Identifying SARS-CoV-2 mutations in Beijing and transmission rates in China in 2020

Sarah Gao

November 19, 2020

Background and Overview

In late 2019, the first cases of coronavirus disease 19 (COVID-19) caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) virus were reported in Wuhan, China in December 2019. Since then, the virus has spread rapidly around the world and has resulted in more than 1.4 million deaths worldwide as of November, 27th 2020 (COVID-19 data repository by the center for systems science and engineering (csse) at johns hopkins university, 2020). Researchers around the globe have identified key mutations that are predominant in different locations, suggesting that the virus has been evolving over time to form location-specific strains (Pachetti, 2020). Some have hypothesized that such mutations could cause evolutionary advantages, such as increased transmission rates (Korber, 2020).

The goal of this report was to identify mutations of the SARS-CoV-2 virus found in samples taken from positive patients in a Beijing hospital, compare them with known mutations identified by previous researchers, and analyze epidemiological data on China's response to the virus at a national level. I also examined government response data as how a country reacts to the pandemic can certainly affect transmission rates as well.

I used a pipeline written in bash to process the sequence data and identify variants from the reference Wuhan genome. Then, I used R scripts to tally commonly found variants, analyze epidemiological data such as total confirmed cases and effective reproduction rate, and understand the Chinese government's response to the pandemic over time.

Previous studies showed that a mutation in the gene encoding the S spike protein has created a SARS-CoV-2 variant that has become the most prevalent form in many geographies, suggesting that this strain could have an evolutionary advantage (Korber, 2020). While there is no definitive understanding of how known mutations affect the virus' transmission or infection capacity, it is certainly an important area of study to understand how the virus will change over time, and in turn, how we may more effectively combat it. Analyses of this Beijing dataset show that there were no more S gene mutations than one would expect assuming equal rates of mutation across the genome. However, there are many other factors that can affect how a region copes with the ever-changing COVID-19 pandemic other than the potential outcomes of viral mutations, including population and government responses. I found that the number of cases across the country as a whole has reduced dramatically since spring of 2020, despite fluctuating effective reproduction rates, as the Chinese government has reacted quickly and strictly in response to spikes.

Methods

SARS-CoV-2 Sequence Data

I downloaded the Beijing dataset from the NCBI Database on Wednesday November 18th, 2020. This dataset was found originally by looking through NCBI's SRA BioProjects and filtering for Illumina sequences only. The 102 samples in the dataset were taken from SARS-CoV-2 positive patients at Beijing Ditan Hospital Capital Medical University between January 29th and April 17th, 2020 via three isolation sources: feces, pharyngeal swab, and sputum (Figure 1) (Du, 2020).

I then downloaded the SARS-CoV-2 Wuhan reference genome from NCBI and the respective annotation `gff`. When downloading the dataset, an additional, smaller `fastq` file was included without the usual "_1" or "_2" suffix. For the purposes of this report, these have been removed along with all reverse reads.

Sequence Data Quality

I ran FastQC on each downloaded `fastq` file to get a summary of the quality of each sample (Andrews). Overall, all the samples passed `fastqc` per base sequence quality and per sequence quality checks (Images 1-3). All samples had either a sequence length of 150bp or 76bp. Given the uniform distribution of the FastQC sequence length plot, one would assume that some processing had already occurred upstream to trim sequences of other lengths. It's possible, for example, that the sequencing machine itself threw out fragments during processing so as to keep only full-length reads.

Data Processing

I used Trimmomatic on each downloaded `fastq` file to throw out bad sequences and trim areas of poor quality (Bolger, 2014). For faster turnaround, I used TrimmomaticSE for single ends rather then using a paired end approach. For my parameters, I used 4 computing threads, converted quality scores to Phred-33, had a leading and trailing value of 5 bases, and a sliding window of size 8 with a minimum phred score of 25. Reads below a length of 100bp were dropped. Of the 102 sequences, 73 did not clear the Trimmomatic step due to having a length of 76bp and I proceeded with 29 sequences.

I then aligned each read against the reference genome using the Burrows-Wheeler Aligner (Li, 2013). These aligned `bam` files were then converted to `sam` and then sorted using samtools (Li, 2009). I ran `bcftools mpileup` to determine read coverage of positions in the genome per base and then `bcftools call` to call the single nucleotide polymorphisms (SNPs) for each file. Finally, I filtered each file to remove short variants and retain only major variants.

R Analyses

I used R scripts to tally observed variants in the sample set as well as to calculate expected SNP counts (Figure 2). Expected SNP counts were calculated by assuming equal rates of mutation across all the genes and using individual gene length proportional to the total length of all the identified genes combined.

I used the `vcfR` package to load and clean up the vcf data (Knaus and Grünwald, 2017). I conducted much of my R analyses using the `dplyr` package (Wickham, 2020). Plots were created using `ggplot2` (Wickham, 2016). For the Beijing epidemiological data, I connected to an API that pulls data from Oxford's COVID-19 database (Guevarra, 2020)(Mahdi, 2020). For the nationwide epidemiological data for China, I used Our World in Data's COVID-19 dataset (Appel, 2020).

Results and Discussion

Beijing Sequence Data

Number of Mutations

By mapping the variants to the gene locations as identified by the Wuhan reference genome, I found that there were more SNPs in the samples in the N and S gene areas (Figure 2). However, if we examine the lengths of these genes, we find that the N and S genes are the longest (Table 1). To understand if there were indeed more mutations in these regions, I calculated the expected number of variants within each gene assuming equal rates of mutation based on their length proportional to the total length of all identified genes (Figure 2). From this, we can see that observed SNPs within the N gene is indeed higher than what we would expect with equal mutation rates. The S gene, on the other hand, actually had fewer variants than expected. Significance was not been calculated for these values.

Mapping Mutations

Previous studies have identified positions at which mutations have been commonly found in samples taken around the world. Some have been found in specific geographic regions, suggesting regional strains with distinctive mutation patterns (Pachetti, 2020).

One strain in particular has emerged in prominence around the world, usurping the previously predominant D614 strain. This increased prevalence arose at multiple geographic levels globally, suggesting that this variant confers an evolutionary advantage over the previously established strain and has been positively selected for. This variant has a change within the Spike protein encoding gene, changing amino acid D614 to G614 through a A-to-G mutation at position 23403. G614 variants almost always also have 3 other distinguishing mutations within coding regions: C14408T within the ORF1b gene that causes an amino acid change in RNA-dependent RNA polymerase (RdRp P323L), C3037T within ORF1a which is silent, and C241T in the 5' UTR (Korber, 2020)(Yang, 2020)(Dorp, 2020). Table 2 shows these common mutations and which genes they're found in.

I mapped the Beijing dataset's variants to these mutation positions and found that there were mutations characteristic of the D614G strain (Figure 3). Interestingly, the mutations characteristic of the D614G strain (241, 3037, 14408, and 23403) appeared with nearly the exact same frequency, with the exception being the mutation at 23403 having 6 occurrences rather than 5. This lends credence to the idea that these mutations occur in tandem and may serve as markers of the strain. That these have been found in samples from China also suggests that this variant may have become more prevalent despite the original strain of the Wuhan reference genome having already been previously established in the region. This is further supported by the work by Korber et al. examining the increased frequency of the D614G variant in Mainland China over time (Korber, 2020).

Epidemiology

Transmission Rates

While Korber et al. hypothesized that the increased frequency of the G614 strain was due to its higher rate of transmission, there is still no clear evidence that this variant is any more infectious or transmissible than the D614 strain (Grubaugh, 2020)(Dorp, 2020). It is evident, however, that transmission rates are affected by the behavior of the regional governments and populace.

China was the first country to report cases of SARS-CoV-2 infections and had spikes in cases much earlier on before the virus spread to other countries. In Figure 4, we can see how Beijing experienced a large spike in confirmed cases in January and February, a smaller one in March, and then another large one in June before plateauing (Mahdi, 2020). The large spike in January-February followed a national trend based on daily new case rates in China (Figure 5) (Appel, 2020). China as a whole experienced a smaller spike in July as well.

Interestingly, when we examine the effective reproduction rate (R) of the virus throughout the year, the spikes in R seem to correlate with the spikes in case numbers we found in Beijing and across the whole country (Figure 6). Since the last spike in July, the R value has remained mostly under 1, correlating with the gradual decrease in case numbers seen in August.

Government Response

The Chinese government's response to the pandemic may help to partially explain China's current low case numbers. The government put relatively severe restrictions on movement from late January to early February. Other researchers have shown that within-city movement and R were highly correlated during this time, suggesting a high efficacy of this response (Ainslie, 2020). A strong and prolonged government response throughout the summer may also partially account for the drop in new cases as well as the prolonged period with $R < 1$. This lends support to the idea that transmission of the virus throughout a region can be mitigated by a stronger government response and adherence to social distancing practices.

Images

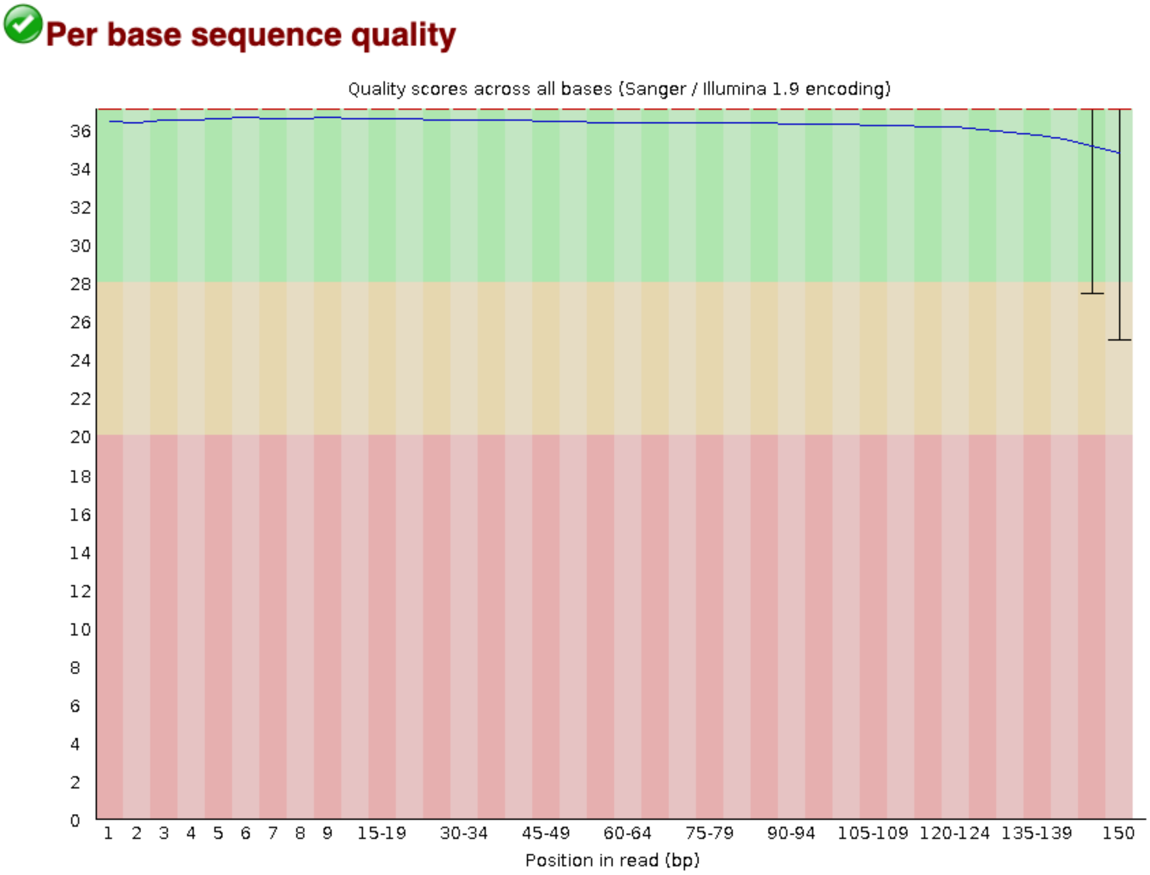

Image 1: An example of one sample's FastQC quality report showing per base sequence quality.

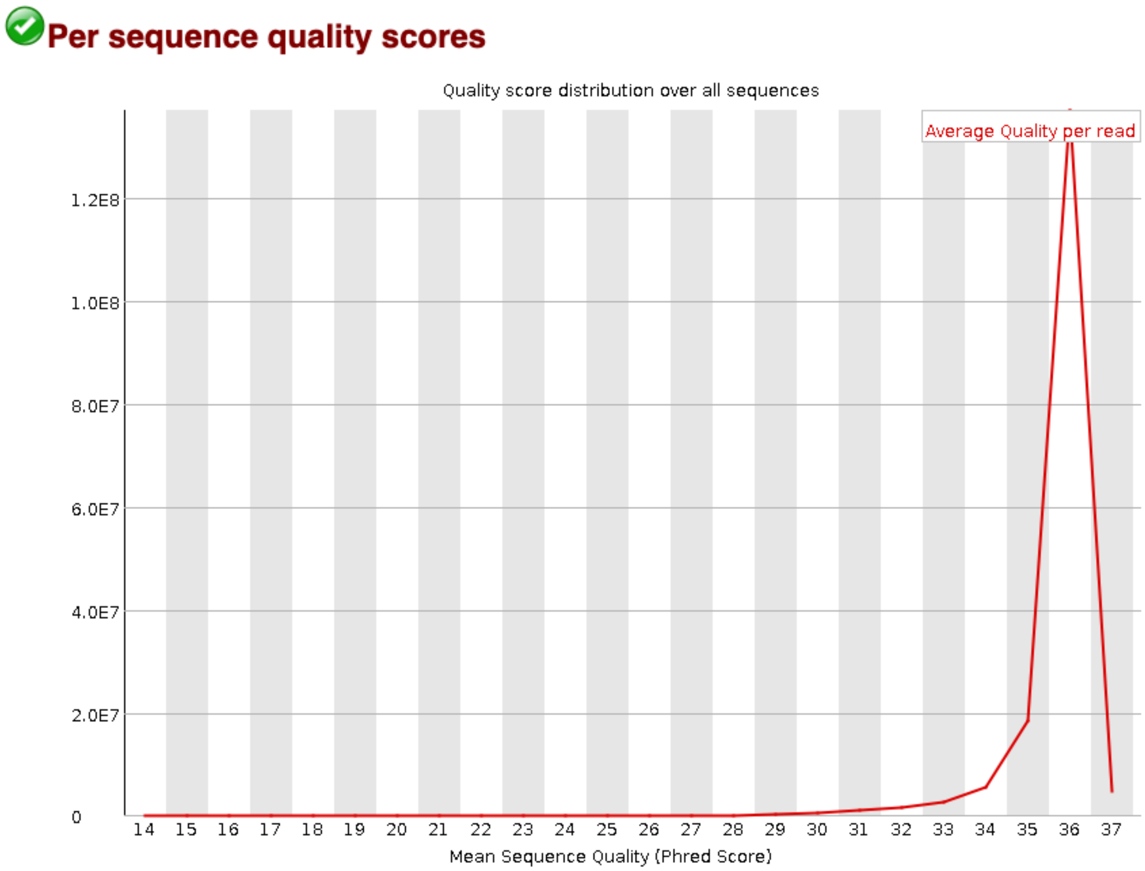

Image 2: The FastQC per sequence quality scores of the same sample.

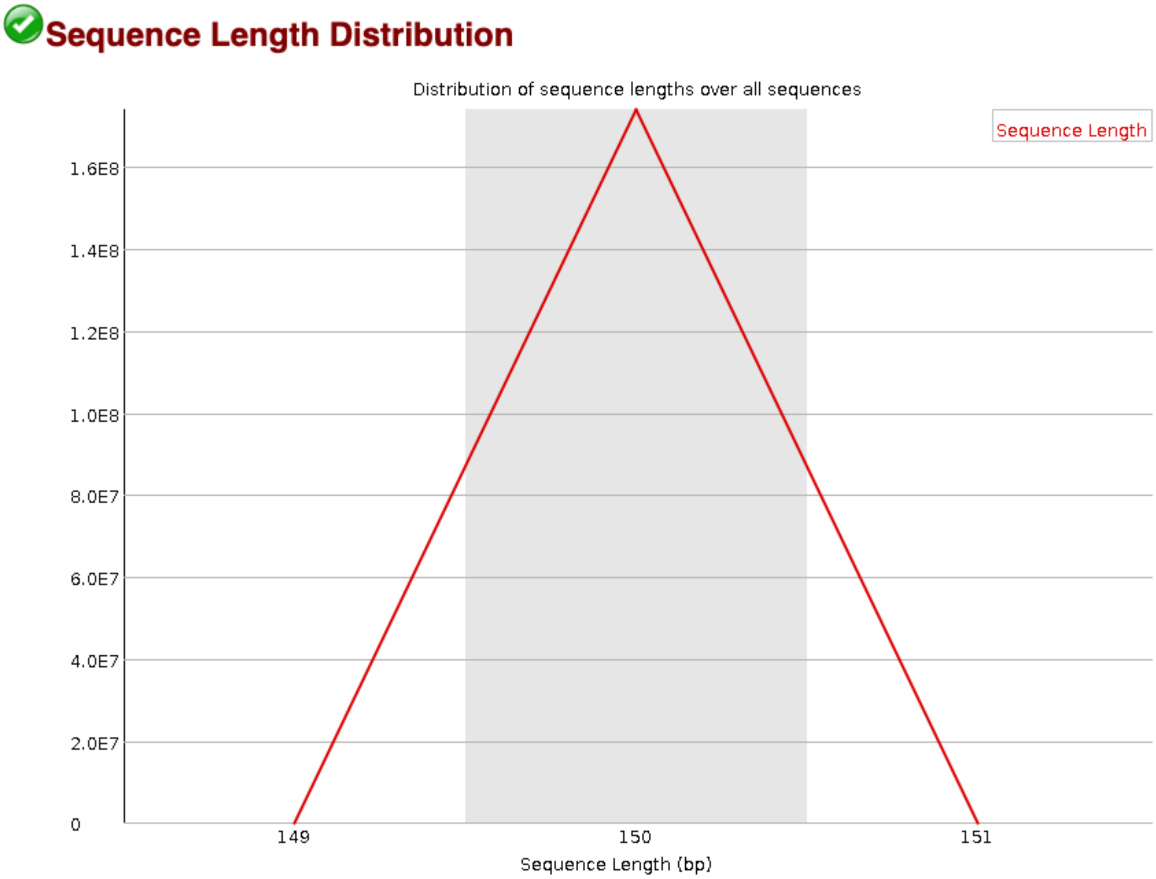

Image 3: The sequence length distribution report of the same sample.

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: Gene names, locations, and lengths in the SARS-CoV-2 genome.

Mutation Position	Gene Name
241	5’ UTR
1397	ORF1a
2891	ORF1a
3036	ORF1a
3037	ORF1a
8782	ORF1a
11083	ORF1a
14408	ORF1b
17746	ORF1b
17747	ORF1b
17857	ORF1b
18060	ORF1b
23403	S
26143	ORF3a
26144	ORF3a
28144	ORF8
28881	N

Table 2: Mutation positions and the genes they are found within (Pachetti, 2020)(Korber, 2020)(Yang, 2020). ***Note:*** Positions 3036, 17746, and 26143, which were identified by Pachetti et al., were not identified in the analyses by van Dorp et al. Instead, adjacent position 3037, 17747, and 26144 were identified, with a relatively large number of isolates having the mutation with high quality reads, which have been included here.

Figures

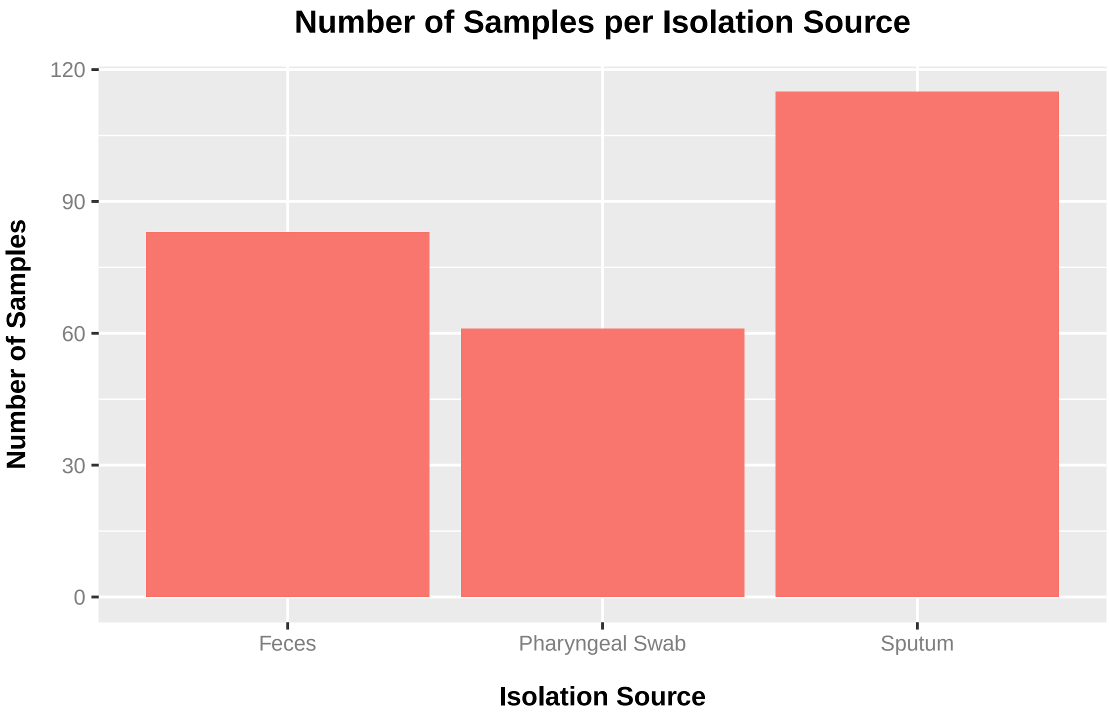

Figure 1: Samples were isolated from, in order from most to fewest, sputum, fecal, and pharyngeal swab sources.

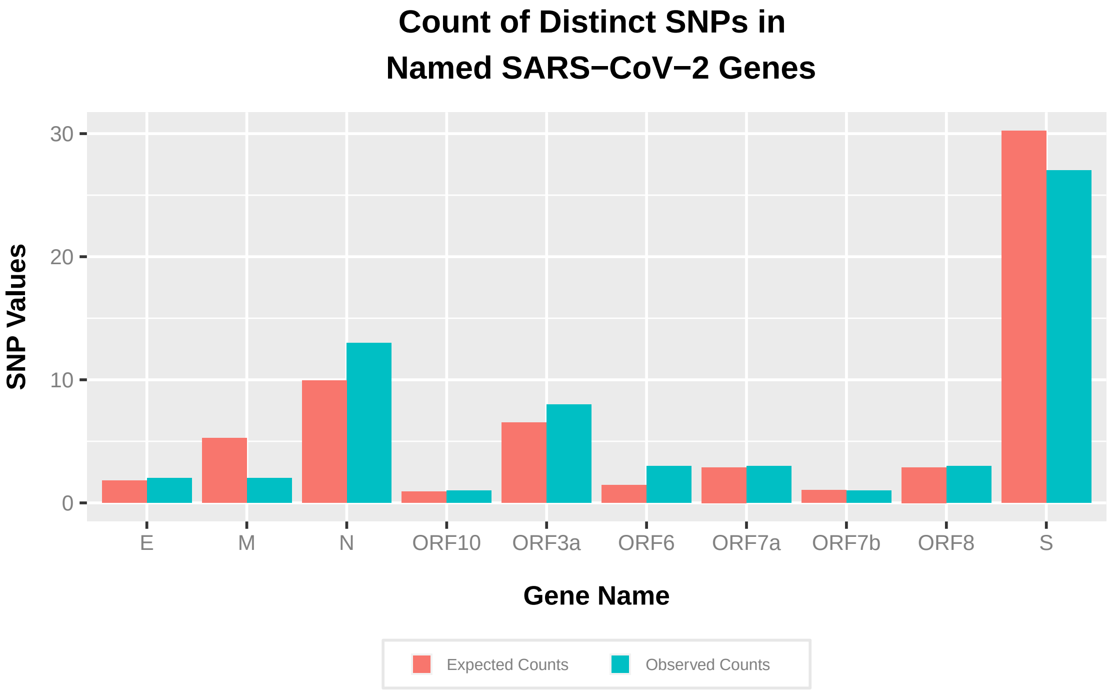

Figure 2: Comparison of expected SNP counts and observed SNP counts based on equal rate of mutation across all genes.

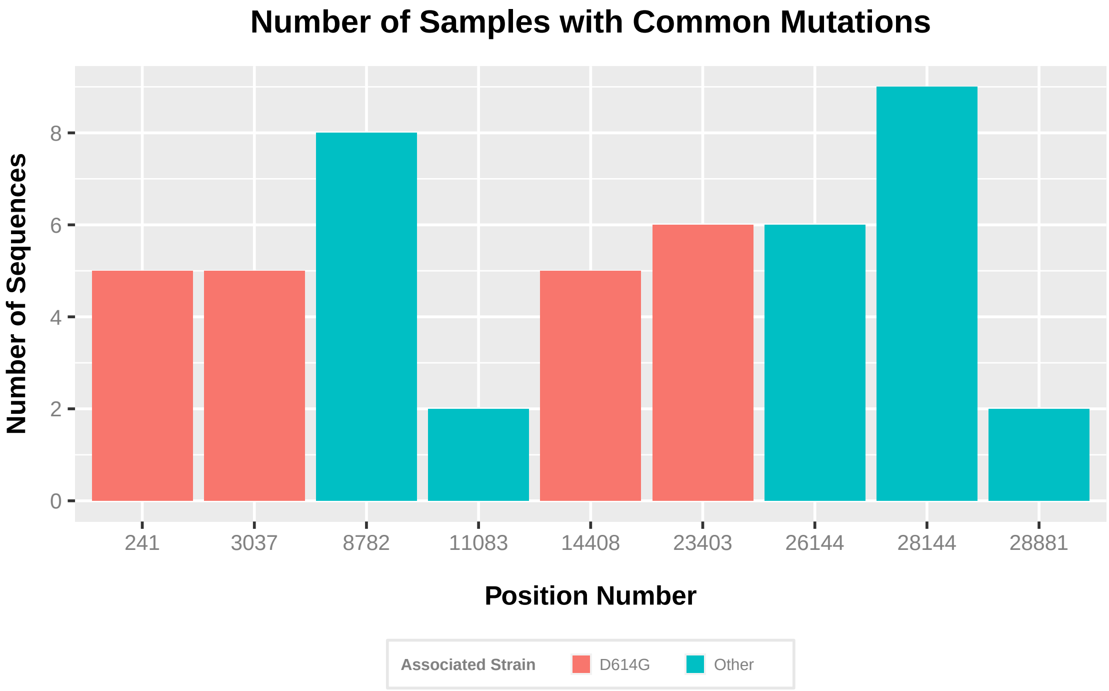

Figure 3: Number of samples from this specific Beijing data set with common mutations found in samples globally. ***Note:*** Positions 3036, 17746, and 26143, which were identified by Pachetti et al., were not identified in the analyses by van Dorp et al (Pachetti, 2020)(Korber, 2020). Instead, adjacent position 3037, 17747, and 26144 were identified, with a relatively large number of isolates having the mutation with high quality reads, which have been included here.

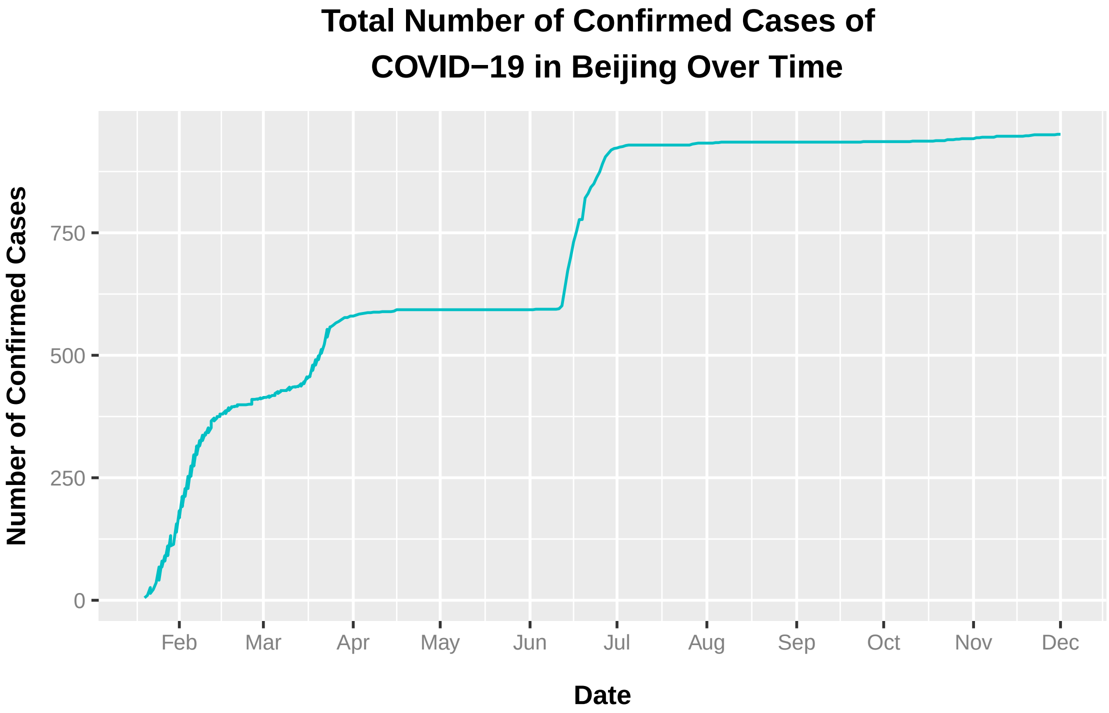

Figure 4: Total confirmed cases in Beijing over time, starting from January 20th, 2020.

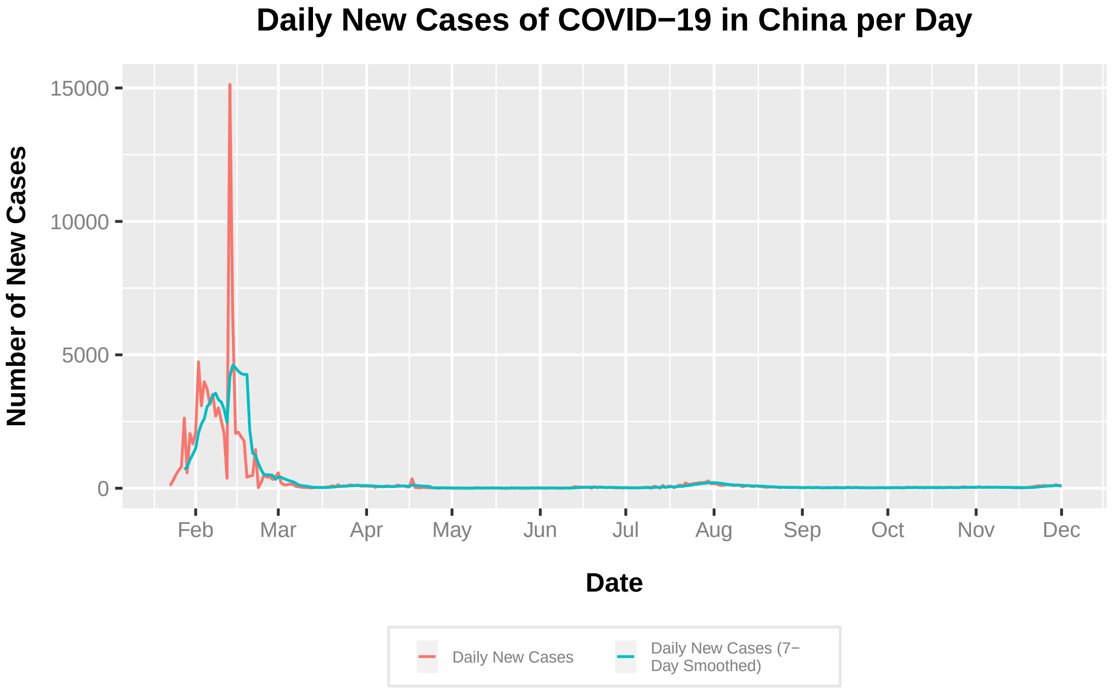

Figure 5: New daily confirmed cases over time in China. The red line represents the raw number of daily new cases whereas the blue line represents daily confirmed cases with 7-day smoothing.

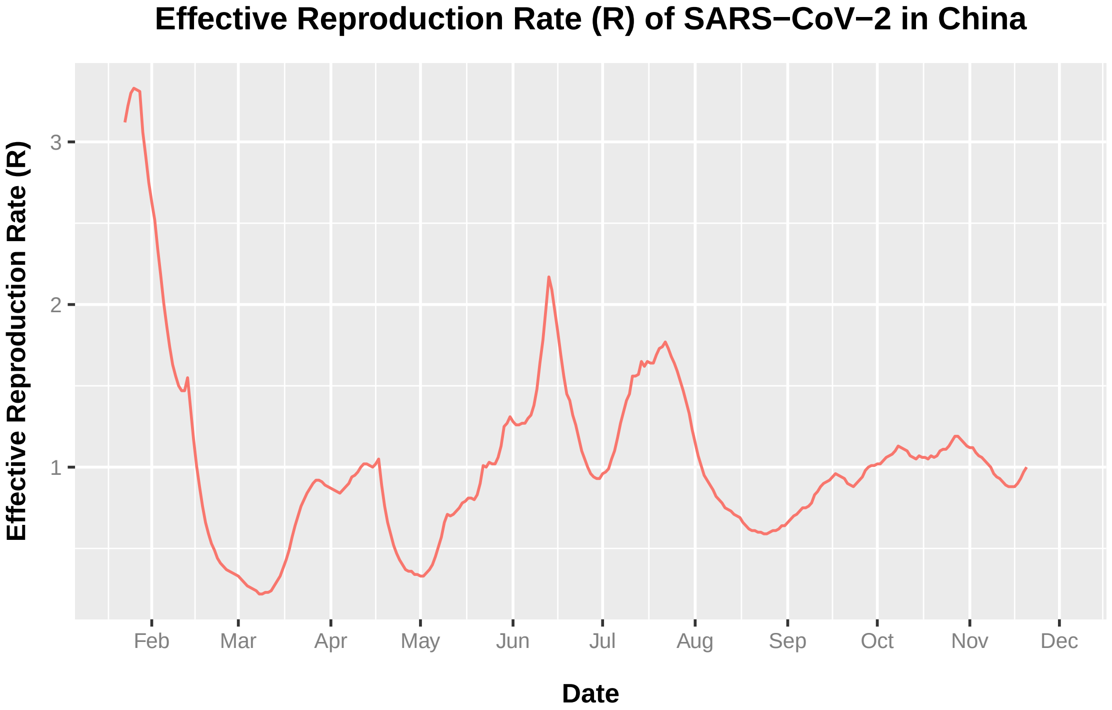

Figure 6: Effective reproduction rate of the SARS-CoV-2 virus in China over time. R indicates how many new infections one infected person causes on average. If R is below 1, the number of cases will gradually decrease whereas if it's greater than 1, cases will increase.

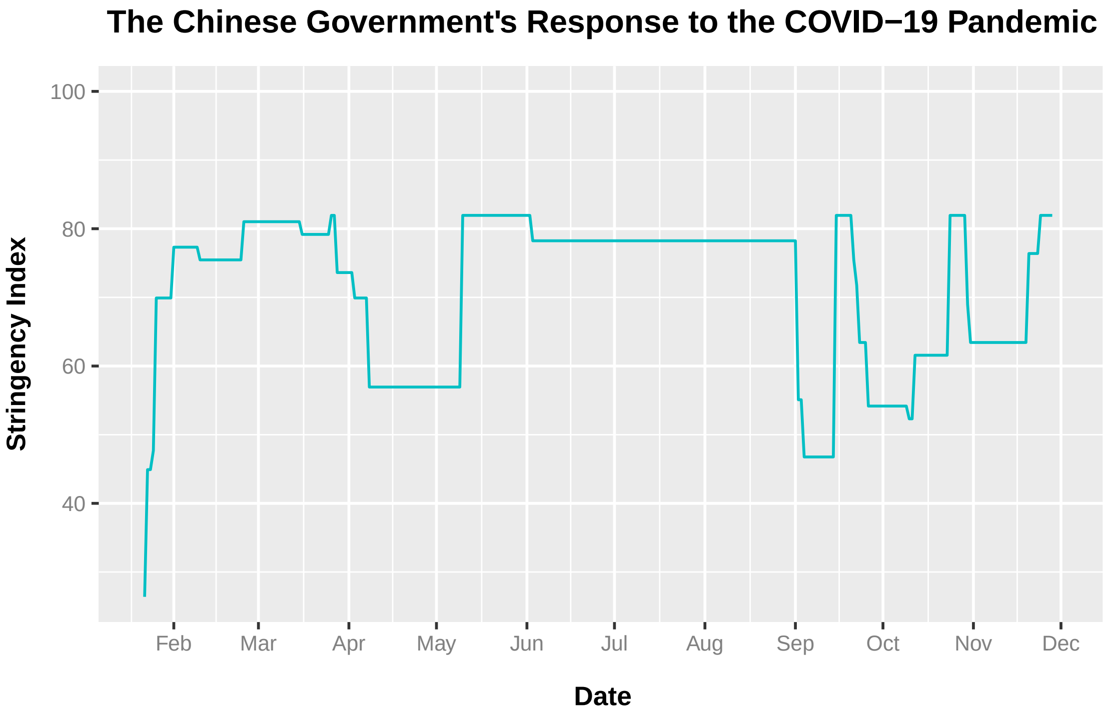

Figure 7: The Chinese government's response to the pandemic measured by a stringency index over time. This index is compositely based on 9 response indicators, including school and workplace closures, contact tracing, and face coverings, and is scaled from 0 to 100 with 100 being the strictest response (Appel, 2020).

Sources Cited

Ainslie,K. et a. (2020) Evidence of initial success for china exiting covid-19 social distancing policy after achieving containment. *Wellcome open research*, **5**.

Andrews,S. FastQC: A quality control tool for high throughput sequence data [online].

Appel,C. et a. (2020) Our world in data covid-19 dataset.

Bolger,A. et a. (2014) Trimmomatic: A flexible trimmer for illumina sequence data.

COVID-19 data repository by the center for systems science and engineering (csse) at johns hopkins university (2020) Johns Hopkins University.

Dorp,L. et a. van (2020) No evidence for increased transmissibility from recurrent mutations in sars-cov-2. *Nature Communications*, **11**.

Du,P. et a. (2020) Genomic surveillance of covid-19 cases in beijing. **11**.

Grubaugh,N. et a. (2020) Making sense of mutation: What d614g means for the covid-19 pandemic remains unclear. *Cell*, **182**, 794–795.

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

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

Korber,B. et a. (2020) Tracking changes in sars-cov-2 spike: Evidence that d614g increases infectivity of the covid-19 virus. *Cell*, **182**, 812–827.e19.

Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with bwa-mem.

Li,H. et a. (2009) The sequence alignment/map (sam) format and samtools.

Mahdi,A. et a. (2020) Oxford covid-19 database: A multimodal data repository for better understanding the global impact of covid-19 University of Oxford.

Pachetti,M. et a. (2020) Emerging sars-cov-2 mutation hot spots include a novel rna-dependent-rna polymerase variant. *Journal of Translational Medicine*, **18**, 1–9.

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

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

Yang,H. et al (2020) Analysis of genomic distributions of sars-cov-2 reveals a dominant strain type with strong allelic associations. *Proceedings of the National Academy of Sciences*, **117**, 30679–30686.