

Analysis of SARS-CoV-2 genome evolutionary patterns

Shubhangi Gupta,¹ Deepanshu Gupta,² Sonika Bhatnagar^{1,2}

AUTHOR AFFILIATIONS See affiliation list on p. 18.

ABSTRACT The spread of SARS-CoV-2 virus accompanied by public availability of abundant sequence data provides a window for the determination of viral evolutionary patterns. In this study, SARS-CoV-2 genome sequences were collected from seven countries in the period January 2020–December 2022. The sequences were classified into three phases, namely, pre-vaccination, post-vaccination, and recent period. Comparison was performed between these phases based on parameters like mutation rates, selection pressure (dN/dS ratio), and transition to transversion ratios (Ti/Tv). Similar comparisons were performed among SARS-CoV-2 variants. Statistical significance was tested using Graphpad unpaired t -test. The analysis showed an increase in the percent genomic mutation rates post-vaccination and in recent periods across all countries from the pre-vaccination sequences. Mutation rates were highest in NSP3, S, N, and NSP12b before and increased further after vaccination. NSP4 showed the largest change in mutation rates after vaccination. The dN/dS ratios showed purifying selection that shifted toward neutral selection after vaccination. N, ORF8, ORF3a, and ORF10 were under highest positive selection before vaccination. Shift toward neutral selection was driven by E, NSP3, and ORF7a in the after vaccination set. In recent sequences, the largest dN/dS change was observed in E, NSP1, and NSP13. The Ti/Tv ratios decreased with time. C→U and G→U were the most frequent transitions and transversions. However, U→G was the most frequent transversion in recent period. The Omicron variant had the highest genomic mutation rates, while Delta showed the highest dN/dS ratio. Protein-wise dN/dS ratio was also seen to vary across the different variants.

IMPORTANCE To the best of our knowledge, there exists no other large-scale study of the genomic and protein-wise mutation patterns during the time course of evolution in different countries. Analyzing the SARS-CoV-2 evolutionary patterns in view of the varying spatial, temporal, and biological signals is important for diagnostics, therapeutics, and pharmacovigilance of SARS-CoV-2.

KEYWORDS SARS-CoV-2, genome, mutation rates, selection pressure, evolution, vaccination

Coronavirus disease 2019 (COVID-19) emerged as an epidemic with its initial spread in Wuhan (China) and was declared as a pandemic in March 2020 by the World Health Organization. Over 650 million confirmed cases of COVID-19, with more than 6 million deaths have been reported worldwide by WHO as of 15 May 2022 (1). In India, the first case of COVID-19 was reported in January 2020. An exponential increase in the number of cases from March 2020 to September 2020 marked the first wave. The second wave was marked by a sudden increase in COVID-19 cases from 12 March 2021 and high death rates (2). The Omicron variant of the causative SARS-CoV-2 characterized the third wave of COVID-19 with higher transmissibility but lower death rates (3).

SARS-CoV-2 has continuously evolved over time, resulting in numerous genetic variants across the globe (4). These variants are classified into three main categories, i.e., variant under monitoring (VUM), variant of interest (VOI), and variant of concern (VOC)

Editor Cyprian C. Rossetto, University of Nevada
Reno, Reno, Nevada, USA

Address correspondence to Sonika Bhatnagar,
sbhatnagar@nsut.ac.in.

The authors declare no conflict of interest.

See the funding table on p. 19.

Received 26 June 2023

Accepted 20 November 2023

Published 10 January 2024

Copyright © 2024 Gupta et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

(5). The VUM includes the GH/G (S gene D614G variant) progeny. VOI includes the Mu (lineage B.1.621) and Lambda (C.37), whereas VOC includes the alpha (lineage B.1.1.7), beta (lineage B.1.351), gamma (lineage P.1), delta (lineage B.1.617.2), and omicron variants (lineages BA1, BA2, BA4, and BA5) (5, 6).

The S protein is responsible for SARS-CoV-2 attachment and entry by binding to the ACE2 (7–10). Continuous mutations in the Spike increase virus adaptability for escaping vaccine treatment resulting in high survival rates and spread of the virus (11–14). Multiple mutations in the omicron variant S protein may also influence its interaction with ACE2, thereby leading to antibody escape (15, 16). Mutations in the S region of SARS-CoV-2 have led to epitope loss, resulting in escape from the vaccine treatment, the most frequent mutation being D614G (17, 18). D614G is also found in the S region of clades G/GR/GRY/GH/GV that has high human host infectivity rate due to efficient transmission (19, 20).

The SARS-CoV-2 genome contains 14 ORFs. Of these, *ORF1a* encodes the NSPs 1–11, while *ORF1b* encodes NSP12–16. Together, NSPs 1–16 form the replicase-transcriptase complex. This is followed by 13 ORFs encoding the four main structural proteins, namely, S, E, M, N, and interspersed by 9 accessory factors (21). Various studies have been performed on genomic mutations in the SARS-CoV-2 virus. The ratio of $dN/dS > 1$ indicates positive selection and has been reported in the S glycoprotein (22, 23). Comparative analysis of SARS-CoV-2, SARS-CoV, and MERS-CoV showed a positive evolution model along with higher dN/dS due to the dominance of dN (24). Studies of mutations in the diagnostic targets in COVID-19 have suggested that the N gene has the highest number of mutations (25, 26).

Analysis of 469 genome sequences from Indian patients led to the identification of 536 dN and dS mutations in the six genes; *ORF1ab*, S, N, *ORF3a*, *ORF7a*, and *ORF8* (27). A broad analysis showed 33 different mutations in 837 Indian SARS-CoV-2 whole-genome sequence isolates, of which 18 were unique to India. S, N, NSP3, NSP12, and NSP2 coding genes showed novel mutations and dN was found to be more than dS by approximately threefold (28). Modeling of the epidemic with different strains and mutations showed the emergence of a virus with higher transmissibility and evolutionary adaptations (29).

In this study, we calculated the genomic rates of mutation and dN/dS ratios in SARS-CoV-2 genome sequences taken from seven countries and showed that they increased with time, whereas the Ti/Tv decreased. Similarly, these parameters were also estimated for different known SARS-CoV-2 variants, where Omicron variant sequences showed highest mutation rate as compared to the other variants, but delta showed the highest dN/dS ratio. The highest mutating protein along with their individual dN/dS ratios and mutation rates were determined within each country. NSP3, S, N, and NSP12b had the highest genomic mutation rates both in before and after vaccination phase. NSP3 showed highest genomic mutation rates before vaccination, which was replaced by S in the after vaccination, while a significant rise was observed in the NSP4 genomic mutation rate. N, ORF8, ORF3a, and ORF10 were under strong positive selection pressure before vaccination, whereas after vaccination, E, ORF7a, and NSP3 showed the highest increase in the dN/dS ratio. The recent sequences showed the highest dN/dS ratios in E, NSP1, and NSP13. The estimated properties showed similar patterns in genetic variability across all geographic regions despite the use of different vaccine technologies. This implies that the forces of evolution have been uniform across multiple parameters. However, there is a definite change in the pattern of mutations with time. While the highly mutating and positively selected genes are important for pharmacovigilance and vaccination, the negatively selected ones are important for diagnostics. This is the only study that has compared these different mutational parameters comprehensively with time and geographical regions, and in different variants both at the whole genome and gene level.

MATERIALS AND METHODS

Collection of genomic sequences of SARS-CoV-2

To perform a critical analysis of the data, genome sequences of SARS-CoV-2 were retrieved in FASTA format for different countries using the GISAID database (30, 31). The countries taken into consideration included India, England, Canada, Italy, France, USA (Washington), and the Netherlands. Additionally, genome sequences of different SARS-CoV-2 variants like Alpha, Beta, Gamma, Lambda, Mu, Delta, and Omicron were also retrieved.

The accuracy of genome sequences retrieved from the database was ensured using filters like complete sequence, high coverage, and complete collection date. The sequences downloaded were high coverage, having <1% Ns (unidentified nucleotides) and <0.05% unique amino acid mutations. In all the sequences, insertions and deletions were accepted only when verified by the submitter.

Classifying the genome sequences retrieved

SARS-CoV-2 genome sequences submitted in the GISAID database from the seven countries were downloaded from three distinct periods. The genome sequences were classified into three categories as follows:

- The sequences considered pre-vaccination sequences had the sample collection date of at least 1 month before the start date of vaccination taken from the website of WHO and from the Government vaccination websites of the concerned countries. The data sets were taken in triplicates.
- The start date of vaccination was different for each country considered. After the date of start of vaccination, a buffer period is required for the vaccine to reach the population. Therefore, sequences collected 5–6 months after the date of start of vaccination were considered as after vaccination sequences. The data sets were taken in triplicates.
- A set of sequences that constitute a recent data set were also taken. As the available number of recent sequences were less in number for many countries, triplicate data sets for this period could not be taken.
- Additionally, approximately 1,000 sequences of each SARS-CoV-2 variant were obtained using the “Variant” filter of GISAID.

Sequence alignment and mutational data analysis

The first reported sequence, Wuhan-Hu-1 (Accession ID: NC045512), was taken as the reference for pre-vaccination, post-vaccination, and recent sequences. Thus, Wuhan-Hu-1 sequence was added as a reference for the alignment. The mutation rates of all the genomic sequences were calculated using COVID-19 genome annotator, an online web-based tool that aligns the input sequences in “nucmer” alignment tool and processes the alignment output using UNIX along with R scripts (32). The aligned sequences were compared with the reference genome (Wuhan-Hu-1) and amongst each other to find information about the mutational sites. The Jupyter notebook (33) was used in association with the “pandas” python library (34) to visualize and analyse the data obtained. The following values were calculated:

- Percentage mutation rate: The mutation rate has been calculated using the following equation:

$$\text{Mutation Rate (\%)} = \frac{\text{Total no. of nucleotide mutations}}{\text{lg} * \text{gs}} \times 100$$
 where “lg” refers

to the length of the data set taken, “gs” refers to the length of the genome sequence and the total no. of mutations estimated with respect to number of single nucleotide mutations (35). Mutation rate is the measure that refers to the frequency of mutations per generation in the population or in an organism (36).

- b. Selection pressure has been calculated as the dN/dS ratio. If dN/dS ratio exceeds unity ($dN/dS > 1$), the mutations are said to be occurring under positive selection which promotes the accumulation of beneficial mutations, whereas if dN/dS ratio is below unity ($dN/dS < 1$), the mutations are said to be occurring under negative selection that promotes mutations that are favouring selective removal of deleterious alleles (37, 38). A mutation table was obtained after the multiple sequence alignment was performed by the COVID-19 genome annotator. The table included the varclass information, i.e., the type/class of mutation that occurred in the input sample sequences. Under varclass, the mutations described as SNP were taken as observed non-synonymous mutations (N_m), whereas SNP_silent were taken observed synonymous mutations (S_m). Further, the selection pressure dN/dS were determined by normalizing the estimated N_m and S_m with the overall expected number of non-synonymous (N_{ref}) and synonymous (S_{ref}) sites in the Wuhan-Hu-1 reference genome. N_{ref} and S_{ref} sites were determined using a biopython script in which the codon substitution table from *T. gojobori* was implemented (39). Therefore, dN/dS ratio was calculated as: $\frac{dN}{dS} = \frac{N_m/N_{ref}}{S_m/S_{ref}}$

- c. The mutation table obtained from the COVID-19 genome annotator also provided information about mutation type/class (varclass) occurring in the respective protein region (protein). Both the “varclass” and “protein” information was considered together to calculate the number of SNP (N_m) and SNP_silent (S_m) for each SARS-CoV-2 protein. Furthermore, the selection pressure dN/dS of each protein were determined by normalizing the estimated N_m and S_m of the protein with their respective expected number of non-synonymous (N_{ref}) and synonymous (S_{ref}) sites within the Wuhan-Hu-1 reference genome. These N_{ref} and S_{ref} sites were determined using a biopython script. Therefore, dN/dS ratio was

$$\text{calculated as } \left(\frac{dN}{dS} \right)_{\text{of protein}} = \frac{\left(\frac{N_m}{N_{ref}} \right)_{\text{of the protein}}}{\left(\frac{S_m}{S_{ref}} \right)_{\text{of the protein}}}$$

- d. The Ti/Tv in the RNA virus has been calculated as the ratio of Ti ($C \leftrightarrow U$ and $A \leftrightarrow G$) to Tv ($A \leftrightarrow C$, $A \leftrightarrow U$, $G \leftrightarrow C$, and $G \leftrightarrow U$), determined for the pre-, post-, and recent vaccination sequence groups using the data obtained from the COVID-19 genome annotator. The mutation table data from the genome annotator provided the refvar (sequence at the mutation site on reference genome) and qvar (sequence at the mutation site on sample sequence) mutation information. The number of Ti mutations C-U, U-C, A-G, and G-A was counted using Microsoft Excel. Similarly, C-A, A-C, U-A, A-U, G-U, U-G, C-G, and G-C changes were counted as the total number of Tv . Therefore, the Ti/Tv ratio was calculated as:

$$Ti/Tv = \frac{\text{Total no. of transitions}}{\text{Total no. of transversions}}$$

Statistical analysis

The statistical significance of the differences between the means was carried out. The unpaired t -test was performed on the % mutation rates, selection pressure and Ti/Tv ratios using GraphPad QuickCalcs (<https://www.graphpad.com/quickcalcs/ttest1.cfm>).

RESULTS AND DISCUSSION

In this work, we have carried out comparison of the SARS-CoV-2 genome sequences from seven different demographic regions, namely, India, France, England, Canada, Italy, Netherlands, and USA (Washington, D.C.). A total of 74,870 retrieved sequences were classified into three different time periods:

- a. Pre-vaccination phase: Broadly, the pre-vaccination sequences taken were ranging from January 2020 to December 2020 (GISAID Identifier: EPI_SET_230517 nt).
- b. Post-vaccination phase: The post-vaccination sequences taken were ranging from May 2021 to April 2022 (GISAID Identifier: EPI_SET_230517va).
- c. Recent period: The recent period sequences were taken from June 2022 to December 2022 (GISAID Identifier: EPI_SET_230518oq).
- d. Additionally, a total of 7,209 sequences of different variants were also analyzed. As new variants emerged sequentially over time, collection dates of the sequences considered for the SARS-CoV-2 variants ranged from November 2020 to January 2022 (GISAID Identifier: EPI_SET_230518xv).

The sequences were compared for % mutation rates, dN/dS ratio and Ti/Tv ratio. The analysis was carried out for the whole genome as well as for every individual gene.

Evolutionary patterns in different geographical locations and time periods

- a. Increase in mutation rates: The % genomic mutation rates estimated for each country in the pre-vaccination, post-vaccination, and the recent period are listed in Table 1. The genome mutation rates were calculated as (0.04 ± 0.02) % in the pre-vaccination phase. The average mutation rates increased to (0.17 ± 0.05) % in the post-vaccination period. A comparable increase was seen in all the seven countries studied. The mutation rates for sequences from a more recent interval were higher than the post-vaccination period. The mutation rate was found to be (0.28 ± 0.01) %, consistently across all the seven countries. Thus, there was an average increase of three- to fourfold in the % genomic mutation rates in all the countries after vaccination and six- to sevenfold increase in the recent period in comparison with mutation rates in the pre-vaccination sequences. Figure. 1a shows the comparison of percent mutation rates within the countries in pre-vaccination, post-vaccination, and recent period. There was a similar increase in the % genomic mutation rates in the post-vaccination period in comparison with the pre-vaccination period across all the seven countries studied. This has increased further in the recent sequences. Despite the large variation in mutation rates in each country, there was a definite increase in the post-vaccination sequences especially in the post-vaccination period. In the recent sequences, the mutation rates show a still higher trend in all the seven countries studied. The significance of the difference between the means was verified using unpaired *t*-test (*P*-value 0.0001; confidence interval of 95%) and is shown in Table 2. The differences in percent genomic mutation rates were found to be extremely statistically significant between the three phases with a two-tailed *P*-value of less than 0.0001.

In RNA viruses, proofreading incapability of RNA polymerases accounts for their high mutation rates. Natural selection further results in increased adaptability and faster replication of viruses referring to their high mutation rates (40, 41). However, the highly conserved nsp14 in the *Coronaviridae* family has a proofreading function that may be a crucial factor for variation in the large and complex viral genome (42). Absolute estimates

TABLE 1 Data comparison of different countries before and after vaccination along with the recent data

Countries	India	France	England	Canada	Italy	Netherlands	USA (Washington DC)	Total sequences Average of properties
Properties								
Date of start of vaccination	16/01/2021	27/12/2020	08/12/2020	14/12/2020	27/12/2020	06/01/2021	17/12/2020	–
Vaccine used (Technology)	Covishield/ Covaxin	Pfizer	Pfizer	Pfizer	Pfizer	Pfizer	Pfizer/Moderna	–
No. of sequences before vaccination (Jan 2020)–(1Dec 2020)	5,022	5,071	5,114	5,178	4,531	4,964	5,090	34,970
No. of sequences after vaccination (May 2020–April 2022)	5,270	5,282	5,070	5,199	4,953	5,029	5,217	36,020
No. of recent sequences (June 2022–Dec 2022)	572	531	555	591	540	533	558	3,880
% Mutation rates (before)	0.05 ± 0.02	0.04 ± 0.02	0.03 ± 0.01	0.04 ± 0.01	0.05 ± 0.03	0.05 ± 0.02	0.04 ± 0.02	0.04 ± 0.02
% Mutation rates (after)	0.17 ± 0.06	0.18 ± 0.06	0.17 ± 0.06	0.18 ± 0.06	0.17 ± 0.06	0.18 ± 0.06	0.18 ± 0.06	0.17 ± 0.05
% Mutation rates (recent)	0.29	0.28	0.28	0.29	0.26	0.28	0.27	0.28 ± 0.01
Ratio of mutation rates (after/before)	3.40	4.50	5.67	4.50	3.40	3.60	4.50	4.22 ± 0.76
Ratio of mutation rates (recent/before)	5.80	7.00	9.33	7.25	5.20	5.60	6.75	6.70 ± 1.28
dN/dS ratio (before)	0.44 ± 0.03	0.50 ± 0.07	0.53 ± 0.08	0.55 ± 0.07	0.42 ± 0.09	0.40 ± 0.06	0.53 ± 0.04	0.48 ± 0.06
dN/dS ratio (after)	1.13 ± 0.26	0.90 ± 0.35	1.17 ± 0.24	0.94 ± 0.33	0.92 ± 0.30	0.92 ± 0.35	0.95 ± 0.31	0.99 ± 0.11
dN/dS ratio (recent)	0.90	0.91	0.96	0.90	0.84	0.91	0.82	0.90 ± 0.05
dN/dS ratio (after/before)	2.57	1.80	2.20	1.71	2.20	2.30	1.80	2.08 ± 0.32
dN/dS ratio (recent/before)	2.04	1.82	1.81	1.64	2.00	2.27	1.55	1.88 ± 0.25
Ti/Tv ratio (before)	3.33 ± 0.24	3.44 ± 1.18	3.54 ± 0.61	4.81 ± 0.26	3.71 ± 2.03	3.33 ± 1.32	4.23 ± 1.28	3.77 ± 1.10
Ti/Tv ratio (after)	2.00 ± 0.06	1.93 ± 0.15	2.08 ± 0.03	1.96 ± 0.14	1.95 ± 0.15	1.97 ± 0.18	2.08 ± 0.11	2.00 ± 0.12
Ti/Tv ratio (recent)	2.15	2.11	2.05	2.22	2.18	2.14	2.18	2.15 ± 0.05
Ti/Tv ratio (after/before)	0.60	0.56	0.59	0.41	0.53	0.60	0.50	0.54 ± 0.06
Ti/Tv ratio (recent/before)	0.64	0.61	0.58	0.46	0.58	0.64	0.51	0.57 ± 0.06
% GC content (before)	37.93 ± 0.03	37.96 ± 0.02	37.92 ± 0.05	37.97 ± 0.01	37.92 ± 0.02	37.96 ± 0.02	37.95 ± 0.04	37.94 ± 0.02
% GC content (after)	37.81 ± 0.06	37.85 ± 0.09	37.84 ± 0.07	37.92 ± 0.03	37.84 ± 0.08	37.88 ± 0.08	37.81 ± 0.10	37.85 ± 0.07
% GC content (recent)	37.81	37.82	37.76	37.89	37.76	37.82	37.85	37.82 ± 0.05

^aThe prominent results reflecting change in mutation rates (%), dN/dS and Ti/Tv ratio from the pre-vaccination period for each country are shown in bold along with the overall change.

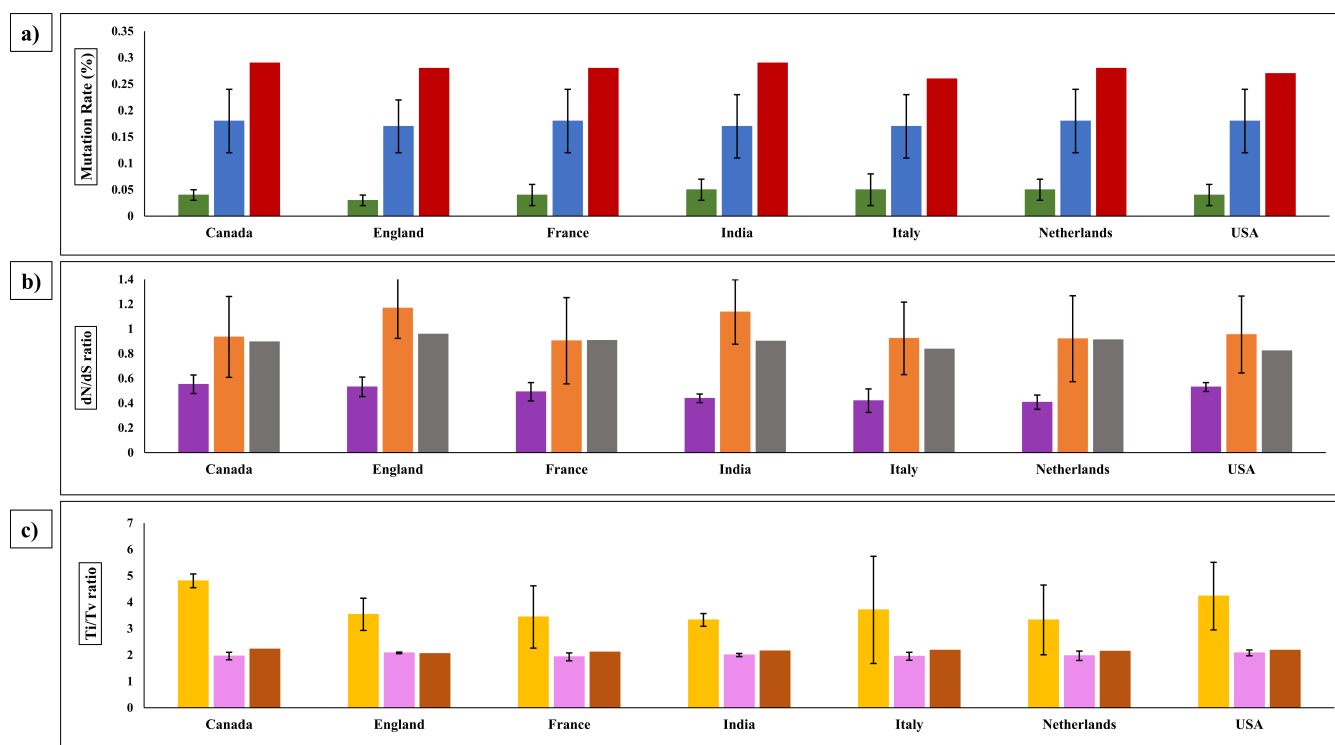


FIG 1 Comparison of mutation rates, dN/dS and Ti/Tv within the countries in the three phases. (a) Mutation rates with respect to before (green), after vaccination (blue), and recent period (red); (b) Selection pressure (dN/dS ratio) with respect to before (violet), after vaccination (orange), and recent period (gray); (c) Transition to transversion (Ti/Tv) ratio with respect to before (yellow), after vaccination (pink), and recent period (brown). Black margins in the graph represent the standard deviation in the before and after vaccination.

TABLE 2 Unpaired *t*-test calculations for the significance testing of the determined country-wise parameters

Group Comparison	Parameters of comparison	t-value	Degrees of freedom (df)	Mean difference	Standard error of difference	95% confidence interval (C.I.)	Two-tailed P-value
Pre-vaccination to Post vaccination period	Mutation rates	452.3	70,988	-0.13	0.000	-0.1306 to -0.1294	<0.0001
	dN/dS ratio	763.7		-0.51	0.001	-0.5113 to -0.5087	<0.0001
	Ti/Tv ratio	303.5		1.77	0.006	1.7585 to 1.7815	<0.0001
	% GC content	231.4		0.09	0.000	0.0892 to 0.0908	<0.0001
Post-vaccination and Recent period	Mutation rates	136.7	39,898	-0.11	0.001	-0.1116 to -0.1084	<0.0001
	dN/dS ratio	50.4		0.09	0.002	0.0865 to 0.0935	<0.0001
	Ti/Tv ratio	77.1		-0.15	0.002	-0.1538 to -0.1462	<0.0001
	% GC content	25.9		0.03	0.001	0.0277 to 0.0323	<0.0001
Pre-vaccination and Recent period	Mutation rates	737.3	38,848	-0.24	0.000	-0.2406 to -0.2394	<0.0001
	dN/dS ratio	420.1		-0.42	0.001	-0.4220 to -0.4180	<0.0001
	Ti/Tv ratio	91.7		1.62	0.018	1.5853 to 1.6547	<0.0001
	% GC content	287.2		0.12	0.000	0.1192 to 0.1208	<0.0001

of mutation rates in coronaviruses ranged from 0.67 to 1.33×10^{-5} per site per year in infectious bronchitis virus to 0.44 – 2.77×10^{-2} per site per year in mouse hepatitis virus. Comparatively, the global SARS-CoV-2 mutation rate was estimated to be moderate at 6×10^{-4} per site per year (43). The more than twofold increase in mutation rates with time in SARS-CoV-2 as observed in this study would bring the genomic mutation rates at par with the reported genomic mutation rates for non-coronaviruses and the other RNA viruses like Influenza virus and HIV-1 (44).

- **Mutation rates in SARS-CoV-2 proteins:** The pattern of mutations in proteins within the SARS-CoV-2 genome sequences was analyzed for each protein in the pre-vaccination, post-vaccination, and recent period (Table 3). As shown, NSP3, S, N, and NSP12b coding regions had the highest % mutation rates in the pre-vaccination period. There was little variation observed in the highest mutating genes in different countries studied. Previously, NSP3 was reported as the mutation hotspot of SARS-CoV-2 genome and mostly found to be co-mutating with NSP12b due to its involvement in replication and transcription complexes (45–49). However, in the post-vaccination and recent period, the highest mutating genes remained the same, where S was observed to uniformly have the highest % mutation rates followed by N and NSP3. From the pre-vaccination sequences, S showed gradual increase in the % mutation rates till the recent period across all the countries. S contains the receptor-binding domain (RBD) which is highly variable, and through its mutational changes, it is known to affect viral replication, transmission and is also involved in immune escape (50, 51). In comparison with S, N has lower mutation rates and is comparatively stable (52). N is considered as a vital hotspot for mutations mainly in its serine-rich domain due to its involvement in viral replication and packaging (53, 54) and can be considered a novel target for vaccine design (55). Comparison of mutation patterns of all the proteins in Table 4 showed that NSP4 had shown low mutation rates before vaccination but rose to be the fifth highest mutating protein post-vaccination. In recent sequences, it uniformly occupied the fourth position and overtook NSP12b in all the countries. NSP4 along with NSP6 and NSP3 transmembrane proteins, induce the development of double membrane vesicles (DMVs) by reorganizing the endoplasmic

TABLE 3 Data comparison of highly mutating genes before, after vaccinations, and recent period in each country along with their respective selection pressures

Country	Before vaccination (Jan 2020)–(Dec 2020)			After vaccination (May 2021–April 2022)			Recent period (June 2022– Dec 2022)		
	Highly mutating proteins	Selection pressure (dN/dS)	Mutation rate (%)	Highly mutating proteins	Selection pressure (dN/dS)	Mutation rate (%)	Highly mutating proteins	Selection pressure (dN/dS)	Mutation rate (%)
India	NSP3	0.16 ± 0.02	0.009 ± 0.002	fruiS	5.00 ± 1.00	0.05 ± 0.04	S	8.56	0.12
	S	0.63 ± 0.02	0.007 ± 0.003	NSP3	0.44 ± 0.14	0.02 ± 0.002	NSP3	0.26	0.03
	NSP12b	1.35 ± 0.31	0.005 ± 0.001	N	5.30 ± 1.93	0.01 ± 0.002	NSP4	1.04	0.02
England	NSP3	0.1 ± 0.03	0.005 ± 0.003	S	6.08 ± 0.80	0.05 ± 0.04	S	7.12	0.11
	S	1.27 ± 0.37	0.005 ± 0.002	NSP3	0.56 ± 0.26	0.02 ± 0.002	NSP3	0.22	0.03
	NSP12b	1.28 ± 0.80	0.004 ± 0.001	N	9.80 ± 4.6	0.01 ± 0.001	N	2.50	0.02
Canada	NSP3	0.08 ± 0.09	0.007 ± 0.004	S	6.16 ± 2.06	0.06 ± 0.04	S	6.58	0.12
	S	1.99 ± 1.10	0.005 ± 0.001	NSP3	0.35 ± 0.14	0.02 ± 0.003	NSP3	0.28	0.03
	NSP12b	1.57 ± 0.74	0.004 ± 0.001	N	3.52 ± 2.05	0.01 ± 0.002	NSP4	1.01	0.02
France	NSP3	0.08 ± 0.04	0.007 ± 0.003	S	4.4 ± 0.73	0.05 ± 0.04	S	6.02	0.11
	NSP12b	1.71 ± 1.18	0.006 ± 0.002	NSP3	0.43 ± 0.16	0.02 ± 0.003	NSP3	0.26	0.02
	S	3.60 ± 0.64	0.005 ± 0.002	N	6.13 ± 2.04	0.01 ± 0.002	N	1.95	0.02
Italy	S	2.17 ± 0.25	0.009 ± 0.005	S	4.27 ± 0.50	0.05 ± 0.04	S	5.46	0.10
	NSP3	0.09 ± 0.04	0.009 ± 0.004	NSP3	0.41 ± 0.15	0.02 ± 0.003	NSP3	0.26	0.02
	NSP12b	1.71 ± 1.18	0.005 ± 0.001	N	5.96 ± 2.00	0.01 ± 0.002	N	2.11	0.02
Netherlands	NSP3	0.09 ± 0.02	0.009 ± 0.004	S	4.50 ± 0.42	0.06 ± 0.04	S	6.36	0.11
	S	2.24 ± 0.74	0.007 ± 0.003	NSP3	0.37 ± 0.11	0.02 ± 0.003	NSP3	0.30	0.02
	NSP12b	1.07 ± 0.45	0.007 ± 0.003	N	4.73 ± 2.27	0.01 ± 0.001	N	2.61	0.02
USA (Washington DC)	NSP3	0.09 ± 0.03	0.005 ± 0.003	S	5.21 ± 1.14	0.05 ± 0.04	S	4.77	0.11
	S	2.01 ± 1.21	0.004 ± 0.003	NSP3	0.40 ± 0.14	0.02 ± 0.003	NSP3	0.28	0.02
	NSP12b	1.05 ± 0.46	0.004 ± 0.003	N	4.14 ± 3.35	0.01 ± 0.002	NSP4	1.04	0.02

TABLE 4 Highest (top) to lowest (bottom) protein-wise contribution to mutation rates in all the countries before, after vaccination and in the recent period

Before vaccination (Jan 2020)–(Dec 2020)										After vaccination (May 2021 – April 2022)					Recent period (June 2022 – Dec 2022)															
Canada	England	France	India	Italy	Netherland USA (WA)					Canada	England	France	India	Italy	Netherland USA (WA)															
					ds					ds					ds															
NSP3	NSP3	NSP3	NSP3	S	NSP3	NSP3	S	NSP3	NSP3	S	NSP3	NSP3	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
S	S	NSP12b	S	NSP3	S	NSP3	S	NSP3	NSP3	NSP3	NSP3	NSP3	NSP3	NSP3	NSP3	NSP3	NSP3	NSP3	NSP3	NSP3	NSP3	NSP3	NSP3	NSP3	NSP3	NSP3	NSP3	NSP3	NSP3	NSP3
NSP12b	NSP12b	S	NSP12b	NSP12b	NSP12b	NSP12b	NSP12b	NSP12b	NSP12b	NSP12b	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
5'UTR	N	5'UTR	N	N	N	5'UTR	N	5'UTR	ORF3a	NSP13	NSP4	NSP4	NSP4	NSP12b	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	N	N
NSP2	5'UTR	ORF3a	5'UTR	5'UTR	ORF3a	NSP13	NSP13	NSP13	NSP13	NSP13	N	5'UTR	NSP12b	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	M	M	M
ORF3a	NSP2	N	ORF3a	M	5'UTR	N	5'UTR	N	5'UTR	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2
N	ORF3a	NSP2	NSP14	NSP2	M	NSP2	M	NSP2	NSP2	NSP2	M	NSP6	3'UTR	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6
NSP13	NSP6	NSP13	NSP2	NSP16	NSP13	NSP13	NSP13	NSP13	NSP13	NSP13	ORF3a	M	NSP6	3'UTR	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6
NSP14	NSP5	M	NSP4	NSP1	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	3'UTR	NSP4	NSP12b	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4
NSP4	NSP14	NSP14	M	ORF10	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP14	3'UTR	NSP14	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4
NSP6	NSP15	NSP4	ORF8	ORF3a	ORF8	ORF8	NSP4	NSP4	NSP4	NSP4	ORF8	NSP13	NSP13	NSP13	NSP13	NSP13	NSP13	NSP13	NSP13	NSP13	NSP13	NSP13	NSP13	NSP13	NSP13	NSP13	NSP13	NSP13	NSP13	NSP13
M	M	NSP6	NSP6	NSP6	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4	NSP4
NSP15	NSP13	NSP5	3'UTR	3'UTR	NSP1	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15
NSP1	NSP4	NSP1	NSP13	NSP13	NSP14	NSP14	NSP14	NSP14	NSP14	NSP14	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16
3'UTR	3'UTR	3'UTR	NSP1	ORF8	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP5	NSP5	NSP5	NSP5	NSP5	NSP5	NSP5	NSP5	NSP5	NSP5	NSP5	NSP5	NSP5	NSP5	NSP5	NSP5	NSP5	NSP5	NSP5	NSP5
NSP8	NSP1	NSP15	NSP16	NSP5	NSP5	ORF10	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6	NSP6
NSP5	ORF8	NSP16	NSP15	NSP15	NSP14	NSP15	NSP15	NSP15	NSP15	NSP15	ORF7a	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15	NSP15
ORF8	NSP16	NSP16	NSP8	NSP4	NSP8	NSP8	NSP8	NSP8	NSP8	NSP8	M	NSP9	M	ORF6	ORF6	ORF6	ORF6	ORF6	ORF6	ORF6	ORF6	ORF6	ORF6	ORF6	ORF6	ORF6	ORF6	ORF6	ORF6	ORF6
ORF10	ORF7a	ORF7a	NSP5	NSP15	3'UTR	NSP1	NSP15	NSP15	NSP15	NSP15	NSP1	NSP1	NSP1	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2	NSP2
ORF7a	ORF10	ORF10	ORF7a	NSP9	NSP5	NSP8	NSP5	NSP5	NSP5	NSP5	NSP8	NSP8	NSP8	NSP8	NSP8	NSP8	NSP8	NSP8	NSP8	NSP8	NSP8	NSP8	NSP8	NSP8	NSP8	NSP8	NSP8	NSP8	NSP8	NSP8
NSP16	NSP8	NSP9	ORF6	ORF7a	NSP9	ORF6	NSP9	NSP9	NSP9	NSP9	ORF6	NSP1	NSP1	NSP1	NSP1	NSP1	NSP1	NSP1	NSP1	NSP1	NSP1	NSP1	NSP1	NSP1	NSP1	NSP1	NSP1	NSP1	NSP1	NSP1
NSP7	NSP7	NSP7	NSP7	NSP8	ORF7a	ORF10	ORF7a	ORF10	ORF10	ORF10	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E
ORF6	E	ORF7b	E	NSP7	ORF7b	NSP9	NSP9	NSP9	NSP9	NSP9	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16	NSP16
NSP9	NSP10	NSP8	ORF7b	ORF7b	E	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7
E	ORF7b	E	NSP9	NSP10	ORF6	E	NSP8	NSP10	NSP10	NSP10	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7
NSP10	NSP9	ORF6	ORF10	E	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10	NSP10
ORF7b	ORF6	NSP10	NSP10	ORF6	NSP7	NSP7	NSP7	NSP7	NSP7	NSP7	ORF7b	ORF10	ORF10	ORF10	ORF10	ORF10	ORF10	ORF10	ORF10	ORF10	ORF10	ORF10	ORF10	ORF10	ORF10	ORF10	ORF10	ORF10	ORF10	ORF10
NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a
NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a
NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a
NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a
NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a
NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a
NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a
NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a
NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a
NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a
NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a
NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a
NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a
NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a
NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a
NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a
NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a
NSP12a	NSP12a	NSP12a	NSP																											

reticulum of the host cell, and therefore it may be co-mutating with NSP3 (56, 57). In contrast with this, ORF10, NSP7, and NSP10 were the lowest mutating in all the three periods for all the countries taken into consideration. This was in line with previous studies where they were found to be conserved within the SARS-CoV-2 genome with few or no mutations (47, 58–60).

- b. **Increasing Selection pressure:** The dN/dS ratio in each country for the pre-vaccination, post-vaccination and the recent period are listed in Table 1. The global dN/dS ratio in the SARS-CoV-2 genome during the pre-vaccination period was calculated as 0.48 ± 0.06 , thus indicating a purifying selection pressure. The observed value is similar to previous studies showing that SARS-CoV-2 genome depicts overall purifying selection pressure (61–63) and study reporting overall SARS-CoV-2 genome dN/dS of around 0.55 (61). The dN/dS ratio increased to 0.99 ± 0.11 in the post-vaccination period in various demographic regions, suggesting SARS-CoV-2 genomes showing overall neutral selection. The recent period sequences showed an average dN/dS of 0.90 ± 0.05 . The comparison of dN/dS ratios in all the countries for pre-vaccination, post-vaccination and recent period is shown in Fig. 1b. There was more variation in the dN/dS value of sequences taken from the post-vaccination period in comparison before vaccination. Taken together, a definite increase in dN/dS is seen in post-vaccination sequences but the data from recent sequences was inconclusive because of small sample size. The differences in the mean dN/dS ratios were found to be extremely statistically significant between the three phases with a unpaired two-tailed t -test P -value of less than 0.0001 (Table 2).

The findings are in line with previous studies suggesting significant purifying selection which has been the primary factor driving the SARS-CoV-2 evolution and ongoing strong positive selection observed within certain sites of the SARS-CoV-2 genome (64). In a previous study, it has been suggested that increase in the selection pressures can cause increase in the mutation rates measured (65), which is also justified in our study as both mutation rates and dN/dS ratios are increasing with time. Combined with the persistence of recurrent infections in immunocompromised individuals, this may have induced the selection of viruses with lower pathogenicity or virulence and higher transmission (66).

- **Selection pressure in SARS-CoV-2 proteins:** The dN/dS ratios of the proteins in all the countries are shown in Table 5. SARS-CoV-2 proteins with the strongest positive selection pressures in the pre-vaccination period were N, ORF8, ORF3a, and ORF10. In a change from the pre-vaccination period, the post-vaccination marked substantial increase in E, ORF7a and NSP3. The recent phase sequences that E, NSP1, and NSP13 had and the highest dN/dS ratios. A continuous trend of positive selection was identified in S in all the three phases. Previous studies during the year 2020 have identified ORF10 to be under strong positive selection, along with ORF3a, ORF8 and N also depicting much higher number of dN (62, 67, 68). ORF8 is involved in host-pathogen interactions through its nine encoded proteins (69), ORF3a is involved in interference with host ion channels by encoding viroporin (70), while overexpression of ORF10 is known to downregulate IFN-1 expression, leading to suppression of the antiviral innate immune response (71). However, most viral proteins are multifunctional, and their other roles are likely to be discovered in future. ORF7a is responsible for triggering the NF-kappa B pathway and pro-inflammatory expression of cytokines (72). Also, evidence of strong positive selection pressures was observed in NSP3 from a 2021 study (73). NSP3 and its interaction with N is known to affect SARS-CoV-2 replication and

TABLE 5 Highest (top) to lowest (bottom) protein-wise contribution to selection pressure (dN/dS ratio) in all the countries before, after vaccination and in the recent period

Before vaccination										After vaccination					Recent period				
Canada	England	France	India	Italy	Netherlan USA (WA)					Canada	England	France	India	Italy	Netherlan USA (WA)				
					ds					ds					ds				
ORF8	N	ORF10	ORF3a	ORF10	ORF10	ORF3a	E	E	ORF7a	ORF7a	E	E	E	E	E	E	E	E	
ORF3a	ORF3a	N	N	N	ORF3a	ORF8	S	ORF7a	NSP3	E	NSP3	ORF8	ORF8	S	S	NSP1	NSP1	NSP1	
NSP7	NSP6	NSP7	NSP12b	S	S	N	ORF8	NSP3	S	S	ORF7a	ORF7a	S	NSP1	NSP1	S	S	S	
NSP8	ORF10	S	ORF8	NSP12b	NSP13	NSP13	NSP3	S	ORF8	NSP3	ORF8	S	ORF7a	N	NSP13	NSP13	NSP13	NSP14	
N	S	NSP13	NSP6	NSP13	N	NSP16	NSP14	ORF8	E	NSP13	S	NSP3	N	NSP13	N	ORF10	N	NSP13	
ORF10	NSP12b	NSP12b	NSP8	NSP8	NSP12b	S	NSP13	NSP1	NSP1	NSP14	NSP13	NSP1	NSP13	NSP14	NSP14	NSP8	NSP14	N	
NSP12b	NSP8	NSP4	NSP2	ORF3a	ORF7a	NSP5	ORF7a	NSP14	NSP1	NSP13	NSP1	NSP13	NSP14	NSP7	M	NSP4	ORF10	ORF10	
NSP16	NSP2	ORF3a	NSP15	ORF7a	NSP9	NSP12b	M	ORF10	ORF10	M	NSP4	NSP14	NSP16	ORF10	NSP6	NSP14	ORF8	ORF8	
S	NSP7	ORF7a	S	ORF8	NSP6	NSP2	NSP4	NSP13	NSP4	ORF8	NSP14	NSP4	M	NSP4	ORF10	NSP6	ORF8	NSP4	
NSP2	NSP5	ORF8	E	NSP4	NSP8	E	ORF3a	M	NSP14	NSP4	M	M	NSP1	NSP8	ORF8	NSP14	NSP4	M	
NSP15	NSP9	E	ORF7b	NSP14	NSP10	ORF10	NSP1	NSP4	ORF3a	NSP12b	ORF3a	ORF3a	NSP4	ORF8	NSP4	NSP6	NSP6	NSP16	
NSP6	ORF6	NSP8	NSP16	NSP5	NSP5	ORF7b	NSP6	NSP6	ORF3a	M	NSP6	ORF10	ORF3a	M	NSP2	NSP7	NSP12a	NSP7	
ORF7a	ORF7a	ORF6	ORF7a	E	NSP4	NSP6	NSP16	N	NSP6	NSP6	NSP10	NSP6	NSP10	NSP10	NSP16	NSP16	NSP7	NSP16	
ORF7b	NSP13	NSP2	ORF10	NSP6	NSP2	ORF7a	NSP8	NSP12b	NSP6	ORF10	ORF10	N	NSP3	NSP12b	NSP12b	NSP2	ORF3a	NSP3	
NSP13	E	NSP9	NSP7	NSP2	NSP7	NSP9	N	NSP16	NSP16	N	N	NSP15	NSP12b	ORF7a	NSP15	NSP12b	NSP3	NSP15	
E	NSP10	NSP16	NSP5	NSP7	NSP14	NSP1	NSP15	ORF7b	NSP15	ORF7b	NSP8	NSP12b	NSP6	NSP3	ORF6	NSP15	NSP6	NSP8	
NSP14	ORF8	NSP5	NSP13	NSP15	NSP15	NSP7	NSP12b	NSP6	NSP8	NSP15	NSP16	NSP16	ORF10	NSP2	NSP3	NSP15	NSP15	ORF6	
NSP5	NSP15	NSP15	NSP9	ORF6	ORF7b	NSP8	NSP10	NSP15	NSP12b	NSP2	NSP12b	ORF6	NSP8	NSP10	NSP3	NSP16	ORF6	NSP12b	
NSP9	NSP14	NSP1	NSP4	NSP9	E	ORF6	NSP2	NSP8	ORF7b	NSP16	NSP15	ORF7b	NSP15	NSP15	NSP5	NSP5	ORF3a	NSP2	
NSP4	NSP16	NSP10	NSP3	NSP3	NSP3	NSP4	ORF6	ORF6	ORF6	ORF6	ORF6	NSP2	ORF7b	ORF6	ORF7a	NSP7	NSP10	ORF3a	
NSP10	NSP4	NSP6	NSP10	ORF7b	ORF8	NSP14	ORF10	NSP2	NSP10	NSP10	ORF7b	NSP10	ORF6	NSP10	NSP5	NSP8	NSP5	ORF7a	
ORF6	ORF6	NSP14	ORF6	NSP10	ORF6	M	ORF7b	NSP10	NSP7	NSP8	NSP2	NSP5	NSP2	NSP7	NSP10	NSP5	NSP10	NSP5	
NSP3	M	NSP3	NSP1	NSP16	M	NSP3	NSP5	NSP5	NSP2	NSP7	NSP5	NSP8	NSP7	NSP6	NSP8	ORF7a	NSP6	NSP6	
NSP1	NSP3	ORF7b	NSP14	NSP1	NSP16	NSP10	NSP7	NSP7	NSP5	NSP5	NSP7	NSP7	NSP5	ORF7b	NSP9	NSP8	ORF7b	ORF7b	
ORF7a	NSP1	M	M	M	NSP1	NSP15	NSP9	NSP9	NSP9	NSP9	NSP9	NSP9	NSP9	NSP9	NSP7	ORF7b	NSP9	NSP9	
5'UTR	5'UTR	5'UTR	5'UTR	5'UTR	5'UTR	5'UTR	5'UTR	5'UTR	5'UTR	5'UTR	5'UTR	3'UTR	5'UTR	5'UTR	3'UTR	3'UTR	NSP9	NSP9	
3'UTR	3'UTR	3'UTR	3'UTR	3'UTR	3'UTR	3'UTR	3'UTR	3'UTR	3'UTR	3'UTR	3'UTR	3'UTR	3'UTR	3'UTR	3'UTR	3'UTR	NSP9	3'UTR	
-	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	-	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	NSP12a	-	-	5'UTR	-	NSP12a	

its pathogenesis (74). In the pre-vaccination phase, NSP1 was under negative selection pressure with $dN/dS < 1$. Also, in a previous study where NSP1 showed negatively selected sites only and it is designed to be evolutionary conserved for its functional requirement of host ribosomal complex binding (75, 76). *E* and *S* indicated strong positive selection pressures in a 2021 study (73). *E* is involved in the expression of a small multi-functional protein that has an important role in host and virus interaction through ion-channel activity. Mutations within *E* are associated with a reduction in virulence (77, 78). NSP13 has also shown strong evidence of positive selection which may be justified due to its role in inhibiting antiviral immunity by hijacking host deubiquitinase USP13 and suppresses type I IFN response through contact with TBK1 (64, 79). *S* has shown continuous strong positive selection in our study in all the three phases as supported by other studies (64, 80). NSP1 showed a large jump in the dN/dS ratio with a drastic increase in the number of dN after vaccination. NSP1 is suggested to suppress the host innate immunity by interacting with the 40S ribosomal subunit (76, 81, 82). NSP9 was observed to be the most common gene having $dN/dS < 1$ showing purifying selection. NSP9 is said to be highly conserved and involved in viral RNA synthesis (83).

- c. **Decrease in Ti/Tv ratio:** The estimated Ti/Tv ratios for each country in the pre-vaccination, post-vaccination, and the recent period are listed in Table 1. The Ti/Tv ratio in the pre-vaccination data set was found to be 3.77 ± 1.10 . The value is in agreement with the range of Ti/Tv of 2.0 to 5.5 calculated in previous studies (84, 85). The Ti/Tv ratio decreased in the post-vaccination period to 2.00 ± 0.12 and remained similar with the recent period sequences, i.e., 2.15 ± 0.05 . The differences in the mean Ti/Tv ratios were found to be highly statistically significant between the three phases with P -value of less than 0.0001 (Table 2). Figure. 1c shows the comparison of Ti/Tv ratios within the countries in pre-vaccination, post-vaccination, and recent period. As observed from the figure, the proportion of Tv were found to be increased post-vaccination. As Tv are said to have a dominant contribution toward dN within the genome, an increase in Tv can be directly

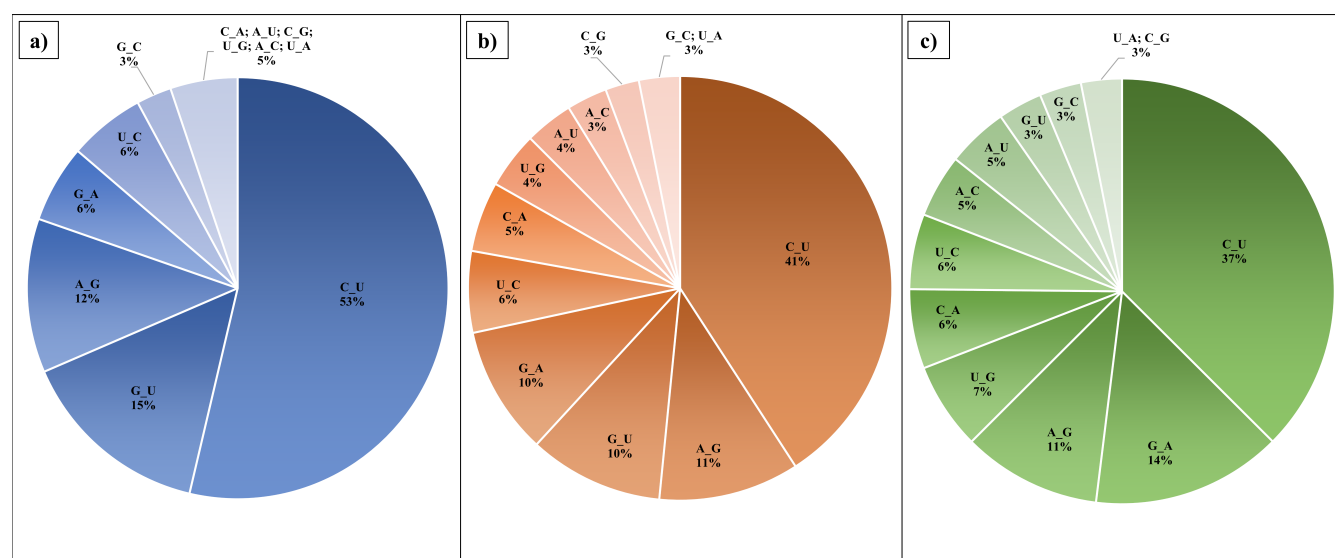


FIG 2 Percentage proportions of transition and transversion mutations in the countries in the three phases. (a) Before vaccination (pre-vaccination period); (b) after vaccination (post-vaccination period), and (c) recent period.

TABLE 6 Data comparison between genomic and spike protein among different SARS-CoV-2 variant sequences^a

Variants	No. of sequences	No. of genomic mutations	Genomic mutation rate (%)	Spike protein mutation rate (%)	Genomic dN/dS ratio	Spike dN/dS ratio	Ti/Tv ratio
Alpha	1,056	33,299	0.12	0.03	0.73	14.60	<i>1.68</i>
Beta	1,005	31,174	<i>0.10</i>	0.03	0.92	4.46	2.07
Delta	997	40,231	0.14	0.03	1.43	5.62	2.00
Gamma	1,003	37,597	0.13	0.04	0.60	10.75	1.81
Lambda	1,002	33,305	0.12	0.03	<i>0.58</i>	<i>1.42</i>	2.80
Mu	1,011	40,142	0.13	0.03	0.77	6.32	2.15
Omicron	1,135	71,411	0.22	0.10	1.01	4.46	2.00
Overall Average			0.14	0.04	0.86	6.8	2.07

^aThe highest mutation rates, dN/dS and Ti/Tv ratios are shown in bold while the lowest are shown as italics.

related to increase in dN. Thus, a decrease in Ti/Tv ratio is correlated with an increase in the dN/dS ratio. Previous studies have shown that *Ti* saturate more rapidly with time in comparison with *Tv*, thus causing Ti/Tv ratios to decline with evolutionary time (86–89).

- Effect of *Ti* and *Tv* patterns on CpG ratios: The percentage of *Ti* and *Tv* in the pre-vaccination, post-vaccination and recent period are shown in Fig. 2a through c respectively. The most frequent *Ti* and *Tv* observed was C→U and G→U respectively in the pre- and post-vaccination period. The C→U and G→U substitutions result in the deterioration of CpG content. In the post-vaccination period, the C→U substitutions decreased by 12% from the pre-vaccination period, whereas in recent period, a slight decrease of 4% was observed from the post-vaccination period. In comparison with C→U, G→U substitutions deteriorated at a slower rate, i.e., 5% reduction post-vaccination and about 7% reduction in recent period from post-vaccination period. In the recent period sequences, C→U remained the most frequent *Ti*, whereas U→G was the most frequent *Tv*. A previous study that evaluated the sequences of the period January 2020 to March 2021 reported a drastic increase in C→U and G→U substitutions in the initial phase of SARS-CoV-2 infection, which eventually stabilized at a point of time and then decreased till March 2021 (90). Table 1 shows the %GC content in the genome sequences from each country for pre-vaccination, post vaccination and recent period. It can be observed that the %GC content slightly reduced after vaccination, whereas negligible reduction was seen in the recent period genome sequences. As a result, the overall decline in GC content is minimal and appears to be stabilizing in future. This is confirmed by another study which depicted minute variations in the CpG content in the SARS-CoV-2 genome sequences and concluded that CpG content reduced at a faster rate in the initial time of evolution which may slow down to become steady with the evolution in human host (91). Thus, these mutations were attributed to rapid evolution after transmission to the human host (92–94). The differences in the mean % GC content were found to be highly significant as tested a two-tailed *P*-value of less than 0.0001 (Table 2). It has also been reported that the CpG motifs are lost in the SARS-CoV-2 sequences, which may enable antiviral response escape through TLR7 (91, 95). Higher U→G has been reported in the SARS-CoV-2 genome sequences collected between January 2022 and July 2022 (96). This further suggested that the amount of C→U and G→U substitutions are likely to decrease upon increase in divergence indicating that some portion

TABLE 7 Highest (top) to lowest (bottom) contribution of the proteins toward the mutation rates (%) in SARS-CoV-2 variants

Alpha	Beta	Delta	Gamma	Lambda	Mu	Omicron
S	S	S	S	S	S	S
NSP3	NSP3	NSP3	NSP3	N	NSP3	NSP3
NSP12b	ORF3a	N	N	NSP3	ORF8	NSP4
ORF8	NSP2	NSP4	NSP12b	NSP4	ORF3a	N
N	5'UTR	NSP12b	3'UTR	NSP12b	3'UTR	M
NSP2	N	ORF7a	ORF3a	NSP6	NSP14	NSP5
5'UTR	NSP12b	5'UTR	NSP13	ORF8	NSP13	NSP12b
NSP6	NSP5	3'UTR	NSP6	3'UTR	N	ORF3a
3'UTR	NSP6	NSP6	NSP1	NSP5	NSP12b	ORF6
NSP13	ORF8	ORF3a	ORF8	5'UTR	NSP4	NSP15
ORF3a	E	NSP13	5'UTR	M	NSP15	NSP6
NSP4	NSP13	NSP14	NSP9	ORF3a	NSP6	3'UTR
NSP15	NSP4	M	NSP2	NSP2	5'UTR	5'UTR
NSP14	NSP14	ORF8	NSP14	NSP14	NSP10	NSP14
NSP8	M	NSP2	NSP4	NSP13	NSP2	ORF7b
NSP5	3'UTR	ORF7b	NSP15	ORF7a	NSP16	E
NSP9	NSP15	NSP9	NSP5	NSP16	ORF7a	NSP13
M	NSP16	NSP15	NSP10	NSP15	M	NSP1
ORF7a	ORF7b	NSP7	NSP16	NSP1	NSP1	NSP9
NSP16	ORF7a	NSP5	M	NSP8	NSP8	NSP10
NSP1	NSP1	NSP1	ORF7a	ORF6	NSP5	NSP2
NSP10	NSP8	NSP16	NSP8	NSP9	NSP9	NSP16
ORF10	NSP9	ORF6	E	NSP7	ORF6	ORF8
E	NSP7	NSP8	ORF7b	NSP10	NSP7	ORF7a
NSP7	NSP10	E	NSP7	ORF10	ORF7b	ORF10

of these mutations are knocked out by purifying selection, irrespective of the origin of SARS-CoV-2 (90, 97). Reduction in C→U and G→U substitutions together with the increase in U→G may eventually lead to an increase in the CpG content in the SARS-CoV-2 genome. Higher CpG content has been linked to the attenuation of the virus (98, 99).

Evolutionary patterns across different SARS-CoV-2 variants

The genomic mutation rates (%), dN/dS and Ti/Tv ratios were also determined for each SARS-CoV-2 variant (shown in Table 6).

- a. **Mutation rate (%) analysis:** The variant-wise genomic mutation rate analysis data suggest Omicron variant has the highest mutation rates among all variants (0.22%) followed by Delta (0.14%). Beta variant displayed the lowest mutation rate (0.10%). Previous studies showed Omicron to have high genomic mutation rate along with the accumulation of highest number of mutations (100–102).
 - It has been suggested that the S mutations are positively selected to generate new variants of SARS-CoV-2 that have improved overall fitness (66, 103). To evaluate this, the average number of mutations, mutation rates and dN/dS ratios were determined for the S region in each variant. The average number of S mutations were calculated as a percentage of total number of S mutations in the genome. The resulting S mutations percentage for Alpha, Beta, Delta, Gamma, Lambda, Mu, and Omicron were 29.4%, 30.5%, 22.7%, 33.7%, 25.3%, 26.3%, and 47.8% respectively. As seen in Table 6, the highest

mutation rates in the whole genome as well as in *S* are seen in the Omicron variant. *S* mutation rates of all the variants are similar except for Omicron, which has the highest *S* mutation rate of 0.1. This has also been observed in a previous study (100). The mutation rate patterns for proteins in each SARS-CoV-2 variant is depicted in Table 7.

- b. dN/dS ratio analysis: As dN/dS ratios exceed unity for each SARS-CoV-2 variant, positive selection is responsible for promoting dN in the sequences. From Table 6, the average genomic dN/dS of all SARS-CoV-2 variants is estimated to be around 0.86 compared to the Wuhan reference strain which is similar with another study that estimated it to be 0.8446 (104). Also, it was observed that Delta variant had the highest dN/dS ratio (1.43) showing positive selection, followed by Omicron variant (1.01) showing neutral selection. Although the number of mutations is found to be higher in Omicron variant, the ratio of dN/dS is higher in the Delta variant. Delta variant was found to be highly virulent in comparison to the Omicron variant, which evolved to be more transmissible and less virulent (105–107). Therefore, the genomic dN/dS seems to correlate well with increased virulence. It has been argued that as virulence causes hindrance of transmission between hosts, dN/dS ratios may work to decrease virulence—thereby increasing transmission between hosts (108, 109).
- Virulent genes have been observed to be highly subject to dN and, therefore, are under strong positive selection (110). Therefore, the protein-

TABLE 8 Highest (top) to lowest (bottom) contribution of the proteins toward the selection pressure (dN/dS ratios) in SARS-CoV-2 variants

Alpha	Beta	Delta	Gamma	Lambda	Mu	Omicron
ORF8	E	ORF7b	S	NSP5	ORF8	E
N	ORF3a	ORF7a	ORF8	ORF7b	ORF3a	NSP14
S	S	N	E	NSP12a	NSP6	NSP1
NSP13	N	S	ORF7b	S	S	N
ORF3a	NSP16	ORF6	ORF3a	ORF3a	NSP13	NSP6
ORF10	NSP5	M	NSP10	N	NSP4	S
NSP8	NSP12b	NSP12b	NSP13	NSP8	ORF7b	NSP13
E	ORF7a	NSP13	ORF7a	NSP3	NSP12b	NSP4
NSP10	ORF7b	ORF3a	N	NSP4	E	M
NSP4	NSP1	E	NSP6	ORF7a	NSP1	NSP16
ORF6	NSP4	NSP14	NSP4	ORF6	N	NSP8
M	NSP2	ORF10	ORF10	NSP9	ORF10	NSP3
NSP16	NSP13	ORF8	NSP8	E	NSP2	NSP15
NSP14	NSP6	NSP4	NSP15	NSP2	NSP3	ORF7a
NSP6	NSP3	NSP8	NSP16	NSP10	ORF7a	NSP12b
NSP3	ORF6	NSP3	NSP5	NSP14	NSP5	ORF3a
ORF7a	NSP15	NSP2	M	NSP6	NSP9	NSP5
NSP5	NSP14	NSP16	NSP2	NSP12b	ORF6	ORF8
NSP9	NSP9	NSP10	NSP12b	ORF10	NSP7	ORF6
ORF7b	NSP7	NSP6	ORF6	NSP1	NSP16	NSP2
NSP15	M	NSP1	NSP3	NSP15	NSP8	ORF10
NSP1	ORF8	NSP15	NSP7	NSP7	NSP15	NSP7
NSP7	NSP10	NSP7	NSP14	NSP13	M	NSP10
NSP12b	NSP8	NSP5	NSP1	NSP16	NSP14	ORF7b
NSP2	ORF10	NSP9	NSP9	ORF8	NSP10	NSP9
ORF8	E	NSP12a	NSP12a	M	NSP12a	NSP12a

wise contribution to dN/dS was also determined in each of the variants and shown in Table 8. The unique genes having highest dN/dS in the highly virulent Delta variant were ORF7b and ORF7a while the unique genes contributing most to dN/dS in the Omicron variant were E, NSP14, and NSP1. The accessory protein ORF7b has been reported to mediate the cellular apoptosis caused by Tumour Necrosis Factor α (111). The ORF7a protein initiates autophagy and helps in virus replication (112). Taken together, these proteins impair the host cell immune response and cellular function (113).

- The dN/dS ratio of *S* was greater than 1 for all the variants. This is in line with the previous findings which suggest that *S* protein coding gene had the dN/dS ratio greater than 1 in all the SARS-CoV-2 variants (114). Alpha variant *S* protein showed the highest dN/dS ratio whereas Lambda variant *S* protein had the lowest dN/dS ratio. However, dN/dS values of both genome and *S* are comparatively lower in Omicron variant. Thus, there seems to be no correlation between dN/dS ratios of *S* and genomic dN/dS ratios, virulence, or transmission. That Omicron and Delta variant have the highest and the lowest *S* mutations, respectively, was also remarked in a previous study (114). The contribution of the SARS-CoV-2 proteins toward the mutation rates and dN/dS ratios in each variant has been shown in Tables 7 and 8, respectively.
- c. Ti/Tv ratio analysis: Lambda variant was found to have the lowest dN/dS ratio but the highest Ti/Tv ratio. A negative correlation has previously been observed between Ti/Tv ratio and dN/dS ratio under positive selection in the comparative analysis of SARS-CoV, SARS-CoV-2 and MERS-CoV for all the nucleotide substitution models (24). The reason for this could be explained by the fact that Tv substitutions favour dN. Previous studies have also remarked that higher number of Ti in the Lambda variant would remark lesser dN (86–89).

Conclusion

To the best of our knowledge, this is the only study that has compared the mutation rates, dN/dS ratios and Ti/Tv ratios during pre-, post-vaccination, and the recent periods in different geographical areas after the emergence of COVID-19. While individual studies find support from literature, this remains a comprehensive study on genomic parameters of SARS-CoV-2 in different regions and time periods.

The mutation rates were observed to be increased from the before vaccination period to the recent period in each country. The dN/dS ratio has increased with time, signifying accumulation of non-synonymous mutations favoring dN. However, the Ti/Tv ratio depicted significant decrease over time and may be correlated with viral attenuation as suggested earlier. Based on these parameters, the unpaired *t*-test helped in confirming that differences amongst the three phases were extremely statistically significant (P -value < 0.0001). As *S* is the main target for the vaccine, it has been proposed that mutations in this protein are driven by natural selection (66). While the number of mutations was observed to be increased in *S* in the Omicron variant, most of them were dS mutations.

A higher dN/dS ratio is also seen in another structural protein *N*. A 2020 communication proposed *N* as a vaccine target in view of its low mutation rates (115). Indeed, the current diagnostic kits also target *N* recognition. This may need to be re-evaluated in the light of the high mutation rates as well as dN/dS ratios of *N* as observed in our study. Conversely, ORF10, NSP7, and NSP10 coding genes had the lowest mutation rates across all the considered demographic regions and could further be targeted and explored for SARS-CoV-2 diagnostics and therapeutics.

Amongst the SARS-CoV-2 variants, Delta and Omicron had the highest dN/dS ratio and mutation rates respectively. The Delta variant has dN and correlates with high virulence in comparison with the Omicron variant, which is less virulent but has higher transmissibility. It was also observed that the pattern of highest to lowest gene-wise dN/dS ratios is unique in each variant genome.

The possible reasons for the observed increase in genomic mutation rates and selection pressure along with implications for vaccine efficacy against variants were considered. It was noted that:

- a. Vaccination is expected to slow down the number of infections, thus limiting the number of variants that the virus can explore (116). High rates of mutation may facilitate viral immune evasion and reduce the efficacy of the vaccine against infection and transmission, finally leading to reduced protection of the vaccine against several disease outcomes (117–119). The SARS-CoV-2 vaccines have indisputably lowered the COVID-19 disease burden in terms of infection, severity, and mortality (120–122). Therefore, it is likely that similar or higher mutation rates of the SARS-CoV-2 genome might have been recorded in the absence of a vaccine.
- b. The premise that increases in genomic mutation rates and selection pressures causes increased virus fitness and immune escape variants has also been debated as most of the mutations are deleterious (123, 124). It has been remarked that high mutation rates may even hinder the emergence of new highly adapted viruses as they do not allow viruses with advantageous genotypes to linger long enough to become fixed in a viral population (125, 126). Other causes for high mutation rates in RNA viruses have been explored, such as selection for more robust viral population or faster replication time (127).
- c. Differences in the evolutionary parameters like mutation rates and selection pressures in the SARS-CoV-2 genome before and after vaccination as observed in this study can be attributed to a range of factors associated with the virus, host immunity, or the environment. Furthermore, host immunity arising from natural causes as well as vaccination is expected to affect the evolution. The generation of new mutations in a viral population is not favored when the virus is well adapted to its surroundings since most mutations become harmful due to purifying selection. So, mutations that become more common can either be neutral and fixed by genetic drift or beneficial and fixed by positive selection (128).
- d. Pathogen fitness is characterized by an increased rate of spread in a population and has been described as a combination of increased infectivity, increased transmissibility, and a long infectious period. The public availability of genome sequences of SARS-CoV-2 variants provides a unique window to examine the changes in genomic parameters vis-à-vis the pathogen fitness. Pathogens can adapt to immunity by many mechanisms and such adaptation depends both on the new variants available in the environment at that time as well as their fitness in the host type. The over-representation of a variant in primed hosts indicates that it is immunity-adapted (129). modelled the evolution of the pathogen during vaccination campaigns by dividing it into two phases:
 - There is an initial phase during which most of the population is immunologically naïve. The first short phase favors the selection of generalist immunity-adapted variants that are equally effective against naïve and primed individuals.

- During the later second phase, most of the population acquires immunity either by natural infection or vaccination. This second phase favors specialist variants immunity-adapted to primed hosts.

The alpha, delta, and probably omicron can be classified as generalists based on epidemiological studies. It is predicted they would have spread regardless of vaccination. In future, specialist variants are predicted to appear, whereas immune-facilitated variants are rare (129–131). However, the duration required by the virus to acquire the number of mutations required for adaptation to the natural or vaccine acquired immunity cannot be predicted *a priori*.

- e. Vaccine-driven evolution is noted to occur in pathogens where its effects do not suppress infection, replication, or transmission sufficiently (132). Recent evidence for SARS-CoV-2 vaccines shows that the vaccine is not 100% effective against infection and disease severity. Furthermore, this effect has waned with time and emergence of new variants like omicron (130). While an array of diverse sub-lineages of the omicron variant have now emerged with the BQ and XBB sub-variants showing increased evasion of neutralizing antibodies (133), COVID-19 vaccines also elicit a specific T-cell response that may have an additional role in host protection (134). Parallels have also been made between these and the human influenza vaccines that are seasonally updated due to ongoing antigenic drift (129, 134). Most of the novel variants, in this case, remain partially inhibited by vaccination (135).
- f. While the currently approved SARS-CoV-2 vaccines continue to provide significant protection against severe disease and death (136), continuous monitoring and tracking for emergence of new fitter SARS-CoV-2 variants combined with molecular epidemiological surveillance is required.

ACKNOWLEDGMENTS

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. The authors would like to acknowledge Ms. Esther John in preparing the biopython code for calculation of expected *N* and *S* sites in the reference genome (Wuhan-Hu-1).

This research was funded by ICMR to SB under the project BMI/12(44)/2021.

S.G. and D.G. did the data collection and analysis. S.G. prepared the results and wrote the manuscript. S.B. conceived the idea and supervised the work.

AUTHOR AFFILIATIONS

¹Department of Biological Sciences and Engineering, Computational and Structural Biology Laboratory, Netaji Subhas University of Technology, Dwarka, New Delhi, India

²Division of Biotechnology, Computational and Structural Biology Laboratory, Netaji Subhas Institute of Technology, Dwarka, New Delhi, India

AUTHOR ORCID_s

Shubhangi Gupta  <http://orcid.org/0000-0001-6349-806X>

Sonika Bhatnagar  <http://orcid.org/0000-0002-8818-4240>

FUNDING

Funder	Grant(s)	Author(s)
Indian Council of Medical Research (ICMR)	BMI/12(44)/2021	Sonika Bhatnagar

AUTHOR CONTRIBUTIONS

Shubhangi Gupta, Data curation, Formal analysis, Validation, Writing – original draft
| Deepanshu Gupta, Data curation, Methodology, Software, Validation, Visualization |
Sonika Bhatnagar, Conceptualization, Supervision, Writing – review and editing

DATA AVAILABILITY

Raw data were generated from GISAID. Derived data supporting the findings of this study are available from the corresponding author S.B. on request. The findings of this study for pre-vaccination period are based on metadata associated with 34,970 sequences available on GISAID up to 31 December 2020 and accessible at doi:[10.55876/gis8.230517nt](https://doi.org/10.55876/gis8.230517nt). The findings of this study for post-vaccination period are based on metadata associated with 36,020 sequences available on GISAID from 15 May 2021 to 15 June 2022 and accessible at doi:[10.55876/gis8.230517va](https://doi.org/10.55876/gis8.230517va). The findings of this study for recent period are based on metadata associated with 3,880 sequences available on GISAID from 15 June 2022 to 05 December 2022 and accessible at doi:[10.55876/gis8.230518oq](https://doi.org/10.55876/gis8.230518oq). The findings of this study for pre-vaccination period are based on metadata associated with 7,209 sequences available on GISAID from 30 December 2019 to 09 January 2022 and accessible at doi:[10.55876/gis8.230518xv](https://doi.org/10.55876/gis8.230518xv).

REFERENCES

- World Health Organization. 2022. COVID-19 weekly epidemiological update.
- Thayer WM, Hasan MZ, Sankhla P, Gupta S. 2021. An interrupted time series analysis of the lockdown policies in India: a national-level analysis of COVID-19 incidence. *Health Policy Plan* 36:620–629. <https://doi.org/10.1093/heapol/czab027>
- Sapkal G, Kant R, Dwivedi G, Sahay RR, Yadav PD, Deshpande GR, Singh R, Nyayanit DA, Patil DY, Shete-Aich AM, Zaman K, Chaudhari AK, Gupta N, Panda S, Abraham P, Bhargava B. 2022. Immune responses against different variants of SARS-CoV-2 including omicron following 6 months of administration of heterologous prime-boost COVID-19 vaccine. *J Travel Med* 29:taac033. <https://doi.org/10.1093/jtm/taac033>
- Gaurav S, Pandey S, Puvar A, Shah T, Joshi M, Joshi C, Kumar S. 2020. Identification of unique mutations in SARS-CoV-2 strains isolated from India suggests its attenuated pathotype. *bioRxiv*. <https://doi.org/10.1101/2020.06.06.137604>
- Ortega MA, García-Montero C, Fraile-Martínez O, Colet P, Baizhaxynova A, Mukhtarova K, Alvarez-Mon M, Kanatova K, Asúnsolo A, Sarria-Santamera A. 2022. Recapping the features of SARS-CoV-2 and its main variants: status and future paths. *J Pers Med* 12:995. <https://doi.org/10.3390/jpm12060995>
- Bansal K, Kumar S. 2022. Mutational cascade of SARS-CoV-2 leading to evolution and emergence of omicron variant. *Virus Res* 315:198765. <https://doi.org/10.1016/j.virusres.2022.198765>
- Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, Zhang Q, Shi X, Wang Q, Zhang L, Wang X. 2020. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* 581:215–220. <https://doi.org/10.1038/s41586-020-2180-5>
- Laurini E, Marson D, Aulic S, Fermeglia M, Pricl S. 2020. Computational alanine scanning and structural analysis of the SARS-CoV-2 spike protein/angiotensin-converting enzyme 2 complex. *ACS Nano* 14:11821–11830. <https://doi.org/10.1021/acsnano.0c04674>
- Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, Geng Q, Auerbach A, Li F. 2020. Structural basis of receptor recognition by SARS-CoV-2. *Nature* 581:221–224. <https://doi.org/10.1038/s41586-020-2179-y>
- Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. 2020. Structural basis for the recognition of the SARS-CoV-2 by full-length human Ace2. *Science* 367:1444–1448. <https://doi.org/10.1126/science.abb2762>
- Gupta D, Sharma P, Singh M, Kumar M, Ethayathulla AS, Kaur P. 2021. Structural and functional insights into the spike protein mutations of emerging SARS-CoV-2 variants. *Cell Mol Life Sci* 78:7967–7989. <https://doi.org/10.1007/s00018-021-04008-0>
- Triveri A, Serapian SA, Marchetti F, Doria F, Pavoni S, Cinquini F, Moroni E, Rasola A, Frigerio F, Colombo G. 2021. SARS-CoV-2 spike protein mutations and escape from antibodies: a computational model of epitope loss in variants of concern. *J Chem Inf Model* 61:4687–4700. <https://doi.org/10.1021/acs.jcim.1c00857>
- Beeraka NM, Sukocheva OA, Lukina E, Liu J, Fan R. 2022. Development of antibody resistance in emerging mutant strains of SARS CoV-2: impediment for COVID-19 vaccines. *Rev Med Virol* 32:e2346. <https://doi.org/10.1002/rmv.2346>
- De Maio N, Walker CR, Turakhia Y, Lanfear R, Corbett-Detig R, Goldman N. 2021. Mutation rates and selection on synonymous mutations in SARS-CoV-2. *Genome Biol Evol* 13:1–14. <https://doi.org/10.1093/gbe/evab087>
- Wang R, Chen J, Hozumi Y, Yin C, Wei G-W. 2022. Emerging vaccine-breakthrough SARS-CoV-2 variants. *ACS Infect Dis* 8:546–556. <https://doi.org/10.1021/acsinfecdis.1c00557>
- Starr TN, Greaney AJ, Addetia A, Hannon WW, Choudhary MC, Diggins AS, Li JZ, Bloom JD. 2021. Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science* 371:850–854. <https://doi.org/10.1126/science.abb9302>
- Gupta AM, Chakrabarti J, Mandal S. 2020. Non-synonymous mutations of SARS-Cov-2 leads EPITOPE loss and segregates its variants. *Microbes Infect* 22:598–607. <https://doi.org/10.1016/j.micinf.2020.10.004>
- Isabel S, Graña-Miraglia L, Gutierrez JM, Bundalovic-Torma C, Groves HE, Isabel MR, Eshaghi A, Patel SN, Gubbay JB, Poutanen T, Guttman DS, Poutanen SM. 2020. Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein mutation now documented worldwide. *Sci Rep* 10:14031. <https://doi.org/10.1038/s41598-020-70827-z>
- Cherian S, Potdar V, Jadhav S, Yadav P, Gupta N, Das M, Rakshit P, Singh S, Abraham P, Panda S, et al. 2021. Convergent evolution of SARS-CoV-2 spike mutations, L452R, E484Q and P681R, in the second wave of COVID-19 in Maharashtra, India. *bioRxiv*. <https://doi.org/10.1101/2021.04.22.440932>

20. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, et al. 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>
21. Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, O'Meara MJ, Rezelj VV, Guo JZ, Swaney DL, 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>
22. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, Ludden C, Reeve R, Rambaut A, COVID-19 Genomics UK (COG-UK) Consortium, Peacock SJ, Robertson DL. 2021. SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* 19:409–424. <https://doi.org/10.1038/s41579-021-00573-0>
23. Zhan X-Y, Zhang Y, Zhou X, Huang K, Qian Y, Leng Y, Yan L, Huang B, He Y. 2020. Molecular evolution of SARS-CoV-2 structural genes: evidence of positive selection in spike glycoprotein. *bioRxiv*. <https://doi.org/10.1101/2020.06.25.170688>
24. Sohpal VK. 2021. Comparative study: nonsynonymous and synonymous substitution of SARS-CoV-2, SARS-CoV, and MERS-CoV genome. *Genomics Inform* 19:e15. <https://doi.org/10.5808/gi.20058>
25. Wang R, Hozumi Y, Yin C, Wei G-W. 2020. Genomics mutations on COVID-19 diagnostic targets. *Genomics* 112:5204–5213. <https://doi.org/10.1016/j.ygeno.2020.09.028>
26. Rahman MS, Islam MR, Alam A, Islam I, Hoque MN, Akter S, Rahaman MM, Sultana M, Hossain MA. 2021. Evolutionary dynamics of SARS-CoV-2 nucleocapsid protein and its consequences. *J Med Virol* 93:2177–2195. <https://doi.org/10.1002/jmv.26626>
27. Das JK, Sengupta A, Choudhury PP, Roy S. 2021. Characterizing genomic variants and mutations in SARS-CoV-2 proteins from Indian isolates. *Gene Rep* 25:101044. <https://doi.org/10.1016/j.genrep.2021.101044>
28. Sarkar R, Mitra S, Chandra P, Saha P, Banerjee A, Dutta S, Chawla-Sarkar M. 2021. Comprehensive analysis of genomic diversity of SARS-CoV-2 in different geographic regions of India: an endeavour to classify Indian SARS-CoV-2 strains on the basis of co-existing mutations. *Arch Virol* 166:801–812. <https://doi.org/10.1007/s00705-020-04911-0>
29. Sridhar A, Yagan O, Eletreby R, Levin SA, Plotkin JB, Poor HV. 2021. “Leveraging a multiple-strain model with mutations in analyzing the spread of COVID-19” ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Toronto, ON, Canada, p 8163–8167. <https://doi.org/10.1109/ICASSP39728.2021.9414595>
30. Elbe S, Buckland-Merrett G. 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* 1:33–46. <https://doi.org/10.1002/gch2.1018>
31. Khare S, Gurry C, Freitas L, Schultz MB, Bach G, Diallo A, Akite N, Ho J, Lee RT, Yeo W, Curation Team GC, Maurer-Stroh S. 2021. GISAID's role in pandemic response. *China CDC Wkly* 3:1049–1051. <https://doi.org/10.46234/ccdcw2021.255>
32. Mercatelli D, Triboli L, Fornasari E, Ray F, Giorgi FM. 2021. Coronapp: a web application to annotate and monitor SARS-CoV-2 mutations. *J Med Virol* 93:3238–3245. <https://doi.org/10.1002/jmv.26678>
33. Randles BM, Pasquetto IV, Golshan MS, Borgman CL. 2017. “Using the jupyter notebook as a tool for open science: an empirical study” 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL); Toronto, ON, Canada. <https://doi.org/10.1109/JCDL.2017.7991618>
34. McKinneyW. 2011. Pandas: a foundational python library for data analysis and statistics 14:1–9.
35. Pathan RK, Biswas M, Khandaker MU. 2020. Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model. *Chaos Solitons Fractals* 138:110018. <https://doi.org/10.1016/j.chaos.2020.110018>
36. Heddle JA, Athanasiou K. 1975. Mutation rate, genome size and their relation to the rec concept. *Nature* 258:359–361. <https://doi.org/10.1038/258359a0>
37. Lv L, Li G, Chen J, Liang X, Li Y. 2020. Comparative genomic analyses reveal a specific mutation pattern between human coronavirus SARS-CoV-2 and bat-CoV RaTG13. *Front Microbiol* 11:584717. <https://doi.org/10.3389/fmicb.2020.584717>
38. Kryazhinskiy S, Plotkin JB. 2008. The population genetics of dN/dS. *PLoS Genet* 4:e1000304. <https://doi.org/10.1371/journal.pgen.1000304>
39. Gojovori TJG. 1983. Codon substitution in evolution and the “saturation” of synonymous changes. *Genetics* 105:1011–1027. <https://doi.org/10.1093/genetics/105.4.1011>
40. Peck KM, Luring AS, Sullivan CS. 2018. Complexities of viral mutation rates. *J Virol* 92:e01031–17. <https://doi.org/10.1128/JVI.01031-17>
41. Kamp C, Wilke CO, Adami C, Bornholdt SJC. 2002. Viral evolution under the pressure of an adaptive immune system: optimal mutation rates for viral escape. *Complexity* 8:28–33. <https://doi.org/10.1002/cplx.10067>
42. Robson F, Khan KS, Le TK, Paris C, Demirbag S, Barfuss P, Rocchi P, Ng W-L. 2020. Coronavirus RNA proofreading: molecular basis and therapeutic targeting. *Mol Cell* 80:1136–1138. <https://doi.org/10.1016/j.molcel.2020.11.048>
43. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CCS, Boshier FAT, Ortiz AT, Balloux F. 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol* 83:104351. <https://doi.org/10.1016/j.meegid.2020.104351>
44. Zhao Z, Li H, Wu X, Zhong Y, Zhang K, Zhang Y-P, Boerwinkle E, Fu Y-X. 2004. Moderate mutation rate in the SARS coronavirus genome and its implications. *BMC Evol Biol* 4:1–9. <https://doi.org/10.1186/1471-2148-4-21>
45. Gupta V, Haider S, Verma M, Singhvi N, Ponnusamy K, Malik MZ, Verma H, Kumar R, Sood U, Hira P, Satija S, Singh Y, Lal R. 2021. Comparative genomics and integrated network approach unveiled undirected phylogeny patterns, co-mutational hot spots, functional cross talk, and regulatory interactions in SARS-CoV-2. *mSystems* 6:e00030-21. <https://doi.org/10.1128/mSystems.00030-21>
46. Behboudi E, Tabarraei A, Tahamtan A, Kalani MR, AJJoP M. 2022. Assessment and molecular docking of SARS-CoV-2 NSP3 and NSP12 mutants in Iranian patients in Golestan province 10:16370–16380.
47. Kaushal N, Gupta Y, Goyal M, Khaiboullina SF, Baranwal M, Verma S. 2020. Mutational frequencies of SARS-CoV-2 genome during the beginning months of the outbreak in USA. *Pathogens* 9:565. <https://doi.org/10.3390/pathogens9070565>
48. Fibriani A, Stephanie R, Alfiantie AA, Siregar ALF, Pradani GAP, Yamahoki N, Purba WS, Alamanda CNC, Rahmawati E, Rachman RW, Robiani R, Ristandi RB. 2021. Analysis of SARS-CoV-2 genomes from west java. *Viruses* 13:2097. <https://doi.org/10.3390/v13102097>
49. Abdullahi IN, Emeribe AU, Ajayi OA, Oderinde BS, Amadu DO, Osuji AI. 2020. Implications of SARS-CoV-2 genetic diversity and mutations on pathogenicity of the COVID-19 and biomedical interventions. *J Taibah Univ Med Sci* 15:258–264. <https://doi.org/10.1016/j.jtumed.2020.06.005>
50. Salleh MZ, Derrick JP, Deris ZZ. 2021. Structural evaluation of the spike glycoprotein variants on SARS-CoV-2 transmission and immune evasion. *Int J Mol Sci* 22:7425. <https://doi.org/10.3390/ijms22147425>
51. Mengist HM, Kombe Kombe AJ, Mekonnen D, Abebaw A, Getachew M, Jin T. 2021. Mutations of SARS-CoV-2 spike protein: implications on immune evasion and vaccine-induced immunity. *Semin Immunol* 55:101533. <https://doi.org/10.1016/j.smim.2021.101533>
52. Thakur S, Sasi S, Pillai SG, Nag A, Shukla D, Singhal R, Phalke S, Velu GSK. 2022. SARS-CoV-2 mutations and their impact on diagnostics, therapeutics and vaccines. *Front Med (Lausanne)* 9:815389. <https://doi.org/10.3389/fmed.2022.815389>
53. Ye Q, West AMV, Silletti S, Corbett KD. 2020. Architecture and self-assembly of the SARS-CoV-2 nucleocapsid protein. *Protein Sci* 29:1890–1901. <https://doi.org/10.1002/pro.3909>
54. Johnson BA, Zhou Y, Lokugamage KG, Vu MN, Bopp N, Crocquet-Valdes PA, Kalveram B, Schindewolf C, Liu Y, Scharf D, Plante JA, Xie X, Aguilar P, Weaver SC, Shi P-Y, Walker DH, Routh AL, Plante KS, Menachery VD, Shih S-R. 2022. Nucleocapsid mutations in SARS-CoV-2 augment replication and pathogenesis. *PLoS Pathog* 18:e1010627. <https://doi.org/10.1371/journal.ppat.1010627>
55. Hajnik RL, Plante JA, Liang Y, Alameh M-G, Tang J, Bonam SR, Zhong C, Adam A, Scharf D, Rafael GH, Liu Y, Hazell NC, Sun J, Soong L, Shi P-Y, Wang T, Walker DH, Sun J, Weissman D, Weaver SC, Plante KS, Hu H. 2022. Dual spike and nucleocapsid mRNA vaccination confer protection against SARS-CoV-2 omicron and delta variants in preclinical models. *Sci Transl Med* 14:eabq1945. <https://doi.org/10.1126/scitranslmed.abq1945>

56. Fooladinezhad H, Shahidi M, Mahmanzar M, Mahdavi B, Tokhanbigli S, Sisakht MM, Moradi B, Ganjalikhany MR, Rahimian K, Ganjalikhani-Hakemi M. 2022. SARS-CoV-2 NSP3, NSP4 and NSP6 mutations and epistasis during the pandemic in the world: evolutionary trends and natural selections in six continents. *Epidemiology*. <https://doi.org/10.1101/2022.05.22.22275422>
57. Sakai Y, Kawachi K, Terada Y, Omori H, Matsuura Y, Kamitani WJV. 2017. Two-amino acids change in the NSP4 of SARS coronavirus abolishes viral replication. *Virology* 510:165–174. <https://doi.org/10.1016/j.virol.2017.07.019>
58. Wang H, Rizvi SRA, Dong D, Lou J, Wang Q, Sopipong W, Su Y, Najar F, Agarwal PK, Kozielski F, Haider S. 2022. Emerging variants of SARS-CoV-2 NSP10 highlight strong functional conservation of its binding to two non-structural proteins, NSP14 and NSP16. *Biophysics*. <https://doi.org/10.1101/2022.12.23.521761>
59. Patro LPP, Sathyaseelan C, Uttamrao PP, Rathinavelan TJB. 2020. Global variation in the SARS-CoV-2 proteome reveals the mutational hotspots in the drug and vaccine candidates. *bioRxiv*. <https://doi.org/10.1101/2020.07.31.230987>
60. Omotoso OE, Babalola AD, Matareek A. 2021. Mutational hotspots and conserved domains of SARS-CoV-2 genome in African population. *Beni Suef Univ J Basic Appl Sci* 10:11. <https://doi.org/10.1186/s43088-021-00102-1>
61. Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, Trebes A, Andersson M, Otecko N, Wise EL, Moore N, et al. 2021. SARS-CoV-2 within-host diversity and transmission. *Science* 372:eabg0821. <https://doi.org/10.1126/science.abg0821>
62. Shishir TA, Jannat T, Naser IB. 2022. Genomic surveillance unfolds the SARS-CoV-2 transmission and divergence dynamics in Bangladesh. *Front Genet* 13:966939. <https://doi.org/10.3389/fgene.2022.966939>
63. Li X, Giorgi EE, Marichannegowda MH, Foley B, Xiao C, Kong X-P, Chen Y, Gnanakaran S, Korber B, Gao F. 2020. Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci Adv* 6:eabb9153. <https://doi.org/10.1126/sciadv.abb9153>
64. Hou Y, Zhao S, Liu Q, Zhang X, Sha T, Su Y, Zhao W, Bao Y, Xue Y, Chen HJG, Proteomics B. 2022. Ongoing positive selection drives the evolution of SARS-CoV-2 genomes. *GPB* 20:1214–1223. <https://doi.org/10.1016/j.gpb.2022.05.009>
65. Ochoa G, Harvey I, Buxton H. 2000 Optimal mutation rates and selection pressure in genetic algorithms, p 315–322
66. López-Cortés GI, Palacios-Pérez M, Velez-HF, Hernández-Aguilar M, López-Hernández GR, Zamudio GS, José MV. 2022. The spike protein of SARS-CoV-2 is adapting because of selective pressures. *Vaccines (Basel)* 10:864. <https://doi.org/10.3390/vaccines10060864>
67. Schuster NA. 2020. Characterization and structural prediction of the putative ORF10 protein in SARS-CoV-2. *bioRxiv*. <https://doi.org/10.1101/2020.10.26.355784>
68. Velazquez-Salinas L, Zarate S, Eberl S, Gladue DP, Novella I, Borca MV. 2020. Positive selection of ORF1ab, ORF3a, and ORF8 genes drives the early evolutionary trends of SARS-CoV-2 during the 2020 COVID-19 pandemic. *Front Microbiol* 11:550674. <https://doi.org/10.3389/fmicb.2020.550674>
69. Vinjamuri S, Li L, Bouvier M. 2022. SARS-CoV-2 ORF8: one protein, seemingly one structure, and many functions. *Front Immunol* 13:1035559. <https://doi.org/10.3389/fimmu.2022.1035559>
70. Zhang J, Ejikemuwa A, Gerzanich V, Nasr M, Tang Q, Simard JM, Zhao RY. 2022. Understanding the role of SARS-CoV-2 ORF3A in viral pathogenesis and COVID-19. *Front Microbiol* 13:854567. <https://doi.org/10.3389/fmicb.2022.854567>
71. Li X, Hou P, Ma W, Wang X, Wang H, Yu Z, Chang H, Wang T, Jin S, Wang X, Wang W, Zhao Y, Zhao Y, Xu C, Ma X, Gao Y, He H. 2022. SARS-Cov-2 Orf10 suppresses the antiviral innate immune response by degrading MAVS through Mitophagy. *Cell Mol Immunol* 19:67–78. <https://doi.org/10.1038/s41423-021-00807-4>
72. Su C-M, Wang L, Yoo D. 2021. Activation of NF-KB and induction of proinflammatory cytokine expressions mediated by ORF7a protein of SARS-CoV-2. *Sci Rep* 11:13464. <https://doi.org/10.1038/s41598-021-92941-2>
73. Emam M, Oweda M, Antunes A, El-Hadidi M. 2021. Positive selection as a key player for SARS-CoV-2 Pathogenicity: insights into ORF1ab, S and E genes. *Virus Res* 302:198472. <https://doi.org/10.1016/j.virusres.2021.198472>
74. Li P, Xue B, Schnicker NJ, Wong L-Y, Meyerholz DK, Perlman S. 2023. Nsp3-N interactions are critical for SARS-CoV-2 fitness and virulence. *Proc Natl Acad Sci U S A* 120:e2305674120. <https://doi.org/10.1073/pnas.2305674120>
75. Lo Presti A, Rezza G, Stefanelli P. 2020. Selective pressure on SARS-CoV-2 protein coding genes and glycosylation site prediction. *Heliyon* 6:e05001. <https://doi.org/10.1016/j.heliyon.2020.e05001>
76. Schubert K, Karousis ED, Jomaa A, Scaiola A, Echeverria B, Gurseler L-A, Leibundgut M, Thiel V, Mühlemann O, Ban N. 2020. SARS-CoV-2 Nsp1 binds the Ribosomal mRNA channel to inhibit translation. *Nat Struct Mol Biol* 27:1094. <https://doi.org/10.1038/s41594-020-00533-x>
77. Pervushin K, Tan E, Parthasarathy K, Lin X, Jiang FL, Yu D, Vararattana-vech A, Soong TW, Liu DX, Torres J. 2009. Structure and inhibition of the SARS coronavirus envelope protein ion channel. *PLoS Pathog* 5:e1000511. <https://doi.org/10.1371/journal.ppat.1000511>
78. Nieto-Torres JL, DeDiego ML, Verdiá-Báguena C, Jimenez-Guardeño JM, Regla-Nava JA, Fernandez-Delgado R, Castaño-Rodríguez C, Alcaraz A, Torres J, Aguilera VM, Enjuanes L, Denison MR. 2014. Severe acute respiratory syndrome coronavirus envelope protein ion channel activity promotes virus fitness and pathogenesis. *PLoS Pathog* 10:e1004077. <https://doi.org/10.1371/journal.ppat.1004077>
79. Guo G, Gao M, Gao X, Zhu B, Huang J, Luo K, Zhang Y, Sun J, Deng M, Lou Z. 2021. SARS-CoV-2 non-structural protein 13 (nsp13) hijacks host deubiquitinase USP13 and counteracts host antiviral immune response. *Signal Transduct Target Ther* 6:119. <https://doi.org/10.1038/s41392-021-00509-3>
80. Fang Z, Peng L, Lin Q, Zhou L, Yang L, Feng Y, Ren P, Renauer PA, Park JJ, Zhou X, Wilen CB, Chen S. 2022. Heterotypic vaccination responses against SARS-CoV-2 omicron BA.2. *Immunology*. <https://doi.org/10.1101/2022.03.22.485418>
81. Thoms M, Buschauer R, Ameisemeier M, Koepke L, Denk T, Hirschenberger M, Kratzat H, Hayn M, Mackens-Kiani T, Cheng J, Straub JH, Stürzel CM, Fröhlich T, Berninghausen O, Becker T, Kirchhoff F, Sparrer KMJ, Beckmann R. 2020. Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2. *Science* 369:1249–1255. <https://doi.org/10.1126/science.abc8665>
82. Vann KR, Tencer AH, Kutateladze TG. 2020. Inhibition of translation and immune responses by the virulence factor Nsp1 of SARS-CoV-2. *Signal Transduct Target Ther* 5:234. <https://doi.org/10.1038/s41392-020-00350-0>
83. Alteri C, Fox V, Scutari R, Burastero GJ, Volpi S, Faltoni M, Fini V, Granaglia A, Esperti S, Gallerani A, Costabile V, Fontana B, Franceschini E, Meschiari M, Campana A, Bernardi S, Villani A, Bernaschi P, Russo C, Gualdi G, Mussini C, Perno CF. 2022. A proof-of-concept study on the genomic evolution of SARS-CoV-2 in molnupiravir-treated, paxlovid-treated and drug-naïve patients. *Commun Biol* 5:1376. <https://doi.org/10.1038/s42003-022-04322-8>
84. Hakmaoui A, Khan F, Liacini A, Kaur A, Berka Y, Machraoui S, Soualhine H, Berka N, Rais H, Admou B. 2021. Relevant SARS-CoV-2 genome variation through six months of worldwide monitoring. *Biomed Res Int* 2021:5553173. <https://doi.org/10.1155/2021/5553173>
85. Rosenberg MS, Subramanian S, Kumar S. 2003. Patterns of transitional mutation biases within and among mammalian genomes. *Mol Biol Evol* 20:988–993. <https://doi.org/10.1093/molbev/msg113>
86. Duchêne S, Ho SYW, Holmes EC. 2015. Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. *BMC Evol Biol* 15:36. <https://doi.org/10.1186/s12862-015-0312-6>
87. Rosenberg MS, Subramanian S, Kumar S. 2003. Patterns of transitional mutation biases within and among mammalian Genomes. *Mol Biol Evol* 20:988–993. <https://doi.org/10.1093/molbev/msg113>
88. Purvis A, Bromham L. 1997. Estimating the transition/transversion ratio from independent pairwise comparisons with an assumed phylogeny. *J Mol Evol* 44:112–119. <https://doi.org/10.1007/pl00006117>
89. Wakeley J. 1996. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends Ecol Evol* 11:158–162. [https://doi.org/10.1016/0169-5347\(96\)10009-4](https://doi.org/10.1016/0169-5347(96)10009-4)

90. Forni D, Cagliani R, Pontremoli C, Clerici M, Sironi M. 2022. The substitution spectra of coronavirus genomes. *Brief Bioinform* 23:bbab382. <https://doi.org/10.1093/bib/bbab382>
91. Kumar A, Goyal N, Saranathan N, Dhamija S, Saraswat S, Menon MB, Vivekanandan P, Leitner T. 2022. The slowing rate of cpG depletion in SARS-CoV-2 genomes is consistent with adaptations to the human host. *Mol Biol Evol* 39:msac029. <https://doi.org/10.1093/molbev/msac029>
92. Matyášek R, Kovařík A. 2020. Mutation patterns of human SARS-CoV-2 and bat RaTG13 coronavirus genomes are strongly biased towards C>U transitions, indicating rapid evolution in their hosts. *Genes (Basel)* 11:761. <https://doi.org/10.3390/genes11070761>
93. Panchin AY, Panchin YV. 2020. Excessive G-U transversions in novel allele variants in SARS-CoV-2 genomes. *PeerJ* 8:e9648. <https://doi.org/10.7717/peerj.9648>
94. Roy C, Mandal SM, Mondal SK, Mukherjee S, Mapder T, Ghosh W, Chakraborty RJG. 2020. Trends of mutation accumulation across global SARS-CoV-2 genomes: implications for the evolution of the novel coronavirus. *Genomics* 112:5331–5342. <https://doi.org/10.1016/j.ygeno.2020.11.003>
95. Ravi V, Swaminathan A, Yadav S, Arya H, Pandey RJV. 2022. SARS-CoV-2 variants of concern and variations within their genome architecture: does nucleotide distribution and mutation rate alter the functionality and evolution of the virus? *Viruses* 14:2499. <https://doi.org/10.3390/v14112499>
96. Gurjar P, Karuvantevida N, Rzhepakovsky IV, Khan AA, Khandia R. 2023. A synthetic biology approach for vaccine candidate design against Delta strain of SARS-CoV-2 revealed disruption of favored codon pair as a better strategy over using rare codons. *Vaccines (Basel)* 11:487. <https://doi.org/10.3390/vaccines11020487>
97. Pybus OG, Rambaut A, Belshaw R, Freckleton RP, Drummond AJ, Holmes EC. 2007. Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. *Mol Biol Evol* 24:845–852. <https://doi.org/10.1093/molbev/msm001>
98. Trus I, Udenze D, Berube N, Wheler C, Martel M-J, Gerds V, Karniychuk U. 2019. CpG-recoding in Zika virus genome causes host-age-dependent attenuation of infection with protection against lethal heterologous challenge in mice. *Front Immunol* 10:3077. <https://doi.org/10.3389/fimmu.2019.03077>
99. Burns CC, Campagnoli R, Shaw J, Vincent A, Jorba J, Kew O. 2009. Genetic inactivation of poliovirus infectivity by increasing the frequencies of CpG and UpA dinucleotides within and across synonymous capsid region codons. *J Virol* 83:9957–9969. <https://doi.org/10.1128/JVI.00508-09>
100. Kudriavtsev AV, Vakhrusheva AV, Novoseletsky VN, Bozdaganyan ME, Shaitan KV, Kirpichnikov MP, Sokolova OS. 2022. Immune escape associated with RBD omicron mutations and SARS-CoV-2 evolution dynamics. *Viruses* 14:1603. <https://doi.org/10.3390/v14081603>
101. Zhang X, Wu S, Wu B, Yang Q, Chen A, Li Y, Zhang Y, Pan T, Zhang H, HeX. 2021. SARS-CoV-2 Omicron strain exhibits potent capabilities for immune evasion and viral entrance. *Sig Transduct Target Ther* 6:430. <https://doi.org/10.1038/s41392-021-00852-5>
102. Kandeel M, Mohamed MEM, Abd El-Lateef HM, Venugopala KN, El-Beltagi HS. 2022. Omicron variant genome evolution and phylogenetics. *J Med Virol* 94:1627–1632. <https://doi.org/10.1002/jmv.27515>
103. Magazine N, Zhang T, Wu Y, McGee MC, Veggiani G, Huang W. 2022. Mutations and evolution of the SARS-CoV-2 spike protein. *Viruses* 14:640. <https://doi.org/10.3390/v14030640>
104. Bai H, Ata G, Sun Q, Rahman SU, Tao S. 2023. Natural selection pressure exerted on “silent” mutations during the evolution of SARS-CoV-2: evidence from codon usage and RNA structure. *Virus Res* 323:198966. <https://doi.org/10.1016/j.virusres.2022.198966>
105. Brüssow HJMB. 2022. COVID - 19: omicron—the latest, the least virulent, but probably not the last variant of concern of SARS - CoV - 2 15:1927–1939.
106. Martins M, do Nascimento GM, Nooruzzaman M, Yuan F, Chen C, Caserta LC, Miller AD, Whittaker GR, Fang Y, Diel DG. 2022. The omicron variant BA. 1.1 presents a lower pathogenicity than B. 1 D614G and delta variants in a feline model of SARS-CoV-2 infection. *J Virol* 96:e0096122. <https://doi.org/10.1128/jvi.00961-22>
107. Iuliano AD, Brunkard JM, Boehmer TK, Peterson E, Adjei S, Binder AM, Cobb S, Graff P, Hidalgo P, Panaggio MJ, Rainey JJ, Rao P, Soeteber K, Wacaster S, Ai C, Gupta V, Molinari N-AM, Ritchey MD. 2022. Trends in disease severity and health care utilization during the early omicron variant period compared with previous SARS-CoV-2 high transmission periods—United States. *MMWR Morb Mortal Wkly Rep* 71:146–152. <https://doi.org/10.15585/mmwr.mm7104e4>
108. Geoghegan JL, Holmes EC. 2018. The phylogenomics of evolving virus virulence. *Nat Rev Genet* 19:756–769. <https://doi.org/10.1038/s41576-018-0055-5>
109. Willemsen A, Zwart MP, Elena SF. 2017. High virulence does not necessarily impede viral adaptation to a new host: a case study using a plant RNA virus. *BMC Evol Biol* 17:25. <https://doi.org/10.1186/s12862-017-0881-7>
110. Merrikh CN, Merrikh H. 2018. Gene inversion potentiates bacterial evolvability and virulence. *Nat Commun* 9:4662. <https://doi.org/10.1038/s41467-018-07110-3>
111. Yang R, Zhao Q, Rao J, Zeng F, Yuan S, Ji M, Sun X, Li J, Yang J, Cui J, Jin Z, Liu L, Liu Z. 2021. SARS-CoV-2 accessory protein ORF7B mediates tumor necrosis factor- α -induced apoptosis in cells. *Front Microbiol* 12:654709. <https://doi.org/10.3389/fmicb.2021.654709>
112. Hou P, Wang X, Wang H, Wang T, Yu Z, Xu C, Zhao Y, Wang W, Zhao Y, Chu F, Chang H, Zhu H, Lu J, Zhang F, Liang X, Li X, Wang S, Gao Y, He H. 2023. The ORF7A protein of SARS-CoV-2 initiates autophagy and limits autophagosome-lysosome fusion via degradation of SNAP29 to promote virus replication. *Autophagy* 19:551–569. <https://doi.org/10.1080/15548627.2022.2084686>
113. García-García T, Fernández-Rodríguez R, Redondo N, de Lucas-Rius A, Zaldívar-López S, López-Ayllón BD, Suárez-Cárdenas JM, Jiménez-Marín Á, Montoya M, Garrido JJ. 2022. Impairment of antiviral immune response and disruption of cellular functions by SARS-CoV-2 ORF7A and ORF7B. *iScience* 25:105444. <https://doi.org/10.1016/j.isci.2022.105444>
114. Nikolaidis M, Papakyriakou A, Chlichlia K, Markoulatos P, Oliver SG, Amoutzias G. 2022. Comparative analysis of SARS-CoV-2 variants of concern, including omicron, highlights their common and distinctive amino acid substitution patterns, especially at the spike ORF. *Viruses* 14:707. <https://doi.org/10.3390/v14040707>
115. Dutta NK, Mazumdar K, Gordy JT, Dutch RE. 2020. The nucleocapsid protein of SARS-CoV-2: a target for vaccine development. *J Virol* 94:e00647-20. <https://doi.org/10.1128/JVI.00647-20>
116. Sah P, Vilches TN, Moghadas SM, Fitzpatrick MC, Singer BH, Hotez PJ, Galvani A. 2021. Accelerated vaccine Rollout is imperative to mitigate highly transmissible COVID-19 variants. *EclinicalMedicine* 35:100865. <https://doi.org/10.1016/j.eclinm.2021.100865>
117. Kupferschmidt K. 2021. Viral evolution may herald new pandemic phase. *Science* 371:108–109. <https://doi.org/10.1126/science.371.6525.108>
118. Van Egeren D, Novokhodko A, Stoddard M, Tran U, Zetter B, Rogers M, Pentelute BL, Carlson JM, Hixon M, Joseph-McCarthy D, Chakravarty A. 2021. Risk of rapid evolutionary escape from biomedical interventions targeting SARS-CoV-2 spike protein. *PLoS One* 16:e0250780. <https://doi.org/10.1371/journal.pone.0250780>
119. Saad-Roy CM, Morris SE, Metcalf CJE, Mina MJ, Baker RE, Farrar J, Holmes EC, Pybus OG, Graham AL, Levin SA, Grenfell BT, Wagner CE. 2021. Epidemiological and evolutionary considerations of SARS-CoV-2 vaccine dosing regimes. *Science* 372:363–370. <https://doi.org/10.1126/science.abg8663>
120. Gupta S, Cantor J, Simon KI, Bento AI, Wing C, Whaley CMJHA. 2021. Vaccinations against COVID-19 may have averted up to 140,000 deaths in the United States. *Health Aff (Millwood)* 40:1465–1472. <https://doi.org/10.1377/hlthaff.2021.00619>
121. Pritchard E, Matthews PC, Stoesser N, Eyre DW, Gethings O, Vihta K-D, Jones J, House T, VanSteenHouse H, Bell I, Bell JI, Newton JN, Farrar J, Diamond I, Rourke E, Studley R, Crook D, Peto TEA, Walker AS, Pouwels KB. 2021. Impact of vaccination on new SARS-CoV-2 infections in the United Kingdom. *Nat Med* 27:1370–1378. <https://doi.org/10.1038/s41591-021-01410-w>
122. Yang J, Han M, Wang L, Wang L, Xu T, Wu L, Ma J, Wong G, Liu W, Gao GF, Bi Y. 2022. Relatively rapid evolution rates of SARS-CoV-2 spike gene at the primary stage of massive vaccination. *Biosaf Health* 4:228–233. <https://doi.org/10.1016/j.bshealth.2022.07.001>
123. Allman B, Koelle K, Weissman D, Jain K. 2022. Heterogeneity in viral populations increases the rate of deleterious mutation accumulation. *Genetics* 222:iyac127. <https://doi.org/10.1093/genetics/iyac127>

124. Regoes RR, Hamblin S, Tanaka MM. 2013. Viral mutation rates: modelling the roles of within-host viral dynamics and the trade-off between replication fidelity and speed. *Proc Biol Sci* 280:20122047. <https://doi.org/10.1098/rspb.2012.2047>
125. Johnson T, Barton N. 2002. The effect of deleterious alleles on adaptation in asexual populations. *Genetics* 162:395–411. <https://doi.org/10.1093/genetics/162.1.395>
126. Peck J. 1994. A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex. *Genetics* 137:597–606. <https://doi.org/10.1093/genetics/137.2.597>
127. Elena SF, Sanjuán R. 2005. Adaptive value of high mutation rates of RNA viruses: separating causes from consequences. *J Virol* 79:11555–11558. <https://doi.org/10.1128/JVI.79.18.11555-11558.2005>
128. Flores-Alanis A, Cruz-Rangel A, Rodríguez-Gómez F, González J, Torres-Guerrero CA, Delgado G, Cravioto A, Morales-Espinosa R. 2021. Molecular epidemiology surveillance of SARS-CoV-2: mutations and genetic diversity one year after emerging. *Pathogens* 10:184. <https://doi.org/10.3390/pathogens10020184>
129. Day T, Kennedy DA, Read AF, Gandon S. 2022. Pathogen evolution during vaccination campaigns. *PLoS Biol* 20:e3001804. <https://doi.org/10.1371/journal.pbio.3001804>
130. Wu N, Joyal-Desmarais K, Ribeiro PAB, Vieira AM, Stojanovic J, Sanuade C, Yip D, Bacon SL. 2023. Long-term effectiveness of COVID-19 vaccines against infections, hospitalisations, and mortality in adults: findings from a rapid living systematic evidence synthesis and meta-analysis up to December, 2022. *Lancet Respir Med* 11:439–452. [https://doi.org/10.1016/S2213-2600\(23\)00015-2](https://doi.org/10.1016/S2213-2600(23)00015-2)
131. Kirsebom FCM, Andrews N, Sachdeva R, Stowe J, Ramsay M, Lopez Bernal J. 2022. Effectiveness of chadox1-S COVID-19 booster vaccination against the omicron and delta variants in England. *Nat Commun* 13:7688. <https://doi.org/10.1038/s41467-022-35168-7>
132. Kennedy DA, Read AF. 2017. Why does drug resistance readily evolve but vaccine resistance does not? *Proc Biol Sci* 284:20162562. <https://doi.org/10.1098/rspb.2016.2562>
133. Wang Q, Iketani S, Li Z, Liu L, Guo Y, Huang Y, Bowen AD, Liu M, Wang M, Yu J, Valdez R, Luring AS, Sheng Z, Wang HH, Gordon A, Liu L, Ho DD. 2023. Alarming antibody evasion properties of rising SARS-CoV-2 BQ and XBB subvariants. *Cell* 186:279–286. <https://doi.org/10.1016/j.cell.2022.12.018>
134. Yewdell J. 2021. Antigenic drift: understanding COVID-19. *Immunity* 54:2681–2687. <https://doi.org/10.1016/j.immuni.2021.11.016>
135. Carrat F, Flahault A. 2007. Influenza vaccine: the challenge of antigenic drift. *Vaccine* 25:6852–6862. <https://doi.org/10.1016/j.vaccine.2007.07.027>
136. World Health Organization. 2023. Statement on the antigen composition of COVID-19 vaccines.