Variant Analysis of SARS-CoV-2 in Zimbabwe between December 2020 and January 2021

Audrey Koti

09 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 (**???**). This influx of variants was likely caused by increased travel between South Africa, where the B.1.351 variant (**???**) was first reported, and Zimbabwe during the December holiday season. A surveillance of the SARS-CoV-2 genome also identified the C.2 and A.23.1 variants in Zimbabwe between December 2020 and January 2021 (**???**). The 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. However, the occurrence of the C.2 and A.23.1 variants declined in Zimbabwe at the end of the surveillance in January 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 - 40 (**???**). 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 - 40 years have a better chance of recovering from COVID-19 than those older. 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 transmission. »><«»«< alternative allele and this age group and sex, 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 (**???**). I carried out the analysis of these variants in `RStudio`(RStudio Team, 2020) using R functions and packages such as `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 Distinct SNPs in SARS-CoV-2 Genes

Using a script written by Professor Naupaka 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`(Wickham *et al.*, 2020) package.

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

I used the `facet_wrap` function to visualize the occurrence of unique alternative alleles at each position by age and by sex using `ggplopt2`(Wickham, 2016). I also counted up these alternative alleles to find out which one is the most commonly occurring allele among different ages and sexes and represented the findings in a table.

- https://kjhealy.github.io/covdata/
- https://github.com/como-ph/oxcovid19
- https://ropensci.org/blog/2020/10/20/searching-medrxivr-and-biorxiv-preprint-data/
- https://covidtracking.com/data/api
  - `readr::read_csv("https://api.covidtracking.com/v1/states/daily.csv")`

Results and Discussion

```
## Error in dimnames(x) <- dn: length of 'dimnames' [2] not equal to array extent
```

```
## Error in is.data.frame(stacked_vcf): object 'stacked_vcfs' not found
```

Figures

```
## Error in filter(., !is.na(gene)): object 'vcf_with_metadata' not found
```

Figure 1: Figure 1 shows the locations of unique SNPs for each SARS-CoV-2 gene. # N and S genes have more unique SNPs in the set of samples analyzed.

```
## Error in filter(., !is.na(gene)): object 'vcf_with_metadata' not found
```

Figure 2: Figure 2 shows which SNPs are commonly found in different patient ages.

```
## Error in filter(., !is.na(gene)): object 'vcf_with_metadata' not found
```

Figure 3: Figure 3 shows which SNPs are commonly found in each sex.

Figure 4: Figure 4 counts the number of unique alternative alleles in the S gene to determine the most common one in patients of different ages.

Figure 5: Figure 5 counts the number of unique alternative alleles in the S gene to determine the most common one in each sex.

Tables

Gene Name	Start	End	Length
S	21563	25384	3821
ORF3a	25393	26220	827
E	26245	26472	227
M	26523	27191	668

Gene Name	Start	End	Length
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 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.

Table 2: Table 2 represents the count of each alternative allele in the S gene.

Sources Cited

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.