MODULE TSS1:

TRANSCRIPTION START SITES INTRODUCTION (BASIC)

JAMIE SIDERS, MEG LAAKSO & WILSON LEUNG

Lesson Plan:

Title	**Identifying transcription start sites for Peaked promoters using chromatin landscape, promoter motifs and Celniker.**
Objectives	• Review the TSS Annotation workflow • Explain how the 9-state track, DNase I hypersensitive sites and TSS (Celniker) data can help to understand the chromatin landscape surrounding a gene and identify transcription start sites • Classify the promoter for the *Antp* gene as peaked, broad or intermediate
Pre-requisites	• TSS Module Primer - *Optional* • Review of Module 2 – Transcription Part I can be helpful.
Order	• Understanding histone modifications and DNase I hypersensitive sites • Review of promoter structure, promoter motifs, and consensus sequences • Explanation of methods and data used to produce the Celniker TSS predictions
Homework	• None
Class Instruction	• Discuss the questions: How can our understanding of transcription and promoter structure be used to find a TSS? • Work through the Genome Browser examples ○ Conclude by challenging students to identify the TSS of the *Antp* gene
Associated Videos	• None

GOALS FOR THE TSS MODULES

The four TSS modules will introduce you to the methods by which a core promoter can be classified in *D. melanogaster*, and a TSS position and/or TSS search region can be defined in the target species. The main objectives of the four modules are as follows:

Module 1 will introduce you to basic skills that will enable you to identify a single TSS in a Peaked promoter for the Antennapedia gene in Drosophila melanogaster.

Module 2 will analyze a sequence alignment between exon 1 of the sevenless (sev) gene in D. melanogaster and a D. eugracilis scaffold in order to identify the best TSS for the sev ortholog in D. eugracilis.

Module 3 will cover the use of additional evidence tracks (CAGE, RAMPAGE, RNA Pol II X-ChIP-Seq, and multiple sequence alignments) to identify one or more TSSs within a promoter.

Module 4 will provide examples of Broad promoters, and show how they are different from Peaked promoters.

COMPARATIVE ANNOTATION OF TRANSCRIPTION START SITES BY THE GENOMICS EDUCATION PARTNERSHIP

The fourth chromosome of *D. melanogaster* is unusual in that its DNA is both highly repetitive and tightly packaged in comparison to the other autosomes. Despite this high level of DNA compaction, fourth chromosome genes are expressed at levels similar to genes on the euchromatic regions of the other autosomes.

One possible explanation for the expression levels of fourth chromosome genes could be the presence of unique motifs in the promoters of these genes. The Genomics Education Partnership (GEP) is doing research that compares the promoter regions of fourth chromosome genes with genes located on the third chromosome. Through comparison, it might be possible to identify motifs for fourth chromosome genes that are distinct from other genes in the *Drosophila* genome. The first step toward this goal is to manually annotate the transcription start sites for fourth chromosome genes in a group of closely related *Drosophila* species. It is the goal of the TSS modules 1-4 to assist in understanding the process and evidence used in the annotation of a TSS.

The GEP TSS annotation strategy is predicated on parsimony with *D. melanogaster*. The annotation strategy shown in **Figure 1** consists of two steps: 1.) characterize the promoter in the *D. melanogaster* ortholog, and 2.) annotate the TSS position or TSS search regions in the target species (*e.g.*, *D. eugracilis*).

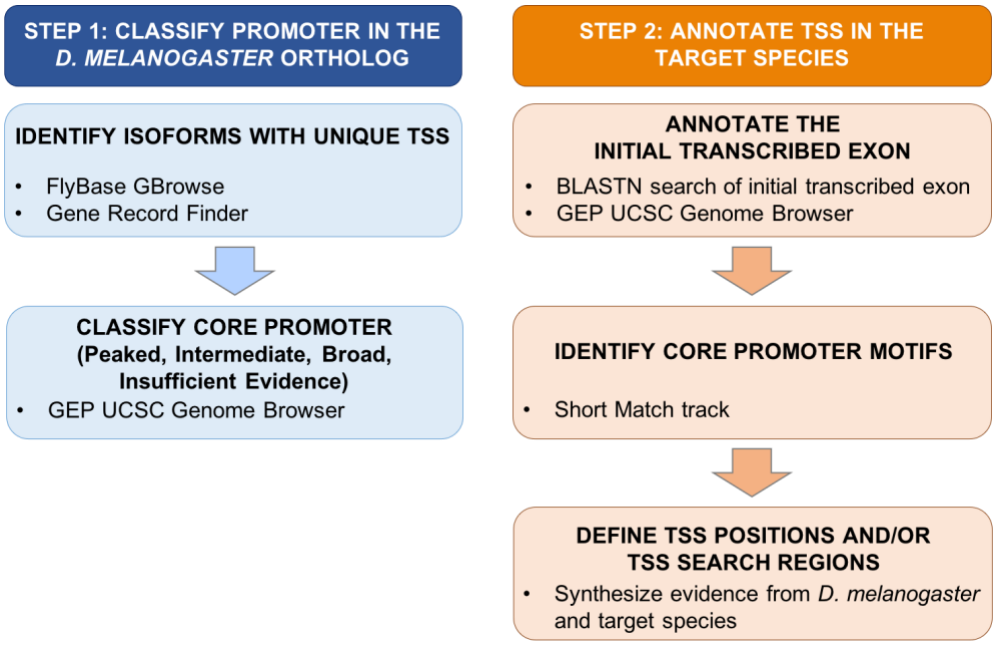

Figure 1 The TSS annotation workflow

INVESTIGATION OF THE ANTENNAPEDIA GENE IN DROSOPHILA MELANOGASTER

This module will utilize the *Antennapedia* (*Antp*) gene in *D. melanogaster* to familiarize you with novel tracks in the Genomics Education Partnership (GEP) UCSC Genome Browser, including the 9-state track, the DHS tracks and TSS (Celniker) tracks. These tracks can aid you in investigating the promoter region of a gene of interest. At the end of the module you will learn how promoters can be classified into one of three possible categories: Peaked, Broad or Intermediate.

EXERCISE 1: USING THE GENOME BROWSER TO INVESTIGATE THE *ANTENNAPEDIA* GENE AND ITS ISOFORMS

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.

2. Once you are at the GEP UCSC Genome Browser, click on the "Genome Browser" link on the left sidebar (**Figure 2**).

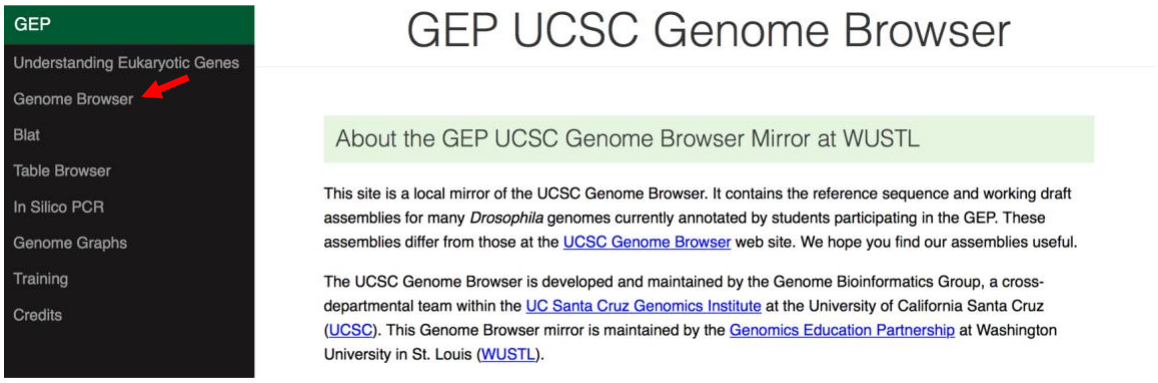

Figure 2 Click on the "Genome Browser" link on the GEP UCSC Genome Browser home page to access the Genome Browser Gateway page.

3. Select "**D. melanogaster**" under "Represented Species", "**Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6)**" under "D. melanogaster Assembly", enter "**Antp**" into the "Position/Search Term" field, and then click on the "**GO**" button (**Figure 3**).

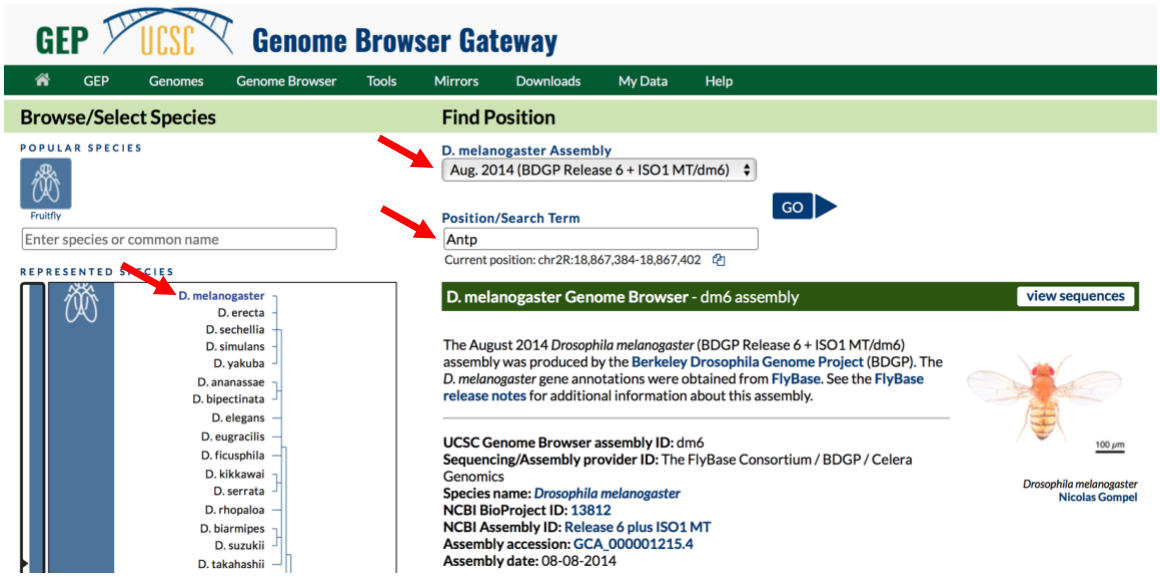

Figure 3 Specify the genome, assembly, and search term on the Genome Browser Gateway page.

4. Click on the link to the E isoform of *Antp* (**Antp-RE**) in the resulting window. The Genome Browser now shows the genomic location of the *Antp* gene. Using the evidence tracks in the Genome Browser, answer the questions below.

Q1. What is the genomic location of the *Antp* gene in the *D. melanogaster* genome?

Q2. How many isoforms does the *Antp* gene have in *D. melanogaster*? List the isoforms below.

Q3. Is the *Antp* gene on the plus or minus strand?

Q4. Upon analyzing the different isoforms of the *Antp* gene, do you expect that all isoforms will utilize the same TSS? Explain.

Q5. Open a new tab on your browser. Navigate to FlyBase (http://flybase.org/) and type "Antp" into the "Jump to Gene" text box to obtain more information about the *Antp* gene.

What is the full name of the *Antp* gene?

According to the "Gene Snapshot" section of the *Antp* gene record, what are the biological functions of the *Antp* gene?

Based on the description of the *Antp* gene in FlyBase, do you expect this gene to be expressed ubiquitously throughout development, or expressed only in specific tissues and developmental stages?

5. Navigate back to the *Antp* gene on the GEP UCSC Genome Browser. The *Antp* gene is located on the minus strand. To make it easier to interpret the evidence tracks, we will reverse complement the entire chromosome sequence. Click on the **"reverse"** button located in the display controls below the Genome Browser image (**Figure 4**, red arrow).

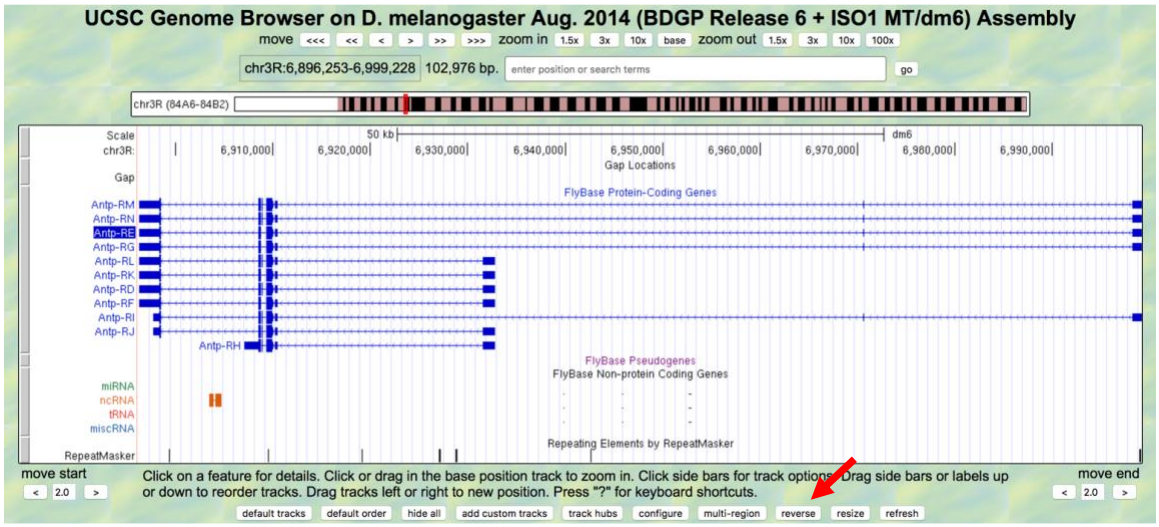

Figure 4 Screenshot of *Antp* gene from *D. melanogaster.* Click on the "reverse" button to reverse complement the entire chr3R sequence.

EXERCISE 2: INVESTIGATION OF THE 9-STATE TRACKS

Recent work by the modENCODE project has characterized the "chromatin landscape" of the *Drosophila* genome in two cell lines: the embryonic cell line S2 and the neuronal cell line BG3. This work classified chromatin into 9 "states" (**Figure 5**).

The 9-state Model provides insights into whether a particular gene is more, or less, likely to be transcribed. For example, state 1 is associated with active promoters while states 7 and 8 are associated with heterochromatin.

State	Description	Color
1	Active promoter/transcription start site region	
2	Actively transcribed exon	
3	Actively transcribed intron (enhancer)	
4	Other open chromatin	
5	Actively transcribed exon on the male X chromosome (dosage compensation)	
6	Region of Polycomb-mediated repression	
7	Heterochromatin	
8	Heterochromatin-like region embedded in euchromatin	
9	Transcriptionally silent intergenic euchromatin	

Figure 5 The 9-state Model produced by modENCODE project, which summarizes the epigenomic landscape of the *D. melanogaster* genome in S2 and BG3 cells. Each state is assigned a different color in the 9-state evidence tracks on the GEP UCSC Genome Browser for *D. melanogaster*. The 9-state model data evidence tracks can be found in the Genome Browser under the 'Chromatin Domains' display heading as BG3 9-state (R5) and S2 9-state (R5), where R5 refers to Release 5.

One state which may be unfamiliar is state 6, associated with genes that are regulated by Polycomb-group (PcG) proteins. PcG proteins act to silence gene expression via a variety of mechanisms, including methylation of histone proteins. Genes whose expression is regulated by PcG proteins tend to be genes that are associated with development (*e.g.*, homeotic genes), or genes that show tissue-specific expression.

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.

   Configure the display modes for your browser as follows:

   - Under "Mapping and Sequencing Tracks"
     - Base Position: **full**
   - Under **"**Chromatin Domains**"**
     - BG3 9-state (R5): **dense**
     - S2 9-state (R5): **dense**
   - Under "Genes and Gene Prediction Tracks"
     - FlyBase Genes: **pack**

2. Zoom out 3x, then 1.5x (**Figure 6**).

   Q6. What does the data in the BG3 and S2 9-state tracks tell you about the chromatin landscape in the region of chromosome 3R in which the *Antp* gene is located?

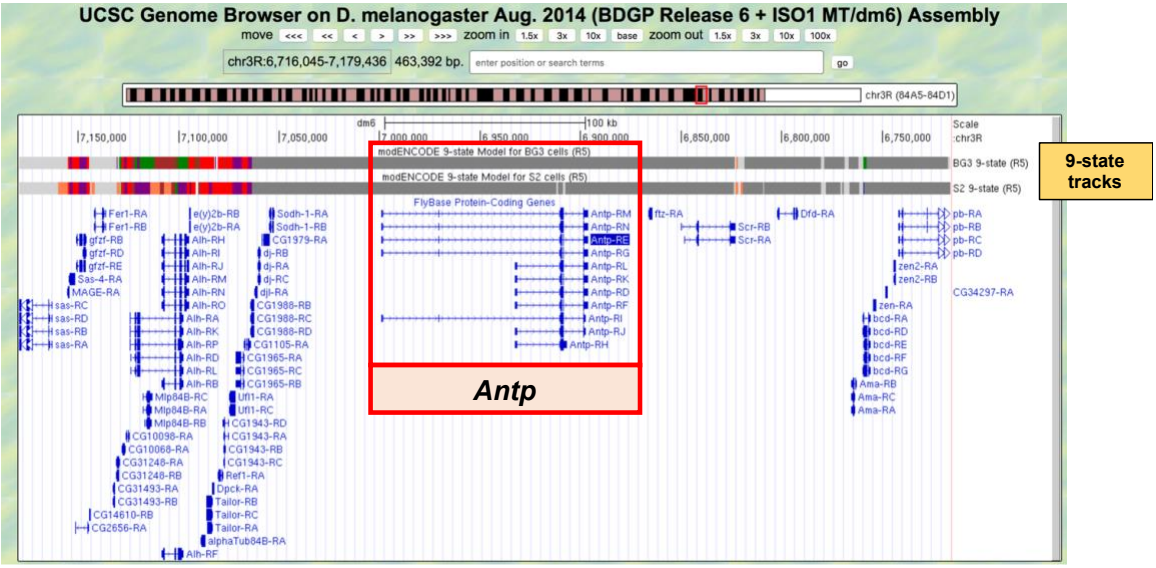

Figure 6 The GEP UCSC Genome Browser view of the *Antp* gene and the 9-state tracks for BG3 and S2 cells (top) and the reverse complement of the genomic region surrounding the *Antp* gene on chr3R after zooming out 4.5x (bottom).

Q7. Is the chromatin landscape compatible with what you found out regarding the function(s) of the *Antp* gene?

SUMMARY: Initial analysis of the *Antp* gene in *D. melanogaster* using FlyBase and the GEP UCSC Genome Browser tells us that that the *Antp* gene codes for a transcription factor that plays an important role in patterning the body plan along the anterior/posterior axis. The 9-state model suggests that *Antp* is inactive in S2 and BG3 cells due to Polycomb-mediated gene silencing. Although the 9-state model suggests that *Antp* is located in an "inactive" region of the genome, the *Antp* gene could still be actively transcribed in these cell lines. We will look at additional lines of evidence to determine whether this region of the genome is transcriptionally active.

DNase I Hypersensitive Sites (DHS)

Regions of the genome with "loosely packed" chromatin are sensitive to cleavage by an enzyme known as DNase I. DNase I is an endonuclease, meaning it is able to cleave covalent bonds in the backbone of double stranded DNA. DNase I can only cut 'free' DNA, that is, DNA that is loosely packed. Note that areas with loosely packed DNA would not only be accessible to DNase I, but also to proteins involved in regulating

transcription. *Therefore regions sensitive to cleavage by DNase I also tend to be regions that are enriched in transcription factor binding sites and transcription start sites*.

Regions of chromatin that are hypersensitive to cleavage by DNase I are referred to as **D**Nase I **H**ypersensitive **S**ites (DHS). These regions can be identified using tracks on the UCSC Genome Browser under the 'Expression and Regulation' heading.

There are two types of DHS tracks that can be utilized as pieces of evidence in the process of identifying the promoter region for a gene: the "DHS Read Density" track and the "DHS Positions" track.

> **DHS Read Density**: Displays relative DNase I sensitivity for an area of the genome being investigated in comparison to other areas of the genome.
>
> **DHS Position**: Shows specific ***genomic positions*** that have significantly higher sensitivity to DNase I than other sites in the genome. This track gives a discrete location of a 'high-magnitude' DHS site.

EXERCISE 3: INVESTIGATION OF THE DHS TRACKS

Let's analyze the DHS evidence tracks for the *Antp* gene.

1. Return to the web browser tab with the GEP UCSC Genome Browser and add the following evidence tracks:
    - Under "Expression and Regulation"
        - Detected DHS Positions (Cell Lines) (R5): **dense**
        - DHS Read Density (Cell Lines) (R5): **full**
2. Click on the "**refresh**" button
3. Enter "**chr3R:6,996,689-7,002,286**" into the "enter position or search terms" text box, and then click on the "**go**" button to navigate to the region surrounding the TSS of the E, G, I, M, and N isoforms of *Antp*.

The Genome Browser now shows both the DHS Read Density and Detected DHS Positions track (**Figure 7**).

> **Q8. How many distinct DHS Positions are located in this region? Are the DHS Read Density and DHS Positions consistent across the three cell lines? Are the DHS Positions in agreement with the FlyBase annotations for the M, N, E, G and I isoforms of *Antp*?**

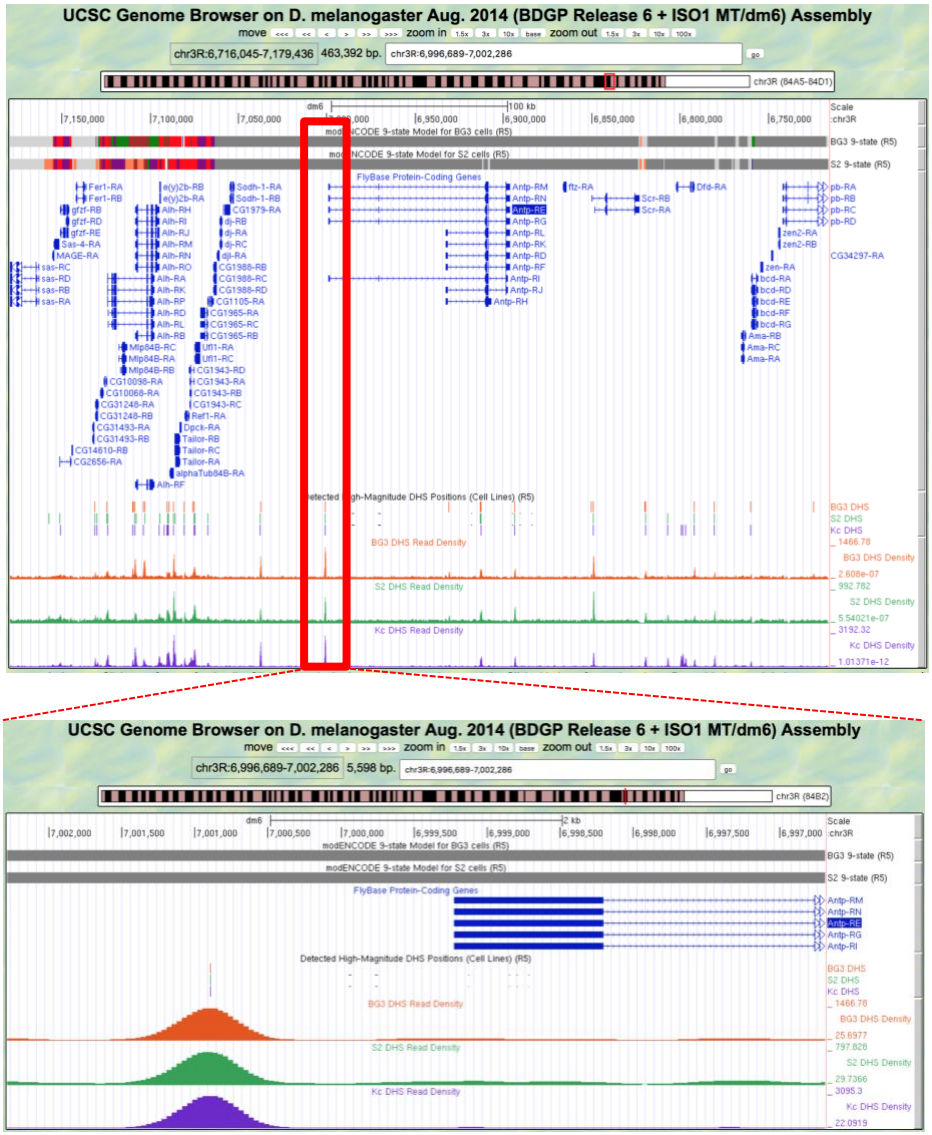

Figure 7 DHS Positions and DHS Read Density tracks on the GEP UCSC Genome Browser. (Top) Genome Browser view of the DHS tracks for the genomic region surrounding the *Antp* gene. (Bottom) Genome Browser view of the DHS tracks for the region surrounding the TSS of the M, N, E, G and I isoforms of *Antp*.

Enter "**chr3R:6,928,927-6,934,891**" into the "enter position or search terms" text box of the GEP UCSC Genome Browser to view the 5' UTR of the remaining isoforms of *Antp*.

Q9. What does the DHS Read Density and Detected DHS positions suggest about the genomic region just upstream of the TSS for the D, F, H, J, K, and L isoforms of *Antp*?

SUMMARY: The DHS Read Density and DHS positions tracks can provide further insights into the chromatin landscape in a genomic area of interest by defining the accessible regions of chromatin. Analysis of the DHS tracks show a DHS site located just upstream of the 5' UTR for the M, N, E, G and I isoforms of *Antp*. Hence this genomic region has relatively low nucleosome density and might contain transcription factor binding sites that facilitate the expression of the M, N, E, G and I isoforms of *Antp*.

USING TSS CELNIKER DATA AS EVIDENCE FOR TSS ANNOTATIONS

The modENCODE project performed a study to systematically characterize the promoter motifs and transcription start sites on a genome-wide basis in *D. melanogaster* embryos; this endeavor was lead by a research scientist by the name of Susan Celniker. The analyses were extensive and used multiple experimental methods to determine the number and the distribution of TSSs in the promoters of genes during embryonic development.

The data obtained from the Celniker and colleagues indicated that *although many promoters in Drosophila have a single TSS, the majority of genes have promoters that allow for transcription initiation at multiple sites — which means their promoters are classified as either Intermediate or Broad*. Broad promoters are also prevalent in other species, including in the human genome. The data also indicated that different TSSs found within a broad promoter could be used preferentially in different developmental stages or tissues.

TSS data obtained by Celniker and colleagues are available via the "TSS (Celniker) (R5)" track on the GEP UCSC Genome Browser (under the "Expression and Regulation" section).

EXERCISE 4: ANALYSIS OF THE CELNIKER TSS MODENCODE DATA

In this exercise, we will use the TSS (Celniker) data to further support the TSS annotation for the longer isoforms of *Antp* (i.e. isoforms E, G, I, M, and N).

1. Navigate back to the web browser tab with the GEP UCSC Genome Browser. Reconfigure your browser as follows:
    - Base Position: **full,** BG3 9-state (R5): **dense,** S2 9-state (R5): **dense,** FlyBase Genes: **pack,** Detected DHS Positions (Cell Lines) (R5): **dense,** DHS Read Density (Cell Lines) (R5): **full**
    - Short Match: **hide**

- **Under "Expression and Regulation"**
    - TSS (Celniker) (R5): **dense**

2. Enter "**chr3R:6,999,172-6,999,267**" into the "enter position or search terms" text box to navigate to the start of the FlyBase annotations for the longer isoforms of *Antp* (**Figure 8**).

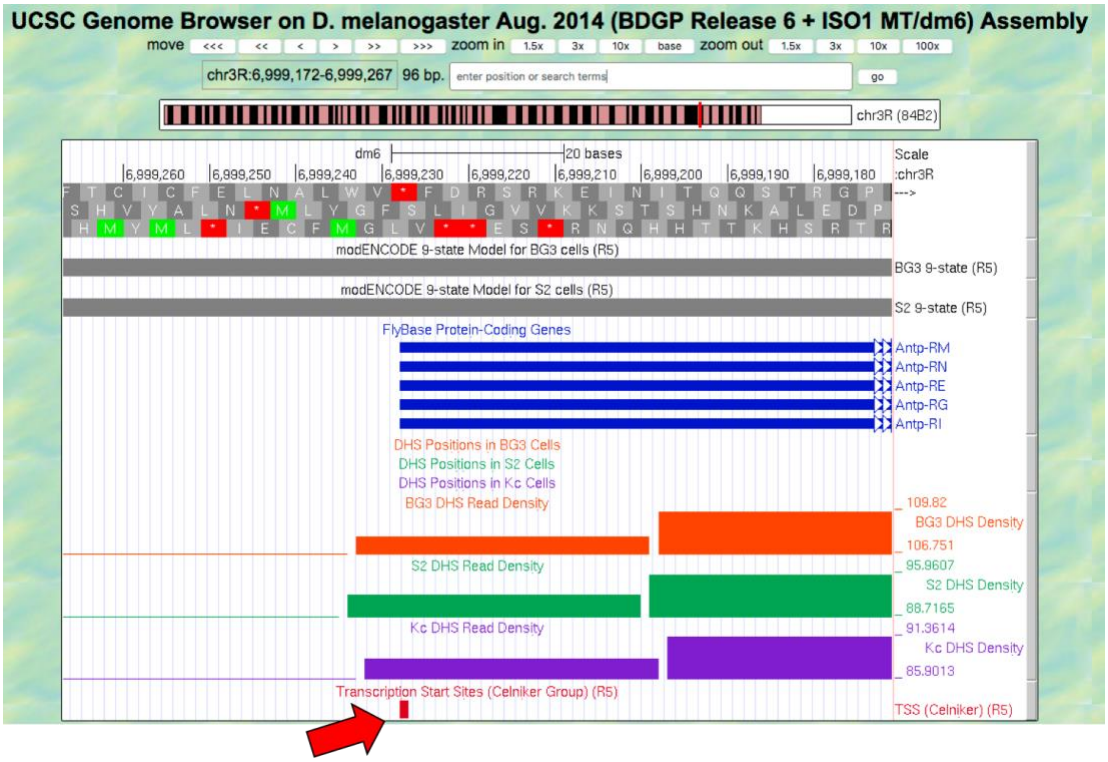

Figure 8 The Celniker TSS prediction (red arrow) in the genomic region surrounding the start of the E, G, I, M, and N isoforms of *Antp*.

Q10. Analyze the "FlyBase Genes" and the "TSS (Celniker) (R5)" data tracks. At which coordinate does the "TSS (Celniker) (R5)" data predict as the TSS?

CLASSIFYING A PROMOTER

So far, we have investigated the chromatin landscape derived from the 9-State Model, DHS data, and the modENCODE Celniker TSS predictions for the *Antp* gene in *D. melanogaster*. All of this information can be used as evidence for the annotation of a transcription start site (although, as you will see in Modules 2 and 3, there is a plethora of other evidence that can also be utilized). The exercises in this module have provided evidence for the annotation of a single TSS at position 6,999,228 of chromosome 3R for the longer isoforms of *Antp* in *D. melanogaster* (i.e., isoforms E, G, I, M, and N).

Before annotating the TSS for a gene in a target species (e.g., *D. eugracilis*), one must first classify the core promoter of the *D. melanogaster* ortholog as Peaked, Intermediate, Broad, or Insufficient Evidence.

For the GEP TSS annotation projects, each core promoter is classified into one of four categories based on the number of "TSS (Celniker) (R5)" annotations and the number of DHS positions within a 300bp window (**Table 1**).

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

Q11. Based on the criteria listed in Table 3 above, how would you classify the promoter for the longer isoforms of *Antp*?

Q12. Perform a similar analysis on the shorter isoforms of *Antp* (i.e. isoforms D, F, H, J, K, L). What is the putative TSS annotation and how would you classify the promoter for these isoforms?

SUMMARY: Analysis of promoter regions and classification of the shape of the core promoter can be challenging. Whereas coding region gene annotations have well-defined rules to guide the determination of exon coordinates (open reading frames, splice sites, etc.), TSS annotations are more ambiguous. This ambiguity could be attributed to our limited understanding of the factors that regulate transcription initiation, and limitations of the current experimental techniques for determining the TSSs. However, part of the ambiguity may also reflect the biology, where the organism can utilize multiple transcription start sites within a promoter. Promoters can vary widely between genes, and different isoforms from the same gene could also have different types of promoters. Promoter diversity makes sense when one thinks about the fact that each gene must be expressed at a specific development stage, in a particular cell type, and at a specific level.