Analyzing SARS-CoV-2 Genome Sequencing Data from Brazil Based on Patient Geolocation and Sex

Chaitrali Deshpande

17 December, 2021

Background and Overview

COVID-19 pandemic dominated the year 2020 and 2021 with no sign of relief as new variants like Omicron despite the increasing vaccination rate throughout the world. It has completely changed the way people interact with each other as the COVID-19 is an air-borne, contagious disease. First discovered in December 2019 in Wuhan, China, COVID-19 is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Symptoms of COVID-19 include fever, chills, shortness of breath, exhaustion, headache, sore throat, nausea, and diarrhea. The United States of America (US), India, and Brazil are the top three countries with the most COVID-19 cases and deaths caused by COVID-19. The pandemic in Brazil is especially troubling considering the smaller population of Brazil compared to the US and India. Between January 3rd and December 10th, 2021, there were 22,167,781 confirmed cases of COVID-19 in Brazil, with 616,251 deaths attributed to the disease. Brazil averaged over 1000 deaths per day by June 2021 (Souza *et al.*, 2020). The foreigners traveling from Europe to the different states of Brazil are considered the main reason for the widespread of the disease. While the north and northeast regions of Brazil are highly vulnerable due to the lack of infrastructure, the metropolitan areas in Sao Paulo and Rio de Janeiro contributed significantly to the COVID-19 cases and deaths.

I recently came across an article (Sanchez, 2021) on how Brazil is becoming the 'natural laboratory' for COVID-19 and a breeding ground for new variants of SARS-CoV-2. The percentage of COVID-19 cases caused by the Delta (B.1.617.2) variant increased from 2.3 percent in June 2021 to 23.6 percent in July 2021. The significant increase in spread prompted the World Health Organization to call it a variant of concern. The Delta variant is considered more contagious than other variants because of the high number of mutations to the spike protein of the virus that alters the amino acid sequence it encodes (Katella, 2021), (Callaway and others, 2021).

Therefore, for the current study, I chose to work with SARS-CoV-2 genome sequencing data obtained between June-July 2021 and from the patients residing in municipalities in Brazil that are worst affected by COVID-19. I extracted genome sequencing data from a study in Brazil and compared them to reference the SARS-CoV-2 genome to count the gene mutations after data processing. Then I performed graphical data analysis to evaluate whether the gene mutations are affected by gender and geography. Though this study could not reach any conclusive finding due to the limited number of genome sequences available from a shorter time period, it builds a solid platform for future studies when more data becomes available.

Methods

Genome Sequence Data

I downloaded the genome sequencing data of COVID-19 patients in municipalities of Brazil that were dominantly affected by COVID-19 states from the NCBI (National Center for Biotechnology Information)

database. It can be found on the BioProject Resource database of NCBI titled 'SARS-CoV-2 genome sequencing Brazil' with Accession Number: PRJNA774631 and SRA Study: SRP343167. It has 66 sequence data obtained using the Illumina platform from populous municipalities in Brazil during June-July 2021.

Genome Sequence Processing Pipeline

The genome sequencing data extraction and processing were automated by using a Bash script. The Bash script performs the following tasks in sequence while saving the processed files at every step for future use.

- Download of the metadata: Using SraRunTable, I downloaded the metadata for Brazil from the NCBI website. The downloaded raw sequences were pushed through 'fasterq-dump', a program written by NCBI to make the downloading of raw sequences feasible.
- Eliminate low quality and unnecessary data: 'fastqc' tool is used to check the quality of raw data. The raw sequencing data contains several redundant data. 'trimmomatic' is used to trim off the primers, delete low-quality base pairs and bad reads.
- Sorting the dataset: I used 'bwa' (Burrows-Wheeler Aligner) that creates the index and maps the short reads against the reference genome. Later, the data is passed through 'samtools' and 'bamtools', which sort and process the mapped reads, respectively.
- Data visualization and Perl script: To visualize the processed and mapped data, I used a Perl script that takes the bam files as input and processes them into 'vcf' files. The mapped and processed reads can be viewed in 'IGV'.

Brazil COVID-19 Pandemic Statistics

In addition to the genome data, I scoured the internet to obtain municipality-level statistics regarding the number of COVID-19 cases, deaths, and vaccination rates in Brazil in the period between June-July 2021. The Github repository (Cota, 2020) maintains temporal statistical data for the COID-19 pandemic in Brazil. I extracted the data relevant to this project i.e. data on 1st June 2021 using excel and copied it in a 'csv' file so that it can be imported by the R script.

R Packages

I imported the 'vcf' files created by the the bash script in R to perform analysis and visualization. For this purpose, I used a variety of R tools and packages. I used readr (Wickham and Hester *et al.*, 2021) to read and write data and output 'csv' files, while 'knittr' and 'tinytex' are used to create the final report files in pdf format from the R markdown file. I used janitor (Firke, 2021) to clean up unnecessary data files in my data. 'Dplyr' (Wickham and François *et al.*, 2021) and 'tidyr' (Wickham, 2021) are used to parse and categorize data. I used 'ggplot2' (Wickham, 2016) with 'ggthemes' (Arnold, 2021) for the graphical analysis of the data and 'kableExtra' (Zhu, 2021) to make the tables look nicer.

Results and Discussion

The raw genome sequencing data passed the quality checks of the 'fastqc' tool. To maintain the quality of the sequences, I selected high-quality base pairs and good reads by trimming the sequences less than 100 in length. Table 1 lists the number of sequences per geographic location (municipalities) and host sex analyzed in this study. It can be noted from Table 1 that, for some locations, the sample count is 1 from either a male host or a female host. As it would no possible to compare the mutation data per host sex for these locations, they were excluded from the analysis.

Table 2 contains the total number of COVID-19 cases per 100k inhabitants recorded as of June 1, 2021, in the municipalities for which the genome sequencing data was extracted in this study. The table also lists the vaccination rate in these regions as of June 1, 2021. I chose this date as this coincides with the start of the period for which we have the genome sequencing data available. this data was obtained from the Github repository (Cota, 2020).

Mutations and Host Sex, Geographic Location

Figure 1 illustrates the number of distinct Single Nucleotide Polymorphisms (SNPs) in named SARS-CoV-2 virus genes found in this study. The figure also distinguishes the count of distinct mutations per host sex. The number of mutations in genes S and N are considerably higher than the rest of the genes for both male and female hosts. The mutation count for ORF7b is the lowest. The length of the genes S and N are among the longest genes of SARS-CoV-2 virus as listed in Table 3. This explains the higher number of SNPs in these genes, as the rate of mutation in them will be faster as they undergo the transcription and translation processes.

I found it interesting to note that the number of distinct SNPs in gene S is more than twofold higher in males compared to females. We can say that males are more susceptible to COVID-19 and provide a favorable host. This is also evident from the higher number of COVID-19 infections among males than females, globally.

Figure 2 compares the number of total mutations between male and female hosts from the different municipalities. There appears to be a significant difference between the mutation rate in the genome of the virus from female and male hosts. However, I also observed that the numbers of samples from female and male hosts are significantly different for many municipalities (Table 1). That would skew the comparison between the total count of mutations for male and female hosts, the total mutation count is divided by the number of samples for that host sex. This averaged count plotted in Figure 3 provides a better comparison and clearly depicts that average counts are approximately similar in male and female in municipalities.

Figure 4 compares the number of distinct SNPs observed in genome sequences from male and female hosts residing in different municipalities. Though the difference in the count ranges from 2 to 6 among the municipalities; this difference is small compared to the total number to make any conclusive remarks about this correlation. Please note, in this figure the SNPs are included regardless of the genes they belong to. Thus, the number of distinct SNPs is much higher compared to Figure 1.
From figures 2 through 4, we can conclude that sex or geographic location is not a factor for the extent of SARS-CoV-2 mutations observed in the genome sequences studied here.

Mutations and Number of Recorded COVID-19 Cases, Vaccination Rate

Figure 5 plots the count of total SNPs (throughout the genome sequence irrespective of the host sex) found in this study vs. the total cases per 100k inhabitants as of June 1, 2021, for all municipalities. Visually, there does not seem to be any correlation between the two. Similarly, Figure 6 illustrates that in the current study we did not find that vaccination rate has any effect on mutations in SARS-CoV-2. However, before reaching any conclusion, we have to note that the data available to us is very limited and also from a very short period of time. Moreover, the vaccination rates are also not high enough to have any impact on gene mutation. So it should not be interpreted that vaccination does not prevent mutations in SARS-CoV-2 virus.

Conclusion

In this study, I demonstrated the successful development of a pipeline to retrieve, process, and analyze genome-sequencing data from the database stored on the Internet. This pipeline which makes use of various bash and R scripts can be used to perform more studies with minimal changes. The analysis results of the current study show that number of distinct SNPs in highly mutating gene S of SARS-CoV-2 virus

is considerably more in male hosts compared to female hosts. However, if the total number of SNPs is considered, there is no correlation between the mutations and host sex or geographical location. Similarly, the data shows no impact of total COVID-19 cases per 100k inhabitants and vaccination rate on mutations. However, the limited genome sequencing data studied here prevents us from making any concrete conclusions.

Figures

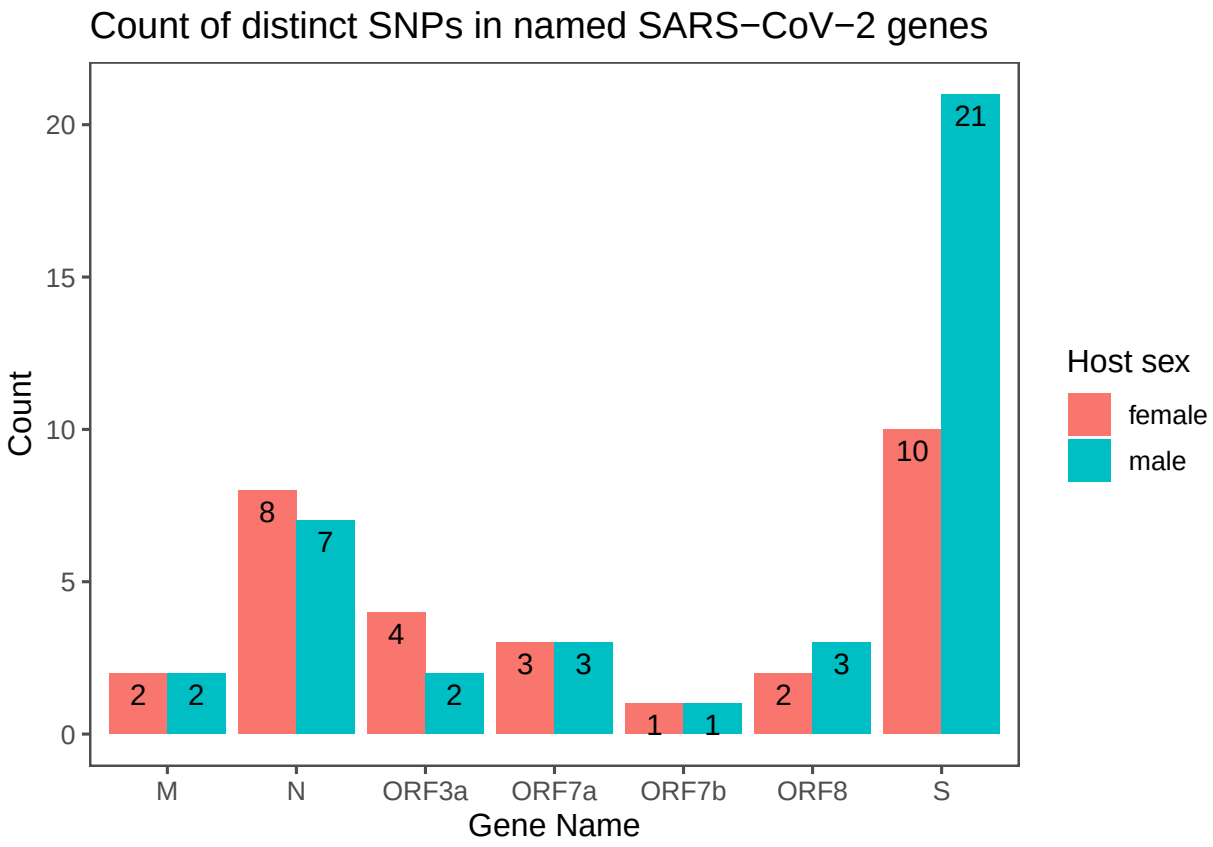

Figure 1: Significant difference in the count of distinct SNPs in gene S found in female and male hosts. Males seem to be more vulnerable.

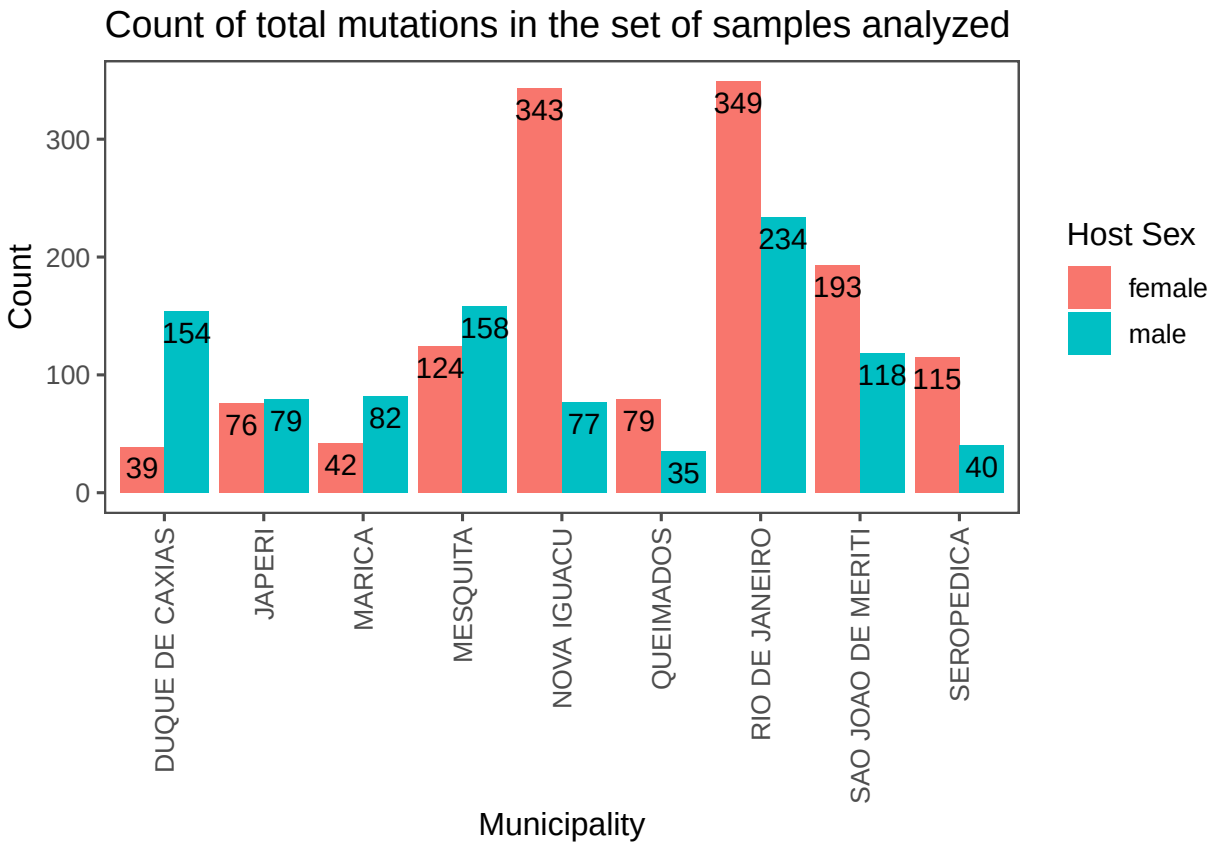

Figure 2: Significant differences between male and female hosts as well as municipalities. But this might not be a fair comparison due to the different number of samples analyzed.

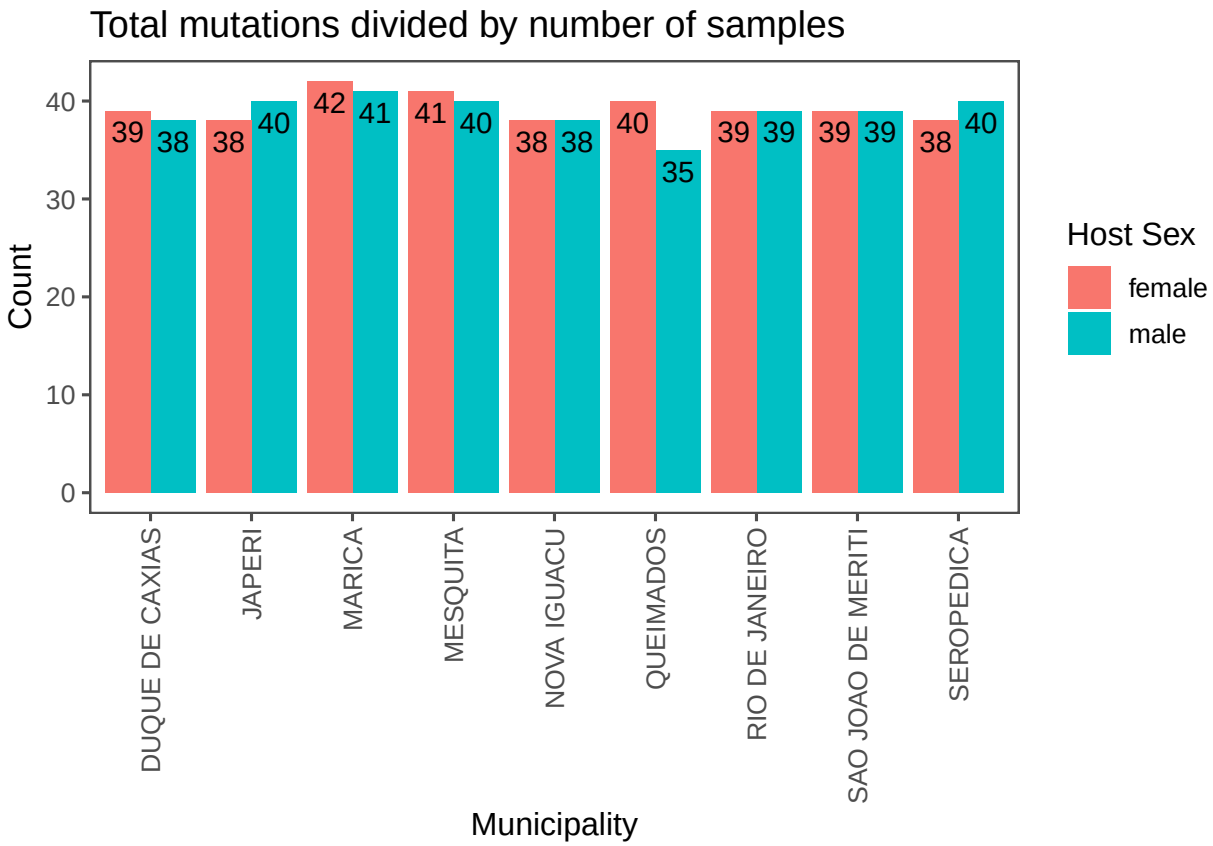

Figure 3: Better to compare average mutations per number of samples. This plot illustrates no correlation between mutations and geographic location or host sex.

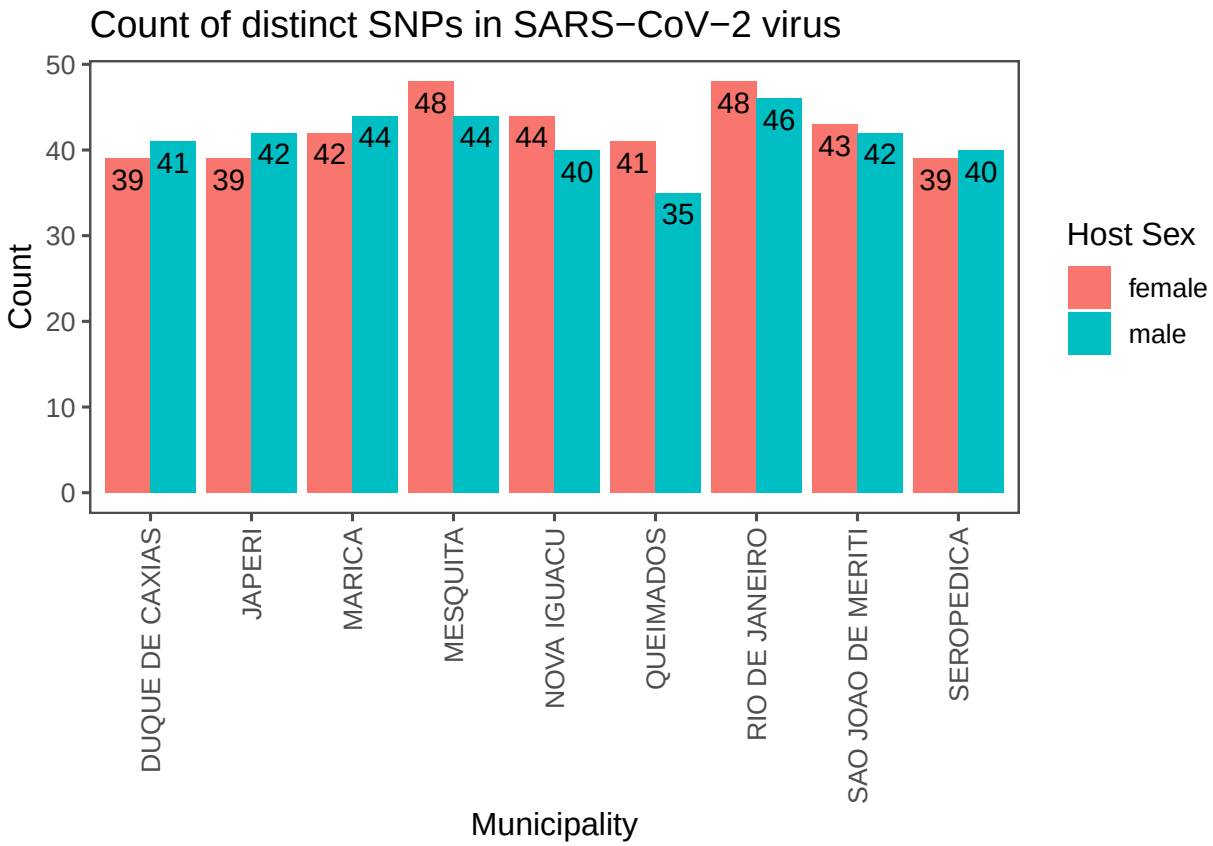

Figure 4: The count of distinct SNPs in the samples analyzed are similar irrespective of host sex and geographic location.

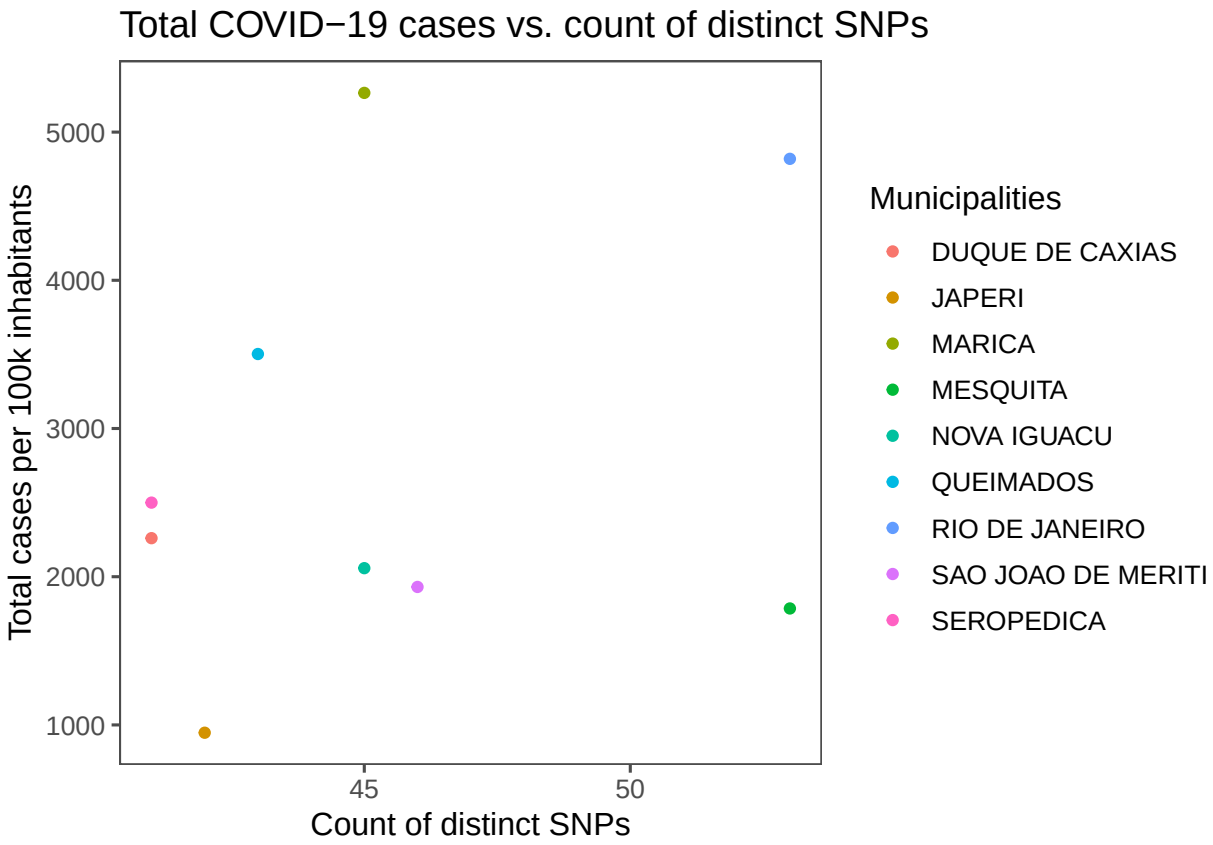

Figure 5: No correlation between COVID-19 spread and gene mutations.

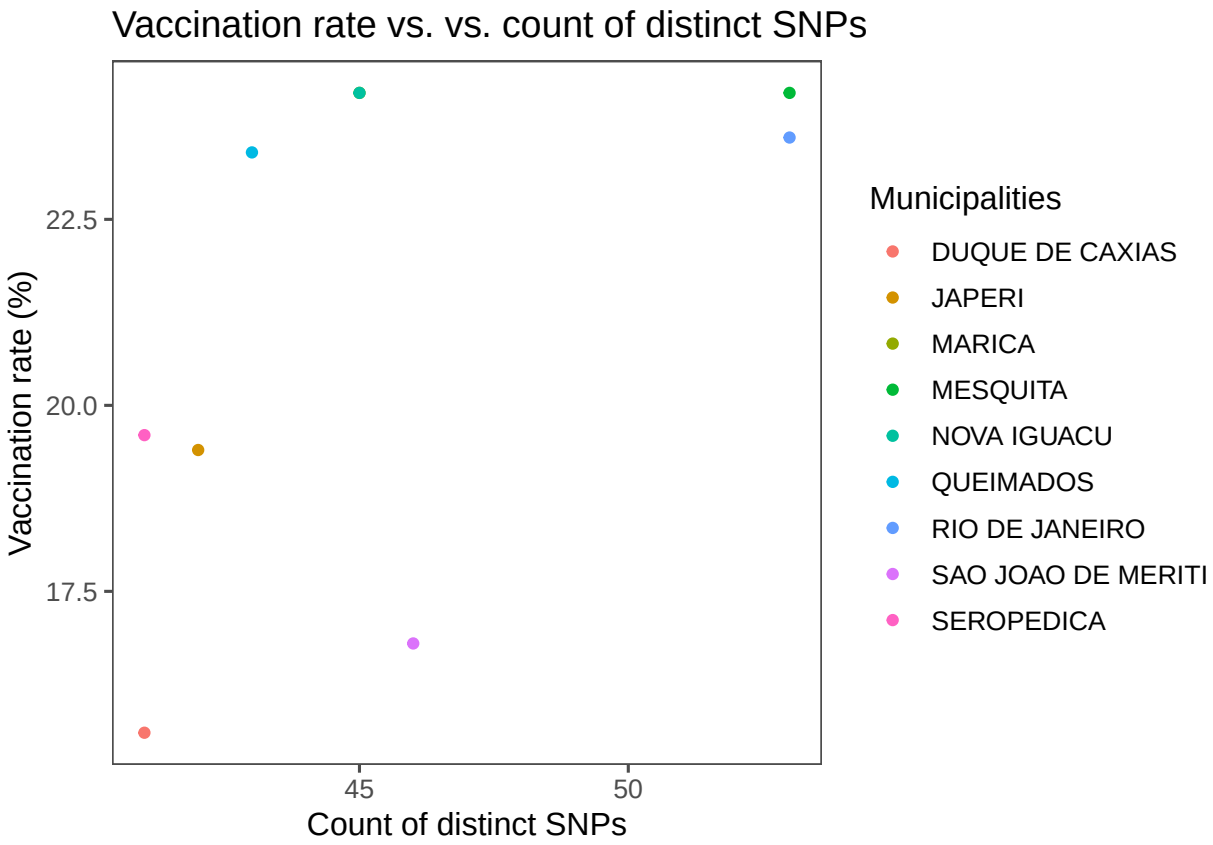

Figure 6: The vaccination rate doesn't seem to affect distinct SNP count.

Tables

Municipality	Total samples	Female samples	Male sample
RIO DE JANEIRO	15	9	6
NOVA IGUACU	11	9	2
SAO JOAO DE MERITI	8	5	3
MESQUITA	7	3	4
DUQUE DE CAXIAS	5	1	4
JAPERI	4	2	2
SEROPEDICA	4	3	1
MARICA	3	1	2
QUEIMADOS	3	2	1
ITABORAI	1	1	0
ITAGUAI	1	1	0
NITEROI	1	0	1
NOVA BASSANO	1	0	1
SANTO ANTONIO DE PADUA	1	0	1
XAMBIOA	1	1	0

Table 1: Number of genome sequences from different municipalities per host sex. I excluded the municipalities where only a single sample was available from the analysis.

Date	Municipality	Total Cases per 100k Inhabitants	Fully Vaccination Rate (%)
6/1/2021	DUQUE DE CAXIAS	2260.0487	15.6
6/1/2021	JAPERI	947.3546	19.4
6/1/2021	MARICA	5264.5705	24.2
6/1/2021	MESQUITA	1785.7143	24.2
6/1/2021	NOVA IGUACU	2057.6989	24.2
6/1/2021	QUEIMADOS	3502.7017	23.4
6/1/2021	RIO DE JANEIRO	4819.5124	23.6
6/1/2021	SEROPEDICA	2499.9702	19.6
6/1/2021	SAO JOAO DE MERITI	1930.9864	16.8

Table 2: The Total number of cases per 100k inhabitants and vaccinations rates in the municipalities under investigation as of June 1st, 2021 (Cota, 2020).

Gene Name	Sequence Start	Sequence End	Gene 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 3: 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.

Sources Cited

Arnold,J.B. (2021) Ggthemes: Extra themes, scales and geoms for 'ggplot2'.

Callaway,E. and others (2021) The mutation that helps delta spread like wildfire. *Nature*, **596**, 472–473.

Cota,W. (2020) Monitoring the number of COVID-19 cases and deaths in brazil at municipal and federative units level. *SciELOPreprints:362*.

Firke,S. (2021) Janitor: Simple tools for examining and cleaning dirty data.

Katella,K. (2021) 5 things to know about the delta variant. *Yale Medicine News*.

Sanchez,C.M. (2021) Experts see brazil as breeding ground for new covid-19 variants.

Souza,W.M. de *et al.* (2020) Epidemiological and clinical characteristics of the covid-19 epidemic in brazil. *Nature human behaviour*, **4**, 856–865.

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

Wickham,H. (2021) Tidyr: Tidy messy data.

Wickham,H. *et al.* (2021) Dplyr: A grammar of data manipulation.

Wickham,H. *et al.* (2021) Readr: Read rectangular text data.

Zhu,H. (2021) KableExtra: Construct complex table with 'kable' and pipe syntax.