Analysis of SARS-CoV-2 variants in Zimbabwe

Audrey Koti

13 May, 2021

Background and Overview

Zimbabwe recorded its first COVID-19 case in March 2020 and by the end of the year, sequencing results showed an increase in the occurrence of variants in the SARS-CoV-2 genome found in samples collected nationwide (Mashe, Takawira, and Oliveira Martins *et al.*, 2021). This influx of variants was likely caused by increased travel between South Africa, where the B.1.351 variant (Mashe, Takawira, and Gumbo *et al.*, 2021) was first reported, and Zimbabwe during the December holiday season. Although the government issued social-distancing and lockdown measures, these measures were not followed or enforced strictly. This may be why several variants spread among the population as shown by the data from the surveillance carried out between the end of 2020 and January 2021. This surveillance of the SARS-CoV-2 genome also identified the C.2 and A.23.1 variants in addition to the B.1.351 variant. However, the occurrence of the C.2 and A.23.1 variants declined in Zimbabwe towards the end of January 2021 (Mashe, Takawira, and Gumbo *et al.*, 2021). All three SARS-CoV-2 variants are associated with increased transmission. Moreover, the B.1.351 variant is also associated with higher mortality as observed in South Africa. At the time the study was conducted, the B.1.351 variant was found in 69% of the 107 cases sequenced in December 2020 and in 95% of the 104 cases sequenced in January (ZimFact, 2021). This report was an analysis of the occurrence of mainly single nucleotide polymorphisms (SNPs) in the SARS-CoV-2 genes in people of different ages and sex. Most COVID-19 patients from Zimbabwe were male and more than half of the cases reported occurred among people of ages 20 to 40 (Mashe, Takawira, and Oliveira Martins *et al.*, 2021). This may be why Zimbabwe's deaths were only about 2.9% of the deaths in neighboring South Africa, despite Zimbabwe having little enforcement of the guidelines to curb transmission, as people in the age group of 20 to 40 years have a better chance of recovering from COVID-19 than those older. Data show that most of the deaths occurred in January 2021 after an increase in the presence of the B.1.351 variant (Reuters, 2021). Variant analysis is important in understanding how SARS-CoV-2 mutates and how that affects public health measures. I analyzed the distinct SNPs in SARS-CoV-2 genes grouped by age and for each sex. Next, I looked at the unique alternative alleles at each position of the sequence sorted by age and by sex. Finally, I counted the alternative alleles in the gene for the spike protein to determine the most common allele that may be the most responsible for increased transmission. I found that there is a strong association between the "T" alternative allele and males aged 30 to 50, which may have led to the higher number of COVID-19 cases among this group of people in Zimbabwe.

Methods

This report looked at the dominant variants of the SARS-CoV-2 genes that are of concern in Zimbabwe. I obtained the data from the National Center for Biotechnology Information (NCBI, 2021). I carried out the analysis of these variants in `RStudio` (RStudio Team, 2020) using R functions and packages such as `vcfR` (Knaus and Grünwald, 2016) to load in the VCF files I needed to work with, `ggplot2` (Wickham, 2016) to plot figures, and `dplyr` (Wickham *et al.*, 2020) to group by variables as well as count up observations from the data.

Analysis of COVID-19 cases in Zimbabwe

I used `ggplot2` (Wickham, 2016) to represent, on a bar plot filled by the host sex, the number of COVID-19 cases in male and female patients 18 years and older. This was done to better understand which age group and sex would have better data for variant analysis.

Analysis of Distinct SNPs in SARS-CoV-2 Genes

Using a script written by Professor Zimmerman, I looked at the number of distinct single nucleotide polymorphisms within each named gene to determine where the majority of the variants were found. The number for each gene was tallied up after grouping by the position of the variations in the sequence. This involved using functions such as `filter()`, `tally()` and `group_by()` from the `dplyr` package (Wickham *et al.*, 2020). I then used the a scatter plot to visualize the occurrence of SARS-CoV-2 genes in the two sexes to confirm whether or not the genes with more unique SNPs were the same for each sex. After determining the gene with the most unique SNPs, I counted the variations in that particular gene to understand the spread of these variations in males and females of different ages. I then created a line graph showing when the variations increased in Zimbabwe in the time the surveillance was carried out. I used the `ymd()` function from the `lubridate` package (Grolemund and Wickham, 2011) to load the collection date in the correct format.

Analysis of Distinct SNPs in SARS-CoV-2 Genes by Sex and Age

I used a table made by Professor Zimmerman to look at the sizes of each gene. Based on the longest gene with the most unique SNPs, I analyzed the most common unique SNP in the different patient sexes and ages to determine how the occurrence of this SNP was distributed. I created a table that counts the number of times the alternative alleles occur in the gene and plotted a bar plot for the distribution of the most commonly occurring allele in male and female patients aged 18 and older using `gglopt2` (Wickham, 2016). Lastly, I created a table to show whether this same alternative allele was also the most commonly occurring in patients under the age of 18, who represent the minority of COVID-19 cases in Zimbabwe.

Results

Although the surveillance from which the data this report analyses suggested that most COVID-19 patients were in the age group of 20 to 40, the demographics plot of this analysis showed that most of the patients were aged 25 to 65 as shown in Figure 1. The majority of these patients were male. Figure 2 shows that variations in the SARS-CoV-2 genome were present mostly in the N and S genes, with the S gene having the most variations. These genes were found to be the longest genes with 3,821 and 1,259 base pairs respectively as represented in Table 1. The S gene variations were more common among males aged 30 to 65 years and also females around the age of 35 as Figure 4 shows. This analysis also showed in Figure 5 that the S variations increased dramatically around late December 2020 to January 2021. Most mutations - present more in males as indicated by Figure 3 - were a change to the “T” allele from either the “C” or the “G” reference alleles as shown in Figure 6. Table 2 shows that the “T” allele was the most common allele occurring 745 times in the study. The next common allele was the “G” allele, which was present 537 times. Table 2 also shows that most of the mutations were single nucleotide polymorphisms as these appeared 1844 times out of 1,851 in the patients. From Table 3, we learned that the “T” allele was also the most common allele in patients under 18 years. However, this group of the study had 6 out of 8 of the characteristic mutations associated with the B.1.351 variant. The “TTA” and the “TTTA” mutations present in the rest of the sample population were not reported in this group of patients under 18.

Discussion

Zimbabwe has a young population working mostly in neighboring South Africa where the B.1.351 SARS-CoV-2 variant was first reported. The majority of Zimbabwe's COVID-19 patients likely reside and/or work in South Africa and visit Zimbabwe especially during holidays. The data showed that most people infected with SARS-CoV-2 were the young to middle-aged people with ages ranging from 25 to 65. The majority of Zimbabweans working in South Africa are males and this likely explains why the data showed that most COVID-19 patients from Zimbabwe were male. The B.1.351 variant is associated with eight characteristic mutations on the binding domain of the S gene that codes for the spike protein, making the mutated virus more efficient in human transmission. The eight mutations of this variant lead to changes from the amino acids asparagine to tyrosine, glutamate to lysine and lysine to asparagine. These changes are likely what makes the spike protein more efficient at binding human cells upon infection. Variants can also make diagnosis harder and lead to more severe illness that is difficult to treat (Otu *et al.*, 2021). This report showed an increase in the presence of S gene mutations among Zimbabwean patients beginning at the end of December 2020 until January 2021. This is the time Zimbabweans working in outside the country mostly return to Zimbabwe to visit families. It is highly likely that increased travel between Zimbabwe and other countries, in particular South Africa, caused an increase of COVID-19 cases. Zimbabwe announced strict measures to curb the disease early in the pandemic. However, these measures were not strictly enforced and the enforcement officers may have been overwhelmed by the increased holiday travel. This probably explains the spike in cases and even deaths in January 2021 (Reuters, 2021). Zimbabwe's ports of entry had strict screening measures and people entering the country were required to quarantine for at least two weeks (Makurumidze, 2020). However, because of reduced containment measures, there were increased interactions among Zimbabweans that may have allowed transmission of the virus from those returning to those already at home. Zimbabwe's economy is largely informal and lockdown measures have proven impractical for the survival of most Zimbabweans. The spread of COVID-19 was probably exacerbated by increased interactions at the beginning of the year 2021 as people of working age, mostly those above 25, returned to work. The higher SNP counts found in the S and N genes may be related to the larger sizes of these genes. The "T" allele was also the most common one among patients under the age of 18. This age group did not have the mutations "TT" or "TTTA" unlike those aged 25 and older. This may just be because there were much fewer COVID-19 patients under 18. The "C" or "G" to T" mutation is likely the most favorable for increased transmission as this alternative allele was more than four times more common than the next most common alternative allele in patients over 25 years. More studies on the effects of the B.1.351 variant need to be carried out to explain the significant differences in the number of cases and deaths between Zimbabwe and South Africa, which have quite similar population densities of 38 and 49 people per square kilometer respectively.

Figures

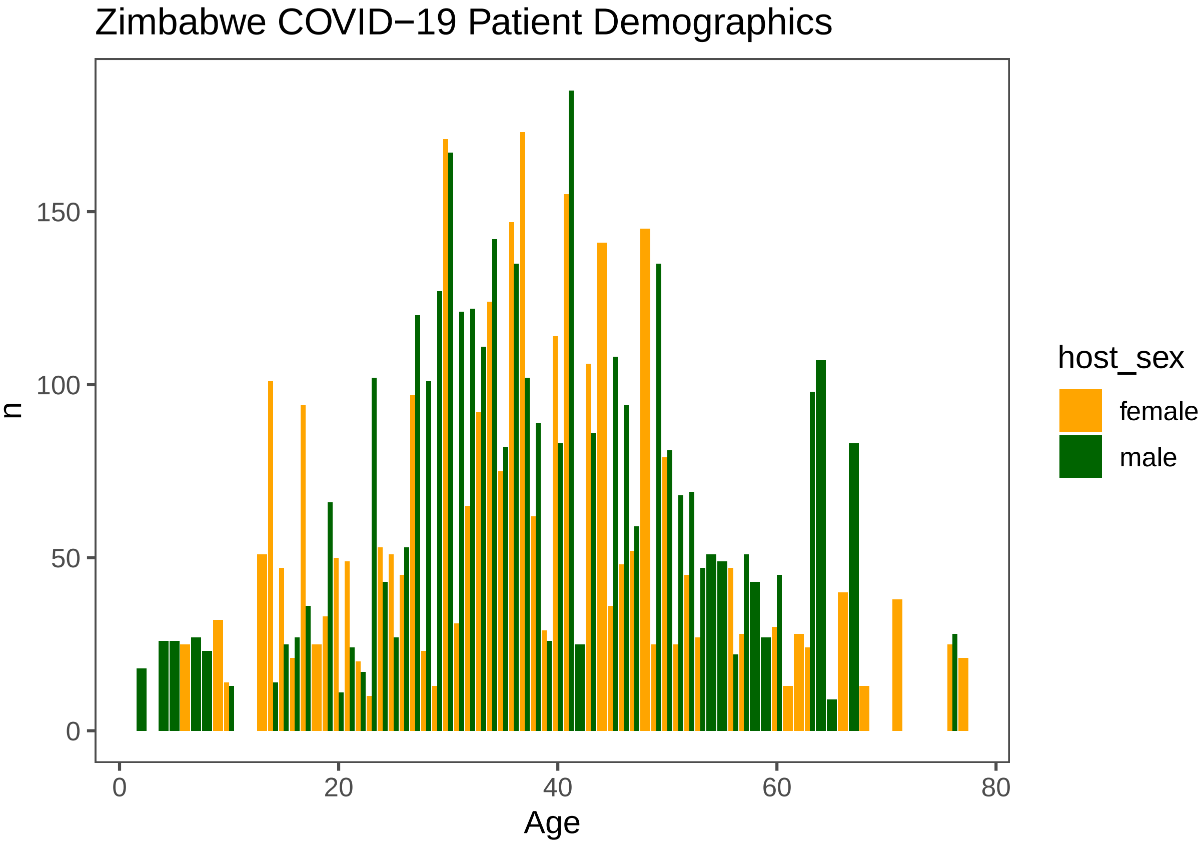

Figure 1: Figure 1 represents the demographics of COVID-19 patients in Zimbabwe.

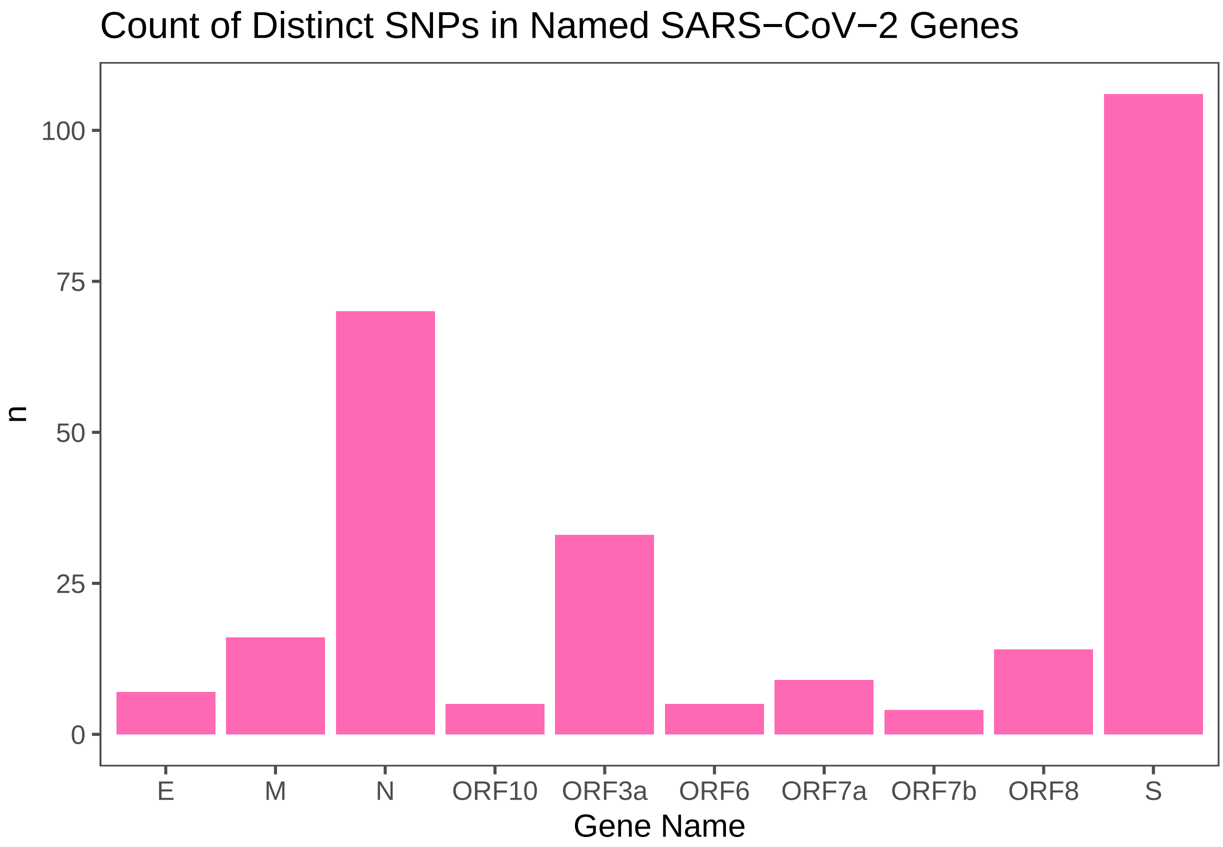

Figure 2: Figure 1 shows the locations of unique SNPs for each SARS-CoV-2 gene.

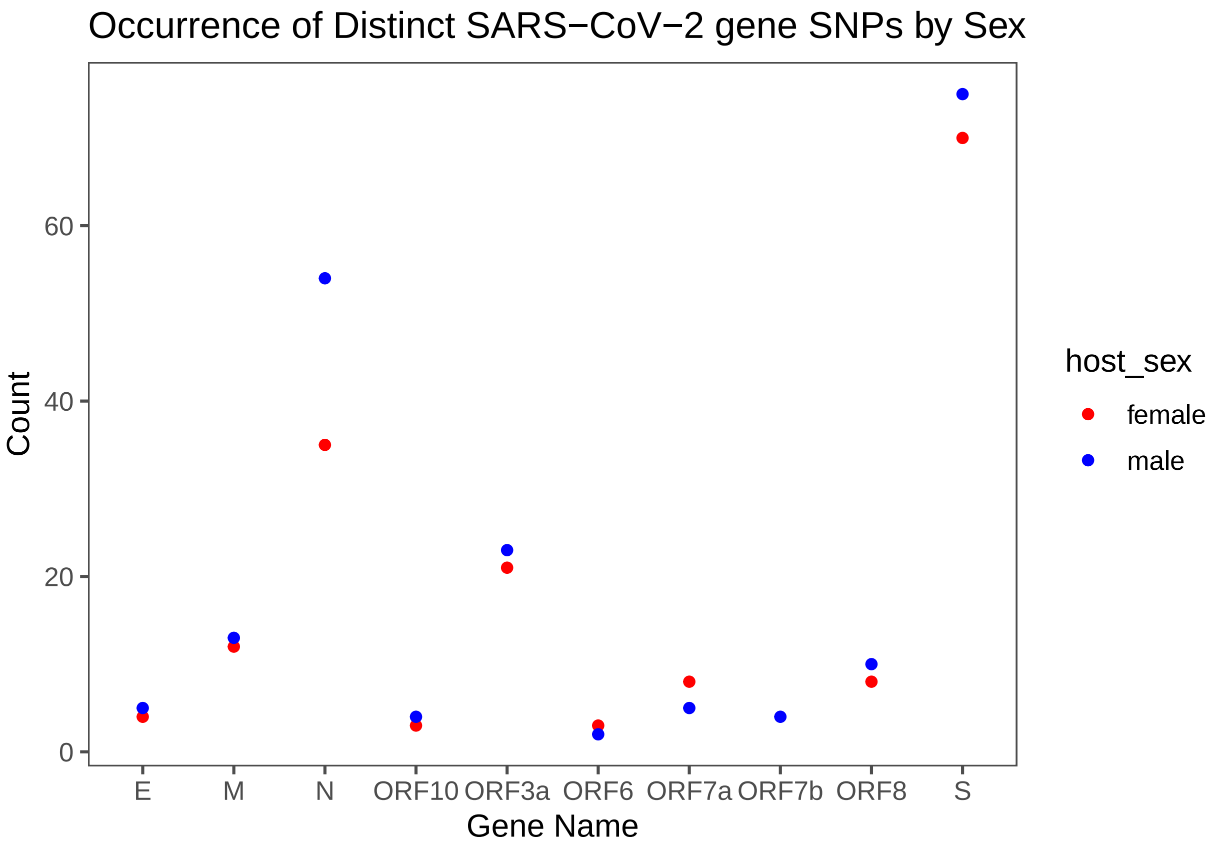

Figure 3: Figure 3 shows the distribution of distinct SNPs in each sex.

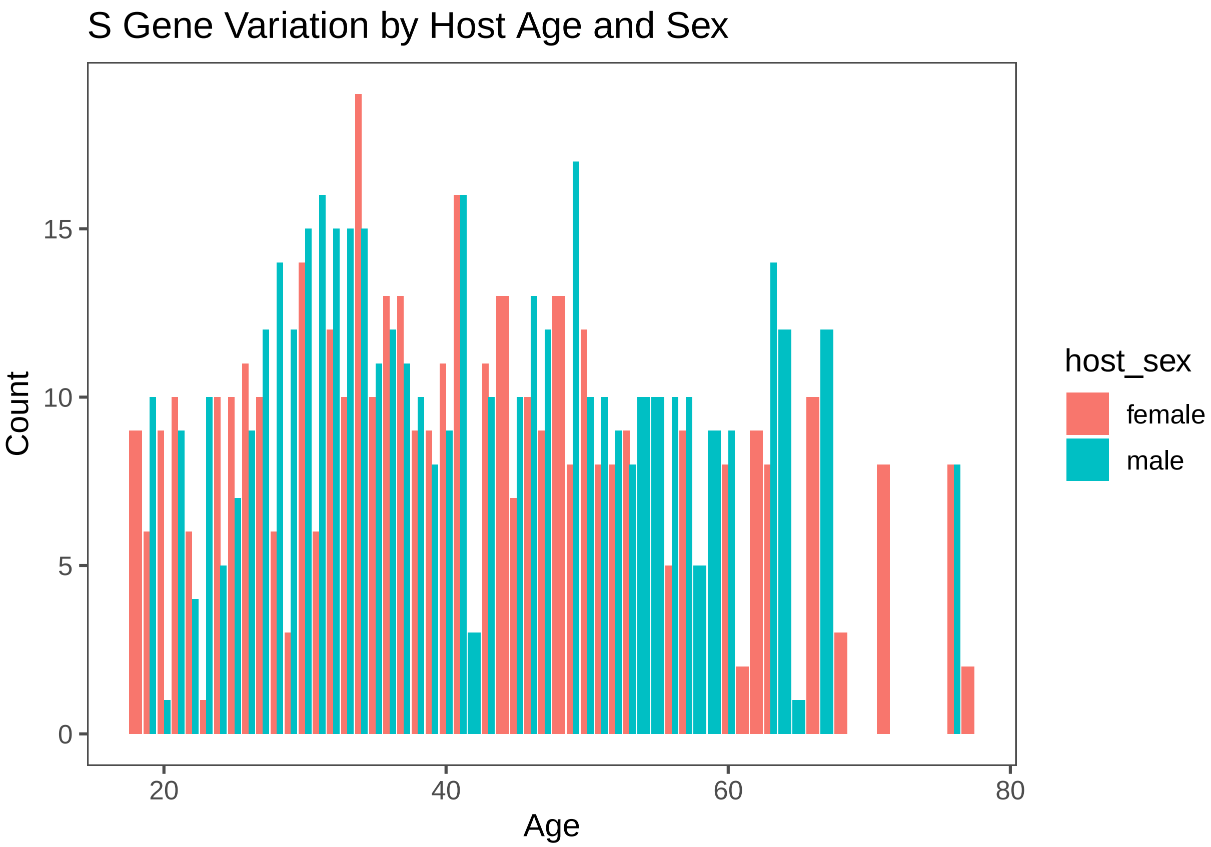

Figure 4: Figure 4 shows the count of the S gene variations in patients older than 18 years.

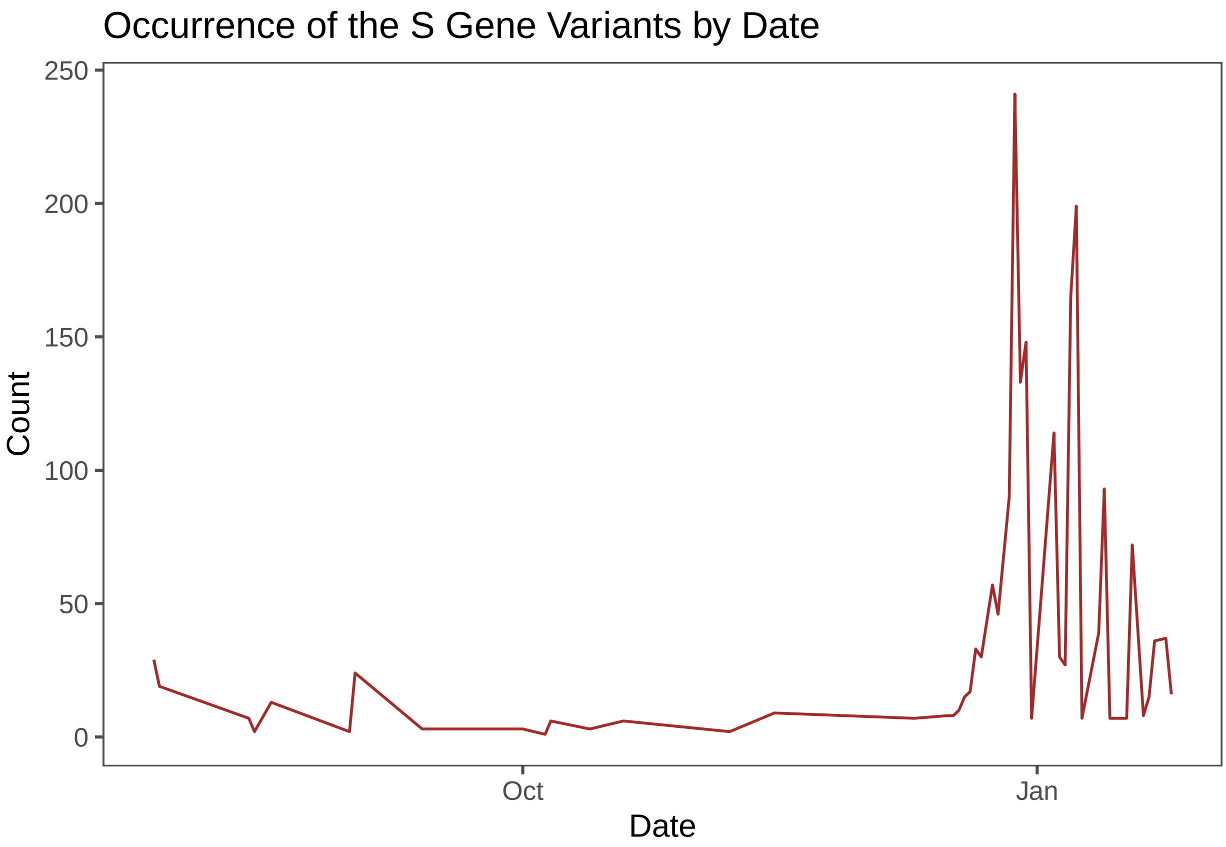

Figure 5: Figure 5 shows the normalized occurrence of S gene variants each day between December 2020 and January 2021.

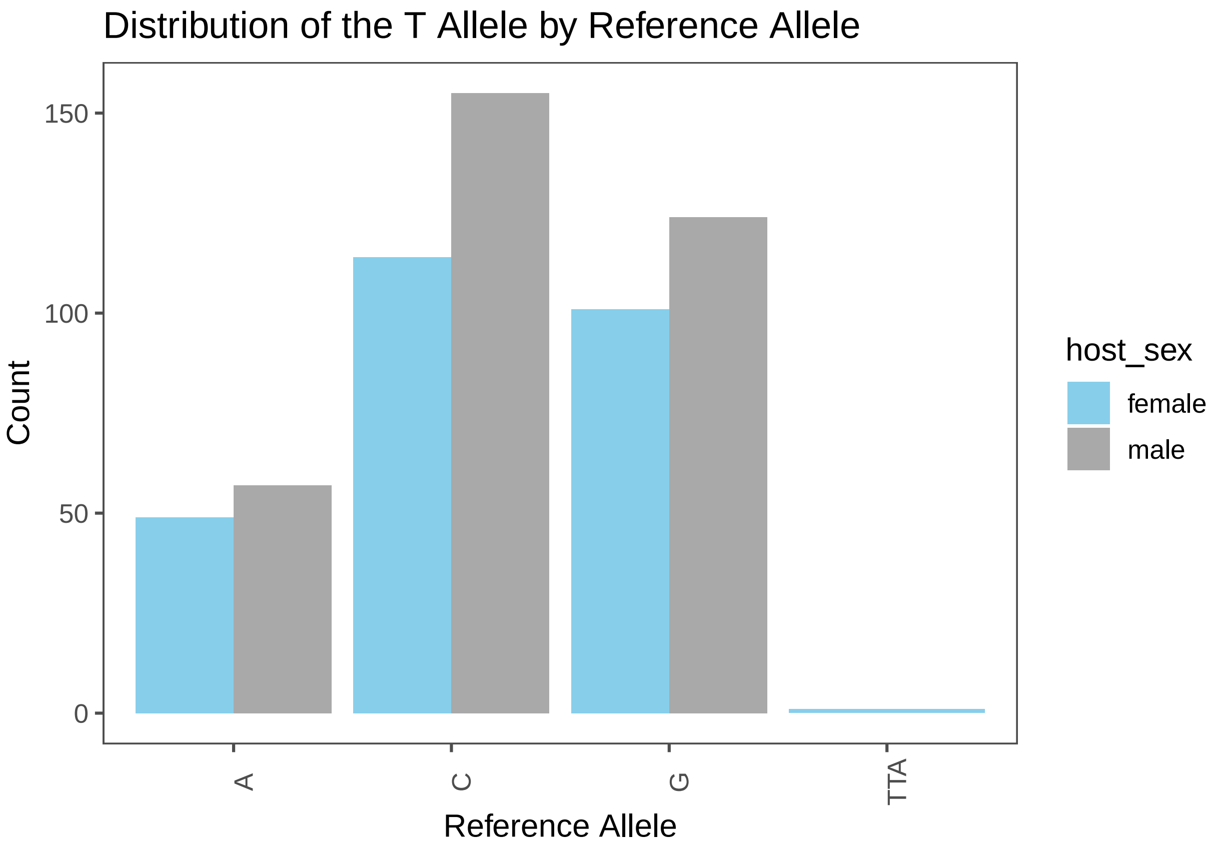

Figure 6: Figure 6 shows the change from represented reference alleles to the "T" allele in the S gene in male and female patients between 25 years and 65 years of age.

Tables

Gene Name	Start	End	Length
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: Table 1 shows the length of each SARS-CoV-2 gene.

Gene Name	Alternative Allele	Count
S	A	169
S	ACTTTAC	148
S	C	245
S	G	537
S	T	745
S	TT	2
S	TTA	1
S	TTTA	4

Table 2: Table 2 shows the count of alternative alleles found on the S gene.

Gene Name	Alternative Allele	Count
S	A	17
S	ACTTTAC	14
S	C	23
S	G	49
S	T	65
S	TTA	1

Table 3: Table 3 shows the count of alternative alleles in the S gene in patients under 18.

Sources Cited

Grolemund,G. and Wickham,H. (2011) Dates and times made easy with lubridate. *Journal of Statistical Software*, **40**, 1–25.

Knaus,B.J. and Grünwald,N.J. (2016) VcfR: An r package to manipulate and visualize VCF format data. *BioRxiv*.

Makurumidze,R. (2020) Coronavirus-19 disease (covid-19): A case series of early suspected cases reported and the implications towards the response to the pandemic in zimbabwe. *Journal of Microbiology, Immunology and Infection*, **53**, 493–498.

Mashe,T. *et al.* (2021) Surveillance of sars-cov-2 in zimbabwe shows dominance of variants of concern. *The Lancet. Microbe.*

Mashe,T. *et al.* (2021) Genomic epidemiology of the sars-cov-2 epidemic in zimbabwe: Role of international travel and regional migration in spread. *medRxiv.*

NCBI (2021) Genomic surveillance of sars-cov-2 in zimbabwe reveals the dominance of variants of concern.

Otu,A. *et al.* (2021) Africa needs more genome sequencing to tackle new variants of sars-cov-2. *Nature Medicine*, 1–2.

Reuters (2021) COVID-19 tracker: Zimbabwe.

RStudio Team (2020) RStudio: Integrated development environment for r RStudio, PBC., Boston, MA.

Wickham,H. (2016) Ggplot2: Elegant graphics for data analysis Springer-Verlag New York.

Wickham,H. *et al.* (2020) Dplyr: A grammar of data manipulation.

ZimFact (2021) New coronavirus variants in zimbabwe.