Module TSS4:

Annotation of Broad Transcription Start Sites (Advanced)

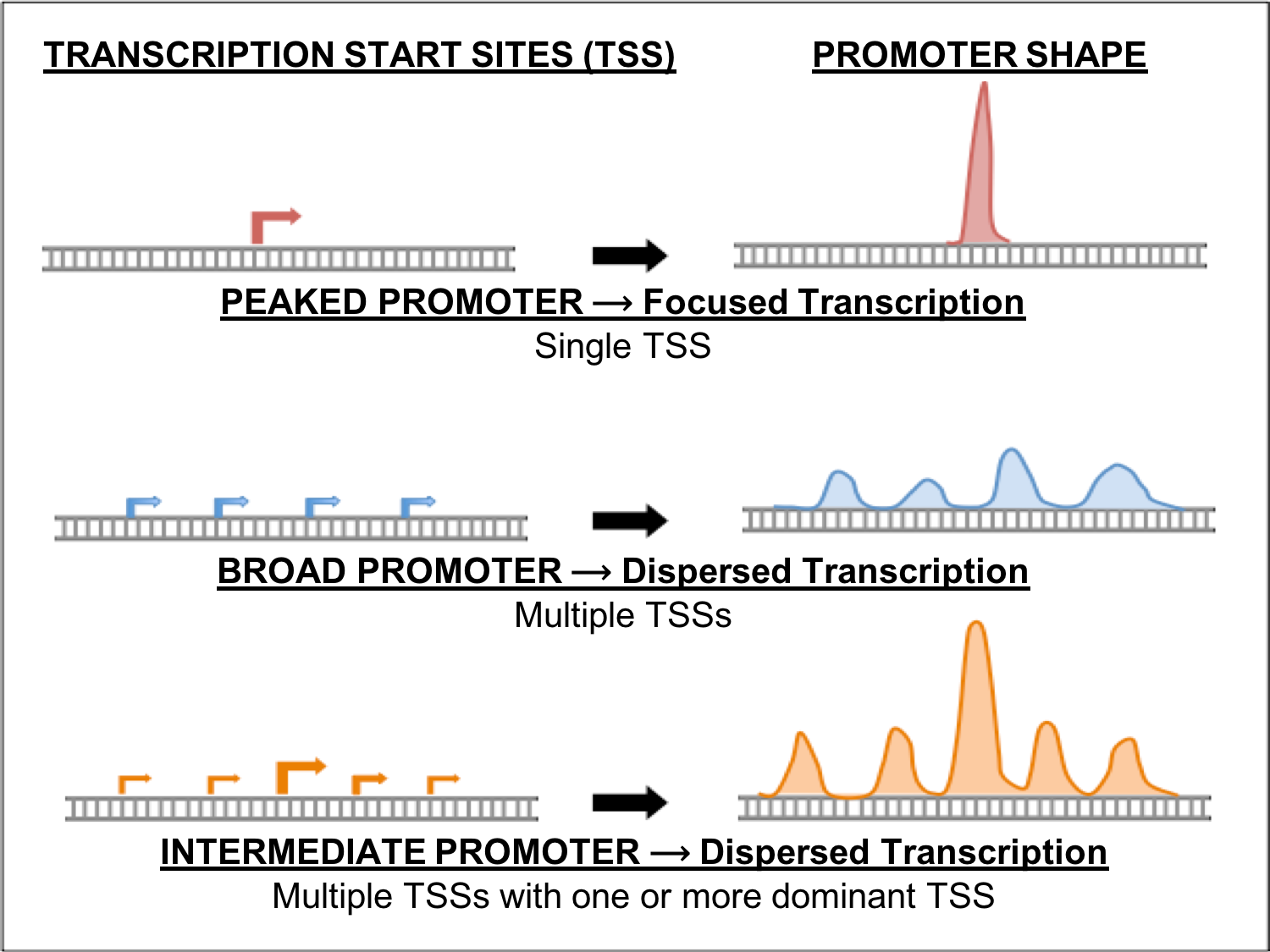
#### Michael Foulk

Lesson Plan:

|  |  |
| --- | --- |
| **Title** | **Identifying transcription start sites for Broad promoters using available combinations of evidence.** |
| **Objectives** | * Characterize a broad TSS in a *D. melanogaster* ortholog * Use landmarks to set the boundaries of TSS search regions * Use to CAGE and RAMPAGE data to confirm the characterization of the D*. melanogaster* TSS * Identify narrow and wide TSS search regions when BLASTn fails to localize the first transcribed exon * Use other lines of evidence to determine a wide TSS search region even when BLASTn succeeds in localizing the first transcribed exon * Use the available evidence to annotate a TSS search region |
| **Pre-requisites** | Modules TSS 1-3 |
| **Order** | * Overview of the challenges of annotating broad TSS * Annotate a broad TSS gene when BLASTn fails to localize the first transcribed exon *(*gw*-*RI) * Annotate a broad TSS gene when BLASTn succeeds in finding the first transcribed exon |
| **Homework** | None |
| **Class Instruction** | * How are peaked and broad promoters different? * How to use landmarks to set the boundaries of TSS search regions? * What is the relative ranking of pieces of evidence that can be used to identify narrow and wide TSS search regions of broad promoters? * Work through the Genome Browser examples of three isoforms of *gw* (RI, RB and RJ) * Conclude by challenging students to identify the TSS search region for the RA, RF and RE isoforms of *gw* and the TSS for *CG33941*. |
| **Associated Videos** | * RNA Seq and TopHat Video: <https://youtu.be/qepVXEsfLMM> * Short Match Video: <https://youtu.be/eoeWufgcdvg> |

# **Applying various lines of evidence to rationally identify a TSS search region for genes with intermediate or broad promoters**

In the previous modules in this series, we learned how to access information used to annotate a search region for the Transcription Start Sites (TSS) for genes in the GEP project. These modules were primarily focused on genes that had relatively simple peaked promoters (**Fig.1**). However, often genes that have intermediate or broad promoters are more difficult to annotate and several lines of evidence must be used to define a search region for the TSS. In particular, F element genes tend to have a higher probability of having broad promoters and can be challenging to annotate. Moreover, on occasion you will run into a peaked promoter that is similarly difficult to annotate and the strategy presented in this module can also be applied to these genes. This module will use several examples to illustrate strategies that you can use to annotate these genes using the available evidence.

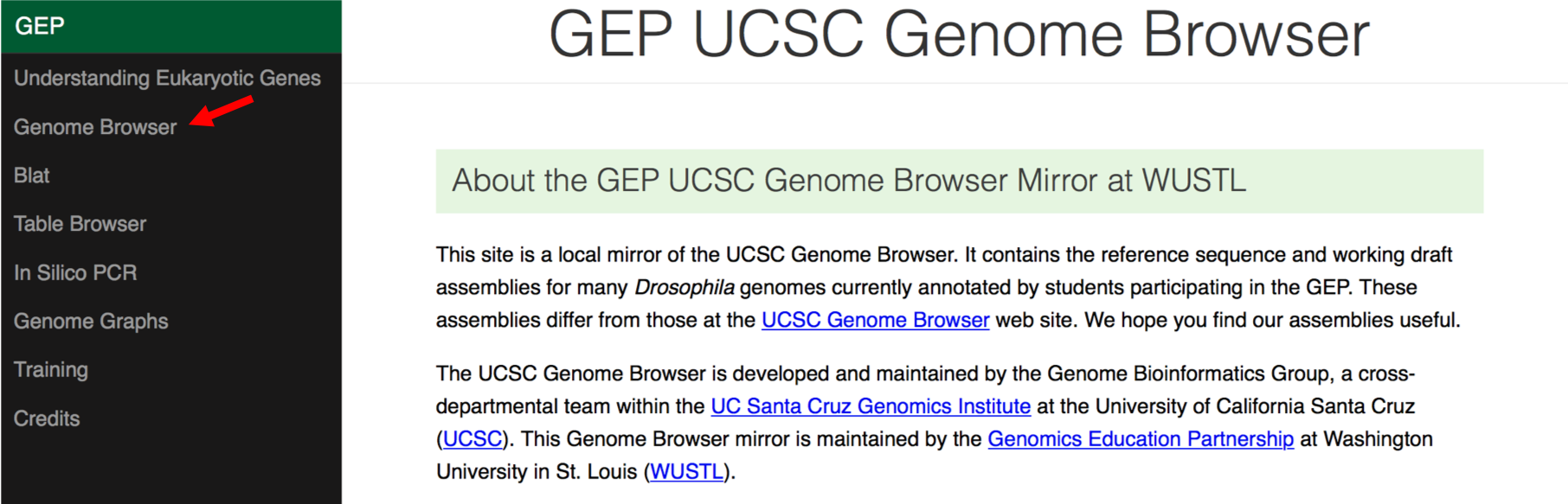


**Figure 1: Three major types of core promoters (Peaked, Broad, and Intermediate)**

# Exercise 1: Classification of intermediate/broad promoters in *D. melanogaster*

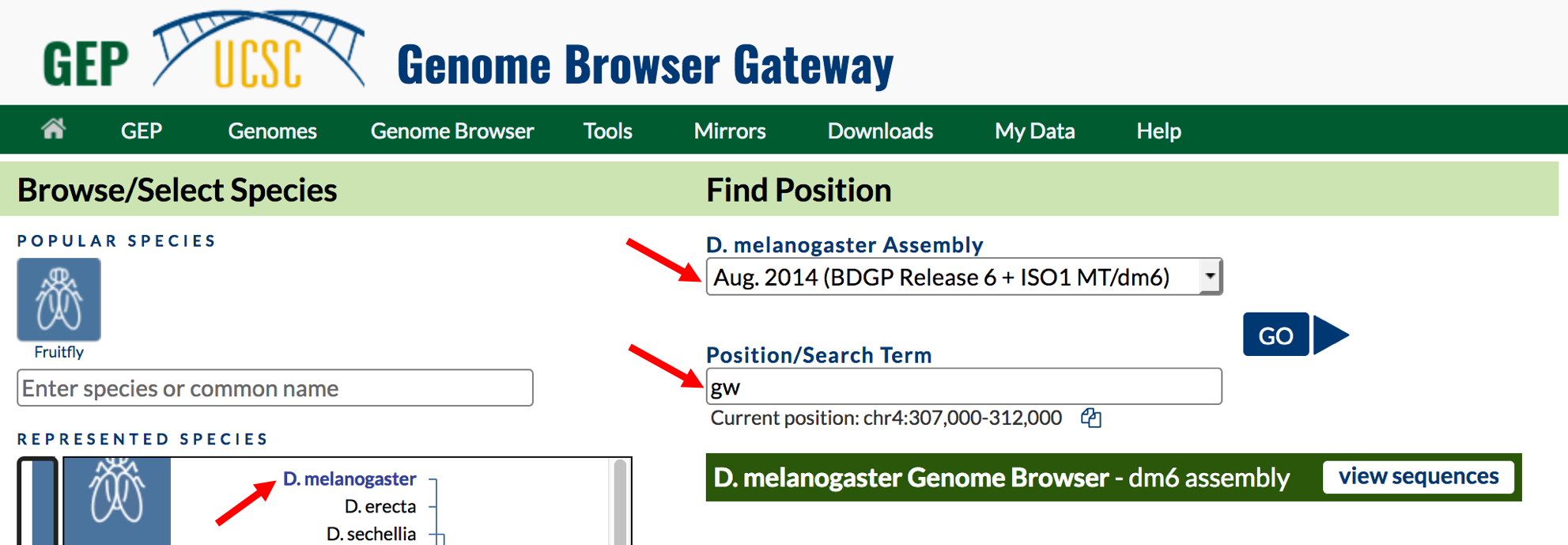
Let’s begin by classifying the promoter type of a *D. melanogaster* ortholog of one isoform of the *gawky* gene (gw-RI).

1. Open a new web browser window and go to the Genomics Education Partnership (GEP) UCSC Genome Browser Mirror at <http://gander.wustl.edu/>. This Genome Browser was developed by the Genome Bioinformatics Group at the University of California Santa Cruz (UCSC) and customized by the GEP to facilitate the annotations of multiple *Drosophila* species (**Fig. 2**).



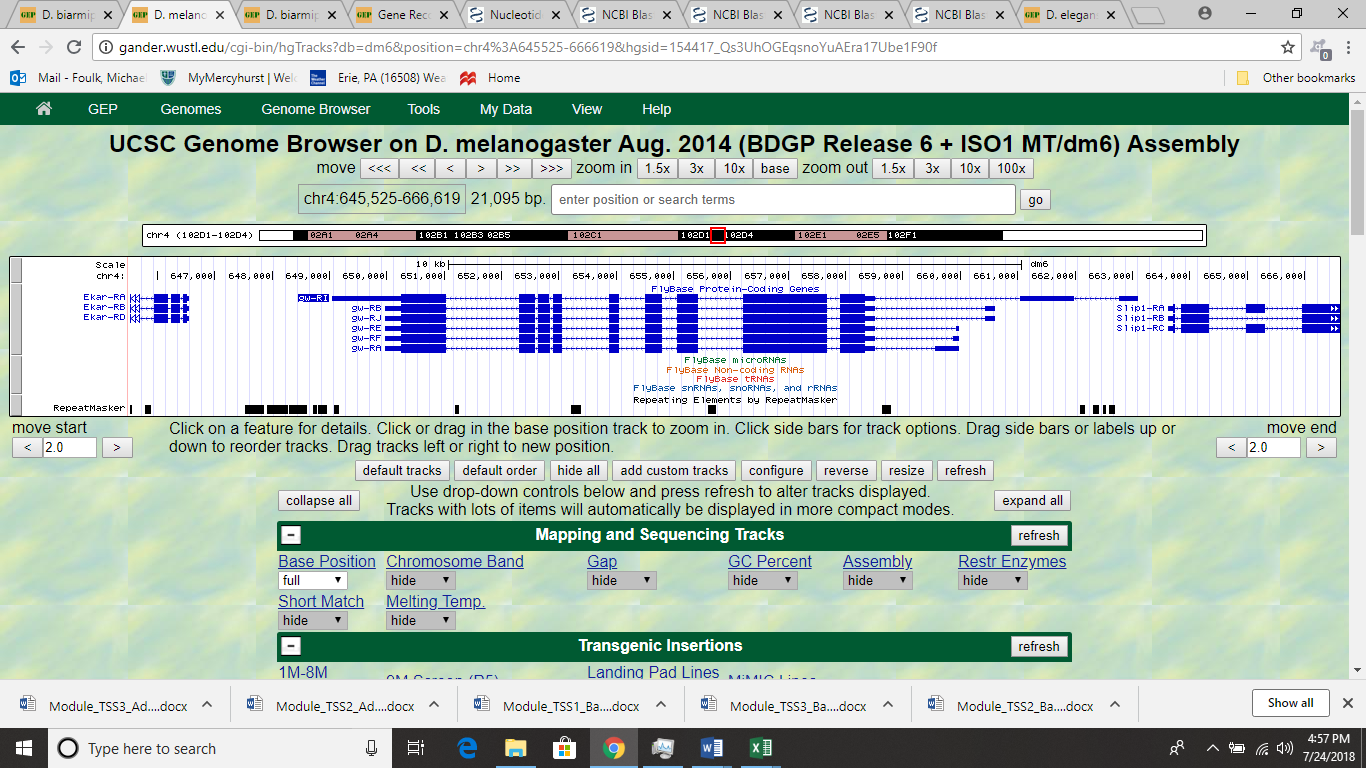
**Figure 2: Access the GEP UCSC Genome Browser Gateway page using the “Genome Browser” link.**

1. To navigate to the genomic region surrounding the RI isoform of the *gawky* (*gw*) gene in *Drosophila melanogaster*, select “***D. melanogaster***” under “Represented Species”, select “**Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6)**” under “D. melanogaster Assembly”, and then enter “**gw**” under the “Position/Search Term” field. Click on the “**GO**” button (**Fig. 3**).



**Figure 3: Change the settings on the Genome Browser Gateway page to navigate to the RI isoform of the *gw* gene in the *D. melanogaster* release 6 assembly.**

1. The resulting page shows a list of the *gw* isoforms (6) present in *D. melanogaster*. Click on the link for [gw-RI at chr4:649041-663103](http://gander.wustl.edu/cgi-bin/hgTracks?position=chr4:649041-663103&hgsid=154417_Qs3UhOGEqsnoYuAEra17Ube1F90f&fbgenes=pack&hgFind.matches=gw%2DRI,). The *gw* gene is located on chromosome 4 and the RI isoform is located between nucleotides 649,041 and 663,103.
2. Zoom out 1.5X. The resulting genome browser window will show the genomic region spanned by the gw-RI (**Fig. 4**). The *gw* gene is encoded on the minus strand. In previous modules, we have used the reverse button to flip the orientation of the chromosome in order to make it easier to interpret the evidence tracks. However, in this module, we will keep the browser in the original orientation. It is useful to practice interpreting the evidence in either orientation.

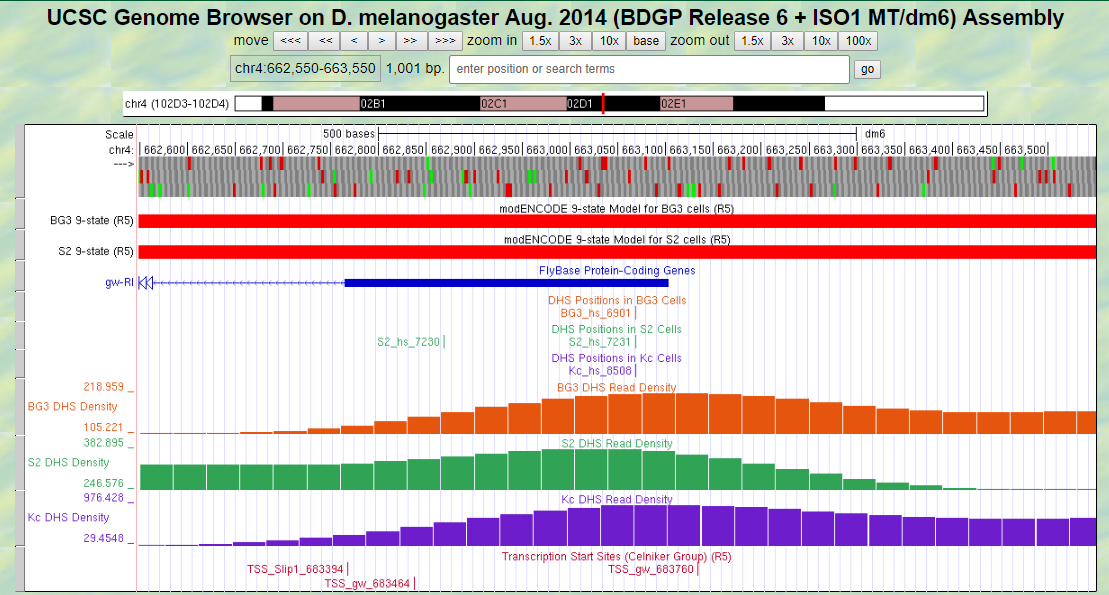
**Figure 4: The UCSC browser window displaying the genomic region surrounding the *gw* gene in *D. melanogaster*. Click on the “hide all” button to hide all of the evidence tracks.**

1. Because the Genome Browser remembers the previous display settings, we will hide all the evidence tracks and then enable only the subset of tracks that we need. Click on the **“hide all”** button located below the Genome Browser image (**Fig. 4**). Next, we will display the evidence tracks that will allow us to characterize the type of promoter.

Configure the display modes as follows:

* Under “Mapping and Sequencing” Tracks
* Base Position: **full**
* Under “Chromatin Domains” Tracks
* BG3 9-state (R5): **dense**
* S2 9-state (R5): **dense**
* Under “Gene and Gene Prediction” Tracks
* FlyBase Genes: **pack**
* Under “Expression and Regulation”
* TSS (Celniker) (R5): **pack**
* Detected DHS Positions (Cell Lines) (R5): **pack**
* DHS Read Density (Cell Lines) (R5): **full**

1. Since we want to characterize the promoter of the gw-RI isoform we will zoom in on that region. Either highlight the farthest upstream exon to zoom in or type **chr4:662550-663550** into the “**enter positions or search terms**” text box, and then click the “**go**” button (**Fig. 5**).



**Figure 5: Genome browser view of the first exon of gw-RI in *D. melanogaster* showing the evidence used to characterize the promoter shape.**

**Q1.** **According the 9-state models, how is the chromatin state of the region surrounding the first exon of gw-RI classified?** (Hint: under “Chromatin Domains” click on the blue/purple 9-state links, either the BG3 or S2, to see a chart that describes the different colors displayed in these tracks.)

**Q2. How many DNase hypersensitive sites are located in this region?**

**Q3. How many TSSs sites were predicted in the Celniker data in this region?**

**Q4. Based on this data how should we classify the promoter of the *D. melanogaster* gw-RI isoform?** (Use Table 1 below to remind yourself of how to use this data to characterize the promoter.)

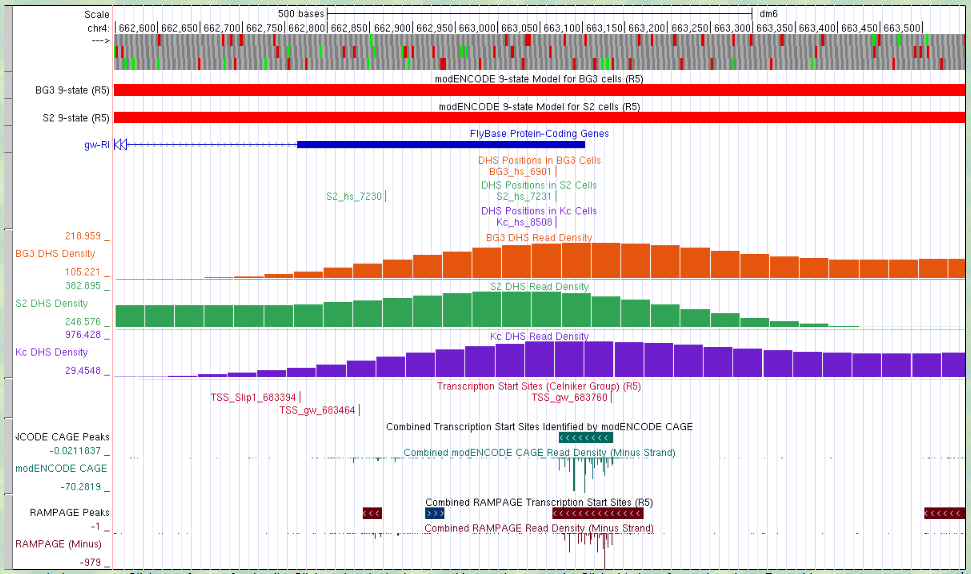
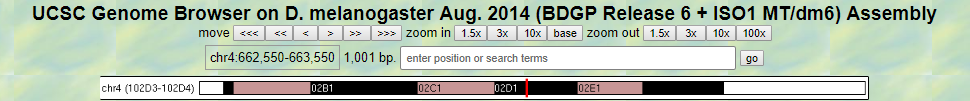
|  |  |
| --- | --- |
| TABLE 1: CLASSIFICATIONS OF DROSOPHILA PROMOTERS FOR THE GEP | |
| PEAKED | * One annotated TSS with no DHS position * No annotated TSS with one DHS position * One annotated TSS with one DHS position |
| INTERMEDIATE | * Zero or one annotated TSS with multiple DHS positions * Multiple annotated TSS with zero or one DHS positions |
| BROAD | * Multiple annotated TSS with multiple DHS positions |
| INSUFFICIENT  EVIDENCE | * No annotated TSS and no DHS positions |

**Table 1: For the GEP TSS annotation projects, the core promoter is classified into one of four categories based on the number of TSS (Celniker) annotations and the number of DHS positions within a 300 bp window.**

1. In addition to the DHS sites and Celniker TSS predictions, we can now utilize both CAGE and RAMPAGE data to support (or not depending on the data) the characterization of the promoter. To view this data, we will add these tracks to the browser while keeping the tracks we loaded above.

Configure the display modes as follows:

* Under “Expression and Regulation”
* Click on the blue “Combined modENCODE CAGE TSS” link
* Change the “Maximum display mode” to “**full**”
* Scroll down to the “List subtracks” section
  + Change the “modENCODE CAGE (Plus)” display mode to “**hide**”
  + Scroll up to the top of the page, and then click on the “**Submit**” button.
* Click on the blue “Combined RAMPAGE TSS (R5)” link
* Change the “Maximum display mode” to “**full**”
* Scroll down to the “List subtracks” section
* Change the “RAMPAGE (Plus)” display mode to “**hide**”
* Scroll up to the top of the page, and then click on the “**Submit**” button



**Figure 6: Genome browser view of the first exon of gw-RI in *D. melanogaster* showing the evidence used to characterize the promoter shape including CAGE and RAMPAGE data.**

**Q5. Do the CAGE and RAMPAGE data match with the locations of the Celniker predicted TSSs?** (Use the browser to zoom in for a closer look if necessary.)

**Q6. Are the CAGE and RAMPAGE data consistent with a peaked promoter or with a broad promoter shape?**

**Q7. Do the CAGE and RAMPAGE data support or not support the promoter shape that we determined using the DHS sites and Celniker predicted TSSs above? Explain.**

# Ranking lines of evidence for determining the boundaries of TSS search regions

As part of the GEP TSS annotation projects, we marshal different lines of evidence to define the boundaries of the TSS search region for any given gene. Ideally, the data from these different sources should be in agreement. However, not all lines of evidence are equal. The different lines of evidence that can be used to establish these boundaries are ranked below from most to least reliable.

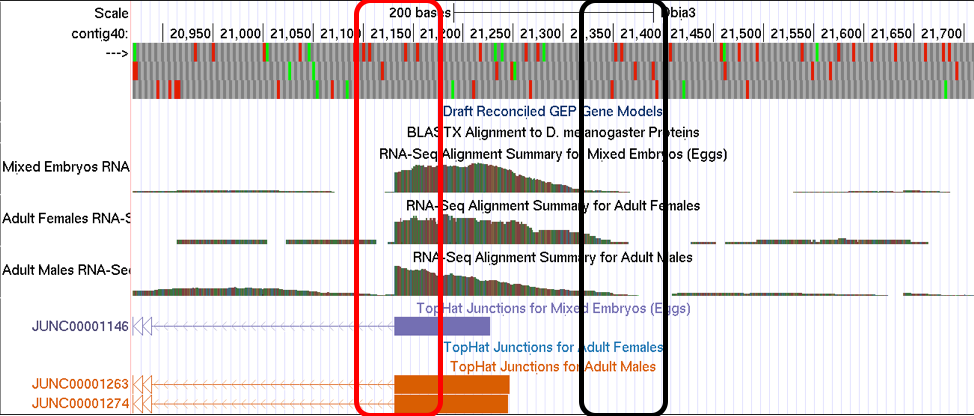
1. BLASTn alignment of the 1st transcribed *D. melanogaster* exon
2. RNA polII X-ChIP-seq
3. RNA-seq (including TopHat junctions)
4. Conservation

**1. BLASTn alignment of the 1st transcribed *D. melanogaster* exon:** The first and best evidence for determining the location of the TSS for a GEP project gene is sequence similarity to the first transcribed *D. melanogaster* exon. A BLASTn alignment can be used to define the TSS search region if it satisfies the following criteria: 1.) has a low E-value; 2.) has a long match between the subject and query sequences; 3.) requires an extrapolation of less than 150 bp to estimate the TSS position; and 4.) is in concordance with other evidence tracks on the genome browser. If the BLASTn alignment satisfies these criteria, then one can define the boundaries of the TSS search region as +/- 300 bp from the initial 5’ nucleotide.

However, sometimes BLASTn fails to identify the location of the 1st transcribed *D. melanogaster* exon (when it fails one or more of the criteria listed above). Fortunately, other lines of evidence can be utilized to define a TSS search region.

**2. RNA polII X-ChIP-seq:** RNA polymerase II is often enriched near the TSS of a gene. Therefore, the RNA polII X-ChIP-seq peaks can be used as landmarks for defining a TSS search region. The boundary of a search region should be set at the boundary of the peak called by MACS2 (the boundaries of the grey bar in the track).

**3. RNA-seq (including TopHat junctions):** The RNA-seq tracks can be very useful for determining a TSS search region. Often when BLASTn fails to identify the location of the 1st transcribed *D. melanogaster* exon, the RNA-seq and TopHat tracks can be used to identify the approximate location of this exon. Use the presence of TopHat reads and RNA-seq enrichment to zoom in on the initial exon. Moreover, these tracks can be used to define a TSS search region. A tail of continuous RNA-seq enrichment upstream of the initial exon can be used to define the upstream boundary of a search region. In addition, TopHat junctions can be used to define the downstream boundary of a search region because the TSS must be upstream of the splice site of the first exon. See examples of both of these types of regions in the boxed areas in **Figure 7**.



**Figure 7: Browser image of a region illustrating the tapering of RNA-seq reads at the 5’ end of the first transcribed exon (black box) and the 3’ splice site of the first exon indicated by the sharp drop-off of RNA-seq reads and the presence of TopHat splice junctions (red box).**

**4. Conservation:** The boundaries of highly conserved regions can serve as landmarks for defining the boundaries of a TSS search region.

# Exercise 2: Using evidence based landmarks to define the boundaries of a TSS search region

As discussed above, when BLASTn fails to identify the location of the 1st transcribed *D. melanogaster* exon, other lines of evidence can be used to define a TSS search region. When using these other lines of evidence, it is important to use clearly delineated landmarks in the evidence tracks to demarcate the TSS search region in order to ensure consistency between different student annotators. In other words, we want to avoid defining the boundaries of a search region arbitrarily. In this exercise we will use an example to practice using landmarks found in the various data tracks to establish TSS search region boundaries that are well supported by experimental evidence.

The genomic region we will investigate is in the vicinity of the TSS of the *myo* gene on the *D. biarmipes* Aug. 2013 (GEP/Dot) contig40. To navigate to this region, return to the Genome Browser Gateway page by following the directions above or by clicking on the “**Genomes**” link in the navigation bar at the top of the browser page. Click on “***D. biarmipes***” under “Represented Species”, select “**Aug. 2013 (GEP/Dot)**” under “D. biarmipes Assembly”, enter “**contig40**” under the “Position/Search Term” field, and then click on the “**GO**” button to navigate to this contig.

Click “**Hide all**” and configure the display modes as follows:

* Under “Expression and Regulation”
  + Base Position: **full**
* Under “Genes and Gene Prediction Tracks”
  + Reconciled Gene Models: **pack**
  + D. mel Transcripts: **pack**
* Under “RNA Seq Tracks
  + RNA-Seq Alignment Summary: **show**
  + RNA-Seq TopHat: **pack**
* Under “Expression and Regulation”
  + RNA PolII Peaks: **pack**
  + RNA PolII Enrichment: **full**
* Under “Comparative Genomics”
  + Conservation: **pack**
* Click on any of the **refresh** buttons.
* Finally, to zoom in to the region surrounding the TSS of the *myo* gene, type “**contig40:20,400-22,100**” into the “**enter position or search terms**” text box and click the “**go**” button.

**9**

**7**

**5**

**6**

**10**

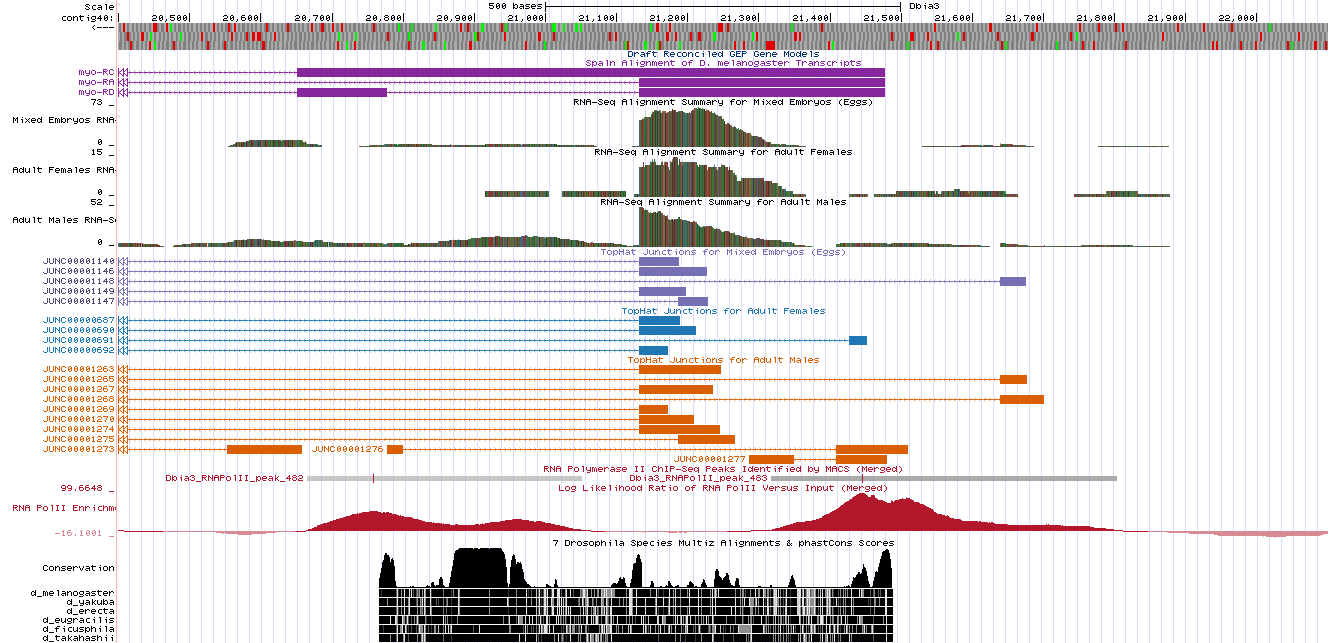
**3**

**4**

**8**

**2**

**1**



**Figure 8: Genome browser window of the region surrounding the TSS of the *myo* gene in *D. biarmipes* showing the tracks necessary for identifying a TSS search region when BLASTn fails to locate the 1st transcribed *D. melanogaster* exon. Lines are drawn from the top to ten landmarks that may or may not be used to define a TSS search region.**

**Q8. In Figure 8, numbered lines are drawn to ten locations in the vicinity of the TSS of the *myo* gene on contig40 in *D. biarmipes*. List which of the lines point to landmarks that are supported by evidence and which lines point to positions that are not supported by evidence. Explain your decision in each case.**

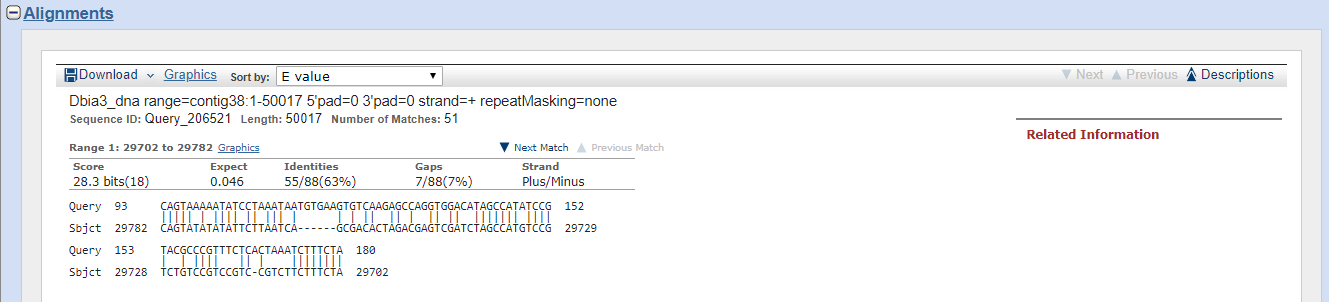
**Q9. Using the landmarks you identified above, define a narrow and a wide search region for the *myo* TSS. Describe the landmarks you used to determine these search regions.** (Use the browser to zoom in and out to identify the precise boundaries to the nucleotide.)

# Exercise 3: Determining a TSS search region for the gw-ri gene.

In exercise 1 we determined that the TSS of the RI isoform of the *gw* gene in *D. melanogaster* had a broad shape. Now that we have learned how to use evidence-based landmarks to set the boundaries of a TSS search region, let’s apply that knowledge to define a TSS search region for the gw-RI gene in the *D. biarmipes* GEP project sequence.

First, we will determine if we can locate the 1st transcribed *D. melanogaster* exon using BLASTn.

1. Open a new browser tab and navigate to the NCBI BLAST home page at <https://blast.ncbi.nlm.nih.gov/Blast.cgi>. Click on the “Nucleotide BLAST” image under the “Web BLAST” section.
2. Check the box “Align two or more sequences”. An “Enter Subject Sequence” text box will appear.
3. Open a new browser tab and navigate the Gene Record Finder on the GEP website (<http://gander.wustl.edu/~wilson/dmelgenerecord/index.html>). Enter “***gw***” into the search term box and click “Find Record”
4. Scroll down and click on the “Transcript Details” tab. By looking at the Exon Usage Map, we can see that the first transcribed exon of the RI isoform of *gw* is exon 1. Scroll down to the exon table and click on the first row (FlyBase ID = 1). A window will open that contains the sequence of exon 1.
5. Copy and paste the exon 1 sequence for gw-RIinto the “Enter Query Sequence” text box
6. The *gw* gene is encoded on the F element in *D. biarmipes* so to navigate to this sequence, simply type “**contig38**” into the “**enter position or search terms**” text box and click the “**go**” button.
7. Retrieve the DNA sequence of this contig by clicking “**View**” in the navigation bar at the top of the page and then clicking on “**DNA**”. Copy the entire sequence and paste it into the “Enter Subject Sequence” text box.
8. Reconfigure the BLAST settings as follows:
   * Program Selection: **Somewhat similar sequences (blastn)**
   * Under Algorithm parameters:
     + Word size = **7**
     + Match/Mismatch Scores = **1,-1**
     + Gap Costs = **Existence:2 Extension:1**
     + Filter = **Uncheck Low complexity regions**
9. Check “Show results in a new window” and click “BLAST”



**Figure 9: The best BLASTn alignment result between *D. melanogaster* gw-RI exon 1 (query) and *D. biarmipes* contig38 (subject).**

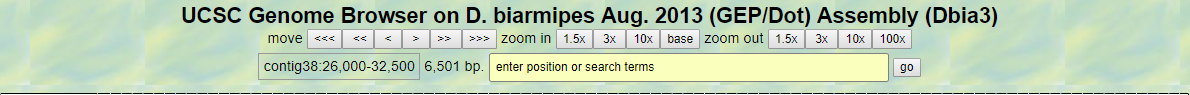
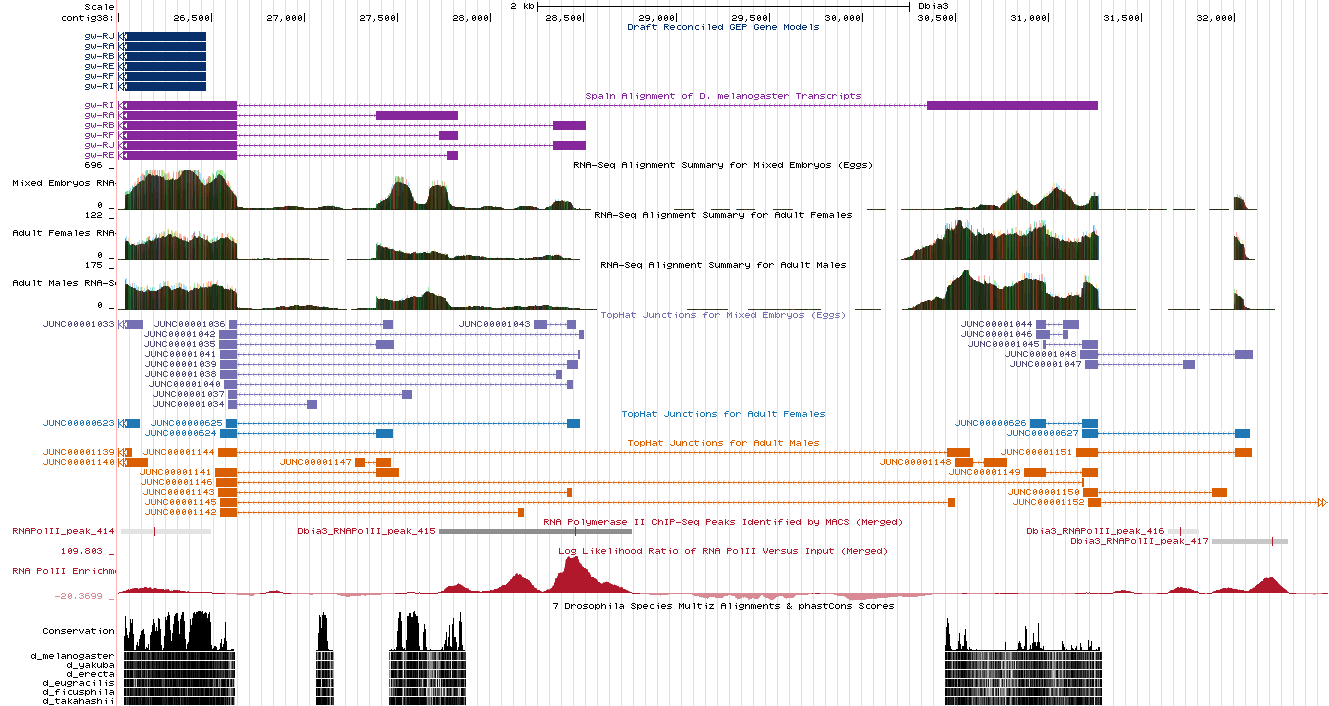
**Q10. Was BLASTn able to locate the gw-RI exon 1 in the *D. biarmipes* contig38 project sequence? Explain why or why not.**

**!!SPOILER ALERT!!**

So (as you might have guessed, because otherwise why would we be doing this particular gene!), because BLASTn fails to localize the 1st transcribed *D. melanogaster* exon for the RI isoform of the *gw* gene in the *D. biarmipes* contig38 project sequence we will have to use other pieces of evidence to determine a TSS search region for this gene.

**!!SPOILER ALERT!!**

1. Since we configured the browser to display the tracks we will need to identify TSS search region boundaries in the previous exercise, and the UCSC browser remembers our settings, you should not need to make any changes. However, if you changed any of the browser settings reconfigure them according to the directions in Exercise 2 above.
2. Now let’s investigate the 5’ end of the *gw* gene. First, let’s take a broad overview of the region in order to get ourselves oriented. Enter “**contig38:26,000-32,500**” into the “**enter position or search terms**” and click the “**go**” button.



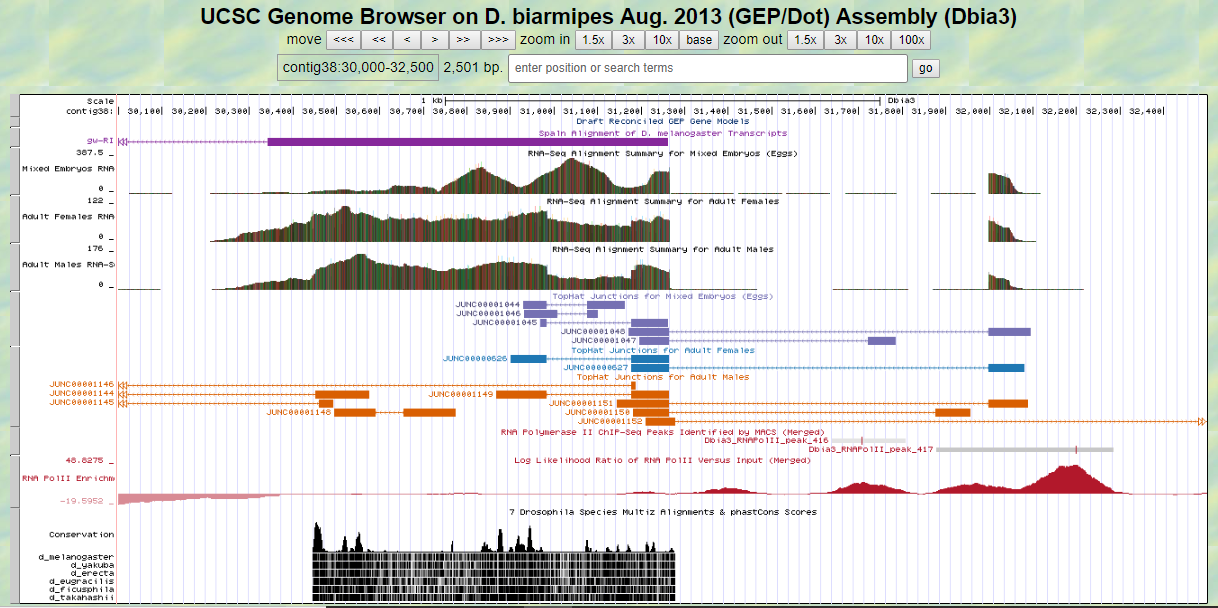
**Figure 10: Genomic region surrounding the 5’ end of the *gw* genes on *D. biarmipes* contig38 with the tracks needed to identify a TSS search region displayed.**

Now we can see displayed the beginnings of the coding DNA sequence (CDS) of the *gw* isoforms (blue tracks at the top) along with the Spaln2 alignment of *D. melanogaster* genes (purple tracks) (**Fig. 10**). Spaln2 is a program that aligns the RNA sequences of many genes from a species (in this case *D. melanogaster*) against the genome of another (in this case the GEP *D. biarmipes* project sequences). The Spaln alignments can help us zero in on a region that may contain the TSS for a gene, but as mentioned above should not be used to set the boundaries of a search region.

**Q11. At approximately, which position does the coding DNA sequence (CDS) of all of the *gw* isoform begin?**

**Q12. According to the Spaln alignment, what are the approximate coordinates of the furthest upstream exon of the RI isoform of the *gw* gene (gw-RI)?**

1. Using the Spaln alignment to guide us, let’s zoom in to that region further upstream. Enter “**contig38:30,000-32,500**” into the “**enter position or search terms**” and click the “**go**” button.

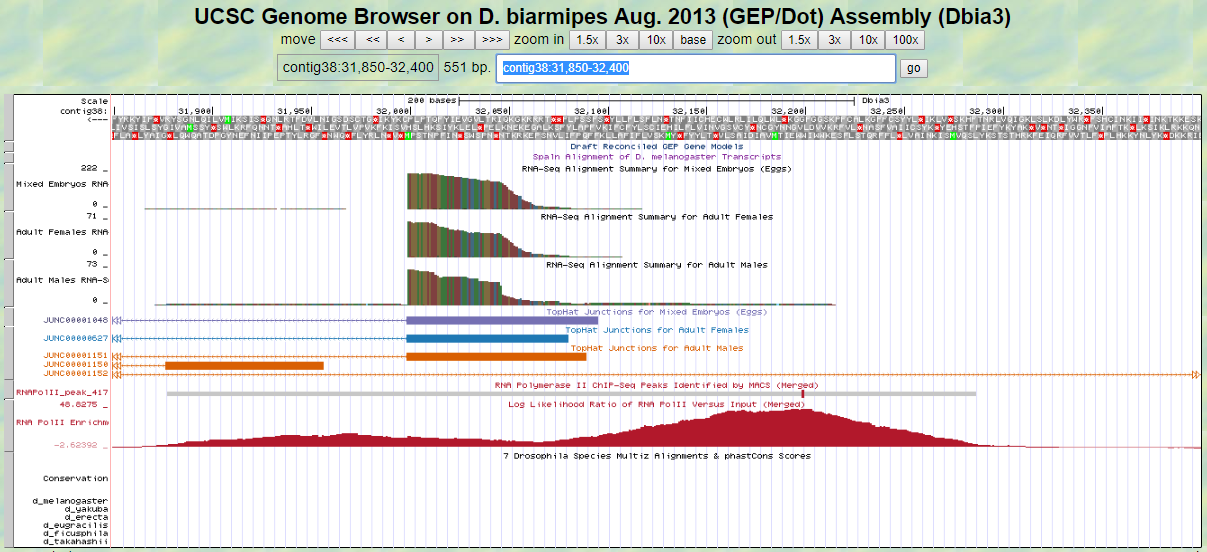


**Figure 11: Browser window showing the region surrounding the furthest upstream Spaln alignment of gw-RI.**

Now we can clearly see the region surrounding the furthest upstream Spaln alignment to the gw-RI isoform.

**Q13. Can we conclude that the 5’ end of the first gw-RI exon is located near the Spaln alignment or is other evidence present that would lead us to conclude that the first exon is actually in a different position. Explain.**

1. Next, zoom into the upstream region by entering “**contig38:31,850-32,400**” into the “**enter position or search terms**” and click the “**go**” button.



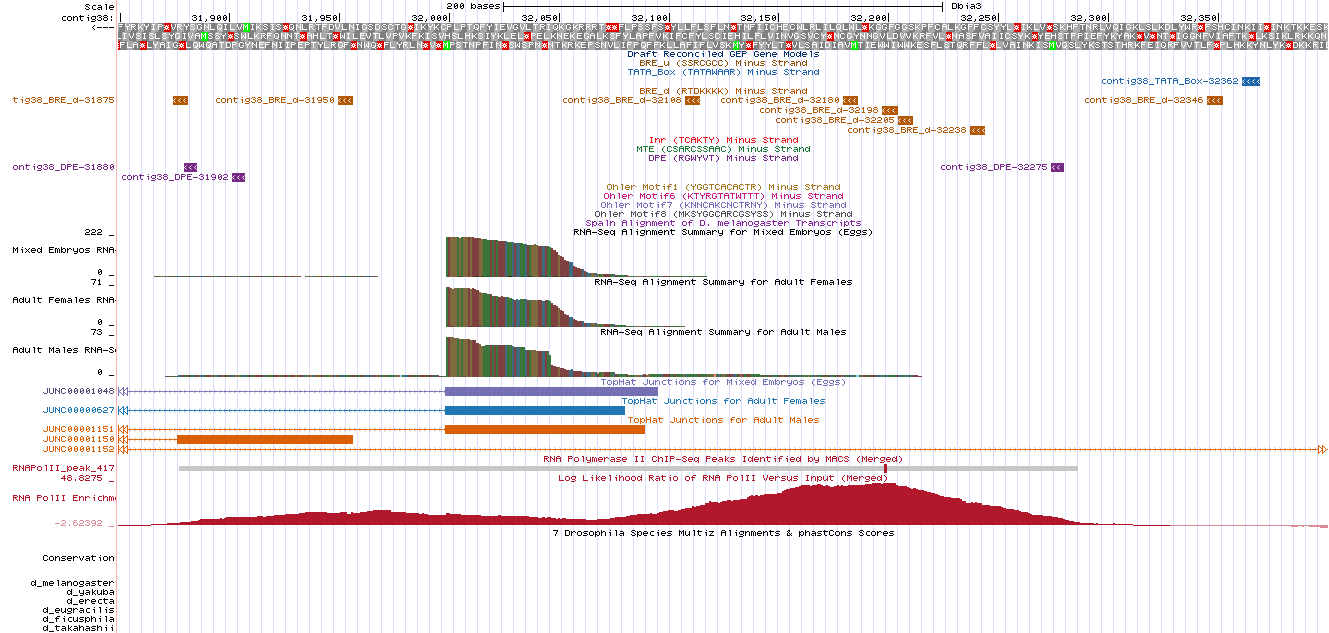
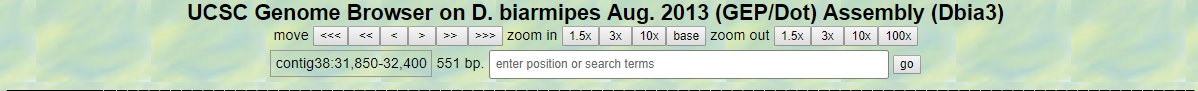
**Figure 12: Browser window showing the region surrounding the first transcribed exon of the *D. biarmipes* gw-RI isoform.**

**Q14. Based on the data available in this region, designate a narrow TSS search region for the *D. biarmipes* gw-RIisoform. Describe the landmarks you used to determine these search regions.** (Use the browser to zoom in and out to identify the precise boundaries to the nucleotide.)

**Q15. For the gw-RI isoform would it make sense to use the entire RNA polII peak to designate a wide TSS search region. Explain why or why not.**

1. Finally, a nice feature of the *D. biarmipes* projects is that all of the consensus promoter motifs have been identified and assembled into an individual track that can be displayed in the browser. Therefore, we can visualize all of these motifs in a single step, without needing to search for each of them individually using Short Match. There are individual tracks for the motifs found on either the plus or minus strand. Since the *gw* gene is encoded on the minus strand of contig38 we will only display the motifs on that strand. Bring up this track by configuring the browser display modes as follows:
   * Under “Mapping and Sequence Tracks”
     + Core Promoter Motifs (minus): **pack**
   * Click “**refresh**”

**Figure 13: Browser window for the region surrounding the first transcribed exon of the *D. biarmipes* gw-RI isoform showing all consensus promoter motif on the minus strand.**



**Q16. List all of the consensus promoter motifs found in the narrow TSS search region you identified in the previous question (include the start position of each motif).**

Conclusion

In this module we have learned how to utilize different lines of evidence to annotate a TSS search region for a gene with a broad promoter, particularly when the BLASTn search fails to locate the 1st transcribed *D. melanogaster* exon in your project sequence. The strategy presented here utilizes the evidence provided by the RNA polII X-ChIP-seq data, the RNA-seq and TopHat data and the Conservation data to rationally define the boundaries of a TSS search region. This strategy should be applicable to a wide range of GEP project TSS annotation problems.

If you would like further practice annotating this type of TSS, try doing *CG33941* in either *D. biarmipes* or *D. elegans* projects.