COMP 5970/6970-004
Computational Biology: Genomics and Transcriptomics
Project 2

Haynes Heaton

Spring, 2022

You will find the reference genome for the original strain of SARS-Cov2 (which is the virus which causes the disease COVID-19) in the files on canvas named SARS-COV2.fasta. You will also find sequencing reads from a recent real sample in the file SRR18775434.fastq. For this project you are going to need to install some software. This might be tricky on windows so maybe find a google collab or a linux lab for this project if you are not on mac or linux.

- samtools - http://www.htslib.org/download/
- minimap2 - can be installed via anaconda with `conda install -c bioconda minimap2` or cloned and compiled from https://github.com/lh3/minimap2
- freebayes - can be installed via homebrew with `brew tap brewsci/bio && brew install freebayes` or bioconda with `conda install -c bioconda freebayes` or cloned and compiled from https://github.com/freebayes/freebayes

And the project will have a few steps...

1. map fastq reads to the reference genome using minimap2
2. convert sam output to bam using samtools
3. sort bam file with samtools
4. index sorted bam file with samtools
5. call variants on bam file with freebayes
6. find the spike protein region from the genome annotation file GCF_009858895.2_ASM985889v3_genomic.gff in the files section on canvas (note, the gff format is 1 indexed)
7. visualize the reads in IGV in the region of the spike protein https://software.broadinstitute.org/software/igv/ and take a screen shot
8. find the amino acid sequence of the spike protein (feel free to use biopython or an online tool)
9. find the sequence of the spike protein in the sample using the variant calls (note, the vcf format is 1 indexed), pyvcf may be useful here, or you can parse the vcf yourself. (`conda install -c conda-forge pyvcf`).

10. for each of these variants that alter the spike protein gene, find the amino acid change if any and identify if they are synonomous, non synonymous, or nonesense. For any non-synonymous mutations, identify if there is a major amino acid difference (acid/base, polar/non-polar). You will have to make sure to make the edit (remove ref allele, add alt allele) and report the new in frame amino acid(s) at that location.

For the functional element to this, you might want to use the following snippet.

```
functional = {"D":"acidic", "E": "acidic", "G": "polar",
"F":"nonpolar", "L": "nonpolar", "S": "polar", "Y": "polar",
        "C": "polar", "W":"nonpolar", "L":"nonpolar", "P": "nonpolar",
        "H": "basic", "Q":"polar", "R": "basic", "I":"nonpolar",
        "M": "nonpolar", "T": "polar", "N": "polar", "K":"basic",
        "S":"polar", "R":"basic", "V":"nonpolar","A":"nonpolar"}
```

Deliverables

- vcf file
- IGV screenshot
- variant changes in spike protein and their properties as discussed above
- amino acid changes in spike protein
- amino acid sequence of reference and sample spike protein
- a short analysis of each variant's effect on the amino acid sequence (synonymous or polar -> nonpolar or for indels it might be something like ['polar','polar'] -> ['polar'])

I wanted to include a structural analysis with alphafold, but unfortunately the free google collab GPUs don't have enough memory to fold a protein as long as the spike protein.