Senior Bioinformatics Data Scientists - technical exercise

The exercise comprises 3 parts, and has a total of 6 questions. Please provide the best answer you can to each question.

Part A

Data files

- `Pf_M76611.pf7.200_samples.vcf` . This file, in VCF (Variant Call Format), represents genetic variation of the mitochondrial genome of Plasmodium falciparum across over 200 samples (the full dataset comprises more than 20,000 samples)

Questions

A1. Write a Python script to calculate **allele frequencies** for each variant in the file. Allele frequency for a variant is the *proportion* of samples carrying the *reference* allele and each *alternative* allele. The output of the script should be a table with 4 columns: chromosome , position, allele list (including the reference allele) and proportion list. For example:

```
Pf_M76611  5956  T,G,A      0.8,0.1,0.1
```

Part B

This exercise looks into SARS-CoV-2 samples sequenced by COVID-19 Genomics UK Consortium (COG-UK) during 2021. You will be given data from 200 samples sequenced in the months of January and August 2021 and have a look at their inherent variation.

Data files

- sample_mutations.csv - A Comma-separated Values (CSV) with 3 fields: *sample_name*, *Mutations*, *Month*. The *Mutations* field contains a list of positions in that sample which are mutated with respect to the reference genome. For example, `orf1ab:A1708D` should be interpreted as “in the orf1ab gene, the amino acid at position 1708 is mutated from A to D”. For each sample, all mutations carried by the sample are listed in the *Mutations* field, **i.e. the sample carries the reference allele for all genes+positions not listed**. The *Month* field refers to the month in 2021 that the sample was collected.
- barcode_ref.json - encodes a data structure for creating a mutation profile, or “barcode”, for a sample. It lists 100 key positions in genes at which mutation has been implicated in conferring observable changes to the virus (e.g. its ability to

infect). For example, `"4": ["ORF3a", 26]` should be interpreted as “position 4 of the profile/barcode comprises the 26th amino acid of the gene ORF3a”.

- unknown_sample_mutations.txt - a text file containing a mutation string for a sample with an unknown sampling month.

Questions

For the following questions please return a document containing your methodology and findings, including any figures. Please include your documented code separately, in the form of raw scripts or Jupyter notebooks. Please make your working and reporting as clear and concise as possible.

C1. Using the “sample_mutations.csv” and “barcode_ref.json” determine a profile/barcode for each sample. In this exercise, each barcode element should be “R” (for “reference”) if the sample is carrying the reference allele at that Gene:Position, and “A” (for “alternate”), if the sample is carrying a mutant allele at that Gene:Position.

C2. Determine whether the sample from “unknown_sample_mutations.txt” is more likely to have been sampled in January or August (hint: do the barcodes cluster with samples from either month?).