Sequence Analysis and Omics – Coursework Assignment 1

PART A

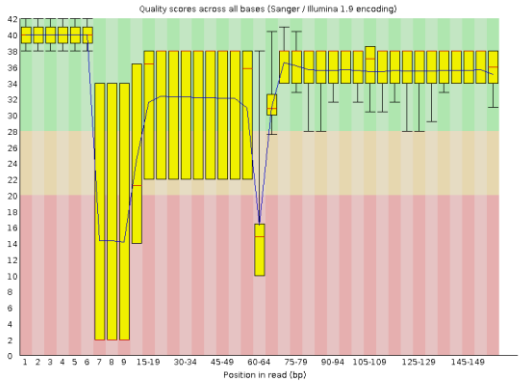
All outputs and files are saved in the following directory: /d/projects/u/as004/omics_coursework_1. Commands are run assuming that this is the starting directory.

Question a)

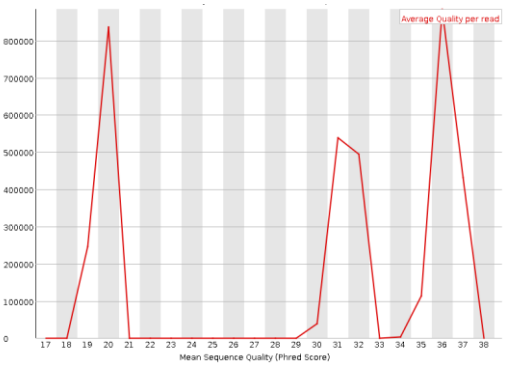
```
cd /d/projects/u/as004/omics_coursework_1
mkdir data
mkdir outputs
mkdir outputs/fastqc
cp /d/in4/u/ubcg71a/teaching/sao/data/final_merge_synthetic_reads.fq ./data

module load fastqc/v0.11.9
fastqc ./data/final_merge_synthetic_reads.fq  -o ./outputs/fastqc
```

Sequence quality is poor towards the start of the sequence (figure 1), and is overall poor for around half of samples, as the mean Phred score splits into two distinct peaks (figure 2). Per base quality is especially poor towards the beginning of the sequences, however improves greatly past base 7 with parallel lines for each base around the 25% composition mark (figure 3). There is also a very high number of low confidence calls (figure 4), also indicating biased sequence composition and general low quality near the start of the sequences. Perhaps trimming the start of the sequences will improve overall quality, as an adapter sequence may still be included.

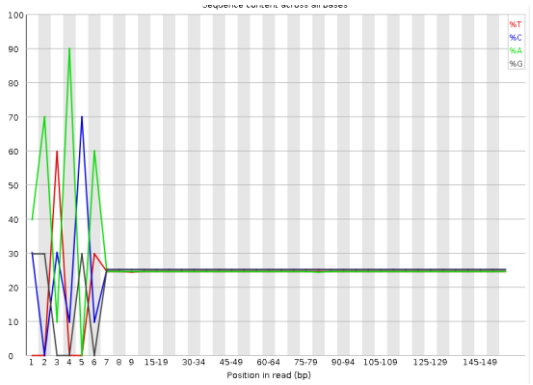
1



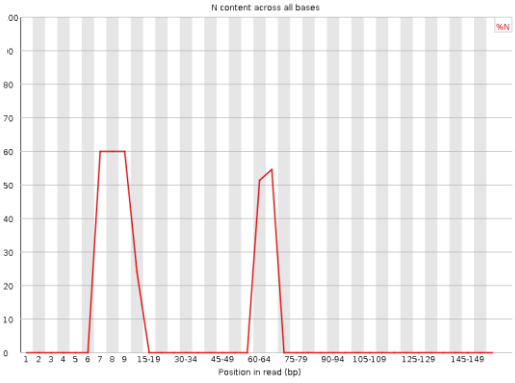
2



3



4



Question b)

```
mkdir ./outputs/trimmed
module load python/v3

cutadapt \
```

```
      -g positive=GATACA \
      -g negative=AGTAGT \
      -g bq=CACACA \
      -g long=AAACCC \
      -o "./outputs/trimmed/trimmed-{name}.fastq.gz" \
      ./data/final_merge_synthetic_reads.fq
```

Question c)

```
mkdir ./outputs/mapped
mkdir ./data/genome
module load bowtie/v2-2.4.2

cp /d/in4/u/ubcg71a/teaching/sao/genomes/AFPN02.1/AFPN02.1_merge.fasta ./data/genome
bowtie2-build ./data/genome/AFPN02.1_merge.fasta ./data/genome/AFPN02.1

bowtie2 --end-to-end \
        -x  ./data/genome/AFPN02.1 \
        -q ./outputs/trimmed/trimmed-negative.fastq.gz \
        -S ./outputs/mapped/neg_AFPN02.1.sam \
        >& ./outputs/mapped/neg_AFPN02.1_end_to_end_stats.txt
bowtie2 --end-to-end \
        -x  ./data/genome/AFPN02.1 \
        -q ./outputs/trimmed/trimmed-positive.fastq.gz \
        -S ./outputs/mapped/pos_AFPN02.1.sam \
        >& ./outputs/mapped/pos_AFPN02.1_end_to_end_stats.txt

cat ./outputs/mapped/neg_AFPN02.1_end_to_end_stats.txt
1077923 reads; of these:
  1077923 (100.00%) were unpaired; of these:
    1072073 (99.46%) aligned 0 times
    5144 (0.48%) aligned exactly 1 time
    706 (0.07%) aligned >1 times
0.54% overall alignment rate
```

Negative file aligns very poorly with the genome, very few reads align (only 0.54% align). Positive file aligns very well in comparison (99.06% overall alignment rate):

```
cat ./outputs/mapped/pos_AFPN02.1_end_to_end_stats.txt
1100714 reads; of these:
  1100714 (100.00%) were unpaired; of these:
    10379 (0.94%) aligned 0 times
    1032839 (93.83%) aligned exactly 1 time
    57496 (5.22%) aligned >1 times
99.06% overall alignment rate
```

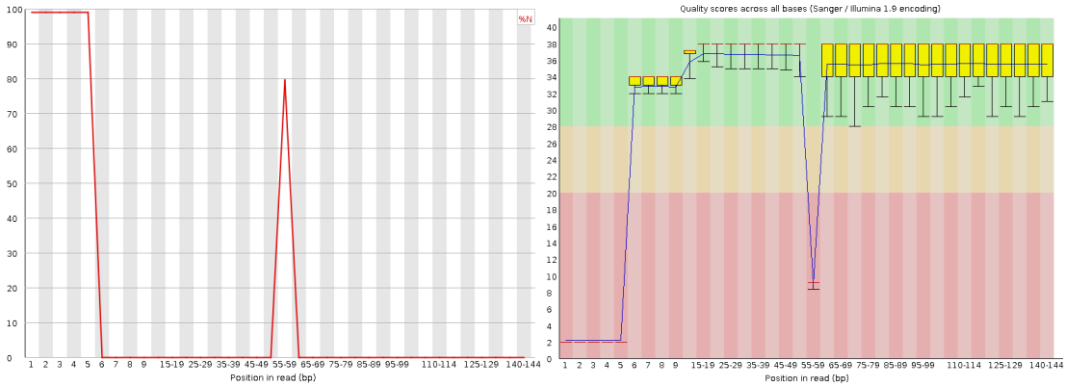
Question d)

```
mkdir ./outputs/negative_remapping
module load samtools

samtools sort ./outputs/mapped/neg_AFPN02.1.sam >
./outputs/negative_remapping/neg_map1.bam
samtools index ./outputs/negative_remapping/neg_map1.bam
samtools stats ./outputs/negative_remapping/neg_map1.bam >
./outputs/negative_remapping/neg_map1_stats.txt
samtools flagstat ./outputs/negative_remapping/neg_map1.bam >
./outputs/negative_remapping/neg_map1_flagstat.txt

cat ./outputs/negative_remapping/neg_map1_flagstat.txt
cat ./outputs/negative_remapping/neg_map1_stats.txt
```

From the samtools stats file, the average sequence length is 58 (max of 14 most common length is 59). The average quality is 31.1 which is low, but not too low to be useable data.

```
fastqc outputs/mapped/neg_AFPN02.1.sam -o ./outputs/negative_remapping
```

From the FASTQC file, base 59 is very low quality (Phred score of 8) and a lot of unknown/N bases are between 55-59. Bases 1-5 are also very poor quality. There also appears to be a slight GC content bias, however this is minor. It seems the issue is with very poor calling of base 59, so as most sequences are this length or shorter (from the samtools stat file), trimming bases longer than 58 might also help mapping. This base might be poor quality as this is the last base of most of the sequences in the file, and sequence quality tends to deteriorate towards the end of the sequence.



Separately, Bowtie2 also has a built-in quality filter, QSEQ filter (--qc-filter), which can be used to remove any remaining low quality reads.

<u>Re-mapping 1 (trimming sequences longer than 58 bases long):</u>

```
bowtie2 --end-to-end \
        -x  ./data/genome/AFPN02.1 \
        -q ./outputs/trimmed/trimmed-negative.fastq.gz \
        --qc-filter \
        --trim-to 58 \
        -S ./outputs/negative_remapping/neg_map2.sam \
        >& ./outputs/negative_remapping/neg_map2_end_to_end_stats.txt

cat ./outputs/negative_remapping/neg_map2_end_to_end_stats.txt
1077923 reads; of these:
  1077923 (100.00%) were unpaired; of these:
    8526 (0.79%) aligned 0 times
    1017600 (94.40%) aligned exactly 1 time
    51797 (4.81%) aligned >1 times
99.21% overall alignment rate
```

Alignment is much improved by trimming the long sequences, with a low number of reads that still do not align with the genome. Comparatively, the trimming removed a small number of sequences (22,791 sequences, 2% of the total), so some information may be lost however the dramatic increase in mapping counters this small loss of information.

Alternatively, the alignment could be improved by adjusting parameters for the alignment itself, in case the genome build is not an exact match for these reads:

1. Could lower penalty for mismatches (--mp 3 instead of –mp 6),
2. Could allow mismatches in the seed alignment for additional flexibility (so -N 1 instead of -N 0)

<u>Re-mapping 2 (altering the alignment parameters and trimming sequences):</u>

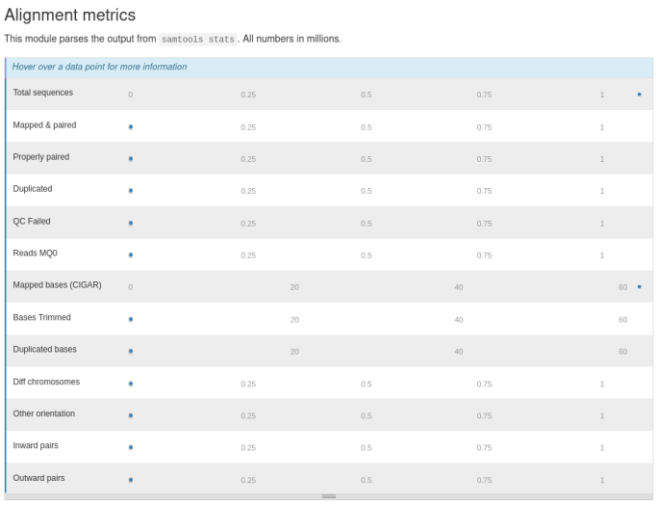
```
bowtie2 --end-to-end \
        -x  ./data/genome/AFPN02.1 \
        -q ./outputs/trimmed/trimmed-negative.fastq.gz \
        --qc-filter \
        --trim-to 58 \
        --mp 3 \
        -N 1 \
        -S ./outputs/negative_remapping/neg_map3.sam \
        >& ./outputs/negative_remapping/neg_map3_end_to_end_stats.txt

cat ./outputs/negative_remapping/neg_map3_end_to_end_stats.txt
1077923 reads; of these:
  1077923 (100.00%) were unpaired; of these:
    2949 (0.27%) aligned 0 times
    1005052 (93.24%) aligned exactly 1 time
    69922 (6.49%) aligned >1 times
99.73% overall alignment rate
```

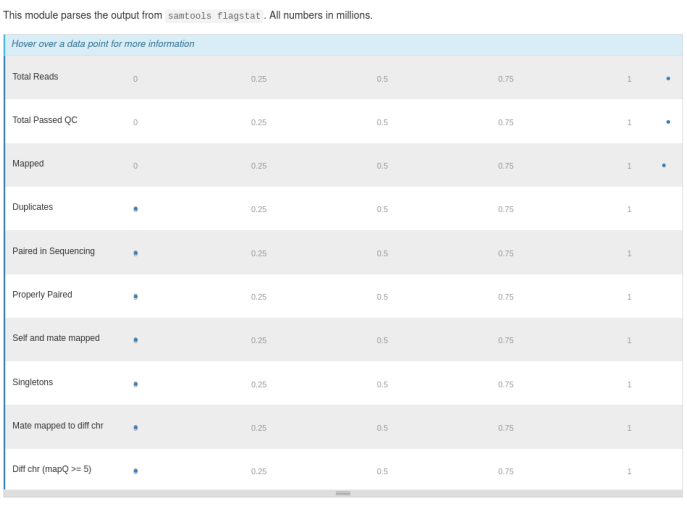
This slightly improves the overall alignment rate when used with the trim-to flag, and reduces the number of reads which do not align at all, but it also increases the number of reads which align to multiple locations, so arguably producing a lower quality mapping. As the increase in overall alignment is small (0.52% increase), I would not use the additional alignment parameters due to the increase in reads aligning to multiple positions, and would keep the stricter alignment rules supplied as default.

Therefore, the final chosen alignment for the negative file is re-mapping 1. Statistics for this mapping are provided below.

```
mkdir ./outputs/negative_remapping_final

samtools sort ./outputs/negative_remapping/neg_map2.sam >
./outputs/negative_remapping_final/neg_map2.bam
samtools stats ./outputs/negative_remapping_final/neg_map2.bam >
./outputs/negative_remapping_final/neg_map2_stats.txt
samtools flagstat ./outputs/negative_remapping_final/neg_map2.bam >
./outputs/negative_remapping_final/neg_map2_flagstat.txt

multiqc ./outputs/negative_remapping_final
```

Sample Name	M Reads Mapped	Error rate	M Non-Primary	M Reads Mapped	% Mapped	M Total seqs
neg_map2_flagstat	1.1					
neg_map2_stats		13.84%	0.0	1.1	99.2%	1.1





PART B

BV-BRC files can be found here https://www.bv-brc.org/workspace/sallen10@bvbrc - I have shared the home directory for user sallen10 (my account) with the user 'irilenia', which will hopefully allow viewing access!

Question 1

a) FastQC report questions
   a. Sequence/read length is 76, all reads are the same length.
   b. Read data is of very high quality; per base sequence quality shows all base positions have a quality above 34 with a slight dip at the start of the sequence.
   c. Other modules in the FastQC report show adapter sequences are retained in the reads, with increasing adapter content towards the end of the read and overrepresentation of TruSeq Adapter, Index 10. There is also sequence duplication. All of these point towards lack of diversity and bias in the sequence library, likely due to adapter dimers being present (which are a contaminant), hence the bias in the per base sequence content.
   d. 38%

b) Genome assembly report questions
   a. Lineage: BF.7, variant: Omicron
   b. Gene 'S' has the most non-synonymous unique SNPs (26 SNPs)
   c. nsp6 (which starts at 10969 and ends at 11830, so the insertion at 11287 falls within this feature)
   d. EMBOSS Stretcher Alignment:

i.

```
#=======================================
#
# Aligned_sequences: 2
# 1: Consensus_assembly.ivar_threshold_0.6_quality_20
# 2: NC_045512.2
# Matrix: EBLOSUM62
# Gap_penalty: 12
# Extend_penalty: 2
#
# Length: 29903
# Identity:   28565/29903 (95.5%)
# Similarity: 28565/29903 (95.5%)
# Gaps:          99/29903 ( 0.3%)
# Score: 159064
#
#
#=======================================
```

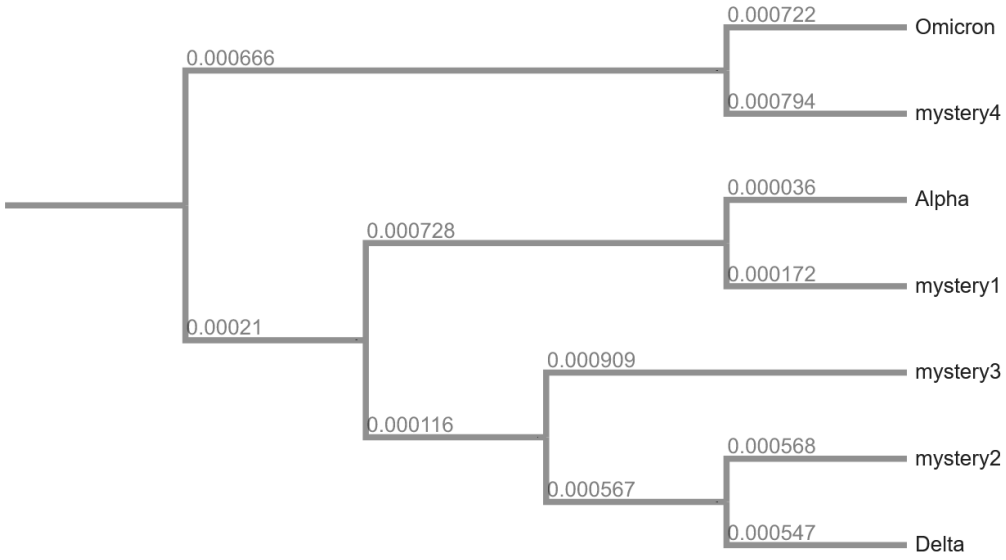
ii.

```
Consensus_ass  11247 TATTATGACATGGTTGGATATGGTTGATACTAGTTTG---------AAGC  11287
                     |||||||||||||||||||||||||||||||||||||||||         ||||
NC_045512.2    11251 TATTATGACATGGTTGGATATGGTTGATACTAGTTTGTCTGGTTTTAAGC  11300
```

The genome assembly from BV-BRC is labelled as 'Consensus_ass' in the EMBOSS alignment

Question 2

a) Cladogram display version, with the numbers on each branch representing the branch length/evolutionary distance:



b) Mystery genome 3 is closest to the Alpha genome, as per the pairwise distances there is less evolutionary time between the Alpha genome and the mystery 3 genome:
   a. Distance to Omicron = 0.000722+0.000666+0.00021+0.000116+0.000909 = 0.002623
   b. Distance to Delta = 0.000547+0.000567+0.000909 = 0.002023
   c. Distance to Alpha = 0.000036+0.000728+0.000116+0.000909 = **0.001789**

Question 3

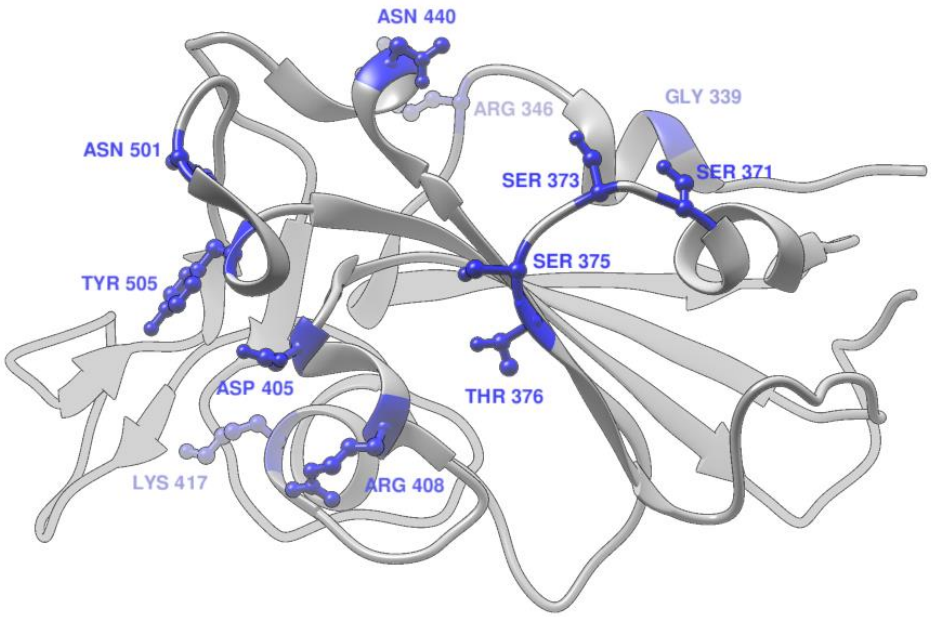


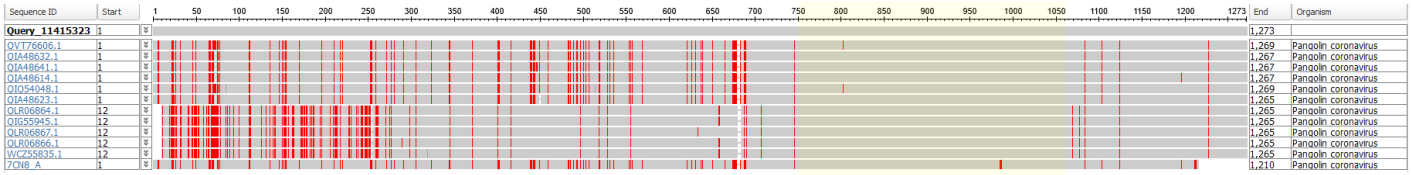
Figure 1: Partial structure (Thr333 – Pro527) of the novel coronavirus spike receptor-binding domain, with SNPs identified in COVID-19 genome SRR23290383 (https://www.ebi.ac.uk/ena/browser/view/SRR23290383) highlighted in blue. SNP positions highlighted: G339D, R346T, S371F, S373P, S375F, T376A, D405N, R408S, K417N, N440K, N501Y, Y505H. Structure derived from PDB entry 6lzg (https://doi.org/10.2210/pdb6LZG/pdb).

Question 4

a) The *Homo sapiens* hits are to PDB 7Y71_A and 7TLZ_J. These are structures which have the spike protein and also the human antibody fragments bound to them, hence the alignment to these entries in BLAST.

b) Pangolin sequences

   a. Region is from approximately 750 to 1050 (highlighted in yellow below) – very few differences in the sequences and in the alignment to the pangolin sequences in this region.



   b. Conserved sequence shown in red in PDB structure 6zb5:

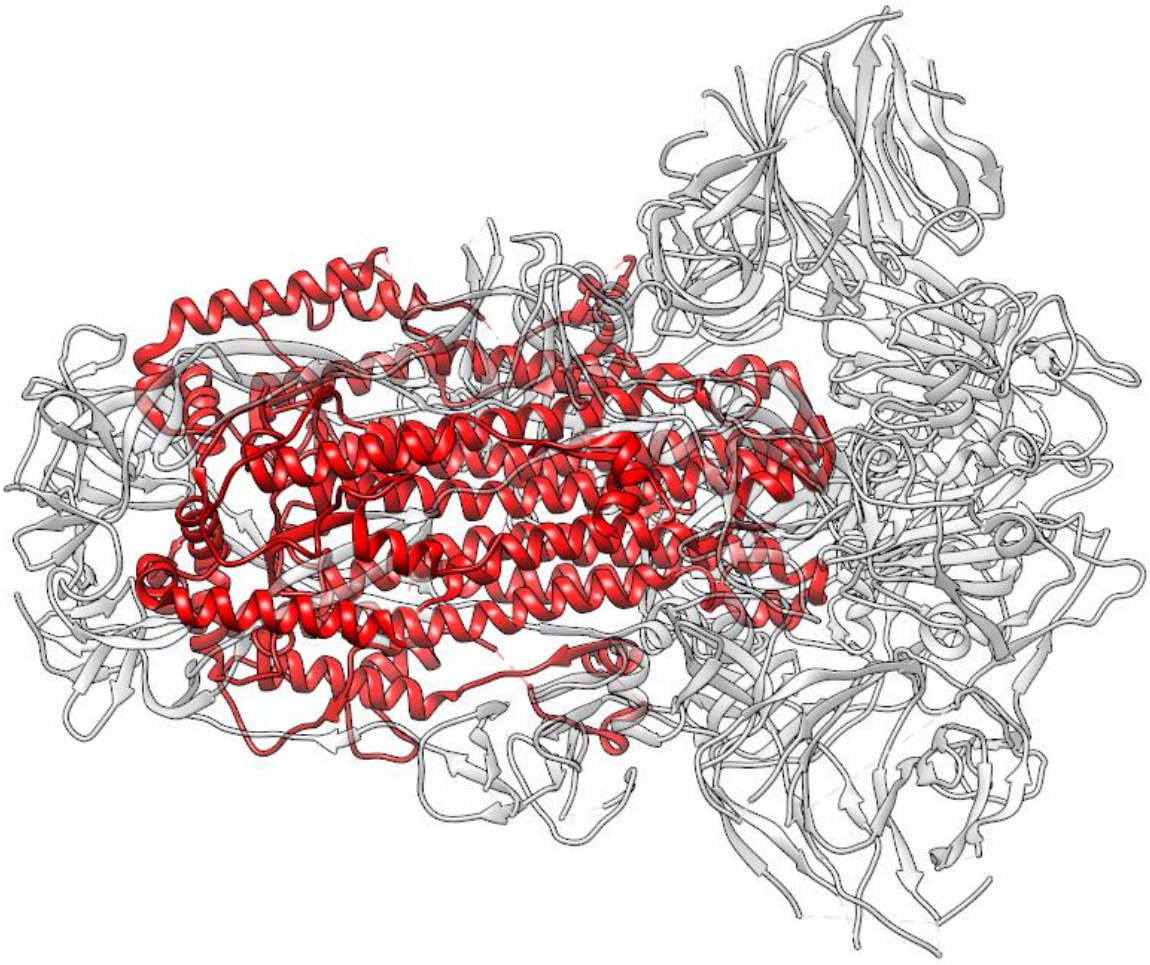


Figure 2: PDB structure 6zb5 (https://doi.org/10.2210/pdb6ZB5/pdb) with conserved region 750-1050 highlighted in red for all three chains. Conserved region was demonstrated when the spike protein from reference genome NC_045512.2 was compared to the 12 pangolin structures identified through NCBI Blast above.