Transcription Start Sites Project Report

Sample TSS report for *onecut*

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

Project details

Project name: contig35
Project species: *D. biarmipes*
Date of submission: 12/26/2023
Number of genes in project: 3

Does this report cover TSS annotations for all of the genes or is it a partial report? Partial report
If this is a partial report, please indicate the region of the project covered by this report:
From base 12,051 to base 21,700

Transcription start sites (TSS) report form

Gene name (e.g., *D. biarmipes eyeless*): *D. biarmipes onecut*

Gene symbol (e.g., *dbia_ey*): *dbia_onecut*

Name(s) of isoform(s) with unique TSS	List of isoforms with identical TSS
onecut-RA	*onecut-RB*

Complete this report form for each gene in your project. Copy and paste this form to create as many copies as needed.

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

Isoform TSS report

Complete an Isoform TSS report (through page 7) for each unique TSS listed in the table above. If the gene has more than one unique TSS, copy and paste this form to create as many copies as needed.

Gene-isoform name (e.g., *dbia_ey-RA*): *dbia_onecut-RA*
Names of the isoforms with the same TSS as this isoform:
dbia_onecut-RB

Type of core promoter in *D. melanogaster* (see table below):
(Peaked / Intermediate / Broad / Insufficient Evidence)
Peaked

The type of core promoter is defined by the number of TSS annotated by the Celniker group at modENCODE and the number of DHS positions:

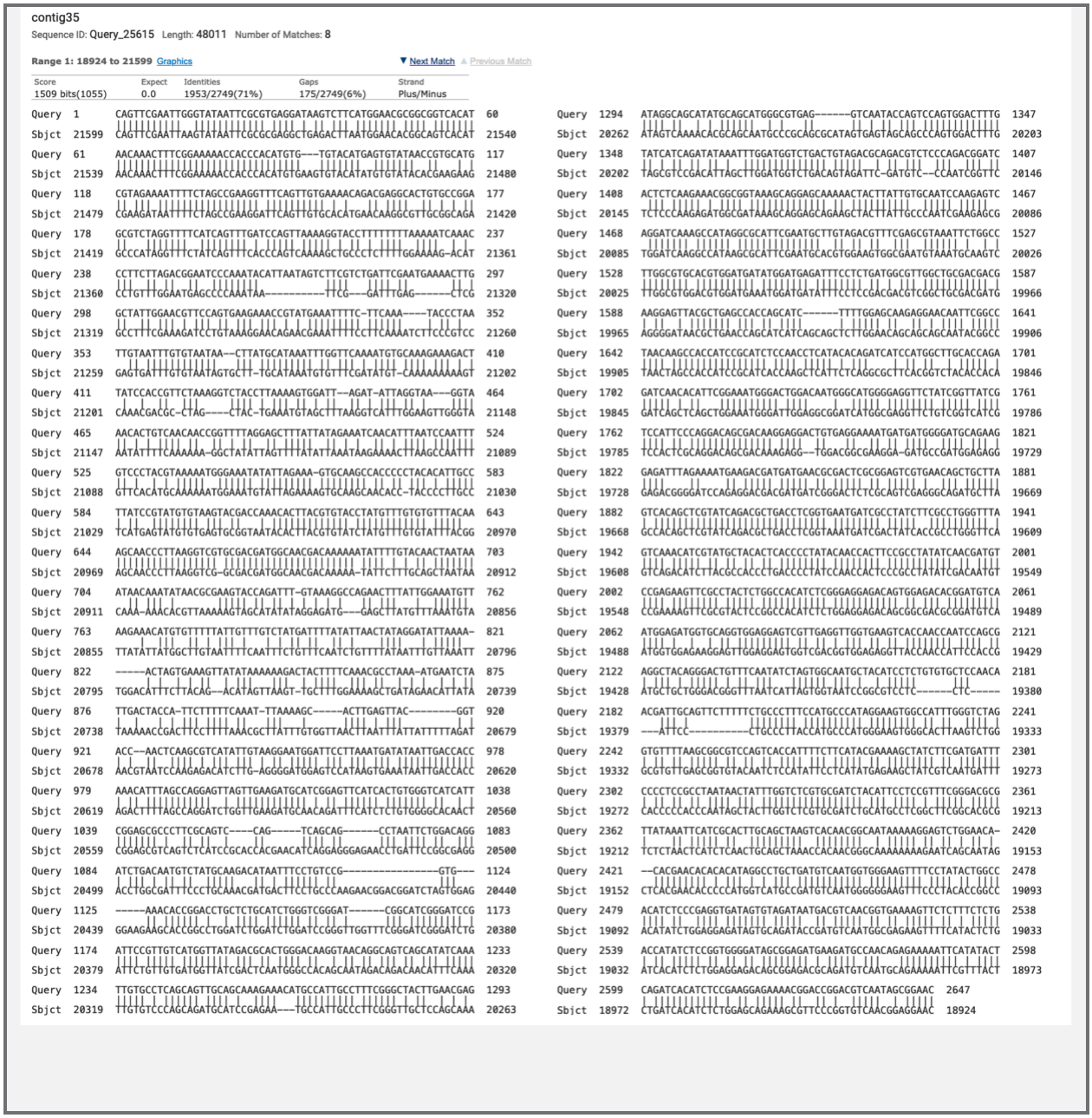
Type of core promoter	# annotated TSS	# DHS positions
Peaked	1	0
	0	1
	1	1
Intermediate	≤ 1	> 1
	> 1	≤ 1
Broad	> 1	> 1
Insufficient Evidence	0	0

1. Annotate the first transcribed exon

Coordinates of the first transcribed exon based on *blastn* alignment:
21,599–18,924

Does the *blastn* alignment cover the entire *D. melanogaster* first transcribed exon? If not, specify the parts of the *D. melanogaster* exon that are missing from the *blastn* alignment.
Yes, the *blastn* alignment covers the entire length (2647 bp) of the first transcribed exon onecut:3

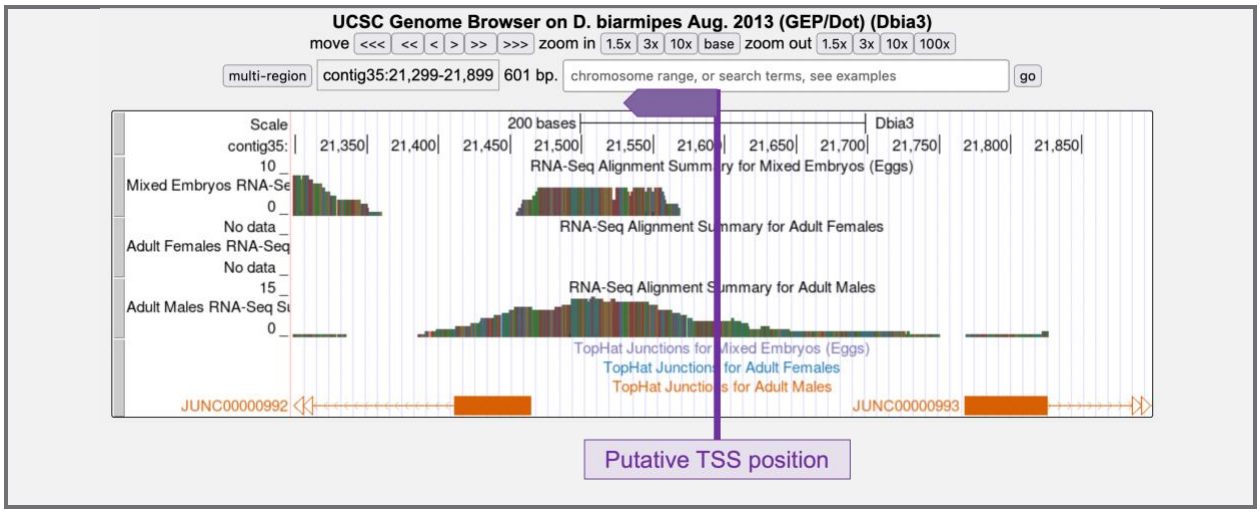
If the TSS annotation is supported by *blastn* alignment of the initial transcribed exon against the contig sequence, **paste a screenshot of the *blastn* alignment into the box below:**



2. Turn on RNA-Seq evidence tracks

If the TSS annotation is supported by RNA-Seq read coverage or splice junction predictions (e.g., TopHat, regtools), **paste a Genome Browser screenshot of the region surrounding the putative TSS (± 300 bp) with the following evidence tracks:**

1. RNA-Seq Alignment Summary or RNA-Seq Coverage
2. RNA-Seq TopHat or Splice Junctions



If the RNA-Seq evidence tracks indicate the TSS position, list it here: NA (see explanation below)
If the RNA-Seq evidence tracks indicate a TSS search region, list it here: contig35:21,599–21,750

3. Turn on comparative genomics tracks

If the TSS annotation is supported by sequence conservation with other *Drosophila* species, **paste a screenshot of the multiple sequence alignment (e.g., from Clustal Omega, ROAST) into the box below:**

4. Define the TSS search region(s)

Note:
If the *blastn* alignment to the initial transcribed exon satisfies the criteria listed on page 8 of Module TSS4 (i.e., a long match with low E-value, requires extrapolation of less than 150 bp to the estimated TSS position, alignment is in concordance with other evidence tracks), then you can define the TSS search region as +/- 300 bp from the initial 5' nucleotide. For example, if the estimated TSS position is located at position 1500, then the narrow TSS search region would be placed at 1200-1800.

If you cannot estimate the TSS position based on the *blastn* alignment to the initial transcribed exon, then you can define the TSS search region(s) based on the experimental data (e.g., RNA-Seq, RNA PolII ChIP-Seq) and the conservation track for the target species. If part of the TSS search region is only weakly supported by the available evidence, then please specify both a **wide** and a **narrow** search region. For example, if the region at 1500-2000 shows high RNA-Seq read coverage but there is very low RNA-Seq coverage from 1000-1499, then you will report "**1000-2000**" as the wide search region and "**1500-2000**" as the narrow search region.

Enter "Insufficient evidence" if the narrow search region cannot be determined based on the available evidence.

Coordinates of the narrow TSS search region:

21,599–21,750

Coordinates of the wide TSS search region:
(Enter "NA" if the narrow TSS search region is defined based on the *blastn* alignment to the initial transcribed exon. Enter "Insufficient evidence" if a wide search region cannot be defined based on the available evidence)

Insufficient evidence

Describe the evidence used to define the TSS search region(s) (e.g., RNA-Seq and Conservation tracks in this species, RAMPAGE data from *D. melanogaster*):

The TSS search region is defined based on RNA PolII enrichment, RNA-Seq read coverage, sequence conservation among seven *Drosophila* species, and the Inr motif at 21,630 (see the explanations in section 6 below).

5. Search for core promoter motifs

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

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

For TSS annotations where you can only define a TSS search region (and not a single coordinate), you should report all motif instances within the narrow TSS search region. If you did not report a narrow TSS search region due to insufficient evidence, report the motif instances in the wide TSS search region.

Coordinates of the motif search region
Your project (e.g., contig10:1500-2000): contig35:21,299-21,899
Orthologous region in *D. melanogaster*: chr4:607,349-607,949

Record the **orientation and the start coordinate** (e.g., +10000) of each motif match below. (Enter "**NA**" if there are no motif instances within the search region.)

Note: Highlight (in yellow) the motif instances that support the TSS annotation above.

Core promoter motif	Your project	*D. melanogaster*
BRE[u]	NA	NA
TATA Box	NA	NA
BRE[d]	-21409, -21494, -21606, -21611, -21640, -21650, -21686, -21695, -21818	-607557, -607635, -607656, -607661, -607663, -607666, -607695, -607705, -607745, -607748, -607757
Inr	-21399, -21448, -21595, -21630, -21883	-607452, -607501, -607645, -607675
MTE	NA	NA
DPE	-21361, -21446, -21809	-607499, -607831, -607905
Ohler_motif1	NA	NA
DRE	NA	NA
Ohler_motif5	NA	NA
Ohler_motif6	NA	NA
Ohler_motif7	NA	NA
Ohler_motif8	NA	NA

6. Summarize all of the evidence that supports the TSS annotation postulated above.

Coordinate(s) of the TSS position(s):
Based on *blastn* alignment: 21,599
Based on core promoter motifs (e.g., Inr): 21,598
Based on other evidence (please specify): NA

Note: If the *blastn* alignment for the initial transcribed exon is a partial alignment, you can **extrapolate the TSS position** based on the number of nucleotides that are missing from the beginning of the exon. (Enter "Insufficient evidence" if you cannot determine the TSS position based on the available evidence.)

Were you able to define a TSS position based on the available evidence? Yes; at 21,599

If so, indicate whether the evidence listed below support the TSS position.
If not, indicate whether the evidence listed below support the TSS search region(s).

Evidence type	Support	Refute	Neither
blastn alignment of the initial exon from *D. melanogaster*	☒	☐	☐
RNA PolII ChIP-Seq	☐	☒	☐
RNA-Seq coverage and splice junctions	☐	☒	☐
Core promoter motifs	☒	☐	☐
Sequence conservation with other *Drosophila* species (e.g., "Conservation" track on the Genome Browser)	☐	☒	☐
RefSeq Genes, N-SCAN PASA-EST, and Augustus TSS predictions	☐	☐	☒
Other (please specify) TSS identified by the RAMPAGE datasets in *D. melanogaster*	☒	☐	☐

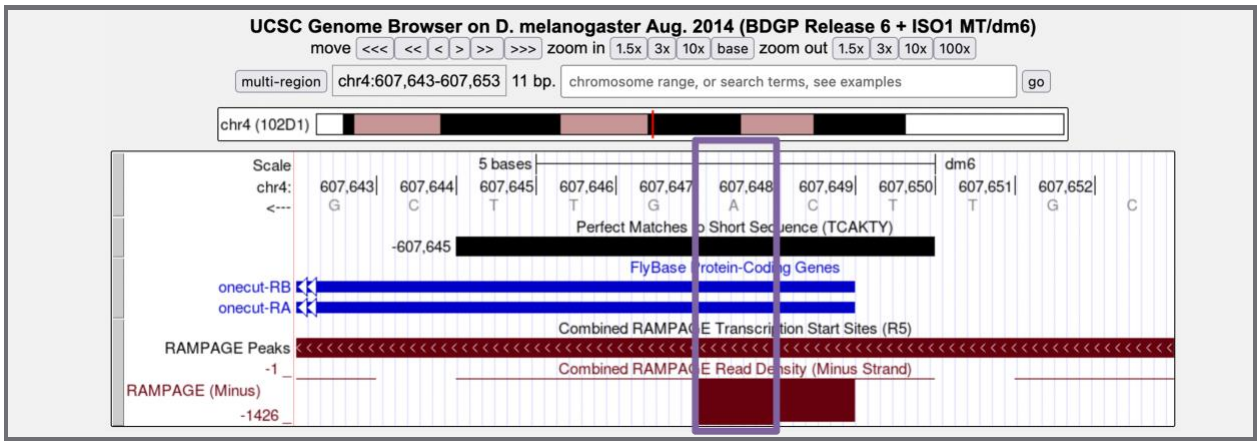
Note: The evidence type refutes the TSS annotation only if it **suggests an alternate TSS position**. For example, the presence of RNA-Seq read coverage upstream of the annotated TSS indicates that the TSS is located further upstream and it would be considered to be evidence against (i.e., Refute) the annotated TSS. In contrast, the lack of RNA-Seq read coverage is a negative result and it neither supports nor refutes the TSS annotation (i.e., Neither).

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

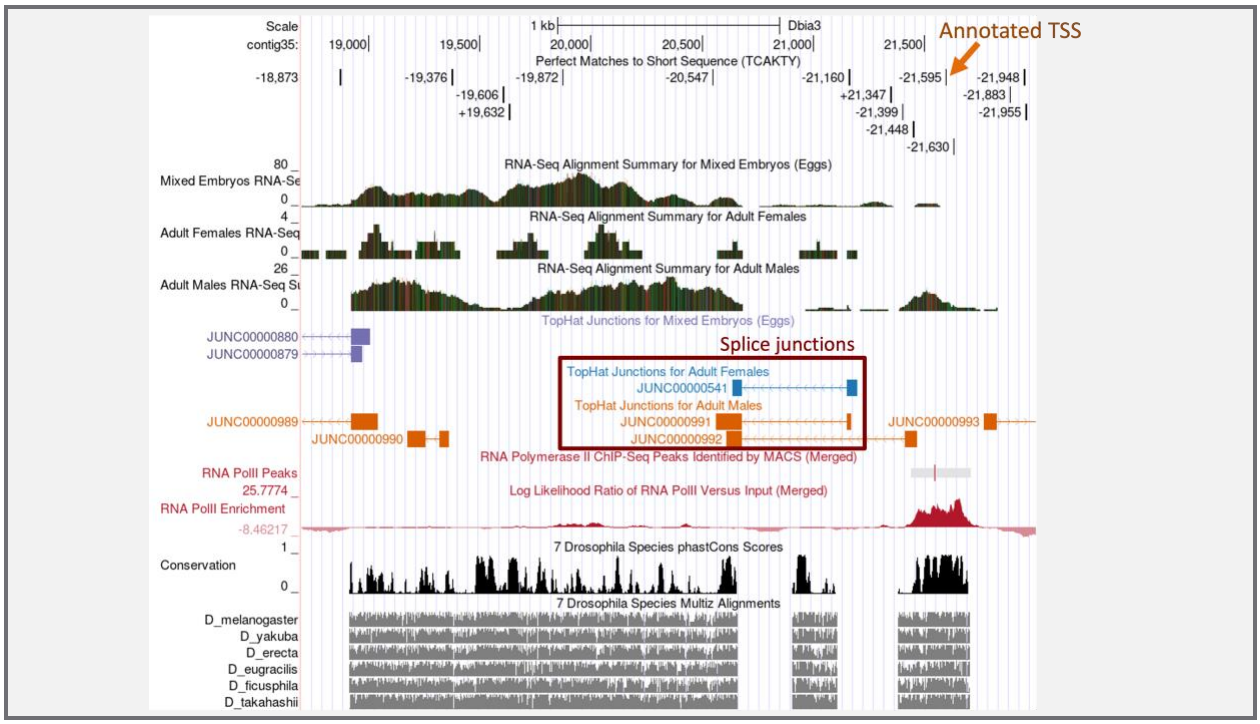
The annotation of the TSS position for the *D. biarmipes onecut* gene is based on minimizing the change in the size of the initial transcribed exon compared to the *D. melanogaster* ortholog (i.e., parsimony). The *blastn* alignment of the initial transcribed exon of the A isoform of *onecut* (onecut:3) from *D. melanogaster* against the genomic sequence of *D. biarmipes* contig35 shows a full-length alignment and the exon is highly conserved (with an E-value of 0.0 and a sequence identity of 71%; see screenshot above). The *blastn* alignment placed the TSS of *onecut* at 21,599. There is also an Inr

motif at 21,595–21,600. Because the Inr motif is found at -2 of the TSS, the Inr motif would place the TSS of *onecut* at 21,598.

Examination of the Combined RAMPAGE TSS datasets for *D. melanogaster* suggest there are two strong TSS in *onecut*. The TSS at 607,649 is consistent with the *D. melanogaster* FlyBase annotations for the A and B isoforms of *onecut* while the TSS at 607,648 is consistent with the Inr motif. Both TSS have high RAMPAGE read density but the TSS at 607,648 is stronger than the TSS at 607,649. Hence the TSS of *onecut* for *D. biarmipes* is placed at 21,599 based on parsimony with *D. melanogaster*.

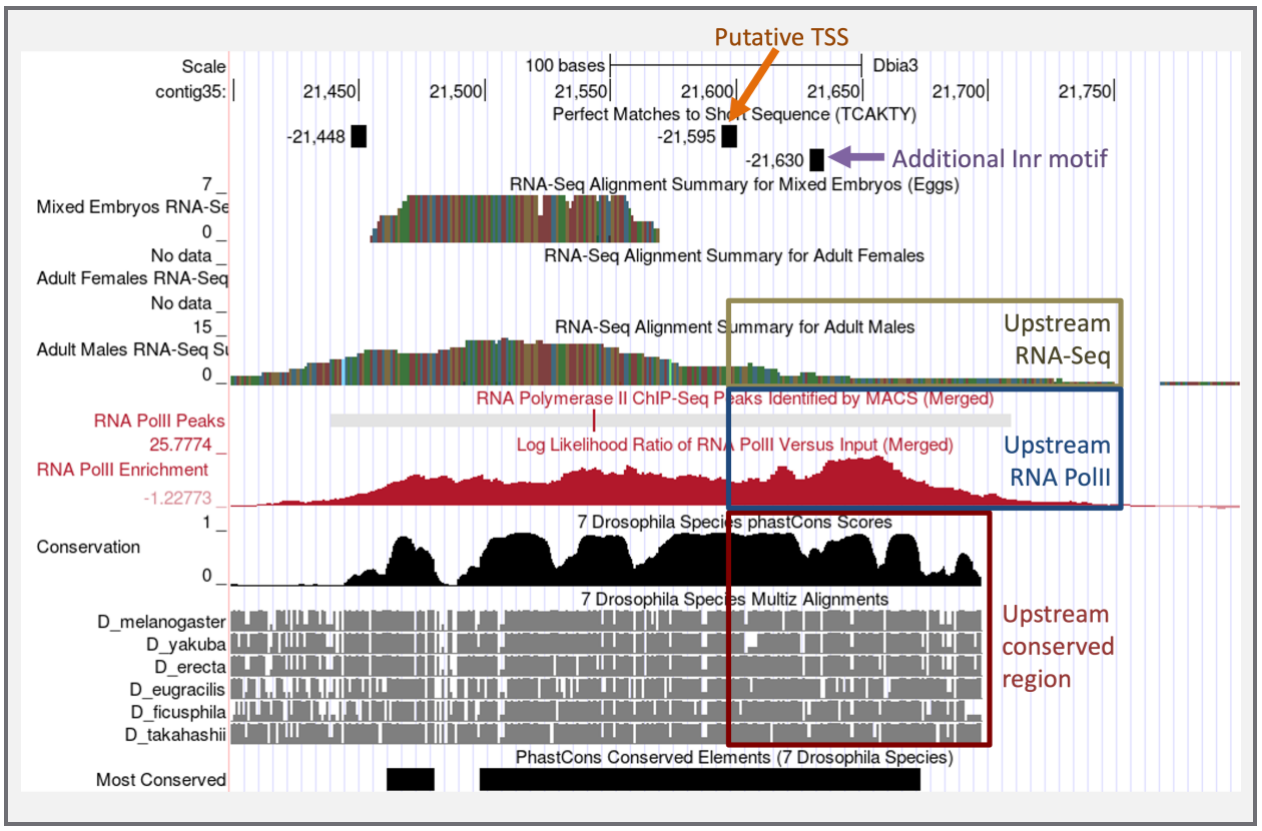


Examination of the region surrounding the *blastn* alignment to exon onecut:3 (i.e., 21,599–18,924) shows that the available RNA-Seq data generally supports the proposed TSS position at 21,599. For example, the TopHat junctions in adult females and adult males (JUNC00000541 and JUNC00000991, respectively) are consistent with the splice junction between exons onecut:1 and onecut:2 in the B isoform of *onecut*.

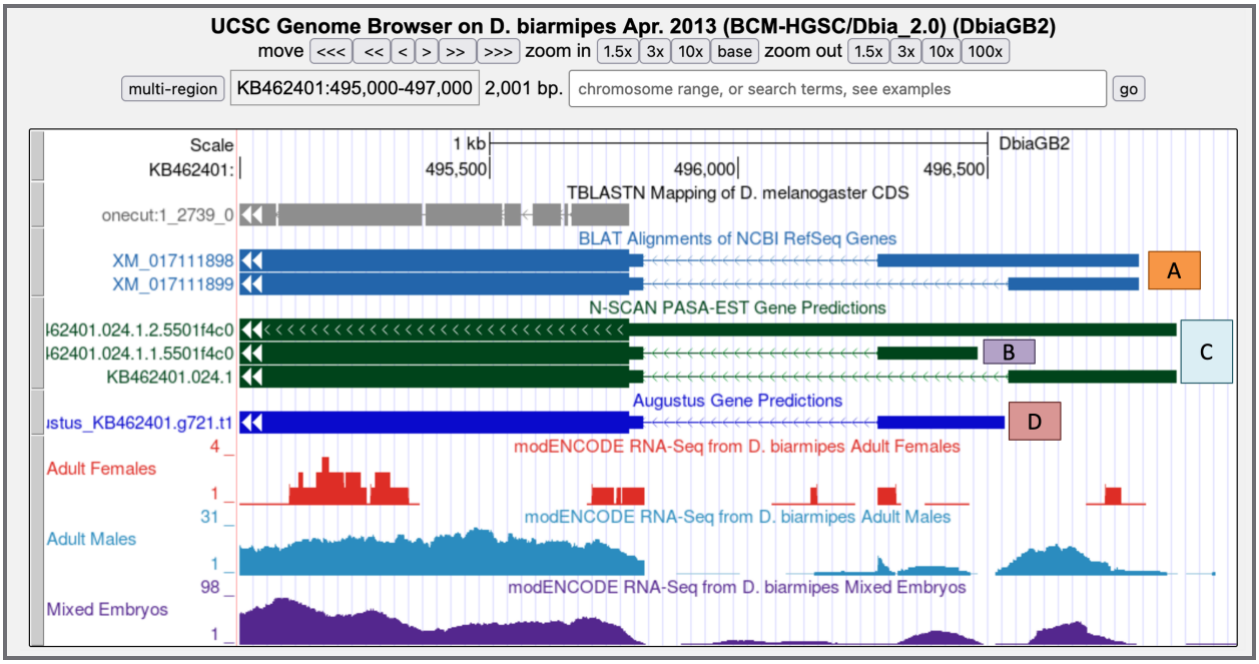


While the RNA-Seq evidence is generally in congruence with the proposed TSS position, the RNA-Seq alignment summary from the adult males sample shows additional RNA-Seq read coverage upstream (i.e., further to the right) of the annotated TSS.

Similarly, the "RNA PolII Peaks" and the "RNA PolII Enrichment" tracks indicate the presence of RNA PolII upstream of the annotated TSS at 21,599. This region also has an Inr motif at 21,630. According to the "Conservation" and the "Most Conserved" tracks, part of this upstream region (21,600–21,697) is highly conserved with the orthologous regions from six *Drosophila* species (*D. melanogaster*, *D. yakuba*, *D. erecta*, *D. ficusphila*, *D. eugracilis*, and *D. takahashii*). Collectively, the available evidence suggests that the TSS of *onecut* might be located further upstream of the proposed TSS for *onecut* in contig35.



Examination of the RefSeq genes, N-SCAN, and Augustus gene predictions in the *D. biarmipes* April 2013 (BCM-HGSC/Dbia_2.0) assembly shows that each tool assigned different position(s) as the TSSs for *onecut* (labeled A–D in the figure below). Due to these inconsistencies and the high error rate associated with TSS predictions, these TSS predictions were not used to define the TSS position or the TSS search region for the *onecut-RA* ortholog in *D. biarmipes*.



Collectively, the narrow TSS search region for *onecut-RA* is defined at contig35:21,599–21,750 to account for the RNA-Seq read coverage, RNA PolII ChIP-Seq data, conservation among seven *Drosophila* species, and the location of the additional Inr motif.