Understanding Eukaryotic Genes

Release 0.0.1

Laakso et al. 2017

May 27, 2019

Modules

CHAPTER 1

Module 1

Module 1: Introduction to the Genome Browser: What is a Gene?

Author Joyce Stamm (University of Evansville)

Last Update May 27, 2019

Version 0.0.1

1.1 Introduction to the Genome Browser

Genes encode information that our cells use to carry out their functions. In particular, protein-coding genes provide the cell with the information to make messenger RNAs (mRNAs), which are then used to make proteins. In this module, we will use a web-based visualization tool called a Genome Browser to explore the structure of a eukaryotic gene, and obtain a basic understanding of how this information is stored and used. In subsequent modules, you will learn more about the details of these biological processes, and use the Genome Browser to examine the experimental data that provide evidence for a detailed gene structure. The protein-coding genes in eukaryotes (higher organisms, with a cell nucleus) are much more complex than the protein-coding genes in prokaryotes (bacteria, organisms without a nucleus). We are still trying to figure out all of the details!

1. Start by watching the Genome Browser video
2. Open a web browser and navigate to a custom version of the Genome Browser. The browser was developed by the Genome Bioinformatics Group at the University of California Santa Cruz (UCSC). The custom version is at http://gander.wustl.edu. Click on the `Genome Browser` link on the left menu (Figure 1.1).
3. Change the following fields in the "Genome Browser Gateway" section (Figure 1.2):
    - Select `D. melanogaster` under the "REPRESENTED SPECIES" field. This will allow you to view the genome of the insect *Drosophila melanogaster*.
    - Confirm that `Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6)` is in the "Assembly" field. This is the version of the *D. melanogaster* genome that you will view. The **genome assembly** is simply the **genome sequence** produced after chromosomes have been fragmented, those fragments have been sequenced, and the resulting sequences have been put back together. A genome assembly is **updated** when

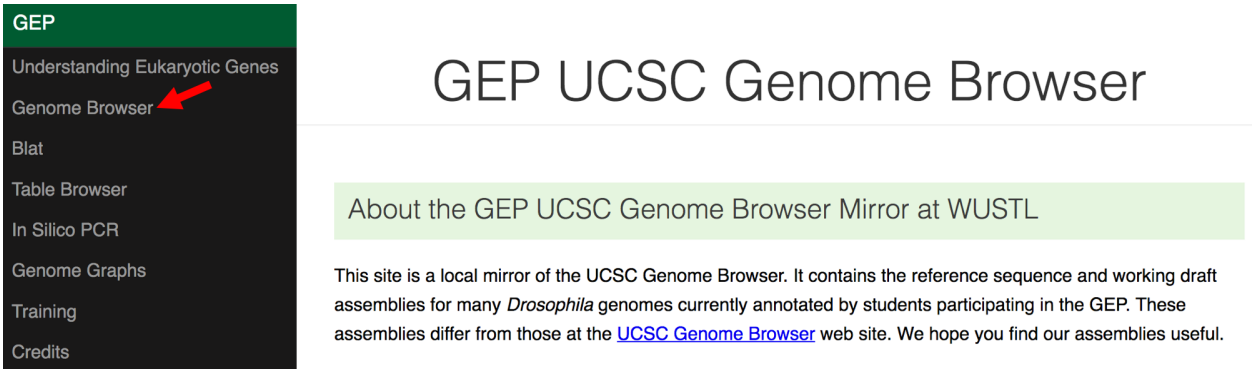


Figure 1.1.: Access the Genome Browser gateway page using the "Genome Browser" link.

DNA has been sequenced that allows gaps to be filled. It may also be updated when a new assembling algorithm is released. The August 2014 *Drosophila melanogaster* (BDGP Release 6 + ISO1 MT/dm6) assembly was produced by the Berkeley Drosophila Genome Project (BDGP).

- Enter `chr3L` into the "Position/Search Term" text box so that you can view the left (L) arm of chromosome 3 (chr3).

4. Click on the `GO` button.

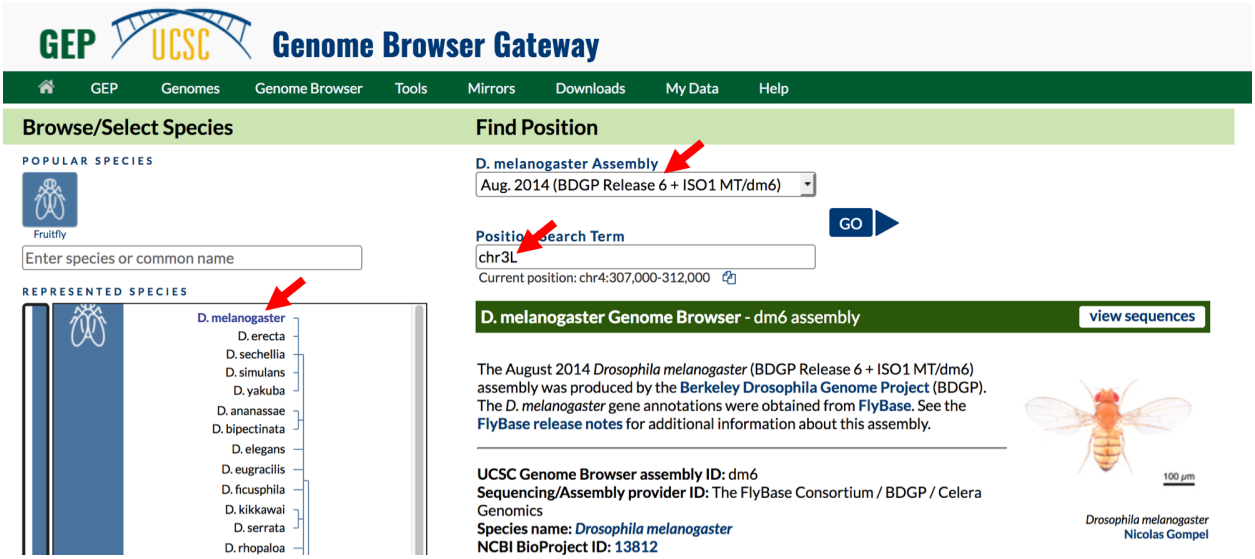


Figure 1.2.: Configure the Genome Browser Gateway page to view the sequence for the left arm of chromosome 3 in *D. melanogaster*.

5. The next screen can be divided into four major sections (Figure 1.3):
   - A top green toolbar is used to navigate to the different tools provided by the Browser.
   - Navigation Controls allow us to navigate or zoom to different parts of the genome.
   - A genomic features panel (the white area) shows the locations of the different genomic features within the portion of the genome (e.g. chr3L) specified by the label next to the "enter position or search terms" text box
   - A Display Controls section may be used to manipulate how much detail is visible in the genomic features panel of the Genome Browser. To match the screenshot in Figure 1.3:

- Scroll down in this section to the bar labeled "Mapping and Sequencing Tracks", go to "Base Position", and select `dense` from the drop-down menu.
- Scroll down to "Genes and Gene Prediction Tracks", go to "FlyBase Genes" and select `squish` from the drop-down menu.
- Check that all other tracks are set to `hide`, and then click on any `refresh` button to update the genomic features panel.

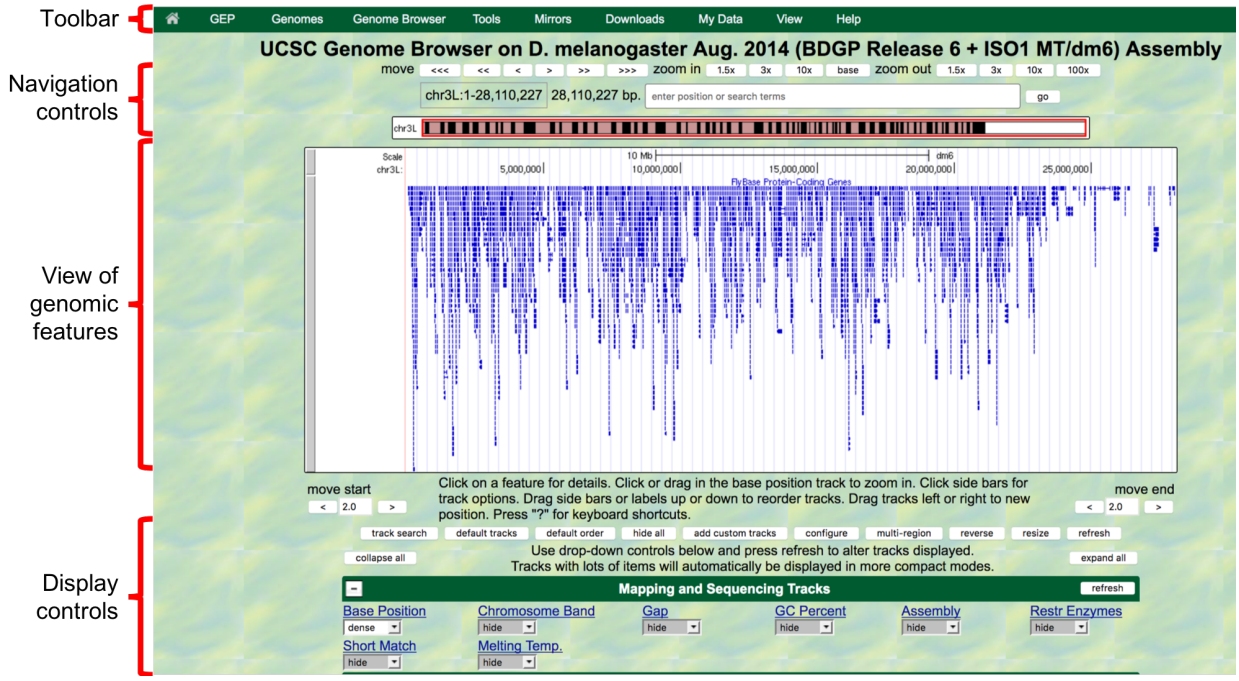


Figure 1.3.: The four major sections of the Genome Browser.

You can use the buttons in the "Navigation control" section to navigate to different parts of the genome. You can zoom in to a region by clicking on one of the buttons next to the "zoom in" label (i.e. 1.5x, 3x, 10x, base). Similarly, you can zoom out by clicking on the buttons next to the "zoom out" label. Alternatively, you can enter the genome *coordinates* into the "enter position or search terms" field and then click on the "go" button to navigate to a specific region in the genome assembly.

The "size" field next to the "enter position or search terms" text box (red arrow in Figure 1.4) shows the total size of the genomic region that you are viewing. In this case, the "size" field shows that chr3L (i.e. the left arm of chromosome 3) in *Drosophila melanogaster* has a total length of ~28 million *base pairs* (bp). We will learn more about the key functionalities of the Genome Browser in subsequent modules. For now, we will focus on the large white rectangle shown on this page; this contains a graphical representation of the genomic features (e.g. protein coding genes, percent GC content) of chr3L mapped against the DNA sequence, which is embedded in the top line of the white box.

The different types of features (also known as "**tracks**" or "**evidence tracks**") are separated by a title and are often shown in different colors. What types and how many tracks are shown in the view of genomic features is controlled by the display controls at the bottom. The view shown on Figure 1.4 displays only some of tracks in the "Gene and Gene Prediction tracks", and all the other tracks in other sections (transgenic insertions, chromatin domains, ChIP seq tracks, Expression and Regulation, etc.) are "hidden". More information about evidence tracks is available in the Tracks video.

We can examine the region under the blue title labeled "**FlyBase Protein-Coding Genes**" to estimate the number of protein-coding genes on chr3L. In this track each gene is represented by a set of blue boxes connected by thin blue lines. There are clearly fewer blue boxes at the right side of the image compared to the left, which suggests that genes

are not uniformly distributed along the chromosome (Figure 1.4).

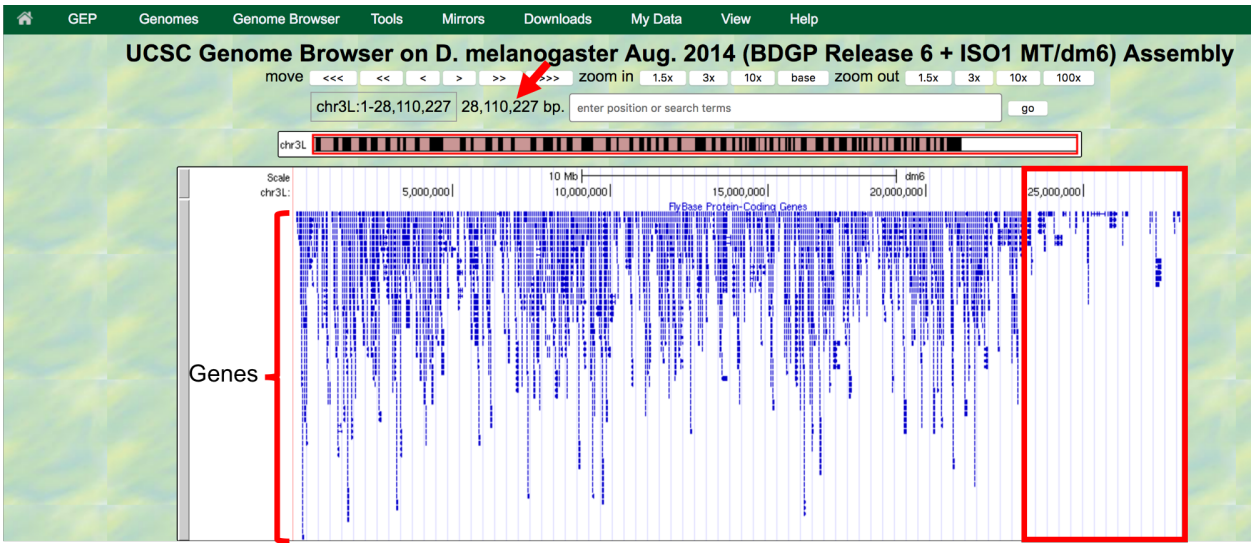


Figure 1.4.: Genome Browser shows that the entire *D. melanogaster* chr3L sequence has a length of ~28 million base pairs (red arrow) and that the right end of the chromosome has low gene density (red box).

In the genome browser, each chromosome may be organized into smaller projects called contigs (for contiguous sequences). In this next part, we will examine contig1, a much shorter region in the left arm of chromosome 3.

6. Click on the `Genomes` link on the top toolbar to return to the Genome Browser Gateway page.
7. Change the assembly option to `July 2014 (Gene)` and verify that the "position" field has been set to `contig1` (Figure 1.5).
8. Click on the `GO` button.

The "size" field now has the value "**size 11,000 bp**", which means that contig1 has a total length of 11,000 bp.

To further explore the features on contig1, we will examine the results from two of the available tracks.

9. Scroll down to the "Display controls" section (i.e. green bars) to the bar labeled "Mapping and Sequencing Tracks" and verify that the display mode under the "Base Position" track is set to `dense` and the "FlyBase Genes" track is set to `pack`.
10. The display mode for all other evidence tracks should be set to `hide` (Figure 1.6).
11. Click on any `refresh` button to update the Genome Browser image.

Explore the contig1 genomic region using these tracks on the Genome Browser. You will observe distinct groups of connected boxes. These connected boxes and lines are genes, and their names are indicated on the left. Connected boxes and lines that are stacked vertically represent alternative forms of a gene, called *isoforms*. Answer the following questions:

Question 1

How many genes are there in contig1?

Question 2

What are the names of these genes?

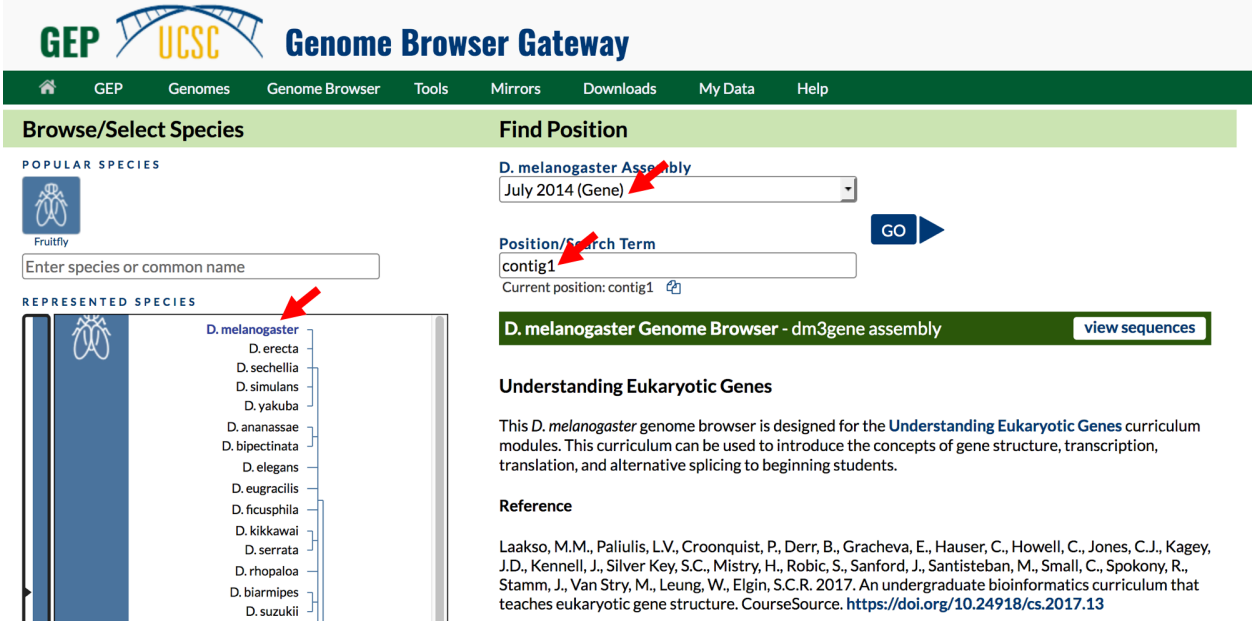


Figure 1.5.: Return to the Genome Browser Gateway page and then select the "July 2014 (Gene)" assembly.

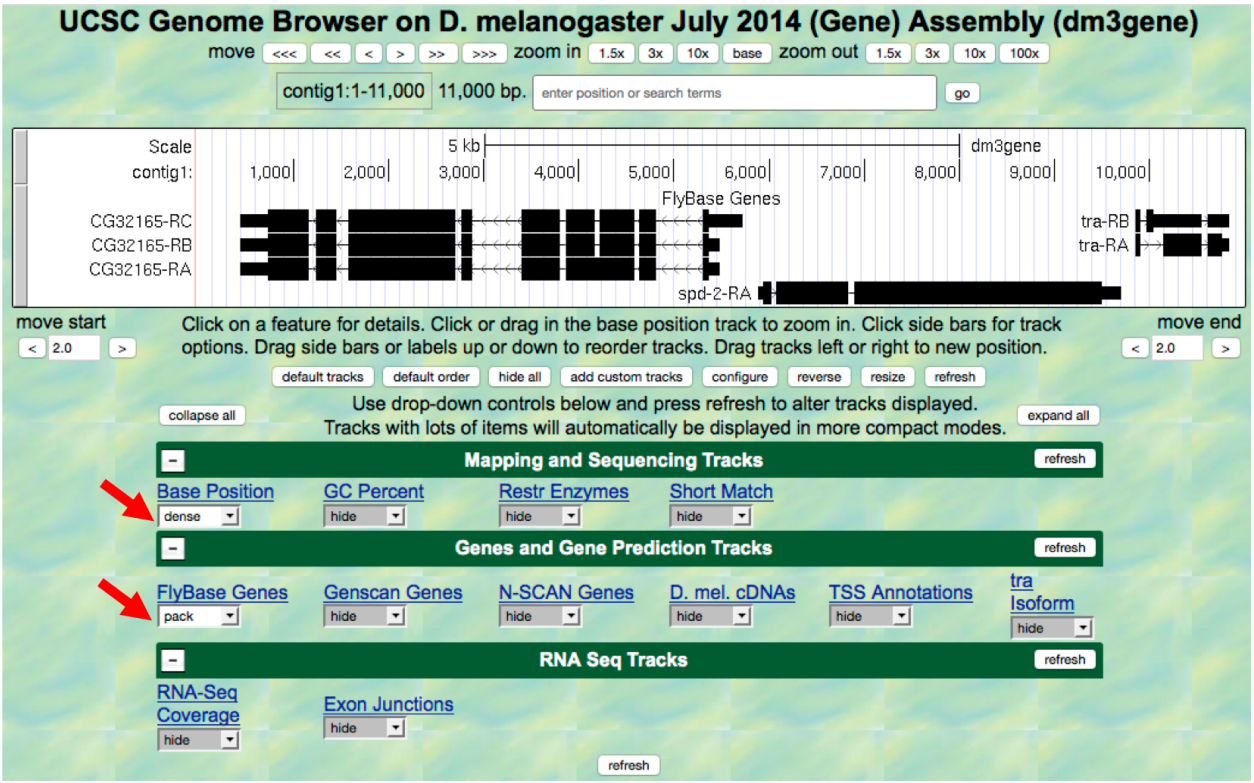


Figure 1.6.: Verify the display settings for the "July 2014 (Gene)" assembly.

Question 3

Which gene has the largest span (i.e. the largest distance between the start and end of the gene)?

12. Now let's examine the gene at the end of this contig more closely. Type `contig1:9,841-9,870` into the "enter position or search terms" text box and then click on `go`. (Note that you don't need to use commas when entering base positions). The Genome Browser image will update to show only bases 9,841 to 9,870 of contig1. Note the letters that appear just below the base position numbers. These letters correspond to the nucleotide at each position. For example, both forms of the *tra* gene, tra-RA and tra-RB, begin with a T at position 9,851 (Figure 1.7).

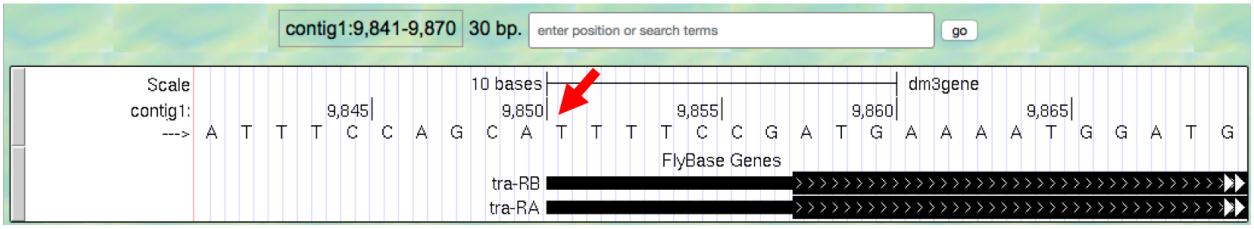


Figure 1.7.: The Base Position track shows the underlying genomic sequence for a region when you zoom in.

13. Look at the left end of the display, under the word "contig1". The arrow here is pointing to the right. When you click on the `--->` arrow, the arrow will switch orientation and point to the left (Figure 1.8, top). In addition, the nucleotides in the "Base Position" track will also change from black to grey. Clicking on the `<---` arrow again will return it to its original orientation (Figure 1.8, bottom).

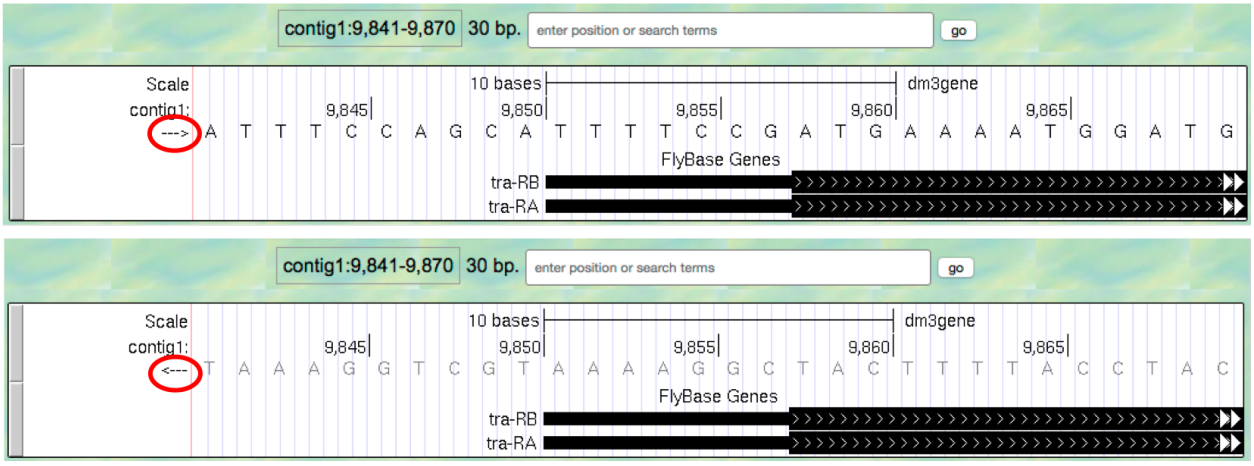


Figure 1.8.: Click on the arrow to change the nucleotides shown on the base position track.

Question 4

What is the relationship between the bases displayed when the arrow is pointed to the left versus when it is pointed to the right?

Question 5

Why do you think the bases are displayed in this way in the Genome Browser?

Both forms of the *tra* gene begin at 9,851 and they have the same prefix ("tra") but different suffixes ("-RB" and "-RA", respectively). The prefix corresponds to the name of the gene (*tra*) in *D. melanogaster* while the two suffixes indicate that there are two different versions (i.e. isoforms) of this gene. We will examine the differences between these two isoforms later. For now, we will focus our analysis on the A isoform of *tra* (tra-RA).

1.2 Genes are composed of exons and introns

14. To see the entire *tra* gene, type `contig1:9,800-10,860` in the "enter position or search terms" text box and click `go` (Figure 1.9). Alternatively, you can use the buttons next to the "zoom out" label and the arrows next to the "move" label to adjust the display.

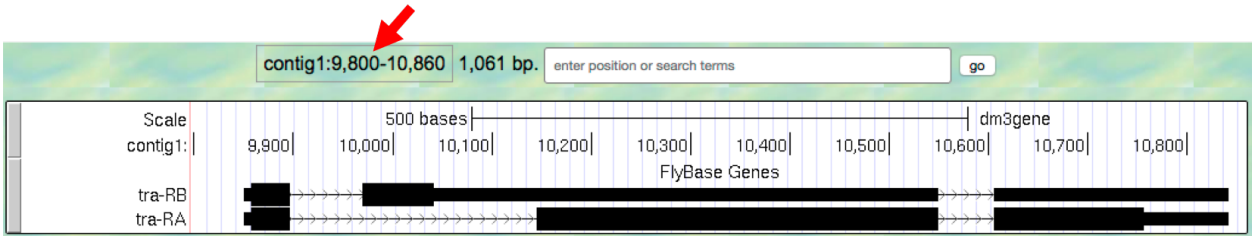


Figure 1.9.: The genomic region surrounding the *tra* gene.

15. Carefully examine the tra-RA isoform. Notice that the isoform consists of black blocks that are connected by lines. On the lines are arrowheads that point from left to right. The black blocks are the *exons* (expressed regions of the gene; Figure 1.10). To use the information stored in a gene, a cell uses the DNA sequence as a template to produce a molecule called a messenger RNA (mRNA). This process is called *transcription*. You will see in *Module 2* that while the initial transcript (product of transcription) is continuous, copying all the DNA, only exon sequences are retained in the processed mRNAs. The lines connecting the blocks are the *introns* (intervening regions of the gene). These sequences will be removed during the production of *mature mRNAs*. The arrows on the lines denote the direction of transcription (or orientation) of the gene.

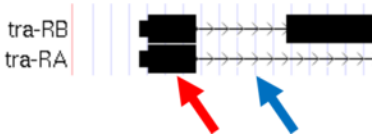


Figure 1.10.: The black boxes (indicated by the red arrow) are the exons and the lines connecting the blocks (indicated by the blue arrow) are the introns.

Question 6

How many exons does tra-RA contain?

Question 7

How many introns does tra-RA contain?

1.3 Genes provide the information to make proteins

The mRNA sequence contains the information that the cell needs to make proteins. You will learn more about this process in *Module 5*. Here we will use the Genome Browser to examine the basic features of a protein.

16. Return to the Genome Browser, and type `contig1:9,850-9,875` into the "enter position or search terms" text box.
17. Scroll down to the "Mapping and Sequencing Tracks" section and change the display mode for the `Base Position` track to `full` (Figure 1.11).
18. Click on the `refresh` button to update your display.

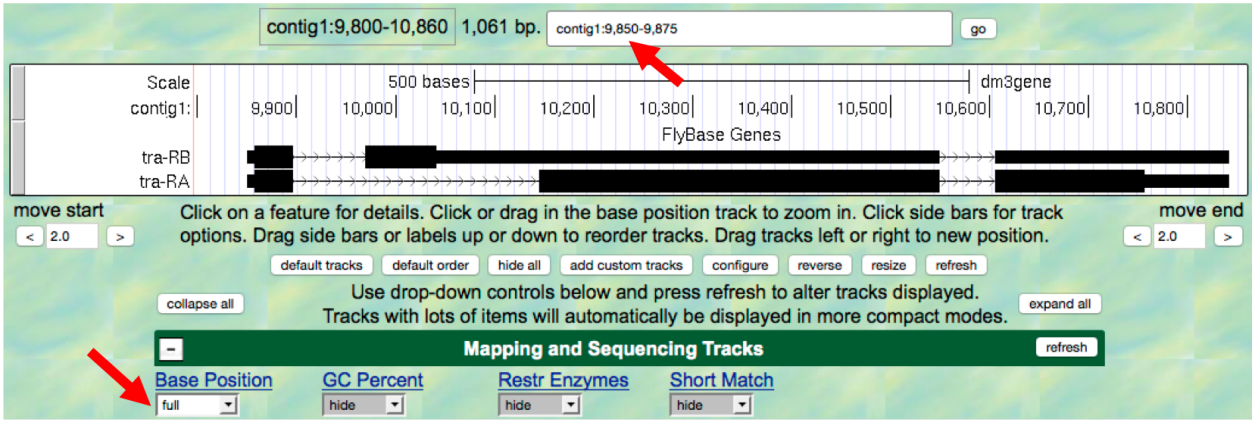


Figure 1.11.: Examine the "Base Position" track in the "full" display mode.

Proteins are made up of *amino acids*, and the mRNA provides the information for the amino acid sequence. This information is read by the cell in groups of three bases, with each three-base group (i.e. *codon*) specifying an amino acid. The Genome Browser uses single-letter abbreviations to represent each amino acid. These are shown on your Genome Browser as three new rows of information directly below the DNA sequence (Figure 1.12).

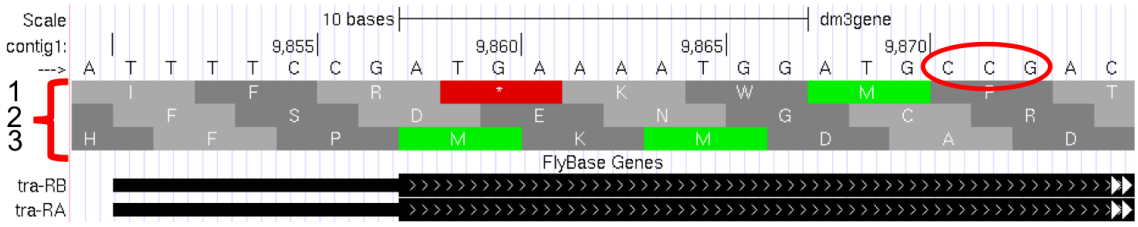


Figure 1.12.: Three new rows appear beneath the nucleotide sequence when the Base Position track is in "full" mode.

Question 8

Why do you think it takes three lines to display the amino acid information?

Tip: Remember that a codon is specified by three bases, e.g. CCG = Proline (circled in Figure 1.12).

Module 5 will have more details about *translation*, the process of copying the information from mRNA into protein. For now, we will just identify the beginning and the end of the protein. You should see three codons that are highlighted

in green (one in row 1 and two in row 3). These codons all correspond to the amino acid M (i.e. Methionine). This amino acid is almost always used to start a protein. There is only one codon that can code for Methionine: **ATG**.

The first M on the third row of amino acids (at 9,858-9,860) corresponds to the start of the protein for the A isoform of *tra*. The position of this Methionine also coincides with the transition of the thinner rectangle to the thicker rectangle. Hence the thick rectangles denote coding sequence — the parts of the exon that carry information about the protein sequence and are the translated parts — while the thin blocks indicate regions that are part of the exon but do not carry protein sequence information, or the untranslated parts (Figure 1.13).

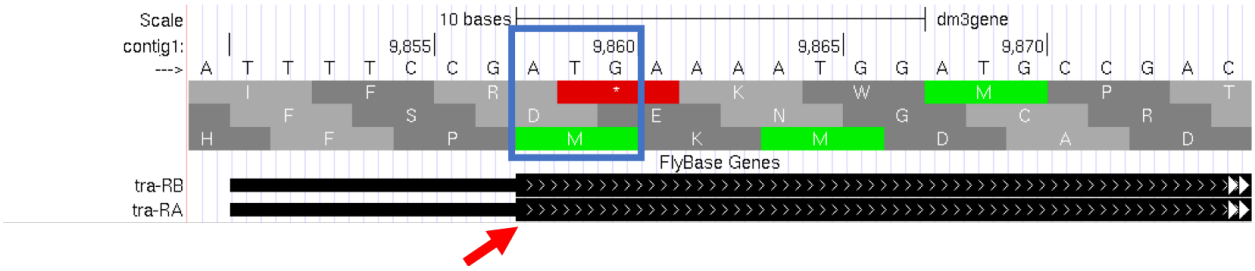


Figure 1.13.: The location of the initial Methionine for the A isoform of *tra*.

Let's examine the other end of the protein. There are three special codons (known as *stop codons*) that signal the end of the protein. These codons (TGA, TAA and TAG) are indicated by an asterisk "*" and are highlighted in red in the "Base Position" track.

19. Type `contig1:10,740-10,765` into the "enter position or search terms" text box and then click on the `go` button. Note the stop codon (*) at position 10,754-10,756, specified by the bases **TGA**, in the second row of amino acids (Figure 1.14). This is the last codon before the transition from the thick exon block to the thinner one. The Genome Browser therefore shows that a part of the mRNA extends beyond the end of the protein-coding region. This is a general property of mRNAs: they contain extra sequences both before and after the protein-coding sequence. These sequences, at the *5'* and *3'* end of the protein-coding sequences, are called the 5' and 3' *UTRs* (untranslated regions) respectively.

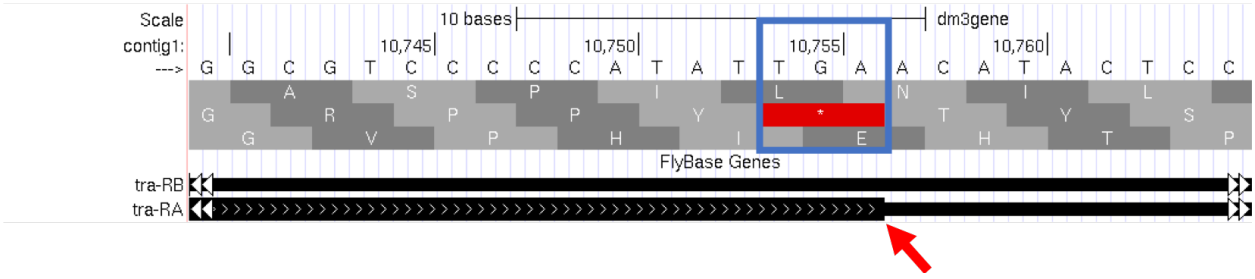


Figure 1.14.: The end of the translated region for the A isoform of *tra*.

1.3.1 Genes have directionality

As you saw above, the sequence of the codons in the A isoform of *tra* are read from left to right relative to the orientation of contig1. This also means that the start of the protein is located toward the left of the end of the gene. However, recall that DNA is double-stranded, and that the two strands run in opposite directions to each other (i.e. they are **antiparallel**). It turns out that, like the *tra* gene here, some genes are read on the DNA strand conventionally termed the "top strand" (from left to right), while other genes are read on the "bottom strand" (from right to left). We will examine one such example next.

20. Type `contig1:5,350-5,375` into the "enter position or search terms" text box and then click on the `go` button. This region contains the start of the protein-coding region of the *CG32165* gene. However, there are

no Methionines (green boxes) at the transition point between the thin and thick rectangles (Figure 1.15, top). However, note that the arrows in the thicker part of the indicated exon point from right to left, indicating that this gene is read from the bottom strand.

21. Click on the `arrow` beneath the "contig1" label in the "Base Position" track so that it points in the same direction as indicated for the gene in this region. This will complement the sequence and allow you to read the bases of the "bottom" strand of DNA. Remember that the codons on this strand must be read from right to left. Now you can see that there is a start codon in this region, the corresponding green M amino acid (at 5,365-5,367) in the third row (Figure 1.15, bottom).

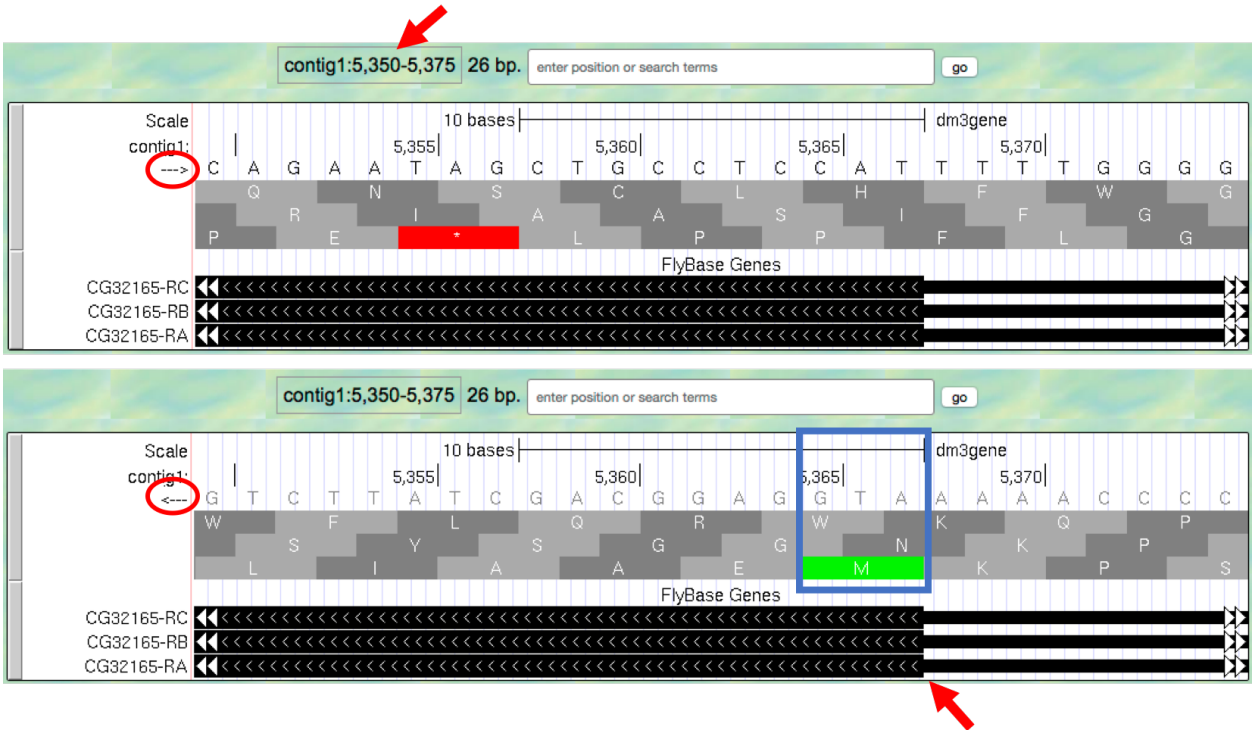


Figure 1.15.: Examine the start of the coding region for a gene on the minus strand.

1.4 Coding exons are translated in a single reading frame

The combination of the directionality (with two alternative directions) and the three rows in the "Base Position" track means that there are six different ways to translate a genomic region, i.e. to determine the sequence of amino acids from a DNA sequence. These different ways to translate a genomic region are known as reading *frames*.

22. To illustrate this concept, change the "enter position or search terms" text box to `contig1:1-12` and then click `go` in order to zoom in to the first 12 nucleotides of the contig1 sequence.

23. Click on the `arrow` underneath the "contig1" label in the "Base Position" track so that it points to the right (Figure 1.16).

The first row (frame +1) begins at the **first** nucleotide in contig1 and the first amino acid (P) is derived from the codon **CCC**. The second row (frame +2) begins at the **second** nucleotide in contig1 and the codon **CCG** also codes for the amino acid P. The third row (frame +3) begins at the **third** nucleotide in contig1 and the codon **CGG** corresponds to the amino acid R (Figure 1.17). Because a codon is comprised of 3 nucleotides, the codon beginning at the fourth nucleotide (GGT) is again in frame +1.

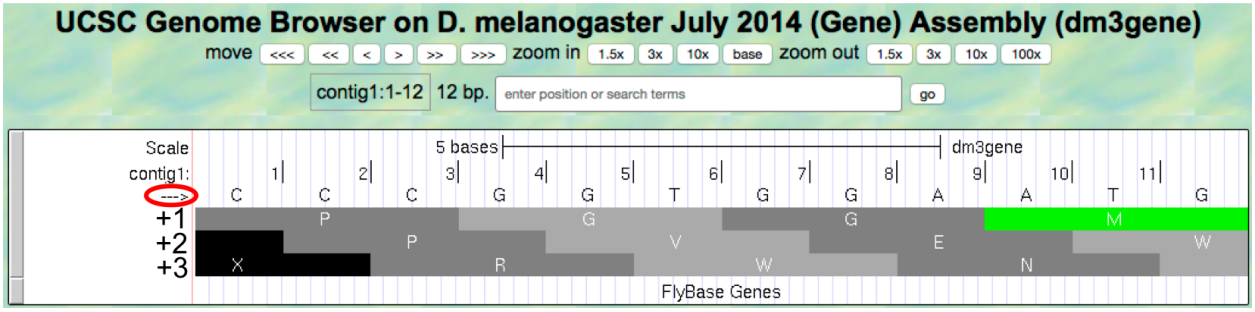


Figure 1.16.: Examine the "Base Position" track for the first 12 bases of contig1 in the top strand.

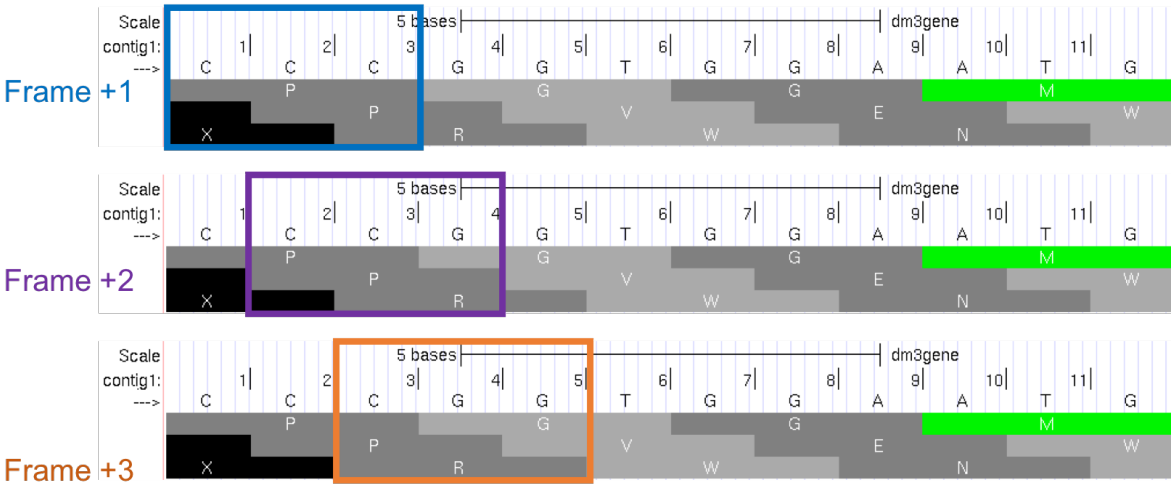


Figure 1.17.: Interpreting the reading frame using the Base Position track.

Examination of the "Base Position" track at the beginning of the contig shows that the three positive reading frames are numbered relative to the start of the contig1 sequence. Similarly, the three reading frames on the bottom strand are numbered relative to the end of the contig1 sequence (i.e. the beginning of the reverse complement of the contig sequence). Because contig1 has a total length of 11,000bp, we will change the "enter position or search terms" field to `contig1:10,989-11,000` so that we can examine the last 12 nucleotides of this contig.

24. Click on the `arrow` underneath the "contig1" label so that it points to the left (Figure 1.18).

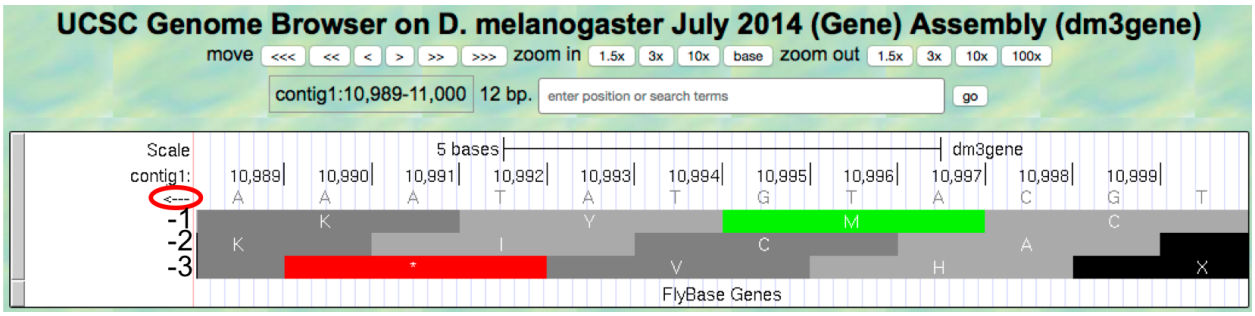


Figure 1.18.: Examine the "Base Position" track for the last 12 nucleotides of contig1 in the bottom strand.

Because we are examining the reverse complement of the contig1 sequence, we need to read the nucleotide and amino acid sequences on the "Base Position" track from right to left. The first row (frame -1) begins at the last nucleotide (11,000) of contig1 and the codon **TGC** codes for the amino acid C. The second row (frame -2) begins at the penultimate nucleotide at 10,999 and the codon **GCA** codes for the amino acid A. The third row (frame -3) begins at 10,998 and the codon **CAT** corresponds to the amino acid H (Figure 1.19).

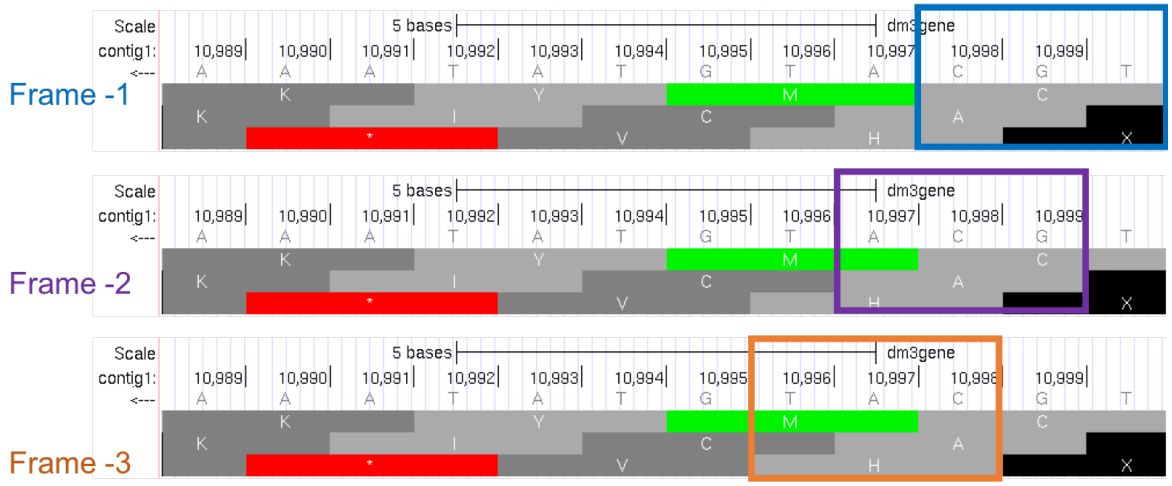


Figure 1.19.: Using the "Base Position" track to interpret the reading frames on the bottom strand.

25. Now that we understand how to interpret the reading frame information using the "Base Position" track, we can investigate the coding regions of the *tra* gene more closely. Change the "enter position or search terms" field to `contig1:9,800-9,960` and then click on the `go` button.

26. Click on the `arrow` underneath the "contig1" label in the "Base Position" track so that we can examine the translations of the top strand (running left to right) (Figure 1.20).

Our previous analysis shows that there is a *start codon* (green rectangle in the "Base Position" track) in the third row that corresponds to the transition from the thin to the thick rectangles (Figure 1.13). Hence the coding part of the first exon of the A isoform of *tra* is said to be "in frame +3". Notice that there is also an open reading frame (*ORF* — stretch of codons uninterrupted by stop codons) that overlaps with the thick box in the second row (frame +2) but

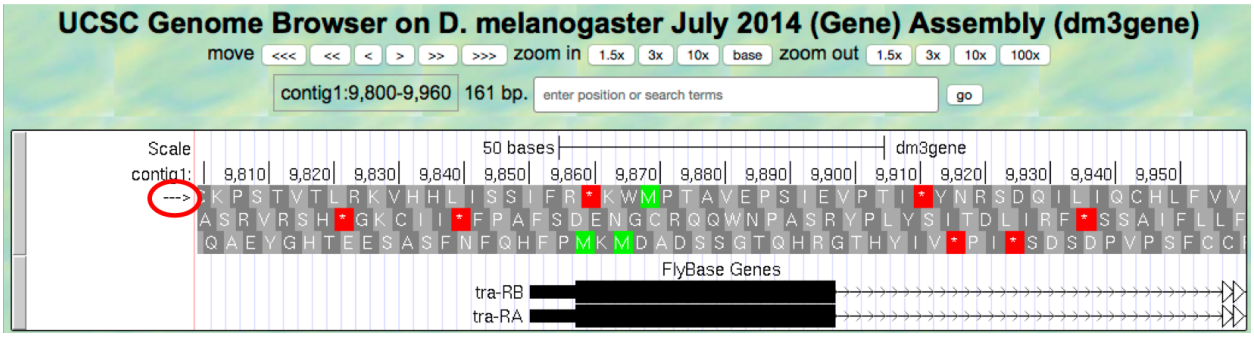


Figure 1.20.: The genomic region surrounding the first exon of tra-RA.

there are no start codons that overlap with the thick box. In contrast, the first row (frame +1) contains a start codon, but the thick box also overlaps with a stop codon (red star). When we examine the region downstream of the black boxes, we find that there are stop codons in all three reading frames. However, these stop codons do not interrupt the open reading frame of the first exon because they occur in the region of the arrowed lines (i.e. the first intron, see blue arrows in Figure 1.21).

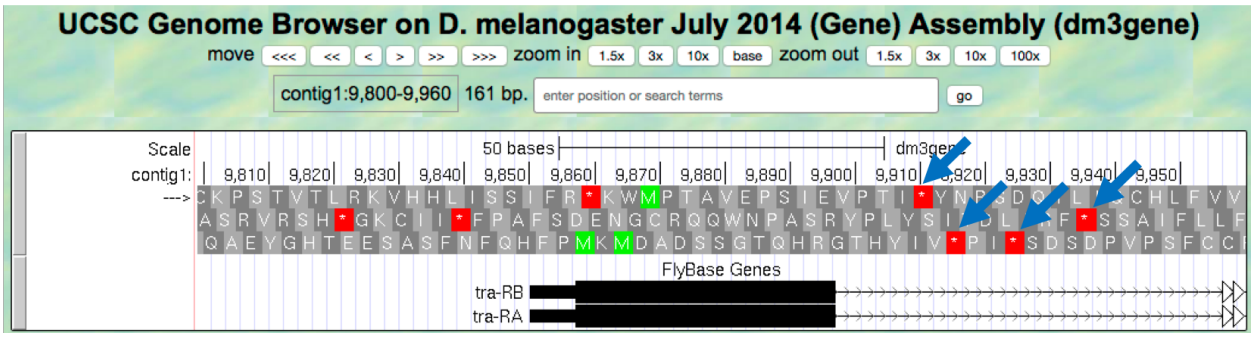


Figure 1.21.: Stop codons (red stars) are found in all three reading frames in the first intron of tra-RA.

27. Change the "enter position or search terms" field to `contig1:10,100-10,600` so that we can examine the second *coding exon* of the A isoform of *tra* to determine its reading frame.

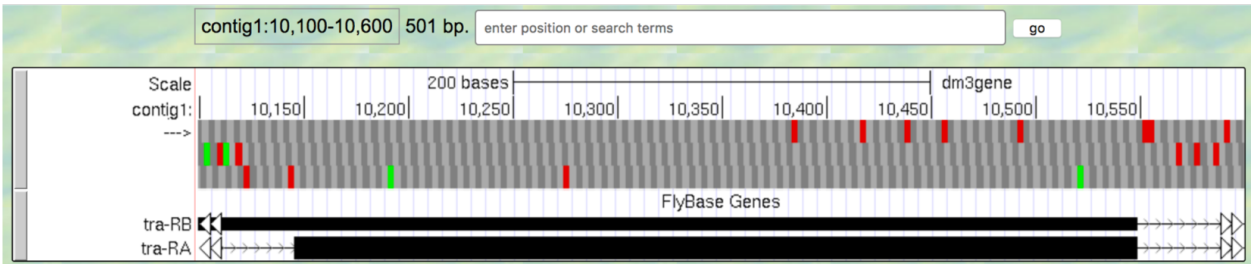


Figure 1.22.: The genomic region surrounding the second coding exon of tra-RA.

Question 9

Based on the screenshot shown in (Figure 1.22), which reading frame contains the amino acid sequence for the second coding exon of tra-RA?

28. Change the "enter position or search terms" field to `contig1:10,550-10,900` so that we can examine the

region surrounding the last coding exon of the tra-RA isoform (Figure 1.23). Based on our previous analysis, we know that there is a stop codon in the second row that corresponds to the transition from the translated (thick rectangle) to the untranslated (thinner rectangle) region of the mRNA (Figure 1.15). Hence the last coding exon of tra-RA is in frame +2 (Figure 1.23).

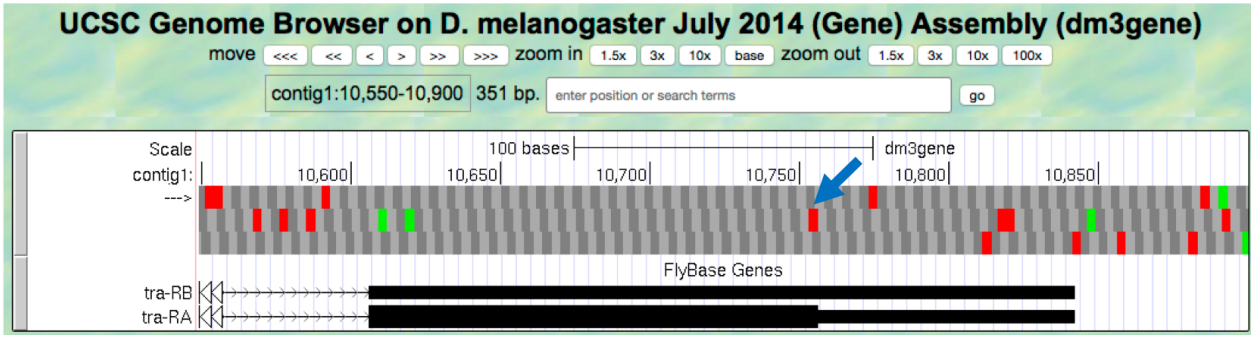


Figure 1.23.: The terminal coding exon of tra-RA is in frame +2.

Question 10

Does frame +2 have an ORF in the coding region of this exon? What about frame +1 and frame +3?

Question 11

Given that 3 of the 64 possible codons are stop codons, what is the chance of having a stop codon at any given position, assuming that the sequence is random?

Note: You might have noticed that the initial coding exon of tra-RA is in frame +3 while the last coding exon is in frame +2. We will learn more about mRNA processing in subsequent modules that will explain this apparent discrepancy.

1.5 Conclusion

In this lesson, you have learned to use the basic navigation features of the Genome Browser to examine the basic structure of a eukaryotic gene. To summarize:

- Genes provide the information to make proteins. This information is captured by transcribing the DNA to make RNA, and is carried on the mRNA in the form of three-base groups called codons.
- Genes are composed of exons and introns. Exons are regions retained in the processed mRNA, and are represented by black blocks in the browser, while introns are the regions that are removed during the process of creating the final mRNA, and are represented by lines connecting the blocks.
- The codon ATG in DNA (AUG in mRNA) specifies the amino acid M (Methionine) and is highlighted in green on the "Base Position" track of the Genome Browser. The first Methionine provides the starting signal for protein synthesis.
- The codons TAA, TAG, and TGA in DNA (UAA, UAG, and UGA in mRNA) encode the stop codon (*) and are highlighted in red on the "Base Position" track of the Genome Browser. The stop codons provide the ending signal for protein synthesis.

- Genes may be read either from left to right (top strand of the DNA), or from right to left (bottom strand of the DNA). Arrows on a gene indicate its directionality.
- Each row in the "Base Position" track (set on `full`) corresponds to a different reading frame. Different coding exons for a transcript can be in different reading frames.

29. To practice using the browser and reinforce the above concepts, examine the third gene in this contig (spd-2-RA):

Question 12

How many exons and introns are present in this gene?

Question 13

What is the orientation of this gene relative to contig1? How do you know? Where are the start codon and the stop codon — give the base position numbers (coordinates) of the start and the stop codon)?

You have now completed Module 1, and are ready to move on to *Module 2*.

1.6 Bonus question

Take a little time to explore some of the other evidence tracks on the browser.

Bonus Question 14

While looking at contig1 (size 11,000 bp), put the "GC Percent" track on `full`. What sort of pattern do you see, relative to the map of the genes? What can you conclude about gene structure?

CHAPTER 2

Module 2

Module 2: Transcription, Part I: From DNA sequence to transcription unit

Author Maria S. Santisteban (University of North Carolina - Pembroke)

Last Update May 27, 2019

Version 0.0.1

2.1 Investigation 1: Identify the Transcription Unit

2.1.1 Introduction (Investigation 1)

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.

2.1.2 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* (*transformer*) gene and the very end of the previous gene *spd-2* (*spindle defective 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 2.1)

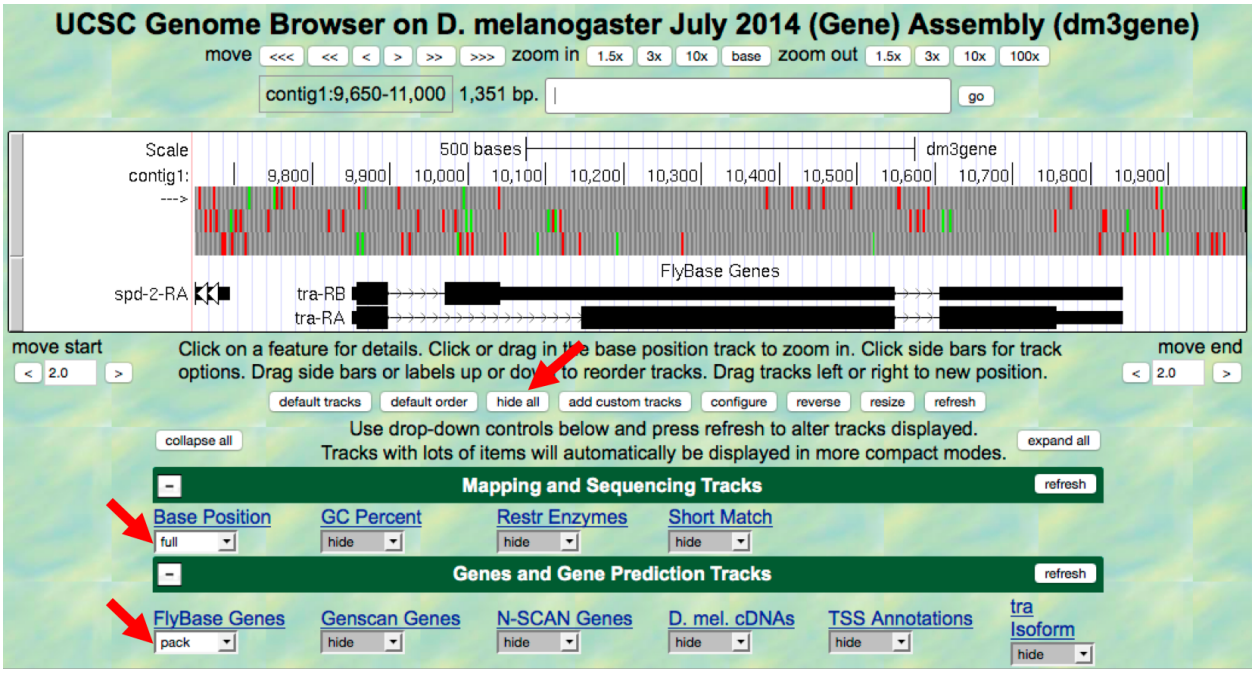


Figure 2.1.: Configuring the display modes for the evidence tracks surrounding the *tra* gene.

Tip: 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`.

2.1.3 Identifying the transcription unit for the *tra* gene

5. 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 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 2.2). These represent two alternative forms known as **isoforms** of the *transformer* (*tra*) gene product.

6. 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

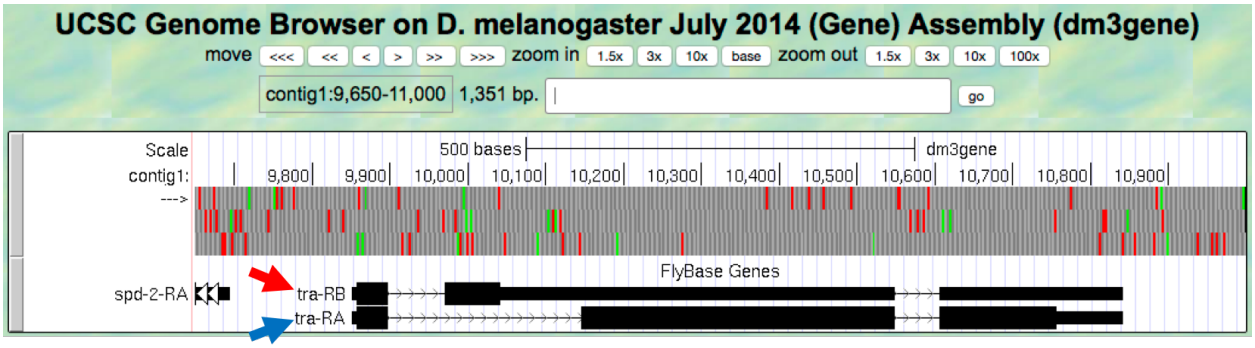


Figure 2.2.: FlyBase annotated isoforms A (blue arrow) and B (red arrow) of *tra* in *D. melanogaster*.

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*, 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.

Question 1

What is the span — the start and end base positions — of the 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 2.3).

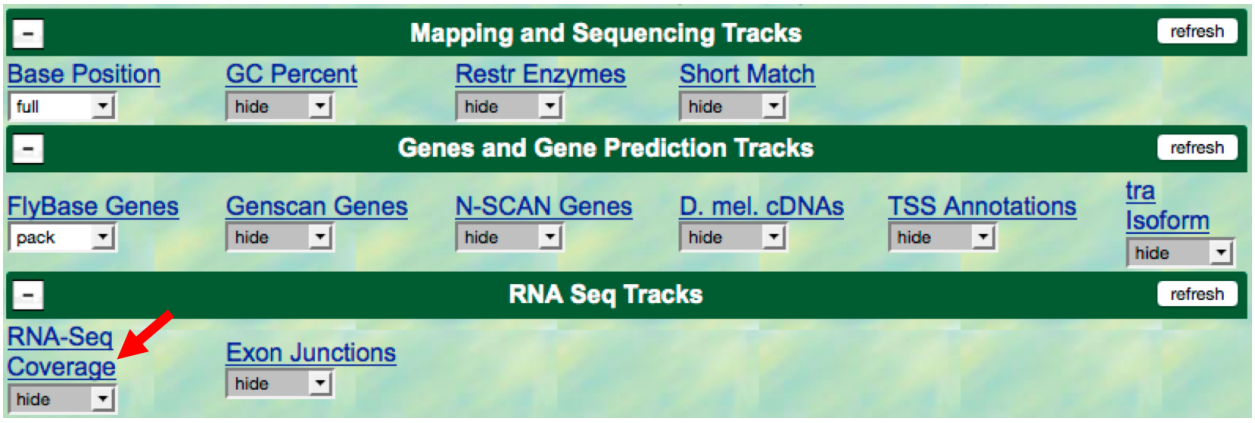


Figure 2.3.: Click on the "RNA-Seq Coverage" to configure the display settings for this evidence track.

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 2.4):

- 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)

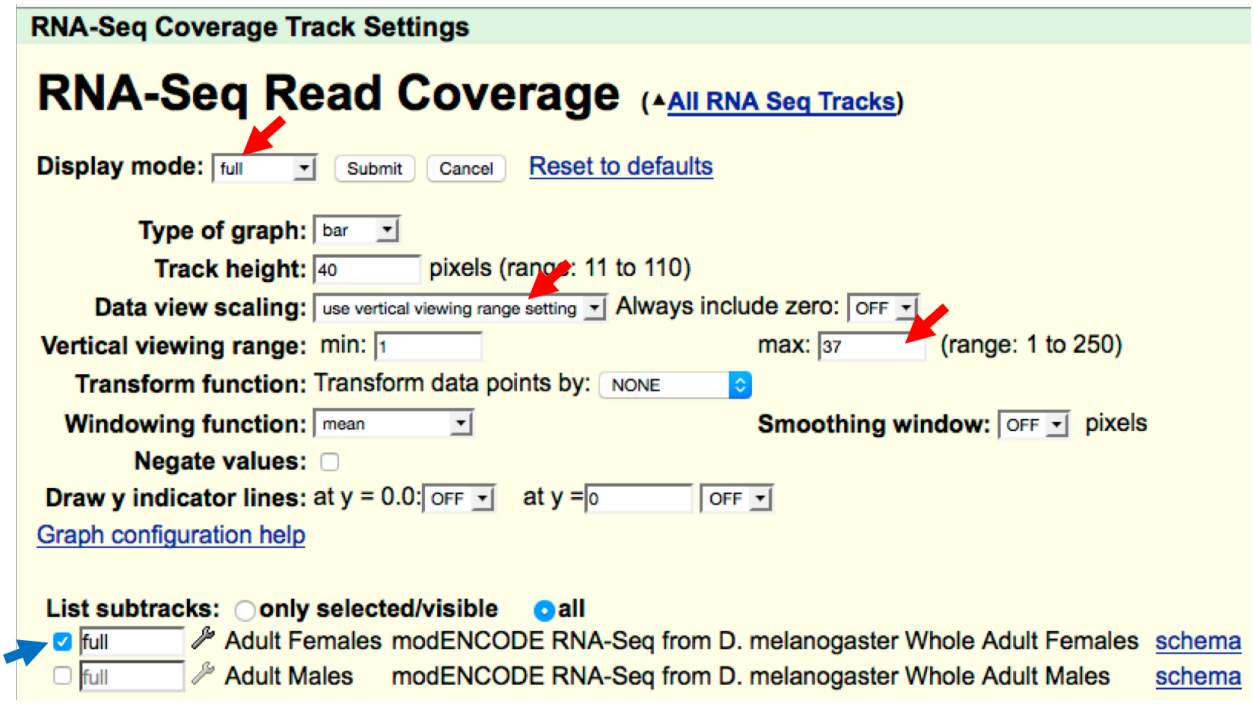


Figure 2.4.: Manually define the viewing range for the RNA-Seq Read Coverage track (red arrows) and select only the sub-track of interest (i.e., Adult Females, blue arrow).

Note: By default, the RNA-Seq Coverage track will auto-scale based on the read depth (that is, the number of reads) 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, 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*.

---

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 2.5). The y-axis corresponds to the number of RNA-Seq reads from whole adult females that have been mapped to each genomic position of this portion of contig1.

Question 2

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

Question 3

Most of the RNA-Seq reads come from mature (processed) RNA. Can you use this data to suggest where introns are

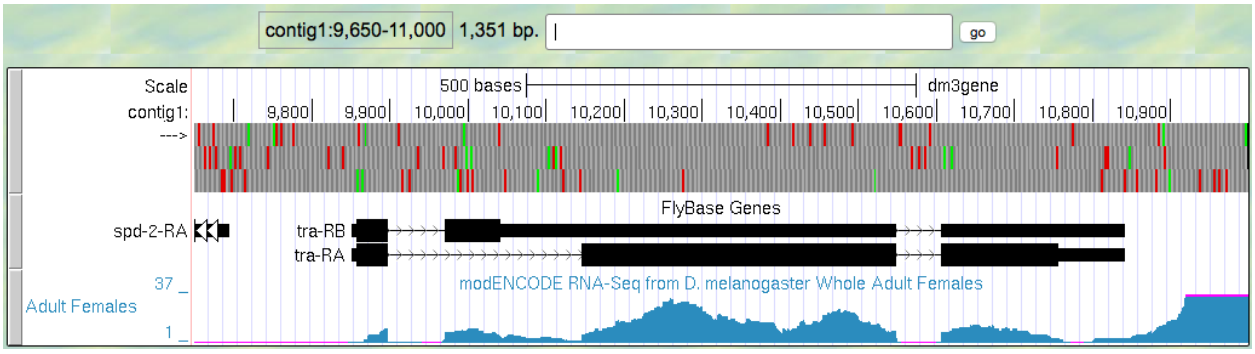


Figure 2.5.: RNA-Seq read coverage track (blue) for the *D. melanogaster* adult female sample.

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.

2.2 Investigation 2: Identify the 5' end of the transcription unit

2.2.1 Introduction (Investigation 2)

Previous studies have identified sequence motifs that are enriched in the region surrounding a gene's Transcription Start Site (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 2.6).

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

---

Question 4

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 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 the tra TSS location.

[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

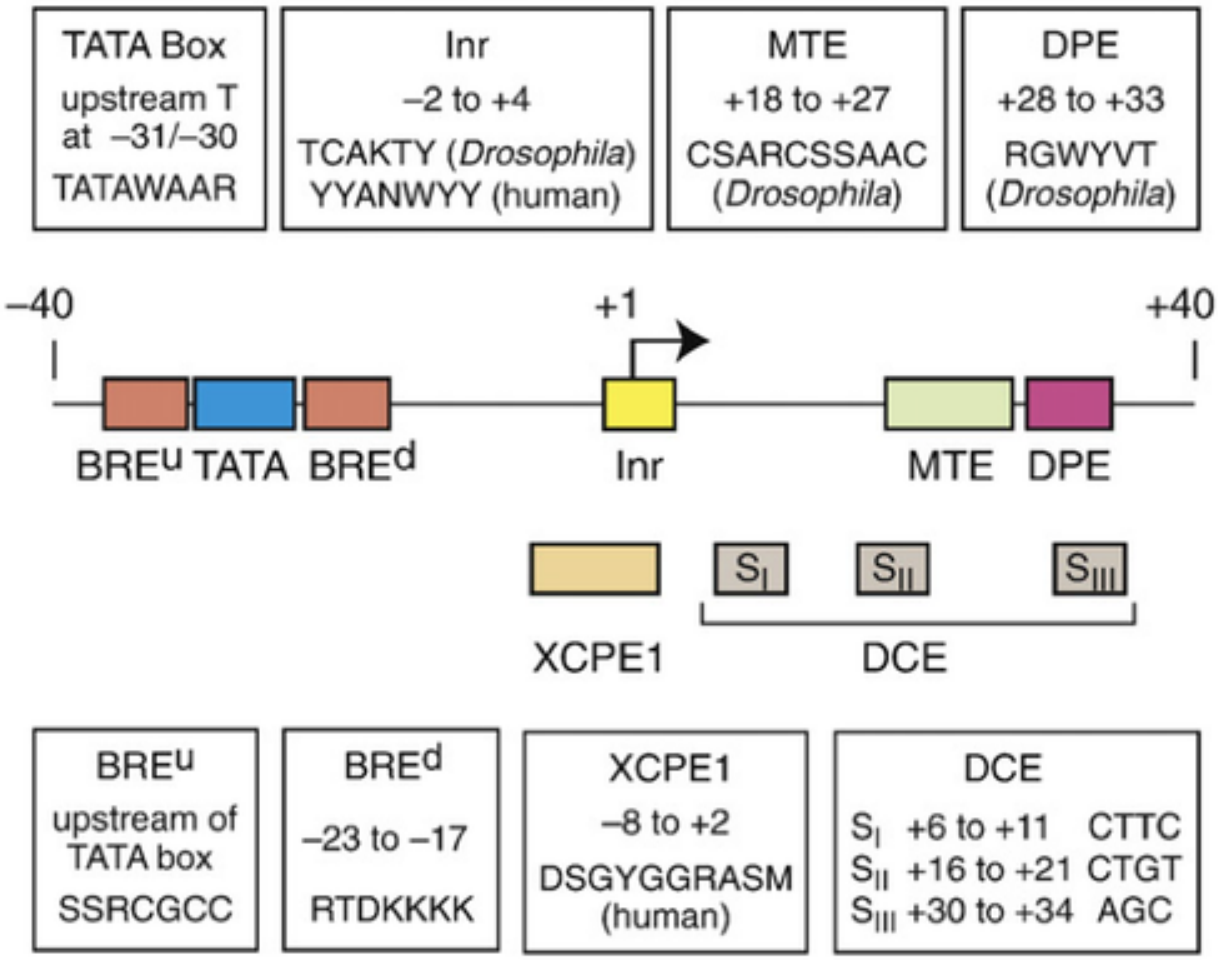


Figure 2.6.: Motifs that are enriched near the transcription start sites of many eukaryotic genes (Juven-Gershon T and Kadonaga JT, 2010[1]). Note that the motifs are often "degenerate," N = any base, R = purine (either A or G), Y = pyrimidine (C or T), K = keto (T or G), M = amino (C or A), S = strong (G or C), W = weak (A or T), V = A/G/C (not T), D = A/G/T (not C).

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 `contig1:9,700-9,900` 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 2.7).

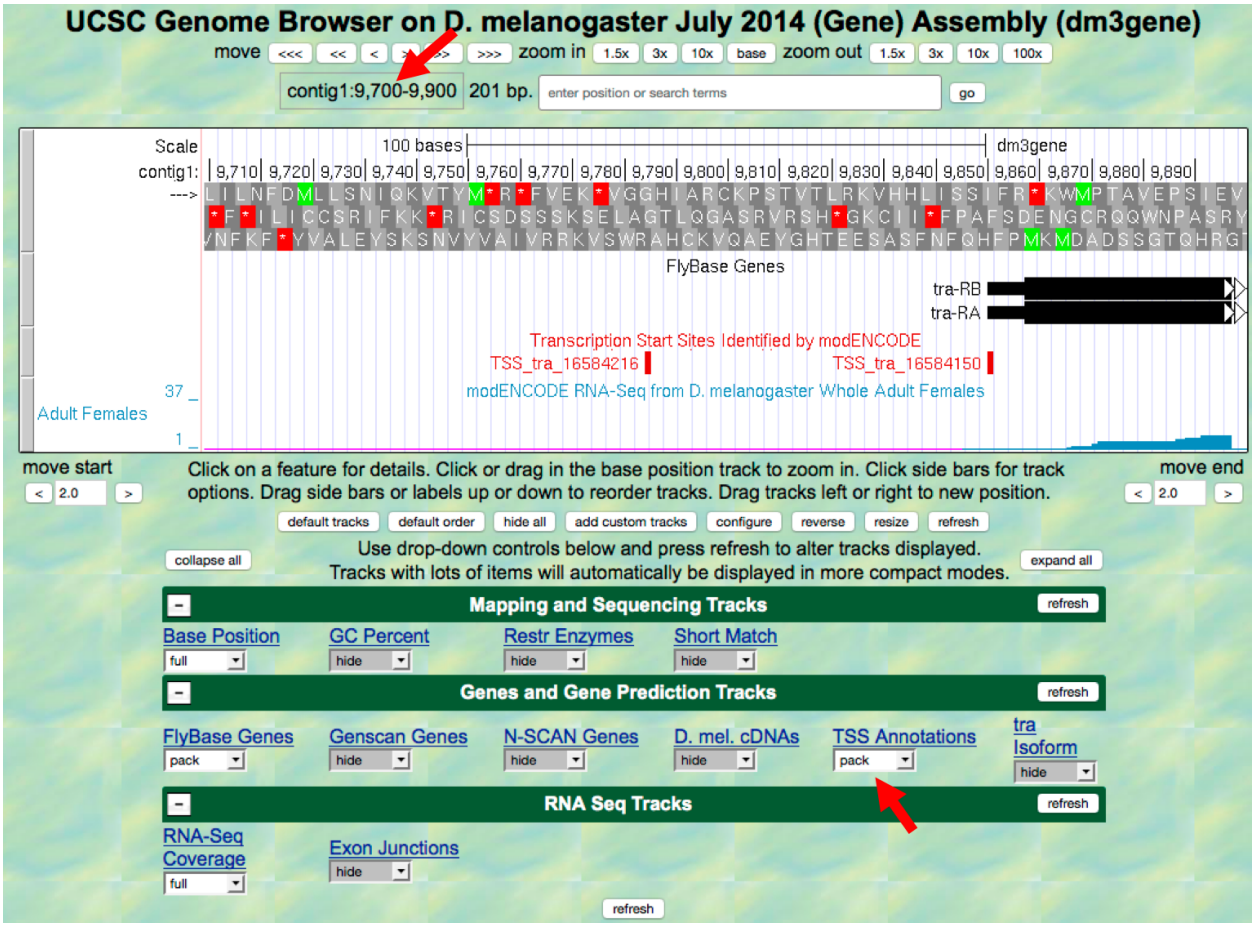


Figure 2.7.: Annotated TSS's in the region surrounding the start of the *tra* transcripts

Question 5

How many TSS sites were identified using this technique?

Question 6

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

2. 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 2.8).

Question 7

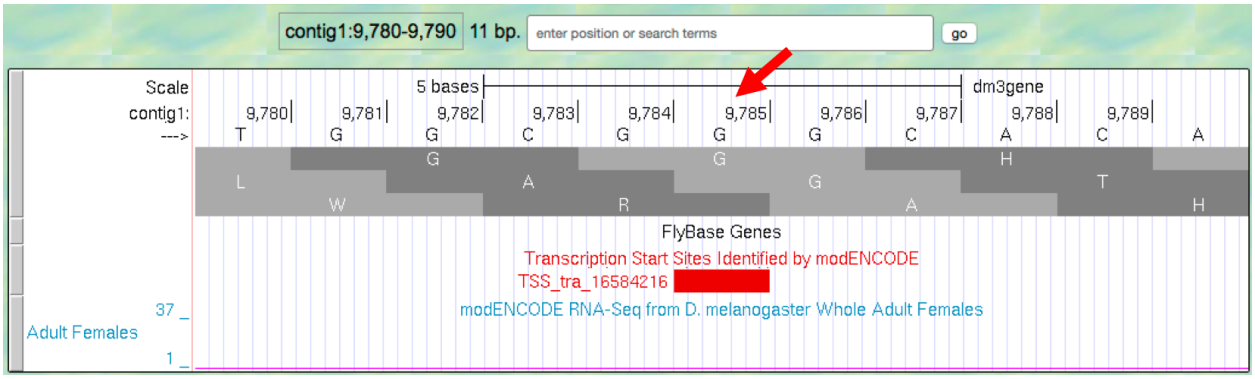


Figure 2.8.: Determine the position for the annotated TSS "TSS_tra_16584216".

What is the coordinate for TSS_tra_16584216?

3. Now let's zoom in to the second TSS site, `TSS_tra_16584150` (Figure 2.9).

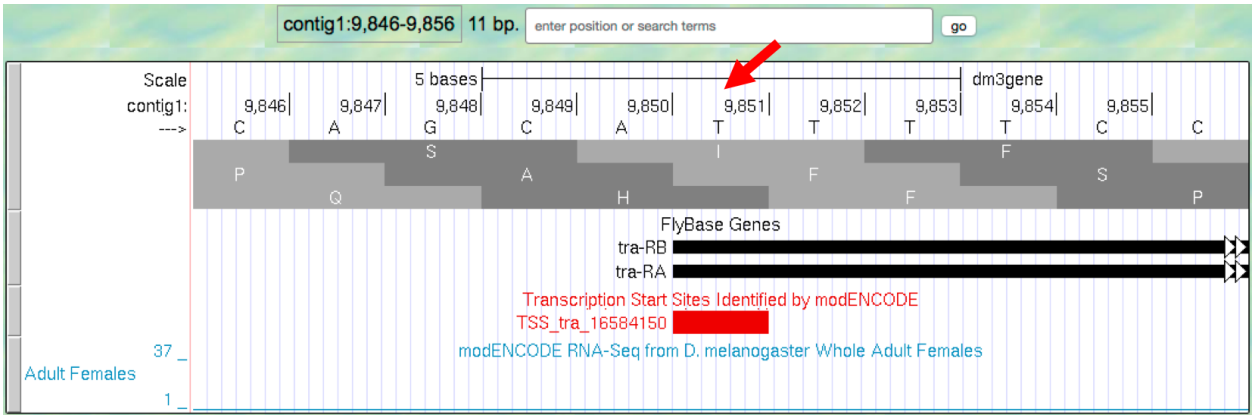


Figure 2.9.: Determine the position for the annotated TSS "TSS_tra_16584150".

Question 8

What is the coordinate for this TSS?

4. We will use the Genome Browser to gather additional evidence to identify the most likely TSS. First, let's search for the Inr motif 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 motif 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 `contig1:9,700-9,900` and then click `go`.
5. To learn more about the Short Match functionality, watch the Short Match video. 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 2.10). 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**Y**rimidine) corresponds to either C or T.

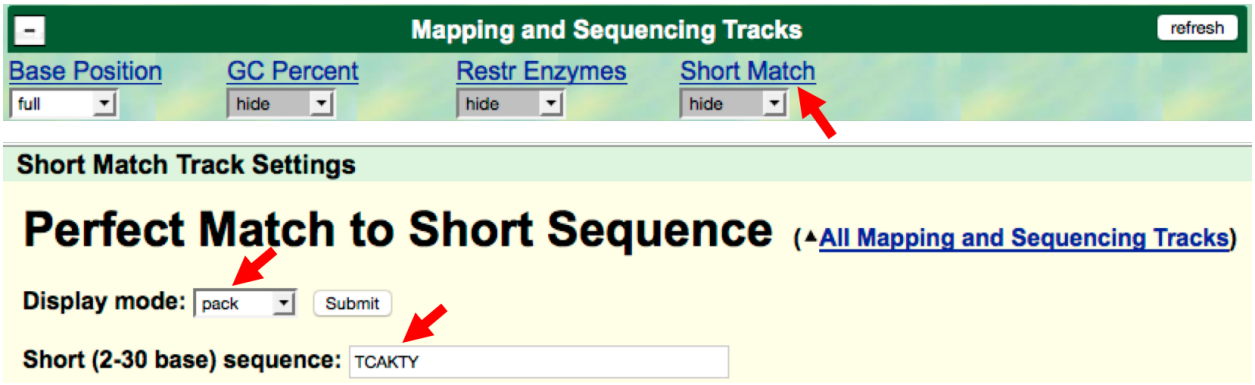


Figure 2.10.: Configure the "Short Match" track to search for the Initiator (Inr) motif.

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.

Question 9

Are there any perfect matches to the Inr consensus sequence in the region between 9,700-9,900? What are the coordinates and orientation of these matches?

Question 10

Which base position(s) would you assign as the TSS of the *tra* gene based on the available evidence? Describe your reasoning.

Question 11

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 a different promoter region. Navigate to the Genome Browser Gateway page by clicking on the `Genomes` tab at the top of the page, and select the *D. melanogaster* `Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6)` assembly. Change the "Position/Search Term" field to `chr2R:18,867,350-18,867,430` and then click on the `go` button.

7. Click on `hide all` and then enable the tracks listed below.
8. Under "Mapping and Sequencing Tracks":
   - Base Position: `full`
   - Short Match: `pack`
9. Search for **TCAKTY**, the Inr consensus sequence.
   - Click on the `Short Match` link under "Mapping and Sequencing Tracks."
   - Type `TCAKTY` in the "Short (2-30 base) sequence" field.
   - Click on the `submit` button.

10. Under "Genes and Gene Predictions Tracks":
    - FlyBase Genes: `pack`
11. Under "Expression and Regulation":
    - TSS (Embryonic) (R5): `pack`
12. Click on a `refresh` button. Record the position(s) and orientation(s) of any matches to the Inr motif.
13. Repeat the search for `TATAWAAR` (the TATA Box motif).

Question 12

Are there any perfect matches to the Inr consensus sequence (Figure 2.11)? What are the coordinates and orientation of these matches? What about the TATA Box motif? Are these signals in good agreement with the beginning of the transcription unit?

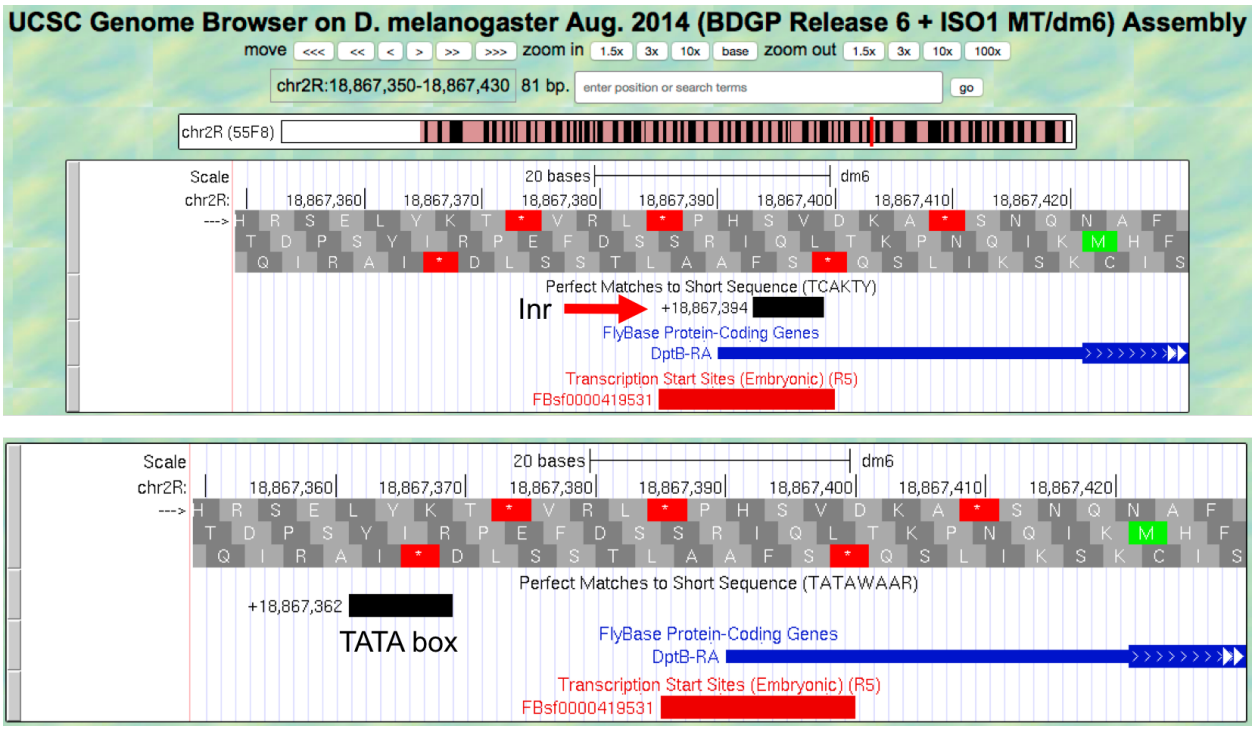


Figure 2.11.: USE of the "Short Match" track to search for the Inr and TATA box motifs.

2.3 Investigation 3: Map the 3' end of the transcription unit

2.3.1 Introduction (Investigation 3)

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 thymine versus RNA, which has uracil as the pyrimidine base that pairs with A.

Question 13

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 which direction?

Question 14

RNA polymerase binds to the **promoter sequence on the template strand**, constructing the transcribed mRNA in which direction?

Note: 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[2]. 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 be further studied in *Module 3* and *Module 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 signal (AATAAA)[2]. 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 2.12).

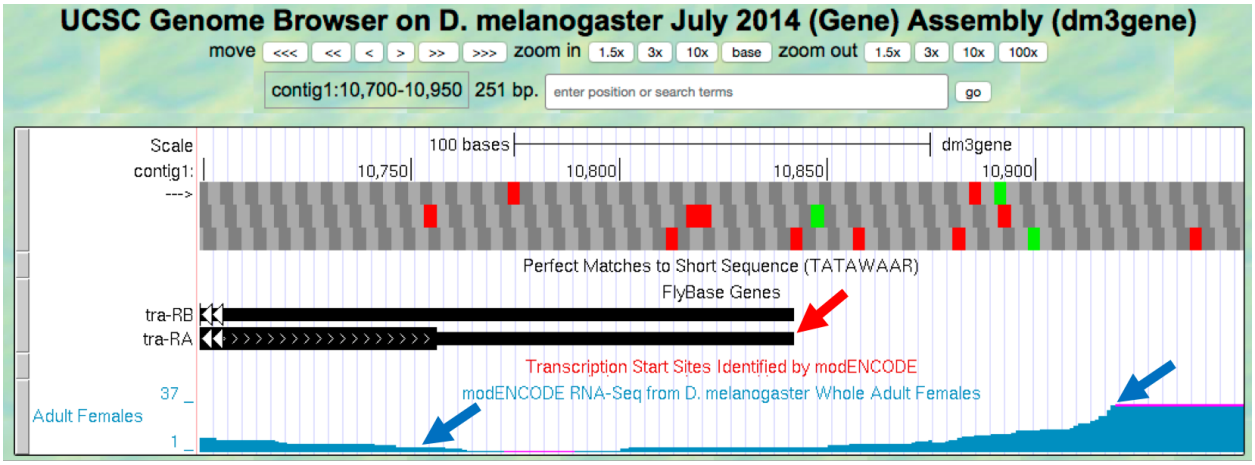


Figure 2.12.: Comparison of the transcription end site annotated by FlyBase (red arrow) versus changes in RNA-Seq read coverage in whole adult females (blue arrows).

Question 15

[2] This signal is also referred to as the poly-A signal because a poly-A tail is added to the mRNA at its 3' end. In the next Module we will refer to it as a poly-A signal.

At which base position do you see a substantial decrease in RNA-Seq read coverage in the whole adult female sample?

Question 16

What is the coordinate of the 3' end 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 consensus 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 2.13). Click on the `Submit` button.

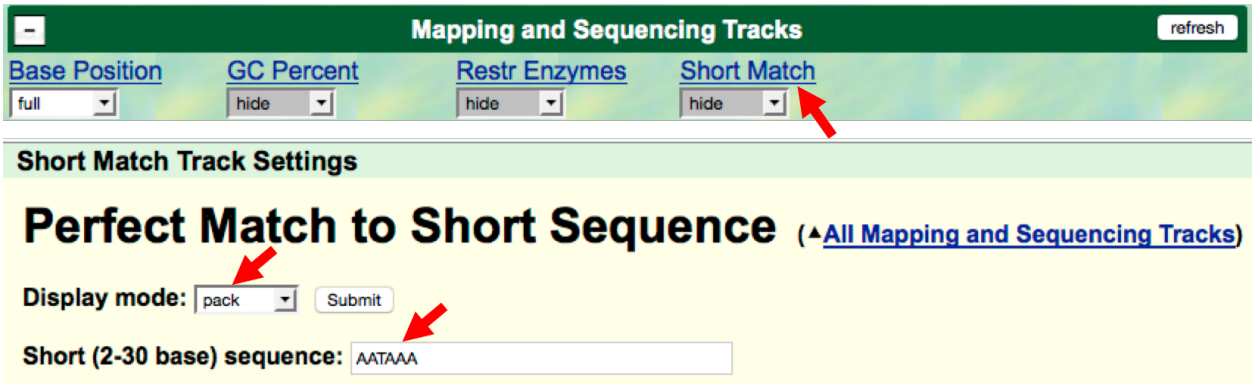


Figure 2.13.: Use the "Short Match" track to search for the mRNA termination signal.

Question 17

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

Question 18

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

Question 19

Is the sequence(s) you found in the question above contained within the 3' untranslated 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.

Question 20

Based on your analysis above, which position is the best choice for the termination signal? Describe your reasoning.

2.4 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 transcription 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:

Question 21

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.

Question 22

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

Question 23

If these regions with no RNA-Seq coverage occur within an initial transcript, what could have happened to these RNA sequences?

2.5 Footnotes

CHAPTER 3

Module 3

Module 3: Transcription Part II: What happens to the initial (pre-mRNA) transcript made by RNA pol II?

Authors S. Catherine Silver Key (North Carolina Central University) and Chiyedza Small (Medgar Evers College CUNY)

Last Update May 27, 2019

Version 0.0.1

3.1 Introduction

In *Module 2*, you identified the transcription start site (TSS) for the A *isoform* of the *tra* gene (tra-RA). In this module, we will explore each of the three steps of *pre-mRNA* processing.

3.1.1 Setting up our Browser page (review):

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. As you may remember from *Module 1*, contig1 is derived from chr3L in the *D. melanogaster* genome. This contig contains three different genes (*CG32165*, *spd-2*, and *tra*). Enter `contig1:9,500-11,000` into the "enter position or search terms" textbox and then click on the `go` button to navigate to the genomic region surrounding the *tra* gene.
3. Because the Genome Browser remembers your previous display settings, you should click on the `default tracks` button to reset the display to the default settings. Change the display mode for the "Base Position" track to `full` and verify that the "FlyBase Genes" track is set to `pack`. Click on the `refresh` button.
4. Scroll down to the "RNA Seq Tracks" section and then click on the `RNA-Seq Coverage` link. Change the track display settings to the following, as we did in *Module 2*:

- 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 subtracks" section, select BOTH the `Adult Females` and the `Adult Males` subtracks (Select the check box next to subtrack to turn the subtrack on.)
- Click on the `Submit` button. Verify that the RNA-Seq Coverage track on the browser page is set to `full`.

3.2 Investigation: mRNA processing

The processing of pre-mRNA into mRNA involves three key steps (Figure 3.1)

- The addition of a **5' cap**
- The addition of a **3' poly(A) tail**
- The removal of introns through **splicing**

Removal of the *introns* during this process results in adjacent *exons* being brought together in the final mRNA message.

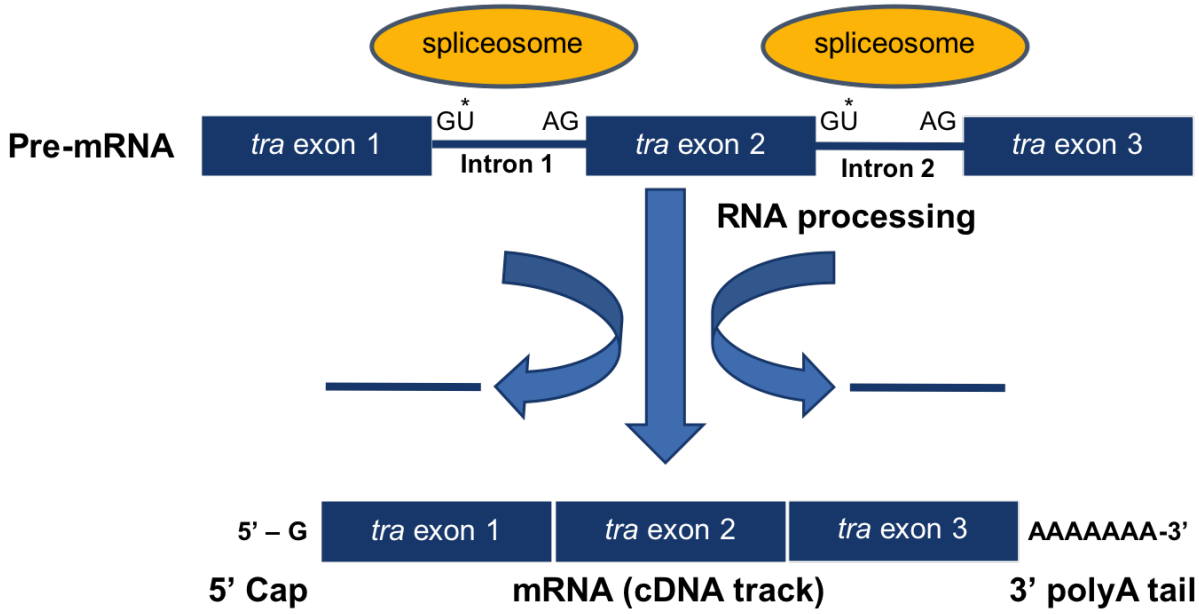


Figure 3.1.: Diagram of mRNA processing that converts a pre-mRNA to a processed mRNA.

3.2.1 Addition of a 5' cap

The **first step in pre-mRNA** processing occurs at the *5'* end of a messenger RNA. Recall that mRNA is synthesized in a 5' to 3' direction, so the 5' end of the mRNA was synthesized first. Let's examine the beginning of the *tra* gene. Type `contig1:9,825-9,870` into the "enter position or search terms" textbox and then click on the `go` button.

In *Module 2*, we identified the **transcription start site** (TSS) of the A isoform of *tra* at position **9,851**.

To show the TSS's that have been annotated by the modENCODE project, scroll down to the "Genes and Gene Prediction Tracks" and change the display mode for the "TSS Annotations" track to `pack`, and click `refresh`. The modENCODE project looked for TSSs by using a chemical method to tag the special structure that occurs at 5' ends of transcript, fishing out the RNA molecules that carried these tags, and mapping the sequence back to the genome, a method called "CAGE" (cap analysis of gene expression).

In addition, we will also display the "D. mel. cDNAs" track (also under the "Genes and Gene Prediction Tracks" section); change this to `pack`. This track shows the alignment of *D. melanogaster* *cDNAs* (complementary DNAs, made by copying the mRNA) that have been sequenced by the Berkeley Drosophila Genome Project (BDGP). Click on the `refresh` button (Figure 3.2). These two tracks both provide an analysis based on the RNA population, and mapping the positions of these sequences indicates where the transcript started.

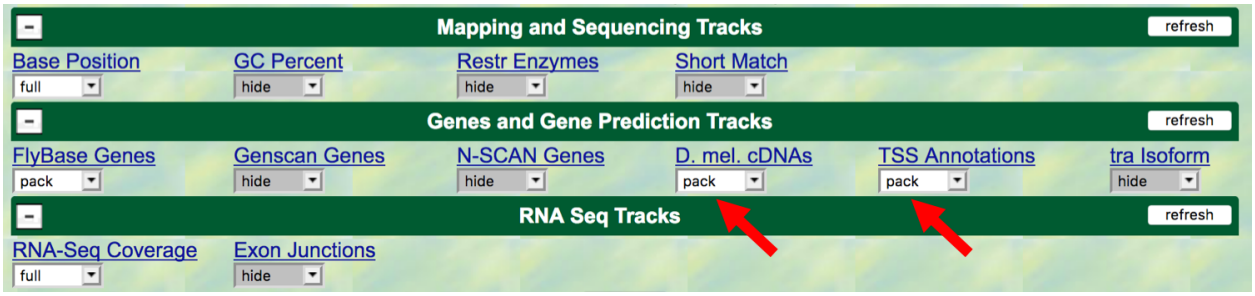


Figure 3.2.: Change the display modes of the "D. mel. cDNAs" and "TSS Annotations" tracks to "pack".

Remember from *Module 2* that we also found a match to TCAKTY, a common initiation signal just upstream, at 9,834 (display this using the "Short Match" track). All of these pieces of evidence argue for a TSS in this region.

The new Genome Browser image (Figure 3.3) shows the 5' end of the pre-mRNA transcript (i.e. the start of *transcription*) based on the CAGE experiment (modENCODE track) with the additional lines of support. On this end of the pre-mRNA, a modified guanine nucleotide (7mG) is added to the nucleotide at position 9,851, forming the **5' cap**. Note that this additional nucleotide is **NOT** visible in the DNA track. It is added AFTER the transcript is made. This is the **first step** in **pre-mRNA processing**: capping.

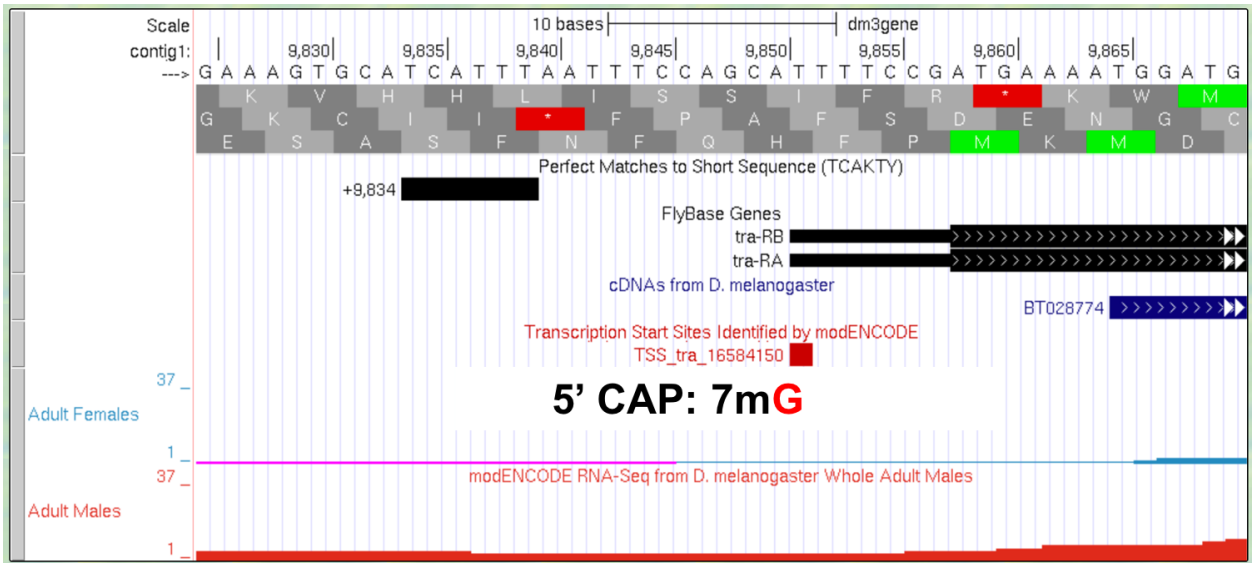


Figure 3.3.: Addition of a 5' cap to the 5' end of the transcript.

Question 1

What is the coordinate of the first nucleotide that is transcribed? In the DNA sequence, is it an A, C, T or G?

Question 2

What are the coordinates for the start codon that codes for the first amino acid of the A isoform of the *tra* gene? (Assume reading frame +3.)

Question 3

The region of the transcript from the 5' cap to the nucleotide just upstream of the start codon is called the 5' untranslated region (5'UTR) because it is part of the transcript that is not translated. How long (in ribonucleotides) is the 5'UTR?

3.2.2 Addition of a 3' poly(A) tail

The **second step** in pre-mRNA processing is **polyadenylation**.

5. To view the *3'* end of the tra-RA gene, change the "enter position or search terms" field to `contig1:10,633-11,000` and then click on the `go` button.

Polyadenylation means that **many** (poly) **adenine nucleotides** (ribonucleotides) are added to the 3' end of the pre-mRNA **AFTER** transcription termination. This generates a *poly-A tail* (typically ~20 to ~250 As) that will be retained in the final mRNA but it is not present in the "Base Position" track of the Genome Browser. This is because the poly-A tail does not exist in the DNA template but is simply added to the RNA by a special polymerase as a long run of adenine nucleotides.

Our previous analysis in *Module 1* has shown that the last *coding exon* of tra-RA is in frame +2 and the *stop codon* is located at 10,754-10,756. We can use the Genome Browser to determine the end of the tra-RA transcript indicated by the cDNA track (in blue). (Note that this aligns with the cDNA although there is some discrepancy between the two as to the exact end of the transcript.)

Question 4

How long (in base pairs) is this 3' untranslated region (3'UTR) as indicated by the cDNA track (in blue)?

Question 5

Zoom into the 3' end of the FlyBase Gene, near the termination site. What is the longest stretch of A nucleotides that you observe?

Question 6

Do your findings support the conclusion that the poly(A) sequence observed in the mature mRNA transcript is not in the template DNA?

6. Perform a "Short Match" search for the poly-A signal (`AATAAA`) using the protocol you learned in *Module 2*. This search should place the poly-A signal at 10,818-10,823 (Figure 3.4). As mentioned in *Module 2*, the transcript is cleaved 11 to 30 nucleotides downstream of the poly(A) signal sequence, and then 150-200 adenines

are added to the pre-mRNA. The nucleotides between the stop codon and the end of the poly-A tail comprise the 3' *UTR*.

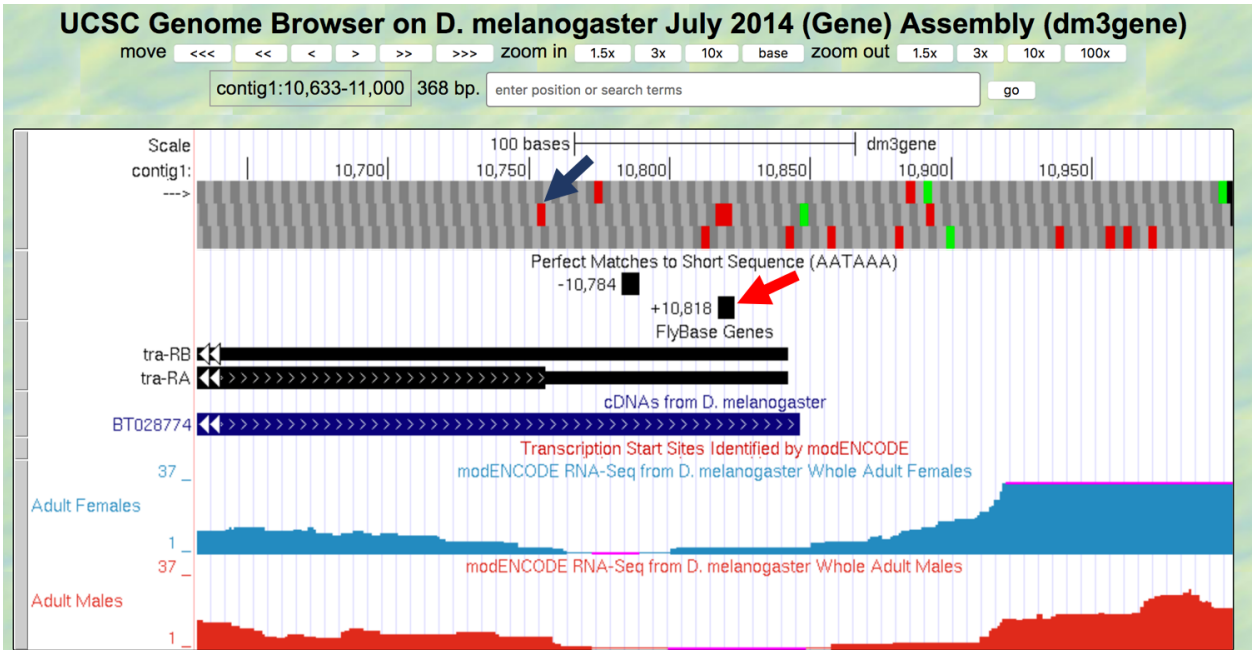


Figure 3.4.: Previous analysis placed the stop codon for the A isoform of *tra* in frame +2 (blue arrow) and the poly-A signal at 10,818-10,823 (red arrow).

We can see the polyadenylation sequence that is associated with the processed mRNA by examining the cDNA (BT028774) that has been aligned to this region.

7. Click on the `BT028774` feature under the "cDNAs from *D. melanogaster*" track and then click on the `View details of parts of alignment within browser window` link (Figure 3.5).

The next figure shows the actual alignment between the *D. melanogaster* cDNA BT028774 and the genomic sequence in contig1 (Figure 3.6). Nucleotides that are identical between the two sequences are shown in blue capital letters while nucleotides that differ are shown as black lowercase letters. The light blue *bases* denote the start and the end of the gap in the alignment. The side-by-side alignment shows the pairwise alignment between the cDNA (top) and the contig1 sequence (bottom) within the viewing region (i.e. contig1:10,633-11,000).

Question 7

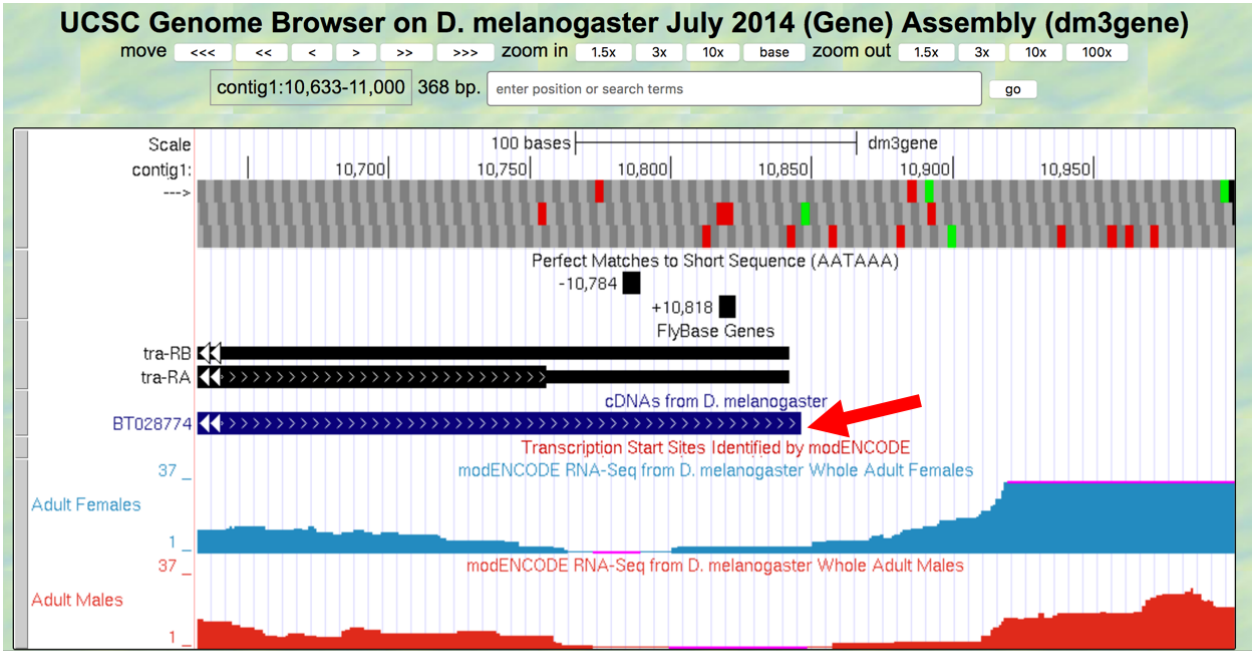
Scroll up to the `cDNA BT028774` area. After which coordinate (number in the cDNA) do you see the polyadenylation track (in lower case black letters)?

Question 8

How many "A" ribonucleotides have been added to the *tra* mRNA (represented in the cDNA)?

Question 9

Locate the `AATAAA` termination signal in the cDNA sequence. How many nucleotides 3' of the final "A" in the signal sequence does the poly(A) run start? (This number should be between 11–30 nucleotides.)



cDNAs from D. melanogaster (BT028774)

GenBank Record: BT028774

BT028774/Genomic Alignments

BROWSER	SIZE	IDENTITY	SCAFFOLD	STRAND	START	END	QUERY	START	END	TOTAL
browser	852	99.6%	contig1	+	9865	10846	BT028774	1	852	885

View details of parts of alignment within browser window.

View table schema

Go to D. mel. cDNAs track controls

Figure 3.5.: Examine the alignment of *D. melanogaster* cDNA BT028774 against contig1.

Alignment of BT028774

BT028774
BT028774 in browser window
D. melanogaster.contig1
together

cDNA BT028774

```
TGGATGCCGA CAGCAGTGGA ACCCAGCATC GAGTGTCATA TTGTGTGAAA  50
TGTGAAATGG ATGAGAATGA ACGCTGGACG ACCAGACAGA GAAAGAGAAG  100
CTAGGACAAT AGGACTCTCA ACTGCGCATT ACGTGGATTC CGTCTCCGAC  150
GATGCGCCAA ACACTATGCG TTAGATGCAT TGACTAACCG CACACTCTTT  200
TCACATAGAT TCCCGTGGCT CAAGGTCTCG ATCCCGGCGA GAAAGAGAAT  250
ACCATGGGCG ATCAAGCGAG AGGGACAGCA GAAAGAAGGA GCACAAGATT  300
CCGTACTTTG CAGACGAGGT TCGAGAACAG GATCGGTTGA GAAGATTGCG  350
CCAAAGAGCG CACCAATCCA CCAGACGCAC TCGCTCCAGA TCCAGATCAC  400
AGTCGTCCAT CAGAGAGAGC AGGCACAGAA GGCATCGCCA GCGCTCTAGG  450
AGCCGCAATC GCAgCCGAAG TCGCAGCAGT GAACGAAAAC GCCGTCAACg  500
gAGCCGAAGT CGCAGCAGTG AACGAAGACG CCGTCAACGG AGCCCGCATC  550
GGTATAATCC TCCGCCAAAG ATCATCAACT ACTATGTGCA AGTGCCACCA  600
CAGGATTTCT ACGGGATGTC TGGcATGCAG CAAAGTTTTG GATACCAAAG  650
GCTACCACGT CCTCCGCCGT TTCCACCGGC CCCCTACAGA TACCGCCAGC  700
GACCGCCGTT CATTGGAGTT CCGCGATTTG GCTACAGAAA CGCGGGGCGT  750
CCCCCATATT GAACATACTC CATTCGACAT AATTACAAAT TTATTTACAT  800
TCGTGTGTTT TGTACATTAA AATAATAAAT ACATATATAT TTCAAACTAA  850
AAcaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaa
```

Genomic contig1 :

```
CACCACAGGA TTTCTACgtt agtagttatt tgtattgaaa caaataacaa  10582
atctaactta aatcgatttt gcagGGGATG TCTGGtATGC AGCAAAGTTT  10632
TGGATACCAA AGGCTACCAC GTCCTCCGCC GTTTCCACCG GCCCCCTACA  10682
GATACCGCCA GCGACCGCCG TTCATTGGAG TTCCGCGATT TGGCTACAGA  10732
AACGCGGGGC GTCCCCCATA TTGAACATAC TCCATTCGAC ATAATTACAA  10782
ATTTATTTAC ATTCGTGTGT TTTGTACATT AAAATAATAA ATACATATAT  10832
ATTTCAAACT AAAAatggag aacataaaat attgcaaacg gagaacgttg  10882
attaaatcat gagcaaatgc agcaaagaag gcggcaaaaa aagggtcaca  10932
cgagttagga agtg
```

Side by Side Alignment

```
00639 tggataccaaaggctaccacgtcctccgccgtttccaccggccccctaca 00688
>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>
10633 tggataccaaaggctaccacgtcctccgccgtttccaccggccccctaca 10682

00689 gataccgccagcgaccgccgttcattggagttccgcgatttggctacaga 00738
>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>
10683 gataccgccagcgaccgccgttcattggagttccgcgatttggctacaga 10732

00739 aacgcggggcgtccccccatattgaacatactccattcgacataattacaa 00788
>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>
10733 aacgcggggcgtccccccatattgaacatactccattcgacataattacaa 10782

00789 atttatttacattcgtgtgttttgtacattaaaataataaatacatatat 00838
>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>
10783 atttatttacattcgtgtgttttgtacattaaaataataaatacatatat 10832

00839 atttcaaactaaaa 00852
>>>>> ||||||||||||||| >>>>>
10833 atttcaaactaaaa 10846
```

Figure 3.6.: Alignment of the *D. melanogaster* cDNA BT028774 with the end of contig1.

3.2.3 Removal of introns through splicing

The final step in pre-mRNA processing is *splicing* out of introns and merging adjacent exons into one continuous open reading frame so that the mRNA is ready for *translation* into a protein.

8. Change the "enter position or search terms" field to `contig1:9,870-10,170` and then click on the `go` button to navigate to the first intron of the tra-RA transcription (Figure 3.7).

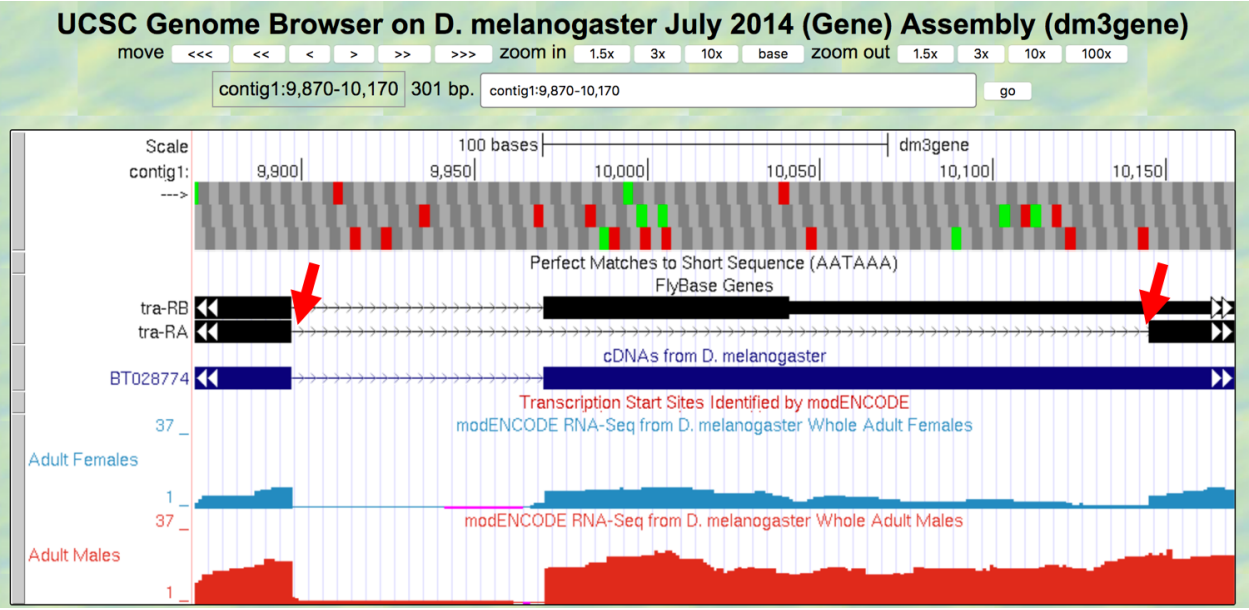


Figure 3.7.: The genomic region surrounding the first intron (red arrows) of tra-RA.

9. Zoom in to the region near the end of the first exon of tra-RA.

Question 10

Which two nucleotides are found just after the end of the first exon of tra-RA? Repeat this determination, identifying the two nucleotides at the start of intron 2 of tra-RA.

These two nucleotides are a signal for **splicing** to occur at the 5' end of an intron; these represent the first two bases of the intron, often called the donor site (or 5' splice site).

Question 11

At which base does exon 1 end?

---

10. Zoom out and then zoom in to the region near the beginning of the second exon of tra-RA.

Question 12

Which two nucleotides are found right before the start of tra-RA exon 2?

---

11. Zoom out and then zoom in to the region near the beginning of the third exon of tra-RA.

Question 13

Which two nucleotides are found right before the start of tra-RA exon 3?

These two nucleotides are the signal for **splicing** out of the 3' end of the intron, often called the acceptor site (or 3' splice site). These represent the last two bases of the intron.

Question 14

At which base does exon 2 of tra-RA begin? What is its coordinate?

3.3 Conclusions

In this module, we learned about the three key steps that are involved in converting the pre-mRNA into a *mature mRNA*:

1. The addition of a **5' cap**
2. The addition of a **3' poly(A) tail**
3. The removal of introns through **splicing**

Note: Introns are removed during this process and adjacent exons are brought together in the mRNA message.

After mRNA processing, the mature mRNA (tra-RA) can now exit the nucleus so that it can be translated into a protein (tra-PA) by the cytoplasmic ribosomes.

CHAPTER 4

Module 4

Module 4: Removal of introns from pre-mRNA by splicing

Author Meg Laakso (Eastern University)

Last Update May 27, 2019

Version 0.0.1

4.1 Investigation 1: Examining RNA-Seq data

We will continue to focus on *isoform* A of *transformer* (referred to as tra-RA). Here we will focus on data from experiments that assess the RNA population in cells. This data can be used to help us identify *exons* and *introns* for the gene under study.

All RNAs in the cell are collectively known as the "transcriptome," as almost all RNA is produced by *transcription* from a DNA template. (In some cases, RNA is made from an RNA template.) The transcriptome includes messenger RNAs, ribosomal RNAs, transfer RNAs, and other RNAs that have specialized functions in the cell.

RNA can be harvested from cells or a whole organism like *Drosophila* and converted to DNA, then sequenced to produce RNA-Seq (RNA Sequencing) data. First, extracted mRNA that has been fully spliced is copied back to DNA with the enzyme called **reverse transcriptase**. Short fragments of the copied or complementary DNA are sequenced, and then these segments are mapped back to the genome. By analyzing the mapping data, it is possible to know which and how many messenger RNAs have been synthesized.

This is a powerful technique that allows us to see when and where different genes are expressed. This kind of information can help researchers and clinicians know which genes are expressed in different types of cancer, for example. Here and in the next modules we are going to use RNA-Seq data to explore how the *transformer (tra)* gene is expressed in male vs. female *Drosophila*.

RNA-Seq data can indicate where transcription occurs, to the exact nucleotide. The number of RNA-Seq fragments that map to a given site also tells us how many copies of the RNA are present in the sample. Remember from *Module 3* that the initial RNA transcripts are quickly processed to remove the introns. Hence in total RNA from a cell, sequences

from exons will be much more abundant than sequences from introns. We will use RNA-Seq data to help us find the exon-intron boundaries for tra-RA, the isoform of the *tra* gene that is expressed in female fruit flies.

1. Open a web browser on your computer. Internet Explorer, Mozilla Firefox, Safari, or Chrome will work for this investigation. Go to the GEP Mirror of the UCSC Genome Browser 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.

Note: Reminder: Configure the Genome Browser Gateway page as follows:

1. Select `D. melanogaster` under "REPRESENTED SPECIES"
2. Select `July 2014 (Gene)` under "*D. melanogaster* Assembly"
3. Enter `contig1` into the "Position/Search Term" text box
4. Click on the `GO` button

As you will remember, this section of DNA is 11,000 *base pairs* long (Figure 4.1) and is part of the left arm of chromosome 3, which is about 28,100,000 bp long. If your Browser window is showing other evidence tracks, reset by clicking on `default tracks`.

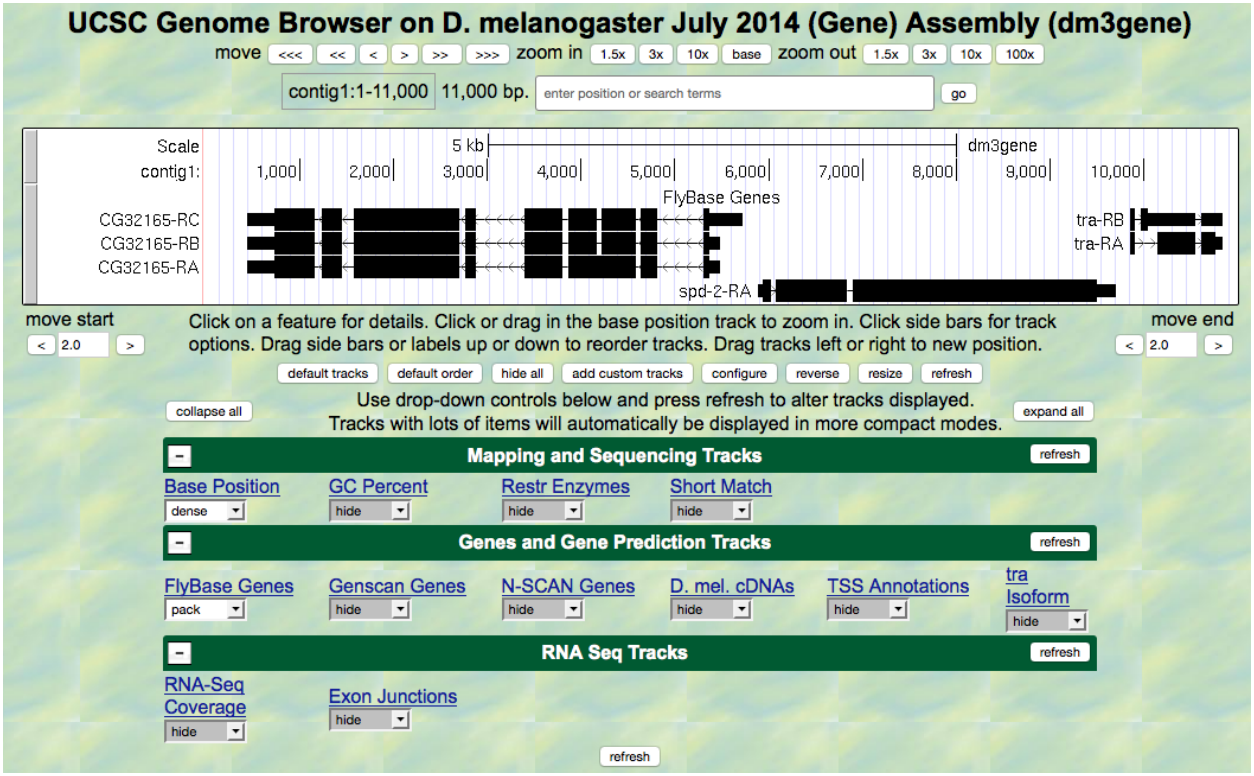


Figure 4.1.: A screen shot of the "contig1" project.

2. Let's start by setting the evidence tracks we want to see. Click on the `hide all` button. Then open only the tracks that will provide data for this investigation.
3. Change the display mode for the "RNA-Seq Coverage" track to `full`. Click on one of the `refresh` buttons.

You will see blue and red histograms representing RNA-Seq data generated using RNA samples from adult females and adult males, respectively. This allows us to infer the pattern of RNA synthesis, and to look for similarities and differences between the sexes. Both similarities and differences are apparent!

4. Zoom in to view only the *tra* gene by entering `contig1:9,800-10,900` in the "enter position or search terms" text box (Figure 4.2).

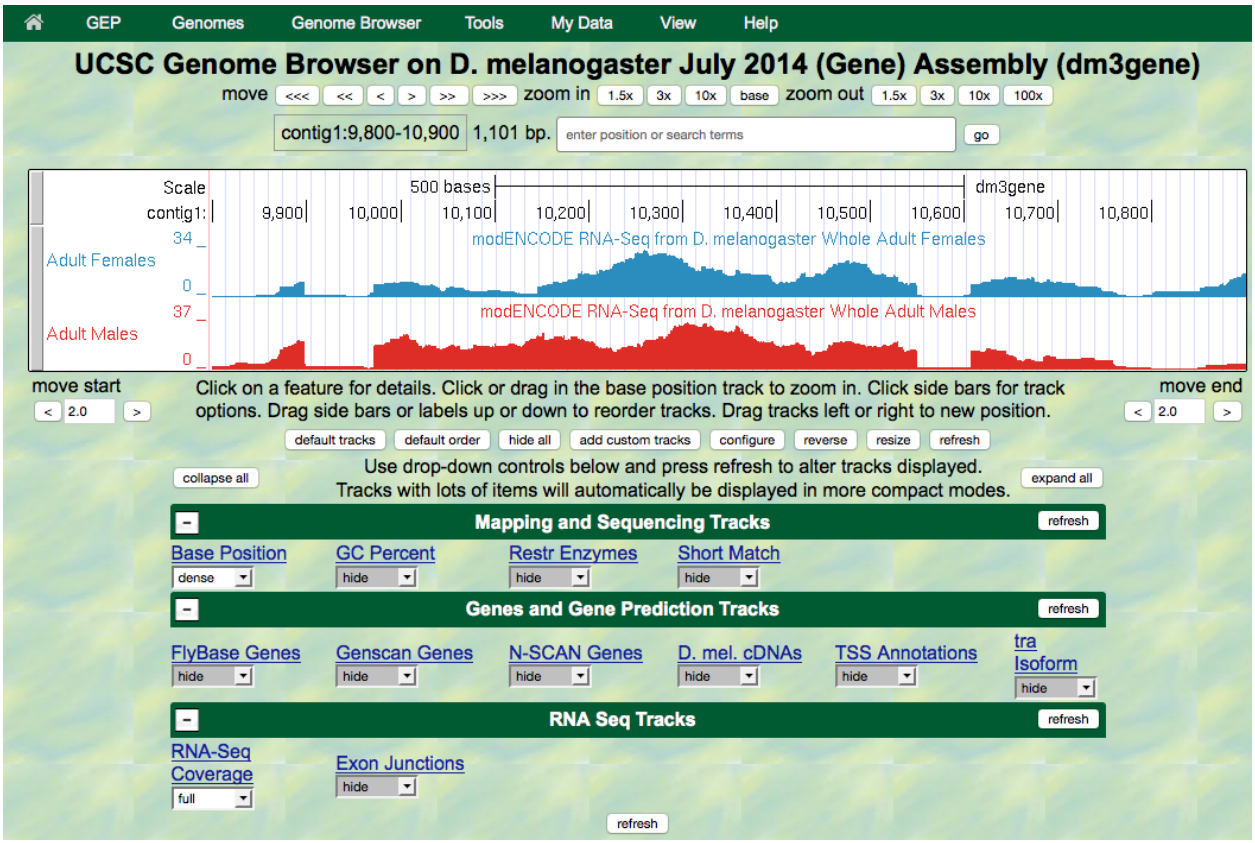


Figure 4.2.: Histograms representing RNA-Seq data in *transformer* (*tra*).

5. We are now looking at the region of chromosome 3 where the *tra* gene is located. Compare the blue and red histograms for Adult Females and Adult Males. Note that there are numbers on the y-axis that show how many RNA reads (sequences from transcripts) map to that position.

Question 1

List two ways in which the histograms are similar.

Question 2

List one way that the histograms differ (other than color).

To recap: The blue histogram represents the sequenced messenger RNA from female fruit flies, and represents the form of the *tra* gene referred to as isoform A (tra-RA). The red histogram represents the RNA-Seq data from male fruit flies. The mRNA in males is different from that in females, and this second isoform of the *tra* gene is called isoform B (tra-RB).

Question 3

Recall that our other evidence (see *Module 3*) indicates that tra-RA extends from 9,851 to 10,846. How many gaps do

you see in the histograms in this interval?

Each gap corresponds to an intron — there is very little RNA-Seq signal for that part of the gene because the intronic RNA was spliced out before the mRNA was collected and sequenced.

Question 4

How many exons do you see?

Remember that the exon is the "expressed" part of the gene and there will be either a sharp, or broad, peak in the RNA-Seq histogram.

Question 5

Do females or males make more *transformer* mRNA or do they express it at about the same level?

6. Now that you've examined the evidence yourself, let's go back and review.

   Gaps (introns) Figure 4.3 is a screen shot of the genome browser with gaps circled. Note that there are 2 gaps in females, and 2 gaps in males. The first gap in females looks a little strange because it doesn't have clean boundaries, suggesting a mixed population of processed transcripts.

   Exons Brackets have been drawn underneath the RNA-Seq data for Adult Female flies, corresponding to the three exons that are expressed.

   Isoforms Note that isoform A and isoform B of the *tra* gene are the result of *alternative splicing*. For other genes, isoforms may have **different transcription start sites**. Isoforms of a gene always have different mRNA sequences, but they may have the same protein sequence. To learn more about isoforms and genes, watch the Genes and Isoforms video.

Question 6

Using the information you've gathered so far, make a diagram of the tra-RA (female specific) isoform with 3 exons and 2 introns. Represent exons as rectangular boxes and introns as lines connecting the boxes. Number each exon and intron (start from left with "exon 1").

Question 7

Where is the promoter in relation to the exons and introns? Mark the putative Transcription Start Site in your diagram with a bent arrow pointing in the direction of transcription.

4.2 Investigation 2: Identifying splice sites

We will again focus on tra-RA to identify splice sites, that is, the exact nucleotides where *splicing* occurs to remove introns from the *pre-mRNA*.

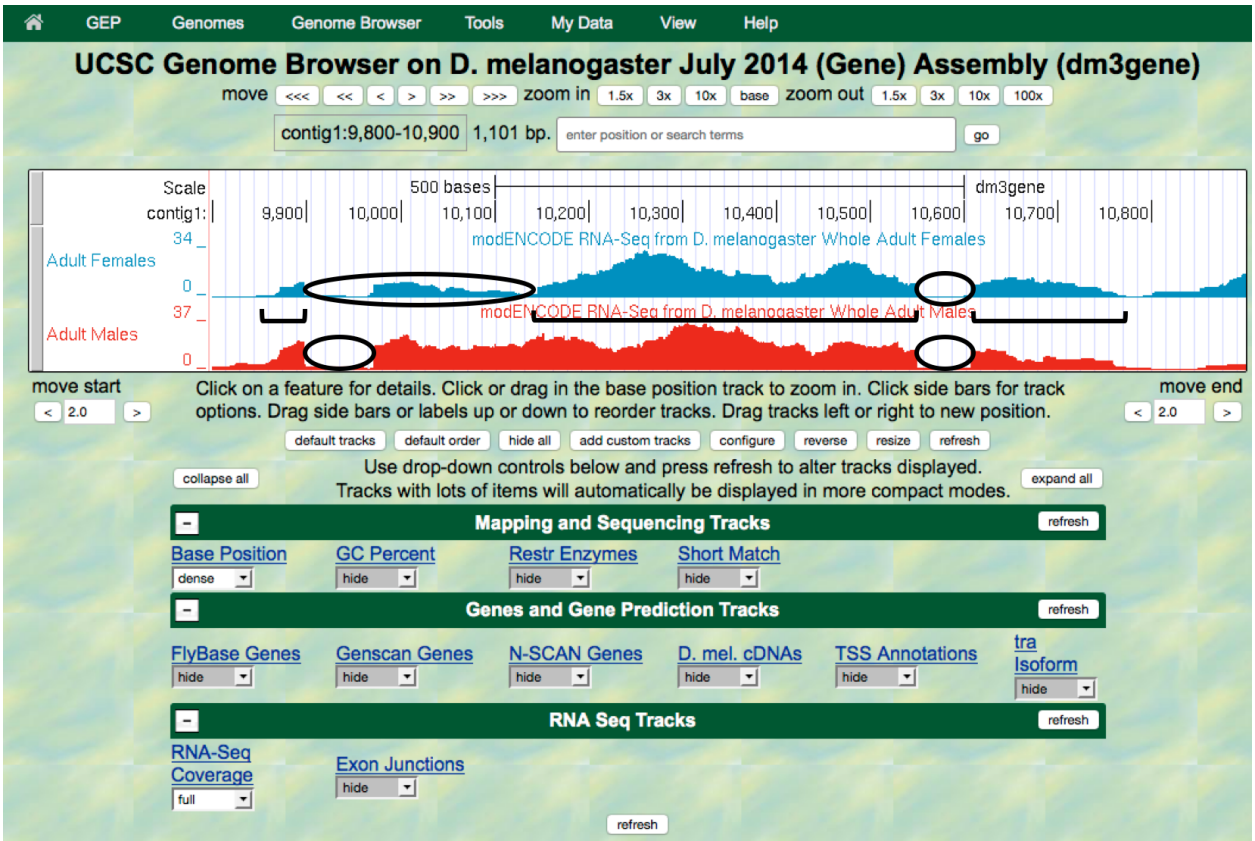


Figure 4.3.: In this screen shot of the *tra* gene, introns have been circled and exons have brackets underneath them.

4.2.1 Background

Two software programs, called TopHat and Bowtie, use the RNA-Seq data to graphically represent the exon junctions. The resulting graphic on the genome browser coincidentally looks something like a little bowtie (two small boxes connected by a thin line) (Figure 4.4). The boxes represent the sequenced mRNA (the exons), and the line represents a gap (the intron). The exon junction can be inferred when the first part of a sequenced fragment from the RNA population matches (for example) DNA positions 50-100 and the second part of the same fragment matches DNA positions 200-250; the RNA from positions 101-199 must have been taken out of the middle!

Short sequences are present at the beginning and end of each intron that allow the spliceosome — the molecular machinery that cuts out introns — to precisely remove the intron, leaving only the exon sequences in the *mature mRNA*. The first two nucleotides of the intron are the *splice donor site* and almost always the nucleotides "GT". The last two nucleotides of the intron are the *splice acceptor site* and almost always the nucleotides "AG". (Recall your observations in *Module 3*.) For more information on RNA-Seq and the search for *splice junctions*, watch the RNA-Seq and TopHat video

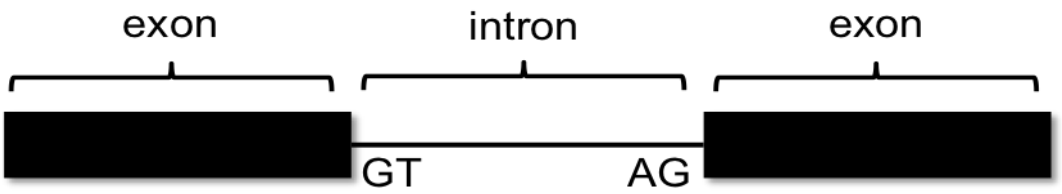


Figure 4.4.: A diagram of intron-exon junctions.

4.2.2 Use the Genome Browser to identify splice sites

1. Using the same Genome Browser page, reset the Browser by clicking on `hide all`. Then open the tracks that will provide the information we want for Investigation 2. (See the beginning of *Investigation 1* for a reminder on how to get the Genome Browser page.)
2. Change the display mode for the "Base Position" track to `dense`, and then click on `refresh`. (Note that you will not be able to see the DNA sequence until you "zoom in".)
3. Change the display mode for the "RNA-Seq Coverage" track to `full`, and then click on `refresh`.
4. You will again see blue and red histograms representing the RNA-Seq data (indicating the amount of mRNA synthesized) in females and males, respectively. We will focus on the blue histogram (Adult Females) again. As we did in *Module 3*, let's customize the RNA-Seq track by setting the "Data view scaling" field to `use vertical viewing range setting` and the "max" field under "Vertical Viewing range" to `37`. Remember that you gain access to these settings by clicking on the `RNA-Seq Coverage` link under the RNA-Seq Tracks green bar.
5. Change the display mode for the "Exon Junctions" track to `full`, and then click on `refresh`. These rectangular boxes joined by a thin black line will help you identify the exon-intron boundaries.

Our graphical output viewer will look like the screen shot below (Figure 4.5).

6. Zoom in to the area that is circled — click and drag the cursor just above the numbers, or use zoom buttons.
7. Set the screen so that you can see about 15–20 nucleotides, as shown in the example below (Figure 4.6).

The blue histogram stops at the end of exon 1. The last three nucleotides of exon 1 are "G-A-G".

Question 8

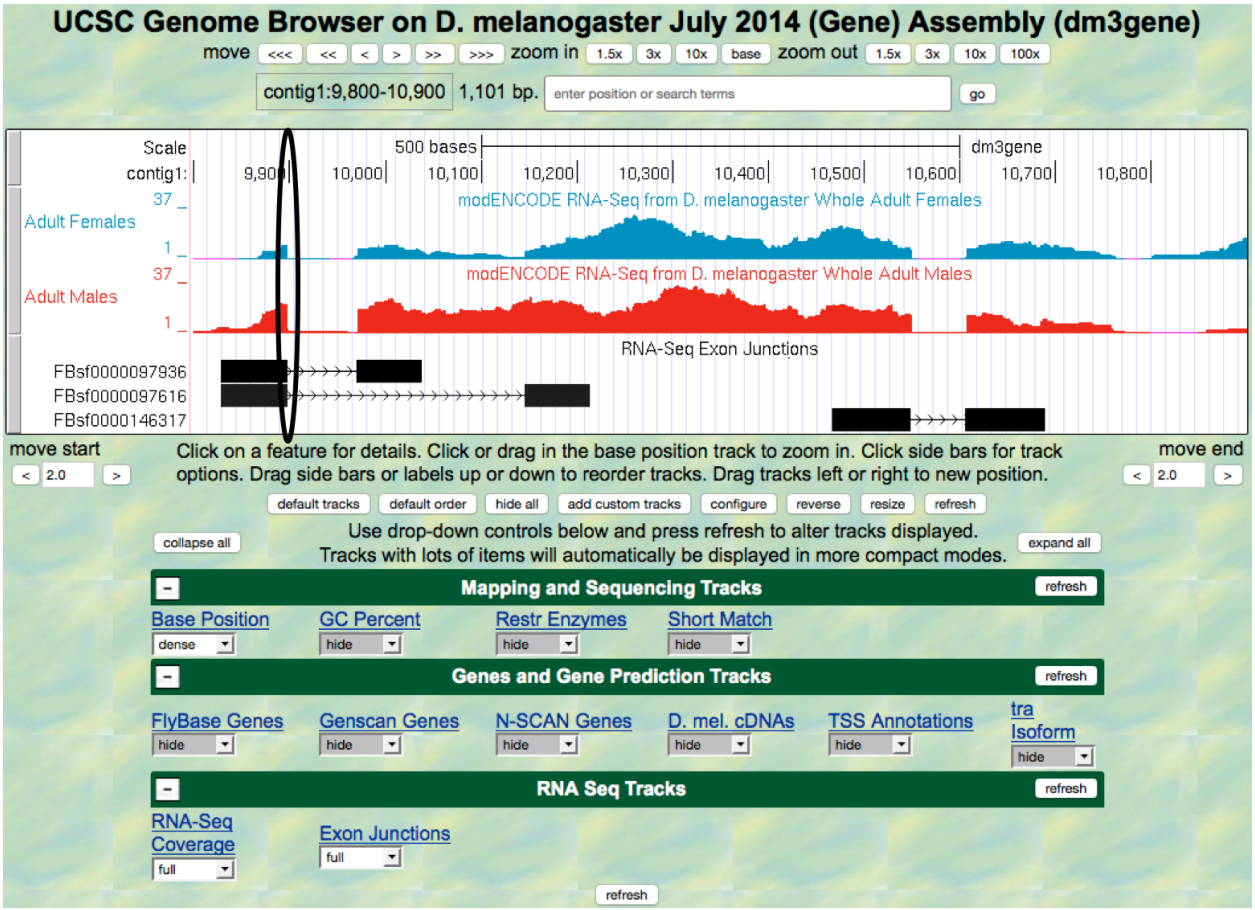


Figure 4.5.: View of the *tra* RNA-Seq data with the splice donor site in intron 1 circled.

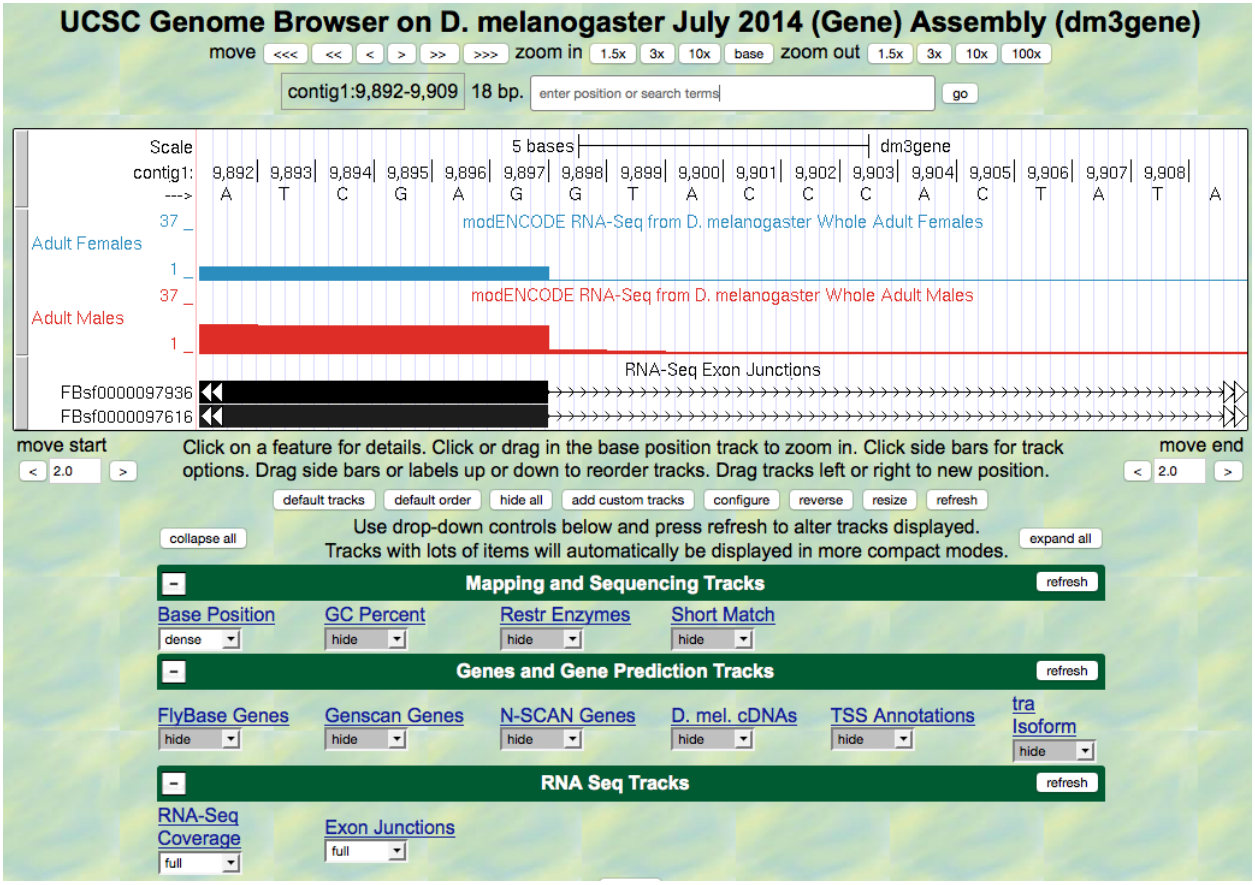


Figure 4.6.: TopHat data (in black) for adult males and adult females indicating an exon junction.

What is the coordinate of the last nucleotide in exon 1?

Question 9

What are the first two nucleotides of intron 1?

This is called the *5'* splice site or "splice donor site."

8. Zoom out so that you can see all of the *tra* gene. Let's use TopHat to help us find the *3'* end of the intron. Examine the Exon Junctions track. The first intron-exon junction predicted by TopHat (black) seems to align with the red histogram data from males; the second junction aligns better with the blue histogram data from females (Figure 4.7).

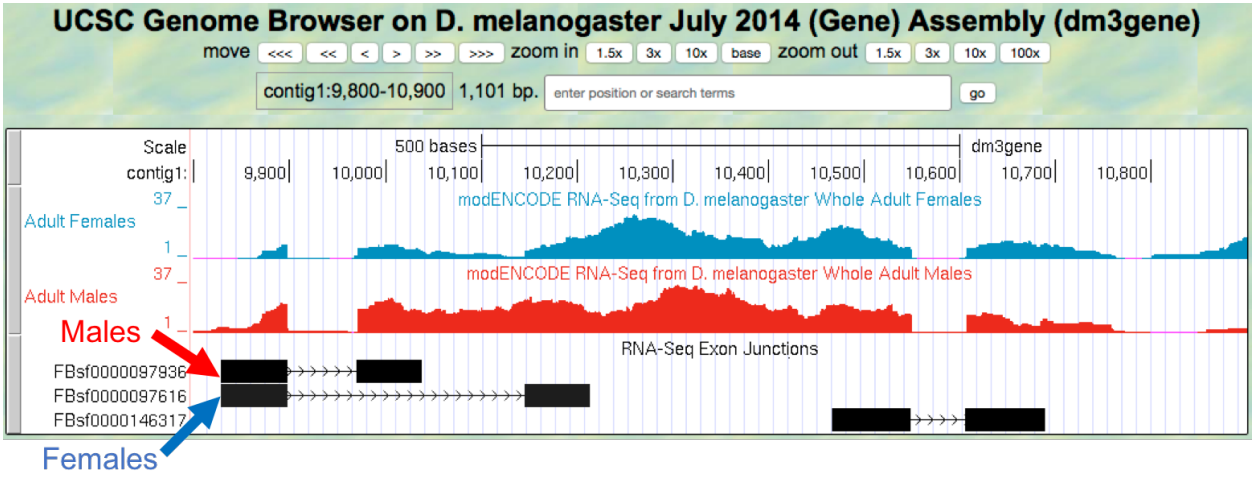


Figure 4.7.: TopHat and RNA-Seq data for males and females.

9. Let's examine the 3' end of intron 1 more closely. Change the "enter position or search terms" field to `contig1:10,140` and then click `go`. Zoom out 10x and then 3x (Figure 4.8).

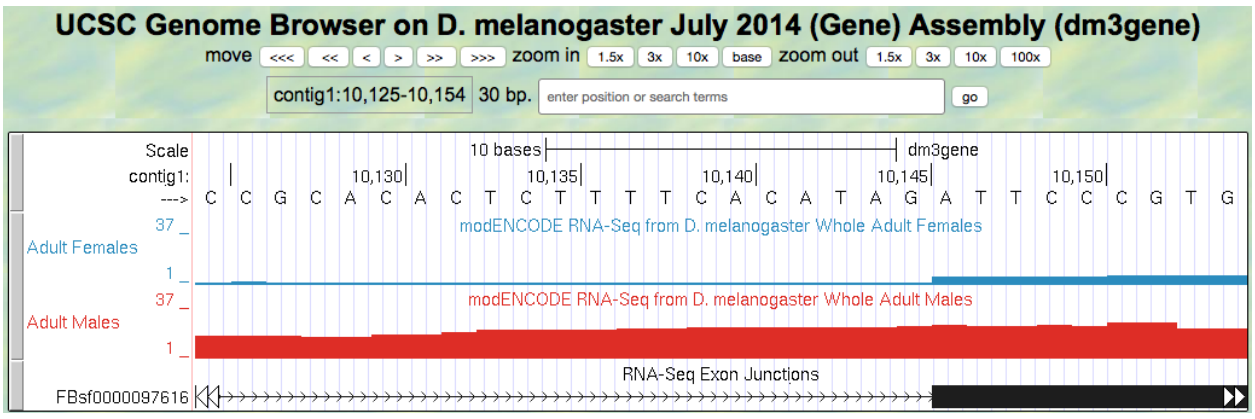


Figure 4.8.: Graphical viewer centered on the junction between intron 1 and exon 2 of the tra-RA (female specific) isoform.

Question 10

What are the last two nucleotides of the female tra-RA intron 1?

This is called the 3' splice site or "splice acceptor site".

Question 11

What is the coordinate of the first nucleotide in the female tra-RA exon 2?

4.3 Investigation 3: Identify the splice sites for intron 2

We can use the same approach to map the exon-intron boundaries of the second intron.

Review steps #2–5 in Investigation 2. Then repeat the process to answer the following questions about the 5' splice donor and 3' splice acceptor sites for the tra-RA (female specific) intron 2.

1. Zoom into the sequence surrounding the 5' splice donor for the tra-RA (female specific) intron 2.

Question 12

What are the last three nucleotides of the female tra-RA exon 2?

Question 13

What is the coordinate of the last nucleotide in the female tra-RA exon 2?

Question 14

What are the first two nucleotides of the female tra-RA intron 2?

2. Zoom out as needed so that you can see all of intron 2. Use the TopHat evidence track to find the 3' end of intron 2.
3. Click and drag so that the end of intron 2 is centered in the viewer. Then zoom in so that you can see the nucleotide sequence.

Question 15

What are the last two nucleotides of the female tra-RA intron 2?

Question 16

What is the coordinate of the first nucleotide in the female tra-RA exon 3?

Question 17

Using the information you've gathered so far, make a graphical picture of the tra-RA (female specific) isoform with 3 exons and 2 introns. Number each exon and intron at the corresponding DNA coordinates. Add the coordinates for

first and last nucleotide of the exons that you have found so far. Add the sequences of the splice donor and splice acceptor sites at the appropriate locations.

Question 18

Where do you think the promoter is located in relation to your gene model? What evidence do you have to support your idea, using the evidence tracks we have displayed (Base Position, RNA-Seq Coverage, Exon Junctions)?

Question 19

Bonus question! Support your hypothesis by gathering additional data. Recall our explorations in *Module 2* and *3*. You might want to open the tracks "D. mel. cDNAs," and "TSS Annotations," both in `full`. What type of evidence is shown by each of these tracks (refer to *Module 2*)? Finally, to see tra-RA as currently annotated in FlyBase, open the "tra Isoform" track on `full`, or to see both isoforms open the "FlyBase Genes" track in `pack`. Do these results support your model? Do any ambiguities remain?

4.4 Homework: Determining splice sites for the *spd-2* gene

1. Open a web browser on your laptop. Internet Explorer, Mozilla Firefox, Safari, or Chrome will work for this investigation. Go to the GEP Mirror of the UCSC Genome Browser 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.

Note: Reminder: Configure the Genome Browser Gateway page as follows:

1. Select `D. melanogaster` under "REPRESENTED SPECIES"
2. Select `July 2014 (Gene)` under "*D. melanogaster* Assembly"
3. Enter `contig1` into the "Position/Search Term" text box
4. Click on the `GO` button

As you will remember, this section of DNA is 11,000 base pairs long (Figure 4.9) and is part of the left arm of chromosome 3, which is about 28,100,000 bp long. If your Browser window is showing other evidence tracks, reset by clicking on `default tracks`.

2. Let's start by setting the evidence tracks we want to see. Click on the `hide all` button, then open only the tracks that will provide data for this investigation:
   - Base Position: `dense`
     - Note that you will not be able to see the DNA sequence until you zoom in. If the Base Position is changed to `full`, you can see the *amino acid* tracks also.
   - RNA-Seq Coverage: `full`
     - You will see blue and red histograms representing RNA-Seq data generated using RNA samples from adult females and adult males, respectively.
   - Exon Junctions: `full`
     - The exon junctions will help you to find the splice donor and splice acceptor sites.

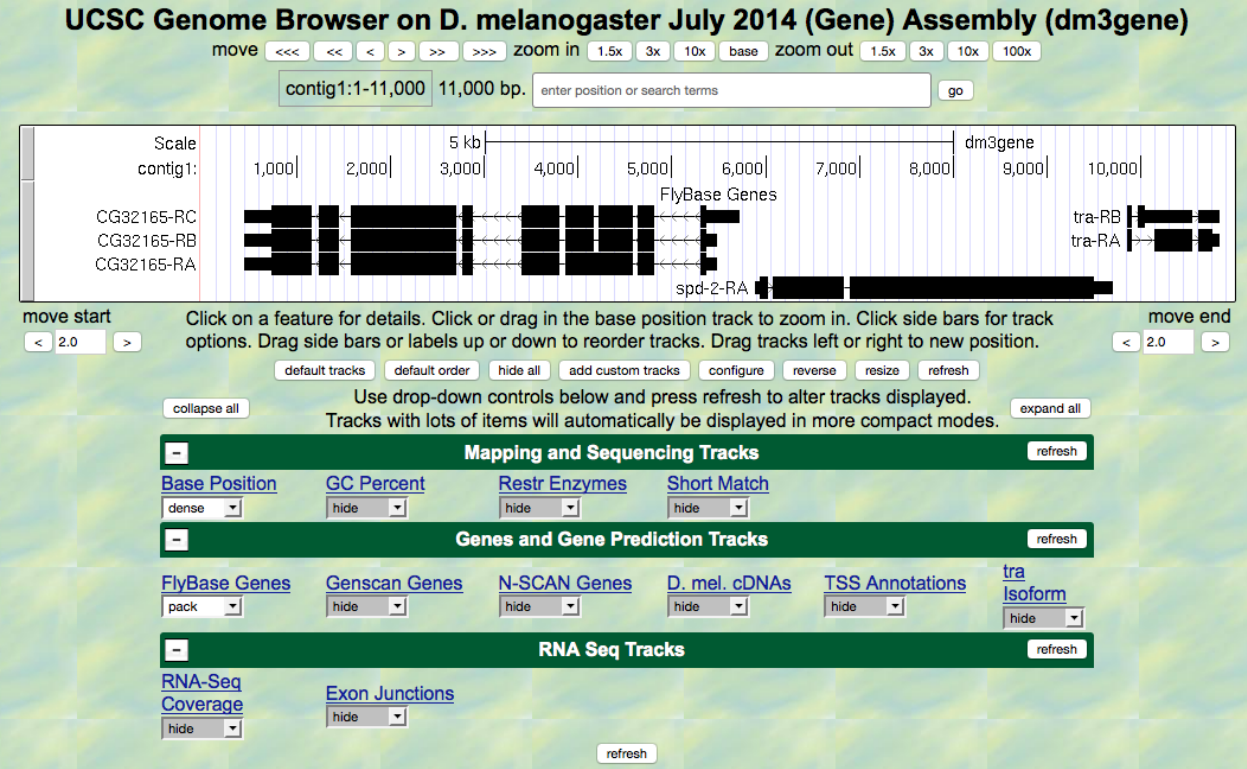


Figure 4.9.: A screen shot of the "contig1" project.

- D. mel. cDNAs: `full`
  - This track was used extensively in previous modules, and is useful for confirming the sequence of the mature mRNA. Remember that a *cDNA* is a DNA copy of an mRNA.

3. Zoom in to view only the *spd-2* gene by entering `contig1:5,750-9,800` in the "enter position or search terms" text box.

We are now looking at the region of chromosome 3 where the *spd-2* gene is located.

Question 1

How many exons does *spd-2* have?

Question 2

How many introns does *spd-2* have?

4.4.1 Part 1: Identifying splice sites for Intron 1

You remember from class that short sequences are present at the beginning and end of each intron that allow the spliceosome to precisely remove each intron, leaving only the exon sequences in the mature mRNA. The first two nucleotides of the intron are "GT" and the last two nucleotides are "AG" (Figure 4.10).

4. Zoom in to the end of exon 1. Set the screen so that you can see about 15–20 nucleotides.

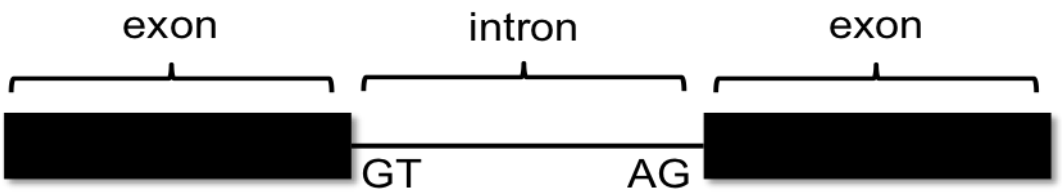


Figure 4.10.: A bowtie diagram

Question 3

What are the last three nucleotides of exon 1? What is the coordinate of the last nucleotide in exon 1?

Question 4

What are the first two nucleotides of intron 1?

5. Zoom out so that you can see all of intron 1, and use TopHat to help find the end of the intron. Examine the Exon Junctions track. Then zoom in so that you can see the nucleotide sequence.

Question 5

What are the last two nucleotides of intron 1?

Question 6

What is the coordinate of the first nucleotide in exon 2?

4.4.2 Part 2: Identifying splice sites for Intron 2

Let's use the same approach to map the exon-intron boundaries for intron 2. Review the steps you used in Part 1, then repeat the process to answer the following questions about the 5' splice donor and 3' splice acceptor sites for intron 2.

6. Zoom in to the sequence surrounding the 5' splice donor for intron 2.

Question 7

What are the last three nucleotides of exon 2?

Question 8

What is the coordinate of the last nucleotide in exon 2?

Question 9

What are the first two nucleotides of intron 2?

7. Zoom out as needed so that you can see all of intron 2. Use TopHat to find the end of intron 2.
8. Click and drag so that the end of intron 2 is centered in the viewer. Then zoom in so that you can see the nucleotide sequence.

Question 10

What are the last two nucleotides of intron 2?

Question 11

What is the coordinate of the first nucleotide in exon 3?

Question 12

Using the information you've gathered so far, make a graphical picture of the *spd-2* gene with 3 exons and 2 introns. Number each exon and intron. Add the coordinates for first and last nucleotide of the exons that you have found so far. Add the sequences of the splice donor and splice acceptor sites at the appropriate locations. Finally, add a bent arrow for the transcription start site.

CHAPTER 5

Module 5

Module 5: Translation: The need for an Open Reading Frame

Authors Carina Endres Howell (Lock Haven University of Pennsylvania) and Leocadia Paliulis (Bucknell University)

Last Update May 27, 2019

Version 0.0.1

5.1 Investigation 1: Examining Open Reading Frames in *tra*

5.1.1 Introduction: review of reading frames

In this exploration, we will continue to focus on the *transformer* gene (referred to as tra-RA or just *tra*), and will learn about how the *tra* mRNA is translated into a string of *amino acids*.

Given that DNA is double-stranded, and that the genetic code is based on triplets (3 consecutive bases), there are six possible reading *frames*. One can determine a reading frame by dividing the sequence of nucleotides in DNA or RNA into a set of consecutive, non-overlapping triplets. There are three possible reading frames (read *5'* → *3'*) in the forward direction on the top strand of DNA, and three (read 5' → 3') in the reverse direction on the complementary bottom strand of the same DNA molecule. Hence, there are six possible reading frames for each gene (see illustration in *Module 1*).

Once it is determined in which direction a particular gene is transcribed (for review see *Module 2* and *3* on transcription), there remain three choices for the reading frame. To determine which of these reading frames is used during *translation*, evidence such as the presence of an *initiation codon* and the absence of *stop codons* is used. As you learned in *Module 1*, the initiation codon is ATG in the coding DNA strand (AUG in the mRNA) and specifies the amino acid methionine. Additional triplets code for the other 19 amino acids, and three triplets are stop codons, causing termination of translation. These stop codons are TAA, TAG and TGA in DNA, or UAA, UAG and UGA when found in mRNA). An Open Reading Frame is a string of consecutive codons that is uninterrupted by stop codons. Every mRNA contains one *ORF* that is translated by the ribosome from start codon to stop codon.

5.1.2 Investigate reading frames for the *tra* gene

1. Go to the UCSC Genome Brower Mirror site at http://gander.wustl.edu/ and follow the instructions given in *Module 1* to open `contig1` of *Drosophila melanogaster* using the `July 2014 (Gene)` assembly.

2. The screen below will appear (Figure 5.1). As you will remember, this section of DNA is 11,000 *base pairs* long and is a small part of the left arm of chromosome 3, which is about 28,100,000 bp long.

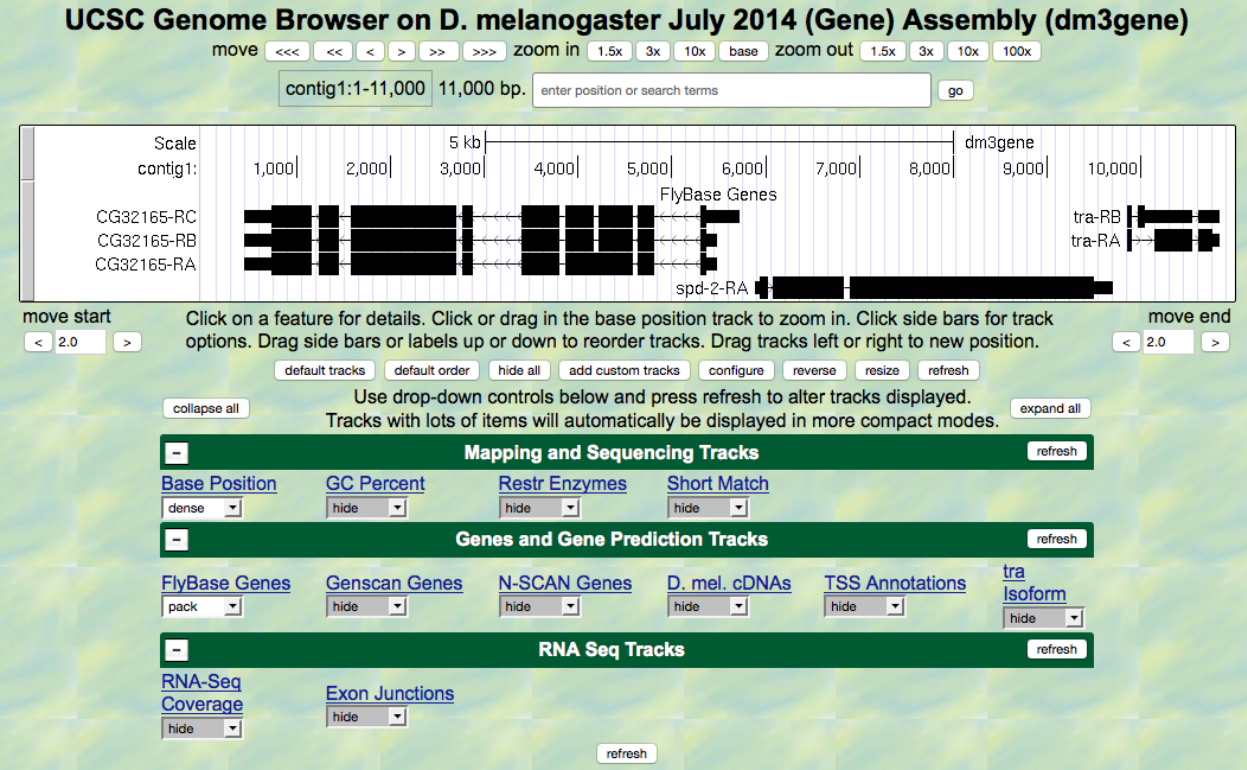


Figure 5.1.: View of the region of the *tra* gene on Chromosome 3.

3. Zoom in to view only the first *exon* of the tra-RA gene by entering `contig1:9,840-9,920` in the "enter position or search terms" text box and hitting the `go` button.

4. Open only the tracks that will provide information for this investigation. Set the Base Position track to `full`. Now we can see the amino acid tracks as well, giving the results from conceptual translation.

5. There are three possible reading frames for this transcript in the forward direction, here indicated by the numbers 1, 2 and 3 (red arrow) (Figure 5.2).

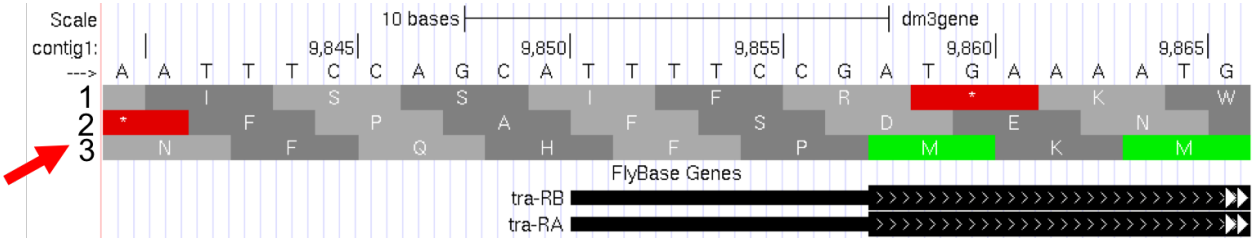


Figure 5.2.: Three possible reading frames for the *tra* gene in the forward direction.

6. Three reading frames are possible in the forward direction, as one could start translating with the first *base*, the

second base or the third base. (Starting with the fourth base is equivalent to starting with the first base, only missing one *codon*.) The DNA "top strand" is read from left to right (indicated by the arrow in the browser, right under the word "contig1" that looks like this `--->`). If you click on that arrow, the three reading frames in the reverse direction will appear, as the "bottom strand" is read from right to left. As you learned in *Module 1*, genes have directionality. The *tra* gene is read from the "top strand" (left to right).

7. Notice that the third reading frame has a green M codon (methionine) at the location where the thick black rectangle indicates the first *coding exon* or *CDS* (Coding DNA Sequence) of tra-RA mRNA. Remember that the codon for methionine (ATG in DNA) is the start signal, the first codon used in translation. This gives us our first piece of evidence that reading frame 3 is the one used in translation of the first CDS of the tra gene. For simplicity let's call this first CDS "CDS1" to distinguish it from other CDS's in the *tra* gene. Note that there is a stretch of RNA transcript upstream (to the left) of the ATG; this is the 5'*UTR* (5' untranslated region), found at the 5' end of all eukaryotic mRNAs.
8. Next carefully examine reading frame 2. Notice that in this reading frame there is no codon for methionine (no start codon) in the region that maps to the first exon. This gives us evidence that reading frame 2 is probably not being used during translation of CDS1 of the *tra* gene.
9. Finally, look at reading frame 1. Notice that there is a stop codon at the beginning of that reading frame (indicated by a red box with an asterisk in it). This evidence indicates that reading frame 1 probably is not being used during translation of CDS1 of the *tra* gene.
10. Let's move on to looking at the reading frames for exon 2 of tra-RA. Zoom in to view only the second exon of the *tra* gene by jumping to `contig1:10,120-10,570` using the "enter position or search terms" text box. Remember that both the RNA-Seq data and the *cDNA* data have been used to map the positions of exons.

Question 1

First examine reading frame 1. Are there any stop codons in the reported exon? If there are early stop codons, do you think this is the reading frame used during translation?

Question 2

Examine reading frame 2. Are there any stop codons in this reading frame within the exon?

Question 3

Examine reading frame 3. Are there any stop codons in the reported exon?

Question 4

Using the evidence above, which reading frame maintains an Open Reading Frame (ORF) across exon2 of tra-RA? Is this the same reading frame as that used for exon 1?

---

11. Finally, take a look at exon three (`contig1:10,600-10,850`). We anticipate that since this is the last CDS there will be one or more stop codons; one of which will mark the site of translation termination. The 3'UTR (3' untranslated region) extends downstream from this point to the site of poly(A) addition where the last exon ends (see *Module 3*). Here all three reading frames have a significant ORF followed by one or more stop codons in the exon.

Note: How can we figure out which reading frame is correct? We will investigate this in the next section by looking

specifically at the *splice junction*.

5.2 Investigation 2: Construct the gene model for tra-RA

We can combine what we know about reading frames with what we know about *splicing* to learn exactly how tra-RA is put together. We'll note where the start codon, splice sites, and stop codon are so we can construct a gene model. Then, in *Module 6*, we'll use these same types of information to solve some mysteries about tra-RB.

1. Using the same Genome Browser page, reset the Browser by clicking on `hide all`. Then open the tracks that will provide the information we want for Investigation 2:
   - Base Position: `full`
     - Note that you will not be able to see the DNA sequence or amino acid tracks until you zoom in.
   - FlyBase Genes: `pack`
   - RNA-Seq Coverage: `full`
     - You will see blue and red histograms representing the RNA-Seq data (indicating the amount of mRNA synthesized) in females and males, respectively.
     - We will focus on the blue histogram (Adult Females) again. As we did in *Module 3*, let's customize the RNA-Seq track:
       * Click on the `RNA-Seq Coverage` label under the RNA-Seq Tracks green bar found in the bottom section of the page.
       * Set the "Data view scaling" field to `use vertical viewing range setting`
       * Set the "max" "Vertical Viewing range" to `37`
       * Under the "List sub-tracks" section, unselect the `Adult Males` track
   - Exon junctions: `full`
     - These rectangular boxes joined by a thin black line will help us identify the exon-intron boundaries.

5.2.1 Identify the start codon

2. Let's find the start codon for tra-RA. Zoom in on where the FlyBase Genes track shows that the translation starts (where the tracked black box gets thicker for the tra-RA *isoform*) as seen in (Figure 5.3)

Question 5

Give the coordinates for the entire start codon for tra-RA (start codon coordinates should be three consecutive numbers, for example: nucleotides 212–214).

Question 6

Which reading frame should we follow along to see the predicted amino acid sequence of tra-RA?

Question 7

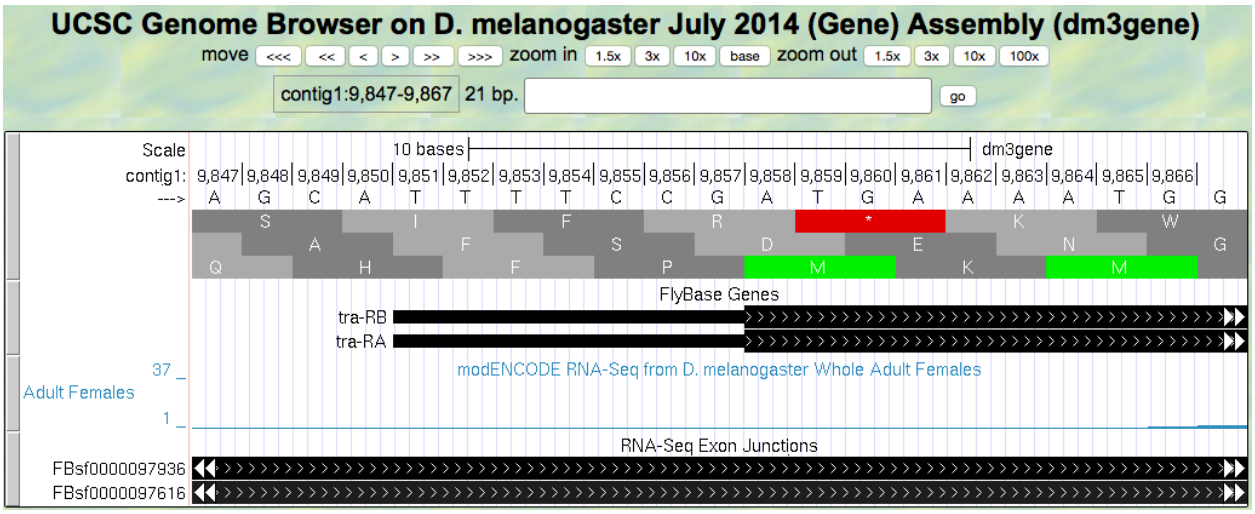


Figure 5.3.: Translation start region of the *tra* gene.

Zoom out to see the entire exon. Are there any stop codons in this reading frame in the first exon?

5.2.2 Identify the splice sites for intron 1

3. Now zoom in and find the last base of the first exon for tra-RA using your RNA-Seq data and Exon Junctions data (Figure 5.4).

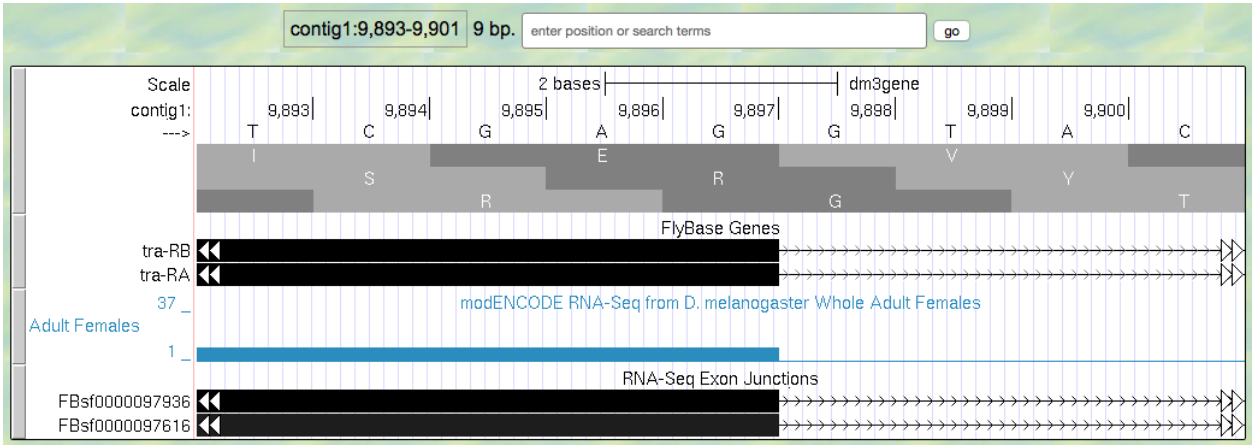


Figure 5.4.: Region at the end of Exon 1 of the *tra* gene.

Question 8

Give the coordinates for the very last base of the first exon.

So that we can follow the polypeptide through until we identify the stop codon, we need to figure out which reading frame we should follow in the second exon. This is not as easy as you might think, because eukaryotes don't always read in the same reading frame when looking at the genome. You saw an example of this previously in *Module 1*. Sometimes we can infer the correct reading frame given the pattern of start and stop codons within the region of the

exon, identified by RNA-Seq data. But that sort of information does not always give a definitive answer — there may be more than one possible reading frame for a given exon. To figure out which reading frame is being translated at exon 2, we need to check the end of the first exon to see how many bases of the last codon are present before the 5' splice site consensus sequence. To do this, look closely at reading frame 3, just before the splice site (Figure 5.5).

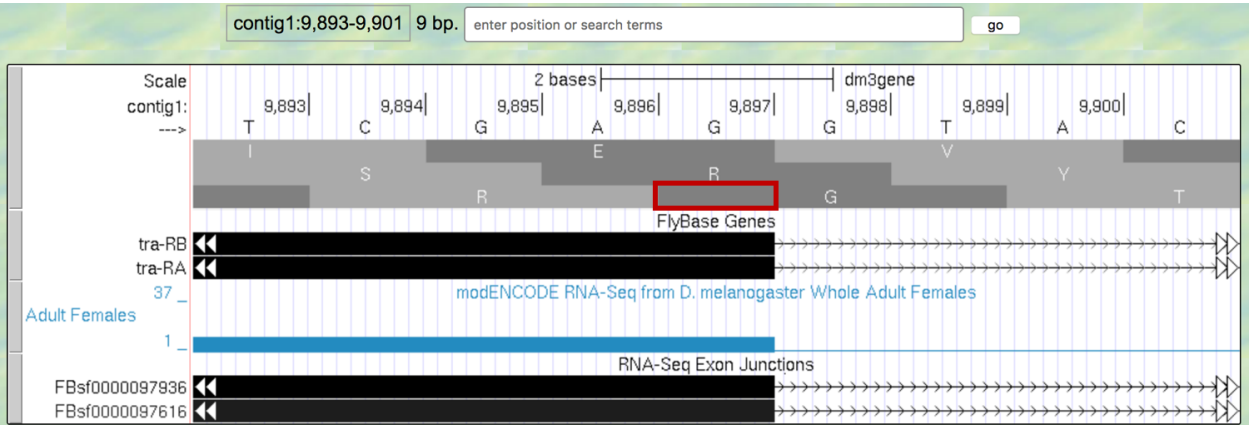


Figure 5.5.: An extra nucleotide between the last complete codon and the splice donor site for Exon 1 of the *tra* gene.

Note that the splice site cuts off the last codon of the first exon after just one base (**as indicated by the red box in** Figure 5.5). Therefore, we would say this exon has a **"phase 1" end because there is a partial codon at the end of the exon that is 1 base long.**

Note: If there were a fully completed codon before the splice site, it would be in **phase 0**, and if there were two bases before the splice site, it would be in **phase 2**.

For this exon with a *phase* 1 end we will need two more bases from the next exon to complete the codon. Knowing this we can identify the reading frame that will be used in the second exon. Navigate to the 3' splice site of *intron* 1 (i.e., the location where the first intron ends and the second exon begins; Figure 5.6). To review splicing and the concept of phase, watch the Splicing and Phase video.

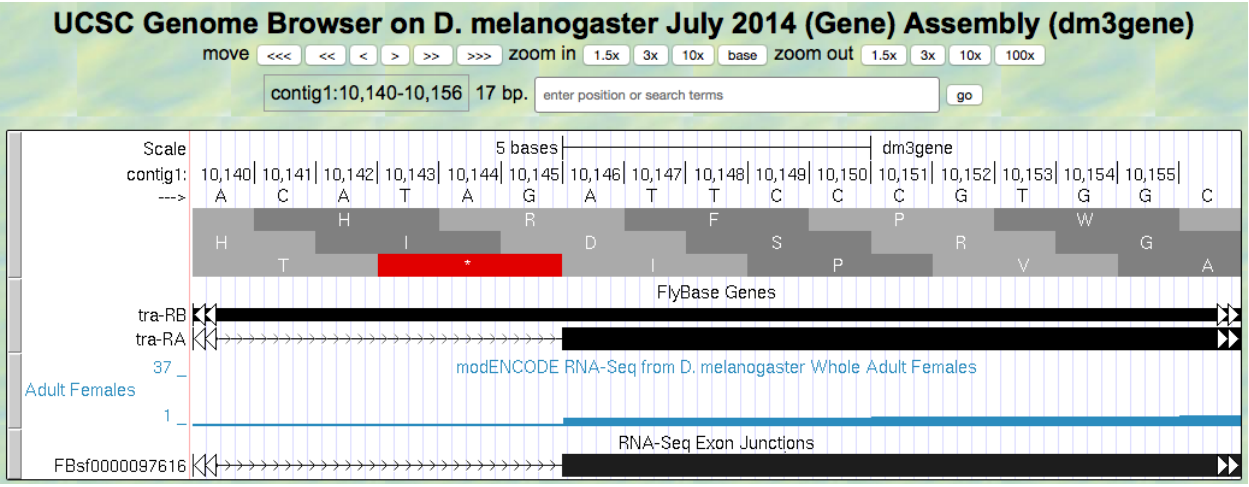


Figure 5.6.: Region at the beginning of Exon 2 of the *tra* gene

Question 9

Knowing that exon 1 ends with a partial codon of 1 base, what reading frame is being used in the second exon?

Question 10

Based on the evidence you see in the browser, give the coordinates for the first base of the second exon of tra-RA.

Question 11

Do you observe an appropriate splice acceptor site just upstream within the intron?

Now we will be using reading frame 2, because, after the splice site, there are two bases left in the codon. These two bases plus the one base left from the first exon make a complete codon.

4. Next, zoom out and look at reading frame 2 for all of exon 2 of tra-RA. You can see that there are no stop codons in this reading frame, which lends support to our conclusion that this is the proper reading frame.

5.2.3 Identify the splice sites for intron 2

5. Now, let's do the same for the 5' splice site of intron 2 for tra-RA. Zoom in on that splice site (Figure 5.7).

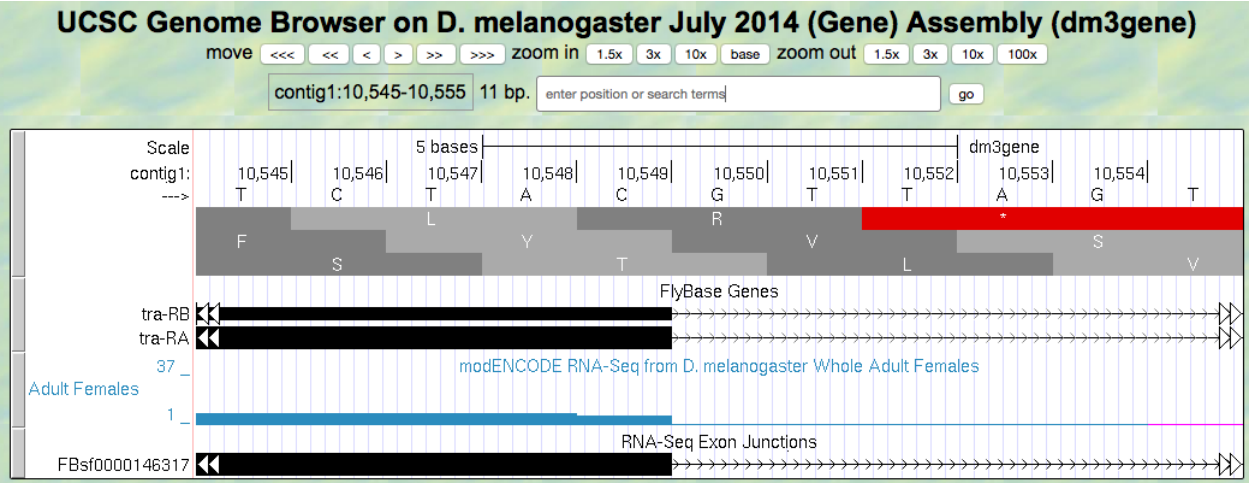


Figure 5.7.: Region at the end of Exon 2 of the *tra* gene.

Question 12

Give the coordinate of the base prior to the 5' splice site of intron 2.

Question 13

How many bases are left in the codon before the splice site, i.e. is this phase 0, phase 1, or phase 2?

6. Now navigate to the start of the final exon (Figure 5.8).

Question 14

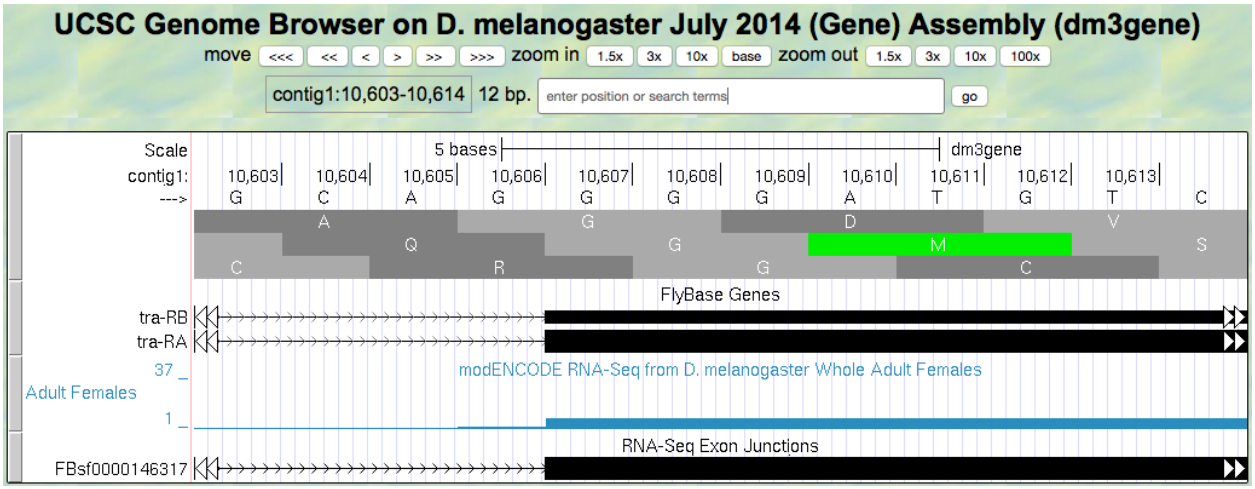


Figure 5.8.: Region at the beginning of Exon 3 of the *tra* gene.

Locate the 3' splice site of Intron 2. Give the coordinate of the first base in exon 3 for tra-RA.

Question 15

Which reading frame is being translated in the final exon?

5.2.4 Identify the stop codon

7. Now locate the first stop codon in the translated reading frame. Stop codons are shown as red boxes with asterisks (red arrows) as shown in Figure 5.9.

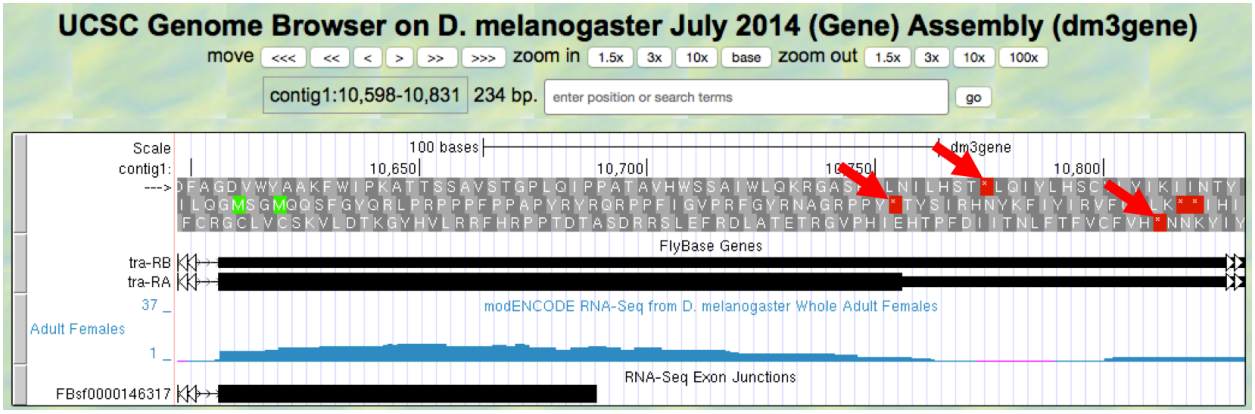


Figure 5.9.: Region at the end of Exon 3 of the *tra* gene.

Question 16

Give the coordinates for the bases in the stop codon.

5.2.5 Construct the complete gene model

Let's consolidate all the data we found above in one place:

Question 17

Gene model for tra-RA:

- Coordinate for start of translation: _____
- Coordinate for last base of exon 1: _____
- Coordinate for first base of exon 2: _____
- Coordinate for last base of exon 2: _____
- Coordinate for first base of exon 3: _____
- Stop codon coordinates: _____

Take the coordinate information above to draw a map of tra-RA using rectangles to represent exons and connecting lines to represent introns. Label the ends of the exons with the appropriate coordinates and indicate the transcription start site for the tra-RA initial transcript. Below this map, provide a map of the processed mRNA after intron removal. Below this map, indicate the regions that are translated into a protein. Give precise coordinates. Color coding may be helpful.

In *Module 6*, we will compare this model of tra-RA with a model of tra-RB.

Question 18

To cement your knowledge of gene structure, you could construct a similar map of the *spd-2* gene. How many exons does this gene have? How many introns? How many isoforms? Use the same approach to determine the coordinates for the exons, and the coordinates for the coding region (another name for the region that is translated).

CHAPTER 6

Module 6

Module 6: Alternative splicing

Author Leocadia Paliulis (Bucknell University)

Last Update May 27, 2019

Version 0.0.1

6.1 Investigation 1: Construct the gene model for tra-RB

In this investigation, we will focus on tra-RB, the second *isoform* of the *tra* gene, and will explore how multiple different mRNAs and polypeptides can be encoded by the same gene. The story of tra-RB is an exciting story of sex, *alternative splicing*, and poison *exons*!

1. To begin, open a web browser
2. Go to the UCSC Genome Browser Mirror site at http://gander.wustl.edu/ and follow the instructions given in *Module 1* to open `contig1` of *Drosophila melanogaster*, using the `July 2014 (Gene)` assembly rather than the "Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6)" assembly. Once you are on the Genome Browser page, set "Base Position" (under the "Mapping and Sequencing Tracks" bar) to `full` so that you will be able see the three possible reading *frames* (remember that you will not see individual *bases* or *amino acids* until you zoom in, though). Also set "FlyBase Genes" (under the "Genes and Gene Predictions" bar) to `full`. Don't forget to click on one of the `refresh` buttons to see your changes.
3. Enter the following coordinates into the "enter position or search terms" text box: `contig1:9,700-11,000` and hit the `go` button to get a good view of the *tra* gene (Figure 6.1).
4. Let's consider what we know about tra-RA and learn more about tra-RB.

Question 1

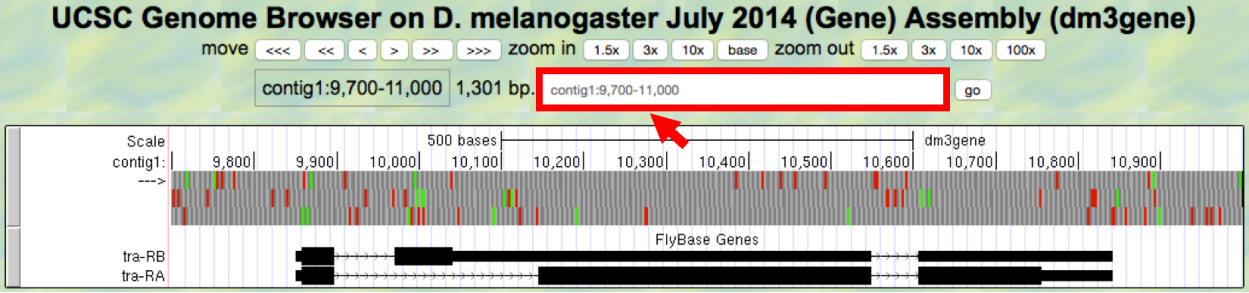


Figure 6.1.: Center the browser on the *tra* gene.

Given that exons are shown by the black boxes, and introns are shown by thin lines with arrowheads in the FlyBase Genes track, what does this tell us about the first intron of tra-RB compared to that of tra-RA?

5. Now let's look at the patterns of *transcription*. Scroll down to "RNA Seq Tracks", click on the `RNA-Seq Coverage` link. Change the track display settings as we did in *Module 2*:
   - Set the "Display mode" to `full`
   - Set the "Data view scaling" field to `use vertical viewing range setting`
   - Set the "max" field under "Vertical viewing range" to `37`
   - Check both `Adult Females` and `Adult Males` under "List subtracks"
   - Hit the `Submit` button (Figure 6.2)

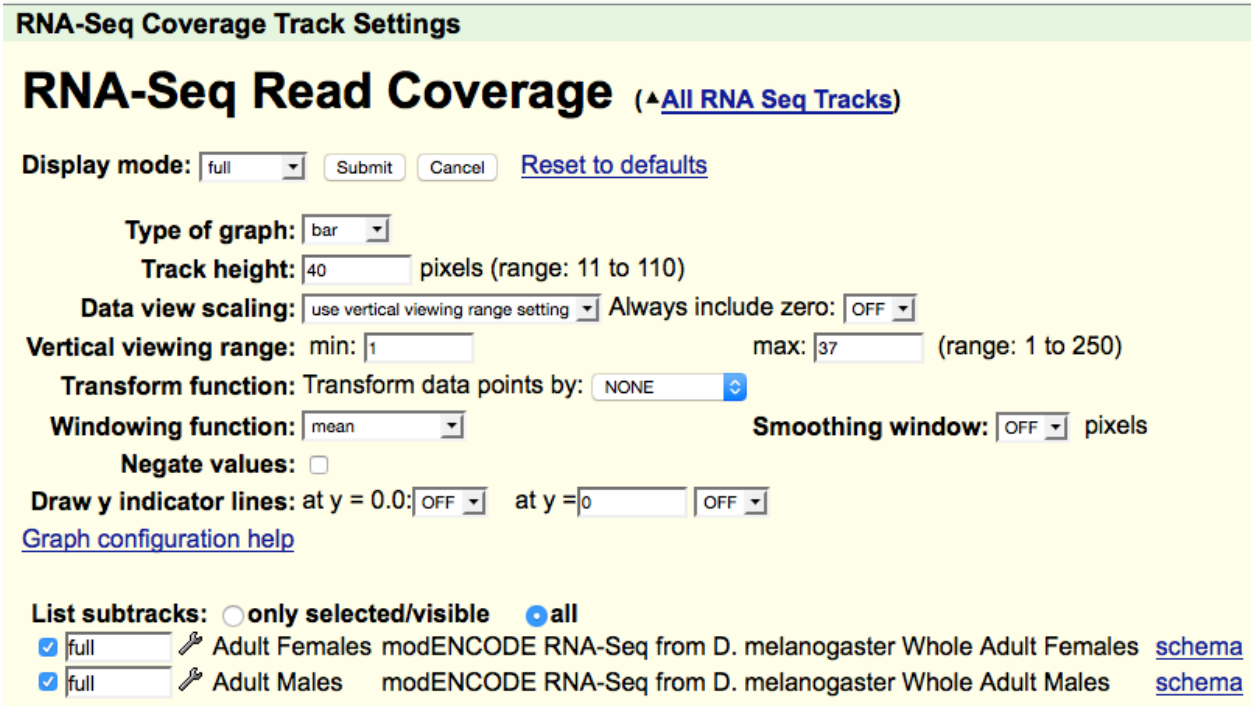


Figure 6.2.: Enter settings for RNA Seq tracks.

6. Back on the browser main page

- Under "RNA Seq Tracks", change the display mode for the "Exon Junctions" track to `full`, then hit `refresh`.

Tip: To review the use of RNA Seq data, watch the RNA Seq and TopHat video

Now we can see the RNA-Seq data for males (red) and females (blue). Recall that peaks in RNA-Seq Read Coverage tracks usually correspond to the regions of the genome that are being transcribed. These two samples generally show similar RNA-Seq read coverage along the entire span of the *tra* gene. However, the adult female sample shows substantially lower RNA-Seq read coverage at around 9,971-10,145 (red box in Figure 6.3). We can also see the RNA-Seq Exon Junctions track, which shows the location of splice sites supported by the RNA-Seq data (as you saw in *Module 4*). Recall that the black boxes in the FlyBase Genes track are exons and the thin lines with arrowheads show the locations of the *introns*. Notice that the diagrams for the first and second RNA-Seq Exon Junctions tracks have the same *5'* splice site but different *3'* splice sites. Let's see what we can find out about these splice sites.

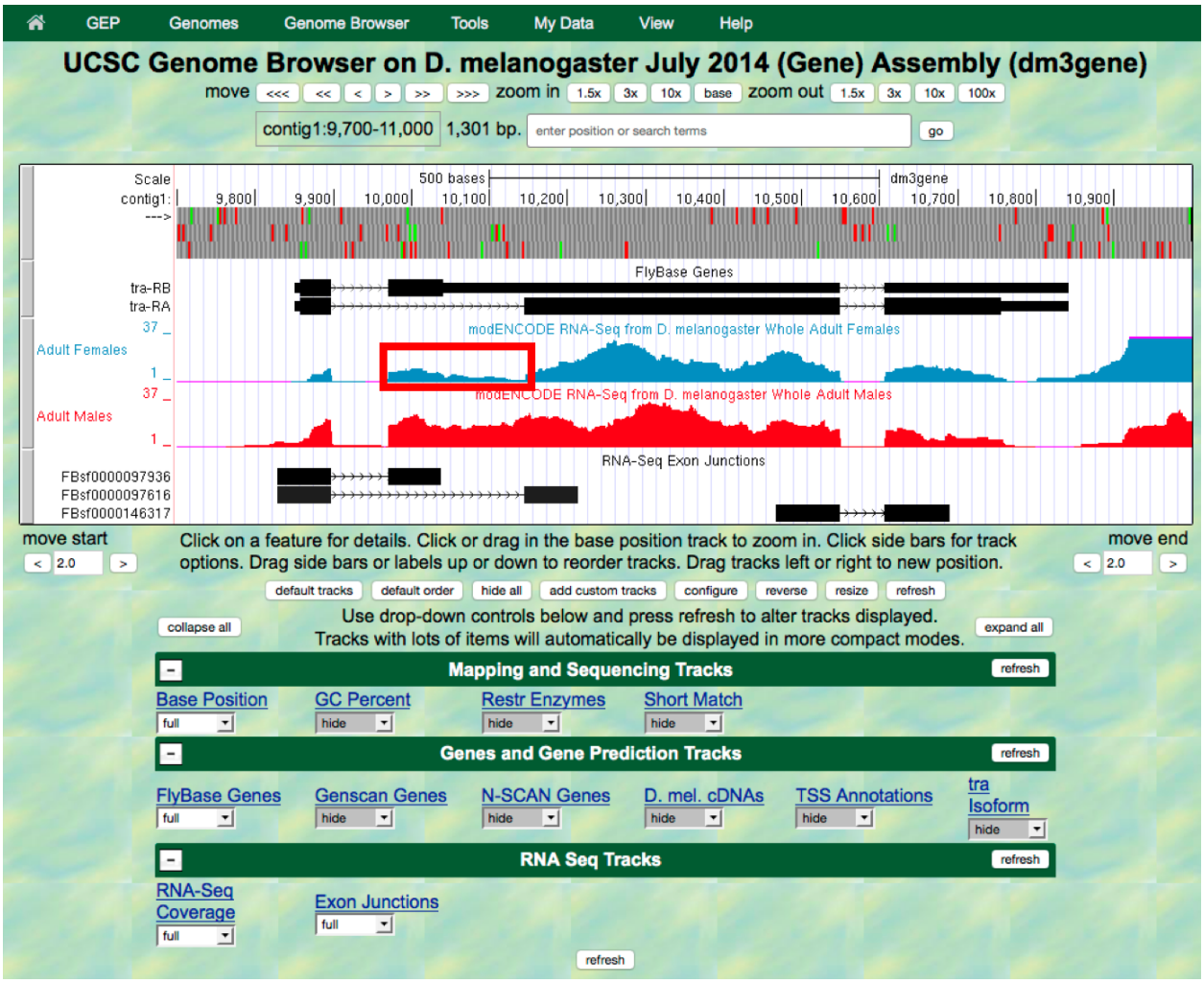


Figure 6.3.: Browser showing gene predictors, RNA-Seq tracks, and RNA-Seq exon junctions for male and female *Drosophila melanogaster*.

7. First, we need to establish the reading frame for the first exon. Zoom in on the 5' end of the transcript around position `contig1:9850-9860`.

Question 2

Given what you know about the initiation of translation, which of the 3 possible reading frames is used for both the tra-RA and tra-RB products?

8. Now zoom in on the location of the 5' splice site at the end of the first exon in both tra-RA and tra-RB (Figure 6.4). We will also be thinking about the concept of *phase* here. To review *splicing* and phase, watch the Splicing and Phase video.

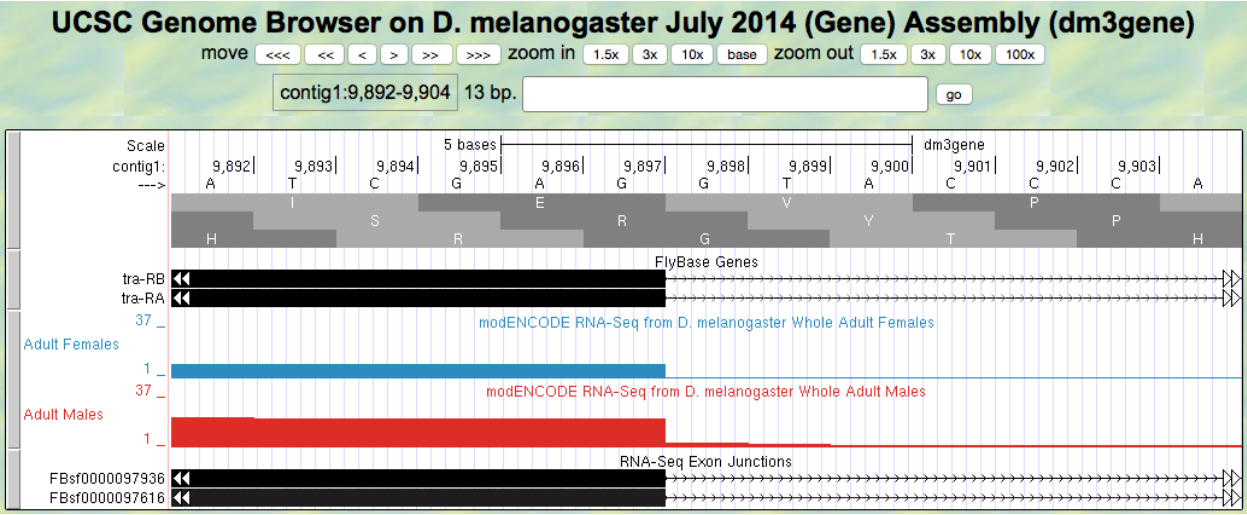


Figure 6.4.: Zoom in on the first 5' splice site.

Question 3

Give the coordinate for the last base of the first exon for tra-RA.

Question 4

Give the coordinate for the last base of the first exon for tra-RB.

Question 5

What is the consensus sequence for the 5' splice site (donor site)?

Question 6

What are the coordinates for the 5' splice site in tra-RA?

Question 7

What are the coordinates for the 5' splice site in tra-RB?

Question 8

What is the phase at this splice site?

9. Now zoom out and zoom in on the start of the second exon in tra-RB, just after the 3' splice site (Figure 6.5). We can identify the second exon by the RNA-Seq data, in particular using the RNA-Seq Exon Junctions data.

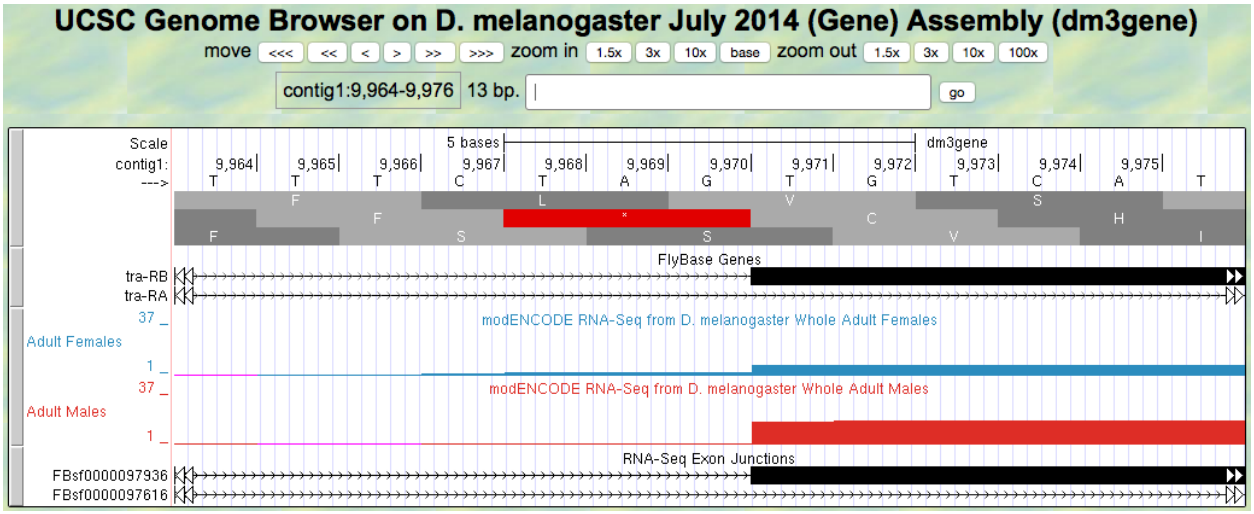


Figure 6.5.: Zoom in on the start of second exon in tra-RB.

Question 9

What are the coordinates for the first base of the second exon in tra-RB?

Question 10

What is the consensus sequence for the 3' splice site?

Question 11

What are the coordinates for the 3' splice site in intron 1 of tra-RB?

Question 12

What phase do we anticipate?

Question 13

Given this, what is the reading frame for tra-RB exon2?

Question 14

Does this make sense, given the location of stop codons?

10. Now zoom out and zoom in on the 3' splice site for tra-RA. (Figure 6.6). This can be identified from the RNA-Seq data, particularly the RNA-Seq Exon Junctions.

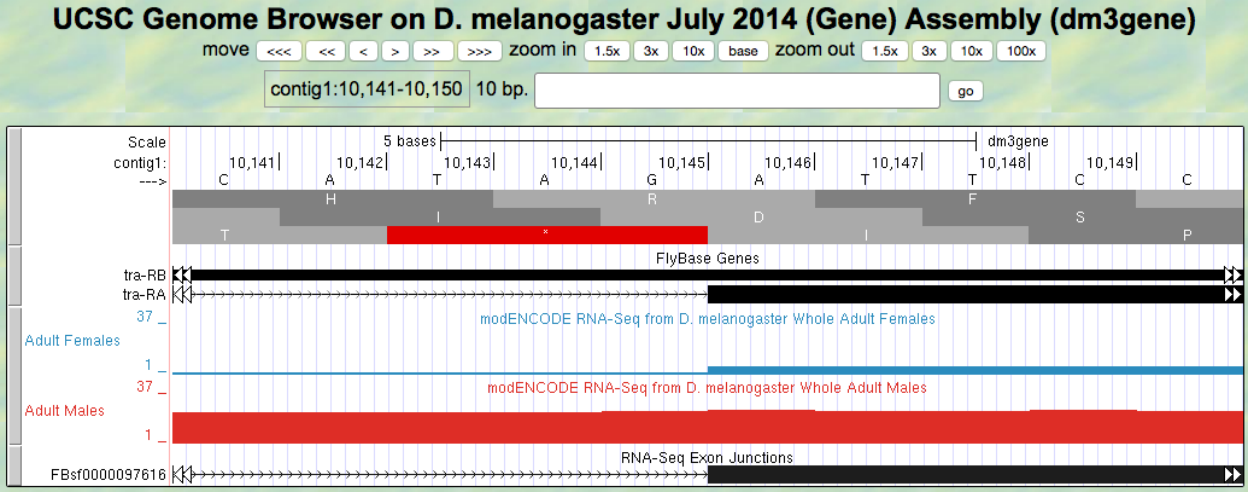


Figure 6.6.: Zoom in on start of second exon for tra-RA.

Question 15

What are the coordinates for the first base of the second exon in tra-RA?

Question 16

What is the consensus sequence for the 3' splice site?

Question 17

What are the coordinates for that sequence in intron 1 of tra-RA?

Question 18

Given the phase at the donor site, what phase are we looking for here?

Question 19

Given this, what is the reading frame for tra-RA exon 2?

Question 20

Does this make sense, given the location of stop codons?

The 3' acceptor site for the second intron in tra-RA is found inside the second exon of tra-RB. This intron is *alternatively spliced*. Alternative splicing is one way eukaryotes produce different proteins from the same coding regions of DNA. Here the alternative decision is made in a sex-specific manner; male fruit flies have targeted the spliceosome to use the first 3' acceptor site identified by the RNA-Seq Exon Junction data, while female fruit flies have targeted the spliceosome to use the second 3' acceptor site identified. This change in splicing has profound effects — in fact, it drives the programming of male and female characteristics in the developing fly. To review alternative splicing, watch the Genes and Isoforms video.

11. Reset your browser by entering `contig1:9,700-11,000` into the "enter position or search terms" text box and hit `go`. Let's analyze the consequences of this alternative splicing on production of a protein product.

Question 21

From your analysis of the A isoform of *tra* in *Module 5*, how many amino acids does the tra-RA protein product have?

Now look at the tra-RB isoform:

Question 22

Write down the coordinates for exon 1.

Question 23

Given the reading frame that you established for tra-RB, does translation continue through exon 2, or is it terminated by a stop codon?

Question 24

Write down the coordinates for the translated portion of exon 2.

Question 25

How many amino acids does the protein translated from the tra-RB isoform have?

Question 26

Is it likely that the protein translated from tra-RB could play the same functional role played by the protein translated from tra-RA?

Note: The Tra protein has an important function in female *Drosophila*, and is itself a splicing factor that regulates splicing. Careful *annotation* of genes, as we have done here, can provide many insights into biological control mechanisms.

6.2 Investigation 2: Polypeptides produced from each isoform of *tra*

Now that we know that *tra* is alternatively spliced to make two isoforms, tra-RA and tra-RB, and that males express one isoform while females express the other, let's try to figure out how alternative splicing affects the polypeptides produced from translating these mRNAs. To do this, we need to produce a gene model for tra-RB and compare it to the gene model for tra-RA that you constructed in *Module 5*, showing where the *start codons* and *stop codons* appear in each isoform.

Use what you learned in *Module 5* to construct a gene model for tra-RB. Locate the start codon, splice sites, and the stop codon. Construct the gene model below.

Question 27

Gene model for tra-RB:

- Coordinate for start of translation: _____
- Coordinate for last base of exon 1: _____
- Coordinate for first base of exon 2: _____
- Coordinate for last base of exon 2: _____
- Coordinate for first base of exon 3: _____
- Stop codon coordinates: _____

6.2.1 Points for discussion

- How does the polypeptide translated from the tra-RB isoform differ from the polypeptide translated from the tra-RA isoform? What are the consequences of these differences on protein function?
- Discuss how the bigger mRNA leads to creation of a smaller polypeptide!!
- Consider how alternative splicing could allow many different proteins to be encoded by the same gene.
- Based on the gene structure of the two isoforms of *tra* shown in the "FlyBase Genes" track, provide a hypothesis that could explain this difference in RNA-Seq read coverage between the adult males sample and adult females sample.

CHAPTER 7

Module 1

Module 1: Introduction to the Genome Browser: What is a Gene? (UCSC Assembly Hub Version)

Author Joyce Stamm (University of Evansville)

Last Update May 27, 2019

Version 0.0.1

7.1 Introduction to the Genome Browser

Genes encode information that our cells use to carry out their functions. In particular, protein-coding genes provide the cell with the information to make messenger RNAs (mRNAs), which are then used to make proteins. In this module, we will use a web-based visualization tool called a Genome Browser to explore the structure of a eukaryotic gene, and obtain a basic understanding of how this information is stored and used. In subsequent modules, you will learn more about the details of these biological processes, and use the Genome Browser to examine the experimental data that provide evidence for a detailed gene structure. The protein-coding genes in eukaryotes (higher organisms, with a cell nucleus) are much more complex than the protein-coding genes in prokaryotes (bacteria, organisms without a nucleus). We are still trying to figure out all of the details!

1. Start by watching the Genome Browser video
2. We will use the Genome Browser developed by the Genome Bioinformatics Group at the University of California Santa Cruz (UCSC) to examine different genomic regions. Open a web browser and navigate to the landing page for the "Understanding Eukaryotic Genes" modules at the CyVerse Data Store. Click on the `Understanding Eukaryotic Genes Assembly Hub` link to access the Assembly Hub (Figure 7.1).
3. Change the following fields in the "Genome Browser Gateway" section (Figure 7.2):
    - Click on the `Fruitfly` icon under the *POPULAR SPECIES* field. This will allow you to view the genome of the insect *Drosophila melanogaster.*
    - Confirm that `Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6)` is in the *D. melanogaster Assembly* field. This is the version of the *D. melanogaster* genome that you will view. The **genome assembly** is simply the **genome sequence** produced after chromosomes have been fragmented, those fragments

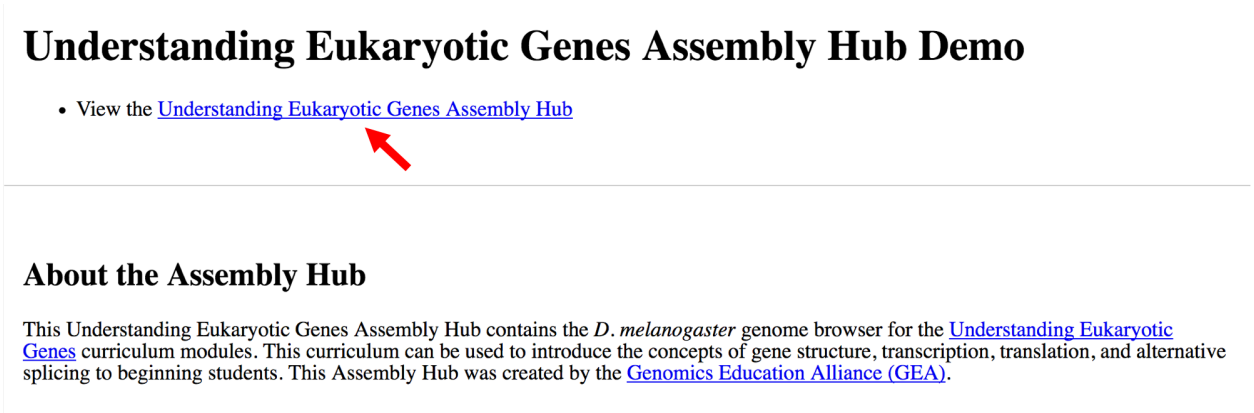


Figure 7.1.: Click on the "Understanding Eukaryotic Genes Assembly Hub" link to access the Genome Browser.

have been sequenced, and the resulting sequences have been put back together. A genome assembly is **updated** when DNA has been sequenced that allows gaps to be filled. It may also be updated when a new assembling algorithm is released. The August 2014 *Drosophila melanogaster* (BDGP Release 6 + ISO1 MT/dm6) assembly was produced by the Berkeley Drosophila Genome Project (BDGP).

- Click on the `GO` button.

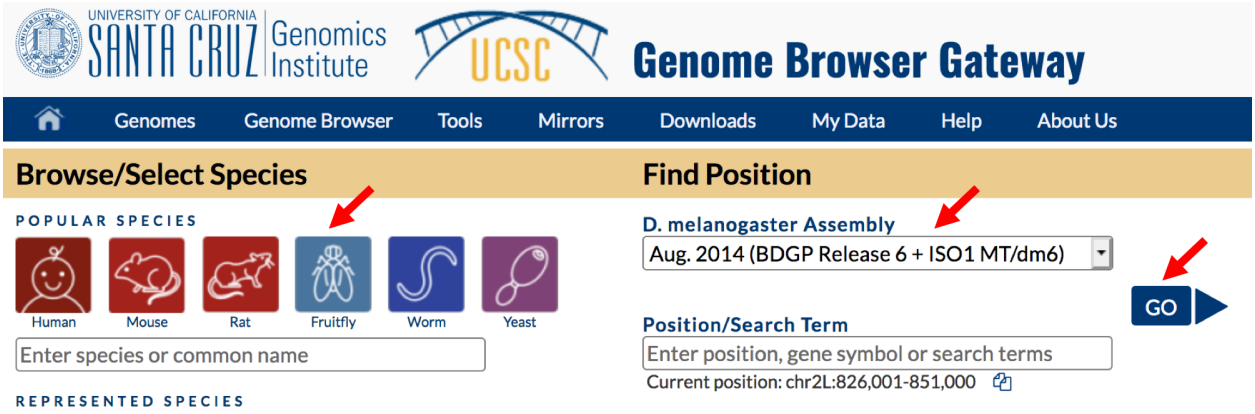


Figure 7.2.: Configure the Genome Browser Gateway page to view the *D. melanogaster* Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6) genome assembly.

4. The next screen can be divided into four major sections (Figure 7.3):
   - A top blue toolbar is used to navigate to the different tools provided by the Browser.
   - Navigation Controls allow us to navigate or zoom to different parts of the genome.
   - A genomic features panel (the white area) shows the locations of the different genomic features within the portion of the genome (e.g. chr3L) specified by the label next to the "enter position or search terms" text box
   - A Display Controls section may be used to manipulate how much detail is visible in the genomic features panel of the Genome Browser.
5. To match the screenshot shown in Figure 7.3:
   - Click on the `hide all` button to hide all the evidence tracks.

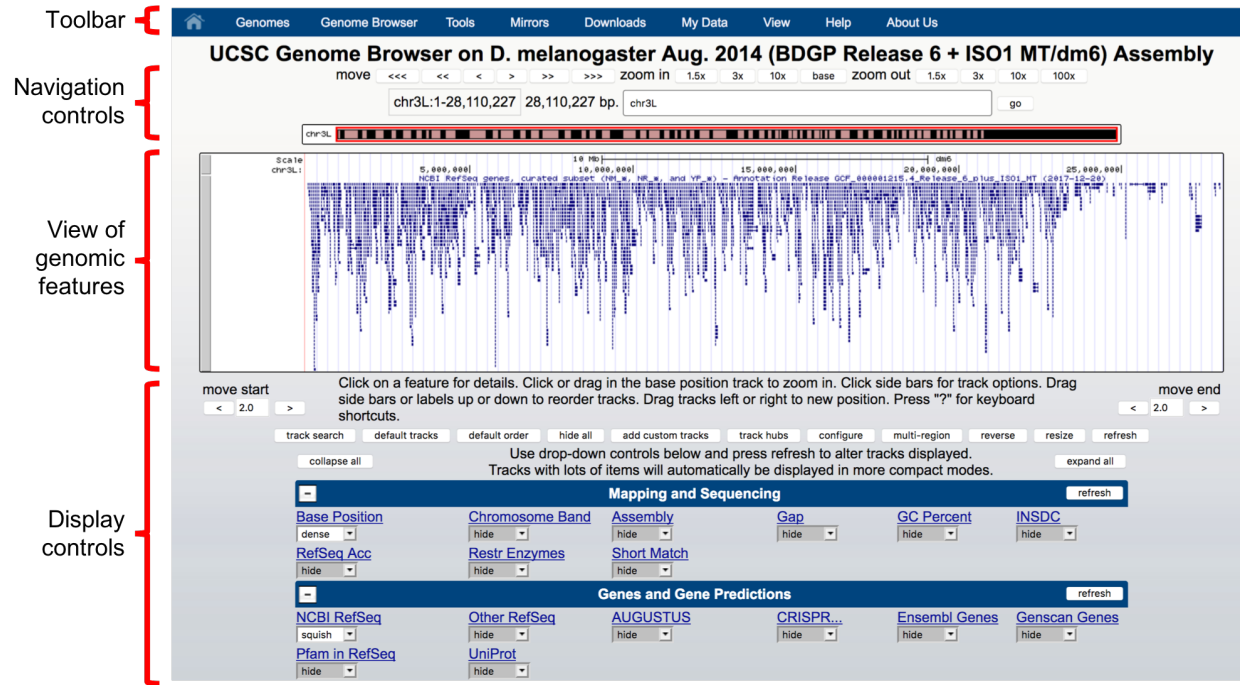


Figure 7.3.: The four major sections of the Genome Browser.

- Enter `chr3L` into the *enter position, gene symbol or search terms* text box, and then click on the `Go` button (Figure 7.4).

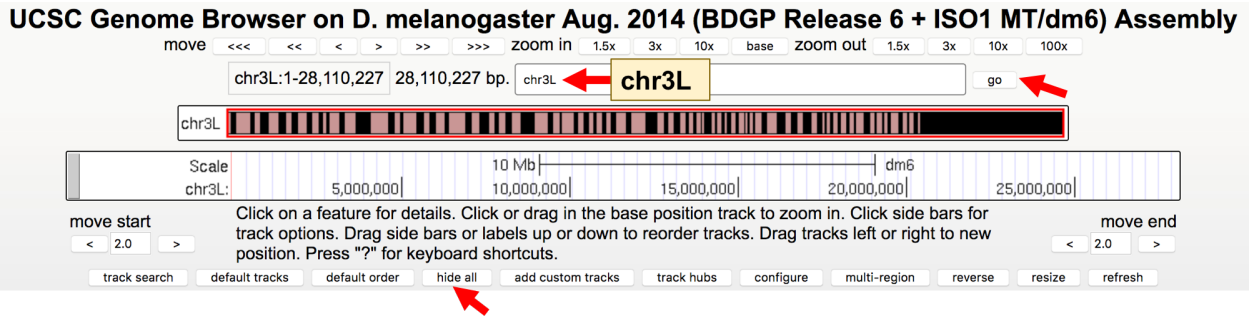


Figure 7.4.: Hide all the evidence tracks and then navigate to chr3L of *D. melanogaster*.

- Scroll down to the bar labeled "Mapping and Sequencing" in the Display Controls section, go to *Base Position*, and select `dense` from the drop-down menu.
- Scroll down to the bar labeled "Genes and Gene Predictions", and then click on the `NCBI RefSeq` link. Select `squish` from the *Maximum display mode* drop-down menu, then select `squish` for the *RefSeq Curated* subtrack (Figure 7.5). Click on the `Submit` button to update the genomic features panel.

You can use the buttons in the "Navigation controls" section to navigate to different parts of the genome. You can zoom in to a region by clicking on one of the buttons next to the *zoom in* label (i.e. 1.5x, 3x, 10x, base), and you can zoom out by clicking on the buttons next to the *zoom out* label. Alternatively, you can enter the genome *coordinates* into the "enter position or search terms" field and then click on the `go` button to navigate to a specific region in the genome assembly.

The "size" field next to the "enter position or search terms" text box (red arrow in Figure 7.6) shows the total size of the genomic region that you are viewing. In this case, the "size" field shows that chr3L (i.e. the left arm of chromosome

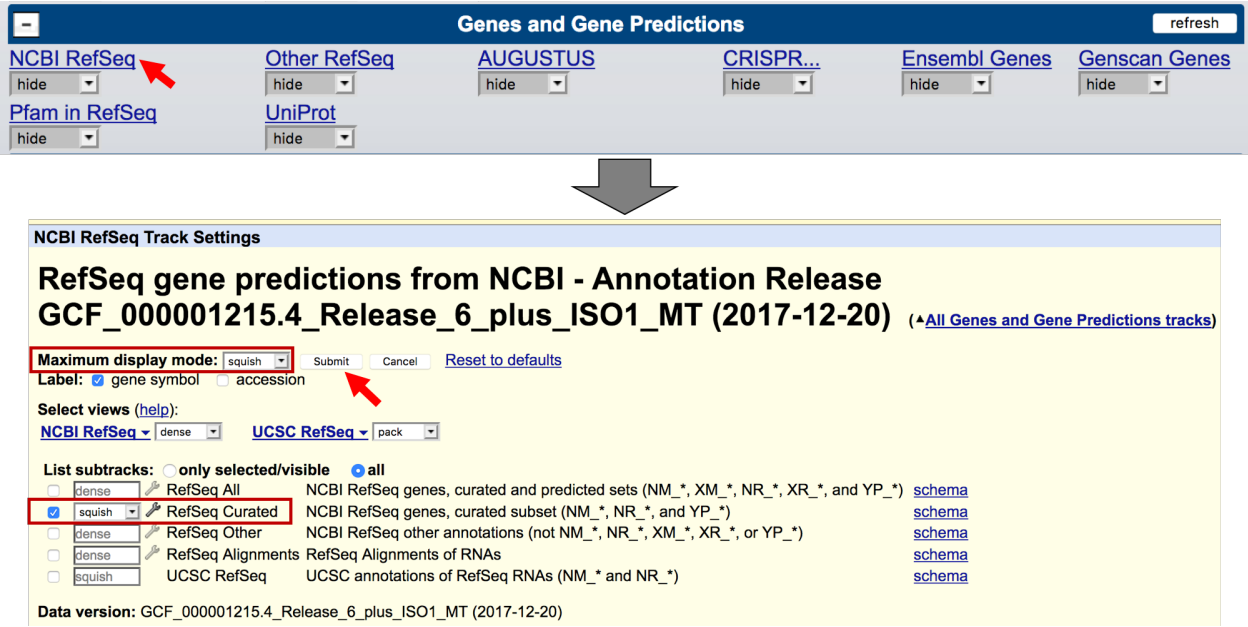


Figure 7.5.: Configure the "NCBI RefSeq" track to display the "RefSeq Curated" track in the squish display mode.

3) in *Drosophila melanogaster* has a total length of ~28 million *base pairs* (bp). We will learn more about the key functionalities of the Genome Browser in subsequent modules. For now, we will focus on the large white rectangle shown on this page; this contains a graphical representation of the genomic features (e.g. protein coding genes, percent GC content) of chr3L mapped against the DNA sequence, which is embedded in the top line of the white box.

The different types of features (also known as "**tracks**" or "**evidence tracks**") are separated by a title and are often shown in different colors. What types and how many tracks are shown in the view of genomic features is controlled by the display controls at the bottom. The view shown on Figure 7.6 displays only one of the tracks in the "Gene and Gene Predictions" section, and all the other tracks in other sections (mRNA and EST, Expression and Regulation, Comparative Genomics, etc.) are "hidden". More information about evidence tracks is available in the Tracks video.

We can examine the region under the blue title labeled *NCBI RefSeq genes, curated subset* to estimate the number of protein-coding genes on chr3L. In this track each gene is represented by a set of blue boxes connected by thin blue lines. There are clearly fewer blue boxes at the right side of the image compared to the left, which suggests that genes are not uniformly distributed along the chromosome (Figure 7.6).

In the genome browser, each chromosome may be organized into smaller projects called contigs (for contiguous sequences). In this next part, we will examine contig1, a much shorter region in the left arm of chromosome 3.

6. Click on the `Genomes` link on the top toolbar to return to the Genome Browser Gateway page.
7. Scroll to the top of the "REPRESENTED SPECIES" section and click on the `Understanding Eukaryotic Genes` link next to the *Hub Genomes* label. Verify that `July 2014 (Gene)` is selected under the *Understanding Eukaryotic Genes Hub Assembly* field. Change the *Position/Search Term* field to `contig1` (Figure 7.7).
8. Click on the `GO` button.

The "size" field now has the value "**11,000 bp**", which means that contig1 has a total length of 11,000 bp.

To further explore the features on contig1, we will examine the results from two of the available tracks.

9. Scroll down to the bar labeled *Mapping and Sequencing Tracks* in the "Display controls" section. Verify the display mode under the *Base Position* track is set to `dense` and the *FlyBase Genes* track is set to `pack`.
10. The display mode for all other evidence tracks should be set to `hide` (Figure 7.8).

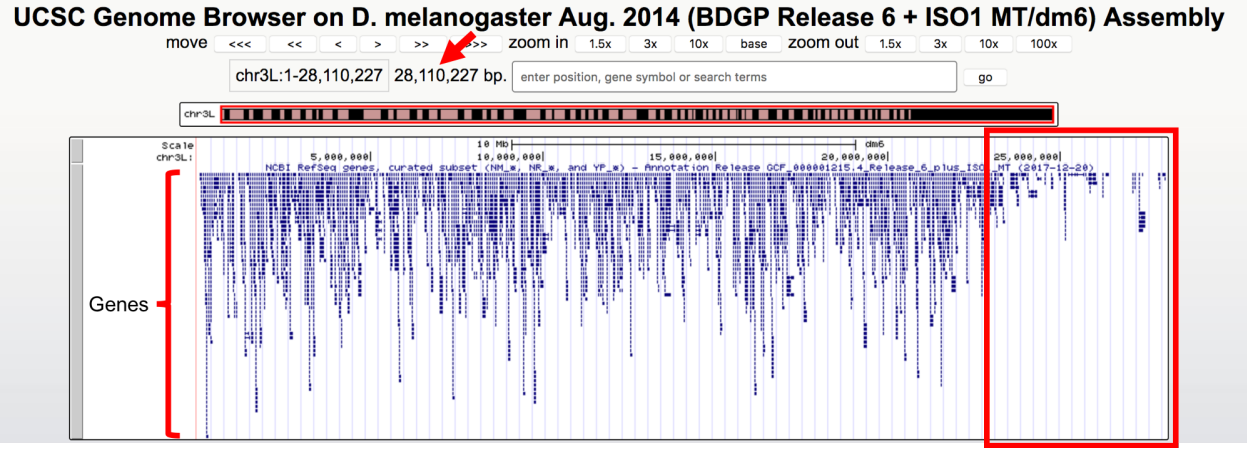


Figure 7.6.: Genome Browser shows that the entire *D. melanogaster* chr3L sequence has a length of ~28 million base pairs (red arrow) and that the right end of the chromosome has low gene density (red box).

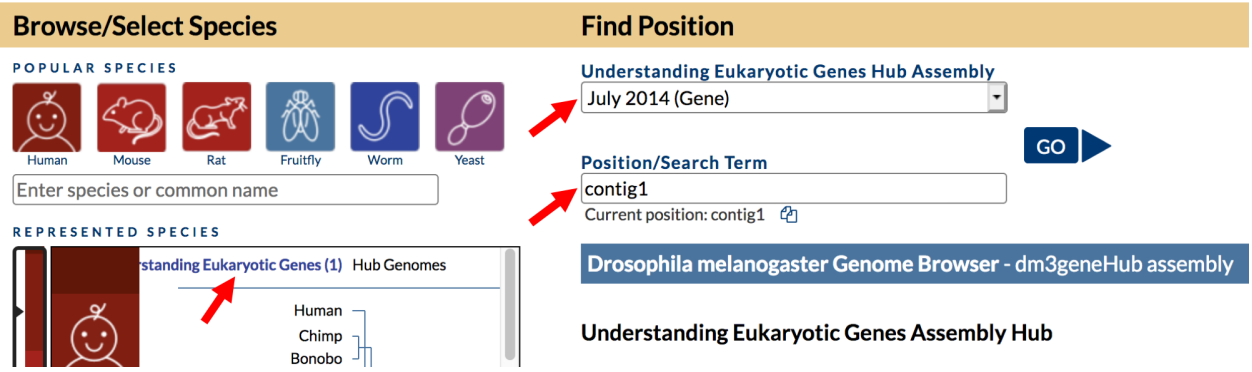


Figure 7.7.: Return to the Genome Browser Gateway page and then select the "July 2014 (Gene)" assembly.

11. Click on any `refresh` button to update the Genome Browser image.

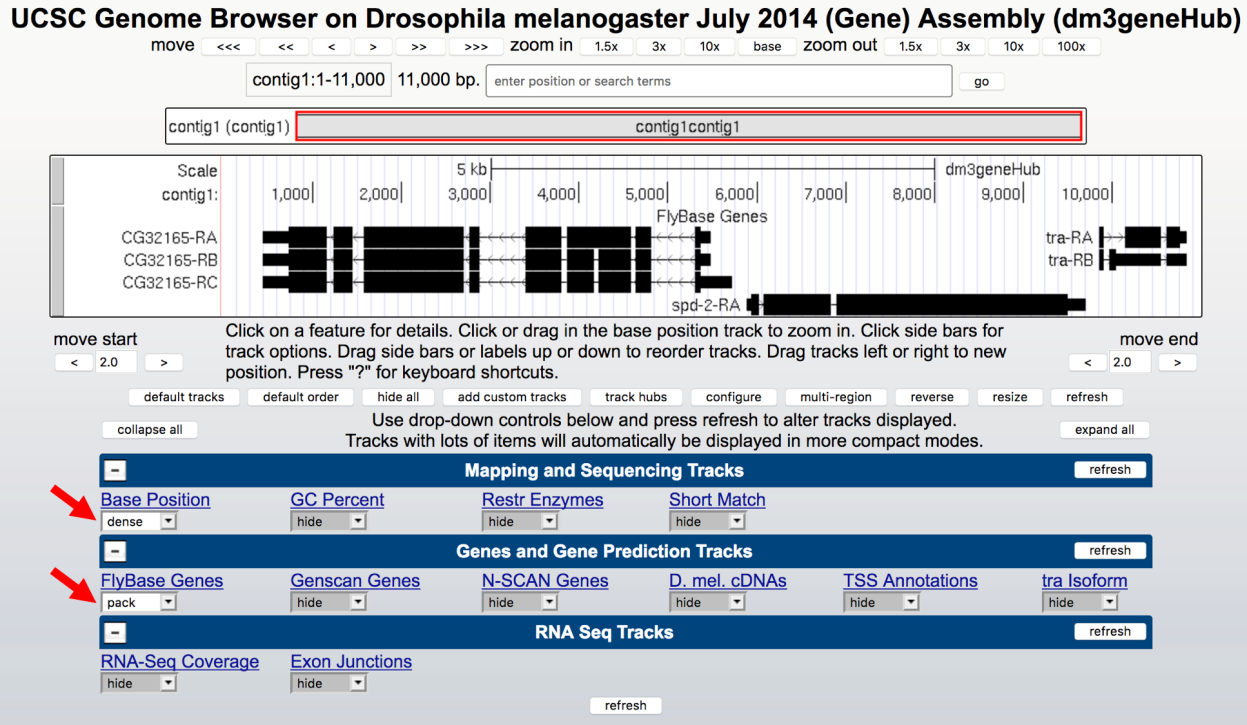


Figure 7.8.: Verify the display settings for the "July 2014 (Gene)" assembly.

Explore the contig1 genomic region using these tracks on the Genome Browser. You will observe distinct groups of connected boxes. These connected boxes and lines are genes, and their names are indicated on the left. Connected boxes and lines that are stacked vertically represent alternative forms of a gene, called *isoforms*. Answer the following questions:

Question 1

How many genes are there in contig1?

Question 2

What are the names of these genes?

Question 3

Which gene has the largest span (i.e. the largest distance between the start and end of the gene)?

12. Now let's examine the gene at the end of this contig more closely. Type `contig1:9,841-9,870` into the *enter position or search terms* text box and then click on `go`. (Note that you don't need to use commas when entering base positions). The Genome Browser image will update to show only bases 9,841 to 9,870 of contig1. Note the letters that appear just below the base position numbers. These letters correspond to the nucleotide at each position. For example, both forms of the *tra* gene, tra-RA and tra-RB, begin with a T at position 9,851 (Figure 7.9).

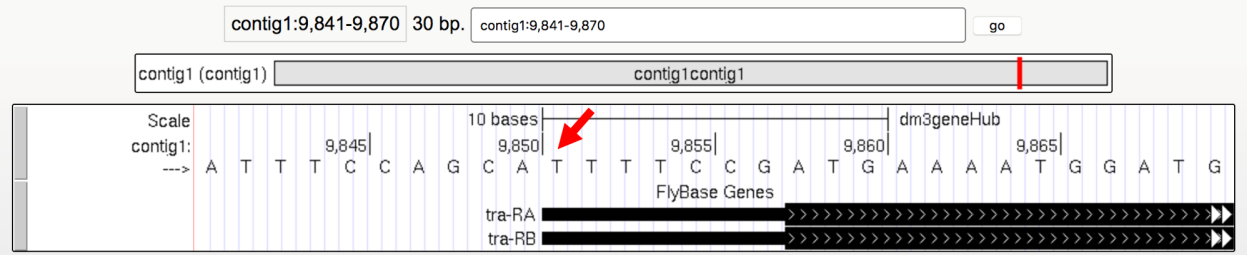


Figure 7.9.: The Base Position track shows the underlying genomic sequence for a region when you zoom in.

13. Look at the left end of the display, under the word *contig1*. The arrow here is pointing to the right. When you click on the −−−> arrow, the arrow will switch orientation and point to the left (Figure 7.10, top). In addition, the nucleotides in the "Base Position" track will also change from black to grey. Clicking on the <−−− arrow again will return it to its original orientation (Figure 7.10, bottom).

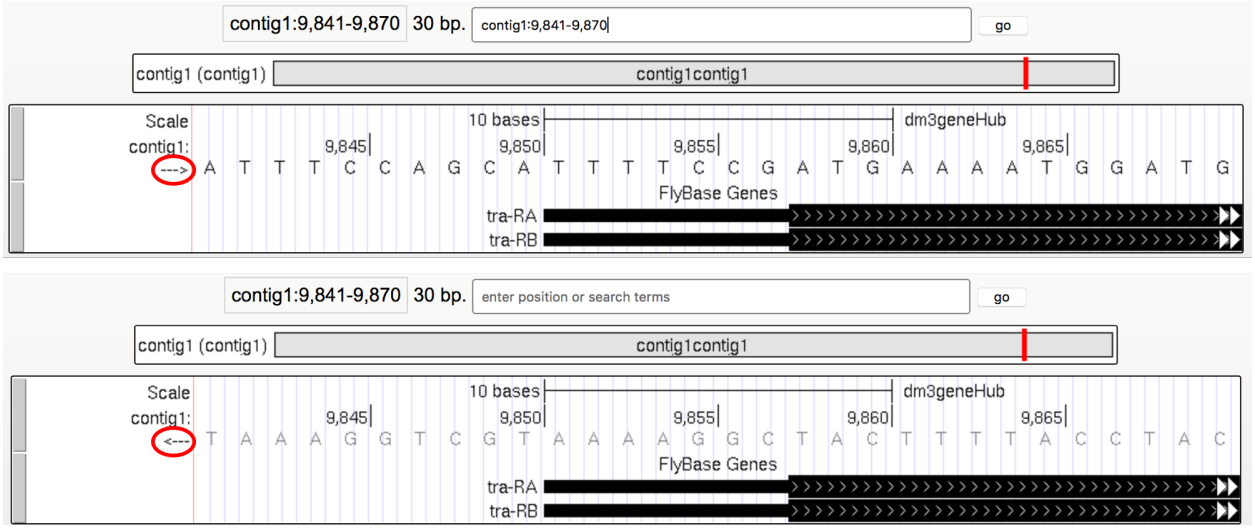


Figure 7.10.: Click on the arrow to change the nucleotides shown on the base position track.

Question 4

What is the relationship between the bases displayed when the arrow is pointed to the left versus when it is pointed to the right?

Question 5

Why do you think the bases are displayed in this way in the Genome Browser?

Both forms of the *tra* gene begin at 9,851 and they have the same prefix ("tra") but different suffixes ("-RB" and "-RA", respectively). The prefix corresponds to the name of the gene (*tra*) in *D. melanogaster* while the two suffixes indicate that there are two different versions (i.e. isoforms) of this gene. We will examine the differences between these two isoforms later. For now, we will focus our analysis on the A isoform of *tra* (tra-RA).

7.2 Genes are composed of exons and introns

14. To see the entire *tra* gene, type `contig1:9,800-10,860` in the *enter position or search terms* text box and click `go` (Figure 7.11). Alternatively, you can use the buttons next to the "zoom out" label and the arrows next to the "move" label to adjust the display.

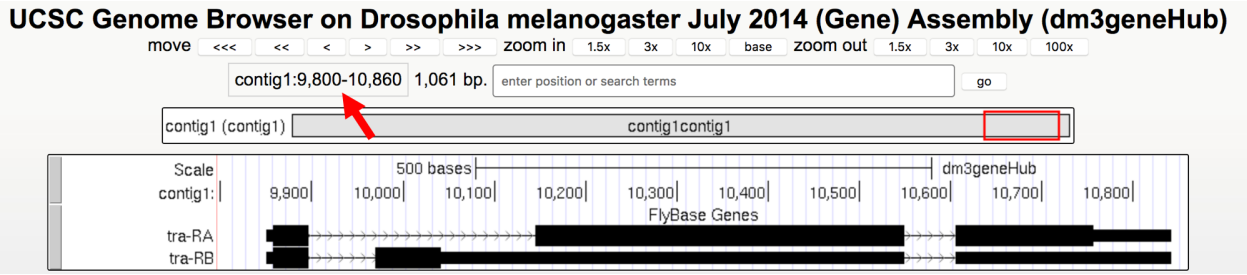


Figure 7.11.: The genomic region surrounding the *tra* gene.

15. Carefully examine the tra-RA isoform. Notice that the isoform consists of black blocks that are connected by lines. On the lines are arrowheads that point from left to right. The black blocks are the *exons* (expressed regions of the gene; Figure 7.12). To use the information stored in a gene, a cell uses the DNA sequence as a template to produce a molecule called a messenger RNA (mRNA). This process is called *transcription*. You will see in *Module 2* that while the initial transcript (product of transcription) is continuous, copying all the DNA, only exon sequences are retained in the processed mRNAs. The lines connecting the blocks are the *introns* (intervening regions of the gene). These sequences will be removed during the production of *mature mRNAs*. The arrows on the lines denote the direction of transcription (or orientation) of the gene.

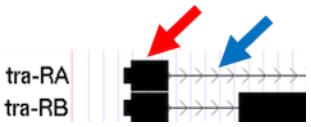


Figure 7.12.: The black boxes (indicated by the red arrow) are the exons and the lines connecting the blocks (indicated by the blue arrow) are the introns.

Question 6

How many exons does tra-RA contain?

Question 7

How many introns does tra-RA contain?

7.3 Genes provide the information to make proteins

The mRNA sequence contains the information that the cell needs to make proteins. You will learn more about this process in *Module 5*. Here we will use the Genome Browser to examine the basic features of a protein.

16. Type `contig1:9,850-9,875` into the *enter position or search terms* text box of the Genome Browser.

17. Scroll down to the "Mapping and Sequencing Tracks" section and change the display mode for the *Base Position* track to `full` (Figure 7.13).

18. Click on the `refresh` button to update your display.

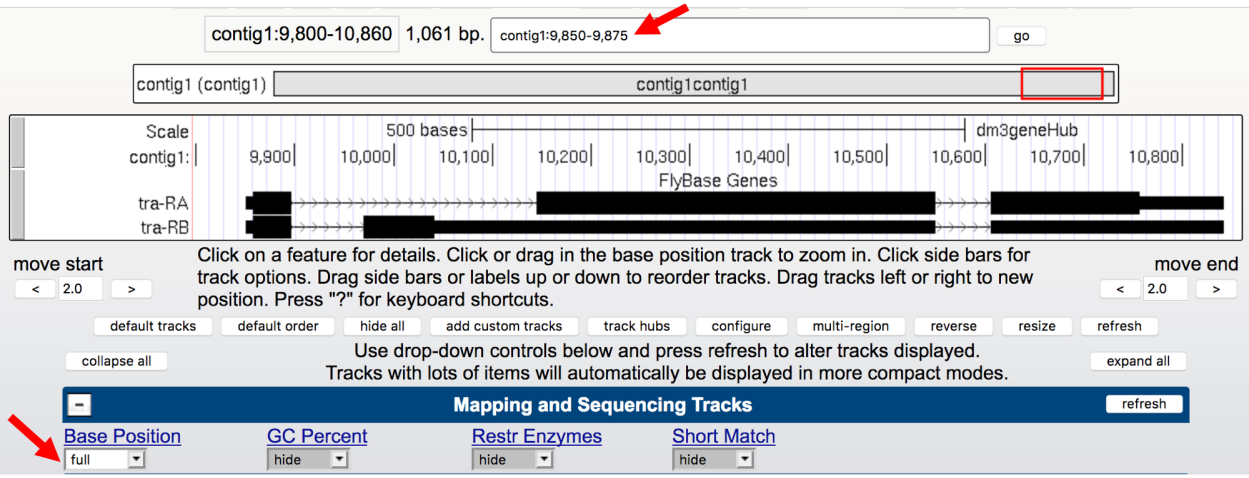


Figure 7.13.: Examine the "Base Position" track in the "full" display mode.

Proteins are made up of *amino acids*, and the mRNA provides the information for the amino acid sequence. This information is read by the cell in groups of three bases, with each three-base group (i.e. *codon*) specifying an amino acid. The Genome Browser uses single-letter abbreviations to represent each amino acid. These are shown on your Genome Browser as three new rows of information directly below the DNA sequence (Figure 7.14).

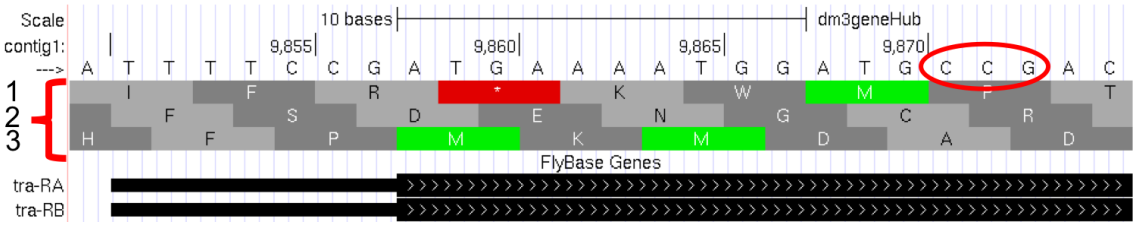


Figure 7.14.: Three new rows appear beneath the nucleotide sequence when the "Base Position" track is in "full" mode.

Question 8

Why do you think it takes three lines to display the amino acid information?

Tip: Remember that a codon is specified by three bases, e.g. CCG = Proline (circled in Figure 7.14).

Module 5 will have more details about *translation*, the process of copying the information from mRNA into protein. For now, we will just identify the beginning and the end of the protein. You should see three codons that are highlighted in green (one in row 1 and two in row 3). These codons all correspond to the amino acid M (i.e. Methionine). This amino acid is almost always used to start a protein. There is only one codon that can code for Methionine: **ATG**.

The first M on the third row of amino acids (at 9,858-9,860) corresponds to the start of the protein for the A isoform of *tra*. The position of this Methionine also coincides with the transition of the thinner rectangle to the thicker rectangle. Hence the thick rectangles denote coding sequence — the parts of the exon that carry information about the protein

sequence and are the translated parts — while the thin blocks indicate regions that are part of the exon but do not carry protein sequence information, or the untranslated parts (Figure 7.15).

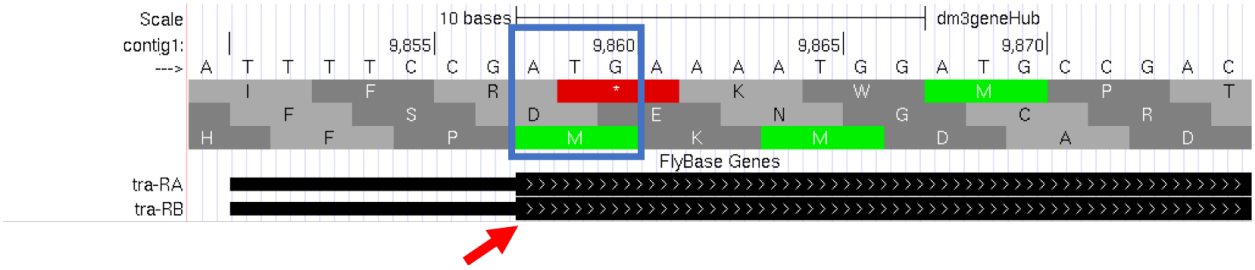


Figure 7.15.: The location of the initial Methionine for the A isoform of *tra*.

Let's examine the other end of the protein. There are three special codons (known as *stop codons*) that signal the end of the protein. These codons (TGA, TAA and TAG) are indicated by an asterisk "*" and are highlighted in red in the "Base Position" track.

19. Type `contig1:10,740-10,765` into the *enter position or search terms* text box and then click on the `go` button. Note the stop codon (*) at position 10,754-10,756, specified by the bases **TGA**, in the second row of amino acids (Figure 7.16). This is the last codon before the transition from the thick exon block to the thinner one. The Genome Browser therefore shows that a part of the mRNA extends beyond the end of the protein-coding region. This is a general property of mRNAs: they contain extra sequences both before and after the protein-coding sequence. These sequences, at the *5'* and *3'* end of the protein-coding sequences, are called the 5' and 3' *UTRs* (untranslated regions) respectively.

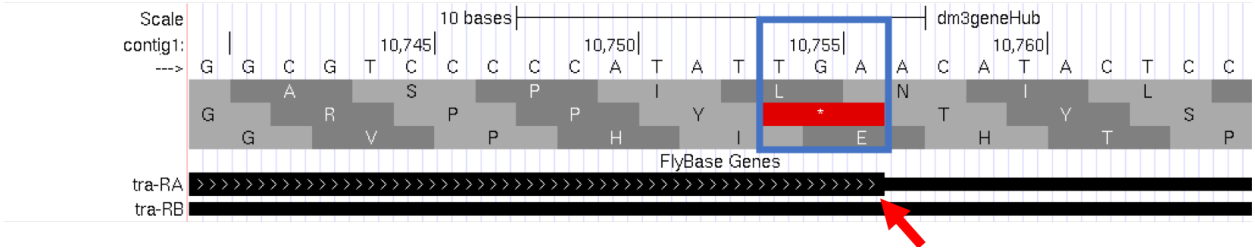


Figure 7.16.: The end of the translated region for the A isoform of *tra*.

7.3.1 Genes have directionality

As you saw above, the sequence of the codons in the A isoform of *tra* are read from left to right relative to the orientation of contig1. This also means that the start of the protein is located toward the left of the end of the gene. However, recall that DNA is double-stranded, and that the two strands run in opposite directions to each other (i.e. they are **antiparallel**). It turns out that, like the *tra* gene here, some genes are read on the DNA strand conventionally termed the "top strand" (from left to right), while other genes are read on the "bottom strand" (from right to left). We will examine one such example next.

20. Type `contig1:5,350-5,375` into the *enter position or search terms* text box and then click on the `go` button. This region contains the start of the protein-coding region of the *CG32165* gene. However, there are no Methionines (green boxes) at the transition point between the thin and thick rectangles (Figure 7.17, top). However, note that the arrows in the thicker part of the indicated exon point from right to left, indicating that this gene is read from the bottom strand.

21. Click on the `arrow` beneath the *contig1* label in the "Base Position" track so that it points in the same direction as indicated for the gene in this region. This will complement the sequence and allow you to read the bases of

the "bottom" strand of DNA. Remember that the codons on this strand must be read from right to left. Now you can see that there is a start codon in this region, the corresponding green M amino acid (at 5,365-5,367) in the third row (Figure 7.17, bottom).

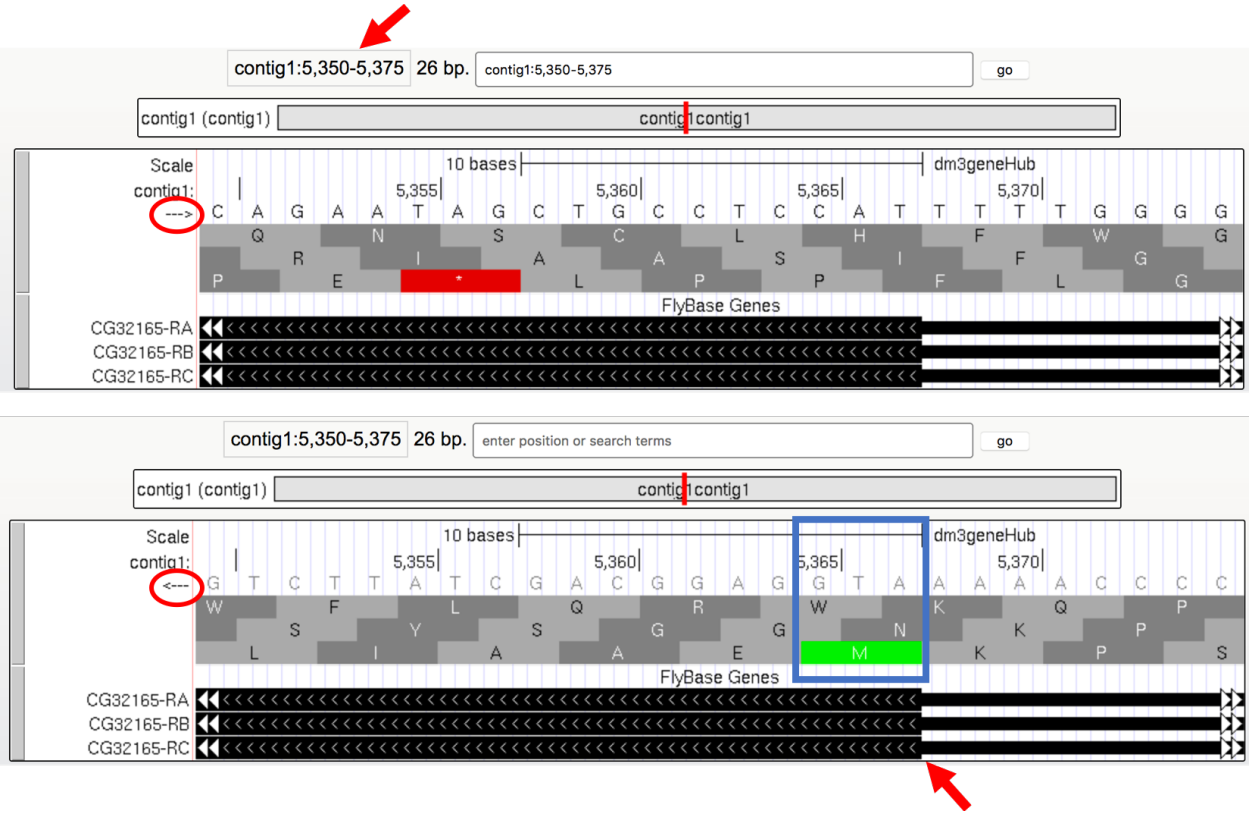


Figure 7.17.: Examine the start of the coding region for a gene on the minus strand.

7.4 Coding exons are translated in a single reading frame

The combination of the directionality (with two alternative directions) and the three rows in the "Base Position" track means that there are six different ways to translate a genomic region, i.e. to determine the sequence of amino acids from a DNA sequence. These different ways to translate a genomic region are known as reading *frames*.

22. To illustrate this concept, change the *enter position or search terms* text box to `contig1:1-12` and then click `go` in order to zoom in to the first 12 nucleotides of the contig1 sequence.

23. Click on the `arrow` underneath the *contig1* label in the "Base Position" track so that it points to the right (Figure 7.18).

The first row (frame +1) begins at the **first** nucleotide in contig1 and the first amino acid (P) is derived from the codon **CCC**. The second row (frame +2) begins at the **second** nucleotide in contig1 and the codon **CCG** also codes for the amino acid P. The third row (frame +3) begins at the **third** nucleotide in contig1 and the codon **CGG** corresponds to the amino acid R (Figure 7.19). Because a codon is comprised of 3 nucleotides, the codon beginning at the fourth nucleotide (GGT) is again in frame +1.

Examination of the "Base Position" track at the beginning of the contig shows that the three positive reading frames are numbered relative to the start of the contig1 sequence. Similarly, the three reading frames on the bottom strand are numbered relative to the end of the contig1 sequence (i.e. the beginning of the reverse complement of the contig sequence). Because contig1 has a total length of 11,000bp, we will change the *enter position or search terms* field to

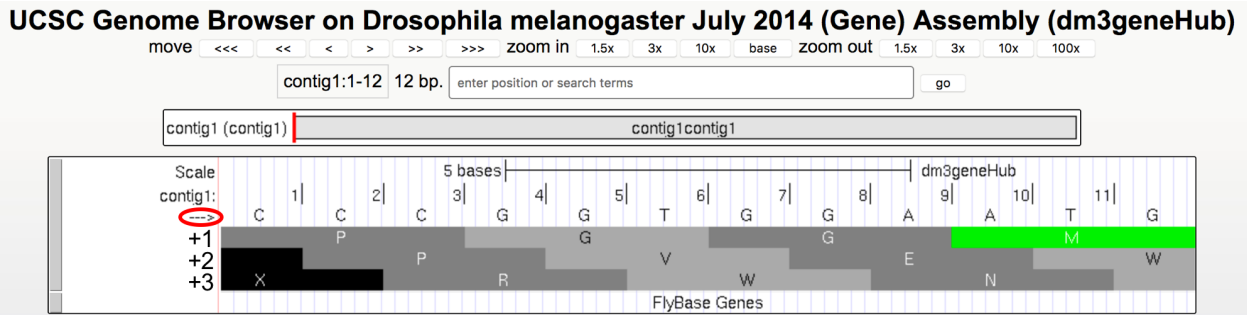


Figure 7.18.: Examine the "Base Position" track for the first 12 bases of contig1 in the top strand.

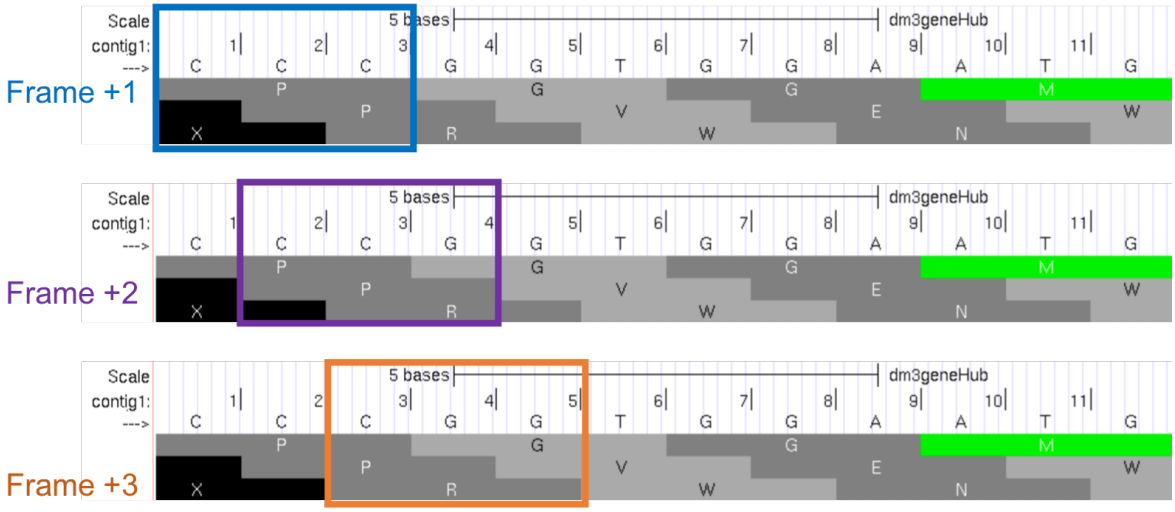


Figure 7.19.: Using the "Base Position" track to interpret the reading frames on the top strand.

`contig1:10,989-11,000` and then click on the `go` button so that we can examine the last 12 nucleotides of this contig.

24. Click on the `arrow` underneath the *contig1* label so that it points to the left (Figure 7.20).

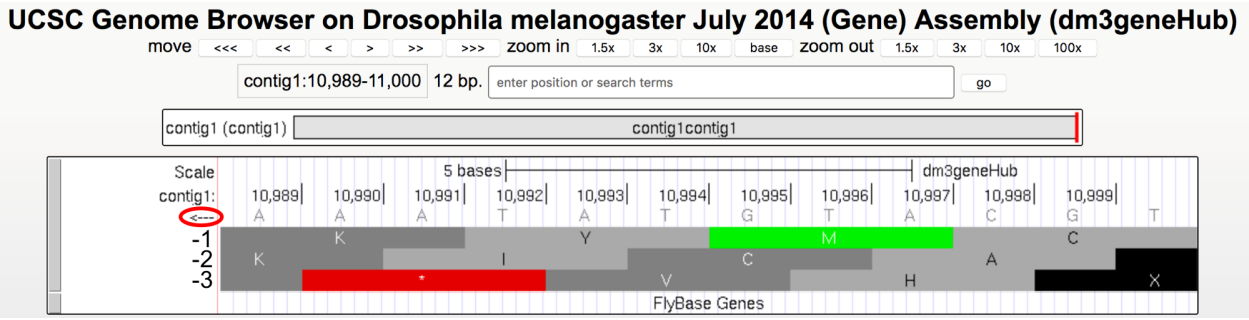


Figure 7.20.: Examine the "Base Position" track for the last 12 nucleotides of contig1 in the bottom strand.

Because we are examining the reverse complement of the contig1 sequence, we need to read the nucleotide and amino acid sequences on the "Base Position" track from right to left. The first row (frame -1) begins at the last nucleotide (11,000) of contig1 and the codon **TGC** codes for the amino acid C. The second row (frame -2) begins at the penultimate nucleotide at 10,999 and the codon **GCA** codes for the amino acid A. The third row (frame -3) begins at 10,998 and the codon **CAT** corresponds to the amino acid H (Figure 7.21).

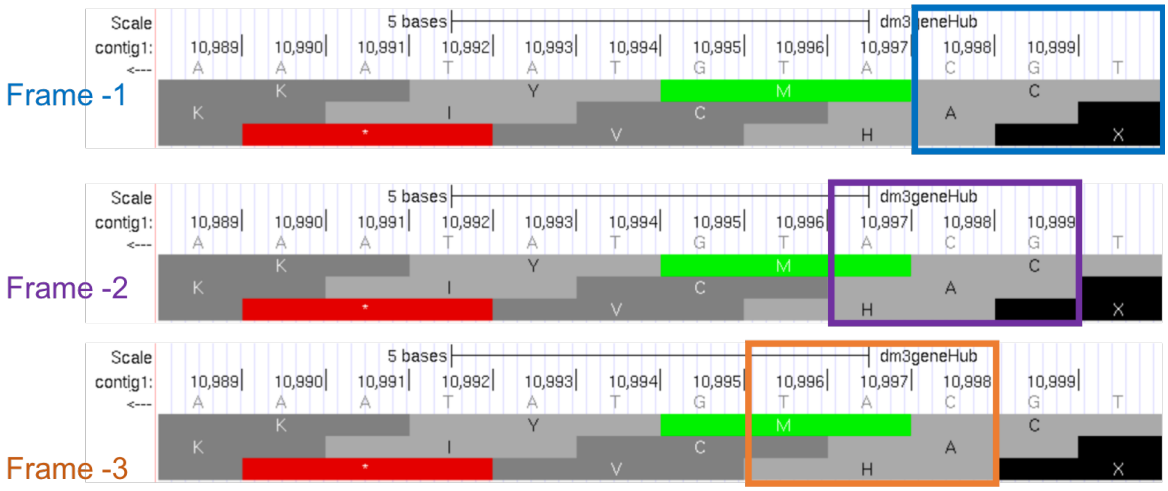


Figure 7.21.: Using the "Base Position" track to interpret the reading frames on the bottom strand.

25. Now that we understand how to interpret the reading frame information using the "Base Position" track, we can investigate the coding regions of the *tra* gene more closely. Change the *enter position or search terms* field to `contig1:9,800-9,960` and then click on the `go` button.

26. Click on the `arrow` underneath the *contig1* label in the "Base Position" track so that we can examine the translations of the top strand (running left to right) (Figure 7.22).

Our previous analysis shows that there is a *start codon* (green rectangle in the "Base Position" track) in the third row that corresponds to the transition from the thin to the thick rectangles (Figure 7.15). Hence the coding part of the first exon of the A isoform of *tra* is said to be "in frame +3". Notice that there is also an open reading frame (*ORF* — stretch of codons uninterrupted by stop codons) that overlaps with the thick box in the second row (frame +2) but there are no start codons that overlap with the thick box. In contrast, the first row (frame +1) contains a start codon, but the thick box also overlaps with a stop codon (red star). When we examine the region downstream of the black

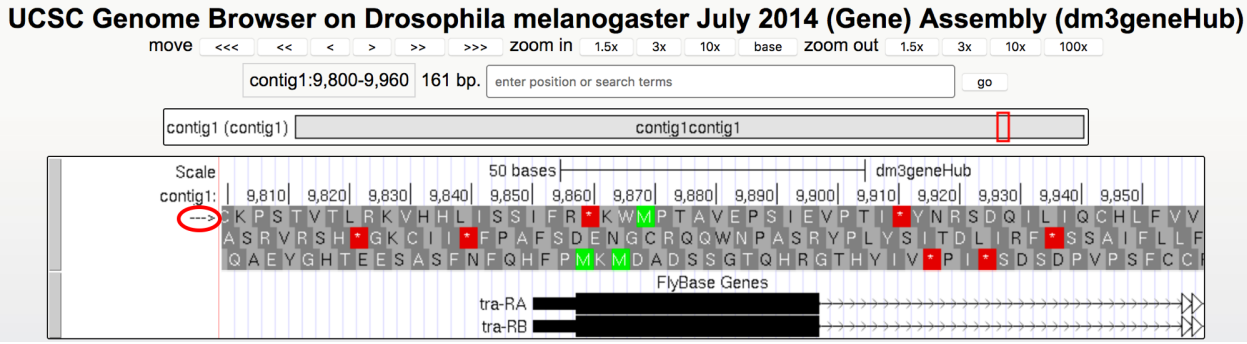


Figure 7.22.: The genomic region surrounding the first exon of tra-RA.

boxes, we find that there are stop codons in all three reading frames. However, these stop codons do not interrupt the open reading frame of the first exon because they occur in the region of the arrowed lines (i.e. the first intron, see blue arrows in Figure 7.23).

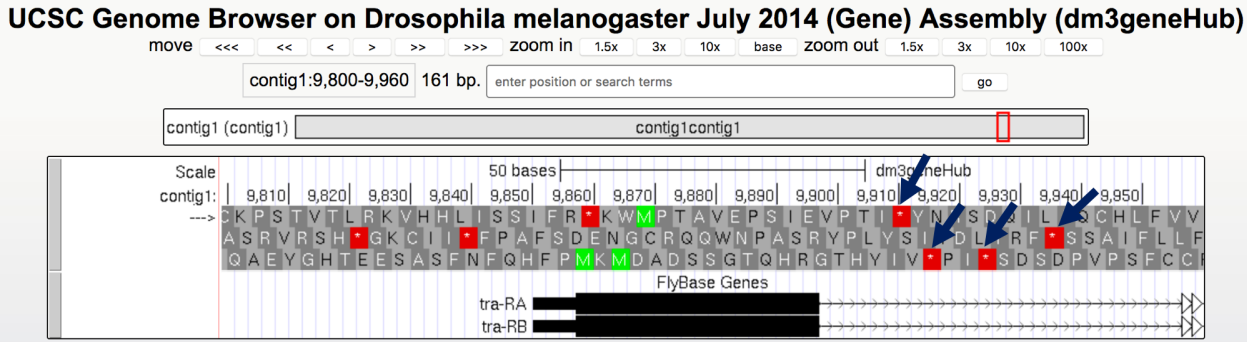


Figure 7.23.: Stop codons (red stars) are found in all three reading frames in the first intron of tra-RA.

27. Change the *enter position or search terms* field to `contig1:10,100-10,600` so that we can examine the second *coding exon* of the A isoform of *tra* to determine its reading frame.

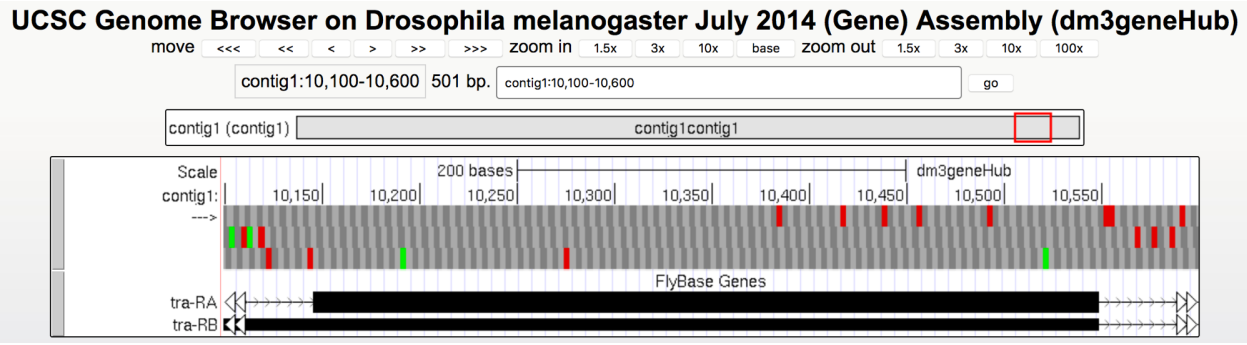


Figure 7.24.: The genomic region surrounding the second coding exon of tra-RA.

Question 9

Based on the screenshot shown in (Figure 7.24), which reading frame contains the amino acid sequence for the second coding exon of tra-RA?

28. Change the *enter position or search terms* field to `contig1:10,550-10,900` so that we can examine the region surrounding the last coding exon of the tra-RA isoform (Figure 7.25). Based on our previous analysis, we know that there is a stop codon in the second row that corresponds to the transition from the translated (thick rectangle) to the untranslated (thinner rectangle) region of the mRNA (Figure 7.16). Hence the last coding exon of tra-RA is in frame +2 (Figure 7.25).

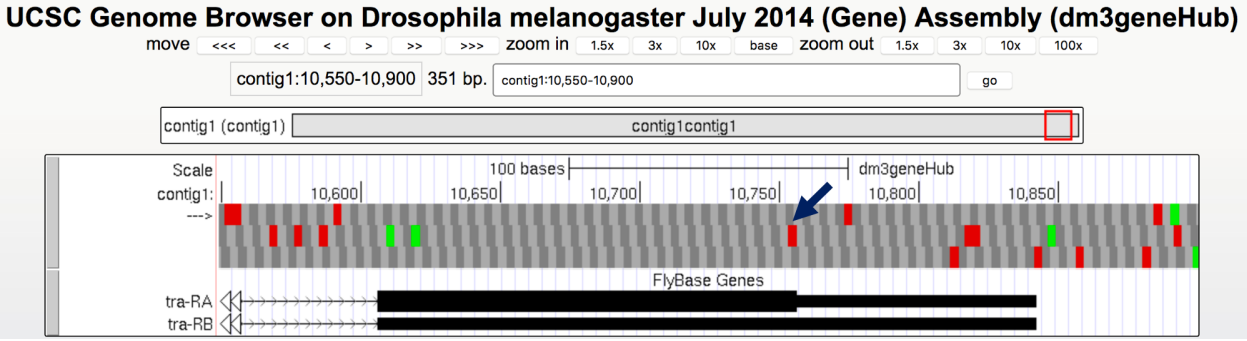


Figure 7.25.: The terminal coding exon of tra-RA is in frame +2.

Question 10

Does frame +2 have an ORF in the coding region of this exon? What about frame +1 and frame +3?

Question 11

Given that 3 of the 64 possible codons are stop codons, what is the chance of having a stop codon at any given position, assuming that the sequence is random?

Note: You might have noticed that the initial coding exon of tra-RA is in frame +3 while the last coding exon is in frame +2. We will learn more about mRNA processing in subsequent modules that will explain this apparent discrepancy.

7.5 Conclusion

In this lesson, you have learned to use the basic navigation features of the Genome Browser to examine the basic structure of a eukaryotic gene. To summarize:

- Genes provide the information to make proteins. This information is captured by transcribing the DNA to make RNA, and is carried on the mRNA in the form of three-base groups called codons.
- Genes are composed of exons and introns. Exons are regions retained in the processed mRNA, and are represented by black blocks in the browser, while introns are the regions that are removed during the process of creating the final mRNA, and are represented by lines connecting the blocks.
- The codon ATG in DNA (AUG in mRNA) specifies the amino acid M (Methionine) and is highlighted in green on the "Base Position" track of the Genome Browser. The first Methionine provides the starting signal for protein synthesis.

- The codons TAA, TAG, and TGA in DNA (UAA, UAG, and UGA in mRNA) encode the stop codon (*) and are highlighted in red on the "Base Position" track of the Genome Browser. The stop codons provide the ending signal for protein synthesis.
- Genes may be read either from left to right (top strand of the DNA), or from right to left (bottom strand of the DNA). Arrows on a gene indicate its directionality.
- Each row in the "Base Position" track (set on `full`) corresponds to a different reading frame. Different coding exons for a transcript can be in different reading frames.

29. To practice using the browser and reinforce the above concepts, examine the third gene in this contig (spd-2-RA):

Question 12

How many exons and introns are present in this gene?

Question 13

What is the orientation of this gene relative to contig1? How do you know? Where are the start codon and the stop codon — give the base position numbers (coordinates) of the start and the stop codon)?

You have now completed Module 1, and are ready to move on to *Module 2*.

7.6 Bonus question

Take a little time to explore some of the other evidence tracks on the browser.

Bonus Question 14

While looking at contig1 (size 11,000 bp), put the *GC Percent* track on `full`. What sort of pattern do you see, relative to the map of the genes? What can you conclude about gene structure?

CHAPTER 8

Module 1 Instructor Resources

8.1 Lesson Plan

8.1.1 Title

- Introduction to the Genome Browser: What is a Gene?

8.1.2 Objectives

- Demonstrate basic skills in using the UCSC Genome Browser to navigate to a genomic region and to control the display settings for different evidence tracks.
- Explain the relationships among DNA, pre-mRNA, mRNA, and protein.

8.1.3 Pre-requisites: knowledge of...

- DNA structure (base composition, anti-parallel double-stranded helix, base-pairing properties)
- Chromosome structure (a chromosome is a continuous DNA molecule, basic understanding of chromosome arms)
- Protein structure (proteins are made up of amino acids)

8.1.4 Order

- Warm Up
- Investigation
- Exit

8.1.5 Homework

- Discuss the question: What is a gene? (Discuss with a partner, then as a class.) Emphasize the *function* of a gene; consider how the structure of the gene is related to its function.
- Work through the genome browser investigation, with pauses to discuss the answers to the questions.
- Conclude with an emphasis on the main points:
- Genes may run in either direction on a chromosome;
- Genes are represented on the genome browser as blocks connected by lines;
- Eukaryotic genes are made up of protein-coding exons (the blocks) connected by introns;
- Proteins usually begin with a Methionine (M) and end at a stop codon (*)

8.1.6 Associated Videos

- Genome Browser video
- Tracks video

8.2 Module 1 Resources

The evidence tracks in the Genome Browser are grouped into three categories:

- **Mapping and Sequencing Tracks** Basically these contain the information obtained from sequencing that region of the chromosome. These tracks show the As, Ts, Cs, and Gs (`Base Position`) and matches to particular sequences of interest (`Short Match`) or to the cleavage sites for different "Restriction Enzymes".
- **Genes and Gene Prediction Tracks** These tracks show the genes as they are reported in the Drosophila Database (`FlyBase Genes`), and as predicted by a couple of computer programs (`Genscan Genes`, `N-SCAN Genes`). It also contains the transcription start site (TSS) annotations (`TSS Annotations`) and *D. melanogaster* cDNAs that have been mapped to contig1.
- **RNA Seq Tracks** These tracks show the results of sequencing mRNAs derived from a particular tissue and developmental time point. Most of the RNA-Seq reads are derived from processed mRNAs (where the introns have been removed). The mRNAs are broken into smaller fragments (e.g. via nebulization) prior to sequencing (usually using the Illumina HiSeq platform). The short reads (~100–125bp) are then mapped against the *D. melanogaster* genome. The y-axis of the `RNA-Seq Coverage` track shows the number of reads that has been mapped to each position of the contig (x-axis); this provides an estimate of the expression level. The `Exon Junctions` track shows the predicted locations of the introns. This track is derived from the subset of RNA-Seq reads that map partially to each of two adjacent exons (i.e. spliced RNA-Seq reads).

Tip: For manipulating tracks, students may need to be reminded to read carefully what is immediately under the displayed tracks: Click on a feature for detail. Click sidebars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to a new position.

CHAPTER 9

Module 2 Instructor Resources

9.1 Lesson Plan

9.1.1 Title

- Transcription Part I: From DNA sequence to transcription unit

9.1.2 Objectives

- Describe how a primary transcript (pre-mRNA) can be synthesized using a DNA molecule as the template.
- Explain the importance of the 5' and 3' regions of the gene for initiation and termination of transcription by RNA polymerase II
- Identify the beginning and end of a transcript using the capabilities of the genome browser (RNA-Seq, Short Match)

9.1.3 Pre-requisites

- Structure of a gene (*Module 1*)

9.1.4 Order

- 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

9.1.5 Homework

- None

9.1.6 Class Instruction

- 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 *tra* gene. How long is the pre-mRNA?
- Conclude by challenging students to think about these questions:
  - 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 transcription 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?

9.1.7 Associated Videos

- RNA Seq and TopHat video
- Short Match video

CHAPTER 10

Module 3 Instructor Resources

10.1 Lesson Plan

10.1.1 Title

- Transcription Part II: What happens to the initial (pre-mRNA) transcript made by RNA pol II?

10.1.2 Objectives

- Explain how the transcript generated by RNA polymerase II (the pre-mRNA) is processed to become mature mRNA, using the sequence signals identified in *Module 2*.
- Use the genome browser to analyze the relationships among:
    1. pre-mRNA
    2. 5' capping
    3. 3' polyadenylation
    4. Splicing
    5. mRNA

10.1.3 Pre-requisites

- *Module 1* and *Module 2*
- Define pre-mRNA as the RNA that results from the process of transcription; this initial transcript includes exons and introns

10.1.4 Order

- Warm Up Discussion
- Investigation

10.1.5 Homework

- None

10.1.6 Class Instruction

- Discuss the question: What happens to the initial (pre-mRNA) transcript made by RNA pol II? Does it leave the nucleus "as is"? Or do changes have to occur? (Hint: introns vs. exons) (Discuss with a partner then as a class).
- Mini-presentation illustrating that during pre-mRNA processing, three events occur:
    1. 5' capping
    2. 3' polyadenylation
    3. Splicing out of introns
- Work through the genome browser investigation, with pauses to discuss the answers to the questions.
- Conclude with emphasis on main points
- Pre-mRNA is processed using 3 steps:
    1. 5' capping
    2. 3' polyadenylation
    3. Removal of introns through splicing (via spliceosome)

CHAPTER 11

Module 4 Instructor Resources

11.1 Lesson Plan

11.1.1 Title

- Removal of introns from messenger RNA by splicing

11.1.2 Objectives

- Identify splice donor and acceptor sites that are best supported by RNA-Seq and TopHat splice junction predictions
- Utilize the canonical splice donor and splice acceptor sequences to identify intron-exon boundaries.

11.1.3 Pre-requisites

- *Module 1*
- *Module 2*
- *Module 3*

11.1.4 Order

- Review mRNA processing; Investigation 1
- Define isoform
- Introduce consensus sequences; Investigation 2
- Discuss pre-mRNA vs. mRNA
- Explore how to find exon-intron junction using tra-RA as the example; Investigation 3

11.1.5 Homework

- Students will identify intron-exon junctions for spd-2-RA

11.1.6 Class Instruction

- Review pre-mRNA processing using appropriate figures from the textbook or *Module 3*
- Investigation 1: Students familiarize themselves with RNA-Seq data
- Review consensus sequences for splice donor and splice acceptor sites
- Investigation 2: Students find splice donor and acceptor for intron 1
- Review RNA splicing of intron 1 using tra-RA as the example
- Investigation 3: Students find remaining splice donor and splice acceptor for intron 2
- Discuss length of pre-mRNA vs. length of spliced mRNA
- Identify isoforms with different TSSs, or alternative splicing patterns

11.1.7 Associated Videos

- RNA-Seq and TopHat video
- Genes and Isoforms video

CHAPTER 12

Module 5 Instructor Resources

12.1 Lesson Plan

12.1.1 Title

- Translation: The need for an Open Reading Frame

12.1.2 Objectives

- Determine the codons for specific amino acids and identify reading frames by looking at the Base Position track in the genome browser
- Assemble exons to maintain the open reading frame (ORF) for a given gene
- Define the phases of the splice donor and acceptor sites and describe how they impact the maintenance of the ORF
- Identify start and stop codons of an assembled ORF

12.1.3 Pre-requisites

- *Module 1*
- *Module 2*
- *Module 3*
- *Module 4*
- Overview of the ribosome, tRNAs, and associated proteins involved in translation (Initiation Factors, Elongation Factors and Release Factors)
- Overview of the DNA codon table

12.1.4 Order

- Warm Up/Review of Pre-requisites
- Investigation 1
- Investigation 2
- Exit

12.1.5 Homework

- None

12.1.6 Class Instruction

- Review the process of translation: Overview of the ribosome, tRNAs, and associated proteins involved in translation (Initiation Factors, Elongation Factors and Release Factors)
- Review the DNA codon table
- Work through the activities using the Genome Browser, with pauses to discuss the answers to the questions.
- Conclude with emphasis on main points:
  - mRNAs are translated into amino acids using triplet codons
  - Identification of ORFs
  - The ORF must be maintained across splice sites to generate a working mRNA
  - The assembled ORF begins with a start codon and ends with a stop codon.

12.1.7 Associated Videos

- Splicing and Phase video

CHAPTER 13

Module 6 Instructor Resources

13.1 Lesson Plan

13.1.1 Title

- Alternative splicing

13.1.2 Objectives

- Demonstrate how alternative splicing of a gene can lead to different mRNAs.
- Show how alternative splicing can lead to the production of different polypeptides and result in drastic changes in phenotype.

13.1.3 Pre-requisites

- *Module 1*
- *Module 2*
- *Module 3*
- *Module 4*
- *Module 5*

13.1.4 Order

- Introduce students to tra-RB, a second isoform of *tra*.
- Investigation 1
- Investigation 2

- Discussion

13.1.5 Homework

- None included. Students could analyze a second gene on the browser using the work they have done on *tra* as a template.

13.1.6 Class Instruction

- Introduce tra-RB
- Discuss differences between tra-RB and tra-RA. Reinforce concept of isoform.
- Investigation 1: How can there be different mRNAs encoded in the same gene?
- Investigation 2: Examine the *tra* polypeptides by looking at the three possible reading frames. Review concept of reading frame and introduce phase if not previously introduced. Students will construct a gene model for tra-RB, using sequence information, and RNA-Seq data as evidence.
- Discussion of gene models/wrap-up

13.1.7 Associated Videos

- Genes and Isoforms video
- Splicing and Phase video
- RNA Seq and TopHat video

CHAPTER 14

Glossary

Glossary of terms

Authors Margaret Laakso, Carina Howell, Cathy Silver Key, Leocadia Paliulis, Maria Santisteban, Chiyedza Small, Joyce Stamm, and Elena Gracheva

Last Update May 27, 2019

Version 0.0.1

3' Refers to the third carbon of the nucleic acid sugar moiety to which additional nucleotides may be added by polymerase, often used to refer to that end of a single-stranded DNA or RNA molecule where the 3' carbon is unattached to an adjacent nucleotide; cf. 5'.

5' Refers to the fifth carbon of the nucleic acid sugar moiety, to which the triphosphate is attached in a nucleotide triphosphate, often used to refer to that end of a single-stranded DNA or RNA molecule where the 5' carbon's phosphate group is unattached to an adjacent nucleotide; cf. 3'.

alternative splicing The inclusion or exclusion of certain exons in the splicing reactions that determine the sequences included in the final mRNA product. This mechanism is utilized to generate a series of closely related protein isoforms, which differ by the inclusion or exclusion of the particular protein domains encoded by those exons. Alternative splicing is directed by RNA-binding proteins that block, or stimulate, utilization of a particular splice site.

amino acid The basic building block of proteins, a small molecule with a -C-C- core, an amino group at one end and a carboxylic acid group at the other end. The basic structure can be represented as NH2-CHR-COOH, where R can be any of 20 different moieties, including acidic, basic, or hydrophobic groups.

annotation Gene annotation is the process of indicating the location, structure, and identity of genes in a genome. As this may be based on incomplete information, gene annotations are constantly changing with improved knowledge. Gene annotation databases change regularly, and different databases may refer to the same gene/protein by different names, reflecting a changing understanding of protein function.

antisense strand Also called the negative, template, or non-coding strand. This strand of the DNA sequence of a single gene is the complement of the 5' to 3' DNA strand known as the sense, positive, non-template, or coding strand. The term loses meaning for longer DNA sequences with genes on both strands.

base Although formally incorrect (the nitrogenous base which gives each nucleotide its name is only part of the nucleotide), this is often used as a synonym for "nucleotide."

base pair (base pairing) The hydrogen bonding of one of the bases (A, C, G, T, U) with another, as dictated by the optimization of hydrogen bond formation in DNA (A-T and C-G) or in RNA (A-U and C-G). Two polynucleotide strands, or regions thereof, in which all the nucleotides form such base pairs are said to be complementary. In achieving complementarity, each strand of DNA can serve as a template for synthesis of its partner strand- the secret of DNA replication's extremely high accuracy and thereby of inheritance.

cDNA "complementary DNA," a double-stranded DNA molecule prepared in vitro by copying an RNA molecule back into DNA using reverse transcriptase. The RNA component of the resulting RNA-DNA hybrid is then destroyed by alkali, and the complementary strand to the remaining DNA strand synthesized by DNA polymerase. The resulting double-stranded DNA can be used for cloning and analysis.

CDS "Coding sequence", that part of the DNA sequence of a gene which is translated into protein.

coding exon In a gene, any exon which contains some part of the CDS; in contrast, an exon which has no part translated into protein is called a "non-coding exon."

coding strand In a gene, the DNA strand that has the sequence found in the RNA molecule. Also called the sense, positive, or non-template strand.

codon The sequence of three nucleotides in DNA or RNA that specifies a particular amino acid.

coordinates Numerical position within a biological sequence, e.g. the first base in a DNA sequence would have the coordinate 1.

exon An exon is a contiguous segment of eukaryotic DNA that corresponds to a portion of the mature (processed) RNA product of that gene. Exons are found only in eukaryotic genomes, and are separated by introns. Although the introns are transcribed with the exons, the latter are spliced out and discarded during RNA processing.

frame A frame is a single series of adjacent nucleotide triplets in DNA or RNA: one frame would have bases at positions 1, 4, 7, etc. as the first base of sequential codons.

There are 3 possible reading frames in an mRNA strand and six in a double stranded DNA molecule due to the two strands from which transcription is possible. Different computer programs number these frames differently, particularly for frames of the negative strand, so care should be taken when comparing designated frames from different programs.

initiation codon (start codon) The first codon of a coding sequence. In eukaryotes this is almost always ATG, which codes for Methionine.

intron Non-coding sections of a eukaryotic nucleic acid sequence found between exons. Introns are removed ("spliced out") of mRNA after transcription and before the molecule is exported to the cytoplasm for translation; cf. exon.

isoform Alternate forms of a gene that are produced by alternative splicing of a particular mRNA, or different transcription start sites. Isoforms of a gene always have different mRNA sequences, but they may have the same protein sequence.

mature mRNA Messenger RNA that has been completely processed; it has a 7-methylguanosine cap at its 5' end, a poly (A) tail at its 3' end, and has all its introns spliced out.

non-coding strand Also called the negative, template, or anti-sense strand. This strand of the DNA sequence of a single gene is the complement of the 5' to 3' DNA strand known as the sense, positive, non-template, or coding strand. The term loses meaning for longer DNA sequences with genes on both strands.

ORF "Open Reading Frame", a long stretch of codons in the same reading frame uninterrupted by stop codons; an ORF may reflect the presence of a gene.

phase The phase describes the number of bases between the end of the exon (defined by the splice site) and the full codon nearest that splice site. The number of bases between the adjacent full codon and an exon/splice site can

be either 0, 1 or 2. The phase of an upstream exon will determine which frame is translated in the downstream exon as it will indicate how many bases are used after the acceptor splice site to create a full codon of 3 bases.

poly(A) tail About 250 nucleotides of adenylate residues that are post-transcriptionally added by poly (A) polymerase to the 3' end of eukaryotic mRNA following cleavage of the newly synthesized RNA about 20 nucleotides downstream of an AAUAAA signal sequence.

pre-mRNA The initial transcript from a protein-coding gene is often called a pre-mRNA and contains both introns and exons. Pre-mRNA requires the addition of a 5' cap and 3' poly (A) tail and the removal of introns to produce the final mRNA molecule containing only exons.

promoter A segment of DNA to which RNA polymerase binds to initiate transcription of the downstream gene(s).

read A raw DNA sequence.

splicing The process by which introns are removed and exons are joined to produce a mature, functional RNA from a primary transcript. Some RNAs are self-splicing, but most require a specific ribonucleoprotein complex to catalyze the reaction.

splice acceptor site The boundary between an intron and the exon immediately downstream (i.e., on the 3' side of the intron).

splice donor site The boundary between an intron and the exon immediately upstream (i.e., on the 5' side of the intron).

splice junction Either a splice acceptor site or a splice donor site.

stop codon (termination codon) A codon that specifies the termination of peptide synthesis; sometimes called "nonsense codons," since they do not specify any amino acid.

transcription The process of copying one strand of a DNA double helix by RNA polymerase, creating a complimentary strand of RNA called the transcript.

translation The process by which codons in an mRNA are used by the ribosome to direct protein synthesis.

UTR "Untranslated region", a segment of DNA (or RNA) which is transcribed and present in the mature mRNA, but not translated into protein. UTRs may occur at either or both the 5' and 3' ends of a gene or transcript.

CHAPTER 15

References

References

The GEA version of the "Understanding Eukaryotic Genes" modules are based on the following publication on CourseSource:

- Laakso, M.M., Paliulis, L.V., Croonquist, P., Derr, B., Gracheva, E., Hauser, C., Howell, C., Jones, C.J., Kagey, J.D., Kennell, J., Silver Key, S.C., Mistry, H., Robic, S., Sanford, J., Santisteban, M., Small, C., Spokony, R., Stamm, J., Van Stry, M., Leung, W., Elgin, S.C.R. 2017. An undergraduate bioinformatics curriculum that teaches eukaryotic gene structure. CourseSource. https://doi.org/10.24918/cs.2017.13

Index