Identifying and Tracking the Evolution of Mutations in the SARS-CoV-2 Virus

Lavanya Venkatesan

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Biological Sciences

Roderick Jensen (Committee Chair)
Kevin Lahmers
Birgit Scharf

05/10/2021
Blacksburg, Virginia

Keywords: Pandemic, Pathogenic, Diagnosis, Mutation, Evolution, Sequencing, Database, Frequency

Identifying and Tracking the Evolution of Mutations in the SARS-CoV-2 Virus

Lavanya Venkatesan

ABSTRACT

SARS-CoV-2 is caused by a pathogenic and highly transmissible beta coronavirus leading to severe infections in immuno-compromised individuals. This study first evaluates the primers used in the Reverse Transcription Polymerase Chain Reaction (RT-PCR) to detect SARS-CoV-2 by understanding how mutations might affect the primer efficiency with the SARS-CoV-2 sequences. Mutations on the Spike protein of SARS-CoV-2 are the most important as the spike protein mediates the viral entry into host cells. This study tracks the course of mutations on the spike protein by focusing on the haplogroups of the sequences across the world. A comprehensive database linking three important, currently available databases is curated as part of this study to fill the gaps caused by sequencing errors. Further, this study exploits the data generated by the Illumina and Oxford Nanopore next generation sequencing methods to study the evolution of mutations in a single Septuagenarian patient over an infection period of 102 days using the gene analysis software Geneious Prime.

Identifying and Tracking the Evolution of Mutations in the SARS-CoV-2 Virus

Lavanya Venkatesan

GENERAL AUDIENCE ABSTRACT

A novel corona virus named Severe Acute Respiratory Syndrome Corona Virus 2 (SARS-CoV-2) has taken down the entire world by causing Covid-19 pandemic. Initially detected in Wuhan, China, the virus has now made its way to more than 200 countries with a heavy death toll. Understanding the virus through mutation tracking has now become the top priority of researchers. Most researchers depend on quality viral sequence datasets to identify and track mutations. One aim of this study is to provide a comprehensive dataset linking the GISAID (Global Initiative on Sharing All Influenza Data), NCBI (National Center for Biological Information) and the SRA (Sequence Read Archive) sequences. The dataset can be used for genome analysis and mutation tracking which provides important insights for vaccine design and improving diagnostic assays. This study provides an analysis of viral mutations in in the genomic regions targeted by commonly used primers in the RT-PCR tests for SARS-CoV-2 that may affect the efficiency of detection. This study also uses the haplogroup information of people across the world to track widespread mutations on the S gene of SARS-CoV-2 as it may be associated with increased transmissibility. To track the course of mutations in SARS-CoV-2, it is important to analyze the sequencing data provided by next generation sequencing methods. We present a case study to investigate the course of SARS-CoV-2 mutations in a single septuagenarian patient over a period of 102days using the Sequence Read Archive (SRA) data generated by two Next Generation Sequencing methods and compare the advantages that one has over the other.

Contents

ABSTRACT ii
GENERAL AUDIENCE ABSTRACT iii
INTRODUCTION 1
Mutations in the RT-PCR primers: 4
Mutation tracking in the S gene of SARS-CoV-2: 6
Analysis of the available datasets: 7
Viral Quasi Species: 8
Comparison of Next Generation Sequencing Methods: 9
MATERIALS AND METHODS 11
Methods for Evaluating Mutations in RT-PCR primers: 11
Methods to Analyze the Mutations on S gene of SARS-CoV-2: 14
Methods to integrate the NCBI, GISAID and SRA databases: 15
Methods for Identifying Viral Quasi Species: 16
Methods to Identify the Advantages and Disadvantages of the Next Generation Sequencing methods: 16
RESULTS 17
Results for primer design 17
Results for S gene mutation tracking: 23
Results for Database Comparison 24
Results for Viral Quasi Species 27
Results for Next Generation Sequencing 31
DISCUSSION 33
REFERENCES 39
APPENDIX 42

Chapter 1

INTRODUCTION

The severe acute respiratory coronavirus 2 (SARS- Cov-2) first emerged in December 2019 in Wuhan, China, in a seafood market. The pathogenic virus is capable of causing respiratory and intestinal infections both in animals and humans. SARS-CoV-2 is said to have originated in bats and transmitted to humans via an unknown intermediate host [1]. Morphologically, SARS-CoV-2 viruses are enveloped and contain a non-segmented, positive sense single stranded ribonucleic acid (RNA).

The virus which spreads at a rapid rate is now a threat for most countries of the world. The virus affects everyone but is deadliest in the older people with underlying medical conditions or immunocompromised individuals. Masks and social distancing were found to be one of the main ways to control the spread of the virus which is transmitted through respiratory droplets when a person coughs or sneezes. Vaccine design and diagnostic methods to detect and treat the Covid-19 has become one of the most important projects for the scientists around the world.

Coronaviruses belong to the Coronaviridae family, the order Nirovirales and the genus Coronavirus. Coronam being the latin for “crown”, the positive stranded RNA virus appears with a crown like structure when observed under the electron microscope. They are classified into four different genera, namely, Alphacoronavirus, Betacoronavirus, Gammacoronavirus and Deltacoronavirus. To confirm that the causative agent of COVID-19 is a Betacoronavirus, Chinese scientists sequenced the metagenomic RNA from the virus isolated from bronchoalveolar lavage fluids of pneumonia patients [2]. The isolated complete viral genome was found to be 29,903 nucleotides long and was

analyzed phylogenetically to reveal that it had 89.1% sequence similarity to a Betacoronavirus isolated in China from bats previously [3]. This outbreak signifies the fact that coronaviruses that infect animals can infect humans and then get transmitted among the population by respiratory droplets from sneezing, coughing or talking. From a sea food market in Wuhan, where the transmission first started, the virus has now made its way all over the world.

The structure of SARs-CoV-2 genome has four main structural proteins- the spike (S) glycoprotein, membrane (M) glycoprotein, the nucleocapsid (N) glycoprotein, and the envelope (E) glycoprotein along with a number of accessory proteins. [4] The spike glycoprotein is a trimeric protein with each monomer about 180 kDa, containing S1 and S2 subunits. SARS-CoV-2 uses the spike glycoprotein to mediate membrane fusion and viral entry into host cells. While the S1 subunit mediates the attachment, the S2 subunit mediates the membrane fusion[5]. The most abundantly expressed M protein stabilizes the nucleocapsid (N-protein- RNA complex) by binding to the N protein. Bound to the nucleic acid component of the virus, the N protein is involved in processes like replication cycle and host cell cellular response to viral infections in the viral genome. Of all the major structural proteins, the E glycoprotein is the smallest with abundant expression inside the infected cell, but in the virion envelope only a small portion of this glycoprotein is incorporated[6].

As the number of cases around the world increased, it became very important to have an organized record of the sequenced genomes for research. GISAID (Global Initiative on Sharing All Influenza Data) made this possible by promoting the sharing of all coronavirus data related to COVID-19. GISAID served as the public database that in

addition to providing the genetic sequence, provides us with the metadata that contains the geographical information, clinical and epidemiological data [7]. The first Covid-19 sequence was submitted to the NCBI GenBank on January 5, 2020 by Dr.Yong-Zhen Zhang and scientists from Fudan University, Shanghai and was first revealed publicly on the internet on January 11, 2020 and one day later five additional sequences from across China (Chinese CDC, Wuhan Institute of Virology and Chinese Academy of Medical Sciences & Peking Union Medical College) were posted on the GISAID website [8]. This provided the first dataset for the researchers around the world to start working on the coronavirus.

Coincidentally, the spread of Covid-19 intensified as travel between cities was at its peak in China during the beginning of the lunar New Year. From Wuhan, the novel coronavirus spread to other cities in the Hubei province and subsequently to other parts of China covering all of the 34 Chinese provinces in one month. Cases that were increasing in hundreds everyday soon accelerated to thousands by January. On 30th January, the novel coronavirus outbreak was declared as a public health emergency of International concern by the World Health Organization (WHO)[9].

Based on the published viral genome sequence, US Centers for Disease Control (CDC) and the World Health Organization (WHO) released a set of primers in January 2020[10] [11], which they deemed suitable for RT-PCR detection of SARS-CoV-2 viral samples. In addition a Korean group published a paper in March 2020 where they released a set of primers that they described to be suitable for an accurate and low-cost RT-PCR [12]. The Korean protocol was later adopted by the Virginia Tech Schiffert Health Center Molecular Diagnostics Laboratory[13].

This study begins by providing a detailed documentation of the suitability of the CDC, WHO and Korean RT-PCR primers by aligning them with viral sequences and analyzing them using a sequence analysis software named Geneious Prime to identify emerging mutations that may affect detection efficiency. This study also provides an account of the infamous D614G mutation in the viral spike S protein and provides a comparison of datasets on SARS-CoV-2 that are currently available for public research.

In an effort to understand the viral quasi species, this study documents the course of mutations in a septuagenarian patient affected by SARS-CoV-2. We have provided a detailed description of the different mutations that occurred in the patient during a course of 102 days. The next generation sequencing data provided by Illumina and Oxford Nanopore sequencing technologies have been analyzed to provide an account of the different mutations in the sequenced genome. As part of this study, we have compared the two sequencing technologies to understand their advantages and disadvantages. To give a better insight into the goals of the project, the specific aims of this research and their importance are highlighted below.

Mutations in the RT-PCR primers:

Vaccine design and diagnostic methods to detect and treat COVID-19 became one of the most important projects for scientists. In today's world the classical detection methods like immunoassays have been replaced by Real-Time Polymerase Chain Reaction (RT-PCR). One of the most used methods to amplify a DNA of interest in diagnostic testing is Polymerase Chain Reaction (PCR). A good PCR amplification depends to a great extent on the choice of primers with the best priming efficiency. Complementarity between primers and template is essential for a successful PCR and the

mismatches are the main reason for a reduced priming efficiency. Every mismatch, irrespective of its location within the primer sequence, will result in a decreased thermal stability of the primer- template duplex, thus potentially affecting PCR amplification. Specificity is also an important characteristics when designing a primer because mismatches between primer and the template sequences may result in biased results as well as PCR failure [14][15].

RT-PCR has widespread applications in detecting viral infected human specimens. Detection of HSV1 and HSV2 strains were made possible using RT-PCR using TaqMan probes and was very successful [16]. RT-PCR has also been instrumental in studying the interactions a virus has with its host which has been a very useful information when we think about designing vaccines in the case of COVID-19 [17]. Even though the RT-PCR technique enables specific and quantifiable detection of the viral genomes, the inherent genetic variability of the viruses makes it essential that the viral variants are recognized by the primer sequences. For viruses to be efficiently amplified, there is a need for primers that will accurately amplify the viral genomes containing conserved nucleic acid sequences. For an oligonucleotide to be an efficient primer, there are several important factors like association and dissociation kinetics of the primer template complexes at temperatures of annealing and extension [18][19].

This study will address the issue of efficiency problem of the CDC, WHO and Korean primers by analyzing their mismatches with the template sequences arising from mutations in the viral sequences using the sequence analysis software- Geneious Prime.

Mutation tracking in the S gene of SARS-CoV-2:

With the rapid increase in the number of cases, scientists were able to better understand the mutations that emerge in the course of evolution of the virus. To understand better viral detection, pathogenesis, and vaccination evasion, knowledge of the mutation rate is very important as mutation rate is the ultimate source of genetic variation [20].

Of the new mutations in the coronavirus genome, the most important ones are those on the regions of genome that encoded the spike protein. SARS-CoV uses the Spike (S) protein to enter the host cells and attaches its Receptor binding Domain (RBD) to the ACE2 protein on the cell surface [21]. One mutation that emerged early and quickly became fixed in the S gene and was found across most of the SARS-CoV-2 genomes sequenced around the world was the D614G mutation [22]. The viruses with the amino acid change in their 614^{th} position emerged dominant raising the question as to what this meant for the Covid-19 pandemic.

Initially it was hypothesized that that the rapid spread of the D614G was associated with increased infectivity. Later, studies confirmed that D614G mutation causes increase in transmission by accumulation of the viral loads in the upper respiratory tract of Covid-19 patients. In addition to this, the study minimized the role of D614G in affecting vaccine efficacy and antibody therapies. According to an experiment, the D614G variant resulted in higher infectious titers in the nasal washes and trachea but not the lungs. Even though the study eliminates the possibility that D614G mutation compromises the efficacy of vaccines, it says that viral transmissibility is increased. This

means that people infected by strains having the D614G mutation can spread the infection at a higher rate [23].

In some of the newer variants like B.1.1.7, there are 17 mutations that result in amino acid changes and 8 of these 17 mutations are present in the S gene that codes for the spike protein of the virus, making the strain more infectious than the other strains of SARS-CoV-2.[24] If the standard RT-PCR method that was used to identify the non-mutated form of the virus is used to test for the B.1.1.7 variant, it might not detect the S gene, resulting in a 'S gene dropout'. But as the RT-PCR technique is designed to detect the other regions of the virus (N gene, E gene, RdRp gene, etc.,), the result of the RT-PCR might still be positive indicating the presence of SARS-CoV-2. While this might not be the perfect way to detect the B.1.1.7 variant, researchers are checking the samples with S gene dropouts for this particular variant [25]. This signifies the importance of tracking the mutations on the S gene and designing primers specific to the S gene of SARS-CoV-2 as it will help us identify the variants carrying these critical mutations accurately.

This study will provide review of the D614G mutation on S gene with an emphasis on its origin and spread.

Analysis of the available datasets:

In our study, we tried to trace down to the first instance of D614G mutation to understand its spread. We started to build a database consisting of 42,329 sequences downloaded from GISAID out of which 34,987 sequences have the D614G mutation. When we took a close look at the sequences in the database, many sequences are incomplete or have sequencing errors. This problem can be potentially solved by linking the complete sequences to the NCBI Nucleotide database and using the raw data obtained

from Illumina and Oxford Nanopore Sequencing in the NCBI SRA (Sequence Read Archive) database.

Currently, the NCBI has an elaborate database of 155,200 Covid sequences [26]. In addition, NCBI has SRA (Sequence Read Archive) records that are short reads of DNA sequencing data (typically fewer than 1,000 bases) for Covid genomes. This study attempted to create a curated database that has the NCBI and SRA accession numbers corresponding to accession numbers in the GISAID database. By doing this, we would be able to provide a valuable dataset that has no ambiguity with respect to the sequencing information. If a particular clade has poorly sequenced data in one database, the researchers can look up for the equivalent sequence in the other two databases which will provide them with the necessary information to carry out their research.

To achieve this, the basic BLAST operation was used to identify the equivalent NCBI sequence for every GISAID sequence. Next, the metadata of the SRA sequences were analyzed to assign them to the appropriate NCBI sequences. Finally, providing a comprehensive table containing accession numbers from all three databases was one of the main aims of this work. This effort would help to resolve issues related to poorly sequenced data as it provides researchers with more than one platform that has access to data.

Viral Quasi Species:

Viral quasi species refers to the population structure where closely related genomes undergo a continuous process of genetic variation, competition, and selection. Due to the lack of proofreading mechanisms in their RNA dependent RNA polymerases, the replication mechanisms of positive, single stranded RNAs are prone to errors. These

errors result in the creation of genetically distinct genetic variations. This range of mutations that have occurred in a gene or at a specific locus in a gene is referred to as mutant spectra. These mutant spectra are the target of evolutionary events in RNA viral populations. Viral cells have the innate ability to infect a particular cell, tissue or host species referred to as cell tropism, tissue tropism and host tropism, respectively. The mutant spectra is the starting point for important events like their ability to change their cell tropism or host range in the biology of RNA viruses [27]. The target of evolutionary events is usually the mutant spectra. Studies on mutant clouds give a good picture on the role on RNA viruses in pathogenesis of viral infection and disease [28]. We observed this phenomenon in the SARS-CoV-2 genome very early on when we analyzed the SRA data of the illumina sequence of a strain from Wisconsin (SRA Study- SRR1114075) using Geneious Prime.

In this project, we present a detailed case study of the evolution of mutations in a patient affected by SARS-CoV-2. The case study details the course of mutations in a septuagenarian patient using the raw data collected at 23 time points in 101 days [29]. The mutation frequencies of the different mutations that arise in the sequence data of the viral genome are analyzed to provide an understanding about the course of mutations in the patient.

Comparison of Next Generation Sequencing Methods:

The diagnostic method currently used in most of the labs to test the presence of SARS-CoV-2 is RT-PCR. Although RT-PCR is an effective technique to give useful information about the presence of the virus, the Next Generation Sequencing methods are very useful in providing information the appearance of new mutations in the viral

genome and the presence of about co-infections of different strains or quasi-species. The two most commonly used Next Generation Sequencing techniques are Illumina and Oxford Nanopore sequencing. These techniques help to track the course of the pandemic by helping us understand the transmission routes and help us keep up with the new mutations that arise on the viral genome and ultimately help in tracking the rate of viral evolution. Understanding these perspectives can help in designing therapeutics to combat the pandemic.

In this study, we compare the Sequence Read Archive data generated by both Oxford Nanopore and Illumina sequencing methods to track and understand the course of evolution of the viral strain in a septuagenarian patient. In addition to giving interesting results about the course of mutations in the patient, the study also helps to understand the advantages that one next generation sequencing method has over the other.

We have used our initial findings as a tool to investigate this case study to understand the possibility of one patient being infected by more than one strain. Our study compares the Illumina and Oxford Nanopore raw data to reveal interesting information about the course of mutations that developed in the patient.

Chapter 2

MATERIALS AND METHODS

Methods for Evaluating Mutations in RT-PCR primers:

The coronavirus genomes used for analysis were downloaded from GISAID (Global Initiative on Sharing All Influenza Data). 42,329 sequences downloaded from GISAID were then integrated into a database using the command line operation "Makeblastdb". Makeblastdb, a part of the new blast+package from the NCBI is used to create a local BLAST (Basic Local Alignment Search Tool) database from these FASTA files. Every sequence is associated with a unique identifier which is the GISAID accession ID in this case. For the command to be executed, the identifier should begin with a ">" symbol on the definition line and should contain no spaces [30]. There are three parameters required to complete a "Makeblastdb" command, namely, 1) "-in" where the location of the input file containing all the nucleotide sequences in FASTA format should be provided, 2) "-out" where the name of the output file should be mentioned, and 3) "-dbtype" where the type of the database i.e., the information for whether the database contains nucleotide or protein sequences should be set. The output of this code gives a single database with all the nucleotide sequences are present in FASTA format.

Our first goal of the project was to perform the BLAST operation between the CDC (Center for Disease Control and Prevention), WHO (World Health Organization) and Korean primers and probes with the sequences in the database to evaluate the primers. BLAST is a tool that helps to find similar regions between a query nucleotide sequence to the subject sequences in a database. NCBI (National Center for Biological Information) offers a wide variety of BLAST algorithms that can be used against

different databases. The default parameters used by BLAST can be customized to our specific needs. BLAST is the best tool to evaluate the biological significance of an alignment as it provides the statistical information for every alignment as 'expect ratio'. Given below are the list of CDC, WHO and Korean Forward and Reverse primers that were analyzed for their priming efficiency. The CDC and Who primer sets also included a Taqman Probe sequence to improve specificity. The Korean Primer design was based on the SYBR green RT-PCR protocol and only require Forward and Reverse Primers.

List of CDC primers and probes

Table 1-CDC- N1 gene

Forward	GAC CCC AAA ATC AGC GAA AT
Reverse	TCT GGT TAC TGC CAG TTG AAT CTG
Probe	ACC CCG CAT TAC GTT TGG TGG ACC

Table 2 – CDC- N2 gene

Forward	TTA CAA ACA TTG GCC GCA AA
Reverse	GCG CGA CAT TCC GAA GAA
Probe	ACA ATT TGC CCC CAG CGC TTC AG

List of WHO primers and probes

Table 3- WHO – N gene

Forward	ACAGGTACGTTAATAGTTAATAGCGT
Reverse	ATATTGCAGCAGTACGCACACA
Probe	ACACTAGCCATCCTTACTGCGCTTCG

Table 4 – WHO – E gene

Forward	CACATTGGCACCCGCAATC
Reverse	ACTTCCTCAAGGAACAACATTGCCA
Probe	GAGGAACGAGAAGAGGCTTG

Table 5- WHO – RdRp gene

Forward	ATGAGCTTAGTCCTGTTG
Reverse	CTCCCTTTGTTGTGTTGT
Probe	AGATGTCTTGTGCTGCCGGTA

List of Korean primers and probes

Table 6- Korean Primers- E-1 gene

Forward	TTCGGAAGAGACAGGTACGTT
Reverse	CACACAATCGATGCGCAGTA

Table 7– Korean Primers- N-2 gene

Forward	GCTGCAATCGTGCTACAACT
Reverse	TGAACTGTTGCGACTACGTG

All the above primers were subjected to BLASTn with the 42,329 sequences in our database to determine how many of these primers are an exact match (100% sequence identity) to the sequences in the database and how many are non-exact matches (less than 100% sequence identity). This was done in an effort to understand which of the primers would have the best priming efficiency and give rise to a minimal number of false negative results. The result of the power shell code used to perform a BLAST search of

the primer sequences which gave an output file with both exact and non-exact matches. To reduce redundancies and scalability reasons (to reduce errors while working in huge amounts of data), a Python script was used to separate the contents of the output into two folders - one with sequences that were an exact match by matching the '%' given in the title and the other with sequences that are a non-exact match. Following this, the non-exact match sequences were imported for analysis to Geneious Prime to understand their mutations.

'Map to Reference' is a feature offered by Geneious Prime that helps to assemble a sequence or a list of sequences to another. Once the alignment is done, Geneious Prime will produce a contig containing the reference sequence at the top of the alignment and all the other sequences listed below it. Each of the imported sequences is aligned to the contig independently and then the pairwise alignments are combined to generate the contig. There are several fine-tuning options available to increase the stringency of consensus. The contig also generates a "Sequence Logo" plot that can give a graphical representation of the consensus and diversity between the reference sequence and the other sequences that have been aligned to the reference sequence [31]. By stacking nucleotides on top of each other, the SeqLogo plot gives the relative frequency of each character with the height of each nucleotide proportional to its relative frequency. The total height of the stacked nucleotides is determined by the information contained in that position.

Methods to Analyze the Mutations on S gene of SARS-CoV-2:

For understanding the mutations present of the Spike protein of the SARS-CoV-2 genome, the S gene of the SARS-CoV-2 genome, between nucleotides 21563 to 25384

was subjected to BLASTn with all the sequences in the database. Again, the result was a single file containing the exact and non-exact matches. These matches were sorted using the Python code and the non-exact matches were imported into Geneious Prime for further analysis. The non-exact match sequences were mapped to the reference genome (NC_045512.2) of SARS-CoV-2 sequence to identify new mutations on the spike protein of the SARS-CoV-2 genome. To track down the first instance of the D614G mutation, all the sequences that were non-exact matches to the S gene were sorted according to their haplogroup. The metadata information for all the sequences on the GISAID database was downloaded. Then, the metadata for the non-exact match sequences was extracted using a Python script and imported to Microsoft Excel for analysis. These sequences were then sorted according to their haplogroups and dates to understand if the D614G mutations were predominantly present in a particular haplogroup.

Methods to integrate the NCBI, GISAID and SRA databases:

The two major repositories of SARS-CoV-2 sequences are NCBI and GISAID databases. In many cases, the sequences or metadata downloaded from one of the two databases would have sequencing errors or be incomplete. Finding the equivalent for a sequence from one database in the other seemed to be the most logical way to approach this problem. As this correlation was not readily available, an effort was made to create a comprehensive table that contained the GISAID and SRA equivalent accession numbers for sequences available on NCBI. Correlating SRA data with the NCBI and GISAID accession numbers can be extremely beneficial as it helps to take a deeper look into the course of mutations in the sequences by analyzing the raw data. A Python script was developed to isolate the GISAID and SRA accession numbers available in the metadata

information of the NCBI sequences was used to curate a table containing the most comprehensive information.

Methods for Identifying Viral Quasi Species:

A case study to understand the course of mutations in a septuagenarian patient infected with SARS-CoV-2 for 101 days [29] was performed. The Sequence Read Archive data was obtained from the NCBI repository for both Illumina and Oxford Nanopore Sequencing methods. Out of the 23 recorded time points, six time points were chosen to analyze the raw data. The reads were uploaded to Geneious and mapped to the reference genome of SARS-CoV-2. New mutations were observed to gradually appear in the course of the infection. The mutation frequency for each mutation at each time point was recorded to analyze the mutations. This helped to understand if the infection led to the emergence of new quasi-species or was a result of a mixture of strains during the course of the infection.

Methods to Identify the Advantages and Disadvantages of the Next Generation Sequencing methods:

The SRA data containing the sequencing information generated by Illumina and Oxford Nanopore Sequencing methods were uploaded to Geneious Prime for analysis. The raw data reads were aligned to the reference genome for SARS-CoV-2 (NC_045512) for analyzing the course of mutations present on SARS-CoV-2. The advantages and disadvantages conferred by each method was understood by observing the alignment between the SRA reads and the reference genome as good alignment gives an accurate picture of the nature and course of mutations present.

Chapter 3

RESULTS

Results for primer design

The BLAST results showed there were a significant number of sequences that were non-exact matches (less than 100% sequence identity) with the CDC, WHO and Korean primer sequences. To investigate the mutations in the non-exact sequences, they were analyzed using Geneious by aligning them with the primer sequences to find the exact location and nature of the mutation. These results can be used to modify the primers to keep up with the mutations on the SARS-CoV-2 genomes thereby helping prevent the false negative results while performing RT-PCR to diagnose the presence of SARS-CoV-2 in an individual.

Table 8- Number of Exact and Non-Exact match Sequences for CDC, WHO and Korean Primers

Name of Primer	Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences
N1-Forward-CDC	42329	42142	61	126
N1-Reverse-CDC	42329	41979	227	123
N1-Probe-CDC	42329	41367	857	105
N2-Forward-CDC	42329	42198	75	56
N2-Reverse-CDC	42329	42248	33	48
N2-Probe-CDC	42329	42100	148	81
E-Forward-WHO	42329	42295	12	22
E-Reverse- WHO	42329	42263	34	32
E-Probe-WHO	42329	42236	50	43
N-Forward-WHO	42329	42242	45	42
N-Reverse-WHO	42329	42113	132	84
N-Probe-WHO	42329	42111	132	86
RdRp-Forward-WHO	42329	90	42199	40
RdRp-Reverse-WHO	42329	42164	120	45
RdRp-Probe-WHO	42329	120	42059	150
E1-Forward-Korean	42329	42198	46	85
E1-Reverse-Korean	42329	2	42096	231
N2-Forward-Korean	42329	42032	111	186
N2-Reverse-Korean	42329	42195	117	17

Having an account of the exact and non-exact matches (Table -8) might help in evaluating the efficiency of the CDC, WHO and Korean primers. The primers with lower number of non-exact matches will have a better priming efficiency than the others. The table also shows the number of sequences that are missing meaning that they were not able to be detected by the Python code. This might be because the sequences had ‘N’ or gaps in the place of nucleotides or sometimes because of ambiguity in the accession numbers of these sequences. A few sequences on GISAID website have gaps in the accession numbers making it difficult to detect them. In addition to the number of non-exact matches, it is also necessary to understand the nature of the mutations that cause the

non-exact matches. The mutations might either occur in different regions of the primer or they may be fixed in one position of the primer or maybe a combination of both. It is very important to understand the nature of the mutations as having mutations in multiple regions of the primer might make it less efficient.

CDC PRIMERS

Table 9 - Frequency of mutations in the non-exact match sequences for N-2 gene (Forward) of CDC primers

S.No	Position	Mutation	Frequency
1	1	T-G	1
2	4	C-T	12
3	8	C-T	28
		C-N	2
		C-Y	1
4	9	A-T	1
		A-N	1
5	12	G-K	1
		G-N	1
6	13	G-T	1
7	14	C-T	2
		C-N	1
8	15	C-S	1
9	16	G-T	22
		G-S	1
		G-R	1
		G-N	1
		G-K	1
10	18	A-M	1

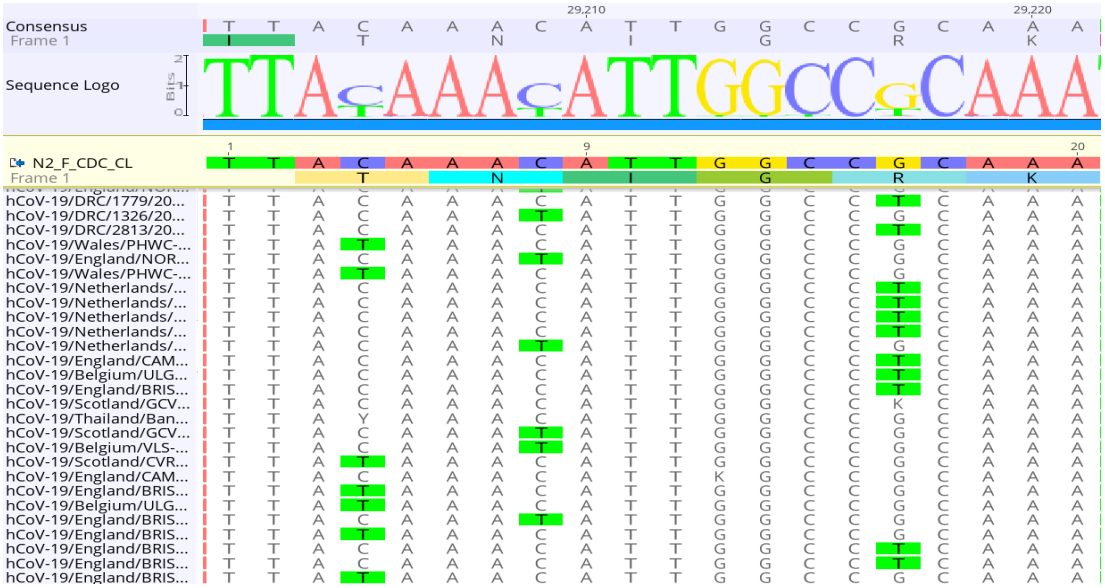

Fig-1: Geneious analysis displaying the Seqlogo plot and mutations on N2-gene primer (Forward) of CDC PRIMERS

From Table – 9, it can be seen that there are 3 mutations that are dominantly present in the N2- gene primer (Forward) of CDC primers. Even though there are 3 mutations in this primer, they are not present in the same sequence and so they might not be able to cause a sequence to be not detected by this primer. The SeqLogo plot (Fig-1) clearly shows that the mutations in positions 4, 8 and 16 are widespread.

Table 10 - Frequency of mutations in the non-exact match sequences for E - gene (Probe) of WHO primers

S.No	Position	Mutation	Frequency
1	2	C-T	1
2	7	G-N	1
3	9	C-T	38
4	13	C-T	1
5	14	T-Y	1
6	15	T-N	1
7	19	G-N	1
8	20	C-T	2
9	22	C-T	1
10	23	T-N	1

11	24	T-Y	1

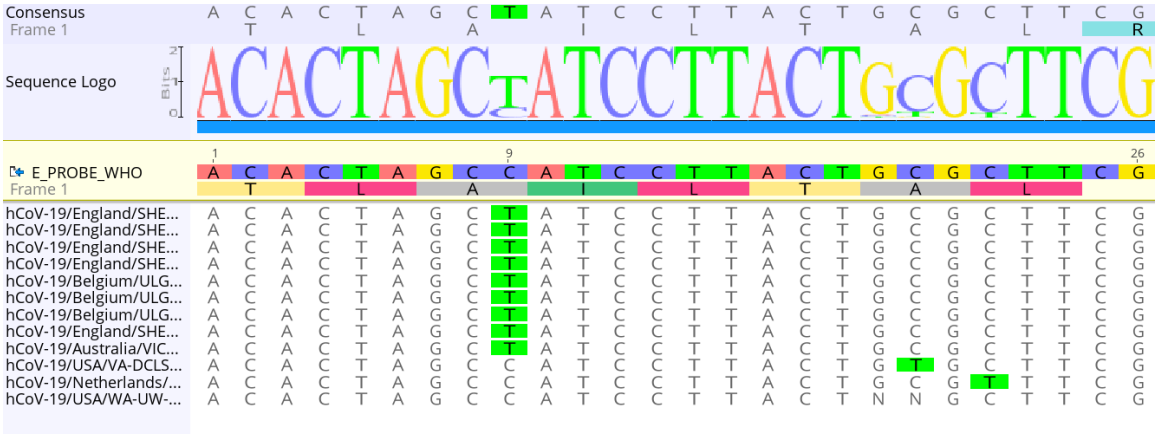

Fig-2: Geneious analysis displaying the Seqlogo plot and mutations on E-gene (Probe) of WHO.

According to Table 10, there is only one significant mutation (C-T) in position 9 that is present in 38 of the 50 non-exact match sequences making it one of the most efficient probes. This can be verified from the SeqLogo plot displaying the dominant mutation that is present across sequences in position 9.I

KOREAN PRIMERS

Table 11 - Frequency of mutations in the non-exact match sequences for E1 - gene (Reverse) of Korean Primers

S.No	Position	Mutation	Frequency
1	1	C-T	1
		C-G	2
2	2	A-T	1
		A-W	1
3	3	C-A	1
4	4	A-T	2
5	5	C-G	1
6	6	A-C	1
		A-G	1
		A-T	1
7	7	A-C	1
		A-T	2
8	8	T-C	2
		T-A	1
9	9	C-A	1

		C-T	1
		C-S	3
10	10	G-T	2
11	11	A-R	1
12	12	T-A	42068
		T-C	2
		T-N	1
13	13	G-A	1
		G-C	1
		G-T	1
14	14	C-A	2
15	16	C-A	1
		C-G	1
16	17	A-C	1
		A-T	1
17	18	G-T	2
		G-A	1
18	19	T-G	2
		T-C	1
19	20	A-C	1
		A-N	1

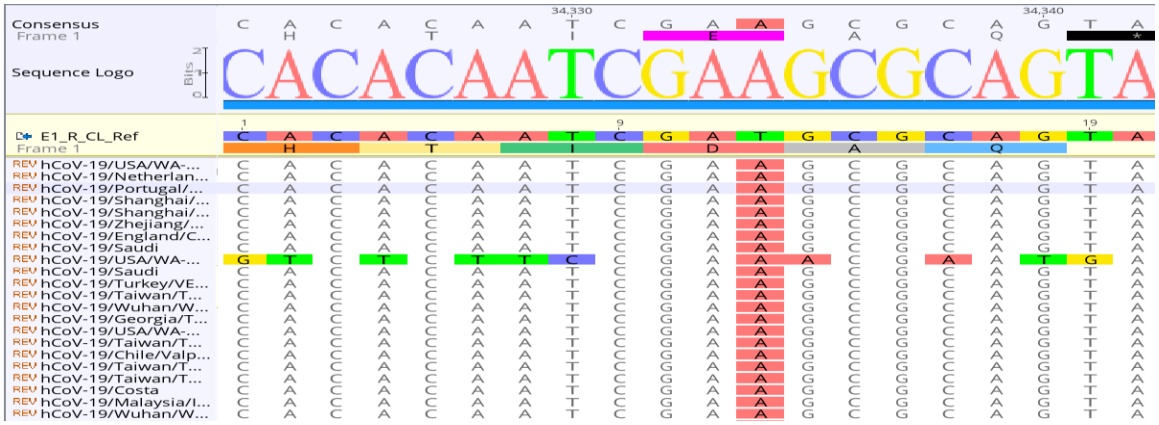

Fig-3: Geneious analysis displaying the Seqlogo plot and mutations on E1-gene (Reverse) of Korean Primer

Even though there are 19 mutations present in the E-1 gene (Reverse) of Korean primers, there is only one significant mutation in position 12. All the other mutations must be sequencing errors as they do not show up in significant numbers across sequences. This shows the importance of recording the frequency of mutations and the

SeqLogo plot. Even though this primer has 42096 non-exact matches, 42068 of those are fixed in position 12. This means that in spite of the large number of the non-exact matches, this primer might still be one of the more efficient primers for RT-PCR.

Results for S gene mutation tracking:

The S gene of SARS-CoV-2 is 3,499 nucleotides long. A BLAST search was performed against the database of 42,329 SARS-CoV-2 sequences obtained from GISAID. There were 34987 sequences that were non-exact matches (less than 100% sequence identity). These sequences were analyzed further by aligning them with the S gene sequence in Geneious to reveal that most of these non-exact matches contained the D614G mutation.

Table 12 - Frequency of mutations in the non-exact match sequences for S gene of SARS-Cov-2

Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences
42329	7029	34987	313

It was very interesting to note that most of the non-exact match sequences had the D614G mutation in the S gene. The strains carrying the D614G variant were proven to be associated with increased transmissibility. This made it important to understand the course of the mutation and trace back to the first instance of the mutation by analyzing the haplogroups of the mutation. Tracking the haplogroups of the infected people can help track the mutations. Analyzing the haplogroups of the non-exact match sequences can help us determine how many of these sequences contain the D614G mutation, as all

the sequences bearing this variant must belong to the same haplogroup. 'Clade' and 'Pangolin Lineage' in Fig-4 contain the information for haplogroups of the non-exact match sequences. All sequences belonging to Clade 'G' or Pangolin Lineage 'B' are the sequences containing the D614G mutation. Of the 34987 sequences that were not an exact match to the S gene, 30,337 sequences belonged to the clade G meaning that they belonged to the D614G variant. We traced back the first instance to a male patient from Germany which was discovered on 31st January 2020 (EPI_ISL_406862). The comprehensive table containing the list of haplogroups for all the sequences in the database containing the D614G mutation can be found here.

Clade	Country of origin	Gisaid Accessio	Pangolin lin	Region exposure	Country Exposur	Length	Host	Age	Sex	Date Submitt
G	Germany	EPI_ISL_406862	B	Europe	Germany	29782	Human	?	Male	1/31/2020 0:00
GR	Germany	EPI_ISL_412912	B.1.1	Europe	Italy	29756	Human	?	?	2/28/2020 0:00
G	Brazil	EPI_ISL_412964	B.1	Europe	Italy	29890	Human	61	Male	2/28/2020 0:00
G	Finland	EPI_ISL_412971	B.1	Europe	Italy	29776	Human	24	Female	3/1/2020 0:00
GR	Mexico	EPI_ISL_412972	B.1.1	Europe	Italy	29903	Human	35	Male	3/1/2020 0:00
G	Italy	EPI_ISL_412973	B.1	Europe	Italy	29903	Human	38	Male	3/1/2020 0:00
GR	Switzerland	EPI_ISL_413020	B.1.1	Europe	Switzerland	29864	Human	37	Male	3/3/2020 0:00
GR	Switzerland	EPI_ISL_413021	B.1.1	Europe	Switzerland	29875	Human	25	Male	3/3/2020 0:00
GR	Switzerland	EPI_ISL_413022	B.1.1	Europe	Italy	29845	Human	27	Male	3/3/2020 0:00
GR	Switzerland	EPI_ISL_413023	B.1.1	Europe	Italy	29788	Human	26	Female	3/3/2020 0:00
GR	Switzerland	EPI_ISL_413024	B.1.1	Europe	Switzerland	29871	Human	30	Male	3/3/2020 0:00
GR	United Kingdom	EPI_ISL_413221	B.1.1	Europe	Italy	29782	Human	51	Male	3/4/2020 0:00
G	Italy	EPI_ISL_413489	B.1	Europe	Italy	29887	Human	38	Female	3/5/2020 0:00
GR	Nigeria	EPI_ISL_413550	B.1.1	Europe	Italy	29759	Human	?	Male	3/6/2020 0:00
G	United Kingdom	EPI_ISL_413555	B.1	Europe	Italy	29782	Human	57	Male	3/7/2020 0:00
GR	United Kingdom	EPI_ISL_413556	B.1.1.1	Europe	Italy	29782	Human	26	Male	3/7/2020 0:00
GR	Netherlands	EPI_ISL_413565	B.1.1	Europe	Italy	29703	Human	?	?	3/7/2020 0:00
G	Netherlands	EPI_ISL_413566	B.1	Europe	Netherlands	29786	Human	?	?	3/7/2020 0:00
GR	Netherlands	EPI_ISL_413569	B.1.1	Europe	Italy	29786	Human	?	?	3/7/2020 0:00
GR	Netherlands	EPI_ISL_413570	B.1.1	Europe	Netherlands	29786	Human	?	?	3/7/2020 0:00
GR	Netherlands	EPI_ISL_413571	B.1.1	Europe	Netherlands	29786	Human	?	?	3/7/2020 0:00
GH	Netherlands	EPI_ISL_413572	B.1	Europe	Netherlands	29786	Human	?	?	3/7/2020 0:00
GR	Netherlands	EPI_ISL_413574	B.1.1	Europe	Netherlands	29786	Human	?	?	3/7/2020 0:00
G	Netherlands	EPI_ISL_413575	B.1	Europe	Netherlands	29786	Human	?	?	3/7/2020 0:00
GR	Netherlands	EPI_ISL_413579	B.1.1	Europe	Netherlands	29786	Human	?	?	3/7/2020 0:00
GR	Netherlands	EPI_ISL_413584	B.1.1	Europe	Netherlands	29786	Human	?	?	3/7/2020 0:00
GR	Netherlands	EPI_ISL_413587	B.1.1	Europe	Netherlands	29786	Human	?	?	3/7/2020 0:00
G	Netherlands	EPI_ISL_413588	B.1	Europe	Netherlands	29786	Human	?	?	3/7/2020 0:00
G	Netherlands	EPI_ISL_413589	B.1	Europe	Netherlands	29786	Human	?	?	3/7/2020 0:00
G	Netherlands	EPI_ISL_413591	B.1	Europe	Netherlands	29786	Human	?	?	3/7/2020 0:00
GH	Taiwan	EPI_ISL_413592	B.1	Europe	Italy	29870	Human	66	Female	3/7/2020 0:00
G	Luxembourg	EPI_ISL_413593	B.1.5	Europe	Italy	29786	Human	?	?	3/7/2020 0:00
G	Finland	EPI_ISL_413602	B.1	Europe	Italy	29808	Human	40	Male	3/8/2020 0:00
GR	Finland	EPI_ISL_413604	B.1.1	Europe	Italy	29777	Human	59	Male	3/8/2020 0:00
GR	Portugal	EPI_ISL_413647	B.1.1	Europe	Italy	29885	Human	60	Male	3/9/2020 0:00
G	Portugal	EPI_ISL_413648	B.1.74	Europe	Portugal	29837	Human	33	Male	3/9/2020 0:00
G	Switzerland	EPI_ISL_413996	B.1	Europe	Switzerland	29872	Human	70	Male	3/10/2020 0:00

Fig-4: Clade and Haplogroup data for non-exact match sequences of S gene of SARS-COV-2

Results for Database Comparison

On 11/18/2020, the NCBI database containing 41,861 sequences and the GISAID database containing 42329 sequences were BLASTed against each other to get an account of the sequences that are there in both the databases, but with their different

accession numbers. The BLAST output displays the query sequence from NCBI with its equivalent GISAID sequence. Further the NCBI sequences may be linked to their respective SRA accessions by using the metadata information available on the SRA website. In this way, the GISAID and NCBI accessions can be mapped to the SRA accessions which will help improve the data quality by filling in the gaps in cases of improperly sequenced data.

Table 13 – Number of Sequences in the GISAID, NCBI and SRA databases used for analysis

GISAID	NCBI	SRA
42329	41861	113217

```
Query= MN938384.1 |Severe acute respiratory syndrome coronavirus 2 isolate
2019-nCoV_HKU-SZ-002a_2020, complete genome|China

Length=29838
                                                                    Score   Total  Query    E     Max
Sequences producing significant alignments:                        (Bits)   Score  cover   Value  Ident

hCoV-19/Shenzhen/HKU-SZ-002/2020|EPI_ISL_406030|2020-01-10          59644    59644  100%   0.0    100%

Query= MN975262.1 |Severe acute respiratory syndrome coronavirus 2 isolate
2019-nCoV_HKU-SZ-005b_2020, complete genome|China

Length=29891
                                                                    Score   Total  Query    E     Max
Sequences producing significant alignments:                        (Bits)   Score  cover   Value  Ident

hCoV-19/Shenzhen/HKU-SZ-005/2020|EPI_ISL_405839|2020-01-11          59750    59750  100%   0.0    100%


Query= MN908947.3 |Severe acute respiratory syndrome coronavirus 2 isolate
Wuhan-Hu-1, complete genome|China

Length=29903
                                                                    Score   Total  Query    E     Max
Sequences producing significant alignments:                        (Bits)   Score  cover   Value  Ident

hCoV-19/Wuhan/Hu-1/2019|EPI_ISL_402125|2019-12-31                   59774    59774  100%   0.0    100%
```

Fig - 5: Result of performing a BLAST comparison between the GISAID and NCBI database

Figure 5 shows some examples of the BLAST code that displays the equivalent GISAID accession number for the NCBI accession numbers. While this method confirmed the presence of equivalent sequences in the two databases, there were only a

very few instances with clear results like the ones mentioned in Fig-5. One of the main problems associated with the method was the presence of gaps and 'N' in the sequences from GISAID database that gave rise to multiple hits when a BLAST search was performed with the NCBI sequences. Another problem was the lack of clarity in a few accession numbers for sequences in the GISAID website. There were sequences that contained 'hyphens' or 'dots' in places of numbers making it tough to assign those sequences to a particular sequence on the NCBI database. Because of these problems, the BLAST method did not work very well and so we used a Python script to analyze the metadata of the SRA sequences to link the databases.

To obtain the equivalent SRA accession numbers, the metadata for all the sequences on the SRA database was analyzed using a Python script to check for NCBI and GISAID accession numbers associated with the SRA accession numbers. As providing the accession numbers are up to the user's discretion while uploading the SRA data on NCBI website, it was not possible to find the equivalent SRA accession numbers for most sequences. However, we curated a table containing the metadata for all the available sequences which can be found here.

Run	AvgSpot	BioProject	Instrument	Organism	Platform	ReleaseDa	geo_loc_name_	Collection_Date	SRA_accessio	Host	Genbank_	Gisaid_id
SRR12245326	492	PRJNA631061	Illumina MiSeq	SARS-CoV-2	Illumina	7/17/2020	USA:CA	3/15/2020	SRP267191	Homo sapiens	MT614462	EPI_ISL_429807
SRR12245319	493	PRJNA631061	Illumina MiSeq	SARS-CoV-2	Illumina	7/17/2020	USA:CA	3/15/2020	SRP267191	Homo sapiens	MT614468	EPI_ISL_413455
SRR12245309	496	PRJNA631061	Illumina MiSeq	SARS-CoV-2	Illumina	7/17/2020	USA: MA	3/14/2020	SRP267191	Homo sapiens	MT614477	EPI_ISL_427515
SRR12245305	298	PRJNA631061	Illumina MiSeq	SARS-CoV-2	Illumina	7/17/2020	USA: MA	3/15/2020	SRP267191	Homo sapiens	MT614481	EPI_ISL_416491
SRR12245304	493	PRJNA631061	Illumina MiSeq	SARS-CoV-2	Illumina	7/17/2020	USA: MA	3/15/2020	SRP267191	Homo sapiens	MT614482	EPI_ISL_417878
SRR12245293	492	PRJNA631061	Illumina MiSeq	SARS-CoV-2	Illumina	7/17/2020	USA: MA	3/15/2020	SRP267191	Homo sapiens	MT614492	EPI_ISL_416661

Fig - 6: Mapping SRA, NCBI and GISAID accession numbers

By using the metadata of the SRA sequences, 944 GISAID sequences were linked to the SRA accession numbers and 4547 NCBI accession numbers were linked to the SRA accession numbers. The only reason for such limited number of sequences being linked is

that NCBI does not require the researchers who submit their SRAs to mention the corresponding NCBI accession. The lack of this information makes it impossible to link all the NCBI sequences to corresponding SRA and GISAID sequences.

Results for Viral Quasi Species

The Illumina and Oxford Nanopore sequencing data was analyzed to understand the course of mutations in a septuagenarian patient affected by SARS-CoV-2 [29]. This was done in an effort to understand if the patient was affected by a single strain of SARS-CoV-2 or if the infection was caused by a mixture of strains. The sequencing data was uploaded to Geneious Prime and the frequency of mutations relative to the SARS-CoV-2 Reference Genome (NC_045512) were recorded for both Illumina and Oxford Nanopore data which can be found here. There was a total of 62 mutations, some of which were fixed right from the first time point, a few others on the course of getting fixed with a gradual increase in the number of mutations between one time point and the next. There were also a few mutations that showed up suddenly in the final two time points.

We will focus primarily on the mutations clearly identified using the Illumina sequencing data. First, there were 11 mutations between nucleotides 21600-25033 located in the S gene. In the 23403 position of the nucleotide, we can see the dominance of the infamous D614G mutation. Other than that, the mutation at 23731 position of the nucleotide also appears to be a fixed mutation.

SNO	POSITION	MUTATION	SRA20-SRR13206521	5/13/2020	SRA3- SRR13206484	6/22/2020	SRA1-SRR13206482	7/5/2020	SRA15-SRR13206496	8/9/2020	SRA10-SRR13206491	8/19/2020	SRA5-SRR13206486	8/21/2020
			FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS
1	21600	G-T	T-2, G-734	736	T-2, G-2237	2239	T-3, A-4,G-1392	1399	T-8,G-58	66	T-33, C-1, G-503	537	T-1303, G-45, C-2	1350
2	21635	C-A,C-T	A-1, C-465	466	A-1, T-1, C-1197	1199	A-3,C-883	886	0	0	0	0	A-108,T-35, G-3, C-595	741
3	21752	U-G	G-2, A-1,T-1652	1657	G-0, A-4, T-992	996	0	0	0	0	G-50, C-1, T-738	789	G-589, T-25, A-2	616
4	22160	U-C	C-348, T-55	403	0	0	C-2, T-2137	2139	C-10, T-91	101	C-228, T-581	809	C-18, T-1413	1431
5	22524	A-G	G-0, C-3,A-1390	1393	G-0, T-1, A-1331	1332	G-924, A-1109	2035	G-1, C-1, A-484	486	0	0	0	0
6	22550	C-T	T-4, A-2, C-1368	1394	T-2, C-1328	1330	T-0,C-1,A-2038	2039	T-3, C-482	485	T-56, C-848	904	T-1133, C-38	1171
7	23403	A-G	G-2342,A-1,C-1,T-2	2346	G-74, A-6	80	G-2030, T-1	2031	G-3177,T-1	3178	G-1162, A-1,T-1	1164	G-1251, A-2	1253
8	23642	G-T	T-203, A-11, C-1, G-2713	2933	T-1, C-1341	1342	T-203, A-3,C-3,C-2549	2758	T-55, G-1599	1654	T-72, C-2, G-1643	1717	T-25, A-2, G-1012	1039
9	23650	U-C	C-148, A-20, G-2, T-2757	2927	C-46, A-2, G-3, T-1883	1934	C-140, A-6, G-6, T-2599	2751	C-40, A-4,G-3	1608	C-42, A-5, T-1671	1718	C-28, A-3, T-1008	1039
10	23731	C-T	T-1528, C-3	1531	T-970, A-2	972	T-1403, A-1,G-1	1405	T-833, A-1	834	T-857, C-4, G-1	862	T-1133, C-38	1171
11	25033	G-A	A-5,G-3215	3220	A-2, C-1, T-1, G-1714	1718	0	0	A-78,T-1, G-1516	1595	A-116, T-1, G-1510	1627	A-1037, G-41	1388

Fig - 7: List of mutations in the S gene

In addition to the two fixed S gene mutations, there were mutations in positions 241 (C to T), 1620 (A to G), 3037(C to T), 4002 (C to T), 6604 (A to G), 10097(G to A), 13236 (C to T), 14408 (C to T), 27059(C to T), 28881(G to A), 28882 (G to C), and 28883(G to T) that all appeared to be fixed in the SARS-CoV-2 sequences taken at six time points from the patient as shown in Fig. 7.

There are a few mutations that are in the course of getting fixed by gradually increasing in numbers between one point and the next as shown in Fig. 8. In addition to this, there are a few mutations that suddenly appear in the final two time points and dominate the other bases at the last time point. There are also a few cases shown in Fig. 10 where mutations that were initially dominant appear to disappear or gradually diminish.

SNO	POSITION	MUTATION	SRA20-SRR13206521	5/13/2020	SRA3- SRR13206484	6/22/2020	SRA1-SRR13206482	7/5/2020	SRA15-SRR1320649	8/9/2020	SRA10-SRR13206491	8/19/2020	SRA5-SRR13206486	8/21/2020
			FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS
1	241	C-T	T-3210, G-1, C-2	3213	T-971	971	T-290, C-1	291	T-2947, C-1	2948	T-1549	1549	T-1283, G-1	1284
2	1620	A-G	G-2067, T-5, A-3	2075	G-1987, C-1, T-1	1989	G-1862, T-3	1865	G-3431	3431	G-1498, A-2, T-1	1501	G-1378,A-1	1379
3	3037	C-T	T-1903	1903	T-1580, A-1	1581	T-1858, A-1	1859	T-1604, C-1	1605	T-1430, A-1	1431	T-947	947
4	4002	C-T	T-447	447	T-1036, G-1	1037	T-1140,A-1	1141	T-109, C-1	110	T-800, C-4	804	T-1091	1091
5	6604	A-G	G-713, T-2	713	G-964	964	G-788,A-1	789	G-314	314	G-563	563	G-12	12
6	13536	C-T	T-843, A-2, G-1	846	T-956, C-1	957	T-1741, A-3, C-1	1745	T-181	181	T-1236, A-2, G-1	1239	T-2132, A-2,G-1	2135
7	14408	C-T	T-3035, A-1,C-1	3037	T-1878	1878	T-2208,G-2, A-1	2211	T-1224, G-1, C-1	1226	T-1454	1454	T-1009	1009
8	23403	A-G	G-2342,A-1,C-1,T-2	2346	G-74, A-6	80	G-2030, T-1	2031	G-3177,T-1	3178	G-1162, A-1,T-1	1164	G-1251, A-2	1253
9	23731	C-T	T-1528, C-3	1531	T-970, A-2	972	T-1403, A-1,G-1	1405	T-833, A-1	834	T-857, C-4, G-1	862	T-1133, C-38	1171
10	27059	C-T	T-1314, A-1, G-1, C-1	1317	T-1777	1777	T-3747, A-1	3748	T-726	726	T-1646, A-2	1648	T-3256,C-3	3259
11	28881	G-A	A-3288, C-5, G-2	3295	A-2129, G-1	2130	A-3288, C-5, G-2	3295	A-3288, C-5, G-2	3295	A-858,C-2,T-2	862	A-3288, C-5, G-2	3295
12	28882	G-A	A-3289, C-3, G-2, T-1	3295	A-2126, C-1, G-3	2130	A-3289, C-3, G-2, T-1	3295	A-3289, C-3, G-2, T-1	3295	A-854,C-1,G-1,T-4	860	A-3289, C-3, G-2, T-1	3295
13	28883	G-C	C-3292, G-1, T-1	3294	C-2128, A-2	2130	C-3292, G-1, T-1	3294	C-3292, G-1, T-1	3294	C-861,G-1	862	C-3292, G-1, T-1	3296

Fig - 8: List of fixed mutations in six different time-points in the Illumina Sequencing data

SNO	POSITION	MUTATION	SRA20-SRR13206521	5/13/2020	SRA3- SRR13206484	6/22/2020	SRA1-SRR13206482	7/5/2020	SRA15-SRR13206496	8/9/2020	SRA10-SRR13206491	8/19/2020	SRA5-SRR13206486	8/21/2020
			FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS
1	2236	U-C	C-7,G-2,T-1122	1131	C-2, G-2, T-1533	1537	C-2, G-3, A-1,T-1357	1363	C-116, T-515, A-2	633	T-935, A-1, C-82	1018	C-1733, T-50	1783
2	10097	G-A	A-185, G-22	207	A-895, T-1, G-197	1093	A-355,G-84	439	A-59, G-3	62	A-159, G-56	215	A-614,G-111	725
3	11770	A-G	G-4,T-1,C-1,A-3055	3061	G-4, C-1, T-2, A-2362	2369	G-1,T-1,A-2908	2910	G-372, A-467	839	G-99,-1, A-1591	1691	G-1867, A-198, T-1	2066
4	25033	G-A	A-5,G-3215	3220	A-2, C-1, T-1, G-1714	1718	0	0	A-78,T-1, G-1516	1595	A-116, T-1, G-1510	1627	A-1037, G-41	1388
5	26333	C-T	T-4, C-1792	1796	T-2, A-1, C-1077	1080	T-0, A-1,C-1644	1645	T-396, C-633	1029	T-67,C-777	844	T-1059, C-28	1087
6	27618	U-C	0	0	0	0	C-1, A-2,T-2240	2243	C-22, T-46	68	C-52,A-1,T-353	406	C-1327,T-35	1362

Fig - 9: List of mutations that show a gradual increase indicating that they maybe in the course of getting fixed in the six different time-points in the Illumina Sequencing data

SNO	POSITION	MUTATION	SRA20-SRR13206521	5/13/2020	SRA3- SRR13206484	6/22/2020	SRA1-SRR13206482	7/5/2020	SRA15-SRR13206496	8/9/2020	SRA10-SRR13206491	8/19/2020	SRA5-SRR13206486	8/21/2020
			FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS
1	5590	U-C	0	0	C-0,G-1,T-1511	1512	C-2, T-851	853	0	0	C-54, T- 597	651	C-885, T-38, A-1	924
2	13527	U-C	C-3,A-1,T-842	846	C-2, T-955	957	C-2, A-2, G-6, 1736	1746	C-1,A-1.T-2,G-6534	6538	C-1,G-2754	2755	C-1101, T-1023	2133
3	14776	G-T	T-0,A-6,G-3583	3589	T-2, G-2085	2087	T-5,G-3507	3514	0	0	T-42,G-2273	2317	T-1369, G-1321	2694
4	21600	G-T	T-2, G-734	736	T-2, G-2237	2239	T-3, A-4,G-1392	1399	T-8,G-58	66	T-33, C-1, G-503	537	T-1303, G-45, C-2	1350
5	15814	G-A	A-5,G-1707	1712	A-0, T-2, G-992	994	A-2,G-1430	1432	0	0	A-67,G-895	962	A-593, G-19	612
6	21752	U-G	G-2, A-1,T-1652	1657	G-0, A-4, T-992	996	0	0	0	0	G-50, C-1, T-738	789	G-589, T-25, A-2	616
7	22550	C-T	T-4, A-2, C-1368	1394	T-2, C-1328	1330	T-0,C-1,A-2038	2039	T-3, C-482	485	T-56, C-848	904	T-1133, C-38	1171
8	26634	G-T	T-1, A-3, C-1865	1869	T-0, A-1, G-990	991	0	0	0	0	T-6, A-12,G-984	1002	G-331, T-132	483
9	26647	G-A	A-1, G-1870	1871	0	0	0	0	0	0	0	0	G-335, A-133	487

Fig - 10: List of mutations that emerge suddenly in the final time-points

SNO	POSITION	MUTATION	SRA20-SRR13206521	5/13/2020	SRA3- SRR13206484	6/22/2020	SRA1-SRR13206482	7/5/2020	SRA15-SRR13206496	8/9/2020	SRA10-SRR13206491	8/19/2020	SRA5-SRR13206486	8/21/2020
			FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS	FREQUENCY	READS
1	5406	U-G	G-677, T-340	1017	G-1, T-1034	1035	G-1, T-1091	1092	0	0	G-225, A-2, T-576	803	G-22, T-950	972
2	12043	C-T	T-2080, A-3, C-2002	4085	T-0, A-3, C-2140	2143	T-1, A-1	1139	T-2, A-3, G-2, C-4504	4511	T-586, A-2, G-1, C-1534	2123	0	0
3	20150	A-G	G-536, A-44	580	G-4,A-360	364	0	0	0	0	G-69, A-177	246	G-7, A-362	369
4	22160	U-C	C-348, T-55	403	0	0	C-2, T-2137	2139	C-10, T-91	101	C-228, T-581	809	C-18, T-1413	1431
5	27408	U-C	C-1317, A-1, G-1, T-737	2056	0	0	C-2,T-1687	1689	C-169, A-1, T-2527	2697	C-281, G-3,T-764	1048	C-14,T-552	566
6	27459	G-T	T-637,A-1,G-1,C-1316	2055	T-1, G-1822	1823	T-3,G-2767	2770	T-2055,A-1,T-2055	2699	T-667, G-555	1222	T-1,G-1239	1240

Fig - 11: List of mutations that are initially fixed but disappear at later time points

In order to identify the viral quasi species and to find whether a mixture of strains have infected the patient [29], the course of the mutations were tracked. The mutations were divided into four categories- 1) Mutations that are fixed from the first time point (Fig – 8), 2) Mutations that showed a gradual increase from the earlier time points to the final time points indicating that they maybe in the course of getting fixed (Fig – 9), 3)Mutations that emerge suddenly in the final time points which were not present in the parent strain in the earlier time points (Fig – 10), 4) Mutations that are initially fixed in the parent strain but disappear or diminish in the subsequent time points. Certain time

points contain zero indicating that there was no data available to analyze in those time points.

Results for Next Generation Sequencing

Of the two next generation sequencing techniques used for the longitudinal case study above, Illumina was better as the data was clear which made the analysis easy. The long reads of Oxford Nanopore sequences made it difficult for importing them into Geneious Prime. The Illumina sequences aligned in a better way compared to the Nanopore sequences making the analysis easier.

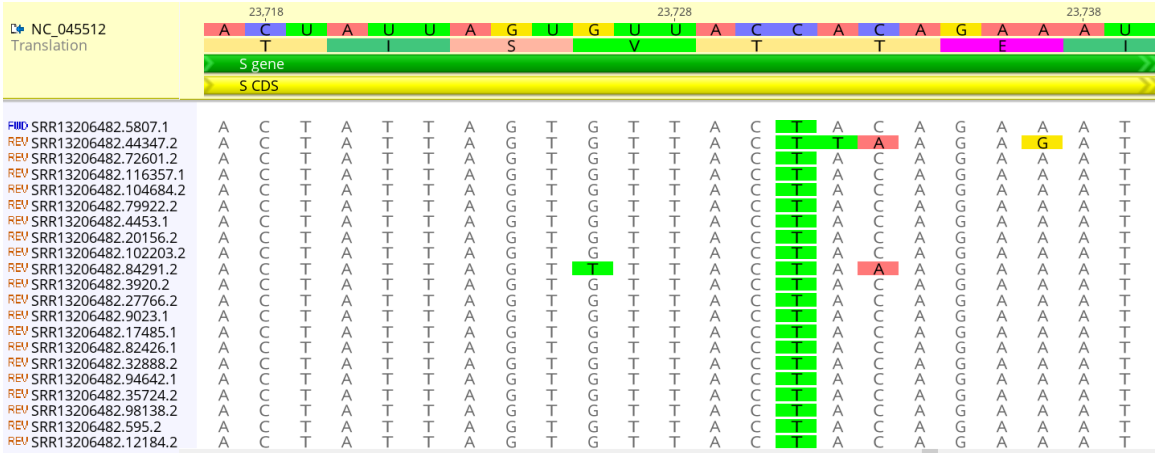

Fig. 12: Geneious plot showing Illumina reads of the time point 7/5/2020 between nucleotides 23717 and 23739

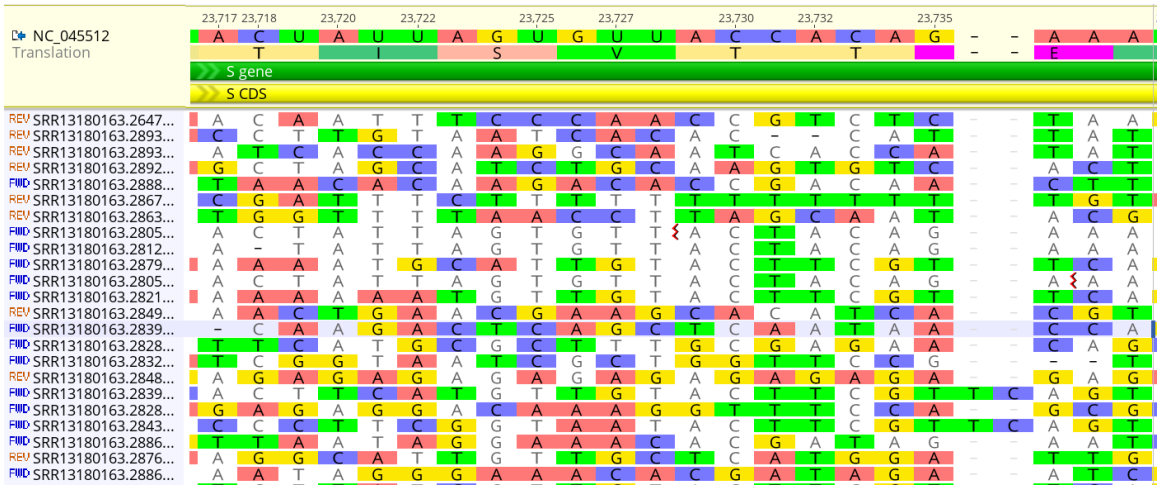

Fig. 13: Geneious plot showing Oxford Nanopore reads of the time point 7/5/2020 between nucleotides 23717 and 23739

Clarity of data is one of the key factors that helps in analysis and interpretation of data. While the fixed mutation (C to T) in position 23732 (Fig – 12) can be very clearly identified using the Illumina data assembled using Geneious Prime, the Oxford Nanopore data does not give a clear picture. While it looks like there is a C-T mutation in position 23731 (Fig – 13), it is not clear and evident as the reads have not assembled well with the reference sequence for SARS-CoV-2. The original study that analyzed the patient data for specific variants using Oxford Nanopore and Illumina sequencing technologies has concluded that the Oxford Nanopore technology is not suitable for detecting minority variants.

Chapter 4

DISCUSSION

Nucleic acid-based assays depend on efficient hybridization between the target sequence and the primers. Whenever there is a mismatch between the primer and template sequences, it might affect the efficiency of the PCR reaction by affecting the stability of the duplex and by affecting the extent to which amplification occurs in a reaction. Primers are designed for efficient amplification of different regions of the SARS-CoV-2 genome. The available primers for SARS-CoV-2 target the various genes coding for the nucleocapsid protein, the envelope protein, the membrane protein, etc., that can reveal key information about the viral genome. It has been documented that generally, for diagnostic assays for influenza and other viral strains, achieving a complete complementarity of the primer and probe sets with the template sequence is complicated [32]. There have been documented results to show that between two to four mismatches in the primer template pairs do not have a considerable impact on the RT-PCR. But, if the number of mismatches is between five and six, the PCR product yield would reduce in comparison to the homologous template [12]. In our study, we have analyzed the CDC, WHO and Korean primers to understand their priming efficiency. All three groups of primers have shown the presence of both transition and transversion mutations when they were aligned with 42329 SARS-CoV-2 sequences. If we consider the N1 and N2 CDC primers and probes, the N1 primer has 61 non-exact matches in the forward primer, 227 non-exact matches in the reverse primer and 857 non-exact matches in the probe. The N2 primer on the other hand has only 75, 33 and 148 non-exact matches for the forward primer, reverse primer and probe respectively making it a better primer than the N1-

CDC primer. When we observe the SeqLogo plot of the N2-CDC primer (Fig – 1), it is seen that there are 3 mutations fixed across the primer giving rise to the 75 non-exact sequences that were detected. Whenever the primer has more than two mismatches, the priming efficiency may reduce depending on the nature of the mutations. So, instead of just taking into account the number of non-exact matches, it is very important to analyze the SeqLogo plots to understand the nature and spread of the mutation, which reveals important details about the priming efficiency. In addition to the N1 and N2 primers, the CDC initially had N3 primers and probes. which showed the maximum number of non-exact matches during our initial analysis. In our initial analysis it had the greatest number of non-exact matches with 41910, 42216 and 42278 non-exact matches for the forward primer, reverse primer and probe respectively (data not shown). It was therefore not surprising that the CDC soon removed the N3 primers and probes from their website confirming that they showed a large number of primer-template mismatches. Overall, the WHO primers showed much better results than the CDC and Korean primers with respect to the number of mismatches. This might be because the researchers considered more conserved regions during primer design. Among the WHO primers, the E gene primers have 12, 34 and 50 non-exact matches for the forward primer, reverse primer and probe respectively making it one of the best primers of all the commercially available primers. When we take a look at the Seqlogo plot of the E- gene probe (Fig – 2) of WHO, we observe that even though there are 35 sequences that are non-exact matches, the mutation is significantly observed only in one location. As the mutations are not spread across different locations in the primer, it might not have a poor priming efficiency. One of the most important reasons for analyzing the Korean primers is that they are currently being

used by Virginia Tech to carry out RT-PCR. These primers have also been used in the State of Virginia initiative to develop a scalable and versatile test for diagnosing COVID-19 in rural communities using RT-PCR in collaboration with Virginia Tech [13]. It is notable that the Korean primer for E gene has a mutation fixed in 42096 (Fig – 3) sequences in the reverse primer. Even though this mutation gives a very high number of non-exact matches, the primer seems to not have many mutations apart from this making it efficient.

Investigating and tracking the mutations on the S gene is specifically important because the spike glycoprotein is used by SARS-CoV-2 to enter the host cells and attach to the ACE2(Angiotensin Converting enzyme 2) protein on the cell surface. A very common approach to track mutations is using the haplogroups of the people infected. The haplogroup table can help us find the first instance of the mutation and trace its spread to understand the course of the mutation. With our analysis, we have discovered that 34,987 out of the 42,329 sequences in the database carry mutations on S gene. In an effort to find the number of sequences bearing the D614G variant out of the 34,987 sequences, we investigated the haplogroup of all the mismatches. The analysis revealed that 30337 that belonged to clade G in the haplogroup table contained the D614G variant. To understand further the spread of this variant, we examined the first instance of the D614G variant of the SARS-CoV-2 and traced it back to the sequenced genome of a male patient from Germany that was submitted on 31st January 2020 (EPI_ISL_406862).

One of the common problems encountered during analyzing the sequencing data is the lack of clarity in the data as many sequences have gaps and sequencing errors at certain nucleotide positions that are critical for analysis. There have been cases with

ambiguity in the accession numbers of sequences when they were downloaded from GISAID with portions of the six-digit accession numbers being replaced by dots (.) or hyphens (-). So, it became very important to find a way to link equivalent sequences present in different databases like NCBI, GISAID and SRA. One major problem encountered while linking the databases was the non-availability of certain important information in the metadata as they were user- defined. For example, NCBI did not require the users who submitted SRA data to submit the NCBI accession numbers corresponding to the sequence. While a few of the users submitted this information, most of the entries did not contain the NCBI accession numbers making it very difficult to link the data. Initially we tried to BLAST the GISAID and NCBI databases together to find the equivalent sequences (Fig - 5). While this helped to find a few sequences, there was a problem with improperly sequenced data with large number of bases containing gaps or 'N'. So, we chose to analyze the metadata of all the sequences available for the SRA sequences which contained the information for NCBI and GISAID accession numbers in an effort to link the databases. Again, we encountered the problem that the user-defined data did not contain complete metadata information. We were successful in finding the accession numbers in all three databases for a few sequences (Fig – 6) but mostly we were able to link the SRA to either GISAID or NCBI accession numbers depending on the one mentioned in the user-defined metadata. This comprehensive database can be very useful as analyzing the metadata is the most efficient method to understand the course of mutations in a patient.

In an effort to identify viral quasi species in SARS-CoV-2 and to answer the question of whether a mixture of strains can affect an individual, we tracked the course of

mutations in a septuagenarian patient affected by SARS-CoV-2 [29]. We identified 11 mutations in the S gene of the patient of which one was the D614G mutation. The data reveals that there are multiple mutations in the S gene and in other genes of the patient that are in the course of getting fixed. Fixed mutations are those mutations that are present in most of the reads of the Next Generation sequencing data when they are mapped to the reference genome of SARS-CoV-2 (NC_045512). In addition to the 13 fixed mutations in the Illumina sequencing data found in the viral genome (Fig-7), there were a few mutations that were in the course of getting fixed (Fig - 8). These mutations have mostly not been present in the first time point at which the viral genome was sequenced but are present in subsequent time points indicating that they may be mutations that arose after the patient were admitted in the hospital. There are also some mutations that arose in some of the final time points at which the viral genome was sequenced before the patient passed away (Fig – 9). Analyzing the raw data gave rise to the understanding of the sudden spike of certain mutations and the mutations that gradually increased but were not present initially when the patient was admitted. These mutations might not belong to the original parent strain that infected the patient. There are also some mutations that were initially present in a high number in the sequencing data from the first time point at which the viral genome was sequenced but diminish quickly or even disappear in certain cases (Fig – 10). These maybe cases where another strain without the mutations might have infected the individual.

The raw data generated by two sequencing methods, namely the Illumina and Oxford Nanopore sequencing methods were analyzed to understand the course of mutations in the patient infected with SARS-CoV-2. We identified that the reads

generated by Illumina were easier to align and had a better alignment when compared to the Oxford Nanopore data. The Oxford Nanopore data did not align with the reference genome of SARS-CoV-2 accurately in Geneious leading to many reads showing mutations in places where they would otherwise not. The original study concluded that the Illumina sequencing technology was better than Oxford Nanopore technology in detecting all the variants as the Oxford Nanopore technology was not suitable for detecting minority variants. The original study speculates the presence of ‘spatially distinct viral populations’ as both population-genetic and small-animal studies have shown the lack of reassortment between influenza virus within a single host during an infection. [29]. This is in accordance with our study as we identified mutations that appeared suddenly in the final time points and we also saw the loss of mutations that were fixed in the initial parent strain.

Chapter 5

REFERENCES

[1] "On the origins of SARS-CoV-2," vol. 27, no. January, p. 41591, 2021, doi: 10.1038/s41591-020-01199-0.

[2] B. Hu, H. Guo, P. Zhou, and Z.-L. Shi, "Characteristics of SARS-CoV-2 and COVID-19," *Nat. Rev. Microbiol.*, vol. 19, no. 3, pp. 141–154, 2021, doi: 10.1038/s41579-020-00459-7.

[3] D. Hu *et al.*, "Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats," *Emerg. Microbes Infect.*, vol. 7, no. 1, p. 154, Sep. 2018, doi: 10.1038/s41426-018-0155-5.

[4] I. Astuti and Ysrafil, "Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An overview of viral structure and host response," *Diabetes Metab. Syndr.*, vol. 14, no. 4, pp. 407–412, 2020, doi: 10.1016/j.dsx.2020.04.020.

[5] X. Ou *et al.*, "Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV," *Nat. Commun.*, vol. 11, no. 1, p. 1620, 2020, doi: 10.1038/s41467-020-15562-9.

[6] D. Schoeman and B. C. Fielding, "Coronavirus envelope protein: current knowledge," *Virol. J.*, vol. 16, no. 1, p. 69, 2019, doi: 10.1186/s12985-019-1182-0.

[7] S. Elbe and G. Buckland-Merrett, "Data, disease and diplomacy: GISAID's innovative contribution to global health," *Glob. Challenges*, vol. 1, no. 1, pp. 33–46, 2017, doi: 10.1002/gch2.1018.

[8] L. E. Gralinski and V. D. Menachery, "Return of the Coronavirus: 2019-nCoV," *Viruses*, vol. 12, no. 2, p. 135, Jan. 2020, doi: 10.3390/v12020135.

[9] "s41579-020-00459-7 @ www.nature.com." [Online]. Available: https://www.nature.com/articles/s41579-020-00459-7.

[10] R. Diseases and D. Control, "2019-Novel Coronavirus (2019-nCoV) Real-time rRT-PCR Panel Primers and Probes 2019-Novel Coronavirus (2019-nCoV) Real-Time rRT-PCR Panel Primer and Probes Note : Oligonucleotide sequences are subject to future changes as the 2019-Novel Coronavirus," 2020.

[11] I. Pasteur, "Protocol : Real-time RT-PCR assays for the detection of SARS-CoV-2," pp. 1–3, 2020.

[12] J. Won *et al.*, "Development of a Laboratory-safe and Low-cost Detection Protocol

for SARS-CoV-2 of the Coronavirus Disease 2019," vol. 29, no. 2, pp. 1–13, 2020.

[13] Ceci A *et al.*, "Development and Implementation of a scalable and versatile test for COVID-19 diagnostics in rural communities," *medRxiv*, p. 2021.03.01.21252679, 2021, [Online]. Available: https://doi.org/10.1101/2021.03.01.21252679.

[14] C. Christopherson, J. Sninsky, and S. Kwok, "The effects of internal primer-template mismatches on RT-PCR: HIV-1 model studies," *Nucleic Acids Res.*, vol. 25, no. 3, pp. 654–658, Feb. 1997, doi: 10.1093/nar/25.3.654.

[15] N. Arnheim and H. Erlich, "POLYMERASE CHAIN REACTION STRATEGY," *Annu. Rev. Biochem.*, vol. 61, no. 1, pp. 131–156, Jun. 1992, doi: 10.1146/annurev.bi.61.070192.001023.

[16] M. Weidmann, U. Meyer-König, and F. T. Hufert, "Rapid detection of herpes simplex virus and varicella-zoster virus infections by real-time PCR," *J. Clin. Microbiol.*, vol. 41, no. 4, pp. 1565–1568, Apr. 2003, doi: 10.1128/jcm.41.4.1565-1568.2003.

[17] J. Legoff *et al.*, "Real-time PCR quantification of genital shedding of herpes simplex virus (HSV) and human immunodeficiency virus (HIV) in women coinfected with HSV and HIV," *J. Clin. Microbiol.*, vol. 44, no. 2, pp. 423–432, Feb. 2006, doi: 10.1128/JCM.44.2.423-432.2006.

[18] H. Kimura *et al.*, "Quantitative analysis of Epstein-Barr virus load by using a real-time PCR assay," *J. Clin. Microbiol.*, vol. 37, no. 1, pp. 132–136, Jan. 1999, doi: 10.1128/JCM.37.1.132-136.1999.

[19] S. Kwok *et al.*, "Effects of primer-template mismatches on the polymerase chain reaction: Human immunodeficiency virus type 1 model studies," *Nucleic Acids Res.*, vol. 18, no. 4, pp. 999–1005, 1990, doi: 10.1093/nar/18.4.999.

[20] E. Callaway, "The coronavirus is mutating - does it matter?," *Nature*, vol. 585, no. 7824, pp. 174–177, 2020, doi: 10.1038/d41586-020-02544-6.

[21] W. Tai *et al.*, "Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine," *Cell. Mol. Immunol.*, vol. 17, no. 6, pp. 613–620, 2020, doi: 10.1038/s41423-020-0400-4.

[22] B. Korber *et al.*, "Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus," *Cell*, vol. 182, no. 4, pp. 812-827.e19, 2020, doi: 10.1016/j.cell.2020.06.043.

[23] J. A. Plante *et al.*, "Spike mutation D614G alters SARS-CoV-2 fitness," *Nature*, vol. 592, no. 7852, pp. 116–121, 2021, doi: 10.1038/s41586-020-2895-3.

[24] N. L. Washington, S. White, K. M. S. Barrett, E. T. Cirulli, A. Bolze, and J. T. Lu, "S gene dropout patterns in SARS-CoV-2 tests suggest spread of the H69del/V70del mutation in the US," *medRxiv*, p. 2020.12.24.20248814, Jan. 2020, doi: 10.1101/2020.12.24.20248814.

[25] “coronavirus-test-variant-contagious-uk-gene @ www.theverge.com.” [Online]. Available: https://www.theverge.com/2020/12/30/22206522/coronavirus-test-variant-contagious-uk-gene.

[26] “0c67586dc254a17478d618c5d3925f1d94a23de6 @ www.ncbi.nlm.nih.gov.” [Online]. Available: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

[27] E. Domingo, J. Sheldon, and C. Perales, “Viral quasispecies evolution,” *Microbiol. Mol. Biol. Rev.*, vol. 76, no. 2, pp. 159–216, Jun. 2012, doi: 10.1128/MMBR.05023-11.

[28] H. E. V. Genotype, S. Agarwal, P. Baccam, R. Aggarwal, and S. Veerapu, “crossm Novel Synthesis and Phenotypic Analysis of Mutant Clouds for,” vol. 92, no. 4, pp. 1–17, 2018.

[29] S. Gayed *et al.*, “SARS-CoV-2 evolution during treatment of chronic infection,” vol. 65, no. December 2020, 2021, doi: 10.1038/s41586-021-03291-y.

[30] National Center for Biotechnology Information Bethesda MD USA, “Building a BLAST database with local sequences. https://www.ncbi.nlm.nih.gov/books/NBK279688/,” no. Md, 2008, [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK279688/.

[31] K. K. Dey, D. Xie, and M. Stephens, “A new sequence logo plot to highlight enrichment and depletion,” *BMC Bioinformatics*, vol. 19, no. 1, p. 473, 2018, doi: 10.1186/s12859-018-2489-3.

[32] R. Stadhouders, S. D. Pas, J. Anber, J. Voermans, T. H. M. Mes, and M. Schutten, “The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5’ nuclease assay,” *J. Mol. Diagn.*, vol. 12, no. 1, pp. 109–117, Jan. 2010, doi: 10.2353/jmoldx.2010.090035.

Chapter 6

APPENDIX

1. Mutations in the RT-PCR primers

Table 14- Exact and Non-exact matches for the N1-gene (Forward primer) of CDC

Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences
42329	42142	61	126

Table 15- Frequency of mutations in the non-exact match sequences for N-1 gene (Forward primer) of CDC

S.No	Position	Mutation	Frequency
1	3	C-A	4
		C-T	2
		C-Y	2
2	4	C-T	7
		C-Y	2
3	5	C-A	2
		C-T	7
		C-Y	1
4	6	C-A	4
		C-N	1
		C-M	1
5	7	A-G	3
6	8	A-C	1
7	14	G-T	22
		G-R	2
		G-K	1
8	15	C-Y	1
9	16	G-R	1
10	17	A-T	1
		A-R	1
		A-N	1

Table 16 - Exact and Non-exact matches for the N1-gene (Reverse primer) of CDC

Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences

42329	41979	227	123

Table 17 - Frequency of mutations in the non-exact match sequences for N-1 gene (Reverse primer) of CDC

S.No	Position	Mutation	Frequency
1	3	T-C	5
		T-A	1
2	4	G-C	8
3	5	G-A	1
		G-T	1
4	6	T-Y	1
		T-N	1
5	7	T-A	1
6	9	C-A	2
7	10	T-N	1
8	11	G-A	1
9	12	C-A	2
		C-T	1
10	13	C-A	3
11	14	C-T	1
12	15	G-A	12
		G-T	39
		G-R	1
13	16	T-C	2
		T-Y	1
		T-W	1
		T-N	2
14	17	T-N	1
15	18	G-A	5
		G-T	1
16	19	A-G	1
		A-T	1
17	20	A-T	1
18	21	T-C	114
		T-Y	4
		T-N	3
19	22	C-A	5
		C-Y	1
		C-N	1
		C-M	2
20	23	T-Y	1
		T-W	1

Table 18 - Exact and Non-exact matches for the N1-gene (Probe) of CDC

Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences
42329	41367	857	105

Table 19- Frequency of mutations in the non-exact match sequences for N-1 gene (Probe) of CDC

S.No	Position	Mutation	Frequency
1	1	A-N	1
2	2	C-Y	4
		C-N	1
3	3	C-T	707
		C-Y	27
		C-N	5
4	4	C-T	1
		C-Y	1
5	5	C-A	2
		C-T	2
		C-Y	1
6	6	G-A	3
		G-T	6
		G-N	1
7	7	C-T	3
		C-Y	2
8	8	A-C	1
		A-R	1
9	9	T-C	4
		T-G	2
10	11	A-G	2
11	12	C-T	4
		C-Y	3
12	13	G-A	12
		G-T	3
		G-R	1
		G-N	1
		G-K	1
13	14	T-A	2
14	17	G-T	19
		G-A	1
		G-K	3
15	18	G-T	20
		G-K	2

16	19	T-A	2
17	20	G-A	6
		G-C	1
18	21	G-T	8
		G-A	3
19	22	A-G	1
20	23	C-Y	1

Table 20 - Exact and Non-exact matches for the N2-gene (Reverse) of CDC

Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences
42329	42248	33	48

Table 21 - Frequency of mutations in the non-exact match sequences for N-2 gene (Reverse) of CDC

S.No	Position	Mutation	Frequency
1	4	C-M	3
2	6	A-R	1
3	7	C-A	1
		C-N	1
4	9	T-N	1
5	10	T-C	1
6	11	C-T	1
7	12	C-S	1
		C-N	2
8	13	G-A	17
		G-R	2
		G-N	1
9	18	G-R	1

Table 22 - Exact and Non-exact matches for the N2-gene (Probe) of CDC

Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences
42329	42100	148	81

Table 23 - Frequency of mutations in the non-exact match sequences for N-2 gene (Probe) of CDC

S.No	Position	Mutation	Frequency

1	7	T-C	6
2	8	G-T	3
3	9	C-N	5
4	10	C-T	6
		C-N	1
		C-Y	1
5	12	C-T	2
		C-Y	1
6	13	T-C	46
		T-Y	5
		T-M	1
7	15	G-A	2
		G-T	1
8	16	C-T	7
		C-Y	2
9	18	C-T	2
		C-N	1
10	19	T-C	2
		T-G	1
11	20	T-Y	1
12	21	C-Y	1
		C-N	1
13	22	A-T	1

Table 24 - Exact and Non-exact matches for the E- gene (Forward) of WHO

Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences
42329	42295	12	22

Table 25 - Frequency of mutations in the non-exact match sequences for E - gene (Forward) of WHO

S.No	Position	Mutation	Frequency
1	4	G-K	1
2	5	G-N	1
		G-T	1
3	6	T-C	1
4	7	A-G	1
5	10	T-Y	1
6	11	A-G	1
7	16	G-K	1
8	17	T-N	1
9	18	T-W	1

10	19	A-N	1
11	20	A-N	1
12	21	T-W	1
13	24	C-M	1

Table 26 - Exact and Non-exact matches for the E- gene (Reverse) of WHO

Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences
42329	42263	34	32

Table 27 - Frequency of mutations in the non-exact match sequences for E - gene (Reverse) of WHO

S.No	Position	Mutation	Frequency
1	5	T-N	1
2	6	G-A	3
		G-N	1
3	9	G-C	1
		G-A	1
4	10	C-G	1
		C-M	1
5	11	A-N	1
6	12	G-A	10
7	14	T-R	1
		T-N	1
8	15	C-Y	1
9	16	G-A	2
		G-S	1
		G-K	1
10	17	C-A	3
11	18	A-W	1
12	19	C-N	1

Table 28 - Exact and Non-exact matches for the E- gene (Probe) of WHO

Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences
42329	42236	50	43

Table 29 - Frequency of mutations in the non-exact match sequences for E - gene (Probe) of WHO

S.No	Position	Mutation	Frequency
1	2	C-T	1
2	7	G-N	1
3	9	C-T	38
4	13	C-T	1
5	14	T-Y	1
6	15	T-N	1
7	19	G-N	1
8	20	C-T	2
9	22	C-T	1
10	23	T-N	1
11	24	T-Y	1

Table 30- Exact and Non-exact matches for the N-gene (Forward) of WHO

Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences
42329	42242	45	42

Table 31 - Frequency of mutations in the non-exact match sequences for N - gene (Forward) of WHO

S.No	Position	Mutation	Frequency
1	3	C-Y	3
2	4	T-C	1
3	7	G-A	3
		G-R	3
		G-S	1
4	8	G-K	2
5	9	C-T	1
6	11	C-T	1
		C-N	1
7	12	C-A	4
		C-T	4
8	14	G-K	2
9	15	C-T	14
		C-N	2

Table 32 - Exact and Non-exact matches for the N-gene (Reverse) of WHO

Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences
42329	42113	132	84

Table 33 - Frequency of mutations in the non-exact match sequences for N - gene (Reverse) of WHO

S.No	Position	Mutation	Frequency
1	3	C-Y	3
2	4	T-C	1
3	7	G-A	3
		G-R	3
		G-S	1
4	8	G-K	2
5	9	C-T	1
6	11	C-T	1
		C-N	1
7	12	C-A	4
		C-T	4
8	14	G-K	2
9	15	C-T	14
		C-N	2

Table 34 – Exact and Non-exact matches for the N- gene (Probe) of WHO

Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences
42329	42111	132	86

Table 35 - Frequency of mutations in the non-exact match sequences for N - gene (Probe) of WHO

S.No	Position	Mutation	Frequency
1	3	T-N	1
2	4	T-N	1
		T-Y	1
3	5	C-T	2
		C-N	4
4	6	C-T	14
5	7	T-C	7

		T-A	1
		T-Y	1
6	12	A-C	1
		A-T	1
7	15	G-T	12
		G-N	2
		G-K	3
8	18	G-A	1
		G-T	1
		G-R	1
		G-K	1
9	20	C-T	4
		C-Y	1
10	21	G-A	10
		G-T	2
		G-N	1
11	22	G-T	2

Table 36 - Frequency of mutations in the non-exact match sequences for RdRp - gene (Forward) of WHO

Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences
42329	90	42199	40

Table 37 - Frequency of mutations in the non-exact match sequences for RdRp - gene (Forward) of WHO

S.No	Position	Mutation	Frequency
1	1	G-N	2
2	5	A-N	16
		A-G	5
		A-R	1
3	7	T-C	1
4	8	G-T	101
		G-N	2
		G-K	1
5	9	G-A	9
		G-K	1
6	12	A-C	1
		A-T	1
7	15	G-T	12
		G-N	2

		G-K	3
8	18	G-A	1
		G-T	1
		G-R	1
		G-K	1
10	21	G-A	10
		G-T	2
		G-N	1
11	22	G-T	2

Table 38 - Frequency of mutations in the non-exact match sequences for RdRp - gene (Reverse) of WHO

Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences
42329	42164	120	45

Table 39 - Frequency of mutations in the non-exact match sequences for RdRp - gene (Reverse) of WHO

S.No	Position	Mutation	Frequency
1	2	T-C	2
2	4	G-A	2
3	7	A-N	4
4	8	A-N	2
		A-T	2
5	12	T-N	4
6	17	T-C	2
		T-Y	2
7	19	G-A	2

Table 40 - Frequency of mutations in the non-exact match sequences for RdRp - gene (Probe) of WHO

Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences
42329	120	42059	150

Table 41 - Frequency of mutations in the non-exact match sequences for RdRp - gene (Probe) of WHO

S.No	Position	Mutation	Frequency

1	12	A-T	12
		A-C	42916
2	19	G-A	42207
		G-C	1
		G-R	1

Table 42 - Frequency of mutations in the non-exact match sequences for E1 - gene (Forward) of Korean Primers

Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences
42329	42198	46	85

Table 43 - Frequency of mutations in the non-exact match sequences for E1 - gene (Forward) of Korean Primers

S.No	Position	Mutation	Frequency
1	3	C-T	1
		C-Y	1
2	4	G-A	5
		G-T	1
		G-K	2
3	5	G-A	1
		G-C	1
		G-N	1
4	7	A-T	1
5	8	G-A	1
		G-K	1
6	10	G-C	1
		G-A	1
7	12	C-T	18
		C-Y	4
8	14	G-T	1
		G-K	1
9	15	G-T	1
		G-N	1
10	16	T-A	1
11	17	A-G	1
12	20	T-Y	1

Table 44 - Frequency of mutations in the non-exact match sequences for N2 - gene (Forward) of Korean Primers

Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences
42329	42032	111	186

Table 45 - Frequency of mutations in the non-exact match sequences for N2 - gene (Forward) of Korean Primers

S.No	Position	Mutation	Frequency
1	1	T-G	1
2	4	C-T	36
		C-Y	3
		C-N	1
3	5	A-N	2
4	8	C-T	28
		C-Y	1
		C-N	2
5	10	T-Y	1
6	12	G-N	1
		G-K	1
7	13	G-T	1
		G-N	1
8	14	C-T	2
		C-N	2
9	15	G-T	22
		G-S	1
		G-R	1
		G-N	2
		G-K	1
10	18	A-N	1
		A-M	1

Table 46 - Frequency of mutations in the non-exact match sequences for N2 - gene (Reverse) of Korean Primers

Total Number of Sequences in Database	No. of Exact Sequences	No. of non-exact Sequences	Number of Missing Sequences
42329	42195	117	17

Table 47 - Frequency of mutations in the non-exact match sequences for N2 - gene (Reverse) of Korean Primers

S.No	Position	Mutation	Frequency
1	2	G-R	3
2	3	T-C	1
3	4	A-N	1
4	5	G-A	15
		A-R	1
5	6	C-A	1
6	7	A-N	1
7	9	G-A	13
		G-R	10
8	10	A-G	1
9	12	T-A	3
		T-K	1
10	13	G-R	1
11	14	C-A	51
12	15	A-T	1
13	16	G-A	9
		G-T	1
14	17	A-N	1
15	20	T-W	1

2. Blast code used to evaluate mutations in the primers and S gene

PS C:\Users\> blastn -db latest_db_corrected -query C:/research/DATABASE/BLAST/sequences/VT_primers/E1_F_VT.fasta -task blastn -out RDRP3_R_CL_VT_thesis.txt -outfmt 0 -sorthits 3 -sorthsps 3 -num_descriptions 422329 -num_alignments 422329

3. Python code used to sort exact and non-exact match sequences

```
list1,list2,anomaly = [],[],[]
f = open("C:/Users/lavan/S2_F_CL_VT_thesis.txt")
count = 0
for i in f:
   i = i.replace(" ","")
   if("hCoV" in i and "100%" not in i[len(i)-5:] and len(i)>1 and ">" not in i):
      print(i.split("|")[1])
      if(len(i.split("|")[1]) == 14):
         list1.append((i.split("|")[1]))
      else:
         anomaly.append((i.split("|")[1]))
      count+=1
print(count)
```

```
if(len(anomaly)>0):
    print("File Anomaly detected !!! ",end="")
    print(anomaly)
from shutil import copy,move
from sys import exit
count = 0
for i in list1:
    src = "C:/Latest_db/"+i+".fasta"
print(src)
    dest =
"C:/Automated_Seq/VT_primers/S2_F_CL_VT_thesis/Non_exact/"
    source = src
    target = dest
    try:
        copy(source, target)
        count += 1
    except IOError as e:
        print("Unable to copy file. %s" % e)
        exit(1)
    except:
        print("Unexpected error:", sys.exc_info())
        exit(1)
print(count)
```