ELUCIDATION OF PERSISTENT MUTATIONAL LANDSCAPE IN SARS-CoV-2 MAIN PROTEASE: A STRUCTURAL BIOINFORMATICS ANALYSIS

by

Winnie Wezi Mkandawire

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Master of Science

in

Bioinformatics and Computational Biology

May 2021

APPROVED:

Prof Dmitry Korkin

Prof Celia Schiffer

ACKNOWLEDGEMENT

First and Foremost, I am extremely grateful to my best advisors, Prof. Dmitry Korkin and Prof. Celia Schiffer for their invaluable advice, continuous intellectual, physical, psychosocial, and financial support, and patience during my Masters study. Their immense knowledge and plentiful experience and guidance have encouraged me in all the time of my academic research and personal daily life. My gratitude extends to the US Government of State through the Fulbright Scholarship for the funding opportunity to undertake my studies at the Department of Bioinformatics and Computational Biology, Worcester Polytechnic Institute (WPI).

I would also like to thank Dr. Shurong Hou, Dr. Florian Leidner and Prof. Elizabeth Ryder for their technical support and mentorship in my study. Let me also extend my appreciation to my special friend Katy Monopoli and all my lab mates in the Korkin Lab at WPI and Schiffer Lab at University of Massachusetts Medical School. It is their kind help and support that have made my study and life in the USA wonderful.

Finally, I would like to express my gratitude to my dear best friend Trusting Inekwe, my friend Raymond Magambo, my son Ryan and my son's father, Reuben Moyo. Without their tremendous understanding and encouragement throughout my journey, it would be impossible for me to complete my study.

Abstract

SARS-CoV-2 is a major public health burden whose spread and severity has dramatically increased causing overwhelming rise in morbidity and mortality alongside economic crisis worldwide. The virus's Main protease (Mpro) enzyme is one of the drug targets that has been widely studied to combat coronaviruses. This protease plays a crucial role in the process of viral maturation and replication, therefore, inhibiting it would reduce viral load and thus alleviate symptom intensity. The design of robust inhibitors against the Mpro requires characterization of the fixed viral genomic mutational landscape and the populations of conformations it engenders. Studies suggest that fixed or pervasive mutations indicate evolution or adaptation of virus to its niche. Thus, central to the success of drug design efforts is developing an understanding of structural and functional variations in the enzyme, particularly as mutations become persistent and new strains emerge. In this work, we attempted to detect if the human SARS-CoV-2 Mpro is undergoing selective pressure due to pervasive mutations and tried to infer the collective effects of fixed positively selected mutations on protease functionality. We analyzed global population isolates of human SARS-CoV-2, downloaded from the GISAID database as of February 9th 2021. Overall, mutations were seen at 169 sites with each enzyme having 1-6 changes – with 16 positions showing significant persistent variations subjected to selection pressure and 11 variants having positive selection. Interestingly, these mutations showed a trend towards substitution for larger and more hydrophobic residues when compared with the wild type SARS-CoV-2 sequence. Additionally, when mapped onto the 3D structure of the reference protein, 3 of the 11 positively selected significant variations were located closer to the active site. Using *in silico* approaches, we speculate that these mutations may have beneficial effects to the protease functionality and hence signify adaptation of SARS-CoV-2 to the human niche. This study will help uncover evolutionary mechanisms of adaptation and resistance in SARS-CoV-2 Mpro that can be targeted with inhibitors designed to be robust to the resistance and in turn help treat this deadly infection.

KEYWORDS: SARS-CoV-2, Main Protease, spatial-temporal, mutation, pervasive, selection, molecular dynamics

Table of Contents

CHAPTER 1 – INTRODUCTION

1.1. Background

1.1.1. Burden of Severe Acute Respiratory Syndrome CoronaVirus-2 (SARS-CoV-2)

Coronaviruses (CoVs) are enveloped, positive-sensed, single stranded RNA viruses belonging to the Coronaviridae family (1). Based on their phylogenetic analyses and antigenic properties, CoVs have been categorized as (2): (a) alpha-CoVs, mainly responsible for gastrointestinal disorders; (b) beta-CoVs, that include: (i) bat coronavirus, (ii) the human severe acute respiratory syndrome (SARS) virus, (iii) the Middle Eastern respiratory syndrome (MERS) virus; (c) gamma-CoVs, causing infections in avian species. CoVs's variants are usually associated with different outcomes with some getting associated with outbreaks, while which are continuously circulating, cause respiratory infections that range from common cold to much more serious infections. The most well-known of these CoVs is the SARS-CoV-1, which between 2002 and 2003 was responsible to cause an outbreak that spread around the world and resulted in over 8000 cases and 774 deaths, with a case fatality rate of around 9% to 11% (2).

In 2012, a novel CoV, MERS-CoV, causing severe respiratory symptoms was identified (3). In December 2019, a novel beta-CoV (SARS-CoV-2) emerged, first detected in Wuhan, China (4) and rapidly spread worldwide, causing an ongoing pandemic – COVID-19 (4,5). Although the sequence of SARS-CoV-2's RNA genome is highly similar to that of SARS-CoV-1, SARS-CoV-2 is believed to have arisen independently from a bat coronavirus (6), to which it shares 96% similarity (7). As of February 9th, 2021, more than 110 million people were infected with the virus worldwide, with more than 2.5 million deaths reported https://covid19.who.int/. The spread and severity of the virus has been dramatically increasing across the world and the burden that this pandemic is causing on people's livelihood and the economy globally is so overwhelming (8,9).

Drastic measures designed to limit the rate of new infections (9) have resulted in global economic problems, which have affected many livelihoods, even exacerbating food insecurity (10). This led to an increase of research into potential drugs alongside clinical trials owing to the rapid generation of genomic sequence data (11,12) and the well-timed availability of 3D structural data. Fundamental research is key to understanding the pathogen's strategies such that more informed decisions can be made about clinical interventions.

Early work on SARS-CoV-2 virus has enabled *in silico* studies suggesting potential solutions to the COVID-19 pandemic using various techniques, including the use of molecular modeling, network-analysis (13–16) and machine learning (17–21); this is with the help of experimentally determined structures, genomic data and annotations. Collectively, the goal is to design potential antivirals that support ongoing vaccination efforts. No antiviral drugs that are able to reduce SARS-CoV-2 mortality in clinical settings are yet known, although extensive efforts are underway to discover them or repurpose existing ones to inhibit key viral proteins. However, these efforts are hampered by fragmentary knowledge of viral structural biology and symptomatology of the disease (22,23), while the death toll and the number of infections keeps on rising. Thus, time is of the essence for the discovery of effective antiviral drugs in addition to the approved vaccines (24). It is crucial to better characterize parts of the viral mechanisms in the viral lifecycle (Figure 1) to better understand the behavior of the virus.

1.1.2. Life Cycle and Genome Organization of SARS-CoV-2

Viral infection is initiated by the interaction between the Spike (S) protein and human angiotensin-converting enzyme 2 (ACE2), followed by subsequent endocytosis or membrane fusion. The S protein comprises two subunits: S1 subunit which contains the receptor binding domain (RBD) and binds to N-terminal ACE2; and the the S2 subunit which mediates virus-host membrane fusion. S proteins are cleaved by the host cell furin protease and transmembrane

serine protease 2 (TMPRSS2) at the S1/S2 boundary and the S2′ position. Proteolytic cleavage at the S1/S2 boundary is thought to promote TMPRSS2-dependent entry into the target cells (25,26).

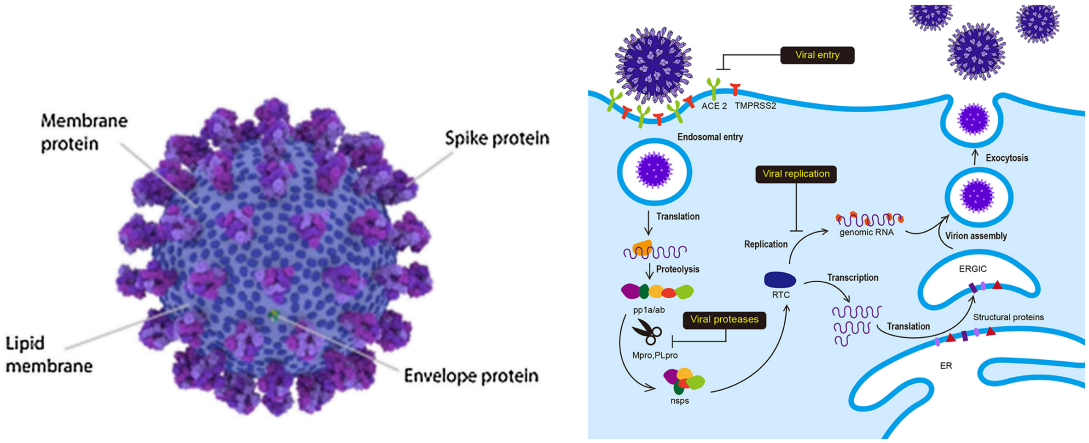


Figure 1. Viral lifecycle of SARS-CoV-2. Source: Jeong et al., Frontiers in Microbiology, July 2020

After the release of the viral polycistronic RNA into the cytoplasm, the replicase gene comprising open reading frames (ORFs) 1a and 1ab is directly translated into either replicase polyprotein pp1a (non-structural proteins (nsp) 1-11) or pp1ab (nsp1-16) and autoproteolytically cleaved into 16 non-structural proteins (nsp1-16) by two ORF1a-encoded protease domains: the main protease – Mpro (also called $3CL^{pro}$) and papain-like protease – PL^{pro} (Figure 2). These two viral proteases participate in this extensive proteolytic cleavage. Then the synthesis of the full-length genome (replication) or discontinuous mRNAs (transcription) is done through the mediation of a replicase-transcriptase complex (RTC) – a cytoplasmic enzyme complex (27). Structural and accessory proteins are subsequently translated from these transcripts, and new viruses assemble by budding into the lumen of the endoplasmic reticulum-Golgi intermediate compartment (ERGIC) and are eventually secreted (28).

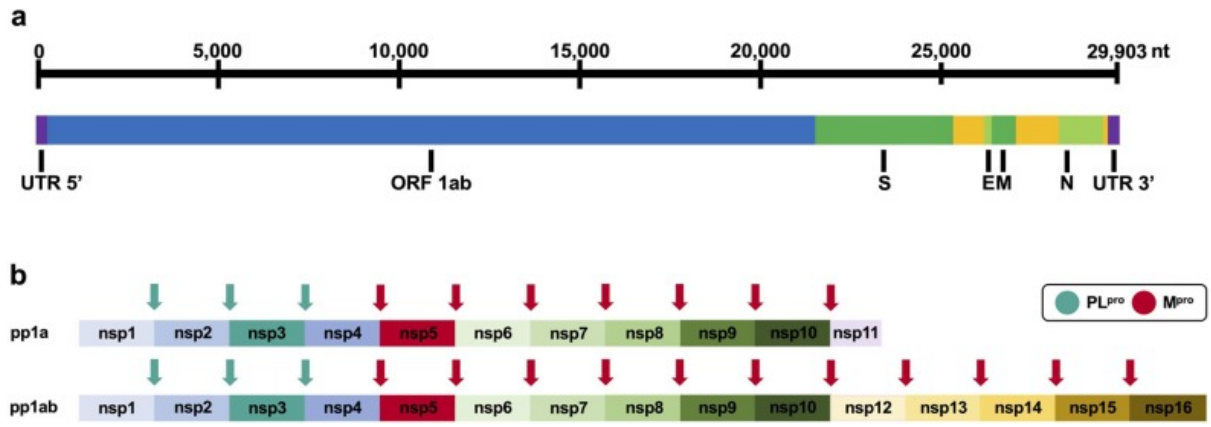


Figure 2. (a) Organization of the RNA genome of SARS-CoV-2 with selected genes. (b) Schematic representation of polyprotein cleavage sites of SARS-CoV-2. The papain-like protease PL^{pro} cleaves at 3 distinct sites. The main protease Mpro cleaves at 11 distinct sites. Source: Ullrich et al., Bioorg. Med. Chem. Lett, 2020

1.1.3. Viral Proteases and their Role in Viral Infection

The identification of drugs to treat coronaviruses is an urgent global need; researchers all around the world are exploiting different strategies targeting both viral and host factors essential for the pathogen replication to block one or more steps of its life cycles, and their efforts have already produced a great amount of data and studies. Viral proteases are one of the targets of direct-acting agents in antiviral therapy due to their vital role in viral replication and maturation (29). Also called a peptidases or proteinases, viral proteases are encoded by the genetic material (DNA or RNA) of viral pathogens (Figure 2). Their role is to perform proteolysis, that is, protein catabolism by hydrolysis of peptide bonds in viral polyprotein precursors or in cellular proteins.

For catalytic action, proteases possess an active site that consists of a binding site and a catalytic site, both constituting residues. The binding site consist of residues that form temporary bonds with the substrate and the catalytic site consist of either serine, cysteine or aspartic acid which catalyze peptide bond cleavage (30–32). The active site in the enzyme is usually a groove or pocket located in a deep tunnel within the enzyme. The residues of the catalytic site are typically very close to the binding site, and some residues can have dual roles, in both binding and catalysis. Once the substrate is bound and oriented in the active site,

catalysis can begin. The active site has conserved sequence motifs extending for up to ten residues (32). Selective recognition of these sequence patterns by a complementary substrate binding site of the enzyme ensures a high degree of specific recognition and cleavage. Studies have revealed proteases as potential druggable targets due to their site-specific binding and cleavage as well as their sequence motif conservation across coronaviruses (31–38).

In SARS-CoV-2 infection, two viral proteases: the main protease (Mpro) and the Papain-like protease (PLpro) play a pivotal role inside the human host cells through mediating viral replication and viral protein maturation, as illustrated in Figure 1. These proteases have been widely studied as potential SARS-CoV-2 drug targets (33,37). SARS-CoV-2 Mpro show the highest degree of conservation across coronaviruses fostering the identification of broad-spectrum inhibitors(33,35,39). SARS-CoV-2 PLpro on the other hand, is poorly characterized and not equally conserved, limiting the identification of broad-spectrum agents (40).

1.1.3.1. Overview of SARS-CoV-2 Main protease (Mpro): Structure and Function

SARS-CoV-2 Mpro, listed as one of potential best drug targets in treating SARS-CoV-2 virus by the World Health Organization, has been widely studied by researchers worldwide. This is mainly due to the similarities in active site and mechanisms with the related pathogenic beta-coronaviruses from previous epidemics of SARS-CoV and MERS-CoV (41). Mpro is a conserved drug target present in all members of the *Coronaviridae* subfamily (42,43) and is highly similar to its SARS-CoV counterpart (41). SARS-CoV-2 Mpro has been exceptionally regarded as one of the best targets due to the fact that it is specific to recognizing and cleaving sequences at Leu-Gln↓(Ser, Ala, Gly) (↓ marks the cleavage site (16). Besides, because no human proteases with a similar cleavage specificity are known, designing inhibitors targeting Mpro is unlikely to be toxic (21, 34), which reduces the chances of accidentally targeting host proteins.

Functionally, Mpro plays an crucial role in the process of viral maturation (17), cleaving the large precursor replicase polyprotein 1ab to produce 16 non-structural proteins, Figure 2 (17,44). SARS-CoV-2 Mpro is a cysteine protease which functions as a homodimer, mainly comprises three domains (I-III) and characterized by a non-canonical histidine-cysteine (Cys-His) dyad, Figure 3 (17). Homo-dimerisation has been shown to play an integral role in the catalytic activity of Mpro (45).

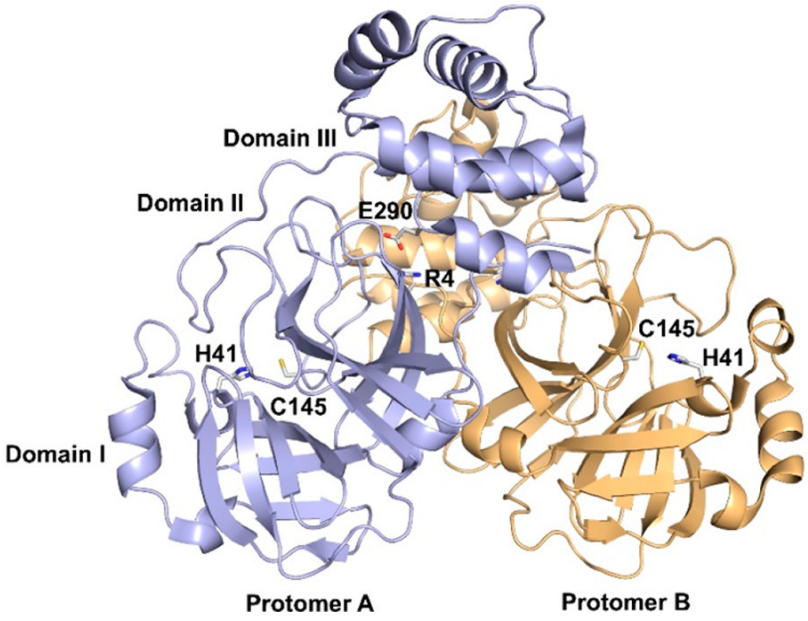


Figure 3: Structure of SARS-CoV-2 Mpro. The X-ray structure (PDB 6Y2G) is shown as a ribbon model with the bound inhibitor removed. Protomers A (light-blue) and B (light-orange) associate into a dimer stabilized by residues Glu290 and Arg4 that forms a salt bridge, while the substrate binding site resides at the interface of domains I and II. The catalytic residues Cys145 and His41 are highlighted.

Since this key enzyme is functionally relevant for viral replication, inhibiting it would be integral as viral load may be reduced and thus alleviate symptom intensity. Studies have shown that approaches similar to this were adopted in managing viral infections such as HIV (46–48). Attempts were made to try to use HIV protease inhibitors against SARS-CoV-2 but they proved to be ineffective as SARS-CoV-2 Mpro and HIV Mpro differ markedly (49–51). HIV protease is an aspartic protease and the active site comprises one residue from each monomer; thus it

functions only as a dimer, whereas Mpro is a 3CL cysteine protease that is likewise most active in the dimeric state, although each monomer has its own catalytic dyad (52). Since 3CL cysteine proteases are characterized by a chymotrypsin-like fold and a cysteine-histidine catalytic dyad in the active site, SARS-CoV-2 Mpro is different from HIV Mpro in terms of structure and chemical mechanisms. While the general strategy of seeking protease inhibitors is hence viable for both SARS-CoV-2 and HIV, drug development for the former depends on characterizing this novel enzyme by elucidating adaptive evolutionary mechanisms that may be caused by mutations accumulated over time (53).

1.1.4. SARS-CoV-2 Viral Evolution caused by Selective Pressure

In viruses, mutations occur constantly, and certain advantageous variations can be selected for over time while deleterious variations fade away – through natural selection (54). Although some mutations change native protein function that may assist in the replication of the virus, some may have potential to cause disease while others may prevent the process of reproduction and other mutations are neutral (55). The pervasive or fixed mutations that are capable of making the virus more or less virulent or transmissible are assumed to be under selective pressure and can be used to trace the spread of the virus around the world. Therefore, only some mutations give the virus some advantage and others may have the potential to decrease the efficacy of vaccines through altering the ability of antibodies and T cells to detect pathogens (24). In the absence of selective pressure, viruses can remain stable in their host. However, when selective pressure is exerted on the viral populations, they can evolve rapidly. Thus, the key to understanding impact of mutations in a rapidly evolving virus, like SARS-CoV-2, is by elucidating the selective pressure of these mutations over time, as this helps in signifying if adaptive viral evolution is taking place (56).

The SARS-CoV-2 genome is RNA-based, and viruses from this category have been reported to have increased rates of mutation (54). For instance, in HIV this has led to several

levels of classification of the virus, in which certain strains can manifest different transmissibility patterns and show differing responses to existing therapies (57,58); the same has been observed with SARS-CoV-2, Figure 4. In SARS-CoV-2, as a new RNA virus affecting humans, a recent host shift likely decreases its fitness and impels the virus to adapt to the new host environments and public health interventions (39,59). Natural selection may act on the transmissibility and virulence of this virus through adaptive dynamics of specific genomic mutations, which has been observed in Ebola, Zika and other viruses (39,59).

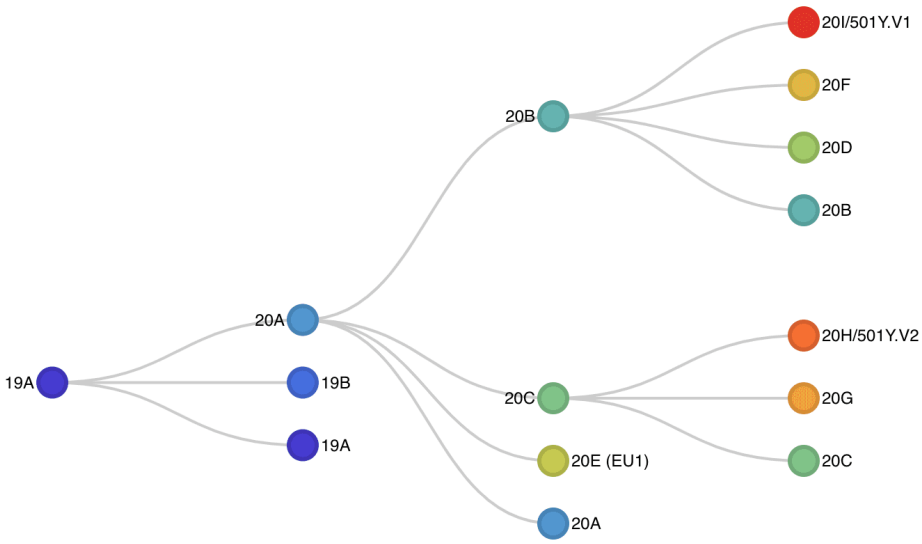


Figure 4: Schematic diagram showing hierarchical relationships among SARS-CoV-2 clades. This diagram illustrates how different SARS-CoV-2 clades evolved from the first clade since onset of the pandemic to February 9th 2021.

From the data gathered from the GISAID database (11) and real-time sub-sample estimates of genetic relatedness from the Nextstrain web resource (60), it can be observed that the virus is evolving within the human host. SARS-CoV-2 has circulated globally since its first outbreak in late December, 2019, and accumulated many genetic polymorphisms within the short period, giving rise to different clades (Figure 4). As of February 9th, 2021, new strains of SARS-CoV-2 were reported from England, Brazil and South Africa, and continue to spread across the world at an alarming rate (61). These new strains have been characterized with high transmission as they are more contagious which implies more hospitalizations and

consequently more deaths. One important aspect is that despite introduction of SARS-CoV-2 vaccines, mutations in these new strains may lessen the impact of the vaccine, thus may require the vaccine to be modified for it to be more effective. Therefore, designing robust direct-acting antivirals targeting essential viral enzymes such as the Mpro is vital for treating SARS-CoV-2 virus as this may be an essential backstop to manage illness resulting from vaccine ineffectiveness and uptake in human hosts. Thus, identifying pervasive mutations that indicate adaptive selection in these enzymes is vital for designing robust inhibitors. It is well known that pervasive mutations are indicative of selective pressures that lead to rapid adaptation causing viral resistance. In this study, we focused on identifying these fixed mutations, predicting whether they are positively or negatively selected and further inferring their functional impacts.

1.1.5. Techniques used for Screening Selective Pressure

Many computational methods have been developed for identifying if variants are under selective pressure and further predicting if the mutation is beneficial or deleterious. Some methods focus on non-coding regions (62–64) and others focus on coding regions to predict the effects of single amino acid variants – non-synonymous single-nucleotide polymorphisms (nsSNPs), or single amino acid substitutions (SAAS) on aspects such as protein function (65–70), structure (71), stability (72–74) and binding affinity (75). Our study focuses on nsSNPs and thus in this subsection, we describe *in silico* methods used to screen if they are under selective pressure and further predict their functional effect.

A common approach to determining the selection pressures that have shaped genetic variation involves estimating the rates of nonsynonymous (dN) and synonymous (dS) substitutions where if dN is significantly different from dS, the assumption is there is evidence for a non-neutral evolution. This method is preferable and has been widely used as it does not

make assumptions regarding demographic history of the population, unlike methods that test for neutral evolution (68,76,77), which compares estimates of effective population size obtained using different measures of genetic variation. Early studies relied upon the average dN/dS ratio for an amino acid site, either using distance-based methods (78,79) or maximum likelihood methods (80,81). Later, due to the inability of the methods to have statistical power to detect positive selection, new methods were devised to fix this. These methods were: the counting methods – that count the number of nonsynonymous and synonymous substitutions along the phylogeny; the random effects models – that assume a distribution of rates across sites and infer the rate at which individual sites evolve given this distribution; and the fixed effects models – that estimate the ratio of nonsynonymous to synonymous substitutions on a site-by-site basis (68).

As Pond and Frost noted, (81) "counting methods are attractive as they are computationally fast and hence can be applied to large data sets and do not involve making any assumptions regarding the distribution of rates across sites. However, they may lack power, especially for data sets comprising a small number of sequences or low divergence, as the power of the test is limited by the total number of inferred substitutions at a site. In addition, counting the number of changes between ancestral states may underestimate the true number of substitutions, and hence, the number of changes inferred using this approach may not accurately reflect the rate at which a site is evolving."

Most people have used a method of the random effects model called the empirical Bayes (82) which is based on the maximum likelihood estimates of the rate parameters to infer the site-by-site substitutions as it is computationally fast inexpensive. However, this method was observed to give misleading results in estimation of the rate distribution for small datasets – as they gave large error (81). In addition, they make assumption regarding the distribution of

rates across sites and gives a precise result when the assumed and the true distribution of rates are similar, which is not easily achieved. Like counting methods, fixed effects models make no assumption regarding the distribution of rates across sites (83). Since they fit substitution rates on a site-by-site basis, these models have been reported to give more accurate (less biased) representation of substitution rates at a site than counting and random effects methods. However, fixed effects models are typically slower than counting methods and may be difficult to fit due to the large number of parameters (81).

In our study, we employed the fixed effect model as it incorporates a general model of codon substitution (81), which allows us to rule out spurious results based on biased nucleotide frequencies. This method identified amino acid sites undergoing either positive, negative or neutral selection and positively selected variants were mapped on protein structure to prepare for functional inference using homology modeling and molecular dynamics.

1.1.6. Protein Modeling and Dynamics for Functional Inference

1.1.6.1. Homology Modeling

The structure of proteins is the basis for understanding the molecular mechanism and interactions at the atomic level. There are multiple experimental methods to determine protein structures including X-ray crystallography, nuclear magnetic resonance (NMR) and recently revolutionized cryogenic electron microscopy (cryo-EM). Using these methods, over 160 thousand structures (according to Protein Data Bank - PDB) have been solved and enabled breakthroughs in research and education. However, determining structures of proteins experimentally could be challenging and time consuming.

Homology, or template-based, modeling, which is complementary to experimental methods listed above, is a powerful computational modeling approach with the goal to construct an atomic or near-atomic resolution model of a target protein. Based on the query

protein sequence as an input, the method builds protein model based on the related homologous protein that has experimental three-dimensional structure (84,85). The quality or accuracy of the structural model is highly dependent on the similarity between the template and target proteins, which can be evaluated based on the quality of the sequence alignment, and template structure (86). Errors usually increase with decreasing sequence identity between the template and target sequences. The regions where has no template reference, for instance loop regions, are generally less accurate compared to the rest of the model (87,88).

Nevertheless, molecular modeling provides valuable insights for studying the molecular properties of protein molecules and their interactions with binding partners (substrates, peptides, inhibitors and proteins). The findings or hypothesis derived from molecular models could be later verified through experimental studies. Molecular modeling in addition to experimental structural determinations has significantly minimized the "structure knowledge gap" between the number of protein sequences and small number of known structures (84) and this has helped infer effects of non-synonymous mutations on protein function.

1.1.6.2. Molecular Dynamics Simulations

Proteins undergo conformational changes to perform their biological function. Hence, understanding protein dynamics is critical for understanding function. Molecular dynamics (MD) simulation is a computational method that enables studying protein dynamics by following their conformational changes through a period of time. Since the simulation is done at the atomic level, they usually are computationally slow and expensive. Nowadays, the speed by which MD simulations can be performed has been greatly increased thanks to the availability of supercomputing clusters and increased parallelization of calculations using powerful graphics processing unit (GPU) technology. The typical analyses that are analyzed in these simulations include dynamics (RMSFs), intermolecular interactions (hydrogen bonds,

van der waals contacts), and electrostatics surface analysis among others. In our study, due to time constraints, we only evaluated RMSD and RMSF. A detailed description of how the simulation was done is explained in the Methods section. The other MD analyses are for future work.

1.1.7. Scope of the Thesis

Molecular modeling is an important tool for guiding inhibitor discovery, making it possible to evaluate large numbers of candidate drugs *in silico* to select experimental targets; however, standard approaches screen against only one version of the protein, typically the reference or wild-type (WT) sequence (53). In a viral population, mutations accumulate rapidly, generating a *mutational landscape*. Cross and his colleagues wrote that "the design of robust inhibitors that can protect against the multiple strains encountered in clinical settings requires characterization of this sequence space and the populations of conformations it engenders. Furthermore, effective and rapid response to future emerging coronavirus diseases requires both *in silico* screening and experimental testing of antiviral agents and a validated library of relatively general inhibitors that can be used as a basis for the development of specialized therapeutics. Central to the success of that effort will be developing an understanding of structural and functional variation in SARS-CoV-2 proteins, particularly as mutations accumulate and new strains emerge" (53).

As of May 2021, limited literature exists with regard to the spatial and temporal mutational behavior of clinically relevant mutations in Mpro and their potential impact on the functionality of the protease. One study showed a trend toward substitution for more hydrophobic residues versus the WT protein and suggested differences in active site flexibility and cohesion (89). However, the investigations were carried out early in the pandemic and are therefore based on small sample sizes and limited geographic distribution. In addition, no study

has been done to understand the selection of fixed mutations that have persisted throughout the pandemic and are common across the world as research studies suggest that positively selected fixed mutations signify adaptation of organisms to their niches. It is thus important for epidemic trend prediction and disease control to understand whether natural selection is actively driving the adaptive evolution of SARS-CoV-2 Mpro during the pandemic. If the mutations are under positive selection, further research is warranted to identify the functional mutants contributing to the evolving epidemiological and pathogenic characteristics.

In this work we studied the dynamics of fixed Mpro mutations under positive selective pressure from global population isolates of SARS-CoV-2 as of February 6th 2021. Pervasive mutations were the focus of our study as literature suggests they are indicative of adaptive evolution of viruses to their niche (16,17,51–59). We then attempted to speculate the effects of these mutations on SARS-CoV-2 protease functionality.

1.2. Aims & Objectives

AIM 1: *To screen for selection in SARS-CoV-2 Mpro sites.* This was achieved by detecting spatiotemporal variations that showed persistence or pervasion according to geographic and real-time mutations in SARS-CoV-2 Mpro across the globe, since onset of the pandemic. Our focus was on positively selected SARS-CoV-2 Mpro sites.

AIM 2: *To elucidate preliminary structural analysis of positively selected mutations observed in AIM 1*. This was accomplished by mapping the mutations onto the structure and conducting a preliminary structural analysis to infer potential impact on the SARS-CoV-2 Mpro. Preliminary MD simulations of some of the variants was also done to support the structural analysis.

1.3. Hypotheses

Our hypothesis was that SARS-CoV-2 Mpro has fixed (pervasive) mutations that are undergoing selection signifying adaptive evolution of the virus to the human niche.

1.4. Rationale

Although SARS-CoV-2 Mpro is one of the best therapeutic targets for treating SARS-COV-2 infections, the protease has been suggested to be undergoing selective pressure that may possibly impact inhibitor design. This study uses *in silico* bioinformatics approaches to elucidate fixed or persistent mutations that are indicative of adaptive evolution of the virus to the human niche. Identifying amino acids that are undergoing adaptive evolution through *in silico* prediction methods is not only important to answering many questions that are critical for epidemic trend prediction and disease control, but can significantly reduce the labor, time, and cost of wet-lab experiments. In addition, this analysis will help in developing hypotheses with regards to identification of functional mutants potentially contributing to the evolving epidemiological and pathogenic characteristics, which can then be experimentally validated.

CHAPTER 2 – METHODS

2.1. Problem Formulation

SARS-CoV-2 Mpro has been observed to be prone to mutations suggestive of natural selection taking place (77,98). These mutations represent a reservoir of spontaneous and fixed mutations that cause both asymptomatic and symptomatic infections, and define whether the protease is becoming positively or negatively selected. Spontaneous mutations usually do not persist over time while fixed mutations usually persist over time and are indicative of either positive or negative selection. A persistent or fixed positively selected mutation signifies that the mutation has some functionally beneficial effects to the virus, and are indicative of viral adaptation to its niche (59,77,95). Negative selection on the other hand, signifies that the mutation has functionally detrimental effects to the virus; these mutations are usually removed. In this study we identified fixed mutations that were undergoing selective pressure in SARS-CoV-2 Mpro and tried to speculate on their effects on functionality of the protease. Functionally characterizing these mutations by in vitro studies is time-consuming and expensive. Thus, in silico approaches are adopted to solve this problem, by predicting the effect of amino acid (aa) variants on protein function. This section describes the methods used in this study to achieve the objectives.

2.2. Methods Pipeline

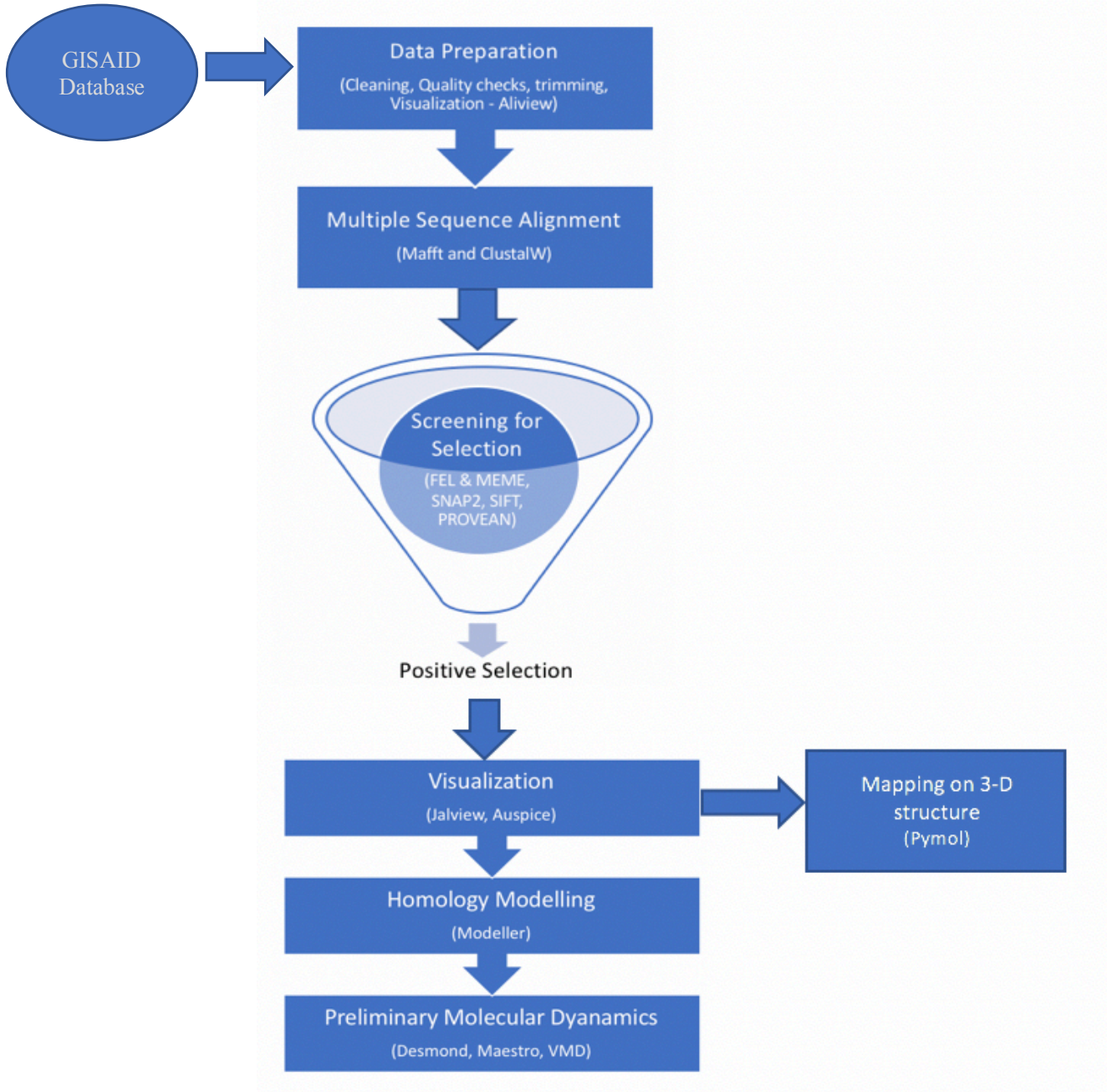


Figure 5: Methods pipeline employed in study

2.2.1. Data Collection and Preprocessing

We downloaded and studied 421,993 complete, high coverage sequences from the Global Initiative on Sharing All Influenza Data (GISAID, https://www.gisaid.org) – one of the repositories of globally distributed SARS-CoV-2 sequences (99). This database shares sequence data on potentially pandemic infectious viruses, as well as methods for sequencing and relevant geographic and clinical information.

Then SARS-CoV-2 Mpro reference gene sequence (from the WuhanSARS-CoV2 reference genome, accession number NC_045512.2) was queried against the downloaded genome sequences to identify the location of the gene. Then SARS-CoV-2 Mpro sequences were extracted 100 nucleotides upstream and downstream. Extended sequences were used to avoid misalignment at the end of the genes due some errors in the sequences. A python script was written for cleaning and checking the quality of the sequences. Sequences that had more than 5 Ns, and those that had 'x' instead of nucleotide bases were removed from the analysis. Inspection, sorting, deletion and merging of the cleaned sequences was done in AliView (100). This resulted in 419,120 sequences for downstream analysis.

2.2.3. Multiple sequence alignment & Visualization

Cleaned sequences were aligned using both *Mafft* and *ClustalW* for validity purposes. The upsteam and downsteam nucleotides of the gene were removed from the alignments; non-reference sequences appeared extended. All the gaps responsible for long range frameshifts were also removed. Then all perfectly aligned sequences were visualized using JalView (101). JalView tool helped provide visual analytics that helped judge the degree of similarity among the aligned sequences.

2.2.4. Sequence Analysis

This subsection describes how the 419,120 aligned sequences were analyzed to achieve the Aims & Objectives and to test the Hypotheses.

2.2.4.1. Mutation Analysis

We wrote a python script to identify mutations across SARS-CoV-2 Mpro sequences and calculated the mutation rate using the formula described by (102).

$$\textit{Mutation rate } (\mu) = [(r_2/N_2) - (r_1/N_1)] \times \ln (N_2/N_1)$$

$$= (f_2 - f_1) \times \ln (N_2/N_1)$$

where r_1 = observed # of mutants at time point 1;

r_2 = observed # of mutants at the next time point 2;

and N_1 and N_2 are the #s of viral sequences at time points 1 and 2

while f_1 and f_2 are the mutant frequencies at points 1 and 2.

Our script identified both synonymous and non-synonymous mutations. For the non-synonymous mutations, we further observed their distribution according to whether they are fixed or spontaneous throughout the global SARS-CoV-2 sequences by observing each of the mutants according to their geographical and spatial distribution using *Auspice* tool (103). We then used the fixed non-synonymous mutations for selection analyses.

2.2.4.2. Selection Analysis

As next-generation sequencing projects generate massive genome-wide sequence variation data, bioinformatics tools are being developed to provide computational predictions on the functional effects of sequence variations and narrow down the search of causal variants for disease phenotypes (69). There are several prediction tools that focus on studying the deleterious effects of single amino acid substitutions through examining amino acid

conservation at the position of interest among related sequences. In this study, we used an algorithm that blended two statistical approaches, that uses both Mixed Evolutionary Model of Evolution (MEME) (104) and Fixed Effects Likelihood (FEL) (68) to predict the selection pressure of the variants. The algorithm was run in the *Hypothesis Testing using Phylogenies* (HyPhy) software package (105).

MEME: MEME (Mixed Effects Model of Evolution) employs a mixed-effects maximum likelihood approach to test the hypothesis that individual sites have been subject to episodic positive or diversifying selection. In other words, MEME aims to detect sites evolving under positive selection under a proportion of branches (106).

"For each site, MEME infers two mutation (ω) rate classes and corresponding weights representing the probability that the site evolves under ω rate class, at a given branch. Importantly, to infer ω rates, MEME infers a single α (dS) value and two separate β (dN) values, β− and β+, which share the same α, per site. For both the null and alternative model, MEME enforces the constraint $\beta- \leq \alpha$. The β+ parameter is therefore the key difference between null and alternative models: In the null model, β+ is constrained as in the null model: $\beta+ \leq \alpha$, but β+ is not constrained in the alternative model. Ultimately, positive selection for each site is inferred when $\beta+ > \alpha$ and shown to be significant using the likelihood ratio test" (106).

FEL: FEL (Fixed Effects Likelihood) uses a maximum-likelihood (ML) approach to infer non-synonymous (dN) and synonymous (dS) substitution rates on a per-site basis for a given coding alignment and corresponding phylogeny. This method assumes that the selection pressure for each site is constant along the entire phylogeny (106).

After optimizing branch lengths and nucleotide substitution parameters, FEL fits a MG94xREV model (81) to each codon site to infer site-specific nonsynonymous and synonymous (dN and dS, respectively) substitution rates. Hypothesis testing is then conducted

on a site-specific basis, using the Likelihood Ratio Test, to ascertain if dN is significantly greater than dS.

We also used three other *in silico* computational tools: Screening for Non- Acceptable Polymorphisms (SNAP2) (66), PROVEAN (70), and The Sorting Intolerant from Tolerant (SIFT) (65,107) to validate our algorithm.

Provean: uses a versatile alignment-based score as a new metric to predict the damaging effects of variations not limited to single amino acid substitutions but also in-frame insertions, deletions, and multiple amino acid substitutions (69). This alignment-based score measures the change in sequence similarity of a query sequence to a protein sequence homolog before and after the introduction of an amino acid variation to the query sequence (108).

SNAP2: uses a feed-forward neural network-based method for the prediction of the functional effects of non-synonymous SNPs by incorporating evolutionary information (residue conservation within sequence families), predicted aspects of protein structure (secondary structure, solvent accessibility), and other relevant information (66).

SIFT: uses sequence homology to compute the likelihood that an amino acid substitution will have an adverse effect on protein function. The underlying assumption is that evolutionarily conserved regions tend to be less tolerant of mutations, and hence amino acid substitutions or insertions/deletions in these regions are more likely to affect function (107).

2.2.4.2.1. Novelty of the Blended Algorithm

The blended FEL-MEME algorithm identifies protein residue sites that have mutants that have persisted with time and checks their dN/dS ratio and uses statistical tests to test for significance of the predicted result. This is different from the other tools that only predicts selection based on a specific time point. For our analysis, we selected variants based on FEL-MEME blended algorithm and validated them using SNAP2, SIFT and PROVEAN tools. Since all the tools described above have different methods of screening for selection, the mutations that showed

same result on at least 3 of these tools were considered as the true result. Since our focus was on positively selected mutants, 11 mutants (variants) were identified and were used for further functional inferences.

2.2.5. Mapping Mutations on Structure

The 11 positively selected fixed variants were mapped on SARS-CoV-2 Mpro structure, PDB ID: 6LU7, using PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.). The 6LU7 structure was chosen because it is one of the SARS-CoV-2 reference structures. The mutations were also analyzed with regard to their location on the structure, inferred contribution to the surrounding residues and possibly to the protease functionality. We then analyzed whether they are located in an evolutionary conserved area and also their solvent accessibility using *Consurf* software (109) which identifies evolutionary constrained sites and also identifies whether the residues are buried inside or exposed to the solvent surrounding the protein.

2.2.6. Homology Modeling of Fixed Positively Selected mutations

For homology modeling, we used SARS-CoV-2 Mpro structure, PDB ID: 6LU7 (resolution 2.1 Å), as a template and developed a SARS-CoV-2 Mpro sequence containing the 11 positively selected fixed variants, as the query sequence. Modeling was done using MODELLER 10.1 (110) and visualization of our 6 models was done in PyMOL. A python script was written with regards to the procedures for modeling. The best model was chosen based on having the highest RMSD and the least B-factor (called 11-mutant-SARS-CoV-2 Mpro) and was prepared for molecular dynamics simulations.

2.2.7. Preliminary Molecular Dynamics Simulations

This computational work was performed using Schrödinger software (**Maestro** 11.4, **Schrodinger** 2020-4). Molecular Dynamics (MD) simulation in explicit solvent was run for the homodimeric form of the 11-mutant structure, in its free (unbound) state. The coordinates

of the solute atoms from the structure were processed with the Schrodinger's Desmond program in order to add the missing H atoms and to assign the molecular mechanics parameters.

2.2.7.1. Preparation of 11-Mutant SARS-CoV-2 Mpro Structure for MD

The SARS-CoV-2 virus protein structure was prepared in the Protein Preparation Wizard and Prime module of the Schrödinger suite to remove defects such as missing hydrogen atoms, inappropriate bond order assignments, charge states, alignments of several groups, and missing side chains. Steric hindrance and strained bonds/angles were removed through restrained energy minimization, permitting movement in heavy atoms up to 0.3 Å.

2.2.7.2. Molecular Dynamics (MD) Simulation Process

We performed a 100ns MD simulation for the 11-mutant SARS-CoV-2 Mpro structure Desmond software – a software that integrates temperature, pressure and volume system (111). In Schrodinger's Maestro program, there is a system builder called TIP3P water model which contains water molecule; we used this model to solvate the system spaced at 1Å in a cubic box filled with water. Then 0.15 NaCl was added to the system to imitate the concentration of ions physiologically and selected the steepest descent method for energy minimization. To describe the potential energy of the system, we used the OPLS_2005 force field (112). We chose this force field because it has been widely recommended for simulation of proteins. Using an orthorhombic box with buffer dimensions of 10 Å × 10 Å × 10 Å and temperature set at 300 K with standard pressure (1.01325 bar), we run the MD simulation for 100ns. We set the energy (kcal/mol) to be recorded at every 1.2ns interval.

After the system showed equilibration, we performed preliminary MD trajectory analysis using the trajectories generated during 100 ns MD simulation. We used the period of 100 ns which has been shown to be adequate time for the rearrangements of Cα atoms of SARS-CoV-2 Mpro. Trajectories were loaded and analyzed in VMD 1.9.4 software and our preliminary analysis was with regards to Root M

CHAPTER 3 – RESULTS

This chapter shows results from our data analysis with reference to the Aims & Objectives. We first outline the findings in three ways: exploratory analysis of the SARS-CoV-2 viral genomes, spatial-temporal mutational dynamics of SARS-CoV-2 Mpro, and structural analysis of relevant mutations that showed positive selection and preliminary molecular dynamics simulation analysis.

3.1. Exploratory Distribution of Human SARS-CoV-2 Viral Genome Sequences

A total of 421,993 human SARS-CoV-2 virus genome sequences were downloaded from the GISAID database as of February 9th, 2021, out of which 419,420 sequences remained for analysis. GISAID database receives sequences worldwide, and in this study, we analyzed 7,202 sequences from Africa, 54,054 from Asia, 213,904 from Europe, 121,600 from North America, 14,453 from Oceania and 8,206 from South America.

About 61% of the sequences were contributed by Europe alone followed by North America – which mostly comprise the united states. Africa had the least amount of sequences submitted to GISAID.

Not all the countries in each of the continents shown in Figure above submitted their sequences to GISAID database. Besides, some countries submitted more sequences than other countries, for instance, in North America, more than 80% of the sequences came from the United States and in Europe, and more than 60% of the sequences came from England. This disproportionate number of sequences required appropriate sampling; analyzing it without sampling is subject to bias,

and therefore all the analysis would be biased as well because the areas which had more samples would overshadow the ones with fewer samples and the results would not be a true representation of the whole population. We thus used subsampling approaches to remove potential sampling biases in order to ensure that regions and time-periods are appropriately included for visualization.

3.1.1. Spatiotemporal Distribution of SARS-CoV-2 Clades

We analyzed viral 4200 subsampled genomes using Nextstrain's Auspice tool (103) to visualize how SARS-CoV-2 virus has evolved and spread across the globe. As of February 9th, 2021, there were 12 SARS-CoV-2 clades using the nomenclature assigned by the GISAID. Since its emergence in late 2019, SARS-CoV-2 has diversified into several co-circulating variants which were grouped into clades due to their specific signature mutations (Figure 4). As time elapsed, some of the earlier clades are seen to be phasing away, for instance, as seen in Figure 6, proportion of circulating SARS-CoV-2 clade 19A and 19B is seen to be reducing while new clades are evolving. The same is seen in Figure 7 which shows a maximum

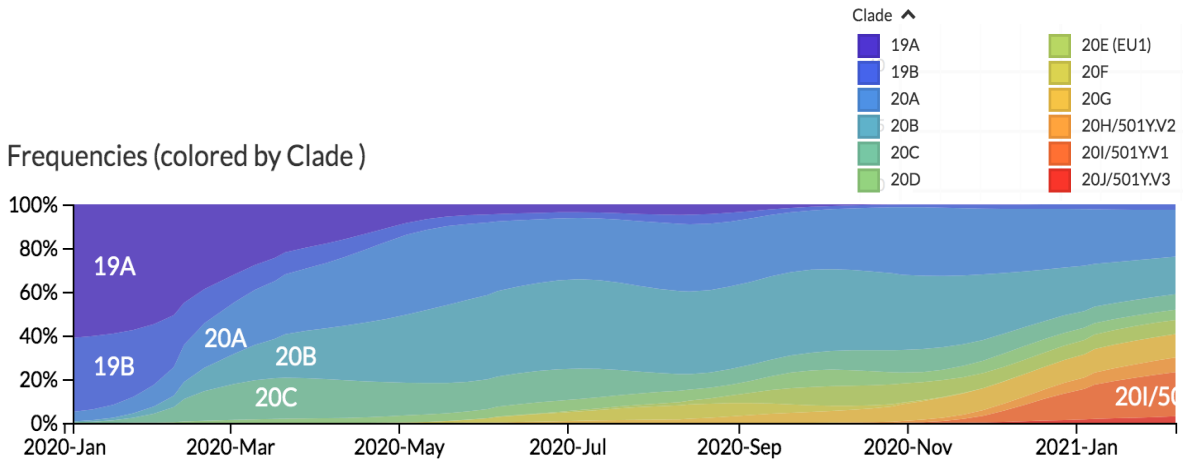


Figure 6: Temporal distribution of SARS-CoV-2 virus clades

likelihood phylogenetic tree of evolutionary relationships of SARS-CoV-2 since the beginning of the COVID-19 pandemic. Clades 20 H, I and J are all seen to have emerged between early December, 2020 and January 2021 while all other clades, except for the earliest ones (19A and 19B), are still seen in circulation in larger numbers.

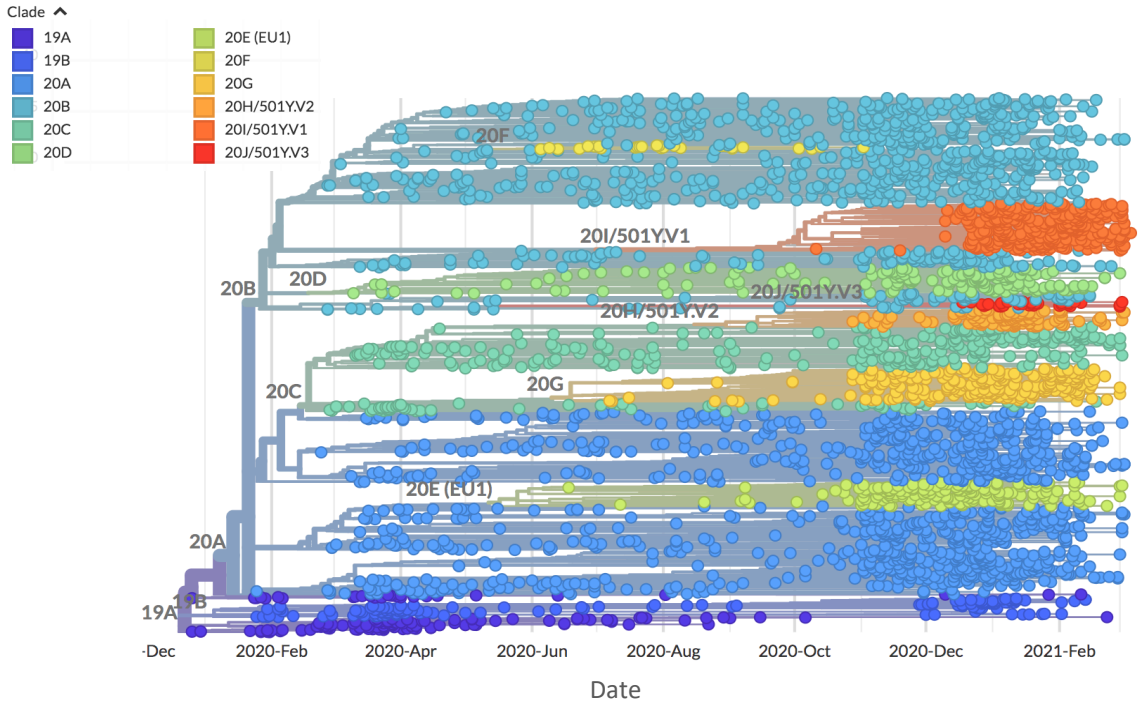


Figure 7: Phylogenetic temporal analysis of SARS-CoV-2 genomes according to clades. The length of the branches represents the distance in time. Each circle represents a sample and the color of the circle represents the clade that the sample belong to.

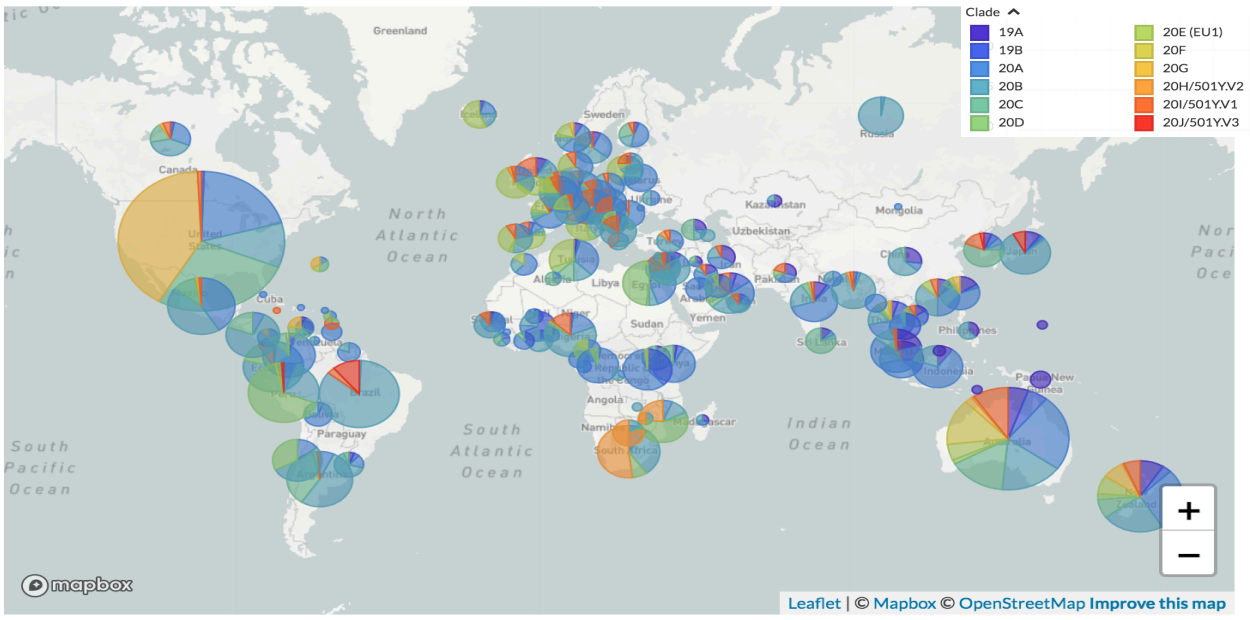


Figure 8: Geographic Distribution of SARS-CoV-2 clades – new clades seen in almost all continents. Each circle is centered on an individual country, the color indicates SARS-CoV-2 clade and the radius or size of the circle is proportional with the number of genomes submitted to GISAID from that country as of February 9th, 2021

From the map in Figure 8, overall, countries in Asia have a higher proportion of 19A and 19B clades which are the two earliest SARS-CoV-2 clades while Europe, North America and Oceania had a mixture of all clades and South America dominating with 20B and 20C. USA had a higher proportion of 19B, 20G and 20D clades, with 20A and 20J dominating in South America – Brazil. One of the three latest clades, 20H had a higher proportion in South Africa, with another latest clade, 20I, seen to be spread-out across Western and Southern Europe, and the last latest clade, 20J with seen predominantly in Brazil. This distribution is consistent with the fact that most recent SARS-CoV-2 cases, which had clinical manifestations that were different from the previous clades, were reported in Brazil, South Africa and some parts of Europe. By comparing the spatial and temporal distribution of the clades associated with each viral genome, we can characterize how COVID-19 is spreading across the world. For instance, the clades 19A, 19B, 20A, 20D and 20E seem to be found in almost all continents inferring to possible transmission from one country or continent to the other due to genetic drift and migration (39,77,95,96,113–115).

3.1.2. Temporal Mutational Dynamics of SARS-CoV2

Figure 9 shows an overview of the mutations accumulating in each of the 4,200 subsampled genomes as of February 9th 2021. We used one of the first Wuhan genomes (NCBI accession number NC_045512.2) as the reference to number the sites and genome structure in the Nextstrain's *auspice* tool; the phylogeny is rooted relative to early samples from Wuhan. Overall, the mutational frequency distribution shows that as time passes, there is an accumulation of mutations with genome samples in new clades having more mutations than earlier clades. However, it was interesting to find that most of these mutations were mostly from other parts of the SARS-CoV-2 viral genome and SARS-CoV-2 Mpro showed very

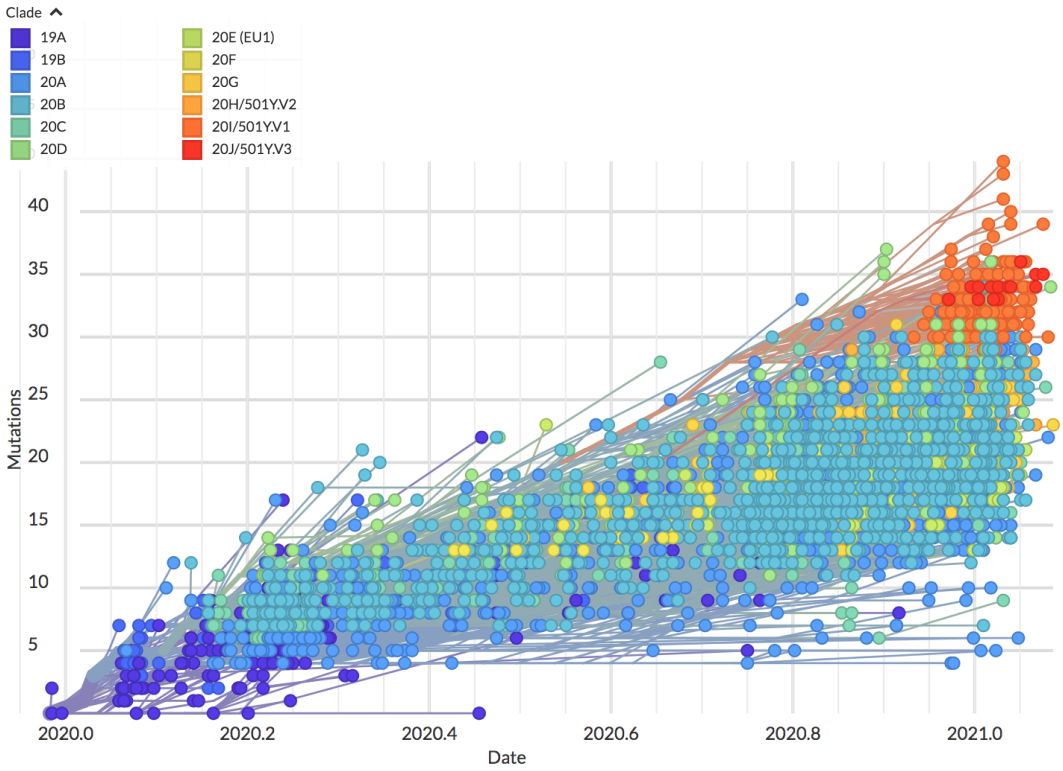


Figure 9: SARS-CoV-2 Mutational time-resolved phylogenetic tree. The length of the branches represents the distance in time. Each circle represents sample and the color of the circle represents the clade that the sample belong to.

little variation that could hardly be seen when plotted. This is in concurrence with our observation – also supported by literature (30,38,53)– that SARS-CoV-2 Mpro is highly conserved across the human SARS-CoV-2 viral genomes as it showed 93.8% identity during our analysis. Thus, to better understand the spatial and temporal distribution, we had to study specific mutations across the genomes from the onset of the pandemic till February 9th, 2021.

3.2. Proportion of SARS-CoV-2 Mpro Mutations across the Human Population

SARS-CoV-2 Mpro has 306 residues in total. Overall, we observed non-synonymous mutations at 169 residue sites of the protease with each enzyme having 1-6 substitutions per site. Figure 10 shows the structure of SARS-CoV-2 Mpro (PDB ID: 6LU7) displaying an overview of all mutations across the human SARS-CoV-2 genome sequences downloaded from the GISAID database mapped on the structure. As seen on Figure 10, it seems these mutations were spread out across the

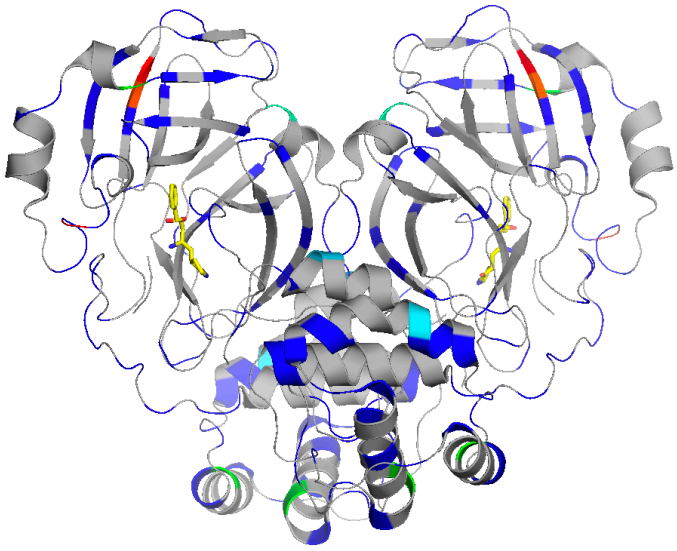


Figure 10: SARS-CoV-2 Mpro structure showing all observed non-synonymous mutations. All mutations are shown in blue, cyan and orange colors and Color intensity is proportional to the frequency where the brighter the color, the higher the frequency of the mutation.

structure. Using the formula explained in 2.2.4.1. Mutation **Analysis** subsection of the CHAPTER **2 – METHODS** chapter, overall SARS-CoV-2 Mpro mutation rate as of February 9th 2021 was calculated to be 0.31 substitutions per site.

3.3. Selection Analysis of SARS-CoV-2 Mpro sites

This subsection attempts to achieve the first aim of the study as stipulated in the Aims & Objectives. We screened all the observed mutations for selection using the techniques written in the 2.2.4.2. Selection **Analysis** subsection of the CHAPTER **2 – METHODS** chapter. Using our hybrid algorithm—that comprised FEL and MEME statistical methods to simultaneously screen for pervasiveness as well as selection of the variants, 16 SARS-CoV-2 Mpro residue sites were identified as undergoing either positive or negative selection, Figure 11. One interesting observation is that residues 144 and 145, which are found in the active site of SARS-CoV-2

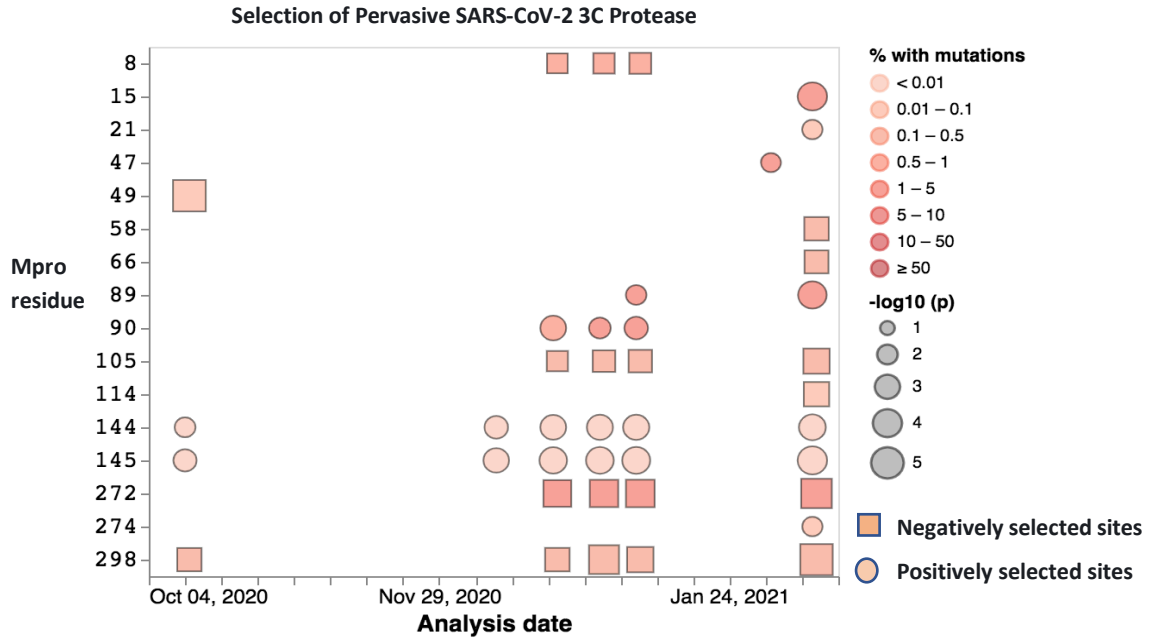


Figure 11: Selection Analysis of SARS-CoV-2 Mpro. Each site that reached significance at $p \leq 0.01$ at least once during our selection analyses is shown on the plot. Larger circles = smaller p-values. Lines indicate the range of detection (to make it easier to see the range of detection). The color of the circle indicates the frequency of non-reference alleles. Showed fixed mutations. Have significant FEL and MEME P-values

Mpro, were also observed to be undergoing selection. Interestingly, our hybrid algorithm identified residue 145 – which is the main catalytic residue, to be undergoing positive selection though with frequency <0.01 but with a pervasive trend across pandemic period. This is unusual since mutation of a catalytic residue implies a cease in functionality of the protease and thus assumes negative selection; however, the fact that this mutation is observed

to be persistent, with its dN/dS ratio significantly greater than 1, it is placed to be identified as undergoing positive selection. Nevertheless, this uncertainty requires further study. Since the focus of our study was mainly on positively selected SARS-CoV-2 Mpro mutations which showed persistence throughout the pandemic until February 9th, 2021, we selected them for further study.

3.3.1. Prediction of Effects of the Persistent Positively Selected Variants on SARS-CoV-2 Mpro Functionality using PROVEAN, SNAP2 and SIFT Tools

We compared the persistent positively selected variants that were detected by the hybrid algorithm with 3 in silico bioinformatics tools: PROVEAN, SNAP2 and SIFT to predict whether the substitutions are beneficial or harmful to the Mpro functionality (Table 1).

VARIANT	FEL/MEME	PROVEAN	SNAP2	SIFT	Final Prediction
G15S	H/B	H	H	H	**H**
T21I*	B	B	B	B	**B**
E47A	H/B	H	H	H	**H**
G71S*	B	B	H	B	**B**
L89F*	H/B	B	B	B	**B**
K90R	H/B	H	H	H	**H**
P96L/S*	B	B	H	B	**B**
S144E*	B	B	B	B	**B**
A191V*	H/B	H	H	B	**B**
A234V*	B	B	B	B	**B**
N274D/S	H/B	H	H	H	**H**

Table 1: Prediction of Variant Effect on SARS-CoV-2 Mpro Function

B represents to beneficial effect; H represents harmful effect;
*indicates the variant was selected for MD simulation

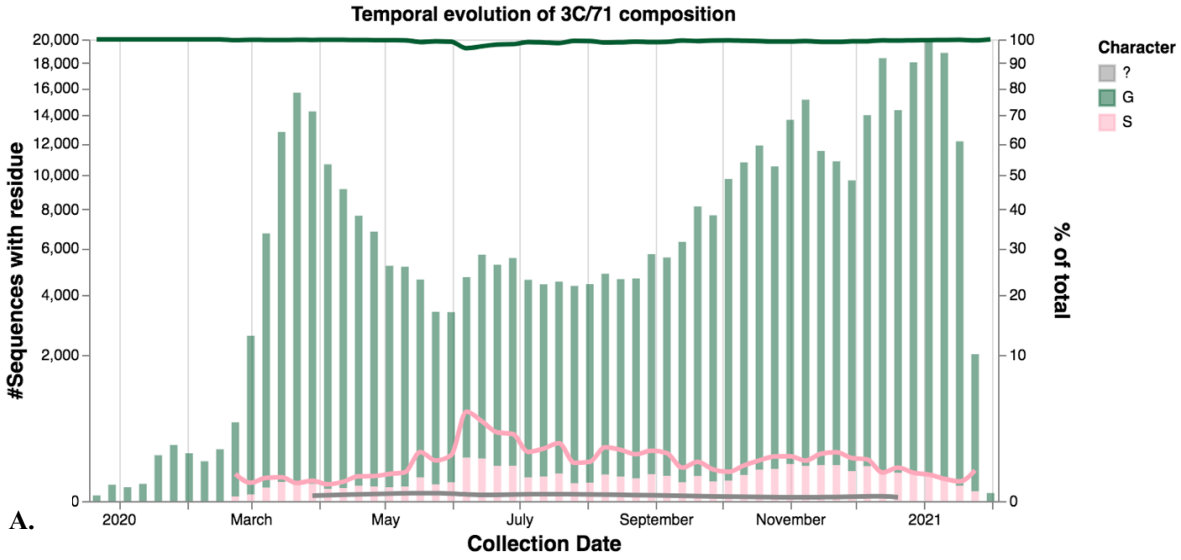
The difference in the predictions of the tools may be because the tools have different approaches of determining selective pressure of protein sites (see 2.2.4.2. Selection Analysis). For instance, FEL and MEME gave different results because FEL identifies episodic mutations while MEME identifies pervasive mutations. This is why we compared the results with the

other 3 approaches. We then selected 7 of the 11 variants for homology modelling and molecular dynamics simulation as they showed same prediction on all the tools.

3.3.2. Temporal and Geographical Distribution of Positively-Selected Mpro Variants

We explored further each of the 11 variants that showed positive selection on either FEL or MEME in our pipeline. These variants' trend varied according to the geographic locations and time-based trends with the significant variations being fixed throughout the sequences; the rest faded out and new substitutions occurred as the pandemic progressed. Here we show the temporal and geographical distribution of 3 variants that showed interesting trends with regards to the geographic and time-resolved distribution. The rest of the variants can be seen in *supplementary* table 1.

3.3.2.1. A closer look at the G71S Variant Observed across Sars-Cov2 population - Temporal and Geographical



A.

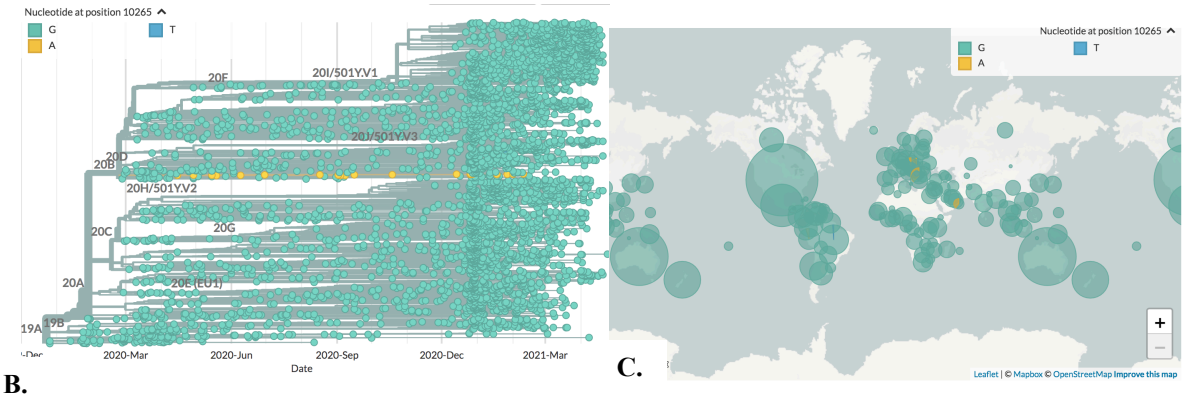


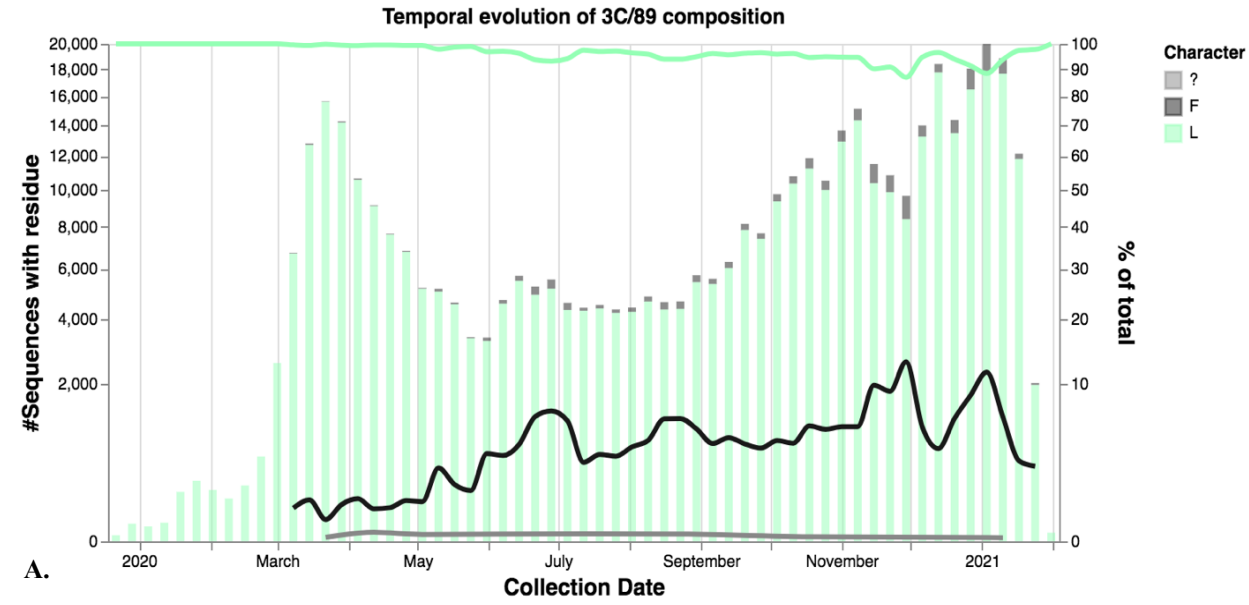
Figure 12: Spatial-temporal dynamics of G71S variant. A. Temporal trend of G71S variant from pandemic onset to February 9th 2021. Green lines represent number of sequences with residue G71 and pink lines represent number of sequences with mutant residue S71 (B.) time-resolved phylogenetic tree of nucleotide substitution with reference to clades. The colors represent nucleotide changes at position 10265. (C.) geographic location of the variant

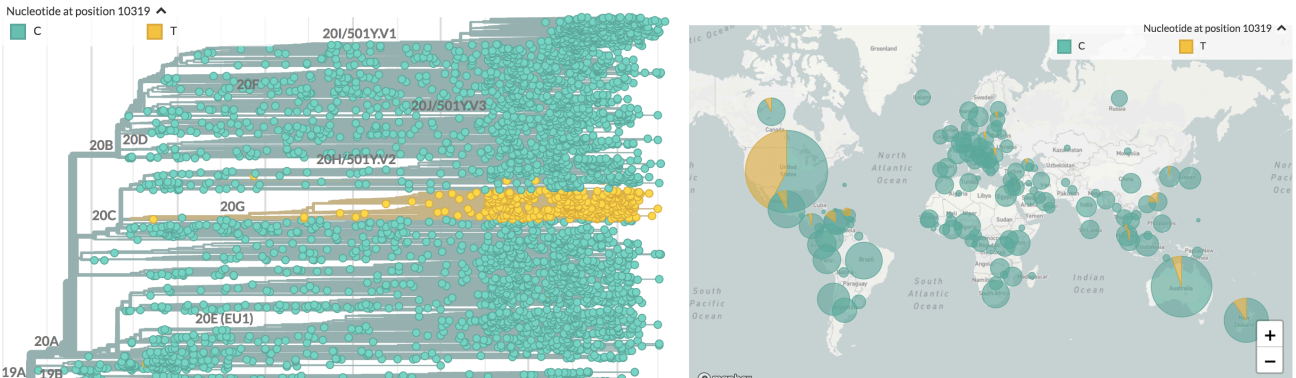
Fi

indication that the variant is undergoing selection. Since this variant was predicted as being positively selected, it suggests that this may be beneficial to the protease functionality.

3.3.2.2. A closer look at the L89F Variant Observed across hSars-Cov2 population - Temporal and Geographical

L89F is one of the variants that was detected to be undergoing positive selection as seen on the graph shown in Figure 13. This variant shows to have a persistent trend since around the





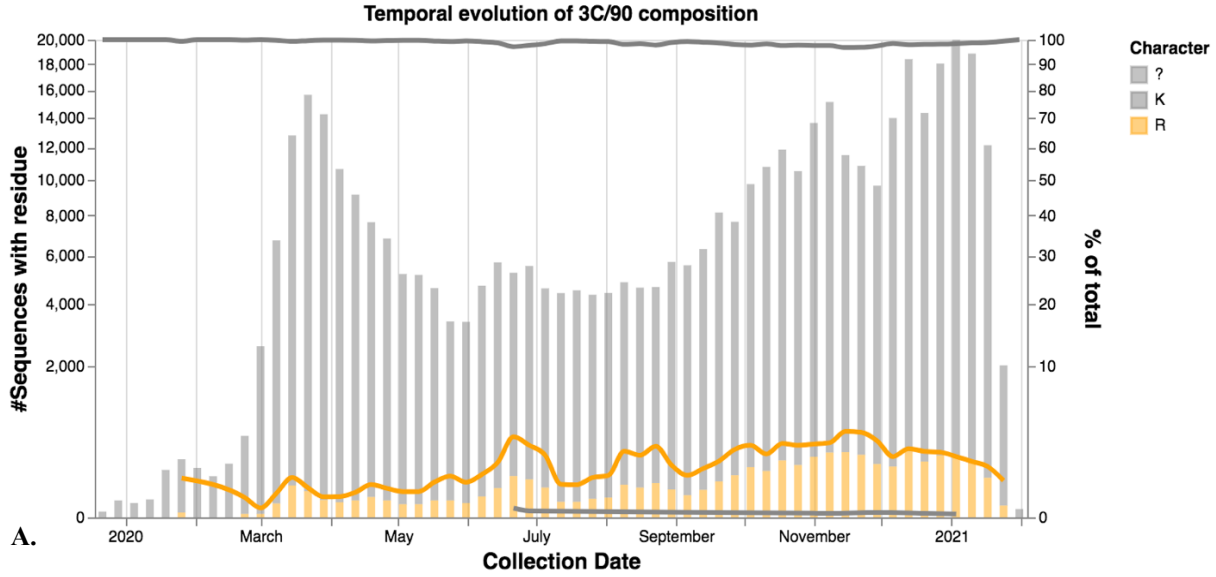
B. C.

*Figure 13: Spatial-temporal dynamics of L89F variant. (*A.) Temporal trend of L89F variant from pandemic onset to February 9th 2021. Light green lines represent number of sequences with residue L89 and grey represent number of sequences with residue F89 (B.) time-resolved phylogenetic tree of nucleotide substitution with reference to clades. The colors represent nucleotide changes at position 10319 (C.) geographic location of the variant

the beginning of March, 2020 till the time we did our sequence analysis, Figure 13A. This variant has shown pervasiveness and thus an indication of selection; and since it has been predicted as undergoing positive selection, it implies the mutation has beneficial effects to the viral protease. Another interesting observation is that this variant is seen predominantly in clade 20G than is seen in any other clade (Figure 13B) and is mostly distributed in North America and some parts of Australia and New Zealand (Figure 13C).

3.3.2.3. A closer look at the K90R Variant Observed across hSars-Cov2 population - Temporal and Geographical

Another interesting pervasive variant as seen in Figure 14 is K90R which is mostly seen in



B. C.

Figure 14: Spatial-temporal dynamics of K90R variant. (A.) Temporal trend of K90R variant from pandemic onset to February 9th 2021. Grey lines represent number of sequences with residue K90 and orange lines represent number of sequences with residue R90 (B.) time-resolved phylogenetic tree of nucleotide substitution with reference to clades. The colors represent nucleotide changes at position 10323 (C.) geographic location of the variant

clade 20H and predominantly found in southern part of Africa and spread in some parts of North and South America, Asia, Europe, Australia and New Zealand.

3.4. Structural Analysis of Persistent Positively-selected SARS-CoV2 Mpro Mutations

This subsection illuminates structural analysis of the positively selected SARS-CoV-2 Mpro variants in an attempt to achieve the second Aims & Objectives. We begin by mapping the persistent positively selected mutations on structure, do homology modelling and perform preliminary molecular dynamics simulations.

3.4.1. Mapping Persistent Positively Selected Mutations on Structure

Figure 15 shows an overview of the positively selected sites mapped onto the SARS-CoV-2 Mpro structure. Interestingly, we observed that about 63% (7 out of 11) of the mutations are

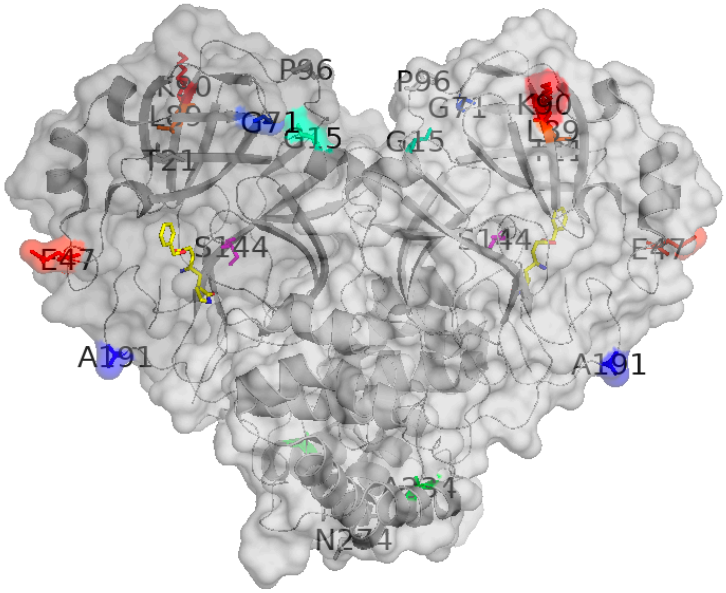


Figure 15: SARS-CoV-2 Mpro structure showing persistent positively selected mutations

located in domain I of the protease as seen in Figure 16A; one mutation – S144 is in Domain II and formed covalent bond with a catalytic Cysteine (C145), Figure 16B; one mutation – A191 is found in the loop region connecting Domain I and II and is one of the residues

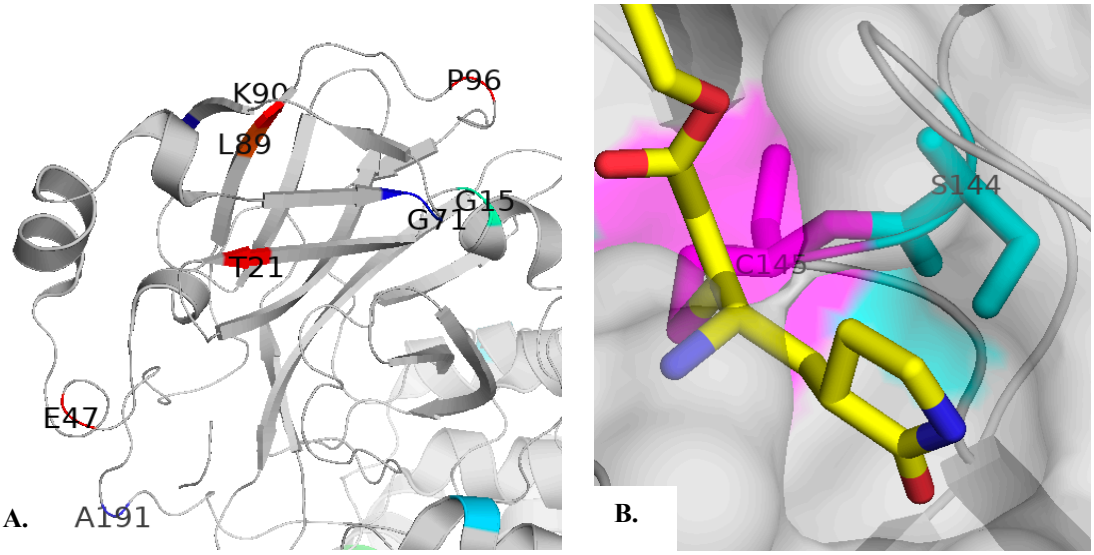


Figure 16: SARS-CoV-2 Mpro structure showing persistent positively selected mutations in Domains I and II. (A.) 7 residues under positive selection located in Domain I. (B.) Residue S144 covalently bonded to catalytic C145. The yellow compound is a ligand.

forming the substrate binding cleft creating the active site cavity (33,98,116)

Figure 17; and the other two mutations – A234 and N274 are located in the helicase domain (III) of Mpro structure. Another interesting observation is that residue E47, found in domain was seen to be covalently bonded to S46 which is also part of the substrate binding cleft just like A191 as seen in

Figure 17.

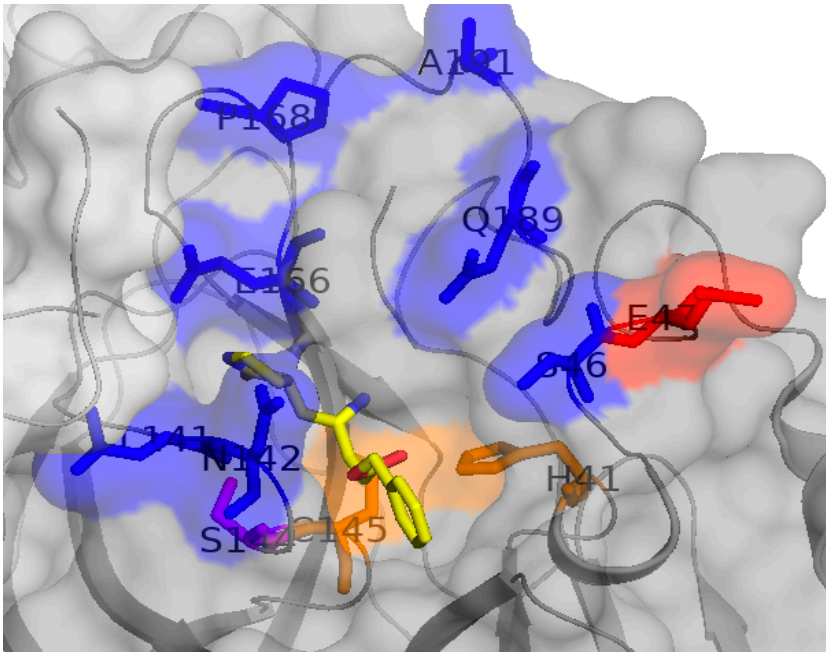


Figure 17: SARS-CoV-2 Mpro structure showing substrate binding cleft with residue A191 being part of it and residue E47 forming a covalent bond with residue S46. Residues in blue color are the ones forming the substrate binding cleft

3.4.2. Solvent Accessibility and Hydrophobicity change

According to their solvent accessible area, overall, 7 of the 11 persistent positively selected variants were exposed or located on the surface of the protease structure with 5 of them located in domain I of the protease and the remaining two located in the helical domain and loop region

Table 2:Solvent Accessibility of Persistent Positively selected SARS-CoV-2 Mpro variants

respectively. An interesting observation was seen on two domain I residues – T21 and L89, which were both buried and both show that they are mutating to being more hydrophobic. T21A variant shows that residue T21 is losing its solubility and becomes more hydrophobic as it mutates to residue Alanine. Since T21 is located on the beta strand of the domain and is surrounded by hydrophobic residues V20, F65 and L66 behind it (Figure 18), we may

Variant	**Solvent Accessibility (Consurf)**	**Hydrophobicity change (Kyte and Doolittle)**
G15S	Exposed	Both hydrophilic
T21I	Buried	Lose solubility becomes hydrophobic
E47A	Exposed	Changes from being hydrophilic to neutral
G71S	Exposed	Both hydrophilic
L89F	Buried	Both hydrophobic
K90R	Exposed	Both hydrophilic
P96L	Exposed	Lose solubility becomes hydrophobic
S144E	Buried	Both hydrophilic
A191V	Exposed	Changes from neutral to being hydrophobic
A234V	Buried	Changes from neutral to being hydrophobic
N274D	Exposed	Both hydrophilic

speculate that this mutation may be beneficial to the protease by providing more or better packing in the helix of the domain. The same was also observed in variant L89F where the

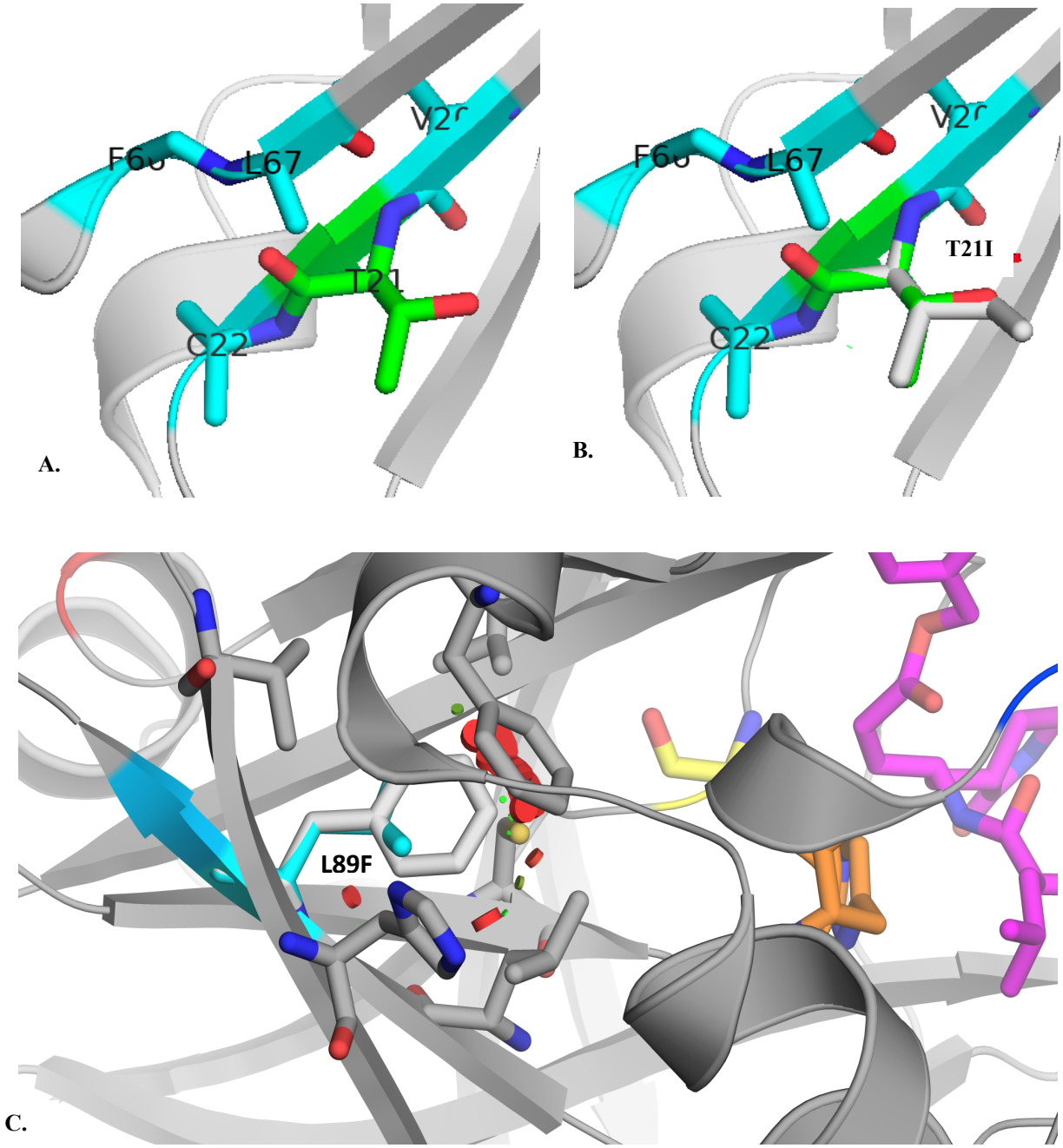


Figure 18: SARS-CoV-2 variants T21I and L89F.

mutation can be speculated to be of benefit for providing better packing with Phe66; Leu87, His80, Val20 pocket (hydrophobic). This residue is also observed to be located at the back of the catalytic active site. Another interesting observation is seen on variant A191V, a surface residue located on the loop region. Despite residue A191 being one of the residues forming

the active site cavity, it is observed bonding with some atom in another protein surrounding the protease, a property which is lost when it mutates to V191, Figure 19.

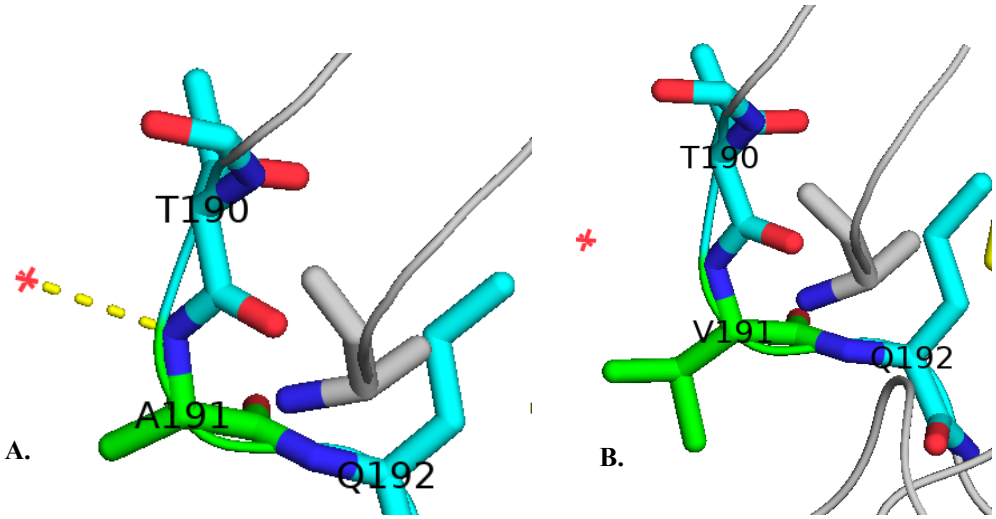


Figure 19: Structure of SARS-CoV-2 Mpro showing A191V. (A.) Nitrogen atom of A191 bonding with a surrounding atom. (B.) V191 showing no bond with any atom surrounding protease

3.4.3. Preliminary Molecular Dynamics simulations of positively selected variants

As a preliminary investigation of the molecular dynamics simulation of the 7mutant_Mpro structure created after homology modeling, we only analyzed two metrics: Root Mean Squared Deviation (RMSD) and Root Mean Squared Fluctuation (RMSF). We calculated the RMSD using VMD 1.9.4. software as seen in Figure 20. To sample the structural stability of the Mpro during the simulations, we measured the deviation of each structure from the starting coordinates of our modelled 7mutant_Mpro structure after a superposition on the protein Cα atoms. As the figure shows, the RMSD provides evidence that the simulated system reached convergence of the structural drift by sampling a local potential energy minimum.

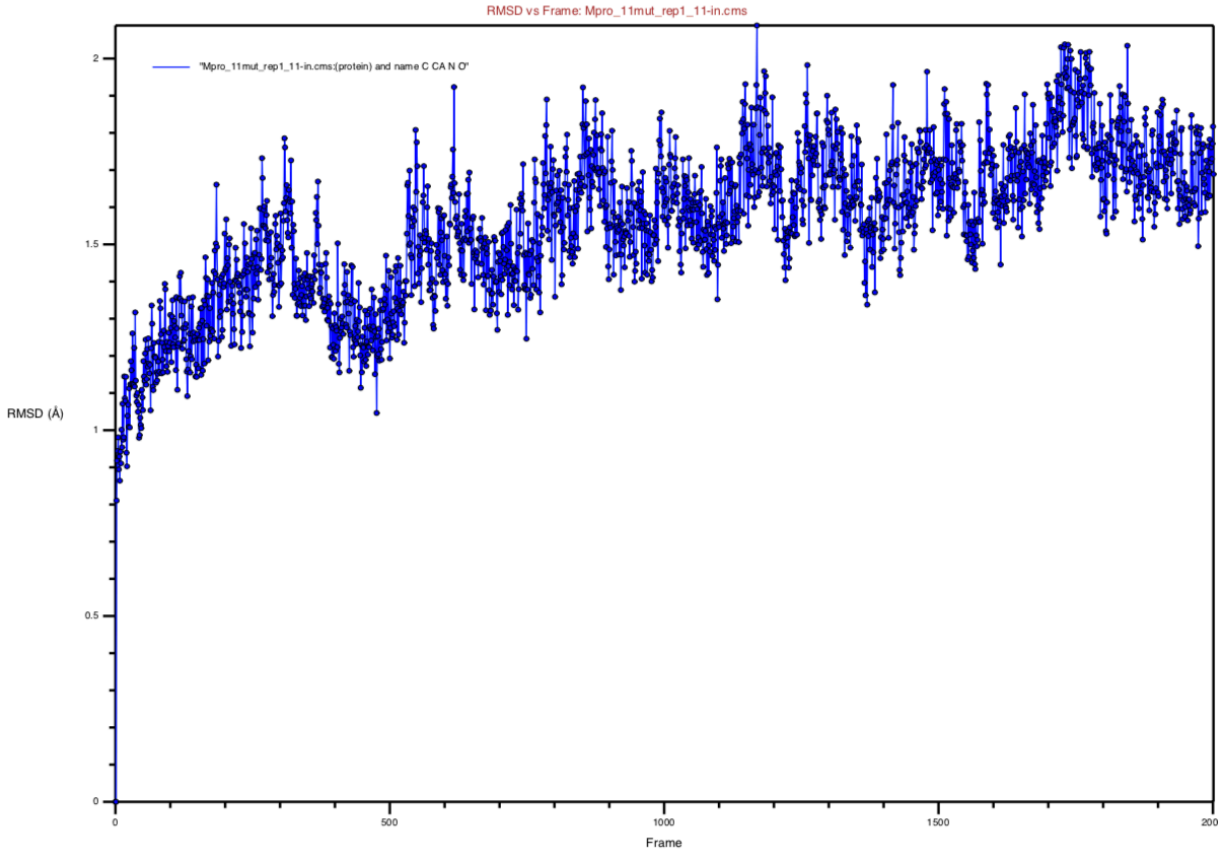


Figure 20: Root mean square deviation (RMSD) of the Cα in the 100 ns molecular dynamics (MD) simulation for the 7_mutant_SARS-CoV-2 Mpro dimer

We also calculated RMSF of both chain A and chain B in the trajectory to measure the flexibility of the residues, Figure 21. The results show that there are some specific regions

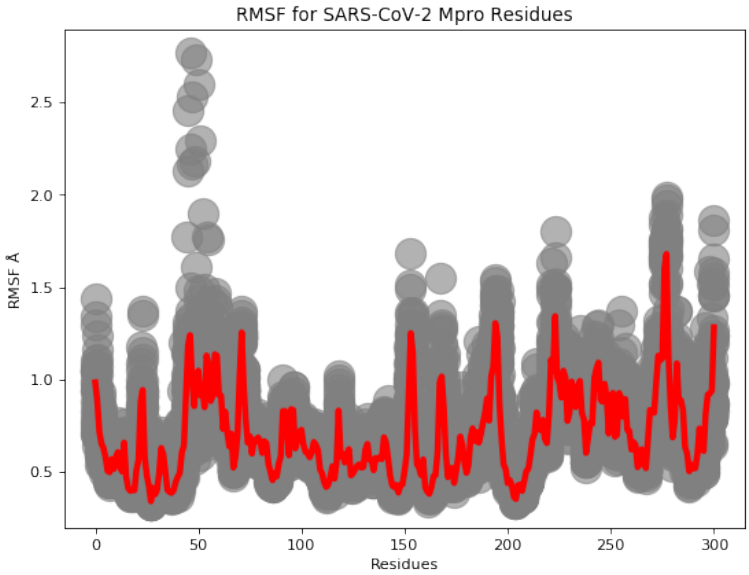


Figure 21: RMSF of 7_mutant_SARS-CoV-2 Mpro dimer. Higher RMSF values indicate greater flexibility during the MD simulation

that showed more flexibility than the other regions. As seen in the figure, domains I and III seems to have more flexibility of the residues than in domains II and the loop region. This makes sense since domain II is the catalytic domain and the loop region connects domains I and II and some of its residues form a catalytic cavity. In addition, residues 1-8 also shows low RMSF, and this can be because these residues are located between chains A and B of the Mpro.

CHAPTER 4 – DISCUSSION AND FUTURE WORK

Most of the spatiotemporal mutational studies conducted on SARS-CoV-2 Mpro, as of February 9th, 2021, focused on one-time snapshot of the observed mutation frequencies. Little has been done with regards to elucidating whether mutations were persistent throughout the pandemic or if they just surfaced at one point for a certain period and then disappeared. Literature suggests that persistent mutations signify natural selection which is an indicator of evolution (76,83,114,117–119). If these persistent mutations alter gene activity or protein function, they can introduce different traits in an organism; if the trait is advantageous and helps the virus survive and reproduce, then it means the virus is becoming evolutionary adaptive to its host. This becomes a challenge if these mutations result in certain viral epidemic, pathological or antimicrobial characteristics as this may have a negative impact on the control and treatment of the epidemic. With a focus on persistent positively selected mutations, we screened SARS-CoV-2 Mpro mutations and attempted to map the mutations on structure and infer the possible functions these mutations may have on protease functionality. This chapter discusses the findings of the analysis stipulated in the CHAPTER **3 – RESULTS** section.

4.1. Screening for Selection in Persistent SARS-CoV-2 Mpro Mutations

Generally, SARS-CoV-2 Mpro was observed to have mutations at 169 residue sites with each enzyme having 1-6 substitutions per site; these mutations comprised of both spontaneous and persistent ones. Since persistent mutations are an indication of selective pressure, we screened the mutations using a hybrid algorithm that comprised of both FEL and MEME to identify both pervasive and episodic selection residue sites. For validity purposes, we also used three other *in silico* bioinformatics tools that have been widely used to screen residues for selection. These three tools screen for episodic selection as they do not track the pattern of the mutation trend. In addition, *dN*/*dS* cannot distinguish between purifying

selection on synonymous codons and positive selection on amino acids (120), thus it would be integral to use bioinformatics tools that can determine positive selection on amino acid level.

In this regard, we found that most residue sites were not under selection with the exception of 16 residues sites. The presence of pervasive mutations in Mpro signifies that this protease may be under selective pressure and thus may need further analyses to elucidate the impacts such mutations will have on viral fitness. This concurs with one study (121) that screened for selection in all SARS-CoV-2 genes and found that the nsp5 gene – that codes for Mpro, was among the genes getting subjected to selective pressure. We observed that out of the 16 residue sites that were identified as to be under selective pressure, 8 were positively selected and interestingly showed a pattern of belonging to some specific clades. For instance, variants G71S, L89F and K90R were predominantly found in clades 20B, 20G and 20H respectively.

Another interesting observation was that since clades 20G and 20H were mostly found in the USA and South Africa as seen in Figure 8, it was also observed that the same countries were dominated with the L89F and the K90R mutations respectively, Figure 14. This is interesting since the variants that were used in naming these clades were mainly from the Spike protein and it is not yet known whether these variants are evolving into subclades of their ancestor clade. Furthermore, since clade 20H is one of the highly transmissible clades that caused high mortality in South Africa, it is not yet known whether variant K90R contributed to any epidemic or pathologic manifestation in this region as most studies focused on the Spike protein (34,122–124).

4.2. Structural Inference of Positively Selected SARS-CoV-2 Mpro Mutations

Literature suggest that the likelihood of mutations getting fixed and subsequently causing adaptive evolution in a genome depends on various factors such as the fitness of the

phenotype (the result of positive selective pressure) or the position of the residues in the three-dimensional (3D) structure (125). In this regard, after mapping the 11 persistent positively selected variants on 3D structure of SARS-CoV-2 Mpro, we found that these mutations were mostly located on the loops and turns of Domain I of the structure; a small number of mutations were seen in the catalytic as well as the helical domains, see Figure 15.

A similar observation was seen on all mutations that were identified in our analysis regardless of their selective pressure, where the regions that showed predominant mutations were the loops and turns of the 3D structure. This concurs with previous studies investigating protein tolerance studies that showed large mutability of amino acids in loops and turns compared to other regions such as α-helices and β-sheets (125,126,127). One of the functions of loops and turns is to provide structural stability of the protein, and thus, highly mutable persistent positively selected flexible loop mutations may signify a structural conformation that will favor fitness of the protease. Several authors have posed this association as being a result of solvent accessibility of the residues (125,126, 124) while other suggest this connection as an outcome of a stronger selective constraint of ordered proteins (127).

In this study, we looked at the exposure of the persistent positively selected residues on the surface of the SARS-CoV-2 Mpro 3D structure. We found that more than 50% (7 out of 11) of the positively selected residues were exposed to the surface out of which 4 did not change their solvent accessibility form of being hydrophilic and the remaining 3 lost their solubility. Interestingly, two of the residues that lost solubility are located near the active site. One residue, A191, which forms part of the active site cavity, is exposed to the surface of the protease and was observed to be interacting with other protein (Figure 19A). Mutating this residue to V191 resulted in no interaction forming between the residue with the other protein as the residue becomes more hydrophobic on the surface. Thus, we may speculate that this

mutation may affect binding affinity and substrate processing and specificity as well as structural flexibility of the protease. Covalently bound to residue S46 which also forms part of the active site cavity, E47A is another interesting variant that is exposed to the surface of the protease. In contrast to residue A191, residue E47 does not interact with any protein outside the structure; however, being bonded to residue S46 may infer that its mutation may have a certain impact on the binding affinity as well as structural flexibility of the protease. Nevertheless, more computational and experimental analyses may be of much benefit to support these claims. Biologically, Proline is one of the special residues important for structural integrity due to its kink shape. In SARS-CoV-2 Mpro, residue P96 forms part of a loop in domain I that is exposed to the surface (see supplemental Figures). Our analysis showed that this residue had its variant P96L which indicated that the mutant became more hydrophobic than residue P96 also suggesting some structural flexibility of the protease.

A further notable finding was that 3 out of 4 of the persistent positively selected variants – comprising T21I (Figure 18A and B), L89F (Figure 18C) and A234V (Supplementary Figures) that were identified as being buried inside the protease 3D structure all indicated the property of becoming more hydrophobic after mutating them. Residue A234 is located in domain III of the protease 3D structure and its change to V234 may imply more structural support to the dimeric state of the protease. Residue T21 is located on the β-sheets of domain I and if mutated to I21, it loses its solubility to becoming more hydrophobic. Likewise, variant L89F which is located on β-sheets, behind the active site maintains its hydrophobic state and can also be speculated to maintaining structural integrity by giving support to its surrounding hydrophobic residues.

Another exceptional discovery was that one of the catalytic residues C145 and its covalently bonded residue S144 were both observed be undergoing persistent positive selection. We found that Cysteine residue was mutating to Phenylalanine which implied a

change from being nucleophilic to becoming hydrophobic and in turn losing the binding affinity to substrate or ligand. Variant S144E implies no solubility change as both Serine and Glutamate are hydrophilic. This finding is not only surprising, as literature has found that mutating catalytic residue leads to denaturation of the protease, but also suspicious as to how such a mutation can be identified as being positively selected. On the other hand, despite these residues showing mutation frequency < 0.01, the fact that they showed persistence motivates a more detailed examination.

4.3. Limitations of Study and Future Work

One of the main limitations of this project was with regards to sampling. Since our dataset comprised of more than 50% data from Europe and North America, our analysis may not really be a representation of the global situation. However, during sampling, our analysis included all submitted sequences from the other regions that had less data and we only sampled from the regions that had more data. In addition, since we chose methods of selection that do not necessarily need metadata, we did not consider the environment in which viral population lives as this is integral to the selection of traits. Some differences introduced by variants may help viruses survive in one setting but not in another—for example, with regards to SARS-CoV-2, pathological characteristics are shown in some people and not in others. Thus, since we did not include metadata in our analysis, we may miss out on some important analyses that may give a greater insight into understanding evolution in Mpro. Further study on this may also be essential as it is evident that SARS-CoV-2 is symptomatic in other people and asymptomatic in others. Thus, we suggest that more in-depth evolutionary studies are needed to understand the genetic mechanisms that may affect the development of therapeutic and preventive tools, like antivirals and vaccines.

CHAPTER 5 – CONCLUSION

The computational study presented in this thesis reports the analysis of human SARS-CoV-2 Mpro genome sequences with regards to selective pressure of persistent mutations, mapping of positively selected mutations on structure (PDB ID: 6LU7) and performing preliminary molecular dynamics simulation to determine flexibility of the structurally modelled protease with reference to the positively selected fixed mutations. Our motivation stemmed from literature that suggest that fixed or persistent mutations signify that the residue site may be undergoing selective pressure which means the site may be under either positive or negative selection. In addition, literature from experimental studies suggests that positively selected residues are beneficial to the viral fitness and survival thus signify adaptive evolution of the virus. Our findings revealed persistent positive selection of residue sites– including those at and surrounding the active site. Overall, these mutations appear to result in systematic variations in both global structural flexibility and active site integrity and functionality. These signify that SARS-CoV-2 Mpro may currently be subjected to selection for enhanced global flexibility and that currently circulating Mpro variants represent a reservoir of phenotypic diversity in active site structure and dynamics that could facilitate an evolutionary response to certain classes of protease inhibitors.

REFERENCES

1. Moriyama M, Hugentobler WJ, Iwasaki A. Seasonality of Respiratory Viral Infections. Annual Review of Virology. 2020;7(1):83–101.

2. Gupta SP. Viral Proteases and Their Inhibitors. Academic Press; 2017. 518 p.

3. Lin M-H, Moses DC, Hsieh C-H, Cheng S-C, Chen Y-H, Sun C-Y, et al. Disulfiram can inhibit MERS and SARS coronavirus papain-like proteases via different modes. Antiviral Research. 2018 Feb 1;150:155–63.

4. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020 Mar;579(7798):265–9.

5. DiMaio D, Enquist LW, Dermody TS. A New Coronavirus Emerges, This Time Causing a Pandemic. Annual Review of Virology. 2020;7(1):iii–v.

6. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nature Medicine. 2020 Apr;26(4):450–2.

7. Hoffmann M, Kleine-Weber H, Krüger N, Müller M, Drosten C, Pöhlmann S. The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells. bioRxiv. 2020 Jan 31;2020.01.31.929042.

8. CDC. Cases, Data, and Surveillance [Internet]. Centers for Disease Control and Prevention. 2020 [cited 2021 Jan 16]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html

9. Rothan HA, Byrareddy SN. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. Journal of Autoimmunity. 2020 May 1;109:102433.

10. Lancker WV, Parolin Z. COVID-19, school closures, and child poverty: a social crisis in the making. The Lancet Public Health. 2020 May 1;5(5):e243–4.

11. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges. 2017;1(1):33–46.

12. Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP, Ostapchuck Y, et al. Virus Variation Resource – improved response to emergent viral outbreaks. Nucleic Acids Research. 2017 Jan 4;45(D1):D482–90.

13. Morselli Gysi D, Do Valle Í, Zitnik M, Ameli A, Gan X, Varol O, et al. Network Medicine Framework for Identifying Drug Repurposing Opportunities for COVID-19. arXiv e-prints. 2020 Apr 1;2004:arXiv:2004.07229.

14. Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discovery. 2020 Mar 16;6(1):1–18.

15. Cava C, Bertoli G, Castiglioni I. In Silico Discovery of Candidate Drugs against Covid-19. Viruses. 2020 Apr;12(4):404.

16. Joshi RS, Jagdale SS, Bansode SB, Shankar SS, Tellis MB, Pandya VK, et al. Discovery of potential multi-target-directed ligands by targeting host-specific SARS-CoV-2 structurally conserved main protease. Journal of Biomolecular Structure and Dynamics. 2020 Apr 24;0(0):1–16.

17. Dai W, Zhang B, Jiang X-M, Su H, Li J, Zhao Y, et al. Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease. Science. 2020 Jun 19;368(6497):1331–5.

18. Das S, Sarmah S, Lyndem S, Roy AS. An investigation into the identification of potential inhibitors of SARS-CoV-2 main protease using molecular docking study. Journal of Biomolecular Structure and Dynamics. 2020 May 2;0(0):1–11.

19. Jin Z, Du X, Xu Y, Deng Y, Liu M, Zhao Y, et al. Structure of M pro from SARS-CoV-2 and discovery of its inhibitors. Nature. 2020 Jun;582(7811):289–93.

20. Shah B, Modi P, Sagar SR. In silico studies on therapeutic agents for COVID-19: Drug repurposing approach. Life Sciences. 2020 Jul 1;252:117652.

21. Randhawa GS, Soltysiak MPM, Roz HE, Souza CPE de, Hill KA, Kari L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. PLOS ONE. 2020 Apr 24;15(4):e0232391.

22. Mousavizadeh L, Ghasemi S. Genotype and phenotype of COVID-19: Their roles in pathogenesis. Journal of Microbiology, Immunology and Infection [Internet]. 2020 Mar 31 [cited 2021 Jan 17]; Available from: http://www.sciencedirect.com/science/article/pii/S1684118220300827

23. Liang Y, Wang M-L, Chien C-S, Yarmishyn AA, Yang Y-P, Lai W-Y, et al. Highlight of Immune Pathogenic Response and Hematopathologic Effect in SARS-CoV, MERS-CoV, and SARS-Cov-2 Infection. Front Immunol [Internet]. 2020 [cited 2021 Jan 17];11. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2020.01022/full

24. Pardi N, Weissman D. Development of vaccines and antivirals for combating viral pandemics. Nature Biomedical Engineering. 2020 Dec;4(12):1128–33.

25. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell. 2020 Apr 16;181(2):271-280.e8.

26. Walls AC. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. :19.

27. Sawicki SG, Sawicki DL, Siddell SG. A contemporary view of coronavirus transcription. J Virol. 2007 Jan;81(1):20–9.

28. Hogue B, Machamer C. Coronavirus Structural Proteins and Virus Assembly. Nidoviruses. 2008 Jan 1;179–200.

29. Anand K, Palm GJ, Mesters JR, Siddell SG, Ziebuhr J, Hilgenfeld R. Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra α-helical domain. The EMBO Journal. 2002 Jul 1;21(13):3213–24.

30. Coronavirus Main Proteinase (3CLpro) Structure: Basis for Design of Anti-SARS Drugs | Science [Internet]. [cited 2021 Apr 27]. Available from: https://science.sciencemag.org/content/300/5626/1763.abstract?casa_token=JdEBdJEZ2jwAAAAA:mM9t0n8b14ZXCEt9pgMGN3z6N9nOIOCWTb8XbN31vwlknA1Pls0AiLUHNt9eBWCHBqCRem3227coAQ

31. Seth S, Batra J, Srinivasan S. COVID-19: Targeting Proteases in Viral Invasion and Host Immune Response. Front Mol Biosci [Internet]. 2020 [cited 2021 Apr 27];7. Available from: https://www.frontiersin.org/articles/10.3389/fmolb.2020.00215/full

32. Gioia M, Ciaccio C, Calligari P, De Simone G, Sbardella D, Tundo G, et al. Role of proteolytic enzymes in the COVID-19 infection and promising therapeutic approaches. Biochem Pharmacol. 2020 Dec;182:114225.

33. Anand K, Ziebuhr J, Wadhwani P, Mesters JR, Hilgenfeld R. Coronavirus Main Proteinase (3CLpro) Structure: Basis for Design of Anti-SARS Drugs. Science. 2003 Jun 13;300(5626):1763–7.

34. Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L, et al. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. Science. 2020 Apr 24;368(6489):409–12.

35. Yoshino R, Yasuo N, Sekijima M. Identification of key interactions between SARS-CoV-2 main protease and inhibitor drug candidates. Scientific Reports. 2020 Jul 27;10(1):12493.

36. Lokhande KB, Doiphode S, Vyas R, Swamy KV. Molecular docking and simulation studies on SARS-CoV-2 Mpro reveals Mitoxantrone, Leucovorin, Birinapant, and Dynasore as potent drugs against COVID-19. J Biomol Struct Dyn. :1–12.

37. Lin S, Shen R, He J, Li X, Guo X. Molecular Modeling Evaluation of the Binding Effect of Ritonavir, Lopinavir and Darunavir to Severe Acute Respiratory Syndrome Coronavirus 2 Proteases. bioRxiv. 2020 Feb 18;2020.01.31.929695.

38. Cannalire R, Cerchia C, Beccari AR, Di Leva FS, Summa V. Targeting SARS-CoV-2 Proteases and Polymerase for COVID-19 Treatment: State of the Art and Future Opportunities. J Med Chem [Internet]. 2020 Nov 13 [cited 2021 Apr 27]; Available from: https://doi.org/10.1021/acs.jmedchem.0c01140

39. Day T, Gandon S, Lion S, Otto SP. On the evolutionary epidemiology of SARS-CoV-2. Curr Biol. 2020 Aug 3;30(15):R849–57.

40. Petushkova AI, Zamyatnin AA. Papain-Like Proteases as Coronaviral Drug Targets: Current Inhibitors, Opportunities, and Limitations. Pharmaceuticals. 2020 Oct;13(10):277.

41. Chen YW, Yiu C-PB, Wong K-Y. Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CL pro) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. F1000Res [Internet]. 2020 Apr 9 [cited 2021 Jan 17];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7062204/

42. Xue X, Yu H, Yang H, Xue F, Wu Z, Shen W, et al. Structures of Two Coronavirus Main Proteases: Implications for Substrate Binding and Antiviral Drug Design. Journal of Virology. 2008 Mar 1;82(5):2515–27.

43. Yang H, Xie W, Xue X, Yang K, Ma J, Liang W, et al. Design of Wide-Spectrum Inhibitors Targeting Coronavirus Main Proteases. PLOS Biology. 2005 Sep 6;3(10):e324.

44. Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L, et al. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. Science. 2020 Apr 24;368(6489):409–12.

45. Chen S, Hu T, Zhang J, Chen J, Chen K, Ding J, et al. Mutation of Gly-11 on the dimer interface results in the complete crystallographic dimer dissociation of severe acute respiratory syndrome coronavirus 3C-like protease: crystal structure with molecular dynamics simulations. J Biol Chem. 2008 Jan 4;283(1):554–64.

46. Deeks SG, Smith M, Holodniy M, Kahn JO. HIV-1 Protease Inhibitors: A Review for Clinicians. JAMA. 1997 Jan 8;277(2):145–53.

47. Sham HL, Kempf DJ, Molla A, Marsh KC, Kumar GN, Chen C-M, et al. ABT-378, a Highly Potent Inhibitor of the Human Immunodeficiency Virus Protease. Antimicrobial Agents and Chemotherapy. 1998 Dec 1;42(12):3218–24.

48. Chandwani A, Shuter J. Lopinavir/ritonavir in the treatment of HIV-1 infection: a review. Ther Clin Risk Manag. 2008 Oct;4(5):1023–33.

49. Luan B, Huynh T, Cheng X, Lan G, Wang H-R. Targeting Proteases for Treating COVID-19. J Proteome Res. 2020 Nov 6;19(11):4316–26.

50. Chang K-O, Kim Y, Lovell S, Rathnayake AD, Groutas WC. Antiviral Drug Discovery: Norovirus Proteases and Development of Inhibitors. Viruses. 2019 Feb;11(2):197.

51. Dayer MR. Old Drugs for Newly Emerging Viral Disease, COVID-19: Bioinformatic Prospective. 2020 Mar 10 [cited 2021 Jan 2]; Available from: https://arxiv.org/abs/2003.04524v1

52. Chen H, Wei P, Huang C, Tan L, Liu Y, Lai L. Only One Protomer Is Active in the Dimer of SARS 3C-like Proteinase*,. Journal of Biological Chemistry. 2006 May 19;281(20):13894–8.

53. Cross TJ, Takahashi GR, Diessner EM, Crosby MG, Farahmand V, Zhuang S, et al. Sequence characterization and molecular modeling of clinically relevant variants of the SARS-CoV-2 main protease. bioRxiv. 2020 May 15;2020.05.15.097493.

54. Duffy S. Why are RNA virus mutation rates so damn high? PLOS Biology. 2018 Aug 13;16(8):e3000003.

55. Zuckerkandl E, Pauling L. Molecules as documents of evolutionary history. Journal of Theoretical Biology. 1965 Mar 1;8(2):357–66.

56. Olson-Manning CF, Wagner MR, Mitchell-Olds T. Adaptive evolution: evaluating empirical support for theoretical predictions. Nat Rev Genet. 2012 Dec;13(12):867–77.

57. Abecasis AB, Wensing AM, Paraskevis D, Vercauteren J, Theys K, Van de Vijver DA, et al. HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. Retrovirology. 2013 Jan 14;10(1):7.

58. Kosakovsky Pond SL, Smith DM. Are All Subtypes Created Equal? The Effectiveness of Antiretroviral Therapy against Non—Subtype B HIV-1. Clinical Infectious Diseases. 2009 May 1;48(9):1306–9.

59. Geoghegan JL, Holmes EC. The phylogenomics of evolving virus virulence. Nature Reviews Genetics. 2018 Dec;19(12):756–69.

60. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018 Dec 1;34(23):4121–3.

61. Zilber A. Fauci warns of “ominous” COVID-19 strains from Brazil, South Africa [Internet]. Mail Online. 2021 [cited 2021 Jan 19]. Available from: https://www.dailymail.co.uk/news/article-9157213/Dr-Fauci-warns-ominous-strains-COVID-19-Brazil-South-Africa.html

62. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012 Jun;6(2):80–92.

63. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010 Aug;7(8):575–6.

64. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010 Aug 15;26(16):2069–70.

65. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;31(13):3812–4.

66. Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Research. 2007 Jun 1;35(11):3823–35.

67. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. Curr Protoc Hum Genet. 2013 Jan;0 7:Unit7.20.

68. Kosakovsky Pond SL, Frost SDW. Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. Molecular Biology and Evolution. 2005 May 1;22(5):1208–22.

69. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLOS ONE. 2012 Oct 8;7(10):e46688.

70. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015 Aug 15;31(16):2745–7.

71. Schaefer C, Rost B. Predict impact of single amino acid change upon protein structure. BMC Genomics. 2012 Jun 18;13(4):S4.

72. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics. 2009 Oct 1;25(19):2537–43.

73. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W306-310.

74. Capriotti E, Fariselli P, Calabrese R, Casadio R. Predicting protein stability changes from sequences using support vector machines. Bioinformatics. 2005 Sep 1;21 Suppl 2:ii54-58.

75. Dehouck Y, Kwasigroch JM, Rooman M, Gilis D. BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations. Nucleic Acids Research. 2013 Jul 1;41(W1):W333–9.

76. Fay JC, Wu C-I. Hitchhiking Under Positive Darwinian Selection. Genetics. 2000 Jul 1;155(3):1405–13.

77. Liu Q, Zhao S, Hou Y, Zhao W, Bao Y, Xue Y, et al. Ongoing natural selection drives the evolution of SARS-CoV-2 genomes [Internet]. Epidemiology; 2020 Sep [cited 2021 Jan 4]. Available from: http://medrxiv.org/lookup/doi/10.1101/2020.09.07.20189860

78. Comeron JM. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J Mol Evol. 1995 Dec 1;41(6):1152–9.

79. Yang Z, Nielsen R. Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. Molecular Biology and Evolution. 2000 Jan 1;17(1):32–43.

80. Muse SV, Gaut BS. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol. 1994 Sep;11(5):715–24.

81. Pond SK, Muse SV. Site-to-Site Variation of Synonymous Substitution Rates. Molecular Biology and Evolution. 2005 Dec 1;22(12):2375–85.

82. Yang Z, Nielsen R, Goldman N, Pedersen AM. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 2000 May;155(1):431–49.

83. Suzuki R, Arita T. Interactions between learning and evolution:: The outstanding strategy generated by the Baldwin effect. Biosystems. 2004 Nov 1;77(1):57–71.

84. Schwede T. Protein modeling: what happened to the “protein structure gap”? Structure. 2013 Sep 3;21(9):1531–40.

85. Baker D, Sali A. Protein structure prediction and structural genomics. Science. 2001 Oct 5;294(5540):93–6.

86. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J. 1986 Apr;5(4):823–6.

87. Chung SY, Subbiah S. A structural explanation for the twilight zone of protein sequence homology. Structure. 1996 Oct 15;4(10):1123–7.

88. Khor BY, Tye GJ, Lim TS, Choong YS. General overview on structure prediction of twilight-zone proteins. Theor Biol Med Model [Internet]. 2015 Sep 4 [cited 2021 Apr 27];12. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4559291/

89. Amamuddy OS, Verkhivker GM, Bishop ÖT. Impact of emerging mutations on the dynamic properties the SARS-CoV-2 main protease: an in silico investigation. bioRxiv. 2020 May 29;2020.05.29.123190.

90. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infection, Genetics and Evolution. 2020 Sep 1;83:104351.

91. Boni MF, Lemey P, Jiang X, Lam TT-Y, Perry BW, Castoe TA, et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nature Microbiology. 2020 Nov;5(11):1408–17.

92. Liu S, Shen J, Fang S, Li K, Liu J, Yang L, et al. Genetic Spectrum and Distinct Evolution Patterns of SARS-CoV-2. Front Microbiol [Internet]. 2020 [cited 2021 Jan 4];11. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2020.593548/full

93. Nakagawa S, Miyazawa T. Genome evolution of SARS-CoV-2 and its virological characteristics. Inflammation and Regeneration. 2020 Aug 10;40(1):17.

94. Naqvi AAT, Fatima K, Mohammad T, Fatima U, Singh IK, Singh A, et al. Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease. 2020 Oct 1;1866(10):165878.

95. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. National Science Review. 2020 Jun 1;7(6):1012–23.

96. Li T, Liu D, Yang Y, Guo J, Feng Y, Zhang X, et al. Phylogenetic supertree reveals detailed evolution of SARS-CoV-2. Scientific Reports. 2020 Dec 22;10(1):22366.

97. Tracking evolution of SARS-CoV-2 virus mutations [Internet]. ScienceDaily. [cited 2021 Jan 4]. Available from: https://www.sciencedaily.com/releases/2020/10/201026114157.htm

98. Jin Z, Du X, Xu Y, Deng Y, Liu M, Zhao Y, et al. Structure of M pro from SARS-CoV-2 and discovery of its inhibitors. Nature. 2020 Apr 9;1–5.

99. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance. 2017 Mar 30;22(13):30494.

100. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 2014 Nov 15;30(22):3276–8.

101. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009 May 1;25(9):1189–91.

102. Pope CF, O'Sullivan DM, McHugh TD, Gillespie SH. A Practical Guide to Measuring Mutation Rates in Antibiotic Resistance. Antimicrobial Agents and Chemotherapy. 2008 Apr 1;52(4):1209–14.

103. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018 Dec 1;34(23):4121–3.

104. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK. Detecting Individual Sites Subject to Episodic Diversifying Selection. PLOS Genetics. 2012 Jul 12;8(7):e1002764.

105. Pond SLK, Frost SDW, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005 Mar 1;21(5):676–9.

106. VEG. HyPhy - Hypothesis Testing using Phylogenies [Internet]. [cited 2021 May 6]. Available from: http://veg.github.io/hyphy-site/methods/selection-methods/

107. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Research. 2012 Jul 1;40(W1):W452–7.

108. Choi Y. A fast computation of pairwise sequence alignment scores between a protein and a set of single-locus variants of another protein. In: Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine [Internet]. New York, NY, USA: Association for Computing Machinery; 2012 [cited 2021 Apr 25]. p. 414–7. (BCB '12). Available from: https://doi.org/10.1145/2382936.2382989

109. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res. 2016 Jul 8;44(Web Server issue):W344–50.

110. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Protein Sci. 2016 Nov 1;86:2.9.1-2.9.37.

111. Bowers K, Chow E, Xu H, Dror R, Eastwood M, Gregersen B, et al. Molecular dynamics---Scalable algorithms for molecular dynamics simulations on commodity clusters. Supercomputing, 2006. SC'06. Proceedings of the ACM/EEEI. 2006. 84 p.

112. Hajzer V, Fišera R, Latika A, Durmis J, Kollár J, Frecer V, et al. Stereoisomers of oseltamivir – synthesis, in silico prediction and biological evaluation. Org Biomol Chem. 2017 Feb 22;15(8):1828–41.

113. Lam TT-Y. Tracking the Genomic Footprints of SARS-CoV-2 Transmission. Trends in Genetics. 2020 Aug 1;36(8):544–6.

114. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infection, Genetics and Evolution. 2020 Sep 1;83:104351.

115. Mercatelli D, Giorgi FM. Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Front Microbiol [Internet]. 2020 [cited 2021 Jan 4];11. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2020.01800/full

116. Xue X, Yu H, Yang H, Xue F, Wu Z, Shen W, et al. Structures of Two Coronavirus Main Proteases: Implications for Substrate Binding and Antiviral Drug Design. Journal of Virology. 2008 Mar 1;82(5):2515–27.

117. Robinson JA, Brown C, Kim BY, Lohmueller KE, Wayne RK. Purging of Strongly Deleterious Mutations Explains Long-Term Persistence and Absence of Inbreeding Depression in Island Foxes. Current Biology. 2018 Nov;28(21):3487-3494.e4.

118. Kauzmann W. Some factors in the interpretation of protein denaturation. Adv Protein Chem. 1959;14:1–63.

119. user-guide.pdf [Internet]. [cited 2021 Apr 25]. Available from: https://dasher.wustl.edu/chem430/software/pymol/user-guide.pdf

120. Spielman SJ, Wilke CO. The Relationship between dN/dS and Scaled Selection Coefficients. Mol Biol Evol. 2015 Apr;32(4):1097–108.

121. Velazquez-Salinas L, Zarate S, Eberl S, Gladue DP, Novella I, Borca MV. Positive Selection of ORF1ab, ORF3a, and ORF8 Genes Drives the Early Evolutionary Trends of SARS-CoV-2 During the 2020 COVID-19 Pandemic. Front Microbiol [Internet]. 2020 [cited 2021 May 4];11. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2020.550674/full

122. Tu H, Avenarius MR, Kubatko L, Hunt M, Pan X, Ru P, et al. Distinct Patterns of Emergence of SARS-CoV-2 Spike Variants including N501Y in Clinical Samples in Columbus Ohio [Internet]. Genomics; 2021 Jan [cited 2021 May 3]. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.01.12.426407

123. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. medRxiv. 2020 Dec 22;2020.12.21.20248640.

124. Wibmer CK, Ayres F, Hermanus T, Madzivhandila M, Kgagudi P, Oosthuysen B, et al. SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. Nature Medicine. 2021 Apr;27(4):622–5.

125. Moutinho AF, Trancoso FF, Dutheil JY. The Impact of Protein Architecture on Adaptive Evolution. Molecular Biology and Evolution. 2019 Sep 1;36(9):2013–28.

126. Lio P, Goldman N, Thorne JL, Jones DT. PASSML: combining evolutionary inference and protein secondary structure prediction. Bioinformatics. 1998 Sep 1;14(8):726–33.

127. Guo HH, Choe J, Loeb LA. Protein tolerance to random amino acid change. PNAS. 2004 Jun 22;101(25):9205–10.

128. Choi I-G, Kim S-H. Evolution of protein structural classes and protein sequence families. PNAS. 2006 Sep 19;103(38):14056–61.