Mutations in SARS-CoV-2 and its Effect in Viral Replication and Entry

Lewis Hendrianto

06 December, 2022

Background and Overview

The World Health Organization (WHO) declared coronavirus 19 (COVID-19), a pandemic on March 11, 2020 (Katella, 2021). Severe acute respiratory syndrome coronavirus 2, also known as SARS-CoV-2, is a virus from the coronavirus family. SARS-CoV-2 causes this disease. COVID-19 is a respiratory disease primarily transmitted through aerosols. Preventative and therapeutic approaches are in development to combat this virus, and an mRNA vaccine has become widely available for the public (Wang *et al.*, 2020). Before understanding the function of the virus, it is essential to understand the structure of a virus. In general, each virus contains a viral genome that resides in a shell called the capsid. Additionally, the coronavirus contains a lipid membrane envelope that surrounds the capsid. Spike glycoproteins surround the lipid envelope, functioning as a gateway to enter the target host. The spike proteins bind to a specific receptor to the host, for SARS-CoV-2, its target is the Angiotensin-converting enzyme 2 (ACE2). Upon binding, the virus will permeate the host cell's membrane, inserting the viral genome. The viral genome gets replicated and expressed, producing viral proteins infect other cells ultimately causing COVID-19 (Intro to viruses (article) | viruses, n.d.). Moreover, the spike proteins could undergo mutations, affecting is permeability and transmissibility towards the target host. Consequently, the SARS-CoV-2 has faced several mutations, acquiring SNPs in multiple positions. With the data collected through BioProject PRJNA793894, it has been identified that the S:665Y substitution causes these effects. This report will take the data and examine how the SNPs affected the viral entry and replication of the newly obtained variants. On November 8, 2021 the first confirmed sample of the Omicron variant is discovered in South Africa. In less than a month, the variant of concern (VOC) made its way across the world to the United States on December 1st, 2021. Rapidly, the Omicron variant became one of the most dominant strains in the world because of the SNP in the spike protein (Omicron variant: What we know so far about this covid-19 strain, 2022). As of November 2022, Omicron is responsible for over 97% of COVID-19 transmissions according to the Centers for Disease

Control and Prevention. As a result, this VOC takes in account for the majority of COVID-19 cases and deaths since its emergence (CDC covid data tracker, n.d.). The primary goal of this analysis were to identify the position of each SNP and interpret how they enhanced the transmissibility of this VOC compared to the original variant. Although the data collected from the biosample was from June 2021 through November 2021, the report discovered how the SNP has impacted the United States in recent times. The approach of the report was accomplished through a combination of bash scripts, R code, and Perl code. The provided code took the downloaded BioProject's SRA RunTable metadata and executed several ramifications. The 89 Illumina samples in the RunTable was deconstructed, trimmed, and polished to produce VCF files that exposed the SNPs, quality scores, and other interpretable data that allowed a close analysis. Finally, a Makefile compiled the analysis and produced the report as a PDF.

Methods

The BioProject was selected and extracted through the National Center for Biotechnology Information (NCBI) website with parameters to isolate COVID-19 data. For an appropriate analysis, the chosen dataset was preferred to have various metadata and SRA experiments. Additionally, the SRA experiments had to be Illumina data for analysis approach to work. Upon choosing an adequate BioProject to work with, the accession number was put through the NCBI's SRA Run Selector. The metadata of the SRA RunTable was downloaded to begin the analysis. The analysis approach used a series of bash scripts that dependently functions with the preceding scripts. For this particular dataset, there were Oxford Nanopore data, and a script was created to remove the Nanopore data. In total, the fifteen scripts processed the RunTable for a complete analysis. The scripts were compiled in a Makefile, accessible through the terminal. Upon the input `make`, the parsing begins. Also, an R Markdown file was needed to produce the final report with functional figures/tables and text. The first script in the pipeline set up the anticipated directories was ran. The script produced thirteen total directories, stored in two places. A `/data/` folder was made along with a `data/` folder. Eleven of the directories were stored in the first folder, and the rest in the latter. After the directory set up, the series of scripts first used a fasterq-dump. The fasterq-dump from the SRA tools software will download the fastq files for each SRA run ID from the RunTable file, parsing it. The script also removed the reverse reads and stored it in a directory for raw fastq files. A second script downloaded the reference genome for the fasta file with the `curl` R code and stored it in its own directory. Following, a separate script will download the annotation gff, using `curl` and extracting the genome annotation with `gunzip`, storing it in its directory. Taking the fastq files, a script will run `fastqc` on the files. The processed fastq files are then trimmed with

a script that used `TrimmomaticSE`, throwing out the bad sequences. Aside from the trimmed fastq files, the next script will take the reference genome and index the file for Burrows-Wheeler Aligner (BWA). A script that runs BWA took the given files and aligns the each sample's reads as inputs to the reference genome, using `bwa mem`, storing the output as a .sam file. The .sam files are converted to .bam files with the next script, using the `samtools view` command. The .bam files are sorted with another script that used the `samtools sort -o` command, storing it in a new directory. The sorted bamfiles undergone a flagstat script. The script ran `flagstats` on the sorted files with the `samtools flagstat` command, extracting information from the sorted files, storing the stats in as a text file in a directory. The next script in the pipeline called the variants in the sorted bam files, and calculated the read coverage of the chromosomal positions in the entire genome which used the `bcftools mpileup` command. The script ultimately extracted information on the read coverage for each base, storing it in the made directory. Subsequently, another script took the bcf files, and exposed the SNPs for each file, using the `bcftools call -ploidy 1` command, saving it as a .vcf file. The next script used a perl script to filter out the short variants and SNPs on the .vcf file. The script used `vcfutils.pl varFilter` to execute the function. These series of scripts completed the whole analysis of the single RunTable, designating each script to save their files in a separate directory. The last remaining bash script drove the rendering of an Rmarkdown file. It required four arguments to run: the Rmd file, path to the gff annotation, path to the directory of processed vcf files, and the SRA RunTable with metadata. With a simple command of `make` in the terminal, the Makefile was set into motion, executing all of the scripts. In the end, a pdf document was produced displaying the report of the SARS-CoV-2 analysis.

Results

The BioProject collected 89 different samples, from four different isolates at the Icahn School of Medicine at Mount Sinai, located in New York, United States. The isolates originated from a nasal wash or nasal turbinate from *Mesocricetus auratus* and viral supernantants with pneumocyte infected cells and without (Table 1). The samples were collected in the months of June 2021 and November 2021. Although a number of the collection dates were before the identification of Omicron, the same SNPs were present in the collection dates during Omicron's emergence. Shortly after, the BioProject was released on January 3rd, 2021, past Omicron's first invasion into the United States. Regarding the spike protein gene *S*, the SNP from nucleotide C to T remained conserved in both month's of the collection date, occurring at the same chromosomal position 23525. Correspondingly, the other SNPs present and the SNP for the *ORF8* gene also remained conserved (Table 2). Parsing through the SRA RunTable with the bash scripts, it was determined that the

vast majority of SNPs in the sample were reliable and certain. The vast majority of the samples possessed SNPs that were confirmed. The samples almost consistently had a score greater than 220, with the some scoring in the range of 190 to 220 (Fig. 1, Fig. 2). For the latter, they were ruled out of the analysis due to the poor quality scores of 134 or lower. The poor SNPs mainly occurred at position 23525, with one changing to base pair A at 18883, and a change to C at 29737 (Table 3). It has been established that there are four certain SNPs that have occurred in SARS-CoV-2. The most common SNP changes the nucleotide C to T, and the other changing from T to C.However, only one SNP affects the spike protein's gene, at 23525. As for the other affected gene, the 28144 SNP of base pair C is shown to alter the ORF8 gene. SNPs at position 8782 and 18060 contains the same SNP as the spike protein SNP, but it is said to not affect any particular gene (Table 4). It has been identified that the SNPs are found in the Omicron variant. As a result, the deaths starting from December 1st, 2021, the day Omicron was identified in the united states, were recorded until November 23, 2022. In the time span, there has been 292, 973 confirmed deaths in the United States due to COVID-19 (Fig. 3). Although the deaths do not explicitly identify the variant to be the cause of death, we should consider that over 97% confirmed cases of COVID-19 are from Omicron since its emergence.

Discussion

With the SNP data collected by the Icahn School of Medicine, there is a couple of theories that revolve around the Omicron variant and it's spike protein. Following the timeline of Omicron, the variant was first identified in South Africa on November 8, 2021. The first detected of Omicron in the United States was December 1st, 2021 (Timeline of the SARS-COV-2 omicron variant, 2022). However, the first recordings of the SNPs were on June 8, 2021, and ending on November 11, 2021. There is a speculation that the Omicron variant has been around since the beginning of June due to the exact SNPs being present in both collection dates. Nonetheless, some SNPs may have been missed in the BioProject that differentiated the integrity of the SNPs, but it is worth noting that the SNPs remained conserved on both collection periods. Moreover, there were four quality SNPs that were identified in the analysis, but only one of the SNPs directly affected the gene responsible for the spike protein, *S*. The other affected gene was *ORF8*. The *ORF8* gene encodes for it's protein, ORF8. According to a published research paper, SARs-CoV-2 uses the protein to change MHC-I expression, affecting the cell's immune response (Zhang *et al.*, 2021). The MHC-I protein, or major histocompatibility protein 1, is responsible for aiding the immune system in recognizing foreign substances. Specifically, the protein binds to the pathogens and presents them to T cells, as a result, the

immune system eliminates the pathogen (Immunobiology, 5th edition: The immune system in health and disease, 2001). Therefore, the presumption that mutations in *ORF8* gene disrupts the protein's opsonization ability, allowing viral replication uncontested. Prior to replication, the virus must find its target. The virus accomplishes this through its spike protein. For SARS-CoV-2, the spike protein binds to the human Angiotensin-Converting Enzyme 2 (ACE2) receptor. The formation of the S protein-ACE2 complex allows the virus to gain entry to the host cell and integrate its viral genome. It has been established that SARS-CoV-2 has a strong binding affinity to the receptor, even stronger than the original SARS-CoV (Xie *et al.*). Conceivably, the SNP in the S protein increases the binding affinity for the ACE2 receptor. In addition, the two SNPs that do not reside within the genes, linked SNPs (or indicative SNPs) may play a role in the enhancement. According to the University of Utah, the linked SNPs do not affect a protein's function, but they correspond to a certain disease or drug response. The linked SNPs are said to be hallmarks, or genetic markers of their disease (Making SNPs make sense, n.d.). Thus, the causative SNPs, the polymorphisms that affect the genes, created a SARS-CoV-2 variant that can permeate and replicate more effectively than others. As a direct consequence, the Omicron dominates in infection rate and death rate (Omicron has caused higher increase in u.s. Daily death count than delta variant, 2022).

Figures

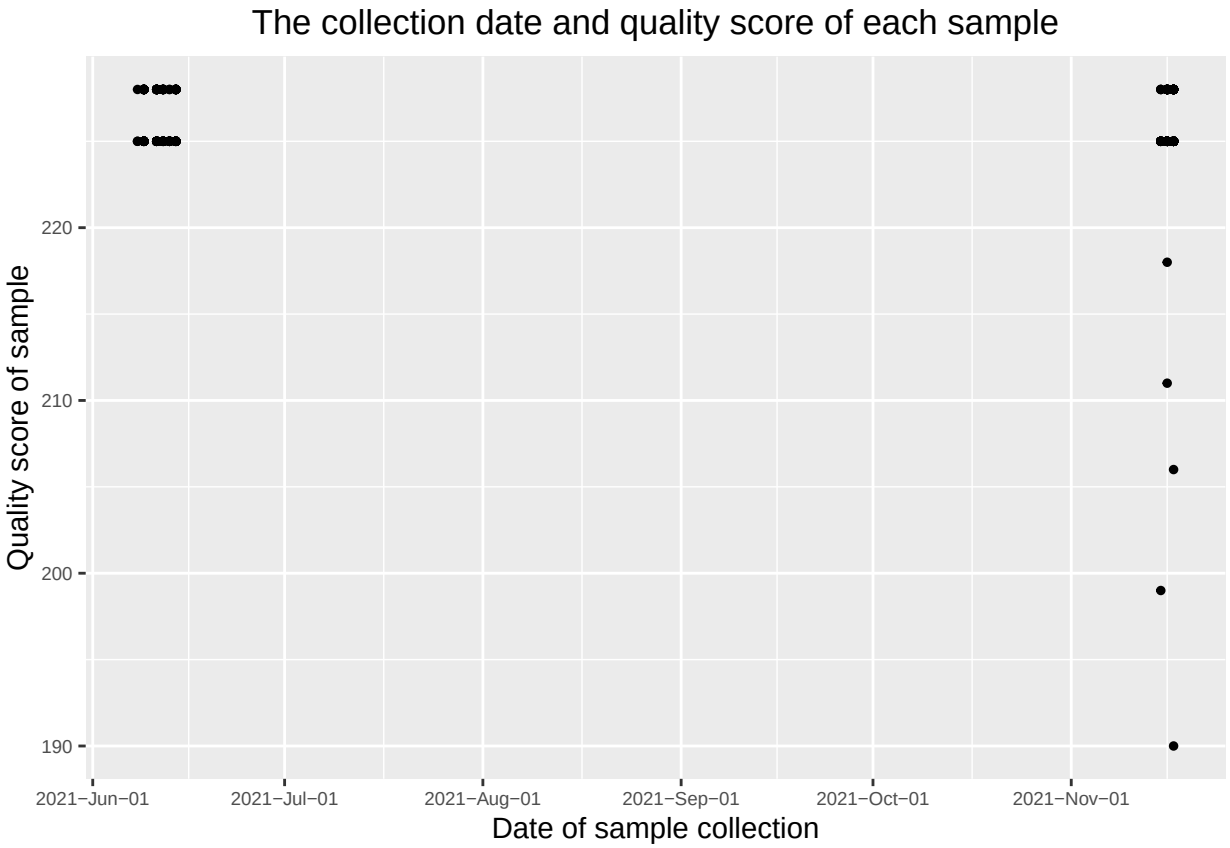

Figure 1: The collection date of the samples and their quality scores

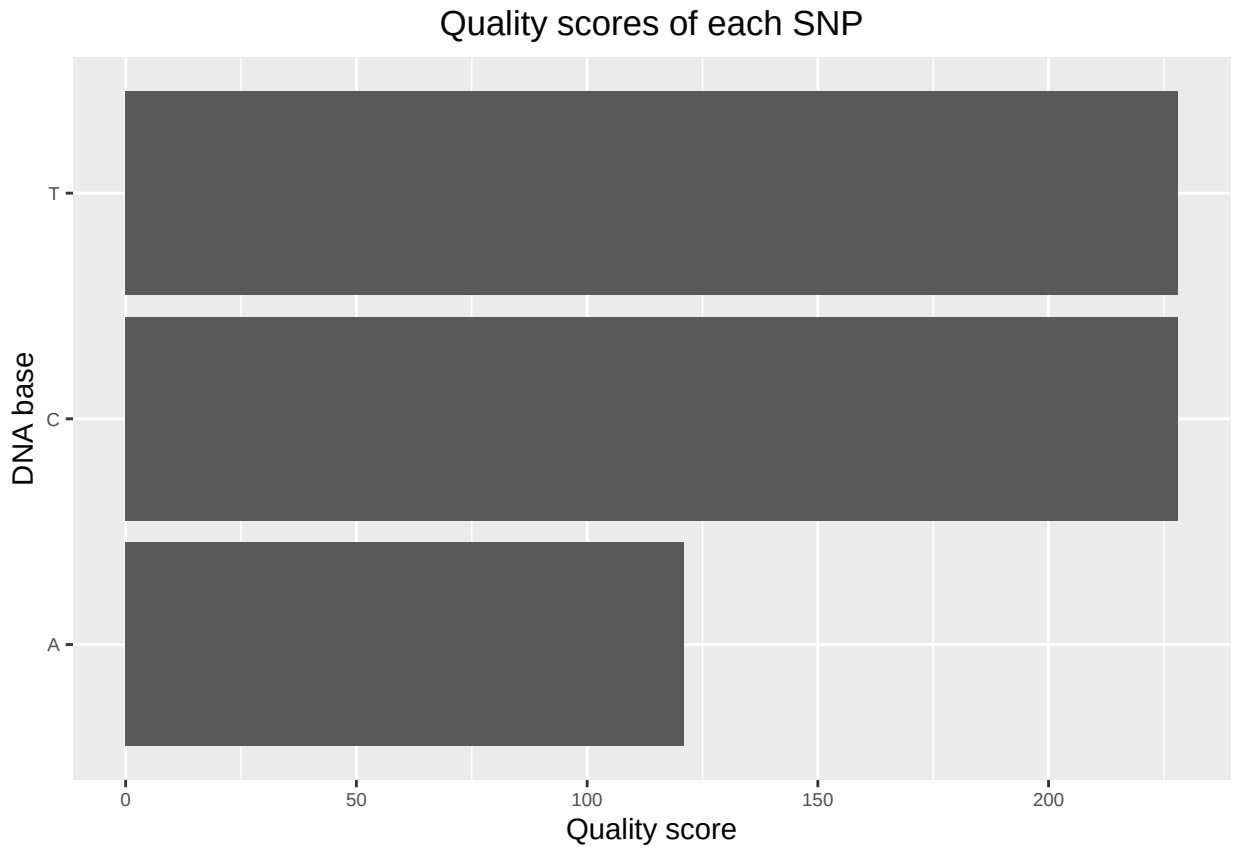

Figure 2: The quality score of each SNP identified in SARS-CoV-2

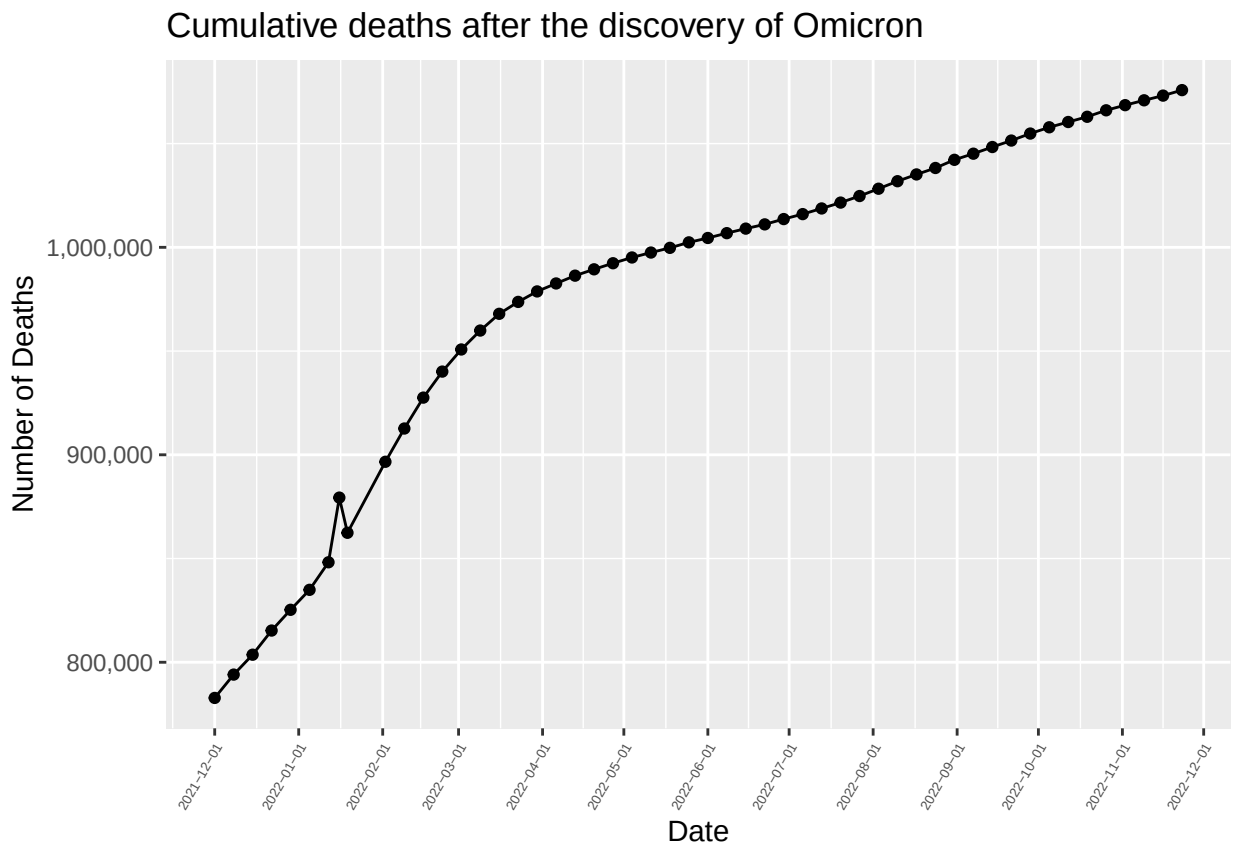

Figure 3: The cumulative deaths since the emergence of Omicron in the United States

Tables

Sample	Isolation source	Location
SRR17415128	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415129	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415130	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415131	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415132	Viral supernatants	USA: New York
SRR17415133	Viral supernatants	USA: New York
SRR17415134	Viral supernatants	USA: New York
SRR17415135	Viral supernatants	USA: New York
SRR17415136	Viral supernatants	USA: New York

Sample	Isolation source	Location
SRR17415137	Viral supernatants	USA: New York
SRR17415138	Viral supernatants	USA: New York
SRR17415139	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415148	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415150	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415151	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415152	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415153	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415154	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415155	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415156	Viral supernatants	USA: New York
SRR17415157	Nasal Turbinate from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415158	Nasal Turbinate from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415159	Nasal Turbinate from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415160	Nasal Turbinate from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415161	Nasal Turbinate from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415162	Nasal Turbinate from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415163	Nasal Turbinate from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415164	Nasal Turbinate from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415165	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415166	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415167	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415168	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415169	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415170	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415171	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York

Sample	Isolation source	Location
SRR17415178	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415188	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415189	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415190	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415191	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415192	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415193	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415194	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415195	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415196	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415197	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415198	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415199	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415200	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415201	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415202	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415203	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415204	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415205	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415206	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415207	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415208	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415209	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415210	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415211	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415221	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York

Sample	Isolation source	Location
SRR17415232	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415236	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415237	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415238	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415239	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415240	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415241	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415242	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415243	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415244	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415245	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415246	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415247	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415248	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415249	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415250	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415251	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415252	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415263	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415275	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415286	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415297	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415298	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415299	Nasal Wash from Syrian Golden Hamster (Mesocricetus auratus)	USA: New York
SRR17415302	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415303	Viral supernantants from pneumocyte infected cells	USA: New York

Sample	Isolation source	Location
SRR17415304	Viral supernantants from pneumocyte infected cells	USA: New York
SRR17415305	Viral supernantants from pneumocyte infected cells	USA: New York

Table 1: The isolation source and location of each sample in the experiment

Collection Date	Release Date	SNP	Gene	Position
2021-06-08	2022-01-03	T	NA	8782
2021-06-08	2022-01-03	T	NA	18060
2021-06-08	2022-01-03	C	ORF8	28144
2021-06-09	2022-01-03	T	NA	8782
2021-06-09	2022-01-03	T	NA	18060
2021-06-09	2022-01-03	T	S	23525
2021-06-09	2022-01-03	C	ORF8	28144
2021-06-11	2022-01-03	T	NA	8782
2021-06-11	2022-01-03	T	NA	18060
2021-06-11	2022-01-03	T	S	23525
2021-06-11	2022-01-03	C	ORF8	28144
2021-06-12	2022-01-03	T	NA	8782
2021-06-12	2022-01-03	T	NA	18060
2021-06-12	2022-01-03	T	S	23525
2021-06-12	2022-01-03	C	ORF8	28144
2021-06-13	2022-01-03	T	NA	8782
2021-06-13	2022-01-03	T	NA	18060
2021-06-13	2022-01-03	T	S	23525
2021-06-13	2022-01-03	C	ORF8	28144
2021-06-13	2022-01-03	A	NA	18883
2021-06-14	2022-01-03	T	NA	8782
2021-06-14	2022-01-03	T	NA	18060
2021-06-14	2022-01-03	T	S	23525
2021-06-14	2022-01-03	C	ORF8	28144
2021-06-14	2022-01-03	C	NA	29737
2021-11-15	2022-01-03	T	NA	8782
2021-11-15	2022-01-03	T	NA	18060
2021-11-15	2022-01-03	C	ORF8	28144
2021-11-15	2022-01-03	T	S	23525
2021-11-16	2022-01-03	T	NA	8782
2021-11-16	2022-01-03	T	NA	18060
2021-11-16	2022-01-03	C	ORF8	28144
2021-11-16	2022-01-03	T	S	23525
2021-11-17	2022-01-03	T	NA	8782
2021-11-17	2022-01-03	T	NA	18060
2021-11-17	2022-01-03	T	S	23525
2021-11-17	2022-01-03	C	ORF8	28144

Table 2: The conserved SNPs and their collection timeframe

Chromosome position	SNP	Quality score
18883	A	121
23525	T	111
23525	T	114
23525	T	86
23525	T	55
23525	T	134
23525	T	56
29737	C	97

Table 3: The excluded SNPs that posessed a low quality score

Chromosome position	Ref. Nucleotide	SNP	Gene affected
8782	C	T	NA
18060	C	T	NA
23525	C	T	S
28144	T	C	ORF8
29737	G	C	NA
18883	G	A	NA

Table 4: The SNPs and their reference nucleotide, and the affected gene

Sources Cited

CDC covid data tracker (n.d.) *Centers for Disease Control and Prevention.*

Immunobiology, 5th edition: The immune system in health and disease (2001) Garland Publishing.

Intro to viruses (article) | viruses (n.d.) *Khan Academy.*

Katella,K. (2021) Our pandemic year-a covid-19 timeline. *Yale Medicine.*

Making SNPs make sense (n.d.) *University of Utah.*

Omicron has caused higher increase in u.s. Daily death count than delta variant (2022) *PBS.*

Omicron variant: What we know so far about this covid-19 strain (2022) *University of California, Davis.*

Timeline of the SARS-COV-2 omicron variant (2022) *Wikipedia.*

Wang,M.-Y. *et al.* (2020) SARS-COV-2: Structure, biology, and structure-based therapeutics development. *Frontiers in Cellular and Infection Microbiology*, **10**.

Xie,Y. *et al.* Spike proteins of SARS-COV and SARS-COV-2 utilize different mechanisms to bind with human ACE2. *Frontiers in Molecular Biosciences*, **7**.

Zhang,Y. *et al.* (2021) The ORF8 protein of SARS-COV-2 mediates immunue evasion through down-regulating MHC-i. *Proceedings of the National Academy of Sciences*, **118**.