Analysis of Shanghai Data from NCBI SARS-CoV-2 SRA Bioproject ID 627662

Mikaela Kuan

13 May, 2021

Background and Overview

COVID-19 has been a huge problem for a year and a half now. It has taken the lives of millions of people worldwide, devastating many friends and families. The first observed case of COVID-19 was seen in Wuhan, China from a 55-year-old man in November 2019 (Bryner, 2020). From there, the SARS-CoV-2 virus has spread from that one location to all over the world. The race to fight this disease and to vaccinate most of the population has been a struggle due to the virus's high spread rate and lack of vaccine production, respectively. As more people catch the disease, the virus is more likely to mutate from person to person. This can lead to problems such as increased COVID-19 cases and decreased efficiency of the vaccines. Countries, such as China, has taken quick action to reduce the number of cases in their country by implementing strict stay-at-orders. Unfortunately, recent events in India have led to a rise in the number of cases. In this study, I wanted to look at the hospitalization records and the type of mutations in the SNPs in Shanghai from January 2020 to February 2020 by using the R packages dplyr (Wickham *et al.*, 2020), ggplot2 (Wickham, 2016), vcfR (Knaus and Grünwald, 2017), and oxcovid19 (Guevarra *et al.*, 2020). This study shows that variants in Shanghai may have effects on the susceptibility of COVID-19.

Methods

NCBI SRA SARS-CoV-2 BioProject Data SNP Position Frequencies

NCBI SRA SARS-CoV-2 BioProject ID 627662 was taken from the NCBI website. The NCBI data was analyzed using the Makefile Dr. Zimmerman wrote to run the SRARunTable.txt file downloaded from the site and turned into vcf files using the vcfR package. The vcf files were analyzed using Dr. Zimmerman's code that gave a metadata set that was used to analyze the SNP positions and the quality scores of the top four most frequent SNPs, using dplyr. The SNP positions were filtered out for those with a frequency greater than five (Table 1).

NCBI SRA SARS-CoV-2 BioProject Data Quality Scores

The quality scores of the NCBI data were analyzed using the R package ggplot2 to visualize the frequencies of the quality scores. The quality scores came from the processed vcf files from Dr. Zimmerman's code.

NCBI SRA SARS-CoV-2 BioProject Data Collection Frequency

Using the NCBI data, the frequency of the samples from January to February was analyzed and graphed in R, using the packages dplyr and ggplot2. The sample frequencies came from the process vcf files from Dr. Zimmerman's code.

Oxford COVID-19 Database for Hospitalization Data

The R package oxcovid19 was used to obtain the COVID-19 database for hospitalization records. The hospitalization data for the city Shanghai and the country China was visualized using ggplot and dplyr. The oxcovid19 was downloaded from GitHub onto the R server.

Results

NCBI SRA SARS-CoV-2 BioProject Data SNP Position Frequencies

The SNPs positions were analyzed to see the frequencies of each SNP position in the NCBI SRA SARS-CoV-2 BioProject ID 627662 (Table 1). The most common SNPs were in positions 8782 and 28144 with both of them having a frequency of 34. The next most frequent SNPs were in positions 11083 and 29742 with both having a frequency of 11. The rest of the SNP positions had a frequency of less than 10. There were 295 SNPs found in this database; however, SNP positions with frequencies less than or equal to five were filtered out of the data to place more focus on the more frequent SNPs.

The top four most frequent SNPs—8782, 28144, 11083, and 29742—were analyzed for their changes in the nucleotides (Table 2). The SNP positions 28114, 8782, and 11083 had one mutation at their location, while the SNP position 29742 had two mutations at that one spot. The SNP position 28114 had a change from a thymine to a cytosine, 8782 had a change from cytosine to thymine, and 11083 had a change from guanine to thymine. The changes in positions 28114 and 8782 represents a transition mutation, while the changes in position 11083 represents a transversion mutation. The SNP position 29742 had two different changes: guanine to adenine (transition mutation) and guanine to thymine (transversion mutation). There was a greater number of transition mutations than transversion mutations when looking at the four common SNPs.

NCBI SRA SARS-CoV-2 BioProject Data Quality Scores

The quality scores of the SNPs from the whole dataset were analyzed to see the distribution of the quality scores (Figure 1). The distribution of the quality scores leaned towards the right with the majority of the scores being greater than 200. The scores that were less than 200 were not considered high quality.

The top four most frequent SNP positions—28144, 8782, 29742, and 11083—were analyzed for their quality scores (Table 3). The SNP positions of 28114 and 29742 had quality scores over 200 and SNP position 8782 had all but one of its samples with a quality score over 200. SNP position 11083 had three of its eleven samples with a quality score over 200. SNP position 11083 had many samples with low quality scores.

NCBI SRA SARS-CoV-2 BioProject Data Collection Frequency

The NCBI dataset was analyzed to visualize the number of collections obtained from January 25 to February 15 of 2020 (Figure 2). There was a huge spike in the number of collections on February 1; however, the number soon dropped after that day. The greatest amount from the data was 86 collections on February 1, while the lowest amount was 2 on February 12, and 14. There was missing collection days on January 26 and 27 and February 3 and 13.

Oxford COVID-19 Database for Hospitalization Data

The Oxford COVID-19 database was analyzed for the hospitalization records in China to see if China's hospitalization records match up with the collection records from the NCBI data. China's hospitalization was analyzed from January 25 to February 15 of 2020 (Figure 3). The hospitalization in China grew exponentially as the days moved forward.

The hospitalization records of Shanghai were analyzed to see if there was a difference in China's hospitalization records. The hospitalization records of Shanghai measured the days from January 25 to February 15 of 2020 (Figure 4). The hospitalization in Shanghai grew exponentially and then looks as if it is leveling off around February 8.

Discussion

Looking at the SNP position frequencies, the 28144 T>C change is a change from the amino acid leucine to serine (Yin, 2020). This mutation is a change from a nonpolar amino acid to a polar amino acid. This mutation can be seen in RNA primase for nsp8. In the SARS-CoV, the nsp8 protein interacts with the ORF6 accessory protein (Kumar *et al.*, 2007). Although not specifically in SARS-CoV-2, I would assume that the nsp8 protein interacts with the ORF6 accessory proteins. Therefore, this mutation may be useful for increasing the affinity for the proteins to interact with one another. The 8782 C>T change is a synonymous mutation (Yin, 2020). Although a synonymous mutation does not affect the amino acid, the nucleotide change can have affects later on if the virus mutates again. The 29742 G>A/T change is a change in the 3'UTR and may affect how the viral RNA is folded (Mishra *et al.*, 2020). Interestingly, the 29742 G>T change was found in mostly in Asia, while the G>A change was found mostly in North America and Asia (Chan *et al.*, 2020). The 11083 G>T change affects the ORF1 protein in SARS-CoV-2 from the amino acid leucine to phenylalanine. This mutation was identified in a superspreaders strain in SARS-CoV-2 and was also identified as the infectious strain in the Diamond Princess cruise ship (Lopez-Rincon *et al.*, 2020; Yang *et al.*, 2020). However, because the 11083 samples were not high quality (>200) and only three samples were high quality, it may be hard to determine if the 11083 variant could have been from Shanghai, China. The low quality scores of 11083 may be due to overlapping samples of different variants, making the 11083 samples not as confident in the quality.

With the NBCI sample collection frequency, the number of samples collected from January to February did not match the hospitalization records in China. I assumed that the collections would be similar to that of the hospitalization records. There is a spike in the collection data that may be due to how many people came to the hospital that day. There is less collection after the spike because there may have been less people or not enough room for hospitalization in that hospital. Because the China hospitalization records did not match the collection data, I wanted to see if the Shanghai hospitalization records would be any different. However, there was not much change in the Shanghai hospitalization trends from the China hospitalization data. The difference between the China and Shanghai hospitalization data was that the China hospitalization was more linear, while the Shanghai hospitalization was more sigmoidal. The sigmoidal trend may be due to Shanghai starting their level 1 public health emergency response on January 24. The lockdown may have led people to constantly be in close closed corridors with many people, thus the increase after January 24. If the 11083 variant is taken into account, then its infectiousness could also have an effect on the increased hospitalization.

Overall, this study shows that variants from the Shanghai samples taken from NCBI may have effects on the likelihood of acquiring the disease.

Figures

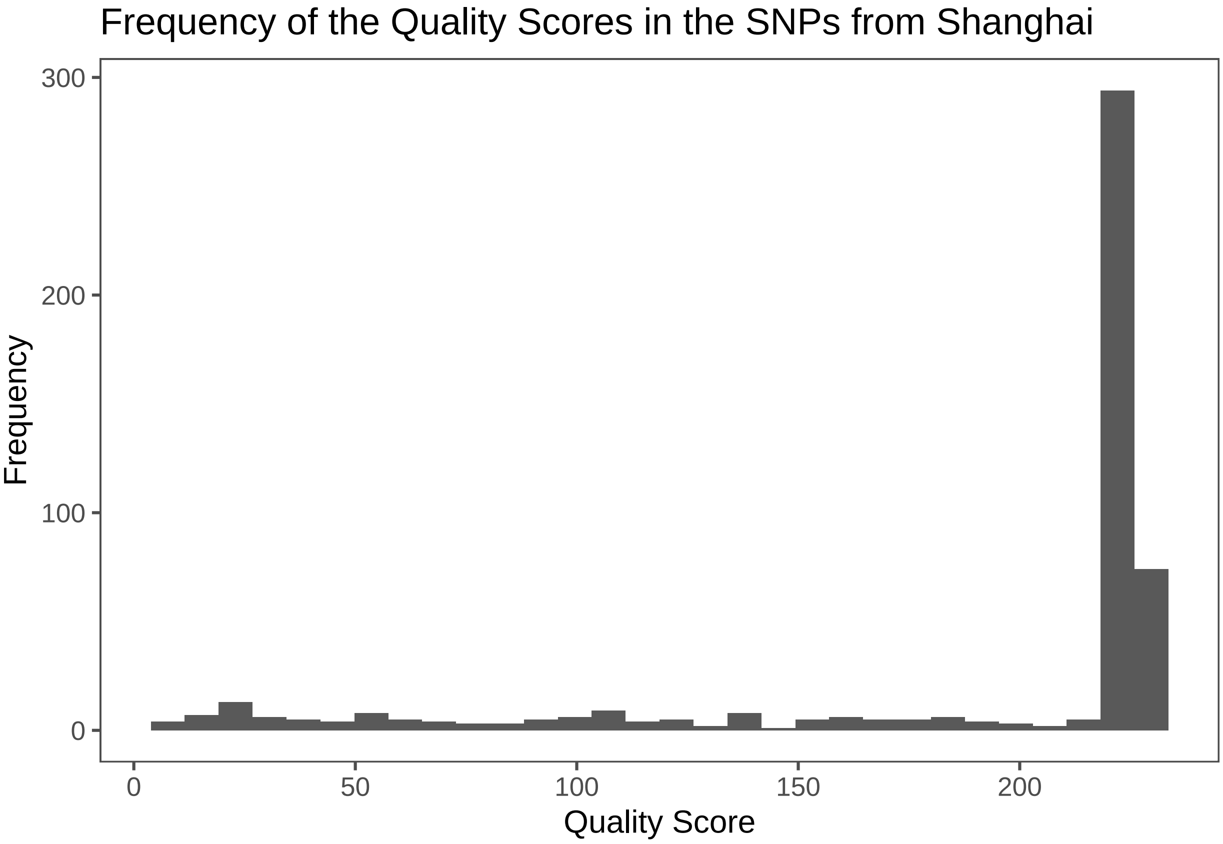

Figure 1. The frequencies of the quality scores from Shanghai data. Quality scores less than 200 are not of high quality.

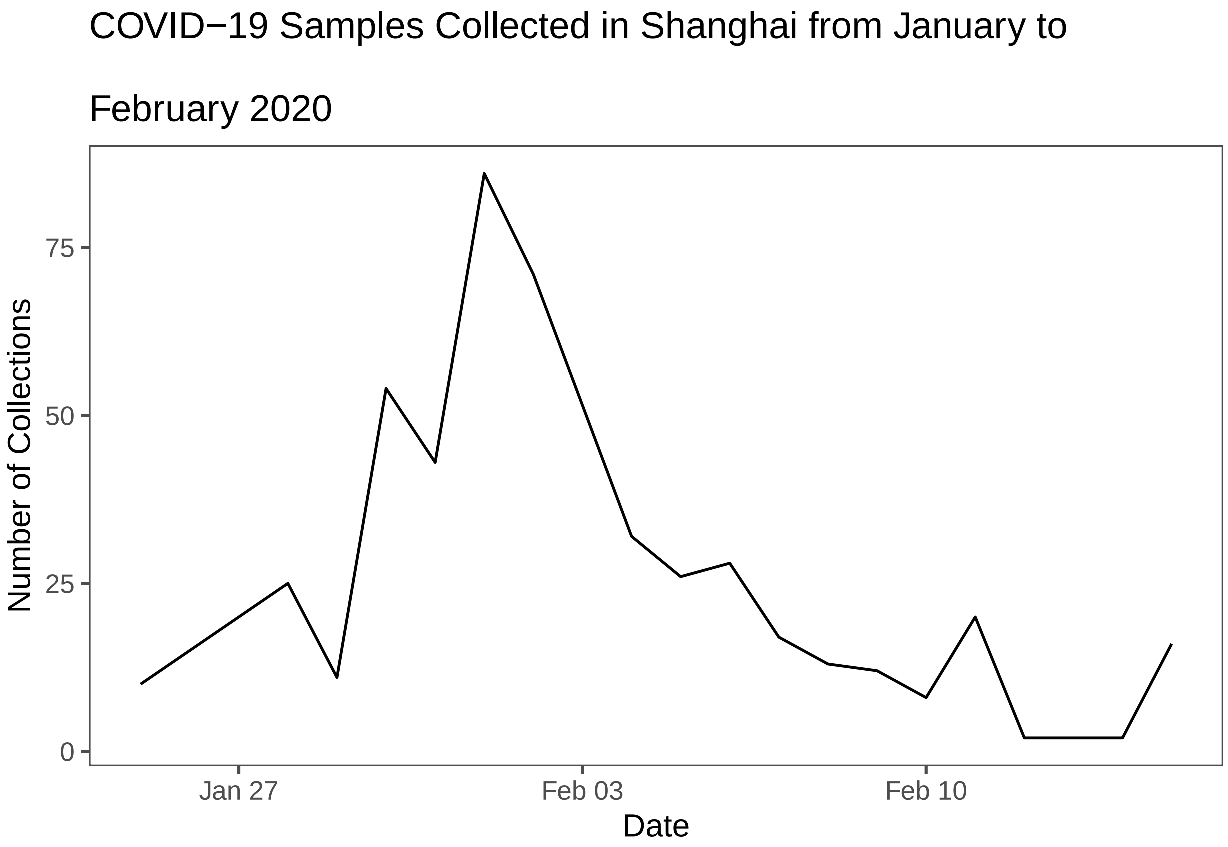

Figure 2. The frequency of samples collected in Shanghai from late January to early February of 2020. There is a peak in the beginning of February.

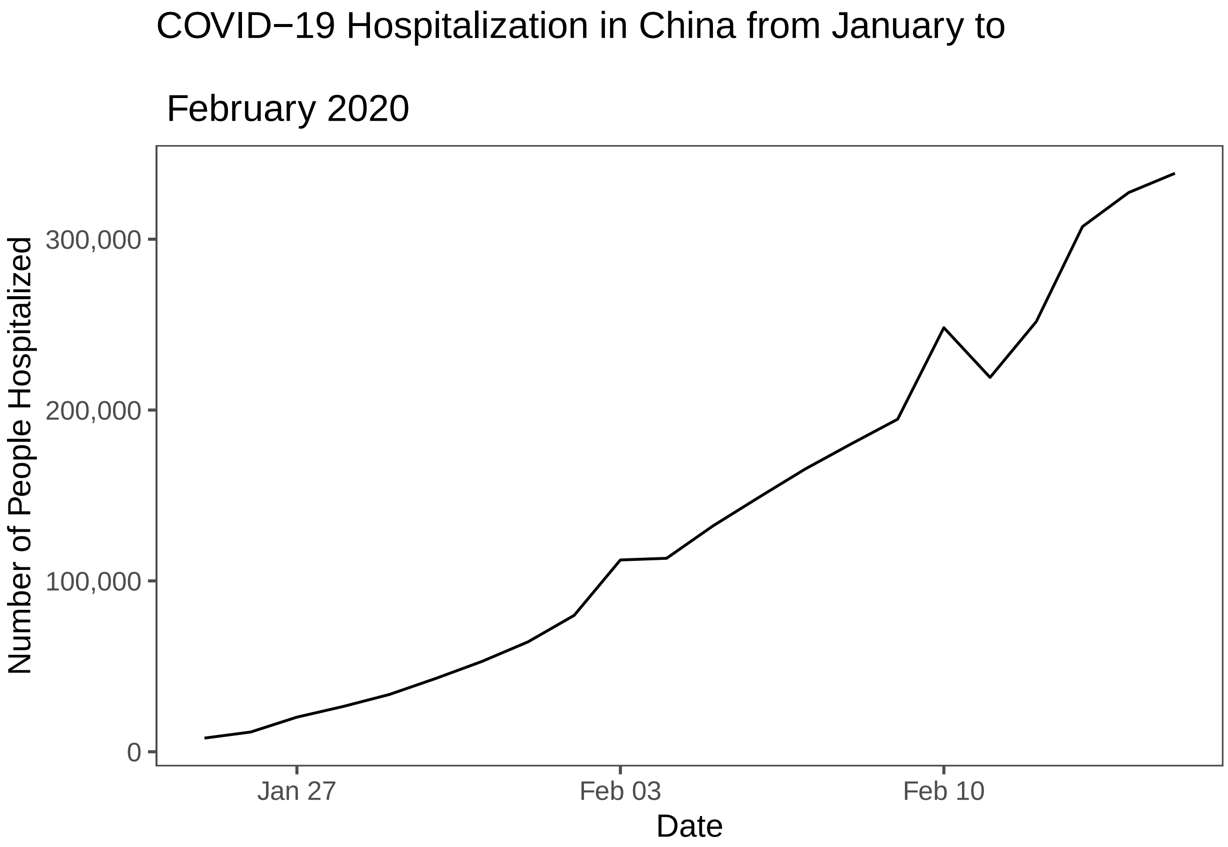

Figure 3. Hospitalization records in China from late January to early February of 2020. As the days went on, the number of people hospitalized increased.

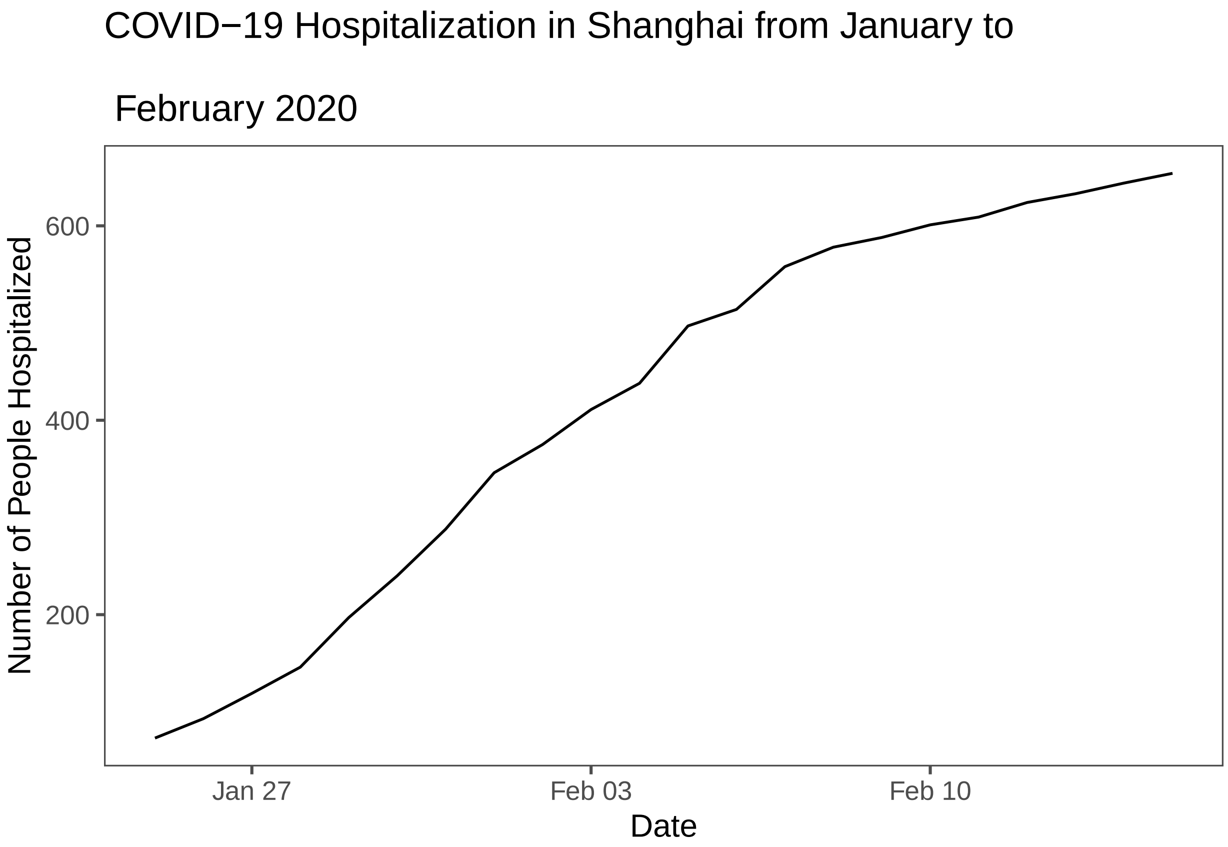

Figure 4. Hospitalization records in Shanghai from late January to early February of 2020. As the days went on, the numbers of cases grew.

Tables

Table 1. Frequency of each SNP position. The highest frequency was the SNPs in position 28144 and 8782. SNPs with frequencies less than five were removed from the table.

SNP Position	Frequency
241	6
1397	7
2293	8
3037	6
8782	34
11083	11
23403	6
24034	6
26144	9
26729	6
28077	6
28144	34
28688	7
29095	6
29449	6
29742	11

Table 2. Frequency of the SNPs and its different changes of the top four frequent SNPs. Transition mutations (pyrimidine to pyrimidine change, or purine to purine change) were more common than transversion mutations (pyrimidine to purine change and vice versa).

SNP Position	Reference Nucleotide	Altered Nucleotide	Frequency
28144	T	C	34
8782	C	T	34
29742	G	A	4
29742	G	T	7
11083	G	T	11

Table 3. The quality score of the top 4 most frequent SNPs in the Shanghai dataset. SNPs in position 28144, 8782, and 29742 had much better quality scores than SNPs in position 11083.

SNP Position	Quality Score	Frequency
28144	225	31
28144	228	3
8782	225	32
8782	228	1
8782	65	1
29742	225	7
29742	228	4
11083	105	1
11083	107	1
11083	117	1
11083	139	1
11083	150	1
11083	169	1
11083	194	1
11083	210	1
11083	225	2
11083	65	1

Sources Cited

Bryner,J. (2020) 1st known case of coronavirus traced back to november in china. *LiveScience*.

Chan,A.P. *et al.* (2020) Conserved genomic terminals of sars-cov-2 as coevolving functional elements and potential therapeutic targets. *MSphere*, **5**.

Guevarra,E. *et al.* (2020) Oxcovid19: An r api to the oxford covid-19 database.

Knaus,B.J. and Grünwald,N.J. (2017) VCFR: A package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, **17**, 44–53.

Kumar,P. *et al.* (2007) The nonstructural protein 8 (nsp8) of the sars coronavirus interacts with its orf6 accessory protein. *Virology*, **366**, 293–303.

Lopez-Rincon,A. *et al.* (2020) A missense mutation in sars-cov-2 potentially differentiates between asymptomatic and symptomatic cases.

Mishra,A. *et al.* (2020) Mutation landscape of sars-cov-2 reveals five mutually exclusive clusters of leading and trailing single nucleotide substitutions. *bioRxiv*.

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

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

Yang,X. *et al.* (2020) Genetic cluster analysis of sars-cov-2 and the identification of those responsible for the major outbreaks in various countries. *Emerging microbes & infections*, **9**, 1287–1299.

Yin,C. (2020) Genotyping coronavirus sars-cov-2: Methods and implications. *Genomics*, **112**, 3588–3596.