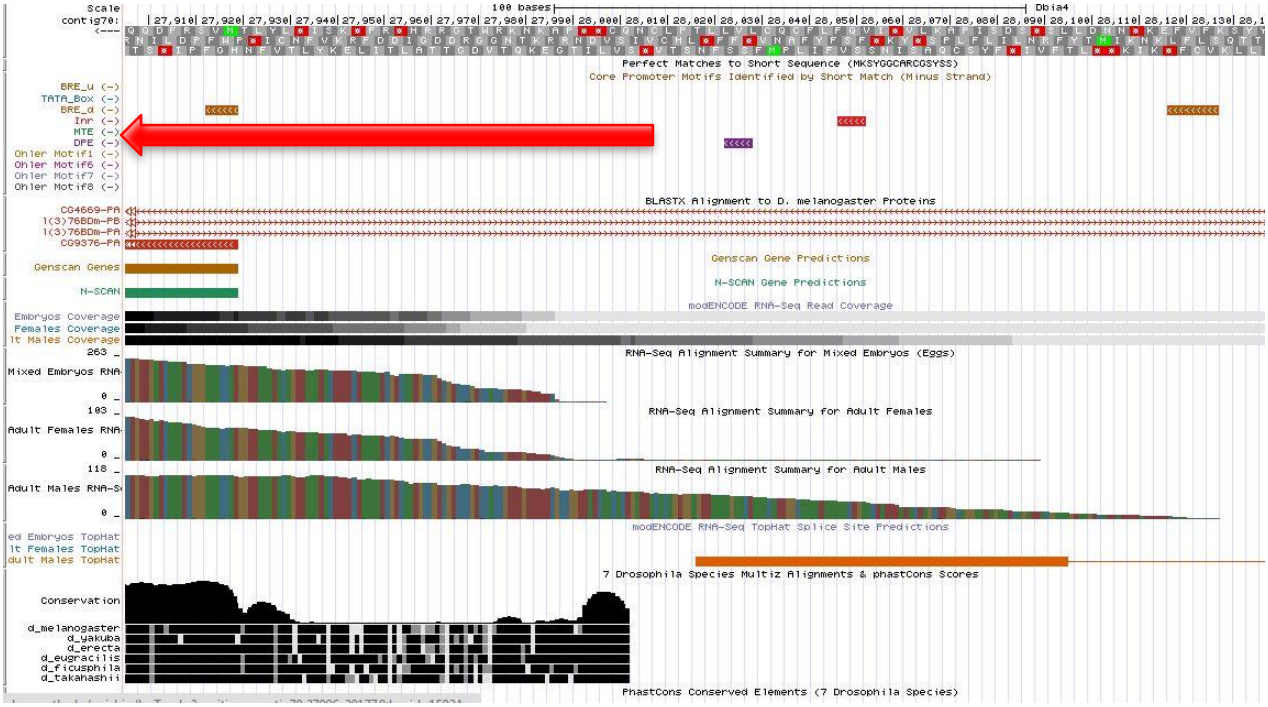
Score	Expect	Identities	Gaps	Strand
1016 bits(710)	0.0	876/1026(85%)	19/1026(1%)	Plus/Minus

```
Query  4      TCAGTGTTGATAGGCGGCGAAATACAAACGAAAGGCAGGAAAATATTGTCCAAGTTTAAG  63
              ||||||||||||  |||||||| |||||||    | | |||  |       ||  || | |
Sbjct  28004  TCAGTGTTGATAACCGGCGAAAAACAAACGGTGGACGGGACGACGG----CA--TTGACG  27951

Query  64     ATTTACCAAAATTGTATTTATCTGCATTTAATCATGGTTTCCCGATTCGACCAACAAAAG  123
              ||||   ||||  ||||||| ||| |||||| ||||||||||||||| || ||||| |||
Sbjct  27950  ATTTTAGAAAA--GTATTTAACTGTATTTAACCATGGTTTCCCGATTTGATCAACAGAAG  27893

Query  124    ACCAAGCACCAGCTCTTCCACGTGTCCCTCAGCCTCGCCACCATCCTGATCTGCATGGCG  183
              |||||||||||||||||||||||||||||||||||| |||||||||||||||||||||||
Sbjct  27892  ACCAAGCACCAGCTCTTCCACGTGTCCCTCAGCCTGGCCACCATCCTGATCTGCATGGCG  27833

Query  184    TACATGCAGTATCGCCGGAATTGGGCACATCTCGGCAGCTTCTGGGATTCTCTTGTCGTC  243
              |||||||||||||| ||||| ||||| |||||||||||||||||||||||| | || |||
Sbjct  27832  TACATGCAGTATCGGCGGAACTGGGCGCATCTCGGCAGCTTCTGGGATTCTTTGGTGGTC  27773

Query  244    CCTATTGTTTTCCTCGGAGAGCTGCTAAAAGTTGTTTTGGCTCGCTTCTACGGGAAGGTT  303
              || |||||||| ||||| ||||| || |||||| |  |||| |||||||| ||||| |||
Sbjct  27772  CCGATTGTTTTTCTCGGTGAGCTTCTGAAAGTTATACTGGCCCGCTTCTATGGGAAAGTT  27713

Query  304    GAGGATGGCGTCCTAACCGCCAAACAGCGCCAGAAAAAGAACTCGTACTTTACGCCGCGG  363
              |||||||||||||| || |  |||||||||||||||||  |||||||||||| ||||||
Sbjct  27712  GAGGATGGCGTCCTGACTGTTAAACAGCGCCAGAAAAAGGCCTCGTACTTTACACCGCGG  27653

Query  364    GAGCTCCTCGGAGGATTCACCCTGCAGTTCCTGTGTACACTCCTCTACGCCTTTATCTGC  423
              |||||||||| |||||||| |||||||| || || || || || ||||||||||||||||
Sbjct  27652  GAGCTCCTCGGGGGATTCACTCTGCAGTTTCTTTGCACGCTGCTTTACGCCTTTATCTGC  27593

Query  424    ATAATTTTGGGAGCCCCGGTGCTGGGCAACTATGAGCAGACCTTCGTCTTGGCCTTACTG  483
              || |||||||||||||||||||||||||||| |||||||||| ||| ||||||||||
Sbjct  27592  ATCATTTTGGGAGCCCCGGTGCTGGGCAACTACGAGCAGACCTTTGTCCTGGCCTTACTT  27533

Query  484    ATGACCTTGTTGACGGTGTCACCCACGGTTTTCCTCCTCGGTGGCGGAGGAGCACTACAA  543
              |||||  || |||||||||| || || |||||||| ||||| ||||| ||||| || ||
Sbjct  27532  ATGACACTGCTGACGGTGTCGCCAACTGTTTTCCTGCTCGGGGGCGGCGGAGCCCTCCAG  27473
```

```
Query  544    GTGTGCTTCTGCGAGAAACCGGACTTTGTGACCAAGTGCGAGGACACGGCTCTGAATCTG  603
              || |||||||| ||||| ||||||||||||||||||||||||||| ||||| ||||| |||
Sbjct  27472  GTTTGCTTCTGTGAGAAGCCGGACTTTGTGACCAAGTGCGAGGATACGGCGCTGAACCTG  27413

Query  604    TTCAAGTACAATGCACTGGGCGGTATTCTGGGAGCTTGGGCCGGGAGCGTGGTCGCTCCA  663
              || |||||||||| |||||||| ||||||||| || |||||||| || ||||| |||||
Sbjct  27412  TTTAAGTACAATGCGCTGGGCGGGATTCTGGGCGCCTGGGCCGGAAGTGTGGTAGCTCCT  27353

Query  664    CTAGACTGGGGACGTGACTGGCAGGCTTATCCCATTCCCAACGTGATCGGAGCACTGCTG  723
              ||||||||||||| |||||||||| || || ||||| || ||||| ||||||||  ||
Sbjct  27352  CTAGACTGGGGACGCGACTGGCAGGCATACCCTATTCCGAATGTGATTGGAGCACTATTG  27293

Query  724    GGAAGCGCTCTGGGCAATATATACGCTTGTACGCATGTCCTCTACGCCACAGCTCGAGTT  783
              || |||||||||| || |||||||| ||||| || ||||| || ||||| || || |||
Sbjct  27292  GGCAGCGCTCTGGGTAACATATACGCCTGTACACACGTCCTTTATGCCACGGCCCGTGTT  27233

Query  784    TACATGACCAAGAAACGCACTTAAATCTAATAATAAAA-CTCTTCATTCTGCCACCAAGA  842
              || |||| |||||||||||||||| ||   ||||||  || || || ||||||||||||
Sbjct  27232  TATATGAGCAAGAAACGCACTTAAACCTTGCAATAAATGCTATTAATCCTGCCACCAAGA  27173

Query  843    TTATTACGTTGCGCTGCCACTGAAATCCAC----TCATCCACTAAGAACCACGTCATCAT  898
              ||||||||||||||||| ||||| |||||    ||||||||||||||||||||||||||
Sbjct  27172  TTATTACGTTGCGCTGCCTCTGAATTCCACACACTCATCCACTAAGAACCACGTCATCAT  27113

Query  899    CCCGTCTAGTCAGCAGCTTAGGAACTACCAAAAAGCAATTACCACTTTAATGTTACATGT  958
              | || |||||||||||||||||||||||||||||||||||| || ||||| ||||||| |
Sbjct  27112  CTCGACTAGTCAGCAGCTTAGGAACTACCAAAAAGCAATTAACAATTTAAAGTTACATAT  27053

Query  959    TTAATCATATTACTAGCGTGGATTATTAATGAATTCTTTTTAAATCAATCAA-TAATAAA  1017
              |||| ||| ||||||| || |  ||||||   |||||||||||||||||| |||||||
Sbjct  27052  TTAAGCATTTTACTAGTGTAG--TATTAA---ATTCTTTTTAAATCAATCAATTAATAAA  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 (±2kb) 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 coding ▾ exons based on Augustus ▾ show all ▾ bases

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

```
Alignment block 1 of 1 in window, 27896 - 28002, 107 bps
B D      D. biarmipes  ctgttgatcaaatcgggaaaccatggttaaatacagttaaat---acttttctaaaatcgtcaa------
B D   D. melanogaster  ttgttggtcgaatcgggaaaccatgattaaatgcagataaatac-aattttggtaaatcttaaacttgga
B D         D. yakuba  ttgttggtcaaaacgggaaaccattgttaaatgcagataaaaac-aattttgttaaatcttaga------
B D         D. erecta  ttgttggtcaaatcgggaaaccatggttaaatgcagataaatac-aattttgttaaatcttagacttggt
B D     D. eugracilis  ctgttggtcaaatcgggaaaccatgattaaatgcaataaaatat-agttttctaaaacctccga------
B D     D. ficusphila  ttgttggtcaaatcgggaaaccatgattaaatgcagttaaatct-aaattt-tacaatctt---------
B D     D. takahashii  ctgttggtcaaatcgggaaaccatggttaaatgcagttcaatagaattttctgaagttccttaa------

         D. biarmipes  tgccg--------tcgtcccgtccaccgtttgtttttcgccggttatcaacact
      D. melanogaster  caata--------ttttcctgcctttcgtttgtatttcgccgcctatcaacact
         D. yakuba     tgtcgcttatatttttttctctcttccgtttgtatttcgccggttatcaacact
         D. erecta     tgata--------ttttcctgtcttccgtttatatttcgccggttatcaacact
       D. eugracilis   tgct---------tcttcctgccctccgtttatttttcgccggttatcaacact
       D. ficusphila   tgccg--------tctttctgactgccgtttgatttttgccgattatcaacact
       D. takahashii   tgccg--------ttttcctgacatccgtttgttttacgcccgttatcaacact
```

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 (±300bp) 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	*D. melanogaster*
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

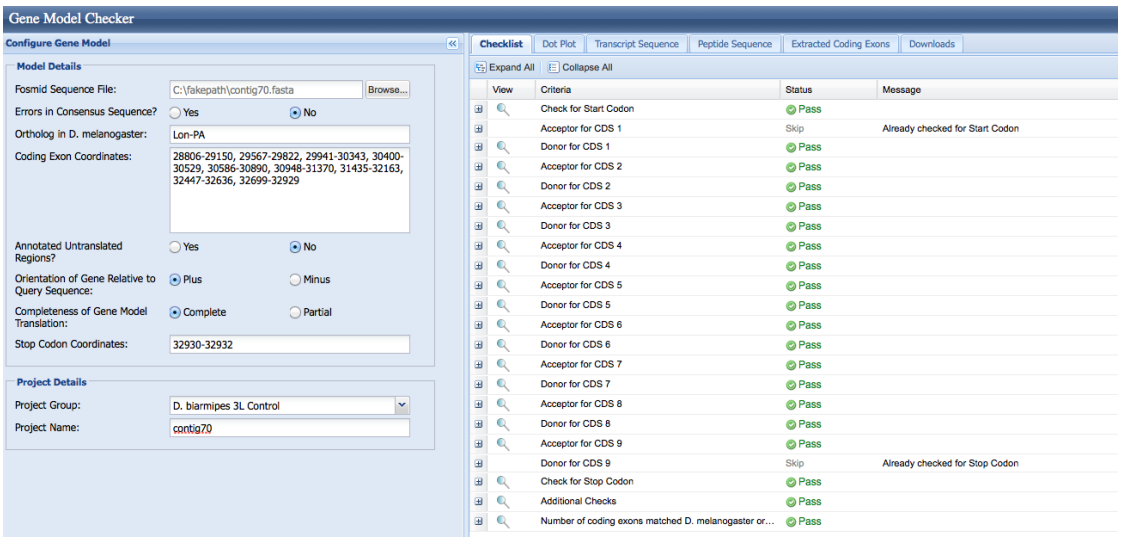
_____

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.



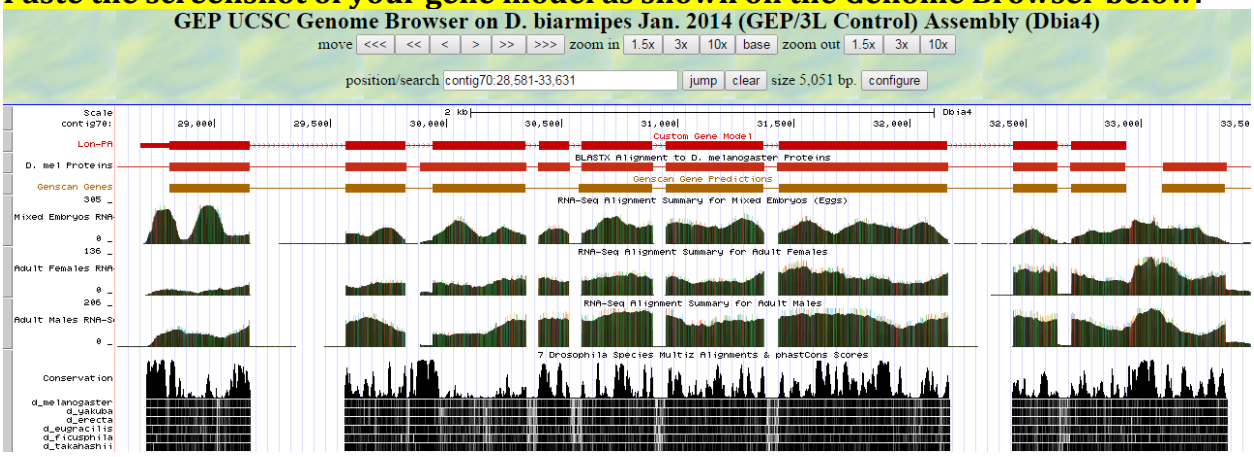
2. View the gene model on the Genome Browser

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)

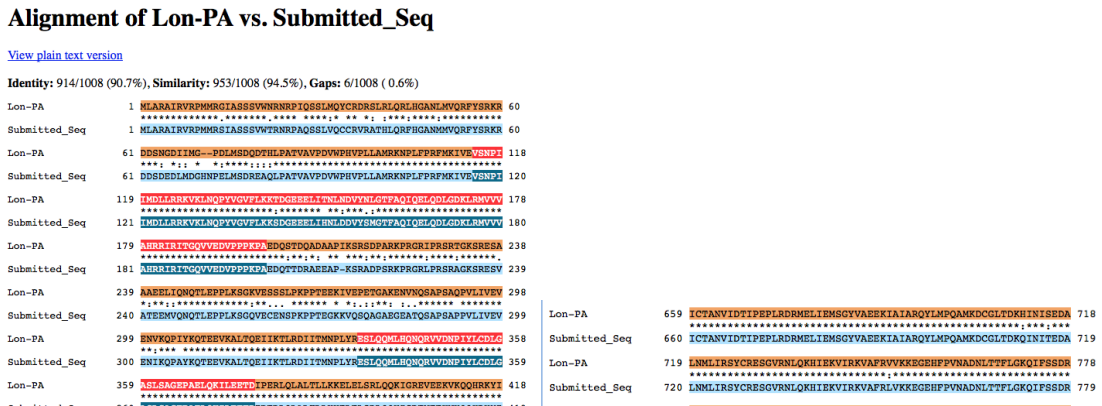
18. At least one gene prediction track (e.g. Genscan)
19. At least one RNA-Seq track (e.g. RNA-Seq Alignment Summary)
20. A comparative genomics track
    (e.g. Conservation, D. mel. Net Alignment, 3-way, 5-way or 7-way multiz)

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*

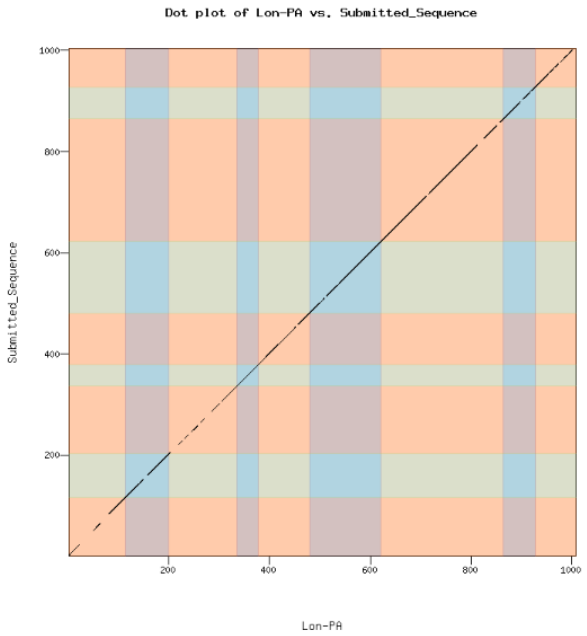
Show an alignment between the protein sequence for your gene model and the protein sequence from the putative *D. melanogaster* ortholog. You can use the protein alignment generated by the Gene Model Checker or you can generate a new alignment using BLAST 2 Sequences (*bl2seq*). **Paste a screenshot of the protein alignment below:**



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 <u>vertical and horizontal gap</u> 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

_____

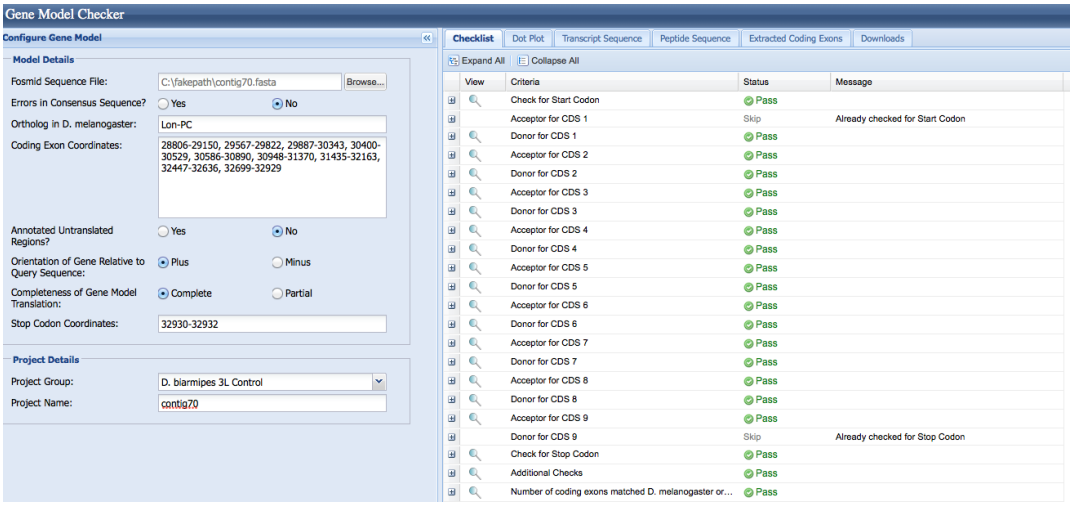
Is the 5' end of this isoform missing from the end of project: <u>No</u>
If so, how many exons are missing from the 5' end: _____
Is the 3' end of this isoform missing from the end of the project: <u>No</u>
If so, how many exons are missing from the 3' end: _____

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 <u>original project sequence</u>. Include the VCF file you have generated

above when you submit the gene model to the Gene Model Checker. The Gene Model Checker will revise the submitted exon coordinates automatically using this VCF file.

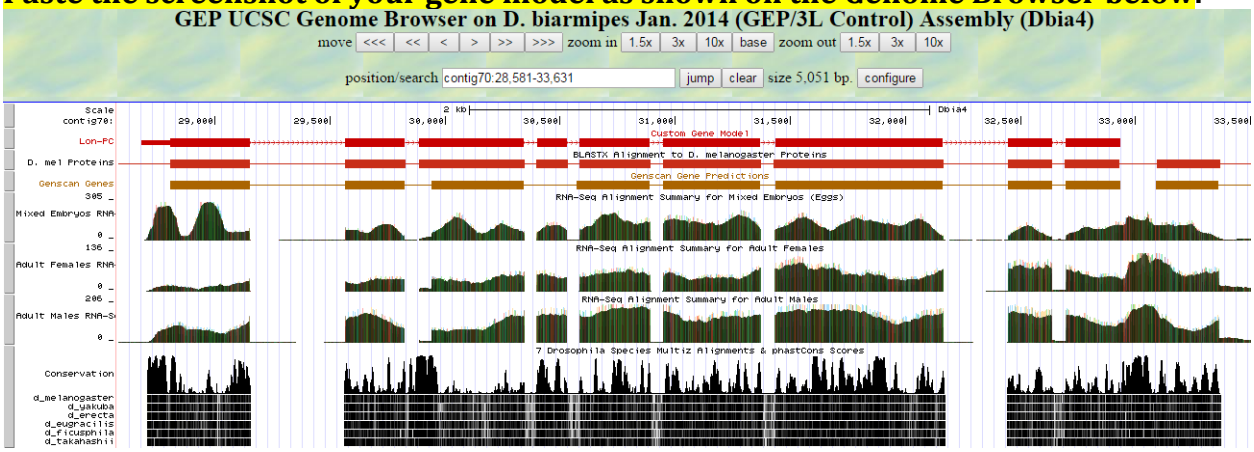


2. View the gene model on the Genome Browser

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.

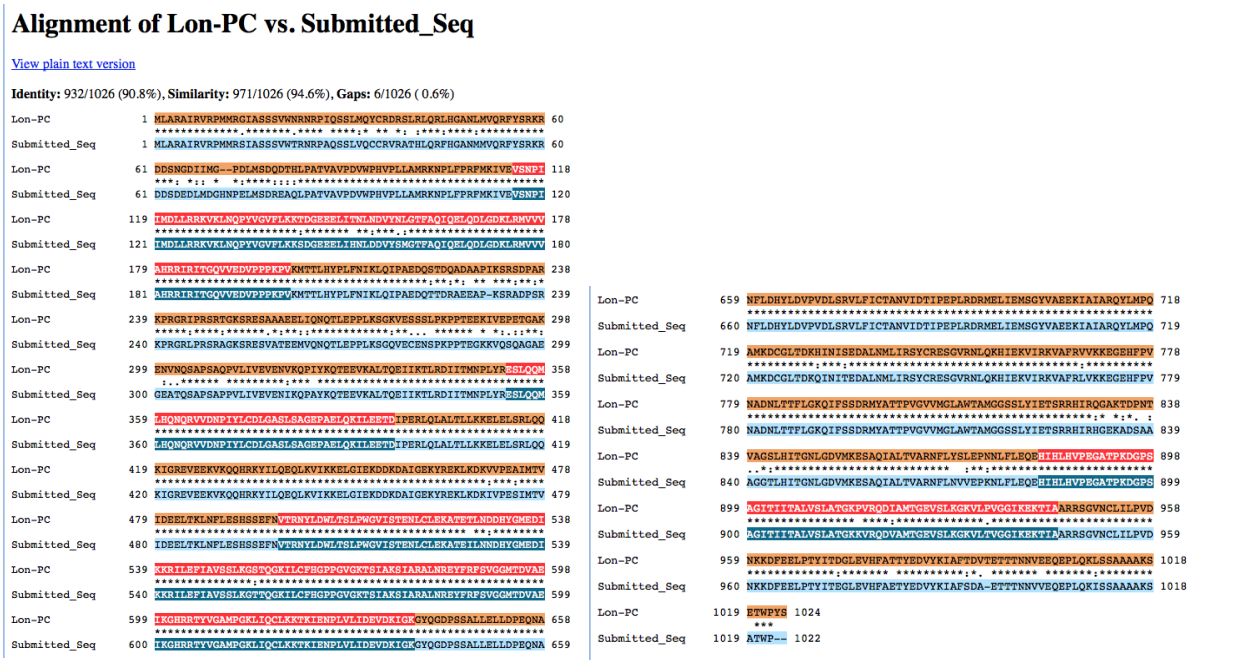
5. A sequence alignment track (D. mel Protein or Other RefSeq)
6. At least one gene prediction track (e.g. Genscan)
7. At least one RNA-Seq track (e.g. RNA-Seq Alignment Summary)
8. A comparative genomics track
   (e.g. Conservation, D. mel. Net Alignment, 3-way, 5-way or 7-way multiz)

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*

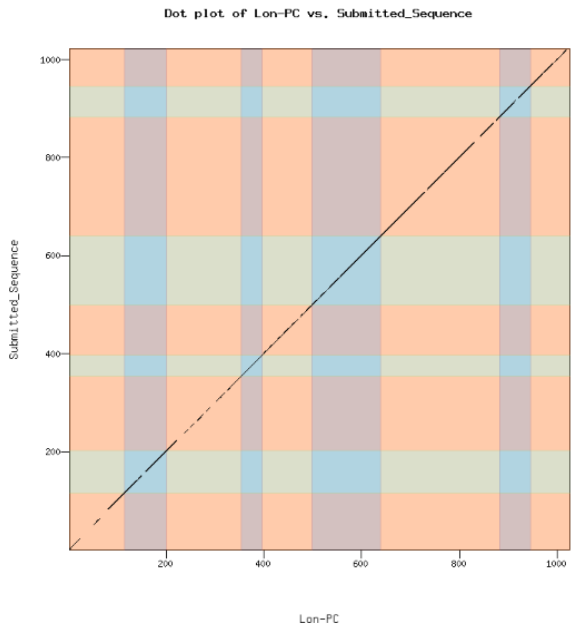
Show an alignment between the protein sequence for your gene model and the protein sequence from the putative *D. melanogaster* ortholog. You can use the protein alignment generated by the Gene Model Checker or you can generate a new alignment using BLAST 2 Sequences (*bl2seq*). **Paste a screenshot of the protein alignment below:**



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 <u>OPTIONAL</u> 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):

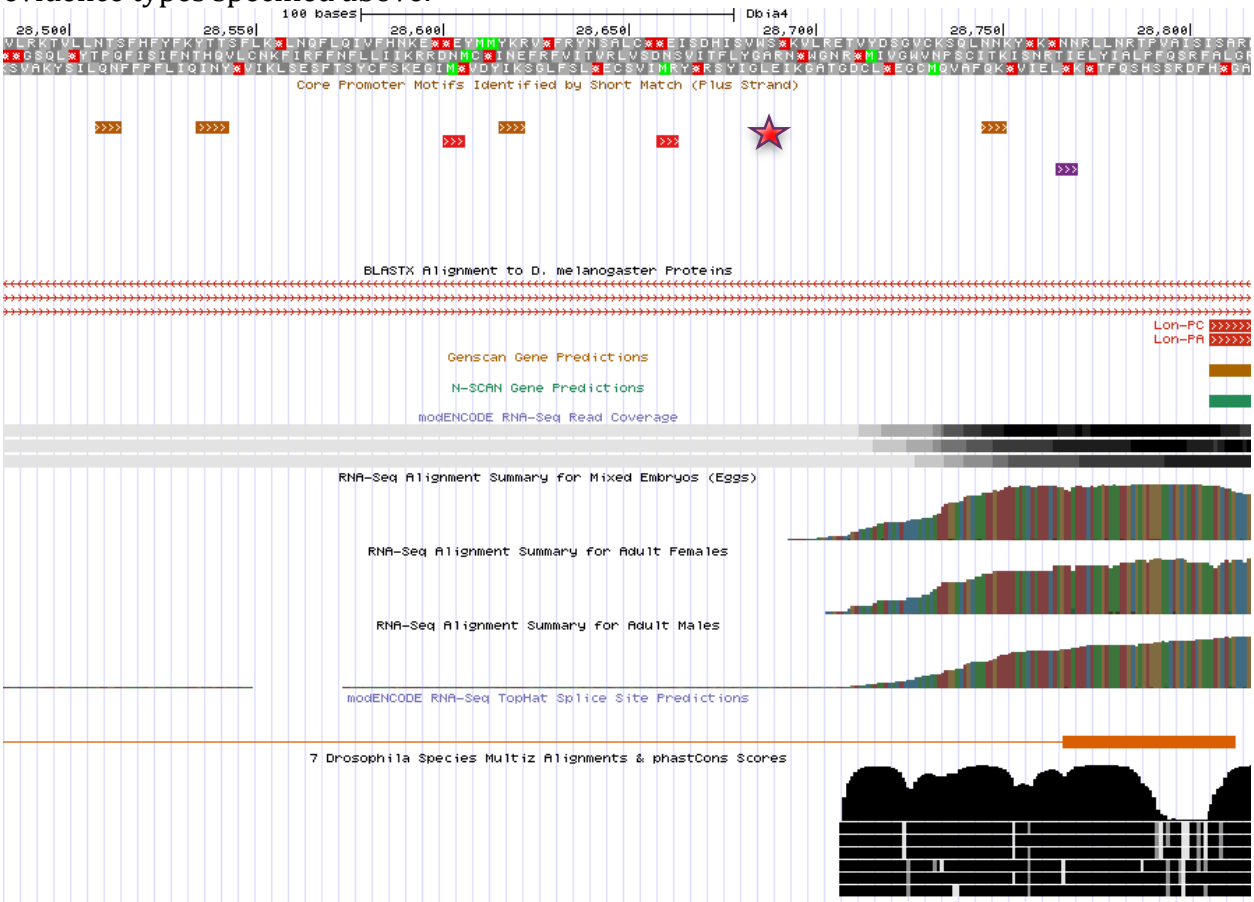
Gene-isoform name (i.e. dbia_ey-RA): dbia_Lon-PA
Names of the isoforms with the same TSS as this isoform: dbia_Lon-PC
Type of core promoter: (Peaked or Broad): Broad
Coordinates of the first transcribed exon: 28,683-29,150
Coordinate(s) of TSS position(s): 28,683
Coordinate(s) of TSS search region(s): 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 *D. melanogaster*	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: lcl|Query_39683 **Length:** 40000 **Number of Matches:** 11

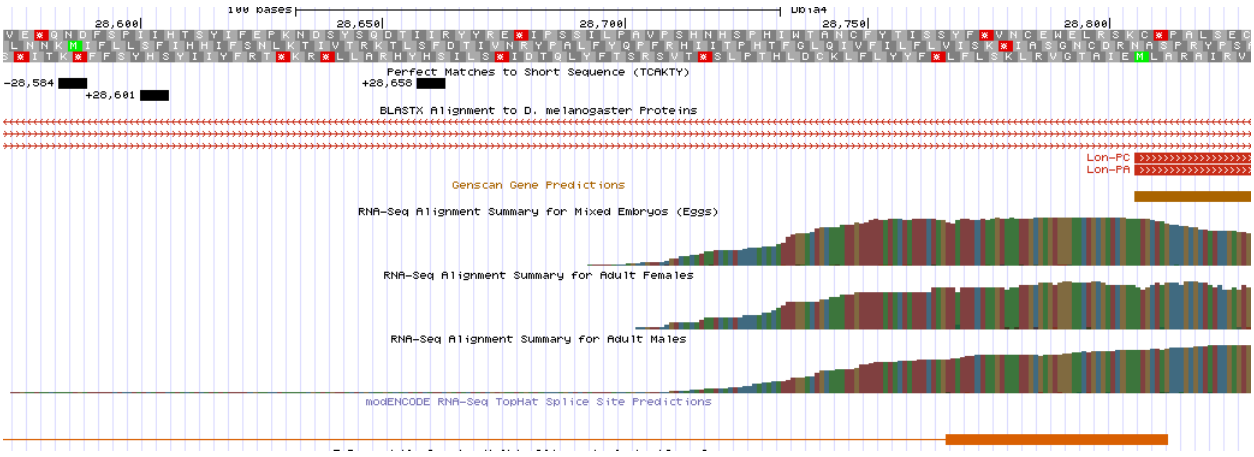
Range 1: 28693 to 29150 Graphics ▼ Next Match ▲ Prev

Score	Expect	Identities	Gaps	Strand
416 bits(290)	6e-119	377/458(82%)	11/458(2%)	Plus/Plus

```
Query  10     ATTTCACCAGCCGT-----CACATAATCACTCACCACACATTTGGACTGCAAATTGTTTT  64
              ||||  |||||||||     ||||||||||||| ||||||||||||||||||||||||||
Sbjct  28693  ATTTTACCAGCCGTTCCGTCACATAATCACTCCCCACACATTTGGACTGCAAATTGTTTT  28752

Query  65     TCTACCATTTCTAGTTATTTCTAAGTAAATTGCGAGTGGATATTGCTTTCAATATGTTAG  124
              | ||| ||||||||||||||||||||||||||||||||||  | |||  || | ||| |||
Sbjct  28753  TATACTATTTCTAGTTATTTCTAAGTAAATTGCGAGTGGGAACTGCGATCGAAATGCTAG  28812

Query  125    CCCGCGCTATCCGAGTGCGCCCCATGATGCGTGGCATCGCCTCGTCGTCAGTGTGGAACC  184
              |||||||||||||||||| || ||||||||  |||||||||||||||| ||||||| ||
Sbjct  28813  CCCGCGCTATCCGAGTGCGTCCTATGATGCGGAGCATCGCCTCGTCGTCGGTGTGGACCC  28872

Query  185    GGAATCGTCCCATTCAGAGTTCCCTGATGCAATACTGCCGGGATCGGTCGTTGCGCCTCC  244
              | || ||||||   |||||||||||| |||||| |||||||| |||| ||  || |||||
Sbjct  28873  GCAACCGTCCCGCCCAGAGTTCCCTGGTGCAATGCTGCCGGGTTCGGGCGACGCACCTCC  28932

Query  245    AGCGGCTCCACGGAGCCAATTTGATGGTGCAGCGCTTCTACAGCCGCAAGCGGGATGATT  304
              ||||  | || ||||| ||  |||||||| || || ||||||||||||||||||| || |
Sbjct  28933  AGCGATTTCATGGAGCAAACATGATGGTCCAACGTTTCTACAGCCGCAAGCGGGACGACT  28992

Query  305    CCAACGGGGATATTAT-TATGG-----GACCCGATCTTATGTCCGATCAAGATACCCATC  358
              || ||| |||| | ||  | ||        ||||| || ||||||||||  ||  ||||
Sbjct  28993  CCGACGAGGATCTCATGGACGGTCATAATCCCGAGCTAATGTCCGATCGGGAAGCCCAGT  29052

Query  359    TTCCGGCAACTGTGGCGGTGCCGGACGTGTGGCCACATGTTCCGTTGTTGGCCATGCGCA  418
              | ||||| ||||| |||||||||||| |||||||||||||| ||| ||||||||||||| |
Sbjct  29053  TGCCGGCCACTGTTGCGGTGCCGGATGTGTGGCCACATGTCCCGCTGTTGGCCATGCGAA  29112

Query  419    AGAATCCTCTCTTTCCCCGCTTTATGAAGATAGTGGAG  456
              ||||||| |||||||| ||||| |||||||| || |||
Sbjct  29113  AGAATCCCCTCTTTCCGCGCTTCATGAAGATTGTAGAG  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 (±2kb) 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  ccgtcacataatcactcctcacacatttggactgcaaattgttttttttaccatttctagatatttctaa
B D    D. takahashii  ccgtcacataatcactcctcacacatttggt-tgcaaattg-tttttataccatttctagttatttctaa

        D. biarmipes  gtaaattgcgagtgggaactgcgatcgaaatgctagcccgcgctatccgagtgcgt
     D. melanogaster  gtaaattgcgagtggatattgctttcaatatgttagcccgcgctatccgagtgcgc
           D. yakuba  gtaaattgcgagtggatattgctttcaatatgttagcccgcgctatccgagtgcgc
           D. erecta  gtaaattgcgagtggatactgctttcaatatgttagcccgcgctatccgagtgcgc
       D. eugracilis  gtaaattgcgattgggaattgtcatcaagatgctagcccgcgctatccgagtgcgt
       D. ficusphila  gtaaattgcgagtgggaattgtcatcgaaatgctagcccgcgctattcgagtgcgt
       D. takahashii  gtaaattgcgagtgggaattgccatcgagatgctagcccgcgctatccgagtgcgt
```

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

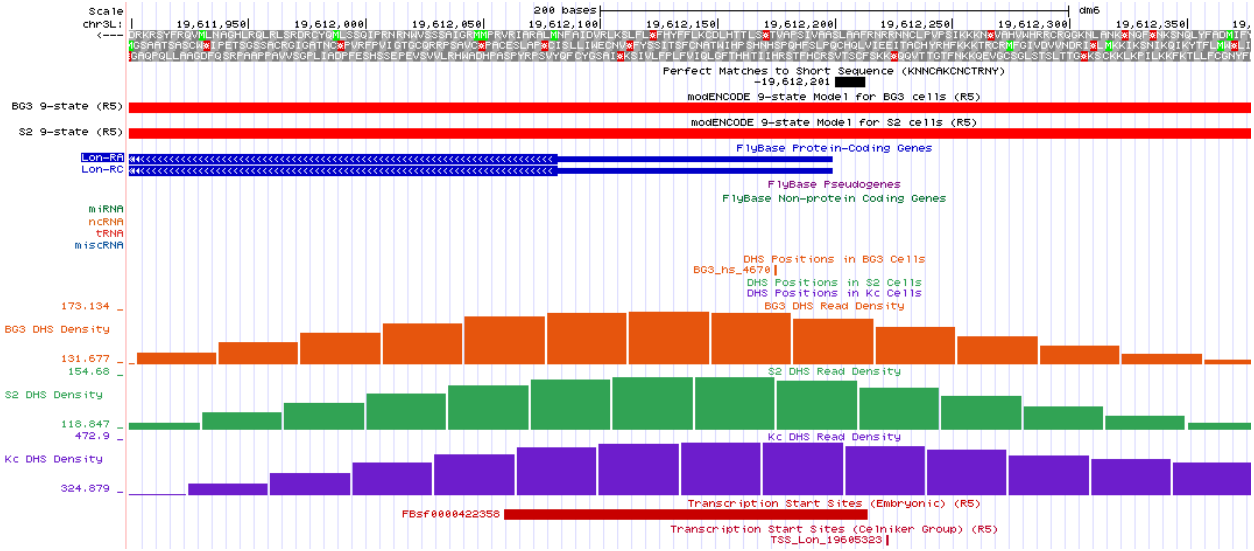
2. Search for core promoter motifs

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 (±300bp) 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	*D. melanogaster*
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

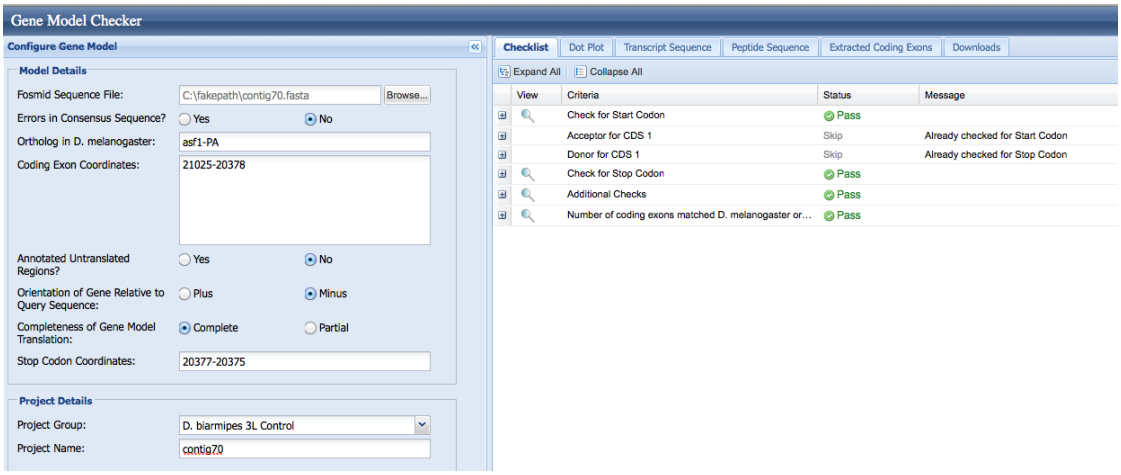
Complete this report form for each unique isoform listed in the table above (copy and paste to create as many copies of this Isoform Report Form as needed):

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:**

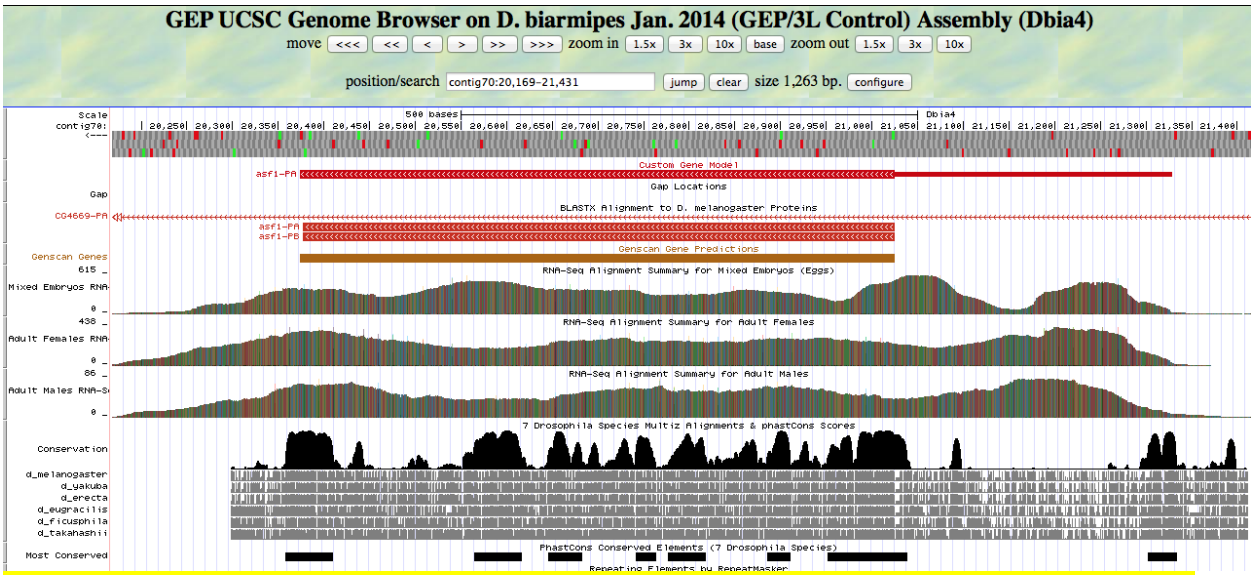
Note: For projects with consensus sequence errors, report the exon coordinates relative to the original project sequence. Include the VCF file you have generated above when you submit the gene model to the Gene Model Checker. The Gene Model Checker will revise the submitted exon coordinates automatically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

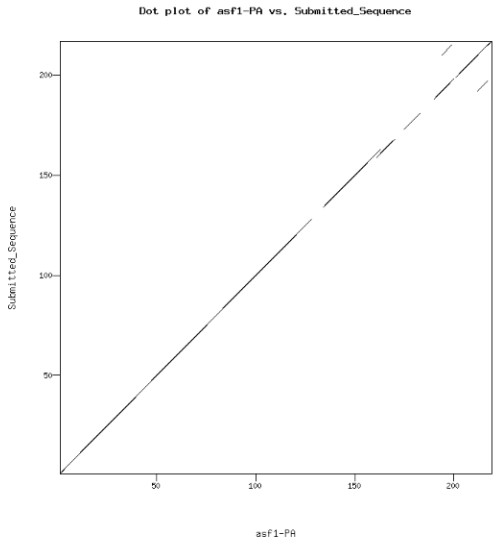
Identity: 206/218 (94.5%), **Similarity:** 210/218 (96.3%), **Gaps:** 2/218 (0.9%)

```
asf1-PA          1 MAKVHITNVVVLDNPSSFFNPFQFELTFECIEELKEDLEWKMIYVGSAESEEHDQVLDTI 60
                   ************************************************************
Submitted_Seq    1 MAKVHITNVVVLDNPSSFFNPFQFELTFECIEELKEDLEWKMIYVGSAESEEHDQVLDTI 60

asf1-PA         61 YVGPVPEGRHIFVFQADPPDVSKIPEPDAVGVTIVLLTCSYRGQEFVRVGYYVNNDYADP 120
                   ************************************************************
Submitted_Seq   61 YVGPVPEGRHIFVFQADPPDVSKIPEPDAVGVTIVLLTCSYRGQEFVRVGYYVNNDYADP 120

asf1-PA        121 EMRENPPTKPLFEKLTRNILASKPRVTRFKINWDYGHINGNGNGVENGHQDEMATDGPST 180
                   *******.****:*********************  ******:.**.******
Submitted_Seq  121 EMRENPPPKPLFDKLTRNILASKPRVTRFKINWDYGHINGN--GVENGHDEEMVTDGPST 178

asf1-PA        181 SEAASAVIHPEDDNSLAMPMENGIKALNENSNSLAMEC 218
                   **.**.*:.*********:******************
Submitted_Seq  179 SEVASVVVQPEDDNSLAMPIENGIKALNENSNSLAMEC 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 <u>vertical and horizontal gap</u> 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 <u>OPTIONAL</u> 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

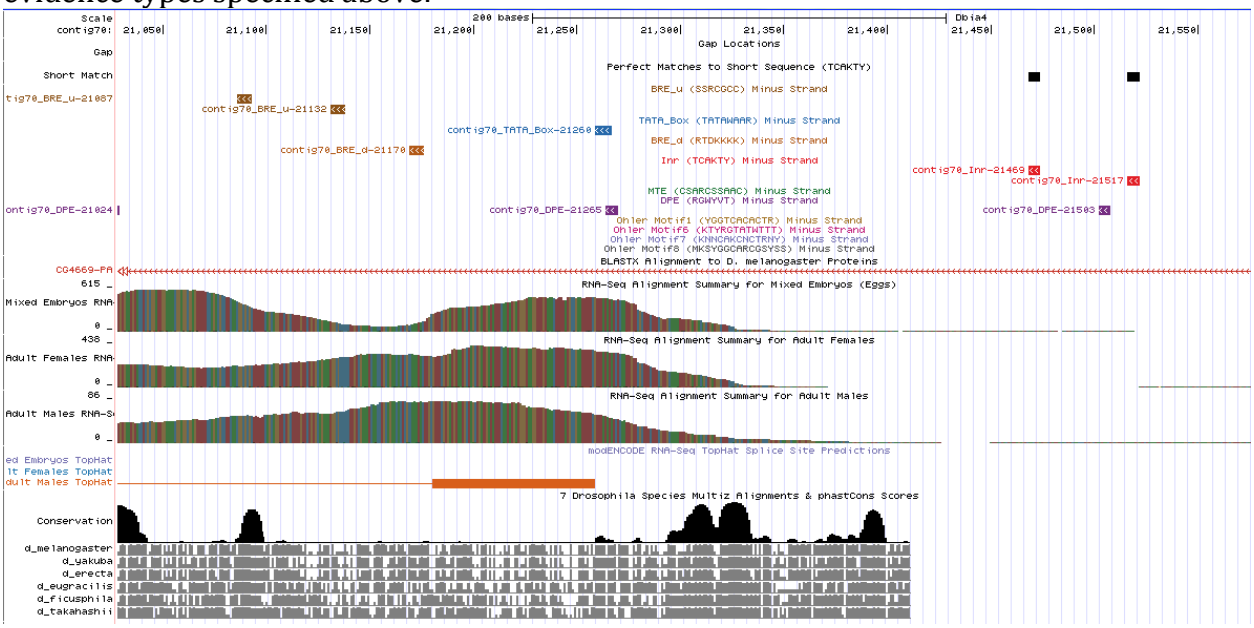
Type of core promoter: (Peaked or Broad): Peaked
Coordinates of the first transcribed exon: 21,329-19,824(asf1-PA); 21,329-21,200(asf1-PB)
Coordinate(s) of TSS position(s): 21,329
Coordinate(s) of TSS search region(s): 21,200-21,500

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 *D. melanogaster*	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:



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

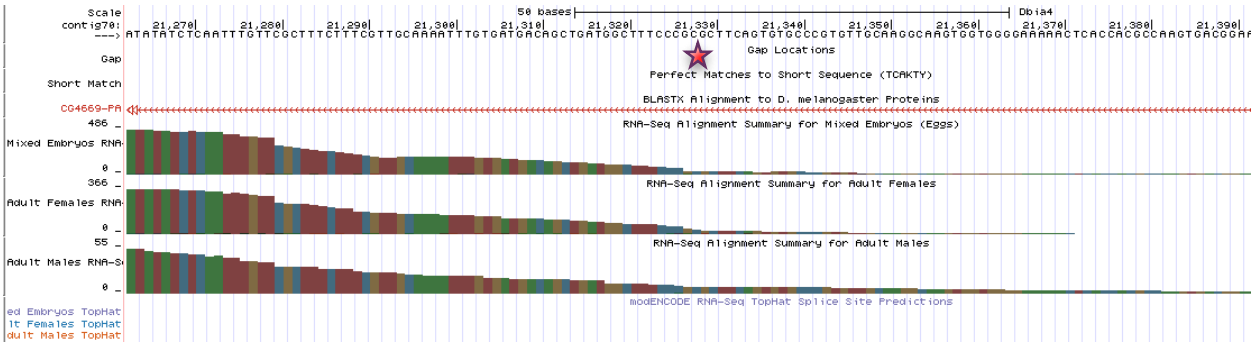
Score	Expect	Identities	Gaps	Strand
900 bits(629)	0.0	929/1191(78%)	60/1191(5%)	Plus/Minus

```
Query  1      CGCGGGAAAGAAGTCAGCTGTCACCAT---TTGTGCGAAGACA---AGCGGAATAATTGC  54
              ||||||||||  ||||||||||  ||    || ||| | || |   ||||  |  |||||
Sbjct  21329  CGCGGGAAAGCCATCAGCTGTCATCACAAATTTTGCAACGAAAGAAAGCGAACAAATTGA  21270

Query  55     GATAAATAAAAGCAGAATAGCGACAGACAAT-TGCACAAACCTGATTCATGTTGTAACCA  113
              ||||  |||||  |  | |||  || ||  | ||  |  |||||||| |  |||  | |    |
Sbjct  21269  GATATATAAACGAAAAATT-CGGCATAAAAAGTTTACAAACCTTACCCATCATATTGAAA  21211

Query  114    AAACAAACTACTAGCATTACAGTACATATATCAAGTATTTGCTGAGCGTTTTGCAGGTGA  173
                | ||| |  ||||||||||  ||  |   ||||||||||  ||| ||| |||||||
Sbjct  21210  TATCAATCAGCTAGCATTACATCACGAAACGCAAGTATTTGGCGAGTGTTCTGCAGGTTG  21151

Query  174    ATTTCGCAAAG------CAATTCCCTAGTCGGACAAAAGCAGAACCT-CTAAAAAAAGCG  226
              ||||||  |||      | ||| |      | ||||||||||||||| || ||||||||||
Sbjct  21150  ATTTCGACAAGGGGGCGCCATTTCTGCACAGAACAAAAGCAGAACCTGCTGAAAAAAGCG  21091

Query  227    CGCAGTCGAAGGAGTCTTTAAAATATCGGTCTACGCTCTGCAATCTCCGAGGTACA----  282
              ||| || ||||| ||||  ||||  || | || || | | |||||  |||||||||
Sbjct  21090  CGCCGTTGAAGGCGTCTGGAAAAACTCCGCCTCCGTTTTCCAATCCGCGAGGTACATTCA  21031

Query  283    --GTCATGGCCAAGGTGCACATCACCAACGTGGTGGTGCTGGACAACCCGAGCAGCTTCT  340
                |||||||||||||||||||||||||| ||||||||||||||||||| ||||||| |
Sbjct  21030  CAGTCATGGCCAAGGTGCACATCACCAATGTGGTGGTGCTGGACAACCCCAGCAGCTTTT  20971

Query  341    TCAACCCCTTTCAGTTCGAACTCACGTTCGAGTGCATTGAGGAGCTAAAAGAGGATCTAG  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.

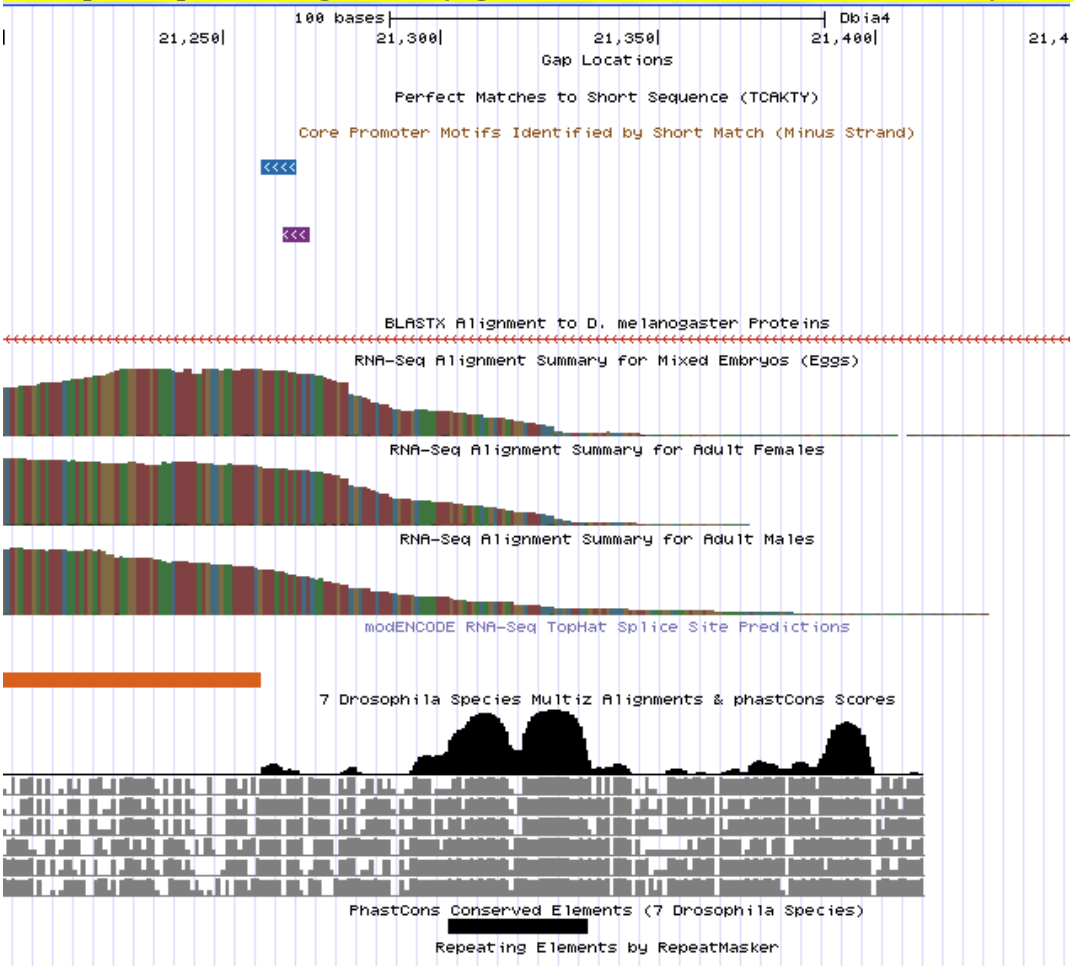
If the TSS annotation is supported by RNA-Seq or RNA polymerase II data, **paste a Genome Browser screenshot of the region around the TSS (±2kb) with the evidence tracks listed below:**

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:**



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 (±300bp) 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	*D. melanogaster*
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

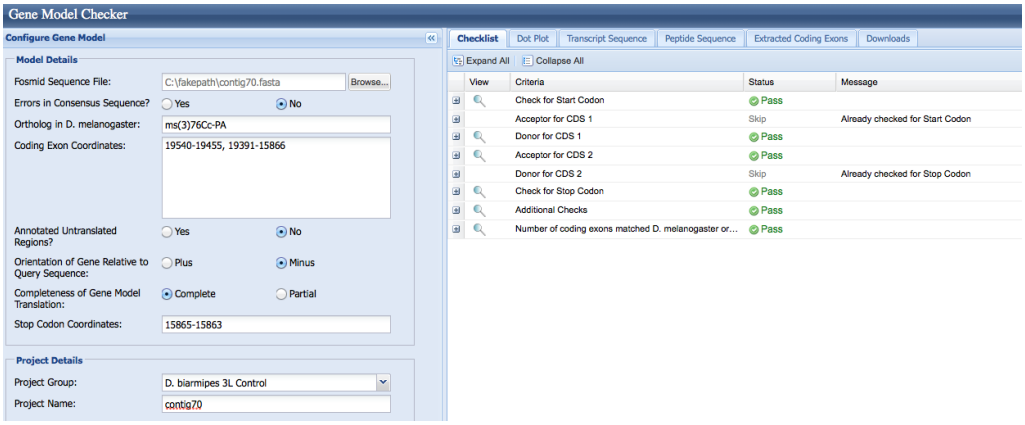
Complete this report form for each unique isoform listed in the table above (copy and paste to create as many copies of this Isoform Report Form as needed):

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.

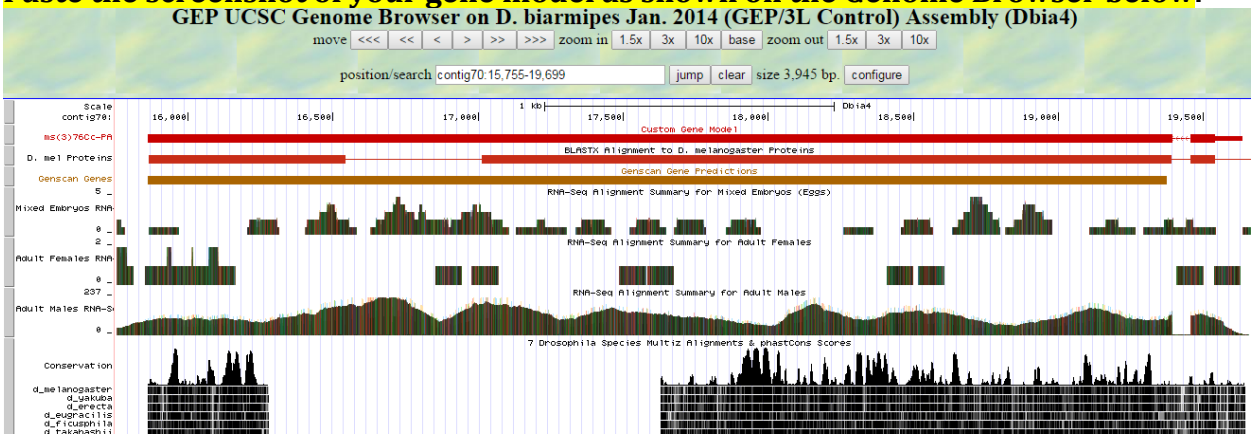


2. View the gene model on the Genome Browser

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.

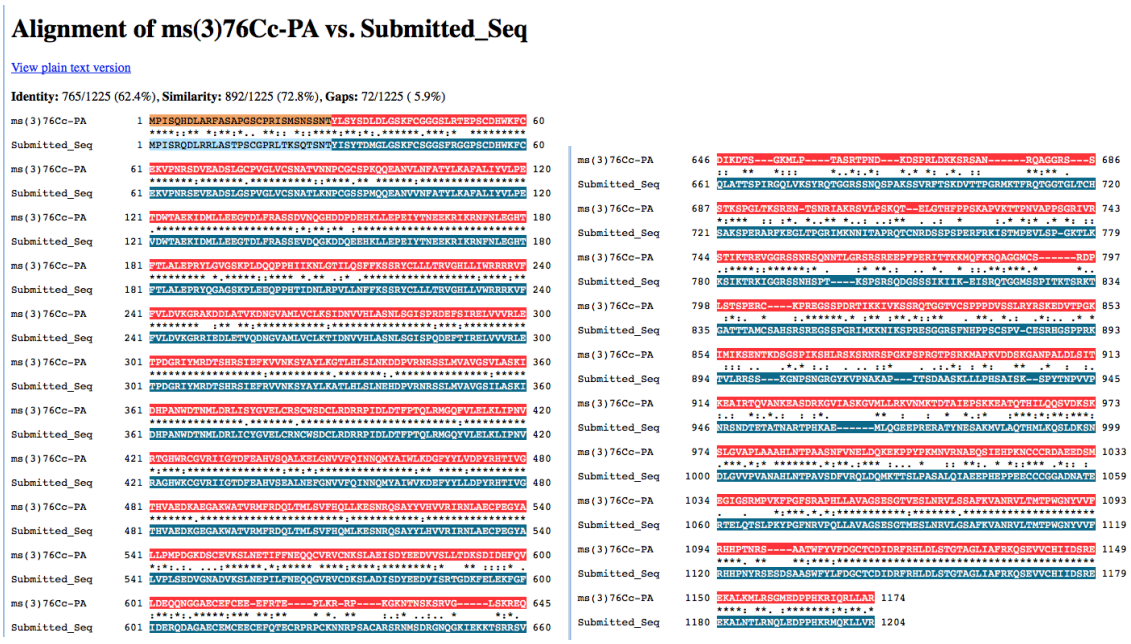
25. A sequence alignment track (D. mel Protein or Other RefSeq)
26. At least one gene prediction track (e.g. Genscan)
27. At least one RNA-Seq track (e.g. RNA-Seq Alignment Summary)
28. A comparative genomics track
    (e.g. Conservation, D. mel. Net Alignment, 3-way, 5-way or 7-way multiz)

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*

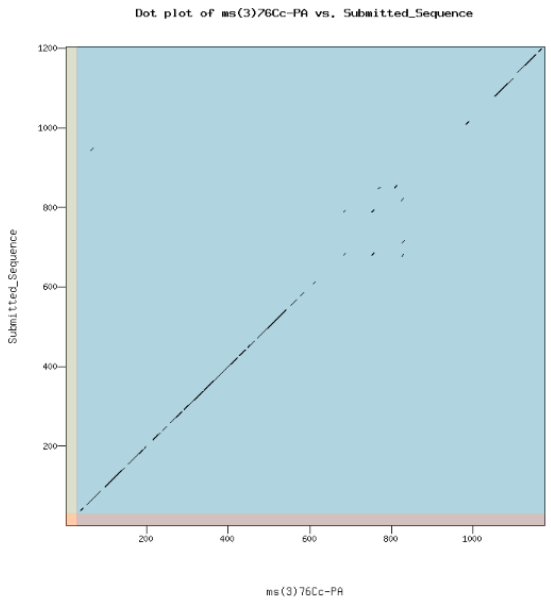
Show an alignment between the protein sequence for your gene model and the protein sequence from the putative *D. melanogaster* ortholog. You can use the protein alignment generated by the Gene Model Checker or you can generate a new alignment using BLAST 2 Sequences (*bl2seq*). **Paste a screenshot of the protein alignment below:**



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 <u>OPTIONAL</u> 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

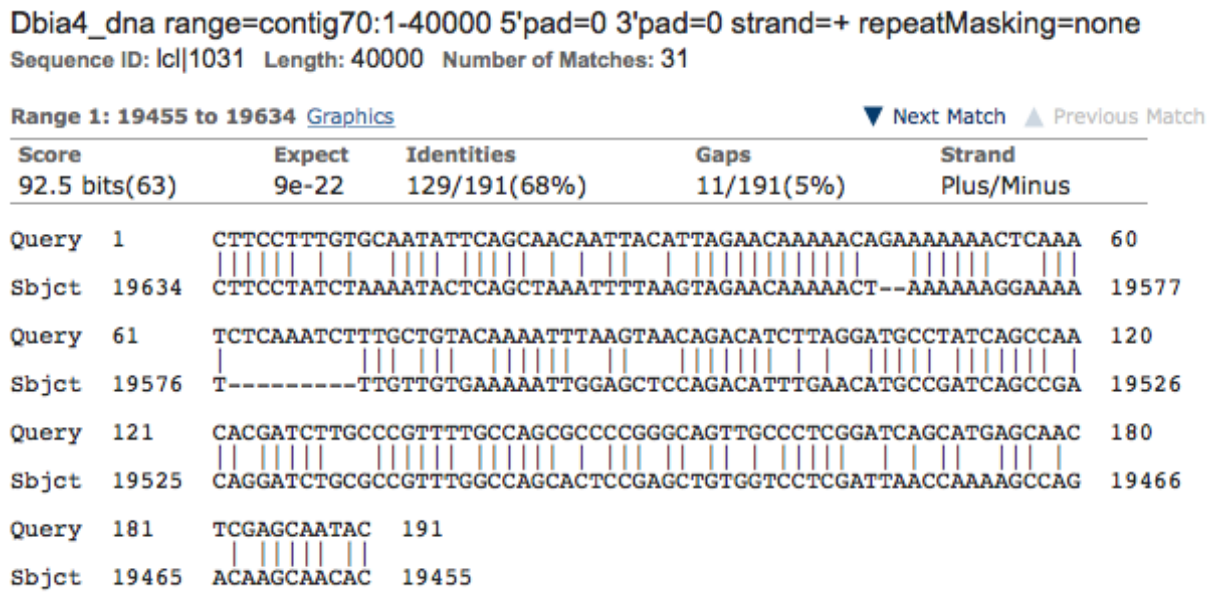
1. Evidence that supports the TSS annotation postulated above

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 *D. melanogaster*	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

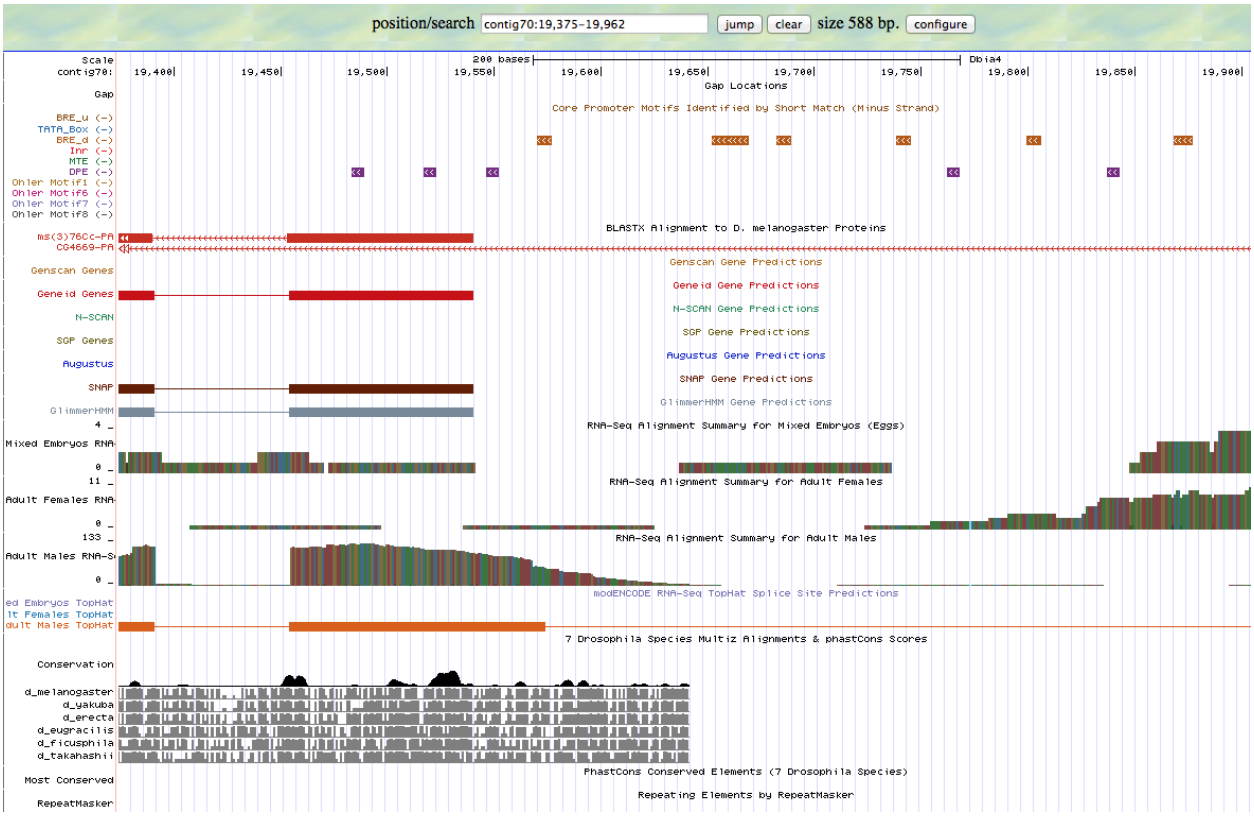
Range 1: 19455 to 19634 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
92.5 bits(63)	9e-22	129/191(68%)	11/191(5%)	Plus/Minus

```
Query  1      CTTCCTTTGTGCAATATTCAGCAACAATTACATTAGAACAAAAACAGAAAAAAACTCAAA  60
              |||||| | |  |||| ||||| | | ||  | ||||||||||||   ||||||   |||
Sbjct  19634  CTTCCTATCTAAAATACTCAGCTAAATTTTAAGTAGAACAAAAACT--AAAAAAGGAAAA  19577

Query  61     TCTCAAATCTTTGCTGTACAAAATTTAAGTAACAGACATCTTAGGATGCCTATCAGCCAA  120
              |        ||| |||  |||||  ||   |||||||  |  |  |||||  |||||||  |
Sbjct  19576  T---------TTGTTGTGAAAAATTGGAGCTCCAGACATTTGAACATGCCGATCAGCCGA  19526

Query  121    CACGATCTTGCCCGTTTTGCCAGCGCCCCGGGCAGTTGCCCTCGGATCAGCATGAGCAAC  180
              ||  |||||   ||||||  ||||||  | |||  ||  ||  | |||||   |  |  ||   |||  |
Sbjct  19525  CAGGATCTGCGCCGTTTGGCCAGCACTCCGAGCTGTGGTCCTCGATTAACCAAAAGCCAG  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 (±2kb) 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  cagtatagctgatgtacctatgaataatcgggaaaaaggattattatttgaatacggcccagtttgggag
B D   D. melanogaster  ccgaatagctgagatacctaaggacgatcaggaggatggatcagt-ttaagcatttcacaaattctggag
B D         D. yakuba  cagaatagctgagatacctaaggacgatcaggagaacggatcagt-aaaagcattagccaaattctggag
B D         D. erecta  ccgaatagctgagatacctaaggacgatcaggaggacggttcagt-ttaaacattagctaaattccggag
B D     D. eugracilis  cagtgtagctatgatatctaggggcaatcaggaacatggatcagttttcgagaaccgcccaatttcagag
B D     D. ficusphila  caacgtagctaatgtatctaagggtaaacagga-cacagataagcactcaaaaaccgctgcaattgggag
B D     D. takahashii  ccgtatagctgatatacctgaggaca------aggaaggatcagttttcgacaatcaccgcattgtgcag

         D. biarmipes  ctctccccacgtgttgcttgtctggcttttggttaatcgaggaccacagctcggagtgctggccaaacgg
      D. melanogaster  tactctccacgtattgctcgagttgctcatgctgatccgagggcaactgcccggggcgctggcaaaacgg
            D. yakuba  tactctccacgtattgctcgaattgctcttgttaatgcg------acagctcggggcgctgaccaaacgg
            D. erecta  tactctccacgtattgctcgagttgcccttagtactgcgaggacaacagctcggggcgctggccaaacgg
        D. eugracilis  taatacccacgtgttgctcgtgtgactcttcgagaacacaggacaacaacttggcgtgctggccaaacgg
        D. ficusphila  cacatctcacgtgttgctagatttgctattgctgaaccgaggacagcagctttgcatgctggccatacgg
        D. takahashii  cgccacccacgtattgctcgaccggcttttggcaactcgaggataacagctcggcgtgctggccaagcgg

         D. biarmipes  cgcagatcctgtcggctgatcggcatgttcaaatgtctggagctccaatttttcacaacaaatttt----
      D. melanogaster  gcaagatcgtgttggctgataggcatcctaagatgtctgttacttaaattttgtacagcaaagatttgag
            D. yakuba  ccaaaatcctgttggctgatcggcatcctcagatgtctgttacttaaattttgtacagcaaaaattt---
            D. erecta  ccaagatcctgttggctgatcgacatcctcagaaatctgttacttttaattttgtacagcaaaaattt---
        D. eugracilis  tgcggttcctgttgactgatcggcatcctgggatgtcttaaaatcgaatttagtacaaaactaatt----
        D. ficusphila  cacagatcctgttggctgatcggcatcct-ggctgtctagaattttaa-tttgcgtaaaaaaaattc---
        D. takahashii  cgcagatcctgttggctgattgccatcctcggatgtctaaaactctaatttgacacaaacaaaattt---

         D. biarmipes  ------c--ctttttttagtttttgttctacttaaaatttagctgagtattttagataggaagaattgat
      D. melanogaster  atttgag-ttttttttctgtttttgttctaatgtaattgttgctgaatattgcacaaaggaagaattgat
            D. yakuba  -tccgagtttttttttcggtttttgttctactggaa-ttttgctgaaaattgcagaaaggaagtattgat
            D. erecta  -tccgag--ttttttttagtttttgttctactggaatttttgctgaaaattgcagaaaggaagaattgat
        D. eugracilis  -ttccac--ttttttcagttttcgttttactagaattttaaatgaatattccagaaaggatgaactgat
        D. ficusphila  -ttacac-----ttttttgttttcgttttactaaaattttagctgaatattgcagaaaggaagatttgat
        D. takahashii  -ttacac--ctttttccgttttccttctacttggaatttagttgagtattgcagaagggaaaatttgat
```

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 (±300bp) 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	*D. melanogaster*
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

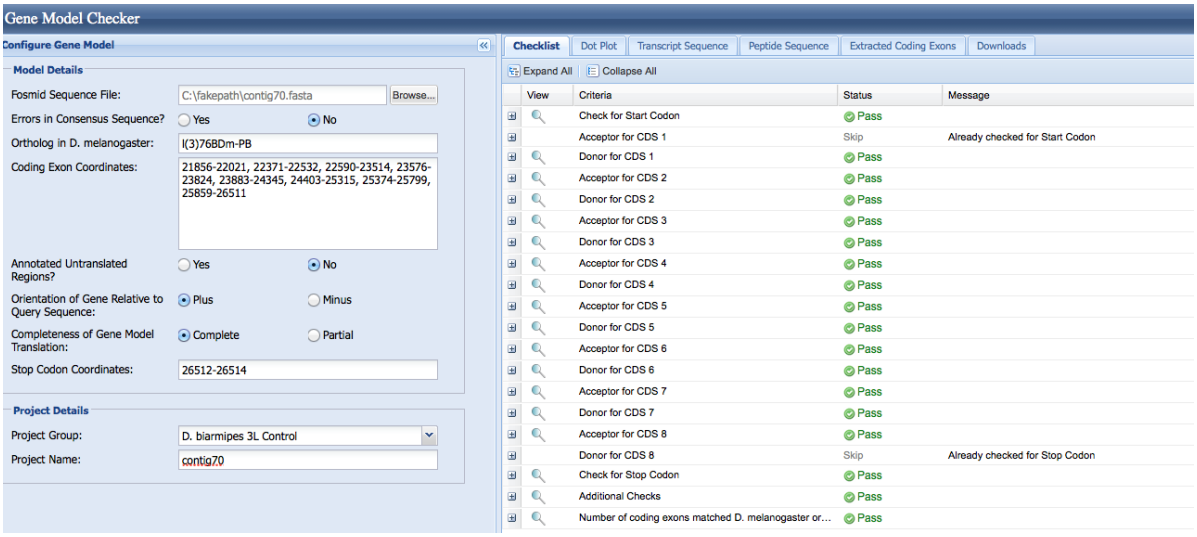
Complete this report form for each unique isoform listed in the table above (copy and paste to create as many copies of this Isoform Report Form as needed):

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.

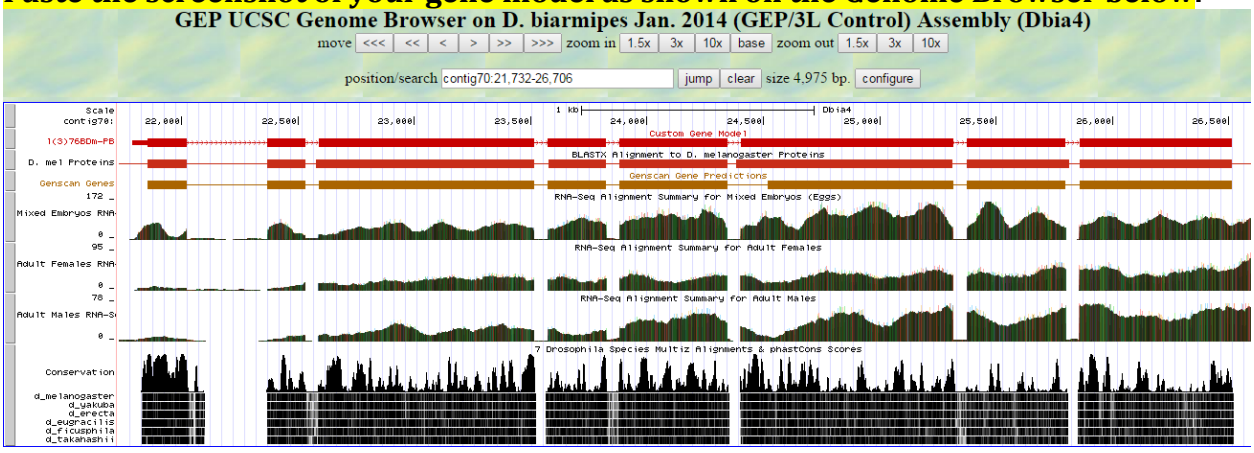


2. View the gene model on the Genome Browser

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)

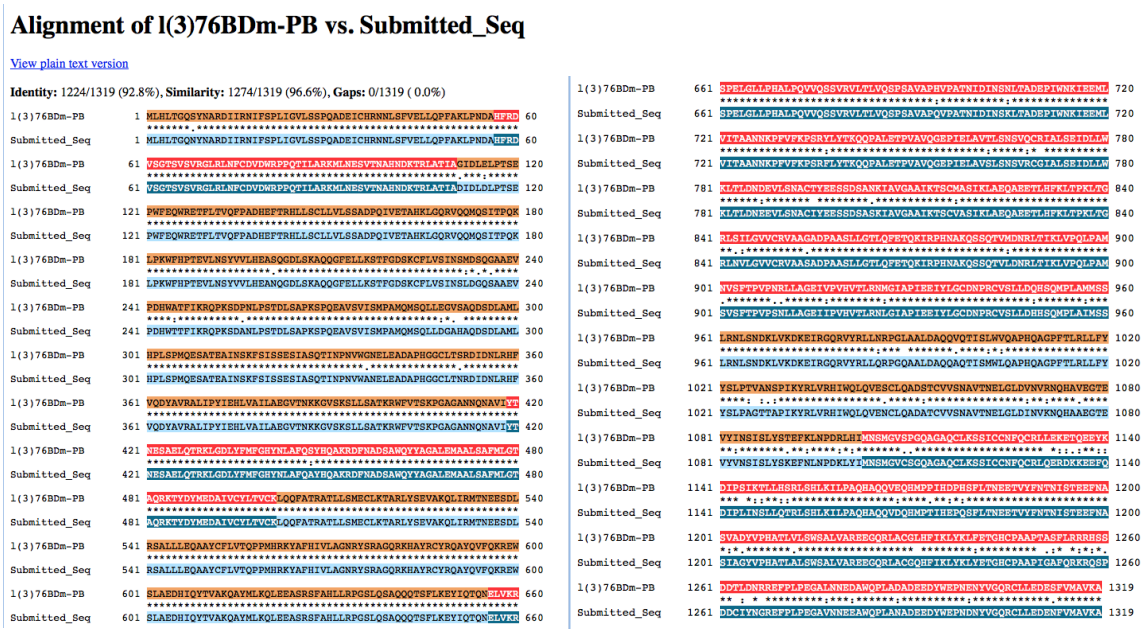
Paste the screenshot of your gene model as shown on the Genome Browser below:



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 l(3)76BDm-PB vs. Submitted_Seq

View plain text version

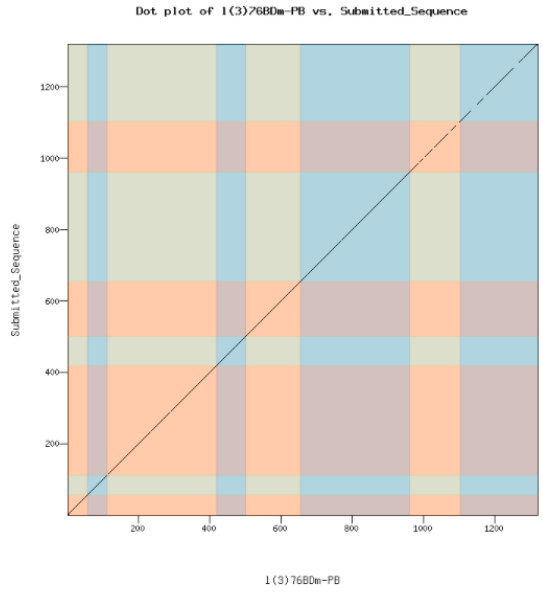
Identity: 1224/1319 (92.8%), **Similarity:** 1274/1319 (96.6%), **Gaps:** 0/1319 (0.0%)

```
l(3)76BDm-PB        1 MLHLTGQSYNARDIIRNIFSPLIGVLSSPQADEICHRNNLSFVELLQPFAKLPNDAHFRD 60
Submitted_Seq       1 MLHLTGQNYNARDIIRNIFSPLIGVLSSPQADEICHRNNLSFVELLQPFAKLPNDAHFRD 60
l(3)76BDm-PB       61 VSGTSVSVRGLRLNFCDVDWRPPQYTLARKMLNESVTNAHNDKTRLAGTAGIDLELPTSE 120
Submitted_Seq      61 VSGTSVSVRGLRLNFCDVDWRPPQYTLARKMLNESVTNAHNDKTRLAGTADIDLDLPTSE 120
l(3)76BDm-PB      121 PWFEQWRETFLTVQFPADHEFTRHLLSCLLVLSSADPQIVETAHKLGQRVQQMQSITPQK 180
Submitted_Seq     121 PWFEQWRETFLTVQFPADHEFTRHLLSCLLVLSSADPQIVETAHKLGQRVQQMQSITPQK 180
l(3)76BDm-PB      181 LPKWFHPTEVLNSYVVLHEASQGDLSKAQQGFELLKSTFGDSKCFLVSINSMDSQGAAEV 240
Submitted_Seq     181 LPKWFHPTEVLNSYVVLHEANQGDLSKAQQGFELLKSTFGDSKCFLVSINSLDQQSAAEV 240
l(3)76BDm-PB      241 PDHWATTFIKRQPKSDPNLPSTDLSAPKSPQEAVSVISMPAMQMSQLLEGVSAQDSDLAML 300
Submitted_Seq     241 PDHWTTFIKRQPKSDANLPSTDLSAPKSPQEAVSVISMPAMQMSQLLDGAHAQDSDLAML 300
l(3)76BDm-PB      301 HPLSPMQESATEAINSKFSISSESIASQTINPNVWGNELEADAPHGGCLTSRDIDNLRHF 360
Submitted_Seq     301 HPLSPMQESATEAINSKFSISSESIASQTINPNVWANELEADAPHGGCLTNRDIDNLRHF 360
l(3)76BDm-PB      361 VQDYAVRALIPYIEHLVAILAEGVTNKKGVSKSLLSATKRWFVTSKPGAGANNQNAVLYT 420
Submitted_Seq     361 VQDYAVRALIPYIEHLVAILAEGVTNKKGVSKSLLSATKRWFVTSKPGACANNQNAVIYT 420
l(3)76BDm-PB      421 NESAELQTRKLGDLYFMFGHYNLAFQSYHQAKRDFNADSAWQYYAGALEMAALSAFMLGT 480
Submitted_Seq     421 NESAELQTRKLGDLYFMFGHYNLAFQAYHQAKRDFNADSAWQYYAGALEMAALSAFMLGT 480
l(3)76BDm-PB      481 AQRKTYDYMEDAIVCYLTVCLLQQFATRATLLSMECLKTARLYSEVAKQLIRMTNEESDL 540
Submitted_Seq     481 AQRKTYDYMEDAIVCYLTVCLLQQFATRATLLSMECLKTARLYSEVAKQLIRMTNEESDL 540
l(3)76BDm-PB      541 RSALLLEQAAYCFLVTQPPMHRKYAFHIVLAGNRYSRAGQRKHAYRCYRQAYQVFQKREW 600
Submitted_Seq     541 RSALLLEQAAYCFLVTQPPMHRKYAFHIVLAGNRYSRAGQRKHAYRCYRQAYQVFQKREW 600
l(3)76BDm-PB      601 SLAEDHIQYTVAKQAYMLKQLEEASRSFAHLLRPGSLQSAQQQTSFLKEYIQTQNELVKR 660
Submitted_Seq     601 SLAEDHIQYTVAKQAYMLKQLEEASRSFAHLLRPGSLQSAQQQTSFLKEYIQTQNELVKR 660

l(3)76BDm-PB      661 SPELGLLPHALPQVVQSSVRVLTLVQSPSAVAPHVPATNIDINSNLTADEPIWNKIEEML 720
Submitted_Seq     661 SPELGLLPHALPQVVQSSVRVLTLVQSPSAVAPQVPATNIDINSKLTADEPIWNKIEEML 720
l(3)76BDm-PB      721 VTTAANNKPFVFKPSRYLYTKQQPALETPVAVQGEPIELAVTLSNSVQCRIALSEIDLLW 780
Submitted_Seq     721 VTTAANNKPFVFKPSRFLYTKQQPALETPVAVQGEPIELAVSLSNSVRCGIALSEIDLLW 780
l(3)76BDm-PB      781 KLTLDNDEVLSNACTYEESSDSANKIAVGAAIKTSCMASIKLAEQAEETLHFKLTPKLTC 840
Submitted_Seq     781 KLTLDNEEVLSNACTYEESSDSASKIAVGAAIKTSCVASIKLAEQAEETLHFKLTPKLTC 840
l(3)76BDm-PB      841 RLSILGVVCRVAAGADPAASLLGTLQFETQKIRPHNAKQSSQTVMDNRLTIKLVPQLPAM 900
Submitted_Seq     841 RLNVLGVVCRVAASADPAASLLGTLQFETQKIRPHNAKQSSQTVLDNRLTIKLVPQLPAM 900
l(3)76BDm-PB      901 NVSFTPVPNRLLAGEIVPVHVTLRNMGIAPIEEIYLGCDNPRCVSLLDQHSQMPLAMMSS 960
Submitted_Seq     901 SVSFTPVPSNLLAGEIIPVHVTLRNLGIAPIEEIYLGCDNPRCVSLLDHHSQMPLAIMSS 960
l(3)76BDm-PB      961 LRNLSNDKLVKDKEIRGQRVYRLLNRPGLAALDAQQVQTISLWVQAPHQAGPFTLRLLFY 1020
Submitted_Seq     961 LRNLSNDKLVKDKEIRGQRVYRLLQRPGQAALDAQQAQTISMWLQAPHQAGPFTLRLLFY 1020
l(3)76BDm-PB     1021 YSLPTVANSPIKYRLVRHIWQLQVESCLQADSTCVVSNAVTNELGLDVNVRNQHAVEGTE 1080
Submitted_Seq    1021 YSLPAGTTAPIKYRLVRHIWQLQVENCLQADATCVVSNAVTNELGLDINVKNQHAAEGTE 1080
l(3)76BDm-PB     1081 VYINSISLYSTEFKLNPDRLHIMNSMGVSPGQAGAQCLKSSICCNPQCRLLEKETQEEYK 1140
Submitted_Seq    1081 VYVNSISLYSKEFNLNPDKLYIMNSMGVCSGQAGAQCLKSSICCNPQCRLQERDKKEEFQ 1140
l(3)76BDm-PB     1141 DIPSIKTLLHSRLSHLKILPAQHAQQVEQHMPPIHDPHSPLTNEETVYFNTNISTEEFNA 1200
Submitted_Seq    1141 DIPLINSLLQTRLSHLKILPAQHAQQVDQHMPTIHEPQSPLTNEETVYFNTNISTEEFNA 1200
l(3)76BDm-PB     1201 SVADYVPHATLVLSWSALVAREEGQRLACGLHFIKLYKLFETGHCPAAPTASFLRRRHSS 1260
Submitted_Seq    1201 STAGYVPHATLALSWSALVAREEGQRLACGQHFIKLYKLYETGHCPAAPIGAFQRKRQSP 1260
l(3)76BDm-PB     1261 DDTLDNRREFPLPEGALNNEDAWQPLADADEEDYWEPNENYVGQRCLLEDESFVMAVKA 1319
Submitted_Seq    1261 DDCIYNGREFPLPEGAVNNEEAWQPLANADEEDYWEPNDNYVGQRCLLEDENFVMAVKA 1319
```

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 <u>OPTIONAL</u> 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):

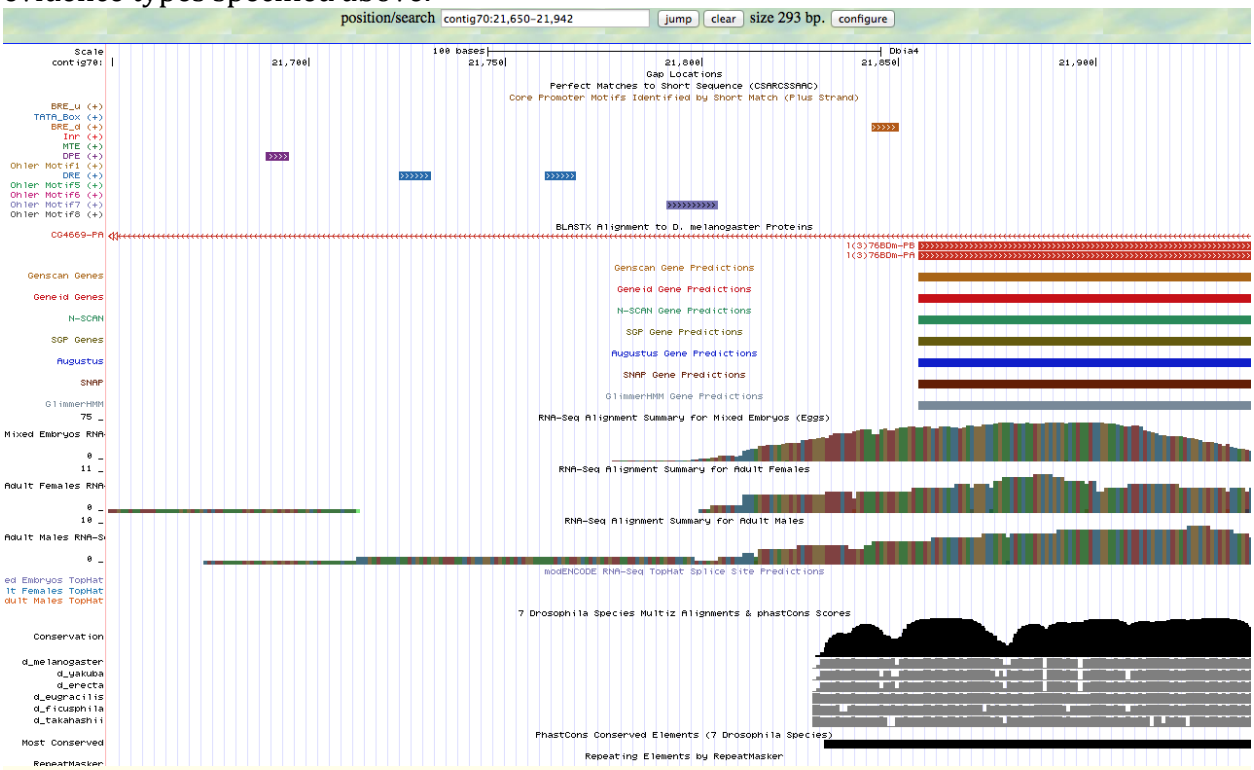
Gene-isoform name (i.e. dbia_ey-RA): dbia_l(3)76BDm-PB
Names of the isoforms with the same TSS as this isoform: dbia_l(3)76BDm-PA
Type of core promoter: (Peaked or Broad): Broad
Coordinates of the first transcribed exon: 21,790-22,021
Coordinate(s) of TSS position(s): 21,785
Coordinate(s) of TSS search 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 *D. melanogaster*	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  ggttttctttgtaaatagttgaagtaatgctgcacctgacgggccagaactacaatgcgcgcgacataat
B D  D. melanogaster  tattttctttgtaaatagttggagtaatgctgcacctgacgggccagagttacaatgctcgcgacatcat
B D         D. yakuba  tattttctttgtaaataattagagtaatgctgcacctgacgggccaaagctacaatgcccgcgacatcat
B D         D. erecta  tattttctttgtaaataattggagtaatgctgcacctgacgggccagagctacaatgcccgcgacatcat
B D     D. eugracilis  ggttttctttgtaaatagttgaagtaatgctgcacctgacgggccagaactacaatgcgcgcgacataat
B D     D. ficusphila  ggttttccctgtaaatagttgaagtaatgctgcacctgacgggtcagaactacaatgcgcgcgatataat
B D     D. takahashii  ggttttctttgtaaatagta-aagtaatgctgcacctgacgggccagagctacaatgcgcgcgacataat

         D. biarmipes  ccggaacatcttctcgccgctgatcggcgtgctgtccagtccg
     D. melanogaster  ccggaacatcttctcgccgctgatcggcgtgctgtccagtcca
         D. yakuba  ccggaacatattctcgccgctgatcggcgtgctgtccagtcca
         D. erecta  ccggaacatcttctcgccgctgatcggcgtgctgtccagtccg
       D. eugracilis  ccggaacatcttctcgccgctgatcggcgtgctgtccagtcct
       D. ficusphila  ccggaacattttctcgccgctgatcggcgtgctgtccagtccg
       D. takahashii  ccggaacatcttctccccactgataggcgtgctgtccagtccg
```

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

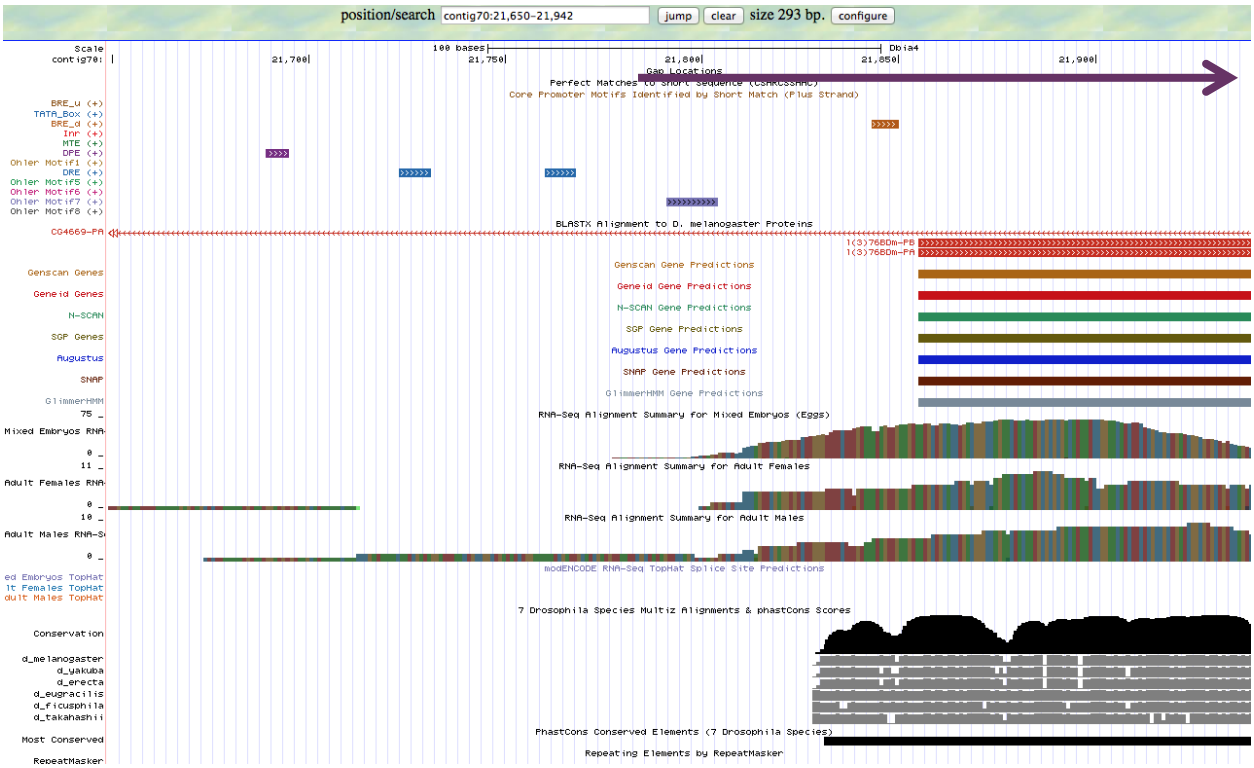
Score	Expect	Identities	Gaps	Strand
255 bits(177)	1e-70	209/237(88%)	5/237(2%)	Plus/Plus

```
Query  6      AGTGCCATCTCCAGCTGCTCCCTGCAATTGTTGTGTGGTGTTCTGTATTTTCTTTGTAAA  65
              ||| ||| ||| ||   ||||| || ||    | |||||  ||  ||||||||||||
Sbjct  21790  AGTTCCAGCTCTAGTGTCTCCCC-CAGTTC----GAGGTGTCTTGGGTTTTCTTTGTAAA  21844

Query  66     TAGTTGGAGTAATGCTGCACCTGACGGGCCAGAGTTACAATGCTCGCGACATCATCCGGA  125
              ||||||  ||||||||||||||||||||||||   |||||||  ||||||||  |||||||
Sbjct  21845  TAGTTGAAGTAATGCTGCACCTGACGGGCCAGAACTACAATGCGCGCGACATAATCCGGA  21904

Query  126    ACATCTTCTCGCCGCTGATCGGCGTGCTGTCCAGTCCACAGGCGGACGAGATTTGCCACC  185
              ||||||||||||||||||||||||||||||||||||||| |||||||||||||  |||||||
Sbjct  21905  ACATCTTCTCGCCGCTGATCGGCGTGCTGTCCAGTCCGCAGGCGGACGAGATCTGCCACC  21964

Query  186    GGAACAACCTCTCCTTCGTGGAACTCCTGCAGCCGTTCGCAAAGTTGCCAAATGATG  242
              | |||||||||||||||||||  ||||||||||||||||||||||||||||||||||
Sbjct  21965  GCAACAACCTCTCCTTCGTGGAGCTCCTGCAGCCGTTCGCAAAGTTGCCAAATGATG  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 (±2kb) 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 (±300bp) 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	*D. melanogaster*

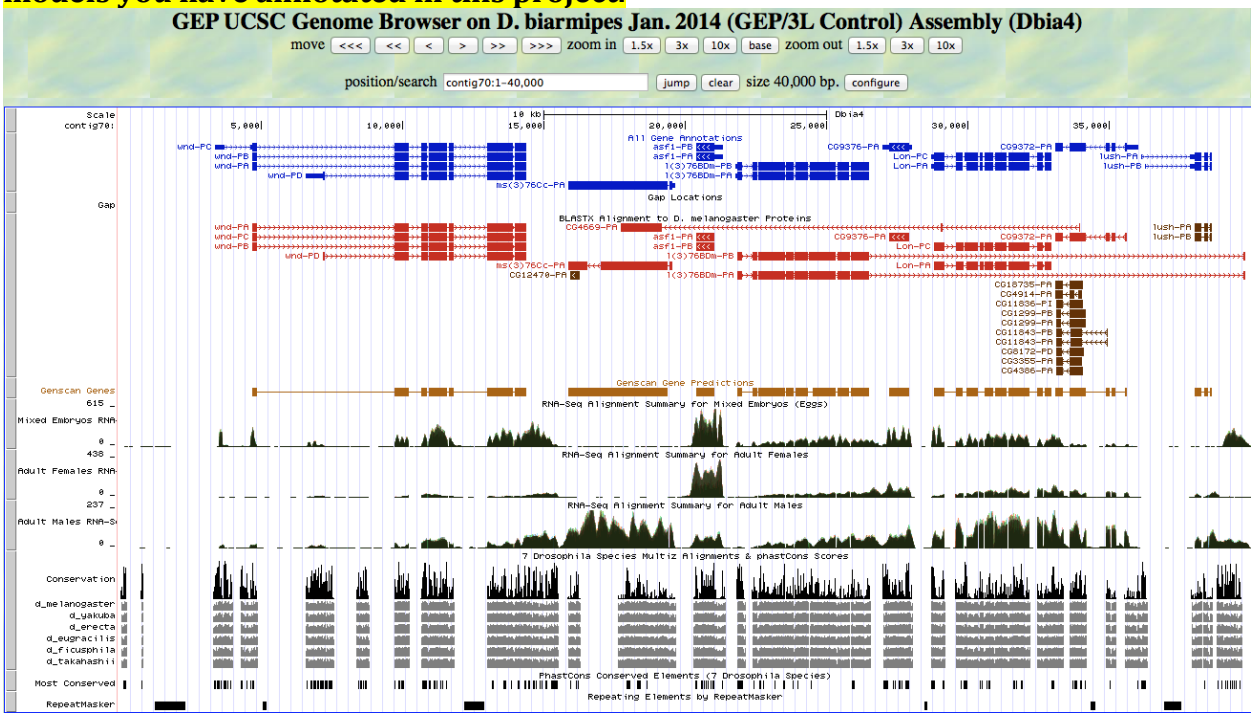
BRE^{u}	NA	NA
TATA Box	NA	NA
BRE^{d}	+21844	-19618677, -19618603, -19618535, -19618484, -19618481, -19618478, -19618476, -19618466, -19618451
Inr	NA	NA
MTE	NA	NA
DPE	+21690	NA
Ohler_motif1	NA	NA
DRE	+21724, +21761	NA
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	+21792	NA
Ohler_motif8	NA	NA

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.

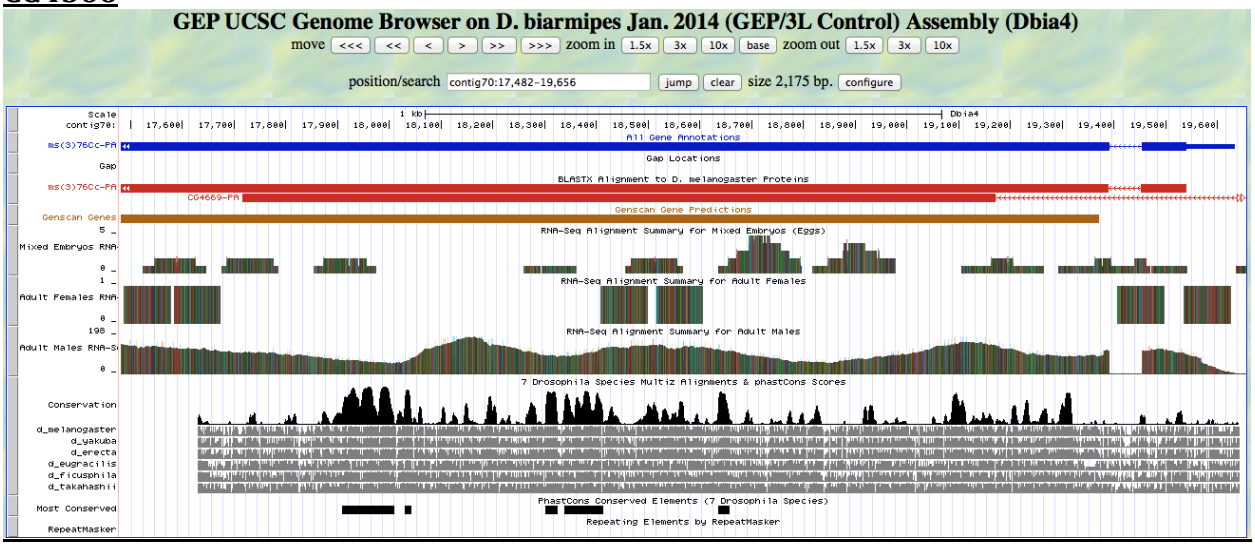


For projects with multiple errors in the consensus sequence, you should combine all the VCF files into a single project VCF file using the Annotation Files Merger (see the Annotation Files Merger User Guide for details). **Paste a screenshot (generated by the Annotation Files Merger) with all the consensus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

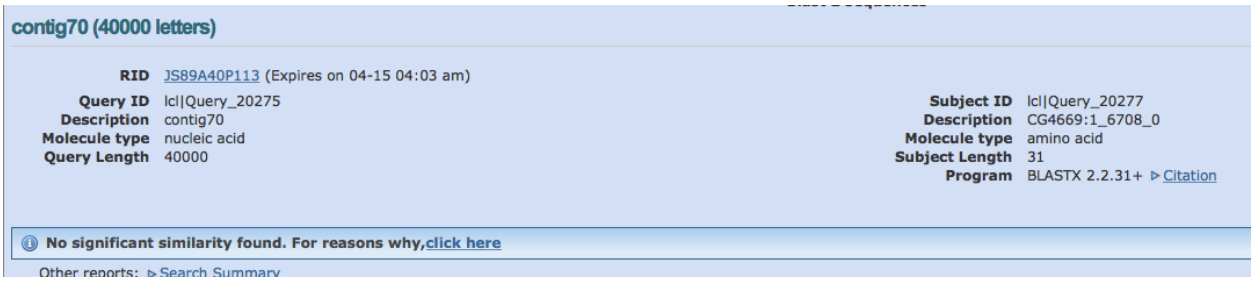
Range 1: 106 to 574 GenPept Graphics ▼ Next Match ▲ Previous Match

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

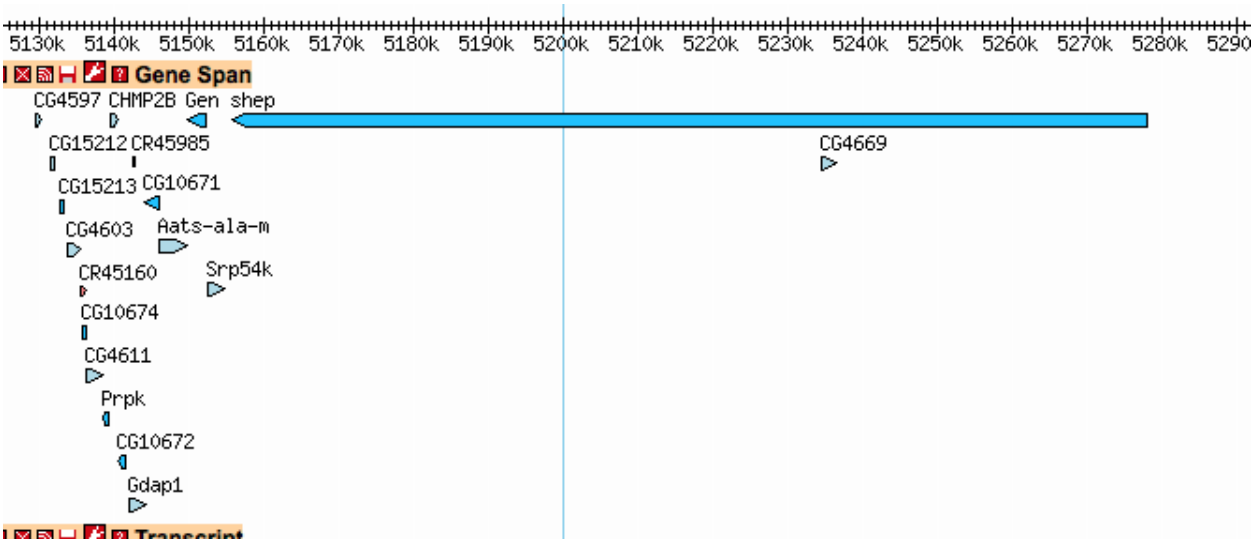
```
Query  1457  VNFATYLKAFALIYVLPEVDWTAEKIDMLLEEGTDLFRASSEVDQGKDDQEEHKLLEPEI  1278
             V FA+ LKAFA  Y     +W  + IDM++ EG  LF  S  +D              P+I
Sbjct  106   VRFASNLKAFAHSYYTRRDEWKPDTIDMVVNEGQVLFNDSENMDPPNASDS------PDI  159

Query  1277  YTNEEKRIKRNFNLEGHTFTLALEPRYQGAGSKPLEEQPPHTIDNLRPVLLNFFKSSRYC  1098
             Y   E+++ R+F +    F + LE  ++  G +       + I NLR +L  FFK ++Y
Sbjct  160   YDENEQKVTRHFKINDIEFAMELEAPFEVYGYE-------NVIRNLRRILRAFFKKAKYG  212

Query  1097  LLLTRVGHLLVWRRRKVFFVLDVKGRRIEDLETVQDNGVAMLVCLKTIDNVVHLASNLSG  918
             +  T   +LL+W+ + V+ VLD+ GR    ++   + G  +L+ LK+ DNVV L    S
Sbjct  213   VFFTPNWYLLIWKEKGVWMVLDLNGRDKNTMKPNNEEGYPLLLGLKSFDNVVWLIKKESY  272

Query  917   ISPQDEFTIRELVVVRLETPD--GRIYMRDTSHRSIEFRVVNKSYAYLKATLHLSLNEHD  744
             +    +F+IRE++VVRL TP   G+ + R+   R  +F V+   YAY+K+ LHL+LN  D
Sbjct  273   LDKNAKFSIREILVVRLATPGSTGQSWEREHGMRMSQFDVIASDYAYVKSNLHLTLNSKD  332

Query  743   PVRNRSSLMVAVGSILASKIDHPANWDTNMLDRLICYGVELCRNCWSDCLRDRRPIDLDT  564
              +RNRS+L VAV + LASKIDHPA WD  M D+++CYGV +C+NCW  C+   +P+DLD
Sbjct  333   ALRNRSALPVAVATALASKIDHPATWDQKMYDKVMCYGVNMCKNCWEPCMDPSKPMDLDD  392

Query  563   FPTQLRMGQYVLELKLIPNVRAGHWKCGVRIIGTDFEAHVSEALNEFGNVVFQINNQMYA  384
             FP Q+R+GQ+V E+ L PN   G WKC       DF   + +AL++   ++FQINNQMY+
Sbjct  393   FPRQIRLGQFVAEIMLTPNAYEGWWKCVPMYKFNDFHLMLEKALDQNDYIIFQINNQMYS  452

Query  383   IWVKDEFYYLLDPYRHTIVGTHVAEDKGEGAKWATVRMFRDQLTMLSVFHQMLKESNRQS  204
             +W K +F YL+DPYRH IVG  + E  GE  K ATVRMF +   +LSVFHQ+L ESNR +
Sbjct  453   LWKKSDFIYLMDPYRHNIVGRILEE--GEDPKSATVRMFGNMDRLLSVFHQILLESNRSA  510

Query  203   AYYLHVVRIRNLAECPEGY--ALVPLSEDVGNADVKSLNEPILFNEQQGVRVCDKSLADI  30
              +++H +RIRN+ ECP G   AL+P  EDV   +V+SLNE I F+E      C + L +I
Sbjct  511   VFHIHTLRIRNITECPPGTAPALLPPDEDV---EVRSLNENIRFDENYD--KCLQELGEI  565

Query  29    SDYEEDVIS  3
             SD+EED++S
Sbjct  566   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.

.

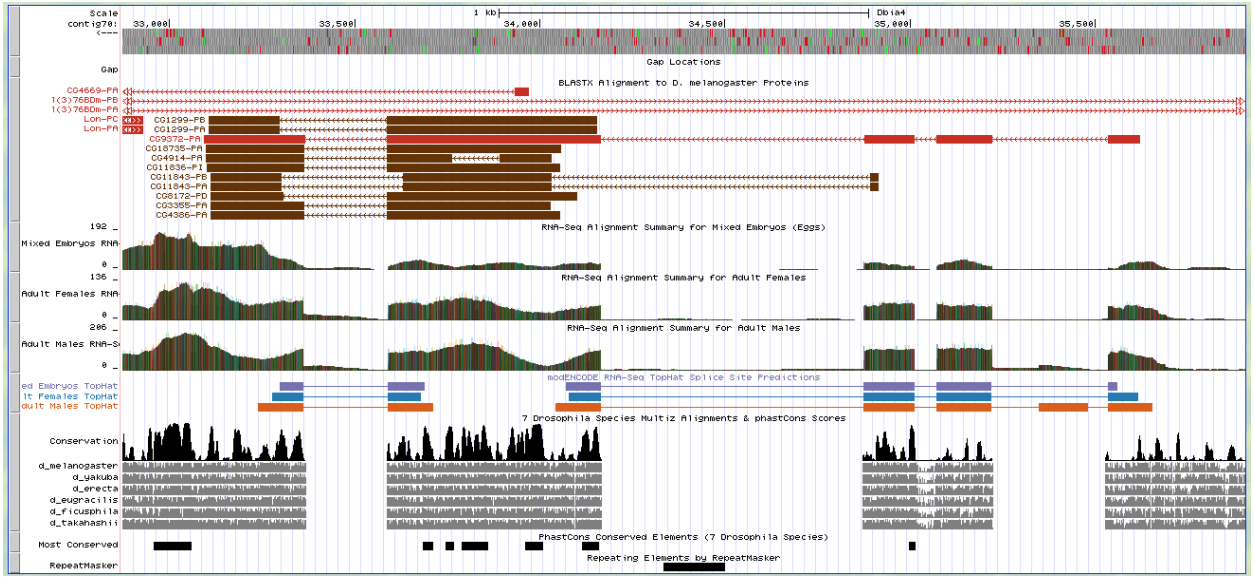


The first exon of CG4366 could not be found in my contig.



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.