Analysis of SARS-CoV-2 Sequences After Infection in Human and Non-Human Primate Tissues

Daniela Dueñas

12 December, 2024

Background and Overview

This is a report on whether SARS-CoV-2 sequences vary depending on the organism it infects. Based on knowledge from the COVID-19 pandemic, scientists understand that SARS-CoV-2 mutates at a rapid rate, much like many other virus. According to an article published in the World Health Organization Bulletin, a total of 5775 SARS-CoV-2 variants were discovered from infected patients between February and May of 2020 (Koyama *et al.*, 2020). This ability to mutate comes down to the virus' genome. SARS-CoV-2 contains genes that encodes for several key proteins. *ORF1ab* is an open-reading frame gene, that produces poly-proteins. These poly-proteins encode for non-structural proteins that are involved with various processes, such as the virus' replication cycle, transcription events, and diminishing the host's immune protection . *ORF8* is another open-reading frame gene that impedes the IFN pathway (which promotes immune responses), and blocks presentation of class 1 MHC on the infected cell for cytotoxic T cell targeting and destruction (Vinjamuri *et al.*, 2022). Structural genes, such as the *S* gene, *M* gene and *N* gene code for the SARS-CoV-2 spike protein, membrane protein, and capsid protein layer (respectively) (Vinjamuri *et al.*, 2022). Previous research has indicated that mutations at these genes may be involved in SARS-CoV-2's increased virulence and continuous presence throughout the world, despite the continuous production of vaccines.

In 2021, a research team at UC Irvine wanted to understand the human body's response to SARS-CoV-2 after an infection. As part of their research, they obtained several human cancer cell lines and either infected those cells with SARS-CoV-2 or a mock infection. The cancer cells allowed the virus an opportunity to replicate within the cells over four days, in which the viral sequences were eventually isolated and analyzed using an Illumina sequencer (Geerling *et al.*, 2022). The scientists were also curious if the type of organism that the virus infects plays a role in the synthesized viral sequence, so they also performed their experiment on *Chlorocebus aethiops* or Grivet monkey cells (Committed to leaving a legacy of hope and tools to build a better tomorrow for all the earth's citizens_2024, 2024). From this analysis, the researchers were able to identify single nucleotide polymorphisms (SNPs) or variations in the viral genetic sequence (Edwards *et al.*, 2007). This data was compiled and organized into a BioProject dataset labelled PRJNA745219 in the NCBI database, which was the data used to develop this report (Transcriptional analysis of SARS-CoV-2 infected human and nonhuman cell lines). To assess the data, it was first downloaded with the associated metadata and then run through a pipeline on a server to obtain the data as a fastq file. `FastQC` was used to run through the file and perform a quality check. The `Trimmomatic` tool was utilized in a shell script to trim the data and remove any possible outliers that would have affected the overall data (Bolger *et al.*, 2014). The `BWA` package aligned the sequence data with reference genomes to identify any SNPs and calculate a quality score of each. The quality score indicated how confident the program was about those SNPs. `BCFtools` converted the resulting information into VCF files for each identified SNP.

The purpose of this data analysis was to compare the various SNPs on the SARS-CoV-2 genome after cellular infection and to understand if the type of organism that SARS-CoV-2 infects has any affect on the resulting SNPs. I found that the majority of the SNPs analyzed affected the ORF1ab gene, regardless of whether the

sequences were synthesized from *H. sapiens* or *C. aethiops*. Due to the importance of this gene in SARS-CoV-2 replication, it suggests that the mutations on ORF1ab assist the virus evade an organism's immune system and is possibly involved with the virus' constant evolution.

Methods

Data Collection

The dataset used in this report (PRJNA745219) was obtained from a list of BioProjects within the National Library of Medicine from the National Center for Biotechnology Information (NCBI) (Transcriptional analysis of SARS-CoV-2 infected human and nonhuman cell lines). This data along with its metadata were downloaded from the SRA Run Selector webpage .

RStudio Packages and Analysis

On RStudio, `vcfR` was used to read and manipulate the VCF files, along with packages `ggplot2` and `dplyr` to develop the figures and tables (Knaus, 2023) (Wickham, 2016) (Wickham *et al.*, 2023). Additional packages called `ggthemes` and `RColorBrewer` were installed to assist with the visualization of the figures (Arnold, 2024) (Neuwirth, 2022). Packages `citr` was used to cite the sources referenced in this report and `knitr` was used to "knit" the RMarkdown report into a PDF file (Aust, 2019) (Xie, 2024a). The TinyTeX package was downloaded to develop LaTeX documents and assist in producing the "knitted" report as a PDF file (Xie, 2024b).

Results and Discussion

To begin the analysis on the data, all VCF files from the pipeline were stacked and combined to create a complete data frame of 116 total SNPs including the associated metadata. With this data frame, the first step of the analysis was to determine the number of SNP samples obtained from *Homo sapien* or *Chlorocebus aethiop* cell cultures. By filtering the data to only include SNPs located on SARS-CoV-2 gene regions, it was determined that 34 SNPs were identified from *H. sapiens* cells, and 8 SNPs from *C. aethiops* (Figure 1). This suggested that SNPs occur most likely in *H. sapiens* cells compared to *C. aethiops*. Although both organisms share about 90% of their genome, the amount of discrepancy on the number of identified SNPs indicated that the non-sharing 10% of the organism's genome may be the determinant (Committed to leaving a legacy of hope and tools to build a better tomorrow for all the earth's citizens_2024, 2024). Next, I wanted to learn more about the SNPs such as their quality score and position on the SARS-CoV-2 genome. Upon inspection, it was revealed that the SNPs quality scores ranged from close to 0 and almost up-to 250 (Figure 2). SARS-CoV-2 SNPs analyzed from *H. sapien* cells had a wide range, about 0 - 230, while the range from *C. aethiop* cells was about 125 - 240 (Figure 2). Although it appears that higher quality SNPs come from *C. aethiop* cells, it's important to remember that *C. aethiops* had a smaller SNP sample number which could affect the spread of quality scores (Figure 1). Since the organism cells were either infected with SARS-CoV-2 or received a mock infection, I wondered how it would affect the SNP quality scores. Interestingly, there did not appear to be any clear distinctions between the cell treatment and SNP quality. SARS-CoV-2-treated cells resulted in both low and high quality SNPs, similarly with mock-treated cells (Figure 2). This could be interpreted as SARS-CoV-2 infection of cells not being necessary to produce SNPs and there might be instances of random cellular SNPs being mistaken for SNPS along the SARS-CoV-2 genome.

To understand if the organism that SARS-CoV-2 infects affects SNP development in the viral genome, various cell lines from both organisms in the study were subjected to SARS-CoV-2 infection and these cell cultures served as reservoirs for viral replication. *H. sapiens* cells included gastric cancer lines AGS and MKN45, kidney cancer line Huh, and lung cancer line A549. The Vero cell lines was derived from the kindey

of *Ceropithecus aethiops*, an African green monkey (Ammerman *et al.*, 2008). However, this cell line is often associated with Grivet monkeys, which is why the data from UC Irvine lists the cells as coming from *Chlorocebus aethiops*, instead . To minimize confusion, this report will continue referring to the non-human primate cell samples as *Chlorocebus aethiops* or *C. aethiops*. The number of identified SNPs isolated from each cell line were tallied together and the results showed that the most number of SNPs were analyzed from the A549 or human lung cell line, with 10 identified SNPs (Figure 3). 2 SNPs were found from each AGS, Huh, and Vero cell lines, while 1 SNP from the MKN45 line (Figure 3). At first glance, it appears that SNPs along the SARS-CoV-2 genome occur most often after infecting human lung cells compared to *C. aethiops* kidney cells, however, it is important to remember that the total number of SNPs from each organism type was not equal.

With this information, I decided to re-analyze the SNP quality scores based on their position along the SARS-CoV-2 genome, but this time looking exclusively at the organism cell lines the SNPs were obtained from. Since a majority of the SNPs with quality scores less than or equal to 100 came from cell samples that had received the mock infection, I filtered the data to only include SNPs with quality scores greater than 100 and were from SARS-CoV-2 infected cell lines (Figure 2) (Figure 4). The results from this deeper analysis showed that, from *H. sapiens* samples, SNPs synthesized from cell lines A549 and AGS were of overall lower quality than the Huh kidney cell line (Figure 1). From the *C. aethiops* samples, all of the SNPs had a quality score greater than 200 (Figure 4). This shows that, although the SARS-CoV-2 sequences were synthesized from different organism samples, the program had greater confidence in the SNPs analyzed from kidney cell lines compared to other cell lines. These results were initially surprising given the fact that SARS-CoV-2 heavily affect the respiratory system. However, SARS-CoV-2 has been found to also affect non-respiratory tissues (Klestova, 2023). The differences in quality scores could be due to the presence or lack of a complete reference genome for each cell line. Alternatively, since SARS-CoV-2 typically travels through and infects the respiratory system first, it's possible that there may have been more than 1 variant of the virus in the lung tissue cells, which could have impeded the program's ability to accurately discern SNPs in the sequences. SARS-CoV-2 sequences obtained from kidney cells may have had "greater quality" SNPs due to there only being 1 variant. Overall, this analysis shows that we cannot assume that increases in the number of SNPs equates to greater SNP quality. Additionally, the produced figure highlights that the SNPs were approximately found at positions 8500, 18000, and 29000 of the SARS-CoV-2 genome (Figure 4).

To understand the importance of those genome positions, I decided to look at the various genes on the SARS-CoV-2 genome, using the data from the gff file. This information was compiled into a table which included the names of the genes, the start and end positions of the genes on the genome, and their base pair length (Table 1). With this information on hand, I tallied the number of SNPs found on each of the genes as listed in the dataset. It was discovered that the SNPs in the data were only found on four of the SARS-CoV-2 genes: ORF1ab, ORF8, S, and M (Figure 5). Of these genes, 7 distinct SNPs were located on the ORF1ab gene, 2 SNPs on the S gene, and 1 SNP each on the ORF8 and M genes (Figure 5). Since ORF1ab and S genes are the two longest genes on the genome, it's possible that more SNPs would be discovered on those regions (Table 1). Next, I analyzed the SNP sequences based on the SARS-CoV-2 gene they were found on to determine if the four listed genes correlate with the position of the SNPs along the genome. The results showed that SNPs found between position 8500 and 18000 were located on the ORF1ab gene, the SNPs on position 23000 were on the S gene, the SNP on position 26500 was on the M gene, and the SNPs at position 28000 to 29000 were found on the ORF8 gene (Figure 6). This confirms that the higher quality SNPs, shown according to the cell line they were synthesized from, were found on the ORF1ab and ORF8 gene (Figure 4) (Figure 6). Althoughb ORF1ab was found to have more identified SNPs than ORF8, the higher quality SNPs on those genes suggest importance to SARS-CoV-2 (Figure 5). The length of the SNP sequences were also analyzed based on their position on the genome and the organism the sequences were obtained from. Overall, the analysis revealed consistent SNP lengths from both *H. sapiens* and *C. aethiops* tissues, except for two instances of SNPs from *H. sapiens* tissues on position 10500 and 26500 (Figure 6). This indicates no relationship between the length of the SNP sequences compared to the number of SNPs identified on the genes and their quality scores.

Finally, I decided to investigate the overall SNP quality scores from each organism sample, depending on which gene they were discovered on. When looking into the SARS-CoV-2 infected *C. aethiops* samples, SNPs were only found on the ORF1ab and ORF8 genes, with the highest average SNP quality score being

from ORF8 (Figure 7). For *H. sapiens* samples, SNPs were found on ORF1ab, ORF8, and the S gene, with the highest average quality score also being found on the ORF8 gene (Figure 7). When compiling all of the SNPs and organizing the information based on the SARS-CoV-2 gene it affects, organism it was synthesized from, and the tissue sample, I discovered that the majority of the identified SNPs from *H. sapiens* and *C. aethiops* samples were on the ORF1ab gene, though the sequences were analyzed from different tissue sources (Table 2). Even with this information, the quality score ranges show that not all of those SNPs had high quality scores which could be an indicator of different SARS-CoV-2 variants with mutations along the same gene. This indicates that ORF1ab is a non-conserved region of the SARS-CoV-2 genome, with many SNPs occurring there.

Based on this analysis, it is clear that ORF1ab is an important part of SARS-CoV-2 evasion of immune responses, whether in *H. sapiens* or *C. aethiops*. Although the large number of SNPs occurring in this region could also have to do with its large gene length, it still suggests that the gene plays a role in the virus' evolution. However, more research would need to be performed on a larger dataset.

Figures

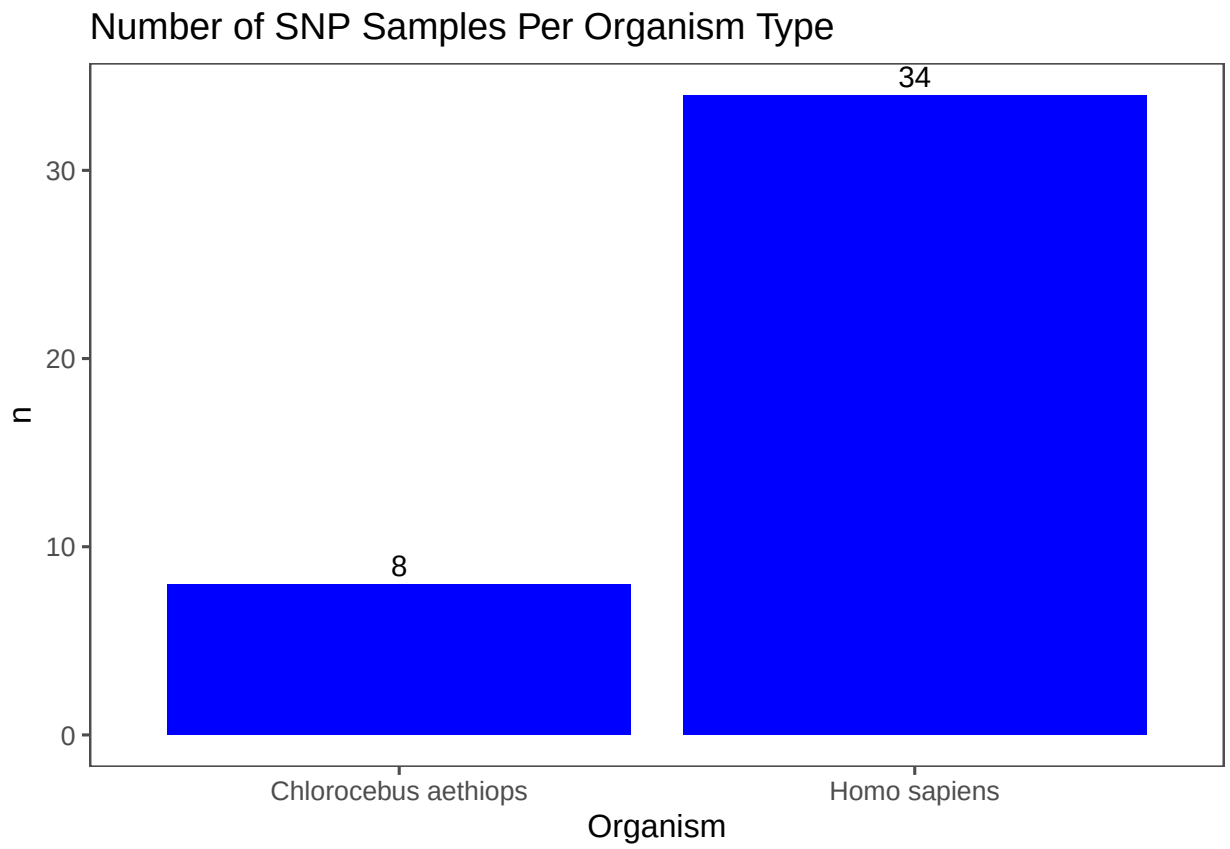

Figure 1: A comparison of the number of SNP samples from *Homo sapiens* and *Chlorocebus aethiops*.

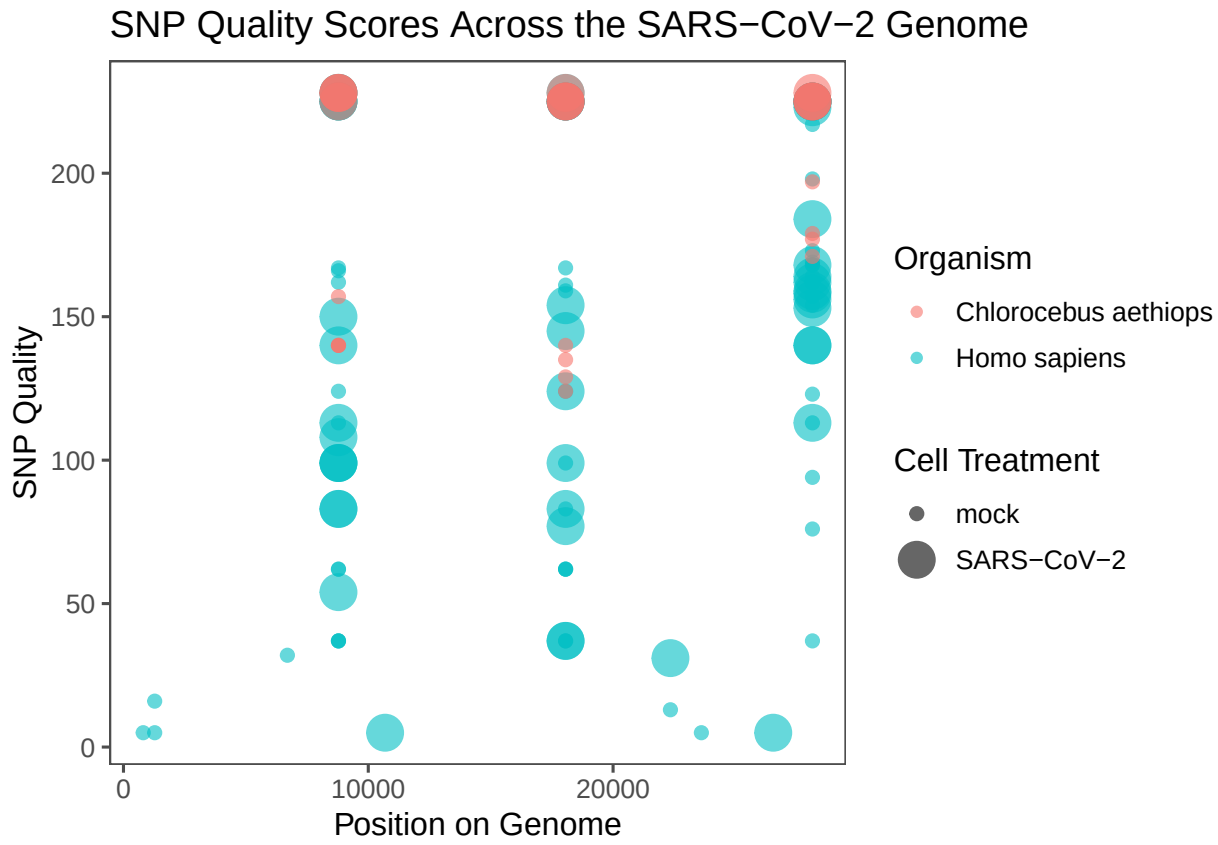

Figure 2: SNP quality across the SARS-CoV-2 genome, isolated from *Homo sapiens* and *Chlorocebus aethiops* cells. The SNPs found in the lower quality region of the graph are mostly from mock infected cells, while low and higher quality SNPs came from SARS-CoV-2 infected cells. SARS-CoV-2 replicated in *H. sapiens* cells had low and high quality SNPs, while mostly high quality SNPs came from *C. aethiops* cells.

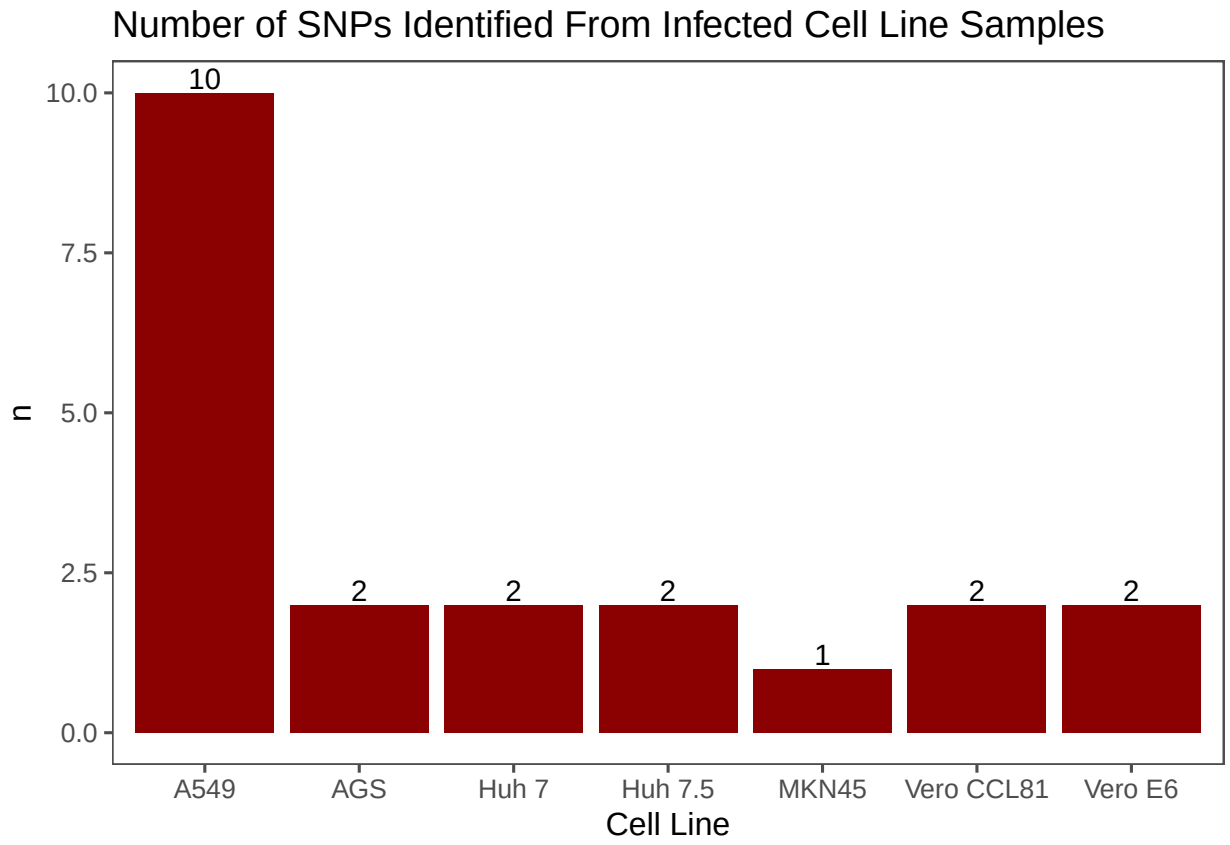

Figure 3: A histogram of the number of distinct SNPs identified from SARS-CoV-2 sequences obtained from *H. sapiens* and *C. aethiops* cell lines. A549, AGS, Huh 7/7.5, MKN45 are *H. sapien* cell lines, while Vero CCL81 and E6 are *C. aethiops* cell lines.

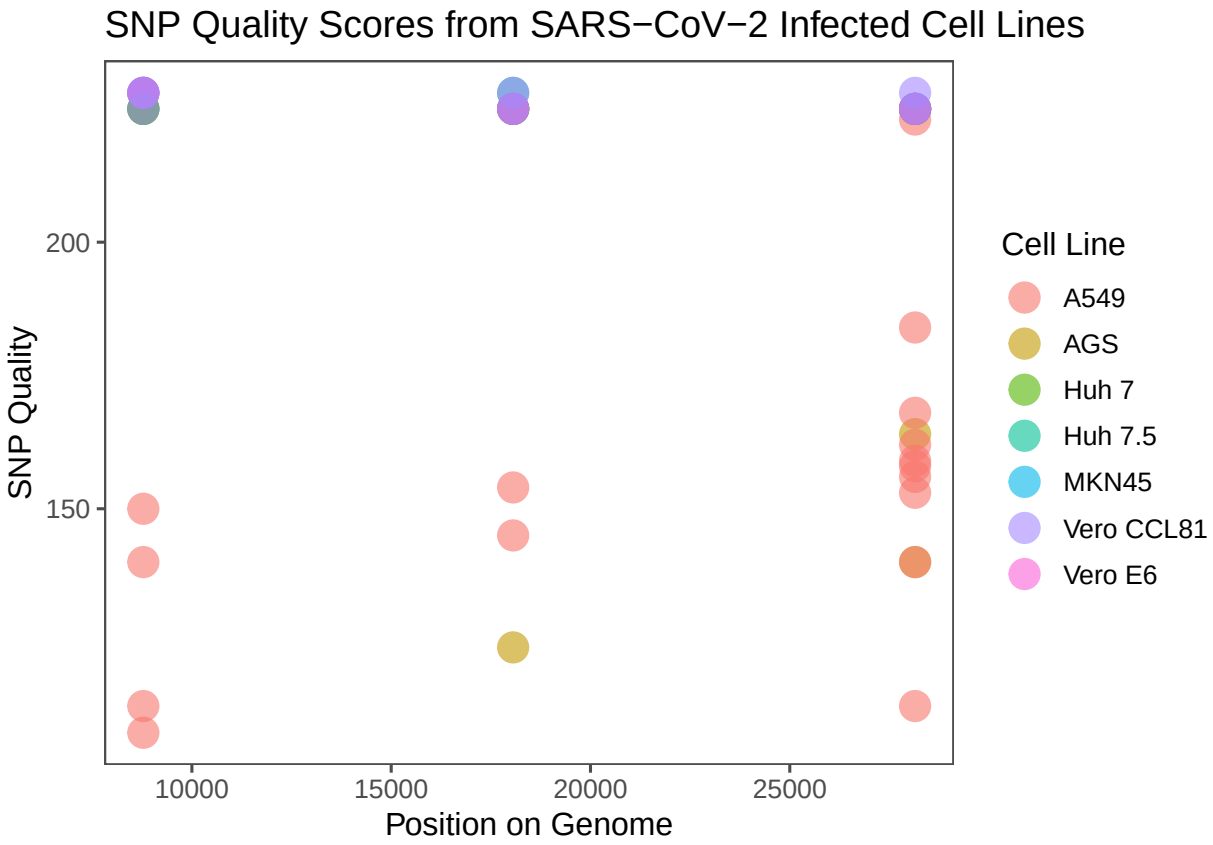

Figure 4: SNP quality scores across the SARS-CoV-2 genome based on the cell line they were synthesized from. A549, AGS, Huh 7/7.5, MKN45 are *H. sapien* cell lines, while Vero CCL81 and E6 are *C. aethiops* cell lines. Only the SNPs with quality scores greater than 100 were included.

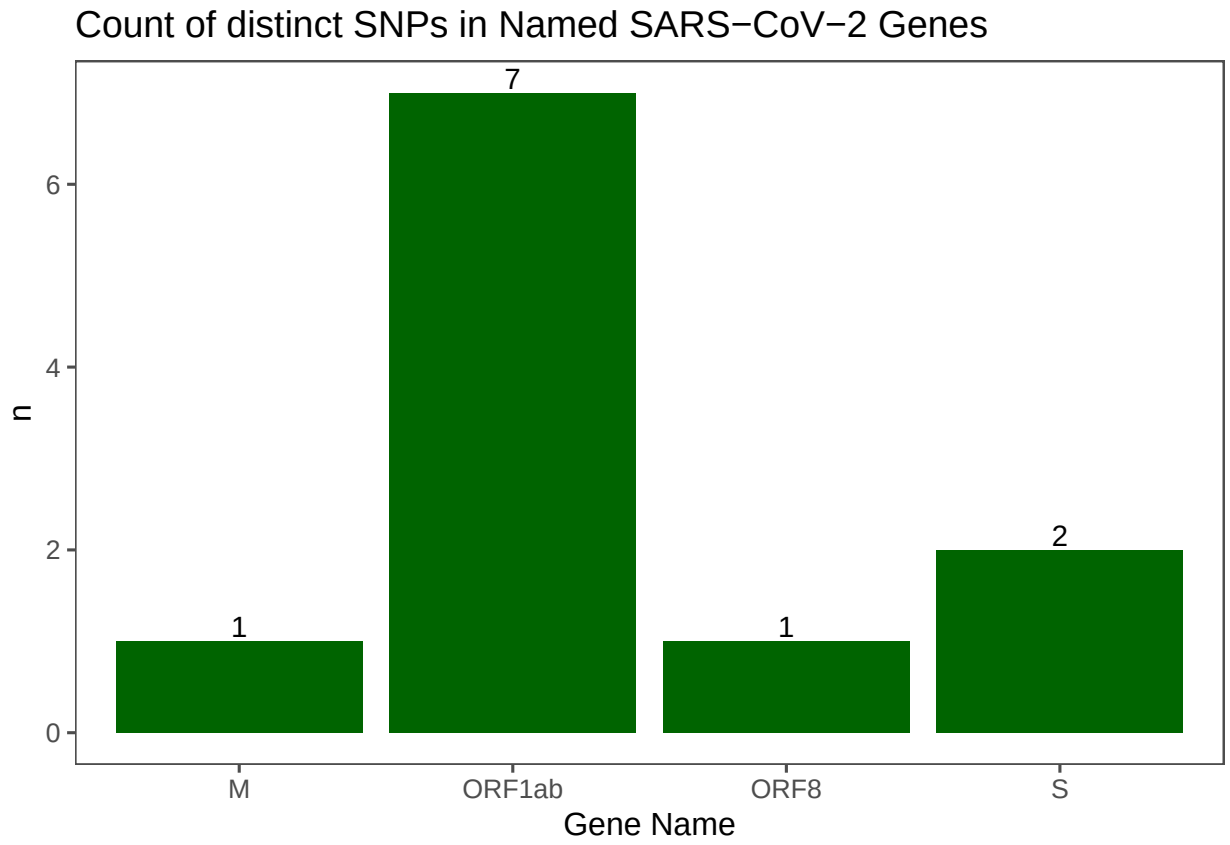

Figure 5: A graph of the number of distinct SNPs from different SARS-CoV-2 Genes. This figure reveals that most of the SNPs are located within the ORF1ab and S gene regions.

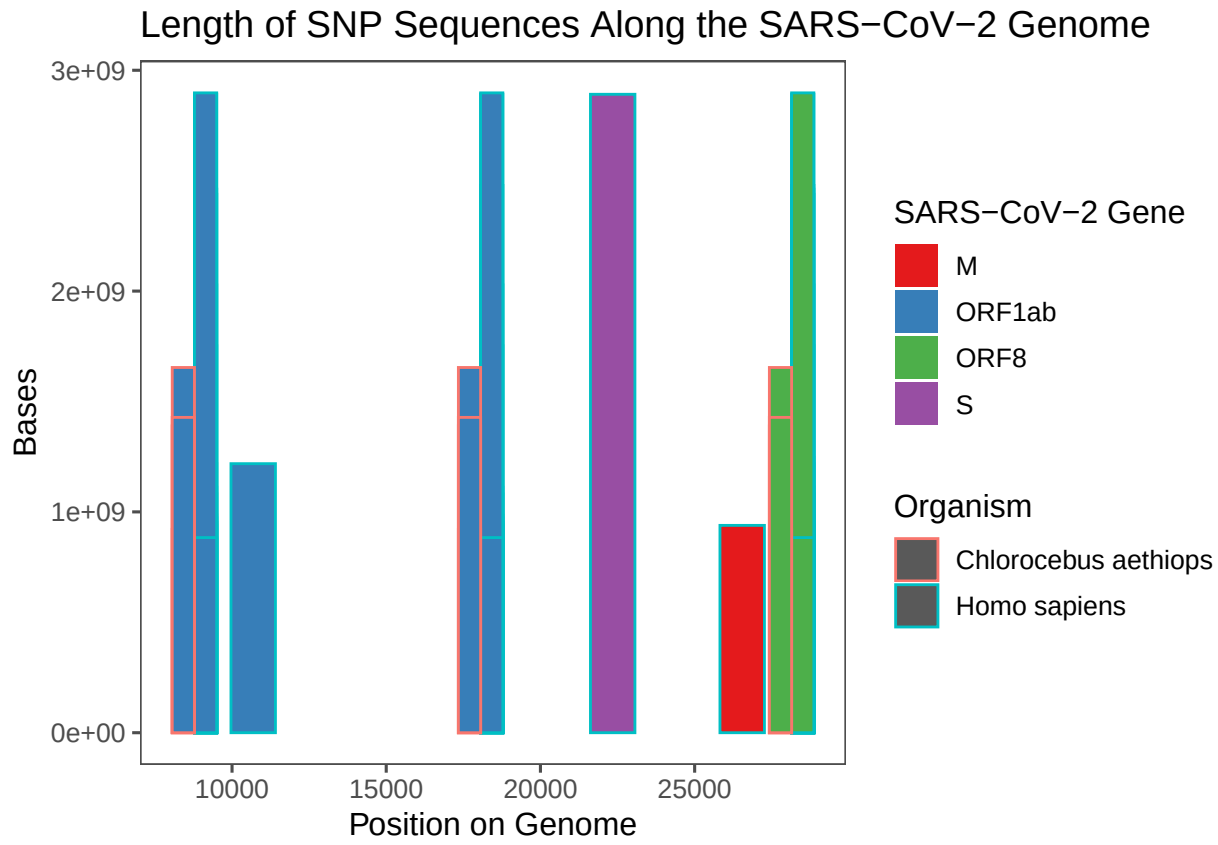

Figure 6: A histogram of the length of SNP sequences depending on their position on the SARS-CoV-2 genome. The colors indicate which SARS-CoV-2 gene the SNPs are located on, and the color outline represents from which organism's cells the SNP readings were obtained from.

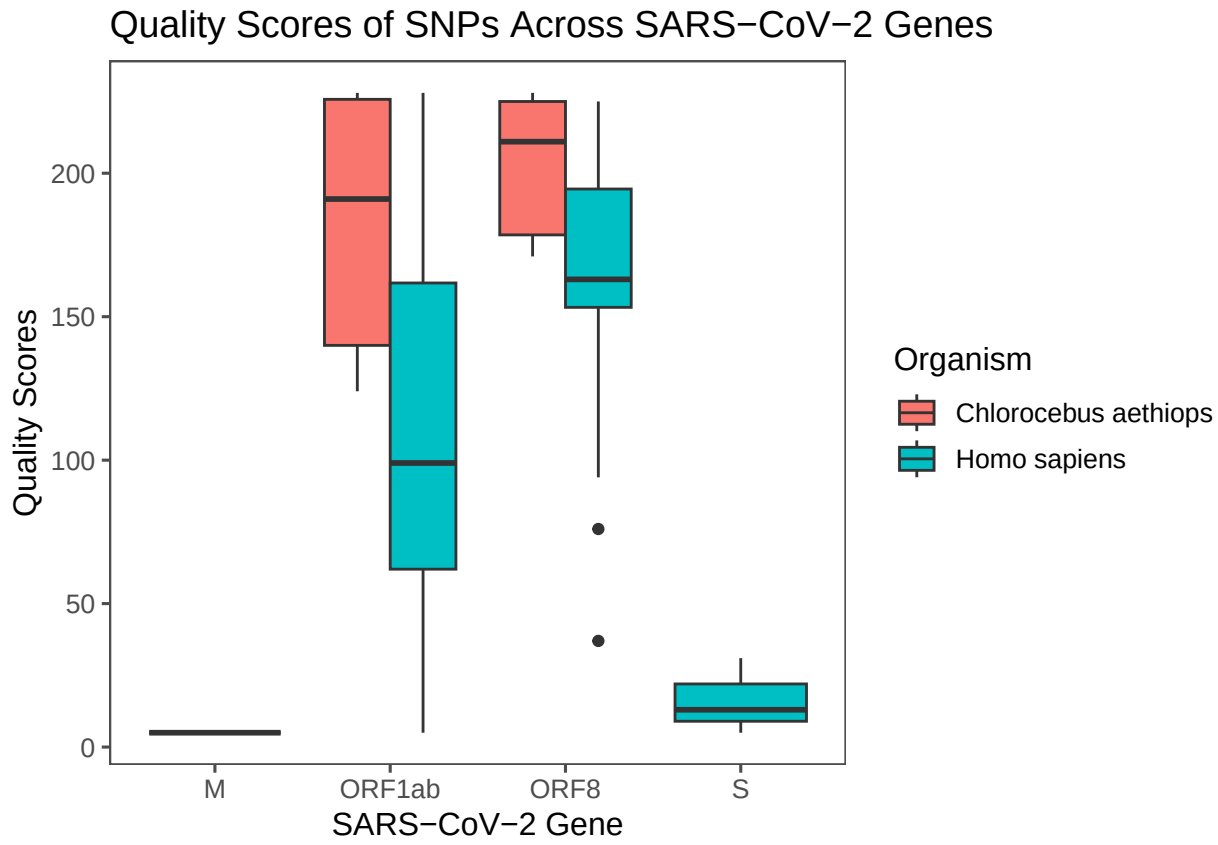

Figure 7: A boxplot comparing the range of and average SNP quality scores on each SARS-CoV-2 gene, synthesized from *H. sapiens* and *C. aethiops* samples.

Tables

Gene Name	Start	End	Length
ORF1ab	266	21555	21289
S	21563	25384	3821
ORF3a	25393	26220	827
E	26245	26472	227
M	26523	27191	668
ORF6	27202	27387	185
ORF7a	27394	27759	365
ORF7b	27756	27887	131
ORF8	27894	28259	365
N	28274	29533	1259
ORF10	29558	29674	116

Table 1: Gene names, locations, and lengths in the SARS-CoV-2 genome. Higher SNP counts in the S and N genes may be related to the larger size of these genes.

Gene Name	Organism	Tissue	Treatment	Number of SNPs
ORF1ab	Homo sapiens	lung	SARS-CoV-2	16
ORF8	Homo sapiens	lung	SARS-CoV-2	10
ORF1ab	Chlorocebus aethiops	kidney	SARS-CoV-2	8
ORF1ab	Homo sapiens	kidney	SARS-CoV-2	8
ORF1ab	Homo sapiens	gastric	SARS-CoV-2	5
ORF8	Chlorocebus aethiops	kidney	SARS-CoV-2	4
ORF8	Homo sapiens	kidney	SARS-CoV-2	4
ORF8	Homo sapiens	gastric	SARS-CoV-2	3
M	Homo sapiens	lung	SARS-CoV-2	1
S	Homo sapiens	gastric	SARS-CoV-2	1

Table 2: A table of all identified SNPs, organized based on the SARS-CoV-2 gene it's found on as well as the organism and the tissue type it was synthesized from. All of the organism's cells received the SARS-CoV-2 infection. The most number of SNPs analyzed from *H. sapien* and *C. aethiops* cells are located on the ORF1ab gene.

Sources Cited

ORF1AB Orf1a polyprotein;orf1ab polyprotein [severe acute respiratory syndrome coronavirus 2] - gene - NCBI.

National Center for Biotechnology Information.

Culture Collections.

Ammerman,N.C. *et al.* (2008) Growth and maintenance of vero cell lines. *Current Protocols in Microbiology*, **11**.

Arnold,J.B. (2024) Ggthemes: Extra themes, scales and geoms for 'ggplot2'.

Aust,F. (2019) Citr: 'RStudio' add-in to insert markdown citations.

Bolger,A.M. *et al.* (2014) Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Committed to leaving a legacy of hope and tools to build a better tomorrow for all the earth's citizens_2024 (2024) *New England Primate Conservancy.*

Edwards,D. *et al.* (2007) What are SNPs? In, Oraguzie,N.C. *et al.* (eds), *Association mapping in plants.* Springer New York, New York, NY, pp. 41–52.

Geerling,E. *et al.* (2022) Roles of antiviral sensing and type i interferon signaling in the restriction of SARS-CoV-2 replication. *Iscience*, **25**.

Klestova,Z. (2023) The effects of SARS-COV-2 on susceptible human cells. *Acta Virologica*, **67**.

Knaus,B.J. (2023) Introduction to vcfR. *Introduction to vcfr.*

Koyama,T. *et al.* (2020) Variant analysis of SARS-CoV-2 genomes. *Bulletin of the World Health Organization*, **98**, 495.

Neuwirth,E. (2022) Colorbrewer palettes [r package RColorBrewer version 1.1-3]. *Home Page.*

Transcriptional analysis of SARS-CoV-2 infected human and nonhuman cell lines *National Center for Biotechnology Information.*

Vinjamuri,S. *et al.* (2022) SARS-CoV-2 ORF8: One protein, seemingly one structure, and many functions. *Frontiers in Immunology*, **13**.

Wickham,H. *et al.* (2023) Dplyr: A grammar of data manipulation.

Wickham,H. (2016) ggplot2: Elegant graphics for data analysis Springer-Verlag New York.

Xie,Y. (2024a) Knitr: A general-purpose package for dynamic report generation in r.

Xie,Y. (2024b) Tinytex: Helper functions to install and maintain TeX live, and compile LaTeX documents.