Annotation Worksheet: Apollo

Due: 06/21/21

These are questions related apollo and annotation. Once it has been completed, I will upload a video lecture explaining all of these topics to my YouTube channel and will provide you with a link. Feel free to type in your answers or print it out and fill it in by hand. Even if you don't have all of this memorized, this can be a good reference sheet to remind you of details should you forget later on down the line. Think of it like a study guide you would fill out for an exam in a class. A few sentences for each answer should be sufficient but this will vary. Use your own discretion but air on the side of being detailed. Feel free to include figures you find online or drawings you make if it is helpful in conveying an idea

Apollo Basics

1.) What is apollo?

2.) What is a gene model?

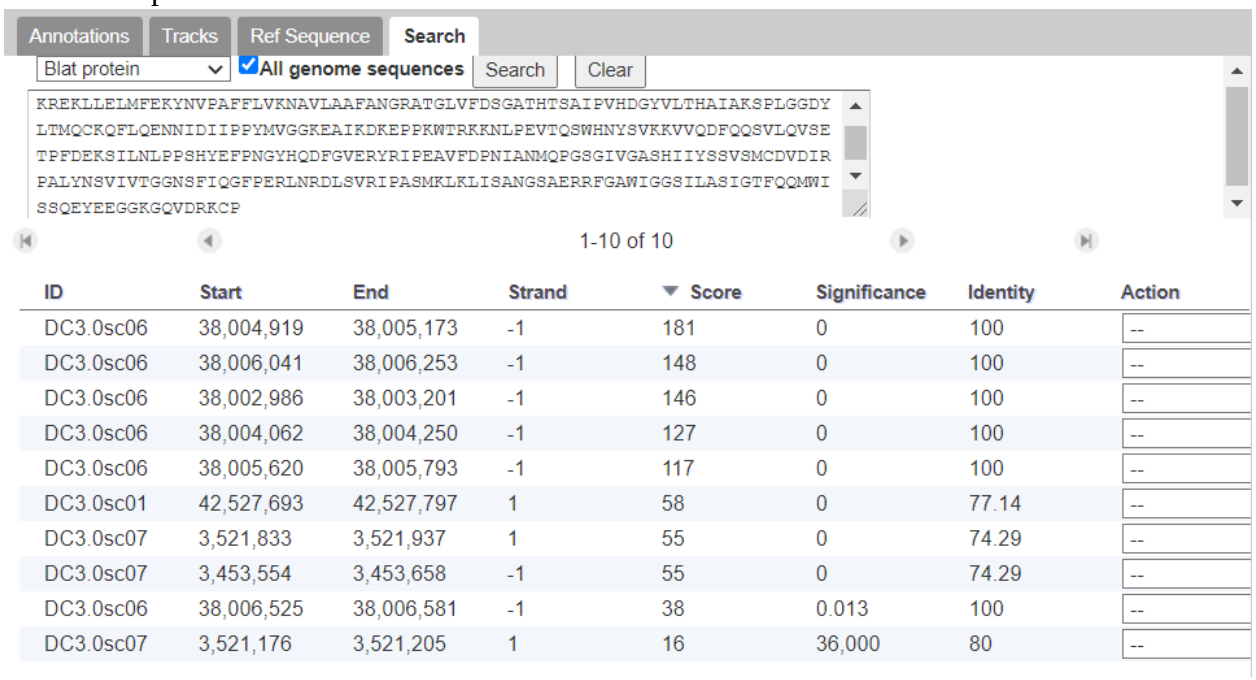
3.) What is a scaffold?

4.) What is the difference between 2.0 and 3.0?

5.) What is the difference between annotation and validation? Why is it necessary to do validation if a gene was previously annotated in an older version of the genome?

6.) What is a blat. How is it different from a BLAST. What sort of information do you get from a blat.

7.) In the following blat, explain what ID, start, score, significance and identity mean and why each are important to consider.

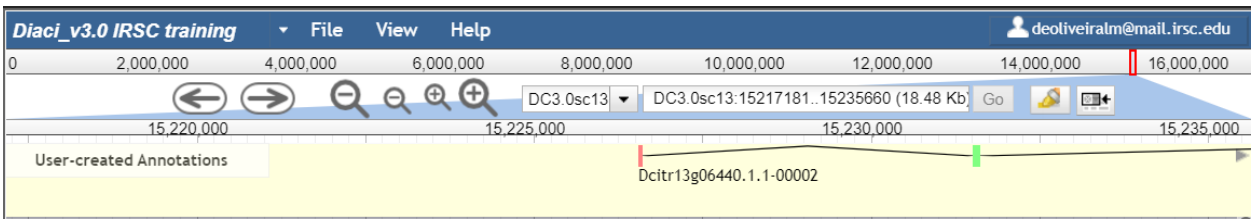


ID	Start	End	Strand	Score	Significance	Identity	Action
DC3.0sc06	38,004,919	38,005,173	-1	181	0	100	--
DC3.0sc06	38,006,041	38,006,253	-1	148	0	100	--
DC3.0sc06	38,002,986	38,003,201	-1	146	0	100	--
DC3.0sc06	38,004,062	38,004,250	-1	127	0	100	--
DC3.0sc06	38,005,620	38,005,793	-1	117	0	100	--
DC3.0sc01	42,527,693	42,527,797	1	58	0	77.14	--
DC3.0sc07	3,521,833	3,521,937	1	55	0	74.29	--
DC3.0sc07	3,453,554	3,453,658	-1	55	0	74.29	--
DC3.0sc06	38,006,525	38,006,581	-1	38	0.013	100	--
DC3.0sc07	3,521,176	3,521,205	1	16	36,000	80	--

8.) What happens if I don't select "all genome sequences" when I blat. What happens if I try blatting a nucleotide if I have inputted a protein sequence?

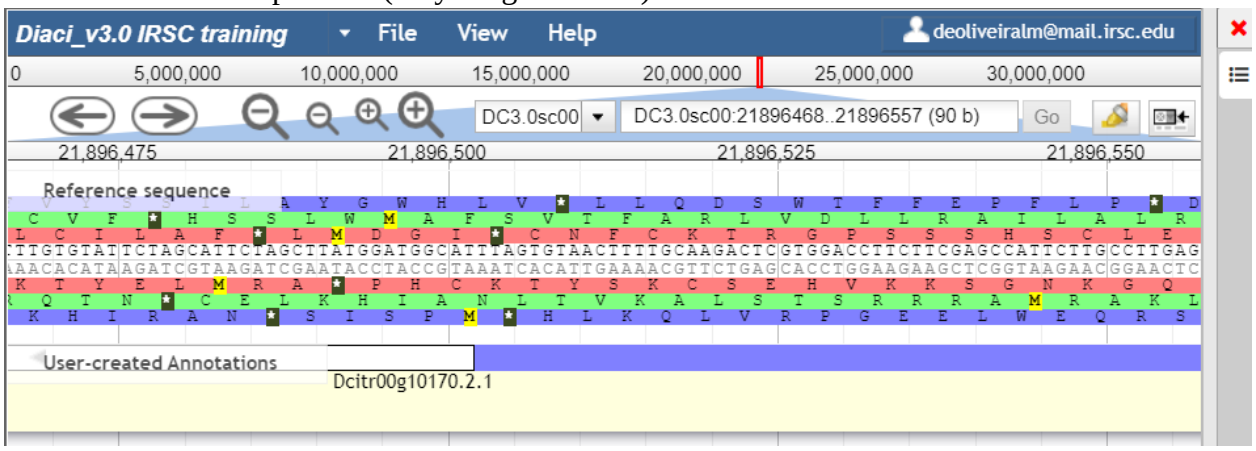
9.) Which is more specific, a peptide blat or a cDNA blat? Give an example of a specific circumstance where you would use a cDNA blat over a peptide blat.

10.) In the following image, circle the location box. Explain what the location tag is and why its useful. Is the location tag specific to the magnification level?

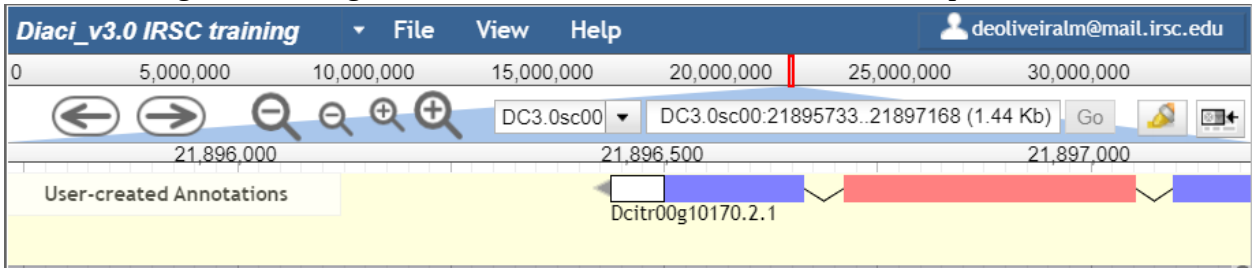


11.) What is an AHRD predicated name? How reliable is it?

12.) Which ORF is being used in this exon? What is the directionality of the gene model? Circle the start or stop codon (only 1 right answer)

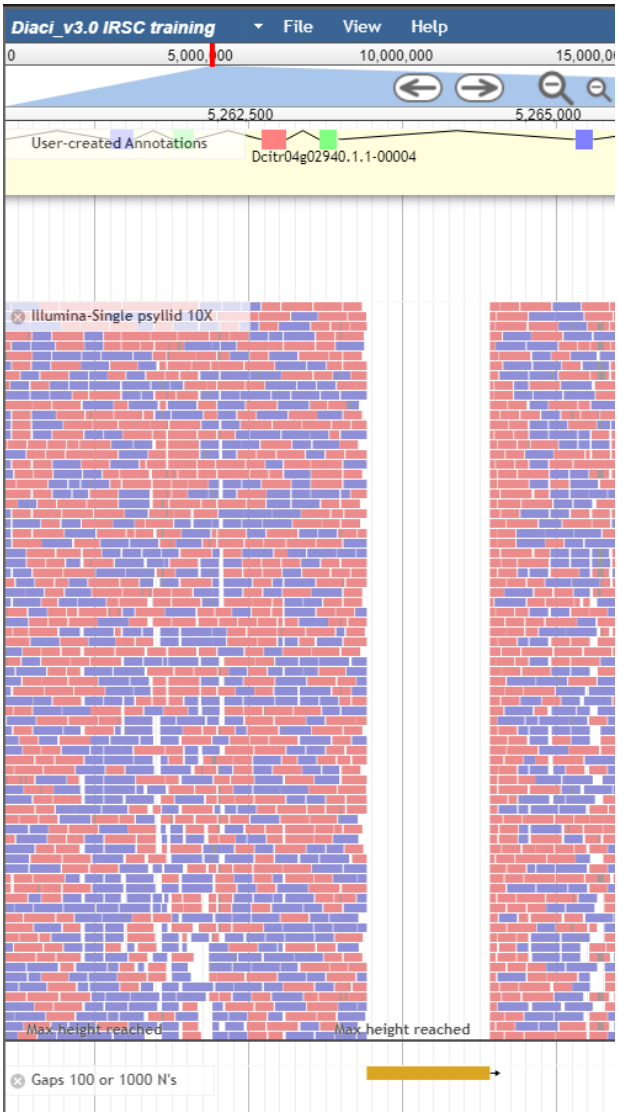


13.) In this fragment of the gene model, label the UTR, Exons and Introns present



14.) What is illumina sequencing? During genome sequencing, is the DNA sequenced as one continuous read? If not, explain how. How might this introduce possible errors?

15.) Explain what occurred here. Do you think it has a major impact on the quality of the gene model as a whole? If not, why?

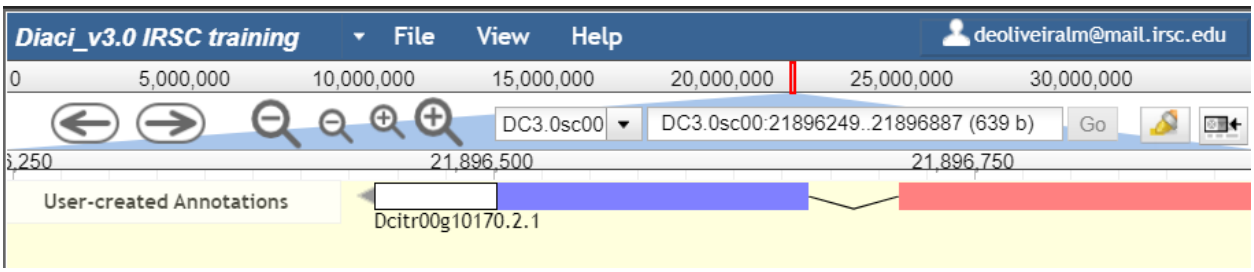


16.) Compare and contrast a duplication and an artifact duplication.

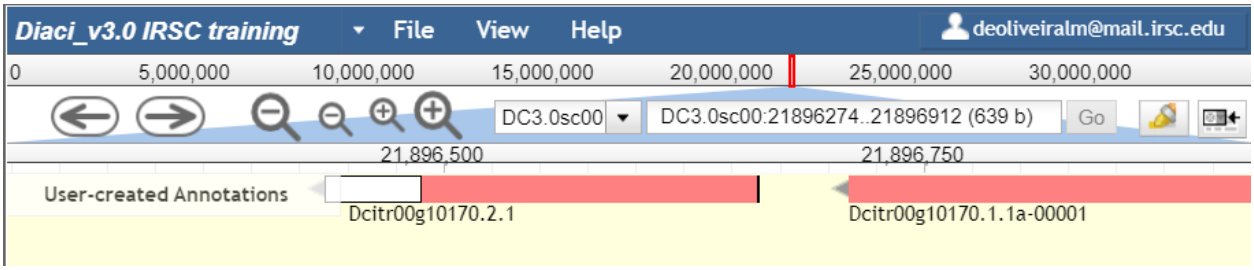
17.) How can you check if a gene model is an artifact duplication? If it ends up being one, what steps should you take?

18.) How would you switch the direction a gene model is reading in the annotator box?

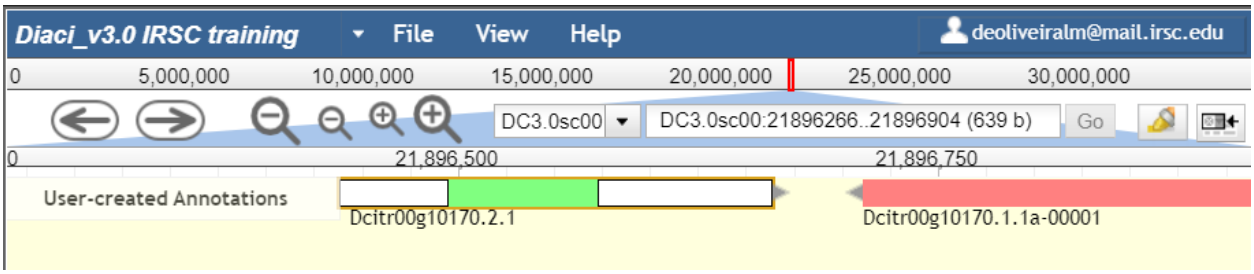
19.) Describe how you would split the following two exons



20.) Explain how you would merge the following 2 exons.



21.) Is it possible to merge to these two exons as they are? If not, what step would you need to take first before attempting to merge them?



22.) What is RNAseq? What is read depth? Does a higher or lower read depth correlate to more confidence in the RNAseq as legitimate evidence?

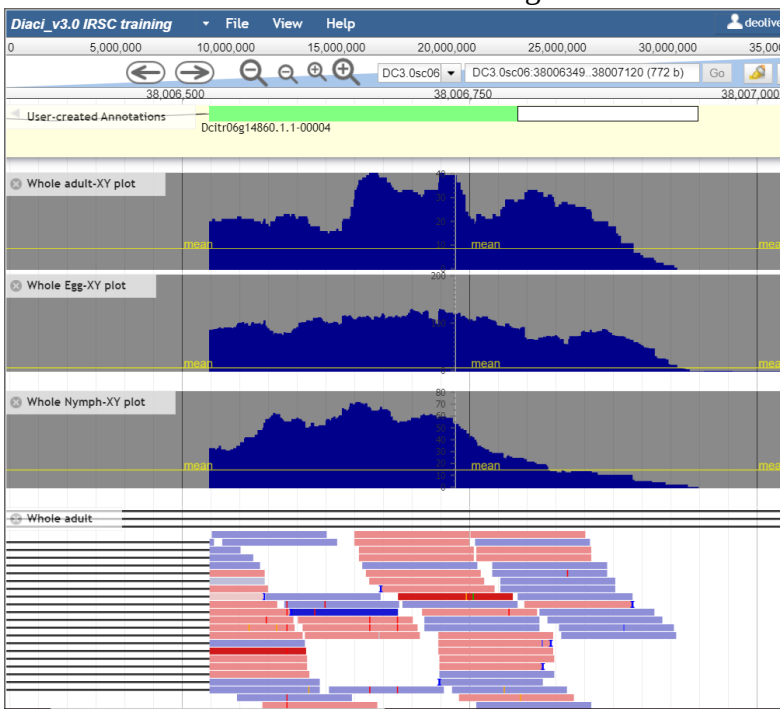
23.) What is the difference between the adult, egg and nymph RNAseq tracks?

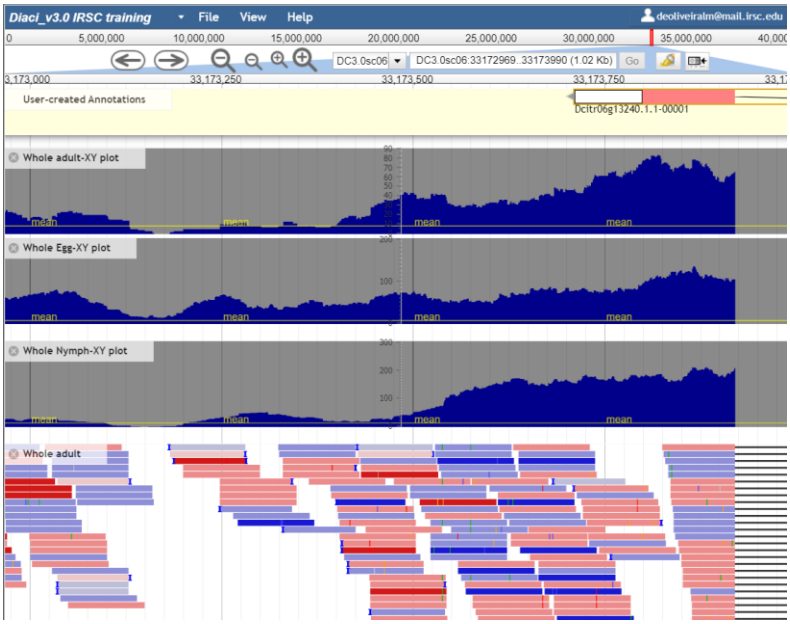
24.) What is the difference (if any) between the RNAseq quantitative XY plots and the mapped reads?

25.) What are the 4 most reliable evidence tracks that have gene models in apollo in decreasing order of reliability? (ties don't matter)

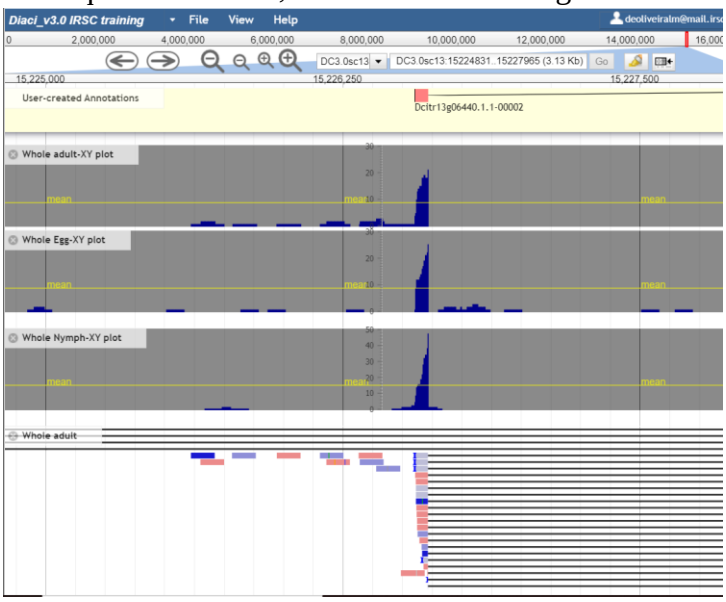
26.) Under what circumstance should we choose a gene model from the other tracks in apollo even if they are “less reliable” than the 4 tracks listed in the previous question?

27.) What is your opinion of the following UTR based on RNAseq data? If its fine explain why and if its not then circle where the UTR should change to.

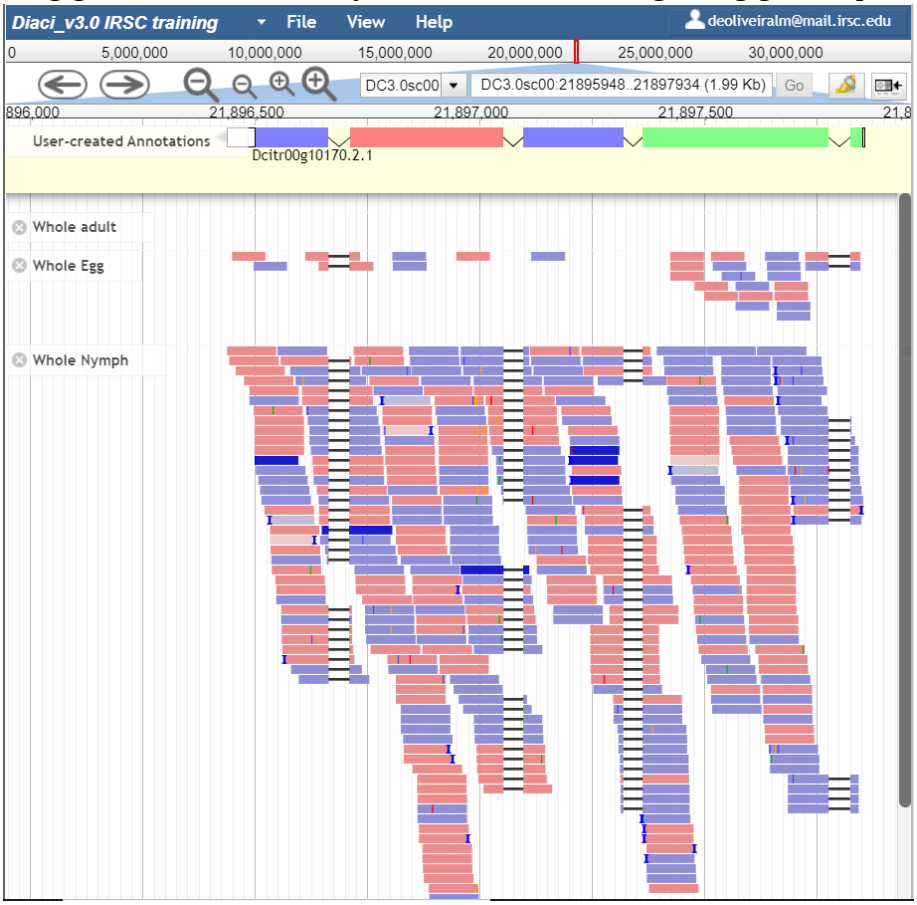




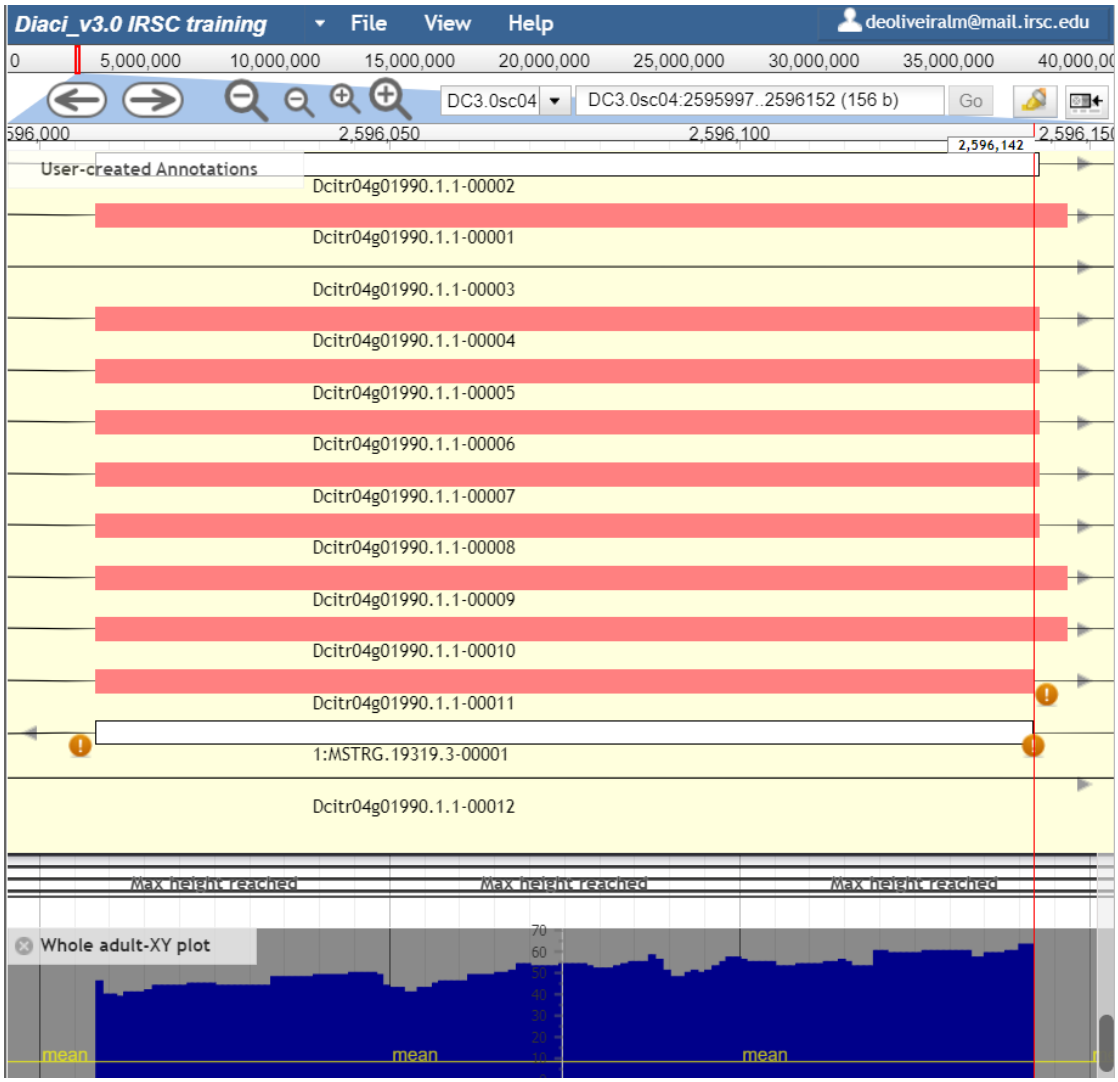
28.) Based on the RNAseq evidence here, does it look like the gene model should end here?



29.) Rate the quality of RNA seq data in descending order of quality between the 3 life stages in the following gene model. What do you think this means regarding gene expression?



30.) Which version of this exon would you choose to include in your final gene model. (Assuming the splice site errors present remain, no new ones are introduced, and the rest of the RNAseq tracks align perfectly to the one shown)



31.) Provide an outline of an example validation workflow. There is room for variance here. Not everyone validates the same. The order may vary but try to include the key elements.