MODULE TSS3: CAGE, RAMPAGE, & RNA POL II X-CHIP-SEQ (ADVANCED)

MEG LAAKSO AND WILSON LEUNG

Lesson Plan:

Title	**Identifying transcription start sites for peaked promoters using experimental data and conservation**
Objectives	• 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) • Analyze the multiple sequence alignment (i.e. Drosophila Conservation track) of exon 1 from multiple Drosophila species
Pre-requisites	Module TSS1 and Module TSS2
Order	• Explain the experimental techniques used to generate CAGE, RAMPAGE, and RNA Pol II X-ChIP-Seq data, and the Celniker TSS predictions • Illustrate how the experimental data and conservation evidence tracks can be used for TSS annotation • Investigation 1: Students find the *sevenless* TSS using CAGE, TSS predictions, RAMPAGE, and RNA Pol II X-ChIP-Seq • Investigation 2: Students use the conservation track (ROAST multiple sequence alignments) to compare sequences from multiple Drosophila species • Analyze all of the evidence in order to identify the best TSS

Class Instruction	• Discuss the questions: What is transcription? How can our understanding of transcription initiation be used to find a TSS?
	• Use the genome browser to introduce TSS evidence tracks
	• 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?

Associated Videos	• RNA Seq and TopHat Video: https://youtu.be/qepVXEsfLMM • Short Match Video: https://youtu.be/eoeWufgcdvg

INTRODUCTION

This module will use the *Drosophila melanogaster sevenless* gene (*sev*) to illustrate how different experimental techniques (CAGE, RAMPAGE, RNA Pol II X-ChIP-Seq), TSS predictions from the Celniker group at modENCODE, and sequence conservation can be used to identify transcription start sites (TSSs).

INVESTIGATION 1: IDENTIFY THE LOCATION OF TSS USING CAGE, RAMPAGE AND RNA POL II X-CHIP-SEQ DATA

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). 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 Investigation 1, we will first use CAGE data to help identify the TSS. We will also use the TSS predictions produced by the Celniker group at modENCODE to identify the TSS. The Celniker TSS predictions were based on the analysis of experimental data from three methods: CAGE, 5' Rapid amplification of cDNA ends (5' RACE), and 5' Expressed Sequence Tags (ESTs).

EXAMINING THE CAGE DATA FOR THE *SEVENLESS* 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 at http://gander.wustl.edu/ (**Figure 1**)**. 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.

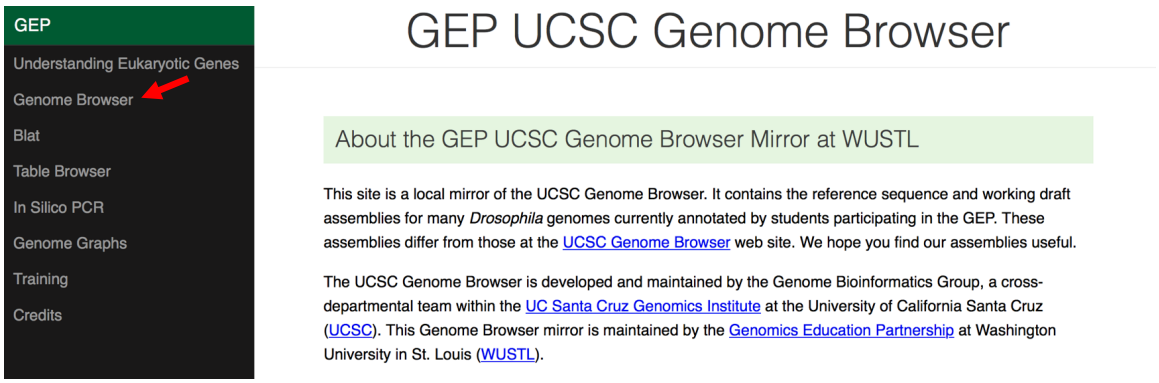

Figure 1 Access the Genome Browser using the "Genome Browser" link.

2. To navigate to the genomic region surrounding the *sevenless* (*sev*) 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 "**sev**" under the "Position/Search Term" field. Click on the "GO" button (**Figure 2**).

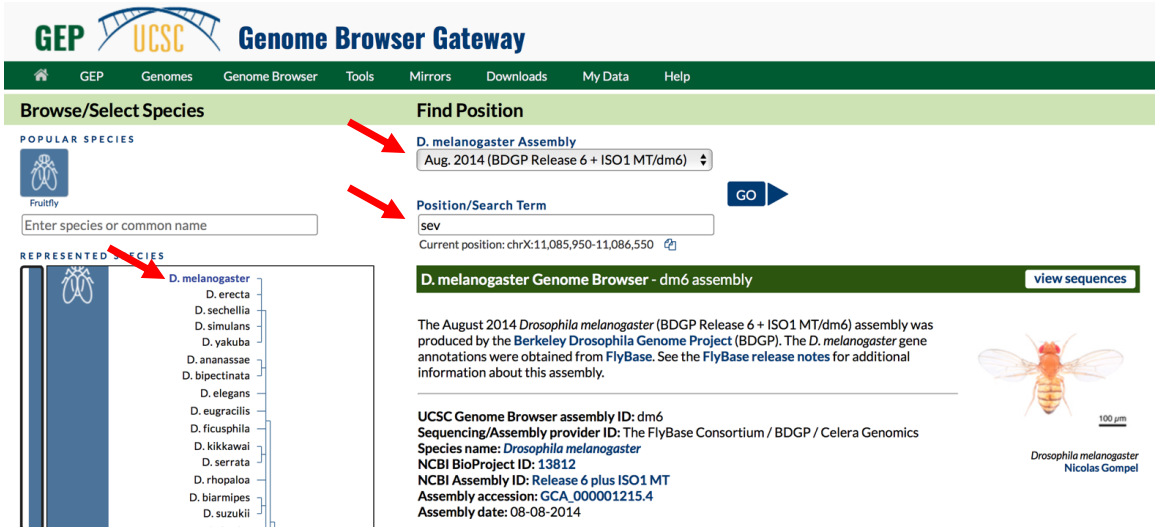

Figure 2 Change the settings on the Genome Browser Gateway page to navigate to the *sev* gene in the *D. melanogaster* release 6 assembly.

3. This region on the X chromosome, denoted **"chrX:11,071,441-11,086,299"** contains the entire *sev* gene. The suffix "-RA" corresponds to the name of the isoform that is associated with the gene. Hence sev-RA corresponds to the A isoform of the *sev* gene. This gene only has one isoform.

4. The *sev* gene is 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 3**).

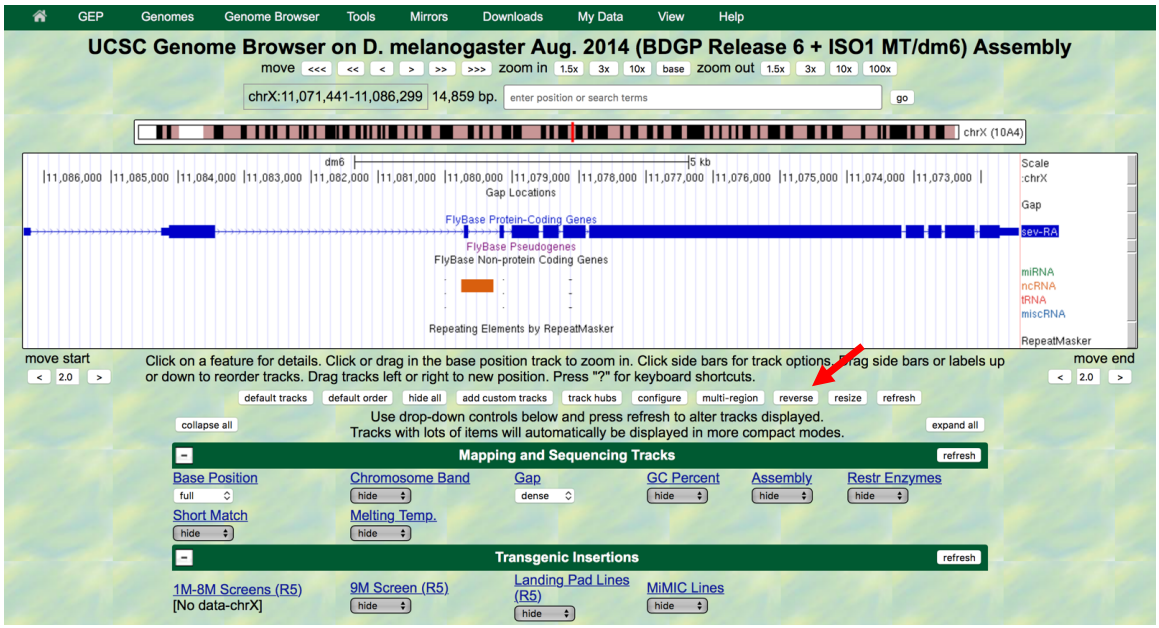

Figure 3 Click on the "reverse" button to reverse complement the entire chrX sequence (red arrow).

5. 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. Then, configure the display modes as follows:

- 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
- Under "Gene and Gene Prediction" Tracks
    - FlyBase Genes: **pack**
- 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.

Note: We will need to zoom in further to see the nucleotide sequence and the amino acid translations. Enter “**chrX:11,086,293-11,086,305**” into the “enter position or search terms” text box, and then click on the “**go**” button.

Note that this exercise requires you to examine the region near exon 1 to determine where transcription begins (**Figure 4**). The start codon for *sev* - where translation begins - is in exon 2.

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

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

Q3. Is the Celniker TSS prediction consistent with the location of the Inr motif? Why or why not? Remember that the Initiator (Inr) motif TC<u>A</u>KTY spans the TSS from -2 to +4, and transcription is predicted to start at the <u>A</u>.

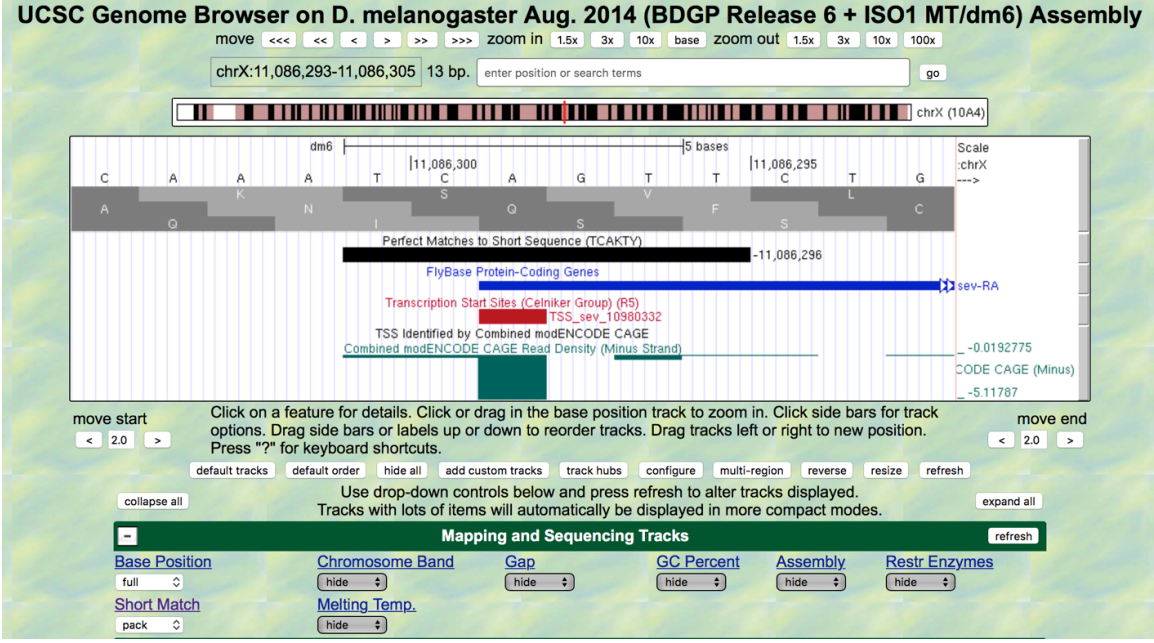

Figure 4 Genome Browser view of the TSS predicted by the Celniker group at modENCODE (red box).

Q4. Based on the Combined modENCODE CAGE TSS track, what is the coordinate of the transcription start site? Remember that this track shows the density of the first base of the CAGE reads, which corresponds to the 5' end of the mRNA.

Q5. Zoom out 10x, and then zoom out another 3x. Compare the CAGE read density for the TSS that you just identified for the *sev* gene, 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 the *sev* gene? Do any of these potential TSSs overlap with the Inr motif?

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

Additional evidence for the TSS locations 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 CAGE[1].

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

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

- Under “Expression and Regulation”
    - TSS (Celniker) (R5): **hide**
    - “Combined modENCODE CAGE TSS: **hide**
    - Click on the blue “Combined RAMPAGE TSS (R5)” link
        - 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 5**)
        - 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

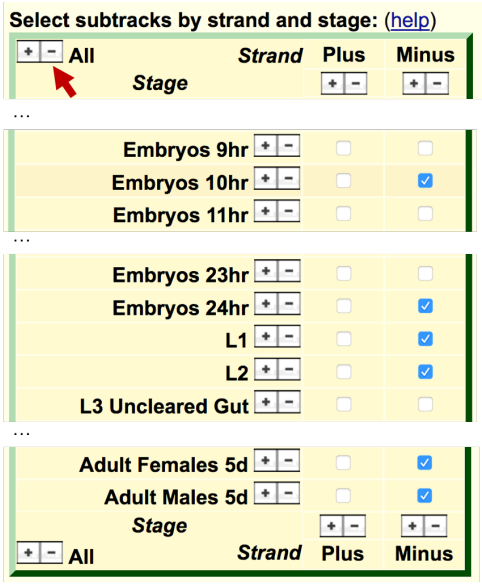

Figure 5 Configure the "RAMPAGE TSS Read Density (R5)" track to show only the data for six developmental stages on the minus strand. Click on the "-" button at the top left corner to unselect all the evidence tracks (red arrow). Select the checkboxes for the developmental stages of interests underneath the "Minus" column label. (The … denotes additional stages that were omitted from the screenshot.)

- Enter "**chrX:11,086,146-11,086,477**" into the "enter position or search terms" text box and then click on the "**go**" button.

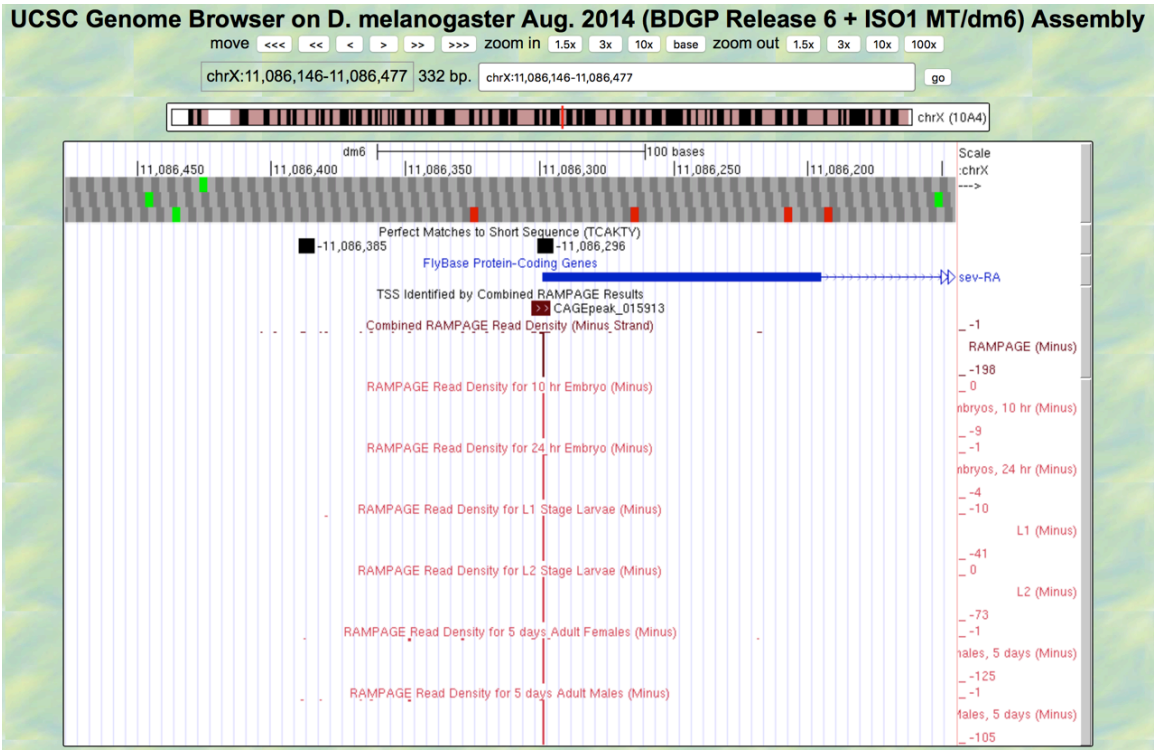

Figure 6 The RAMPAGE read density for all samples combined (dark red), and for six developmental stages (bright red) of *D. melanogaster*.

Q6. How many TSSs are supported by the RAMPAGE evidence track (Figure 6)?

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.

- 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 7**)
    - Scroll up to the top of the page, and then click on the "**Submit**" button to update the Genome Browser display

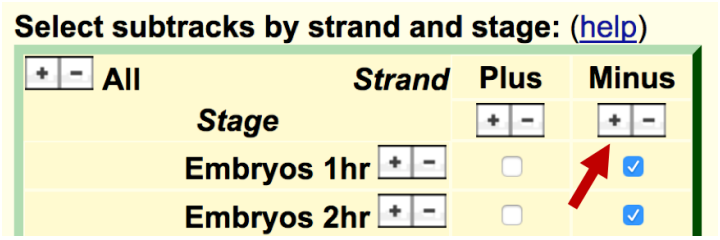

Figure 7 Click on the "+" button underneath the "Minus" column label in the "Select subtracks by strand and stage" section (red arrow) to show the minus strand RAMPAGE data for all of the available developmental stages.

Examination of the RAMPAGE data for all of the available developmental stages shows a more complex picture of transcription initiation for the *sev* gene. Examine **Figure 8** 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 several developmental stages, there is no RAMPAGE signal.

Figure 8 RAMPAGE read density for different stages of development.

Q8. Do you think there is a TSS for each RAMPAGE signal (e.g., examine the RAMPAGE signal for 9-hour embryos)? Why or why not?

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

- Under “RNA Seq Tracks”
  - 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. How many developmental stages show no RAMPAGE signal? How does the lack of RAMPAGE read density relate to the RNA expression levels of the *sev* gene during these developmental stages?

Q10. Based on the analysis of the CAGE and RAMPAGE data, do you think all of the developmental stages use the same TSS?

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 gene[2]. 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.

To examine the X-ChIP-Seq data for the *sev* gene more closely:

- Change the “enter position or search terms” field to “**chrX:11,086,192-11,086,362**” and then click “**go**”.
- Click on the “**hide all**” button to hide all the evidence tracks

- Under “Mapping and Sequencing Tracks”
  - Base Position: **full**
  - Click on the blue “Short Match” link
    - Verify the “**TCAKTY**” sequence is in the “Short (2-30 base) sequence” text box
    - Change the “Display mode” to “**pack**”
    - Click on the “**Submit**” button
- Under “Gene and Gene Prediction Tracks”
  - FlyBase Genes: **pack**
- Under “ChIP Seq Tracks”
  - RNA PolII X-ChIP-Seq: **full**
- 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

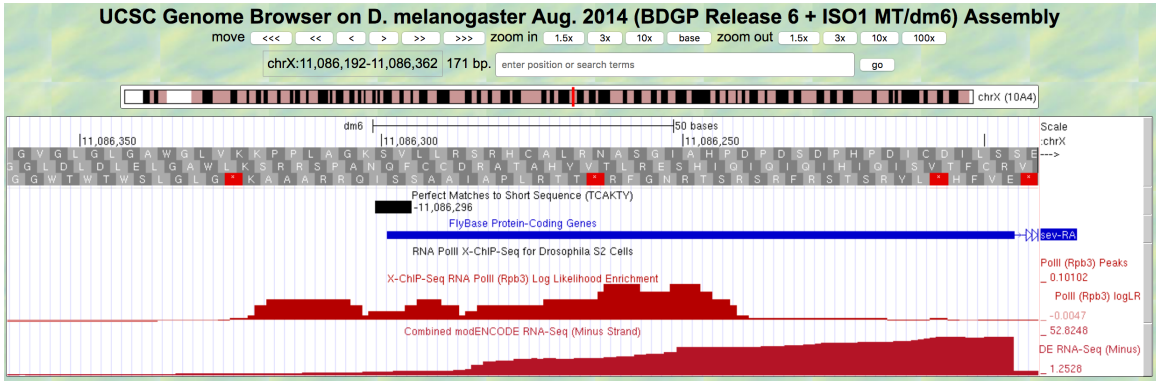

Figure 9 Enrichment of RNA Pol II near the 5’ end of the *sev* gene, and the combined RNA-Seq data on the minus strand.

Q11. According to the X-ChIP-Seq Log Likelihood Enrichment track (Figure 9), what are the coordinates of the region that is enriched in RNA Pol II?

INVESTIGATION 2: IDENTIFICATION OF TSS USING SEQUENCE CONSERVATION DATA

CONSERVATION

Comparison between closely related species of Drosophila can be used to identify conserved sequences near the TSS. In general, we expect the coding regions of a gene will be the most highly conserved (because it is under the strongest selective pressure), followed by the untranslated regions, and regulatory motifs (e.g., transcription factor binding sites). In Investigation 2, you will learn how to configure the genome browser in order to view sequences from 28 species of Drosophila. The sequences have been "aligned"; that is, the sequences are shown in rows to make it easy to compare them.

To examine the Conservation track for the *sev* gene more closely:

- Under "Expression and Regulation"
    - RNA Pol II X-ChIP-Seq: **hide**
- Under "RNA Seq Tracks"
    - Combined modENCODE RNA-Seq (Development) (R5): **hide**
- Under "Comparative Genomics"
    - Drosophila Conservation (28 Species): **full**
    - Change the "enter position or search terms" field to "**chrX:11,086,250-11,086,349**" and then click "**go**".

Figure 10 (page 15) shows a multiple sequence alignment of the region near the TSS of the *sev* gene from 28 species. Note that the sequence near the TSS is highly conserved — the Initiator motif is conserved in *D. melanogaster* and 15 other *Drosophila* species.

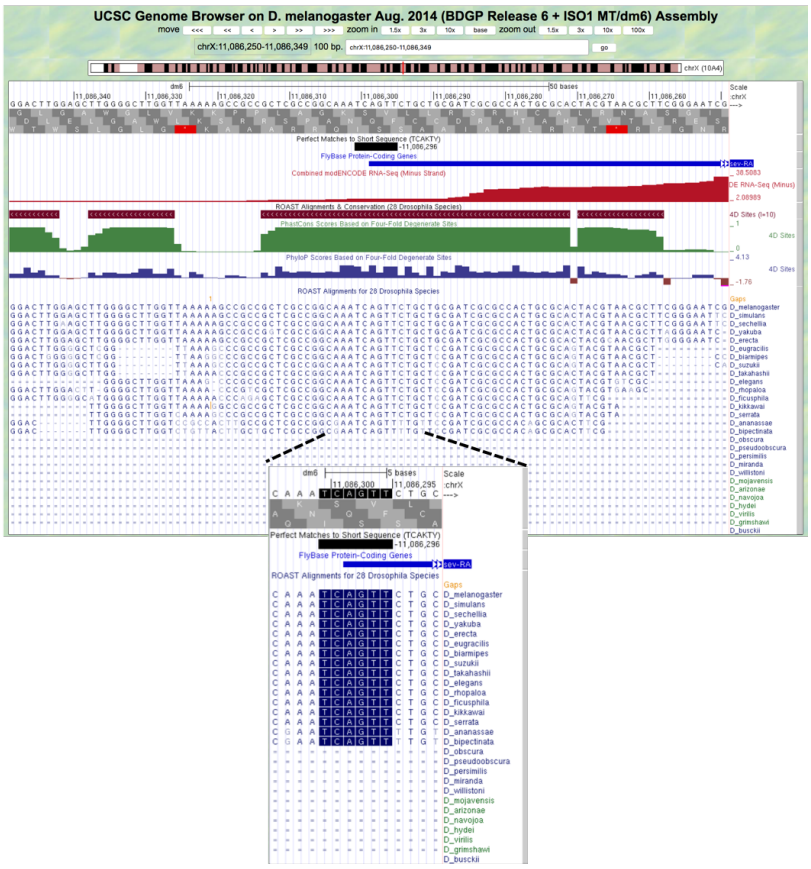

Figure 10 Multiple sequence alignment of 28 *Drosophila* species with conservation scores from PhastCons and phyloP for the region surrounding the TSS of *sev*. (Inset) The Initiator motif in *D. melanogaster* is conserved in 15 other Drosophila species.

Q12. What is the sequence of the Initiator motif for the *sev* gene?

Q13. Based on the location of the conserved Initiator motif, what is the coordinate of the first nucleotide of exon 1?

REFERENCES

1. 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).

2. Skene, P. J. & Henikoff, S. A simple method for generating high-resolution maps of genome-wide protein binding. *Elife* **4,** e09225 (2015).