

MODULE 2: TRANSCRIPTION PART I

MARIA S. SANTISTEBAN

Lesson Plan:

Title	Transcription Part I: From DNA sequence to transcription unit
Objectives	<ul style="list-style-type: none"> Describe how a primary transcript (pre-mRNA) can be synthesized using a DNA molecule as the template. Illustrate the importance of the 5' and 3' regions of the gene for initiation and termination of transcription by RNA polymerase II Learn to identify the beginning and end of a transcript using the capabilities of the genome browser (RNA-Seq, Short Match)
Pre-requisites	Structure of a gene (Module 1)
Order	<ul style="list-style-type: none"> Describe regulatory signals: Transcription Start Site (TSS), and AATAAA sequence (site of transcript cleavage for termination) Investigation 1: Students find the transcript for tra-RA using the genome browser and identify the transcription unit Investigation 2: Students map the 5' end of the transcript Investigation 3: Students map the 3' end of the transcript Review pairing of DNA bases
Homework	None
Class Instruction	<ul style="list-style-type: none"> Discuss the questions: What is transcription? What cellular proteins are required for transcription? How does it work mechanistically? What is/are the products of transcription? (students discuss in pairs, then as a class) Work through the genome browser investigation, then identify where transcription starts and ends for the <i>tra</i> gene. How long is the pre-mRNA? Conclude by challenging students to think about these questions: <ul style="list-style-type: none"> How important is it for RNA polymerase II to recognize the promoter sequence? Do you think it is possible for a gene to have more than one transcriptional start site? How would RNA polymerase II know which one to choose? When would it make a difference in the protein product, and when not?
Associated Videos	<ul style="list-style-type: none"> RNA Seq and TopHat Video: https://youtu.be/qepVXEsfLMM Short Match Video: https://youtu.be/eoeWufqcdvg

INVESTIGATION 1: IDENTIFY THE TRANSCRIPTIONAL UNIT

INTRODUCTION

This module will introduce you to the use of the Genome Browser to illustrate the process of transcription and help you identify regulatory elements, using the *Drosophila melanogaster transformer (tra)* gene as an example. You will use the UCSC Genome Browser Mirror developed by the Genome Education Partnership (GEP), which contains RNA expression data, to identify the different parts of the gene that give rise to pre-mRNA through transcription.

FINDING THE TRANSCRIPT FOR TRA-RA USING THE UCSC GENOME BROWSER MIRROR

1. Open a new web browser window and go to the UCSC Genome Browser Mirror site at <http://gander.wustl.edu/>. Follow the instructions given in module 1 to navigate to the contig1 project in the *D. melanogaster* "July 2014 (Gene)" assembly.
2. To navigate to the genomic region surrounding the *tra* gene, enter "**contig1:9,650-11,000**" into the "enter position or search terms" field located just above the displayed tracks and then click on the "go" button. As you learned in the previous module, you can also use the buttons in the navigation controls section to zoom in, zoom out, and use the arrows to move to different parts of the contig. In addition, you can place your cursor on the "Scale" or the "Base Position" sections of the Genome Browser image and then drag your cursor from the initial position to the end position to zoom into a region of interest.
3. This region from 9,650-11,000 contains the entire *tra* gene and the very end of the previous gene *spd-2*. As described in module 1, the suffix (e.g. -RA) corresponds to the name of the isoform that is associated with the gene. Hence *spd-2-RA* corresponds to the A isoform of the *spd-2* gene.
4. Because the Genome Browser remembers your 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**
 - Under "Gene and Gene Prediction Tracks"
 - FlyBase Genes: **pack**
 - Click on any of the "refresh" buttons to update the display (Figure 1)

Note: Depending on your screen resolution, you may need to zoom in further to see the nucleotides and amino acid translations even if you set the "Base Position" track to full.

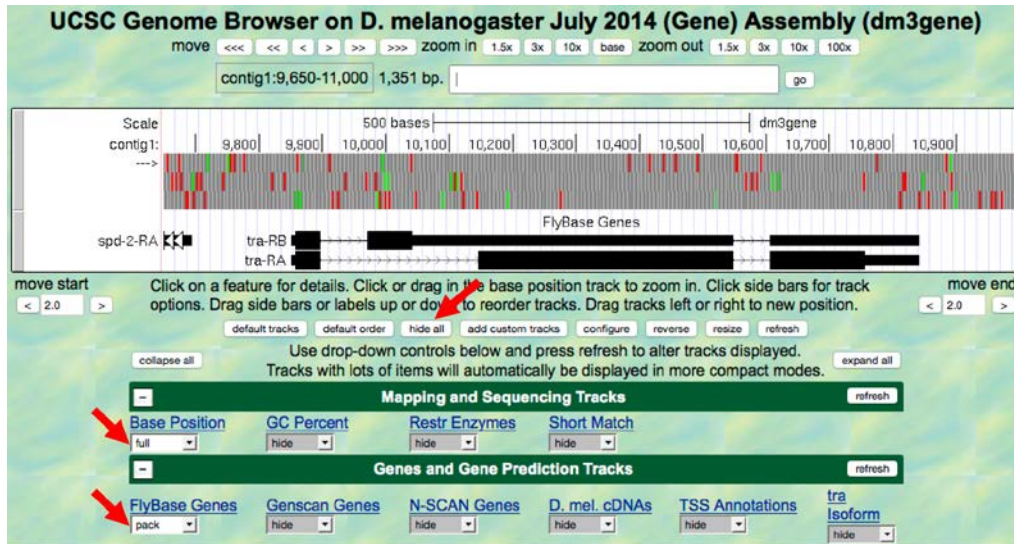


FIGURE 1 CONFIGURING THE DISPLAY MODES FOR THE EVIDENCE TRACKS SURROUNDING THE *TRA* GENE.

IDENTIFYING THE TRANSCRIPTIONAL UNIT FOR THE *TRA* GENE

- Now let's investigate how the string of As, Ts, Cs, and Gs of the DNA sequence in this genomic region give rise to the mRNAs for the *tra* gene. The "FlyBase Genes" track shows the list of protein-coding genes that have been annotated by FlyBase. According to this track, there are actually two different mRNAs (*tra*-RA and *tra*-RB) made from the same DNA sequence (Figure 2). These represent two alternative forms (also known as **isoforms**) of the *transformer* (*tra*) gene product.

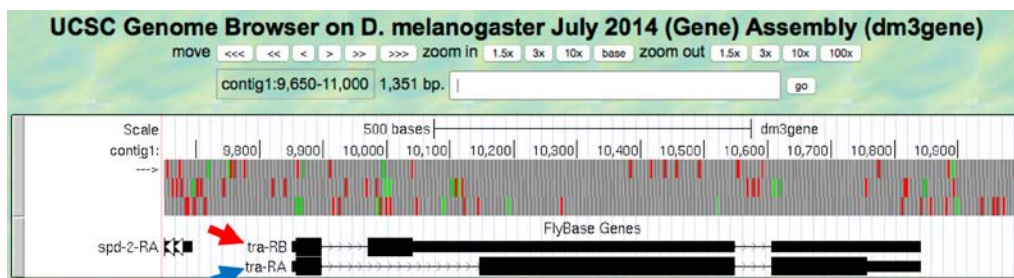


FIGURE 2 FLYBASE ANNOTATED ISOFORMS A (BLUE ARROW) AND B (RED ARROW) OF *TRA* IN *D. MELANOGASTER*.

- For the moment we will focus only on the A isoform of *tra* (*tra*-RA). As you learned in module 1, the black boxes represent the exons (the part of the transcript that makes up the mRNA); the thick black boxes represent the translated regions (i.e. the parts of the exons that contain information that codes for protein) while the thinner black boxes represent untranslated regions (i.e. the part of the exons that do not contain information that codes for protein). Lines that connect multiple boxes together represent introns (i.e. the parts of the transcript that are removed in the production of

a mature mRNA). Collectively, they constitute the **transcription unit**, the part of the gene that is read by RNA polymerase II during transcription. We use the name "transcription unit" rather than "gene" because genes also contain regulatory sequences (promoters and both positive and negative regulatory elements) that are not transcribed. In contrast to prokaryotes, where most of the transcript codes for protein in a single open reading frame (no introns!); in eukaryotes, the transcript contains a lot of extra nucleotides that are not used to form the protein.

Q1.What is the span of (i.e. the start and end base positions) of tra-RA transcription unit?

7. The Genome Browser contains tracks that we can use to visualize the regions of the DNA that are transcribed into RNA. For example, the "RNA Seq Tracks" section contains results from sequencing (mostly mature) mRNAs and then mapping the sequences found in the RNA-Seq reads back to the genome. Hence regions with RNA-Seq read coverage usually correspond to regions in the genome that are being transcribed. To visualize the distribution of these RNA-Seq reads, scroll down to the bottom of the page and then click on the "[RNA-Seq Coverage](#)" link under the "RNA Seq Tracks" section header (Figure 3).

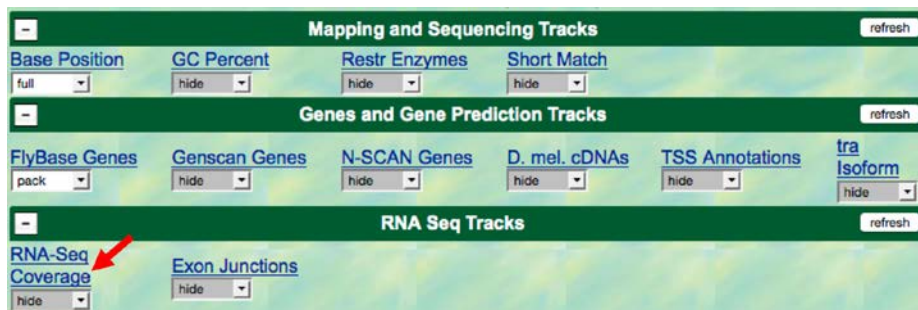


FIGURE 3 CLICK ON THE "RNA-SEQ COVERAGE" TO CONFIGURE THE DISPLAY SETTINGS FOR THIS EVIDENCE TRACK.

8. Using the controls in the "RNA-Seq Read Coverage" page that comes up when you click the "RNA-Seq Coverage" link, we will modify the display settings to the following (Figure 4):
- Change the "Display mode" field to "full"
 - Select the "Data view scaling" field to "use vertical viewing range setting"
 - Change the "max" field under "Vertical viewing range" to 37
 - Under the "List sub-tracks" section, unselect the "Adult Males" track
 - Click on the "Submit" button ("Display mode" line, near the top of the page)

By default, the RNA-Seq Coverage track will auto-scale based on the read depth in the viewing region. The settings above override this setting and manually define the scale to be from 1 to 37. The RNA-Seq Coverage track contains data from mRNA isolated from two separate samples (i.e. from adult males and adult females). Here we unselect the "Adult Males" track so that the Genome Browser will only show the RNA-Seq read coverage from adult females. We will return to the "Adult Males" track in Module 6.

FIGURE 4 MANUALLY DEFINE THE VIEWING RANGE FOR THE RNA-SEQ READ COVERAGE TRACK (RED ARROWS) AND SELECT ONLY THE SUB-TRACK OF INTEREST (I.E. ADULT FEMALES, BLUE ARROW).

9. The Genome Browser image now includes a track in blue with peaks and valleys, labeled "modENCODE RNA-Seq from *D. melanogaster* Whole Adult Females" (Figure 5). The y-axis corresponds to the number of RNA-Seq reads (from whole adult females) that have been mapped to each genomic position of contig1.

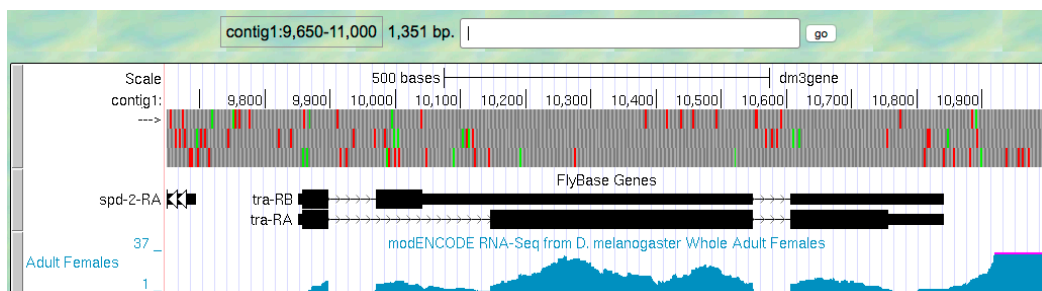


FIGURE 5 RNA-SEQ READ COVERAGE TRACK (BLUE) FOR THE *D. MELANOGASTER* ADULT FEMALE SAMPLE.

Q2.How do the peaks in the RNA-Seq Read Coverage track relate to mRNA abundance?

Q3.Most of the RNA-Seq reads come from mature (processed) RNA. Can you use this data to suggest where introns are located? Are there any regions that seem ambiguous?

In subsequent modules, we will learn more about the mRNA processing that occurs in the nucleus to remove introns prior to translation.

INVESTIGATION 2: IDENTIFY THE 5' END OF THE TRANSCRIPTIONAL UNIT

INTRODUCTION

Previous studies have identified sequence motifs that are enriched in the region surrounding the Transcription Start Sites (TSS). This region is known as the core promoter. By convention, we designate the TSS as +1 and we specify the positions of the sequence motifs with respect to the TSS. For example, the initiator (**Inr**) motif is found at -2 relative to the TSS (2 bp upstream) while the **TATA box** motif is found at -31 or -30 relative to the TSS. Both of these motifs are in the same orientation as the transcript (Figure 6).

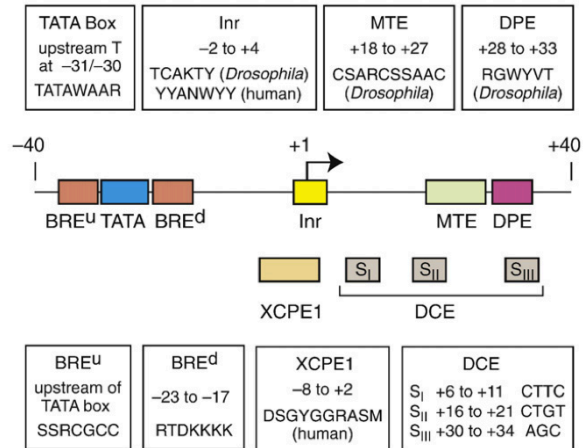


FIGURE 6 MOTIFS THAT ARE ENRICHED NEAR THE TRANSCRIPTION START SITES OF MANY EUKARYOTIC GENES (JUVEN-GERSHON T AND KADONAGA JT¹). NOTE THAT THE MOTIFS ARE OFTEN “DEGENERATE;” N = ANY BASE, R = PURINE (EITHER A OR G), Y = PYRIMIDINE (C OR T), K = KETO (T OR G), M = AMINO (C OR A), S = STRONG (G OR C), W = WEAK (A OR T), V = A/G/C (NOT T), D = A/G/T (NOT C).

In this module, we will review three lines of evidence to determine the TSS position(s) for the *tra* gene. Because RNA-Seq identifies regions of the genome that are being transcribed, we will use the RNA-Seq Coverage track to define the scope of the region to search. The start of the region with RNA-Seq read coverage is the 5' end of the transcript and corresponds to the approximate TSS site, i.e. the beginning of the transcription unit. RNA-Seq data hence becomes our first line of evidence to try to determine the location of the TSS. In other words the information gathered from RNA-Seq will be used to support the choice of the TSS. To learn more about RNA-Seq, watch the RNA-Seq and TopHat video: <https://youtu.be/qepVXEsfLMM>

Q4. Examine the "RNA-Seq Coverage" and the "FlyBase Genes" tracks in the Genome Browser from left to right. At approximately which coordinate (base position) does the RNA-Seq data start for the *tra* gene? (Remember that you can use the navigation controls at the top of the page to zoom in to the region of interest.)

One of the first steps in mRNA processing is the addition of the 5' cap at the beginning of the transcript (we will learn more about capping in the next module). There are experimental techniques that specifically isolate the sequences that are associated with

¹ Juven-Gershon T and Kadonaga JT. (2010) Regulation of gene expression via the core promoter and the basal transcriptional machinery *Developmental Biology* 339:225–229

the 5' cap. These sequences (or “reads”) can then be mapped against the genomic assembly, and the TSSs will show higher read density than the rest of the genome. The modENCODE project summarizes these experimental data to produce a set of predicted TSSs; these predictions are shown in the "TSS Annotations" track. The TSS annotations predicted by modENCODE constitute our second line of evidence to determine TSS location.

1. Before we turn on this track, we will zoom into the region between the end of the previous gene (*spd-2*) and the region where we see RNA-Seq data for *tra-RA*. Change the "enter position or search terms" field to "**contig1:9700-9900**" and then click “go”. **We expect the RNA polymerase to bind and initiate transcription somewhere in this area.** Scroll down to the "Gene and Gene Prediction Tracks" section and change the display mode for the "TSS Annotations" track to "**pack**". Click on a "refresh" button (Figure 7).

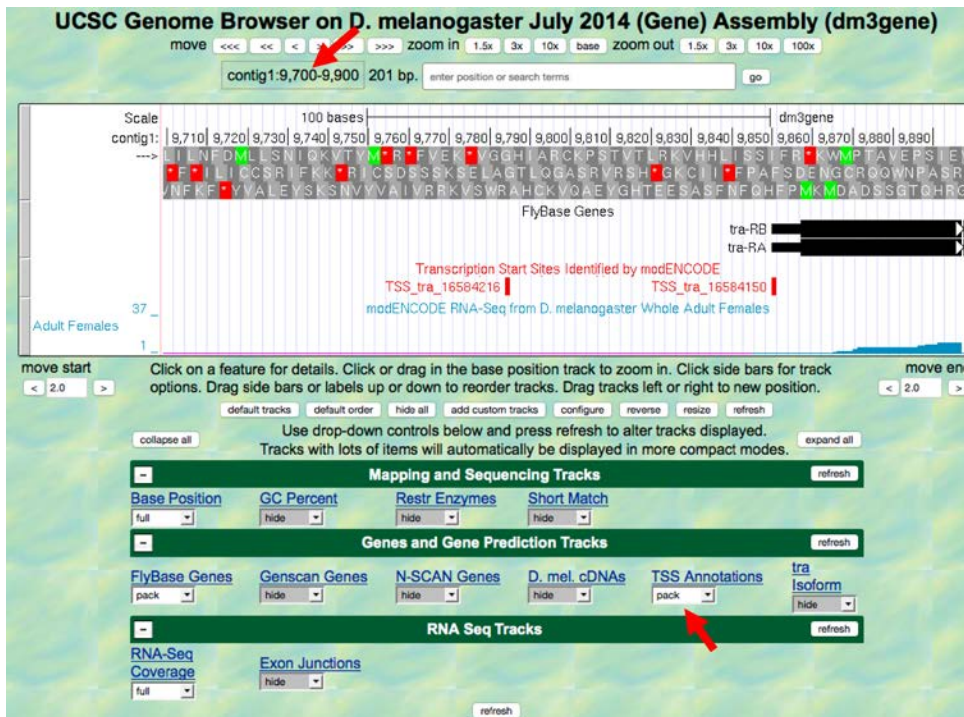


FIGURE 7 ANNOTATED TSS'S IN THE REGION SURROUNDING THE START OF THE *TRA* TRANSCRIPTS

Q5. How many TSS sites were identified using this technique?

Q6. Look at the labels next to each of the annotated TSSs. What are the labels for the TSS sites?

- We will examine each of the annotated TSS's separately to determine their precise coordinates. First, let's zoom in on the feature labeled "TSS_tra_16584216". Look at the ruler in the "Base Position" track to determine the coordinate for this TSS (Figure 8).

Q7.What is the coordinate for TSS_tra_16584216?

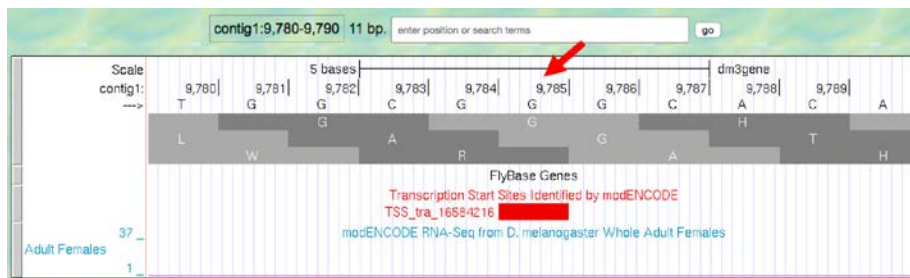


FIGURE 8 DETERMINE THE POSITION FOR THE ANNOTATED TSS "TSS_TRA_16584216".

- Now let's zoom in to the second TSS site, TSS_tra_16584150 (Figure 9).

Q8.What is the coordinate for this TSS?

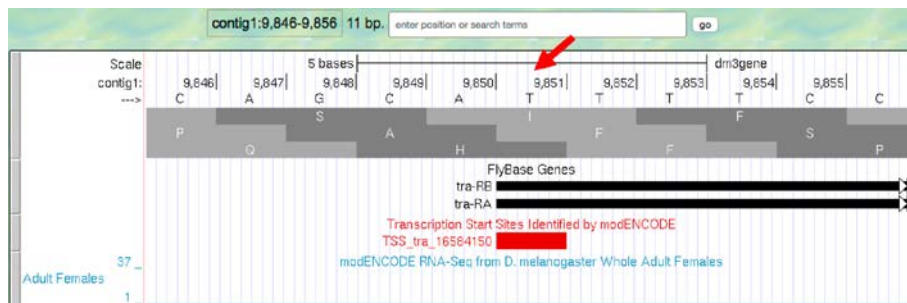


FIGURE 9 DETERMINE THE POSITION FOR THE ANNOTATED TSS "TSS_TRA_16584150".

4. We will use the Genome Browser to gather additional evidence to identify the most likely TSS. First, let's search for the Inr sequence using the "**Short Match**" functionality under "Mapping and Sequencing Tracks". Note that we expect this motif to overlap with the TSS (i.e. from -2 to +4 relative to the TSS). The presence of Inr sequence in the 5' region of the gene will be our third line of evidence to support the most likely TSS location. Change the "enter position or search terms" field to "**contig1:9700-9900**" and then click "go".
 5. To learn more about the Short Match functionality, watch the Short Match video: <https://youtu.be/eoeWufgcdvq>. Scroll down to the "Mapping and Sequencing Tracks" section and click on the "Short Match" link. Change the "Display mode" field to "**pack**" and the "Short (2-30 base) sequence" field to "**TCAKTY**" (Figure 10). Click on the "Submit" button.
- Note: "TCAKTY" is the consensus sequence for the Inr motif, where K (**K**eto) denotes either G or T and Y (**p**Yrimidine) corresponds to either C or T.

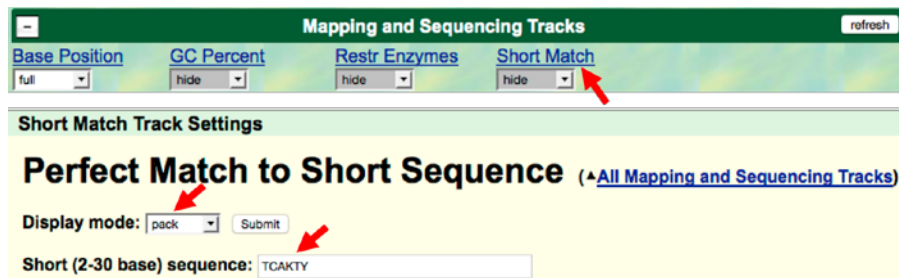


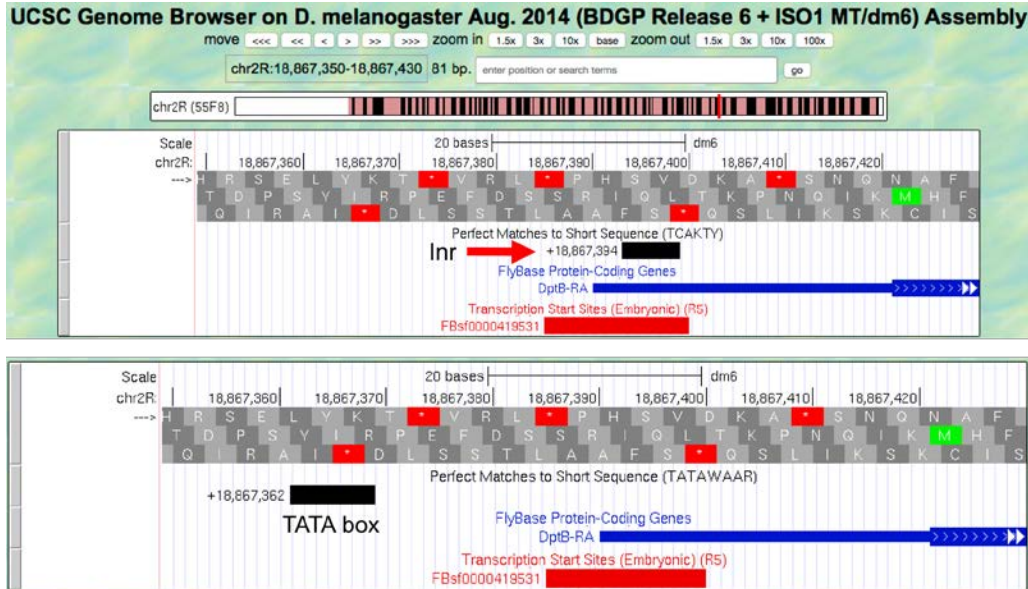
FIGURE 10 CONFIGURE THE "SHORT MATCH" TRACK TO SEARCH FOR THE INITIATOR (INR) MOTIF.

6. Each box in the "Perfect Matches to Short Sequence (TCAKTY)" track corresponds to an instance of the motif. The sign (+ or -) next to each bar denotes the orientation of the match while the number corresponds to the first base of the motif match.
- Q9.** Are there any perfect matches to the Inr consensus sequence in the region between 9700-9900? What are the coordinates and orientation of these matches?
- Q10.** Which base position(s) would you assign as the TSS of the *tra* gene based on the available evidence? Describe your reasoning.

Q11. Is there any ambiguity? In other words, do the three lines of evidence (RNA-Seq tracks, TSS as predicted by the modENCODE data, and the Inr consensus sequence location) point to exactly the same position as being the TSS? If they don't, why might they differ? Could there be more than one TSS?

Let's look at different promoter region. Navigate to the Genome Browser Gateway page and select the *D. melanogaster* " **Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6)**" assembly. Change the "position and search term" field to "**chr2R:18,867,350-18,867,430**" and then click on the "submit" button.

7. Click on "**hide all**" and then enable the following tracks:
Under "Mapping and Sequencing Tracks":
 - Base Position: **full**
 - Short Match: **pack**
8. Search for **TCAKTY** (the Inr motif). Under "Genes and Gene Predictions Tracks":
 - FlyBase Genes: **pack**
9. Under "Expression and Regulation":
 - TSS (Embryonic): **pack**
 - Click on a "refresh" button. Record the position(s) and orientation(s) of any matches to the Inr motif
 - Repeat the search for **TATAWAAR** (the TATA Box motif).



Q12. What do you observe in this case? Are there any perfect matches to the Inr consensus sequence? What are the coordinates and orientation of these matches? What about the TATA Box motif?

INVESTIGATION 3: MAP THE 3' END OF THE TRANSCRIPTIONAL UNIT

INTRODUCTION

After RNA polymerase II has started transcribing a gene (**initiation**), generally with the help of various transcription factors, it will proceed (in a process called elongation) all the way to the termination signal in order to produce a molecule of pre-mRNA. Let's review what we know about the template, and then consider termination.

RNA polymerase II will use the template DNA to synthesize a primary transcript (pre-mRNA) by pairing purine bases with pyrimidine bases. Actually the sequence of nucleotides that you observe on the tracks in the Browser corresponds to the "coding strand" of the DNA (complementary to the template strand); the coding strand is almost identical to that pre-mRNA, except that DNA has _____ versus RNA, which has ____ as the pyrimidine base that pairs with A.

Because DNA is antiparallel, if the coding strand that you see in the browser track runs 5' to 3', then the template strand runs in the _____ direction.

RNA polymerase binds to the **promoter sequence on the template strand**, constructing the transcribed mRNA in the _____ direction, just like the coding DNA strand that you see on the tracks. In fact, polymerases can only add nucleotides to the 3' end (free – OH) of the growing RNA molecule.

Termination of mRNA transcription is different in eukaryotes than in prokaryotes. In eukaryotes, RNA polymerase II passes through one or more **AATAAA** sequences, which lie beyond the 3' end of the coding region (i.e. thick black boxes in the FlyBase Genes track). The pre-mRNA molecule will thus carry the signal AAUAAA. This AAUAAA signal is recognized by a special endonuclease that cuts at a site 11 to 30 nucleotides to its 3' side. As you will learn in the mRNA processing module, a tail of polyriboadenylic acid, poly(A), is added by a special non-template-directed polymerase to the end of the transcript.

Pre-mRNA processing will further studied in modules 3 and 4.

1. We will try to identify the approximate end of the tra-RA transcript using the RNA-Seq data, and will then search the DNA sequence for a termination signal² (AATAAA). Return to the “July 2014 (Gene)” assembly, change the "search terms" field to "**contig1:10,700-10,950**" and then click "submit" to navigate to the 3' end of the *tra* gene. Examine the RNA-Seq read density in the "RNA-Seq Coverage" track (Figure 11).

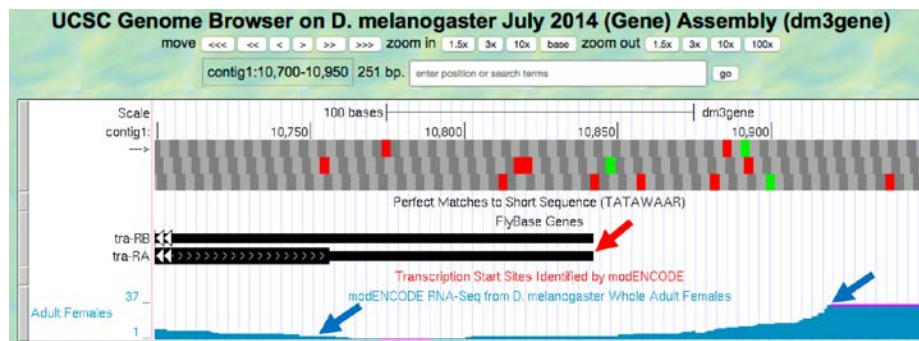


FIGURE 11 COMPARISON OF THE TRANSCRIPTION END SITE ANNOTATED BY FLYBASE (RED ARROW) VERSUS CHANGES IN RNA-SEQ READ COVERAGE IN WHOLE ADULT FEMALES (BLUE ARROWS).

Q13. Where (which base position) do you see a substantial decrease in RNA-Seq read coverage in the whole adult female sample?

² This signal is also referred to as poly-A signal because a poly-A tail is added to its 3' side. In the next module we will refer to it as poly-A signal.

Q14. Where is the 3' end coordinate of the *tra*-RA transcript according to the "FlyBase Gene" track?

You may observe that after decreasing, the amount of RNA-Seq reads in this region starts to increase again, continuing at a higher level to the end of the contig. This is because there is another gene downstream (to the right) very close to *tra*. We can ignore the region (starting at around position 10,900) where the RNA-Seq reads increase.

2. We will now look for a termination signal in this 3' region of the *tra* gene. As we did when searching for the *Inr* sequence, we can use the "Short Match" functionality to search for the **AATAAA** sequence.
3. Click on the "Short Match" link under the "Mapping and Sequence Tracks section". Verify that the "Display mode" is set to **pack** and enter the sequence "AATAAA" into the "Short (2-30 base) sequence" field (Figure 12). Click on the "Submit" button.

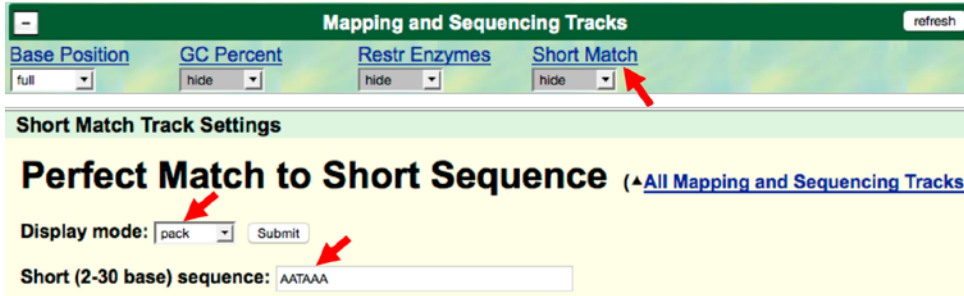


FIGURE 12 USE THE "SHORT MATCH" TRACK TO SEARCH FOR THE MRNA TERMINATION SIGNAL.

Q15. How many matches are there in the search region (i.e. contig1:10,700-10,950)?

Q16. How many of these matches are on the positive (+) strand of the DNA? Remember these sequences, like the *Inr* sequence we discussed before, are strand specific and your gene is on the + strand.

Q17. Is the sequence(s) you found in the question above contained within the 3' untranslated (i.e. non-coding) region of the transcript? (Remember from module 1 that the thick black boxes in the "FlyBase Genes" track represent coding (translated) regions while the thin black boxes represent non-coding (untranslated) regions.

Q18. Based on your analysis above, which position is the best choice for the transcription end site? Describe your reasoning.

Conclusion:

In this lesson, you have seen how the primary transcript (the mRNA molecule) is produced from the template DNA by an RNA polymerase interpreting different signals on the DNA. We saw that DNA sequences upstream of the 5' end (promoter) and near the 3' end (terminator) are important parts of the transcriptional unit. The pre-mRNA molecule from the TSS site to the termination signals will undergo several modifications (processing) in addition to capping that you will learn about in the next few modules.

As discussed above, the reads produced by an RNA-Seq experiment are derived primarily from processed mRNA (not the pre-mRNA). Hence we can explore several additional questions using the RNA-Seq Coverage track:

Q19. Do you see any correlation between the areas with high RNA-Seq read coverage (high peaks) and the different boxes in the tra-RA isoform? Zoom out 10X to get an overview. (Remember that the thick boxes correspond to the coding regions, the thin boxes are the untranslated regions, and the lines with arrows are introns.)

Q20. Where do you see regions in the RNA-Seq coverage data with no coverage at all?

Q21. Where such regions occur within a transcript, what could have happened to these RNA sequences?