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

Answer Sheet

Q1. What is the span — the start and end base positions — of the tra-RA transcription unit?

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

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

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

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

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

Q7. What is the coordinate for TSS_tra_16584216?

Q8. What is the coordinate for this TSS?

Q9. 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?

Q10. Which base position(s) would you assign as the TSS of the *tra* gene based on the available evidence? Describe your reasoning.

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

Q12. Are there any perfect matches to the Inr consensus sequence (Figure 9)? 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?

Q13. At which base position do you see the RNA-Seq read coverage ending in the whole adult female sample? Zoom in close to the beginning of the pink area (no RNA-Seq coverage) in the RNA-Seq track.

Q14. What is the coordinate of the 3' end of the tra-RA transcript according to the "FlyBase Genes" track? You will need to zoom in on the end of the "FlyBase Genes" track.

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

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

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

Q18. Based on your analysis above, which position is the best choice for the termination signal? Describe your reasoning.

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

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

Q21. If these regions with no RNA-Seq coverage occur within an initial transcript, what could have happened to these RNA sequences?