SARS-CoV-2 Sequence analysis Project

You are a bioinformatician working in a public health lab in British Columbia helping to track cases of COVID-19, which is caused by the SARS-CoV-2 virus. Members of your team are working to keep track of where people contract the virus. It is important for them that they have accurate records of who got the virus internationally, from a different province, or from transmission within British Columbia. This is important information needed to inform policy decisions surrounding boarder closures. They are having trouble getting a travel history from one patient and have asked you if you can use phylogenetics to help them figure out where the patient may have contracted the virus (but they are still going to work hard to get the accurate travel history!).

You tell them you'll try, since you already have 9 SARS-CoV-2 sequences that you can compare the mystery sequence to. These sequences include: 3 from British Columbia, 3 from The Maritimes and 3 from Washington State.

In addition to the location where the infection was contracted, you are also interested in seeing been a mutation of interest for many months so you decide to look at all the sequences that you have and see if any of them have this mutation. The nucleotide responsible for this change is at position 23403. If there is an A at that position, it encodes aspartic acid, but if it is a G it encodes glycine.

Before you can start, you feel that it would be a good idea to get a reference sequence. Go to NCBI and search for the SARS-CoV-2 reference sequence.

1. **What is the accession number for this entry (0.5 marks)?**
2. **How long is this sequence (0.5 marks) and how many genes does it have (0.5 marks)?**
3. **What type of genome does this virus have (0.5 marks)?**
4. **The spike protein, which is encoded by the S gene, is a region of interest for many. Use the "Graphics" option next to the "FASTA" button in NCBI to visualize the genome and get a zoomed in region of the S gene. (1 point)**

Now that you have a reference sequence, combine it with the 10 sequences you have already (the 9 from a known location plus your mystery sequence), you should have a total of 11 sequences. You can construct an alignment however you wish, but since this is a time sensitive investigation your co-workers suggest using MAFFT in Galaxy since that seems to run the fastest (and it is also the alignment tool used in NextStrain!)

1. **After constructing an alignment of the sequences, which sequences have the D614G mutation (1 mark)?**
2. **Why are we concerned about mutations in the spike protein (2 marks)?**
3. **Identify two other nucleotide positions that may be of interest when looking for mutations to track (2 marks)**
4. **Include images of the alignment at position 23403 (1 mark) and the positions chosen in Q3 (2 marks).**

Now that you know about this significant mutation, build a tree of all the sequences.

5. **What sequence have you chosen for the root of your tree? (1 mark) Why? (1 mark)**
6. **Based on this tree, where was the mystery sequence most likely picked up? (1 mark)**
7. **Is it surprising that two geographically distinct areas (The Maritimes and British Columbia) have such similar SARS-CoV-2 sequences? (1 mark)**
8. **Are there any sequences that do not cluster with any of the groups? If so, which one? (1 mark).**
9. **Include an image of your phylogenetic tree (1 mark)**

One thing many organizations have done is assign lineages to their SARS-CoV-2 sequences. You can do this too by uploading the fasta file at this website: https://pangolin.cog-uk.io/. This is the Pangolin tool, it has been well adopted by many groups during the pandemic. If you would like more information about the tool you can visit: https://github.com/cov-lineages/pangolin

1. **Once you upload your fasta file, you should be taken to a window with a "Start Analysis" button in the top left-hand corner near the Pangolin logo. Hit this button and you will start to get your lineages. Once this complete, fill out the table with the assigned lineages (1 mark).**

Sequence	Lineage
Reference	
British Columbia 1	
British Columbia 2	
British Columbia 3	
The Maritimes 1	
The Maritimes 2	
The Maritimes 3	
Washington State 1	
Washington State 2	
Washington State 3	
Mystery Sequence	

2. **Do you think using only a lineage assignment would be enough to decide where a SARS-CoV-2 sequence came from? (1 mark) Why or why not? (1 mark)**
3. **Knowing what you know about international transmission, based on what you have seen in this analysis, would you recommend keeping the international boarders closed? (1 mark) Why or why not (1 mark)?**
4. **What would you conclude about this mystery sequence? Significant concern based on mutations?**