COVID-19 in India: An Interdisciplinary Study of Demographic, Structural, and Epidemiological Factors

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computational Natural Sciences by Research

by

Kushagra Agarwal
2018113012
kushagra.agarwal@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2023

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled ”COVID-19 in India: An Interdisciplinary Study of Demographic, Structural, and Epidemiological Factors” by Kushagra Agarwal, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Prof. Nita Parekh

To family, friends and
the Almighty

Acknowledgments

I would like to express my sincere gratitude to all those who have supported me throughout my Master's thesis journey. Firstly, I would like to extend my heartfelt thanks to my advisor, Prof. Nita Parekh, for her invaluable guidance, patience, and encouragement. Her continuous support and understanding have been instrumental in shaping my research work, and I could not have asked for a better mentor. She allowed me to go out and explore my research interests, which has enabled me to find my direction in life. Her guidance has been critical in shaping my career and academic aspirations. I am grateful for her constant support and willingness to go above and beyond to help me achieve my goals.

My family has been my support system and my motivation throughout my Master's thesis journey. I am deeply grateful to my father, who has been a constant source of inspiration and has always encouraged me to pursue my dreams. His constant support, guidance, and belief in me have been a driving force behind my success. My mother has been my pillar of strength, always ready to offer a listening ear, comfort, and support whenever I needed it. I am very grateful to my brother, who has been my role model, and to my grandparents who have been my constant source of love and blessings.

I would like to express my sincere thanks to my batchmate, Shreeya Pahune, who has gone out of her way in helping me complete my thesis work. Her academic insights, feedback, and encouragement were invaluable, and I am grateful for her friendship. I would also like to express my sincere gratitude to my friends, Shantanu, Jashn, Aryamaan, Akshit, Nikunj, Sajal, Manas, Anubhav, Mansi Ramuka, and Hemant Chandak who have been a source of constant support and encouragement throughout my time at IIIT. Their support has not only helped me to achieve my academic aspirations but also to grow as a person. I cherish the moments spent with them and will always be grateful for their contributions.

Finally, I would like to express my sincere gratitude to Prof. Chittaranjan Hens and Subrata Ghosh for their invaluable insights and guidance. Their willingness to sit with me and explain things in detail was of great help to me during my research.

I am truly grateful to all those who have contributed to my academic success and supported me throughout my Master's thesis journey. Their support has been invaluable, and I will always cherish their contributions.

Abstract

In this study, we carried out a comprehensive analysis of SARS-CoV-2 mutations and their spread in India over the past two years of the pandemic (27th Jan 2020 – 8th Mar 2022). The analysis covers four important timelines, viz., the early phase, followed by the first, second, and third waves of the pandemic in the country. Phylogenetic analysis of the isolates indicated multiple independent entries of coronavirus in the country, while principal component analysis identified few state-specific clusters. Genetic analysis of isolates during the first year revealed that though lockdown helped in controlling the spread of the virus, region-specific sets of shared mutations were developed during the early phase due to local community transmissions. We thus report the evolution of state-specific subclades, namely, I/GJ-20A (Gujarat), I/MH-2 (Maharashtra), I/Tel-A-20B, I/Tel-B-20B (Telangana), and I/AP-20A (Andhra Pradesh) that explain the demographic variation in the impact of COVID-19 across states. In the second year of the pandemic, India faced an aggressive second wave while the third wave was quite mild in terms of severity. Here we also discuss the prevalence and impact of different lineages and Variants of Concerns/Interests, viz., Delta, Kappa, Omicron, etc. observed during this period. From the genetic analysis of mutation spectra of Indian isolates, the insights gained into its transmission, geographic distribution, containment, and impact are discussed. Next, we evaluated the impact of some important India-specific mutations on protein function using structure and network-based analysis. Finally, we performed epidemiological modeling of the dominant variants in India during the Second (Delta and Kappa) and Third (Omicron) waves of the pandemic using multi-strain SEIR models to estimate the transmission factor (β) for the different variants. This interdisciplinary study provides valuable insights into the demographic, structural, and epidemiological factors influencing the spread of COVID-19 in India. The findings can inform and aid policymakers and public health officials in responding to future waves of COVID-19 or any other pandemic caused by a novel pathogen.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is a member of the betacoronavirus family that causes the coronavirus disease 2019 (COVID-19) pandemic. The first case of this novel coronavirus was reported in late 2019 in Wuhan, in the Hubei province of China, and has since spread rapidly through human-to-human transmission [83]. The outbreak was declared a pandemic by the World Health Organization on March 11, 2020. The pandemic has had a significant impact on global health and the economy, with over 300 million confirmed cases and more than 5 million deaths reported worldwide as of March 2023.

Coronaviruses are enveloped RNA viruses with a genome size ranging from 26.4–31.7 kilobases (Fig. 1.1) and a diameter of 60-140 nanometers [143]. Because of the huge size of their genome, they have more flexibility in accepting and changing genes, and the spike protein creates a crown-like structure that assists in detection by surface receptors, thus the term "coronavirus" [116]. The first coronavirus (ancestor) would have existed around 8100 BC as per studies involving molecular clock estimates [144]. Human coronaviruses were found in the 1960s, and seven have been identified since then: SARS-CoV, MERS-CoV, NL63, 229E, SARS-CoV-2, OC43, and HKU1 [29, 82, 72]. Over the past two decades, there have been three major coronavirus outbreaks, all of which have resulted in high morbidity rates [104, 113].

The first among them was due to SARS-CoV, which started in Guangdong (China) in 2002. It killed $\sim$800 people out of the $\sim$8000 people infected (fatality 10%) across five continents. Next, MERS-CoV, which originated in the Arabian Peninsula around 2012, killed $\sim$850 people out of the $\sim$2500 infected (fatality 34%) [63, 160]. While SARS-CoV-2, shares $\sim$79% nucleotide sequence similarity to SARS-CoV, and $\sim$ 50% similarity to MERS-CoV [83], its initial case fatality estimates were significantly lower (3.4% to 6.6%) [141]. Both the SARS-CoVs originated from bats, which served as their reservoir hosts [63]. The infection in the cases of SARS-CoV-2 is generally fatal in those over the age of 65 with

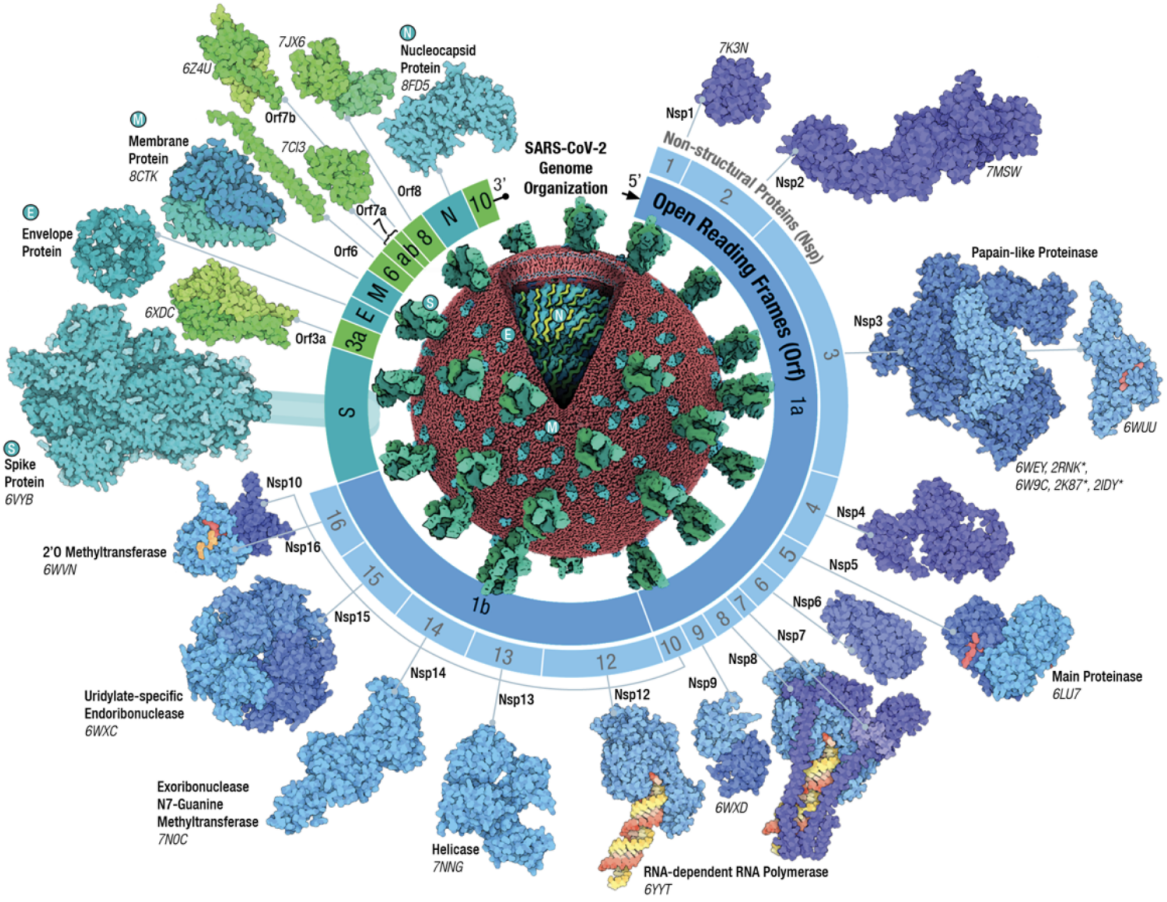


Figure 1.1: Structures of the different proteins in the SARS-CoV-2 virus. Adapted from the architecture of SARS-CoV-2 proteins given by Lubin et al. (2022) [85].

comorbidities such as heart disease, chronic lung disease, obesity, and diabetes [115, 151].

1.1.1 Demographic

The accumulation of mutations in the genome of a virus over time can serve as a molecular clock and provide valuable information about the virus's evolution and origin [112]. Mutations are a primary source of variation, which, along with natural selection, forms the foundation of evolution. Due to their high mutation rate, viruses can evolve rapidly, with the genome of an RNA virus changing up to one million times faster than that of its host [36, 38]. Viral replication machinery's tendency to make errors, exposure of the viral genome to mutagens, and recombination are some of the primary causes of mutation. During an outbreak, RNA viruses can accumulate genetic variations as they spread geographically [111]. Comparing genome sequences from different regions allows for the detection of genetic vari-

ations among viruses, aiding in understanding the disease's epidemiology, determining virulence and disease pathogenicity, and tracking the spread of SARS-CoV-2 between regions [95, 14]. This information can be useful in developing region-specific approaches to reduce the spread and severity of the disease.

Genomic surveillance is an essential tool for tracking the evolution of the virus and identifying novel mutations and variants. Whole-genome sequencing (WGS) can provide a high-resolution view of the genetic diversity of the virus and its transmission dynamics. WGS data can also be used to identify clusters of infections and track the spread of the virus within and between communities. The identification of new mutations and variants can also inform the development of diagnostic tests, treatments, and vaccines.

Several studies have demonstrated the importance of genomic surveillance in the context of the COVID-19 pandemic. An early study of 449 viral genomes, sampled from various countries found that the Indian samples could be categorized into two different clusters, with possible sources of infection from Oceania and Kuwait (cluster A), and countries from South Asia, Europe, and the Middle East (cluster B) [95]. Banu et al. (2020) described a distinct cluster of genomes (Clade I/A3i) which is primarily prevalent in Indian SARS-CoV-2 genomes, with negligible representation outside the country [15]. We too identified this cluster in our analysis (details in Section 2.3.2.3). A study on 210 samples from the southern state of Telangana in India, revealed the presence of unique and distinct mutations in ORF1a, and ORF3a [56]. These mutations were identified to be part of state-specific subclades in clade 20B through our analysis of the first wave of the pandemic in India (Section 3.5.3). Similar demographic studies have been conducted globally too. For example, the emergence of the highly transmissible B.1.1.7 variant in the UK in late 2020 was identified through genomic surveillance [5]. This variant quickly became dominant in the UK and spread globally owing to multiple mutations present in the Spike protein of the virus. Another variant, B.1.351, identified in South Africa, was also monitored due to its potential immune-escape ability [131]. Even the B.1.617.2 (Delta) variant, first identified in India, has several mutations in the spike protein which may impact immune escape. The emergence of these variants highlights the importance of continued genomic surveillance to monitor the evolution of the virus and identify potentially concerning variants.

The identification of novel mutations and variants has also raised concerns about the effectiveness of current diagnostic tests, treatments, and vaccines. For example, some mutations in the spike protein can result in false-negative results in diagnostic tests, while others can reduce the effectiveness of neutralizing antibodies induced by vaccines or previous infections [48]. Therefore, ongoing genomic surveillance efforts are crucial to identify mutations and assess their impact in order to help set up public health policies to reduce the spread of the pandemic.

1.1.2 Protein Structural and Network Analysis

Understanding the implications of mutations in protein structures is crucial for predicting their impact on protein function and disease outcomes. Structural information from molecular dynamics simulations, X-ray crystallography, etc. plays a significant role in providing insights into the alterations in protein structure induced by mutations. Network analysis of protein structures, on the other hand, helps in understanding the structural and functional consequences of these mutations by comparing the inter-residue interactions. Such analyses have been employed to understand the role of mutations in diseases such as cancer and infectious diseases. In particular, the ongoing COVID-19 pandemic caused by SARS-CoV-2 has highlighted the need for understanding the structural and functional consequences of mutations in viral proteins.

Estrada (2020) [40], represented the three-dimensional structure of SARS-CoV-2 proteins as a protein residue network, with each amino acid depicted as a node that is connected to another node if the corresponding residues had an inter-residue interaction (distance less than 7 Angstrom). The study computed topological centralities from this network representation, enabling the identification of key differences between the main protease of SARS-CoV-1 and SARS-CoV-2. This method provided valuable insights into the structural and functional implications of the observed differences in the protein structures of these two viruses.

Overall, the use of structural information and network analysis can provide valuable insights into the impact of mutations on protein structure and function. These methods can help in identifying potential drug targets and understanding the mechanisms of viral pathogenesis in such pandemics.

1.1.3 Epidemiology Modeling

Epidemiological modeling is a powerful tool that has played a crucial role in understanding the spread and impact of infectious diseases, particularly during pandemics. Modeling allows researchers and public health officials to simulate different scenarios and estimate the potential impact of interventions such as vaccination, social distancing, and quarantine. These models also allow for the estimation of key parameters, such as the transmission factor (β), which is a measure of the average number of people that an infected individual will go on to infect [70]).

Estimating beta is essential for understanding the potential for disease spread and for guiding public health interventions. In the case of pandemics, beta can vary widely depending on the virus in question and its method of transmission. For example, a virus that is primarily spread through droplets or aerosols may have a higher beta than one that is spread primarily through contact with contaminated surfaces or objects. Additionally, different strains of a virus may have different betas, which can impact

the potential for disease spread and the effectiveness of control measures [47].

In the case of the ongoing COVID-19 pandemic caused by the SARS-CoV-2 virus, epidemiological modeling has played a critical role in understanding the spread of the disease and in guiding public health interventions. Early models allowed researchers to estimate key parameters such as the basic reproduction number (R_0), which represents the average number of people that each infected individual will go on to infect, and the case fatality rate, which represents the proportion of cases that will result in death [77]. As the pandemic has evolved, models have been refined to incorporate new data and to estimate the potential impact of interventions such as lockdowns, travel restrictions, and vaccination campaigns.

One important use of epidemiological modeling in the context of COVID-19 has been to estimate the transmission factor for different strains of the virus. For example, the Alpha variant (B.1.1.7), which emerged in the UK in late 2020, has been found to be significantly more transmissible than earlier strains of the virus, leading to a surge in cases in many parts of the world [34]). Epidemiological models have been used to estimate the potential impact of such variants on the spread of the disease and to inform public health responses.

1.2 Motivation for current Thesis

India was the fifth country in the world to sequence the viral genome for inclusion in GISAID and has now become the country with the third-highest recorded number of infections. Due to a high population density and overwhelmed healthcare services, India is naturally at a higher risk of COVID-19 community transmission [14]. In this study, we examine the emergence of both region-specific mutations and globally prevalent variants that are dynamically evolving and circulating in different parts of India. Then we perform structural and network-based analysis to study the impact of mutations specific to India during the early phase and the first wave. Finally, we model the spread of COVID-19 at the lineage level for the second and third waves in India in order to estimate their transmission factors.

1.3 Organisation of Thesis

This thesis is organized as follows:

Chapter 2 details the mutations in samples of SARS-CoV-2 during the early phase (27th Jan - 27th May 2020) in India. Unique mutations in India, which are not seen in the rest of the world are discussed and novel state-specific subclades were identified. **Chapter 3** builds up on the previous chapter. We

examine the viral RNA samples over the entire 2 years of the pandemic over 4 different phases. Multiple novel state-specific subclades were identified for the first wave, and the prevalence of important VOCs/VOIs is discussed for the second and third waves in India. **Chapter 4** discusses the impact of India-specific mutations identified in the early phase and the first wave in India. We conducted network and structural analysis on the wild-type and mutated protein structures to evaluate the possible impact of the mutations. Important mutations such as ORF3a Q57H, N P13L, N S194L, and ORF1a A1812D are discussed. **Chapter 5** models the transmission of SARS-CoV-2 in India, with a focus on important states such as Maharashtra and Gujarat, during the second and third waves using variant counts. The dominant variants during the second wave were Delta (B.1.617.2) and Kappa (B.1.617.1), while Omicron (BA.2) was primarily responsible for the third wave. We estimated the transmission coefficient of these variants by constructing multi-strain SEIR models fitted using Markov chain Monte Carlo simulations. **Chapter 6** finally summarizes the main findings of the works described in the thesis and puts forwards the limitations and possible future directions of our work.

Chapter 2

Demographic Analysis of Indian SARS-CoV-2 Isolates during the Early Phase

2.1 Introduction

A large variation in the rate of infectivity and fatality due to COVID-19 is observed across different countries, with a similar trend observed across various states of India. To understand at the genetic level the role of acquired mutations in the circulating SARS-CoV-2 virus and their possible impact on the spread and virulence, a detailed analysis of Indian SARS-CoV2 isolates obtained from GISAID [126] during the early phase of the pandemic (27th Jan – 27th May 2020) is carried out. During the first half of this period (Jan-Mar), positive cases were mainly associated with the travel history of the individuals or their close contacts. Here, the distribution of the virus in different parts of the country based on mutational analysis of the infected individuals with travel history is discussed. The analysis of genetic variations accumulated is expected to indicate the impact of contact tracing, quarantine, and lockdown in containing the spread of COVID-19 during this period [99]. A detailed state-wise distribution of shared mutations and their global distribution across the world is carried out to understand the transmission and virulence of the disease within and between states. We expect this to help in identifying mutations responsible for the large variation in the number of cases and deaths across different states in India during the early period and identify corresponding subclades of Indian isolates.

2.2 Materials and Methods

For this study 705 Indian SARS-CoV-2 viral isolates sequence data was obtained from GISAID [126] corresponding to the period 27th Jan – 27th May 2020. Of these 20 had low-quality genomes (>5% missing bases) and were discarded and only 685 isolates data were considered for analysis. Comparison of 685 Indian SARS-CoV-2 isolates with Wuhan-1 isolate (NC_045512.2) as reference revealed a total of 1279 variations. We conducted a phylogenetic tree analysis of these isolates was carried out using the bioinformatic engine Augur and visualization tool Auspice from Nextstrain [58]. The Nextstrain

pipeline was executed using the Snakemake workflow which includes a multiple sequence alignment using MAFFT [69], followed by creating a maximum likelihood tree using IQ-TREE [94] with 1000 bootstraps. The tree was then time-resolved using TimeTree after refinement. The final step included the inference of clades, mutations (both at the nucleotide and amino acid levels), and ancestral traits to obtain the Newick format tree visualized using Auspice.

2.3 Phylogenetic Analysis

2.3.1 Clade Analysis

The phylogenetic tree of Indian isolates constructed with respect to Wuhan-1 isolate as reference using the Nextstrain pipeline is shown in Figure 2.1. We observe that all the five clades (defined in the new Nextstrain classification scheme) were present in India during this early period with the following distribution: 19A: 264 samples (38.54%), 19B: 44 samples (6.42%), 20A:300 samples (43.80%), 20B: 75 samples (10.95%), 20C: 2 samples (0.29%). The earliest recorded entries of SARS-CoV-2 in the country are both from Kerala with a travel history from Wuhan, China. Acc. ID: EPI_ISL_413522 on 27th Jan 2020 belonging to the root clade 19A and a sample of clade 19B on 31st Jan 2020 (Acc. ID: EPI_ISL_413523). By January end, these two clades had globally spread across most parts of the world, including India. The earliest sample corresponding to clade 20A is dated 3rd March, of a tourist from Italy (Acc. ID: EPI_ISL_420543). Two samples corresponding to clade 20B were observed during the same time, one having contact with another Indian with a travel history from Italy (Acc. ID: EPI_ISL_426179), dated 2nd March, and the other on 29th February (Acc. ID: EPI_ISL_414515), with no state or travel history available. Two samples of clade 20C were observed (Acc. ID: EPI_ISL_435051 and EPI_ISL_435052), dated 13th April in Gujarat with no travel history or contact with anyone with travel history. These results clearly indicate early community spread of the virus in the country.

The highest number of isolates belong to clade 20A (43.80%) in accordance with the global trend, with the next large cluster corresponding to root clade 19A (38.54%), comprising 82% of reported isolates in India till 27th May. Our analysis revealed subclusters of clade 20A and 19A, respectively, with defining mutations. It would be interesting to study the state-wise distribution of this clade along with travel history to assess any community transmission during the lockdown.

2.3.1.1 State-wise Clade Analysis

Of 685 Indian isolates, 658 isolates were considered for state-wise distribution analysis as 27 isolates had no state information available. As can be seen from Figure 2.2 Clade 20A is predominantly observed in Gujarat (178/201), followed by West Bengal (30/45), Odisha (22/46), Madhya Pradesh (16/19), and Maharashtra (16/80), with the earliest reported sample from Gujarat dated 5th April (Acc. ID: EPI_ISL_426414). Multiple independent entries of this clade are indicated based on clusters with

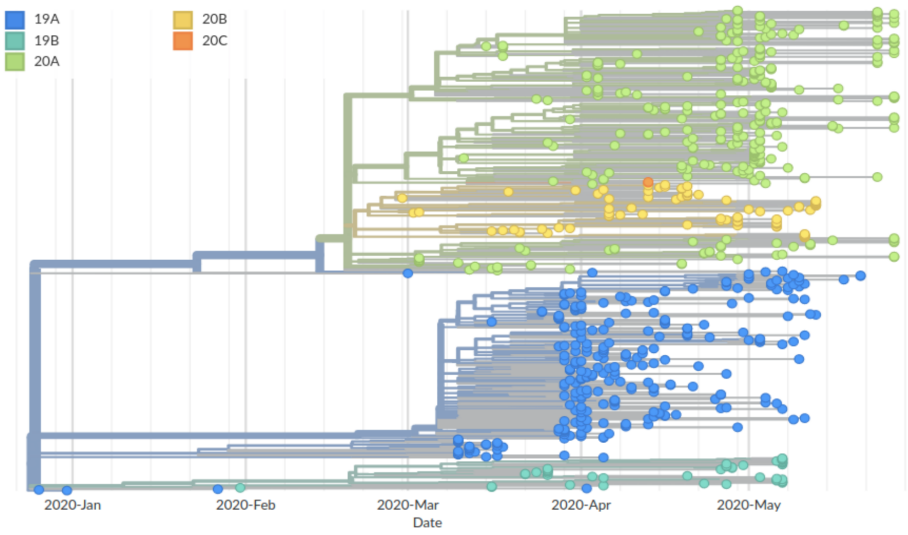


Figure 2.1: Phylogenetic tree obtained with Wuhan-1 isolate as reference depicts the divergence of Indian SARS-COV-2 isolates during the period 27th Jan – 27th May 2020.

shared mutations (discussed later in Section 2.5.1). The second major clade 19A is predominant in Telangana (75/97), followed by Delhi (55/76), Maharashtra (31/80), and Tamil Nadu (19/34). Very few isolates of clade 19B (Odisha (17), Maharashtra (10), Gujarat (8) and West Bengal (5)), clade 20B (Maharashtra (23), Telangana (15), Tamil Nadu (15), and Delhi (7)) and clade 20C (only from Gujarat (2)) are observed. We observed that the genomic sequences of a few isolates from Odisha were shorter in length with missing bases; 1 - 29 bases in 5' UTR (38 isolates) and 29686 - 29903 bases in 3' UTR (39 isolates), which are likely to be due to a sequencing artifact as the majority of these are from the same sequencing laboratory. The isolates from Odisha also contain other deletion regions, 23842 - 24400 bases in the S gene, 26306-26524 bases in the E gene, 27527-28033 bases covering ORF7b and ORF8 genes in 8 samples, and 28462 - 28680 bases in N gene (10 samples), 29000 – 29685 bases covering N and ORF10 genes (15 samples), which may be region-specific.

2.3.2 Mutational Analysis

2.3.2.1 Most frequent mutations

The important mutations in Indian isolates and their distribution state-wise are summarized in Table 2.1. Mutation C241T in the 5' UTR region of ORF1ab is the most common mutation with an incidence in more than half of the Indian isolates. It is highly prevalent in Gujarat (183/201), followed by Maharashtra (39/80) and West Bengal (33/45). Being a non-genic mutation, it does not cause an amino acid change. However, being part of the stem-loop region upstream of the ORF1ab gene, it may involve differential RNA binding affinity to the ribosome and affect the translational efficiency of the

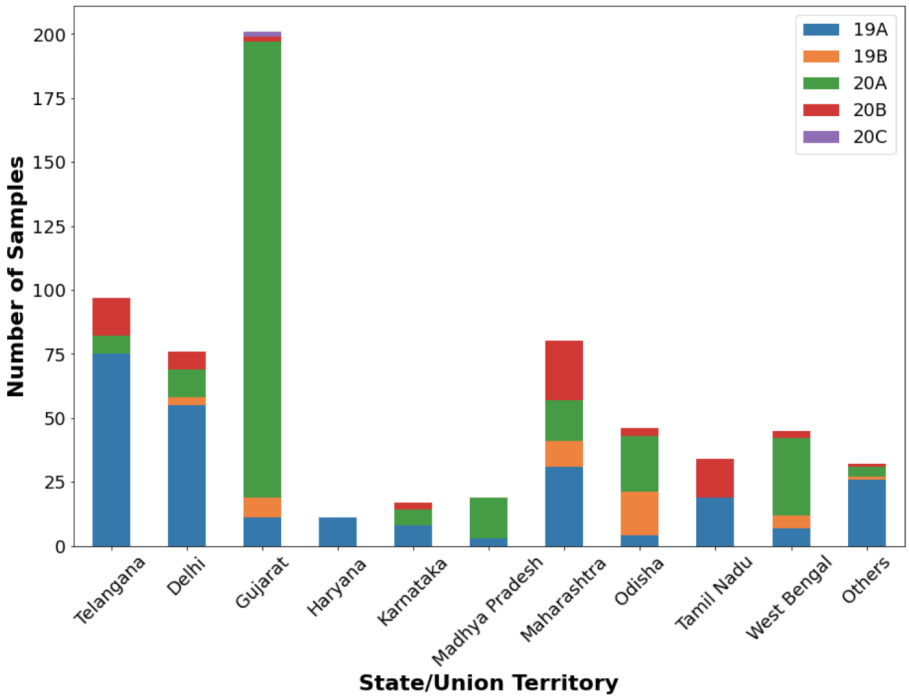


Figure 2.2: State-wise distribution of clades. Clade 20A is predominant in Gujarat while the root clade 19A is observed majorly in Telangana, Delhi, and Maharashtra.

virus, though no change in the RNA structure was observed for the mutant form [10]. Chaudhari et al. (2021) [27] analyzed the role of C241T mutation in the interaction of stem-loop region with the host replication factors, MADP1 Zinc finger CCHC-type and RNA-binding motif 1 (hnRNP1). Molecular docking and molecular dynamics simulations suggest reduced replication efficiency of the virus within the host and may result in less mortality. The D614G mutation (A23403G) in the Receptor Binding Domain (RBD) of Spike protein has been widely studied globally and has a similar distribution in the Indian samples. It was first observed in China and Germany in late January and is now the most prevalent mutation worldwide. It is observed with over 50% representation (373/685) in Indian isolates, and with a high predominance in Gujarat (180/201). Glycine being less bulky than aspartic acid, contributes to a more flexible hinge region in the Spike protein, enabling more efficient receptor-ligand binding. This provided a selective advantage to the virus in infection and transmission, making it predominant all over the world [155]. [108] characterized the replication of D614 and G614 viruses in a primary human airway tissue culture model and showed that the mutation enhances viral replication through increased virion infectivity and enhances the stability of the virus. Another study [33] suggested a possible mechanism for the increased infectivity of G614 variant to it being more resistant to proteolytic cleavage. This stability confers G614 variant its efficient transmission ability. The D614G mutation is found to co-occur with C241T, C3037T, and C14408T [73] and these are the defining mutations for clade 20A in Nextstrain and are predominantly found in North America and Europe. The C14408T

mutation within the RNA-dependent RNA polymerase (RdRp) encoding region of ORF1ab is a missense mutation that leads to an amino acid change from Proline to Leucine (P4715L). It is found in the Nsp12 region which is involved in replication and pathogenesis and is a potential target for antiviral candidates in coronaviruses. It is also observed in >50% of Indian samples with predominance in Gujarat (178/201), Maharashtra (37/80), and West Bengal (33/45). Nucleotide mutation C3037T is a silent mutation in the Nsp3 protein of ORF1ab (F924F) and no functional significance of this mutation is reported. The haplotype defined by these four co-occurring mutations is now globally the dominant form. This haplotype was proposed to be related to the pathogenicity of the virus and correlated with high death rates in Europe [13]. The high prevalence of these mutations in Gujarat with similar death rates suggests probable transmission of the virus to Gujarat from North America and Europe and its containment in the state due to the lockdown.

Table 2.1: Important mutations observed in Indian samples along with their frequency, state distribution. Acronyms used for states: GJ - Gujarat, WB – West Bengal, MH – Maharashtra, TL – Telangana, DL – Delhi, TN – Tamil Nadu.

Variant Description	**Nucleotide Mutation**	**Protein**	**Amino Acid Mutation**	**Frequency (India)**	**Frequency (States)**
Clade 20A	C241T	5' UTR	-	378 (55.2)	GJ (91.0), WB (73.3), MH (48.8)
Maharashtra-specific	C313T	ORF1ab (nsp1)	-	39 (5.7)	MH (25.0)
Subclade I/GJ-20A	C2836T	ORF1ab (nsp3)	-	51 (7.4)	GJ (25.4)
Clade 20A	C3037T	ORF1ab (nsp3)	-	374 (54.6)	GJ (90.6), WB (73.3), MH (48.8)
Maharashtra-specific	C5700A	ORF1ab (nsp3)	A1812D	33 (4.8)	MH (26.2)
Co-occurs with subclade I/A3i	C6310A	ORF1ab (nsp3)	S2015R	44 (6.4)	TL (15.5), DL (10.5)
Subclade I/A3i	C6312A	ORF1ab (nsp3)	T2016K	228 (33.3)	TL (73.2), DL (69.7), TN (55.9)
Co-occurs with subclade I/A3i	G11083T	ORF1ab (nsp6)	L3606F	258 (37.7)	TL (72.2), DL (71.0), TN (55.9)
Subclade I/A3i	C13730T	ORF1ab (nsp12)	A4489V	239 (34.9)	TL (74.2), DL (71.0), TN (55.9)

Clade 20A	C14408T	ORF1ab (nsp12)	P4715L	367 (53.6)	GJ (88.6), WB (73.3), MH (46.2)
Subclade I/GJ-20A	C18877T	ORF1ab (nsp14)	-	126 (18.4)	GJ (52.2)
Co-occurs with subclade I/A3i	C19524T	ORF1ab (nsp14)	-	60 (8.8)	TL (20.6), DL (13.2)
Subclade I/GJ-20A	C22444T	S	-	71 (10.4)	GJ (31.8)
Clade 20A	A23403G	S	D614G	373 (54.4)	GJ (89.6), WB (73.3), MH (48.8)
Subclade I/A3i	C23929T	S	-	224 (32.7)	TL (71.1), DL (69.7), TN (47.0)
Subclade I/GJ-20A	G25563T	ORF3a	Q57H	134 (19.6)	GJ (53.7)
Subclade I/GJ-20A	C26735T	M	-	123 (18.0)	GJ (50.2)
Subclade I/A3i	C28311T	N	P13L	236 (34.4)	TL (76.0), DL (68.4), TN (55.9)
Subclade I/GJ-20A	C28854T	N	S194L	73 (10.6)	GJ (32.8)
Clade 20B	G28881A	N	R203K	75 (10.9)	TN (44.1), MH (28.8), TL (15.5)
Clade 20B	G28882A	N	R203K	75 (10.9)	TN (44.1), MH (28.8), TL (15.5)
Clade 20B	G28883C	N	G204R	75 (10.9)	TN (44.1), MH (28.8), TL (15.5)
Maharashtra-specific	A29827T	3' UTR	-	44 (6.4)	MH (55.0)
Maharashtra-specific	G29830T	3' UTR	-	61 (8.9)	MH (73.8)

Clade 20B defining mutations, G28881A, G28882A, and G28883C, result in two adjacent amino acid changes R203K and G204R in the nucleocapsid (N) protein. This trinucleotide-bloc mutation, 28881-28883: GGG>AAC, is reported to result in two sub-strains of SARS-CoV-2, viz., SARS-CoV-2g and SARS-CoV-2a. The AAC genotype is observed from March globally and in India (75 samples) it is mainly observed in Maharashtra (23/80), Tamil Nadu (15/34), and Telangana (15/97). Mutation RG>KR is observed to disrupt the S-R motif by introducing lysine between them. This may affect the phosphorylation of the SR-rich domain that plays an important role in the cellular localization and

translation inhibitory function of the N protein [11]. Earlier experimental work on SARS-CoV-1 has shown reduced pathogenicity on deleting part of the SR domain [138]. These mutations also reduce the number of miRNA binding sites from seven to three and thereby increase the likelihood of viral infection [88]. The high prevalence of these mutations in Gujarat with similar death rates suggests probable transmission of the virus to Gujarat from North America and Europe. Its containment in the state due to lockdown and local transmission is probably the reason for its prevalence in Gujarat state.

2.3.2.2 India-specific mutations

The diversity plots for SARS-CoV-2 isolates from India and globally are given in Figure 2.3 with sites exhibiting higher entropy in Indian isolates compared to global isolates marked. Four of these, C6312A (T2016K) and C13730T (A4489V) in ORF1ab, C23929T in Spike protein, and C28311T (P13L) in N gene have been reported in an earlier study as India-specific subclade I/A3i defining mutations [15]. Two other India-specific mutations, C6310A (S2015R) and C19524T in ORF1ab are identified to be associated with this subclade and correspond to two branch points of subclade I/A3i with predominance in Delhi and Telangana. Our analysis also revealed mutations C313T (synonymous) and C5700A (A1812D) in ORF1ab which co-occur with high frequency in India compared to their global presence. These mutations define a unique cluster of viral isolates predominantly observed in Maharashtra and Telangana states. The mutation C5700A is observed to branch out of clade 20B primarily, and some clade 20A samples also exhibit this mutation (Figure 2.4). This mutation is unique to the Indian region ($\sim$5% frequency) with no presence globally and its co-occurrence with mutation C313T is observed in 33 Indian samples: Maharashtra (21), Telangana (8), and Gujarat (4). The first instance of this mutation was on 4th April 2020 in Maharashtra, and it is also reported in an earlier study on 90 sequences from western India [103]. Its presence is observed to increase with 10% of Indian samples by July 2020 carrying this mutation, indicating local transmission. Mutation C5700A lies in the Nsp3 region of ORF1ab, which forms multi-subunit assemblies with other nonstructural proteins for the formation of replication transcription complex [76]. It has the ability to alter the surface electrostatic environment in the proximity of Nsp3's viral protease domain [56]. In Figure 2.3, the mutations in green, C2836T, C18877T, C22444T, C26735T, and C28854T are specific to Gujarat viral isolates while those in black, A29827T and G29830T are observed in Maharashtra, and their significance is discussed below in Section 2.5.2.

2.3.2.3 Clade I/A3i

About one-third of Indian isolates (219/685) defined by co-occurring mutations C6312A, C13730T, C23929T, and C28311T are associated with subclade I/A3i that branches out of clade 19A (Figure 2.5). State-wise distribution analysis of this subclade (Figure 2.6) revealed its predominance in Telangana (69/97, $\sim$71%), Delhi (52/76, $\sim$68%), Haryana (11/11) and Tamil Nadu (16/34). The mutation C13730T (ORF1ab: A4489V) lies in the RdRp protein's NiRAN domain which is involved in RNA

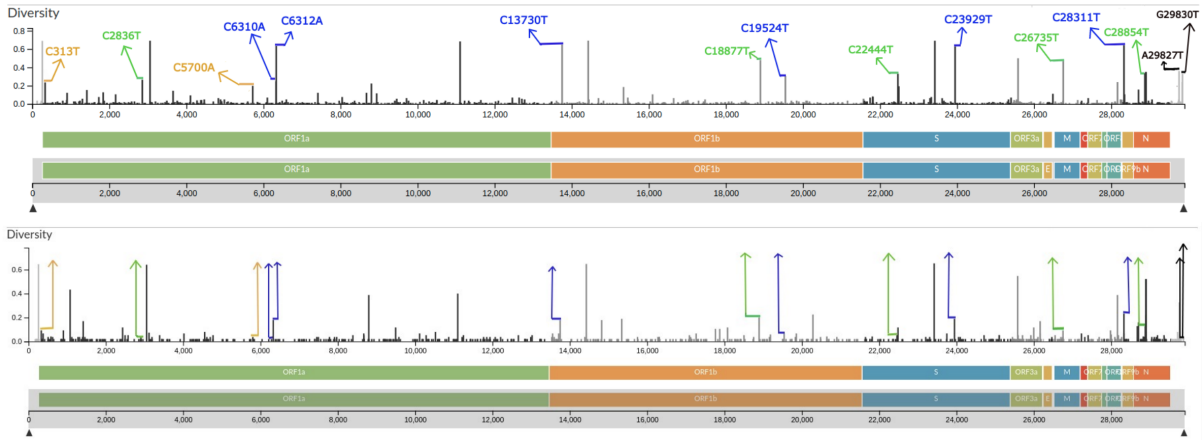


Figure 2.3: The diversity plots shown for isolates from (a) India and (b) World. In (a) 15 mutations predominant in India are marked in Blue (clade 19A): C6310A, C6312A, C13730T, C19524T, C23939T, and C28311T, Green (clade 20A): C2836T, C18877T, C22444T, C26735T, C28854T, Orange (clade 20B): C313T, C5700A, and Black (not specific to any particular clade): A29827T, G29830T.

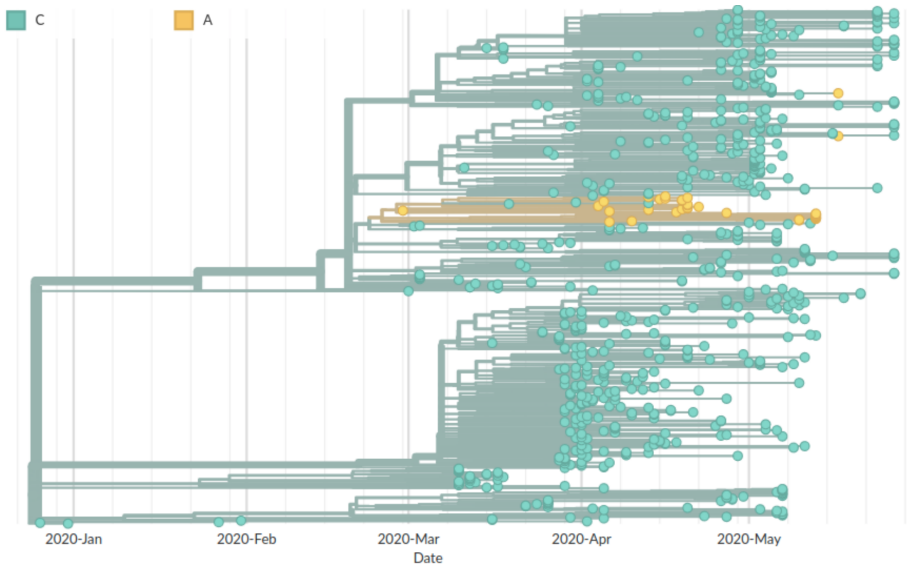


Figure 2.4: The sequences carrying the mutation C5700A are depicted in yellow color on the phylogenetic tree.

binding and nucleotidylation and is essential for viral replication [54]. The nucleocapsid protein (N) is required for viral RNA replication, transcription, as well as genome packing [62, 89]. Mutation C28311T (N: P13L) is located in the protein's intrinsically disordered region and may impair the terminal domain's RNA-binding function [25, 26]. A previous study evaluated how this mutation may alter

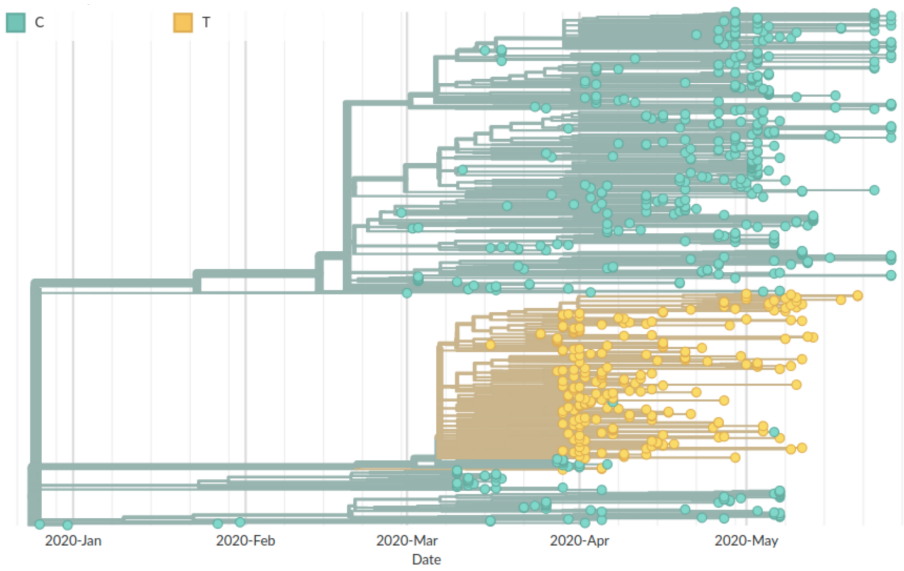


Figure 2.5: The sequences carrying the mutation C23929T are depicted in yellow color on the phylogenetic tree.

protein-protein interaction and proposed its impact on virus stability, potentially contributing to lower pathogenesis [100].

Analysis of subclade I/A3i is important both scientifically and epidemiologically as its defining mutations are found in ∼32% of Indian samples, while outside India its distribution is very low (∼3.5%). This clearly hints at early community transmission due to some super spreader event during March-April, as it is highly unlikely that around one-third of the samples sharing the same set of mutations could have arisen by multiple independent entries with international travel history, especially when its presence globally is negligibly small. The first reported entry of this clade in India is on 16th March 2020 in Telangana from an Indonesian citizen visiting India and globally first incidence of this subclade is in a sample from Saudi Arabia, dated 7th March 2020. It is predominantly found in Asia (mostly Singapore and Malaysia) (Figure 2.6). The coincidence of subclade I/A3i in the country following the Tablighi congregation held during early March, the first reported case from Saudi Arabia on 7th March, and several Indonesian citizens identified with these mutations in Delhi during that period, all hint at the Tablighi congregation event as the likely cause of the spread of this subclade. No isolates belonging to this subclade were observed after May, except for 5 samples in India with the last one dated 13th June 2020 (according to data available in Nextstrain). This indicates that the spread of subclade I/A3i had been largely contained during the lockdown with efforts of contact tracing and quarantine of infected individuals. Similarly, no further global spread of this subclade is observed, probably due to the air travel ban, the efficacy of contact tracing, and quarantining of COVID-19-positive individuals. Thus, genetic

analysis can help in identifying the chain of transmission of infection and the success of measures used in its containment.

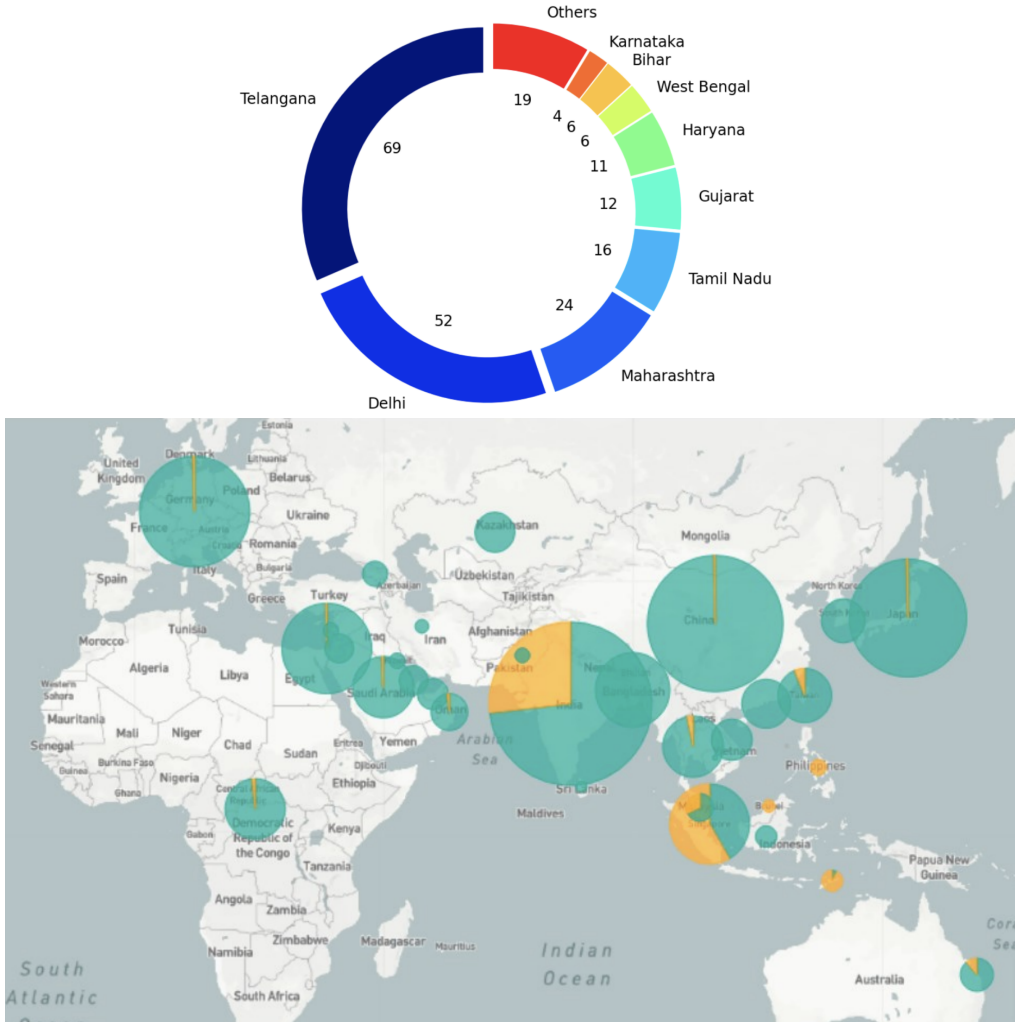


Figure 2.6: (a) State wise distribution of the subclade I/A3i isolates in India. (b) The global distribution (zoomed in to Asia) of the mutation C23929T part of I/A3i is shown.

2.3.2.4 Samples with travel history from foreign nations

With the ban on international flights from March 23rd, 2020, and the lockdown imposed in the country two days later, no new transmissions were likely from outside the country. Further, lockdown would have promoted the accumulation of region-specific mutations. Analysis of shared mutation sets can help in understanding the extent of local transmission and containment of the circulating strains in a localized region. In 14 samples with travel history from Iran, 13 samples had the mutations, G1397A, T28688C, and G29742T. Amongst them, 11 also had mutations, C884T, G8653T, A29879G, A29883T,

and A29901T. These mutations were observed in very few other Indian isolates; Ladakh (6/6), Kargil (1/1), and Maharashtra (1/80) indicating the efficacy in quarantining and contact tracing during the lockdown period. The 11 samples with travel history from Italy contained the mutations C3037T, C241T, C14408T (ORF1ab: P4175L), and A23403G (S: D614G), which are clade 20A defining mutations (in Nextstrain). These mutations are found in 50% of Indian samples and with the highest frequency in Gujarat ($\sim$90% samples). Eight samples from Indonesia sampled in Delhi contained mutations C23929T, C6312A, C13730T, and C28311T. As discussed above, this mutation set, identified as subclade I/A3i, was mainly observed in Telangana, Delhi, and Tamil Nadu but no further increase was seen in the number of samples containing it.

2.4 Principal Component Analysis

Since travel between states was restricted because of the lockdown, an increase in the number of samples with state-specific shared mutations is expected because of local community transmission. To see if any state-specific clustering is observed during this initial period, principal component analysis (PCA) on the mutational profile of Indian isolates was performed. Figure 2.7 shows the plot of the first two principal components that captured over 35% variance in this high-dimensional data. Mutations were sorted based on their loading scores to assess their impact on PC1. Top 10 mutations thus identified are: A23403G (Spike: D614G), C3037T, C241T, C14408T (ORF1ab: P4715L), G11083T (ORF1ab: L3606F), C28311T (N: P13L), C6312A (ORF1ab: T2016K), C13730T (ORF1ab: A4489V), C23929T, and G25563T (ORF3a: Q57H). Not surprisingly, these are also the top 10 most common mutations in Indian isolates. It may be noted from Figure 2.7 that the Gujarat samples shown in 'pink' form a distinct cluster. Samples from Telangana, Delhi, and Haryana states cluster to the right in the plot due to the shared mutation profile of subclade I/A3i. Samples from Maharashtra are scattered throughout the plot, though in closely grouped clusters. A detailed analysis revealed two sets of co-occurring mutations in Maharashtra isolates (discussed later).

2.5 Novel subclades

It is observed that certain countries like Italy, UK, Spain, etc. had a large number of mortality cases. The severity of the circulating strains and the large aging population was proposed to be the probable cause. India is a very vast country, and it would be interesting to study if a similar pattern is observed across its different states. We analyzed the infection and death rates across states and states with high severity of COVID-19 cases. We observe Gujarat (5.12%) and Maharashtra (4.19%) reported higher death rates compared to the country average (2.67%), as of 11th July 2020. The recorded number of deaths in Gujarat is 2008 out of a total of 39194 cases while Karnataka (31105), Telangana (30946), and Uttar Pradesh (32363) with a similar number of cases, recorded much fewer fatalities, 486, 331, and

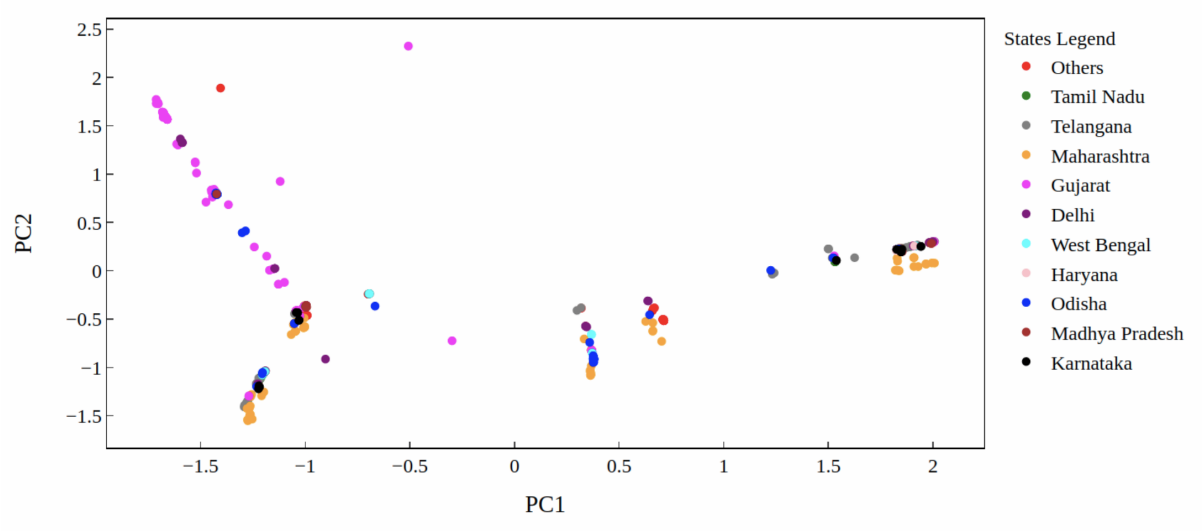


Figure 2.7: PCA plot of 685 samples colored state-wise.

862, respectively. Even some of the worst-hit regions such as Delhi (3.29%) and Tamil Nadu (1.39%) had much lower death rates compared to Gujarat. To understand this significantly large difference in the percentage of deaths in Gujarat at the genetic level, a detailed analysis of the mutational profile of sequences from Gujarat with that of the Rest of India (RoI) was carried out. Tables 2.2 and 2.3 summarize non-synonymous mutations that are over- and under-represented in Gujarat isolates compared to the rest of the country. Clade 20A defining mutations are observed in ∼90% of Gujarat samples. Due to the countrywide lockdown from 25th March 2020 clade 20A and its sub-clusters were localized in the state and are identified as Gujarat-specific mutations. A reverse scenario, that is, under-representation of certain mutations is observed in Gujarat that has a high frequency in RoI, e.g., mutations defining subclade I/A3i and clade 20B.

This difference in the mutational profile of isolates from Gujarat with that of the RoI is clearly seen in the diversity plots for non-synonymous mutations in isolates in Figure 2.8. It may be noted from plots 2.8(a) and 2.8(c) that the frequency of mutations in Gujarat isolates is very different from that of isolates from the whole of India. In contrast, the diversity plots of Telangana (Fig 2.8(b)) and India (Fig 2.8(c)) are quite similar, probably due to subclade I/A3i. The death rates for Telangana and India are also comparable, while that of Gujarat is strikingly different. This analysis suggests that the characteristic mutations of subclade I/A3i may not be deleterious.

2.5.1 Gujarat subclade I/GJ-20A

Analysis of Gujarat isolates revealed a novel subclade of 20A which we refer to as I/GJ-20A (Figure 2.9) and is defined by the shared mutations C18877T, G25563T (ORF3a: Q57H), and C26735T

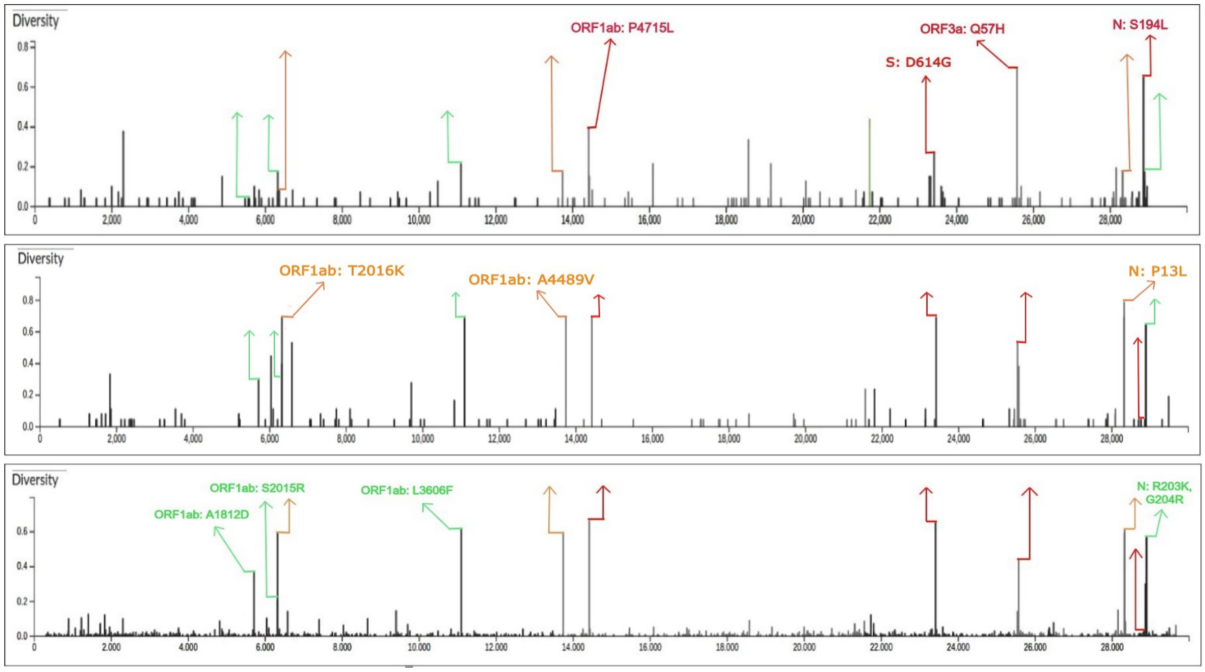


Figure 2.8: Diversity plots for non-synonymous mutations in isolates from (a) Gujarat, (b) Telangana, and (c) India clearly exhibit different sets of mutations.

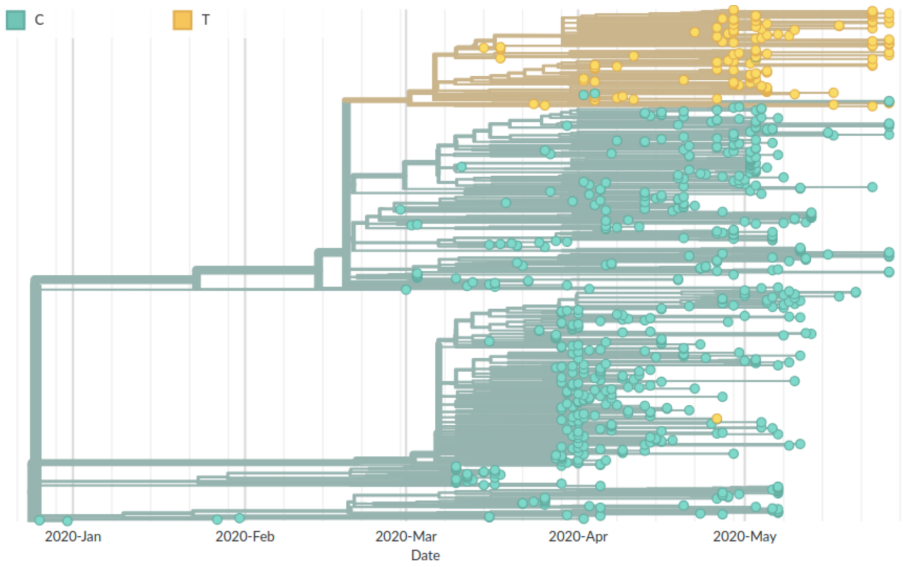


Figure 2.9: The sequences carrying the mutation C18877T are depicted in yellow color on the phylogenetic tree.

(listed in Table 2.2). The high prevalence of these mutations in Gujarat has also been reported in an earlier study [67]. These mutations are under-represented in RoI and their analysis may be helpful in

Table 2.2: Over-represented mutations observed in SARS-CoV-2 isolates from Gujarat (201 samples) compared to the Rest of India, RoI (484 samples) are listed.

Nucleotide Mutation	**Protein**	**Amino Acid mutation**	**Count in Gujarat**	**Frequency in Gujarat (%)**	**Count in RoI**	**Frequency in RoI (%)**
C241T	5' UTR	-	183	91.0	195	40.3
C3037T	ORF1ab (nsp3)	-	182	90.6	192	39.7
A23403G	S	D614G	180	89.6	193	39.9
C14408T	ORF1ab (nsp12)	P4175L	178	88.6	189	39.0
G25563T	ORF3a	Q57H	108	53.7	26	5.4
C18877T	ORF1ab (nsp14)	-	105	52.2	21	4.3
C26735T	M	-	101	50.2	22	4.5
C28854T	N	S194L	66	32.8	7	1.4
C22444T	S	-	64	31.8	7	1.4
C2836T	ORF1ab (nsp3)	-	51	25.4	0	0.0

explaining high death rates in Gujarat compared to the other Indian states. Mutations in ORF3a are of significance as the ORF3a protein is involved in regulating immunological responses in the host, including "cytokine storm". The mutation results in significant conformational changes to the protein and forms a more stable quaternary structure thereby increasing the viral particle release resulting in increased transmission [142, 100]. Other mutations that are part of this subclade are C2836T, C22444T, and C28854T (N: S194L). The S194L mutation due to C28854T resides in the central region of the nucleocapsid protein that is essential for oligomerization and is expected to alter the protein structure [158, 152, 145].

2.5.2 Maharashtra-specific mutations

A similar analysis of Maharashtra isolates when compared to isolates from RoI showed noticeable differences at the genetic level (Table 2.4). Maharashtra with the second highest death rate (4.19%) after

Table 2.3: Under-represented mutations observed in SARS-CoV-2 isolates from Gujarat (201 samples) compared to the Rest of India, RoI (484 samples) are listed.

Nucleotide Mutation	**Protein**	**Amino Acid mutation**	**Count in Gujarat**	**Frequency in Gujarat (%)**	**Count in RoI**	**Frequency in RoI (%)**
C313T	ORF1ab (nsp1)	-	2	1.0	37	7.6
C5700A	ORF1ab (nsp3)	A1812D	4	2.0	29	6.0
C6310A	ORF1ab (nsp3)	S2015R	6	3.0	38	7.8
C6312A	ORF1ab (nsp3)	T2016K	12	6.0	216	44.6
G11083T	ORF1ab (nsp6)	L3606F	16	8.0	242	50.0
C13730T	ORF1ab (nsp12)	A4489V	12	6.0	227	46.9
C19524T	ORF1ab (nsp14)	-	6	3.0	54	11.2
C23929T	S	-	12	6.0	212	43.8
C28311T	N	P13L	13	6.5	223	46.1
G28881A, G28882A, G28883C	N	R203K, G204R	2	1.0	73	15.1
A29827T	3' UTR	-	0	0.0	44	9.1
G29830T	3' UTR	-	0	0.0	61	12.6

Gujarat also recorded the highest number of infections. Two co-occurring mutations in the 3' UTR, A29827T and G29830T, exhibited a prevalence of 55% and 73.75% respectively in the isolates from Maharashtra but their presence was observed in only 2 isolates from RoI. Another set of co-occurring mutations, C313T and C5700A (ORF1ab: A1812D) was observed in Maharashtra with a high frequency

Table 2.4: Mutations observed in SARS-COV-2 isolates from Maharashtra (80 samples) with high frequency compared to the Rest of India (605 samples) are listed.

Nucleotide Mutation	**Protein**	**Amino Acid Mutation**	**Count in Maharashtra**	**Frequency in Maharashtra (%)**	**Count in RoI**	**Frequency in RoI (%)**
C313T	ORF1ab (nsp1)	-	20	25.0	19	3.1
C5700A	ORF1ab (nsp3)	A1812D	21	26.2	12	2.0
A29827T	3' UTR	-	44	55.0	0	0.0
G29830T	3' UTR	-	59	73.8	2	0.3

of 25% and 26.25% respectively but with $<3\%$ in isolates from RoI. These two distinct sets of mutations form two distinct clusters of samples from Maharashtra in the PCA plot. Among these only C5700A (A1812D) is a non-synonymous mutation in the nsp3 protein of ORF1ab and its functional significance has been discussed above in Section 2.3.2.2.

2.6 Conclusion

In this study, demographic analysis of mutations in Indian SARS-CoV-2 isolates was conducted to understand the viral spread in the country during the early phase of the pandemic (27th Jan 2020 – 27th May 2020) and assess the effectiveness of contact tracing, quarantine, and lockdown in controlling its spread. Genetic analysis revealed that though lockdown helped in controlling the spread of the virus, a region-specific set of shared mutations observed indicate local transmission within the states. In this study, we identified a Gujarat-specific novel subclade I/GJ-20A along with two distinct sets of mutations in Maharashtra. This provides a probable explanation of the observed variation in the number of infected cases and death rates across these states. Region-specific sequencing efforts can help in understanding the mutational spectra in local hotspots like Gujarat and Maharashtra and aid in the implementation of state-wise lockdown policies. Due to limited sequencing data, numerous variants are likely to have been missed. A vast country like India needs to improve its sequencing efforts to understand the transmission dynamics of the virus, capture novel variants and assess their clinical impact to follow appropriate containment measures; this would prepare us for future waves of COVID-19.

Chapter 3

Comparative Analysis of SARS-CoV-2 Variants Across Three Waves in India

3.1 Introduction

In this study we attempt to examine the emergence of region-specific mutations and their spread across the country over a period of two years across four different phases (27th Jan 2020 – 8th March 2022) with three waves of the coronavirus disease 2019 (COVID-19) pandemic caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) in India. The first two phases were analysed to show state-specific strains and their importance in terms of transmissibility and severity, as well as the efficiency of contact tracing, quarantine, and lockdown in restricting its spread after the early phase. Multiple novel subclades, I/MH-2, I/TelA-20B, I/TelB-20B, including I/GJ-20A (from early phase) from different states in India were reported for the first wave in India. To understand the distinctions between the later waves, sequences were analysed using their classification as Pangolin lineages. Multiple VOCs/VOIs across the country were analyzed and their potential influence on transmissibility and pathogenicity was studied. Significant mutations of lineages such as Delta, Kappa, and Omicron variations were also analysed.

3.2 Materials and Methods

Building up on our previous work, in this study, we provide a comparative analysis of Indian SARS-CoV-2 isolates for four strategic time points. The early phase (discussed in detail in the previous chapter) was characterized by localised region-specific sets of mutation. During the first wave, we saw higher number of infections although some region-specific sublades continued to prevail. The second wave was quite severe across the country, resulting in a huge number of hospitalisations and deaths. Finally, while having a large number of infections (the majority of which were not recorded owing to being asymptomatic), the third wave was substantially lower in terms of severity and death.

1. Early Phase (27th Jan - 27th May '2020): Dataset-I (685 samples)

2. After 1st wave (till 11th Jan '21): Dataset-II (4708 samples)

3. After 2nd wave (till 1st Oct '21): Dataset-III (45171 samples)

4. After 3rd wave (till 8th Mar '22): Dataset-IV (101527 samples)

The data for Datasets I and II were obtained from GISAID, and the Wuhan sequence (EPI_ISL_402125) was used as a reference. Any incomplete metadata or low-quality genomes (>5% missing bases) have been discarded. To assess the impact of the lockdown during the early phase and after the first wave in India, a comparative analysis is carried out between Datasets I and II. This comparison helps us analyse the survival and spread of state-specific mutations observed in the early phase after the lockdown was lifted. The evolution and spread of various lineages are discussed specifically after the second and third waves in the country. For this analysis, the mutation and lineage frequencies in Datasets III and IV were obtained from COVID-19 CG [28]. Phylogenetic analysis of the samples in Datasets I and II was carried out using the resources Augur and Auspice from Nextstrain [58]. The Nextstrain pipeline was the same as that explained in Section 2.2. The time-resolved phylogenetic tree thus obtained for the first year of COVID-19 (Jan '20 – Jan '21), corresponding to Dataset-II is shown in Fig 3.1.

State-specific clusters based on shared mutations (in Datasets I and II) are identified by performing a principal component analysis (PCA) on the mutational profile of Indian isolates. To assess the impact of India-specific non-synonymous mutations on protein function, SIFT [140] and PROVEAN [31] have been used. SIFT (Sorting Intolerant From Tolerant) uses sequence homology and properties of amino acids to predict whether an amino acid substitution would affect the protein function. PROVEAN (Protein Variation Effect Analyzer) predicts whether an amino acid substitution/indel has an impact on the biological function of a protein based on its homologs searched against the NCBI 'nr' database using BLAST and clustered using CD-HIT. Mutations are defined as deleterious if the PROVEAN score is <-2.5 and SIFT score is <0.05.

3.3 Phylogenetic Analysis

3.3.1 Clade Analysis

As mentioned previously, 5 clades, namely 19A, 19B, 20A, 20B, and 20C had been observed in the early phase of the pandemic in India (Dataset-I). By the end of the first year of the pandemic (Dataset-II), 2 new clades (20E, 20I/501Y.V1) were observed with 20B being the most prominent clade followed by 20A and 19A covering >97% of samples in the country. Of the 4708 samples in Dataset-II, 4678 had state information available (covering 17 states and 2 Union Territories: Delhi and Ladakh). Before lockdown clade 20A was pre-dominant in Gujarat (followed by West Bengal, Odisha, and Madhya Pradesh), and by Jan'21 (Dataset-II) it was well-represented in the majority of states, with higher representation in Gujarat (583/655), Maharashtra (323/1299), Andhra Pradesh (204/281) and West Bengal

Figure 3.1: Time-resolved radial phylogenetic tree for the period (Jan' 2020 – Jan – 2021) with samples colored according to Nextstrain clades: Dark blue: 19A, Sky blue: 19B, Dark green: 20A, Light green: 20B, Yellow: 20C, Orange: 20E (EU1), Red: 20I/501Y.V1.

(178/232). Similarly, Clade 20B with minimal representation in Maharashtra, Tamil Nadu, and Telangana in Dataset-I, showed a tremendous increase in Maharashtra (915/1299) and Telangana (724/970), followed by Karnataka (116/292). The earliest clade 19A (and its India-specific subclade I/A3i [15], which was earlier seen in four states, Telangana, Delhi, Maharashtra, and Tamil Nadu, however, showed a small increase only in Delhi (196/338), suggesting the effect of contact tracing and quarantine during the lockdown in containing their spread. Thus, from the distribution of clades across different states in Figure 3.2, we observe localized transmissions during the early phase and the spread of some of these variants to other states after the lockdown was lifted.

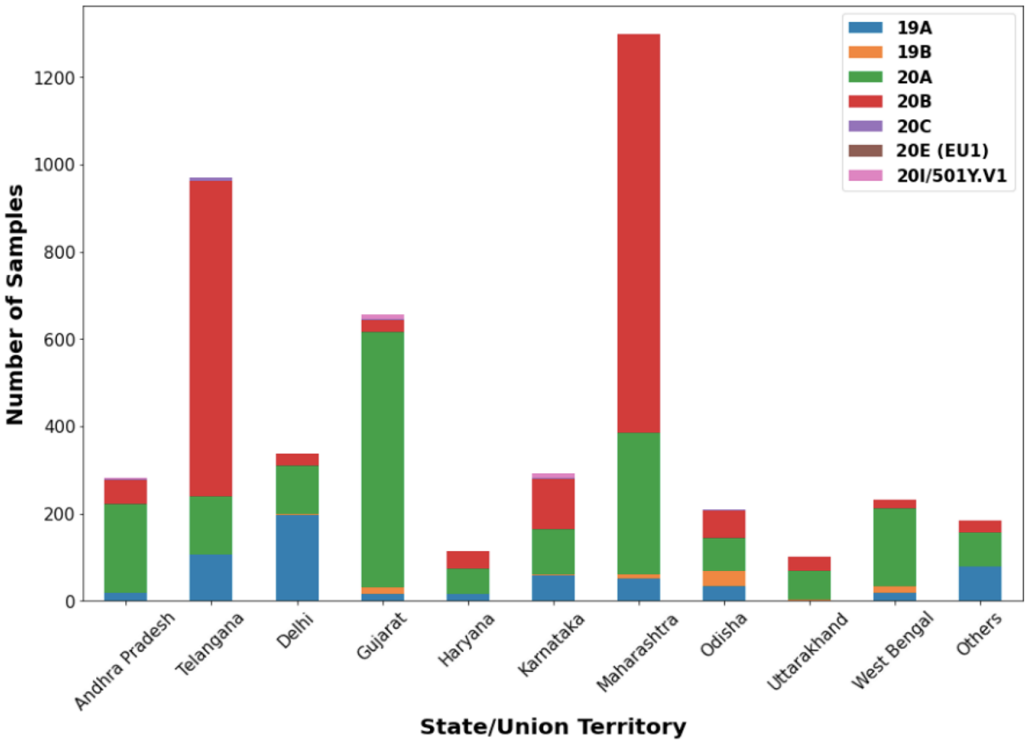


Figure 3.2: State-wise distribution of clades across the country shown for the period 26th Dec' 2019 – 21st Jan' 21 (Dataset-II).

Table 3.1: Frequency of the significant mutations circulating in the first year of the pandemic is given for the four time-points (i.e., in Datasets I, II, III, and IV) to understand their evolution in India. Their functional relevance is assessed using SIFT and PROVEAN. (PROVEAN analysis, N: Neutral, D: Deleterious; SIFT analysis, A: Affects function, T: Tolerated. Del*: Deletion not supported, Failed**: PSI-BLAST could not retrieve enough sequences.)

Mutation	**Amino Acid Mutation**	**Freq in Dataset-I**	**Freq in Dataset-II**	**Freq in Dataset-III**	**Freq in Dataset-IV**	**PROVEAN Analysis**	**SIFT Analysis**
A23403G	S: D614G	373 (54.4%)	4002 (85.0%)	44319 (98.2%)	100624 (99.1%)	0.60 (N)	0.62 (T)
C14408T	ORF1b: P314L	367 (53.6%)	3941 (83.7%)	44087 (97.6%)	100286 (98.8%)	-0.45 (N)	0.23 (T)
C3037T	-	374 (54.6%)	3984 (84.6%)	44094 (97.6%)	100142 (98.6%)	-	-

C241T	-	378 (55.2%)	3973 (84.4%)	43983 (97.4%)	98550 (97.1%)	-	-
G28881A	N: R203K	75 (11.0%)	2077 (44.1%)	7123 (15.8%)	23976 (23.6%)	-1.60 (N)	0.11 (T)
G28883C	N: G204R	75 (11.0%)	2066 (43.9%)	7045 (15.6%)	23892 (23.5%)	-1.66 (N)	0.02 (A)
G28882A	N: R203K	75 (11.0%)	2073 (44.0%)	7043 (15.6%)	23887 (23.5%)	-1.60 (N)	0.11 (T)
G25563T	ORF3a: Q57H	134 (19.6%)	1057 (22.4%)	4678 (10.4%)	5903 (5.8%)	-3.29 (D)	0 (A)
C26735T	-	123 (18.0%)	1039 (22.1%)	4521 (10.0%)	5723 (5.6%)	-	-
C18877T	-	126 (18.4%)	1043 (22.1%)	4462 (9.9%)	5851 (5.8%)	-	-
C28854T	N: S194L	73 (10.7%)	887 (18.8%)	4320 (9.6%)	5412 (5.3%)		
C22444T	-	71 (10.4%)	859 (18.2%)	3829 (8.5%)	4895 (4.8%)	-	-
C2836T	-	51 (7.4%)	314 (6.7%)	562 (1.2%)	603 (0.6%)	-	-
C313T	-	39 (5.7%)	1242 (26.4%)	2008 (4.4%)	2202 (2.2%)	-	-
C5700A	ORF1a: A1812D	33 (4.8%)	1253 (26.6%)	1884 (4.2%)	2024 (2.0%)	-0.75 (N)	0.41 (T)
G11083T	ORF1a: L3606F	258 (37.7%)	662 (14.1%)	1790 (4.0%)	3418 (3.4%)	-1.43 (N)	0.01 (A)
C28311T	N: P13L	236 (34.4%)	557 (11.8%)	641 (1.4%)	16721 (16.5%)	-1.23 (N)	0 (A)
C13730T	ORF1b: A88V	239 (34.9%)	567 (12.0%)	645 (1.4%)	688 (0.7%)	-2.35 (N)	0 (A)
C23929T	-	224 (32.7%)	525 (11.2%)	603 (1.3%)	682 (0.7%)	-	-
C6312A	ORF1a: T2016K	228 (33.3%)	532 (11.3%)	554 (1.2%)	568 (0.6%)	-0.17 (N)	0.59 (T)
C3267T	ORF1a: T1001I	0 (0%)	261 (5.5%)	4499 (10.0%)	5882 (5.8%)	0.22 (N)	0.17 (T)

C21034T	ORF1b: L2523F	0 (0%)	242 (5.1%)	1318 (2.9%)	2223 (2.2%)	-2.01 (N)	0.13 (T)
T28277C	N: S2P	0 (0%)	220 (4.7%)	1350 (3.0%)	2214 (2.2%)	-0.66 (N)	0 (A)
G26173T	ORF3a: E261del	0 (0%)	240 (5.1%)	1383 (3.0%)	2213 (2.2%)	-0.87 (N)	Del*
G28183T	ORF8: S97I	0 (0%)	242 (5.1%)	1370 (3.0%)	2184 (2.2%)	-3.67 (D)	Failed**
C6573T	ORF1a: S2103F	6 (0.9%)	426 (9.0%)	582 (1.3%)	796 (0.8%)	-0.37 (N)	0.04 (A)
C25528T	ORF3a: L46F	6 (0.9%)	418 (8.9%)	518 (1.2%)	591 (0.6%)	-3.30 (D)	0 (A)
G4354A	-	5 (0.7%)	422 (9.0%)	521 (1.2%)	580 (0.6%)	-	-
C9693T	ORF1a: A3143V	8 (1.2%)	291 (6.2%)	432 (1.0%)	542 (0.5%)	-0.23 (N)	1 (T)
C16626T	-	8 (1.2%)	278 (5.9%)	345 (0.8%)	388 (0.4%)	-	-
A4372G	-	5 (0.7%)	141 (3.0%)	185 (0.4%)	213 (0.2%)	-	-
G29474T	N: D401Y	2 (0.3%)	108 (2.3%)	145 (0.3%)	171 (0.2%)	-0.72 (N)	0 (A)
A21551T	ORF1b: N2695L	7 (1.0%)	123 (2.6%)	149 (0.3%)	171 (0.2%)	0.30 (N)	0.01 (A)
A21550C	ORF1b: N2695L	7 (1.0%)	125 (2.7%)	136 (0.3%)	150 (0.2%)	0.30 (N)	0.01 (A)

3.3.2 Mutation Analysis

A total of 1279 variations were identified in Dataset-I (685 sequences) and 7126 variations in Dataset-II (4708 sequences) when compared with Wuhan-1 isolate as reference. The top 20 most frequent mutations in Dataset-II along with their amino acid change (if any) along with their frequencies in Datasets I and II are shown in Figure 3.3. The clade 20A mutation A23403G (D614G) in the Receptor Binding Domain (RBD) of Spike protein, first discovered in late January in the West European region [73] was observed in over 50% of samples in Dataset-I (373/685) with a high prevalence in Gujarat (180/201).

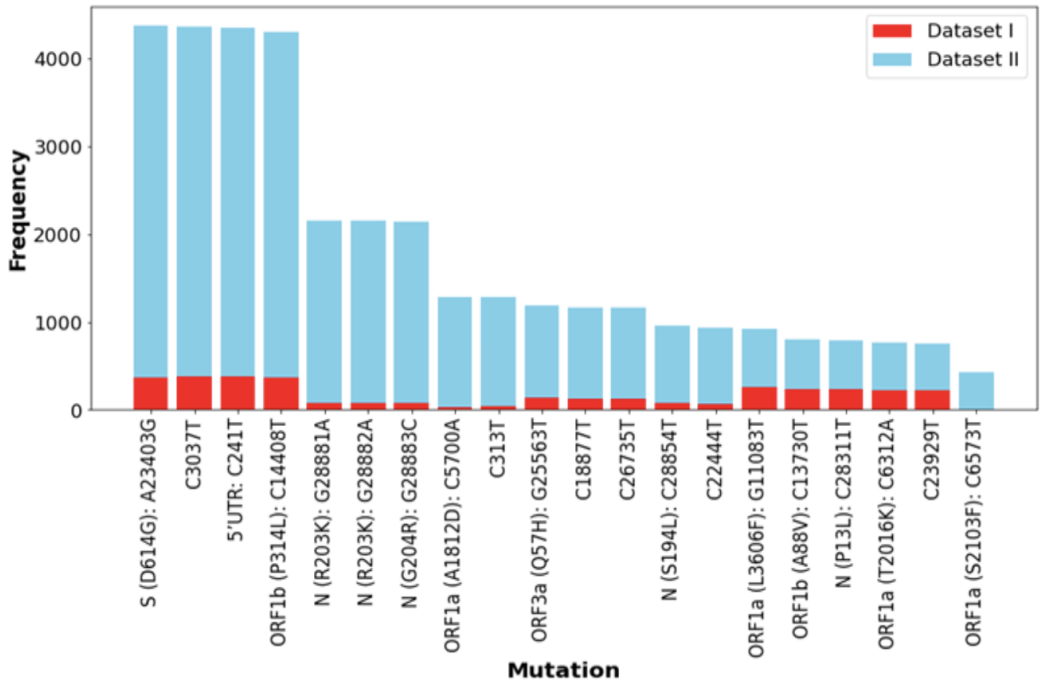


Figure 3.3: Top 20 most frequent mutations during the early phase (red) and after the first wave (blue) in Indian samples shown.

In accordance with [1] it continued to be the most common mutation, occurring in 4002/4708 samples in Dataset-II with predominance in Maharashtra (1238/1299), Telangana (861/970), and Gujarat (619/655) states. Three other mutations corresponding to clade 20A (Figure 3.4), C14408T (ORF1b: P314L, RdRP: P323L), C241T (5' UTR of ORF1ab), and C3037T are observed to co-occur with D614G. According to the predictions of SIFT and PROVEAN, mutations D614G and P314L may not severely affect the protein function (Table 3.1).

The clade 20B (Figure 3.5) tri-bloc mutation, G28881A (R203K, N protein), G28882A (R203K, N protein), and G28883C (G204R, N protein), observed in 75 samples in the early phase became the second most prevalent mutation set after the first wave. Observed in over 2000 samples ($\sim$40%), primarily from Maharashtra (914/1299) and Telangana (720/970), clade 20B was the second most dominant clade after clade 20A. The corresponding amino acid mutations R203K and G204R result in the insertion of a lysine residue in the SR-rich region of nucleocapsid protein, which is involved in viral capsid formulation [26]. Both R203K and G204R are predicted to be neutral by PROVEAN, while SIFT indicates that R203K will be tolerated while G204R may affect protein function.

In Dataset-I, about one-third of the isolates shared the mutations, viz., C6312A (ORF1a: T2016K), C13730T (ORF1b: A88V), C23929T, C28311T (N: P13L) and G11083T (ORF1a: L3606F), branching

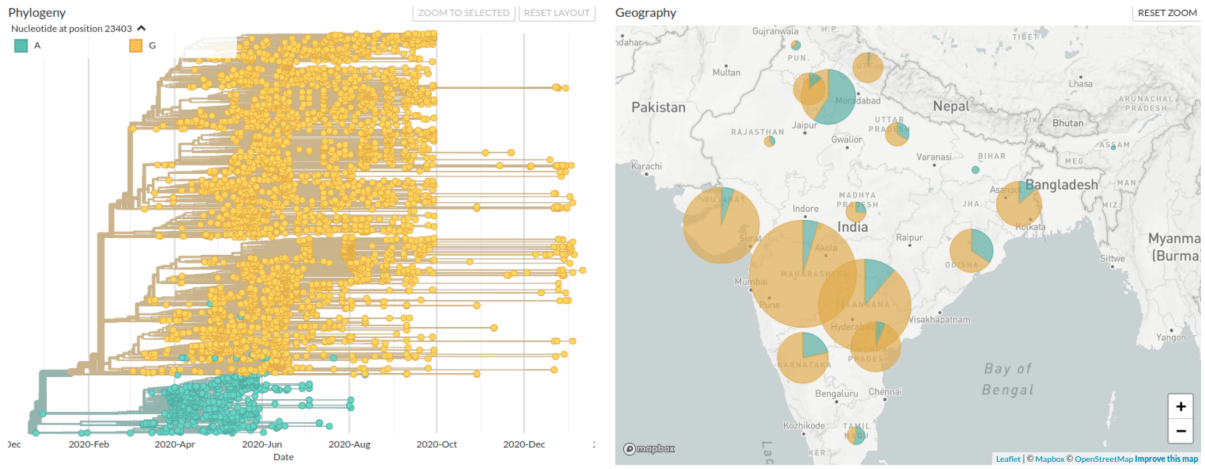


Figure 3.4: Clade 20A is highlighted in yellow on the phylogenetic tree and the geographical map of India.

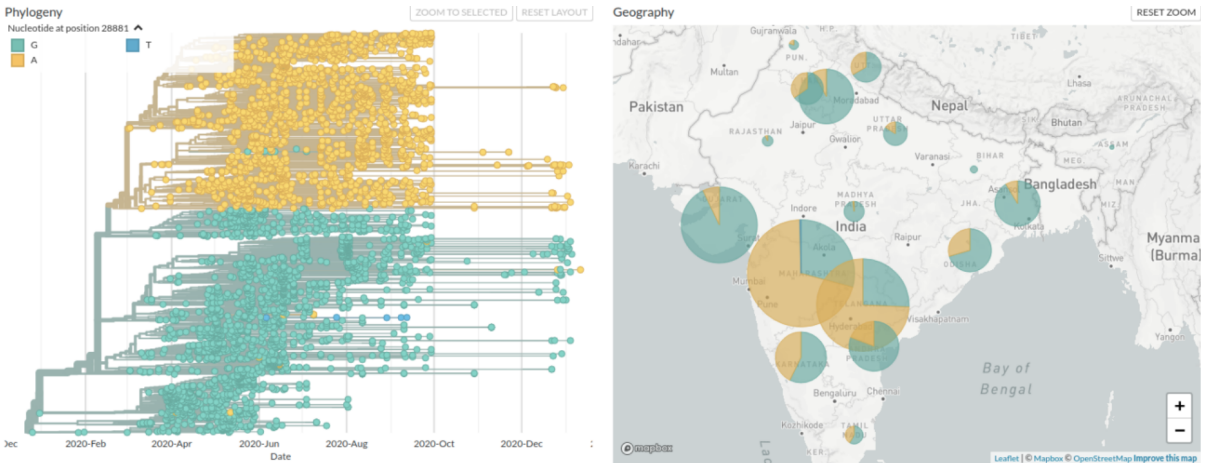


Figure 3.5: Clade 20B is highlighted in yellow on the phylogenetic tree and the geographical map of India.

out of clade 19A. According to SIFT prediction, the non-synonymous mutations L3606F, A88V, and P13L may affect the protein function, while T2016K may be tolerated. In contrast, PROVEAN predicted all the non-synonymous mutations to be neutral, however, the predicted score for A88V being close to the cutoff suggests that it may have some impact on protein function. Since globally the frequency of these 5 mutations is low ($\sim$3.5%), Banu et al. (2020) [15] defined these as India-specific subclade I/A3i (Figure 3.6). Followed by country-wide lockdown, contact tracing, and quarantine, their numbers reduced to $\sim$10.5% after the first wave. Thus, based on genetic analysis one can identify the chain of transmission of a variant strain and the success of measures for its containment.

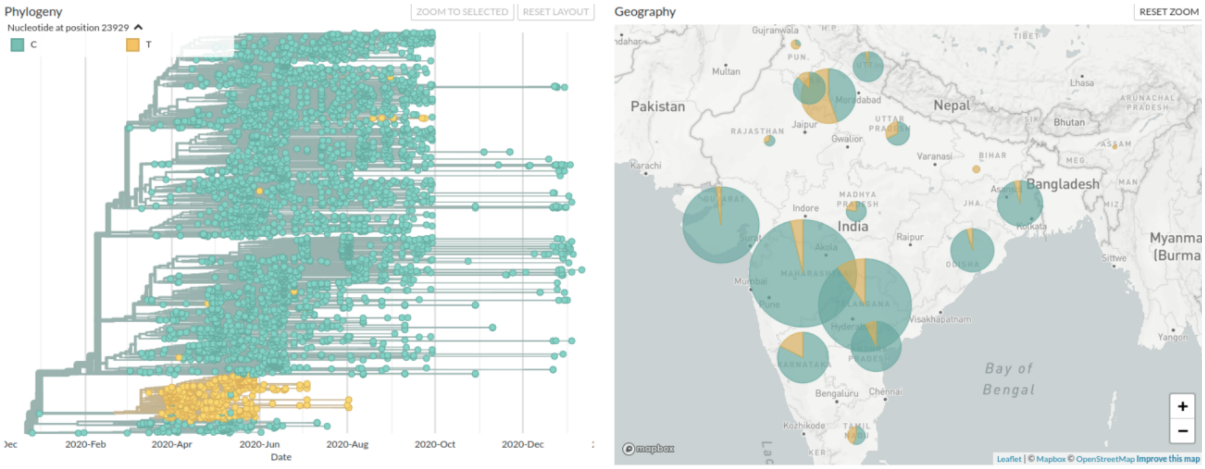


Figure 3.6: Subclade I/A3i is highlighted in yellow on the phylogenetic tree and the geographical map of India.

3.3.2.1 Types of Mutations

Analyzing the distribution of mutations acquired in the viral genomes of Indian SARS-CoV-2 isolates in the first year of the pandemic revealed that the longest protein ORF1a carries the highest number of mutations, followed by ORF1b and Spike protein (Figure 3.7). Mutations in 3' UTR are significantly higher than in 5' UTR. From Figure 3.8, it is observed that transition C>T (1990) intervened by cytosine deaminases and the transversion G>T (1135) introduced by oxo-guanine from reactive oxygen species are the most frequent mutations in the viral genome.

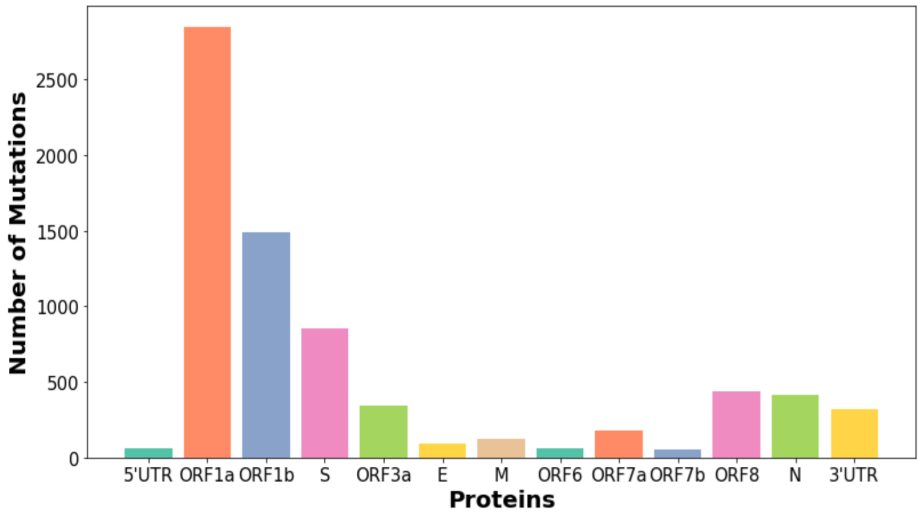


Figure 3.7: Distribution of mutations acquired in Indian SARS-CoV-2 genomes.

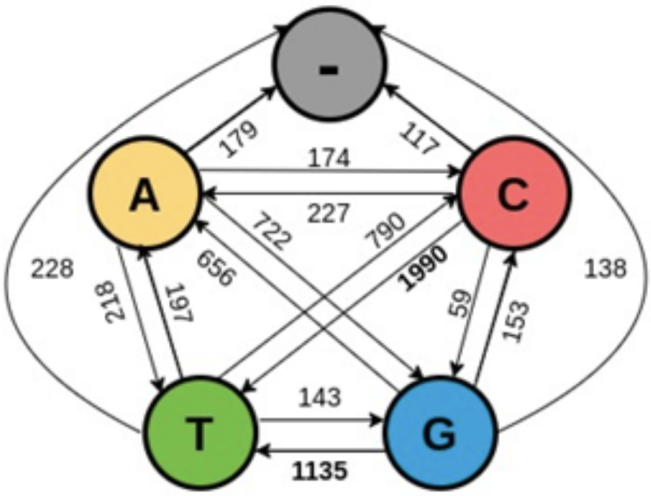


Figure 3.8: Frequency of nucleotide mutations observed in Indian SARS-CoV-2 genomes.

3.4 Principal Component Analysis

Since travel between states was restricted due to the lockdown, an increase in the number of samples with shared mutations is likely because of local community transmission. To gain insight into state-specific clustering, during the initial period and propagation of shared mutations over the first year of the pandemic, principal component analysis (PCA) was performed on the mutational profile comprising 7126 mutations in 4708 isolates (Dataset-II) shown in Figure 3.9. Mutations sorted based on their loading scores to assess their impact on PC1 identified 8 out of the top 12 most common mutations in Indian isolates after the first wave: G28881A, G28882A, G28883C, C313T, C5700A, G25563T, C26735T, and C18877T. The remaining 4 mutations (A23403G, C14408T, C241T, C3037T) occur in more than 4000 samples and hence divide the dataset with lower entropy.

Similar to the early phase, a separate 'pink' cluster is observed. Analysis of this cluster revealed that the samples correspond to the previously identified Gujarat-specific subclade I/GJ-20A. In Maharashtra, the cluster corresponding to mutations C313T and C5700A is enriched. However, that of A29837T and G29830T mutations (seen in the early phase) reduced from nearly 55% and 73.8% to 3.5% and 6.9% respectively. Other significant clusters correspond to the states of Telangana and Andhra Pradesh (discussed later in Sections 3.5.3 and 3.5.4 respectively).

Since some of the mutations are clade-defining mutations (in Nextstrain) PCA results in three distinctly observable clusters: "Top-Left" (clades 19A and 19B), "Bottom-Left" (clades 20A, 20C, and 20E (EU1)), and "Right" (20B and 20I/501Y.V1). From Figure 3.10 it is observed that the Top-Left cluster has a major representation from Delhi and Telangana, the "Bottom-Left" cluster contains sam-

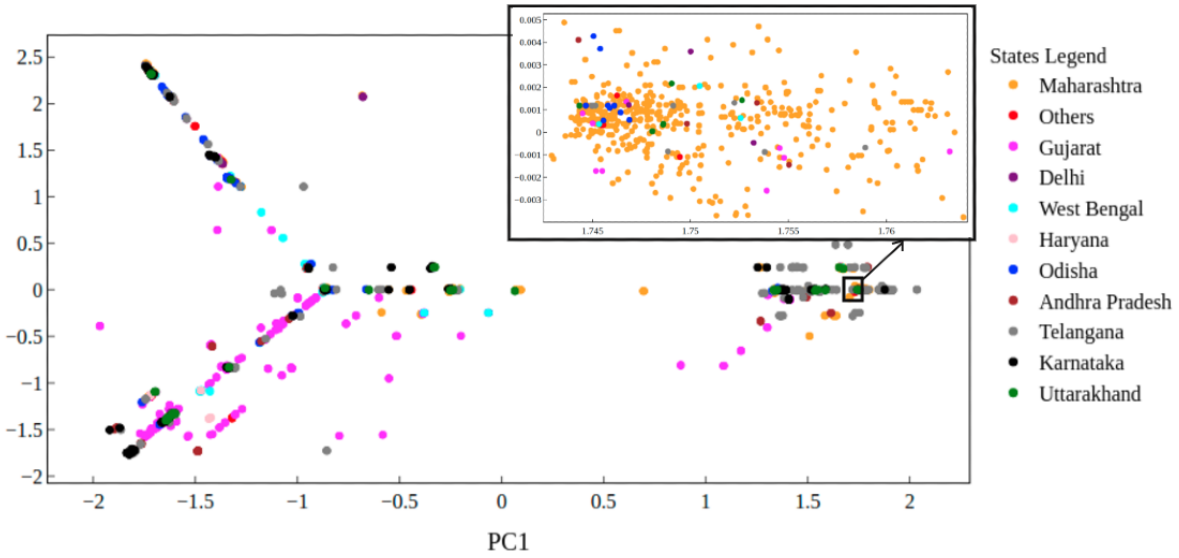


Figure 3.9: PCA plot for 4708 samples (Dataset-II) colored state-wise.

ples from Gujarat, Maharashtra, Andhra Pradesh, and West Bengal, while the "Right" cluster contains isolates from Maharashtra and Telangana. This suggests that there is a distinct distribution of clades across different states which are responsible for state-specific clusters observed in the PCA plots.

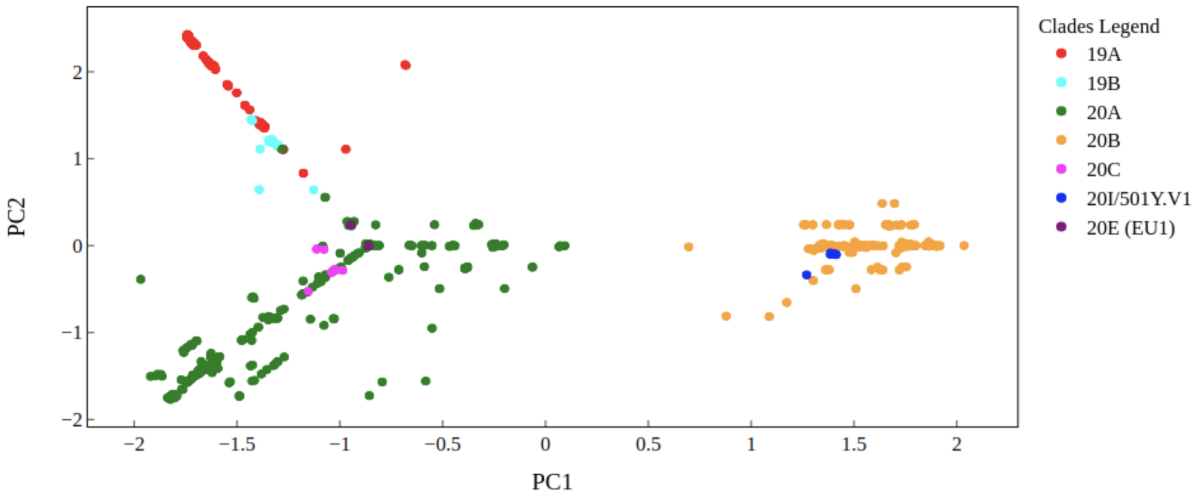


Figure 3.10: PCA plot for 4708 samples (Dataset II) colored by clades defined by Nextstrain.

3.5 Novel subclades

By the end of the first year of the pandemic, mortality rates in all states fell significantly with the country average dropping to 1.44%. With the highest fatality rate of 2.54%, Maharashtra also recorded the maximum number of cases (1971552), while Gujarat despite having much fewer cases (252559), had the second highest fatality rate of 1.72% after the first wave. Telangana reported the lowest mortality rate of 0.54%. To understand this significantly large difference in mortality rates at the genetic level, a detailed analysis of the mutational profile of sequences from these states was carried out.

3.5.1 Gujarat Analysis

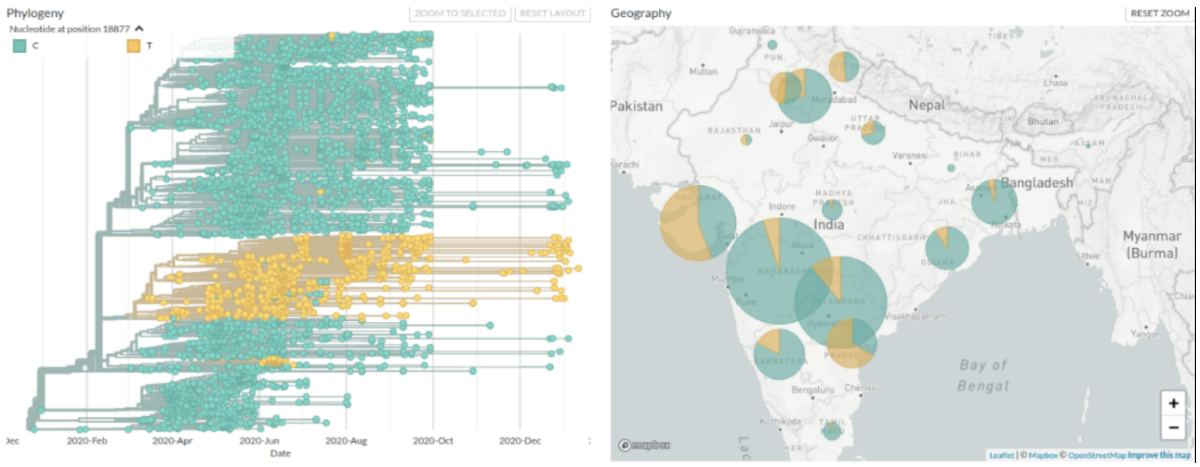


Figure 3.11: Subclade I/GJ-20A marked in yellow on the phylogenetic tree.

To understand the dense clustering of Gujarat samples in PCA plots and the significantly large difference in the percentage of deaths compared to the country average, we compared the mutational profile of Gujarat isolates with those from the Rest of India (RoI). A common set of shared mutations was identified in Gujarat isolates, viz., G25563T (ORF3a: Q57H), C26735T, C18877T, C28854T (N: S194L), C22444T, and C2836T (Figure 3.11). We refer to it as Gujarat-specific subclade I/GJ-20A (discussed in Section 2.5.1). From Table 3.2, during the first wave I/GJ-20A defining mutations continued to dominate in Gujarat ($\sim$56%) and spread to other states ($\sim$16%) after the lockdown was lifted in agreement with an earlier study [67]. Noticeably, C2836T mutation was observed in 42% samples from Gujarat but its representation in RoI was $<$1%. Most mutations that were under-represented in Gujarat isolates in Dataset-I continued to be so in Dataset-II, namely, subclade I/A3i mutations, tri-bloc mutation, and the C313T, C5700A, and G11083T mutations (Table 3.3). Both SIFT and PROVEAN predictions also indicate Q57H mutation has a significant impact on ORF3a protein. The other non-synonymous mutation C28854T (N: S194L) results in an altered structure of the nucleocapsid protein. Based on the

PROVEAN score the mutation is likely to be deleterious, while SIFT score being equal to the threshold score predicts a probable impact on the protein function.

3.5.2 Maharashtra Analysis

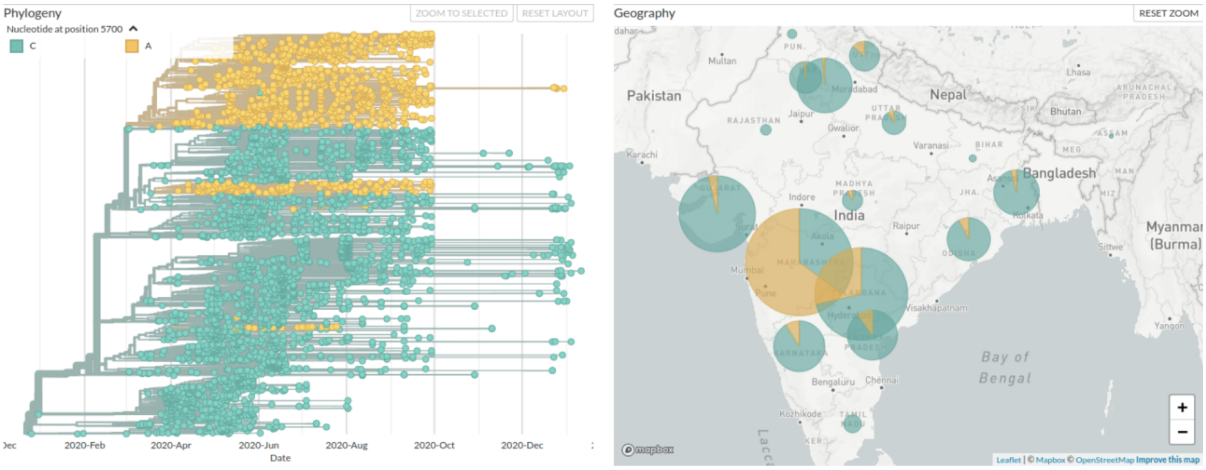


Figure 3.12: Subclade I/MH2 marked in yellow on the phylogenetic tree.

A similar analysis of Maharashtra isolates with that of RoI showed noticeable differences at the genetic level. As mentioned previously, before the first wave, two co-occurring mutations A29837T and G29830T (in 3'UTR of the viral RNA), with an incidence of 55% and 74% respectively in Maharashtra isolates, were found in only 2 samples outside the state. Interestingly, the occurrence of these mutations reduced to 3.5% and 6.9% respectively after the first wave, hinting at the success of lockdown in containing the spread of the virus. Another set of co-occurring mutations, C313T and C5700A (ORF1a: A1812D) were identified with frequencies 25% and 26.3% respectively in Maharashtra, but with $<$3% in isolates from RoI in Dataset-I. By Jan '21 (Dataset-II), this set of co-occurring mutations was over-represented (63.12%) in Maharashtra but under-represented in RoI (10.94%). The increased frequency in the entire country (26.6%) indicates its spread to different states after the lockdown was lifted. They form a distinct subclade I/MH-2 in the phylogenetic tree (Figure 3.12). This mutation set has also been reported by a study on sequences from western India [103]. SIFT and PROVEAN analysis however indicates that the A1812D mutation may not severely affect the biological function of the ORF1a protein.

3.5.3 Telangana Analysis

Two independent subclades emerged from clade 20B in Telangana isolates after the lockdown was lifted, viz., I/TelA-20B (G4354A, C6573T (ORF1a: S2103F), C25528T (ORF3a: L46F)), and I/Tel-B-20B (C9693T (ORF1a: A3143V), C16626T, A4372G, G29474T (N: D401Y)) (Table 3.4). Subclade

Table 3.2: Mutations in Gujarat (GJ) samples with a higher frequency than in Rest of India (RoI) during early phase (Dataset-I) and after first wave (Dataset-II) are given. No. of Gujarat samples: 201 in Dataset-I, 655 in Dataset-II. Rest of India samples: 484 in Dataset-I, 4053 in Dataset-II.

Nucleotide Mutation	**Amino Acid Mutation**	**Dataset-I**		**Dataset-II**	
		Count in GJ	**Count in RoI**	**Count in GJ**	**Count in RoI**
C3037T	-	182 (90.6%)	192 (39.7%)	621 (94.8%)	3363 (83.0%)
A23403G	S: D614G	180 (89.6%)	193 (39.9%)	619 (94.5%)	3383 (83.5%)
C241T	-	183 (91.0%)	195 (40.3%)	617 (94.2%)	3356 (82.8%)
C14408T	ORF1b: P314L	178 (88.6%)	189 (39.0%)	581 (88.7%)	3360 (82.9%)
G25563T	ORF3a: Q57H	108 (53.7%)	26 (5.4%)	369 (56.3%)	688 (17.0%)
C26735T	-	101 (50.2%)	22 (4.5%)	367 (56.0%)	672 (16.6%)
C18877T	-	105 (52.2%)	21 (4.3%)	366 (55.9%)	677 (16.7%)
C22444T	-	64 (31.8%)	7 (1.4%)	293 (44.7%)	566 (14.0%)
C28854T	N: S194L	66 (32.8%)	7 (1.4%)	284 (43.4%)	603 (14.9%)
C2836T	-	51 (25.4%)	0 (0.0%)	274 (41.8%)	40 (1.0%)

Table 3.3: Mutations under-represented in Gujarat (655 isolates) but are well represented in the Rest of India (RoI) (4053 samples) in Dataset II.

Nucleotide Mutations	**Amino Acid Mutations**	**Count in Gujarat**	**Count in RoI**
C313T	-	21 (3.2%)	1221 (30.1%)
C5700A	ORF1a: A1812D	24 (3.7%)	1229 (30.3%)
G11083T	ORF1a: L3606F	19 (2.9%)	643 (15.9%)
C6312A, C13730T, C23929T, C28311T	ORF1a: T2016K, ORF1b: A88V, N: P13L	12 (1.8%)	483 (11.9%)
G28881A, G28882A, G28883C	N: R203K, G204R	38 (5.8%)	2026 (50.0%)

Table 3.4: Telangana-specific mutations in Dataset II are summarized. Total Telangana (Tel) samples: 970 and Rest of India (RoI): 3738.

Nucleotide Mutation	**Protein**	**Amino acid Mutation**	**Count in Telangana**	**Count in RoI**
G4354A	-	-	404 (41.6%)	18 (0.5%)
A4372G	-	-	124 (12.8%)	17 (0.4%)
C6573T	ORF1a	S2103F	406 (41.9%)	20 (0.5%)
C9693T	ORF1a	A3143V	261 (26.9%)	30 (0.8%)
C16626T	-	-	250 (25.8%)	28 (0.8%)
A21550C	ORF1b	N2695L	116 (12.0%)	9 (0.2%)
A21551T	ORF1b	N2695L	113 (11.6%)	10 (0.3%)
C25528T	ORF3a	L46F	401 (41.3%)	17 (0.4%)
G29474T	N	D401Y	92 (9.5%)	16 (0.4%)

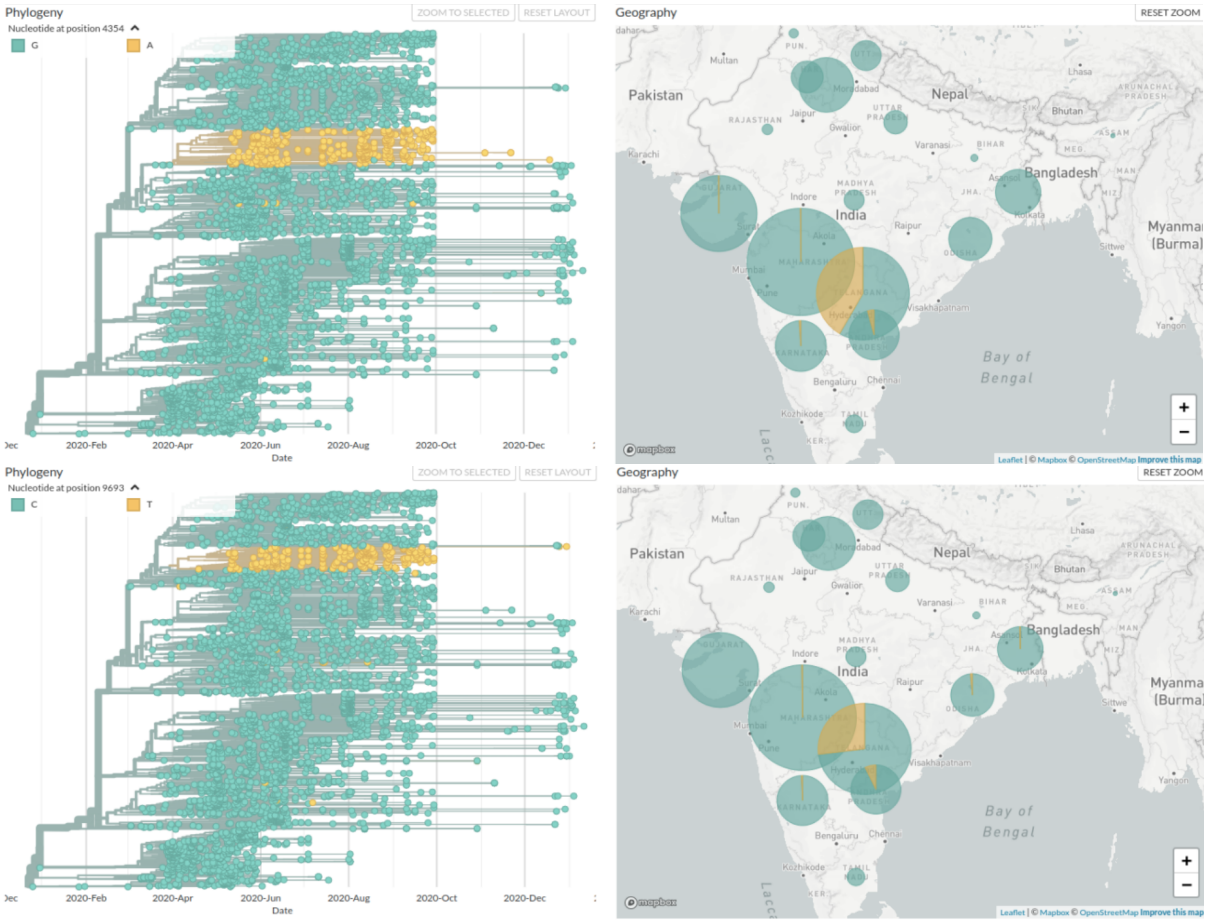


Figure 3.13: Telangana-specific subclades shown in yellow branching out on the phylogenetic tree: (a) I/Tel-A-20B, and (b) I/Tel-B-20B.

I/Tel-A-20B (Figure 3.13 (a)) defining mutations were found in 393 Telangana samples (40.52%) but in only 17 (0.5%) samples from the rest of the country. Similarly, I/Tel-B-20B (Figure 3.13 (b)) had over 25% prevalence in the state but in $<$1% samples from the rest of the country. Apart from these, two adjacent co-occurring mutations A21550C and A21551T resulting in amino acid change N2695L in ORF1b were also found to be Telangana-specific (but not subset any clade) with more than 11% representation in the state but 0.2% in RoI, in accordance with earlier work [56]. This clearly indicates local community transmission within the state. SIFT analysis indicates that the non-synonymous mutations, S2103F, L46F, D401Y, and N2695L may affect the protein function, while PROVEAN score indicates the mutation L46F to be deleterious. According to SIFT mutation D401Y of subclade I/Tel-B-20B, and mutation N2695L may affect the protein function.

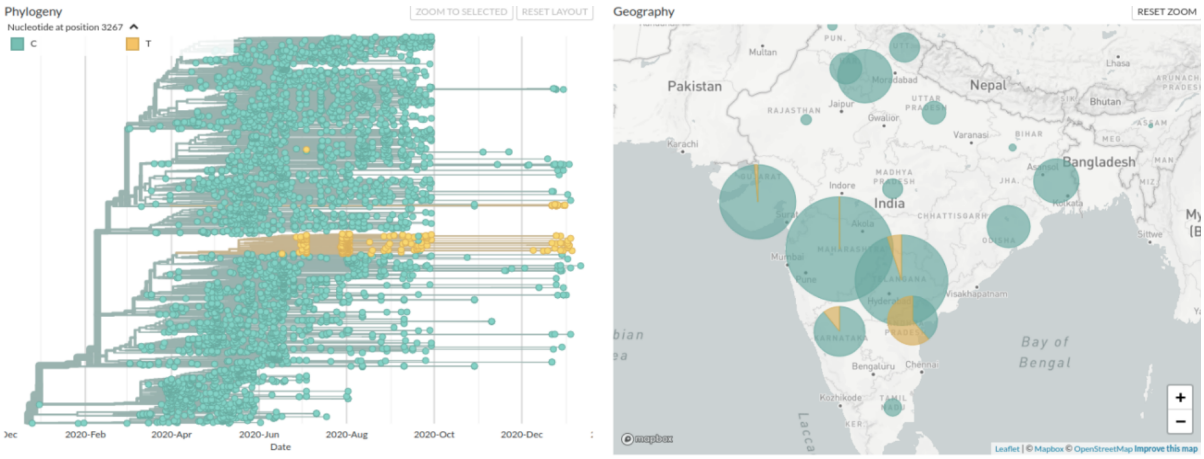


Figure 3.14: Subclade I/AP-20A shown in yellow on the phylogenetic tree.

3.5.4 Andhra Pradesh Analysis

Isolates from Andhra Pradesh (AP) revealed a state-specific subclade, I/AP-20A, defined by the mutations C3267T (ORF1a: T1001I), C21034T (ORF1b: L2523F), G26173T (ORF3a: E261del), G28183T (ORF8: S97I) and T28277C (N: S2P) in Dataset-II (Table 3.5). These mutations co-occur and branch out of clade 20A in the phylogenetic tree (Figure 3.14). It is observed in nearly 53% of AP isolates, with a small presence in the neighboring states of Telangana (4.6%) and Karnataka (6.9%). The mutation G28183T is predicted to be deleterious according to PROVEAN, while mutation T28277C is predicted to affect protein function according to SIFT.

Apart from these four states, our analysis of Dataset-II revealed noticeable region-specific variations. For example, analysis of 292 sequences from Karnataka revealed 3 mutations, C1218T (ORF1a: S318L), C27110T, and T27384C with $>$20% prevalence in the state and $<$1% outside it. Similarly, differences were observed in sequences from West Bengal (G15451A: ORF1ab G662S) and Odisha (A25381C) with close to 9% prevalence in the respective states and nearly negligible presence in RoI.

3.5.5 Correlation Analysis

To assess the correlation between the mutations observed in Indian SARS-CoV-2 isolates, we computed pairwise Pearson correlation coefficients (PCC) between 29 mutations occurring with $>$5% frequency. This resulted in 67 pairs with an absolute PCC value $>$0.5. As expected, the strongest positive correlation was observed between the tri-bloc mutations; G28882A-G28883C (0.996), G28881A-G28883C (0.994), and G28881A-G28882A (0.993). A strong correlation with a mean of 0.990 was also observed for the four clade 20A defining mutations (C241T, C3037T, C14408T, A23403G). The mutations corresponding to subclade I/A3i, viz., C6312A, C11083T, C13730T, and C23929T, exhibited a strong correlation amongst themselves (mean = 0.955) and also with the co-occurring mutation

Table 3.5: Frequencies of Andhra Pradesh (AP) specific mutations are compared with that in the rest of India (RoI) in Dataset II. Total samples in AP: 281, and RoI: 4427.

Nucleotide Mutation	Protein	Amino acid Mutation	Count in AP	Count in RoI
C3267T	ORF1a	T1001I	172 (61.2%)	89 (2.0%)
C21034T	ORF1b	L2523F	170 (60.5%)	72 (1.6%)
G26173T	ORF3a	E261del	170 (60.5%)	70 (1.6%)
G28183T	ORF8	S97I	170 (60.5%)	72 (1.6%)
T28277C	N	S2P	150 (53.4%)	70 (1.6%)

G11083T (mean = 0.807). A strong negative correlation (mean = -0.838) is obtained between the mutations corresponding to clade 20A and those corresponding to subclade I/A3i and G11083T (which are part of clade 19A). C313T and C5700 of subclade I/MH-2 exhibit a strong correlation (0.965) and are moderately correlated with the tri-bloc mutations (mean = 0.662). It is observed that state-specific subclade-defining mutation sets exhibit a strong correlation among themselves: I/GJ-20A (C18877T, C22444T, G25563T, C26735T, C28854T) (mean = 0.897), I/AP-20A (C3267T, C21034T, G26173T, G28183T) (mean = 0.972), I/Tel-A-20B (G4354A, C6573T, C25528T) (mean = 0.982) and I/Tel-B-20B (C9693T, C16626T) (mean = 0.963), confirming their co-occurrence and probably co-dependence. It would also be interesting to investigate the joint impact of these co-occurring mutations on the infectivity or transmissibility of the region-specific isolates.

3.5.6 Prevalence in Second and Third Waves

It is observed that over the past two years SARS-CoV-2 has evolved in humans, exploring the sequence space, and resulting in the selection of variants with improved replication efficiency and transmissibility. Here we discuss the India-specific variants observed in the first year of the pandemic (Datasets I and II) that are still circulating in the country (Dataset-III and IV). It may be noted from Table 3.1 that clade 20A defining mutations continue to dominate with more than 97% prevalence in the country till date, while the frequency of novel India-specific subclades identified after the first wave is significantly reduced, e.g., the frequency of I/GJ-20A and I/MH-2 subclades reduced from 56% and 26.6% respectively after first wave (Dataset-II) to 5% and 2% after third wave (Dataset-IV). Similar analysis of Clade 20B defining mutations indicate that its frequency decreased from nearly 44% after the first wave to 15% after the second wave and then increased again to 23% after the third wave. No new incidence of India-specific subclade I/A3i is observed after the early phase (Dataset-I) and its cumulative presence is $\leq$1% after the second and third waves. However, one of its characteristic muta-

tions, C28311T (P13L), is observed in over 16% of samples after the third wave. The mutation has been reintroduced in the Omicron variant, leading to its increased incidence after the third wave. A similar decreasing trend is observed for state-specific subclades I/Tel-A-20B, I/Tel-B-20B, and I/AP-20A after the second and third waves. Thus, the state-specific mutational analysis indicates that during the lockdown some variants were confined by state boundaries, and these grew in number after the lockdown was lifted, due to local transmission. Limited travel between states resulted in their numbers being restricted majorly to the respective states after the first wave. With the advancement of the pandemic and the introduction of more transmissible variants on the opening of international borders, loss of region-specific mutations is observed. The introduction and spread of new lineages of SARS-CoV-2 after the first wave of the pandemic in India are discussed below.

3.6 Pangolin Lineage Analysis

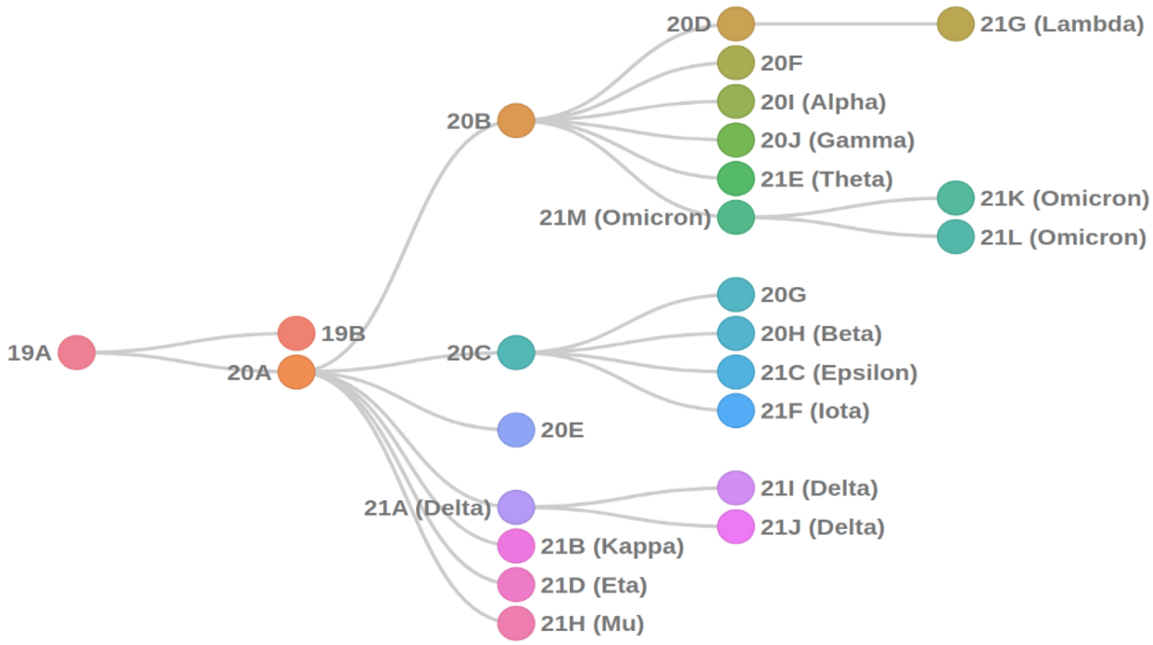


Figure 3.15: The association between the two nomenclatures, Pangolin, and Nextstrain Clade for SARS-CoV-2.

GISAID categorized the virus into clades during the first year based on unique mutations identified at different positions in the genome. Since SARS-CoV-2 has been evolving very quickly, a dynamic nomenclature based on a phylogenetic framework (Figure 3.15) is employed to designate lineages with an active spread, known as Pangolin [114]. Lineages that pose an increased risk to global public health have been termed 'Variants of Interest' (VOI), or 'Variants of Concern' (VOC) (represented using Greek

alphabets) by the World Health Organization (WHO), based on the acquired mutations in the spike protein receptor binding domain that resulted in a substantial increase in its binding affinity with human ACE2 protein and linked to rapid spread in the population. The early phase of the pandemic in India (till May '2020) saw only two lineages, B.6 (Clade 19A) and B.1 (Clade 20A), with 30% and 25% incidence, respectively. After the first wave, lineage B.1.1.306 (Clade 20B) became dominant with close to 27% prevalence followed by B.1 (14%) and B.6 (10%).

3.6.1 Second Wave

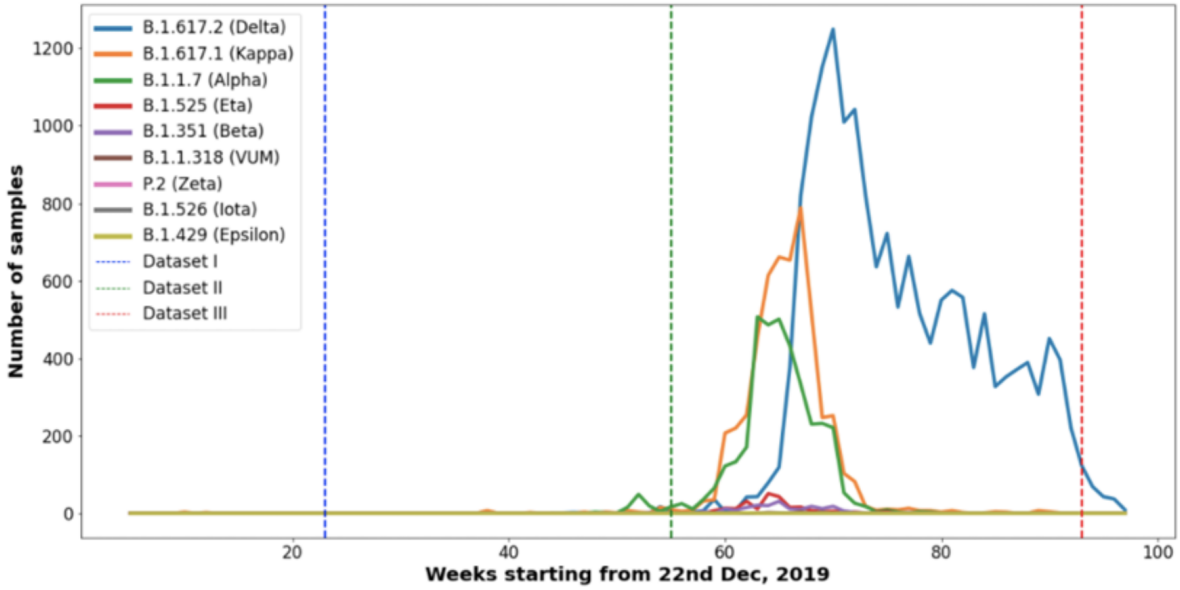


Figure 3.16: Week-wise evolution of all the VOCs, VUMs, and former VOIs observed in Indian isolates. Week 0 corresponds to 22nd - 28th Dec 2019 and Week 97 to 25th - 31st October 2021. Dashed vertical lines correspond to the time periods for which data was collected. The data for analysis is obtained from CoV-GLUE (http://cov-glue.cvr.gla.ac.uk/)

When the world was experiencing huge numbers of COVID-19 cases during the summer of 2020, India luckily missed the first wave. Timely screening of international travelers, quarantine, and contact tracing, followed by a country-wide lockdown on 24th Mar '2020 was helpful. During this period, UK along with other European nations, underwent a major first wave, sparked primarily by clade 20E (EU1), lineage B.1.177 that first emerged in Spain in early summer. It is characterized by the mutations C22227T (M: L93L), C28932T (ORF10: V30L), and G29645T (S: A222V) that exhibit no increased transmissibility. Its rapid spread is primarily attributed to the resumption of travel across Europe coupled with a lack of effective screening and containment measures. The spread of lineage B.1.177 across

Europe is a classic example of how a variant without any selective advantage can become dominant if proper screening and containment strategies are not implemented on time. The USA, like India, had a slow start, but by late 2020, the spread of the virus increased with the variant Epsilon (B.1.427, B.1.429) or clade 21C. It first emerged in California, USA, and by the end of the first wave in Jan '2021, nearly 50% cases were affected by the Epsilon variant. It is characterized by spike protein mutations: D614G, E484K, L452R, S13I, and W152C. In India, the first wave peaked around Sept '2020, and the second wave around Mar-Apr '2021. Since the second wave was very severe with very high infectivity and fatality, analysis of new emerging variants between the first and second waves is important in the Indian context [122]. We observe that the Delta variant (B.1.617.2, clade 21A) emerged by Oct '2020 in India along with the Kappa (B.1.617.1, clade 21B) and Alpha (B.1.1.7, clade 20I) variants, while the Alpha variant was still dominant in both the UK and USA. Figure 3.16 shows the evolution of all the important VOC/VOI/VUM present in the Indian samples across Datasets I, II, and III, wherein, Delta, Kappa, and Alpha variants can be seen dominating in the Indian samples after the first wave.

3.6.1.1 Delta variant

Delta variant was first detected in India in late 2020, swept rapidly through the country, and reached UK and then USA where it became the dominant variant, accounting for 99% of COVID-19 cases and a large number of hospitalizations. It is thought to be twice as contagious with viral loads nearly 1000 times more compared to the previous variants, resulting in enhanced severity, and was flagged as a Variant of Concern (VOC) by WHO. Delta variant is characterized by the spike mutations: T19R, del157/158 (in N-terminal domain of RBD), L452R, T478K (in RBD), D614G, P681R, D950N (in S2 region). Spike mutation P681R, being beside the furin cleavage site, is responsible for the efficient clipping of spike protein and enabling the virus to invade human cells more efficiently, thereby enhancing its infectivity. Mutation L452R, located in the receptor binding domain region (also present in the Epsilon variant), increases the binding affinity of spike protein to the human ACE2 receptor, again leading to increased infectivity. In silico analysis revealed that the substitution of a non-charged Threonine with positively charged Lysine in mutation T478K may be responsible for altering the electrostatic surface of the protein, which in turn may enhance the binding affinity of RBD to ACE2 and the ability of the virus to invade the host cell. These three key mutations, L452R, T478K, and P681R, are responsible for the increased transmissibility and generated immune escape of the Delta variant [133]. The D950N mutation is different than other mutations because of its location outside RBD in an area that helps the virus fuse with human cells, affecting the types of cells (i.e., organs, tissues) the virus infects and the viral load. Further, it is shown by [107] that the vaccine effectiveness is notably lower for the Delta variant compared to the Alpha variant. This may be because of mutations T19R and del157/158 in the N-terminal domain which provides a 'supersite' for antibodies to latch to the virus, making monoclonal antibodies less effective in treating COVID and increases the Delta variant's ability to escape vaccine-generated antibodies. We assessed the functional impact of these mutations using SIFT and PROVEAN and the results are summarized in Table 3.6. Both SIFT and PROVEAN scores indicate that three key

mutations do not have a deleterious effect on spike protein, except for mutation T478K which according to PROVEAN may affect protein function. According to SIFT, mutation D950N is likely to affect protein function, while PROVEAN returned a high negative score, suggesting the mutation may not be deleterious. The D950N spike protein mutation is observed in the Mu variant (B.1.621) and the mutation T478K in the Omicron variant (BA.1, BA.2). These results explain the Delta variant's increased transmissibility and severity, which led to the sudden spike in cases and a near breakdown of the healthcare system during the second wave in India and other countries.

3.6.1.2 Kappa variant

Another prevalent variant during the second wave in India was the Kappa variant (9%), also called 'double mutant' and later designated as lineage B.1.617.1 and a variant of interest (VOI) by WHO. First observed in Maharashtra in Mar '2021, it is characterized by the spike mutations: E154K, E484Q, L452R, D614G, P681R, and Q1071H. By April 2021, the Kappa variant accounted for nearly 35% of all sequenced cases in India and coincided with the rise in daily COVID-19 cases in India. It is an evolutionary ancestor of the Delta variant, with shared mutations E484Q, L452R, D614G, and P681R that provide it with increased transmissibility. However, the variant's impact on severity has not been proven (Table 3.7).

3.6.1.3 Alpha variant

The second wave also saw the introduction of a de-escalated VOI, the "Alpha" variant (7%), lineage B.1.1.7 (clade 20B/501Y.V1). It was first observed in UK in Dec '2020 and has been well characterized for both increased transmissibility [34] and increased severity [52]. During the second wave, it was observed that the northern part of India had a higher dominance of Alpha, while in the southern and central parts of India, Delta, and Kappa variants were rampant [147, 57]. Alpha variant is characterized by the spike mutations: A570D, D614G, D1118H, H69del, N501Y, P681H, S982A, T716I, V70del, and Y145del. Of these, the most prominent mutation is N501Y with tyrosine substitution in RBD allowing additional interaction with human ACE2 at residue 353 and improving binding affinity. This provides the Alpha variant with increased transmissibility. Two other mutations, H69-V70del and P681H, are also shown to have significant potential in influencing the biological characteristics of the virus.

Apart from these highly infectious variants circulating, some of the factors that fueled the second wave in the country are complacency by the government and the public after the first wave, super spreader events such as Kumbh mela, assembly elections, weddings and celebrations, and insufficient vaccination (0.3% of the population with single dose).

Table 3.6: Shared mutations between Pangolin lineages (classified as VOCs/VUMs/VOIs) are listed along with SIFT and PROVEAN scores for functional relevance. Here, P - presence of the mutation; A: Alpha (B.1.1.7), B: Beta (B.1.351), K: Kappa (B.1.617.1), D: Delta (B.1.617.2), O1: Omicron (BA.1), O2: Omicron (BA.2), G: Gamma (P.1), E: Eta (B.1.525), I: Iota (B.1.526), L: Lambda (C.37), M: Mu (B.1.621).

Mutation	**SIFT**	**PROVEAN**	**A**	**B**	**K**	**D**	**O2**	**Other Lineages**
L18F	0.23 (T)	0.78 (N)	-	P	-	-	-	G
H69-	-	0.26 (N)	P	-	-	-	-	O1, E
V70-	-	0.26 (N)	P	-	-	-	-	O1, E
Y144-	-	0.85 (N)	P	-	-	-	-	O1, E
G339D	0.88 (T)	-0.19 (N)	-	-	-	-	P	O1
S373P	0.38 (T)	1.21 (N)	-	-	-	-	P	O1
S375F	0.07 (T)	-0.24 (N)	-	-	-	-	P	O1
K417N	0.56 (T)	0.27 (N)	-	P	-	-	P	O1
N440K	0.7 (T)	-1.66 (N)	-	-	-	-	P	O1
L452R	0.38 (T)	0.56 (N)	-	-	P	P	-	-
S477N	0.84 (T)	-0.03 (N)	-	-	-	-	P	O1
T478K	0.75 (T)	-0.52 (N)	-	-	-	P	P	O1
E484K	0.85 (T)	0.13 (N)	-	P	-	-	-	G, E, I, M
E484A	0.51 (T)	0.73 (N)	-	-	-	-	P	O1
Q493R	0.51 (T)	-0.34 (N)	-	-	-	-	P	O1

Q498R	0.39 (T)	0.06 (N)	-	-	-	-	P	O1
N501Y	0.09 (T)	-0.09 (N)	P	P	-	-	P	G, O1, M
Y505H	0.03 (A)	-0.67 (N)	-	-	-	-	P	O1
D614G	0.62 (T)	0.60 (N)	P	P	P	P	P	G, O1, E, I, L, M
H655Y	0.5 (T)	-0.81 (N)	-	-	-	-	P	G, O1
N679K	0.53 (T)	0.25 (N)	-	-	-	-	P	O1
P681H	0.17 (T)	0.06 (N)	P	-	-	-	P	O1, M
P681R	0.33 (T)	0.74 (N)	-	-	P	P	-	-
A701V	0.4 (T)	0.60 (N)	-	P	-	-	-	I
N764K	0 (A)	-2.84 (D)	-	-	-	-	P	O1
D796Y	1 (T)	0.04 (N)	-	-	-	-	P	O1
D950N	0 (A)	-1.63 (N)	-	-	-	P	-	M
Q954H	0.08 (T)	-2.04 (N)	-	-	-	-	P	O1
N969K	0.09 (T)	-3.07 (D)	-	-	-	-	P	O1

3.6.2 Third Wave

With the end of the second wave in India, the number of deaths decreased drastically. This could also be credited to a large-scale vaccination campaign that took place post-second wave and prepared India for a third wave. The third wave in India (Dec 2021 - Feb 2022) was characterized by mild infections, very few hospitalizations, and mostly home treatments. This indicates that the variant in circulation, while having increased transmissibility, had reduced severity compared to the Delta variant. A comparison of isolates in Datasets III and IV can help in identifying differences between the circulating variants during the second and third waves. The most frequent lineages found in Dataset-IV are shown in Figure 3.17.

Table 3.7: Pangolin lineages labeled as VOC/former VOI, with their associated Nextstrain Clades, frequencies (Dataset IV), and functional impact obtained from "European Centre for Disease Prevention and Control Dashboard (https://www.ecdc.europa.eu/en/covid-19/variants-concern)" is given. The impact of variants is annotated with (v) or (m) to indicate whether the evidence is available for the variant itself (v) or for mutations associated with the variant (m). *VOC, **former VOI

Pangolin Lineage	**Nextstrain Clade**	**WHO Label**	**Frequency**	**Transmissibility**	**Immunity**	**Severity**
B.1.617.2	21A	Delta*	24761 (24.4%)	Increased (v)	Increased (v)	Increased (v)
B.1.1.529, BA.1, BA.2	21K, 21L	Omicron*	14455 (14.2%)	Increased (v)	Increased (v)	Reduced (v)
B.1.351	20H	Beta*	209 (0.2%)	Increased (v)	Increased (v)	Increased (v)
P.1	20J	Gamma*	3 (0.0%)	Increased (v)	Increased (v)	Increased (v)
B.1.617.1	21B	Kappa**	4827 (4.8%)	Increased (v)	Increased (v)	No evidence
B.1.1.7	20I	Alpha**	3559 (3.5%)	Increased (v)	Similar	Increased (v)
B.1.525	21D	Eta**	181 (0.2%)	No evidence	Increased (m)	No evidence
P.2	20B	Zeta**	2 (0.0%)	No evidence	Increased (m)	No evidence
B.1.526	21F	Iota**	1 (0.0%)	No evidence	Increased (m)	No evidence
B.1.427, B.1.429	21C	Epsilon**	1 (0.0%)	Unclear (32)	Increased (v)	No evidence

3.6.2.1 Omicron variant

In Dataset-IV, while the Delta variant (B.1.617.2) continued to be the most prevalent lineage (24%) a new variant, BA.2, which is a subset of the Omicron variant (VOC), grew in the country (11%). Various studies have discussed the increased transmissibility [86] and reduced severity [12, 125] of the Omicron variant. An interesting observation by [110] was that the Omicron variant can evade host immunity induced due to prior infection. This was not the case with earlier variants, namely, Alpha, Beta, and Delta, wherein, previous infection robustly prevented reinfection (90%), while that was not true with Omicron (60%) [4].

Omicron variant is shown to have a completely new serotype because of which a person infected with Omicron does not have protection against infections due to other variants [119]. It contains more than 30 mutations in spike protein compared to Wuhan-1 as a reference [110]. The N501Y mutation enhances binding affinity with the ACE2 receptor, which is a primary influencer of enhanced transmission, and when combined with Q498R, binding affinity increases further, allowing Omicron an easy entry into the host [125]. Mutation K417N, also seen in the Beta variant, has been related to structural changes in spike protein that may improve immune evasion [8]. It shares mutation T478K with the Delta variant which enhances RBD binding affinity and permits immunological escape. Mutations H655Y and N679K in Omicron are situated near the furin cleavage site and can accelerate spike cleavage, making the virus more infectious [61]. Mutation P681H, on the other hand, can increase transmissibility by boosting spike protein cleavage. Due to "S gene target failure", samples infected with Omicron result in a false negative RT-PCR test, which may have further added to its rapid spread [136]. PROVEAN analysis predicted the mutations N764K and N969K in Omicron to be deleterious, while according to SIFT, mutation Y505H may affect the protein function. Along with Omicron variant BA.2, AY.127 (B.1.617.2.127), Kappa variant (B.1.617.1) and AY.112 (B.1.617.2.112), each with 5% prevalence Alpha variant with 4% and BA.1 (subset of Omicron, VOC) with 3% are observed after the third wave.

3.6.2.2 Other variants

It is observed that all the Variants of Concern, namely, Delta (B.1.617.2), Beta (B.1.351), Omicron (B.1.1.529, BA.1, BA.2, BA.3), and Gamma (P.1) are observed in Indian isolates. The Beta variant is highly transmissible and more severe than all earlier variants and is observed in 157 and 209 samples in Datasets III and IV respectively. It was first detected in South Africa in late 2020 where it sparked a second wave before spreading globally. It has been designated as a Variant of Concern (VOC) by WHO. [131] showed that the Beta variant had a selective advantage either due to increased transmissibility or immune escape. Higher viral load, more severity leading to higher hospitalizations and mortality with the Beta variant are probably because of three spike mutations K417N, E484K, and N501Y. This triple mutation is likely to increase RBD-ACE2 receptor interaction and support the escape of the virus from

immune response. Studying its impact on immunity, [87] showed that even two doses of ChAdOx1 vaccine were unable to provide protection against infection by the Beta variant, and it is presumed that natural and vaccine-induced immunity may not provide protection against the Beta variant. The least prevalent VOC in Indian samples is the Gamma variant (P.1), observed in 3 samples after the second wave and no new samples after the third wave. First detected in Brazil in late 2020, it is found with increased transmissibility (but not as transmissible as Alpha/Delta) [45] and exhibits increased severity [52] and immune escape capabilities [35] probably because of shared spike mutations with Alpha and Beta variants.

Currently, there are two Variants of Interest (VOI), Lambda (C.37 alias for B.1.1.1.37) and Mu (B.1.621), and one Variant under Monitoring (VUM), B.1.1.318, observed in 11 samples from Dataset-III with 1 new incidence in Dataset-IV. WHO had formerly listed many Variants of Interest that have been removed from the list, viz., Kappa, Iota (B.1.526), Eta (B.1.525), Epsilon, Zeta (P.2 alias for B.1.1.28.2), and Theta (P.3 alias for B.1.1.28.3). Apart from Theta all other variants have been observed in India; 1 sample each of Iota and Epsilon variants in Datasets III and IV and 2 samples in Dataset-III of Zeta variant with no new incidences in Dataset-IV. The Eta variant (B.1.525) was observed in 152 samples in Dataset-III and 181 samples in Dataset-IV.

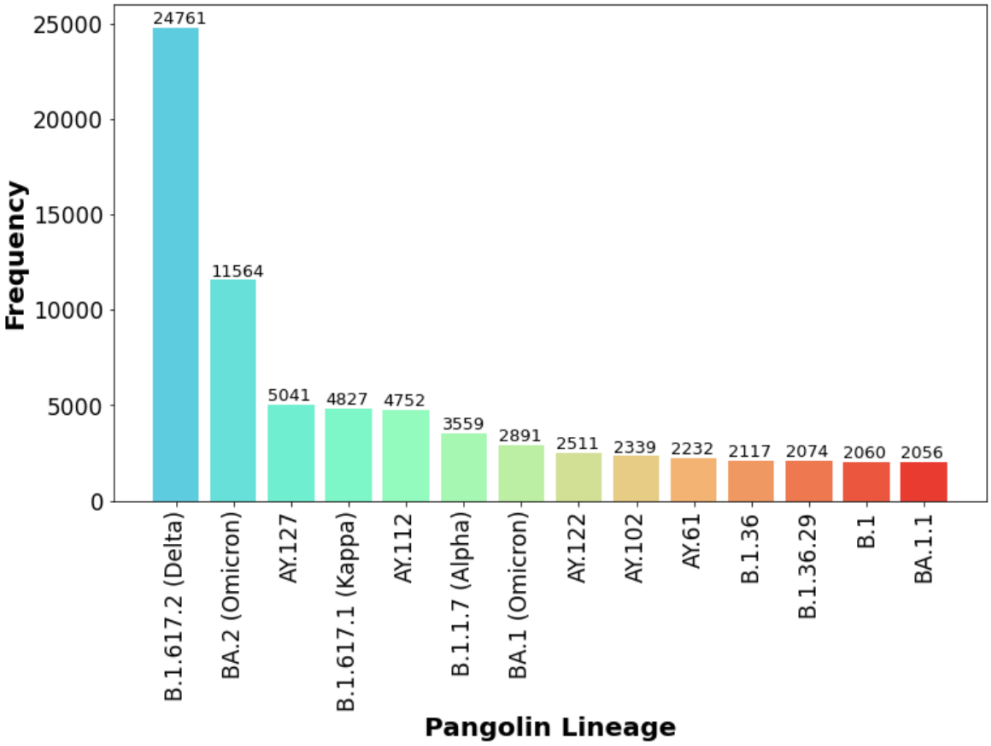


Figure 3.17: Pangolin Lineage frequencies in Dataset-IV.

3.7 Conclusion

To our knowledge this is the most comprehensive study of the mutational profiles of Indian SARS-CoV-2 isolates conducted across four phases of the pandemic (early phase, and after 1st, 2nd, and 3rd waves), considering 101527 isolates. Analysis of the first two phases revealed state-specific strains and their relevance in terms of transmissibility and severity and assessed the effectiveness of contact tracing, quarantine, and lockdown in controlling its spread after the early phase. Genetic analysis revealed that though lockdown helped in controlling the spread of the virus, region-specific set of shared mutations indicates local transmission within the states. This partially explains the observed variation in infectivity and fatality across states. The second wave was very severe in the country and led to a large number of hospitalizations and deaths, while the third wave was comparatively milder. Analysis of important mutations of Delta and Omicron variants is carried out to understand the differences between the waves, apart from immunity provided by vaccines to over 70% adult population compared to only 2% before the second wave. After first-year, international borders had opened and there were a number of new circulating variants in the country. Analysis of various lineages and VOCs/VOIs was performed to understand their spread in the country and possible impact on transmissibility and pathogenicity. Continued sequencing efforts followed by detailed data analysis efforts are required to identify potentially virulent strains and better pre-emptive action to control their spread. Such measures can prepare India for future waves of SARS-CoV-2 and other novel viruses.

Chapter 4

Protein Structural and Network Analysis for India-specific Mutations

4.1 Introduction

In the previous chapters, a number of mutations have been identified in the SARS-CoV-2 virus, some of which are specific to India. However, there is limited information available regarding the impact of these mutations on the structure and function of the virus. To address this gap, we conduct a thorough analysis of the wild-type and mutated protein structures using network and structural analysis techniques. In addition, we also utilise the available literature to gain insights into the impact of the identified mutations and the importance of specific sites in the protein. To further support our analysis, we explore various structural analysis tools such as SAS [93], Missense3D [64], and mCSM-PPI2 [118], which allow us to gain a deeper understanding of the residues that are critical for maintaining the protein structure and function. Our results provide important insights into the impacts of the India-specific mutations on the SARS-CoV-2 virus, which could be useful for developing effective therapeutic strategies against the disease.

4.2 Materials and Methods

4.2.1 Network Analysis

Mutations at key residues can impact the protein functionally by changing the interactions of other residues in the protein as well. To find the most impacted nodes due to a mutation, we compared the network structures for the wild-type protein and the mutated protein.

The first step is to get the wild-type and mutated protein structures. For wild-type proteins, structures are available, but they have many missing residues. As an example, for ORF3a (wild-type) coordinate information is not available for the terminal residues 1-39 and 239-275, and positions 175-180. For the mutant protein, no solved structure is generally available for India-specific mutations. Therefore, we decided to simulate both wild-type and mutated protein structures. We chose trRosetta [37] to simulate

protein structures using a sequence input as it was the best-performing model in the latest CASP-15 experiment. I-TASSER [149] is another frequently used model to perform structure simulation. We found trRosetta to perform better than I-TASSER across different proteins. As an example, for the D614G mutation in the spike protein, I-TASSER resulted in TM-Scores of 0.52 and 0.46 for the wild-type and mutated structures respectively. However, trRosetta was able to get TM-Scores of 0.61 and 0.59 for the simulated wild-type and mutated proteins. We performed our experiments on the trRosetta (transform-restrained Rosetta) web-server which uses deep learning and Rosetta to perform accurate de novo protein structure prediction quickly. In trRosetta, first, a deep neural network is used to predict inter-residue geometries, including distance and orientations, using the amino acid sequence as input. Then these predicted inter-residue geometries are translated into constraints to guide structure prediction based on direct energy reduction, which is done inside the Rosetta framework.

Once the simulated structures for both the wild-type and mutated proteins were available, the next step was to identify nodes that were impacted the most due to the mutation. To do this, we first converted the three-dimensional protein structure into a network, wherein, the amino acid residues represented the different nodes and inter-residue interaction was modeled using edges. To do this, we used NAPS (Network Analysis of Protein Structures) [24] where the C_α atom of an amino acid residue is considered as the node, and an unweighted edge is drawn if the C_α-C_α distance between a pair of residues is within a threshold distance ($\sim$7 A). Such a protein contact network captures the 3D topology of protein structure very well. We compared the two networks (wild-type and mutated) using global and residue-level centrality values.

On the global level, various values were computed such as the degree heterogeneity, subgraph centralities, etc. These values have been previously used to quantify differences between the main protease structures of SARS-CoV-1 and SARS-CoV-2 [40].

1. Number of nodes: n

2. Number of edges: m

3. Average degree $\langle k \rangle$: The degree accounts for a node's direct influence on its nearest neighbors.

$$\langle k \rangle = \frac{2m}{n} \tag{4.1}$$

4. Edge Density (δ): The edge density δ is in the range $[0, 1]$. It is 0 for a network with zero edges and 1 if all the nodes are connected, i.e, the graph is complete.

$$\delta = \frac{2m}{n(n-1)} \tag{4.2}$$

5. Degree Heterogeneity (ρ) [39]: The summation for degree heterogeneity is carried out over all edges in the network. For a network where all the nodes have the same degree (k), $\rho = 0$.

$$\rho = \sum_{(i,j)\in E} \left(k_i^{-1/2} - k_j^{-1/2}\right)^2 \tag{4.3}$$

6. Average Betweenness Centrality $\langle BC \rangle$ [50]: σ_{ikj} is the number of shortest paths between the nodes i and j through k and σ_{ij} is the total number of shortest paths from i to j. It represents the node's importance in information passing through the shortest paths through it.

$$\langle BC \rangle = \frac{1}{n} \sum_{i\neq k\neq j} \frac{\sigma_{ikj}}{\sigma_{ij}} \tag{4.4}$$

7. Average Eigenvector Centrality $\langle EC \rangle$ [19]: It represents the spread of information from nodes beyond the neighbours using walks in the graph. These walks can be of infinite length (l) as well.

$$EC_i = \lim_{l\to \text{inf}} \frac{N_l(i)}{\sum_{j=1}^{n} N_l(j)} \tag{4.5}$$

8. Average Subgraph Centrality $\langle SC \rangle$ [43]: It represents the presence of a node in all possible subgraphs and gives a higher weight to subgraphs of smaller lengths. To compute, we first calculate a matrix $G = exp(A)$.

$$\langle SC \rangle = \frac{1}{n} \sum_{p=1}^{n} G_{pp} \tag{4.6}$$

9. Average Communicability $\langle G_{pq} \rangle$ [42]: It represents the average communication between any two nodes in a network using all possible walks/paths. Again, it gives a higher weight to walks/paths of smaller lengths.

$$\langle G_{pq} \rangle = \frac{2}{n(n-1)} \sum_{p<q} G_{pq} \tag{4.7}$$

10. Average Communicability Angle $\langle \theta_{pq} \rangle$ [41]: It describes the efficiency with which information is passed between any two pairs of nodes using all possible walks/paths in a network.

$$\langle \theta \rangle = \frac{2}{n(n-1)} \sum_{p<q} \theta_{pq} \tag{4.8}$$

where the angle θ_{pq} is defined as,

$$\theta_{pq} = \cos^{-1}\left(\frac{G_{pq}}{\sqrt{G_{pp}G_{qq}}}\right) \tag{4.9}$$

11. Average Long-range Subgraph Centrality $\langle Z_{pp} \rangle$: It represents the presence of a node in all possible subgraphs. However, it gives a higher weight to bigger subgraphs than given in $\langle SC \rangle$. To compute Long-range centralities, we first calculate a matrix Z as follows [44]:

$$Z := \sum_{k=0}^{\infty} \frac{A^k}{k!!} = \frac{1}{2} \left[\sqrt{2\pi}\, \mathrm{erf}\left(\frac{A}{\sqrt{2}} \right) + 2I \right] \exp\left(\frac{A^2}{2} \right) \tag{4.10}$$

where $erf(A)$ is the matrix error function of A. However for ease of calculation, we approximated $erf(A)$ by using $tanh(kA)$, where $k = ln(2)\sqrt{\pi}$ [44].

Now, the average long-range subgraph centrality is:

$$\langle Z_{pp} \rangle = \frac{1}{n} \sum_{p=1}^{n} Z_{pp} \tag{4.11}$$

12. Average Long-range Communicability $\langle Z_{pq} \rangle$: It represents the average communication between any two nodes in a network using all possible walks/paths. Again, it allows longer-range transmission than communicability.

$$\langle Z_{pq} \rangle = \frac{2}{n(n-1)} \sum_{p<q} Z_{pq} \tag{4.12}$$

We also computed certain centrality values for each corresponding node in the wild-type and mutated networks. The nodes which had the highest change (top 1 percentile) for each centrality were chosen as the most impacted nodes. Such nodes were aggregated using different node centralities measures namely: Degree, Closeness, Betweenness, Clustering coefficient, Eigenvector centrality, and Eccentricity (Table 4.1). These nodes could be important to the protein structurally, functionally, or both. We also checked if any of these residues had been reported in the literature to be significant for the protein.

The degree is a fundamental measure in network analysis that counts the number of edges connected to a node. It can be utilized in the context of protein structure networks to locate highly connected residues that are crucial for mediating protein-protein interactions or stabilizing protein folds. Degree and betweenness values can help understand the structural and conservation characteristics of a protein [80]. Closeness is a network measure that quantifies the average distance between a node and all other nodes in the network. It offers significant insights into the local and overall organization of biological networks and might be a valuable tool for anticipating the evolutionary and functional links between proteins. Global geometrical characteristics have been found to be described by closeness centrality, and they are less sensitive to changes in protein conformation [49]. Betweenness centrality is a widely used measure in network analysis that identifies nodes that act as essential intermediaries in communication between other nodes in the network. It measures the number of shortest paths between pairs of nodes that pass through a particular node, indicating the extent to which a node connects different parts of the network. Betweenness centrality in protein structure networks is used to determine important

residues that facilitate communication across various protein domains or subunits as well as to predict functional and evolutionary links between proteins. The clustering coefficient is a network metric that measures the degree to which nodes in a network tend to cluster together. It can be used to find clusters of residues that are closely linked and may represent functional domains or units in protein structure networks. According to the significance of its connections to other highly central nodes in the network, each node in a network is given a score by the network metric known as eigenvector centrality. It can be used to discover residues that are closely related to other significant residues in the protein structure and may be vital for protein stability and function. The complex interactions of amino acid that result in allosteric signaling has been studied and characterized using the eigenvector centrality measure [97]. Eccentricity measures the largest shortest path between a node and all other nodes in the network. It can be used to pinpoint residues that are essential for maintaining the protein's structural integrity. Protein with low eccentricity is subject to stricter regulation, making it more likely that it will have an impact on multiple other proteins [46].

We execute the entire network analysis using a python based automated pipeline which just takes in the simulated PDB files for the wild-type and mutated protein structures and returns all the information discussed above.

Table 4.1: The different node centrality measures (computed using NAPS) used for each residue (node) in the protein (network). Here, V represents the set of all vertices in the network and A represents the adjacency matrix.

Property	**Description**	**Equation**
Degree	The number of a node's immediate neighbours.	$k(u) = \sum_{v \in V} A_{uv}$
Closeness	The inverse of the node's shortest path distance to all other nodes in the network. Here, $dist(u, v)$ represents the shortest path distance between nodes u and v.	$CL(u) = (n-1) / \sum_{v \in V} \text{dist}(u, v)$
Betweenness	The ratio of all shortest paths that pass through a node to all the shortest paths in the network.	$BC(u) = \sum_{s \neq u \in V} \sum_{t \neq u \in V} \sigma_{sut} / \sigma_{st}$

Clustering coefficient	The ratio of the number of connections between neighbours of a node to the total number of connections possible between the neighbours of the node. Here $\kappa(u)$ represents the number of neighbours of u connected by an edge.	$CC(u) = \kappa(u)/\tau(u)$ $where,$ $\tau(u) = k(u)\,(k(u)-1)\,/2$
Eigenvector centrality	The eigenvector component corresponding to the largest eigenvalue of the adjacency matrix.	$x_u = \frac{1}{\lambda}\sum_{v=1}^{N} A_{uv}x_v$
Eccentricity	The shortest path distance between the node to its farthest node in the network.	$ECC(u) = \max_{v \in V}(\text{dist}(u,v))$

4.2.2 Structural Analysis

Mutations have been shown to have an impact on protein folding and stability [81, 2, 135], protein function [128, 20], inter-residue interactions, and protein-protein interactions [132, 98]. Protein structural analysis is required to determine whether a mutation affects the important sites or motifs, secondary structures, or other significant properties which could affect the functions of a protein. Therefore, we performed structural analysis to better understand the impact of India-specific mutations on certain proteins of SARS-CoV-2 using Sequence Annotated by Structure (SAS) [93], Missense3D [64], and mCSM-PPI2 [118] web tools.

SAS [93] annotates a given protein sequence with structural information from similar proteins with a known 3D structure in the Protein Data Bank (PDB). SAS takes as input a protein FASTA file or a user-specified PDB file of an experimental or known protein and performs multiple alignments against every protein in the PDB database, and generates structural data about the features of the input protein. Relevant structural information such as secondary structures, active sites, and contact sites to ligands, metals, or DNAs/RNAs were extracted for the proteins of the SARS-CoV-2 virus. We then checked if these important sites were in close proximity to the mutation site or coincide with the most impacted residues that are reported by NAPS. This would enable us to predict a possible functional impact on the protein which should be verified experimentally.

Missense3D [64] predicts the structural changes introduced by a missense mutation. It generates a comparative model using the Phyre2 server or any input PDB file to determine the structural changes

Table 4.2: The PBD structure used for the structural analysis of each India-specific mutant. Note that for network analysis simulated structures were used.

SARS-CoV-2 Protein	PDB Identifier	Citation
ORF3a	6XDC	[71]
Nucleocapsid	8FD5	[23]
NSP3	6WUU	[120]

brought about by an amino acid substitution. Missence3D generates predictions for many important structural properties including alterations in secondary structure, breakage of salt bridges, hydrogen or disulphide bonds, change of the cavity volume, and replacement of a buried hydrophobic residue with a hydrophilic residue. We aggregated information on any impacted structural properties from the tool for India-specific mutations.

mCSM-PPI2 [118] was used to predict the impact of mutations on protein-protein affinities. This is important as changes in the binding affinity and inter-residue interactions caused by mutations affect the formation of interacting complexes. The change in binding affinity caused by the mutation is calculated as the negative of $\Delta\Delta G_{\mathrm{wt-mt}} = \Delta G_{\text{wild-type}} - \Delta G_{\text{mutant}}$, that is, from mutant to wild-type protein. mCSM-PPI2 uses graph-based structural signatures called mCSM [106] which are then fed into a machine-learning model. mCSM-PPI2 is trained on several databases that describe modifications to protein stability and protein-protein affinities. The web tool takes a PDB file and a point mutation as input and outputs the predicted change in binding affinity (in kcal/mol) as well as an interactive 3D viewer displaying the following inter-residue interactions: clash, van der Waals, aromatic, hydrophobic, hydrogen bonds, carbonyl, ionic, and polar. These inter-residue interactions at the mutation site were compared for the wild-type and mutated proteins for the India-specific mutations.

The PBD structures used for each India-specific mutant analysis are specified in Table 4.2.

4.3 Results and Discussion

4.3.1 ORF3a: Q57H

ORF3a in SARS-CoV-2 is a conserved protein with a homo-tetramer structure that presents itself in a dimer-of-dimer configuration [71] with multiple well-conserved functional motifs. The protein is a viroporin that interferes with ion channel activities to enable replication and release of the virus [154]. Gene Ontology (GO) annotations for the ORF3a protein indicate that it is linked to biological processes and functions related to the activation of host autophagy by viruses, ion transport, and vi-

ral ion channels. Similar outcomes from pathway analysis indicate that the protein may be associated with virion assembly and release (R-HSA-9694322) and attachment and entry (R-HSA-9694614). The ORF3a protein forms homotetrameric ion channels that are localized at endosomes and lysosomes and may lead to lysosome deacidification [92, 55]. This is essential for viral egress as it enables virions to exit safely through lysosomal trafficking to plasma membrane [92, 55]. The ORF3a protein interacts with host HMGB1 and facilitates the interaction of host HMGB1 with BECN1, as well as host RETREG1/FAM134B-dependent reticulophagy, which aids viral infection by inducing endoplasmic reticulum stress, and inflammatory responses [156].

The Q57H mutation in ORF3a (G25563T) was identified as a Gujarat-specific mutation (part of subclade I/GJ-20A) in the previous chapter (Section 3.5.1). The mutation was present in 53.7% samples in Gujarat during the early phase and in just 5.4% samples from the other Indian states. Even after the first wave of the pandemic, its prevalence in Gujarat (56.3%) compared to the rest of the country (17%) remained high. However, due to the natural decline in the prevalence of the I/GJ-20A subclade, its frequency in Dataset III (10.4%) and Dataset IV (5.8%) drastically dropped. It is one of the most widespread mutations in the ORF3a protein present in multiple VOCs/VOIs such as Beta (B.1.351), Epsilon (B.1.429), Iota (B.1.526) and Mu (B.1621) with a stabilizing impact on the virus ($\Delta\Delta G = 0.429$ kcal/mol) [17]. The variant causes a change in vibrational entropy ($\Delta\Delta S$) value of 0.44 kcal/Kmol, and therefore an increase in the flexibility of the protein [65]. The mutation has also been shown to increase the binding affinity of the ORF3a and Spike protein (Orf3a-S) complex in comparison to its wild-type [145], and this increase in binding affinity is associated with the severity of the disease [154]. Patients with the Q57H mutation are more likely to be hospitalized, require ICU admission or experience other serious outcomes [154]. We too observe the same from our demographic analysis as this mutation (part of I/GJ-20A) created havoc in the state of Gujarat during the early phase with nearly twice the death rate compared to the rest of the country. This substitution also becomes a hot spot in the Orf3a-S and Orf3a-Orf8 complexes. This leads to disruptions at drug-targeting sites and drug resistance as the protein-binding interface is altered [145]. The Q57H substitution introduces an early stop codon in the ORF3b protein resulting in a truncated 13 a.a. protein instead of the full-length 57 a.a. ORF3b. This truncated ORF3b loses the ability to inhibit interferon induction, implying increased virus transmission [74]. Although ORF3a plays a vital role in the virus' ion channel activity, the presence of the mutation does not significantly affect the relative ion permeability [71].

The ORF3a protein has five α helices, eight beta sheets, six active sites, and eight ligand interaction sites, as annotated from our SAS analysis (Figure 4.1). Interestingly, the mutation is located towards the end of the first alpha helix and lies very close to K61, I62, I63, and T64 active sites. Additionally, these four active sites have also been predicted to function as ligand-binding sites. A protein's solvent-accessible surface area (ASA), calculated as relative solvent accessibility (RSA) [134], has generally been considered to be essential in determining protein folding and stability. For the mutant residue histidine (H), RSA = 52.1%, whereas the wild-type residue glutamine (Q) is exposed with RSA = 43.9%.

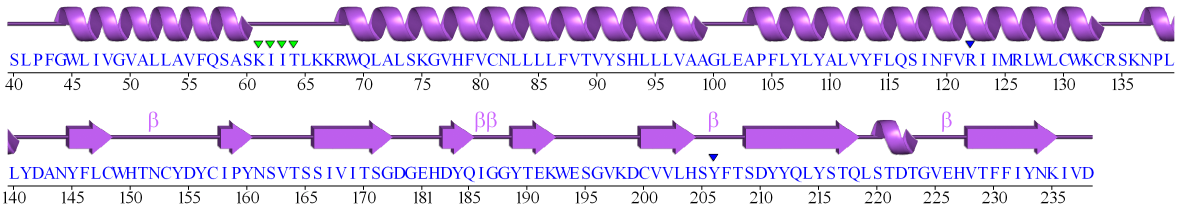


Figure 4.1: SAS annotation secondary structure plot for the ORF3a protein of SARS-CoV-2. The inverted triangles indicate that the residue is an active and/or contact site to ligands.

Apart from a marginal increase in RSA value and no other significant changes, Missense3D analysis predicts that the Q57H mutation does not result in any structural damage. However, the lack of major structural change does not directly imply minimal functional impact. mCSM-PPI2 predicted that the inter-residue interactions of the ORF3a protein undergo certain changes, with the number of hydrogen bonds, polar interactions, and clash interactions increasing. The number of van der Waals and hydrophobic interactions decreases. This may be one of the many causes for increased stability, as hydrogen bonds and other polar bonds have been known to affect the protein stability and binding [101, 121]. The observed increase in RSA value may be attributed to a reduction in hydrophobic bonds and an increase in hydrogen bonds as well as the large size of the mutated residue. Additionally, due to the mutation, glutamine (Q) is replaced with an aromatic amino acid histidine (H) and the introduction of this ring structure may also cause changes in the inter-residue interactions. The presence of an additional Nitrogen atom in the ring of the mutant H57 increases the number of hydrogen bonds and polar interactions. This, however, also results in clash interactions which increased from four in the wild-type to eight in the mutated protein. Six out of these eight interactions take place in the ring structure of H57. It was also predicted that the substitution causes an increase in affinity ($\Delta\Delta G = 0.061$ kcal/mol). The increase in the number of inter-residue hydrogen bonds may be causing this increase in binding affinity.

From the network analysis using NAPS, we found significant changes in global centrality values between the wild-type and mutated structures (Table 4.3). The average betweenness centrality increased by 7.01%, average subgraph centrality by 4.23%, and average communicability by 2.68%. The average eigenvector centrality, however, decreased by 4.86%. Interestingly, both long-range subgraph and communicability centralities increased significantly, 78.79% and 69.36%, respectively. Increased subgraph and communicability centralities (shorter and longer-range) in the mutated protein suggest that the mutated ORF3a protein is topologically more efficient in transmitting information between its residues. Therefore, residues (both short and long distances away) have an increased sensitivity to feel perturbation/thermal oscillation produced in other amino acid residues in the protein. However, increased betweenness centrality and reduced eigenvector centrality indicate that much of this transmission occurs through the shortest path between the residues instead of long walks. At position 57, due to the Q-H substitution, the betweenness centrality increases (10.4%) while the eigenvector centrality signifi-

cantly drops (64%), which is in line with the changes in our global centrality values. Changes in global metrics for the ORF3a protein, upon mutation, also indicate that there is a decrease in diameter from 27 to 26 (could be due to an increase in the betweenness centrality) and a decrease in clustering coefficient from 0.54 to 0.53, implying that the mutant is more compact in comparison to the wild-type protein. This might be due to the increase in the number of polar and hydrogen bonds in the mutant. Comparing centrality measures between wild-type and mutated proteins at individual residue levels resulted in 18 most impacted residues. Among them, nine are in well-conserved motifs (Table 4.4). M1, D2, and L3 all have large eccentricity changes and are located in the N-terminal signal peptide motif (a.a. 1 - 13). I37, which has a high betweenness centrality change, is present in the TRAF3-binding motif that has been linked to the activation of the NF-B and NLRP3 inflammasomes [66, 127]. Residues A72 and K75 (also with high betweenness centrality change) are located in caveolin-binding motifs 1 and 2, respectively, which control the trafficking of ORF3a to the plasma membrane, endosomes, and lysosomes [102]. L147, in the cysteine-rich domain and the third caveolin-binding motif, has a high eigenvector centrality change. Both T176 (high clustering coefficient change) and T179 (high degree and eigenvector centrality change) are within the SGD motif which is required for protein sorting and transporting ORF3a from the Golgi to plasma membranes [130]. I249, one of the 28 conserved residues in ORF3a [17], has a high degree change.

Table 4.3: Percentage change in global centrality values in the mutant compared to wild-type obtained through network analysis using NAPS

Property	**% Change for Q57H**	**% Change for P13L**	**% Change for S194L**	**% Change for A1812D**
Number of nodes	0.00	0.00	0.00	0.00
Number of edges	0.20	-0.30	0.68	1.70
Average degree	0.20	-0.30	0.68	1.70
Edge density	0.20	-0.30	0.68	1.70
Degree heterogeneity	0.62	0.51	0.24	6.49
Average betweenness	7.01	-1.27	-2.07	-13.74
Average eigenvector centrality	-4.86	4.36	3.98	4.12
Average subgraph centrality	4.23	-0.83	0.34	8.28
Average communicability	2.68	-2.87	-3.14	8.97
Average communicability angle	0.13	0.06	0.03	-0.16
Longrange average subgraph centrality	78.79	-49.49	-45.47	77.24
Longrange average communicability	69.36	-46.06	-43.31	90.86

Table 4.4: List of most impacted residues while comparing centrality measures between wild-type and mutated proteins at individual residue levels for the ORF3a: Q57H mutation. Only the residues with functional/structural importance have been listed.

Residue	Importance	Impacted centrality
I249	Conserved residue, Predicted in beta sheet	Degree
M1	Located in the N-terminal signal peptide	Clustering coefficient, Eccentricity
D2	Located in the N-terminal signal peptide	Eccentricity
L3	Located in the N-terminal signal peptide	Eccentricity
I37	Located in the TRAF3-binding motif	Betweenness
A72	Located in the Caveolin-binding motif 1	Betweenness
K75	Located in the Caveolin-binding motif 2	Betweenness
L147	Located in the Caveolin-binding motif 3 and Cysteine-rich domain	Eigenvector centrality
T176	Located in the SGD moitf	Clustering coefficient
T179	Located in the SGD moitf	Degree, Eigenvector centrality

4.3.2 N: P13L

The nucleocapsid (N) protein is a multiple-domain, structurally diverse RNA-binding protein located inside the virion [129]. Its N-terminal domain (NTD) and the C-terminal domain (CTD) are two conserved, independently folded regions that the protein shares with other coronaviruses. It is a crucial protein in the viral life cycle and has functions that mainly involve regulating host-cell cycle progression, host–pathogen interactions, and apoptosis [146]. One of its main roles is packing the positive strand of the viral genome RNA during virion assembly [129]. In order to do this, the N protein interacts with the membrane protein M and the virus's genome [129]. The N protein not only aids in virion assembly but also improves the effectiveness of viral replication and subgenomic RNA transcription [129]. Furthermore, research also suggests that the N protein may influence transforming growth factor-beta signaling by interacting with the host SMAD3 [159]. In terms of molecular function, the top Gene Ontology (GO) annotations were related to DNA and RNA-binding as well as molecular adaptor activity. On the other hand, the GO annotations associated with biological processes included negative regulation of interferon-beta production, viral RNA genome packaging, host-virus interaction, and transcription

regulation. Specifically, the protein's involvement in interferon-beta production regulation implies that it may have a significant impact on the host's innate immune response to the virus [18]. In line with the previous findings, the protein's role in viral RNA genome packaging and transcription regulation highlights its essential role in the virus's replication and propagation. The Nucleoprotein interacts with a range of other molecules during its role in viral replication. It interacts with both monomeric and oligomeric RNA [60]. It also participates in interaction with viral proteins M and E [60, 137].

The P13L mutation is a non-synonymous mutation found in the Nucleoprotein (N protein) of SARS-CoV-2 which co-occurs with subclade I/A3i mutations (Section 2.3.2.3). It was present in nearly 33% samples in India during the early phase with a negligible presence outside India. However, by the end of August, owing to no new infections, the subclade gradually disappeared from Indian isolates. The states of Tamil Nadu, Telangana, Delhi, etc. where I/A3i was prevalent, displayed reduced fatality in the early phase. This hints at possible reduced severity due to infection by the strain. Notably, a recent study identified the N: P13L mutation in nearly all sequences of the Omicron variant analyzed [68]. It has been predicted to be deleterious in nature, thereby impacting protein functionality and having a stabilizing effect on the protein [105]. The mutation results in a distinct (AEGSRGGSQASSRSSSRSRNS) epitope with high antigenicity value [22], i.e., the ability of a virus to bind to specific antibody molecules in the host [139].

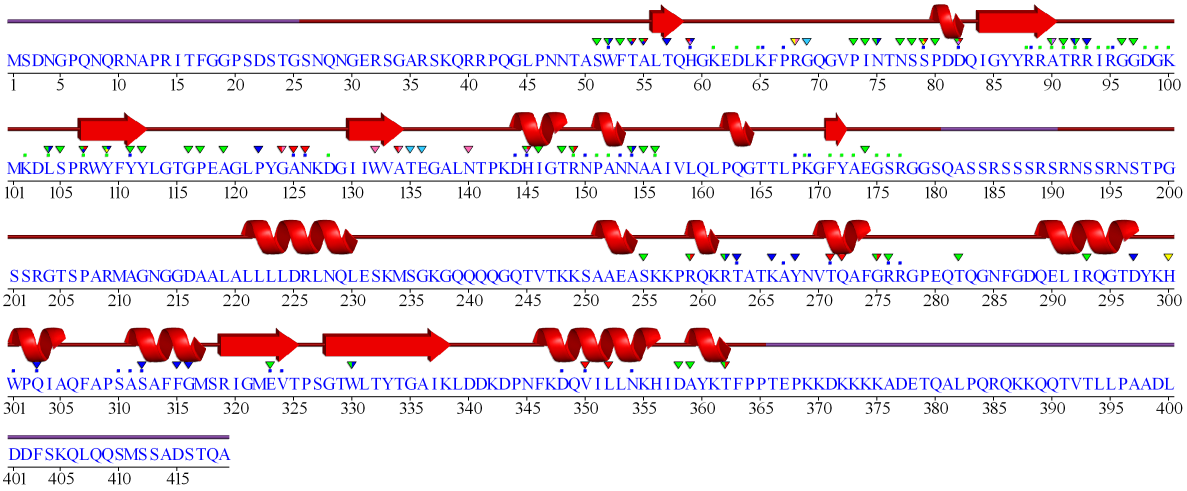


Figure 4.2: SAS annotation secondary structure plot for the Nucleocapsid protein of SARS-CoV-2. The inverted triangles indicate that the residue is an active and/or contact site to ligands. The presence of a green and blue dot means that the residue is a contact site for DNA/RNA interactions and metal ions, respectively.

SAS predicted secondary structures, along with multiple active sites and contact sites for ligands, metals, and DNAs. However, given that the region in the nucleocapsid protein around the mutation is not well-solved, such sites could not be identified near position 13 (Figure 4.2). Missense3D predicted no structural damage to the protein. The substitution of the amino acid proline with leucine led to the expansion of the cavity volume by 7.128 $Å^3$. Cavities have been shown to determine the pressure unfolding of proteins [117], however, in this case, the increase in volume for the mutation is marginal. Additionally, an increase in the residue's relative solvent accessibility (RSA), from 32.3% for the wild-type proline (P) to 40.8% for the mutant residue leucine (L), was observed. mCSM-PPI2 analysis predicts that the P13L mutation causes a decrease in the binding affinity (predicted affinity change, $\Delta\Delta G = -0.056$ kcal/mol). From the inter-residue interactions between the wild-type and mutant proteins, we observed that no other alterations were found apart from the hydrophobic interaction between the mutant L13 and the adjacent T16 residue. This increase in hydrophobic interaction can be attributed to the hydropathy of the amino acids, where L is significantly more hydrophobic than P [109].

From the change in average centrality measures computed for the wild-type and mutated proteins (Table 4.3), we observe a decrement in average betweenness (1.27%), communicability (2.87%), long-range subgraph centrality (49.49%), and long-range communicability (46.06%), and an increase in average eigenvector centrality (4.36%). These results clearly suggest that information passing (both short-range and long-range) through the protein is greatly reduced due to the mutation. The transmissibility through the shortest paths has decreased (negative betweenness centrality change). In contrast, the transmissibility through longer walks between the residues has increased (positive eigenvector centrality change), resulting in a net decrease in communicability between residues. Observations of a reduction in both diameter and shortest path, resulting from global changes in the protein network, indicate that the mutant protein is more densely packed. This change in compactness in homologous proteins can account for differences in folding rates and mechanisms [53]. The betweenness centrality for proline dropped by 3.5% due to the mutation to lysine at position 13. This is also in line with what was observed for the global betweenness centrality (reduced by 1.27%). Through the centrality comparison of individual residues between the wild-type and mutated nucleocapsid proteins using NAPS, we identified key residues that were impacted the most (Table 4.5). A381, a residue with a high degree change, is located in the antibody-binding region (a.a. 365 - 400), which is nearly 100% conserved [23]. A264 (high eigenvector centrality change) forms a strong hydrophobic interaction with S312 between the $\alpha 1$ and $\alpha 4$ helices to hold the two monomers firmly [150]. A336 (both high closeness and eccentricity change) is one of the RNA-binding sites in the SARS-CoV-2 CTD [146]. The following six residues with high values of change: M210 (degree), A211 (clustering coefficient and betweenness), G212 (degree), N213 (degree and betweenness), G215 (betweenness), and A217 (betweenness) are positioned in the junction region (a.a. 210 – 246) of the N protein. This region is essential for RNA-mediated liquid-liquid phase separation (LLPS) [84, 157]. Residues G116 (degree), I146 (clustering coefficient), K266 (eigenvector centrality), and T282 (clustering coefficient) are predicted by SAS to be active sites that also act as

ligand-binding sites. G69, which has high closeness and eccentricity change, was also predicted to be an active site. K56 and D98, with high clustering coefficient changes, function as metal contact and DNA/RNA interaction sites for the nucleocapsid protein, respectively.

Table 4.5: List of most impacted residues while comparing centrality measures between wild-type and mutated proteins at individual residue levels for the N: P13L mutation. Only the residues with functional/structural importance have been listed.

Residue	**Importance**	**Impacted centrality**
A264	Takes part in strong hydrophobic interaction with S312 in α1 helix and α4 helix to firmly fix the two monomers	Eigenvector centrality, Eccentricity
M210	Located in the junction region at 210–246, which has been found to be essential for RNA-mediated LLPS	Degree, Eccentricity
S188	Located in the SR motif, Play a role in RNA binding	Closeness
S208	Located in the SR motif, Play a role in RNA binding	Clustering coefficient
R177	Located in the RGG/RG motif of PRMT1-methylated N protein, Regulates the binding of N protein to its 5'-UTR genomic RNA, Contact site for DNA/RNA interaction	Betweenness
I146	Active Site, Contact site for ligands MES (2-(N-Morpholino)-Ethanesulfonic acid) and DJO ((Phenylmethyl) (2s)-2-(Hydroxymethyl)-2,3-Dihydroindole-1-Carboxylate)	Clustering coefficient
K266	Active site, Contact site for ligand PEG (Di(hydroxyethyl)ether)	Eigenvector centrality, Eccentricity

4.3.3 N: S194L

The S194L variant of the N protein in SARS-CoV-2 is a non-synonymous mutation caused by a C28854T transition [122]. This mutation is located in the linker region of the N protein, which is a central and essential area for its oligomerization [16, 153, 158, 152]. Studies have suggested that this region lacks an organized structure and is highly conserved [96]. The S194L mutation is primarily prevalent in Asia, particularly in countries like India, Bangladesh, and Saudi Arabia [21]. It was identified as one of the mutations in the Gujarat-specific subclade I/GJ-20A (Section 2.5.1). During the early phase, nearly one-third of samples from Gujarat had this mutation, while only 1% samples from the rest of the country carried this mutation. According to various studies, this mutation is associated with fatal outcomes. It is also deleterious in nature, suggesting that it may have a significant role in increasing the virus's pathogenicity, severity as well as the clinical outcome of the disease [124, 91, 78, 90]. This corroborates the death rates discussed in the previous chapter. Gujarat ($\sim$5%) had the highest death rate in the early phase. ORF3a: Q57H which has also been predicted to increase severity co-occurs with this

mutation in the subclade I/GJ-20A. Both of these mutations together might be responsible for nearly twice the fatality rate in Gujarat samples compared to the rest of the country. Furthermore, the S194L variant is associated with symptomatic patients and serves as an additional site that is located outside the spike region for RT-qPCR screening of SARS-CoV-2 samples [16]. A study also observed that the two mutations, S194L in the N gene and the D294D in the S gene, co-evolve in Indian SARS-CoV-2 genome sequences [14].

The nucleocapsid protein region surrounding the S194L mutation remains poorly resolved. As with the P13L mutation, the identification of active sites and other contact sites near the mutant position 194 was not possible Fig. 4.2. Most of the interactions predicted by mCSM-PPI2 for the S194L mutation are with neighboring N196, Q163, and N192 residues. The analysis revealed an increase in the number of hydrophobic and clash interactions while the number of hydrogen bonds and polar interactions decreased. This can be attributed to the loss of an oxygen atom and the gain of three carbon and six hydrogen atoms resulting from the amino acidic substitution of S with L as well as the highly hydrophobic character of L [109]. Moreover, the predicted binding affinity decreased with a $\Delta\Delta G$ of -0.15 kcal/mol. According to Missense3D analysis, the S194L mutation did not result in any structural damage. However, the clash score increased slightly from 23.18 to 25.08, but it was not high enough to trigger a clash alert. Additionally, the relative solvent accessibility (RSA) value of the amino acid at position 194 decreased from 42.3% for S to 28.6% for L, indicating a possible change in the local environment. These results reveal insights about the loss of hydrophilic character and are therefore consistent with the mCSM-PPI2 results. It was also predicted that the mutation led to a reduction in cavity volume by 44.93 $Å^3$, which could have potentially contributed to the observed alteration in protein stability [32].

The summarised average centrality measures (Table 4.3) reveal that the average eigenvector centrality shows an increase of 3.98%, while the average communicability decreases by 3.14%. Additionally, a significant decrease is observed in both the long-range average subgraph centrality (45.47%) and long-range average communicability (43.31%). The increased average eigenvector centrality suggests that the importance of well-connected nodes has become more prominent, while the decreased long-range average subgraph centrality and communicability indicate a disruption in long-range information flow between residues. At position 194, due to the substitution, the degree changes drastically from 4 to 7, with a significant increase in the clustering coefficient. However, the betweenness centrality drops significantly ($\sim$25%) indicating reduced information flow through the node. Other key impacted residues (Table 4.6) include A264 and M210. A264 is involved in strong hydrophobic interactions with S312 in α1 helix and α4 helix to stabilize the N protein [150], and showed significant changes in eccentricity and eigenvector centrality. M210, located in the junction region at 210–246, also exhibited significantly changed degree and eccentricity values. This region has been reported to be crucial for LLPS mediated by RNA [84, 157]. Moreover, S188 and S208, located in the SR motif, were also among the most impacted residues with large changes in closeness and clustering coefficient values, respectively. A

study observed that reduced RNA binding and a change in the protein-RNA populations with different solution characteristics might occur if these positions are mutated [146]. Furthermore, I146 and K266 are active sites and contact sites for ligands. I146 exhibited a significant change in the clustering coefficient, while K266 was amongst the top 1 percentile affected residues for both eigenvector centrality and eccentricity. The R177 residue has been identified as having an affected betweenness centrality and also functions as a contact site for interactions with both DNA and RNA. This residue, which is part of the PRMT1-methylated N protein's arginine-glycine-glycine (RGG/RG) motif, is crucial for regulating how the N protein binds to its 5'-UTR genomic RNA [146].

Table 4.6: List of most impacted residues while comparing centrality measures between wild-type and mutated proteins at individual residue levels for the N: S194L mutation. Only the residues with functional/structural importance have been listed.

Residue	**Importance**	**Impacted centrality**
A381	Significant antibody-binding site	Degree
A264	Takes part in strong hydrophobic interaction with S312 in α1 helix and α4 helix to firmly fix the two monomers	Eigenvector centrality
A336	RNA-binding site in SARS-CoV2 CTD	Closeness, Eccentricity
M210	Located in the junction region at 210–246, which has been found to be essential for RNA-mediated LLPS	Degree
A211	Located in the junction region at 210–246, which has been found to be essential for RNA-mediated LLPS	Clustering coefficient, Betweenness
G212	Located in the junction region at 210–246, which has been found to be essential for RNA-mediated LLPS	Degree
N213	Located in the junction region at 210–246, which has been found to be essential for RNA-mediated LLPS	Degree, Betweenness
G215	Located in the junction region at 210–246, which has been found to be essential for RNA-mediated LLPS	Betweenness
A217	Located in the junction region at 210–246, which has been found to be essential for RNA-mediated LLPS	Betweenness
K56	Contact site for metal	Clustering coefficient
G69	Active Site	Closeness, Eccentricity
D98	Contact site for DNA/RNA interaction	Clustering coefficient
G116	Active site, Contact site for ligand DJU (N,N-Dimethyl-1-(5-Phenylmethoxy-1h-Indol-3-Yl) methanamine) and DJO ((Phenylmethyl) (2s)-2-(Hydroxymethyl)-2,3-Dihydroindole-1-Carboxylate)	Degree

I146	Active site, Contact site for ligand DJO ((Phenylmethyl) (2s)-2-(Hydroxymethyl)-2,3- Dihydroindole-1-Carboxylate)and MES (2-(N-Morpholino)-Ethanesulfonic acid)	Clustering coefficient
K266	Active site, Contact site for ligand PEG (Di(hydroxyethyl)ether)	Eigenvector centrality
T282	Active site, Contact site for ligand GTP (Guanosine-5'-Triphosphate) and 5GP (Guanosine-5'-Monophosphate)	Clustering coefficient

4.3.4 ORF1a: A1812D (NSP3: A994D)

ORF1a is a 7096-length polyprotein. The mutation occurs in the non-structural protein (nsp3) within it (819-2763). Nsp3 is the largest cysteine protease associated with a membrane, containing ten domains [148]. It contains the SARS-unique domain (SUD), which is present in both the SARS-CoVs [3]. During SARS-CoV-2 replication, nsp3 interacts with certain RNA G4s through its SUD domain [75]. It also functions as a protease to separate Nsp1, Nsp2, and Nsp3 from the viral polypeptide by cleaving between Nsp1-Nsp2, Nsp2-Nsp3, and Nsp3-Nsp4 [9]. Moreover, nsp3 is crucial in compromising the immune response by the host as it binds to ATF6 and blocks the transcription factor's stress response [3]. The nsp3 protein is responsible for cleavages that occur at the replicase polyprotein's N terminus [79]. It is also crucial for the rearrangement of the host membrane, which results in the formation of the DMVs, or double-membrane vesicles, that are necessary for viral replication. Studies have shown that nsp3 along with nsp4, and nsp6 can together result in the formation of DMV [6, 7]. By preventing the phosphorylation, dimerization, and subsequent nuclear translocation of the host's transcription factor interferon regulatory factor 3 (IRF3), nsp3 also has an antagonistic effect on the innate immune induction of type I interferon [51, 30]. It also suppresses NF-kappa-B signaling, which is a pathway that regulates immune response within the host cells [51, 30]. Thus, nsp3 is a good drug target candidate due to its critical role in viral reproduction and immune evasion [148].

The missense mutation C5700A (A1812D) in the nsp3 protein of ORF1a results in the substitution of alanine (A) with aspartic acid (D). This mutation predominates in Indian SARS-CoV-2 genomes and is most common in the western states of India [67, 123]. A non-synonymous C313T mutation has been found to co-occur with C5700A in samples from Maharashtra [103]. From our analysis in Sections 2.5.2 and 3.5.2 we identified C5700A as a India-specific mutation found in the state of Maharasthra. Along with C313T, it forms a Maharashtra-specific subclade I/MH-2 during the first wave in India. It was present in 63.12% samples from Maharashtra, while outside the state it was only found in 10.94% samples. Given that during that period Maharashtra had the second highest fatality rate among all the states in India, it can be postulated that the mutation might result in increased severity of infection. Paul

et al. (2020) [103] also reported the same, as compared to asymptomatic people, symptomatic people had a higher percentage of this mutation. Papain-like protease-2 (PLP2) and protease-like non-canonical (PLnc) are important protease domains in the nsp3 protein that play a vital role in the virus's pathogenicity and virulence [59]. The C5700A mutation occurs within these essential viral protease domains and nucleic acid binding (NAR) domains [56].

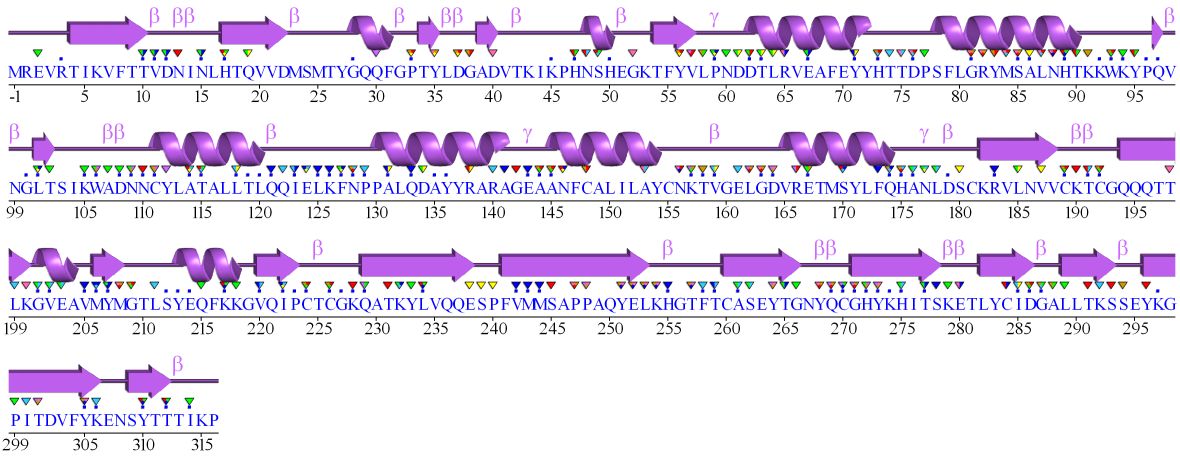


Figure 4.3: SAS annotation secondary structure plot for the Nsp3 protein (residues 745 to 1061) of SARS-CoV-2. The inverted triangles indicate that the residue is an active and/or contact site to ligands. The presence of a green and blue dot means that the residue is a contact site for DNA/RNA interactions and metal ions, respectively.

Nsp3 is a very huge 1945-length protein, so we used a PDB structure that contained the smaller region around the mutation position (994). The 6WUU PDB structure has solved coordinates for the residues 745 to 1061 in Nsp3. The SAS analysis of 6WUU resulted in a structure enriched with active sites, ligand binding sites and contact sites to various metals (Figure 4.3). The mutant position 249 (994-745), located in a beta-strand, is surrounded by several active sites and residues that are contact sites to metal ions. This implies that the mutation in this region may lead to the production of a defective or structurally altered protein. Missense3D indicates that the mutation does not cause any significant structural damage to the nsp3 protein, but it does reduce the clash score slightly from 14.90 in the wild-type to 14.87 in the mutant. The mutation results in a rise in the solvent exposure, from 52.8% to 69.3%. This change can be attributed to the contrasting hydrophobic and hydrophilic properties of A and D, respectively [109]. According to mCSM-PP2, interactions with the residue at position 249 do not change significantly, but there is a loss of one clash interaction between this residue and the carbon atom of neighboring P299 residue, consistent with the results from Missense3D. The mutation results in an increase in binding affinity ($\Delta\Delta G = 0.34$ kcal/mol).

For the simulated structures, we modelled only the region solved by 6WUU as the entire nsp3 protein is very huge (1945 aa). Table 4.3 shows the average centrality measures for the wild-type and mutated proteins. The mutation leads to an increase in degree heterogeneity (6.49%), subgraph centrality (8.28%), long-range subgraph centrality (77.24%), and long-range communicability (90.86%), and a decrease in average betweenness (13.74%). These changes suggest that the mutation leads to a significant increase in information transmission between nsp3 residues. The mutated structure's diameter reduced by 18% and the average shortest path length by 12.4% indicating that the protein became more compact and dense. We also identified key residues that were the most affected by this mutation (Table 4.7). R166 is both an active site and a contact site for two ligands, and it has an increased clustering coefficient. M243, M244, and A261 have a significant change in their eigenvector centrality and also function as active and contact sites for sodium ions, while, C192 (clustering coefficient) is an active site which acts as the contact site for zinc ions. G271 (degree) is a contact site for chloride ions, a ligand binding site for caffeine and proflavin, and participates in a backbone-backbone hydrogen bond with peptide inhibitor VIR250 [120].

Table 4.7: List of most impacted residues while comparing centrality measures between wild-type and mutated proteins at individual residue levels for the ORF1a: A1812D (NSP3: A994D) mutation. Only the residues with functional/structural importance have been listed.

Residue	**Importance**	**Impacted Centrality**
R166	Active site, Conatct site for ligands GYX (N-[(3-Acetamidophenyl)methyl]-1-[(1r)-1-Naphthalen-1-Ylethyl]piperidine-4-Carboxamide) and DZI (3,4,5-Tris(oxidanyl)-N-[(E)-1h-Pyrrol-2-Ylmethylideneamino]benzamide)	Clustering coefficient
C192	Active site, Contact site to Zinc ion	Clustering coefficient
M243	Active site, Contact site to Sodium ion	Eigenvector centrality
M244	Active site, Contact site to Sodium ion	Eigenvector centrality
A261	Active site, Contact site to Sodium ion	Eigenvector centrality
G271	Active site, Contact site to Chloride ion, Contact site for ligands Caffeine and Proflavin, Participates in a backbone-backbone hydrogen bond with peptide inhibitor VIR250	Degree

4.4 Conclusion

This study presents a thorough examination of the effects of India-specific mutations on the SARS-CoV-2 virus's structure and function. By utilizing network and structural analysis in combination with

insights from available literature, we have identified the impact of India-specific mutations on the protein's structure and function. Protein residue networks were built and topologically important centralities were calculated for the wild-type and mutant protein structures. Additionally, we also identified several key residues that play crucial roles within the protein's life cycle like in ligand binding or viral replication, and therefore, significantly impact the protein's stability and function. These findings could be instrumental in developing effective therapeutic strategies against the SARS-CoV-2 virus. Overall, our study provides valuable insights into the effects of India-specific mutations on the SARS-CoV-2 virus, which can aid in understanding the COVID-19 pandemic.

Chapter 5

Epidemiological Analysis of the Second and Third Wave in India

5.1 Introduction

The SARS-CoV-2 virus-caused COVID-19 pandemic has resulted in an unparalleled global health crisis, with India being severely impacted due to its large population. The country has experienced three waves of the pandemic, with the second wave being particularly devastating. To understand the transmission dynamics of the virus and guide policy decisions to mitigate its spread, epidemiological modeling can provide valuable insights. This study aims to model the transmission of SARS-CoV-2 in India, with a focus on important states such as Maharashtra and Gujarat, during the second and third waves using variant counts and confirmed case data. The dominant variants during the second wave were Delta (B.1.617.2), and Kappa (B.1.617.1), while Omicron (BA.2) was primarily responsible for the third wave. We estimate the transmission coefficient of these variants and other critical parameters such as infection duration and time to reinfection by constructing multi-strain SEIR models fitted to the scaled variant data (explained in the next section) using Markov chain Monte Carlo simulations.

Furthermore, in the future, we aim to use the estimated parameters and identify crucial airport nodes in the network to implement intervention strategies that can curb the spread of COVID-19 and similar future pandemics.

5.2 Methods

5.2.1 Data

For the epidemiological modeling, we used 3 different data types:

1. Airport Data: Monthly passenger movement data to and fro from all airports in India aggregated state-wise from January 2018 to September 2021. The data for May 2019 was unavailable and was substituted by averaging the data for April 2019 and June 2019. Source (http://knowindia.net/).

2. Variant Data: Data of SARS-CoV-2 genome samples collected throughout the pandemic. Each data point represents a SARS-CoV-2 genome with the collection date, state, and variant (Delta, Kappa, Alpha, Omicron, or Other) information present. This data was downloaded from GISAID [126] for Indian samples. Fig. 5.2 shows the counts of different variants in India.

3. Cases Data: We also collected daily state-wise data on the number of confirmed COVID-19 cases in India, which allowed us to accurately model the spread of variants. The available sequence data from GISAID were limited, making it difficult to estimate the prevalence of specific variants. To overcome this limitation, we used the number of confirmed cases as a proxy for the true number of infections and scaled up the available sequence data accordingly. For example, if only 100 sequence data were available for a state where 10,000 confirmed cases were reported in that given period, we scaled up the available sequence data by multiplying it by a factor of 100 (i.e., 10,000/100) to estimate the overall prevalence of the variant. The cases data till October 2021 was downloaded from Kaggle (https://www.kaggle.com/datasets/sudalairajkumar/covid19-in-india) and after that was taken from PRS India (https://prsindia.org/covid-19/cases). Fig. 5.1 shows the daily confirmed cases data in the second and third waves in India.

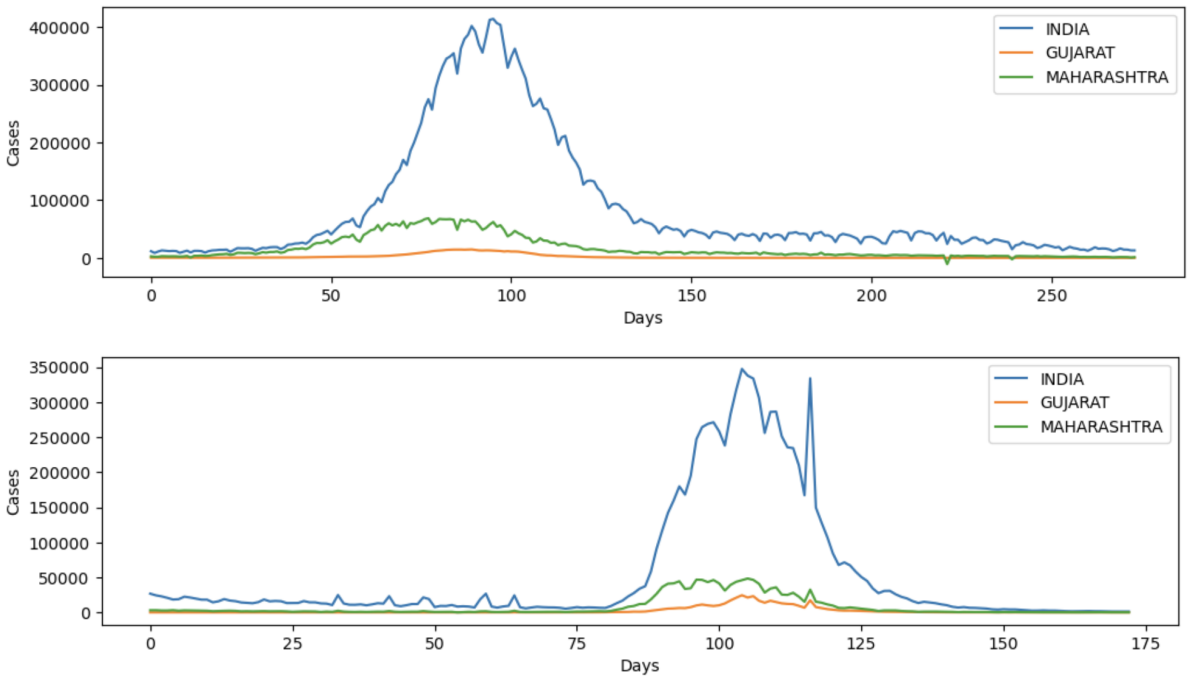


Figure 5.1: Daily confirmed cases data in the a) second and b) third waves in India.

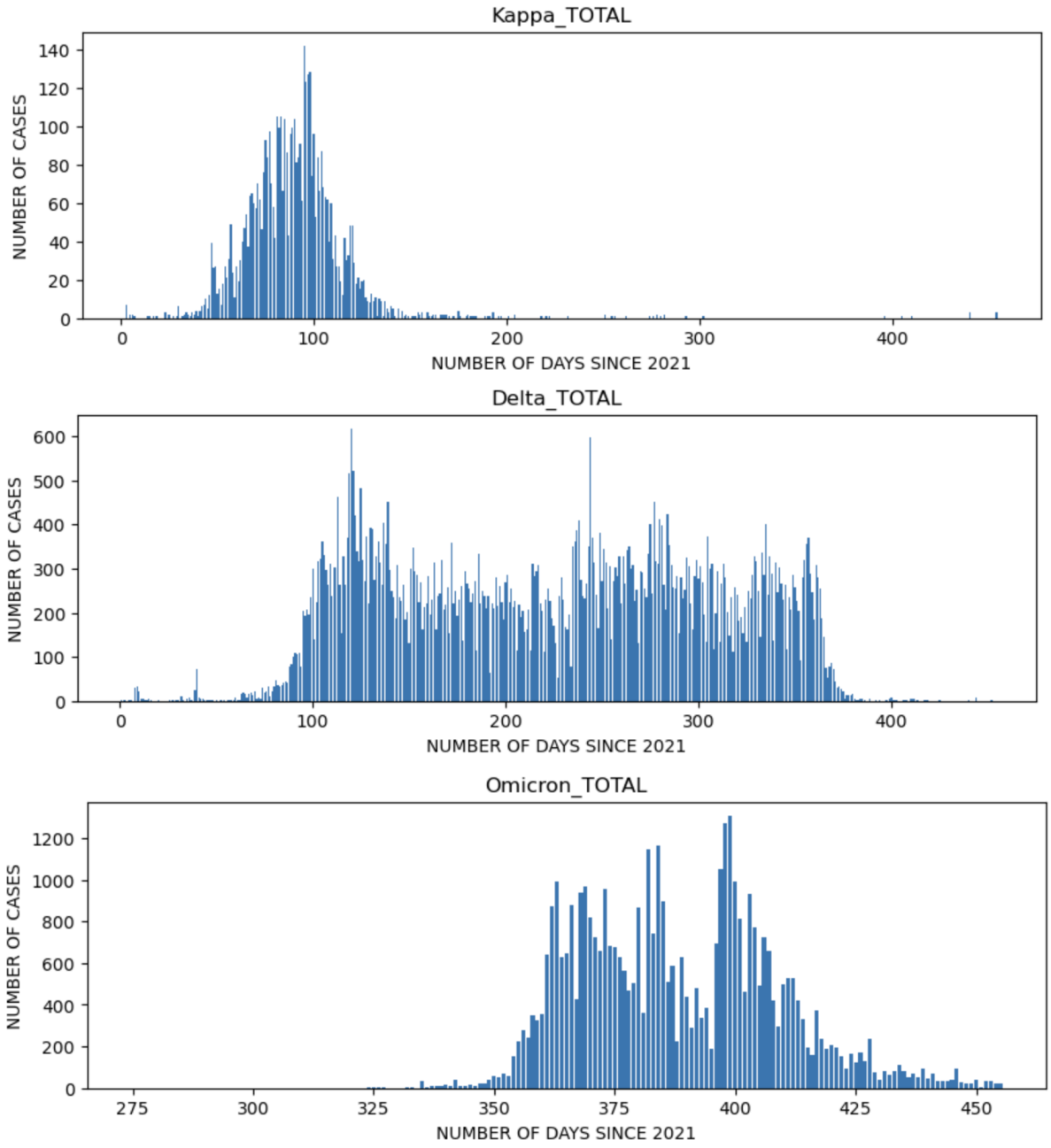


Figure 5.2: Daily counts of different lineages in India from GISAID. a) Kappa in the second wave, b) Delta in the second wave, and c) Omicron in the third wave.

5.2.2 Methodology

Our proposed modeling approach involves developing variant-level SEIR models for each state in India to capture the spread of infection across the country. We conducted this exercise separately for the

second and third waves of the pandemic due to the presence of different circulating variants. During the second wave, the Delta and Kappa variants accounted for around 90% of the cases, while the Omicron variant alone caused over 90% of the cases during the third wave.

Our primary objective is to simulate the pandemic using these models and accurately capture the variant-specific epidemic curves. This will enable us to estimate the β value for each variant, which represents its transmission coefficient.

5.2.2.1 SEIR model

An SEIR model has the following 4 compartments, susceptible, exposed, infectious, and recovered. Now, to be able to include variant information in the SEIR model, we keep a separate E_i and I_i for each variant i. The S and R are kept the same for all variants to model the interaction between variants. The equations are as follows:

$$\frac{dS}{dt} = -S \sum_{i=0}^{N} \beta_i I_i + \delta R \tag{5.1}$$

$$\frac{dE_i}{dt} = S\beta_i I_i - \sigma E_i \tag{5.2}$$

$$\frac{dI_i}{dt} = \sigma E_i - \gamma I_i \tag{5.3}$$

$$\frac{dR}{dt} = \gamma \sum_{i=0}^{N} I_i - \delta R \tag{5.4}$$

where, S represents the total susceptible population, E_i represents the exposed population for variant i, I_i represents the currently infected population for variant i, and finally R represents the total recovered population.

In an SEIR model, a susceptible person (either healthy or previously infected with reduced immunity) can contract the virus by coming into contact with an infected individual. β_i represents the transmission of virus per encounter for variant i, and an encounter occurs when a susceptible person meets a person-infected with variant i, defined as SI_i. Individuals who have contracted the virus but are not yet infectious are placed in the exposed compartment. This intermediate state is characterized by a low viral load, and individuals remain in this compartment for a period of $1/\sigma$ days before transitioning to the infected compartment at a rate of σE_i. It is important to note that the exposed and infected compartments are distinct for each variant i. Assuming that the infection duration is $(1/\gamma)$ these infected people move to the recovered state at a rate of γI_i. In this model, we assume a constant population throughout hence, do not consider cases of birth, death due to COVID-19, or death due to other causes. Finally, the recovered individuals transition back to the susceptible category after $(1/\delta)$ days at a rate

δR, which represents reduced immunity from the virus after a long period allowing reinfection by the same or other variants.

5.2.2.2 Estimating parameters

Curve fitting is a popular technique for estimating the parameters of epidemiological models. The model is fitted to the observed data, and the parameters are changed until the model output closely matches the observed data. To estimate the parameters for our SEIR model, we utilized the pymcmcstat package, a Python-based software program that performs Bayesian statistical inference using Markov-chain Monte Carlo (MCMC) simulations. As the ground truth for model fitting, we used the scaled variants counts, which were produced by scaling the variant data using the cases data. We defined the model structure and likelihood function using pymcmcstat, and it then provided a flexible and efficient technique to investigate the posterior distribution of the parameters using MCMC methods. For the second wave (January 2021 - September 2021) we modeled the Kappa and Delta variants, and for the third wave (October 2021- March 2022) we modeled the Omicron variant. We used mean squared error as the cost function for the likelihood estimator.

While running pymcmcstat-based curve-fitting for the second wave, we allowed the model to estimate, β_1: Transmission coefficient of the Delta variant (initial value = 0.5), β_2: Transmission coefficient of the Kappa variant (initial value = 0.5), σ: Time to move from Exposed to Infected compartment (initial value = 0.1428 or 1/7 days), γ: Duration of infection in an individual (initial value = 0.0476 or 1/21), δ: Time after infection when a recovered individual becomes susceptible to reinfection (initial value = 0.0066 or 1/150), and τ: Time delay between the introduction of the Kappa and Delta variants (initial value = 0). Similarly, while running for the third wave, β represented the transmission coefficient for the Omicron variant, with the other parameters, except τ which was removed (only one variant in the system), were kept the same. PI in the output plots represents the 95% prediction interval, and CI is the 95% confidence interval.

5.3 Results

5.3.1 Scaling variant data using infection cases data

The sequencing rate in India has been very poor (cite) with less than 1% samples being sequenced and has been observed to be fluctuating substantially due to changes in government policies and lockdowns. Sequencing rates have also been observed to vary across states, with certain regions exhibiting higher rates than others. As a consequence, relying solely on variant data to model the spread of infection would be inappropriate. To overcome this issue, we used actual monthly infection case data for each state and scaled up the variants data to match the infection numbers. This had to be done separately for different states and months due to significant spatiotemporal variation in sequencing rates. Fig. 5.3

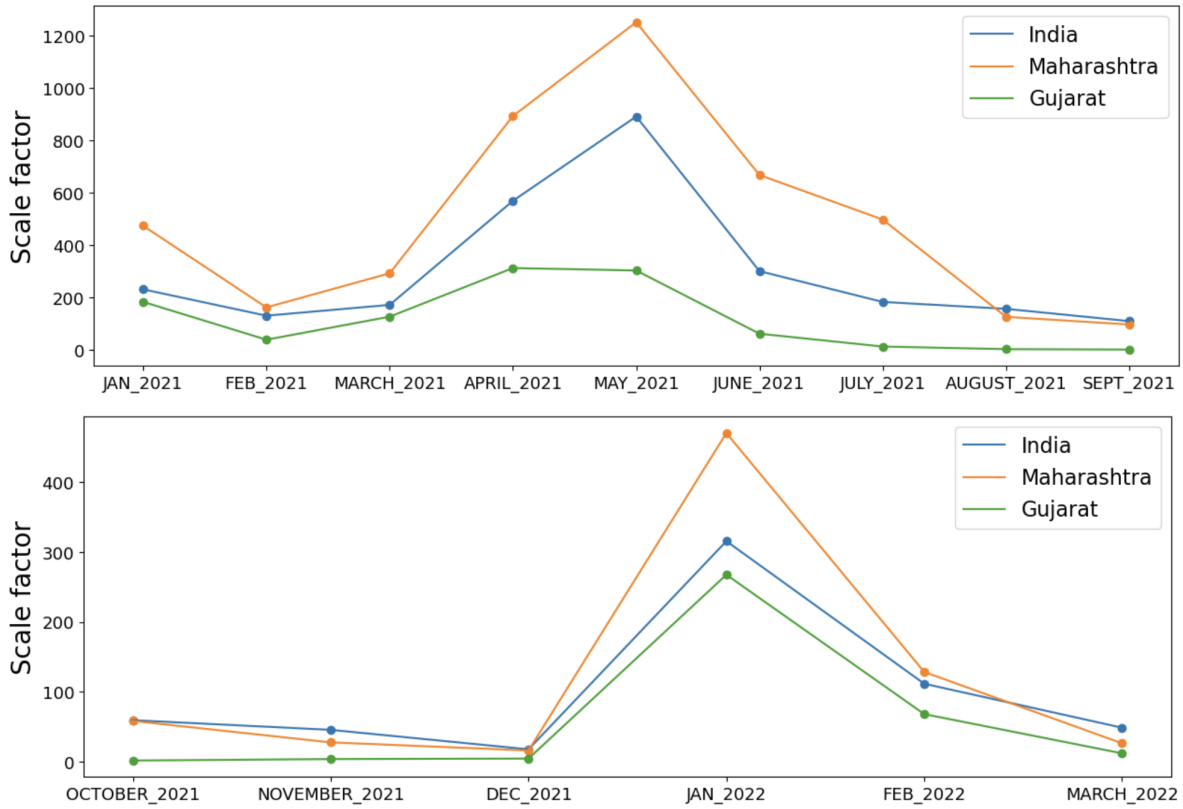


Figure 5.3: The monthly scale factor (total cases/total sequences) for India, Maharashtra, and Gujarat in the a) Second and b) Third waves in India.

shows the monthly scale factor (total cases/total sequences) for different states and the country as a whole for the second and third waves. The spatiotemporal fluctuations described above can be clearly seen from these figures.

Following that, we used these factors to scale up the variation counts based on the time (month) and state of origin of the samples (Fig. 5.4). We can see a clear peak during the second wave in Fig. 5.4 however, the same could not be observed for the unscaled variants data (Fig. 5.2). This is due to the different sequencing rates across different time periods motivating the use of scaling.

Lastly, the daily scaled variant counts were smoothed by taking a 7-day moving average (Fig. 5.5). These scaled counts of variants after taking the 7-day moving average served as ground facts for the SEIR modeling, which was utilized to calculate the transmission coefficient.

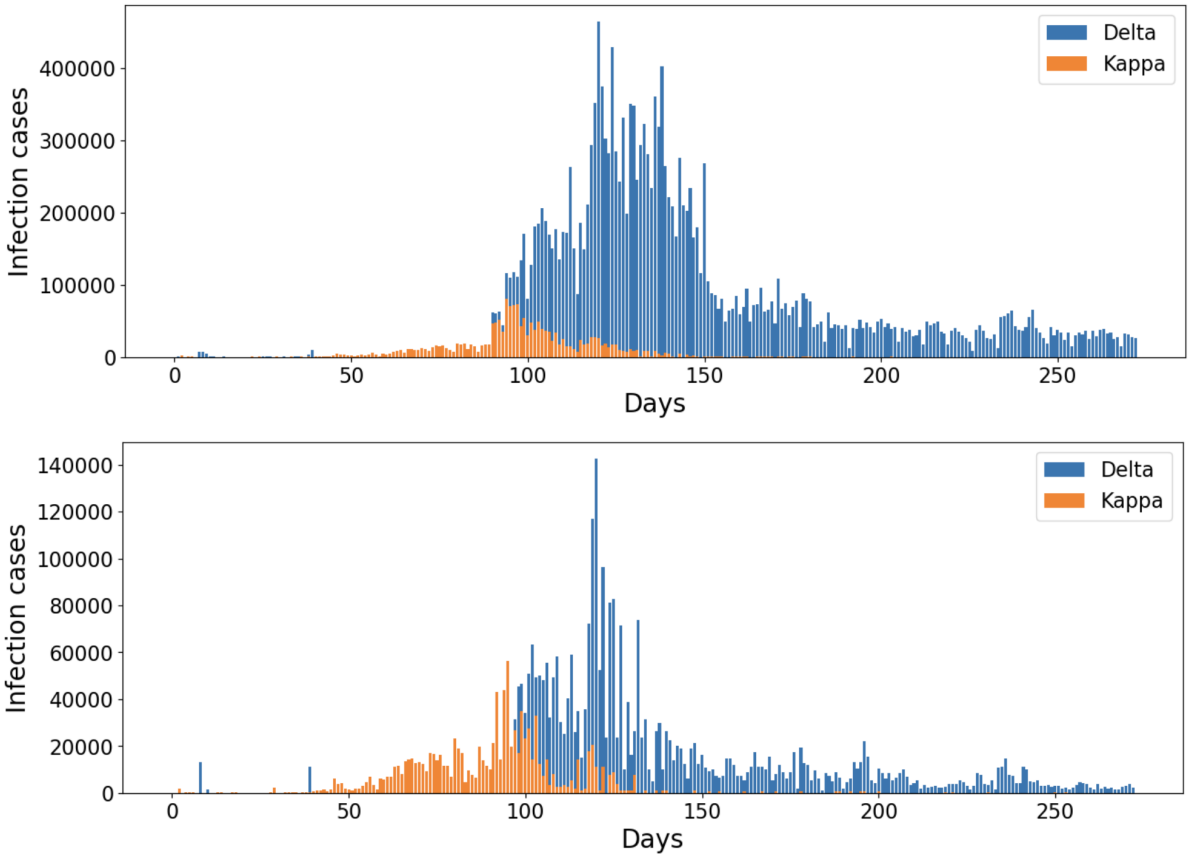


Figure 5.4: Scaled up daily variant counts from a) India and b) Maharashtra during the second wave.

5.3.2 SEIR Model Fitting

Fig. 5.6 illustrates the plots for the 6 compartments (S, E1, E2, I1, I2, and R) during the second wave in India. In this context, E1 and I1 denote the exposed and infected populations for the Delta variant, while E2 and I2 represent the same for the Kappa variant. To model the difference in peaks between the two variants, we used τ which estimates the delay between the introduction of the two variants in India. Delta is known to be a successor of Kappa from the phylogenetic tree, and the parameter τ estimates the time it took for this succession. Plots for I1 and I2 for the state of Maharashtra are also shown in Fig. 5.7. The data indicate that, despite being introduced later (with a peak around May 2021 in India), the Delta variant resulted in a higher number of infections compared to the Kappa variant (which peaked around April 2021 in India). This suggests that a variant already in circulation (Kappa) can be replaced by a more transmissible variant (Delta) if strict containment measures are not employed.

During the third wave, only the Omicron variant was dominant, and therefore, only 4 compartments (S, E, I, and R) were considered. The plots for the infection (I) for India, Maharashtra, and Gujarat are shown in Fig. 5.8. Cases of Omicron in India suddenly spiked around January 2022. However, the

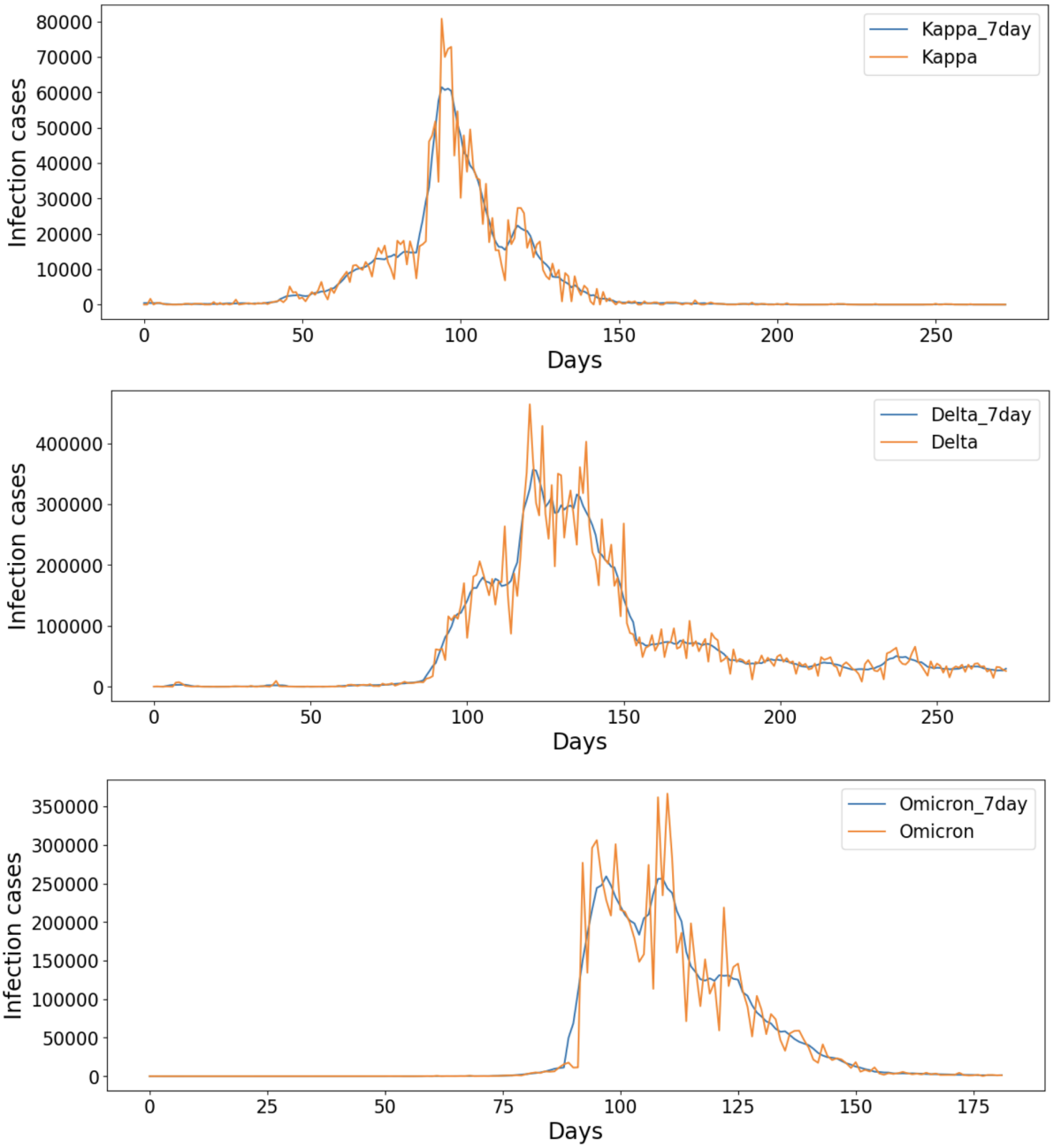
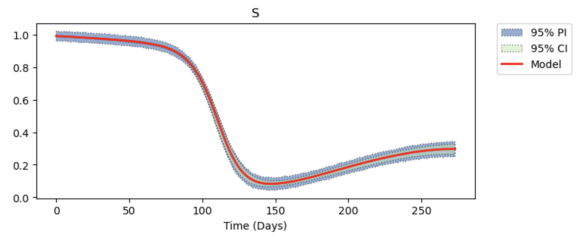
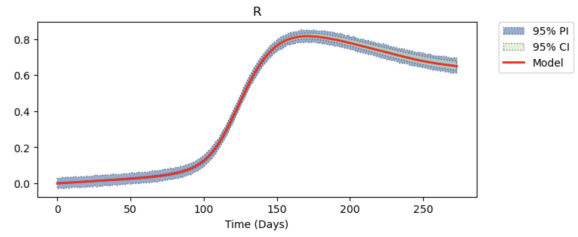


Figure 5.5: Counts after taking the 7-day moving average (in blue) for each variant in India: a) Kappa in the second wave, b) Delta in the second wave, and c) Omicron in the third wave.

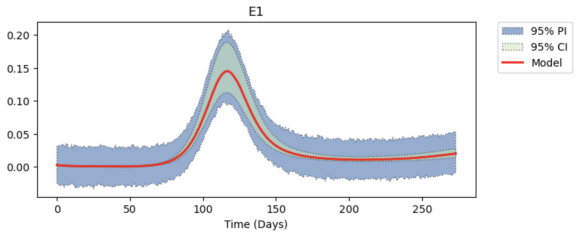
reported confirmed cases of Omicron are significantly lesser owing to the mild and asymptomatic manifestation of the infection. This impacted the scale factor used and therefore the scaled variant counts for Omicron are an under-representation of the actual counts of the variant in the country.



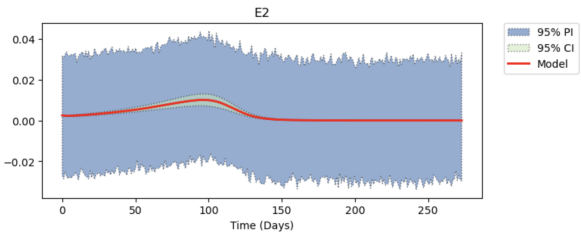
(a) S



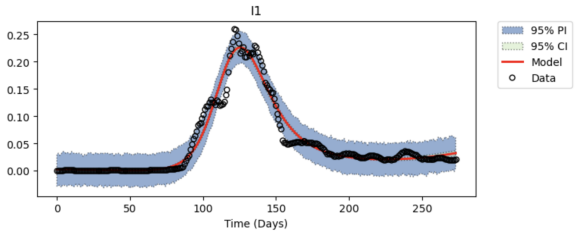
(b) R



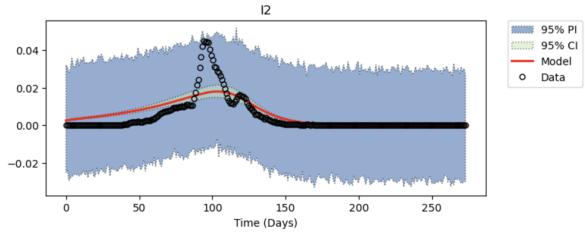
(c) E1



(d) E2

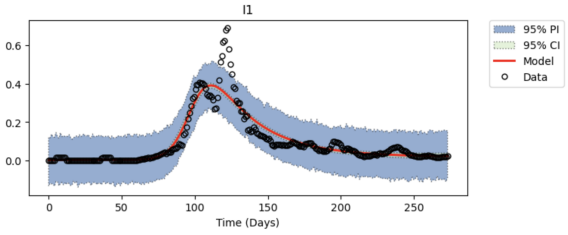


(e) I1

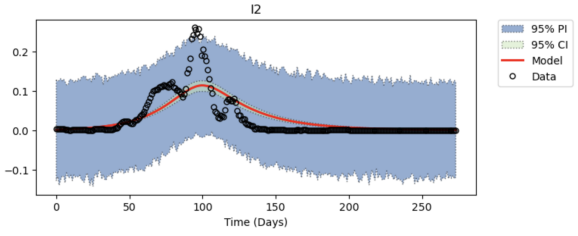


(f) I2

Figure 5.6: The different compartments for the second wave in India. E1 and I1 represent exposed and infected populations for the Delta variant, while, E2 and I2, represent the same for Kappa. The y-axis represents the fraction of the population (multiplied by 1000).

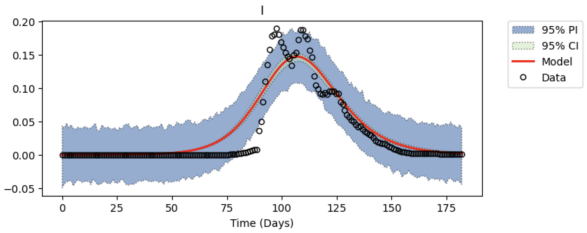


(a) I1

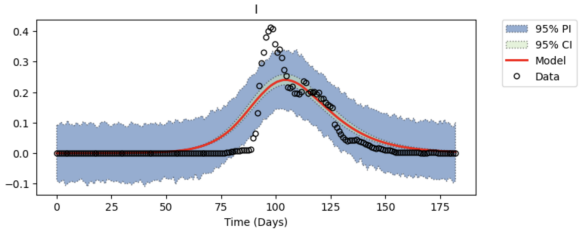


(b) I2

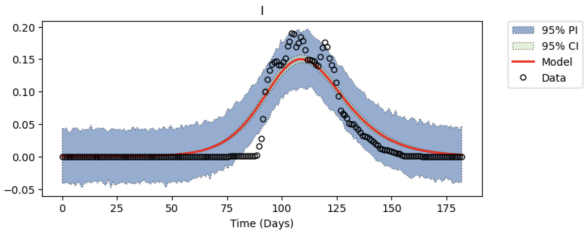
Figure 5.7: Infections I1 (Delta) and I2 (Kappa) for the second wave in Maharashtra. The y-axis represents the fraction of the population (multiplied by 1000).



(a) Infections in India



(b) Infections in Maharashtra



(c) Infections in Gujarat

Figure 5.8: Infections compartment for Omicron in the third wave in a) India b) Maharashtra c) Gujarat. The y-axis represents the fraction of the population (multiplied by 1000).

Table 5.1: Estimated parameters (using pymcmcstat) for the Second and Third waves in India.

Parameter	**Value in Second wave**	**Value in Third wave**
β_{delta}	0.35	-
β_{kappa}	0.12	-
$\beta_{omicron}$	-	0.38
σ	0.17	0.16
γ	0.08	0.13
δ	5.18 x 10^{-3}	5.46 x 10^{-4}
τ	54.32	-

Table 5.1 states the estimated values for the different parameters for the second and third waves in India. The estimated β for Delta (across the entire country, Maharashtra, and Gujarat) are 0.35, 0.39, and 0.37. For Kappa, β is 0.12, 0.10, and 0.19, and finally, for Omicron the estimated β is 0.38, 0.31, and 0.35 respectively. Therefore, we can infer that Kappa was very less transmissible compared to Delta and Omicron which had very similar high β values. Also while, the respective β values for Omicron and Delta seem similar, very few cases were reported in the third wave due to the majorly asymptomatic nature of the mild infection. Therefore, the actual β for Omicron may be even higher than predicted here. The R_0 values can be computed as $R_0 = \beta/\gamma$ which comes out to be 4.375 for Delta, 1.5 for Kappa, and 2.93 for Omicron.

Interestingly, the mild nature of the Omicron variant-induced infection can also be concluded from the γ parameter which represents the inverse of the infection period. For the second wave $1/\gamma$, or the estimated infection period, was nearly 2 weeks, whereas, for the third wave, this dropped to $\sim$1 week. This indicates that infections during the third wave (due to Omicron) lasted for a shorter duration than those during the second wave (mainly due to Delta).

Another very important factor to be considered is the large-scale vaccination drive conducted in India post the second wave. The drive started on 16th January 2021, and by May 2021 (the peak of the second wave in India) less than 20% of the eligible population received the first dose and less than 5% were fully vaccinated. However, by January 2022 (the peak of the third wave in India), nearly 90% of the eligible population had received the first dose and $\sim$70% were fully vaccinated. The immunity conferred to the individuals by the vaccines would have further reduced the cases of the Omicron variant resulting in the seemingly lower β value. The parameter $1/\delta$ denotes the duration required to become susceptible to reinfection after recovery. Since we did not incorporate vaccination data in our model, an increase in population vaccination would lead to a decrease in the probability of reinfections and hence an increase in the value of $1/\delta$. The estimated values of $1/\delta$ for the third wave indicate that the time to reinfection increased by a factor of $\sim$10 as compared to the second wave (where the immunity period was nearly 6 months), which is consistent with our initial hypothesis.

5.4 Conclusion and Future Work

In this chapter, we used the variant counts (using the sequences uploaded on GISAID), and scaled them using confirmed cases data to model the infection in India. We also repeated this for important states such as Maharashtra and Gujarat. Currently, we used multi-strain SEIR models for the second wave and a simple single-strain SEIR model for the third wave. We used the Python-based pymcmcstat package which efficiently searches the parameter space using the MCMC method. On fitting the SEIR model to the variant counts we estimated β, which models the transmission coefficient, for the different strains. The estimated β for Kappa (0.12) was significantly lower than that for Delta (0.35) and Omi-

cron (0.38). We then hypothesize that the actual β parameter for Omicron would be even higher than the one we predicted due to two crucial factors. First, the number of reported cases during the third wave were lower than the actual because the infection was milder and often asymptomatic. Second, during the peak of the third wave, almost 90% of the eligible population had received at least one vaccination dose, whereas, during the peak of the second wave, it was less than 20%. Significant changes could also be seen in other parameters such as γ: inverse of infection duration (increased for third wave), and δ: inverse of time to reinfection (decreased for third wave).

The work on this project is currently in progress. For future work, we intend to utilize the estimated parameters to predict COVID-19 transmission patterns across India using a network model built on the Airport data. Identifying crucial airport nodes can help devise effective interventions, including shutdowns, increased testing, and PPE supplies for passengers. Implementation of these interventions at the identified crucial airport nodes could decelerate the transmission of the virus, providing the government with additional time to respond. Such tactics may establish a precedent for responding to future pandemics.

Chapter 6

Conclusion and Future Directions

6.1 Summary of the Main Findings

In conclusion, this interdisciplinary study of demographic, structural, and epidemiological factors related to COVID-19 in India sheds light on several important aspects of the pandemic. The analysis of SARS-CoV-2 isolates across four phases of the pandemic has provided valuable insights into the spread and mutation of the virus in different regions of the country. It highlights the importance of implementing effective measures such as contact tracing, quarantine, and lockdown to control the spread of the virus, while also recognizing the limitations of these measures in the face of local transmission within states. Furthermore, we examined the impact of the Delta and Omicron variants on the severity of the second and third waves of the pandemic. The findings of the study emphasize the need for continued efforts to monitor the spread of SARS-CoV-2 in the country, identify potentially virulent strains, and take pre-emptive action to control their spread.

In addition to the extensive demographic analysis of SARS-CoV-2 isolates, this study also includes a thorough examination of the effects of India-specific mutations on the structure and function of the virus. By utilizing network and structural analysis techniques and incorporating insights from existing literature, we have identified the impact of these mutations on the protein's structure and function. Our analysis involved constructing protein residue networks and calculating topologically important centralities for both wild-type and mutant protein structures. Such analysis also gives direction to future experimental studies to ascertain the importance of key residues identified.

Finally, the thesis presents a modeling approach to estimate the transmission coefficients of different SARS-CoV-2 strains in India during the second and third waves of the pandemic. The estimated β for Delta and Omicron was higher than that for Kappa, suggesting increased transmission. However, given the milder and often asymptomatic nature of infections during the third wave and the significantly higher vaccination coverage, we hypothesize that the actual β for Omicron is even higher than the one estimated. We also observed changes in other parameters such as γ and δ, which could have significant

implications for the dynamics of the pandemic. These findings highlight the importance of continued modeling efforts to better understand the impact of the pandemic and inform public health strategies.

6.2 Limitations and Future Work

A major limitation of the work is not accounting for vaccination data in our models. As a future direction, we plan to utilize the estimated parameters obtained from our modeling approach to predict the transmission patterns of COVID-19 across India using a network model constructed from airport data. The identification of crucial airport nodes can help develop effective interventions, including shutdowns, increased testing, and adequate personal protective equipment (PPE) supplies for passengers. By implementing these interventions at the identified crucial airport nodes, we can potentially decelerate the transmission of the virus, providing the government with additional time to respond. This approach may establish a precedent for responding to future waves of COVID-19 or any other pandemic caused by a novel pathogen and can be a valuable tool for policymakers and public health officials in preparing for and mitigating the effects of future outbreaks.

Related Publications

1. **Agarwal, K.**, Parekh, N. (2021) Demographic Analysis of Mutations in Indian SARS-CoV-2 Isolates. Poster presented at the joint 29th Intelligent Systems for Molecular Biology Conference and the 20th European Conference on Computational Biology (ISMB/ECCB 2021). doi:10.1101/2021.09.22.461342.

   Full paper peer-reviewed by independent reviewers through PeerRef:

   (a) Prof. Hurng-Yi Wang, National Taiwan University

   (b) Dr. Jyotsnamayee Sabat, Indian Council of Medical Research

   (c) Dr. Parvin Abraham, MIMS Research Foundation

2. **Agarwal, K.**, Parekh, N. (2022) Comparative Analysis of SARS-CoV-2 Variants Across Three Waves in India. Accepted at the 3rd International Conference on Bioinformatics and Data Science (ICBDS 2022). To be published in Advances in Health Sciences Research (Springer Nature). Received the **Best Oral Presentation Award**.

Bibliography

[1] Shweta Alai et al. “Pan-India novel coronavirus SARS-CoV-2 genomics and global diversity analysis in spike protein”. In: *Heliyon* 7.3 (2021), e06564.

[2] Marwan Alfalah et al. “Compound heterozygous mutations affect protein folding and function in patients with congenital sucrase-isomaltase deficiency”. In: *Gastroenterology* 136.3 (2009), pp. 883–892.

[3] Katherine M Almasy, Jonathan P Davies, and Lars Plate. “Comparative host interactomes of the SARS-CoV-2 nonstructural protein 3 and human coronavirus homologs”. In: *Molecular & Cellular Proteomics* 20 (2021).

[4] Heba N Altarawneh et al. “Effect of prior infection, vaccination, and hybrid immunity against symptomatic BA. 1 and BA. 2 Omicron infections and severe COVID-19 in Qatar”. In: *MedRxiv* (2022), pp. 2022–03.

[5] Rambaut Andrew. “Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations”. In: *virological* (2020).

[6] Megan M Angelini et al. “Severe acute respiratory syndrome coronavirus nonstructural proteins 3, 4, and 6 induce double-membrane vesicles”. In: *MBio* 4.4 (2013), e00524–13.

[7] Megan Mary Angelini, Benjamin William Neuman, and Michael J Buchmeier. “Untangling membrane rearrangement in the nidovirales”. In: *DNA and cell biology* 33.3 (2014), pp. 122–127.

[8] Yusha Araf et al. “Omicron variant of SARS-CoV-2: genomics, transmissibility, and responses to current COVID-19 vaccines”. In: *Journal of medical virology* 94.5 (2022), pp. 1825–1832.

[9] Lee A Armstrong et al. “Biochemical characterization of protease activity of Nsp3 from SARS-CoV-2 and its inhibition by nanobodies”. In: *PloS one* 16.7 (2021), e0253364.

[10] Alam ASMRU et al. “Dominant clade-featured SARS-CoV-2 co-occurring mutations reveal plausible epistasis: An in silico based hypothetical model.” In: *Journal of Medical Virology* (2021).

[11] Mustak Ibn Ayub. “Reporting two SARS-CoV-2 strains based on a unique trinucleotide-bloc mutation and their potential pathogenic difference”. In: (2020).

[12] Peter Bager et al. “Reduced risk of hospitalisation associated with infection with SARS-CoV-2 omicron relative to delta: a Danish cohort study”. In: (2022).

[13] Yunmeng Bai et al. “Comprehensive evolution and molecular characteristics of a large number of SARS-CoV-2 genomes reveal its epidemic trends”. In: *International Journal of Infectious Diseases* 100 (2020), pp. 164–173.

[14] Anindita Banerjee et al. “The novel Coronavirus enigma: Phylogeny and mutation analyses of SARS-CoV-2 viruses circulating in India during early 2020”. In: *bioRxiv* (2020), pp. 2020–05.

[15] Sofia Banu et al. “A distinct phylogenetic cluster of Indian severe acute respiratory syndrome coronavirus 2 isolates”. In: *Open forum infectious diseases*. Vol. 7. 11. Oxford University Press US. 2020, ofaa434.

[16] Francisco Barona-Gómez et al. “Phylogenomics and population genomics of SARS-CoV-2 in Mexico during the pre-vaccination stage reveals variants of interest B. 1.1. 28.4 and B. 1.1. 222 or B. 1.1. 519 and the nucleocapsid mutation S194L associated with symptoms”. In: *Microbial Genomics* 7.11 (2021).

[17] Martina Bianchi et al. “SARS-Cov-2 ORF3a: mutability and function”. In: *International journal of biological macromolecules* 170 (2021), pp. 820–826.

[18] Samir Bolıvar et al. “IFN-β plays both pro-and anti-inflammatory roles in the rat cardiac fibroblast through differential STAT protein activation”. In: *Frontiers in pharmacology* 9 (2018), p. 1368.

[19] Phillip Bonacich. “Power and centrality: A family of measures”. In: *American journal of sociology* 92.5 (1987), pp. 1170–1182.

[20] Yana Bromberg, Guy Yachdav, and Burkhard Rost. “SNAP predicts effect of mutations on protein function”. In: *Bioinformatics* 24.20 (2008), pp. 2397–2398.

[21] Ngoc-Niem Bui et al. “The extent of molecular variation in novel SARS-CoV-2 after the six-month global spread”. In: *Infection, Genetics and Evolution* 91 (2021), p. 104800.

[22] Hüseyin Can et al. “In silico discovery of antigenic proteins and epitopes of SARS-CoV-2 for the development of a vaccine or a diagnostic approach for COVID-19”. In: *Scientific reports* 10.1 (2020), p. 22387.

[23] Michael A Casasanta et al. “Structural Insights of the SARS-CoV-2 Nucleocapsid Protein: Implications for the Inner-workings of Rapid Antigen Tests”. In: *Microscopy and Microanalysis* (2022).

[24] Broto Chakrabarty and Nita Parekh. “NAPS: network analysis of protein structures”. In: *Nucleic acids research* 44.W1 (2016), W375–W382.

[25] Chung-Ke Chang et al. “Multiple nucleic acid binding sites and intrinsic disorder of severe acute respiratory syndrome coronavirus nucleocapsid protein: implications for ribonucleocapsid protein packaging”. In: *Journal of virology* 83.5 (2009), pp. 2255–2264.

[26] Chung-ke Chang et al. “The SARS coronavirus nucleocapsid protein–forms and functions”. In: *Antiviral research* 103 (2014), pp. 39–50.

[27] Armi Chaudhari et al. “In-Silico analysis reveals lower transcription efficiency of C241T variant of SARS-CoV-2 with host replication factors MADP1 and hnRNP-1”. In: *Informatics in medicine unlocked* 25 (2021), p. 100670.

[28] Albert Tian Chen et al. “COVID-19 CG enables SARS-CoV-2 mutation and lineage tracking by locations and dates of interest”. In: *Elife* 10 (2021), e63409.

[29] Nanshan Chen et al. “Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study”. In: *The lancet* 395.10223 (2020), pp. 507–513.

[30] Xiaojuan Chen et al. “SARS coronavirus papain-like protease inhibits the type I interferon signaling pathway through interaction with the STING-TRAF3-TBK1 complex”. In: *Protein & cell* 5.5 (2014), pp. 369–381.

[31] Yongwook Choi and Agnes P Chan. “PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels”. In: *Bioinformatics* 31.16 (2015), pp. 2745–2747.

[32] Mateusz Chwastyk et al. “Properties of Cavities in Biological Structures—A Survey of the Protein Data Bank”. In: *Frontiers in Molecular Biosciences* 7 (2020), p. 591381.

[33] Zharko Daniloski et al. “The Spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types”. In: *Elife* 10 (2021), e65365.

[34] Nicholas G Davies et al. “Estimated transmissibility and impact of SARS-CoV-2 lineage B. 1.1. 7 in England”. In: *Science* 372.6538 (2021), eabg3055.

[35] Wanwisa Dejnirattisai et al. “Antibody evasion by the P. 1 strain of SARS-CoV-2”. In: *Cell* 184.11 (2021), pp. 2939–2954.

[36] EJJH Domingo and JJ Holland. “RNA virus mutations and fitness for survival”. In: *Annual review of microbiology* 51.1 (1997), pp. 151–178.

[37] Zongyang Du et al. “The trRosetta server for fast and accurate protein structure prediction”. In: *Nature protocols* 16.12 (2021), pp. 5634–5651.

[38] Siobain Duffy. “Why are RNA virus mutation rates so damn high?” In: *PLoS biology* 16.8 (2018), e3000003.

[39] Ernesto Estrada. “Quantifying network heterogeneity”. In: *Physical Review E* 82.6 (2010), p. 066102.

[40] Ernesto Estrada. “Topological analysis of SARS CoV-2 main protease¡? A3B2 show [editpick]?¿” In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 30.6 (2020), p. 061102.

[41] Ernesto Estrada and Naomichi Hatano. “Communicability angle and the spatial efficiency of networks”. In: *SIAM Review* 58.4 (2016), pp. 692–715.

[42] Ernesto Estrada and Naomichi Hatano. “Communicability in complex networks”. In: *Physical Review E* 77.3 (2008), p. 036111.

[43] Ernesto Estrada and Juan A Rodriguez-Velazquez. “Subgraph centrality in complex networks”. In: *Physical Review E* 71.5 (2005), p. 056103.

[44] Ernesto Estrada and Grant Silver. “Accounting for the role of long walks on networks via a new matrix function”. In: *Journal of Mathematical Analysis and Applications* 449.2 (2017), pp. 1581–1600.

[45] Nuno R Faria et al. “Genomics and epidemiology of the P. 1 SARS-CoV-2 lineage in Manaus, Brazil”. In: *Science* 372.6544 (2021), pp. 815–821.

[46] Yanghe Feng, Qi Wang, and Tengjiao Wang. “Drug target protein-protein interaction networks: a systematic perspective”. In: *BioMed research international* 2017 (2017).

[47] Neil M Ferguson et al. “Strategies for mitigating an influenza pandemic”. In: *Nature* 442.7101 (2006), pp. 448–452.

[48] Daniele Focosi and Fabrizio Maggi. “Neutralising antibody escape of SARS-CoV-2 spike protein: risk assessment for antibody-based Covid-19 therapeutics and vaccines”. In: *Reviews in medical virology* 31.6 (2021), e2231.

[49] David Foutch, Bill Pham, and Tongye Shen. “Protein conformational switch discerned via network centrality properties”. In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 3599–3608.

[50] Linton C Freeman et al. “Centrality in social networks: Conceptual clarification”. In: *Social network: critical concepts in sociology. Londres: Routledge* 1 (2002), pp. 238–263.

[51] Matthew Frieman et al. “Severe acute respiratory syndrome coronavirus papain-like protease ubiquitin-like domain and catalytic domain regulate antagonism of IRF3 and NF-κB signaling”. In: *Journal of virology* 83.13 (2009), pp. 6689–6705.

[52] Tjede Funk et al. “Characteristics of SARS-CoV-2 variants of concern B. 1.1. 7, B. 1.351 or P. 1: data from seven EU/EEA countries, weeks 38/2020 to 10/2021”. In: *Eurosurveillance* 26.16 (2021), p. 2100348.

[53] Oxana V Galzitskaya, Natalya S Bogatyreva, and Dmitry N Ivankov. “Compactness determines protein folding type”. In: *Journal of Bioinformatics and Computational Biology* 6.04 (2008), pp. 667–680.

[54] Yan Gao et al. “Structure of the RNA-dependent RNA polymerase from COVID-19 virus”. In: *Science* 368.6492 (2020), pp. 779–782.

[55] Sourish Ghosh et al. “β-Coronaviruses use lysosomes for egress instead of the biosynthetic secretory pathway”. In: *Cell* 183.6 (2020), pp. 1520–1535.

[56] Asmita Gupta et al. “A comprehensive profile of genomic variations in the SARS-CoV-2 isolates from the state of Telangana, India”. In: *The Journal of General Virology* 102.3 (2021).

[57] Nivedita Gupta et al. “Clinical characterization and genomic analysis of samples from COVID-19 breakthrough infections during the second wave among the various states of India”. In: *Viruses* 13.9 (2021), p. 1782.

[58] James Hadfield et al. “Nextstrain: real-time tracking of pathogen evolution”. In: *Bioinformatics* 34.23 (2018), pp. 4121–4123.

[59] Yu-San Han et al. “Papain-like protease 2 (PLP2) from severe acute respiratory syndrome coronavirus (SARS-CoV): expression, purification, characterization, and inhibition”. In: *Biochemistry* 44.30 (2005), pp. 10349–10359.

[60] Runtao He et al. “Characterization of protein–protein interactions between the nucleocapsid protein and membrane protein of the SARS coronavirus”. In: *Virus research* 105.2 (2004), pp. 121–125.

[61] Md Golzar Hossain et al. “Roles of the polybasic furin cleavage site of spike protein in SARS-CoV-2 replication, pathogenesis, and host immune responses and vaccination”. In: *Journal of Medical Virology* 94.5 (2022), pp. 1815–1820.

[62] Wei-Chen Hsin et al. “Nucleocapsid protein-dependent assembly of the RNA packaging signal of Middle East respiratory syndrome coronavirus”. In: *Journal of biomedical science* 25 (2018), pp. 1–12.

[63] Ben Hu et al. “Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus”. In: *PLoS pathogens* 13.11 (2017), e1006698.

[64] Sirawit Ittisoponpisan et al. “Can predicted protein 3D structures provide reliable insights into whether missense variants are disease associated?” In: *Journal of molecular biology* 431.11 (2019), pp. 2197–2212.

[65] Jobin John Jacob et al. “Evolutionary tracking of SARS-CoV-2 genetic variants highlights an intricate balance of stabilizing and destabilizing mutations”. In: *MBio* 12.4 (2021), e01188–21.

[66] DY Jin, BJ Zheng, and HMV Tang. “Mechanism of inflammasome activation by saRs coronavirus 3a protein: abridged secondary”. In: *Hong Kong Med J* 27.3 Supplement 2 (2021).

[67] Madhvi Joshi et al. “Genomic variations in SARS-CoV-2 genomes from Gujarat: Underlying role of variants in disease epidemiology”. In: *Frontiers in genetics* 12 (2021), p. 586569.

[68] Saathvik R Kannan et al. “Omicron SARS-CoV-2 variant: Unique features and their impact on pre-existing antibodies”. In: *Journal of autoimmunity* 126 (2022), p. 102779.

[69] Kazutaka Katoh et al. “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform”. In: *Nucleic acids research* 30.14 (2002), pp. 3059–3066.

[70] Matt J Keeling and Pejman Rohani. “Stochastic dynamics”. In: *Modeling infectious diseases in humans and animals*. Princeton University Press, 2011, pp. 190–231.

[71] David M Kern et al. “Cryo-EM structure of SARS-CoV-2 ORF3a in lipid nanodiscs”. In: *Nature structural & molecular biology* 28.7 (2021), pp. 573–582.

[72] Nathalie Kin et al. “Genomic analysis of 15 human coronaviruses OC43 (HCoV-OC43s) circulating in France from 2001 to 2013 reveals a high intra-specific diversity with new recombinant genotypes”. In: *Viruses* 7.5 (2015), pp. 2358–2377.

[73] Bette Korber et al. “Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus”. In: *Cell* 182.4 (2020), pp. 812–827.

[74] Joy-Yan Lam et al. “Loss of orf3b in the circulating SARS-CoV-2 strains”. In: *Emerging microbes & infections* 9.1 (2020), pp. 2685–2696.

[75] Marc Lavigne et al. “SARS-CoV-2 Nsp3 unique domain SUD interacts with guanine quadruplexes and G4-ligands inhibit this interaction”. In: *Nucleic acids research* 49.13 (2021), pp. 7695–7712.

[76] Jian Lei, Yuri Kusov, and Rolf Hilgenfeld. “Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein”. In: *Antiviral research* 149 (2018), pp. 58–74.

[77] Qun Li et al. “Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia”. In: *New England journal of medicine* (2020).

[78] Sanket Limaye et al. “Circulation and evolution of SARS-CoV-2 in India: let the data speak”. In: *Viruses* 13.11 (2021), p. 2238.

[79] Holger A Lindner et al. “Selectivity in ISG15 and ubiquitin recognition by the SARS coronavirus papain-like protease”. In: *Archives of biochemistry and biophysics* 466.1 (2007), pp. 8–14.

[80] Tiago JS Lopes et al. “Protein residue network analysis reveals fundamental properties of the human coagulation factor VIII”. In: *Scientific reports* 11.1 (2021), p. 12625.

[81] Mark Lorch et al. “Effects of mutations on the thermodynamics of a protein folding reaction: implications for the mechanism of formation of the intermediate and transition states”. In: *Biochemistry* 39.12 (2000), pp. 3480–3485.

[82] Hongzhou Lu, Charles W Stratton, and Yi-Wei Tang. “Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle”. In: *Journal of medical virology* 92.4 (2020), p. 401.

[83] Roujian Lu et al. “Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding”. In: *The lancet* 395.10224 (2020), pp. 565–574.

[84] Shan Lu et al. “The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein”. In: *Nature communications* 12.1 (2021), p. 502.

[85] Joseph H Lubin et al. “Evolution of the SARS-CoV-2 proteome in three dimensions (3D) during the first 6 months of the COVID-19 pandemic”. In: *Proteins: Structure, Function, and Bioinformatics* 90.5 (2022), pp. 1054–1080.

[86] Frederik Plesner Lyngse et al. “Transmission of SARS-CoV-2 Omicron VOC subvariants BA. 1 and BA. 2: evidence from Danish households”. In: *MedRxiv* (2022), pp. 2022–01.

[87] Shabir A Madhi et al. “Efficacy of the ChAdOx1 nCoV-19 Covid-19 vaccine against the B. 1.351 variant”. In: *New England Journal of Medicine* 384.20 (2021), pp. 1885–1898.

[88] Arindam Maitra et al. “Mutations in SARS-CoV-2 viral RNA identified in Eastern India: Possible implications for the ongoing outbreak in India and impact on viral structure and host susceptibility”. In: *Journal of Biosciences* 45 (2020), pp. 1–18.

[89] Paul S Masters. “Coronavirus genomic RNA packaging”. In: *Virology* 537 (2019), pp. 198–207.

[90] Ranjeet Maurya et al. “SARS-CoV-2 mutations and COVID-19 clinical outcome: mutation global frequency dynamics and structural modulation hold the key”. In: *Frontiers in Cellular and Infection Microbiology* (2022), p. 245.

[91] Priyanka Mehta et al. “Clinico-genomic analysis reveals mutations associated with COVID-19 disease severity: possible modulation by RNA structure”. In: *Pathogens* 10.9 (2021), p. 1109.

[92] Guangyan Miao et al. “ORF3a of the COVID-19 virus SARS-CoV-2 blocks HOPS complex-mediated assembly of the SNARE complex required for autolysosome formation”. In: *Developmental cell* 56.4 (2021), pp. 427–442.

[93] Duncan Milburn, Roman A Laskowski, and Janet M Thornton. “Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis.” In: *Protein engineering* 11.10 (1998), pp. 855–859.

[94] Bui Quang Minh et al. “IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era”. In: *Molecular biology and evolution* 37.5 (2020), pp. 1530–1534.

[95] Mainak Mondal, Ankita Lawarde, and Kumaravel Somasundaram. “Genomics of Indian SARS-CoV-2: Implications in genetic diversity, possible origin and spread of virus”. In: *Medrxiv* (2020), pp. 2020–04.

[96] Sai Narayanan et al. “SARS-CoV-2 Genomes From Oklahoma, United States”. In: *Frontiers in genetics* 11 (2021), p. 612571.

[97] Christian FA Negre et al. “Eigenvector centrality for characterization of protein allosteric pathways”. In: *Proceedings of the National Academy of Sciences* 115.52 (2018), E12201–E12208.

[98] Hafumi Nishi et al. “Cancer missense mutations alter binding properties of proteins and their interaction networks”. In: *PloS one* 8.6 (2013), e66273.

[99] Sarah P Otto et al. “The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic”. In: *Current Biology* 31.14 (2021), R918–R929.

[100] Anastasis Oulas et al. “Generalized linear models provide a measure of virulence for specific mutations in SARS-CoV-2 strains”. In: *PloS one* 16.1 (2021), e0238665.

[101] C Nick Pace et al. “Contribution of hydrogen bonds to protein stability”. In: *Protein Science* 23.5 (2014), pp. 652–661.

[102] Kartika Padhan et al. “Severe acute respiratory syndrome coronavirus Orf3a protein interacts with caveolin”. In: *Journal of General Virology* 88.11 (2007), pp. 3067–3077.

[103] Dhiraj Paul et al. “Phylogenomic analysis of SARS-CoV-2 genomes from western India reveals unique linked mutations”. In: *BioRxiv* (2020), pp. 2020–07.

[104] JSM Peiris et al. “Coronavirus as a possible cause of severe acute respiratory syndrome”. In: *The lancet* 361.9366 (2003), pp. 1319–1325.

[105] Neha Periwal et al. “In silico characterization of mutations circulating in SARS-CoV-2 structural proteins”. In: *Journal of Biomolecular Structure and Dynamics* 40.18 (2022), pp. 8216–8231.

[106] Douglas EV Pires, David B Ascher, and Tom L Blundell. “mCSM: predicting the effects of mutations in proteins using graph-based signatures”. In: *Bioinformatics* 30.3 (2014), pp. 335–342.

[107] Delphine Planas et al. “Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization”. In: *Nature* 596.7871 (2021), pp. 276–280.

[108] Jessica A Plante et al. “Spike mutation D614G alters SARS-CoV-2 fitness”. In: *Nature* 592.7852 (2021), pp. 116–121.

[109] Christelle Pommié et al. “IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties”. In: *Journal of Molecular Recognition* 17.1 (2004), pp. 17–32.

[110] Juliet RC Pulliam et al. “Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa”. In: *Science* 376.6593 (2022), eabn4947.

[111] Oliver G Pybus, Andrew J Tatem, and Philippe Lemey. “Virus evolution and transmission in an ever more connected world”. In: *Proceedings of the Royal Society B: Biological Sciences* 282.1821 (2015), p. 20142878.

[112] Sunil Raghav et al. “Analysis of Indian SARS-CoV-2 genomes reveals prevalence of D614G mutation in spike protein predicting an increase in interaction with TMPRSS2 and virus infectivity”. In: *Frontiers in Microbiology* 11 (2020), p. 594928.

[113] V Stalin Raj et al. “MERS: emergence of a novel human coronavirus”. In: *Current opinion in virology* 5 (2014), pp. 58–62.

[114] Andrew Rambaut et al. “A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology”. In: *Nature microbiology* 5.11 (2020), pp. 1403–1407.

[115] Safiya Richardson et al. “Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area”. In: *Jama* 323.20 (2020), pp. 2052–2059.

[116] Douglas D Richman, Richard J Whitley, and Frederick G Hayden. *Clinical virology*. John Wiley & Sons, 2020.

[117] Julien Roche et al. “Cavities determine the pressure unfolding of proteins”. In: *Proceedings of the National Academy of Sciences* 109.18 (2012), pp. 6945–6950.

[118] Carlos HM Rodrigues et al. “mCSM-PPI2: predicting the effects of mutations on protein–protein interactions”. In: *Nucleic acids research* 47.W1 (2019), W338–W344.

[119] Annika Rössler et al. “Neutralization profile of Omicron variant convalescent individuals”. In: *medRxiv* (2022), pp. 2022–02.

[120] Wioletta Rut et al. “Activity profiling and structures of inhibitor-bound SARS-CoV-2-PLpro protease provides a framework for anti-COVID-19 drug design”. In: *biorxiv* (2020), pp. 2020–04.

[121] Sebastian Salentin et al. “Polypharmacology rescored: Protein–ligand interaction profiles for remote binding site similarity assessment”. In: *Progress in biophysics and molecular biology* 116.2-3 (2014), pp. 174–186.

[122] Arnab Sarkar, Alok Kumar Chakrabarti, and Shanta Dutta. “Covid-19 infection in India: a comparative analysis of the second wave with the first wave”. In: *Pathogens* 10.9 (2021), p. 1222.

[123] Rakesh Sarkar et al. “Comprehensive analysis of genomic diversity of SARS-CoV-2 in different geographic regions of India: an endeavour to classify Indian SARS-CoV-2 strains on the basis of co-existing mutations”. In: *Archives of virology* 166 (2021), pp. 801–812.

[124] SeyedAhmad SeyedAlinaghi et al. “Characterization of SARS-CoV-2 different variants and related morbidity and mortality: a systematic review”. In: *European journal of medical research* 26.1 (2021), pp. 1–20.

[125] Aziz Sheikh et al. “Severity of Omicron variant of concern and vaccine effectiveness against symptomatic disease: national cohort with nested test negative design study in Scotland”. In: (2021).

[126] Yuelong Shu and John McCauley. “GISAID: Global initiative on sharing all influenza data–from vision to reality”. In: *Eurosurveillance* 22.13 (2017), p. 30494.

[127] Kam-Leung Siu et al. “Severe acute respiratory syndrome coronavirus ORF3a protein activates the NLRP3 inflammasome by promoting TRAF3-dependent ubiquitination of ASC”. In: *The FASEB Journal* 33.8 (2019), p. 8865.

[128] Misha Soskine and Dan S Tawfik. “Mutational effects and the evolution of new protein functions”. In: *Nature Reviews Genetics* 11.8 (2010), pp. 572–582.

[129] Silke Stertz et al. “The intracellular sites of early replication and budding of SARS-coronavirus”. In: *Virology* 361.2 (2007), pp. 304–315.

[130] Yee-Joo Tan et al. “A novel severe acute respiratory syndrome coronavirus protein, U274, is transported to the cell surface and undergoes endocytosis”. In: *Journal of virology* 78.13 (2004), pp. 6723–6734.

[131] Houriiyah Tegally et al. “Detection of a SARS-CoV-2 variant of concern in South Africa”. In: *Nature* 592.7854 (2021), pp. 438–443.

[132] Shaolei Teng et al. “Modeling effects of human single nucleotide polymorphisms on protein-protein interactions”. In: *Biophysical journal* 96.6 (2009), pp. 2178–2188.

[133] Dandan Tian et al. “The global epidemic of the SARS-CoV-2 delta variant, key spike mutations and immune escape”. In: *Frontiers in immunology* (2021), p. 5001.

[134] Matthew Z Tien et al. “Maximum allowed solvent accessibilites of residues in proteins”. In: *PloS one* 8.11 (2013), e80635.

[135] Nobuhiko Tokuriki and Dan S Tawfik. “Stability effects of mutations and protein evolvability”. In: *Current opinion in structural biology* 19.5 (2009), pp. 596–604.

[136] Ingrid Torjesen. *Covid-19: Omicron may be more transmissible than other variants and partly resistant to existing vaccines, scientists fear*. 2021.

[137] Ying-Tzu Tseng et al. “SARS-CoV envelope protein palmitoylation or nucleocapid association is not required for promoting virus-like particle production”. In: *Journal of biomedical science* 21.1 (2014), pp. 1–11.

[138] Shaun Tylor et al. “The SR-rich motif in SARS-CoV nucleocapsid protein is important for virus replication”. In: *Canadian journal of microbiology* 55.3 (2009), pp. 254–260.

[139] MHV Van Regenmortel. “Antigenicity and immunogenicity of viral proteins”. In: (2008).

[140] Robert Vaser et al. “SIFT missense predictions for genomes”. In: *Nature protocols* 11.1 (2016), pp. 1–9.

[141] Chen Wang et al. “A novel coronavirus outbreak of global health concern”. In: *The lancet* 395.10223 (2020), pp. 470–473.

[142] Rui Wang et al. “Characterizing SARS-CoV-2 mutations in the United States”. In: *Research square* (2020).

[143] Susan R Weiss and Sonia Navas-Martin. “Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus”. In: *Microbiology and molecular biology reviews* 69.4 (2005), pp. 635–664.

[144] Patrick CY Woo et al. “Discovery of seven novel Mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus”. In: *Journal of virology* 86.7 (2012), pp. 3995–4008.

[145] Siqi Wu et al. “Effects of SARS-CoV-2 mutations on protein structures and intraviral protein–protein interactions”. In: *Journal of medical virology* 93.4 (2021), pp. 2132–2140.

[146] Wenbing Wu et al. “The SARS-CoV-2 nucleocapsid protein: its role in the viral life cycle, structure and functions, and use as a potential target in the development of vaccines and diagnostics”. In: *Virology Journal* 20.1 (2023), pp. 1–16.

[147] Pragya D Yadav et al. “An epidemiological analysis of SARS-CoV-2 genomic sequences from different regions of India”. In: *Viruses* 13.5 (2021), p. 925.

[148] Weizhu Yan et al. “Structural biology of SARS-CoV-2: open the door for novel therapies”. In: *Signal transduction and targeted therapy* 7.1 (2022), p. 26.

[149] Jianyi Yang et al. “The I-TASSER Suite: protein structure and function prediction”. In: *Nature methods* 12.1 (2015), pp. 7–8.

[150] Mei Yang et al. “Structural insight into the SARS-CoV-2 nucleocapsid protein C-terminal domain reveals a novel recognition mechanism for viral transcriptional regulatory sequences”. In: *Frontiers in chemistry* 8 (2021), p. 624765.

[151] Xiaobo Yang et al. “Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study”. In: *The lancet respiratory medicine* 8.5 (2020), pp. 475–481.

[152] I-Mei Yu et al. “Crystal structure of the severe acute respiratory syndrome (SARS) coronavirus nucleocapsid protein dimerization domain reveals evolutionary linkage between corona-and arteriviridae”. In: *Journal of Biological Chemistry* 281.25 (2006), pp. 17134–17139.

[153] Weihong Zeng et al. “Biochemical characterization of SARS-CoV-2 nucleocapsid protein”. In: *Biochemical and biophysical research communications* 527.3 (2020), pp. 618–623.

[154] Jiantao Zhang et al. “Understanding the role of SARS-CoV-2 ORF3a in viral pathogenesis and COVID-19”. In: *Frontiers in microbiology* 13 (2022), p. 854567.

[155] Lizhou Zhang et al. “SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity”. In: *Nature communications* 11.1 (2020), p. 6013.

[156] Xiaolin Zhang et al. “SARS-CoV-2 ORF3a induces RETREG1/FAM134B-dependent reticulophagy and triggers sequential ER stress and inflammatory responses during SARS-CoV-2 infection”. In: *Autophagy* 18.11 (2022), pp. 2576–2592.

[157] Huaying Zhao et al. “Plasticity in structure and assembly of SARS-CoV-2 nucleocapsid protein”. In: *PNAS nexus* 1.2 (2022), pgac049.

[158] Ping Zhao et al. “Immune responses against SARS-coronavirus nucleocapsid protein induced by DNA vaccine”. In: *Virology* 331.1 (2005), pp. 128–135.

[159] Xingang Zhao, John M Nicholls, and Ye-Guang Chen. “Severe acute respiratory syndrome-associated coronavirus nucleocapsid protein interacts with Smad3 and modulates transforming growth factor-β signaling”. In: *Journal of Biological Chemistry* 283.6 (2008), pp. 3272–3280.

[160] Peng Zhou et al. “A pneumonia outbreak associated with a new coronavirus of probable bat origin”. In: *nature* 579.7798 (2020), pp. 270–273.