NGS Bioinformatics
Practical assignment

Module topic: genome assembly
Contact session title: Module10_Day2
Trainer: Eugene Gardner
Participant: *<write your name here>*
Date: *<write today's date here>*

Module 10 Genome assembly Day 2

Please note

- **Hand-in information** please upload your completed assignment to the Vula 'Assignments' tab. Take note of the final hand-in date for each assignment, which will be indicated on Vula.

Please ONLY provide answers to the exercise questions in the practical assignment document. The numbering for the questions in each section are provided below:

2. Comparing Reference Genomes

1. Based on your work during the previous assembly module, can you think of a reason why assembly might not be perfect?

2. Is there an obvious issue with our assembly?

3. Why do you think both ends of the reference genome align to the same part of our assembled genome?

4. What do you think the green segments represent in this image?

5. Why is the red line not centered in the plot and moves up or down?

3. Identifying Repetitive DNA:

1. Can you identify repetitive DNA sequences that are longer than 5 basepairs in this sequence?

```
TATAAATACAATATAATATAACGACGAACAGATATGAAAGTGTTAGAACTAGACATACCA
TTTTTCTGTGAAAAATACTTCAAGCTGTAGTATTATTATTATTGCGCTGCTTAGATGTAGT
```

2. Why do the sections “Retroelements” and “DNA transposons” all have zeros?

3. Approximately what proportion of our genome assembly was masked?

4. Finding Genes

1. How many exons does the gene “1_g” have?

2. Can you think of a simple LINUX command to figure out how many genes GeneMark-ES identified?

3. How many genes did GeneMark find?

4. What is the part of the command 2> /dev/null actually doing?

5. How many genes are in our final set of possible genes (bonafide.gb)?

6. What do you think a limitation of using just 1 chromosome to train our gene finder is?

7. Can you figure out how many genes each approach found?

8. We identified protein coding genes, but can you think of any other types of anno- tations we could find with Augustus?

9. How many exons does this gene have?

10. Do you think there are any issues in the gene structure?

11. How do the predictions for your model versus the default model compare?

12. What are both predictions missing, and why do you think that is?

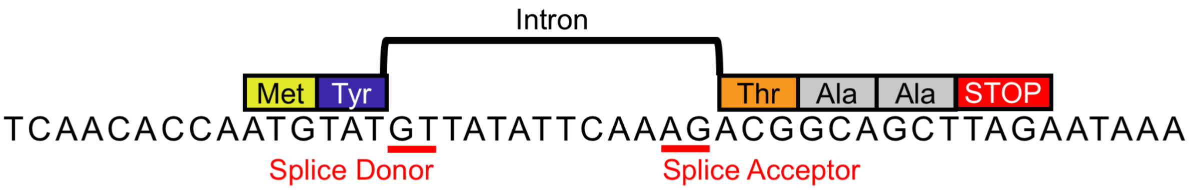
13. How many exons does this gene have?

14. How many introns?

15. How can you extract the same information for another gene by modifying the above command? Report the command and result here.

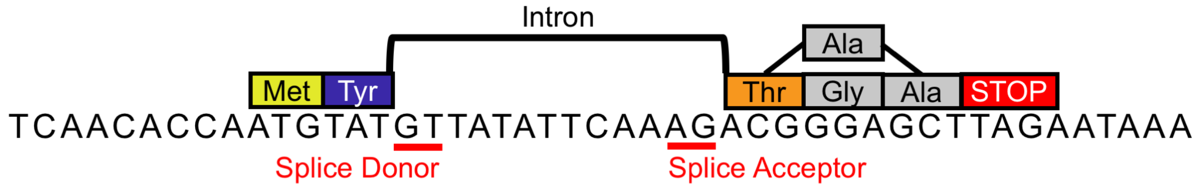
5. Using Comparative Genomes to Identify Genes

1. What is difficult about this alignment?



2. Did you notice something at the end of the alignment that was not in the protein sequence?

3. What was difficult in this example?



4. Do you think this is an issue, or is there something biology-related going on?

5. What do you think the * character represents?

6. What is the gene that we identified in IGV?

7. Can you name a function of this gene and how did you get the answer?