## Module TSS1:

### Transcription Start Sites

**Introduction**

*Jamie Siders, Meg Laakso & Wilson Leung*

---

**Lesson Plan:**

<table>
<thead>
<tr>
<th>Title</th>
<th>Identifying transcription start sites for Peaked promoters using chromatin landscape, promoter motifs and Celniker.</th>
</tr>
</thead>
</table>
| Objectives | • Review the TSS Annotation Workflow  
• Explain how TSS (Celniker) data, DNase I hypersensitive sites, and the 9-state tracks 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 *Understanding Eukaryotic Genes Module 2* can be helpful. |
| Order | • Explanation of methods and data used to produce the Celniker TSS predictions  
• Description of DNase I hypersensitive sites and chromatin landscape  
• Review of promoter classification |
| 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  
  o Conclude by challenging students to identify the TSSs 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 transcription start site (TSS) position and/or TSS search region can be defined in the target species. The main objectives of the three 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 (Antp) gene in Drosophila melanogaster.**

Module 2 will analyze a sequence alignment between exon 1 of the Antennapedia gene, isoform M, in *D. melanogaster* and a *D. eugracilis* scaffold in order to identify the best TSS for the Antp-RM 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, in multiple *Drosophila* species. 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 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 region in the target species (e.g., *D. eugracilis*).
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 TSS (Celniker) track, DHS tracks, and 9-state 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 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 annotation of multiple Drosophila species.

2. Once you are at the GEP UCSC Genome Browser home page, click on the “Genome Browser” link on the left sidebar (Figure 2).
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).

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 isoform E 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. (Note that you can confirm your answer using the GEP Gene Record Finder by typing "Antp" into the Search box. http://gander.wustl.edu/%7ewilson/dmelgenerecord/index.html)

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

Q4. 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?

Q5. Scroll down to the "Genomic Maps" section and click on the "GBrowse" button. Upon analyzing the different isoforms of the Antp gene, do you expect that all isoforms will utilize the same TSS? Explain.

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).

![Image of Genome Browser with red arrow](image)

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

**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 led by research scientist Susan Celniker.\(^1\) 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 2: 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. Click on the “hide all” button located below the Genome Browser image, then configure your browser as follows:
   - Under “Expression and Regulation”
     - TSS (Celniker) (R5): dense
     - Click "refresh"

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 5).

![Image of Genome Browser with TSS prediction](image)

Figure 5  The Celniker TSS prediction (red arrow) in the genomic region surrounding the start of the E, G, I, M, and N isoforms of Antp.

Q6. Analyze the “FlyBase Genes” and the “TSS (Celniker) (R5)” data tracks. Which coordinate does the “TSS (Celniker) (R5)” data predict is the TSS?
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 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.²

Regions of chromatin that are hypersensitive to cleavage by DNase I are referred to as DNase I Hypersensitive Sites (DHS). These regions can be identified using tracks on the GEP 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 a histogram of 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 were experimentally determined to have significantly higher sensitivity to DNase I than genomic background. This track gives a discrete location of ‘high-magnitude’ DHS sites.

**EXERCISE 3: INVESTIGATION OF THE DHS TRACKS**

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

1. Return to 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 tracks (Figure 6).
Figure 6  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 E, G, I, M and N isoforms of Antp.
Q7. How many DHS Positions and read Density peaks are located in this region? Where is/are the DHS Read Density Peaks located relative to the TSS of *Antp* isoforms E, G, I, M, and N? Do you think this DNase I-sensitive region is part of the core promoter?

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*.

Q8. How many DHS Positions and DHS Read Density peaks are located in this region? Where is/are the DHS Read Density Peaks located relative to the TSS of *Antp* isoforms D, F, H, J, K, and L? Do you think the DNase I-sensitive regions are part of the core promoter for these isoforms?

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 upstream of exon 1 for the M, N, E, G and I isoforms of *Antp*. Hence this genomic region has relatively low nucleosome density. Although the region is distal to the TSS, it might contain transcription factor binding sites that facilitate the expression of the M, N, E, G and I isoforms of *Antp*.

Analysis of the genomic region surrounding the TSS for *Antp* isoforms D, F, H, J, K, and L indicate at least two DHS Read Density peaks. One peak overlaps the TSS for exon 1 of isoforms D, F, H, J, K, and L, suggesting this is where the core promoter is located. The second DHS Read Density peak is located in an intron, and may indicate the location of an enhancer element.
EXERCISE 4: 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 7).²

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.

<table>
<thead>
<tr>
<th>State</th>
<th>Description</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Active promoter/transcription start site region</td>
<td>Red</td>
</tr>
<tr>
<td>2</td>
<td>Actively transcribed exon</td>
<td>Orange</td>
</tr>
<tr>
<td>3</td>
<td>Actively transcribed intron (enhancer)</td>
<td>Purple</td>
</tr>
<tr>
<td>4</td>
<td>Other open chromatin</td>
<td>Brown</td>
</tr>
<tr>
<td>5</td>
<td>Actively transcribed exon on the male X chromosome (dosage compensation)</td>
<td>Green</td>
</tr>
<tr>
<td>6</td>
<td>Region of Polycomb-mediated repression</td>
<td>Black</td>
</tr>
<tr>
<td>7</td>
<td>Heterochromatin</td>
<td>Blue</td>
</tr>
<tr>
<td>8</td>
<td>Heterochromatin-like region embedded in euchromatin</td>
<td>Cyan</td>
</tr>
<tr>
<td>9</td>
<td>Transcriptionally silent intergenic euchromatin</td>
<td>Gray</td>
</tr>
</tbody>
</table>

Figure 7 The 9-state model produced by the 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. The 9-state model 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 that 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. Enter “chr3R:6,716,045-7,179,436” into the “enter position or search terms” text box, and then click on the “go” button (Figure 8).

![UCSC Genome Browser on D. melanogaster Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6) Assembly](image)

**Figure 8** The GEP UCSC Genome Browser view of the *Antp* gene and the 9-state tracks for BG3 and S2 cells (box) and the genomic region surrounding the *Antp* gene on chr3R.

Q9. 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?
Q10. *Antennapedia* encodes a transcription factor that is expressed in the thorax during *Drosophila* development. Why is the gene repressed in S2 and BG3 cells?

SUMMARY: Initial analysis of the *Antp* gene in *D. melanogaster* using FlyBase and the GEP UCSC Genome Browser tells us 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 not expressed 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 in S2 and BG3 cells, the *Antp* gene could still be actively transcribed in other cell lines.

CLASSIFYING A PROMOTER

So far, we have investigated the chromatin landscape derived from modENCODE Celniker TSS predictions, DHS data, and the 9-State model 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 three categories based on the number of “TSS (Celniker) (R5)” annotations and the number of DHS positions within a 300bp window (Table 1). Promoters that have no annotated TSS or DHS positions are denoted "Insufficient Evidence".

Q11. Based on the criteria listed in Table 1 below, how would you classify the promoter for the longer isoforms of *Antp*?
TABLE 1: CLASSIFICATIONS OF DROSOPHILA PROMOTERS FOR THE GEP

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEAKED</td>
<td>• One annotated TSS with no DHS position</td>
</tr>
<tr>
<td></td>
<td>• No annotated TSS with one DHS position</td>
</tr>
<tr>
<td></td>
<td>• One annotated TSS with one DHS position</td>
</tr>
<tr>
<td>INTERMEDIATE</td>
<td>• Zero or one annotated TSS with multiple DHS positions</td>
</tr>
<tr>
<td></td>
<td>• Multiple annotated TSS with zero or one DHS positions</td>
</tr>
<tr>
<td>BROAD</td>
<td>• Multiple annotated TSS with multiple DHS positions</td>
</tr>
<tr>
<td>INSUFFICIENT</td>
<td>• No annotated TSS and no DHS positions</td>
</tr>
</tbody>
</table>

Q12. Perform a similar analysis for the shorter isoforms of Antp (i.e., isoforms D, F, H, J, K, L). What is the coordinate for the putative TSS, 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.

REFERENCES
