ORIGINAL RESEARCH

Phyloepidemiology and adaptive evolution of SARS-CoV2 during the first and second wave of COVID-19 in India

Authors: Anuj Tewari [1,*], Rajesh Kumar [1]

Abstract

Introduction. This study aimed to identify the circulating lineages of SARS-CoV-2, prevalent mutations in these lineages, and the selective pressure on the genome of SARS-CoV-2 during the first and second waves of the COVID-19 pandemic in India. **Methods**. We downloaded 1,451 sequences from June 2020 to June 2021 from the National Center for Biotechnology Information (NCBI) and the Global Initiative on Sharing All Influenza Database (GISAID). We identified the lineages using the Pangolin COVID-19 lineage assigner. **Results**. We found 41 circulating lineages in India during the year studied. Thirty-five lineages were circulating during the first wave and twenty during the second wave, including six new lineages. During the first wave in 2020, only one Variant of Concern (Alpha) was found, but during the second wave in 2021, three Variants of Concern (Alpha, Beta, and Delta) were in circulation, as well as one Variant Under Monitoring. The most frequent mutations observed were S: D614G, NSP3: F106F, NSP12b: P314L, ORF3a: Q57H, M: Y71Y, NSP14:C279C, S: D294D, and N: S194L. The ten most mutated samples all belonged to the Delta variant of B.1.617.2 lineage and were found in the second wave. Five mutations in the spike protein (L452R, T478K, E484Q, N501Y, and D614G), responsible for increased transmissibility and reduced neutralization by convalescent sera, were majorly prevalent during the second wave. D614G, L452R, and T478K were present at prevalence rates of 88.25%, 21.04%, and 16.80%, respectively. The major selection was purifying selection, but a few sites in the NSP2, NSP3, NSP13, S protein, ORF3a, and ORF9 evolved under positive selection. **Conclusion**. We report six novel mutations (three in NSP2 [P129A, V381A, V381F], one in NSP3 [P822S], and one in the S protein [Q23R]) that evolved under positive selection pressure.

Key word: lineages, variants of concern, variants under monitoring, selection pressure, non-synonymous substitution.

[1]Department of Veterinary Microbiology, College of Veterinary & Animal Sciences, India. Govind Ballabh Pant University of Agriculture & Technology, Pantnagar, Uttarakhand, India.
*https://orcid.org/0000-0002-6602-6014

Corresponding author:
Anuj Tewari
Address: Pantnagar, Uttarakhand, India.
E-mail: anuj474@gmail.com

Introduction

The severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), which was first identified in Wuhan City in the Hubei Province of China in December 2019, has now spread worldwide (1–3). In India, the first case of COVID-19 was identified on January 27th, 2020 in Kerala (4). During this ongoing pandemic, whole-genome sequencing (WGS) has been employed globally to gain insights into the geographical distribution of variants of SARS-CoV-2. As a result, WGS is a powerful tool to understand the mutations occurring across different segments of the viral genome and the selection pressure affecting various genes.

The B.1.617.1, B.1.617.2, and B.1.617.3 lineages of SARS-CoV-2 were first detected in India in December 2020 and have since become increasingly prevalent. These lineages are distinct and differ in their characteristic mutations (5). Therefore, it is crucial to understand the genetic variation among the different variants circulating in a country to discern mutation patterns among SARS-CoV-2 variants.

The World Health Organization has classified SARS-CoV-2 variants into three groups: "variants of interest" (VOI), "variants of concern" (VOC), and "variants under monitoring" (VUM) (https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/) (accessed on October 12th, 2021). The B.1.1.7 (Alpha), B.1.351 (Beta), P.1 (Gamma), and B.1.617.2 (Delta) Pangolin lineages are categorized as VOCs. C.37 (Lambda) and B.1.621 (Mu) are classified as VOIs. Some of the previously identified VOIs, such as B.1.617.1 (Kappa), B.1.526 (Iota), and B.1.525 (Eta), are now classified as VUMs.

The mutation rate of SARS-CoV-2 has been estimated to be approximately ~6×10^{-4} (CI: $4 \times 10^{-4} \sim 7 \times 10^{-4}$) substitutions/site/year (6). So far, SARS-CoV-2 typically acquires only one to two mutations per month (7), which is largely unremarkable for an RNA virus (6,8,9). However, some non-synonymous mutations, such as L452R, E484Q, P681R, and D614G, in the spike protein of SARS-CoV-2 may increase

transmissibility and generate virus escape mutants from convalescent sera (10–12). Due to the large number of genetic variants of SARS-CoV-2, the virus exists as a quasispecies, as revealed by next-generation sequencing and single nucleotide polymorphism (SNP) analysis (13,14).

Previous studies in India have assessed the mutational profile of SARS-CoV-2 during the second wave of COVID-19, which identified the presence of mutations such as L452R, T478K, E484Q, D614G, and P681R (15,16). It was also found that mutations such as L452R, T478K, and E484Q may result in increased ACE2 binding (15), and P681R may increase the rate of S1-S2 cleavage (15). As of March 25th, 2022, there have been 1,660 new cases of SARS-CoV-2 reported in India, (https://www.worldometers.info/coronavirus/country/india/)

It has been observed that human RNA viruses such as HIV-1, influenza viruses, SARS-CoV, and hepatitis C viruses may undergo adaptive changes to evade the host's immune system and establish an infection in a new host (17). The devastating second wave of COVID-19 in India has resulted in numerous deaths (18). Consequently, questions have arisen regarding the adaptive changes that SARS-CoV-2 has undergone during the pandemic's second wave in India.

SARS-CoV-2 infections across different countries have affected individuals with varying immunological strength, age, sex, and environmental conditions, imposing selection pressure on the virus. A previous study found that positive selection of ORF1ab, ORF3a, and ORF8 genes drove the early evolutionary trends of SARS-CoV-2 during the 2020 COVID-19 pandemic (19). Therefore, this study aimed to detect the genetic diversity and crucial mutational events of SARS-CoV-2 strains collected from different Indian states at different time points from June 2020 to June 2021. The study investigated the various lineages circulating during the pandemic's first and second wave, their mutations, and selection pressure across the SARS-CoV-2 genome.

Materials and methods

Genome sequences retrieval and alignment

In this study, we retrieved a total of 1451 complete and high-coverage SARS-CoV-2 viral genome sequences from the Global Initiative on Sharing All Influenza Database (GISAID) platform and National Center for Biotechnology Information (NCBI). To ensure the quality of the data, only sequences affecting human hosts were selected, and low-quality sequences with more than 5% NNNs and other ambiguous characters were removed using BioEdit 7.2 software. We included only full-length sequences with more than 29000 nucleotides, including the reference sequence MN908947.3, for further analysis. As a result, 1106 sequences were eligible for further analysis, after removing 344 sequences with >5% NNNs and other ambiguous characters.

Multiple sequence alignment and lineage prediction

To analyze the evolutionary trends of SARS-CoV-2 during the first and second waves of COVID-19 in India, we divided the samples into two groups: the first wave, which spanned from June 2020 to December 2020, and the second wave, which started from 15th January 2021 to June 2021, with a peak in April 2021, as reported by Worldometer (https://www.worldometers.info/coronavirus/country/india/)

To align the consensus sequences with the reference sequence of the SARS-CoV-2 Wuhan-Hu-1 isolate (GenBank accession number MN908947.3), we used Multiple Alignment using Fast Fourier Transform (MAFFT) (20). Additionally, we predicted the lineages using the Pangolin COVID-19 lineage assigner (https://pangolin.cog-uk.io/) (5).

Mapping nucleotide substitutions in the genome and expressed proteins of SARS CoV-2

The aligned sequences were mapped with the reference sequence of the SARS-CoV-2 Wuhan-Hu-1 isolate (GenBank accession number MN908947.3) to identify mutations across the genome using an online software tool (21).

Selection pressure amongst the lineages of SARS-CoV-2

To detect positive (diversifying) and negative (purifying) selection, a combination of different analysis methods was employed. Single-Likelihood Ancestor Counting (SLAC) (22), fixed effects likelihood (FEL) (22), and mixed effects model of evolution (MEME) (23) were used to analyze selection pressure using the datamonkey server. The results were expressed as dN/dS at a 0.1 level of significance. These methods utilize maximum likelihood to infer nonsynonymous (dN) and synonymous (dS) substitution rates on a per site basis for a given coding alignment and corresponding phylogeny (24). While SLAC and FEL identified sites experiencing pervasive diversifying or purifying selection, MEME detected sites experiencing both pervasive and episodic diversifying selection (25). The mutated sites exhibiting selective pressure were annotated using Molecular Evolutionary Genetics Analysis (MEGA 11) software.

Prediction of the effect of non-synonymous mutations under positive selective pressure on the functionalities of SARS-CoV-2

To predict the effect of non-synonymous mutations on the functionalities of SARS-CoV-2, the Protein Variation Effect Analyser (PROVEAN) software was used. This analysis assumes that protein sequences that are evolutionarily conserved among living organisms have survived natural selection (26). Mutations were predicted to have a deleterious effect if they had a threshold value below the default value of -2.5.

Phylogenomic analysis

A maximum likelihood tree was generated in the CIPRES Science Gateway using the generalized time reversible (GTR) nucleotide substitution model with 1000 bootstrap

values. The resulting tree was visualized using iTOL (https://itol.embl.de/) and MEGA 11. The phylogenetic tree expressing the Nextstrain clades was created using https://clades.nextstrain.org/.

Results

Out of the total of 1451 downloaded sequences, 344 sequences (23.7%) were removed due to the presence of >5% NNNs and other ambiguous characters. Hence, 1106 sequences (76.3%) were used for further analysis.

Table 1
Lineages of SARS-Co-V2 circulating during 2020 and 2021 in India

| S.No.	Lineages	2020	2021	Total
1	A.0	15	0	15
2	A.7	4	0	4
3	B.0	1	0	1
4	B.1	124	11	135
5	B.1.1	27	11	38
6	B.1.1.101	2	0	2
7	B.1.1.113	1	0	1
8	B.1.1.216	9	7	16
9	B.1.1.306	38	2	40
10	B.1.1.46	6	1	7
11	B.1.1.5	1	1	2
12	B.1.1.7(Alpha) (VOC)	6	69	75
13	B.1.113	75	0	75
14	B.1.143	25	0	25
15	B.1.145	5	0	5
16	B.1.153	1	1	2
17	B.1.184	13	0	13
18	B.1.2	2	0	2
19	B.1.210	34	0	34
20	B.1.222	1	0	1
21	B.1.247	41	0	41
22	B.1.351(Beta)(VOC)	0	8	8
23	B.1.36	56	27	83
24	B.1.36.10	1	2	3
25	B.1.36.16	0	1	1
26	B.1.36.22	0	1	1
27	B.1.36.24	2	1	3
28	B.1.36.29	3	13	16
29	B.1.36.38	1	0	1
30	B.1.36.8	156	9	165
31	B.1.465	4	0	4
32	B.1.523	1	0	1
33	B.1.523	1	0	1
34	B.1.540	2	3	5
35	B.1.617.1(Kappa)(VUM)	0	45	45
36	B.1.617.2(Delta)(VOC)	0	181	181
37	B.1.617.3	0	1	1
38	B.28	1	0	1
39	B.4.7	4	0	4
40	B.6	28	0	28
41	B.6.6	2	0	2
42	None	1	17	18

Lineages circulating during the first and second waves of COVID-19 in India

Lineages circulating during the first and second waves of COVID-19 in India have been comparatively evaluated. From June 2020 to December 2020 (first wave), a total of 35 lineages were in circulation, while from January 2021 to June 2021, a total of 20 lineages were in circulation (Table 1). In 2021 (second wave), fourteen lineages were common as in the first wave, and six new lineages were found to be circulating. Out of the six new circulating lineages, one was VUM B.1.617.1 lineage (Kappa variant), and the other two were VOCs B.1.351 lineage (Beta variant) and B.1.617.2 lineage (Delta variant). In the first wave, one case was from a virus of an undesignated lineage, and in the second wave, seventeen cases were from viruses of undesignated lineage.

In the first wave, six sequences were of VOC-B.1.1.7, whereas in 2021, sixty-nine sequences were found to be of the B.1.1.7 lineage, which was more than ten times higher than the number of B.1.1.7 cases found in the first wave. Another VOC, B.1.351 (Beta variant), did not exist during 2020, but in 2021, eight cases were found to be due to this VOC. The most drastic surge was for the VOC-B.1.617.2 lineage (Delta variant). This VOC did not exhibit any case in the first wave, whereas in the second wave, 181 cases were due to this lineage (Table 1, Fig. 1). At the same time, one VUM-B.1.617.1 lineage (Kappa variant) was also prevalent in the second wave, which was absent in the first wave, but in the second wave, forty-five cases were due to the B.1.617.1 lineage.

Figure 1
Bar diagram showing the prevalent Variants of concern (VOCs) and Variants under monitoring (VUM) during 2020 and 2021 as derived from the whole genome sequencing data from NCBI & GISAID. Frequency of samples of first wave are depicted in the year 2020 and of second wave are depicted in the year 2021
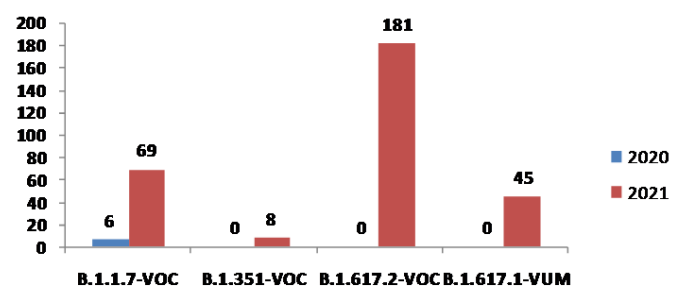

Apart from the VOCs and VUM, other lineages were also prevalent from June 2020-June 2021. During the first wave, the maximum number of cases (156 cases) were due to B.1.36.8 lineage, followed by the B.1 lineage (124 cases), B.1.113 lineage (75 cases), and B.1.36 lineage (56 cases). During the second wave, the maximum number of cases (181 cases) were due to VOC B.1.617.2 lineage (Delta variant), followed by VOC B.1.1.7 lineage (Alpha variant) (69 cases) and VUM B.1.617.1 lineage (Kappa variant) (45 cases) (Table 1, Fig. 1).

Mutational events across the genome segment of SARS-CoV-2 lineages

In this study, we analyzed the mutational events across the genome segment of SARS-CoV-2 lineages circulating in India from June 2020 to June 2021. Out of the 1451 downloaded sequences, 344 were removed due to ambiguous characters and NNNs, leaving 1106 sequences for further analysis. The ten most mutated samples belonged to the Delta variant of B.1.617.2 lineage, all of which were found in the second wave, suggesting the severity of the second wave compared to the first (Fig. 2).

Figure 2
Ten most mutated samples during one year time period from June 2020 to June 2021
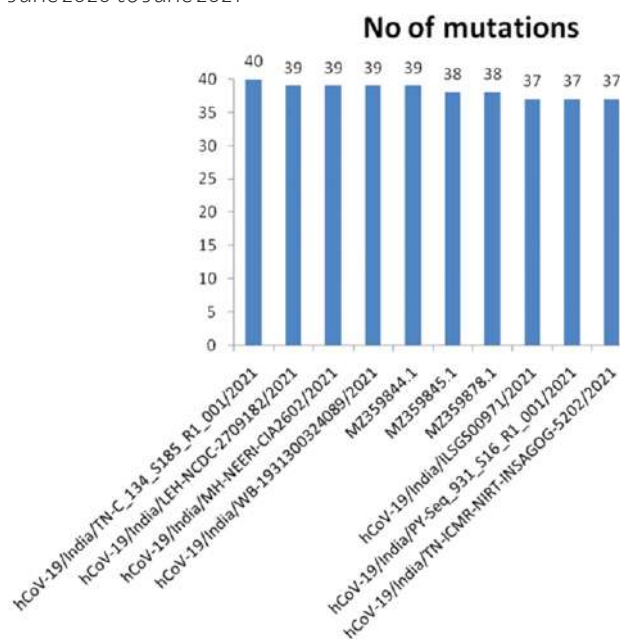
We identified a total of 1031 mutational events from June 2020 to June 2021. The most frequent events per class in decreasing order of their occurrence were SNP, SNP_silent, extragenic, SNP_stop, deletion_frameshift, insertion_frameshift. The most common nucleotide mutations were A23403G, C3037T, C241T, C14408T, C25563T, C26735T, C18877T, C22444T, C28854T, G210T (Fig. 3.a), while the most frequent protein mutations were S: D614G, NSP3:F106F, NSP 12b: P314L, ORF3a: Q57H, M: Y71Y, NSP14C279C, S: D294D, N: S194L (Fig. 3.b).

Prevalence of L452R, T478K, E484Q, N501Y, and D614G mutations in the RBD of SARS-CoV-2

We also investigated the prevalence of mutations in the receptor-binding domain (RBD) of SARS-CoV-2. Out of the 1106 samples analyzed, 1050 had the D614G mutation, with a prevalence rate of 94.93%. The L452R mutation was found in 233 samples (21.06% prevalence rate), T478K in 186 samples (16.81% prevalence rate), N501Y in 96 samples (8.77% prevalence rate), and E484Q in 47 samples (4.24% prevalence rate) (Table 2, Fig. 3C). The frequency of all mutations was significantly higher in the second wave, except for D614G, which was present in both waves and was the most frequent

mutational event observed. L452R and T478K were not present in the first wave but were found in significantly higher numbers during the second wave (Table 2). E484Q and N501Y were sporadically present in the first wave but were present in significantly higher numbers in the second wave.

Figure 3
Frequency of various mutations at nucleotide level and at protein level

Figure 3.a
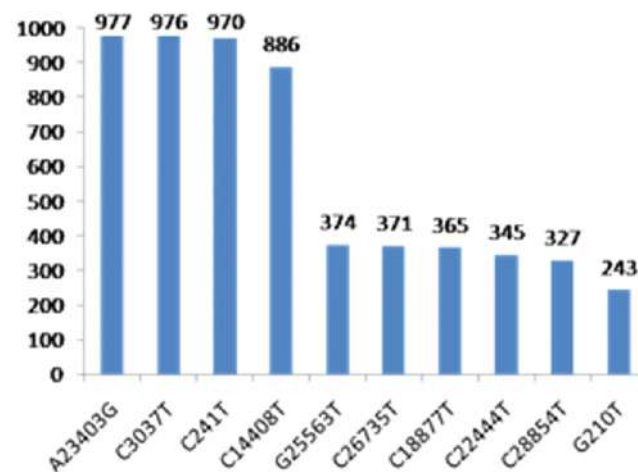*Shows most frequent events occuring at nucleotide level*

Figure 3.b
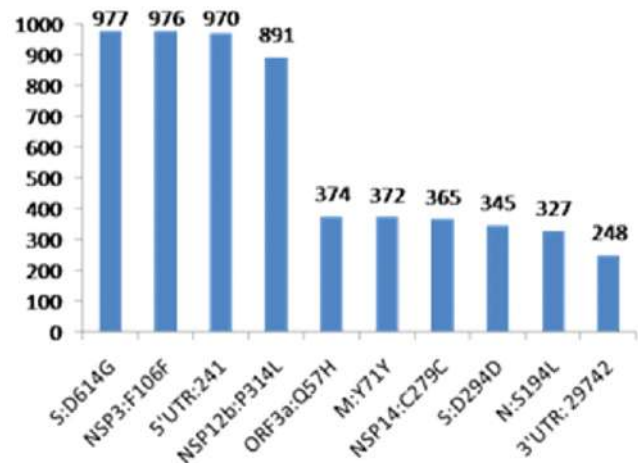*Most frequent events occuring at protein level*

Figure 3.c
Frequency of spike protein mutations responsible for higher transmissibility and reduction in neutralisation by convalescent sera
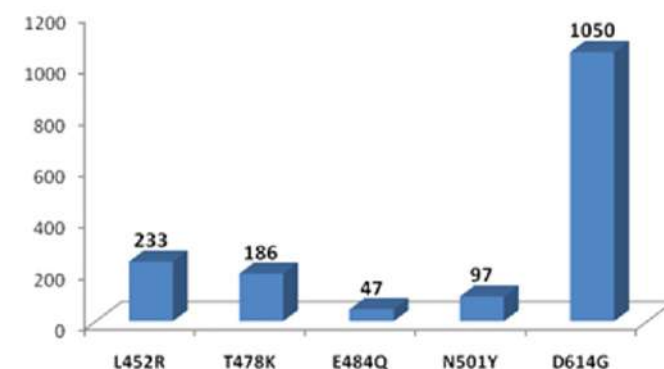
Table 2

Percentage prevalence of the mutations in RBD of spike protein responsible for increased virulence and virus escape mutants from convalescent sera (Total samples=1106)

Protein mutations	Frequency First wave	Frequency Second wave	Overall frequency	Percentage prevalence
L452R	0	233	233	21.06%
T478K	0	186	186	16.81%
E484Q	2	45	47	4.24%
N501Y	7	90	97	8.77%
D614G	638	412	1050	94.93%

Mutations in VOC, VUM, and their comparison with other lineages

The genetic sequences were classified into VUMs, VOCs, and other lineages. A comparison of the mutational events occurring in these groups was presented in Table 2. The most common mutational event observed at the nucleotide level was A23403G, while the S:D614G substitution was the most common at the protein level in the spike protein. Synonymous mutations were observed in NSP3, M, NSP14, and S, which did not result in changes in the amino acid sequence. However, non-synonymous mutations in S, NSP12b, ORF3a, and N indicated that these proteins were under constant selection pressure.

Three types of VOCs were prevalent among all circulating lineages, including the B.1.1.7 (alpha variant), B.1.351 (beta variant), and B.1.617.2 (delta variant) lineages, which accounted for 23.86% of the total sequences analyzed. The most common mutational events shared by all three VOCs were 5'UTR: 241 and NSP3: F106F. Mutations were also observed in S, 5'UTR, ORF3a, ORF7a NSPs, M, E, and N proteins, with the Delta variant having the three most frequent substitutions in the Spike protein (D614G, L452R, and T478K). The Kappa variant had two non-synonymous mutations (L452R & D614G) in the spike protein. The alpha variant exhibited only one mutation, S: D614G, in the spike protein, while the beta variant did not show any mutation in the spike protein. The NSPs of the alpha and beta variants had six mutations, with four and two being silent, respectively. The Delta variant exhibited only one silent mutation in the NSP3:F106F. Mutations in the N protein varied among all three VOCs (Table 3).

One VUM, the B.1.617.1 lineage, was identified, which had two mutations (L452R & D614G) in the spike protein and five mutations in the NSP (Table 3). Among these five mutations, two were silent. The mutations in the spike protein of VUM were D614G and L452R.

A comparative analysis was also performed between other lineages versus VOCs/VUMs. Unique mutations that were found in other lineages but absent in VOCs/VUMs included M: Y71Y, ORF3a: Q57H, NSP14:C279C, S: D294D, N: S194L, and NSP3:C39C. Mutations common in other lineages and VUM/VOCs were S: D614G, NSP3: F106F, and NSP12b:

P314L. The highest percentage of non-synonymous mutations was observed in the Delta variant (88.89%), followed by the Kappa (75.00%), beta (75.00%), alpha (44.45%), and other lineages (44.45%) (Table 3).

Table 3

Most frequent mutational events in the Variants of concern (VOC) and Variants under monitoring (VUM)

SNo	Lineage type	Number of samples	VOC/VOI type	Most frequent nucleotide mutational events	Most frequent protein mutational events	Synonymous and non synonymous mutations percentage
1	B.1.1.7	75	VOC-Alpha variant	A234303G C241T C3037T A23063T A28111G C14676T C23604A C23709T C27972T C28977T	5'UTR:241 NSP3:F106F S:D614G N:S235F NSP12b:P403P NSP12b: T903T NSP2: S36S NSP3: A890D NSP3:F1089F NSP3:I1412T	Synonymous mutation:5/9 x 100=55.55% Non-synonymous mutation: 4/9x100=44.45%
2	B.1.351	8	VOC-Beta variant	A10323G A21801C A22206G A23063T A23403G A2692T C1059T C14408T C23664T C241T	5'UTR:174 5'UTR:241 E:P71L N:T205I NSP12b:P314L NSP2:T629T NSP2:T85I NSP3:F106F NSP3:K837N NSP5:K90R	Synonymous Mutation: 2/8x100=25% Non-synonymous mutation: 6/8x100=75%
3	B.1.617.2	181	VOC-Delta variant	C3037T T26767C A23403G C22995A C25469T G28881T T22917G C241T C27752T G29402T	5'UTR:241 M:I82T NSP3:F106F N:R203M ORF3A:S26L S:D614G S:L452R S:T478K N:D377Y ORF7a:T120I	Synonymous Mutation: 1/9 x 100=11.11% Non-synonymous mutation: 8/9x100=88.89%
4	B.1.617.1	45	VUM-Kappa variant	A11201G A20396G C241T C3037T G210T T22917G A23403G T27638C A24775T C25469T	5'UTR:210 5'UTR:241 NSP15:K259R NSP3:F106F NSP6:T77A S:L452R ORF7a:V82A S:D614G NSP3:T749I NSP3:Y246Y	Synonymous Mutation: 2/8 x 100=25% Non-synonymous mutation: 6/8x100=75%
5	B.1.617.3	1	Nil	A23403G C11322T C14408T C16293T C21618G C23604G C241T C27970T C28472T C3037T	3'UTR:2720 3'UTR:29742 5'UTR:210 5'UTR:241 N:D377Y N:P67S N:R203M NSP12b:P314L NSP13:C19C NSP2:F10F	2/6 x 100 =33.33 % 4/6 x 100=66.67 %
6	Other lineages	796	Nil	A23403G C3037T C241T C14408T C26735T C18877T G25563T C2244T C28854T C2836T	S:D614G NSP3: F106F 5'UTR:241 NSP12b:P314L M:Y71Y ORF3a:Q57H NSP14:C279C S:D294D N:S194L NSP3:C39C	5/9 x 100=55.55% 4/9 x 100=44.45%

Selection pressure amongst the various lineages of SARS CoV-2

A codon-based selection pressure analysis was performed to identify purifying (negative) selection (dN < dS) and positive selection (dN > dS). All sequences were analyzed through the Datamonkey server, and after the removal of stop codons, sites under positive selection with a significance level < 0.1 were annotated (Table 4, Fig. 4a-f). The annotation of sites under positive selection has been provided in Figure S1-S4. The annotation was performed by comparing the substitution with the reference sequence using MEGA 11.

Table 4

Mutations and mutational sites in the SARS CoV-2 evolving under positive selection pressure. dN denontes non-synonymous mutation. dS denotes synonymous mutation. P value is the level of significance for dN/dS >1. The effect of mutation on the SARS CoV-2 is denoted as deleterious or neutral using COVEAN

Region	Site	P[dN/dS]>1	Mutation	Effect
NSP2	129	0.076	P129L	Deleterious
			P129S	Deleterious
			P129A	Deleterious
	381	0.008	V381A	Neutral
			V381F	Neutral
NSP3	822	0.039	P822L	Neutral
			P822S	Neutral
NSP13	77	0.0012	P77L	Deleterious
S protein	23	0.033	Q23R	Neutral
	54	0.042	L54F	Neutral
	95	0.0623	T95I	Neutral
	142	0.0026	G142D	Neutral
			G142S	Neutral
	484	0.0738	E484D	Neutral
			E484Q	Neutral
			E484K	Neutral
	681	0.0454	P681H	Neutral
			P681R	Neutral
ORF3a	26	0.0021	S26L	Neutral
ORF9	13	0.0878	P13L	Neutral
			P13S	Neutral
	194	0.00217	S194L	Deleterious

Using SLAC, a broad picture of the extent of natural selection acting upon the SARS-CoV-2 genome was obtained. Positive selection was present in NSP2 (site numbers 129 and 381), NSP3 (site number 822), NSP13 (site number 77), S protein (site numbers 23, 54, 95, 142, 484, and 681), ORF3a (site number 26), ORF9 (site numbers 13 and 194), and S protein (site number 23) (Table 4). Positive selection was present at a significance level < 0.10 in all these sites. NSP2 site number 129 had mutations P129L, P129S, and P129A. Site number 381 had mutations V381A and V381F. NSP3 had mutations P822L and P822S. NSP13 had a mutation as P77L. S protein exhibited mutations Q23R, L54F, T95I, G142D, G142S, E484D, E484Q, E484K, P681H, and P681R. ORF3a had a mutation S26L. ORF9 had mutations P13L, P13S, and S194L (Table 4). Overall, ORF3a (dN/dS = 1.6) and N (dN/dS = 1.02) were found to be evolving under positive selection (dN/dS > 1). Interestingly, the dN/dS ratio for the Envelope gene was found to be 0.911. Out of the total mutations evolving under positive selection, we found five novel substitutions, out of which three substitutions were in NSP2 (P129A, V381A, V381F), one in NSP3 (P822S), and one in the S protein (Q23R).

In NSP2, NSP3, and NSP13, the P to L substitution was significantly higher in the second wave. Some mutations, such as NSP3: P822L, NSP13: P77L, S: G142D, S: E484K, and ORF3a: S26L, were unique to the second wave and absent in the first wave (Table 5). Mutations NSP2: P129A, NSP2: V381F, NSP3: P822S, S: L54F, and S: E484D were present in the first wave but absent in the second wave. Other mutations were present

in both the first and second waves. Interestingly, the highest number of substitutions were substitutions by L (Leucine), D (Aspartic acid), and R (Arginine) amino acids in NSP2, NSP3, NSP13, S, ORF3a, and ORF9 (Table 5).

Table 5

*Frequencies of mutations in first and second wave evolving under positive selection pressure. Mutations frequency higher in second wave are marked with * sign in front. Mutations frequency higher in first wave are marked with + sign*

Region	Site	Mutation	First wave	Second wave
NSP2	129	P129L *	1	81
		P129S +	9	4
		P129A +	2	0
	381	V381A +	55	33
		V381F +	1	0
NSP3	822	P822L *	0	143
		P822S +	2	0
NSP13	77	P77L *	0	180
S protein	23	Q23R *	12	16
	54	L54F +	109	0
	95	T95I *	6	22
	142	G142D *	0	190
		G142S +	0	1
	484	E484D +	7	0
		E484Q *	2	45
		E484K *	0	8
	681	P681H *	17	82
		P681R *	2	227
ORF3a	26	S26L*	0	227
ORF9 (N protein)	13	P13L +	32	1
		P13S *	1	9
	194	S194L +	293	54

The SARS-CoV-2 genomes underwent analysis using the FEL method, which showed that purifying selection was uniformly distributed throughout the genome (refer to Table S1). Diversifying selection was observed in several locations, including NSP1 (1 site), NSP2 (3 sites), NSP3 (3 sites), NSP6 (1 site), NSP12 (1 site), NSP16 (2 sites), ORF3A (6 sites), N (9 sites), and the S protein (11 sites) (refer to Table S1).

Additionally, the sequences were analyzed using the MEME method to assess Episodic Positive/diversifying selection. This type of selection was identified in NSP2 (1 site), NSP3 (3 sites), NSP6 (1 site), NSP12 (1 site), NSP15 (1 site), NSP16 (2 sites), ORF3A (6 sites), M (1 site), ORF7A (1 site), ORF8 (1 site), ORF9 (7 sites), and the S protein (12 sites) (refer to Table S2).

Figure 4.

*Selective pressure analysis of SARS CoV-2 genome segments coding for NSP2 (4.a), NSP3 (4.b), NSP13 (4.c) ORF3a (4.d), ORF9 (4.e) and S (4.f) using SLAC analysis. Sites under positive selection are shown in the upper panel and sites under negative selection are shown in the lower panel. Sites under positive selective pressure with P value <0.10 are marked with * sign. The number with * sign denotes the site number under positive selection (4.g) Representation of SARS CoV-2 genome. Arrow shows the regions under positive selection pressure. In ORF 1ab - NSP2, NSP3 & NSP13 are under positive selection*

4.a-NSP2

4.b-NSP3

4.c-NSP13

4.d-ORF3a

4.e-ORF9

4.f-S protein

ORF 1ab - NSP2, NSP3 & NSP13 are under positive selection

4.g

Prediction of non-synonymous mutations under positive selective pressure on the functionalities of SARS-CoV-2.

Upon analyzing the effect of mutations on the fitness of the virus, NSP2: P129L, P129S, P129A; NSP13: P77L; ORF9: S194L were found to have a deleterious effect. The remaining mutational events that evolved under positive selection pressure had a neutral effect on the fitness of the virus.

Phylogenomic analysis

The 1106 genomes of SARS-CoV-2 underwent clade and phylogenetic analysis using Nextclade v1.14.0. The analysis identified twelve Nextstrain clades: 19A, 19B, 20A, 20B, 20C, 20G, 20H, 20I, 21B, 21A, 21I, and 21J. Out of these, 21J, 21I, and 21A were categorized under the Delta lineage, 21B as Kappa, 20I as alpha V1, and 20H as Beta V2 (Fig. 5). Upon constructing a Fast tree, we observed that the tree initially divided into three clades, with MN908947.3 (reference strain) in the first clade, and the second and third clades clustering the remaining sequences (Figure S5). Some sequences did not fall into any of the clusters upon subjecting them to clustering. These sequences belonged to lineages B, B.1, B.6, B.28, B.1.1, B.1.113, B.1.1.216, B.1.1.7, B.1.113, B.1.143, B.1.210, B.1.36, B.1.36.29, B.1.36.8, B.1.617.1, B.1.617.2, and B.1.617.3 (Figure S6).

Figure 5

Phylogenetic analysis of SARS-CoV-2 genomes using Nextclade v1.14.0. Twelve Nextstrain clades viz. 19A, 19B, 20A, 20B, 20C, 20G, 20H, 20I, 21B, 21A, 21I and 21J were identified by the phylogenetic analysis
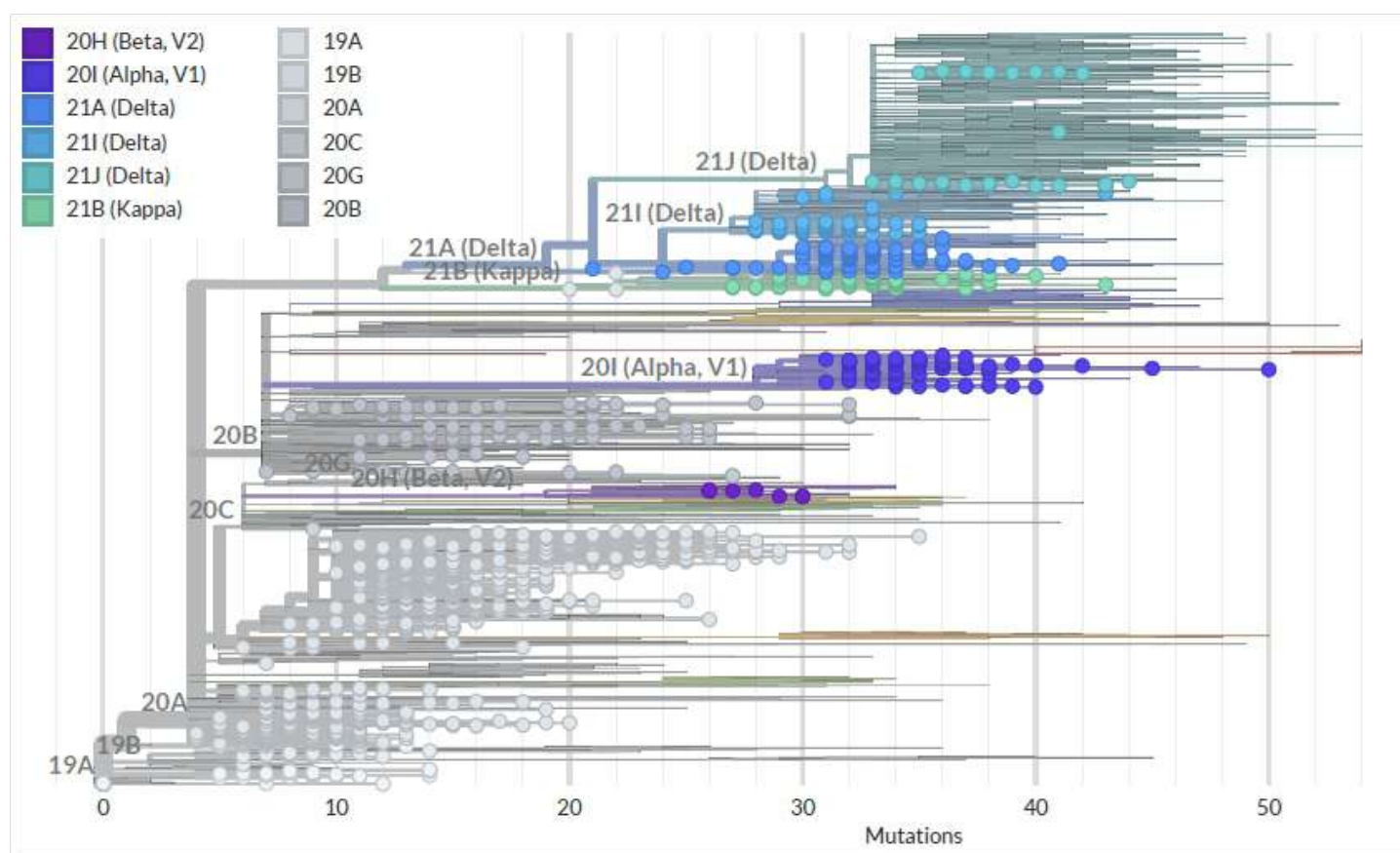
Discussion

In the last two years, SARS-CoV-2 has caused devastating losses to both human and economic resources. Therefore, antiviral drugs have been used as a therapeutic and vaccination as a prophylactic measure to combat the COVID-19 pandemic. Antiviral antibodies elicited by vaccination can be used to combat COVID-19 viral infections, but they can also increase the risk of developing resistant strains due to the rapid mutations of viruses (27). The resistance imparted due to selective pressure is remarkable, as seen in the case of G614 which has become common (28). The viruses may also develop antiviral drug resistance due to the extreme use of antiviral therapeutics, which is one of the factors that imparts selective pressure to the viral genome. Therefore, it is imperative to study the viral genome at regular periods when a novel virus is introduced into a population, followed by extreme use of vaccination and antiviral therapeutics, as seen in the current COVID-19 pandemic. Hence, the present study aimed to determine the prevalent SARS-CoV-2 lineages, their mutations, and selection pressure during the first and second waves of COVID-19 in India.

According to our study, amongst all VOCs, the maximum number of cases in the second wave were predominantly of B.1.617.2 lineage (Delta variant), whereas no cases of B.1.617.2 lineage (Delta variant) were found in the first wave. The movement of people, lifting of lockdowns, public gatherings, and festivals could be reasons why the cases of B.1.617.2 lineage (Delta variant) rose significantly. During the second wave, the B.1.617.2 lineage (Delta variant) was the predominant variant in all outbreaks. Additionally, the cases

from Variants Under Monitoring B.1.617.1 lineage (Kappa variant) and Variants Of Concern (B.1.1.7, B.1.351 & B.1.617.2 lineages) were significantly higher in 2021 than in 2020. This could be because unlocking and public gatherings might have flared up the dissemination of the virus strains. The ten most frequent mutational events observed were exhibited by the B.1.617.2 lineage (Delta variants), which shows the high genetic variance during the second wave. The beta, delta, and kappa variants exhibited a high percentage (75%) of non-synonymous mutations. The higher percentage of the non-synonymous mutations in the beta, delta, and kappa variants shows that these variants are still evolving under constant positive selection pressure and pose a threat to the future. Moreover, all three variants - Delta, Alpha, and Kappa - exhibited mutations in the receptor-binding domain.

SARS-CoV-2 exhibits mutation and diversification as a result of the selection pressure imparted on it (29). However, most mutations do not impart a selective advantage to the virus. Still, some of them may provide a selective advantage to the virus, such as increased transmissibility due to increased receptor binding ability or escaping the host's immune system by conferring changes in the viral structure recognized by antibodies (30). Mutations L452R, E484Q, D614G, and N501Y, responsible for high transmission ability due to increased binding with the receptor and host immune escape, were also present. These mutations were present in significantly higher numbers in the second wave compared to the first wave. The L452R mutation is associated with increased transmission of SARS-CoV-2 and a reduction in neutralization by convalescent plasma and therapeutic antibodies (10). The E484Q mutation (found only in the B.1.617.1 and B.1.617.3 lineages) is associated with a reduction in neutralization by convalescent sera (12), and the D614G mutation is associated with increased transmissibility and is carried by the majority of currently circulating viruses (11).

In a study, it has been reported that serum derived from vaccinated individuals was able to neutralize the B.1.617.1 and B.1.1.7 lineages, with a two-fold reduction in neutralization compared to the vaccine strain virus B.1 (31). In a recent study (32), the major vaccine breakthrough variant was the Delta variant (B.1.617.2 lineage), and the breakthrough events were P384L, K417N, E484K, and N501Y. In our study, the major dominant mutations and co-mutations in the receptor-binding domain (RBD) were L452R, T478K, N501Y, and E484K. These mutations were predicted to be dominant in the near future in earlier studies (32).

The host's immune system exerts constant selective pressure on the virus. Therefore, to study the selective pressure on the SARS-CoV-2 genome, the sequences were subjected to selective pressure analysis. On subjecting the sequences to selective pressure analysis, a few sites were found to be evolving under positive selection. NSP2, which acts to suppress the immune system of the host, has been described. The function of NSP2 is not entirely known, but it is thought to associate with the endosome of the host and disrupt the host cell environment (33). It is also a highly conserved protein among coronaviruses. NSP2 and NSP3 of SARS-CoV are detected not only as matured processed proteins but also as NSP2 and NSP3 precursors (34–36), suggesting the role of precursors in replication. NSP2 interacts with PHB1 and PHB2 host protein complexes, which are involved in mitochondrial biogenesis (37). In our study, we found two sites (site numbers 129 and 381) in NSP2 evolving under positive selection.

NSP3 has two significant jobs. The first is cutting other viral proteins to free them to do their tasks, and the second is removing tags from old proteins that are set for destruction. The removal of tags from old and damaged proteins by NSP3 changes the balance of proteins, thus possibly compromising the cell's ability to fight the virus (38). NSP3 had amino acid substitutions P822L and P822S. The NSP13 helicase had an amino acid substitution P77L evolving under positive selection pressure. The helicase separates double-stranded RNA or DNA with a 5′ 3′ polarity. Besides the helicase activity on double-stranded DNA and RNA, it is also capable of unwinding RNA/DNA duplex. Moreover, it has NTPase activity as well as 5′ mRNA capping activity (39). However, the above substitution was found to have a deleterious effect on virus fitness.

In the spike protein, three mutations (E484D, E484Q, and E484K) were identified at site number 484, which are under positive selection pressure. All these mutations are present in the receptor-binding motif (RBM) of the receptor binding domain (RBD) of the spike protein, which binds to the human ACE-2 receptor. The RBD is a 273 amino acid segment of the whole spike protein, stretching from residue 319 to 591. Ten substitutions were found in the spike protein that is evolving under positive selection pressure, indicating that the SARS-CoV-2 genome is still evolving to evade the host's immune system. Therefore, the emergence of a highly virulent novel variant cannot be ruled out.

ORF3a is involved in the high expression of cytokines and chemokines that cause cell death and is a high pathogenicity factor. Site number 26 of ORF3a had a non-synonymous mutation (S26L) evolving under positive selection, indicating the high virulence of these strains.

ORF9 modulates the host immune response by compromising type I IFN synthesis, interferon signaling, antigen processing and presentation, complement signaling, and inducing IL-6 signaling. Previous studies have suggested that high induction of IL-6 is related to increased viral virulence. Biological activities affected by the production of IL-6 include regulating the expression of macrophage colony-stimulating factor, increasing B-cell IgG production, negatively regulating dendritic cell maturation, and promoting the Th2 response. Site numbers 13 and 194 of ORF9 had non-synonymous mutations evolving under high pressure.

Conclusion

Based on the above facts, it can be inferred that the second wave was more devastating due to variants of concern (VOCs) with high transmissibility. The mutations

present in the spike protein of SARS-CoV-2 made the virus highly transmissible, and the mutations in the NSPs made the host's body more vulnerable to infection by the virus, followed by secondary infection by bacteria and fungus. This may be the possible reason that the second wave had a sharper peak than the first wave, which had a broader peak due to the slow transmission of the virus. The study suggests that viral components responsible for immunosuppression, entry inside the host cell, viral replication, and cytokine storm are targeted by the host's immune response, based on the evidence of positive selection on NSP2, NSP3, NSP13, S protein, ORF3a, and ORF9. However, genomic surveillance needs to be carried out to closely monitor mutational events and selection pressure, as the virus may mutate further to develop escape mutants due to increasing vaccination rates.

Authors' Contribution

The authors confirm contribution to the paper as follows. All authors reviewed the results and approved the final version of the manuscript. All authors agreed to be responsible for all aspects of the work to ensure the accuracy and integrity of the published manuscript. Study conception and design: Anuj Tewari; Data collection: Anuj Tewari; Analysis and interpretation of results: Anuj Tewari; Draft manuscript preparation: Anuj Tewari, Rajesh Kumar. All authors reviewed the results and approved the final version of the manuscript. All authors agreed to be responsible for all aspects of the work to ensure the accuracy and integrity of the published manuscript.

Ethics statement

The authors declare that the published work reflects an investigation and analysis carried out truthfully and completely. The study does not include any human trials/experiments and no ethical approval is needed.

Conflict of interest

The authors declare that they have no conflict of interest.

Availability of data

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Consent for publication

All authors hereby give their consent for publication.

References

1. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. New England Journal of Medicine. 2020;382(13):1199–207.

2. Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, et al. Severe acute respiratory syndrome-related coronavirus: The species and its viruses – a statement of the Coronavirus Study Group. bioRxiv. 2020;

3. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. The Lancet. 2020;395(10223):507–13.

4. Andrews MA, Areekal B, Rajesh K, Krishnan J, Suryakala R, Krishnan B, et al. First confirmed case of COVID-19 infection in India: A case report. Indian Journal of Medical Research. 2020;151(5):490–2.

5. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature Microbiology. 2020;5(11):1403–7.

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

7. Rambaut A, Loman N, Pybus O, Barclay W, Barrett J, Carabelli A, et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology - Virological [Internet]. Virological.org. 2020. p. 1–9. Available from: https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563

8. Domingo-Calap P, Schubert B, Joly M, Solis M, Untrau M, Carapito R, et al. An unusually high substitution rate in transplant-associated BK polyomavirus in vivo is further concentrated in HLA-C-bound viral peptides. PLoS Pathogens. 2018;14(10).

9. Holmes EC, Dudas G, Rambaut A, Andersen KG. The evolution of Ebola virus: Insights from the 2013-2016 epidemic. Nature. 2016;538(7624):193–200.

10. Deng X, Garcia-Knight MA, Khalid MM, Servellita V, Wang C, Morris MK, et al. Transmission, infectivity, and antibody neutralization of an emerging SARS-CoV-2 variant in California carrying a L452R spike protein mutation. medRxiv : the preprint server for health sciences [Internet]. 2021; Available from: http://www.ncbi.nlm.nih.gov/pubmed/33758899%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7987058

11. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, et al. Spike mutation D614G alters SARS-CoV-2 fitness. Nature. 2021;592(7852):116–21.

12. Jangra S, Ye C, Rathnasinghe R, Stadlbauer D, Alshammary

H, Amoako AA, et al. SARS-CoV-2 spike E484K mutation reduces antibody neutralisation. The Lancet Microbe. 2021;2(7):e283–4.

13. Al Khatib HA, Benslimane FM, Elbashir IE, Coyle PV, Al Maslamani MA, Al-Khal A, et al. Within-Host Diversity of SARS-CoV-2 in COVID-19 Patients With Variable Disease Severities. Frontiers in Cellular and Infection Microbiology. 2020;10.

14. Gao R, Zu W, Liu Y, Li J, Li Z, Wen Y, et al. Quasispecies of SARS-CoV-2 revealed by single nucleotide polymorphisms (SNPs) analysis. Virulence. 2021;12(1):1209–26.

15. Cherian S, Potdar V, Jadhav S, Yadav P, Gupta N, Das M, et al. Sars-cov-2 spike mutations, l452r, t478k, e484q and p681r, in the second wave of covid-19 in Maharashtra, India. Microorganisms. 2021;9(7).

16. Singh J, Rahman SA, Ehtesham NZ, Hira S, Hasnain SE. SARS-CoV-2 variants of concern are emerging in India. Nature Medicine. 2021;27(7).

17. Frost SDW, Magalis BR, Kosakovsky Pond SL. Neutral theory and rapidly evolving viral pathogens. Molecular Biology and Evolution. 2018;35(6).

18. Ranjan R, Sharma A, Verma MK. Characterization of the second wave of COVID-19 in India. Current Science. 2021;121(1).

19. Velazquez-Salinas L, Zarate S, Eberl S, Gladue DP, Novella I, Borca M V. Positive selection of ORF1ab, ORF3a, and ORF8 genes drives the early evolutionary trends of SARS-CoV-2 during the 2020 COVID-19 pandemic. Frontiers in Microbiology. 2020;11.

20. Katoh K, Misawa K, Kuma KI, Miyata T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research. 2002;30(14):3059–66.

21. Mercatelli D, Triboli L, Fornasari E, Ray F, Giorgi FM. Coronapp: A web application to annotate and monitor SARS-CoV-2 mutations. Journal of Medical Virology. 2021;93(5):3238–45.

22. Kosakovsky Pond SL, Frost SDW. Not so different after all: A comparison of methods for detecting amino acid sites under selection. Molecular Biology and Evolution. 2005;22(5).

23. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. PLoS Genetics. 2012;8(7).

24. Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL. Datamonkey 2.0: A modern web application for characterizing selective and other evolutionary processes. Molecular Biology and Evolution. 2018;35(3).

25. Spielman SJ, Weaver S, Shank SD, Magalis BR, Li M, Kosakovsky Pond SL. Evolution of viral genomes: Interplay between selection, recombination, and other forces. In: Methods in Molecular Biology. 2019.

26. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS ONE. 2012;7(10).

27. Baum A, Fulton BO, Wloga E, Copin R, Pascal KE, Russo V, et al. Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. Science. 2020;369(6506):1014–8.

28. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. 2020;182(4):812-827.e19.

29. Bakhshandeh B, Jahanafrooz Z, Abbasi A, Goli MB, Sadeghi M, Mottaqi MS, et al. Mutations in SARS-CoV-2; Consequences in structure, function, and pathogenicity of the virus. Microbial Pathogenesis. 2021;154.

30. European Centre for Disease Prevention and Control. SARS-CoV-2- increased circulation of variants of concern and vaccine rollout in the EU / EEA , 14th update Epidemiological situation. In: Rapid Risk Assessment. ECDC: Stockholm; 2021.

31. Yadav PD, Sapkal GN, Abraham P, Ella R, Deshpande G, Patil DY, et al. Neutralization of Variant Under Investigation B.1.617.1 With Sera of BBV152 Vaccinees. Clinical Infectious Diseases. 2021;

32. Wang R, Chen J, Hozumi Y, Yin C, Wei GW. Emerging Vaccine-Breakthrough SARS-CoV-2 Variants. ACS Infectious Diseases. 2022;8(3).

33. Angeletti S, Benvenuto D, Bianchi M, Giovanetti M, Pascarella S, Ciccozzi M. COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis. Journal of Medical Virology. 2020;92(6):584–8.

34. Graham RL, Sims AC, Brockway SM, Baric RS, Denison MR. The nsp2 Replicase Proteins of Murine Hepatitis Virus and Severe Acute Respiratory Syndrome Coronavirus Are Dispensable for Viral Replication. Journal of Virology. 2005;79(21):13399–411.

35. Harcourt BH, Jukneliene D, Kanjanahaluethai A, Bechill J, Severson KM, Smith CM, et al. Identification of Severe Acute Respiratory Syndrome Coronavirus Replicase Products and Characterization of Papain-Like Protease Activity. Journal of Virology. 2004;78(24):13600–12.

36. Schiller JJ, Kanjanahaluethai A, Baker SC. Processing of the coronavirus MHV-JHM polymerase polyprotein: Identification of precursors and proteolytic products spanning 400 kilodaltons of ORF1a. Virology. 1998;242(2):288–302.

37. Kumar R, Verma H, Singhvi N, Sood U, Gupta V, Singh M, et al. Comparative Genomic Analysis of Rapidly Evolving SARS-CoV-2 Reveals Mosaic Pattern of Phylogeographical Distribution. mSystems. 2020;5(4).

38. Corum J, Zimmer C. Bad News Wrapped in Protein: Inside the Coronavirus Genome A String of RNA. The New York Times. 2020;22:1–32.

39. Arya R, Kumari S, Pandey B, Mistry H, Bihani SC, Das A, et al. Structural insights into SARS-CoV-2 proteins. Journal of Molecular Biology. 2021;433(2).

40. Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature. 2020;581(7807):215–20.

41. Fung SY, Yuen KS, Ye ZW, Chan CP, Jin DY. A tug-of-war between severe acute respiratory syndrome coronavirus 2 and host antiviral defence: lessons from other pathogenic viruses. Emerging Microbes and Infections. 2020;9(1):558–70.

42. Kreimendahl S, Rassow J. The mitochondrial outer membrane protein tom70—mediator in protein traffic, membrane contact sites and innate immunity.

International Journal of Molecular Sciences. 2020;21(19).

43. Dominguez Andres A, Feng Y, Campos AR, Yin J, Yang CC, James B, et al. SARS-CoV-2 ORF9c Is a Membrane-Associated Protein that Suppresses Antiviral Responses in Cells. bioRxiv [Preprint]. 2020 Aug;Aug 19(2020.08.18. 256776).

44. Velazquez-Salinas L, Verdugo-Rodriguez A, Rodriguez LL, Borca MV. The role of interleukin 6 during viral infections. Frontiers in Microbiology. 2019;10(MAY).

45. Chomarat P, Banchereau J, Davoust J, Palucka AK. IL-6 switches the differentiation of monocytes from dendritic cells to macrophages. Nature Immunology. 2000;1(6).

46. Yang R, Masters AR, Fortner KA, Champagne DP, Yanguas-Casás N, Silberger DJ, et al. IL-6 IL-6 promotes the differentiation of a subset of naive CD8+ T cells into IL-21-producing B helper CD8+ T cells. Journal of Experimental Medicine. 2016;213(11).

47. Park SJ, Nakagawa T, Kitamura H, Atsumi T, Kamon H, Sawa S ichiro, et al. IL-6 Regulates In Vivo Dendritic Cell Differentiation through STAT3 Activation. The Journal of Immunology. 2004;173(6).

48. Diehl S, Rincón M. The two faces of IL-6 on Th1/Th2 differentiation. Molecular Immunology. 2002;39(9).