NAZARBAYEV
UNIVERSITY

# A high scale SARS-CoV-2 profiling by whole-genome sequencing using Oxford Nanopore Technology in Kazakhstan

Amina Amanzhanova
(B.Sc., Nazarbayev university)

A THESIS SUBMITTED
FOR THE DEGREE OF MASTER OF SCIENCE IN
BIOLOGICAL SCIENCES DEPARTMENT OF
BIOLOGY SCHOOL OF SCIENCE AND HUMANITIES
NAZARBAYEV UNIVERSITY

2022

Student: Amina Amanzhanova (name and signature)
(submission date) 08/04/2022

Student's Supervisor/Advisor: Dos Sarbassov (name and signature)
08/04/2022

# DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis. This thesis has also not been submitted for any degree in any university previously.

Amina Amanzhanova
(Candidate'd full name)
Date: 08.04.2022

# ACKNOWLEDGEMENTS

I would like to express my appreciation to my supervisor Dr. Dos Sarbassov, co-supervisor Dr. Ulykbek Kairov and colleagues in the National Laboratory of Astana, who provided an opportunity, encouragement, and support to perform and complete this thesis work, especially for valuable suggestions, guidance, and expertise to obtain research experience. I am grateful for the chance to be part of your research community.

In addition, I would like to address my gratitude to the Faculties of Biology Department (SSH), specifically, Dr. Ferdinand Molnar, who generously provided his licensed Desmond program, constructive feedback, and suggested Bioinformatics analysis tools, course coordinators, Dr. Tursonjan Tokay, Dr. Otilia Nuta, Dr. Tri Pham, and all other professors whose classes I have taken during the Master's program.

Finally, I would like to acknowledge the support and help of my family, particularly, my daughter Alaina who motivated me to accomplish my Master's degree.

# TABLE OF CONTENTS

**TITLE PAGE** **I**
**DECLARATION** **II**
**ACKNOWLEDGEMENTS** **III**
**TABLE OF CONTENTS** **IV**
**SUMMARY** **V**
**LIST OF TABLES** **VI**
**LIST OF FIGURES AND ILLUSTRATIONS** **VII**
**ABBREVIATIONS** **VII**
**INTRODUCTION** **1**
1.1 Global Pandemic **1**
1.1.1 COVID-19 pandemic in Kazakhstan **1**
1.2 Viral structure **2**
1.2.1 Spike **3**
1.2.2 Site-Directed Mutator (SDM) analysis **4**
1.2.3 Molecular dynamics (MD) analysis **4**
1.3 Taxonomy **6**
1.4 Epidemiology and surveillance **6**
1.5 Viral genome sequencing **8**
1.5.1 Sequencing by Oxford nanopore technology (ONT) **11**
**MATERIALS & METHODS** **13**
2.1 Sample collection **13**
2.2 ONT Library Preparation and Sequencing **13**
2.3 Software setup and installation **14**
2.4 Demultiplexing of multiple barcoded samples **15**
2.5 Variant calling and phylogenetic profiling **15**
2.6 SDM analysis **16**
2.7 Molecular dynamics simulation analysis **16**
**AIMS OF THE THESIS PROJECT** **17**
**RESULTS** **18**
4.1 The 77 genetic sequences upload on GISAID **18**
4.2 Phylogenetic tree **19**
4.3 Lineage distribution **19**
4.4 Frequency of mutations in SARS-CoV-2 **20**
4.5 Spike glycoprotein of EPI_ISL_5465615 **20**
4.6 Protein-Ligand Molecular Dynamics analysis **22**
4.7 Protein-Protein Affinity Change Upon Mutation **23**
**DISCUSSION** **27**
**CONCLUSION** **32**
**REFERENCE LIST** **33**
**APPENDICES** **37**

# SUMMARY

Severe acute respiratory syndrome (SARS-CoV-2) is responsible for the worldwide pandemic COVID-19. The original viral whole-genome was sequenced by a high-throughput sequencing approach from the samples obtained from Wuhan, China. Real-time gene sequencing is the main parameter to manage viral outbreaks because it expands our understanding of virus proliferation, spread, and evolution. Primarily, it relies on the prompt sequencing technique of viral genome immediately from the clinical specimens without the necessity of viral culturing step. Whole-genome sequencing is critical for SARS-CoV-2 variant surveillance, the development of new vaccines and boosters, and the representation of epidemiological situations in the country. A significant increase in the number of COVID-19 cases confirmed in August 2021 in Kazakhstan facilitated a need to establish an effective and proficient system for further study of SARS-CoV-2 genetic variants. The SARS-CoV-2 whole-genome was sequenced using the SARS-CoV-2 ARTIC protocol (EXP-MRT001) by Oxford Nanopore Technologies at the National Laboratory of Astana, Kazakhstan. The 96 samples kindly provided by the Republican Diagnostic Center (RDC) were collected from individuals in Nur-Sultan city diagnosed with COVID-19 in August 2021 using real-time reverse transcription-quantitative polymerase chain reaction (RT-qPCR). All samples had a cycle threshold (Ct) value below 20 with an average Ct value of 13.88. The genomic coverage of the sequenced samples was ~99.8% and 77 out of 96 (80%) samples that passed quality control are deposited in the Global initiative on sharing all influenza data (GISAID). The AY.122 (n=69), B.1.617.2 (n=6), and AY.89 (n=1) delta lineages, and one sample B.1.637 belongs to a separate lineage corresponding to Iota were detected. To the best of our knowledge, this is the first study of SARS-CoV-2 whole genome sequencing by the ONT approach in Kazakhstan. The performed work confirms that ONT sequencing is a robust technique for the monitoring of SARS CoV-2 new variants by its whole genome sequencing. Also, established genomic sequences reveal the novel mutations that alter viral characteristics, particularly, the detected mutations are predicted to increase the spike protein stability by computational tools. The further high-throughput analysis and SARS-CoV-2 surveillance in the country are expected by the GridION at National Laboratory of Astana (NLA) at Nazarbayev University at a larger scale.

# LIST OF TABLES

**Table 1:** Average summary characteristics of sequenced SARS-CoV-2 samples by ONT platform

| GC-content | Depth | Ct value | Females | Males |
|---|---|---|---|---|
| 38.02% | 359.45X | 13.88 | 41 | 36 |

# LIST OF FIGURES AND ILLUSTRATIONS

**Figure 1:** SARS-CoV-2 whole-genome sequencing and analysis workflow.

**Figure 2:** A) The 77 SARS-CoV-2 sequences uploaded on GISAID platform. B) The example of EPI_ISL_5465615 sample information

**Figure 3:** Phylogenetic distribution of ONT sequenced SARS-CoV-2 genomes. A phylogenetic tree was generated by Nexstrain and the sequenced samples are identified according to Nextstrain nomenclature. The samples belonging to 21J (Delta) variants are highlighted with blue circles, whereas earlier origin variant 20C is highlighted with grey circle. Besides, the 21 K omicron variants detected in recent run in March 22 are indicated with red circles.

**Figure 4:** The proportion of clades and Pangolin lineages of n=77 sequenced SARS-CoV-2 samples. The AY.122 (n=69), B.1.617.2 (n=6), and AY.89 (n=1) delta lineages, and one sample B.1.637 belongs to a separate lineage corresponding to Iota were detected in this study.

**Figure 5:** Substitution and deletion mutation frequency in sequenced SARS-CoV-2 genomes. The graph illustrates that substitution mutations primarily occur at ORF1 and S protein region, while deletion mutations predominantly are found in S and ORF8 regions. The spike protein in viral genome is a primary region for mutations.

**Figure 6:** Spike glycoprotein (PDB: 6acc, EM 3.6 Angstrom) with RBD in upward conformation with amino acid changes identified in the query sequences shown as colored balls of sample EPI_ISL_5465615. B. Spike glycoprotein of EPI_ISL_5465615 (PDB: 6acj, EM 4.2 Angstrom) in complex with host cell receptor ACE2 (green ribbon).

**Figure 7:** Protein stability analysis by SDM. A) The protein stability analysis by SDM upon mutation of sequence EPI_ISL_5465615. A positive (blue) and negative (red) values correlate with mutation expected to be stabilizing and destabilizing accordingly. The results output suggests that the rare mutation Q271E slightly increases stability of the spike protein of $\Delta\Delta G=0.27$ kcal/mol and generally most mutations contribute to increase in protein stability. Overall predicted net pseudo $\Delta\Delta G$ is equal to 1.5 kcal/mol. B) The protein stability analysis of barcode 18 (omicron) suggests that overall $\Delta\Delta G$ of spike protein stability increased to 3 kcal/mol.

**Figure 8:** The molecular dynamics simulation output of ACE2-RBD protein interaction. RMSD of ACE2 receptor Cα atoms in accordance with the initial structure. A) ACE2-RBD protein (6MOJ) interaction. B) ACE2-RBD (EPI_ISL_5465615) interaction RMSD plot. C) ACE2-RBD (barcode 18 omicron) interaction RMSD plot.

# ABBREVIATIONS

| | |
|---|---|
| **SARS-CoV-2** | severe acute respiratory syndrome coronavirus 2 |
| **COVID-19** | coronavirus disease 2019 |
| **ICTV** | International Committee on Taxonomy of Viruses |
| **HCoV** | human coronaviruses |
| **NSP** | non-structural proteins |
| **ORF** | open reading frame |
| **ACE2** | angiotensin-converting enzyme 2 |
| **RBD** | receptor-binding domain |
| **NTD** | N-terminal domain |
| **CTD** | C-terminal domains |
| **MERS** | Middle East respiratory syndrome |
| **cDNA** | complementary DNA |
| **RT-qPCR** | real-time quantitative reverse transcription-polymerase reaction |
| **NGS** | next generation sequencing |
| **AMR** | antimicrobial resistance |
| **ONT** | Oxford nanopore technology |
| **SNV** | single nucleotide variants |
| **NLA** | National Laboratory of Astana |
| **GISAID** | Global Initiative on Sharing Avian Influenza Data |
| **SDM** | site directed mutator |
| **ΔΔG** | free energy difference |
| **MD** | Molecular dynamics |
| **RMSD** | Root mean square deviation |

# 1 INTRODUCTION

## 1.1 Global Pandemic

The global coronavirus disease 2019 (COVID-19) pandemic is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that is responsible for the severe acute respiratory syndrome (Wu et al., 2020)(Zhu et al., 2020) that led to the death of 5.8 million people by February 2022. As of February 15, 2022, there have been 412 351 279 confirmed cases of disease, counting 5 821 004 deaths, while as of 13 February 2022, overall, 10 227 670 521 vaccine doses have been administered according to WHO ("WHO Coronavirus (COVID-19) Dashboard" 2022). Originated in Wuhan, the capital of Hubei province in late December 2019, SARS-CoV-2 immediately propagated worldwide, causing infectious pneumonia. Initially, symptoms such as sore throat, fever, weakness, and respiratory distress primarily resembled viral pneumonia. Nevertheless, genomic examination of samples obtained from the infected patients confirmed the novel disease as coronavirus (2019-nCoV) pneumonia (Zhou et al., 2020). Then it was renamed SARS-CoV-2 by the International Committee on Taxonomy of Viruses (ICTV) due to its genetic relevance to earlier verified coronaviruses (Lu et al., 2020). To cease the ongoing COVID-19 pandemic, the research worldwide has focused on advancing basic and clinical studies of the viral infection, therapy, epidemiology, diagnostics, drug, and vaccine development.

### 1.1.1 COVID-19 pandemic in Kazakhstan

The disease has been exponentially spreading around the world, and the first SARS-CoV-2 case in Kazakhstan was confirmed to be on the 16th of March 2020, in Almaty city (Zhalmagambetov et al., 2020). To date, 5 551 314 deaths associated with COVID-19. As of January 2022, in general, 9.5 billion vaccine doses were administered as reported to WHO. Although the government implemented several lockdowns, strict quarantine regimes, and restrictions to prevent massive COVID-19 spread, gaps exist in the interpretation of the clinical and epidemiological characterization of the local pandemic. Previous nation-wide retrospective cohort study distinguished samples of five of the eight global SARS-CoV-2 clades detected in the early stages of pandemics in Kazakhstan. Besides, it was suggested that a unique lineage (B.4.1) arose independently in Kazakhstan. The genomic surveillance is critical in the representation of the genetic diversity of circulating SARS-CoV-2 that in terms reflects the clinical and epidemiological situation in the country (Yegorov et al., 2021). Currently, overall, there are 662 whole genome samples sequenced in Kazakhstan by February 2022 among 8,447,671 genome sequence submissions including n=77 of this study are available on Global Initiative on Sharing Avian Influenza Data (GISAID) platform repository. The scarce of the genomic sequencing limits the epidemiological and health care studies in the country. The SARS-CoV-2 whole genome sequencing by Oxford Nanopore Technology was performed for the first time in Kazakhstan by the National Laboratory of Astana (NLA) at Nazarbayev University (Nur-Sultan, Kazakhstan) and intended to perform further monitoring of viral circulation in the country.

## 1.2 Viral structure

SARS-CoV-2 is a positive-sense single-stranded RNA virus that is composed of approximately 30 kilobase pairs (Wu et al., 2020)(Zhu et al., 2020) and has variable open reading frames (ORFs), substantially resembling human coronaviruses (HCoVs)(Song et al., 2019). The viral genome encodes for structural proteins, to be specific nucleocapsid (N), membrane (M), envelope (E), spike (S), sixteen non-structural proteins, namely NSP1-NSP16, and nine accessory proteins—ORF3a, 3d, 6, 7a, 7b, 8, 9b, 14, and 10 (Yadav et al., 2021). Currently, research studies expand understanding of genomic organization, structural and non-structural proteins that may act as targets of novel drugs for clinical therapeutics (Yadav et al., 2021). The complete genome sequencing facilitated the advancement of RT-PCR assays for SARS-CoV-2 investigation to standardize the diagnostics of the COVID-19 outbreak (Lu et al., 2020).

### 1.2.1 Spike

Spike glycoprotein is a type I membrane protein that constitutes a trimer, attached to the viral membrane via its transmembrane domain, and it occupies the virion surface by its substantial ectodomain. Normally, spike binds to the angiotensin-converting enzyme 2 (ACE2) receptor found on the host cell and structurally rearranges to facilitate membrane fusion. It is highly glycosylated as its each promoter has 22N-linked glycosylation sites. The entire S protein of original Wuhan-Hu-1 strain is composed of 1273 amino acid residues, containing the signal peptide (N-terminus), the S1 receptor binding subunit, and the S2 fusion subunit. The S1 subunit is composed of the N-terminal domain (NTD), receptor-binding domain (RBD) and C-terminal domains (CTD1 and CTD2), whereas S2 subunit consists of fusion peptide (FP), fusion-peptide proximal region (FPPR), heptad repeat 1 (HR1), central helix (CH), connector domain (CD), heptad repeat 2 (HR2), transmembrane segment (TM) and the cytoplasmic tail (CT). The coronaviruses invade the host cells via fusion of viral envelope lipid bilayer and host membrane. The primary viral entrance is catalyzed by spike glycoprotein that is found on virion surface as an antigen and induces the immune response for antibody production, hence it is crucial to study spike protein for development of treatment, diagnostics, and vaccines against the COVID-19 (Zhang et al., 2020).

Besides tracking the viral spread, viral whole genome sequencing enables recognition of arising variants and predominant mutations responsible for virulence, importantly in spike protein region that is highly prone to mutations. Viral spike protein is responsible for infection initiation by facilitating the viral and host cell membrane fusion. According to cryo-EM analysis, spike protein has two distinct conformations, specifically prefusion and postfusion forms. Further studies found that S2 subunit of postfusion conformation was enriched by  N-linked glycans, indicating the defense mechanism against the host immune response by induction of non-neutralizing antibody response to avoid the host immune system and protect from external conditions (Kumar et al., 2021). Besides, spike mutation evolution can cause alterations in its surface that decrease efficiency of vaccine-induced antibodies. More extensive studies of spike protein are expected to expand current understanding of viral infection and immune escape of the SARS-CoV-2.

### 1.2.2 Site-Directed Mutator (SDM) analysis

One of the analyses considered in this study is examining the impact of viral mutation on protein stability that is crucial in understanding its role in disease development. Site Directed Mutator (SDM) is a computation tool that analyzes the changes of amino acid replacements found at certain structural environment that are accepted in homologous proteins of available 3D structures. The quantitative analysis is performed to predict the protein stability regarding the mutations. SDM calculates the change in thermal stability between reference and mutant protein, to be specific, the computational analysis should predict the effect of novel mutations on protein function by measuring the difference in free energies of denaturation of mutation and wild-type proteins. The protein stability is the difference in Gibbs free energy, G, between the folded and unfolded conformations. The larger and positive value of unfolded G represents higher stability of the protein to denaturation (Pandurangan et al. 2022).

### 1.2.3 Molecular dynamics (MD) analysis

Molecular dynamics (MD) simulation is a computational tool applied for analysis of conformational flexibility of molecules. It is important tool for analysis of physical features of the molecular conformation and macromolecular structure functions. Simulations provide extensive information regarding the particle motions as function of time more easily that the actual experiments (Karplus and McCammon, 2002). MD examines the stability of the native state of the proteins and produces a trajectory via numerical integration of motion equation. The generated output trajectory represents dynamical information required for further calculation. Also, conformational changes of model protein, fluctuations, and evaluation of interactions between the proteins are examined (Arnittali et al., 2019). Root mean square deviation (RMSD) is used to measure the similarity between two structures. It is computed by converting and rotating the coordinates of instant structures to superimpose with reference conformation with a maximum alignment. In other words, RMSD is a indicates the deviation from the alignment of the two compared conformations. Since RMSD value represents the difference between the two superimposed structures coordinates, the smaller value suggests the higher similarity between two structures. Mathematical algortihms were developed to determine the most efficient orientation of two structures that provides the least possible RMSD value (Reva et al., 1998). Desmond MD stimulations possess explicit water treatment and receptor flexibility, catching the dynamic character of receptor-ligand interactions. Thus, it is an effective tool for studying the atomic level interactions that are decisive for ligand affinity and selectivity to its receptor. Desmond MD stimulations possess explicit water treatment and receptor flexibility, catching the dynamic character of receptor-ligand interactions. Thus, it is an effective tool for studying the atomic level interactions that are decisive for ligand affinity and selectivity to its receptor.  Such analysis is applicable for prediction of binding interaction and examine structure-activity relationship (SAR) records ("Introducing SID (Simulation Interactions Diagram) | zhiuödinger" 2022).

## 1.3 Taxonomy

History already recognizes two large-scale outbreaks caused by coronaviruses, specifically, SARS and Middle East respiratory syndrome (MERS). Since the SARS outbreak in 2002–2004, a significant number of SARS-associated coronaviruses have been detected in bats that are their indigenous reservoir hosts (Zhou et al., 2020). Besides, coronaviruses have been found in various hosts, such as humans, bats, civets, dogs, mice, cats, camels, and cows (Wang et al., 2006). In early 2007 it has been already known that bats were associated as a natural reservoir for a rising number of transmitting zoonotic viruses including many other viruses that possess high genetic similarity with coronaviruses that cause SARS. The high mutation rate of coronaviruses contributes to the feasibility of human and domestic mammals acquisition of the disease that is also enhanced by legal and illegal wild animal trading at the locations that facilitate cross-species viral transmission which in turn leads to the prompt spread of viral infections globally (Wong et al., 2019). SARS and MERS were caused by genetically distinct coronaviruses that also emerged from bats in poor sanitary markets (Cui et al., 2019). The current COVID-19 outbreak is atypical pneumonia caused by novel SARS-CoV-2 is suggested to originate from zoonotic and cross-species viral transmission between bars and pangolins at Wuhan market, where they were stored near the consumer meat (Chan et al., 2020). Despite that SARS-CoV-2 develops to be less deadly than SAR-CoV and MERS-CoV, it possesses a significantly higher transmissibility rate (Harrison et al., 2020).

## 1.4 Epidemiology and surveillance

Viral genome surveillance is critical to monitor disease transmission during major outbreaks (Gardy et al., 2015). Real-time gene sequencing is a key parameter to manage viral outbreaks because it expands our understanding of the virus proliferation, spread, and evolution (Dudas et al., 2017) (Gardy et al., 2015). Primarily, it relies on the prompt sequencing technique of viral genome immediately from clinical specimens without the necessity of viral culturing step. The Ebola virus epidemic of 2013-2016 demonstrated that viral genome surveillance can yield crucial evidence on Ebola virus progression and facilitate epidemiological examination. Essentially, genome sequencing directly from the clinical samples is more rapid, requires less laborious work, and is less time-consuming than culturing enrichment methods. Metagenomics examines genetic material (DNA or cDNA) directly from the samples for virus identification and diagnostics. The whole-genome sequencing of the Ebola virus immediately from clinical samples skipping the culturing amplification step was feasible due to significantly high levels of viral copy numbers in severe cases. On the other hand, metagenomic analysis directly from the clinical samples may possess some limitations in terms of sensitivity, specificity, limited or the absence of genome coverage while sequencing a low copy number of viral genomic material in comparison with a high abundance of host genetic information (Quick et al., 2017).

SARS-CoV-2 genome sequencing provided important data on the viral mutation rate, transmission dynamics, and its taxonomic origin. Genomic surveillance of SARS-CoV-2 is critical for tracking viral spread in each country, detecting the geographical origin of viral strains, or indication of control measures efficiency, and viral evolution. Besides, the genomic analysis yields vital insights into epidemiological investigations of pandemic evolution. Altogether cumulative investigations facilitated the establishment of nomenclature systems for various SARS-CoV-2 lineages (Rambaut et al., 2020).

The documentation confirming the reinfections suggests that distinct SARS-CoV-2 strains are able to infect the same person (Tillett et al., 2021)(To et al., 2020). Genomic sequencing is required to verify these reinfections and eliminate medical recidivism. Rapid and reliable sequencing techniques in clinical application are crucial for epidemiological supervision (González-recio et al., 2021). Consequently, reliable early detection is crucial in COVID-19 surveillance. Even though the antibody-based detection approach is fast, this method has several limitations, specifically, bacterial contamination, hemolysis, fibrin presence, patient autoantibodies, and promoting false-positive results. Accordingly, the sequence detection method remains to be the most appropriate for COVID-19 diagnostics, and viral mutation rate control. In particular, real-time quantitative reverse transcription-polymerase reaction (RT-qPCR) is the most prominent testing technique for SARS-CoV-2 identification. RT-qPCR is highly specific, fast, and financially affordable, yet it cannot accurately examine amplified gene fragments. Therefore, COVID-19 positive infection is verified by the detection of one or more conservative sites by RT-qPCR. Also, this method possesses a high level of false-negative rates in clinical settings that can cause disease spread via postponed patient isolation and curing, facilitating further viral transmission (Wang et al., 2020).  In association with different sequencing techniques, currently, third-generation sequencing of the SARS-CoV-2 whole genome by Oxford nanopore technology is one of the main approaches. The main advantages of this platform are long genome reads, an optimized analysis pipeline, rapid sequencing, and data collection (Li et al., 2020).

## 1.5 Viral genome sequencing

Currently, clinical diagnostics of infectious disease are based on several laboratory techniques, such as nuclei acid amplification tests, antigen detection, culturing, serologic assays, and direct visualization. Some of these methods can be laborious, for instance, in culture-based analysis, the necessary growth amplification step usually requires from a day to couple of weeks, with respect to the type of the pathogen. The probability of pathogen isolation can be compromised due to fastidious organism presence or prescription of antimicrobial treatment to the patient. Primarily, it is crucial to develop accurate and rapid pathogen and antimicrobial resistance (AMR) characterization methods to facilitate better clinical outcomes. The advances in next generation sequencing (NGS) provides the opportunities of ubiquitous pathogen detection directly from clinical specimens (Goldberg et al., 2015; Forbes et al., 2017).  NGS is comprised of various methods of nucleic acid sequencing, in other words all these approaches aim to sequence DNA or RNA molecules on a vast scale parallel process. Usually, the primary sample preparation for sequencing step is often a labor-consuming process when genetic material is isolated, examined for quality measurement, and genetic library is prepared according to the protocol that requires from hours to days to be accomplished. The sequencing step itself usually proceeds from 1 to 6 days, according to the length of the reads, platform, and the quantity of the obtained data (Petersen et al., 2019).

Viral genome surveillance is critical to monitor disease transmission during major outbreaks (Gardy et al., 2015). Real-time gene sequencing is a key parameter to manage viral outbreaks because it expands our understanding of virus proliferation, spread, and evolution (Dudas et al., 2017) (Gardy et al., 2015). Primarily, it relies on the prompt sequencing technique of viral genome immediately from clinical specimens without the necessity of viral culturing step. The Ebola virus epidemic of 2013-2016 demonstrated that viral genome surveillance can yield crucial evidence on Ebola virus progression and facilitate epidemiological examination. Essentially, genome sequencing directly from the clinical samples is more rapid, requires less laborious work, and is less time-consuming than culturing enrichment methods. Metagenomics examines genetic material (DNA or cDNA) directly from the samples for virus identification and diagnostics. The whole-genome sequencing of the Ebola virus immediately from clinical samples skipping the culturing amplification step was feasible due to significantly high levels of viral copy numbers in severe cases. On the other hand, metagenomic analysis directly from the clinical samples may possess some limitations in terms of sensitivity, specificity, limited or the absence of genome coverage while sequencing a low copy number of viral genomic material in comparison with high abundance of host genetic information (Quick et al., 2017).

SARS-CoV-2 genome sequencing provided important data on the viral mutation rate, transmission dynamics, and its taxonomic origin. Genomic surveillance of SARS-CoV-2 is critical for tracking viral spread in each country, detecting the geographical origin of viral strains, or indication of control measures efficiency, and viral evolution. Besides, the genomic analysis yields vital insights into epidemiological investigations of pandemic evolution. Altogether cumulative investigations facilitated the establishment of nomenclature systems for various SARS-CoV-2 lineages (Rambaut et al., 2020).

The documentation confirming the reinfections suggests that distinct SARS-CoV-2 strains are able to infect the same person (Tillett et al., 2021)(To et al., 2020). Genomic sequencing is required to verify these reinfections and eliminate medical recidivism. Rapid and reliable sequencing techniques in clinical application are crucial for epidemiological supervision (González-recio et al., 2021). Consequently, early reliable detection is crucial in COVID-19 surveillance. Despite the fact that the antibody-based detection approach is fast, this method has several limitations, specifically, bacterial contamination, hemolysis, fibrin presence, patient autoantibodies, and promoting false-positive results. Accordingly, the sequence detection method remains to be the most appropriate for COVID-19 diagnostics, and viral mutation rate control. In particular, real-time quantitative reverse transcription-polymerase reaction (RT-qPCR) is the most prominent testing technique for SARS-CoV-2 identification. RT-qPCR is highly specific, fast, and financially affordable, yet it cannot accurately examine amplified gene fragments. Therefore, COVID-19 positive infection is verified by the detection of one or more conservative sites by RT-qPCR. Also, this method possesses a high level of false-negative rates in clinical settings that can cause disease to spread via postponed patient isolation and curing, facilitating further viral transmission (Wang et al., 2020). In association with different sequencing techniques, currently, third-generation sequencing of the SARS-CoV-2 whole genome by Oxford nanopore technology is one of the prominent approaches. The main advantages of this platform are long genome reads, an optimized analysis pipeline, rapid sequencing, and data collection (Li et al., 2020).

### 1.5.1 Sequencing by Oxford nanopore technology (ONT)

Oxford nanopore technology (ONT) is based on nanopore channels for sequencing, the ionic current passes across the flow cell, so as the nanoscale genetic polymer passes through the nanopore, each nucleotide base is distinguished according to the changes in current (Laver et al., 2015). Before the sequencing, both ends of genomic (DNA or cDNA) polymers are ligated with special adaptors that promote capture and loading capacity of processive enzyme at 5’-end of the strand. Also, adaptors assist strand concentration at the membrane surface adjacent to the nanopore, enhancing strand capture by a thousand-fold. The enzyme is necessary to facilitate unidirectional single-nucleotide movement throughout the length of the strand at a millisecond time scale. At the time of strand capture in the nanopore, the enzyme proceeds along the template strand that passes via the pore, the sensor identifies the ionic current changes induced by alteration of the nucleotide residing in the pore. The ionic current alterations are recognized as distinct events that possess corresponding variance, duration, and mean amplitude. (Jain et al., 2016). ONT aimed to perform long-read sequencing encourages significant improvements in terms of portability, cost, and turnaround time in comparison with accepted short-read sequencing platforms widely used for viral whole genome sequencing (e.g., Illumina). Nevertheless, acquisition of ONT for SARS-CoV-2 monitor has been restricted due to general concerns regarding the sequencing accuracy. To address this issue a recent study performed evaluation of

SARS-CoV-2 whole genome sequencing by ONT and Illumina and compared analytical performance. Regardless of increased error rates detected in ONT reads, consensus-level sequence reads were highly accurate, specifically single nucleotide variants (SNVs) identified at >99% sensitivity and >99% precision over a minimum 60-fold coverage depth, by assuring suitability for viral genome analysis. Besides, ONT sequencing could detect an unexpected diversity of structural variation of viral samples that was supported by verification from short-read sequencing on corresponding samples (Bull et al., 2020). The ONT protocol for viral sequencing via tiled PCR-generated amplicon pools approach was established by the Artic Network (https://artic.network/ ) for whole genome sequencing of Chikungunya, Ebola, and Zika viruses. Then the initial protocol was rapidly modified for prompt sequence analysis of SARS-CoV-2 RNA obtained directly from the clinical specimens, namely oropharyngeal and nasopharyngeal swabs. Additional improvements simplified the library preparation step, reduced hands-on time, and extended sample multiplexing up to 96 which reduced the reagent price to nearly £10 per specimen, increasing the availability of this technique for epidemiologic monitoring and further studies (Brejová et al., 2021).

# 2 MATERIALS & METHODS

The general outline of the study is visualized in the figure below:

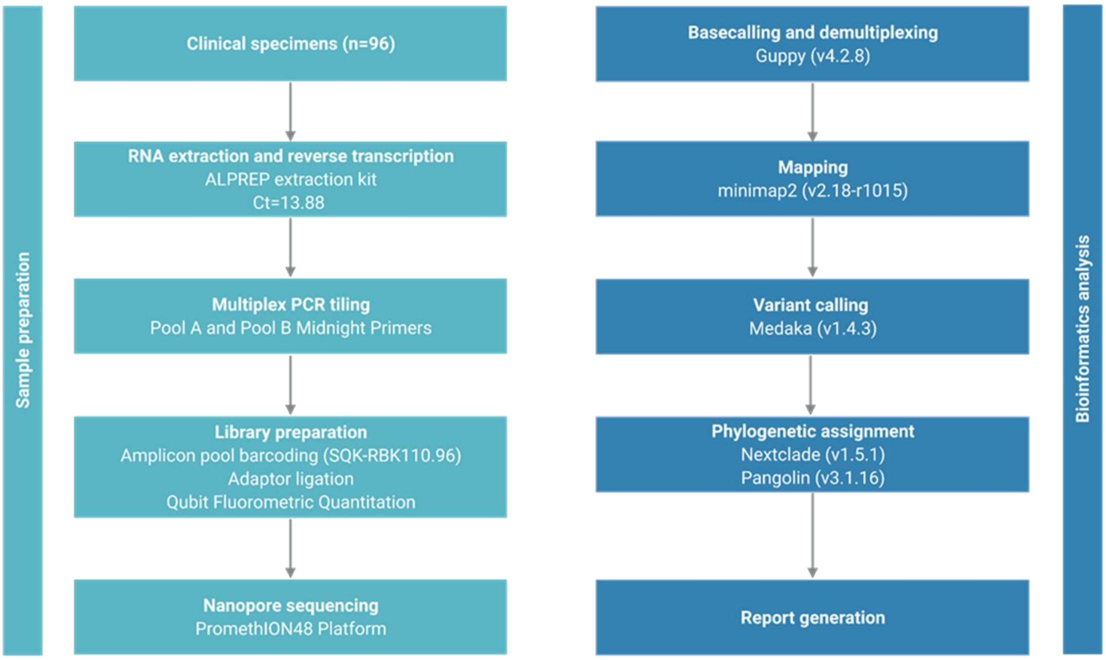


***Figure 1.*** SARS-CoV-2 whole-genome sequencing and analysis workflow.

## 2.1 Sample collection

The 96 samples used in this study were nasopharyngeal swabs kindly provided by the Republican Diagnostic Center (RDC). The nasopharyngeal swab fluid samples (5-10 mL) were obtained from COVID-19 positive patients whose status was laboratory-confirmed by qRT-PCR results in 2-10th of August 2021 at RDC. Viral RNA was isolated from clinical biomaterials using ALPREP extraction kit following manufacturer (Algimed Techno, Belarus) instructions at the RDC laboratory. All samples had a cycle threshold (Ct) value below 20, while the average Ct value of all RNA samples was 13.88 corresponding to a high viral genetic material load.

## 2.2 ONT Library Preparation and Sequencing

ONT library was prepared following the SQK-RBK110.96 protocol (https://community.nanoporetech.com ) and sequenced on the PromethION48 platform using flow cell. The 8 μl RNA samples were reverse transcribed with 2 μl LunaScript RT SuperMix (LS RT) at thermal cycler using the following program: at 25°C for 2 min, at 55°C for 10 min, at 95 °C for 1 min, and at 4°C hold. Midnight RT PCR Expansion (EXP-MRT001) contained separate primer pools (Freed et al., 2020) used for the overlapping tiled PCR reactions spanning the viral genome. The PCR reaction mix for x96 samples contains 241 μl of nuclease-free water, 6 μl of Pool A or Pool B Midnight Primers, and 687 μl of Q5 HS Master Mix (Q5). Two-midnight primer pools were used to anneal to 4.5% of the genome and produce 1200 bp amplicons that overlap by approximately 20 bp. PCR amplification step was carried out under the following conditions: initial denaturation step at 98°C for 30 sec, followed by 35 amplifications at 98°C for 15 s, at 65 °C for 5 min, and at 4°C hold. The addition of rapid barcodes was performed in the 96-well Barcode Attachment Plate by mixing 2.5 μl nuclease-free water, 5 μl pooled PCR products (from pools A and B), and 2.5 μl barcodes from the Rapid Barcode Plate. The reaction was incubated in a thermal cycler at 30°C for 2 minutes and then at 80°C for 2 minutes. Two step lean-up performed using the SPRI beads and 80% ethanol. To measure the concentration of DNA (PCR products and DNA libraries), Qubit dsDNA HS Assay Kit was used for fluorometric measurement of DNA (Thermo Fisher Scientific) on a Qubit 4.0 Fluorometer. The >1400 ng of DNA library was loaded onto a primed PromethION48 flow cell (PAH13359).

### 2.3 Software setup and installation

The ARTIC sequencing data obtained by Midnight protocol was analyzed by the wf-artic bioinformatics pipeline (https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html ). The Nextflow software orchestrates this workflow. The Nextflow and Docker software are required for wf-artic pipeline installation. Nextflow project is available in open access (https://www.nextflow.io/ ) and allows implementation of scientific pipeline in a reproducible manner. The installation of the software on the Linux operating system is straightforward and is supported onGirdION and PromethION devices.

After the demultiplexing step, the sequence reads were processed by ARTIC FieldBioinformatics software that was adapted to analyze FASTQ Nanopore sequences. Besides, the ARTIC pipeline was modified to utilize a primer scheme that specifies the sequencing primers used in Midnight protocol and their genomic localization on the SARS-CoV-2 genome. The wf-artic pipeline classifies the sequenced samples according to Nexclade clastidic analysis and Pangolin strain assignment.

### 2.4 Demultiplexing of multiple barcoded samples

Multiplexed FASTQ format sequence data is required for wf-artic workflow. The sequences can be demultiplexed immediately in MinKNOW software or Guppy software 4.2.8 in a post-sequencing step. Guppy barcoding performs basecalling of all reads and identifies barcodes in the sequence. To prevent re-basecalling, the software copies the reads pertaining to each barcode to the corresponding tag output directory. Since Midnight protocol utilizes a rapid barcoding kit, the demultiplexing step does not need barcodes at both ends of the sequence. In addition, filtering against mid-strand barcodes is not required.

### 2.5 Variant calling and phylogenetic profiling

A medaka is a bioinformatics tool that generates consensus sequences from basecalled data by using a collection of individual sequencing reads against a draft assembly. The variant calling was performed by the set of utilities bcftools 1.12. The pipeline output includes NextClade (https://clades.nextstrain.org/ ) and Pangolin analysis that includes the clade designation according to GISIAD (https://www.gisaid.org/ ) and Pangolin nomenclature. Then consensus sequences were submitted to the GISAID database.

### 2.6 SDM analysis

To run Site Directed Mutator the wild-type structure in PDB format should be submitted into the http://marid.bioc.cam.ac.uk/sdm2/ webserver. A file with a mutation list to be analyzed was provided with indication of mutation position. The query for analysis was submitted and run. The obtained results were downloaded.

### 2.7 Molecular dynamics simulation analysis

To run Desmond simulation (https://www.schrodinger.com/ ), firstly, the structure file in PDB format was imported. The mutations of the sequenced samples were introduced into the wildtype PDB file in PyMOL (https://pymol.org/2/ ) and uploaded into the software. The Protein Preparation Wizard panel was used to prepare the initiate the structure simulation, where ions and molecules (artifacts) were removed, proper bond orders were set, missing side chains and some residues if necessary were filled, various residue protonation states or groups were reoriented to set hydrogen bonding network etc. Next System Builder panel was set to generate the system and ions were distributed to set desired ionic strength in system for simulation. After that the system was relaxed by minimization step in the panel before simulation. The simulation parameters were set for molecular dynamics and the simulation was run for 200 ns period. Then the results were analyzed by analysis tools.

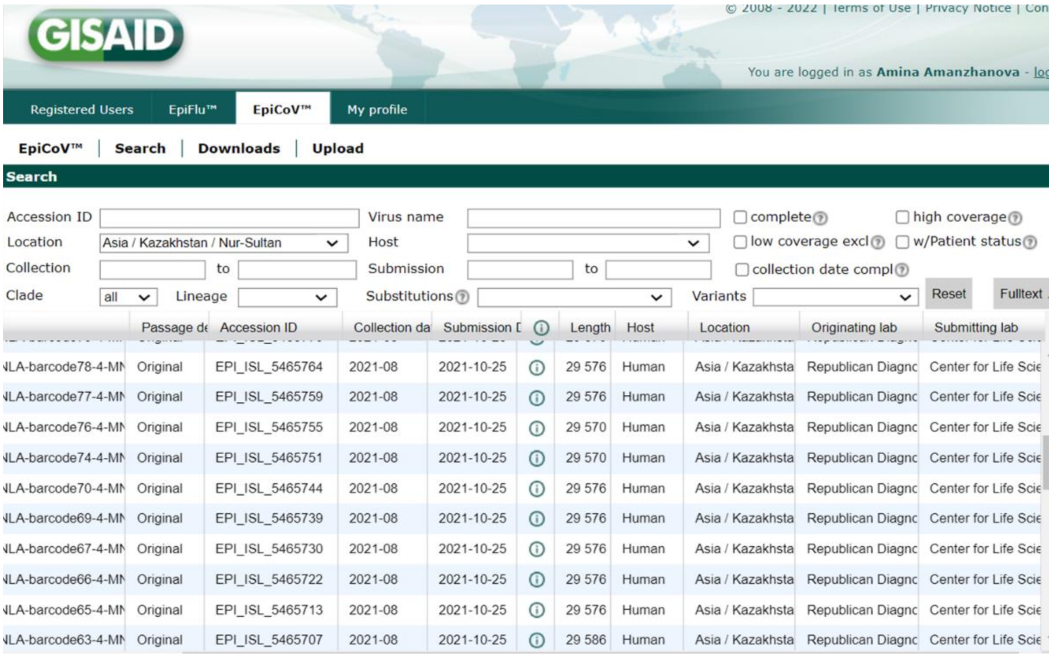
# AIMS OF THE THESIS PROJECT

SARS-CoV-2 sequencing was initiated at the National Laboratory of Astana for several reasons, (1) to verify the feasibility of Oxford nanopore amplicon-based SARS-CoV-2 genome sequencing at our institution; and (2) contribute to SARS-CoV-2 genome surveillance in Nur-Sultan, and (3) establish an optimized protocol for future SARS-CoV-2 monitoring in Kazakhstan. Also, (4) the effect of the mutations in the spike region on protein stability and (5) molecular dynamics of ACE2 and RBD of spike interaction were analyzed by bioinformatic tools. In this study, the 96 SARS-CoV-2 samples obtained from the RT-PCR confirmed COVID-19 positive patients in August 2021 were sequenced by next-generation sequencing technology Oxford nanopore to characterize viral dynamics in the country in connection with the global pandemic.

# RESULTS

## 4.1 The 77 genetic sequences upload on GISAID

**A)**



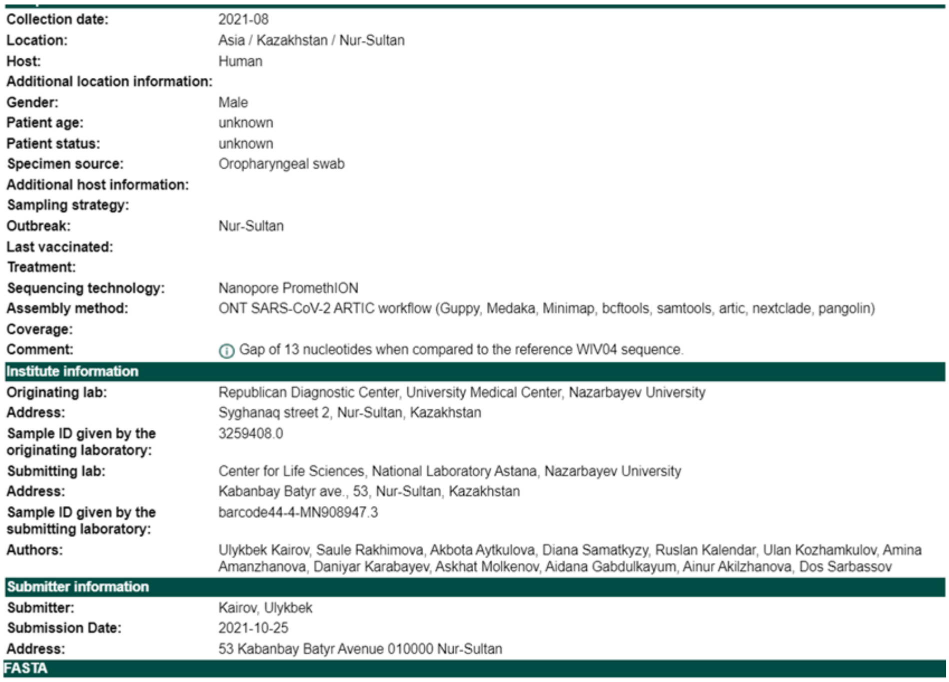
GISAID
© 2008 - 2022 | Terms of Use | Privacy Notice | Con
Registered Users
EpiFlu™
EpiCoV™
My profile
EpiCoV™ | Search | Downloads | Upload
Search
Reset
Fulltext

You are logged in as **Amina Amanzhanova**

Accession ID
Virus name
☐ complete
☐ high coverage
Location: Asia / Kazakhstan / Nur-Sultan
Host
☐ low coverage excl
☐ w/Patient status
Collection to
Submission to
☐ collection date compl
Clade: all
Lineage
Substitutions
Variants

| | Passage de | Accession ID | Collection da | Submission D | ⓘ | Length | Host | Location | Originating lab | Submitting lab |
|---|---|---|---|---|---|---|---|---|---|---|
| NLA-barcode78-4-MN | Original | EPI_ISL_5465764 | 2021-08 | 2021-10-25 | ⓘ | 29 576 | Human | Asia / Kazakhsta | Republican Diagnc | Center for Life Scie |
| NLA-barcode77-4-MN | Original | EPI_ISL_5465759 | 2021-08 | 2021-10-25 | ⓘ | 29 576 | Human | Asia / Kazakhsta | Republican Diagnc | Center for Life Scie |
| NLA-barcode76-4-MN | Original | EPI_ISL_5465755 | 2021-08 | 2021-10-25 | ⓘ | 29 570 | Human | Asia / Kazakhsta | Republican Diagnc | Center for Life Scie |
| NLA-barcode74-4-MN | Original | EPI_ISL_5465751 | 2021-08 | 2021-10-25 | ⓘ | 29 570 | Human | Asia / Kazakhsta | Republican Diagnc | Center for Life Scie |
| NLA-barcode70-4-MN | Original | EPI_ISL_5465744 | 2021-08 | 2021-10-25 | ⓘ | 29 576 | Human | Asia / Kazakhsta | Republican Diagnc | Center for Life Scie |
| NLA-barcode69-4-MN | Original | EPI_ISL_5465739 | 2021-08 | 2021-10-25 | ⓘ | 29 576 | Human | Asia / Kazakhsta | Republican Diagnc | Center for Life Scie |
| NLA-barcode67-4-MN | Original | EPI_ISL_5465730 | 2021-08 | 2021-10-25 | ⓘ | 29 576 | Human | Asia / Kazakhsta | Republican Diagnc | Center for Life Scie |
| NLA-barcode66-4-MN | Original | EPI_ISL_5465722 | 2021-08 | 2021-10-25 | ⓘ | 29 576 | Human | Asia / Kazakhsta | Republican Diagnc | Center for Life Scie |
| NLA-barcode65-4-MN | Original | EPI_ISL_5465713 | 2021-08 | 2021-10-25 | ⓘ | 29 576 | Human | Asia / Kazakhsta | Republican Diagnc | Center for Life Scie |
| NLA-barcode63-4-MN | Original | EPI_ISL_5465707 | 2021-08 | 2021-10-25 | ⓘ | 29 586 | Human | Asia / Kazakhsta | Republican Diagnc | Center for Life Scie |

**B)**



| | |
|---|---|
| **Collection date:** | 2021-08 |
| **Location:** | Asia / Kazakhstan / Nur-Sultan |
| **Host:** | Human |
| **Additional location information:** | |
| **Gender:** | Male |
| **Patient age:** | unknown |
| **Patient status:** | unknown |
| **Specimen source:** | Oropharyngeal swab |
| **Additional host information:** | |
| **Sampling strategy:** | |
| **Outbreak:** | Nur-Sultan |
| **Last vaccinated:** | |
| **Treatment:** | |
| **Sequencing technology:** | Nanopore PromethION |
| **Assembly method:** | ONT SARS-CoV-2 ARTIC workflow (Guppy, Medaka, Minimap, bcftools, samtools, artic, nextclade, pangolin) |
| **Coverage:** | |
| **Comment:** | ⓘ Gap of 13 nucleotides when compared to the reference WIV04 sequence. |
| **Institute information** | |
| **Originating lab:** | Republican Diagnostic Center, University Medical Center, Nazarbayev University |
| **Address:** | Syghanaq street 2, Nur-Sultan, Kazakhstan |
| **Sample ID given by the originating laboratory:** | 3259408.0 |
| **Submitting lab:** | Center for Life Sciences, National Laboratory Astana, Nazarbayev University |
| **Address:** | Kabanbay Batyr ave., 53, Nur-Sultan, Kazakhstan |
| **Sample ID given by the submitting laboratory:** | barcode44-4-MN908947.3 |
| **Authors:** | Ulykbek Kairov, Saule Rakhimova, Akbota Aytkulova, Diana Samatkyzy, Ruslan Kalendar, Ulan Kozhamkulov, Amina Amanzhanova, Daniyar Karabayev, Askhat Molkenov, Aidana Gabdulkayum, Ainur Akilzhanova, Dos Sarbassov |
| **Submitter information** | |
| **Submitter:** | Kairov, Ulykbek |
| **Submission Date:** | 2021-10-25 |
| **Address:** | 53 Kabanbay Batyr Avenue 010000 Nur-Sultan |
| **FASTA** | |

***Figure 2.*** A) The 77 SARS-CoV-2 sequences uploaded on GISAID platform. B) The example of EPI_ISL_5465615 sample information.

## 4.2 Phylogenetic tree

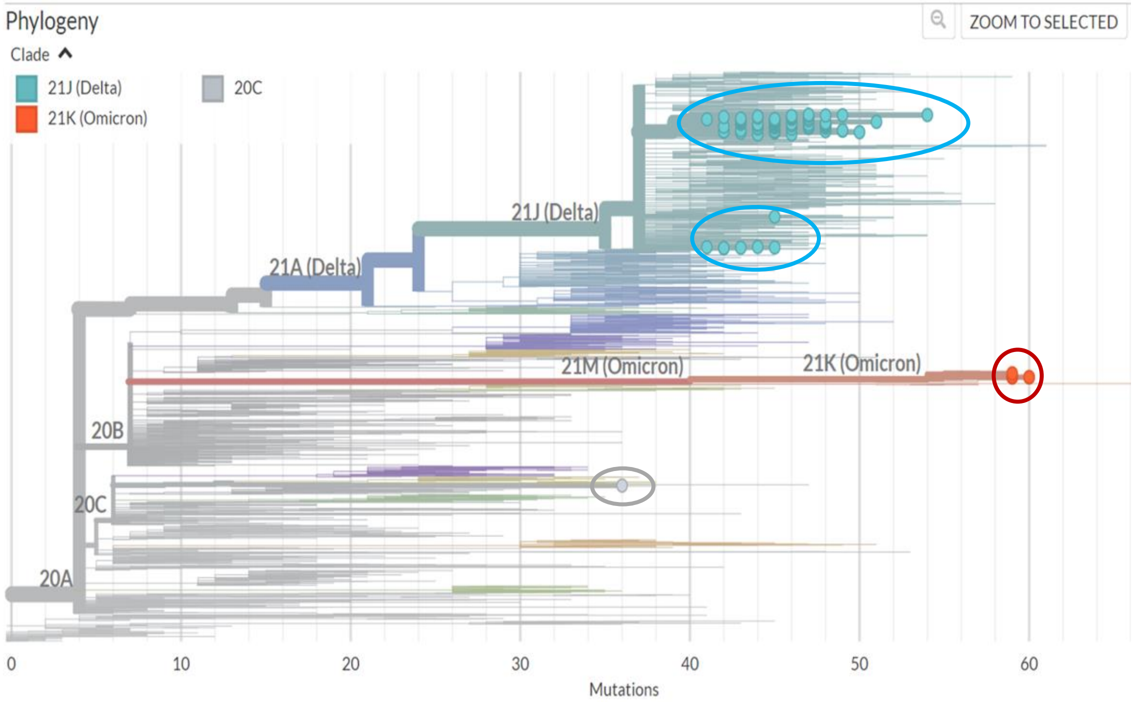


***Figure 3.*** Phylogenetic distribution of ONT sequenced SARS-CoV-2 genomes. A phylogenetic tree was generated by Nexstrain and the sequenced samples are identified according to Nextstrain nomenclature. The samples belonging to 21J (Delta) variants are highlighted with blue circles, whereas earlier origin variant 20C is highlighted with grey circle. Besides, the 21 K omicron variants detected in recent run in March 2022 are indicated with red circles.

## 4.3 Lineage distribution

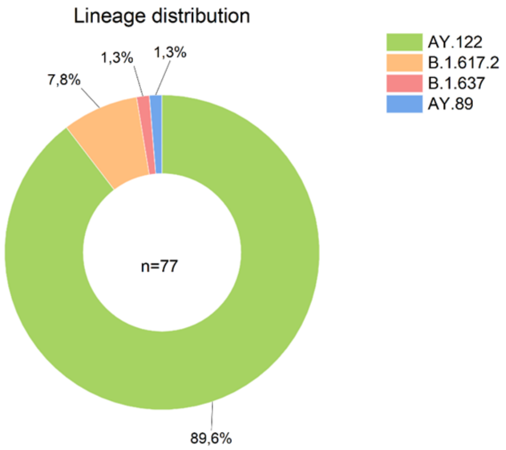


***Figure 4.***The proportion of clades and Pangolin lineages of n=77 sequenced SARS-CoV-2 samples. The AY.122 (n=69), B.1.617.2 (n=6), and AY.89 (n=1) delta lineages, and one sample B.1.637 belongs to a separate lineage corresponding to Iota were detected in this study.

## 4.4 Frequency of mutations in SARS-CoV-2

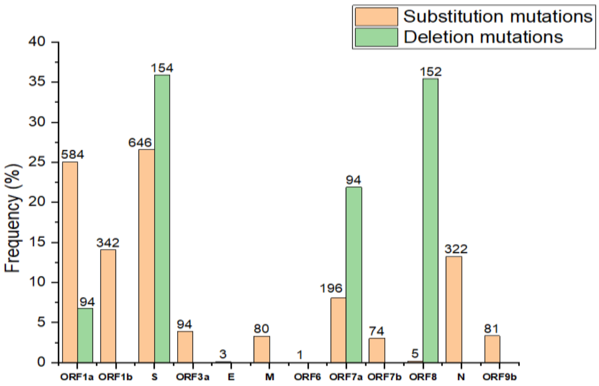


***Figure 5.*** Substitution and deletion mutation frequency in sequenced SARS-CoV-2 genomes. The graph illustrates that substitution mutations primarily occur at ORF1 and S protein region, while deletion mutations predominantly are found in S and ORF8 regions. The spike protein in viral genome is a primary region for mutations.

## 4.5 Spike glycoprotein of EPI_ISL_5465615

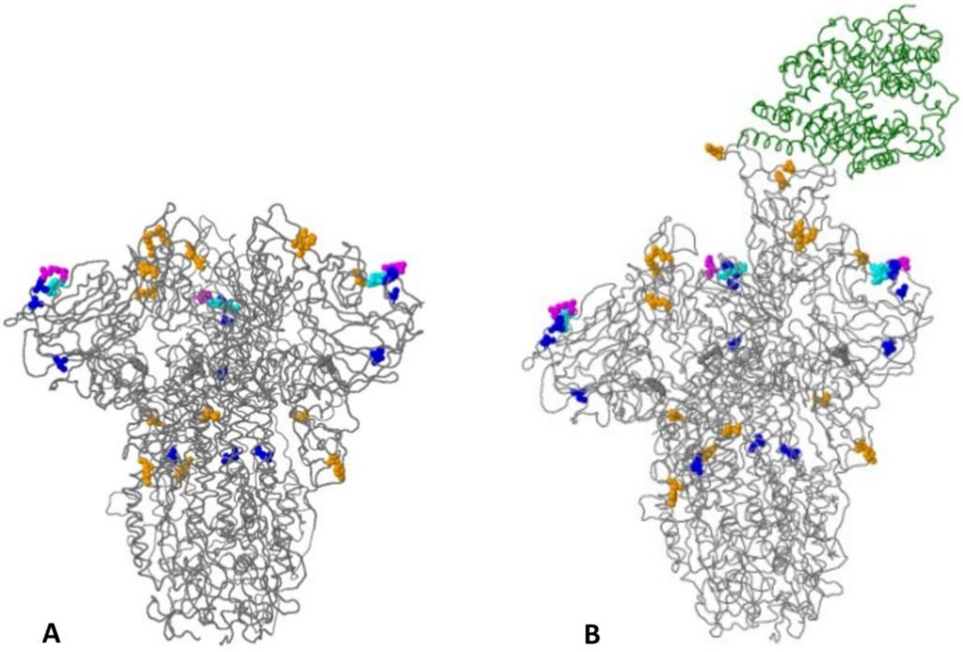
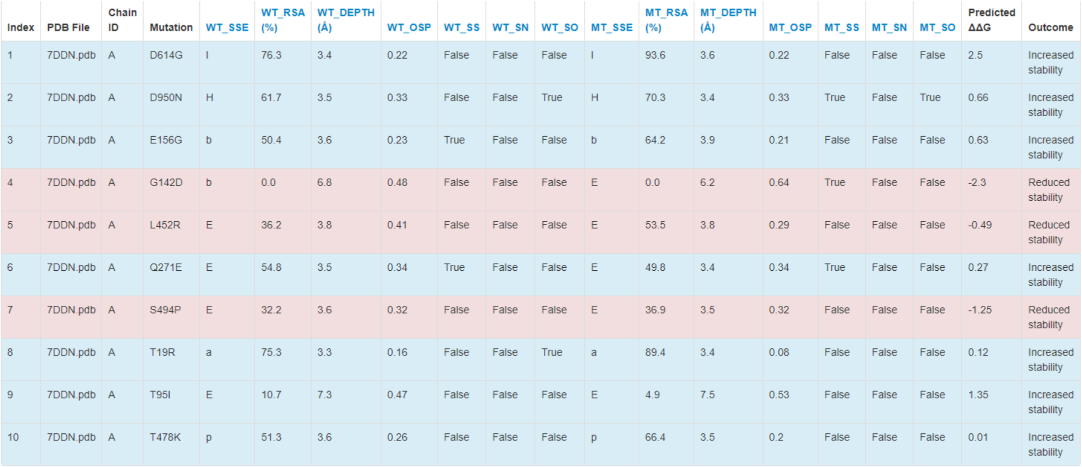


***Figure 6.*** A) Spike glycoprotein (PDB: 6acc, EM 3.6 Angstrom) with RBD in upward conformation with amino acid changes identified in the query sequences shown as colored balls of sample EPI_ISL_5465615. B) Spike glycoprotein of EPI_ISL_5465615 (PDB: 6acj, EM 4.2 Angstrom) in complex with host cell receptor ACE2 (green ribbon). The amino acid alterations in spike glycoprotein of EPI_ISL_5465615 query sequences are distinguished by distinct colors, as black font represents amino acid changes with unknown effects, blue font stands for frequent and epidemiologically significant amino acid alterations, orange font illustrates the mutations associated with host-cell receptor binding or antigenicity, and magenta font indicates the addition or deletion of glycosylation sites.

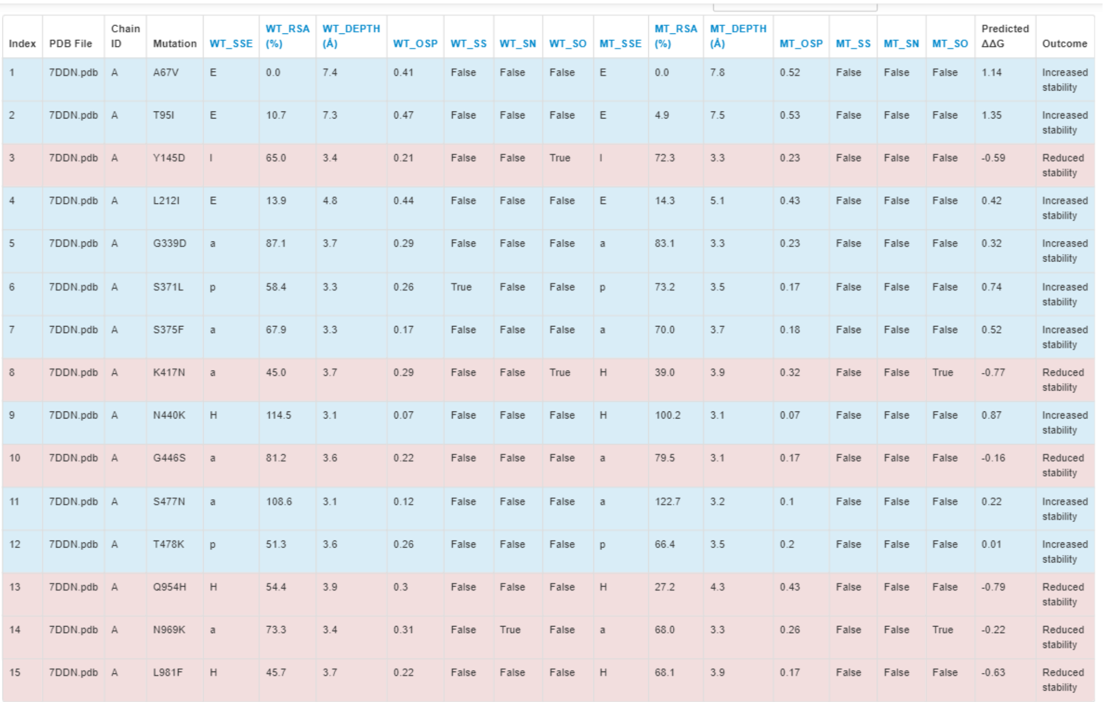
## 4.6 Protein stability analysis by SDM upon mutation

### A)



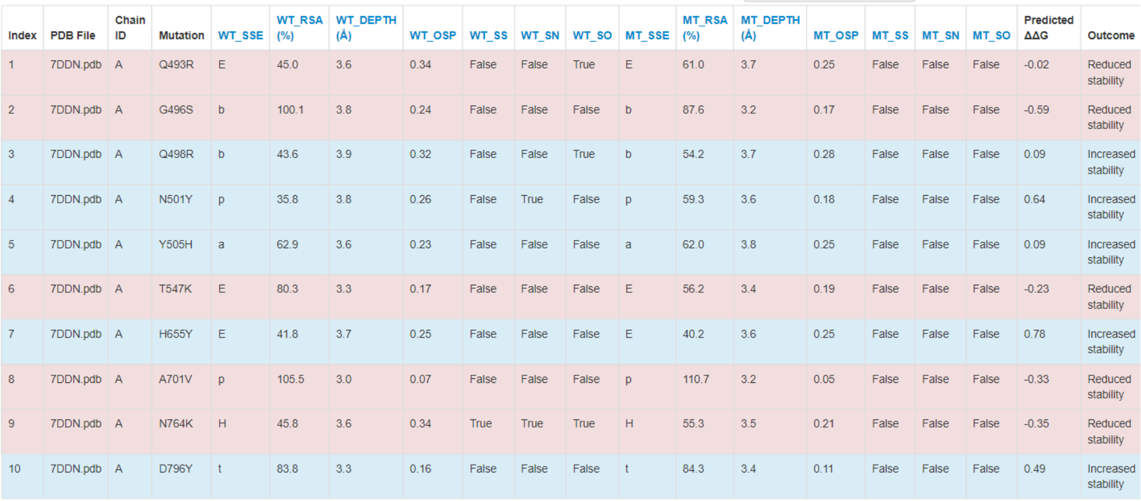
| Index | PDB File | Chain ID | Mutation | WT_SSE | WT_RSA (%) | WT_DEPTH (Å) | WT_OSP | WT_SS | WT_SN | WT_SO | MT_SSE | MT_RSA (%) | MT_DEPTH (Å) | MT_OSP | MT_SS | MT_SN | MT_SO | Predicted ΔΔG | Outcome |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7DDN.pdb | A | D614G | I | 76.3 | 3.4 | 0.22 | False | False | False | I | 93.6 | 3.6 | 0.22 | False | False | False | 2.5 | Increased stability |
| 2 | 7DDN.pdb | A | D950N | H | 61.7 | 3.5 | 0.33 | False | False | True | H | 70.3 | 3.4 | 0.33 | True | False | True | 0.66 | Increased stability |
| 3 | 7DDN.pdb | A | E156G | b | 50.4 | 3.6 | 0.23 | True | False | False | b | 64.2 | 3.9 | 0.21 | False | False | False | 0.63 | Increased stability |
| 4 | 7DDN.pdb | A | G142D | b | 0.0 | 6.8 | 0.48 | False | False | False | E | 0.0 | 6.2 | 0.64 | True | False | False | -2.3 | Reduced stability |
| 5 | 7DDN.pdb | A | L452R | E | 36.2 | 3.8 | 0.41 | False | False | False | E | 53.5 | 3.8 | 0.29 | False | False | False | -0.49 | Reduced stability |
| 6 | 7DDN.pdb | A | Q271E | E | 54.8 | 3.5 | 0.34 | True | False | False | E | 49.8 | 3.4 | 0.34 | True | False | False | 0.27 | Increased stability |
| 7 | 7DDN.pdb | A | S494P | E | 32.2 | 3.6 | 0.32 | False | False | False | E | 36.9 | 3.5 | 0.32 | False | False | False | -1.25 | Reduced stability |
| 8 | 7DDN.pdb | A | T19R | a | 75.3 | 3.3 | 0.16 | False | False | True | a | 89.4 | 3.4 | 0.08 | False | False | False | 0.12 | Increased stability |
| 9 | 7DDN.pdb | A | T95I | E | 10.7 | 7.3 | 0.47 | False | False | False | E | 4.9 | 7.5 | 0.53 | False | False | False | 1.35 | Increased stability |
| 10 | 7DDN.pdb | A | T478K | p | 51.3 | 3.6 | 0.26 | False | False | False | p | 66.4 | 3.5 | 0.2 | False | False | False | 0.01 | Increased stability |

### B)



| Index | PDB File | Chain ID | Mutation | WT_SSE | WT_RSA (%) | WT_DEPTH (Å) | WT_OSP | WT_SS | WT_SN | WT_SO | MT_SSE | MT_RSA (%) | MT_DEPTH (Å) | MT_OSP | MT_SS | MT_SN | MT_SO | Predicted ΔΔG | Outcome |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7DDN.pdb | A | A67V | E | 0.0 | 7.4 | 0.41 | False | False | False | E | 0.0 | 7.8 | 0.52 | False | False | False | 1.14 | Increased stability |
| 2 | 7DDN.pdb | A | T95I | E | 10.7 | 7.3 | 0.47 | False | False | False | E | 4.9 | 7.5 | 0.53 | False | False | False | 1.35 | Increased stability |
| 3 | 7DDN.pdb | A | Y145D | I | 65.0 | 3.4 | 0.21 | False | False | True | I | 72.3 | 3.3 | 0.23 | False | False | False | -0.59 | Reduced stability |
| 4 | 7DDN.pdb | A | L212I | E | 13.9 | 4.8 | 0.44 | False | False | False | E | 14.3 | 5.1 | 0.43 | False | False | False | 0.42 | Increased stability |
| 5 | 7DDN.pdb | A | G339D | a | 87.1 | 3.7 | 0.29 | False | False | False | a | 83.1 | 3.3 | 0.23 | False | False | False | 0.32 | Increased stability |
| 6 | 7DDN.pdb | A | S371L | p | 58.4 | 3.3 | 0.26 | True | False | False | p | 73.2 | 3.5 | 0.17 | False | False | False | 0.74 | Increased stability |
| 7 | 7DDN.pdb | A | S375F | a | 67.9 | 3.3 | 0.17 | False | False | False | a | 70.0 | 3.7 | 0.18 | False | False | False | 0.52 | Increased stability |
| 8 | 7DDN.pdb | A | K417N | a | 45.0 | 3.7 | 0.29 | False | False | True | H | 39.0 | 3.9 | 0.32 | False | False | True | -0.77 | Reduced stability |
| 9 | 7DDN.pdb | A | N440K | H | 114.5 | 3.1 | 0.07 | False | False | False | H | 100.2 | 3.1 | 0.07 | False | False | False | 0.87 | Increased stability |
| 10 | 7DDN.pdb | A | G446S | a | 81.2 | 3.6 | 0.22 | False | False | False | a | 79.5 | 3.1 | 0.17 | False | False | False | -0.16 | Reduced stability |
| 11 | 7DDN.pdb | A | S477N | a | 108.6 | 3.1 | 0.12 | False | False | False | a | 122.7 | 3.2 | 0.1 | False | False | False | 0.22 | Increased stability |
| 12 | 7DDN.pdb | A | T478K | p | 51.3 | 3.6 | 0.26 | False | False | False | p | 66.4 | 3.5 | 0.2 | False | False | False | 0.01 | Increased stability |
| 13 | 7DDN.pdb | A | Q954H | H | 54.4 | 3.9 | 0.3 | False | False | False | H | 27.2 | 4.3 | 0.43 | False | False | False | -0.79 | Reduced stability |
| 14 | 7DDN.pdb | A | N969K | a | 73.3 | 3.4 | 0.31 | False | True | False | a | 68.0 | 3.3 | 0.26 | False | False | True | -0.22 | Reduced stability |
| 15 | 7DDN.pdb | A | L981F | H | 45.7 | 3.7 | 0.22 | False | False | False | H | 68.1 | 3.9 | 0.17 | False | False | False | -0.63 | Reduced stability |

**B)**

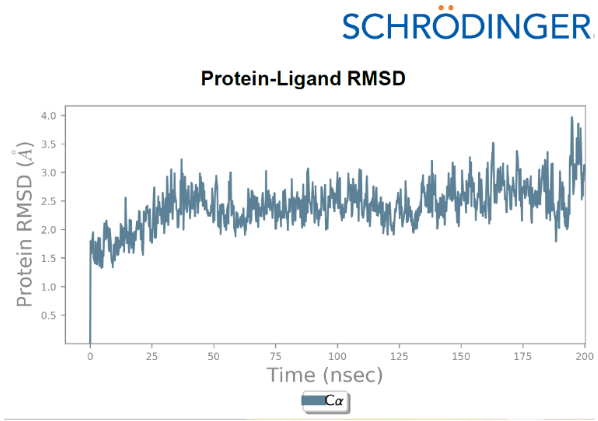


| Index | PDB File | Chain ID | Mutation | WT_SSE | WT_RSA (%) | WT_DEPTH (Å) | WT_OSP | WT_SS | WT_SN | WT_SO | MT_SSE | MT_RSA (%) | MT_DEPTH (Å) | MT_OSP | MT_SS | MT_SN | MT_SO | Predicted ΔΔG | Outcome |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7DDN.pdb | A | Q493R | E | 45.0 | 3.6 | 0.34 | False | False | True | E | 61.0 | 3.7 | 0.25 | False | False | False | -0.02 | Reduced stability |
| 2 | 7DDN.pdb | A | G496S | b | 100.1 | 3.8 | 0.24 | False | False | False | b | 87.6 | 3.2 | 0.17 | False | False | False | -0.59 | Reduced stability |
| 3 | 7DDN.pdb | A | Q498R | b | 43.6 | 3.9 | 0.32 | False | False | True | b | 54.2 | 3.7 | 0.28 | False | False | False | 0.09 | Increased stability |
| 4 | 7DDN.pdb | A | N501Y | p | 35.8 | 3.8 | 0.26 | False | True | False | p | 59.3 | 3.6 | 0.18 | False | False | False | 0.64 | Increased stability |
| 5 | 7DDN.pdb | A | Y505H | a | 62.9 | 3.6 | 0.23 | False | False | False | a | 62.0 | 3.8 | 0.25 | False | False | False | 0.09 | Increased stability |
| 6 | 7DDN.pdb | A | T547K | E | 80.3 | 3.3 | 0.17 | False | False | False | E | 56.2 | 3.4 | 0.19 | False | False | False | -0.23 | Reduced stability |
| 7 | 7DDN.pdb | A | H655Y | E | 41.8 | 3.7 | 0.25 | False | False | False | E | 40.2 | 3.6 | 0.25 | False | False | False | 0.78 | Increased stability |
| 8 | 7DDN.pdb | A | A701V | p | 105.5 | 3.0 | 0.07 | False | False | False | p | 110.7 | 3.2 | 0.05 | False | False | False | -0.33 | Reduced stability |
| 9 | 7DDN.pdb | A | N764K | H | 45.8 | 3.6 | 0.34 | True | True | True | H | 55.3 | 3.5 | 0.21 | False | False | False | -0.35 | Reduced stability |
| 10 | 7DDN.pdb | A | D796Y | t | 83.8 | 3.3 | 0.16 | False | False | False | t | 84.3 | 3.4 | 0.11 | False | False | False | 0.49 | Increased stability |

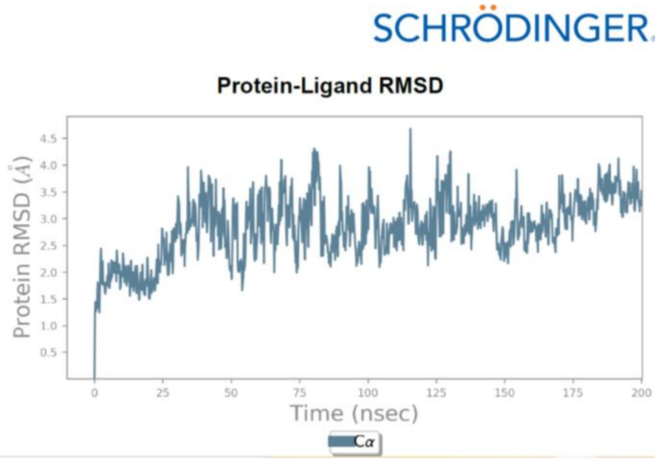
***Figure 7***. Protein stability analysis by SDM. A) The protein stability analysis by SDM upon mutation of sequence EPI_ISL_5465615. A positive (blue) and negative (red) values correlate with mutation expected to be stabilizing and destabilizing accordingly. The results output suggests that the rare mutation Q271E slightly increases stability of the spike protein of ΔΔG=0.27 kcal/mol and generally most mutations contribute to increase in protein stability. Overall predicted net pseudo ΔΔG is equal to 1.5 kcal/mol. B) The protein stability analysis of barcode 18 (omicron) suggests that overall ΔΔG of spike protein stability increased to 3 kcal/mol.

## 4.6 Protein-Ligand Molecular Dynamics analysis

**A)**



**B)**



**C)**

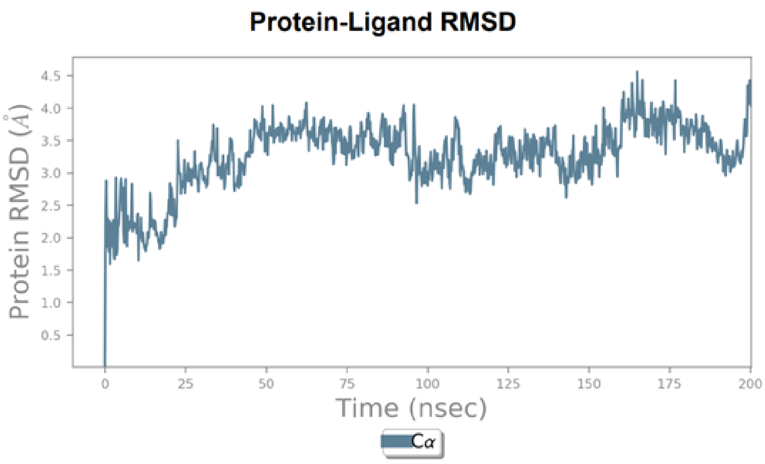
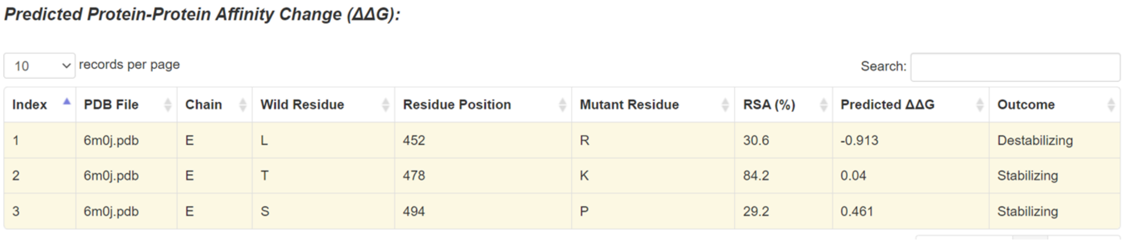


***Figure 8.*** The molecular dynamics simulation output of ACE2-RBD protein interaction. RMSD of ACE2 receptor Cα atoms in accordance with the initial structure. A) ACE2-RBD protein (6MOJ) interaction. B) ACE2-RBD (EPI_ISL_5465615) interaction RMSD plot. C) ACE2-RBD (barcode 18 omicron) interaction RMSD plot.

## 4.7 Protein-Protein Affinity Change Upon Mutation

A)

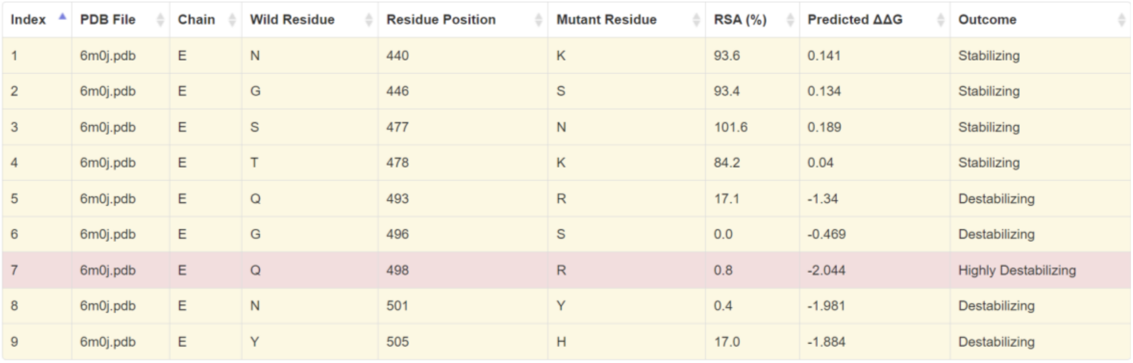


***Predicted Protein-Protein Affinity Change (ΔΔG):***

10 records per page
Search:

| Index | PDB File | Chain | Wild Residue | Residue Position | Mutant Residue | RSA (%) | Predicted ΔΔG | Outcome |
|---|---|---|---|---|---|---|---|---|
| 1 | 6m0j.pdb | E | L | 452 | R | 30.6 | -0.913 | Destabilizing |
| 2 | 6m0j.pdb | E | T | 478 | K | 84.2 | 0.04 | Stabilizing |
| 3 | 6m0j.pdb | E | S | 494 | P | 29.2 | 0.461 | Stabilizing |

B)



| Index | PDB File | Chain | Wild Residue | Residue Position | Mutant Residue | RSA (%) | Predicted ΔΔG | Outcome |
|---|---|---|---|---|---|---|---|---|
| 1 | 6m0j.pdb | E | N | 440 | K | 93.6 | 0.141 | Stabilizing |
| 2 | 6m0j.pdb | E | G | 446 | S | 93.4 | 0.134 | Stabilizing |
| 3 | 6m0j.pdb | E | S | 477 | N | 101.6 | 0.189 | Stabilizing |
| 4 | 6m0j.pdb | E | T | 478 | K | 84.2 | 0.04 | Stabilizing |
| 5 | 6m0j.pdb | E | Q | 493 | R | 17.1 | -1.34 | Destabilizing |
| 6 | 6m0j.pdb | E | G | 496 | S | 0.0 | -0.469 | Destabilizing |
| 7 | 6m0j.pdb | E | Q | 498 | R | 0.8 | -2.044 | Highly Destabilizing |
| 8 | 6m0j.pdb | E | N | 501 | Y | 0.4 | -1.981 | Destabilizing |
| 9 | 6m0j.pdb | E | Y | 505 | H | 17.0 | -1.884 | Destabilizing |

***Figure 9.*** The protein-protein affinity change upon mutation. A) The net spike RBD-ACE2 receptor affinity change upon mutations of EPI_ISL_5465615 is equal to ΔΔG= -0.412 kcal/mol. B) The net spike RBD-ACE2 receptor affinity change upon mutations of barcode 18 omicron is equal to ΔΔG= -7.214 kcal/mol.

All 96 sample sequences used in this study were obtained from Nur-Sultan during the 2-10th of August 2021 and 77 sequences that passed quality control are deposited in the Global initiative on sharing all influenza data (GISAID) (https://www.gisaid.org/hcov19-variants/ ) (figure 2). Accession IDs are included in the appendices section below. The obtained SARS-CoV-2 whole-genome sequences were used to generate a phylogenetic tree by the Nextstrain tool that classifies the differences between the uploaded and reference sequences to allocate our data into clades according to Nextclade nomenclature. It illustrates that 76 samples highlighted with blue circles belong to the 21J (Delta) variant, while one sample (highlighted with grey) is associated with an earlier origin 20C variant that is a genetically distinct subclade of the 20A group that originated at the beginning of 2020. Also, 7 sequences (highlighted with red) that belong to the 21K (omicron) variant were identified in this study (figure 3). Primarily, the majority of the sequenced samples, specifically, 69 viral genomes (or 89.6%) belong to the AY.122 lineage, 6 samples (7.8%) belong to B.1.617.2, one sample (1.3%) belongs to AY.89, and another sequence (1.3%) belongs to the B.1.637 lineage according to the PANGOLIN nomenclature system. The lineage distribution of the 77 viral genomes according to PANGOLIN is shown in figure 4. Viral whole-genome sequencing allows the identification of novel mutations and their frequency. The bar chart plot in figure 5 demonstrates the substitution and deletion mutation frequencies in the analyzed 77 SARS-CoV-2 samples. Substitution mutations are primarily found in S and ORF1 protein regions, 26.6% and 25.05%, respectively, while deletion mutations are predominantly detected in S and ORF8 regions, 35.89% and 35.43%, respectively. Generally, the S gene is the hotspot region for viral mutations in the virus.

The rare mutation in genome EPI_ISL_5465615, specifically, the Q271E mutation, was detected in the spike region. The 3D structural visualization of the spike protein

(EPI_ISL_5465615) and its interaction with the ACE2 receptor was built by CoVsurver (Figure 6A, B). The amino acid alterations in spike glycoprotein of EPI_ISL_5465615 query sequences are distinguished by colors. Specifically, the blue color is associated with amino acid alterations found more than 100 times, while the orange color represents the alterations that affect the phenotype in terms of host-cell receptor binding. Magenta-colored amino acid mutations stand for the addition or elimination of the glycosylation sites, whereas cyan-colored amino acid mutations facilitate the insertion or deletion of amino acid residues. While the effect of black-colored amino acid alterations is not known, they occurred only once in the current set of sequences.

The protein stability analysis upon insertion of specific mutations was performed by a site-directed mutator. A positive (blue) and negative (red) output correspond with mutations predicted to be stabilizing and destabilizing, respectively. The results tables in figure 7 demonstrate that the rare mutation Q271E of EPI_ISL_5465615 (delta) facilitates a slight increase in the stability of the spike protein up to $\Delta\Delta G=0.27$ kcal/mol and, generally, most mutations contribute to an increase in protein stability (figure 7A). The general predicted net pseudo $\Delta\Delta G$ of the protein is equal to 1.5 kcal/mol. The protein stability analysis of barcode 18 (omicron) suggests that the overall $\Delta\Delta G$ of spike protein stability increased to 3 kcal/mol (figure 7B).

The results of molecular dynamics analysis by the Desmond computational tool are demonstrated in figure 8. The molecular dynamics simulation RMSD output of spike RBD-ACE2 interaction plots of 6MOJ (Wuhan original), EPI_ISL_5465615, and barcode 18 (omicron) are compared. The graph of the ACE2 receptor and 6MOJ (PDB) Wuhan original SARS-CoV-2 strain spike protein interaction is shown in figure 8A. It shows that the system reaches equilibrium after 25 ns and the average RMSD is equal to

2.5 Å. Next, figure 8B illustrates the RMSD value stabilization at around 150 ns, and the average RMSD is equal to 3.3 Å. Figure 8C demonstrates weak RMSD value stabilization after 110 ns, and the average RMSD is equal to 3.8 Å.

The protein-protein affinity change upon mutation analysis was performed by the mCSM tool. It was found that the overall spike RBD-ACE2 receptor affinity change or ΔΔG for EPI_ISL_5465615 (delta) is equal to ΔΔG= -0.412 kcal/mol and for the barcode 18 omicron is equal to ΔΔG= -7.214 kcal/mol (figure 9). In other words, the protein-protein interaction stability was higher for the EPI_ISL_5465615 (delta) spike RBD-ACE2 in comparison with omicron protein.

## DISCUSSION

After successful viral whole-genome sequencing, the obtained data was analyzed according to the nCoV-2019 novel coronavirus bioinformatics protocol. Patient and sequenced genome characteristics can be seen in the appendix section. Most of the sequenced genomes had a GC content of around 38% in the coding sequence, which is consistent with previous research studies (Li et al., 2020). All the sequenced samples had coverage above 30X (Table 1). The amplicon length spanned around 200 to 1100 bp (mean 672, median 632). On average, the genomic coverage of the sequenced SARS-CoV-2 samples was ~99.8%, which validates the sequenced data quality. The low number of reads is associated with low RNA quality, and the low-quality reads (n=19) that have more than 3000 missing base pairs were eliminated from further analysis. Phylogenetic analysis was performed by using the Nexstrain protocol for the analysis of SARS-CoV-2 genomes, uploading 77 sequenced samples deposited in GISAID and 7 unpublished samples (21K omicron) from a recent run. The results of the later sequencing in March 2022 are also integrated in this work, but they have not been uploaded to the database yet. All the sequenced samples from March 3rd belong to the 21K or omicron variant and are illustrated in the phylogenetic tree (Figure 3). Out of 77 samples, 76 genomes (98.7%) were clustered under clade 21J, whereas 1 genome (barcode33 4-MN908947.3_new) was clustered under clade 20C, which is a large genetically distinct subclade of 20A that emerged at the beginning of 2020 (Figure 3). The sequencing analysis according to PANGOLIN nomenclature revealed that most of the samples (n=69) belonging to the AY.122 Pango lineage or Delta variant were circulating in Nur-Sultan city. Besides, 7.8% of B.1.617.2 (n=6) and AY.89 (n=1) lineage samples that belong to delta variants were identified (Figure 4). All these variants belong to the delta variant according to WHO nomenclature and GK as designated by GISAID. One B.1.637 sample belongs to a separate lineage corresponding to Iota (original B.1.526) or GH in GISAID nomenclature. Generally, viral whole genome sequencing confirmed that the SARS-CoV-2 variants in the region predominantly belong to the delta strain, which corresponds to a global trend. Essentially, viruses continually evolve as genetic mutations accumulate during the genomic replication step. A lineage is basically a group of viral variants that evolved from a common progenitor. A variant is a viral group that possesses one or more mutations that distinguish it from other viral variants. Since the emergence of the SARS-CoV-2 in December 2019, it has undergone various mutations that alter its characteristics, such as transmissibility, virulence, antigenicity, and vaccine efficiency (Cosar et al., 2021). Most of the mutations do not facilitate significant alterations in virulence. No novel mutations were detected in the analyzed genomes. The substitution mutations in the sequenced samples particularly occur in S and ORF1a, while deletion mutations mostly occur in S and ORF8 regions (Figure 5). Primarily, T19R, G142D, T478K, P681R, and D950N mutations were identified in 98.7% (n=76) of genomes, and L452R and D614G substitution mutations were found in 100% (n=77) spike protein regions. Of the sequenced samples. A488S, P1228L, V167L, T492I, T77A substitution mutations were found in 98.7% (n=76), P1469S substitution mutations were identified in 97.4% (n=75), and K81N substitution mutations in 89.6% (n=69) were found in sequenced genomes of ORF1a region. Notably, according to CoVsurver enabled by GISAID research tool analysis, in sample EPI_ISL_5465615, a Q271E rare substitution mutation was detected in the spike region. This amino acid alteration in spike occurred 75 times, that is <1% of all samples in 14 countries, first Q271E mutation was noted in July 2020 (hCoV-19/USA/ID-IBL-636276/2020) and the most recent one was in a strain collected in December 2021 (hCoV-19/Italy/VEN-IZSVe-21RS8297-8_VR/2021). Deletion mutations, specifically, S:E156- and S:F157- are found in all sequenced genomes in spike region and ORF8:D119- and ORF8:F120- deletions are found in 98.7% (n=76) ORF8 regions of sequenced genomes. Besides, frameshift mutations in the ORF7b: 15-44 region were detected in two genomes (EPI_ISL_5532919, EPI_ISL_5532921).

The spike protein stability analysis was performed to investigate the impact of each amino acid mutation on the pseudo ΔΔG value, which is the difference between the original Wuhan spike RBD (6MOJ PDB) and mutant proteins. The numerical value of each ΔΔG upon mutation is illustrated in figure 7, where the blue color represents the increase in stability while the red color indicates the decrease in protein stability. Since the EPI_ISL_5465615 was found to possess a rare mutation, it was used as a sample analysis by the site-directed mutator webtool. The output of the run suggests that the net increase of predicted ΔΔG was 1.5 kcal/mol, whereas the rare mutation Q271E slightly increased the stability of the spike protein of ΔΔG=0.27 kcal/mol (figure 7A). The sample barcode 18 (omicron) from a later run that has not yet been uploaded on GISAID yet was also considered for comparison between delta and omicron variants. As it can be seen in figure 7B, the net increase in predicted ΔΔG of spike protein is equal to 3 kcal/mol which is two times higher than the delta variant. So, it can be suggested that the higher number of mutations that developed in omicron that developed later in time in comparison to the delta variant facilitated the higher stability of the spike protein by approximately two times. In other words, predicted ΔΔG value specifies if the indicated mutations stabilize the mutated spike protein in its folded conformation or in its proper native structure to be functionally active state, consequently the higher ΔΔG of spike should facilitate higher virulence of the SARS-CoV-2.

The RMSD graph demonstrates the protein evolution (Y-axis) (figure 8). Firstly, all protein domains are aligned on the reference frame backbone, and afterwards the RMSD value is detected according to the atom selection. Since RMSD measures the change in position of carbon backbones from their initial to final conformation, the stability of the protein according to its conformation is found by alterations in accordance with its conformation changes during the stimulation. Recording the protein RMSD provides structural conformation analysis. The plot displays whether the system is equilibrated or not, in other words, if the fluctuations stabilize around an average structure. According to the manual, fluctuations of around 1-3 Å are expected for small globular proteins, while values beyond that suggest that the protein significantly changes its conformation during simulation.

Since the ACE2 receptor is a relatively large protein (~130kDa), the higher RMSD values are expected in our model. Besides, the RMSD value stabilization around a certain value suggests the system's convergence. Apart from being responsible for interaction with the ACE2 receptor, the RBD considered in this analysis is expected to facilitate the prefusion state, so the molecular stability of the protein-ligand complex is considered in this study.

Generally, all ACE2-spike RBD complexes were retained with a Cα RMSD from the initial structure maximum of 4.5Å which is quite higher than literature data (Amanat Ali and Ranjit Vijayan, 2020). It was expected that mutant variants would have lower RMSD values and less fluctuating plots in comparison with the original Wuhan strain RBD (figure 8A), but it was not observed in the obtained graphs. Besides, the molecular dynamics RMSD plot suggests that barcode 18 (omicron) RBD with the ACE2 receptor has a higher RMSD value and weaker stabilization of the graph in comparison to the EPI_ISL_5465615 RBD interaction with the receptor. Thus, the better ACE2 receptor and spike RBD is observed for delta strain than of that in omicron. It was not expected since the omicron variant should possess a higher number of mutation accumulation to facilitate higher affinity and, thus, better interaction with the host ACE2 receptor and a greater infectivity rate in patients.

The following protein-protein affinity change upon mutation analysis of spike RBD-ACE2 EPI_ISL_5465615 (delta) and barcode 18 (omicron) was performed by mCSM. The net spike RBD-ACE2 receptor affinity change, or ΔΔG for EPI_ISL_5465615 (delta) is equal to ΔΔG= -0.412 kcal/mol and for the barcode 18 omicron is equal to ΔΔG= -7.214 kcal/mol (figure 9). It suggests that the protein-protein interaction stability was higher for the EPI_ISL_5465615 (delta), while the mutations found in barcode 18 (omicron) are highly destabilizing. This result supports the previous molecular dynamics plot by Schrodinger. In other words, mutations in the omicron's spike RBD region facilitate destabilization of the binding to the ACE2 receptor. The literature data also supports that omicron RBD demonstrated weaker binding affinity to human ACE2 receptor in comparison with delta variant. Nevertheless, the omicron variant possesses a shorter incubation time and a higher rate of spread. Although the omicron strain has a weaker binding affinity to the ACE2 receptor, it has a higher rate of entry than that of the delta. A recently published study suggests that the SARS-CoV-2 entry rate into the host cell depends not only on thermodynamic properties, such as Gibbs energy, enthalpy, entropy, etc., but also on the kinetic properties, including the binding phenomenological coefficient (binding rate, kon and koff), and immune response (Popovic, 2020).

# CONCLUSION

Globally, many laboratories are proceeding to optimize the whole-genome sequencing of SARS-CoV-2 in terms of cost and efficiency to benefit epidemiological surveillance as the virus is mutating. As of March 2022, there are only a total of 596 viral genomes available on the GISAID platform including 77 sequences from this study submitted from Kazakhstan. Also, a paper draft of this study was submitted to the Frontiers in Genetics journal for publication and further sequencing analysis is expected soon. The whole genome of 96 COVID-19 samples collected in Nur-Sultan, in 2-10th of August were sequenced by ONT at National Laboratory Astana. To the best of our knowledge, this is the first study of SARS-CoV-2 whole-genome sequencing by the ONT approach in Kazakhstan. A significant increase in the number of COVID-19 cases confirmed in August 2021 in Kazakhstan facilitated a need to establish an effective scientific and proficient system for further study of SARS-CoV-2 genetic analysis. This study validates that ONT sequencing approach is a reasonable method for the viral monitoring of novel variants and mutations by whole genome sequencing during the pandemic.

Integration of genomic and phylogenetic examinations in the evaluation of epidemiological situations in the region would facilitate recognition of risk for viral transmission and the introduction of efficient preventive measures. The further high-throughput analysis and SARS-CoV-2 monitoring in Nur-Sultan city are expected by the GridION ONT sequencer at NLA. The results of the whole genome sequencing can significantly support the scientific foundation for the public health measures, thereby facilitating improvement of epidemiological situations and increase of public awareness.

# REFERENCE LIST

Amanat Ali and Ranjit Vijayan (2020). Dynamics of the ACE2 - SARS-CoV/SARS-CoV-2 spike protein interface reveal unique mechanisms. 151–156. doi:https://doi.org/10.1101/2020.06.10.143990.

Arnittali, M., Rissanou, A. N., and Harmandaris, V. (2019). Structure of Biomolecules Through Molecular Dynamics Simulations. *Procedia Comput. Sci.* 156, 69–78. doi:10.1016/j.procs.2019.08.181.

Brejová, B., Boršová, K., Hodorová, V., Čabanová, V., Gafurov, A., Fričová, D., et al. (2021). Nanopore sequencing of SARS-CoV-2: Comparison of short and long PCR-tiling amplicon protocols. *PLoS One* 16. doi:10.1371/journal.pone.0259277.

Bull, R. A., Adikari, T. N., Ferguson, J. M., Hammond, J. M., Stevanovski, I., Beukers, A. G., et al. (2020). Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nat. Commun.* 11, 1–8. doi:10.1038/s41467-020-20075-6.

Chan, J. F. W., Kok, K. H., Zhu, Z., Chu, H., To, K. K. W., Yuan, S., et al. (2020). Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg. Microbes Infect.* 9, 221–236. doi:10.1080/22221751.2020.1719902.

Cosar, B., Karagulleoglu, Z. Y., Unal, S., Ince, A. T., Uncuoglu, D. B., Tuncer, G., et al. (2021). SARS-CoV-2 Mutations and their Viral Variants. Cytokine Growth Factor Rev. doi:10.1016/j.cytogfr.2021.06.001.

Covid19.who.int. 2022. WHO Coronavirus (COVID-19) Dashboard. [online] Available at: <https://covid19.who.int/> [Accessed 5 April 2022].

Cui, J., Li, F., and Shi, Z. L. (2019). Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* 17, 181–192. doi:10.1038/s41579-018-0118-9.

Dudas, G., Carvalho, L. M., Bedford, T., Tatem, A. J., Baele, G., Faria, N. R., et al. (2017). Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature 544, 309–315. doi:10.1038/nature22040.

Forbes, J. D., Knox, N. C., Ronholm, J., Pagotto, F., and Reimer, A. (2017). Metagenomics: The next culture-independent game changer. *Front. Microbiol.* 8, 1–21. doi:10.3389/fmicb.2017.01069.

Freed, N., Vlkova, M., Faisal, M., Silander, O. (2020). Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. Biology Methods and Protocols, 2020, 1–7. doi:10.1093/biomethods/bpaa014.

Gardy, J., Loman, N. J., and Rambaut, A. (2015). Real-time digital pathogen surveillance - the time González-recio, is now. Genome Biol. 16, 15–17. doi:10.1186/s13059-015-0726-x.

Goldberg, B., Sichtig, H., Geyer, C., Ledeboer, N., and Weinstock, G. M. (2015). Making the leap from research laboratory to clinic: Challenges and opportunities for next-generation sequencing in infectious disease diagnostics. *MBio* 6. doi:10.1128/mBio.01888-15.

González-recio, O., Gutiérrez-rivas, M., Peiró-pastor, R., and Aguilera-sepúlveda, P. (2021). Sequencing of SARS-CoV-2 genome using different nanopore chemistries. 3225–3234.Li, Y., Yang, X., Wang, N., Wang, H., Yin, B., Yang, X., et al. (2020). GC usage of SARS-CoV-2 genes might adapt to the environment of human lung expressed genes. Mol. Genet. Genomics 295, 1537–1546. doi:10.1007/s00438-020-01719-0.

Harrison, A. G., Lin, T., and Wang, P. (2020). Mechanisms of SARS-CoV-2 Transmission and Pathogenesis. *Trends Immunol.* 41, 1100–1115. doi:10.1016/j.it.2020.10.004.

Jain, M., Olsen, H. E., Paten, B., and Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17, 1–11. doi:10.1186/s13059-016-1103-0.

Karplus, M., and McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* 9, 646–652. doi:10.1038/nsb0902-646.

Kumar, A., Prasoon, P., Kumari, C., Pareek, V., Faiq, M. A., Narayan, R. K., et al. (2021). SARS-CoV-2-specific virulence factors in COVID-19. *J. Med. Virol.* 93, 1343–1350. doi:10.1002/jmv.26615.

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., et al. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION.

Li, Y., Yang, X., Wang, N., Wang, H., Yin, B., Yang, X., et al. (2020). GC usage of SARS-CoV-2 genes might adapt to the environment of human lung expressed genes. Mol. Genet. Genomics 295, 1537–1546. doi:10.1007/s00438-020-01719-0.

Li, J., Wang, H., Mao, L., Yu, H., Yu, X., Sun, Z., et al. (2020). Rapid genomic characterization of SARS-CoV-2 viruses from clinical specimens using nanopore sequencing. Sci. Rep. 10, 1–10. doi:10.1038/s41598-020-74656-y.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., et al. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395, 565–574. doi:10.1016/S0140-6736(20)30251-8. *Biomol. Detect. Quantif.* 3, 1–8. doi:10.1016/j.bdq.2015.02.001.

Petersen, L. M., Martin, I. W., Moschetti, W. E., Kershaw, C. M., and Tsongalis, G. J. (2019). Third-Generation Sequencing in the Clinical Laboratory: Sequencing. *J. Clin. Microbiol.* 58, 1–10.

Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, K., et al. (2017). Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat. Protoc. 12, 1261–1266. doi:10.1038/nprot.2017.066.

Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., et al. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. 5, 1403–1407. doi:10.1038/s41564-020-0770-5.

Reva, B. A., Finkelstein, A. V., and Skolnick, J. (1998). What is the probability of a chance prediction of a protein structure with an rmsd of 6 Å? *Fold. Des.* 3, 141–147. doi:10.1016/S1359-0278(98)00019-4.

Schrodinger.com. 2022. *Introducing SID (Simulation Interactions Diagram) | Schrödinger*. [online] Available at:

<https://www.schrodinger.com/newsletters/introducing-sid-simulation-interactions-diagram> [Accessed 5 April 2022].

Song, Z., Xu, Y., Bao, L., Zhang, L., Yu, P., Qu, Y., et al. (2019). From SARS to MERS, thrusting coronaviruses into the spotlight. Viruses 11. doi:10.3390/v11010059.

Tillett, R. L., Sevinsky, J. R., Hartley, P. D., Kerwin, H., Crawford, N., Gorzalski, A., et al. (2021). Genomic evidence for reinfection with SARS-CoV-2: a case study. Lancet Infect. Dis. 21, 52–58. doi:10.1016/S1473-3099(20)30764-7.

To, K. K.-W., Hung, I. F.-N., Ip, J. D., Chu, A. W.-H., Chan, W.-M., Tam, A. R., et al. (2020). Coronavirus Disease 2019 (COVID-19) Re-infection by a Phylogenetically Distinct Severe Acute Respiratory Syndrome Coronavirus 2 Strain Confirmed by Whole Genome Sequencing. Clin. Infect. Dis. 2019, 1–6. doi:10.1093/cid/ciaa1275.

Pandurangan, A., Ochoa-Montaño, B., Ascher, D. and Blundell, T., 2022. SDM: Predict effects of mutation on protein stability. [online] Marid.bioc.cam.ac.uk. Available at: <http://marid.bioc.cam.ac.uk/sdm2/> [Accessed 5 April 2022].

Popovic, M. (2022). Strain wars 3: Differences in infectivity and pathogenicity between Delta and Omicron strains of SARS-CoV-2 can be explained by thermodynamic and kinetic parameters of binding and growth. *Microb. Risk Anal.,* 100217. doi:10.1016/j.mran.2022.100217.

Wang, L. F., Shi, Z., Zhang, S., Field, H., Daszak, P., and Eaton, B. T. (2006). Review of bats and SARS. *Emerg. Infect. Dis.* 12, 1834–1840. doi:10.3201/eid1212.060401.

Wang, M., Fu, A., Hu, B., Tong, Y., Liu, R., Liu, Z., et al. (2020). Nanopore Targeted Sequencing for the Accurate and Comprehensive Detection of SARS-CoV-2 and Other Respiratory Viruses. Small 16. doi:10.1002/smll.202002169.

Wong, A. C. P., Li, X., Lau, S. K. P., and Woo, P. C. Y. (2019). Global epidemiology of bat coronaviruses. *Viruses* 11, 1–17. doi:10.3390/v11020174.

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., et al. (2020). A new coronavirus associated with human respiratory disease in China. Nature 579, 265–269. doi:10.1038/s41586-020-2008-3.

Yadav, R., Chaudhary, J. K., Jain, N., Chaudhary, P. K., Khanra, S., Dhamija, P., et al. (2021). Role of structural and non-structural proteins and therapeutic targets of SARS-CoV-2 for COVID-19. Cells 10, 1–16. doi:10.3390/cells10040821.

Yegorov, S., Goremykina, M., Ivanova, R., Good, S. V., Babenko, D., Shevtsov, A., et al. (2021). Epidemiology, clinical characteristics, and virologic features of COVID-19 patients in Kazakhstan: A nation-wide retrospective cohort study. Lancet Reg. Heal. - Eur. 4, 100096. doi:10.1016/j.lanepe.2021.100096.

Zhalmagambetov, B., Madikenova, M., Paizullayeva, S., Abbay, A., and Gaipov, A. (2020). COVID-19 Outbreak in Kazakhstan: Current Status and Challenges. J. Clin. Med. Kazakhstan 1, 6–8. doi:10.23950/1812-2892-jcmk-00763.

Zhang, J., Xiao, T., Cai, Y., and Chen, B. (2020). Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID- 19 . The COVID-19 resource centre is hosted on Elsevier Connect , the company ’ s public news and information .

Zhou, P., Yang, X. Lou, Wang, X. G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270–273. doi:10.1038/s41586-020-2012-7.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. N. Engl. J. Med. 382, 727–733. doi:10.1056/nejmoa2001017.

# APPENDICES

Table 1. Summary characteristics of sequenced SARS-CoV-2 samples by ONT

| | **Sample ID** | **GISAID ID** | **Sample collection date** | **Gender** | **Ct value** | **Lineage** | **Genome length** | **GC-content** | **Depth** | **Ns** |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | hCoV-19/Kazakhstan/NLA/barcode01-4-MN908947.3/2021 | EPI_ISL_5465367 | 02.08.21 | Female | 15,24 | B.1.617.2 | 29 576 | 38% | 372X | 316 |
| **2** | hCoV-19/Kazakhstan/NLA/barcode02-4-MN908947.3/2021 | EPI_ISL_5465373 | 02.08.21 | Male | 11,89 | B.1.617.2 | 29 576 | 38% | 374X | 314 |
| **3** | hCoV-19/Kazakhstan/NLA/barcode03-4-MN908947.3/2021 | EPI_ISL_5465381 | 02.08.21 | Male | 14,87 | B.1.617.2 | 29 576 | 38% | 360X | 315 |
| **4** | hCoV-19/Kazakhstan/NLA/barcode04-4-MN908947.3/2021 | EPI_ISL_5465387 | 02.08.21 | Female | 13,97 | B.1.617.2 | 29 549 | 39% | 367X | 314 |
| **5** | hCoV-19/Kazakhstan/NLA/barcode05-4-MN908947.3/2021 | EPI_ISL_5465398 | 02.08.21 | Male | 13,45 | B.1.617.2 | 29 578 | 38% | 362X | 312 |
| **6** | hCoV-19/Kazakhstan/NLA/barcode06-4-MN908947.3/2021 | EPI_ISL_5465404 | 02.08.21 | Male | 13,18 | B.1.617.2 | 29 578 | 38% | 372X | 313 |
| **7** | hCoV-19/Kazakhstan/NLA/barcode07-4-MN908947.3/2021 | EPI_ISL_5465410 | 02.08.21 | Female | 17,54 | B.1.617.2 | 29 576 | 38% | 365X | 314 |
| **8** | hCoV-19/Kazakhstan/NLA/barcode08-4-MN908947.3/2021 | EPI_ISL_5465417 | 02.08.21 | Male | 13,05 | B.1.617.2 | 29 576 | 38% | 351X | 315 |
| **9** | hCoV-19/Kazakhstan/NLA/barcode09-4-MN908947.3/2021 | EPI_ISL_5465422 | 02.08.21 | Female | 14,91 | B.1.617.2 | 29 576 | 38% | 369X | 314 |
| **10** | hCoV-19/Kazakhstan/NLA/barcode10-4-MN908947.3/2021 | EPI_ISL_5465425 | 02.08.21 | Female | 11,79 | B.1.617.2 | 29 576 | 38% | 369X | 315 |
| **11** | hCoV-19/Kazakhstan/NLA/barcode11-4-MN908947.3/2021 | EPI_ISL_5465431 | 02.08.21 | Male | 14,72 | B.1.617.2 | 29 576 | 38% | 361X | 314 |
| **12** | hCoV-19/Kazakhstan/NLA/barcode12-4-MN908947.3/2021 | EPI_ISL_5465438 | 02.08.21 | Male | 19,10 | B.1.617.2 | 29 584 | 39% | 352X | 306 |
| **13** | hCoV-19/Kazakhstan/NLA/barcode13-4-MN908947.3/2021 | EPI_ISL_5465445 | 02.08.21 | Male | 13,13 | B.1.617.2 | 29 576 | 38% | 367X | 314 |
| **14** | hCoV-19/Kazakhstan/NLA/barcode14-4-MN908947.3/2021 | EPI_ISL_5465452 | 02.08.21 | Male | 8,57 | B.1.617.2 | 29 576 | 38% | 368X | 314 |
| **15** | hCoV-19/Kazakhstan/NLA/barcode15- | EPI_ISL_5465456 | 02.08.21 | Male | 17,83 | B.1.617.2 | 29 576 | 38% | 347X | 315 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 4-MN908947.3/2021 | | | | | | | | | |
| **16** | hCoV-19/Kazakhstan/NLA/barcode16-4-MN908947.3/2021 | EPI_ISL_5465463 | 02.08.21 | Female | 11,65 | B.1.617.2 | 29 576 | 37% | 371X | 314 |
| **17** | hCoV-19/Kazakhstan/NLA/barcode17-4-MN908947.3/2021 | EPI_ISL_5465470 | 02.08.21 | Male | 15,56 | B.1.617.2 | 29 576 | 38% | 363X | 314 |
| **18** | hCoV-19/Kazakhstan/NLA/barcode18-4-MN908947.3/2021 | EPI_ISL_5465480 | 02.08.21 | Male | 10,76 | B.1.617.2 | 29 573 | 38% | 352X | 318 |
| **19** | hCoV-19/Kazakhstan/NLA/barcode19-4-MN908947.3/2021 | EPI_ISL_5465482 | 02.08.21 | Male | 16,30 | B.1.617.2 | 29 581 | 38% | 356X | 309 |
| **20** | hCoV-19/Kazakhstan/NLA/barcode20-4-MN908947.3/2021 | EPI_ISL_5465491 | 03.08.21 | Male | 10,15 | B.1.617.2 | 29 576 | 38% | 370X | 315 |
| **21** | hCoV-19/Kazakhstan/NLA/barcode21-4-MN908947.3/2021 | EPI_ISL_5465494 | 03.08.21 | Female | 13,58 | B.1.617.2 | 29 576 | 38% | 360X | 314 |
| **22** | hCoV-19/Kazakhstan/NLA/barcode22-4-MN908947.3/2021 | EPI_ISL_5465502 | 03.08.21 | Female | 13,63 | B.1.617.2 | 29 576 | 38% | 368X | 314 |
| **23** | hCoV-19/Kazakhstan/NLA/barcode23-4-MN908947.3/2021 | EPI_ISL_5465505 | 03.08.21 | Female | 11,51 | B.1.617.2 | 29 586 | 38% | 371X | 314 |
| **24** | hCoV-19/Kazakhstan/NLA/barcode24-4-MN908947.3/2021 | EPI_ISL_5465507 | 03.08.21 | Female | 12,85 | B.1.617.2 | 29 586 | 37% | 322X | 565 |
| **25** | hCoV-19/Kazakhstan/NLA/barcode25-4-MN908947.3/2021 | EPI_ISL_5465514 | 03.08.21 | Male | 9,92 | B.1.617.2 | 29 576 | 38% | 372X | 314 |
| **26** | hCoV-19/Kazakhstan/NLA/barcode26-4-MN908947.3/2021 | EPI_ISL_5465518 | 03.08.21 | Female | 6,48 | B.1.617.2 | 29 560 | 38% | 317X | 333 |
| **27** | hCoV-19/Kazakhstan/NLA/barcode27-4-MN908947.3/2021 | EPI_ISL_5465523 | 03.08.21 | Female | 17,87 | B.1.617.2 | 29 576 | 38% | 357X | 316 |
| **28** | hCoV-19/Kazakhstan/NLA/barcode28-4-MN908947.3/2021 | EPI_ISL_5465533 | 03.08.21 | Male | 17,85 | B.1.617.2 | 29 576 | 38% | 362X | 316 |
| **29** | hCoV-19/Kazakhstan/NLA/barcode29-4-MN908947.3/2021 | EPI_ISL_5532919 | 03.08.21 | Male | 16,55 | B.1.617.2 | 29 574 | 38% | 367X | 314 |
| **30** | hCoV-19/Kazakhstan/NLA/barcode30-4-MN908947.3/2021 | EPI_ISL_5465539 | 03.08.21 | Male | 13,94 | B.1.617.2 | 29 580 | 38% | 357X | 310 |
| **31** | hCoV-19/Kazakhstan/NLA/barcode31-4-MN908947.3/2021 | EPI_ISL_5465542 | 03.08.21 | Female | 17,94 | B.1.617.2 | 29 571 | 38% | 358X | 320 |
| **32** | hCoV-19/Kazakhstan/NLA/barcode33-4-MN908947.3/2021 | EPI_ISL_5465549 | 04.08.21 | Male | 14,27 | B.1.637 | 29 576 | 38% | 362X | 315 |
| **33** | hCoV-19/Kazakhstan/NLA/barcode34- | EPI_ISL_5465553 | 04.08.21 | Male | 10,85 | B.1.617.2 | 29 576 | 38% | 369X | 314 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 4-MN908947.3/2021 | | | | | | | | | |
| **34** | hCoV-19/Kazakhstan/NLA/barcode35-4-MN908947.3/2021 | EPI_ISL_5465563 | 04.08.21 | Female | 10,09 | B.1.617.2 | 29 576 | 38% | 365X | 315 |
| **35** | hCoV-19/Kazakhstan/NLA/barcode36-4-MN908947.3/2021 | EPI_ISL_5465570 | 04.08.21 | Male | 10,19 | B.1.617.2 | 29 576 | 39% | 370X | 314 |
| **36** | hCoV-19/Kazakhstan/NLA/barcode37-4-MN908947.3/2021 | EPI_ISL_5465574 | 04.08.21 | Female | 7,27 | B.1.617.2 | 29 586 | 38% | 370X | 304 |
| **37** | hCoV-19/Kazakhstan/NLA/barcode38-4-MN908947.3/2021 | EPI_ISL_5465584 | 04.08.21 | Male | 15,01 | AY.39 | 29 586 | 38% | 315X | 553 |
| **38** | hCoV-19/Kazakhstan/NLA/barcode39-4-MN908947.3/2021 | EPI_ISL_5465593 | 04.08.21 | Male | 11,07 | B.1.617.2 | 29 581 | 38% | 373X | 309 |
| **39** | hCoV-19/Kazakhstan/NLA/barcode41-4-MN908947.3/2021 | EPI_ISL_5465598 | 04.08.21 | Female | 12,35 | B.1.617.2 | 29 570 | 38% | 358X | 323 |
| **40** | hCoV-19/Kazakhstan/NLA/barcode42-4-MN908947.3/2021 | EPI_ISL_5465603 | 04.08.21 | Female | 11,08 | B.1.617.2 | 29 565 | 38% | 358X | 326 |
| **41** | hCoV-19/Kazakhstan/NLA/barcode43-4-MN908947.3/2021 | EPI_ISL_5465608 | 04.08.21 | Male | 17,76 | B.1.617.2 | 29 576 | 37% | 338X | 315 |
| **42** | hCoV-19/Kazakhstan/NLA/barcode44-4-MN908947.3/2021 | EPI_ISL_5465615 | 04.08.21 | Male | 16,42 | B.1.617.2 | 29 576 | 38% | 363X | 314 |
| **43** | hCoV-19/Kazakhstan/NLA/barcode45-4-MN908947.3/2021 | EPI_ISL_5465620 | 04.08.21 | Female | 12,05 | B.1.617.2 | 29 576 | 38% | 364X | 314 |
| **44** | hCoV-19/Kazakhstan/NLA/barcode46-4-MN908947.3/2021 | EPI_ISL_5465627 | 04.08.21 | Female | 18,69 | B.1.617.2 | 29 575 | 38% | 347X | 310 |
| **45** | hCoV-19/Kazakhstan/NLA/barcode47-4-MN908947.3/2021 | EPI_ISL_5465636 | 04.08.21 | Female | 7,38 | B.1.617.2 | 29 563 | 38% | 369X | 328 |
| **46** | hCoV-19/Kazakhstan/NLA/barcode49-4-MN908947.3/2021 | EPI_ISL_5465643 | 05.08.21 | Male | 10,81 | B.1.617.2 | 29 582 | 39% | 371X | 308 |
| **47** | hCoV-19/Kazakhstan/NLA/barcode50-4-MN908947.3/2021 | EPI_ISL_5465647 | 05.08.21 | Male | 12,59 | B.1.617.2 | 29 574 | 38% | 348X | 316 |
| **48** | hCoV-19/Kazakhstan/NLA/barcode51-4-MN908947.3/2021 | EPI_ISL_5465654 | 05.08.21 | Female | 15,67 | B.1.617.2 | 29 571 | 38% | 363X | 320 |
| **49** | hCoV-19/Kazakhstan/NLA/barcode52-4-MN908947.3/2021 | EPI_ISL_5465658 | 05.08.21 | Male | 10,63 | B.1.617.2 | 29 576 | 38% | 360X | 314 |
| **50** | hCoV-19/Kazakhstan/NLA/barcode53-4-MN908947.3/2021 | EPI_ISL_5465664 | 05.08.21 | Female | 14,10 | B.1.617.2 | 29 576 | 39% | 367X | 314 |
| **51** | hCoV-19/Kazakhstan/NLA/barcode55- | EPI_ISL_5532920 | 06.08.21 | Female | 17,09 | AY.47 | 29 509 | 38% | 370X | 314 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 4-MN908947.3/2021 | | | | | | | | | |
| **52** | hCoV-19/Kazakhstan/NLA/barcode57-4-MN908947.3/2021 | EPI_ISL_5465668 | 06.08.21 | Female | 16,18 | B.1.617.2 | 29 574 | 38% | 352X | 318 |
| **53** | hCoV-19/Kazakhstan/NLA/barcode58-4-MN908947.3/2021 | EPI_ISL_5465675 | 06.08.21 | Male | 17,10 | B.1.617.2 | 29 562 | 38% | 338X | 313 |
| **54** | hCoV-19/Kazakhstan/NLA/barcode59-4-MN908947.3/2021 | EPI_ISL_5465681 | 06.08.21 | Male | 13,61 | B.1.617.2 | 29 531 | 38% | 371X | 314 |
| **55** | hCoV-19/Kazakhstan/NLA/barcode60-4-MN908947.3/2021 | EPI_ISL_5465687 | 06.08.21 | Male | 19,65 | B.1.617.2 | 29 575 | 38% | 360X | 316 |
| **56** | hCoV-19/Kazakhstan/NLA/barcode61-4-MN908947.3/2021 | EPI_ISL_5465694 | 06.08.21 | Female | 13,03 | B.1.617.2 | 29 576 | 38% | 366X | 314 |
| **57** | hCoV-19/Kazakhstan/NLA/barcode62-4-MN908947.3/2021 | EPI_ISL_5465699 | 06.08.21 | Female | 11,16 | B.1.617.2 | 29 581 | 38% | 368X | 309 |
| **58** | hCoV-19/Kazakhstan/NLA/barcode63-4-MN908947.3/2021 | EPI_ISL_5465707 | 06.08.21 | Female | 15,44 | B.1.617.2 | 29 586 | 38% | 368X | 304 |
| **59** | hCoV-19/Kazakhstan/NLA/barcode65-4-MN908947.3/2021 | EPI_ISL_5465713 | 06.08.21 | Male | 11,01 | B.1.617.2 | 29 576 | 39% | 365X | 315 |
| **60** | hCoV-19/Kazakhstan/NLA/barcode66-4-MN908947.3/2021 | EPI_ISL_5465722 | 06.08.21 | Male | 10,91 | B.1.617.2 | 29 576 | 38% | 366X | 315 |
| **61** | hCoV-19/Kazakhstan/NLA/barcode67-4-MN908947.3/2021 | EPI_ISL_5465730 | 06.08.21 | Female | 17,98 | B.1.617.2 | 29 576 | 37% | 354X | 316 |
| **62** | hCoV-19/Kazakhstan/NLA/barcode68-4-MN908947.3/2021 | EPI_ISL_5532921 | 07.08.21 | Female | 17,73 | B.1.617.2 | 29 562 | 39% | 342X | 327 |
| **63** | hCoV-19/Kazakhstan/NLA/barcode69-4-MN908947.3/2021 | EPI_ISL_5465739 | 07.08.21 | Female | 16,57 | B.1.617.2 | 29 576 | 38% | 348X | 315 |
| **64** | hCoV-19/Kazakhstan/NLA/barcode70-4-MN908947.3/2021 | EPI_ISL_5465744 | 07.08.21 | Female | 12,53 | B.1.617.2 | 29 576 | 38% | 361X | 314 |
| **65** | hCoV-19/Kazakhstan/NLA/barcode71-4-MN908947.3/2021 | EPI_ISL_5532922 | 07.08.21 | Male | 17,34 | B.1.617.2 | 29 556 | 37% | 354X | 325 |
| **66** | hCoV-19/Kazakhstan/NLA/barcode73-4-MN908947.3/2021 | EPI_ISL_5532923 | 07.08.21 | Female | 19,98 | AY.47 | 29 476 | 39% | 323X | 324 |
| **67** | hCoV-19/Kazakhstan/NLA/barcode74-4-MN908947.3/2021 | EPI_ISL_5465751 | 07.08.21 | Male | 11,51 | AY.39 | 29 570 | 38% | 364X | 322 |
| **68** | hCoV-19/Kazakhstan/NLA/barcode75-4-MN908947.3/2021 | EPI_ISL_5532924 | 07.08.21 | Female | 11,92 | B.1.617.2 | 29 485 | 38% | 368X | 314 |
| **69** | hCoV-19/Kazakhstan/NLA/barcode76- | EPI_ISL_5465755 | 08.08.21 | Female | 10,90 | B.1.617.2 | 29 570 | 38% | 350X | 320 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 4-MN908947.3/2021 | | | | | | | | | |
| **70** | hCoV-19/Kazakhstan/NLA/barcode77-4-MN908947.3/2021 | EPI_ISL_5465759 | 08.08.21 | Female | 14,54 | B.1.617.2 | 29 576 | 38% | 353X | 317 |
| **71** | hCoV-19/Kazakhstan/NLA/barcode78-4-MN908947.3/2021 | EPI_ISL_5465764 | 08.08.21 | Female | 14,09 | B.1.617.2 | 29 576 | 38% | 365X | 314 |
| **72** | hCoV-19/Kazakhstan/NLA/barcode79-4-MN908947.3/2021 | EPI_ISL_5465770 | 08.08.21 | Female | 15,60 | B.1.617.2 | 29 570 | 38% | 368X | 320 |
| **73** | hCoV-19/Kazakhstan/NLA/barcode80-4-MN908947.3/2021 | EPI_ISL_5532925 | 08.08.21 | Female | 14,32 | B.1.617.2 | 29 478 | 37% | 354X | 315 |
| **74** | hCoV-19/Kazakhstan/NLA/barcode81-4-MN908947.3/2021 | EPI_ISL_5465778 | 08.08.21 | Male | 14,98 | B.1.617.2 | 29 576 | 38% | 354X | 314 |
| **75** | hCoV-19/Kazakhstan/NLA/barcode82-4-MN908947.3/2021 | EPI_ISL_5465785 | 08.08.21 | Female | 15,61 | B.1.617.2 | 29 576 | 38% | 355X | 315 |
| **76** | hCoV-19/Kazakhstan/NLA/barcode83-4-MN908947.3/2021 | EPI_ISL_5465790 | 08.08.21 | Female | 18,13 | B.1.617.2 | 29 576 | 38% | 362X | 314 |
| **77** | hCoV-19/Kazakhstan/NLA/barcode84-4-MN908947.3/2021 | EPI_ISL_5465795 | 08.08.21 | Female | 12,06 | B.1.617.2 | 29 576 | 38% | 370X | 314 |