Module TSS3: CAGE, RAMPAGE, & RNA Pol II X-ChIP-Seq

Meg Laakso & Wilson Leung

**Title**

* Identifying transcription start sites for peaked promoters using experimental data from sequenced transcripts and RNA polymerase II occupancy

**Objective**

* Use the placement of the Initiator motif; experimental data from CAGE, RAMPAGE, and RNA Pol II X-ChIP-Seq experiments; and TSS predictions from the Celniker group at modENCODE to identify putative transcription start sites (TSSs) and characterize promoter type

**Prerequisites**

* [Module TSS1](https://thegep.org/lessons/jsiders-lesson_with_exercises-tss_module1_introduction/)

**Order**

* Explain the experimental techniques used to generate CAGE, RAMPAGE, and RNA Pol II X-ChIP-Seq data, and the Celniker TSS predictions
* Identify the AntennapediaTSS using CAGE, TSS predictions, RAMPAGE, and RNA Pol II X-ChIP-Seq
* Analyze all of the evidence in order to identify the best TSS and characterize the promoter type

**Class Instruction**

* Discuss the question: How can our understanding of transcription initiation be used to find a TSS?
* Conclude by challenging students to think about these questions:
* For each of the evidence tracks you have investigated, how might a broad promoter differ from a peaked promoter?
* If a gene has multiple isoforms, does it have to have multiple transcription start sites?

**Resources & Tools**

* [RNA-Seq and TopHat Video](https://youtu.be/XD-egRcHYL4)
* [GEP UCSC Genome Browser](https://gander.wustl.edu/)

Table of Contents

[Identify the location of Transcription Start Site using CAGE, RAMPAGE and RNA Pol II X-ChIP-Seq data 2](#_Toc143263863)

[Summary of Cap Analysis of Gene Expression (CAGE) 2](#_Toc143263864)

[Exercise 1: Examining the CAGE Data for the Antennapedia Gene Using the UCSC Genome Browser Mirror 2](#_Toc143263865)

[RNA Annotation and Mapping of Promoters for the Analysis of Gene Expression (RAMPAGE) 5](#_Toc143263866)

[Exercise 2: Examining the RAMPAGE data for the Antennapedia gene 6](#_Toc143263867)

[RNA Pol II ChIP-Seq 9](#_Toc143263868)

[Exercise 3: Examining the X-ChIP-Seq data for the Antennapedia gene using the UCSC Genome Browser Mirror 9](#_Toc143263869)

[Wrapping up 10](#_Toc143263870)

[References 11](#_Toc143263871)

# Identify the location of Transcription Start Site using CAGE, RAMPAGE and RNA Pol II X-ChIP-Seq data

In Module 1 we utilized Celniker data, DNase I hypersensitivity data and the 9-state model track to investigate the transcription start site region for the Antennapedia (*Antp*) gene in *Drosophila melanogaster*. This module will further explore the *Antp* gene in *D. melanogaster* to illustrate how the experimental techniques CAGE, RAMPAGE, and RNA Pol II X-ChIP-Seq can be used to identify or confirm transcription start sites (TSSs) and promoter type.

# Summary of Cap Analysis of Gene Expression (CAGE)

The relative abundance of mRNA transcripts can be determined by many different methods such as RNA-Seq, quantitative PCR, or Serial Analysis of Gene Expression, which you may have studied in class. However, these methods tend to show a bias toward the body of the transcript. Hence, specialized techniques are needed to isolate sequences associated with the 5’ end (the “beginning”) of the transcript.

One technique that can be used to identify a TSS is Cap Analysis of Gene Expression (CAGE).1 This technique takes advantage of the fact that there is a unique nucleotide at the 5’ end of the newly synthesized eukaryotic mRNA, the “5’ cap”. This 5’ cap can be chemically modified, allowing the capped mRNAs to be purified and sequenced.

In Exercise 1, we will use CAGE data to help identify the TSS and cross-reference with TSS predictions produced by the Celniker group at modENCODE to identify the TSS2. The Celniker TSS predictions were based on the analysis of experimental data from three methods: CAGE3, 5’ Rapid amplification of cDNA ends (5’ RACE), and 5’ Expressed Sequence Tags (ESTs).

## Exercise 1: Examining the CAGE Data for the Antennapedia Gene Using the UCSC Genome Browser Mirror

1. Open a new web browser window and go to the Genomics Education Partnership [(GEP) UCSC Genome Browser Mirror](https://gander.wustl.edu/) (Figure 1).

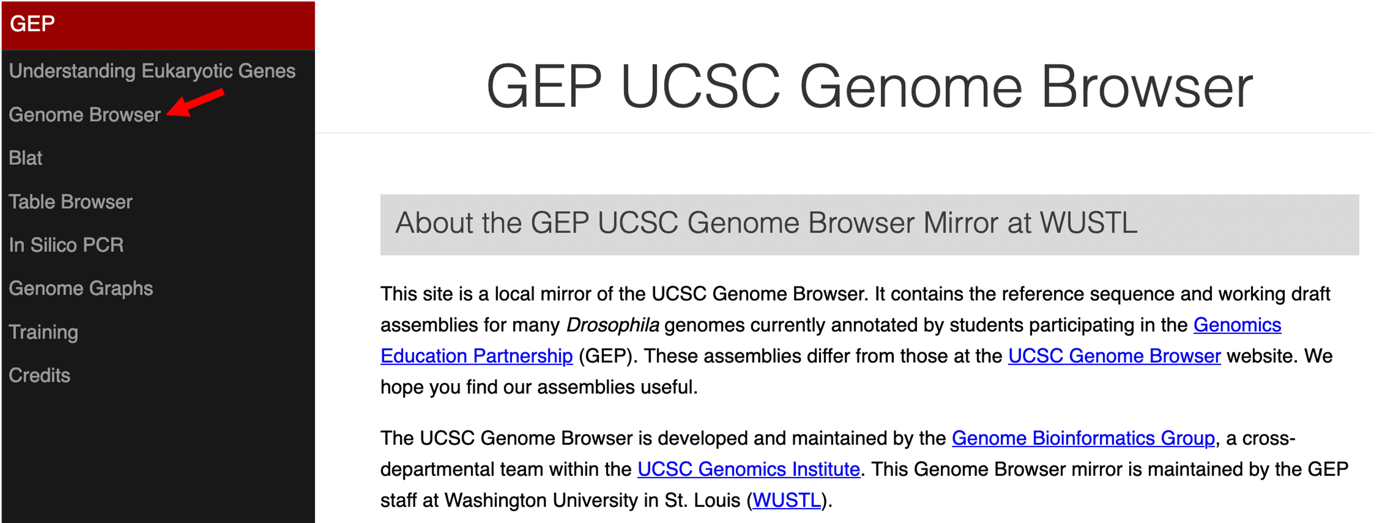


Figure Access the GEP UCSC Genome Browser Gateway page using the “Genome Browser” link.

1. To navigate to the genomic region surrounding the Antennapedia (*Antp*) gene in *Drosophila melanogaster*, select “D. melanogaster” and “Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6)” under the “*D. melanogaster* Assembly” field, and then enter “**Antp**” under the “Position/Search Term” field. Click on the “GO” button (Figure 2).

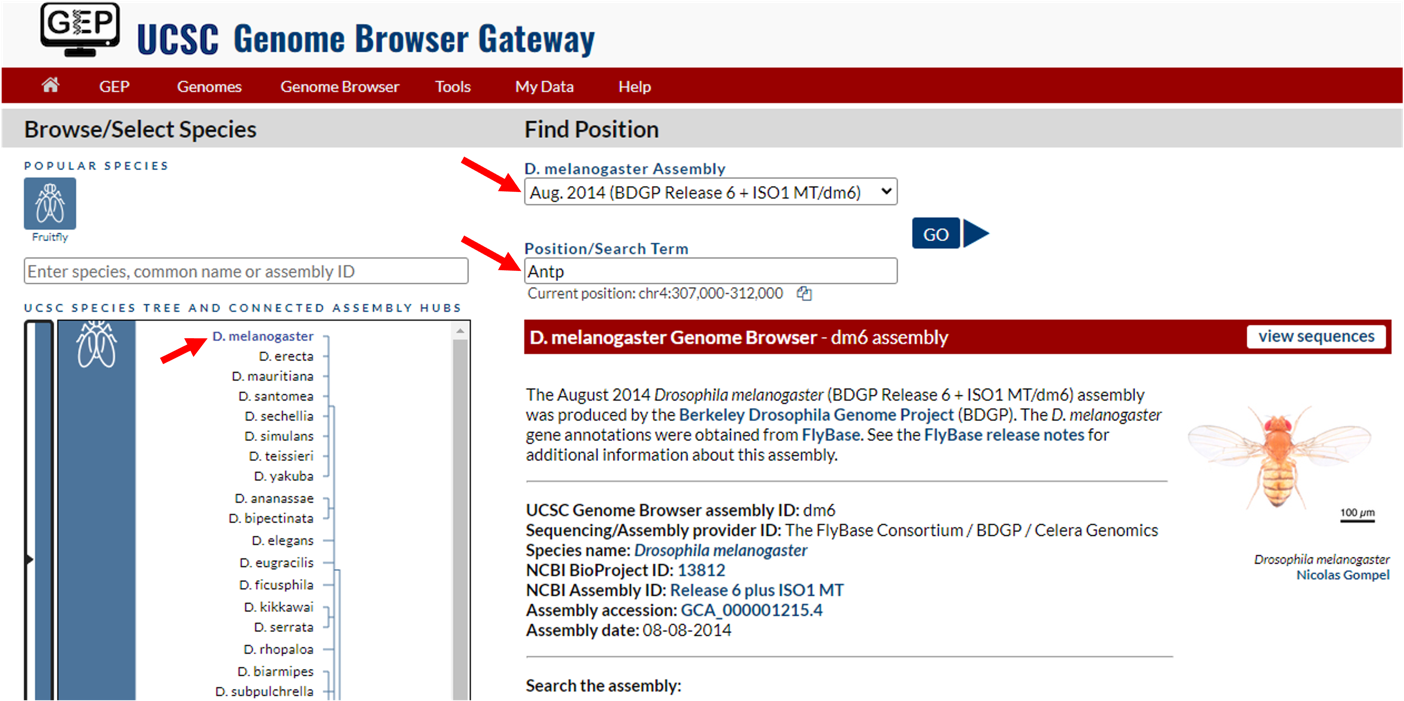


Figure Change the settings on the Genome Browser Gateway page to navigate to the *Antp* gene in the *D. melanogaster* release 6 assembly.

1. Select "Antp-RM at chr3R:6896253-6999228" from the list of FlyBase Protein-Coding Genes to investigate the M isoform of the *Antp* gene.
2. The *Antp* gene is on the minus strand. Click on the “reverse” button located in the display controls below the Genome Browser image.
3. Click on the “hide all” button located below the Genome Browser image. Then, configure the display modes as follows:
4. Under “Mapping and Sequencing Tracks”

* Base Position: “full”
* Click on the blue “Short Match” link
* Enter “**TCAKTY**” (i.e., the Initiator motif) into the “Short (2-30 base) sequence” text box
* Change the “Display mode” to “pack”
* Click on the “Submit” button

1. Under “Genes and Gene Prediction Tracks”

* FlyBase Genes: “pack”

1. Under “Expression and Regulation”

* TSS (Celniker) (R5): “pack”
* Click on the blue “Combined modENCODE CAGE TSS” link
* Change the “Maximum display mode” to “full”
* Scroll down to the “Select views” section, and change the “Peaks” display mode to “pack”
* In the “List subtracks” section, uncheck the box for “modENCODE CAGE (Plus)”
* Scroll up to the top of the page, and then click on the “Submit” button to update the display.

1. Enter “**chr3R:6,999,222-6,999,236**” into the “chromosome range, or search terms, see examples” text box, and then click on the “go” button.

We will examine the region near exon 1 of Antp-RM to determine where transcription begins (Figure 3). This is a non-coding exon and does not contain a start codon for translation.

**Q1. Based on the “FlyBase Protein-Coding Genes” track, what is the coordinate of the transcription start site?**

**Q2. Based on the Transcription Start Sites (Celniker Group) (R5) track, what is the coordinate of the transcription start site?**

**Q3. Is the coordinate of the Transcription Start Sites (Celniker Group) (R5) prediction the same as the Combined modENCODE CAGE Read Density prediction? Remember that the Combined modENCODE CAGE Read Density track shows the density of the first base of the CAGE reads, which corresponds to the 5’ end of the mRNA.**

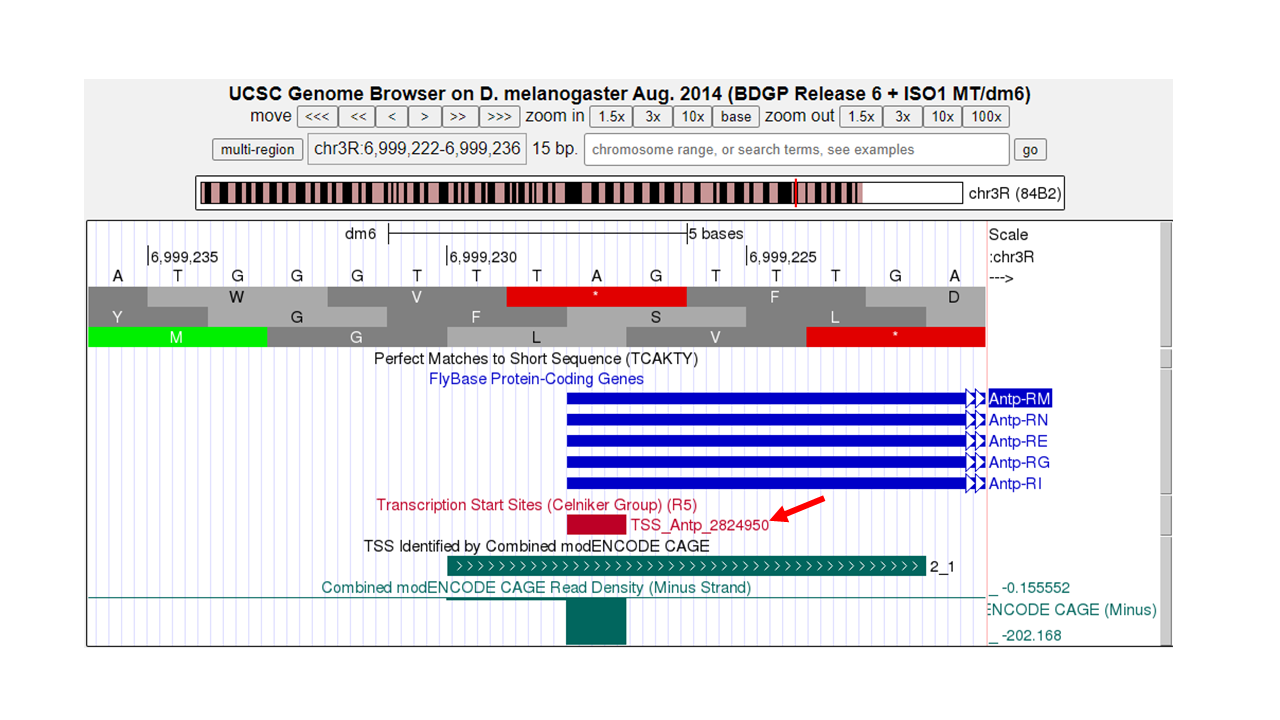


Figure Genome Browser view of the TSS predicted by the Celniker Group (red arrow) and Combined modENCODE Read Density.

**Q4. Is there an Inr motif that agrees with the other TSS predictions?**

**Q5. Zoom out 10x, and then zoom out another 3x. Compare the CAGE read density for the TSS that you just identified for Antp-RM, and the CAGE read densities for other positions in the region. How many positions within this region have CAGE read densities that are similar to the CAGE read density for the TSS of Antp-RM? Are there any Inr motifs within this 450bp region?**

# RNA Annotation and Mapping of Promoters for the Analysis of Gene Expression (RAMPAGE)

Additional evidence for the TSS location(s) can be gathered from RAMPAGE (RNA Annotation and Mapping of Promoters for the Analysis of Gene Expression) data. This technique allows for the identification of the putative TSS with base-pair resolution, and it generally has higher specificity (i.e., signal to noise ratio) than CAGE4.

Importantly, if there are multiple TSSs for a gene (characteristic of genes with an Intermediate or Broad promoter), RAMPAGE will quantify how frequently each TSS is used.

## Exercise 2: Examining the RAMPAGE data for the Antennapedia gene

To examine the RAMPAGE evidence tracks, we will make the following changes to the display modes of the Genome Browser:

1. Under “Mapping and Sequencing Tracks”, change the “Short Match” track to “**hide**”
2. Under “Expression and Regulation”

* TSS (Celniker) (R5): hide
* Combined modENCODE CAGE TSS: hide
* Click on the blue “Combined RAMPAGE TSS (R5)” link
* Change the “Maximum display mode” to “full”
* Scroll down to the “Select views” section
* Change the “Peaks” display mode to “pack”
* In the “List subtracks” section, uncheck the box for “RAMPAGE (Plus)”
* Scroll up to the top of the page, and then click on the “Submit” button
* Click on the blue “RAMPAGE TSS Read Density (R5)” link
* Change the “Display mode” field to ““full”
* Scroll down to the “Select subtracks by strand and stage” section
* Click on the “-” button at the top left corner to unselect all the datasets (Figure 4)
* Select the “Minus” strand checkboxes for the following developmental stages: Embryos 10hr, Embryos 24hr, L1, L2, Adult Females 5d, Adult Males 5d
* Scroll up to the top of the page, and then click on the “Submit” button to update the Genome Browser display

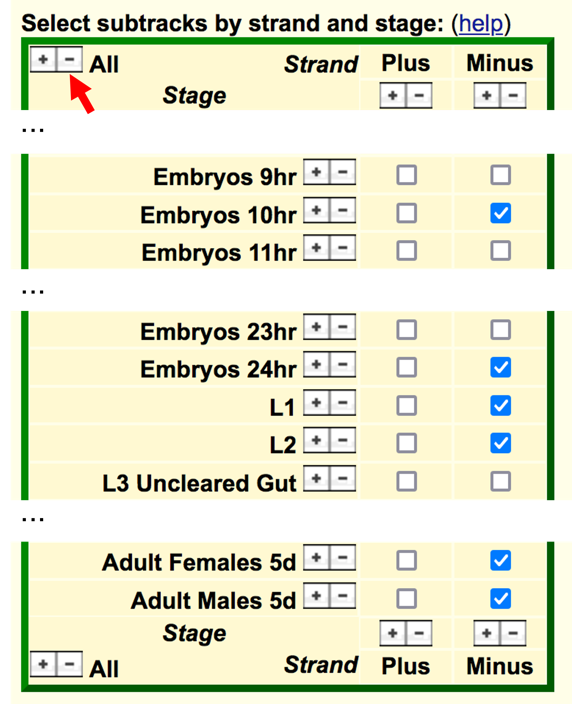


Figure Configure the “RAMPAGE TSS Read Density (R5)” track to show only the data for six developmental stages on the minus strand. Click on the “-” button at the top left corner to unselect all the evidence tracks (red arrow). Select the checkboxes for the developmental stages of interests underneath the “Minus” column label. (The “…” denotes additional stages that were omitted from the screenshot.)

1. Enter “**chr3R:6,999,101-6,999,400**” into the “chromosome range, or search terms” text box and then click on the “go” button.

**Q6. How many TSSs are supported by the RAMPAGE evidence track (Figure 5)?**

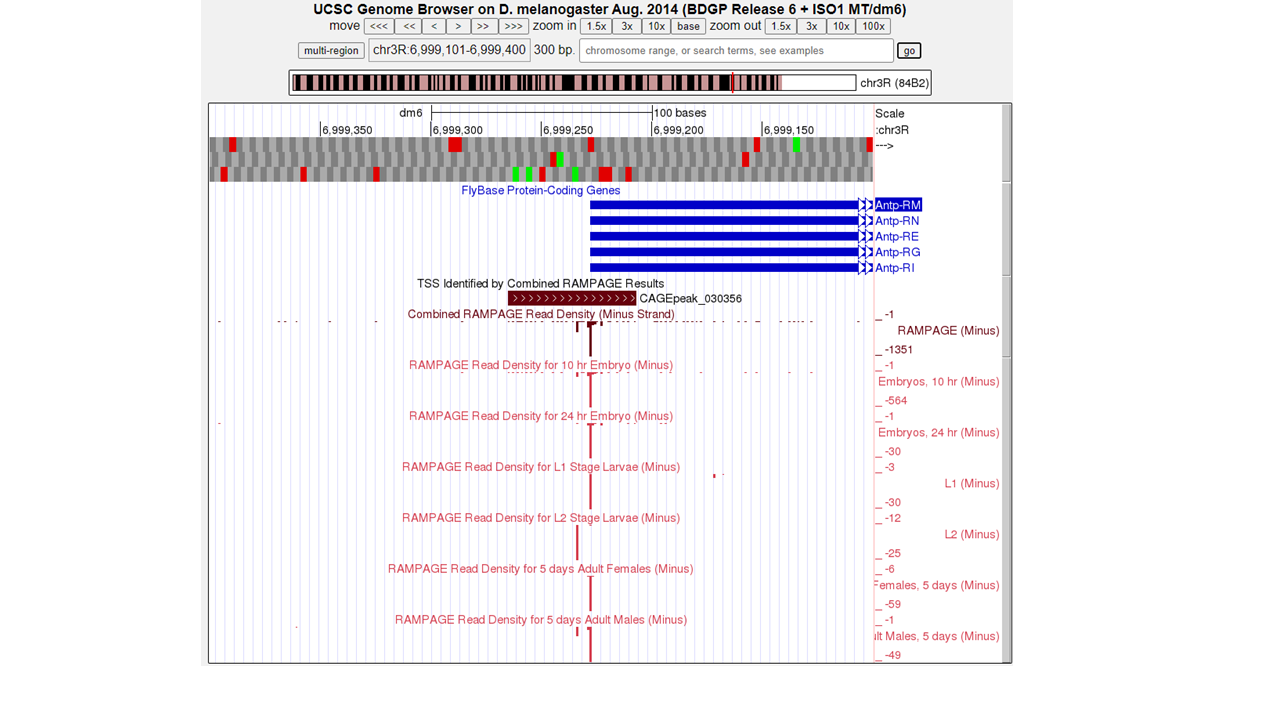


Figure The RAMPAGE read density for all samples combined and for six developmental stages of Antp-RM.

**Q7. Do all of the developmental stages shown use the same TSS?**

To ascertain whether RAMPAGE identified additional TSS in the other developmental stages, we will reconfigure the “RAMPAGE TSS Read Density (R5)” track to show the minus strand RAMPAGE data for all of the available developmental stages.

1. Under “Expression and Regulation”

* Click on the blue “RAMPAGE TSS Read Density (R5)” link
* Scroll down to the “Select subtracks by strand and stage” section
* Click on the “+” button underneath the “Minus” column label (Figure 6)
* At the top of the page, click on the “Submit” button to update the display

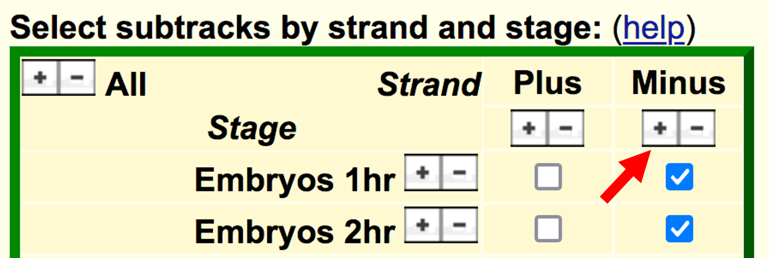


Figure Click on the “+” button underneath the “Minus” column label in the “Select subtracks by strand and stage” section (red arrow) to show the minus strand RAMPAGE data for all of the available developmental stages.

Examination of the RAMPAGE data for all of the available developmental stages shows a more complex picture of transcription initiation for Antp-RM. Examine Figure 7 and your genome browser and note that some developmental stages have a RAMPAGE signal at a different genomic position or have an additional RAMPAGE signal besides the TSS annotated by FlyBase. For other isoforms of *Antp* (or other genes) there may be no RAMPAGE signal for some developmental stages.

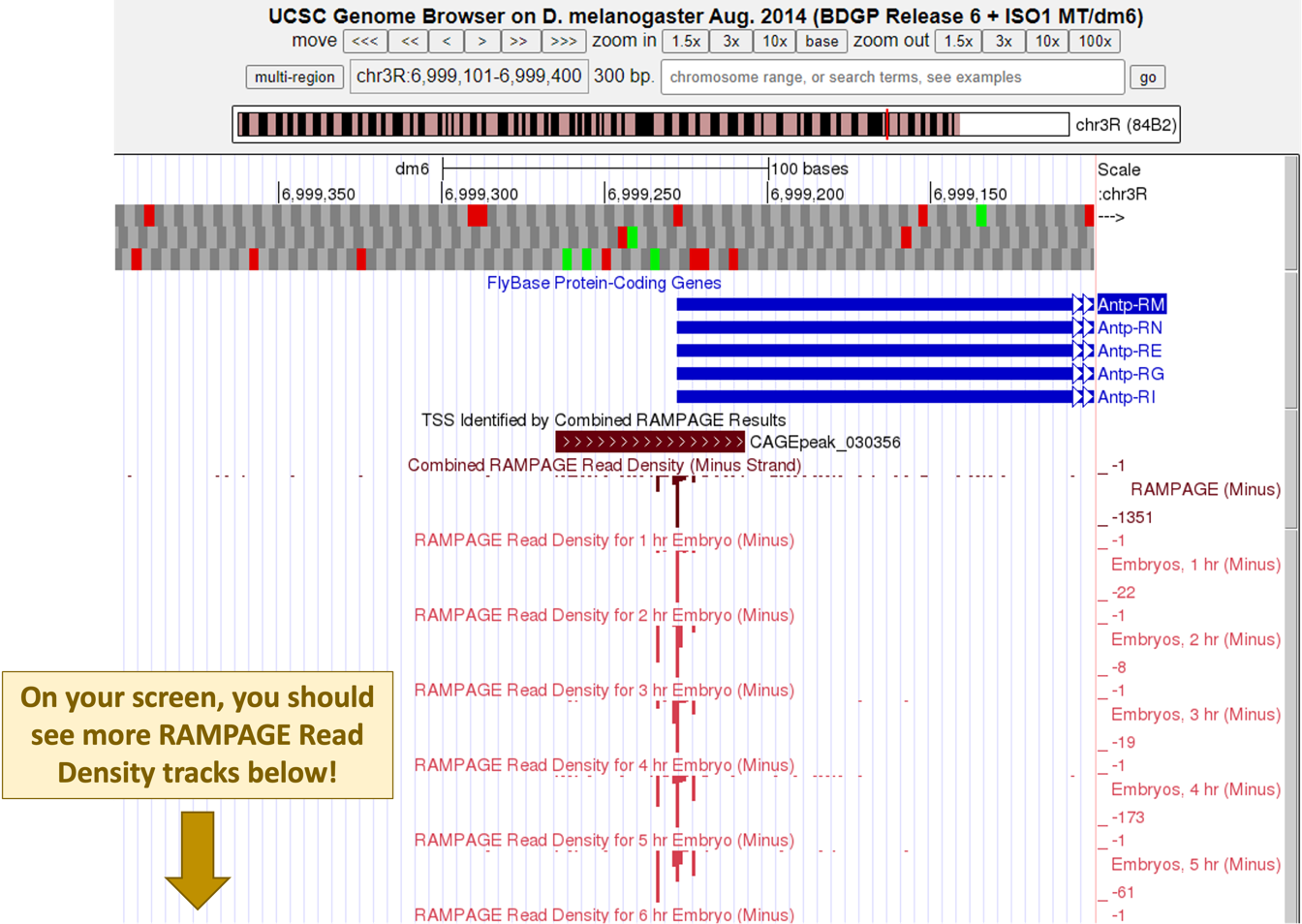


Figure RAMPAGE read density for Antp-RM for different stages of development.

**Q8.** **Based on the RAMPAGE data, do you think that RNA polymerase II initiates transcription at a single site, or at multiple sites for Antp-RM (e.g., examine the RAMPAGE signal for L1 vs. L2, and 1 hr embryo vs. 2 hr embryo)?**

To compare the TSS identified by RAMPAGE with the mRNA expression levels:

1. Under the “RNA Seq Tracks” section, click on the blue “modENCODE RNA-Seq (Development) (R5)” link

* Scroll down to the “Select subtracks by strand and stage” section
* Click on the “**-**” button at the top left corner to unselect all the datasets
* Click on the “+” button underneath the “Minus” column label
* Scroll up to the top of the page, and then click on the “Submit” button to update the Genome Browser display

**Q9.** **For each developmental stage that shows RAMPAGE signal, is there a corresponding RNA-Seq signal?**

# RNA Pol II ChIP-Seq

Chromatin immunoprecipitation followed by high throughput sequencing (ChIP-Seq) is a valuable method that can be used to identify DNA binding sites for proteins such as transcription factors. More recently, X-ChIP-Seq has been used to identify where RNA polymerase II (RNA Pol II) is enriched along the gene5. DNA is crosslinked to histone proteins, then nuclease is added to cleave the linker DNA between nucleosomes. An antibody against one of the proteins in the RNA Pol II complex is used to immunoprecipitate DNA bound by the polymerase. The DNA is then sequenced and mapped to the genome, as with other types of ChIP-Seq.

The results show the genomic regions that are enriched in RNA Pol II, which usually corresponds to the approximate location of TSSs. The RNA Pol II X-ChIP-Seq data can also be used to identify genes whose expression is regulated by RNA Pol II pausing, where RNA Pol II is paused between 30–60nt downstream of the TSS after the initiation of transcription.

## Exercise 3: Examining the X-ChIP-Seq data for the Antennapedia gene using the UCSC Genome Browser Mirror

To examine the X-ChIP-Seq data for the Antp-RM isoform more closely:

1. Change the “chromosome range, or search terms” field to “**chr3R:6,999,151-6,999,320**” and then click “go”.
2. Click on the “hide all” button to hide all the evidence tracks
3. Under “Mapping and Sequencing Tracks”

* Base Position: “full”

1. Under “Genes and Gene Prediction Tracks”

* FlyBase Genes: “pack”

1. Under “ChIP Seq Tracks”

* RNA PolII X-ChIP-Seq: “full”

1. Under "Expression and Regulation"

* TSS (Celniker) (R5): “full”

1. Under “RNA Seq Tracks”

* Click on the blue “Combined modENCODE RNA-Seq (Development) (R5)” link
* Change the “Display mode” field to “full”
* Scroll down to the “List subtracks” section
* Uncheck the box for the “modENCODE RNA-Seq (Plus)” track
* Scroll up to the top of the page, and then click on the “Submit” button to update the Genome Browser display

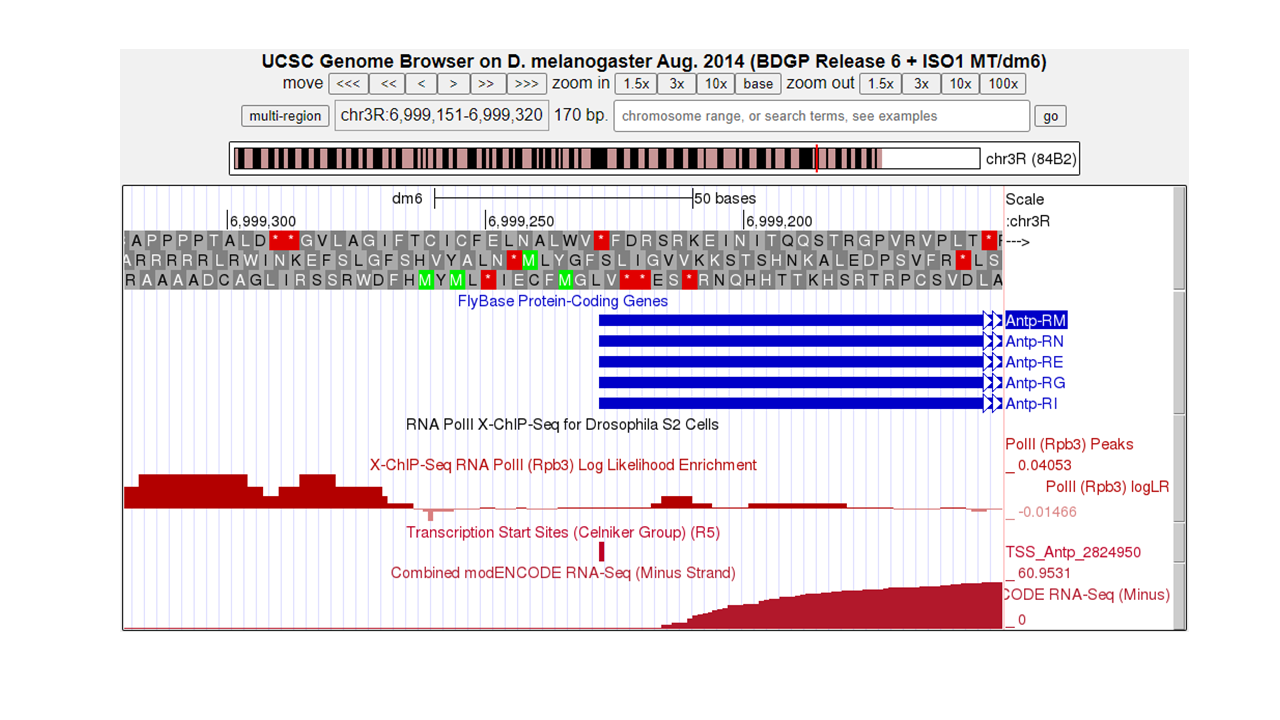


Figure Enrichment of RNA Pol II near the 5’ end of Antp-RM, and the combined RNA-Seq data on the minus strand.

**Q10. According to the X-ChIP-Seq Log Likelihood Enrichment track (Figure 8), is the region immediately upstream of exon 1 enriched in RNA Pol II?**

# Wrapping up

Multiple evidence tracks can be used to identify the TSS for a given isoform of a gene. Examining CAGE, RAMPAGE, and/or X-ChIP-Seq data for genes in *D. melanogaster* may provide additional insight when annotating orthologs in other *Drosophila* species.

# References

1. Kodzius, R. et al. CAGE: cap analysis of gene expression. *Nat. Methods* **3**, 211–222 (2006).
2. Hoskins, R. A. et al. Genome-wide analysis of promoter architecture in *Drosophila melanogaster.* *Genome Res.* **21**, 182–192 (2011).
3. Haberle, V., Forrest, A. R. R., Hayashizaki, Y., Carninci, P. & Lenhard, B. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res*. **43**, e51 (2015).
4. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T. R. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res*. **23**, 169–180 (2013).
5. Skene, P. J. & Henikoff, S. A simple method for generating high-resolution maps of genome-wide protein binding. *eLife* **4**, e09225 (2015).