AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Spectrum

RESEARCH ARTICLE

Time Series Analysis of SARS-CoV-2 Genomes and Correlations among Highly Prevalent Mutations

Neha Periwal,[a] Shravan B. Rathod,[b] Sankritya Sarma,[c] Gundeep S. Johar,[d] Avantika Jain,[a,e] Ravi P. Barnwal,[f] Kinsukh R. Srivastava,[g] Baljeet Kaur,[h] Pooja Arora,[c] Vikas Sood[a]

[a]Department of Biochemistry, SCLS, Jamia Hamdard, New Delhi, India
[b]Department of Chemistry, Smt. S. M. Panchal Science College, Talod, Gujarat, India
[c]Department of Zoology, Hansraj College, University of Delhi, New Delhi, India
[d]Humber College, Toronto, Ontario, Canada
[e]Delhi Institute of Pharmaceutical Sciences and Research, New Delhi, Delhi, India
[f]Department of Biophysics, Panjab University, Chandigarh, India
[g]Division of Medicinal and Process Chemistry, CDRI, Lucknow, Uttar Pradesh, India
[h]Department of Computer Science, Hansraj College, University of Delhi, New Delhi, India

ABSTRACT The efforts of the scientific community to tame the recent pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) seem to have been diluted by the emergence of new viral strains. Therefore, it is imperative to understand the effect of mutations on viral evolution. We performed a time series analysis on 59,541 SARS-CoV-2 genomic sequences from around the world to gain insights into the kinetics of the mutations arising in the viral genomes. These 59,541 genomes were grouped according to month (January 2020 to March 2021) based on the collection date. Meta-analysis of these data led us to identify significant mutations in viral genomes. Pearson correlation of these mutations led us to the identification of 16 comutations. Among these comutations, some of the individual mutations have been shown to contribute to viral replication and fitness, suggesting a possible role of other unexplored mutations in viral evolution. We observed that the mutations 241C>T in the 5′ untranslated region (UTR), 3037C>T in nsp3, 14408C>T in the RNA-dependent RNA polymerase (RdRp), and 23403A>G in spike are correlated with each other and were grouped in a single cluster by hierarchical clustering. These mutations have replaced the wild-type nucleotides in SARS-CoV-2 sequences. Additionally, we employed a suite of computational tools to investigate the effects of T85I (1059C>T), P323L (14408C>T), and Q57H (25563G>T) mutations in nsp2, RdRp, and the ORF3a protein of SARS-CoV-2, respectively. We observed that the mutations T85I and Q57H tend to be deleterious and destabilize the respective wild-type protein, whereas P323L in RdRp tends to be neutral and has a stabilizing effect.

IMPORTANCE We performed a meta-analysis on SARS-CoV-2 genomes categorized by collection month and identified several significant mutations. Pearson correlation analysis of these significant mutations identified 16 comutations having absolute correlation coefficients of >0.4 and a frequency of >30% in the genomes used in this study. The correlation results were further validated by another statistical tool called hierarchical clustering, where mutations were grouped in clusters on the basis of their similarity. We identified several positive and negative correlations among comutations in SARS-CoV-2 isolates from around the world which might contribute to viral pathogenesis. The negative correlations among some of the mutations in SARS-CoV-2 identified in this study warrant further investigations. Further analysis of mutations such as T85I in nsp2 and Q57H in ORF3a protein revealed that these mutations tend to destabilize the protein relative to the wild type, whereas P323L in

Editor Ujjwal Neogi, Karolinska Institutet

Address correspondence to Vikas Sood, vikas1101@gmail.com.

The authors declare no conflict of interest.

Received 4 April 2022
Accepted 3 August 2022
Published 7 September 2022

RdRp is neutral and has a stabilizing effect. Thus, we have identified several comutations which can be further characterized to gain insights into SARS-CoV-2 evolution.

KEYWORDS COVID-19, hierarchical clustering, mutations, Pearson correlation, protein dynamics, SARS-CoV-2

A novel coronavirus first appeared in Wuhan, China, in December 2019 and became a public health emergency of international concern. Since its emergence, the virus has caused catastrophe across the globe. This virus, known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has infected nearly 486 million people and killed more than 6.3 million globally [WHO Coronavirus (COVID-19) Dashboard] as of 13 July 2022. Of the seven known coronaviruses—human coronavirus OC43 (HCoV-OC43), human coronavirus-229E (HCoV-229E), human coronavirus-HKU1 (HCoV-HKU1), human coronavirus-NL63 (HCoV-NL63), severe acute respiratory syndrome coronavirus (SARS-CoV), middle east respiratory syndrome coronavirus (MERS-CoV), and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), (1)—SARS-CoV-2 is highly pathogenic to humans (2). This virus has linear, positive-sense, single-strand RNA (ssRNA) as its genetic material, which is 29,903 bp long and is encapsulated by the nucleocapsid protein, which is one of the four structural proteins, the others being spike, envelope, and membrane proteins (3). Once the virus gains entry into the cell, two viral polyproteins, open reading frame 1a (ORF1a) and ORF1ab proteins, are formed. These polyproteins are then cleaved by the viral proteases into 16 nonstructural proteins, which initiate the process of viral replication and transcription. Apart from the viral nonstructural proteins, SARS-CoV-2 encodes 11 accessory proteins that play a key role in the viral pathogenesis (4).

Among the nonstructural proteins of SARS-CoV, nsp14, along with nsp10 and nsp12, plays a key role in maintaining the integrity of the viral RNA, resulting in fewer mutations than in other RNA viruses (5, 6). Despite the fact that SARS-CoV-2 mutates at a slower pace, this virus has evolved into numerous variants since the onset of the pandemic (7). The continuous evolution of SARS-CoV-2 has hindered the efforts of the scientific community to design vaccines and effective antivirals against it (8). Since mutations are one of the key factors driving the virus' evolution, understanding the kinetics of the mutations is imperative. Several studies have identified a large number of genetic variations, including missense mutations, synonymous mutations, insertions, and deletions, in the genomic sequences of SARS-CoV-2. The most common types of variations along the viral genome are reported to be missense and synonymous mutations (9). Although synonymous mutations may not have a direct impact on protein function, they have the potential to alter codon usage and translational frequency, as well as being able to affect the binding kinetics of microRNAs. Furthermore, it was speculated that the mutations in the 5′ untranslated region (UTR) may alter viral transcription, replication, and folding of the genomic ssRNA sequences (10). Genome analysis of SARS-CoV-2 revealed a substantial mutation bias toward uracil, which might be caused by improved immunogenicity, selection for greater expression, and better mRNA stability (11).

Viral transmission rates are rapidly increasing as the virus evolves. For instance, a single mutation (D614G) in the spike protein has been shown to increase the infectivity of SARS-CoV-2 (12). The appearance of multiple mutations in the same haplotype might lead to possible correlations among these mutations. It has been shown that comutations Y449S and N501Y in the spike protein can lead to reduced infectivity and play a major role in disrupting the antibody-mediated virus neutralization (13). This implies that mutations can have a synergistic effect, resulting in enhanced viral fitness and immune escape. Therefore, understanding the correlations among the mutations in the viral genome might lead to a better understanding of viral pathogenesis and evolution.

Several studies on this topic have been published. Zuckerman et al. analyzed 371 Israeli genomic sequences from February 2020 to April 2020 and observed correlations among identified mutations with that of known clade-defining ones (14). Wang et al. analyzed pairwise comutations in the most frequent 11 missense mutations

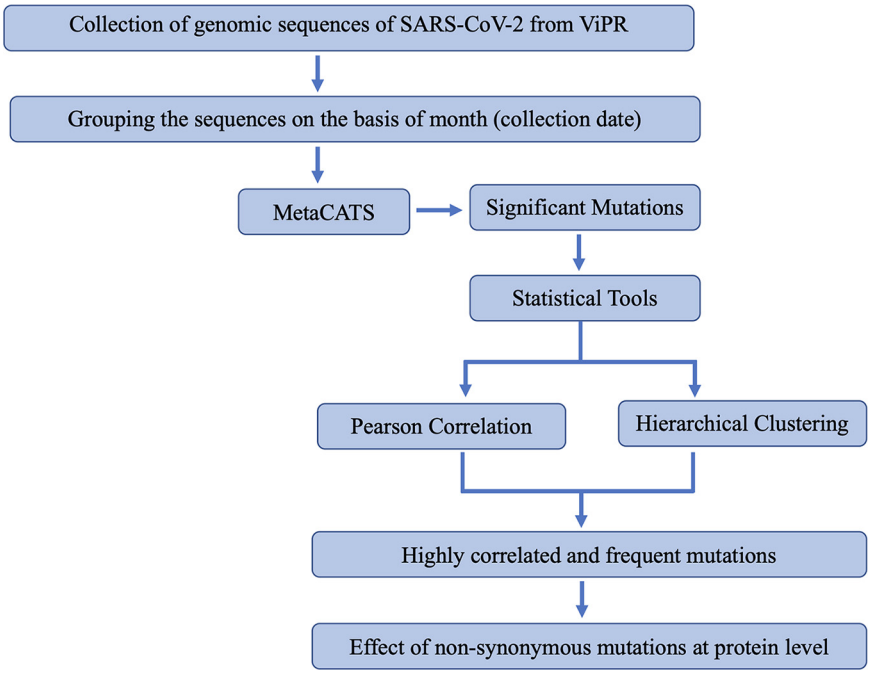

FIG 1 Methodology used in this study. The SARS-CoV-2 complete genome sequences were obtained from the ViPR and grouped by month based on the collection date. The meta-analysis was then performed in a pairwise manner, i.e., comparing January 2020 to each month through March 2021, to identify the highly significant mutations arising among the genomic sequences of SARS-CoV-2. Once the significant mutations were identified, we used Pearson correlation- and hierarchical clustering-based approaches to identify correlations and clusters among the highly significant mutations. The effect of these mutations on the wild-type protein was then studied using several computational tools, including PredictSNP, ENCoM $\Delta\Delta S_{Vib}$, DynaMut, ENCoM $\Delta\Delta G$, mCSM $\Delta\Delta G$, DUET $\Delta\Delta G$, and SAAFEC-SEQ.

that were prevalent in the United States (15). The authors included 12,754 SARS-CoV-2 sequences from the United States and identified missense mutations. In another study, Rahman et al. analyzed 324 complete and nearly complete SARS-CoV-2 genomic sequences which were isolated between 30 March 2020 and 7 September 2020 (16). They identified 3037C>T as the most frequent mutation, as it occurred in 98% of isolates. Though synonymous, this mutation was shown to co-occur with 3 other mutations, including 241C>T, 14408C>T, and 23403A>G. In another study, Chen et al. (17) analyzed 261,323 sequences of SARS-CoV-2 from across the globe to study the evolution of the virus. The authors observed that the initial SARS-CoV-2 M genotype ignited the COVID-19 outbreak. The M genotype harbored two concurrent mutations and was transformed to WE1 by acquiring four additional concurrent mutations (17). The WE1 genotype further evolved into WE1.1 by incorporating three additional concurrent mutations.

Some of the studies mentioned above were performed with SARS-CoV-2 genomic sequences obtained from a specific region, whereas some focused on the few significant missense mutations only. We hypothesized whether a similar trend could be observed with the genomic sequences of SARS-CoV-2 collected from around the world. In order to gain a better understanding of the origin of mutations in SARS-CoV-2 sequences, we analyzed viral genomic sequences in a time series-dependent manner. Meta-analysis of these SARS-CoV-2 genomic sequences led us to the identification of significant mutations. We performed two widely used statistical tools: Pearson correlation, which identified the comutations in the viral genome, and hierarchical clustering, which measured similarities between these mutations and grouped them in clusters. *In silico* protein dynamics was then used for the characterization of the impact of these mutations on their respective proteins (Fig. 1).

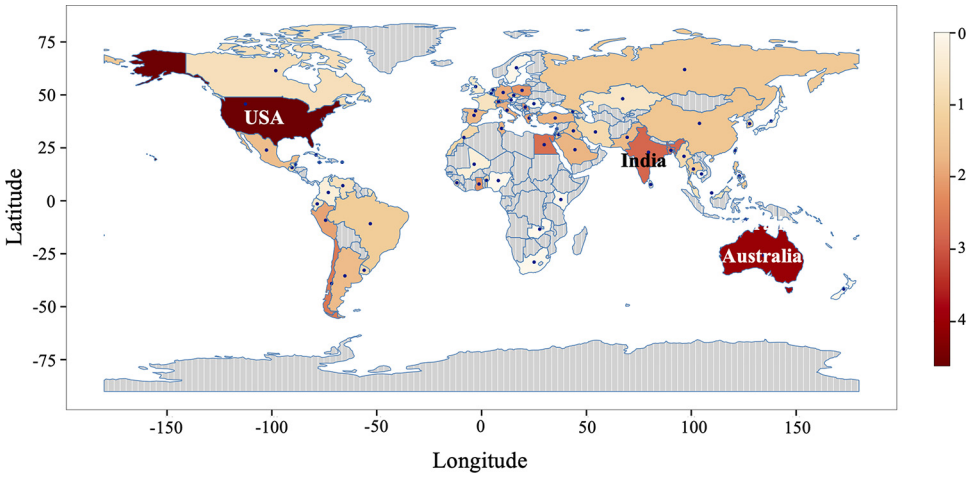

FIG 2 Geographical distribution of the SARS-CoV-2 genomic sequences used in this study. The color bar represents the frequency (log_{10}) of sequences. The areas that contributed the maximum number of genomic sequences in this study are represented by dark red shading, whereas areas that contributed fewer genomic sequences are represented by light orange.

RESULTS AND DISCUSSION

SARS-CoV-2 genomes. In order to understand the kinetics of highly prevalent mutations in the SARS-CoV-2 circulating genomes, we sought to analyze these genomic sequences in a time series manner. There can be a considerable time lapse between sample collection and sample processing; therefore, we used the sample collection month to classify the SARS-CoV-2 genomic sequences. We included a total of 59,541 SARS-CoV-2 genomes that were collected from January 2020 until March 2021 and grouped them by month based on sample collection (see Table S1 in the supplemental material). The number of SARS-CoV-2 genomic sequences from each country was visualized on a world map (Fig. 2). The global distribution of the samples revealed that the majority of the samples were from the United States, followed by Australia, India, and Egypt.

Identification of significant mutations in the SARS-CoV-2 genomes. Once the SARS-CoV-2 genomic sequences were grouped on the basis of the sample collection month, we used the META-CATS algorithm to identify significant mutations among the genomes. This algorithm compares two different data sets to identify significant mutations among them. As the genomic sequences collected at the start of the pandemic tend to be very similar to the parent sequence, all the SARS-CoV-2 genomic sequences collected in January 2020 were grouped together to form a control group. The sequences obtained in subsequent months were then analyzed against the sequences from the control group to identify significant mutations in SARS-CoV-2 genomes of that particular month. We obtained significant mutations for each month except December 2020 (Fig. 3A). Since mutations at the nucleotide level might not lead to changes in amino acids due to the degeneracy of the genetic code, we focused our attention on the mutations at the amino acid level (Fig. 3B to I). We identified 940 unique mutations at the amino acid level which were unevenly distributed among the genome of SARS-CoV-2. Our analysis identified 610, 256, 33, 2, 11, 10, 16, and 2 mutations in the ORF1ab, spike, ORF3a, membrane, ORF6, ORF8, nucleocapsid, and ORF10 proteins of SARS-CoV-2, respectively. As the length of SARS-CoV-2 proteins is highly variable, we calculated the frequency of the mutations at the amino acid level in order to understand their distribution in the viral proteins. We observed that the spike protein had the highest frequency of mutations (20.10%), followed by ORF6 (18%) and ORF1ab (8.59%) (Table 1). We observed that the membrane protein of SARS-CoV-2 had the fewest mutations compared to the other proteins, suggesting that this region might be highly conserved among SARS-CoV-2 variants. The recently emerged SARS-CoV-2

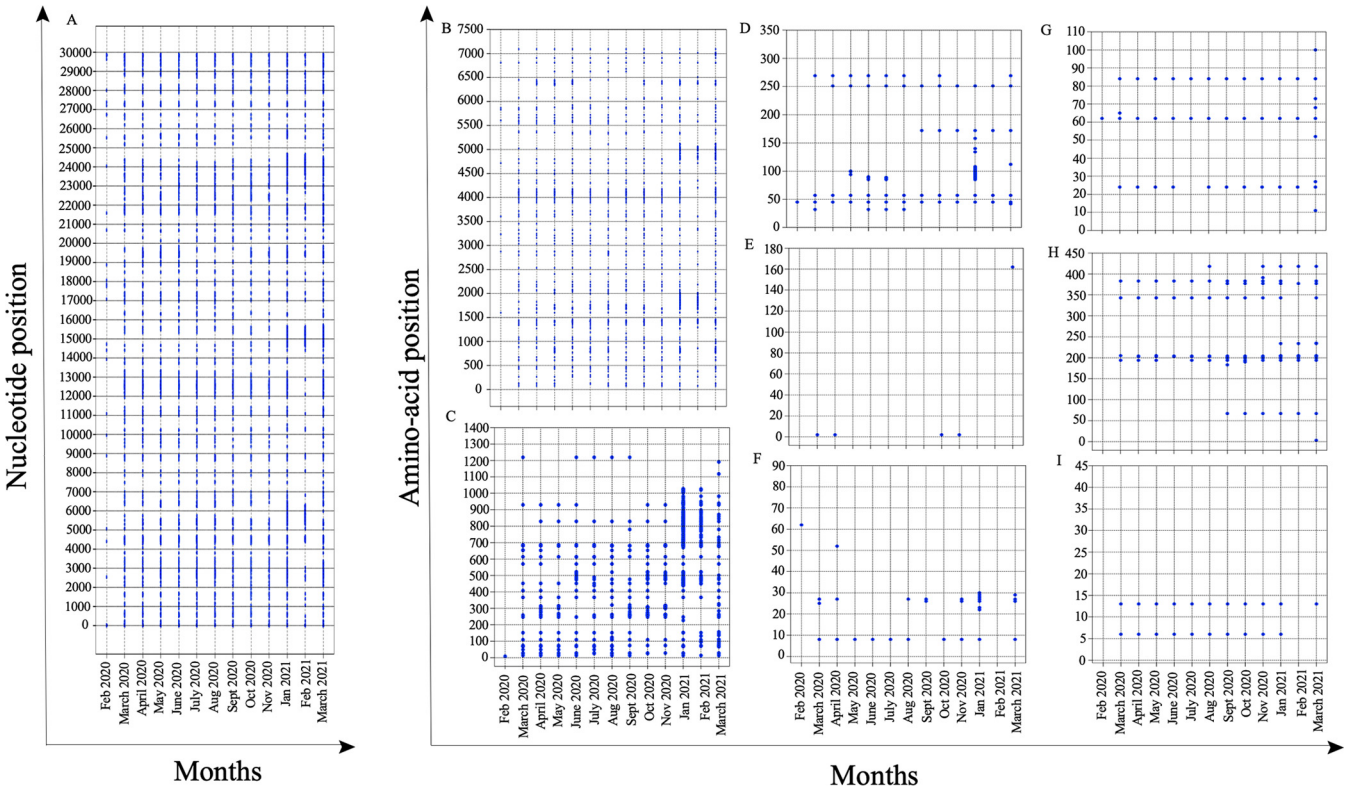

FIG 3 Mutations in the SARS-CoV-2. (A) Mutations at the nucleotide level across the whole genome of SARS-CoV-2. (B to I) Nonsynonymous mutations at the amino acid level for the SARS-CoV-2 proteins (B) ORF1ab, (C) spike protein, (D) ORF3a protein, (E) membrane protein, (F) ORF6 protein, (G) ORF8 protein, (H) nucleocapsid protein, and (I) ORF10 protein. Some mutations appeared early in the pandemic and were consistently present throughout the study.

mutant Omicron has been shown to have more than 40 mutations in its spike protein, suggesting that the protein is highly amenable to mutations (18).

Correlation among significant mutations in SARS-CoV-2 genomes. Co-occurrence of several mutations has been shown to modulate the function of the proteins (19). Therefore, we sought to understand whether there was any correlation among the significant mutations that we identified in this study. For this purpose, we utilized two well-established statistical approaches: Pearson correlation, which measures the correlation coefficient (positive or negative) between two mutations, and hierarchical clustering, which groups similar mutations into clusters.

(i) Pearson correlation coefficient. Analysis of SARS-CoV-2 sequences in a time series manner led us to the identification of several significant mutations. In order to identify the correlation among these mutations, Pearson correlation was performed on a binary matrix, with 1 representing significant mutations and 0 representing no mutations in SARS-CoV-2 genomes. The correlation value ranges from −1.0 to +1.0, with

TABLE 1 Frequency of unique significant mutations in various SARS-CoV-2 proteins

Protein	Length (aa)	No. of unique mutations	Frequency[a]
Orf1ab	7,096	610	8.59
S	1,273	256	20.10
Orf3a	275	33	12
M	222	2	0.90
Orf6	61	11	18.0
Orf8	121	10	8.2
N	419	16	3.8
Orf10	38	2	5.1

[a]Calculated by dividing the number of unique mutations by the length of the respective protein and then multiplying by 100.

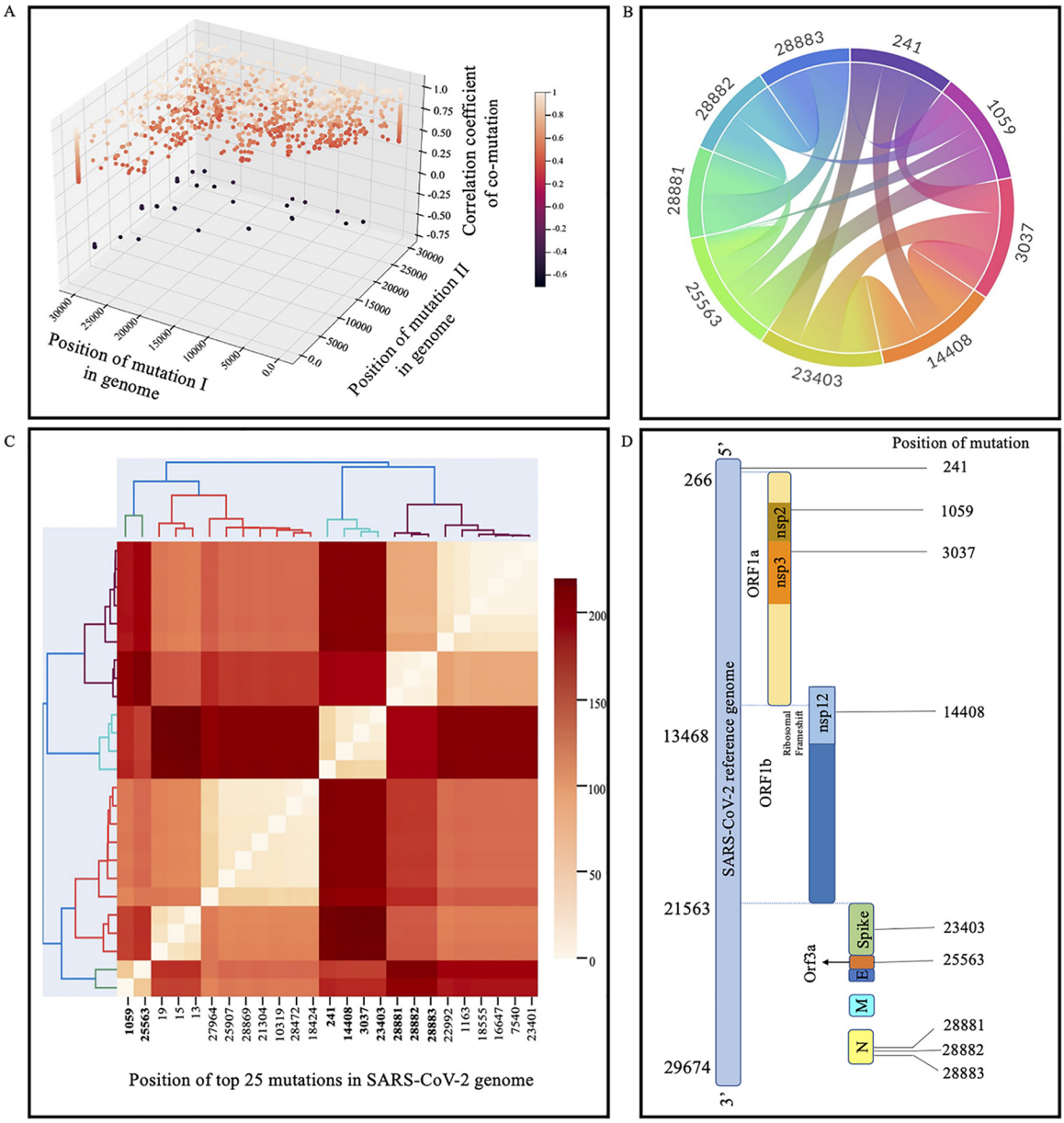

FIG 4 Significant mutations that were correlated and clustered with each other by using statistical approaches (Pearson correlation and hierarchical clustering). (A) Three-dimensional plot showing the comutations having Pearson correlation coefficients with absolute values of >0.4. The degree of correlation is reflected by the color. The correlation coefficient ranges from −1 to +1; absolute values closer to 1 have a higher correlation than those closer to 0. (B) Chord plot representing comutations with absolute correlation coefficients of >0.4 and occurring in more than 30% of the genomes used in the study. (C) Hierarchical clustering of the 25 most significant mutations obtained from ViPR having a frequency of >10%. The *x* axis represents the significant mutations. The colors represent the distances among the mutations. The comutations identified in this study are highlighted in bold. (D) Cartoon showing the positions of mutations in the viral genome.

negative values indicating negative correlation and positive values indicating positive correlation. Additionally, absolute values closer to 1 indicate a very strong correlation. The results obtained from the Pearson correlation were then filtered to obtain only comutations where the absolute value of the correlation coefficient was greater than 0.4. Using this criterion, we obtained 2,205 comutations (Fig. 4A). It was observed that the frequency of the majority of these comutations was very low. For instance, a comutation at positions 21306 and 22995 with an absolute value of the correlation coefficient of >0.4 but occurrence in less than 5% of the genomes might not be of interest. Therefore, we considered only comutations that were present in >30% of genomes for

TABLE 2 Correlations among various unique mutations in SARS-CoV-2 genomes[a]

Position of mutation:		Correlation value of comutation	% of genomes with:		
I	II		Mutation I	Mutation II	Both mutations
241 (5′ UTR)	3037 (nsp3)	0.760	91.054	89.889	87.16
241 (5′ UTR)	23403 (spike)	0.758	91.054	90.082	87.26
241 (5′ UTR)	14408 (RdRp)	0.733	91.054	89.385	86.61
1059 (nsp2)	25563 (ORF3a)	0.863	40.022	46.948	39.33
1059 (nsp2)	28882 (nucleocapsid)	−0.534	40.022	30.234	0.07
1059 (nsp2)	28881 (nucleocapsid)	−0.535	40.022	30.381	0.10
1059 (nsp2)	28883 (nucleocapsid)	−0.535	40.022	30.259	0.06
3037 (nsp3)	23403 (spike)	0.976	89.889	90.082	88.52
3037 (nsp3)	14408 (RdRp)	0.943	89.889	89.385	87.87
14408 (RdRp)	23403 (spike)	0.943	89.385	90.082	87.97
25563 (ORF3a)	28882 (nucleocapsid)	−0.613	46.948	30.234	0.13
25563 (ORF3a)	28881 (nucleocapsid)	−0.614	46.948	30.381	0.17
25563 (ORF3a)	28883 (nucleocapsid)	−0.614	46.948	30.234	0.11
28881 (nucleocapsid)	28882 (nucleocapsid)	0.995	30.381	30.234	29.77
28881 (nucleocapsid)	28883 (nucleocapsid)	0.994	30.381	30.259	29.77
28882 (nucleocapsid)	28883 (nucleocapsid)	0.997	30.234	30.259	29.77

[a]The comutations shown have a correlation value of <-0.4 and >0.4 and are present in $>30\%$ of the genomes. Comutations showing positive correlations are present in the majority of the same genomes. As expected, comutations showing negative correlations are not present in the same genomes.

further study. Using this stringent criterion, we identified 16 comutations that had an absolute value of the correlation coefficient of >0.4, with each mutation of the comutation being present in $>30\%$ of the genomes (Fig. 4B and Table 2). It was further observed that six comutations were present in $>89\%$ of the genomes, suggesting their possible role in viral fitness. Our analysis captured a highly prevalent mutation in the spike protein (D614G) that has nearly replaced the wild-type sequence and is known to increase viral infectivity (12). The identification of the D614G mutation further validated our approach and prompted us to further explore other comutations that were identified.

(ii) Hierarchical clustering. In order to garner confidence and validate our results that were obtained using Pearson correlation, we used another statistical tool to group the significant mutations in clusters. Since hierarchical clustering is a computationally intensive process, we analyzed only the 25 most significant mutations that were present in $>10\%$ of the genomes used in this study (Table 3). Similar to the results obtained from Pearson correlation, hierarchical clustering analysis led to the grouping of the mutations in clusters that possess similarities. Here, we analyzed only clusters in which the frequency of each mutation was greater than 30% of genomes (Fig. 4C). The mutations at positions 241, 14408, 3037, and 23403 in the SARS-CoV-2 genome form a cluster and are the most common concurrent mutations. Since both the statistical tools provided similar results, we then focused our attention on these mutations to acquire an in-depth understanding of them. The positions of the nine mutations in the SARS-CoV-2 genomes are depicted in Fig. 4D.

Frequency and global distribution of highly correlated and frequent significant mutations. Once the comutations that have a correlation coefficient with an absolute value of >0.4 and are present in $>30\%$ of genomes were identified, we sought to investigate the frequency of each mutation in SARS-CoV-2 genomes. There are 9 mutations that constitute the 16 comutations. The nucleotide positions where these mutations occur are 241, 1059, 3037, 14408, 23403, 25563, 28881, 28882, and 28883. Analysis of each position in the genome revealed that mutations at 241, 3037, 14408, and 23403 almost completely replaced the wild-type sequences (Fig. 5A). Analysis of the genome revealed that the major nucleotide change that occurred in around 97% of the mutated population at position 241 in the SARS-CoV-2 genome was 241C$>$T (Fig. 5B). However, in the remaining 3% of the mutated SARS-CoV-2 sequences, 241C$>$A was observed (Fig. 5 and Table 4). Since the frequency of the major mutations was much higher than that of the minor mutation at the same nucleotide position, we considered the major mutation for further study. We investigated the global distribution of all nine mutations

TABLE 3 The 25 most significant mutations and their positions in SARS-CoV-2 genomes

No.	Position	Gene	No. of sequences[a]
1	241	5′ UTR	50,771
2	23403	Spike	50,229
3	3037	ORF1ab	50,121
4	14408	ORF1ab	49,840
5	25563	ORF3a	26,178
6	1059	ORF1ab	22,316
7	28881	Nucleocapsid	16,940
8	28883	Nucleocapsid	16,872
9	28882	Nucleocapsid	16,858
10	27964	ORF1ab	10,985
11	1163	ORF1ab	9,098
12	10319	ORF1ab	8,882
13	18555	ORF1ab	8,772
14	28869	Nucleocapsid	8,770
15	16647	ORF1ab	8,755
16	23401	Spike	8,746
17	7540	ORF1ab	8,729
18	18424	ORF1ab	8,758
19	28472	Nucleocapsid	8,572
20	21304	ORF1ab	8,320
21	25907	ORF3a	8,236
22	22992	Spike	8,146
23	19	5′ UTR	6,637
24	15	5′ UTR	5,662
25	13	5′ UTR	5,069

[a]Number of genomic sequences in which the mutation in present.

among the circulating SARS-CoV 2 genomes and found that the mutation 241T in the 5′ UTR completely replaced the wild-type nucleotide, C241, as early as June-July 2020 (Fig. 6A). Similar trends were observed with the mutations 3037C>T, 14408C>T, and 23403A>G in nsp3, RNA-dependent RNA polymerase (RdRp), and spike proteins, respectively (Fig. 6B to D). The prevalence of these mutations in the SARS-CoV 2 circulating

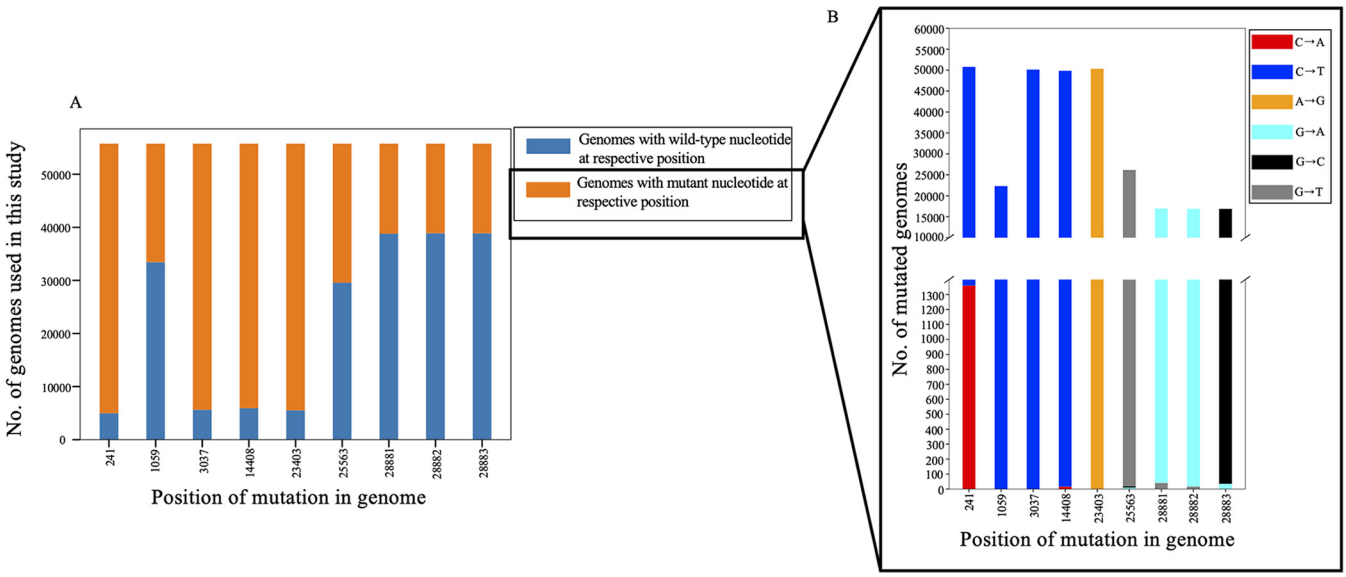

FIG 5 Stacked bar chart showing the distribution of genomes with a given mutation. (A) The total length of the bar represents the total number of SARS-CoV-2 genomic sequences. The length of the blue bar represents the number of genome sequences with wild-type nucleotides at that position; the length of the orange bar represents the number of genome sequences with a given mutation (mutated nucleotide). (B) Bar length represents the number of genomic sequences that have a mutation at a particular position. A position with a single bar indicates that the wild-type nucleotide is substituted by a single nucleotide. A position with multiple bars indicates that the wild-type nucleotide is substituted by more than one nucleotide. The lengths of the different-color bars represent the degree of substitution.

TABLE 4 Frequency of major and minor nucleotide substitutions for a specific mutation

Position of mutation in the genome	Genomic region	Nucleotide before mutation (reference)	Nucleotide(s) after mutation		Amino acid(s) in:		
					Wild-type protein (reference)	Mutant protein	
			Major	Minor		Major	Minor
241	5′ UTR	C241	241T	241A			
1059	nsp2 (ORF1ab)	C1059	1059T		T265	265I	
3037	nsp3 (ORF1ab)	C3037	3037T		F106	106F	
14408	nsp12 (ORF1ab)	C14408	14408T	14408A	P323	323L	323H
23403	Spike	A23403	23403G		D614	614G	
25563	ORF3a	G25563	25563T	25563A/25563C	Q57	57H	57Q/57H
28881	Nucleocapsid	G28881	28881A	28881T	R203	203K	203M
28882	Nucleocapsid	G28882	28882A	28882T	R203	203R	203S
28883	Nucleocapsid	G28883	28883C	28883A	G204	204R	204R

genomes suggests their critical role in viral pathogenesis. Other mutations, including one in nsp2 (1059C>T), one in ORF3a (25563C>T), and three in nucleocapsid protein (28881G>A, 28882G>A, and 28883G>C) showed a mosaic pattern of global distribution that increased over time (Fig. 6E to I).

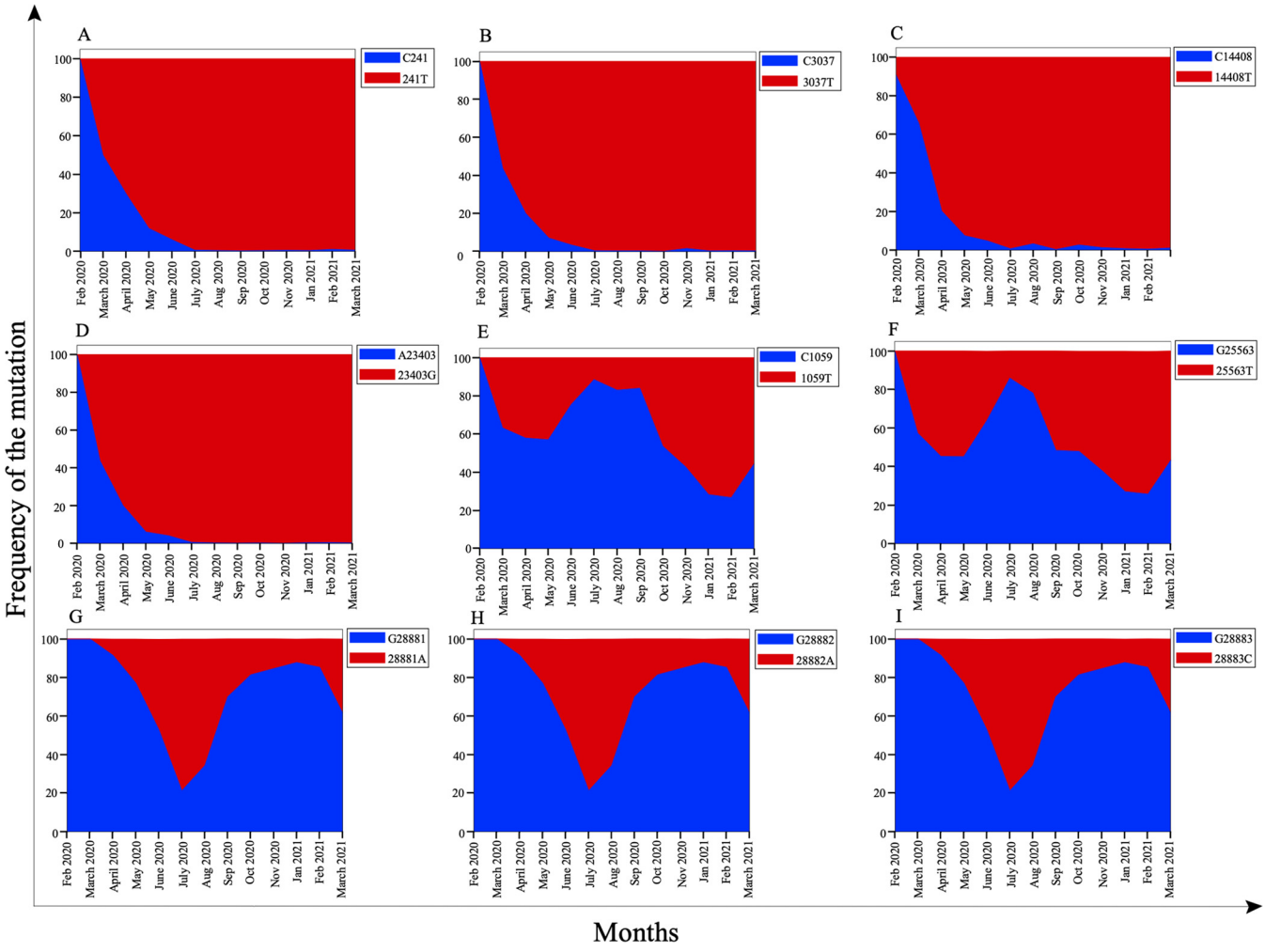

FIG 6 Running monthly counts of the sampled sequences exhibiting significant mutations (that have a correlation coefficient with an absolute value of >0.4 and are present in >30% of the SARS-CoV-2 genomes under study) between 1 February 2020 and 31 March 2021, as follows: (A) 241C>T in the 5′ UTR; (B) 3037C>T in nsp3 (ORF1ab protein); (C) 14408C>T in RdRp (ORF1ab protein); (D) 23403A>G in the spike protein; (E) 1059C>T in nsp2 (ORF1ab protein); (F) 25563G>T in the ORF3a protein; (G) 28881G>A in the nucleocapsid protein; (H) 28882G>A in the nucleocapsid protein; and (I) 28883G>C in the nucleocapsid protein. Blue represents the SARS-CoV-2 genomic sequences having wild-type nucleotides at a particular position, whereas red represents the sequences having mutant nucleotides at a particular position.

Mutation in the 5′ UTR. The untranslated region of the viral genome plays a vital role in viral replication. This region has been shown to form various secondary structures to allow the binding of cellular and viral proteins, thereby regulating the translation of viral proteins (20, 21). Therefore, any mutation in these highly conserved regions has the potential to regulate viral replication. Statistical approaches revealed that the mutation 241C>T was closely correlated with three different mutations, 3037C>T, 14408C>T, and 23403A>G, in the nsp2, RdRp, and spike genes, respectively, of SARS-CoV-2. Remarkably, it can be observed that the correlation coefficient of mutation 241C>T with all the other mutations mentioned above was >0.75, pointing toward a very strong correlation. Additionally, these mutations were found in >89% of the genomes, further suggesting their critical role in viral evolution. These observations were further supported by hierarchical clustering, in which these mutations were clustered together. Our results are in agreement with published studies that have shown similar correlations among these mutations (14). However, these studies were conducted on the genomes of viruses from various countries, including Israel, the United States, and Bangladesh, whereas our analysis was carried out with SARS-CoV-2 genomes obtained globally. The correlation of the 241C>T mutation with frequently occurring mutations in the SARS-CoV-2 genomes points to its role in viral pathogenesis and fitness.

Mutation in nsp2. The protein nsp2 of SARS-CoV-2 was recently shown to be associated with host proteins involved in vesicle trafficking. It was also proposed that targeting the interactions of viral nucleocapsid proteins nsp2 and nsp8 with the host translational machinery might have therapeutic effects (22). Therefore, understanding the dynamics of nsp2 is essential. Our analysis revealed that the mutation 1059C>T in nsp2 was both positively and negatively correlated with other mutations in SARS-CoV-2. As described in Table 2, the mutation 1059C>T (T85I) in the nsp2 gene was positively correlated with 25563G>T (Q57H) in the ORF3a gene with a correlation coefficient of 0.863 and was present in >39% of the genomes, suggesting that the co-occurrence of these mutations might play a role in viral evolution. These observations are in agreement with earlier studies where co-occurrence of 1059C>T with 25563G>T was observed in nearly 70% of COVID-19 cases across the United States (15). Additionally, we observed that the 1059C>T (T85I) mutation in the nsp2 gene was negatively correlated with three mutations—28881G>A (R203K), 28882G>A (R203R), and 28883G>C (G204R)—in the nucleocapsid gene. Though the co-occurrence of these mutations was also established in another study (14), in this study, we show that these mutations are negatively correlated. The negative correlation among these mutations suggests that in a single haplotype, only one of them can occur. The analysis of data revealed the co-occurrence of 1059C>T with 28881G>A, 28882G>A, and 28883G>C in 0.10, 0.07 and 0.06% of the genomes, respectively. Therefore, it would be interesting to further investigate the negative relationship among these mutations under experimental conditions.

Since the T85I (1059C>T) mutation was widespread among nsp2 proteins, we sought to investigate the role of this mutation in the function of this protein. The full-length 3.2-Å crystal structure of nsp2 (PDB ID 7SMW) was solved by combining cryo-electron microscopy (cryo-EM) and the recently developed AI tool AlphaFold2 (23). In the structure, there is a highly conserved zinc binding site, which indicates the role of nsp2 in RNA binding (Fig. 7A). We also studied the T85I mutation in nsp2, in which a polar threonine residue is replaced with a hydrophobic isoleucine. The PredictSNP tool revealed that this mutation is deleterious, with around a 70% confidence score. The ENCoM-based negative vibrational entropy energy ($\Delta\Delta S_{vib}$) value suggests that this mutation confers some degree of flexibility on nsp2. It can be seen that two helices (1: 19 to 28 amino acids [aa] and 2:35 to 45 aa) at the N terminus gain slight flexibility (Fig. 8A, red). Among the six predictors, four predicted a negative free energy change ($\Delta\Delta G$), thereby implying the destabilization of nsp2 (Table 5).

Our results on nsp2 protein stability and flexibility are in accordance with already published reports (15). In the wild-type and mutant proteins, two identical residues

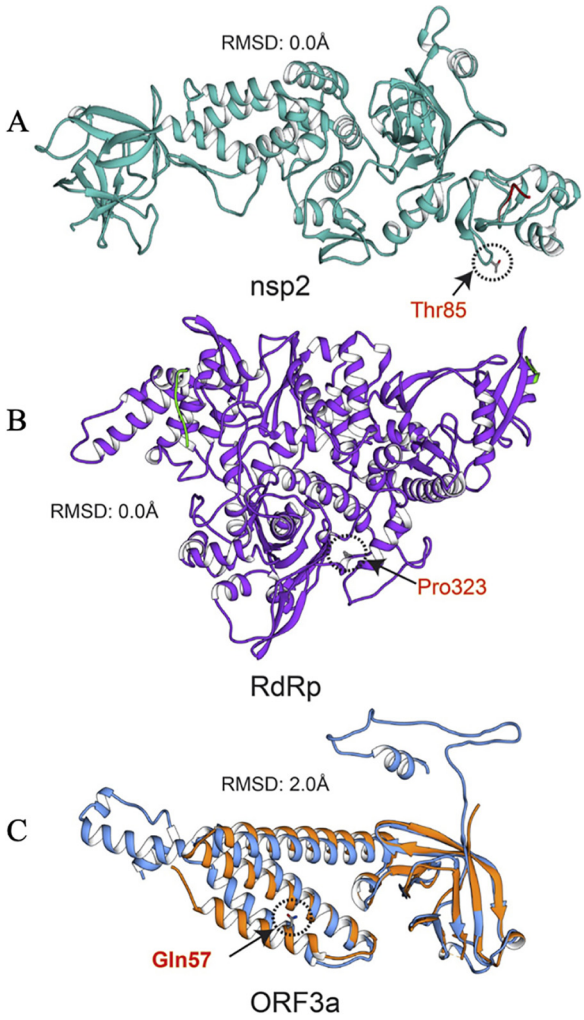

FIG 7 Structure alignments of crystal structure and the modeled proteins. (A) nsp2 crystal and modeled structures are in cyan and red, respectively. (B) RdRp crystal and modeled structures are in purple and red, respectively. (C) ORF3a crystal and modeled structures are in orange and blue, respectively. Mutation positions are circled, and mutant residues are represented in stick form.

(Phe83 and Asn87) interact with the wild-type and mutant residues. In both the structures, van der Waals clashes were observed between the side chain oxygen of Thr85 and aromatic carbons of Phe83 in the wild type and between the side chain methyl group carbon atom of Ile85 and aromatic carbons of Phe83 in the mutant. In the wild type, Thr85 amide group oxygen and nitrogen interact with surrounding amide group atoms of Phe83 and Asn87 through hydrophobic, van der Waals, and polar interactions. However, in the mutant protein, similar interactions were noted, but a polar-van der Waals clash was observed between Asn87 and Ile85. This might be the cause of the predicted instability of the T85I mutation in nsp2. The wild-type and mutant interactions are illustrated in Fig. 8D.

Mutations in nsp3. The nsp3 protein in coronavirus has been shown to antagonize the innate immune responses (24). The mutations in the nsp3 macrodomain region lead to enhanced type I interferon (IFN) responses and reduced viral replication (25). Understanding the dynamics of mutations in nsp3 might provide clues to SARS-CoV-2 evasion of type I IFN signaling. We identified a synonymous mutation, 3037C>T (F106F), that was positively correlated with 241C>T in the 5′ UTR, 23403A>G (D614G) in the spike, and 14408C>T (P323L) in the RdRp of SARS-CoV-2. Though silent, 3037C>T (F106F) was shown to disrupt the mir-197-5p target sequence (26). mir-197-5p was

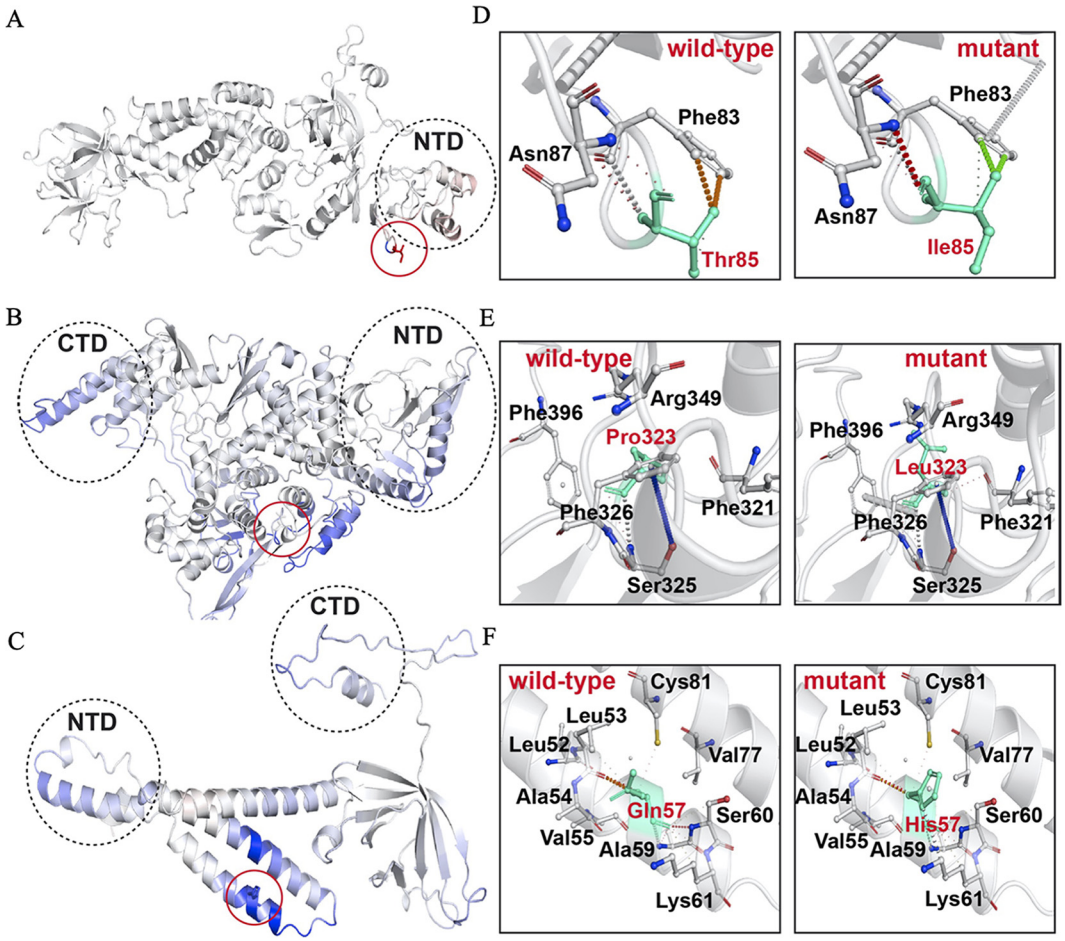

FIG 8 Visual representation of mutant protein dynamicity and intramolecular interactions of wild-type and mutant residues with proximal amino acids. Mutations are shown in stick representation and are circled in red. Red and blue indicate flexibility and rigidity, respectively. The wild type and the mutant residues are shown in cyan. Mutant residues are in red, while interacting residues are in black. Interactions are illustrated in different colors; for further interpretation of interactions, see the web version of the Arpeggio web server. (A to C) Visual representation of nsp2, RdRp, and ORF3a, respectively. (D to F) Intramolecular interactions of wild-type and mutant residues in nsp2, RdRp, and ORF3a, respectively.

shown to be associated with some other viruses also (27–29), indicating its role in viral biology. The 14408C>T (P323L) mutation was shown to increase the mutation rate among SARS-CoV-2 isolates, whereas 23403A>G (D614G) has been shown to contribute to the infectivity of the virus (16). The co-occurrence of all these mutations in >87% of the genomes further points to their critical role in driving viral evolution.

Mutation in RdRp. RdRp (nsp12) of SARS-CoV-2 is important for viral replication and transcription. This protein is also believed to be the most prominent target for potential antiviral drugs (30). Therefore, understanding the mutations in this protein is critical for RdRp-based drug designs. The mutation 14408C>T (P323L) in RdRp was present in >89% of genomes, suggesting that this mutation is now a part of the circulating genomes. Apart from its widespread presence, this mutation was correlated with some other mutations, including 241C>T in the 5′ UTR, 3037C>T in nsp3, and 23403A>G in the spike, with high correlation coefficients of 0.73, 0.94, and 0.94, respectively. Interestingly, 14408C>T and 23403A>G mutations were reported in patients with severe COVID-19, in contrast to those with mild infections, suggesting their possible role in disease severity (31). Owing to the widespread presence of P323L mutation in RdRp, we sought to study its effect on the stability of the wild-type protein.

The 2.83-Å crystal structure of RdRp in complex with nsp7, nsp8, nsp9, and helicase was determined using cryo-electron microscopy (32). The RdRp structure has the RdRp

TABLE 5 Predicted results for the effects of mutation on functionality, stability, and flexibility of respective proteins using PredictSNP, DynaMut, and SAAFEC-SEQ web servers

Protein (mutation)	Confidence score (%) and nature of mutation (PredictSNP)	$\Delta\Delta S_{vib}$ (kcal mol^{-1} K^{-1}) and flexibility (ENCoM)	$\Delta\Delta G$ (kcal mol^{-1})[a]					
			DynaMut	ENCoM	mCSM	SDM	DUET	SAAFEC-SEQ
nsp2 (T85I)	72 (deleterious)	0.021 (increase)	−0.264	−0.017	−0.125	0.460	0.151	−0.93
RdRp (P323L)	83 (neutral)	−0.225 (decrease)	0.732	0.180	−0.261	1.570	0.457	−0.77
ORF3a (Q57H)	76 (deleterious)	−0.117 (decrease)	0.597	0.094	−0.843	0.060	−0.652	−1.03

[a]A negative value indicates a predicted destabilizing effect, while a positive value indicates a stabilizing one.

domain (367 to 920 aa) (33) and N terminus (60 to 249 aa), which adopts the nidovirus RdRp-associated nucleotidyltransferase (NiRAN) structural scaffold (34). Another region (4 to 118 aa) is composed of two helices and five β-strands that are antiparallel. Additionally, the short β-strand (215 to 218 aa) was observed in RdRp, which is highly ordered in SARS-CoV-2 compared to SARS-CoV. This β-strand has contact with other β-strand residues (96 to 100 aa) and thus increases the conformational stability of RdRp in SARS-CoV-2 (33).

The P323L mutation is present on RdRp interface domain, especially in the loop region, which connects the interface domain's three helices to the same domain's three β-strands (Fig. 7B). An earlier study suggested that this mutation enhances the processivity of RdRp (35). It is predicted to be functionally neutral, with a notable confidence score of 83%. This mutation results in a conformationally rigid proline ring being replaced by a flexible side chain containing a leucine residue. Though the wild-type and mutant residues are hydrophobic, their conformational flexibility must be the deciding factor for protein stability and flexibility. Nonetheless, this mutation significantly rigidifies RdRp (Fig. 8B), and $\Delta\Delta S_{vib}$ was also observed to be much lower (Table 5). Results show that this mutation has a strong communication network in RdRp and impacts various helices and β-sheets. The P323L mutation is located in a loop formed by the β-strand (328 to 335 aa) and helix (304 to 320 aa); thus, these two secondary structures gain rigidity. However, a helix-proximal mutation gained greater rigidity than other regions of RdRp. The helices at the N- and C-terminal domains also gained rigidity due to this mutation. All-atom simulation data also suggested that the P323L mutation reduces the flexibility of RdRp, which is in line with our results (36). Three $\Delta\Delta G$ value predictors predicted stabilization and the remaining three predicted destabilization, but the $\Delta\Delta G$ stabilization values are considerably higher than the destabilization values. Mohammad et al. performed 200-ns all-atom molecular dynamics simulation by calculating free energy (ΔG) of the wild-type and mutant RdRp and confirmed that P323L increases the stability of RdRp (36). Hence, this mutation stabilizes the RdRp structure. Analysis of the RdRp wild-type and mutant interaction revealed that there are more interactions in the mutant than the wild type. The wild-type and mutant residues are surrounded by Phe321, Ser325, Phe326, Arg349, and Phe396 residues. In the wild type, only a single polar interaction between Ser325 and Pro323 residues is observed, while in the mutant, two additional hydrogen bonds with Ser325 and Phe326 and polar interaction with Phe349 were observed. Thus, it can be considered that higher stability in the mutant comes from these interactions. Figure 8E shows the interactions in the wild-type and mutant RdRp.

Mutations in the ORF3a protein. ORF3a is the largest accessory protein of SARS-CoV-2 and plays a key role in the viral infection cycle. Moreover, this protein is essential for viral replication, and mutations in this protein are associated with higher mortality rates (37). The mutation 25563G>T (Q57H) in the SARS-CoV-2 ORF3a gene has been shown to be associated with decreased death and increased cases of COVID-19 (38). We further observed that 25563G>T (Q57H) mutation is positively correlated with the 1059C>T mutation in nsp2, whereas it is negatively correlated with 28881G>A, 28882G>A, and 28883G>C mutations in the nucleocapsid gene. Our observations are in agreement with previous studies which identified similar associations within the genomes of SARS-CoV-2 isolated in Israel (14). The ORF3a functional domains are vital for SARS-CoV-2 infectivity, virulence, ion channel synthesis, and the release of the virus

(39). A recent study showed that ORF3a in SARS-CoV-2 has a weaker potential for proapoptotic activity than SARS-CoA ORF3a, which might be linked to the infectivity of the viruses (40). Furthermore, another study confirmed that ORF3a binds to the homotypic fusion and protein sorting (HOPS) complex and prevents autolysosome formation (41). ORF3a is also considered a potential vaccine and drug target (42, 43).

The experimental structure of ORF3a protein (PDB ID 6XDC) was determined using cryo-EM at 2.1-Å resolution. ORF3a protein has three main regions: the N terminus (1 to 39 aa), the cytoplasmic loop (175 to 180 aa), and the C terminus (239 to 275 aa) (44). Figure 7C shows the structure of ORF3a protein. In the Q57H mutation, a glutamine polar residue is replaced by a polar and basic histidine residue. This mutation is situated in the helix region of ORF3a protein and is predicted to be deleterious, having a 76% confidence score (Table 5). Similar observations were reported in earlier studies (39, 45). The $\Delta\Delta S_{vib}$ value implied that the ORF3a protein gains rigidity and becomes less flexible due to the appearance of this mutation. Similar to RdRp, the mutant residue in ORF3a protein also has a wide communication dynamics. This single mutation in the helix increases the rigidity of the whole ORF3a protein (Fig. 8C). Our findings are in agreement with a previous study on the impact of the Q57H mutation in the protein (15).

Based upon the $\Delta\Delta G$ values, this mutation was predicted to be destabilizing (Table 5). The wild-type and the mutant residues are in close proximity to Leu52, Leu53, Ala54, Val55, Ala59, Ser60, Lys61, Val77, and Cys81. In the wild-type protein, two hydrogen bonds are observed in Ser60 and Lys61 amide bond amino groups with the amide carbonyl group of the wild-type residue. These identical hydrogen bonds are also present in the mutant structure. Other types of interactions, such as polar and hydrophobic, were observed in the wild-type and the mutant ORF3a protein. However, there are two new clashes seen in the mutant structure between the histidine ring and Lys61 and mutant amide group and Leu53 residue. Thus, the overall number of clashes has increased in the mutant, and this might be the factor responsible for the destabilization of ORF3a protein under the influence of the Q57H mutation. This mutation was predicted to have significant potential to alter the ORF3a conformation and lead to disruption of intramolecular hydrogen bonds in ORF3a (38). Our findings that Q57H causes destabilization of ORF3a are in agreement with the previous study. The wild-type and the mutant interactions are illustrated in Fig. 8F.

Mutation in the spike protein. Spike protein is a homotrimer protein that studs the surface of SARS-CoV-2, giving it a crown-like shape. The spike protein of SARS-CoV-2 consists of two subunits that are covalently attached to each other. One of the subunits, S1, binds to the ACE2 receptor on the target cells, whereas the S2 subunit helps anchor the spike protein to the cell membrane (46, 47). The D614G mutation in the spike protein has been shown to increase the infectivity of the virus. In our previous study (48), we characterized the effect of D614G mutation on protein activity and suggested that the mutation led to decreased protein stability but enhanced protein movement. In this study, we observed a correlation of the 23403A>G (D614G) mutation in the spike with the mutations 241C>T (5′ UTR), 3037C>T (nsp3), and 14408C>T (RdRp). The presence of these mutations in >96% of the genomes suggests their critical role in viral pathogenesis. The above-mentioned mutations had replaced the wild-type sequences by June-July 2020 (Fig. 6A to D). Therefore, a better understanding of these comutations via further experimentation is urgently required.

Mutations in the nucleocapsid protein. Nucleocapsid protein is one of the most conserved proteins among SARS coronaviruses (49). This protein is known to interact with viral RNA as well as the viral membrane protein to aid virion assembly. This protein is also shown to play a role in regulating host immune responses (50) and cellular apoptosis (51). The nucleocapsid protein of SARS-CoV-2 acts as a viral RNA interference (RNAi) suppressor and has been shown to antagonize cellular RNAi pathways (52). Thus, understanding the role of mutations in modulating the function of this protein becomes important. The mutations 28881G>A, 28882G>A, and 28883G>C in the nucleocapsid are positively correlated with each other. Of these, the mutations at 28881 and 28883 are missense mutations, whereas the mutation at 28882 is synonymous.

Since the mutations in nucleocapsid protein are known to increase the infectivity and virulence of the virus (53), the correlation of these mutations with other mutations warrants further study. Interestingly we observed that these three mutations in the nucleocapsid were negatively correlated with two other mutations, 1059C>T in nsp2 and 25563C>T in ORF3a. It was further observed that a very small number of haplotypes (0.06 to 0.17%) had both mutations in the same genome (Table 2). The absence of these mutations from the same genome implies their possible negative impact on viral evolution and pathogenesis. In our recent study (48), we probed the impacts of these two mutations, 28881 (R203K) and 28883 (G204R), on the nucleocapsid protein structure, function, and dynamics. Though we performed the analysis of each of these mutations separately, a recent study investigated the synergistic effect of these comutations in the nucleocapsid protein function (54). It was observed that these comutations increase the fitness, infectivity, replication, and virulence of SARS-CoV-2. These comutations were shown to increase the phosphorylation of the viral nucleocapsid protein and to confer enhanced resistance to glycogen synthase kinase-3 (GSK-3), thereby leading to efficient viral replication.

Impact of mutations on protein dynamics. The protein structure and functions are significantly altered by the insertion of single-point mutations (55–57). Investigating the structural or functional impacts of point mutations in all proteins can be achieved using a suite of computational tools, including NMA models (58), Gaussian network models (GNM) (59), anisotropic network models (ANM) (60), elastic network models (ENM) (61), discrete molecular dynamics (DMD) (62), all-atom molecular dynamics (AAMD) simulation (63), and protein evolutionary data. Therefore, we employed these tools to probe the effects of mutations on protein structures. The predicted results of the mutations are given in Table 5.

Linear mutual information analysis of the mutants. The normalized linear mutual information (nLMI) gives insight into the protein residue network and dynamic correlation. Figure 9 illustrates the nLMI correlation and correlation difference plots of nsp2, RdRp, and ORF3a proteins along with their mutants. It can be observed that the nsp2 and ORF3a protein residues are strongly correlated (>0.625) compared to RdRp, where correlation among residues is considerably lower (<0.500) (Fig. 9A to C). However, to understand the impacts of every single mutation in each protein, we obtained correlation differences between the wild-type and mutant structures of all the three proteins. In the correlation difference plots, the yellow regions indicate no or a very slight correlation (0.00 to 0.25), whereas cyan regions indicate slightly negative anticorrelation between the residues (Fig. 9D to F). In RdRp and ORF3a mutant structures, residues have significantly less correlation. However, the nsp2 mutant structure's residues are slightly anticorrelated. Thus, the T85I mutation in nsp2 causes a slight disruption in the structure that can be considered destabilization of the nsp2 mutant structure. However, P323L in RdRp and Q57H in ORF3a do not result in notable destabilization of the mutant structures.

Conclusion. Since the onset of the SARS-CoV-2 pandemic in December 2019, the virus has significantly mutated. The mutations in the virus have led to the emergence of mutants that have the capacity to dodge vaccine and antiviral therapies. Therefore, understanding the dynamics of mutations in the viral genome is of utmost importance. To this end, we performed a time series analysis of viral sequences to understand the origin and frequency of significant mutations present in the SARS-CoV-2 circulating genomes. The meta-analysis approach was used to identify significant mutations in the SARS-CoV-2 sequences. Pearson correlation and hierarchical clustering were then used to identify the correlations and the clusters among the significant mutations. We identified 16 comutations that had an absolute Pearson correlation coefficient of >0.4 and were present in >30% of the genomes analyzed in this study. We identified a strong correlation coefficient (>0.73) for the mutations 241C>T in the 5′ UTR with 3037C>T (F106F) in nsp3, 23403A>G (D614G) in spike, and 14408C>T (P323L) in RdRp. The cooccurrence of these mutations was found in >86% of the genomes that were studied, suggesting that these comutations were part of the same haplotypes. The area plot

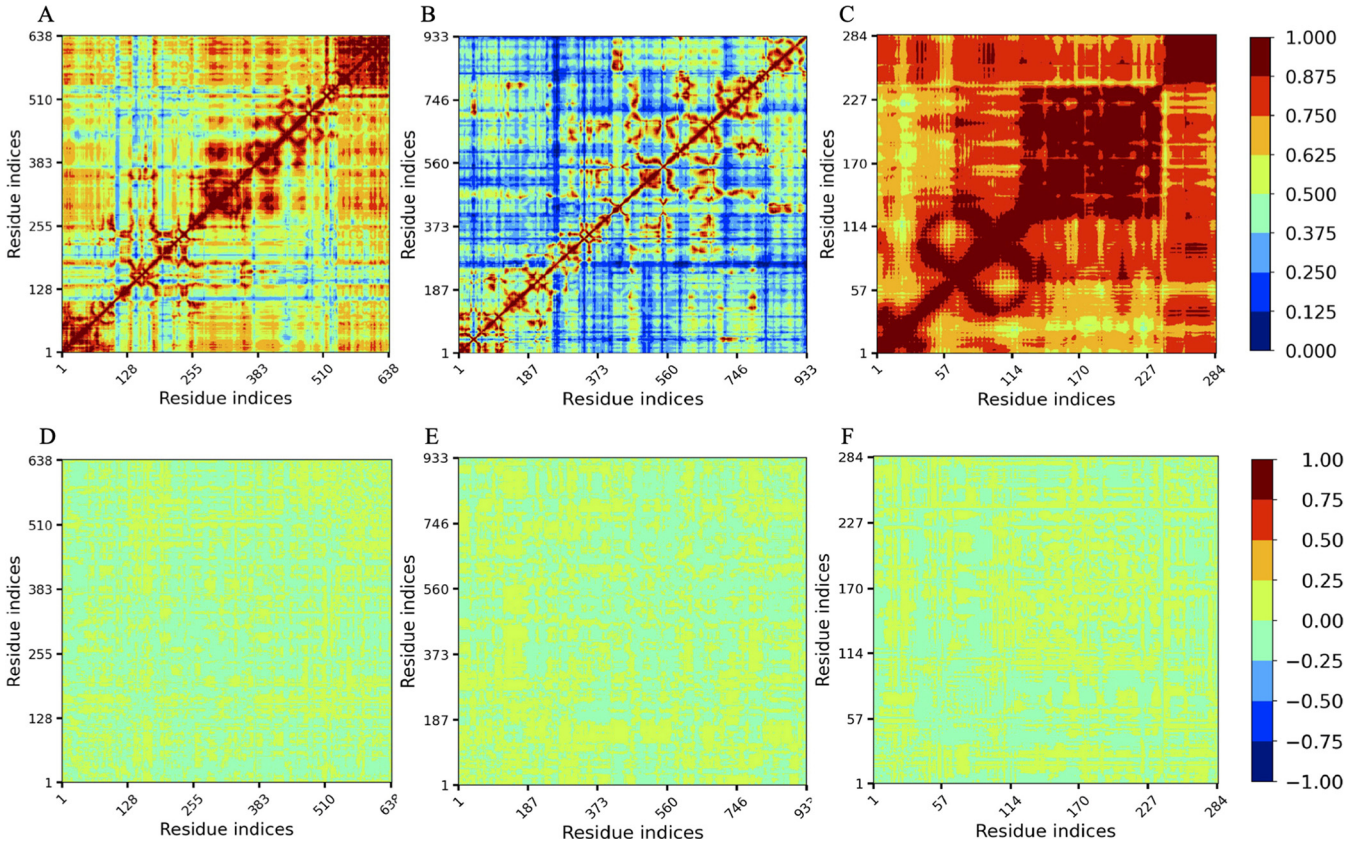

FIG 9 nLMI correlation plots. (A to C) Wild-type nsp2 (A), RdRp (B), and ORF3a (C). (D to F) Correlation difference plots of wild-type and mutant nsp2 (D), RdRp (E), and ORF3a (F). The degree of correlation is indicated by the color.

analysis revealed that these mutations replaced the respective wild-type sequences by June-July 2020. In this study, we were able to capture negative correlations of the mutations 28881G>A, 28882G>A, and 28883G >C in the nucleocapsid gene with mutations, including 1059C>T in nsp2 and 25563G>T in ORF3a, implying that a haplotype will not harbor 28881G>A, 28882G>A, and 28883G>C nucleocapsid mutations along with 1059C>T in nsp2 or 25563G>T in ORF3a. However, the combined effect of these mutations having negative correlations in the viral replication still needs to be investigated.

To investigate the impacts of T85I, P323L, and Q57H mutations in nsp2, RdRp, and ORF3a proteins, respectively, on their structural stability and flexibility, we employed structure and sequence-based tools. $\Delta\Delta G$ and $\Delta\Delta S_{vib}$ were used to evaluate the stability and flexibility of proteins, respectively. From the free energy calculation, T85I and Q57H mutations in nsp2 and ORF3a proteins, respectively, disrupt the residual network in the wild-type protein and destabilize the wild-type protein, while P323L in RdRp brings stability to the wild-type protein by adding new contacts between the residues. Also, consensus predictors were used to predict the impacts of mutation on protein functions. It was noted that T85I in nsp2 and Q57H in ORF3a were found to be deleterious, which implies that they alter the protein functions. However, P323L in RdRp was predicted to be neutral, which suggests that this mutation does not have any impact on protein function. The graph theory-based nLMI correlation was also obtained for the wild-type and mutant structures of three proteins to understand the residue communication in proteins. The nsp2 and ORF3a residues have a greater correlation than with RdRp. The correlation difference plot suggests that compared to nsp2, a significant correlation was observed in RdRp and ORF3a under the influence of the mutations. Thus, it can be observed that the latter two mutations increase the residue correlation in RdRp and ORF3a proteins, while in the case of nsp2, there was no significant correlation

difference observed, indicating a slight or no impact of mutation on the structure of nsp2. The fact that some of the mutations have destabilizing effects but have very high frequency suggests that these might play some role in viral fitness. Further experimentation is required to study the effect of these comutations on viral transmission and pathogenesis.

MATERIALS AND METHODS

SARS-CoV-2 genomic sequences. Since the onset of the SARS-CoV-2 pandemic in 2019, the virus is continuously evolving thereby resulting in the emergence of several variants. The availability of SARS-CoV-2 genomic sequences has been instrumental in understanding viral evolution and pathogenesis. To gain an in-depth understanding of the mutational landscape of SARS-CoV-2, we sought to analyze SARS-CoV-2 genomic sequences in a time series manner. All the SARS-CoV-2 genomic sequences were collected in a monthwise manner (based on the sample collection month) from the Virus Pathogen Resource (ViPR) database (64). The database was accessed on 18 April 2021, and only complete genomic sequences of SARS-CoV-2 were downloaded for further processing. In order to obtain high-quality genomic sequences, only complete genomes, which were around 15% of the total sequences, were used in this study. Additionally, an in-house Python script was written to check for the presence of unusual bases in the sequences included in this study. In the current study, SARS-CoV-2 sequences were collected from January 2020 to March 2021, resulting in 59,541 complete genome sequences.

Meta-analysis of SARS-CoV-2 genomic sequences. Once the SARS-CoV-2 genomic sequences were obtained and categorized by collection month, we performed a meta-analysis on these sequences to identify significant mutations among them. The genomes from various months were analyzed with respect to the genomes collected in the month of January 2020. The genomes obtained during the initial phase of infections tend to be close to the wild type, with few mutations, compared to the genomes collected at the later stages of the infection. For the identification of significant mutations, a metadata-driven comparative analysis tool (META-CATS) was used (65). All the analyses were performed with default settings, and mutations with P values of >0.05 were considered significant. We obtained a considerable number of significant mutations for each month. Notably, SARS-CoV-2 genomic sequences collected in December 2020 did not yield any significant mutations. Therefore, sequences collected in December 2020 were not included for further analysis.

Pearson correlation coefficient. In order to identify the correlation among significant mutations in the SARS-CoV-2 circulating genomes, Pearson correlation was used. Pearson correlation measures the linear correlation between two variables. An empty matrix of 55,759 by 29,903 was created using the NumPy module of Python. In this matrix, the number of rows represents the number of SARS-CoV-2 genomic sequences that were analyzed, and the number of columns represents the length of the reference SARS-CoV-2 genome. In this matrix, the occurrence of mutation at a particular position was represented by the number 1, whereas nonoccurrence of the mutation at a specific position was represented by the number 0. All comparisons were made with the SARS-CoV-2 reference genome, GenBank accession no. NC_045512.2. We used the "corr" method of the pandas library (66) to implement Pearson correlation.

Hierarchical clustering. To validate the results obtained from the Pearson correlation, the hierarchical clustering technique was applied; this method groups similar objects together. Since highly frequent mutations tend to play a critical role in the evolution of the virus (17), the 25 most significant mutations that were present in more than 10% of the genomes were tested for similarity using the hierarchical clustering technique. The figure_factory method from the plotly library of Python was used to perform the hierarchical clustering on the binary matrix of the 25 most significant mutations. The pdist and squareform methods from the SciPy library of Python (67) were used to create the dendrogram with a heat map. The dendrogram, together with the heat map, represents the significant mutations that are clustered. All these analyses were performed in Python version 3.8.5.

Protein structure and model preparation. Once the comutations in SARS-CoV-2 were identified, several computational tools were used to investigate the effect of the mutations (that constitute the comutations) on the respective protein structure. Of these mutations, 241C>T in 5′ UTR does not get translated into an amino acid. The mutations 3037C>T (F016F in nsp3) and 28882G>A (R203R in nucleocapsid protein) are synonymous and hence were not included for further analysis. A detailed analysis of three other mutations, including 23403A>G (D614G in spike protein), 28881G>A (R203K in nucleocapsid protein), and 28883G>C (G204R in nucleocapsid protein) was performed in our recent study (48) and thus not explored in this study. Therefore, in this study, we targeted three mutations—1059C>T (T85I in nsp2), 14408C>T (P323L in RdRp), and 25563G>T (Q57H in ORF3a protein)—to probe their impact on the protein structure and function. The crystal structures of these proteins were obtained from the Protein Data Bank (PDB). Analysis of these protein structures revealed that they had some missing amino acids. Hence, we employed a deep learning-based protein modeling tool, RoseTTAFold (68), to add missing residues to these proteins. The nsp2 protein (PDB ID 7MSW) is 638 amino acids long, and in the crystal structure, the first three residues at the N terminus were missing. RdRp (PDB ID 7CYQ) is 942 amino acids long, and 1 to 3 amino acids at the N terminus and 930 to 942 amino acids at the C terminus are missing in the structure. The ORF3a protein (PDB ID 6XDC) is only 284 amino acids long, but a large number of residues from both termini (1 to 39 aa at the N terminus and 239 to 284 aa at the C terminus) and six residues (175 to 180 aa) are missing in the protein structure. RoseTTAFold modeled all these missing residues in these proteins

except for the nine histidine residues at the C terminus of RdRp. These three modeled proteins were further analyzed for mutation effects.

Functional impact of mutations. To investigate the effect of mutation on protein function, we used the widely popular PredictSNP web server (69). This web tool is composed of a suite of six different predictors, including predictor of human deleterious single nucleotide polymorphism (PhD-SNP), multivariate analysis of protein polymorphism (MAPP), screening of nonacceptable polymorphism (SNAP), polymorphism phenotyping v1 (PolyPhen-1), polymorphism phenotyping v2 (PolyPhen-2) and sorting intolerant from tolerant (SIFT) to predict whether a given mutation is deleterious or neutral. PredictSNP gives a consensus prediction score using these six predictors. These six predictors make use of different methods to predict the nature of the mutation. PhD-SNP, MAPP, SNAP, PolyPhen-1, SIFT, and PolyPhen-2 utilize a support vector machine, physicochemical characteristics and a protein sequence alignment score, a neural network approach, an expert set of empirical rules, a protein sequence alignment score, and naive Bayes, respectively (69), to predict whether a given mutation is deleterious or neutral. To calculate the PredictSNP score, the following equation is employed:

$$\text{PredictSNPScore} = \frac{\sum_{i=1}^{N}(\delta_i S_i)}{\sum_{i=1}^{N} S_i} \tag{1}$$

where δ_i is an inclusive prediction (-1, neutral; $+1$, deleterious), S_i indicates the transformed confidence scores, and N is the number of predictors. The PredictSNP consensus score is between -1 and $+1$, where -1 to 0 corresponds to a neutral and 0 to $+1$ to a deleterious mutation.

Effect of mutations on protein dynamics. The normal mode analysis (NMA)-based DynaMut (70) web tool was utilized to probe the effects of a single mutation in each protein on its stability and flexibility. The folding free energy change ($\Delta\Delta G$) was calculated to exactly predict the stability of the protein under the influence of the mutations. In addition to its own $\Delta\Delta G$ prediction, DynaMut also predicts $\Delta\Delta G$ using NMA-based ENCoM (Elastic network contact model) (71) and other structure-based predictors, like the mutation cutoff scanning matrix (mCSM) (72), site-directed mutator (SDM) (73), and DUET (74). The free energy change indicates the stability of the proteins by measuring the energy difference between the wild-type and mutant proteins. Additionally, DynaMut employs ENCoM to predict vibrational entropy energy ($\Delta\Delta S_{vib}$). The values of $\Delta\Delta S_{vib}$ are calculated for wild-type and mutant proteins by screening their all-atom pair interactions. We utilized the protein sequence-based SAAFEC-SEQ (single amino acid folding free energy changes-sequence) tool (75) to validate the DynaMut predictions for wild-type and mutant proteins. This tool utilizes different protocols, such as protein sequence properties, evolutionary details, and physicochemical properties, to calculate the $\Delta\Delta G$ value.

Linear mutual information. To understand the nature of dynamics and fluctuations in the protein structures, dynamical cross-correlation (DCC)- and LMI-based approaches were employed in this study (76–79). Since DCC cannot calculate the correlation of atoms moving concurrently in perpendicular directions (80), we applied normalized LMI (nLMI) to overcome this limitation. To calculate the nLMI of wild-type and mutant proteins, we employed the Python-based correlation-plus 0.2.1 tool (80), which uses PDB files as the input. During the calculation, the program uses the anisotropic network model (ANM) to generate 100 models of wild-type and mutant proteins, and the correlation is thus obtained using these models. Additionally, we calculated the difference in correlation between wild-type and mutant proteins. To calculate nLMI between residues i and j, the following equation was used;

$$\text{LMI}_{ij} = \frac{1}{2}[\ln(\det C_i) + \ln(\det C_j) - \ln(\det C_{ij})] \tag{2}$$

where $C_i = \langle x_i^T x_i \rangle$, $C_j = \langle x_j^T x_j \rangle$, and $C_{ij} = \langle (x_i, x_j)^T (x_i, x_j) \rangle$; also, $x_i = R_i - \langle R_i \rangle$ and $x_j = R_j - \langle R_j \rangle$, where R_i and R_j are the atom i and j position vectors, det = Determinant of protein residue cross correlation metrices. In the nLMI calculation, the LMI was considered greater than or equal to 0.3 and the distance threshold was less than or equal to 7 Å. The values 0 and 1 indicate no correlation and complete correlation of residues, respectively.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 0.1 MB.

ACKNOWLEDGMENTS

N.P. is thankful to UGC for a Ph.D. fellowship. V.S. received a research grant from UGC, Govt. of India.

S.B.R. is thankful to his chemistry department for providing computational and infrastructure facilities.

N.P., S.S., S.B.R., A.J., and G.S.J. performed the experiments and analyzed the data. B.K. contributed to the statistical experiments. V.S., P.A., B.K., R.P.B., S.B.R., and K.R.S.

designed and supervised the study. N.P., S.B.R., S.S., and V.S. wrote the first draft of the manuscript. V.S., P.A., B.K., S.B.R., R.P.B., and K.R.S. edited and finalized the draft.

We declare no conflict of interest.

REFERENCES

1. Liu DX, Liang JQ, Fung TS. 2021. Human coronavirus-229E, -OC43, -NL63, and -HKU1 (Coronaviridae), p 428–440. *In* Bamford D, Zuckerman M (ed), Encyclopedia of virology. Elsevier, Amsterdam, The Netherlands.
2. Hu B, Guo H, Zhou P, Shi Z-L. 2021. Characteristics of SARS-CoV-2 and COVID-19. Nat Rev Microbiol 19:141–154. https://doi.org/10.1038/s41579-020-00459-7.
3. V'kovski P, Kratzel A, Steiner S, Stalder H, Thiel V. 2021. Coronavirus biology and replication: implications for SARS-CoV-2. Nat Rev Microbiol 19:155–170. https://doi.org/10.1038/s41579-020-00468-6.
4. Redondo N, Zaldívar-López S, Garrido JJ, Montoya M. 2021. SARS-CoV-2 accessory proteins in viral pathogenesis: knowns and unknowns. Front Immunol 12:2698. https://doi.org/10.3389/fimmu.2021.708264.
5. Sevajol M, Subissi L, Decroly E, Canard B, Imbert I. 2014. Insights into RNA synthesis, capping, and proofreading mechanisms of SARS-coronavirus. Virus Res 194:90–99. https://doi.org/10.1016/j.virusres.2014.10.008.
6. Ferron F, Subissi L, Silveira De Morais AT, Le NTT, Sevajol M, Gluais L, Decroly E, Vonrhein C, Bricogne G, Canard B, Imbert I. 2018. Structural and molecular basis of mismatch correction and ribavirin excision from coronavirus RNA. Proc Natl Acad Sci U S A 115:E162–E171. https://doi.org/10.1073/pnas.1718806115.
7. Wang L, Cheng G. 2022. Sequence analysis of the emerging SARS-CoV-2 variant omicron in South Africa. J Med Virol 94:1728–1733. https://doi.org/10.1002/jmv.27516.
8. Christensen PA, Olsen RJ, Long SW, Subedi S, Davis JJ, Hodjat P, Walley DR, Kinskey JC, Ojeda Saavedra M, Pruitt L, Reppond K, Shyer MN, Cambric J, Gadd R, Thakur RM, Batajoo A, Mangham R, Pena S, Trinh T, Yerramilli P, Nguyen M, Olson R, Snehal R, Gollihar J, Musser JM. 2022. Delta variants of SARS-CoV-2 cause significantly increased vaccine breakthrough COVID-19 cases in Houston, Texas. Am J Pathol 192:320–331. https://doi.org/10.1016/j.ajpath.2021.10.019.
9. Rahimi A, Mirzazadeh A, Tavakolpour S. 2021. Genetics and genomics of SARS-CoV-2: a review of the literature with the special focus on genetic diversity and SARS-CoV-2 genome detection. Genomics 113:1221–1232. https://doi.org/10.1016/j.ygeno.2020.09.059.
10. Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, Chang H. 2020. The architecture of SARS-CoV-2 transcriptome. Cell 181:914–921. https://doi.org/10.1016/j.cell.2020.04.011.
11. Rice AM, Castillo Morales A, Ho AT, Mordstein C, Mühlhausen S, Watson S, Cano L, Young B, Kudla G, Hurst LD. 2021. Evidence for strong mutation bias toward, and selection against, U content in SARS-CoV-2: implications for vaccine design. Mol Biol Evol 38:67–83. https://doi.org/10.1093/molbev/msaa188.
12. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, Hastie KM, Parker MD, Partridge DG, Evans CM, Freeman TM, de Silva TI, McDanal C, Perez LG, Tang H, Moon-Walker A, Whelan SP, LaBranche CC, Saphire EO, Montefiori DC, Sheffield COVID-19 Genomics Group. 2020. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 182:812–827. https://doi.org/10.1016/j.cell.2020.06.043.
13. Wang R, Chen J, Wei G-W. 2021. Mechanisms of SARS-CoV-2 evolution revealing vaccine-resistant mutations in Europe and America. J Phys Chem Lett 12:11850–11857. https://doi.org/10.1021/acs.jpclett.1c03380.
14. Zuckerman NS, Bucris E, Drori Y, Erster O, Sofer D, Pando R, Mendelson E, Mor O, Mandelboim M. 2021. Genomic variation and epidemiology of SARS-CoV-2 importation and early circulation in Israel. PLoS One 16:e0243265. https://doi.org/10.1371/journal.pone.0243265.
15. Wang R, Chen J, Gao K, Wei GW. 2021. Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. Commun Biol 4:228. https://doi.org/10.1038/s42003-021-01754-6.
16. Rahman MM, Kader SB, Rizvi SS. 2021. Molecular characterization of SARS-CoV-2 from Bangladesh: implications in genetic diversity, possible origin of the virus, and functional significance of the mutations. Heliyon 7:e07866. https://doi.org/10.1016/j.heliyon.2021.e07866.
17. Chen Y, Li S, Wu W, Geng S, Mao M. 2021. Distinct mutations and lineages of SARS-CoV-2 virus in the early phase of COVID-19 pandemic and subsequent 1-year global expansion. J Med Virol 94:2035–2049. https://doi.org/10.1002/jmv.27580.
18. Tian D, Sun Y, Xu H, Ye Q. 2022. The emergence and epidemic characteristics of the highly mutated SARS-CoV-2 Omicron variant. J Med Virol 94:2376–2383. https://doi.org/10.1002/jmv.27643.
19. Del Veliz S, Rivera L, Bustos DM, Uhart M. 2021. Analysis of SARS-CoV-2 nucleocapsid phosphoprotein N variations in the binding site to human 14-3-3 proteins. Biochem Biophys Res Commun 569:154–160. https://doi.org/10.1016/j.bbrc.2021.06.100.
20. Yang D, Leibowitz JL. 2015. The structure and functions of coronavirus genomic 3′ and 5′ ends. Virus Res 206:120–133. https://doi.org/10.1016/j.virusres.2015.02.025.
21. Miao Z, Tidu A, Eriani G, Martin F. 2021. Secondary structure of the SARS-CoV-2 5′-UTR. RNA Biol 18:447–456. https://doi.org/10.1080/15476286.2020.1814556.
22. Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, O'Meara MJ, Rezelj VV, Guo JZ, Swaney DL, Tummino TA, Hüttenhain R, Kaake RM, Richards AL, Tutuncuoglu B, Foussard H, Batra J, Haas K, Modak M, Kim M, Haas P, Polacco BJ, Braberg H, Fabius JM, Eckhardt M, Soucheray M, Bennett MJ, Cakir M, McGregor MJ, Li Q, Meyer B, Roesch F, Vallet T, Mac Kain A, Miorin L, Moreno E, Naing ZZC, Zhou Y, Peng S, Shi Y, Zhang Z, Shen W, Kirby IT, Melnyk JE, Chorba JS, Lou K, Dai SA, Barrio-Hernandez I, Memon D, Hernandez-Armenta C, et al. 2020. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature 583:459–468. https://doi.org/10.1038/s41586-020-2286-9.
23. Gupta M, Azumaya CM, Moritz M, Pourmal S, Diallo A, Merz G, Jang G, Bouhaddou M, Fossati A, Brilot AF, Diwanji D, Hernandez E, Herrera N, Kratochvil HT, Lam VL, Li F, Li Y, Nguyen HC, Nowotny C, Owens TW, Peters JK, Rizo AN, Schulze-Gahmen U, Smith AM, Young ID, Yu Z, Asarnow D, Billesbolle C, Campbell MG, Chen J, Chen KH, Chio US, Dickinson MS, Doan L, Jin M, Kim K, Li J, Li YL, Linossi E, Liu Y, Lo M, Lopez J, Lopez KE, Mancino A, Moss FR, Paul MD, Pawar KI, Pelin A, Pospiech TH, Puchades C, Remesh SG, Safari M, Schaefer K, et al. 2021. CryoEM and AI reveal a structure of SARS-CoV-2 Nsp2, a multifunctional protein involved in key host processes. bioRxiv. https://doi.org/10.1101/2021.05.10.443524.
24. Devaraj SG, Wang N, Chen Z, Chen Z, Tseng M, Barretto N, Lin R, Peters CJ, Tseng C-TK, Baker SC, Li K. 2007. Regulation of IRF-3-dependent innate immunity by the papain-like protease domain of the severe acute respiratory syndrome coronavirus. J Biol Chem 282:32208–32221. https://doi.org/10.1074/jbc.M704870200.
25. Fehr AR, Channappanavar R, Jankevicius G, Fett C, Zhao J, Athmer J, Meyerholz DK, Ahel I, Perlman S. 2016. The conserved coronavirus macrodomain promotes virulence and suppresses the innate immune response during severe acute respiratory syndrome coronavirus infection. mBio 7:e01721-16. https://doi.org/10.1128/mBio.01721-16.
26. Hosseini Rad SM A, McLellan AD. 2020. Implications of SARS-CoV-2 mutations for genomic RNA structure and host microRNA targeting. Int J Mol Sci 21:4807. https://doi.org/10.3390/ijms21134807.
27. Winther TN, Bang-Berthelsen CH, Heiberg IL, Pociot F, Hogh B. 2013. Differential plasma microRNA profiles in HBeAg positive and HBeAg negative children with chronic hepatitis B. PLoS One 8:e58236. https://doi.org/10.1371/journal.pone.0058236.
28. Yu K, Shi G, Li N. 2015. The function of microRNA in hepatitis B virus-related liver diseases: from dim to bright. Ann Hepatol 14:450–456. https://doi.org/10.1016/S1665-2681(19)31165-2.
29. Hasan M, McLean E, Bagasra O. 2016. A computational analysis to construct a potential post-exposure therapy against pox epidemic using miRNAs in silico. J Bioterror Biodef 7:140. https://doi.org/10.4172/2157-2526.1000140.
30. Jang WD, Jeon S, Kim S, Lee SY. 2021. Drugs repurposed for COVID-19 by virtual screening of 6,218 drugs and cell-based assay. Proc Natl Acad Sci U S A 118:e2024302118. https://doi.org/10.1073/pnas.2024302118.
31. Biswas SK, Mudi SR. 2020. Spike protein D614G and RdRp P323L: the SARS-CoV-2 mutations associated with severity of COVID-19. Genomics Inform 18:e44. https://doi.org/10.5808/GI.2020.18.4.e44.

32. Yan L, Ge J, Zheng L, Zhang Y, Gao Y, Wang T, Huang Y, Yang Y, Gao S, Li M, Liu Z, Wang H, Li Y, Chen Y, Guddat LW, Wang Q, Rao Z, Lou Z. 2021. Cryo-EM structure of an extended SARS-CoV-2 replication and transcription complex reveals an intermediate state in cap synthesis. Cell 184:184–193.E10. https://doi.org/10.1016/j.cell.2020.11.016.
33. Gao Y, Yan L, Huang Y, Liu F, Zhao Y, Cao L, Wang T, Sun Q, Ming Z, Zhang L, Ge J, Zheng L, Zhang Y, Wang H, Zhu Y, Zhu C, Hu T, Hua T, Zhang B, Yang X, Li J, Yang H, Liu Z, Xu W, Guddat LW, Wang Q, Lou Z, Rao Z. 2020. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. Science 368:779–782. https://doi.org/10.1126/science.abb7498.
34. Lehmann KC, Gulyaeva A, Zevenhoven-Dobbe JC, Janssen GMC, Ruben M, Overkleeft HS, van Veelen PA, Samborskiy DV, Kravchenko AA, Leontovich AM, Sidorov IA, Snijder EJ, Posthuma CC, Gorbalenya AE. 2015. Discovery of an essential nucleotidylating activity associated with a newly delineated conserved domain in the RNA polymerase-containing protein of all nidoviruses. Nucleic Acids Res 43:8416–8434. https://doi.org/10.1093/nar/gkv838.
35. Spratt AN, Kannan SR, Woods LT, Weisman GA, Quinn TP, Lorson CL, Sönnerborg A, Byrareddy SN, Singh K. 2021. Evolution, correlation, structural impact and dynamics of emerging SARS-CoV-2 variants. Comput Struct Biotechnol J 19:3799–3809. https://doi.org/10.1016/j.csbj.2021.06.037.
36. Mohammad A, Al-Mulla F, Wei D-Q, Abubaker J. 2021. Remdesivir MD simulations suggest a more favourable binding to SARS-CoV-2 RNA dependent RNA polymerase mutant P323L than wild-type. Biomolecules 11:919. https://doi.org/10.3390/biom11070919.
37. Hyser JM, Estes MK. 2015. Pathophysiological consequences of calcium-conducting viroporins. Annu Rev Virol 2:473–496. https://doi.org/10.1146/annurev-virology-100114-054846.
38. Oulas A, Zanti M, Tomazou M, Zachariou M, Minadakis G, Bourdakou MM, Pavlidis P, Spyrou GM. 2021. Generalized linear models provide a measure of virulence for specific mutations in SARS-CoV-2 strains. PLoS One 16:e0238665. https://doi.org/10.1371/journal.pone.0238665.
39. Issa E, Merhi G, Panossian B, Salloum T, Tokajian S. 2020. SARS-CoV-2 and ORF3a: nonsynonymous mutations, functional domains, and viral pathogenesis. mSystems 5:e00266-20. https://doi.org/10.1128/mSystems.00266-20.
40. Ren Y, Shu T, Wu D, Mu J, Wang C, Huang M, Han Y, Zhang X-Y, Zhou W, Qiu Y, Zhou X. 2020. The ORF3a protein of SARS-CoV-2 induces apoptosis in cells. Cell Mol Immunol 17:881–883. https://doi.org/10.1038/s41423-020-0485-9.
41. Miao G, Zhao H, Li Y, Ji M, Chen Y, Shi Y, Bi Y, Wang P, Zhang H. 2021. ORF3a of the COVID-19 virus SARS-CoV-2 blocks HOPS complex-mediated assembly of the SNARE complex required for autolysosome formation. Dev Cell 56:427–442.E5. https://doi.org/10.1016/j.devcel.2020.12.010.
42. Lu B, Tao L, Wang T, Zheng Z, Li B, Chen Z, Huang Y, Hu Q, Wang H. 2009. Humoral and cellular immune responses induced by 3a DNA vaccines against severe acute respiratory syndrome (SARS) or SARS-like coronavirus in mice. Clin Vaccine Immunol 16:73–77. https://doi.org/10.1128/CVI.00261-08.
43. Zhong X, Guo Z, Yang H, Peng L, Xie Y, Wong T-Y, Lai S-T, Guo Z. 2006. Amino terminus of the SARS coronavirus protein 3a elicits strong, potentially protective humoral responses in infected patients. J Gen Virol 87:369–373. https://doi.org/10.1099/vir.0.81078-0.
44. Kern DM, Sorum B, Mali SS, Hoel CM, Sridharan S, Remis JP, Toso DB, Kotecha A, Bautista DM, Brohawn SG. 2021. Cryo-EM structure of SARS-CoV-2 ORF3a in lipid nanodiscs. Nat Struct Mol Biol 28:573–582. https://doi.org/10.1038/s41594-021-00619-0.
45. Azad GK, Khan PK. 2021. Variations in Orf3a protein of SARS-CoV-2 alter its structure and function. Biochem Biophys Rep 26:100933.
46. Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, Li F. 2020. Cell entry mechanisms of SARS-CoV-2. Proc Natl Acad Sci U S A 117:11727–11734. https://doi.org/10.1073/pnas.2003138117.
47. Jackson CB, Farzan M, Chen B, Choe H. 2022. Mechanisms of SARS-CoV-2 entry into cells. Nat Rev Mol Cell Biol 23:3–20. https://doi.org/10.1038/s41580-021-00418-x.
48. Periwal N, Rathod SB, Pal R, Sharma P, Nebhnani L, Barnwal RP, Arora P, Srivastava KR, Sood V. 2021. In silico characterization of mutations circulating in SARS-CoV-2 structural proteins. J Biomol Struct Dyn Epub ahead of print. https://doi.org/10.1080/07391102.2021.1908170.
49. Bai Z, Cao Y, Liu W, Li J. 2021. The SARS-CoV-2 nucleocapsid protein and its role in viral structure, biological functions, and a potential target for drug or vaccine mitigation. Viruses 13:1115. https://doi.org/10.3390/v13061115.
50. Kopecky-Bromberg SA, Martínez-Sobrido L, Frieman M, Baric RA, Palese P. 2007. Severe acute respiratory syndrome coronavirus open reading frame (ORF) 3b, ORF 6, and nucleocapsid proteins function as interferon antagonists. J Virol 81:548–557. https://doi.org/10.1128/JVI.01782-06.
51. Surjit M, Liu B, Chow VTK, Lal SK. 2006. The nucleocapsid protein of severe acute respiratory syndrome-coronavirus inhibits the activity of cyclin-cyclin-dependent kinase complex and blocks S phase progression in mammalian cells. J Biol Chem 281:10669–10681. https://doi.org/10.1074/jbc.M509233200.
52. Mu J, Xu J, Zhang L, Shu T, Wu D, Huang M, Ren Y, Li X, Geng Q, Xu Y, Qiu Y, Zhou X. 2020. SARS-CoV-2-encoded nucleocapsid protein acts as a viral suppressor of RNA interference in cells. Sci China Life Sci 63:1413–1416. https://doi.org/10.1007/s11427-020-1692-1.
53. Wu H, Xing N, Meng K, Fu B, Xue W, Dong P, Tang W, Xiao Y, Liu G, Luo H, Zhu W, Lin X, Meng G, Zhu Z. 2021. Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2. Cell Host Microbe 29:1788–1801.e6. https://doi.org/10.1016/j.chom.2021.11.005.
54. Johnson BA, Zhou Y, Lokugamage KG, Vu MN, Bopp N, Crocquet-Valdes PA, Kalveram B, Schindewolf C, Liu Y, Scharton D, Plante JA, Xie X, Aguilar P, Weaver SC, Shi P-Y, Walker DH, Routh AL, Plante KS, Menachery VD. 2022. Nucleocapsid mutations in SARS-CoV-2 augment replication and pathogenesis. PLoS Pathog 18:e1010627. https://doi.org/10.1371/journal.ppat.1010627.
55. Prabantu VM, Naveenkumar N, Srinivasan N. 2020. Influence of cisease-causing mutations on protein structural networks. Front Mol Biosci 7:620554.
56. Shakhnovich EI, Gutin AM. 1991. Influence of point mutations on protein structure: probability of a neutral mutation. J Theor Biol 149:537–546. https://doi.org/10.1016/S0022-5193(05)80097-9.
57. Sotomayor-Vivas C, Hernández-Lemus E, Dorantes-Gilardi R. 2022. Linking protein structural and functional change to mutation using amino acid networks. PLoS One 17:e0261829. https://doi.org/10.1371/journal.pone.0261829.
58. Wako H, Endo S. 2017. Normal mode analysis as a method to derive protein dynamics information from the Protein Data Bank. Biophys Rev 9:877–893. https://doi.org/10.1007/s12551-017-0330-2.
59. Erman B. 2006. The Gaussian network model: precise predictions of residue fluctuations and application to binding problems. Biophys J 91:3589–3599. https://doi.org/10.1529/biophysj.106.090803.
60. Eyal E, Yang L-W, Bahar I. 2006. Anisotropic network model: systematic evaluation and a new web interface. Bioinformatics 22:2619–2627. https://doi.org/10.1093/bioinformatics/btl448.
61. Yang L, Song G, Jernigan RL. 2009. Protein elastic network models and the ranges of cooperativity. Proc Natl Acad Sci U S A 106:12347–12352. https://doi.org/10.1073/pnas.0902159106.
62. Shirvanyants D, Ding F, Tsao D, Ramachandran S, Dokholyan NV. 2012. Discrete molecular dynamics: an efficient and versatile simulation method for fine protein characterization. J Phys Chem B 116:8375–8382. https://doi.org/10.1021/jp2114576.
63. Hollingsworth SA, Dror RO. 2018. Molecular dynamics simulation for all. Neuron 99:1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011.
64. Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, Liu M, Kumar S, Zaremba S, Gu Z, Zhou L, Larson CN, Dietrich J, Klem EB, Scheuermann RH. 2012. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res 40:D593–D598. https://doi.org/10.1093/nar/gkr859.
65. Pickett BE, Liu M, Sadat EL, Squires RB, Noronha JM, He S, Jen W, Zaremba S, Gu Z, Zhou L, Larsen CN, Bosch I, Gehrke L, McGee M, Klem EB, Scheuermann RH. 2013. Metadata-driven comparative analysis tool for sequences (meta-CATS): an automated process for identifying significant sequence variations that correlate with virus attributes. Virology 447:45–51. https://doi.org/10.1016/j.virol.2013.08.021.
66. McKinney W. 2011. pandas: a foundational Python library for data analysis and statistics. Python for High Performance and Scientific Computing 14:1–9.
67. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, SciPy 1.0 Contributors. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2.

68. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millán C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD, Read RJ, Baker D. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. Science 373:871–876. https://doi.org/10.1126/science.abj8754.
69. Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, Brezovsky J, Damborsky J. 2014. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput Biol 10:e1003440. https://doi.org/10.1371/journal.pcbi.1003440.
70. Rodrigues CH, Pires DE, Ascher DB. 2018. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res 46:W350–W355. https://doi.org/10.1093/nar/gky300.
71. Frappier V, Najmanovich RJ. 2014. A coarse-grained elastic network atom contact model and its use in the simulation of protein dynamics and the prediction of the effect of mutations. PLoS Comput Biol 10:e1003569. https://doi.org/10.1371/journal.pcbi.1003569.
72. Pires DE, Ascher DB, Blundell TL. 2014. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30:335–342. https://doi.org/10.1093/bioinformatics/btt691.
73. Worth CL, Preissner R, Blundell TL. 2011. SDM—a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res 39:W215–W222. https://doi.org/10.1093/nar/gkr363.
74. Pires DE, Ascher DB, Blundell TL. 2014. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 42:W314–W319. https://doi.org/10.1093/nar/gku411.
75. Li G, Panday SK, Alexov E. 2021. SAAFEC-SEQ: a sequence-based method for predicting the effect of single point mutations on protein thermodynamic stability. Int J Mol Sci 22:606. https://doi.org/10.3390/ijms22020606.
76. Negre CFA, Morzan UN, Hendrickson HP, Pal R, Lisi GP, Loria JP, Rivalta I, Ho J, Batista VS. 2018. Eigenvector centrality for characterization of protein allosteric pathways. Proc Natl Acad Sci U S A 115:E12201–E12208. https://doi.org/10.1073/pnas.1810452115.
77. Penkler DL, Atilgan C, Tastan Bishop Ö. 2018. Allosteric modulation of human Hsp90α conformational dynamics. J Chem Inf Model 58:383–404. https://doi.org/10.1021/acs.jcim.7b00630.
78. Sethi A, Eargle J, Black AA, Luthey-Schulten Z. 2009. Dynamical networks in tRNA: protein complexes. Proc Natl Acad Sci U S A 106:6620–6625. https://doi.org/10.1073/pnas.0810961106.
79. Van Wart AT, Durrant J, Votapka L, Amaro RE. 2014. Weighted implementation of suboptimal paths (WISP): an optimized algorithm and tool for dynamical network analysis. J Chem Theory Comput 10:511–517. https://doi.org/10.1021/ct4008603.
80. Tekpinar M, Neron B, Delarue M. 2021. Extracting dynamical correlations and identifying key residues for allosteric communication in proteins by correlationplus. J Chem Inf Model 61:4832–4838. https://doi.org/10.1021/acs.jcim.1c00742.