MODULE 4: SPLICING

Meg Laakso

Lesson Plan:

| Title | Removal of introns from messenger RNA by splicing |
|-------------------|---|
| 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. Predict the location of the transcription start site and the number of exons and introns in a gene using RNA Seq evidence tracks (Investigation 1). Identify splice donor and acceptor sites using the genome browser (Investigation 2). |
| Pre-requisites | Modules 1-3 |
| 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 |
| Homework | Students will identify intron-exon junctions for spd-2-RA |
| 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 |
| Associated videos | RNA-Seq and TopHat Video: <u>https://www.youtube.com/watch?v=qepVXEsfLMM</u> Genes and Isoforms Video: <u>https://www.youtube.com/watch?v=8jtTp_6vN4M</u> |

INVESTIGATION 1: EXAMINING RNA-SEQ DATA

Learning Objectives:

• Predict the location of the transcription start site and the number of exons and introns in a gene using RNA Seq evidence tracks.

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.

 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 <u>http://gander.wustl.edu</u>. Follow the instructions given in Module 1 to navigate to the contig1 project in the *D. melanogaster* July 2014 (Gene) assembly. *Reminder*: use the dropdown menus on the Genome Browser to select the following:

- 1. Select "D. melanogaster" under the "REPRESENTED SPECIES" field.
- 2. Select "July 2014 (Gene)" in the "Assembly" field.
- 3. Enter "contig1" in the "Position/Search Term" field.
- 4. Then click 'GO' and the screen below will appear (Figure 1).

| Scale 5 kb 4,000 5,000 6,000 7,000 8,000 9,000 10, CG32165-RC FyBase Genes tra-R tra-R tra-R tra-R CG32165-RB CG32165-RB spd-2-RA tra-R tra-R CG32165-RB CG32165-RB spd-2-RA tra-R CG32165-RB Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts. art Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts. default tracks default order hide all add custom tracks configure multi-region reverse resize refresh ollapse all Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes. e Mapping and Sequencing Tracks Short Match ited all e nse and Gene Prediction Tracks fide all ited all ited all ited all Genes and Gene P | | contig1:1 | -11,000 11,0 | 000 bn enter nos | sition or search | terms | | | go | |
|--|--|---|---|---|---|---|--|--|-------------------|-------------------------------------|
| contigit: 1,000 2,000 3,000 4,000 5,000 6,000 7,000 8,000 9,000 10, CG32165-RB FlyBase Genes FlyBase Genes tra-R CG32165-RA spd-2-RA spd-2-RA tra-R CG32165-RB spd-2-RA spd-2-RA spd-2-RA art Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts. multi-region reverse resize refresh ollapse all Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes. e Se Position GC Percent Restr Enzymes Short Match hide nse < Inide Nide Nide Inide Base Genes Genes and Gene Prediction Tracks TSS Annotations tra Is hide | | contig r. | -11,000 11,0 | enter pos | sition of search | | | - | go | |
| CG32165-RC CG32165-RA CG32165-RA CG32165-RA Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts. default tracks default order hide all add custom tracks configure multi-region reverse resize refresh Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes. Mapping and Sequencing Tracks se Position nse v CGenes and Gene Prediction Tracks CGase Genes ck v | | 1,000 | 2,000 | | 4,000 | 5,000 6,000 | 7,000 | | | 10,000 |
| Art Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts. default tracks default order Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes. Base Position GC Percent hide ■ Restr Enzymes Short Match hide ■ Base Genes Genes and Gene Prediction Tracks Base Genes Genscan Genes Nide ■ Nide ■ | CG32165-RB | | | < : < < < < < : < < < < < : < < < < | | | | | | tra-RB + tra-RA →- |
| ollapse all Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes. Mapping and Sequencing Tracks se Position nse • GC Percent hide • Restr Enzymes hide • Short Match hide • Genes and Gene Prediction Tracks Genes and Genes hide • TSS Annotations hide • tracks hide • | | | eature for de | talls. Click or dr | rad in the ba | ase position track to | zoom in. C | | ars for | |
| See Position GC Percent Restr Enzymes Short Match nide nide nide nide Genes and Gene Prediction Tracks Base Genes Genscan Genes N-SCAN Genes nide nide nide nide | r t | rack option osition. Pr | ns. Drag side r <mark>ess</mark> "?" for ke | bars or labels u eyboa <mark>rd shortcu</mark> | up or down t uts. | to reorder tracks. D | J | Ű | | < |
| nse v hide v hid | III > t Gefa | rack option osition. Pr | ns. Drag side ess "?" for ke default order Use | bars or labels u eyboard shortcu hide all add drop-down cont | up or down t uts. I custom tracks trols below a | to reorder tracks. D configure mul- and press refresh t | ti-region re o alter track | verse resi s displayed | ize refre: | sh |
| Base Genes Genscan Genes N-SCAN Genes D. mel. cDNAs TSS Annotations tra Is ck | art t > F defa | rack option osition. Pr | ns. Drag side ess "?" for ke default order Use | bars or labels u eyboard shortcu hide all add drop-down cont with lots of items | up or down t its. I custom tracks trols below a s will automa | to reorder tracks. E configure mut and press refresh t atically be displaye | ti-region re o alter track | verse resi s displayed | ize refre: | < |
| ck v hide v hide v hide v hide v hide v hide v | ollapse all se Position | cack option position. Pr nult tracks | ns. Drag side ess "?" for ke default order Use Tracks v <u>C Percent</u> | bars or labels u eyboard shortcu hide all add drop-down cont with lots of items Mapping Restr I | up or down t uts. I custom tracks trols below a s will automa g and Sequ <u>Enzymes</u> | to reorder tracks. E configure mut and press refresh t atically be displaye tencing Tracks Short Match | ti-region re o alter track | verse resi s displayed | ize refre: | sh |
| | ollapse all se Position | cack option position. Pr nult tracks | ns. Drag side ess "?" for ke default order Use Tracks v <u>C Percent</u> | bars or labels u eyboard shortcu hide all add drop-down cont with lots of items Mapping Restr hide | up or down t uts. d custom tracks trols below a s will automa g and Sequ Enzymes | to reorder tracks. I configure mut and press refresh t atically be displaye tencing Tracks Short Match hide | ti-region re o alter track | verse resi s displayed | ize refre: | sh expand refre |
| RNA Seq Tracks | III + t defa ollapse all se Position rse + Base Genes | rack option osition. Pr uult tracks Gi hi | ns. Drag side ess "?" for ka default order Use Tracks v <u>C Percent</u> de <u></u> | bars or labels u eyboard shortcu hide all add drop-down cont with lots of items Mapping Restr hide Genes ar is N-SCA | up or down t uts. I custom tracks trols below a s will automa g and Sequ Enzymes I nd Gene Pr | to reorder tracks. E configure mult and press refresh to atically be displaye tencing Tracks Short Match hide T ediction Tracks D. mel. cDM | il-region re o alter track ed in more co | verse res s displayed ompact mod | ize refre des. | sh expand refre tra Isofor |

FIGURE 1 A SCREEN SHOT OF 'CONTIG1' PROJECT.

As you will remember, this section of DNA is 11,000 base pairs long 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. Note that you will not be able to see the DNA sequence until you "zoom in" at step 7.
- 3. Change the RNA-Seq coverage to "full," then click "refresh."

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 2**).
- 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.

| move | <<< | << | < > | >> | >>> | zoom in | 1.5x | Зх | 10x b | ase Z | oom ou | 1.5x | Зх | 10x | 100x |
|--|-------------------------------------|--|---|--|--|--|---|---|--|---|--|---|------------------------------------|---------------------------|---|
| | contig | :9,800 | -10,900 | 1,101 | bp. c | ontig1:9,800: | -10,900 | | | | | | | go | |
| Scale contig1 | | ,900 | 10,000 | 500 b; | 0,100 | 10,200 | | 1,300 | 10,400 | | 0,500 | 10,600 | p · | gene 10,700 | 10,800 |
| t Females | | | | | | ENCODE F | | | | | | | | | |
| t Males 0_ | | | | | mo | dENCODE | RNA-Se | eq from D | melano | gaster | Whole A | dult Ma | es | | |
| | | | | | | | | | | | | | | | |
| start 2.0 > | track o | ptions. | | le bars | or lab | or drag in els up <mark>or</mark> ortc <mark>uts</mark> . | | | | | | | | | m < 2.0 |
| 2.0 > | track o | ptions. n. Pres | Drag sic s "?" for fault order Use | le bars keyboa hide | or lab ard sho all down | els up or | down to tracks elow ar | configu | re m refresh | Drag | tracks le | eft or rig verse displag | ht to r resize red. | new refi | |
| 2.0 > de | track o position | ptions. n. Pres | Drag sic s "?" for fault order Use | le bars keyboa hide | or lab ard sho all down ots of it | els up or ortcuts. add custor controls b | down to tracks elow an automat | configu nd press tically be | re m refresh display | Drag | tracks le | eft or rig verse displag | ht to r resize red. | new refi | < 2.0 resh |
| 2.0 > de | track o position | ptions. n. Presa de | Drag sic s "?" for fault order Use | le bars keyboa hide | or lab ard sho all down ots of it Map | els up or ortcuts. add custor controls b tems will a | down to htracks below ar automat Seque | configu configu nd press tically be ncing T | re m refresh display racks | Drag ulti-regin to alte yed in | tracks le | eft or rig verse displag | ht to r resize red. | new refi | < 2.0 resh expand al |
| 2.0 > de collapse all Base Position | track o position | ptions. n. Press de <u>GC P</u> | Drag sic s "?" for fault order Use Tracks | le bars keyboa hide | or lab ard sho e all down ots of it Map | els up or ortcuts. add custor controls b tems will a ping and estr Enzyr | down to n tracks eelow ar automat Seque mes | o reorder configu nd press tically be ncing T Sho hide | refresh display racks | Drag ulti-regin to alte yed in | tracks le | eft or rig verse displag | ht to r resize red. | new refi | < 2.0 resh expand al |
| 2.0 > de collapse all Base Position dense v | track o position fault tracks | ptions. n. Press de <u>GC P</u> hide | Drag sic s "?" for fault order Use Tracks | le bars keyboa hide drop with lo | or lab ard sho all down ots of it Map Re hi Gene | add custom controls b tems will a pping and estr Enzyr | down to htracks below an automat Seque mes ene Prec | o reorder configu nd press tically be ncing T Sho hide | re m refresh display racks rt Matc Tracks | Drag ulti-regin to alte yed in | tracks le | eft or rig verse display mpact | ht to r resize /ed. modes | n <mark>ew</mark> refi | < 2.0 resh expand al refresh |
| 2.0 > de collapse all Base Position dense v FlyBase Gene | track o position fault tracks | ptions. . Press de <u>GC P</u> hide <u>Gens</u> | Drag sic s "?" for fault order Use Tracks ercent | le bars keyboa hide drop with lo | or lab ard sho all down ots of it Map Re hi Gene | add custom controls b tems will a ping and estr Enzyr de s and Ge SCAN Ge de | down to htracks below an automat Seque mes ene Prec | o reorder configu nd press tically be ncing T Sho hide diction | re m refresh display racks rt Matc Tracks | Drag ulti-regin to alte yed in | tracks le on re er tracks more cc | eft or rig verse display mpact | ht to r resize /ed. modes | n <mark>ew</mark> refi | < 2.0 esh expand al refresh tra Isoform |

FIGURE 2 HISTOGRAMS REPRESENTING RNA-SEQ DATA IN TRANSFORMER (TRA).

Q1. List two ways in which the histograms are similar.

a._____b.

Q2. List one way that the histograms differ (other than color).

a._____

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).

Q3. 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.

Q4. 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.

- **Q5.** 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 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: <u>https://www.youtube.com/watch?v=8jtTp_6vN4M</u>

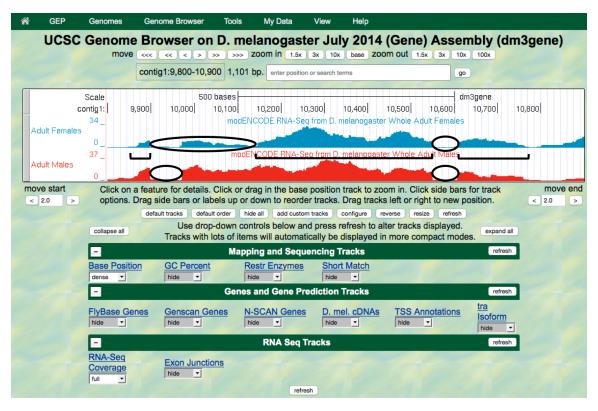


FIGURE 3 IN THIS SCREENSHOT OF THE TRA GENE, INTRONS HAVE BEEN CIRCLED AND EXONS HAVE BRACKETS UNDERNEATH THEM.

Q6. 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').

Q7. 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.

INVESTIGATION 2: IDENTIFYING SPLICE SITES

Learning Objectives:

• Identify splice donor and acceptor sites using the genome browser.

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.

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). 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-Seg and the search splice junctions, RNA-Seq TopHat video: for watch the and https://www.youtube.com/watch?v=qepVXEsfLMM

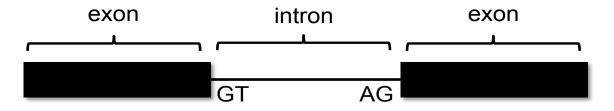
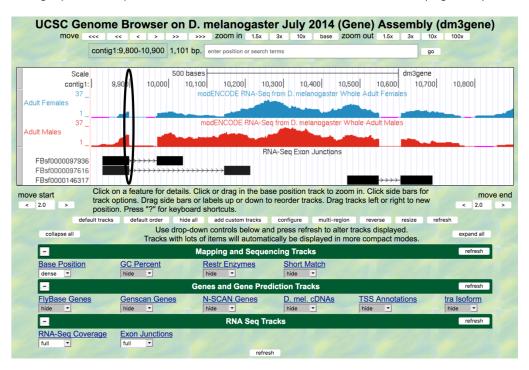


FIGURE 4 A DIAGRAM OF INTRON-EXON JUNCTIONS.

 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. (note: If you stopped after Module 4: Investigation 1, then you may want to return to page 3 for a reminder of how to get to the Genome Browser page.)

- Change the display mode for the "Base Position" track to "dense", and then click "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 "**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 Exon Junctions to "**full**," then click "**refresh**." The 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 5).

FIGURE 5 VIEW OF THE TRA RNA-SEQ DATA WITH THE SPLICE DONOR SITE IN INTRON 1 CIRCLED.

- 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 6).

| move | <<< | << | < > | >> | >>> | zoom | 1. | 5x 3x | 10x | base | e 200 | mout | 1.5x | 3x | 10x | 100x | | |
|--------------------|-----------------|---------------|------------|------------|------------|------------|---------------------------------------|---|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|-------------------------------------|---------------------------------------|---------------|
| | со | ntig1:9,8 | 92-9,90 | 9 18 b | p. con | tig1:9,892 | -9,909 | | | | | | | | go | | | |
| Sca | le | | | | | 5 base | es | | | | | - dm | 3gene | | | | | - |
| contig | 1: 9,89 -> A | 2 9,893 T | 9,894 C | 9,895 G | 9,896 A | 9,897 G | 9,898 G | 9,899 T | 9,900 A | 9,901 C | 9,902 C | 9,903 C | 9,904 A | 9,905 C | 9,906 T | 9,907 A | 9,908 T | , |
| 37 dult Females | - | | | | mod | IENCODI | E RNA: | -Seq fror | n D. m <mark>e</mark> | lanogas | ter Wh | ole Adu | lt Fema | les | | | | |
| 1 37 | - | | | | mo | dENCOE |)F RN4 | A-Sea fra | m D. m | elanoga | ster Wi | nole Ad | ult Male | 15 | | | | _ |
| dult Males | | | | | | | | | | | | | | | | | | |
| 1 | - | | | | | | | RNA-S | Seq Exo | n Junct | ions | | | | | | | |
| FBsf000009793 | | | | | | | \longrightarrow | $\rightarrow \rightarrow \rightarrow$ | $\rightarrow \rightarrow \rightarrow$ | $\rightarrow \rightarrow \rightarrow$ | $\rightarrow \rightarrow \rightarrow$ | $\rightarrow \rightarrow \rightarrow$ | \cdots | $\rightarrow \rightarrow \rightarrow$ | \longrightarrow | \longrightarrow | $\rightarrow \rightarrow \rightarrow$ | \rightarrow |
| FBsf000009761 | 6 🔨 | | | | | - | $\rightarrow \rightarrow \rightarrow$ | $\rightarrow \rightarrow \rightarrow \rightarrow$ | $\rightarrow \rightarrow \rightarrow$ | $\rightarrow\rightarrow\rightarrow$ | $\rightarrow \rightarrow \rightarrow$ | \rightarrow |

FIGURE 6 TOPHAT DATA (IN BLACK) FOR ADULT MALES AND ADULT FEMALES INDICATING AN EXON JUNCTION.

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

Q8. What is the coordinate of the last nucleotide in exon 1?

Q9. 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 7).

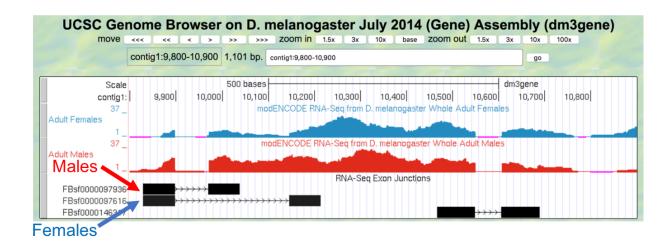


FIGURE 7 TOPHAT AND RNA-SEQ DATA FOR MALES AND FEMALES.

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 8).

| | | 10m | | | | on | | nelanc | - | | - | | (Ge zoom | | | emt _{3x} | 01y 10: | | _ | jen | e) | |
|--------------|------------------------------------|------------|-------------|---------------------------|--------------|------------|------|---------------------------------------|------------|---------|---------------------------|---------|-------------|-------------|-------------|----------------------|------------|---------------|--------|-----|----|----|
| | move | | - | | -10,15 | | - | contig1:10,125 | | 3x | IUX | Dase | 20011 | out | 1.5x | 3x | go | | JX | | | |
| Adult Female | Scale contig1: > 37_ s | c | C G | с | 10,1: A C | 30 : A | с |) bases 10,13 T C T dENCODE | | | 10,14(C A). melar | C, | | 10,1 A (| GA | т | т | 10,15(C C | c c | G | т | G |
| Adult Males | 37_ | | | | | | m | DDENCODE | RNA-Se | eq from | D. mela | inogast | ter Who | le Adu | ılt Male: | 3 | | | | | | |
| FBsf0000 | - 1 097616 | <{\ | · · · · · · | $\rightarrow \rightarrow$ | ····; | · · · · · | ···· | · · · · · · · · · · · · · · · · · · · | R →→→→→ | NA-Seo | q Exon . | Junctio | ns >>>>> | · · · · · | →- ; | | | | | | | >> |

FIGURE 8 GRAPHICAL VIEWER CENTERED ON THE JUNCTION BETWEEN INTRON 1 AND EXON 2 OF THE TRA-RA (FEMALE SPECIFIC) ISOFORM

Q10. What are the last two nucleotides of the female tra-RA intron 1?

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

Q11. What is the coordinate of the first nucleotide in the female tra-RA exon 2?

INVESTIGATION 3: IDENTIFY THE 5' SPLICE DONOR AND 3' SPLICE ACCEPTOR 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

Summary Question 1. What are the last three nucleotides of the female tra-RA exon 2?

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

Summary Question 3. 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 TopHat 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.

Summary Question 4. What are the last two nucleotides of the female tra-RA intron 2?

Summary Question 5. What is the coordinate of the first nucleotide in the female tra-RA exon 3?

Summary Question 6. 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.

Summary Question 7. Where do you think the promoter is located in relation to your gene model? What <u>evidence</u> do you have to support your idea, using the evidence tracks we have displayed (Base Position, RNA-Seq Coverage, Exon Junctions)?

Bonus Question 1. Support your hypothesis by gathering additional data. Recall our explorations in Modules 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?

MODULE 4 HOMEWORK: SPLICING

Meg Laakso

DETERMINING SPLICE SITES FOR THE SPD-2 GENE

 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 <u>http://gander.wustl.edu</u>. Follow the instructions given in Module 1 to navigate to the contig1 project in the *D. melanogaster* 'July 2014 (Gene)' assembly.

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 and is part of the left arm of chromosome 3, which is about 28,100,000 bp long (Figure 9). If your Browser window is showing other evidence tracks, reset by clicking on "default tracks."

| move | <<< << | < > >: | >>>> Z | coom in 1.8 | 5x 3x 1 | 10x base | zoom ou | t 1.5x | 3x 10x | 100x |
|--|--|--|--|--|---|---|--|---|---------------------------------------|--|
| | contig1:1-1 | 1,000 11,00 | 0 bp. enter p | osition or sear | ch terms | | | | go | |
| A 1 | | | 5 kb | | | | | | day Cara a a | |
| Scale | 1.000l | 2.000l | 3.000 | 4.000 | 5.000 | 6.000 | 7.000l | 8,000 | dm3gene 9.000 | 10,000 |
| contig1: | 1,0001 | 2,000 | 3,0001 | 4,000] | | se Genes | 7,0001 | 0,0001 | 9,0001 | 10,0001 |
| CG32165-BC | | | . | | | se denes | | | | tra-RB |
| CG32165-RB | | | | | | | | | | tra-RA → |
| CG32165-RA | | (← | € : | | • • • • • • • | | | | | |
| | | _ | | | | -2-RA | | | | |
| | Click on a fea rack options. position. Pres ult tracks de | Drag side b s "?" for key fault order | ars or labels board short hide all ad | s up or dow cuts. dd custom track | base position n to reorder | re multi- | region re | eft or righ | t to new | m < 2.0 resh |
| | rack options. positio <mark>n. Pres</mark> | Drag side b s "?" for key fault order Use dr | ars or labels board short hide all ac op-down co | s up or dow cuts. dd custom track ntrols belov | base position n to reorder | re multi- | ag tracks l region re alter tracks | eft or righ verse r displaye | t to new resize refr | < 2.0 |
| .0 > defa | rack options. positio <mark>n. Pres</mark> | Drag side b s "?" for key fault order Use dr | ars or labels board short hide all ac op-down co h lots of iter | s up or dow cuts. dd custom track ntrols belov ns will auto | base position n to reorder ks configure v and press | re multi- refresh to displayed | ag tracks l region re alter tracks | eft or righ verse r displaye | t to new resize refr | < 2.0 |
| .0 > defa | rack options. position. Pres ult tracks de | Drag side b s "?" for key fault order Use dr | ars or labels board short hide all ac op-down co h lots of iter Mappin | s up or dow cuts. dd custom track ntrols belov ns will auto | base position in to reorder ks configure v and press matically be quencing Tr | re multi- refresh to displayed | ag tracks l region re alter tracks | eft or righ verse r displaye | t to new resize refr | < 2.0 resh expand all |
| collapse all Base Position | rack options. position. Pres ult tracks de | Drag side b s "?" for key fault order Use dr Tracks wit | ars or labels board short hide all ac op-down co h lots of iter Mappin | s up or dow cuts. dd custom track ntrols below ns will auto ng and Sec | base position in to reorder ks configure v and press matically be quencing Tr | re multi- refresh to displayed racks | ag tracks l region re alter tracks | eft or righ verse r displaye | t to new resize refr | < 2.0 resh expand all |
| collapse all Base Position | rack options. position. Pres ult tracks de <u>GC F</u> | Drag side b s "?" for key fault order Use dr Tracks wit | ars or labels board short hide all ac op-down co h lots of iter Mappin Rest hide | s up or dow cuts. dd custom track ntrols belov ns will auto ng and Sec r Enzymes | base position in to reorder ks configur v and press matically be quencing Ti Sho | recks. Dr remulti- refresh to e displayed racks | ag tracks l region re alter tracks | eft or righ verse r displaye | t to new resize refr | < 2.0 resh expand all |
| collapse all Base Position | rack options. bosition. Pres ult tracks de <u>GC F</u> hide | Drag side b s "?" for key fault order Use dr Tracks wit | ars or labels board short hide all ar op-down co h lots of iter Mappi Rest hide Genes a <u>N-SC</u> | s up or dow cuts. dd custom track ntrols belov ns will auto ng and Sec r Enzymes | base position n to reorder wand press matically be quencing Tr Sho hide Prediction T | recks. Dr refresh to displayed racks rt Match Tracks nel. cDNAs | ag tracks I region re alter tracks in more co | eft or righ verse r displaye | t to new esize refr d. odes. | < 2.0 esh expand all refresh tra Isoform |
| collapse all - dense v | rack options. bosition. Pres ult tracks de <u>GC F</u> hide | Drag side b s "?" for key fault order Use dr Tracks wit | ars or labels board short hide all ar op-down co h lots of iter Mappin Rest hide Genes a | s up or dow cuts. dd custom track ntrols belov ns will auto ng and Sec r Enzymes | base position to reorder ks configured and press matically be quencing Tr Sho hide Prediction 1 | recks. Dr refresh to displayed racks rt Match Tracks nel. cDNAs | ag tracks I region re alter tracks in more co | eft or righ verse r s displaye ompact m <u>S Annota</u> | t to new esize refr d. odes. | < 2.0 esh expand all refresh |
| collapse all collapse all asse Position dense v | rack options. bosition. Pres ult tracks de <u>GC F</u> hide <u>Gens</u> | Drag side b s "?" for key fault order Use dr Tracks wit Percent | ars or labels board short hide all ar op-down co h lots of iter Mappi Rest hide Genes a <u>N-SC</u> | s up or dow cuts. dd custom track ntrols belov ns will auto ng and Sec r Enzymes - and Gene R CAN Genes | base position n to reorder ks configur v and press matically be quencing Tr Sho hide Prediction T | recks. Dr refresh to displayed racks rt Match Tracks nel. cDNAs | ag tracks I region re alter tracks in more co | eft or righ verse r s displaye ompact m <u>S Annota</u> | t to new esize refr d. odes. | < 2.0 esh expand all refresh tra Isoform |

FIGURE 9 A SCREEN SHOT OF '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.
 - 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"
 - Viewing the exon junctions will help you to find the splice donor and splice acceptor sites.
 - 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.

Summary Question 1. How many exons does *spd-2* have?

Summary Question 2. How many introns does *spd-2* have?

HOMEWORK PART 1: IDENTIFYING SPLICE SITES FOR INTRON 1

You remember from class that short sequences are present at the beginning and end of each <u>intron</u> 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 10**).

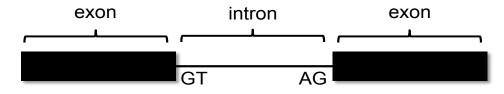


FIGURE 10 A BOWTIE DIAGRAM

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

Summary Question 3. What are the last three nucleotides of exon 1? _____ What is the coordinate of the last nucleotide in exon 1? _____

Summary 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.

Summary Question 5. What are the last two nucleotides of intron 1?

Summary Question 6. What is the coordinate of the first nucleotide in exon 2?

HOMEWORK PART 2: IDENTIFYING THE 5' SPLICE DONOR AND 3' SPLICE ACCEPTOR 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.

Summary Question 7. What are the last three nucleotides of exon 2?

Summary Question 8. What is the coordinate of the last nucleotide in exon 2?

Summary 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.

Summary Question 10. What are the last two nucleotides of intron 2?

Summary Question 11. What is the coordinate of the first nucleotide in exon 3?

Summary 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.