Meta-analysis and structural dynamics of the emergence of genetic variants of SARS-CoV-2

Castonguay, Nicolas; Zhang, Wandong; Langlois, Marc-André

Publisher's version / Version de l'éditeur:
https://doi.org/10.3389/fmicb.2021.676314
Frontiers in Microbiology: Virology, 12, June 2021, pp. 1-19, 2021-06-29

frontiers in Microbiology

ORIGINAL RESEARCH
published: 29 June 2021
doi: 10.3389/fmicb.2021.676314

Meta-Analysis and Structural Dynamics of the Emergence of Genetic Variants of SARS-CoV-2

*Nicolas Castonguay[1], Wandong Zhang[2,3] and Marc-André Langlois[1,4]**

[1] Department of Biochemistry, Microbiology and Immunology, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada, [2] Department of Cellular and Molecular Medicine, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada, [3] Human Health Therapeutics Research Centre, National Research Council Canada, Ottawa, ON, Canada, [4] uOttawa Center for Infection, Immunity and Inflammation (CI3), Ottawa, ON, Canada

The novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged in late December 2019 in Wuhan, China, and is the causative agent for the worldwide COVID-19 pandemic. SARS-CoV-2 is a positive-sense single-stranded RNA virus belonging to the betacoronavirus genus. Due to the error-prone nature of the viral RNA-dependent polymerase complex, coronaviruses are known to acquire new mutations at each cycle of genome replication. This constitutes one of the main factors driving the evolution of its relatively large genome and the emergence of new genetic variants. In the past few months, the identification of new B.1.1.7 (United Kingdom), B.1.351 (South Africa), and P.1 (Brazil) variants of concern (VOC) has highlighted the importance of tracking the emergence of mutations in the SARS-CoV-2 genome that impact transmissibility, virulence, and immune and neutralizing antibody escape. Here we analyzed the appearance and prevalence trajectory over time of mutations that appeared in all SARS-CoV-2 genes from December 2019 to April 2021. The goal of the study was to identify which genetic modifications are the most frequent and study the dynamics of their propagation, their incorporation into the consensus sequence, and their impact on virus biology. We also analyzed the structural properties of the spike glycoprotein of the B.1.1.7, B.1.351, and P.1 variants for its binding to the host receptor ACE2. This study offers an integrative view of the emergence, disappearance, and consensus sequence integration of successful mutations that constitute new SARS-CoV-2 variants and their impact on neutralizing antibody therapeutics and vaccines.

Keywords: SARS-CoV-2, COVID-19, variants, evolution, Coronavirus, B1.351 variant, B.1.1.7 variant, P.1 variant

OPEN ACCESS

Edited by:
Kai Huang,
University of Texas Medical Branch at Galveston, United States

Reviewed by:
Siddappa N. Byrareddy,
University of Nebraska Omaha, United States
Svetlana Khaiboullina,
University of Nevada, Reno, United States

***Correspondence:**
Marc-André Langlois
langlois@uottawa.ca

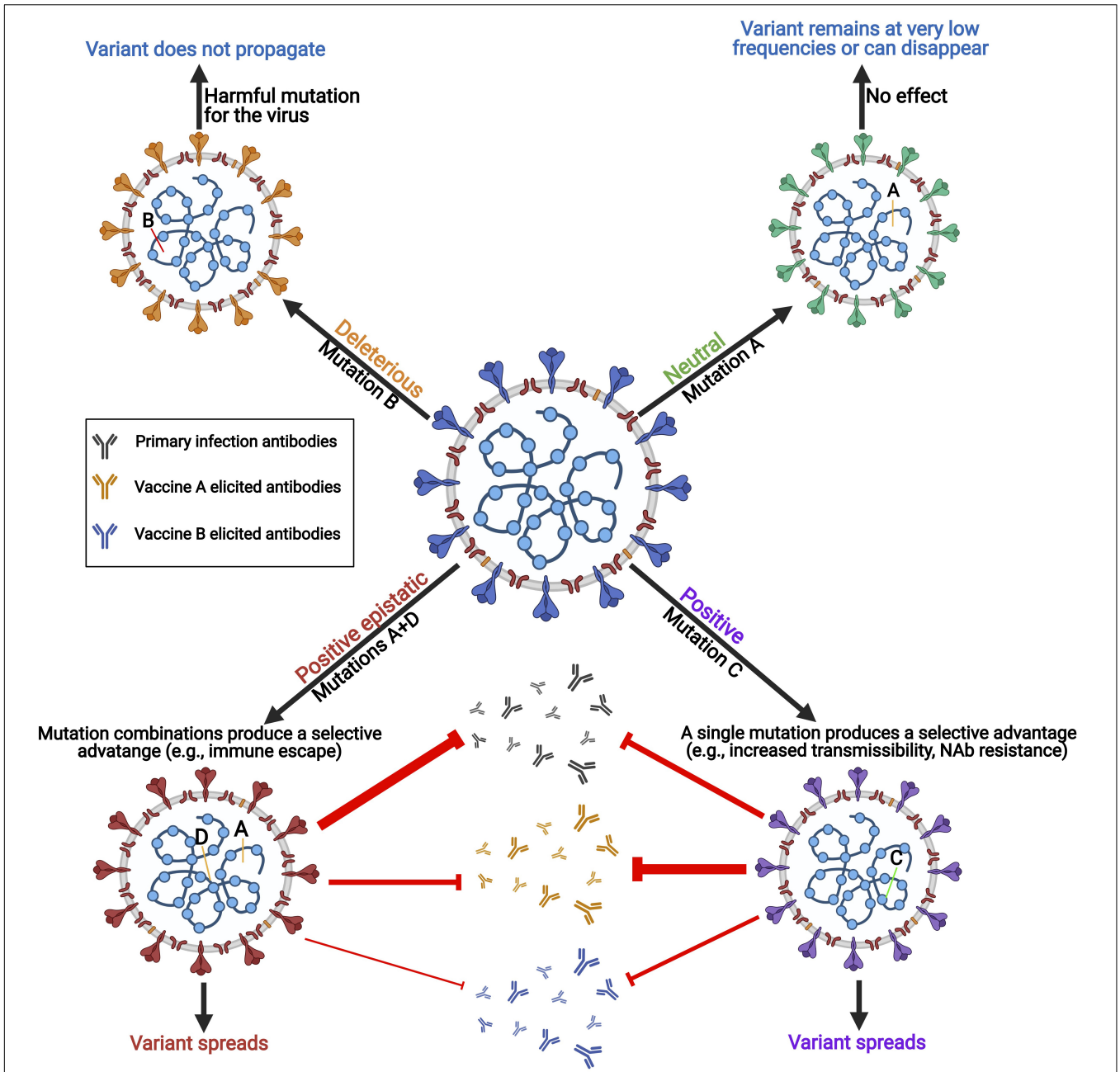
Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *05 March 2021*
Accepted: *27 May 2021*
Published: *29 June 2021*

Citation:
Castonguay N, Zhang W and Langlois M-A (2021) Meta-Analysis and Structural Dynamics of the Emergence of Genetic Variants of SARS-CoV-2. Front. Microbiol. 12:676314. doi: 10.3389/fmicb.2021.676314

HIGHLIGHTS

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the etiological agent of COVID-19, which has caused > 3.4 million deaths worldwide as of April 2021. Mutations occur in the genome of SARS-CoV-2 during viral replication and affect viral infectivity, transmissibility, and virulence. In early March 2020, the D614G mutation in the spike protein emerged, which increased viral transmissibility and is now found in over 90% of all SARS-CoV-2 genomic sequences in GISAID database. Between October and December 2020, B.1.1.7 (United Kingdom), B.1.351 (South Africa), and P.1 (Brazil) variants of concern (VOCs) emerged, which have increased



GRAPHICAL ABSTRACT | Errors are regularly made when SARS-CoV-2 replicates its RNA genome. The viral polymerase complex is error-prone with imperfect proofreading abilities. These errors or mutations often lead to deleterious or neutral effects on the virus. However, sometimes these mutations have a positive effect and create genetic variants of the virus with different features including increased transmissibility, pathogenicity, and immune escape capabilities. When mutations work collaboratively to create a new virus feature, this is called epistasis.

neutralizing antibody escape capabilities because of mutations in the receptor-binding domain of the spike protein. Characterizing mutations in these variants is crucial because of their effect on adaptive immune responses, neutralizing antibody therapy, and their impact on vaccine efficacy. Here we tracked and analyzed mutations in SARS-CoV-2 genes since the beginning of the pandemic and investigated their functional impact on the spike of these three VOCs.

INTRODUCTION

In late December 2019, a new betacoronavirus known as severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) emerged in the city of Wuhan in the province of Hubei, China (Zhu et al., 2020). SARS-CoV-2 is the etiological viral agent for the worldwide COVID-19 pandemic resulting in more than 162 million infected and 3.4 million deaths worldwide as

of April 2021 (Dong et al., 2020; Wu F. et al., 2020). SARS-CoV-2 is an enveloped, positive-sense single-stranded RNA (+ ssRNA) virus with a genome length of 29,811 nucleotides (Kim et al., 2020). The mutation rates of RNA viruses are generally higher than that of DNA viruses because of the low fidelity of their viral RNA polymerases (Drake, 1993; Duffy, 2018). Mutations occur when viral replication enzymes introduce errors in the viral genome resulting in the creation of premature termination codons, deletions, and insertions of nucleotides that can alter open reading frames and result in amino acid substitutions in viral proteins. These mutations, combined with the selective pressure of the human immune system, lead to the selection and evolution of viral genomes (Drake, 1993; Peck and Lauring, 2018). However, coronaviruses are one of the few members of the RNA virus family that possess limited, but measurable, proofreading ability via the 3′- to 5′-exoribonuclease activity of the non-structural viral protein 14 (nsp14) (Minskaia et al., 2006; Eckerle et al., 2010). Coronaviruses are therefore expected to evolve through genetic drift much slower than other RNA viruses that do not have this ability, such as influenza viruses (Eckerle et al., 2010; Ma et al., 2015). Additionally, SARS-CoV-2 and other coronaviruses have low known occurrences of recombination between family members (i.e., genetic shift), and therefore are mostly susceptible to genetic drift (Rausch et al., 2020).

SARS-CoV-2 has reached pandemic status due to its presence on every continent and has since maintained a high level of transmissibility across hosts of various ethnical and genetic backgrounds (Chu et al., 2020; Dong et al., 2020). Moreover, SARS-CoV-2 infections have been reported to naturally infect minks, ferrets, cats, tiger, and dogs, which allows the virus to replicate in completely new hosts and mutate to produce new variants and possibly new strains (Li X. et al., 2020; Shi et al., 2020). In March 2020, the now dominant D614G mutation first emerged in the spike protein (S) of SARS-CoV-2. The S protein is present as a trimer at the surface of the viral envelope and is responsible for attachment of the virus to the human angiotensin-converting enzyme 2 (hACE2), the entry receptor for SARS-CoV-2 into human cells (Walls et al., 2020). Published evidence has now shown that D614G increases viral fitness, transmissibility, and viral load but does not directly affect COVID-19 pathogenicity (Korber et al., 2020; Li Q. et al., 2020; Plante et al., 2020; Zhang et al., 2020). Additionally, emerging evidence indicates that D614G may have epistatic interactions that exacerbate the impact of several other independent mutations (Li Q. et al., 2020). Mutations in the S protein, and particularly in the receptor-binding domain (RBD), are of very high concern given that they can directly influence viral infectivity, transmissibility, and resistance to neutralizing antibodies and T cell responses.

New mutations are frequently and regularly detected in the genome of SARS-CoV-2 through whole genome sequencing; however, very few of these mutations make it into the transmitted viral consensus sequence. The reference strain is generally regarded as the dominant transmitted strain at a given time. Its sequence is determined by aligning large numbers of recently sequenced genomes and establishing a consensus sequence composed of the highest-frequency nucleotide for each position in the viral genome. A genetic variant is a version of the reference strain that has acquired one or several new mutations and acts as the founder for further genetic diversification and evolution. Mutations arise regularly in the reference strain, but few are longitudinally conserved. Genetic variants are therefore the rare successful offshoots of the reference strain.

Some variants rise rapidly in frequency and then collapse and disappear, while others will rise and overtake the frequency of the reference strain and become the new reference. There are three main genetic variants that have emerged in the past few months with sustained upward frequency trajectories. The first is the United Kingdom variant, also known as B.1.1.7/501Y.V1 (B1.1.7). It was first detected in September 2020 and is now present worldwide and poised to become the new reference strain (Rambaut et al., 2020; O'Toole et al., 2021). The South African variant, also known as B.1.351/501Y.V2 (B.1.351), was first reported in October 2020 and is now increasing in prevalence in South Africa, Europe, and North America (O'Toole et al., 2021; Tegally et al., 2021). The Brazilian variant, P.1/501Y.V3 (P.1), was first detected in travelers from Brazil that landed in Japan in January 2021 (Faria et al., 2021). It has since been identified in 42% of specimens in the Amazonian city of Manaus and in the United States at the end of January 2021 (Faria et al., 2021). These three variants are associated with increased resistance to neutralizing antibodies (Nabs), and all possess the N501Y mutation, which is a mutation in the RBD that is critical for the spike protein to interact with hACE2 (Yi et al., 2020; Greaney et al., 2021; Wang et al., 2021). This mutation is reported to cause increased resistance to Nabs, increased transmissibility, and increased virulence in animal models (Gu et al., 2020). In addition to the N501Y mutation, both the South African and Brazil variants possess RBD mutations K417N(T) and E484K, which are also associated with further increased Nabs escape capabilities (Greaney et al., 2021; Wang et al., 2021).

Here we present a retrospective metadata analysis of mutations throughout the SARS-CoV-2 genome that reached at least a 1% worldwide frequency between December 2019 and January 2021. We specifically investigated their frequency trajectory over time and their fixation into the reference sequencing using the Global Initiative on Sharing Avian Influenza Data (GISAID) (Shu and McCauley, 2017). Additionally, we analyzed mutations in the S protein of the B.1.1.7, B.1.351, and P.1 variants and illustrated their impact on molecular interactions between the S protein and hACE2 and their potential impact on Nabs.

MATERIALS AND METHODS

Data Collection and Mutational Analysis

Genomes uploaded to the GISAID EpiCoV^{TM} server database were analyzed from December 1, 2019, to December 31, 2020, and viral sequences with submission dates from December 1, 2019 to January 6, 2021 were selected. We first selected recurring mutations that were present in more than 500 reported

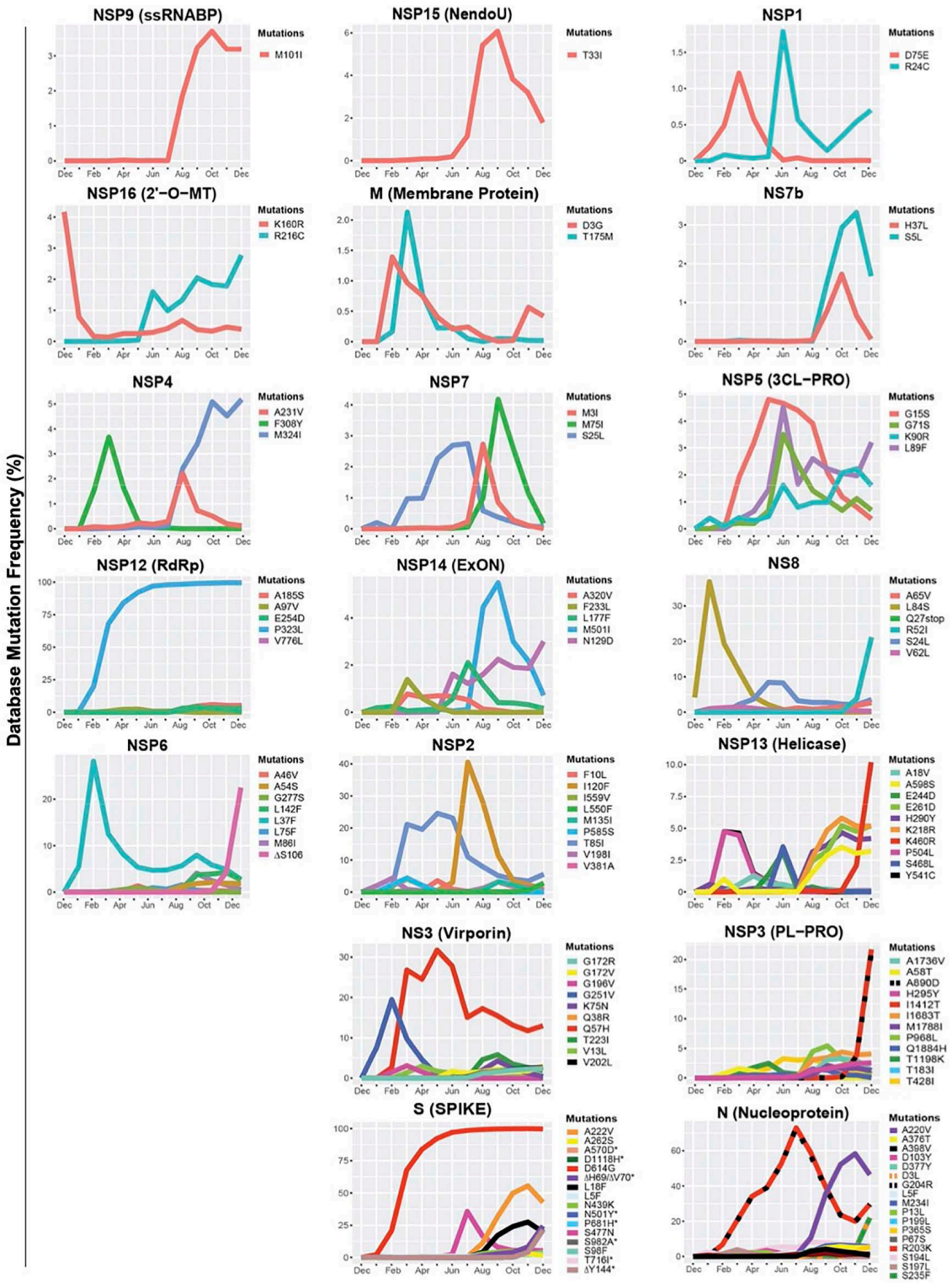


FIGURE 1 | Variations in mutations and mutation frequencies in severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) genes. The occurrence and frequency of mutations in various SARS-CoV-2 genes are presented for the period between December 2019 and January 1, 2021. Plotted are mutations that reached at least a 1% worldwide frequency. SARS-CoV-2 genes are represented with the function of the genes in parentheses. Graphs were generated using RStudio. *Overlapping curves.

TABLE 1 | Non-synonymous mutations in severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) genes with a worldwide frequency ≥ 0.01%.

Protein (gene)	Genome nucleotide mutation	Codon	AA mutation	Frequency (%)*	Effect
S	A23403G	G**A**T to G**G**T	D614G	99.70	Moderate increase in Transmissibility Covid-19 Genomic UK Consortium, 2021
	C22227T	G**C**T to G**T**T	A222V	42.79	No effect Covid-19 Genomic UK Consortium, 2021
	C21615T	**C**TT to **T**TT	L18F	19.86	N/A
	G22992A	A**G**C to A**A**C	S477N	5.36	Spike-hACE2 complex stability and Nab interference Zahradník et al., 2021
	C22879A	AA**C** to AA**A**	N439K	4.09	Antibody escape ability Li Q. et al., 2020; Covid-19 Genomic UK Consortium, 2021
	C21575T	**C**TT to **T**TT	L5F	1.44	N/A
	C21855T	T**C**T to T**T**T	S98F	2.74	N/A
	G22346T	**G**CT to **T**CT	A262S	1.64	N/A
	A23063T	**A**AT to **T**AT	N501Y	80.16	Increase affinity for hACE2 Covid-19 Genomic UK Consortium, 2021
	C23709T	A**C**A to A**T**A	T716I	75.37	N/A
	C23271A	G**C**T to G**A**T	A570D	76.30	N/A
	G24914C	**G**CA to **C**AC	D1118H	74.82	N/A
	C23709A	C**C**T to C**A**T	P681H	79.18	N/A
	T24506G	**T**CA to **G**CA	S982A	74.61	N/A
	21765-21770del	**N/A**	ΔH69/ΔV70	75.13	Antibody escape Covid-19 Genomic UK Consortium, 2021
	21991-21993del	**N/A**	ΔY144	76.91	Decrease infectivity, antibody escape Li Q. et al., 2020
NSP1(ORF1ab)	C335T	**C**GC to **T**GC	R24C	0.38	N/A
NSP2	G1210T	AT**G** to AT**T**	M135I	0.01	N/A
(ORF1ab)	C1059T	A**C**C to A**T**C	T85I	13.95	N/A
	T1947C	G**T**T to G**C**T	V381A	1.08	N/A
	C2453T	**C**TC to **T**TC	L550F	5.93	N/A
NSP3	C7926T	G**C**A to G**T**A	A1736V	0.12	N/A
(ORF1ab)	T7767C	A**T**C to A**C**C	I1683T	0.19	N/A
	C5622T	C**C**T to C**T**T	P968L	0.09	N/A
	G8371T	CA**G** to CA**T**	Q1884H	0.05	N/A
	C4002T	A**C**T to A**T**T	T428I	0.44	N/A
	C5388A	G**C**T to G**A**T	A890D	76.25	N/A
	G8083A	AT**G** to AT**A**	M1788I	0.68	N/A
	C3602T	**C**AC to **T**AC	H295Y	0.30	N/A
NSP4	C9246T	G**C**T to G**T**T	A231V	0.12	N/A
(ORF1ab)	G9526T	AT**G** to AT**T**	M324I	0.14	N/A
NSP5	A10323G	A**A**G to A**G**G	K90R	3.07	N/A
(ORF1ab)	G10097A	**G**GT to **A**GT	G15S	0.35	N/A
	C10319T	**C**TT to **T**TT	L89F	1.44	N/A
NSP6	C11109T	G**C**T to G**T**T	A46V	0.15	N/A
(ORF1ab)	C11396T	**C**TT to **T**TT	L142F	0.02	N/A
	G11083T	TTG to TT**T**	L37F	1.26	N/A
	C11195T	**C**TT to **T**TT	L75F	0.13	N/A
	G11230T	AT**G** to AT**T**	M86I	0.03	N/A
	G11801A	**G**GT to **A**GT	G277S	0.07	N/A
	G11132T	**G**CT to **T**CT	A54S	0.02	N/A
	11288-11296 del	N/A	ΔS106/ΔG3676/ ΔF3677	87.61	N/A
NSP7	G12067T	AT**G** to AT**T**	M75I	0.15	N/A
(ORF1ab)	C11916T	T**C**A to T**T**A	S25L	0.10	N/A
NSP9(ORF1ab)	G12988T	AT**G** to AT**T**	M101I	0.23	N/A
NSP12(ORF1ab)	C14408T	C**C**T to C**T**T	P323L	99.49	Improves processivity by interaction with NSP8 Kannan et al., 2020
	G15766T	**G**TG to **T**TG	V776L	0.14	N/A
	G13993T	**G**CT to **T**CT	A185S	0.09	N/A
	G15598A	**G**TC to **A**TC	V720I	0.17	N/A
	G14202T	GA**G** to GA**T**	E254D	0.01	N/A
	C13730T	G**C**T to G**T**T	A97V	0.35	N/A

(Continued)

TABLE 1 | Continued

Protein (gene)	Genome nucleotide mutation	Codon	AA mutation	Frequency (%)*	Effect
NSP13	C16289T	G**C**T to G**T**T	A18V	0.03	N/A
(ORF1ab)	G17019T	GA**G** to GA**T**	E261D	0.22	N/A
	C17104T	**C**AT to **T**AT	H290Y	0.23	N/A
	C17747T	C**C**T to C**T**T	P504L	0.07	Increases hydrophobicity of 2A domain Ugurel et al., 2020
	C17639T	T**C**A to T**T**A	S468L	0.02	N/A
	A17615G	A**A**G to A**G**G	K460R	15.82	N/A
	A16889G	A**A**A to A**G**A	K218R	0.08	N/A
	G18028T	**G**CA to **T**CA	A598S	0.26	N/A
NSP14	C18998T	G**C**A to G**T**A	A320V	0.01	N/A
(ORF1ab)	C18568T	**C**TC to **T**TC	L177F	0.22	N/A
	G19542T	AT**G** to AT**T**	M501I	0.21	N/A
	A18424G	**A**AT to **G**AT	N129D	1.37	N/A
NSP15(ORF1ab)	C19718T	A**C**A to A**T**A	T33I	0.06	N/A
NSP16	A21137G	A**A**G to A**G**G	K160R	4.69	N/A
(ORF1ab)	A21390G	TT**A** to TT**G**	R216C	1.30	N/A
N	C28932T	G**C**T to G**T**T	A220V	0.38	N/A
	G28580T	**G**AT to **T**AT	D103Y	0.02	N/A
	G28883C	**G**GA to **C**GA	G204R	75.15	Destabilizes N protein Wu S. et al., 2020
	C28311T	C**C**C to C**T**C	P13L	2.00	N/A
	C28869T	C**C**A to C**T**A	P199L	5.55	N/A
	G28882A	A**G**C to A**A**A	R203K	81.96	Destabilizes N protein Wu S. et al., 2020
	C28883A	AG**C** to AA**A**	R203K	81.96	Destabilizes N protein Wu S. et al., 2020
	C28854T	T**C**A to T**T**A	S194L	0.33	Destabilizes N protein Wu S. et al., 2020
	C28977T	T**C**T to T**T**T	S235F	76.34	N/A
	G28280C	**G**AT to **C**TA	D3L	75.53	N/A
	A28281T	G**A**T to C**T**A	D3L	75.53	N/A
	T28282A	GA**T** to CT**A**	D3L	75.53	N/A
	G28975C	AT**G** to AT**C**	M234I	7.58	N/A
	C28472T	**C**CT to **T**CT	P67S	1.38	N/A
	G29399A	**G**CT to **A**CT	A376T	0.09	N/A
	C29366T	**C**CA to **T**CA	P365S	0.15	N/A
	G29402T	**G**AT to **T**AT	D377Y	1.61	N/A
	C28887T	A**C**T to A**T**T	T205I	7.00	N/A
	C29466T	G**C**A to G**T**A	A398V	0.51	N/A
M	A26530G	G**A**T to G**G**T	D3G	0.01	N/A
	C27046T	A**C**G to A**T**G	T175M	0.02	N/A
NS3	G25907T	G**G**T to G**T**T	G172V	1.41	N/A
(ORF3a)	G25617T	AA**G** to AA**T**	K75N	0.02	N/A
	G25563T	CA**G** to CA**T**	Q57H	14.36	N/A
	C26060T	A**C**T to A**T**T	T223I	0.38	N/A
	G25429T	**G**TA to **T**TA	V13L	0.05	N/A
	G25996T	**G**TA to **T**TA	V202L	0.17	N/A
	A25505G	C**A**A to C**G**A	Q38R	0.04	N/A
	G25906C	**G**GT to **C**GT	G172R	5.31	N/A
NS7b	A27866T	C**A**T to C**T**A	H37L	0.01	N/A
(ORF7b)	AT27867A	CA**T** to CT**A**	H37L	0.01	N/A
	C27769T	T**C**A to T**T**A	S5L	0.07	N/A
NS8	C28087T	G**C**T to G**T**T	A65V	0.77	N/A
(ORF8)	T28144C	T**T**A to T**C**A	L84S	0.14	N/A
	C27964T	T**C**A to T**T**A	S24L	1.38	N/A
	G28077T	**G**TG to **T**TG	V62L	0.24	N/A
	C27972T	**C**AA to **T**AA	Q27stop	76.05	Inactivates NS8 Pereira, 2020
	G28048T	A**G**A to A**T**A	R52I	76.00	N/A

**Mutation frequency as of April 30, 2021.*

genomes by August 2020, and another selection was made in January 2021 to capture recurring mutations present in more than 4,000 reported genomes in GISAID. This strategy allowed us to study mutations reaching at least 1% in worldwide frequency. We filtered through 309,962 genomes for the analysis of selected mutations.

The worldwide frequency of the hCoV-19/Wuhan strain, and hCoV-19/D614G, B.1.1.7, B.1.351, and P.1 variants were analyzed from December 1, 2019 to April 30, 2021. For the analysis of the mutations in B.1.1.7, B.1.351, and P.1 variants, we used the GISAID EpiCoV^{TM} server database. Viral sequences on GISAID with submission dates between December 1, 2019 and April 30, 2021 were selected for the analysis, and we filtered through 1,090,689 genomes for the analysis of the variants on GISAID EpiCoV^{TM} server database. Only complete SARS-CoV-2 genomes (28 to 30 kbps) isolated from human hosts were analyzed. MUSCLE alignment tool on UGene and SnapGene was used to determine the nucleotide mutations and codon changes of the non-synonymous and synonymous mutations by sequence alignments to the NCBI

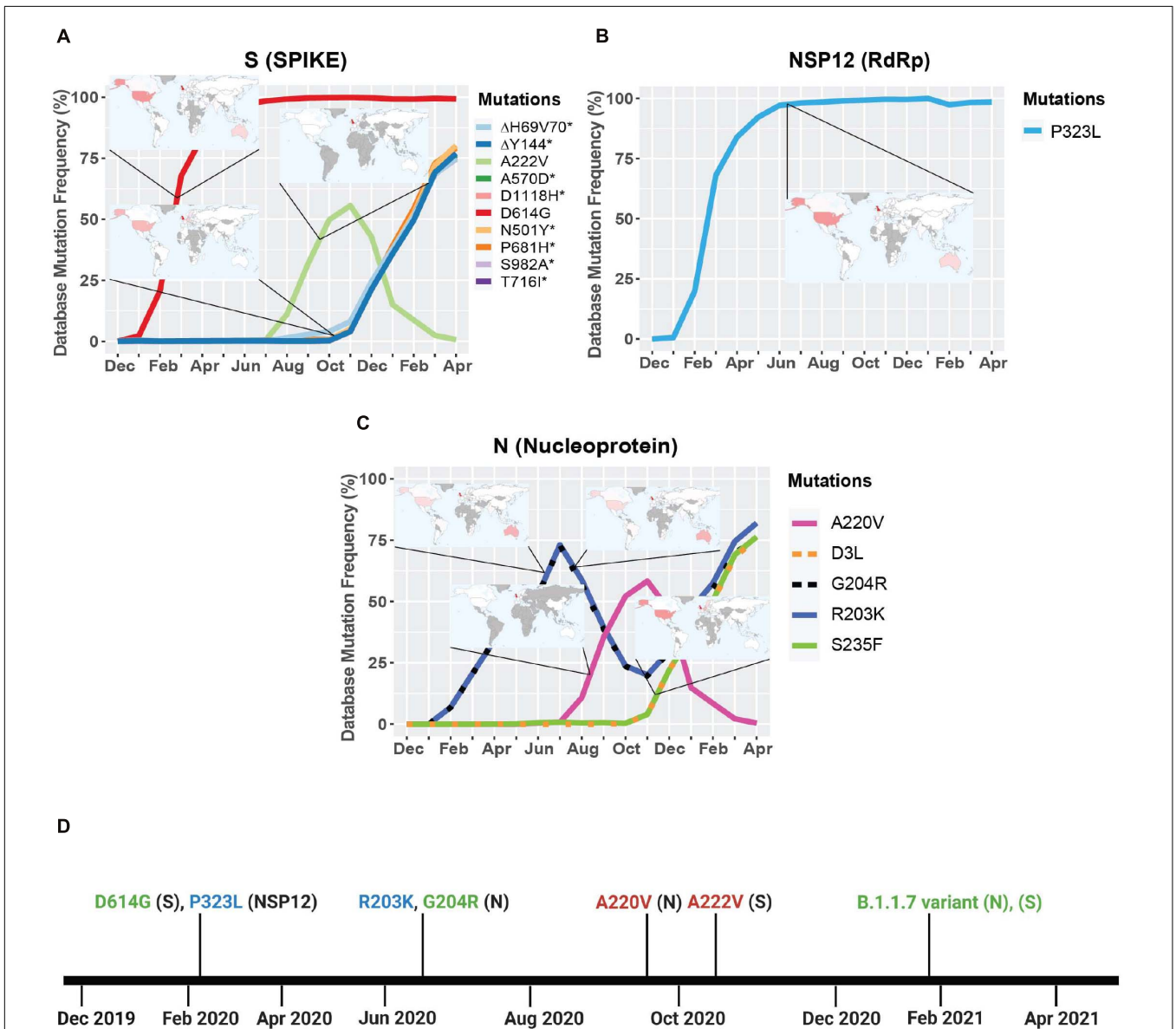


FIGURE 2 | Geographic location and timeline of dominant mutations in NSP12, S, and N genes. **(A)** Frequency of S protein mutations with corresponding geographic maps. **(B)** Frequency of RdRp mutations with corresponding geographic maps. **(C)** Frequency of nucleoprotein mutations with corresponding geographic maps. **(D)** Timeline of the appearance of mutations reaching a frequency higher than 50% worldwide between December 2019 and April 2021. For the geomaps: a low frequency of reported cases of the mutations is represented in white, while higher frequencies are represented from pink to red, and gray represents no data. All maps were taken from Global Initiative on Sharing Avian Influenza Data (GISAID). Graphs were generated using RStudio and Biorender. *Overlapping curves.

SARS-CoV-2 reference genome (NC_045512). All genomes uploaded to the GISAID database that were used in this study are presented in **Supplementary Table 1**. Graphs of mutations and variants were plotted with RStudio using the ggplot2 command. The timeline and genome illustrations were produced in Biorender.

Structural Modeling

Mutations in the spike protein in complex with hACE2 were analyzed using a mutagenesis tool for PyMOL (PDB: 7A94). Visualizations of mutations in the B.1.1.7, B.1.351, and P.1 variants were produced using the spike protein closed conformation (PDB: 6ZGE), interaction with hACE2 (PDB: 7A94), interaction with C102 Nab (PDB: 7K8M), and interaction with C121 Nab (PDB: 7K8X). Figures and rendering were prepared with PyMOL.

RESULTS

Identification of Emerging Mutations in Various Severe Acute Respiratory Syndrome Coronavirus 2 Genes

Emerging mutations in the SARS-CoV-2 genome were investigated to illustrate the fluctuations of these mutations during a period of 12 months. We tracked genome mutations with a worldwide frequency greater than 1% from December 2019 to December 31, 2020 in the GISAID database. Genes NSP8, NSP10, NS6, NS7a, and E are not illustrated in **Figure 1** given that they did not display mutations with frequencies sufficiently high to meet our inclusion criteria during the study period. This is indicative that these are some of the most conserved sequences of the SARS-CoV-2 genome. Viral proteins that have low mutation frequencies could be due to

TABLE 2 | Synonymous, non-synonymous, and deletion mutations in the B.1.1.7 (United Kingdom) variant.

Variant	Protein (gene)	Genome nucleotide mutation	S or NS	AA mutation Rambaut et al., 2020	Domain	Frequency* (%)	Effect
B.1.1.7	*NSP2 (ORF1ab)*	C913T	S			N/A	N/A
	NSP3 (ORF1ab)	C3267T	NS	T183I		76.14	N/A
		C5388A	NS	A890D		76.24	N/A
		C5986T	S			N/A	N/A
		T6954C	NS	I1412T		75.21	N/A
	NSP6 (ORF1ab)	11288-11296del	NS	ΔS106/ ΔG107/ΔF108		87.61	N/A
	NSP12 (ORF1ab)	C14676T	S			N/A	N/A
		C15279T	S			N/A	N/A
	NSP13 (ORF1ab)	C16176T	S			N/A	N/A
	S	21765-21770del	NS	ΔH69/ΔV70	NTD	75.13	Antibody escape Covid-19 Genomic UK Consortium, 2021
		21991-21993del	NS	ΔY144	NTD	76.91	Decreases infectivity, Antibody escape Li Q. et al., 2020
		A23063T	NS	N501Y	RBD	80.16	Increases affinity for hACE2 Wu et al., 2017
		C23271A	NS	A570D	SD1	76.30	
		A23403G	NS	D614G	SD2	99.29	Moderate increase in transmissibility Li Q. et al., 2020; Wu et al., 2017
		C23709A	NS	P681H	SD2	79.18	N/A
		C23709T	NS	T716I		75.37	N/A
		T24506G	NS	S982A	HR1	74.61	N/A
		G24914C	NS	D1118H		74.82	N/A
	NS8 (ORF8)	C27972T	NS	Q27stop		76.05	Inactivation of NS8 Pereira, 2020
		G28048T	NS	R52I		76.00	N/A
		A28111G	NS	Y73C		76.07	N/A
	M	T26801C	S			N/A	N/A
	N	G28280C	NS	D3L		75.53	N/A
		A28281T	NS	D3L		75.53	N/A
		T28282A	NS	D3L		75.53	N/A
		C28977T	NS	S235F		76.34	N/A

**Mutation frequency as of April 30, 2021.*

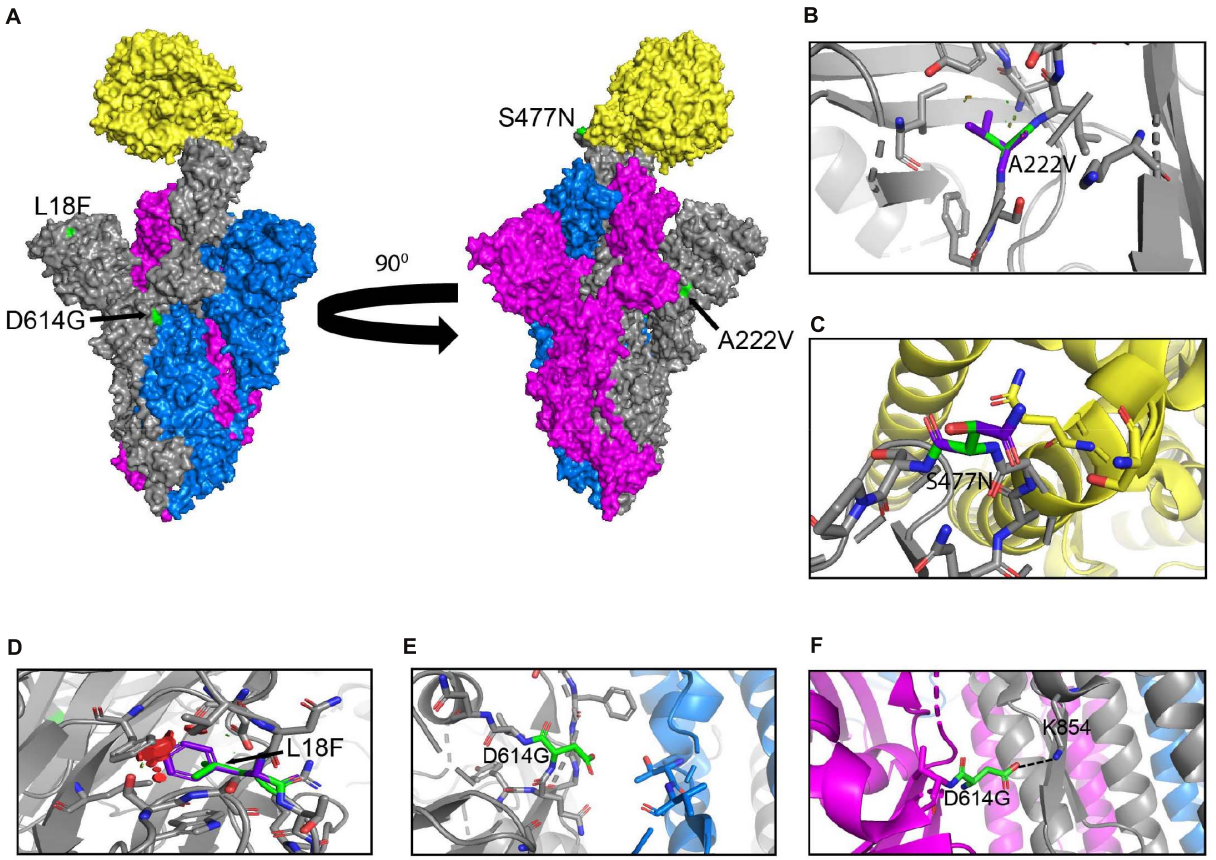


FIGURE 3 | Structural rendering of the most frequent mutations in the S protein. **(A)** Surface representation of human angiotensin-converting enzyme 2 (hACE2) (yellow) in complex with S protein trimers illustrated in gray, blue, and magenta. Interactions of high frequency mutations are presented as follows: **(B)** A222V, **(C)** S477N, **(D)** L18F, **(E)** D614G in the open conformation, and **(F)** D614G in the closed conformation. Reference sequence residues are illustrated in green, and the mutated amino acid is represented in purple. The red markers illustrate the steric clash when the mutations are inserted into the structure. Graphs were generated using PyMOL.

evolutionary constraints. Some of these viral proteins may be evolutionarily conserved and play important roles in virus assembly and stability, genome replication, and viral release from infected cells. Our analysis highlights the fixation of the D614G mutation in the S protein and that of the P323L mutation in the RNA-dependent RNA polymerase (RdRp) (**Figure 1**). These are the only mutations to have successfully become part of the reference sequence as of December 2020. They appeared to have emerged simultaneously in January 2020 and became present in more than 90% of all sequenced genomes by June 2020. Additionally, some other mutations also emerged rapidly but then stabilized in frequency or faded out. For example, Q57H (NS3), R203K (N), and G204R (N) are mutations that emerged rapidly and appeared to have stabilized at a frequency of 15% to 40%. Others like I120F (NSP2), L37F (NSP6), S477N (S), and L84S (NS8) illustrate mutations that emerged rapidly and then faded-out just as quickly. We also demonstrate that most genes in SARS-CoV-2 have mutations with overall frequencies lower than 10% (**Figure 1**). These mutations are also summarized in **Table 1**, which illustrates nucleotide substitution producing the amino acid change, the frequency, and their respective effects. Globally, there is an uneven effort to test for SARS-CoV-2 infections in the population and sequence the viral genomes in those infected. As a result, emerging mutations go unreported until they reach

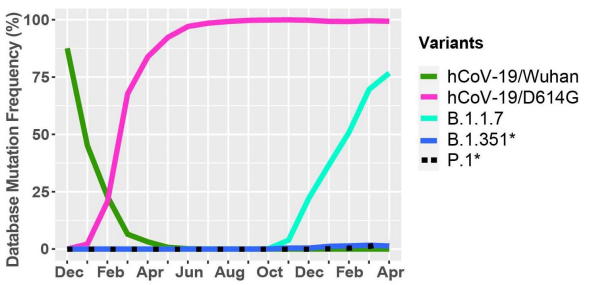


FIGURE 4 | Worldwide frequency of B.1.1.7, B.1.351, P.1 and D614G variants. Database variants frequency were analyzed from December 2019 to April 30, 2021. hCoV-19/Wuhan is the original reference strain and hCoV-19/D614G is the current reference strain containing the D614G mutation in the S protein. B.1.1.7, B.1.351, and P.1 represent the United Kingdom, South African, and Brazilian variants. *Overlapping curves.

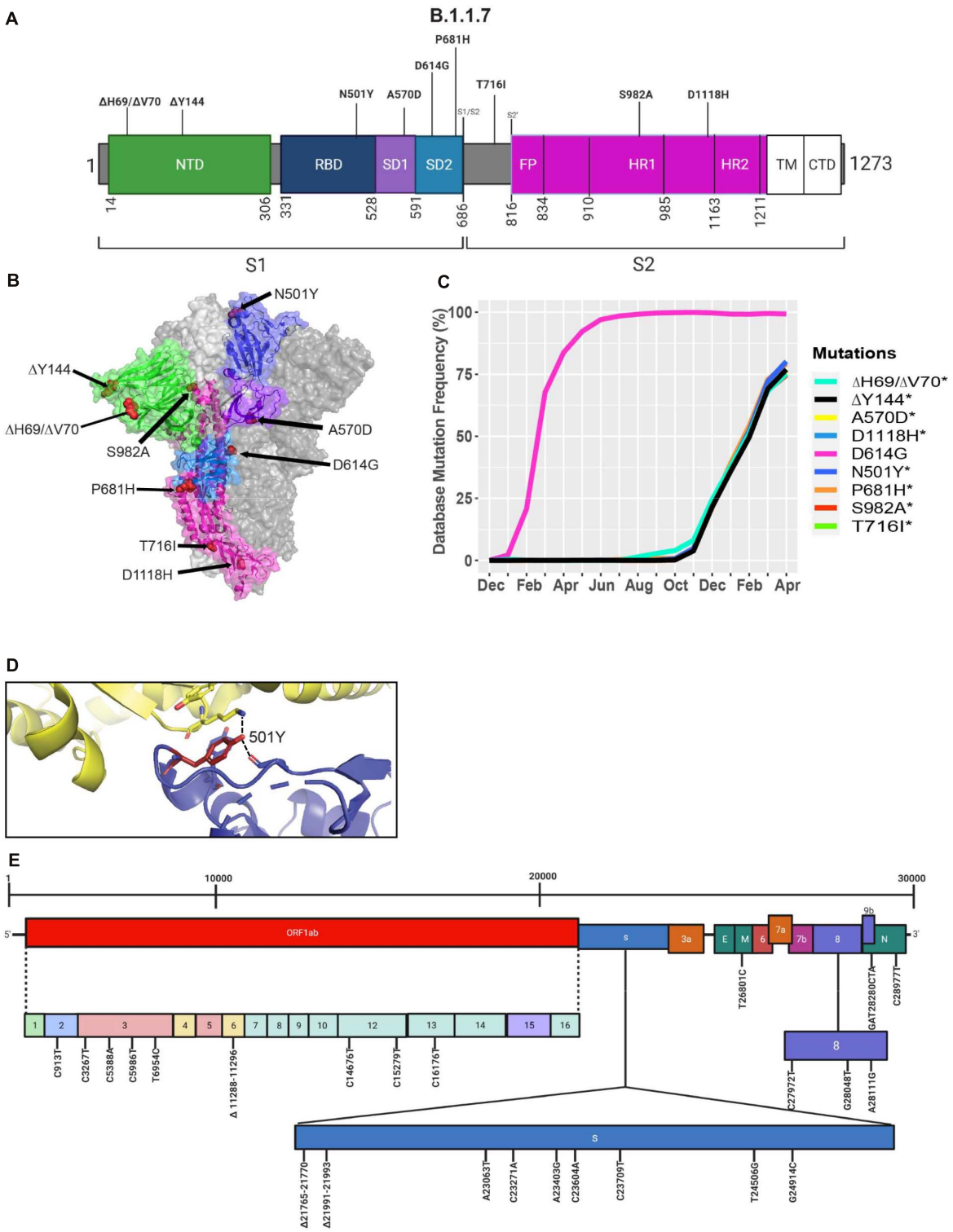


FIGURE 5 | Analysis of mutations in the B.1.1.7 variant. **(A)** Mutation map of the spike protein of B.1.1.7. **(B)** Structural representation of spike with ACE2. B.1.1.7 S protein mutations are presented in red. N-terminal domain (NTD) (green), receptor-binding domain (RBD) (blue), SD1 (purple), subdomain 2 (SD2) (light blue), and S2 (magenta) are illustrated. The other S protein monomers are displayed in gray and white. **(C)** Frequency of the mutations in the S protein, B.1.1.7 variant from December 2019 to April 30, 2021. **(D)** Interaction of the N501Y (red) mutation in the RBD (blue) of S protein with hACE2 (yellow). The dash lines indicate interactions with adjacent residues. **(E)** Genome of the SARS-CoV-2 B.1.1.7 variant with identified nucleotide substitutions and deletions. Graphs were generated using Biorender, PyMOL, and RStudio. *Overlapping curves.

countries with more intensive sequencing capabilities. Therefore, mutations presented here vastly underrepresent the global landscape of mutation frequency dynamics.

Geographic Localization and Timeline of the Viral Genes With Mutations Higher Than 50% Frequency

We next turned our attention to viral genes where at least one mutation reached a frequency higher than 50%. Only three genes met this criterion: NSP12, S, and N. We then took their mutation graphs and added worldwide geographic data provided from GISAID. Geomaps are useful to determine if a mutation is a localized and regional event or is found worldwide. In the S protein, D614G is found worldwide with initially higher reported cases in the United States, United Kingdom, and Australia, probably due to more intense large-scale testing and sequencing, while A222V is mostly reported in the United Kingdom, but has not yet been reported in South America, Central and East African regions (**Figure 2A**). Like D614G, P323L in the RdRp is also found worldwide, with higher reported cases in the United States, United Kingdom, and Australia (**Figure 2B**). The N gene does not yet have successful mutations that have attained reference sequence status, but R203K, G204R, A220V, D3L, and S235F all reached a frequency of 50% or higher during the study period. A220V emerged in August of 2020 and reached a frequency higher than 50% in October of 2020 (**Figure 2D**). Mutations G204R, R203K, and A220V in the N gene are reported at high frequency in the United Kingdom, but were not been detected in South America, Central, East, and South African regions (**Figure 2C**). We also observe the emergence of several mutations in S and N present in the B.1.1.7 variant, which now have a frequency higher than 65% (**Figures 2C,D** and **Table 2**). The analyses of these data illustrate the localization of the most prevalent mutations to date, which appear to be mostly present in Western high-income countries. This, however, is undoubtable attributable to overall more intense testing and sequencing.

Localization and Molecular Interactions of Prevalent S Protein Mutations

Here we illustrate the molecular interactions and spatial localization of the mutations in the S protein. PyMOL was used to model structures of the S protein binding to ACE2, and we analyze the possible effects of specific mutations at given positions in the protein. **Figure 3A** presents an overview of the physical positions of S protein mutations. The A222V mutation has no reported effects on protein stability, neutralizing antibody escape, and affinity for hACE2 (**Table 1**). The substitution from A to V results in a low steric clash between neighboring residues (**Figure 3B**). The S477N substitution in the RBD enables increased stability during hACE2–RBD interactions (**Figure 3C** and **Table 1**). In the N-terminal domain (NTD), mutation L18F leads to a steric clash between neighboring residues (**Figure 3D**). However, this does not appear to impact the stability of the protein given that no such effects have been reported so far (**Table 1**). In the closed conformation, D614 makes an ionic bond with K854 in the S2 subunit of another S protein monomer (**Figure 3F**; Benton et al., 2020). In the open conformation, D614 (or G614) does not make interactions or display steric clashes with neighboring residues (**Figure 3E**).

The Emergence of the B.1.1.7 Variant in the United Kingdom

A new variant was discovered in late 2020 in the United Kingdom that displayed increased affinity to hACE2, virulence, and Nabs escape capabilities (**Figure 4**; Starr et al., 2020; Greaney et al., 2021; Haas et al., 2021; Wang et al., 2021). Here we further investigated the B.1.1.7 variant by looking at S protein mutations of this variant in complex with Nabs and hACE2. We mapped

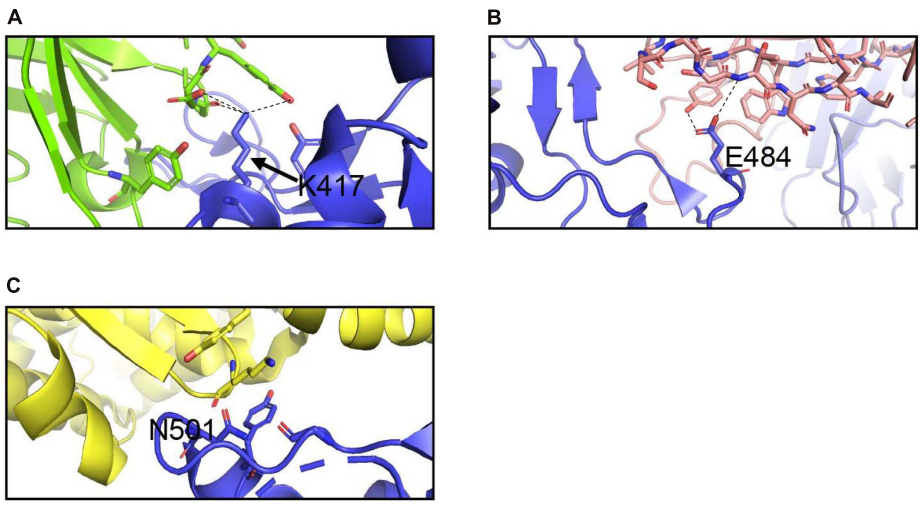


FIGURE 6 | Interactions of K417, E484, and N501 of the S protein with neutralizing antibodies and hACE2. **(A)** Interaction of K417 (blue) with C102 Nab (green) residues. **(B)** Interaction of E484 (blue) with C121 Nab (light pink). **(C)** Interaction of N501 (blue) with hACE2 (yellow) Dashes lines indicate interactions between residues. The graphs were generated using PyMOL.

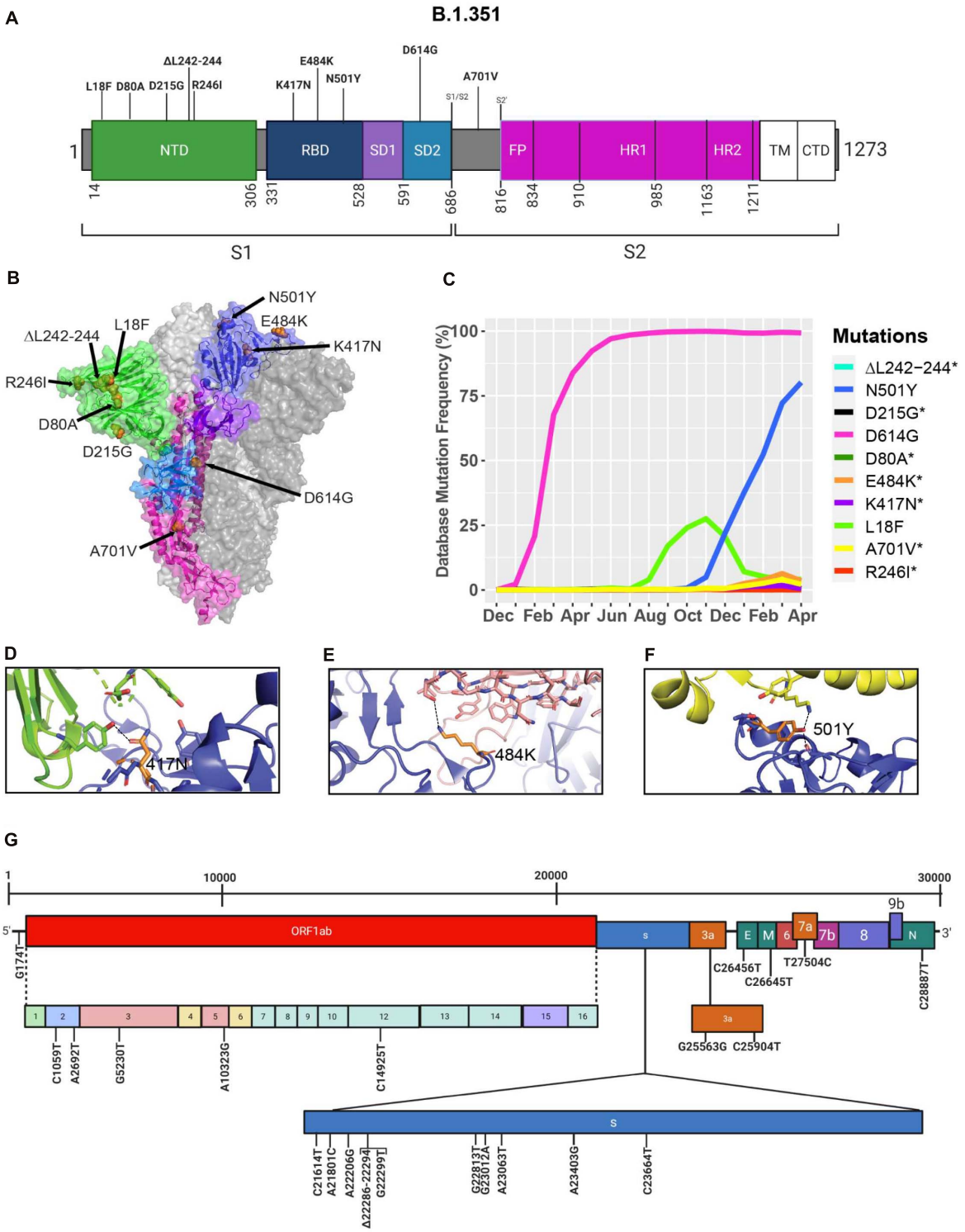


FIGURE 7 | Analysis of mutations in the B.1.351 variant. **(A)** Mutation map of the spike protein of B.1.351. **(B)** Structural representation of spike with ACE2. B.1.352 S protein mutations are presented in orange. NTD (green), RBD (blue), SD1 (purple), SD2 (light blue), and S2 (magenta) are illustrated. The other S protein monomers are illustrated in gray and white. **(C)** Frequency of the mutations in the S protein B.1.351 variant from December 2019 to April 30, 2021. Interaction of **(D)** 417N with C102 Nab (green), **(E)** 484K with C121 Nab (light pink), and **(F)** 501Y with hACE2 (yellow). The mutant residues are illustrated in orange, and the dashed lines represent interactions with adjacent residues. **(G)** Genome of the SARS-CoV-2 B.1.351 variant with identified nucleotide substitutions or deletions. Graphs were generated using Biorender, PyMOL, and RStudio. *Overlapping curves.

the localization of mutations with available Cryo-EM structures of the S protein and assessed the frequency of the variant by interrogating the GISAID database. There are nine mutations in the S protein out of the total 24 mutations in the B.1.1.7 SARS-CoV-2 genome (**Figures 5A,E**). Mutations in the S protein of the B.1.1.7 variant, except D614G, emerged in October of 2020 and reached a worldwide frequency of 70% to 80% in April 2021 (**Figures 4**, **5C** and **Table 2**). N501Y, found in the RBD, can interact with K353 in hACE2 (**Figures 5B,D**, **6C**). The N501Y mutation is associated with an increased affinity to hACE2, along with an increase in infectivity and virulence (**Table 2**). **Figure 5E** illustrates the whole genome of SARS-CoV-2 with all nucleotide substitutions of the B.1.1.7 variant. Substitutions C913T, C5986T, C14676T, C15279T, C16176T in ORF1ab, and T26801C in M protein are all synonymous mutations. Also, the C27972T mutation has a frequency of 75% and produces a premature stop codon in NS8 that inactivates the protein (Q27stop) without obvious consequences (**Figure 5E** and **Table 2**) (Pereira, 2020). These results allow us to better understand the frequencies, localization, and interactions of mutations in the S protein of the B.1.1.7 variant. For viruses, both synonymous and non-synonymous mutations are of importance. Synonymous mutations that do not change the amino acid sequence of proteins can still exercise very important functions. They can play a key role in docking sites for RNA-binding proteins, transcription factors, and primers, or partake in important secondary structures of the viral RNA like functional loops and folds.

The Emergence of the B.1.351 Variant in South Africa

During the spread of the B.1.1.7 variant in the United Kingdom, another variant was emerging in South Africa, known as B.1.351 (O'Toole et al., 2021; Tegally et al., 2021). The GISAID database was used to identify mutations and Cryo-EM structures of the S protein to model the effects of point mutations (**Figure 7**). Most of the mutations in the S protein of the B.1.351 variant are localized in the S1 subunit, with only A701V in the S2 subunit. Additionally, three mutations reside in the RBD,

TABLE 3 | Synonymous, non-synonymous mutations and deletions in the B.1.351 (South Africa) variant.

Variant	Protein (gene)	Genome nucleotide mutation Tegally et al., 2021	S or NS	AA mutation	Domain	Frequency* (%)	Effect
B.1.351	*5′-UTR*	G174T	S			N/A	N/A
	NSP2	C1059T	NS	T85I		13.95	N/A
	(ORF1ab)	A2692T	S			N/A	N/A
	NSP3 (ORF1ab)	G5230T	NS	K837N		1.41	N/A
	NSP5 (ORF1ab)	A10323G	NS	K90R		3.07	N/A
	NSP12 (ORF1ab)	C14925T	S			N/A	N/A
	S	C21614T	NS	L18F	NTD	4.11	N/A
		A21801C	NS	D80A	NTD	0.50	N/A
		A22206G	NS	D215G	NTD	0.51	N/A
		22286-22294del	NS	ΔL242/ΔA243/ ΔL244	NTD	0.51	Antibody escape Weisblum et al., 2020
		G22299T	NS	R246I	NTD	0.004	Antibody escape Weisblum et al., 2020
		G22813T	NS	K417N	RBD	0.51	Antibody escape Wang et al., 2021
		G23012A	NS	E484K	RBD	3.51	Antibody escape Barnes et al., 2020; Covid-19 Genomic UK Consortium, 2021
		A23063T	NS	N501Y	RBD	80.16	Increases affinity for hACE2 Wu et al., 2017
		A23403G	NS	D614G	SD2	99.29	Moderate increase in transmissbility Li Q. et al., 2020; Covid-19 Genomic UK Consortium, 2021
		G23664T	NS	A701V	S1/S2 – S2′	2.15	N/A
	NS3	G25563T	NS	Q57H		14.36	N/A
	(ORF3a)	C25904T	NS	S171L		1.77	N/A
	E	C26456T	NS	P71L		1.42	N/A
	M	C26645T	S				N/A
	NS7a (ORF7a)	T27504C	S				N/A
	N	C28887T	NS	T205I		7.00	N/A

**Mutation frequency as of April 30, 2021.*

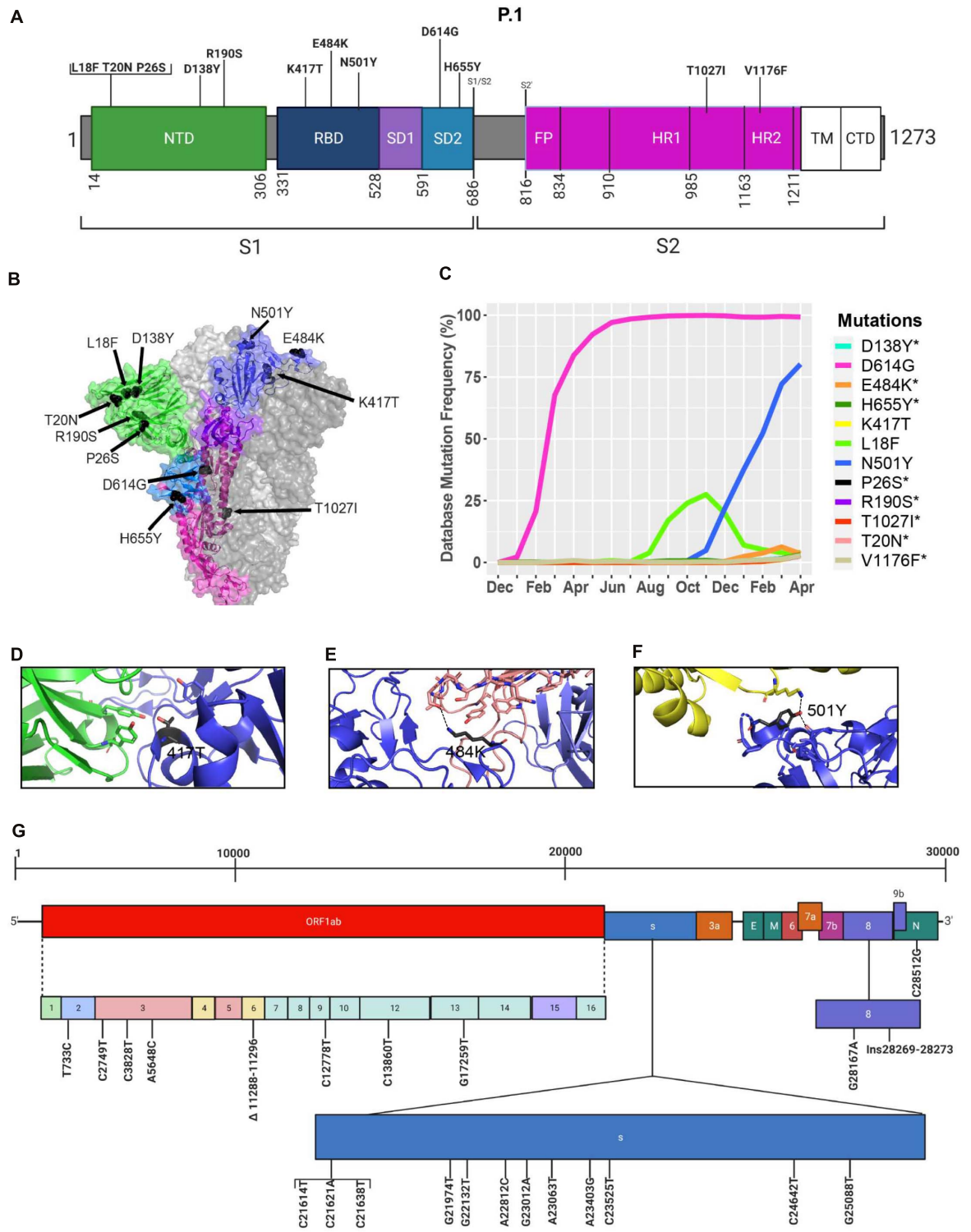


FIGURE 8 | Analysis of mutations in the P.1 variant. **(A)** Mutation map of the spike protein of P.1. **(B)** Structural representation of spike with ACE2. P.1 S protein mutations are presented in black. NTD (green), RBD (blue), SD1 (purple), SD2 (light blue), and S2 (magenta) are illustrated. The other S protein monomers are illustrated in gray and white. **(C)** Frequency of P.1 variant S protein mutations from December 2019 to April 30, 2021. Interaction of **(D)** 417T with C102 Nab (green), **(E)** 484K with C121 Nab (light pink), and **(F)** 501Y with hACE2 (yellow). The mutations are colored in black and interaction with adjacent residues are demonstrated by dashed lines. **(G)** Genome of the SARS-CoV-2 P.1 variant with identified nucleotides substitution, deletions, and insertions. Figures were generated using Biorender, PyMOL, and RStudio. *Overlapping curves.

among which two of them are not found in the B.1.1.7 variant (K417N, E484K) (**Figures 7A,B**). However, this variant contains D614G and N501Y, which are also seen in the B.1.1.7 variant (**Figure 5A**). Furthermore, many of the mutations found in the B.1.351 variant have worldwide frequencies lower than 5%. The exception is D614G, L18F, and N501Y (**Figure 7C** and **Table 3**). In comparison to B.1.1.7 and P.1 variants, the B.1.351 variant has not reached a frequency higher than 2% as of April 2021 (**Figure 4**). By using the mutagenesis tool of PyMOL, we modeled the S protein in complex with hACE2, and with C103 and C121 Nabs, which are human recombinant class I and II neutralizing antibodies, respectively (Barnes et al., 2020). Our *in silico* mutagenesis modeling predicts that B.1.351 mutations in the RBD favor a loss of interactions with C102 and C121 Nabs. Diminished interactions between RBD and C102 are predicted when the K417 is mutated to Asn producing Nabs escape capabilities (**Figures 7D**, **6A**). Diminished interactions with RBD are also predicted with C121 when the E484K mutation is present

TABLE 4 | Synonymous, non-synonymous, and deletions in the P.1 (Brazil) variant.

Variant	Protein (gene)	Genome nucleotide mutation Faria et al., 2021	S or NS	AA mutation	Domain	Frequency* (%)	Effect
P.1	*NSP2*	T733C	S			N/A	N/A
	(ORF1ab)	C2749T	S			N/A	N/A
	NSP3	C3828T	NS	S370L		3.09	N/A
	(ORF1ab)	A5648C	NS	K977Q		3.00	N/A
	NSP6 (ORF1ab)	11288-11296del	NS	ΔS106/ ΔG107/ΔF108		87.61	N/A
	NSP9 (ORF1ab)	C12778T	S			N/A	N/A
	NSP12 (ORF1ab)	C13860T	S			N/A	N/A
	NSP13 (ORF1ab)	G17259T	NS	E341D		2.97	N/A
	S	C21614T	NS	L18F	NTD	4.11	N/A
		C21621A	NS	T20N	NTD	2.77	N/A
		C21638T	NS	P26S	NTD	3.01	N/A
		G21974T	NS	D138Y	NTD	2.95	N/A
		G22132T	NS	R190S	NTD	2.71	N/A
		A22812C	NS	K417T	RBD	2.88	Antibody escape Garcia-Beltran et al., 2021
		G23012A	NS	E484K	RBD	3.51	Antibody escape Barnes et al., 2020; Covid-19 Genomic UK Consortium, 2021
		A23063T	NS	N501Y	RBD	80.16	Increase affinity for hACE2 Covid-19 Genomic UK Consortium, 2021
		A23403G	NS	D614G	SD2	99.29	Moderate increase in transmissibility Li Q. et al., 2020; Covid-19 Genomic UK Consortium, 2021
		C23525T	NS	H655Y	SD2	3.16	N/A
		C24642T	NS	T1027I	S2	2.79	N/A
		G25088T	NS	V1176F	S2	2.79	N/A
	NS8	G28167A	NS	E92K		2.96	N/A
	(ORF8)	Ins28269-28273	S			N/A	N/A
	N	C28512G	NS	P80R		2.94	N/A

**Mutation frequency as of April 30, 2021.*

TABLE 5 | Efficacy of vaccines against SARS-CoV-2 variants.

Vaccine	hCoV-19/Wuhan/WIV-4/2019	B.1.1.7	B.1.351	P.1
Pfizer-BioNTech	95.0% (Polack et al., 2020)	95.3% (Haas et al., 2021; Abu-Raddad et al., 2021)	75.0% (Abu-Raddad et al., 2021)	N/A
Moderna	94.1% (Baden et al., 2020)	N/A	N/A	N/A
Oxford-AstraZeneca	70.4% (Voysey et al., 2021)	70.4% (Emary et al., 2021)	10% (Madhi et al., 2021)	N/A
Johnson & Johnson	66% (Sadoff et al., 2021)	N/A	52.0% and 64.0% (14 and 28 days) (Sadoff et al., 2021)	66% (Sadoff et al., 2021)
Novavax	96% (Shinde et al., 2021)	86% (Shinde et al., 2021)	51% (Shinde et al., 2021)	

(**Figures 7E**, **6B**). As previously mentioned, N501Y is also found in B.1.351, and our modeling predicts that it will have similar effects as those observed with the B.1.1.7 variant (**Figure 7F**). **Figure 7G** illustrates the position of the nucleotide substitutions of the B.1.351 variant.

The Emergence of the P.1 Variant in Brazil

Similar to the B.1.351 variant, the P.1 variant harbors the N501Y and E484K mutations, but position 417 of the S protein displays a threonine (T) instead of a lysine (K) residue (Faria et al., 2021; **Figure 8**). Similar to the United Kingdom and South African variants, we present in **Figure 8C** mutations in the S protein that are characteristic of the P.1 variant and their frequencies from December 2019 to April 30, 2021. The P.1 variant harbors amino acid substitutions L18F, T20N, P26S, D138Y, and R190S in the NTD of the S protein. H655Y and T1027I, V1176F are in the subdomain 2 (SD2) of S1 and in the S2 subunit, respectively (**Figures 8A, B** and **Table 4**). The V1176F mutation is not shown because the structure is unresolved in this region. Overall, P.1-specific mutations have currently worldwide frequencies less than 5% (**Figure 8C** and **Table 4**). D614G, L18F, and N501Y are not specific to P1. Similar to B.1.351, the P.1 variant has a low frequency representing less than 5% of global variants as of April 2021 (**Figure 4**). Here, we also modeled P.1 mutations to investigate interaction alterations with known recombinant neutralizing antibodies (Barnes et al., 2020). The K417T mutation reduces interactions with neighboring residues in the C102 Nab, and therefore, the model predicts, as with K417N in B.1.1.7, an increased ability to escape neutralization (**Figure 8D**). Also, the P.1 variant has E484K and N501Y mutations in the RBD. Modeling predicts that they will have the same effect reported for the B.1.1.7 and B.1.351 variants (**Figures 8E,F**). In **Figure 8G**, we indicate the positions of synonymous, non-synonymous, and deletions in the SARS-CoV-2 P.1 genome.

DISCUSSION

The emergence of new genetic variants that are more transmissible, virulent, and resistant to antibody neutralization has highlighted the importance of studying the function of mutations in the viral genome. The number of sequenced viral genomes uploaded to the GISAID database grew rapidly from 131,417 at the end of September 2020 to 451,913 by January 30, 2021 (Shu and McCauley, 2017). GISAID is a formidable tool for tracking the emergence of mutations, identifying the geographic region where it emerged, and tracking its spread around the globe. Given that the risk mutations and new variants pose for neutralizing antibody therapy and vaccines for SARS-CoV-2, it is crucial to continuously monitor the susceptibility of these variants to neutralization by humoral and cellular immune responses either induced through natural exposure to the reference strain or induced by vaccination (Barnes et al., 2020; Greaney et al., 2021; Wang et al., 2021). Recent reports on the efficacy of the various vaccines have shown that these are variable and often diminished against variants (**Tables 5**, **6**).

TABLE 6 | Resistance of SARS-CoV-2 variants to vaccine-induced antibody neutralization.

Vaccine	Fold reduction in neutralization		
	B.1.1.7	**B.1.351**	**P.1**
Pfizer-BioNTech	[a]2.1 (Garcia-Beltran et al., 2021)	[a](34.5–42.4) (Garcia-Beltran et al., 2021)	[a]6.7 (Garcia-Beltran et al., 2021)
Moderna	[a]2.3 (Garcia-Beltran et al., 2021)	[a](19.2–27.7)(Garcia-Beltran et al., 2021)	[a]4.5 (Garcia-Beltran et al., 2021)
Oxford-AstraZeneca	[b]2.5 (Supasa et al., 2021)	[b]9.0 (Zhou et al., 2021)	[b]2.6 (Dejnirattisai et al., 2021)
Johnson & Johnson	N/A	N/A	N/A
Novavax	[a](0.85 to 20) (Shen et al., 2021b)	[a]13.1 (Shen et al., 2021a)	N/A

[a]*Pseudotyped virus neutralization assay.*
[b]*Live virus neutralization assay.*

Most available vaccines generally remain near fully efficacious against the B.1.1.7 variant (Emary et al., 2021; Haas et al., 2021; Wu et al., 2021). However, the Oxford-AstraZeneca vaccine has displayed compromised efficacy against the B.1.351 variant with only 10% vaccine efficacy (Madhi et al., 2021). Preliminary data with the Pfizer-BioNTech and Moderna mRNA vaccines also show a reduction in efficacy against B.1.351 (Abu-Raddad et al., 2021; Liu et al., 2021; Wu et al., 2021). Furthermore, antibodies induced by the Pfizer-BioNTech and Moderna vaccines appear to display a 6.7-fold and 4.5-fold decrease in neutralization efficacy against the P.1 variant (**Table 6**; Garcia-Beltran et al., 2021). The recently approved Johnson & Johnson adenovirus-based vaccine only requires a single dose, in comparison with the two for the mRNA vaccines and Oxford-AstraZeneca vaccine, and has an efficacy of 66% against the original Wuhan reference strain, 52.0% to 64.0% against the B.1.351 variant and 66% against the P.1 variant (Sadoff et al., 2021). The Novavax vaccine also shows reduced efficacy against the variants with 86% and 51% against the B.1.1.7 and P.1 variants, respectively. Nevertheless, humoral responses are only one component of the adaptive immune response. T-cell responses have not been studied in detail against these variants at this time and still provide robust contribution to overall protection against severe COVID-19 disease.

Studying the effect of individual mutations in a viral protein may not always reflect their true impact on virus features if they appear in combination with other mutations. Epistasis is the combinatory effect of two or more mutations in a genome (Phillips, 2008). Epistasis has previously been intensely studied in the surface protein hemagglutinin (HA) of influenza viruses, and positive epistasis was identified in several regions of the HA receptor-binding domain (Wu et al., 2017). In relation to the S protein of SARS-CoV-2, epistatic mutations could, for instance, allow the S protein to adopt a specific conformation when all the mutations are present, thereby producing a unique folding of the protein and enabling new features. A recent study demonstrated the impact of antigenicity and infectivity of D614G SARS-CoV-2 variants with a combination of different mutations occurring in

the S protein (Li Q. et al., 2020). The study shows that D614G alone only mildly increases infectivity, but in combination with different other mutations in the S protein, together, these can either more intensely increase or decrease viral infectivity. Similar findings have been reported regarding sensitivity to Nabs. D614G alone has undetectable effects on Nabs escape. However, the combination of D614G with other mutations in S can enable Nabs escape (Li Q. et al., 2020). This data suggests that the continuous emergence of epistatic mutations in SARS-CoV-2 will likely be involved in further altering the properties of the virus, including transmissibility, pathogenicity, stability, and Nab resistance.

Our analyses have highlighted that several new fast-spreading mutations had frequency trajectories that eventually plunged or stabilized at low frequencies. Only the D614G in the S protein and the P323L in the RdRp have maintained their presence in the consensus sequence (**Figures 1**, **2**). Analyses of the GISAID database also reveal which countries upload the most sequences to the database and are therefore carrying out the most testing and sequencing. The global frequency of mutations and variants in the database is therefore biased to represent the genetic landscape of the virus in the countries doing the most testing, sequencing, and data sharing. Emergence of new variants may therefore go undetected until they leave their point of origin and enter countries with high testing and sequencing rates. This delayed notification constitutes a major obstacle in identifying and preventing the spread of nefarious variants that are potentially resistant to current vaccines and neutralizing antibody therapy. Another caveat in analyzing sequences in the GISAID database is that consensus sequences are uploaded, but subsequences and quasispecies are generally not included. Therefore, it is possible that mutations in sub-variants may be missed that could contribute to virus replication, pathogenicity, and spread (Rocheleau et al., 2021). A global approach to analyzing both transmitted variants and non-transmitted sub-variants and quasispecies could provide a better understanding of the effects of SARS-CoV-2 mutations.

In conclusion, our metadata analysis of emerging mutations has highlighted the natural fluctuations in mutation prevalence. We also illustrate how mutations sometimes need to co-emerge in order to create a favorable outcome for virus propagation. Tracking mutations and the evolution of the SARS-CoV-2 genome is critical for the development and deployment of effective treatments and vaccines. Thus, it is the responsibility of all countries and governing jurisdictions to increase testing and sequencing, and uploading SARS-CoV-2 genomes to open access databases in real-time. This will result in more accurate information to inform policy and decision makers about interventions required to blunt the global transmission of the virus and ensure that vaccines remain effective against all circulating variants.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.gisaid.org.

AUTHOR CONTRIBUTIONS

NC performed the analysis and drafted the manuscript. WZ contributed to study design, provided guidance, and drafted the manuscript. M-AL designed the study, provided guidance, and drafted the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by a COVID-19 Rapid Response grant to M-AL by the Canadian Institute of Health Research (CIHR) and by a grant supplement by the Canadian Immunity Task Force (CITF).

ACKNOWLEDGMENTS

The authors wish to thank Dr. Sean Li at Health Canada for helpful comments on our manuscript. M-AL holds a Canada Research Chair in Molecular Virology and Intrinsic Immunity.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.676314/full#supplementary-material

REFERENCES

Abu-Raddad, L. J., Chemaitelly, H., and Butt, A. A. (2021). Effectiveness of the BNT162b2 Covid-19 vaccine against the B.1.1.7 and B.1.351 variants. *N. Engl. J. Med.* doi: 10.1056/NEJMc2104974

Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., et al. (2020). Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N. Engl. J. Med.* 384, 403–416.

Barnes, C. O., Jette, C. A., Abernathy, M. E., Dam, K.-M. A., Esswein, S. R., Gristick, H. B., et al. (2020). SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature* 588, 682–687.

Benton, D. J., Wrobel, A. G., Xu, P., Roustan, C., Martin, S. R., Rosenthal, P. B., et al. (2020). Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. *Nature* 588, 327–330. doi: 10.1038/s41586-020-2772-0

Chu, H., Chan, J. F.-W., Yuen, T. T.-T., Shuai, H., Yuan, S., Wang, Y., et al. (2020). Comparative tropism, replication kinetics, and cell damage profiling of SARS-CoV-2 and SARS-CoV with implications for clinical manifestations, transmissibility, and laboratory studies of COVID-19: an observational study. *Lancet Microbe* 1, e14–e23.

Covid-19 Genomic UK Consortium (2021). *COG-UK Report on SARS-CoV-2 Spike Mutations of Interest in the UK*. London: Covid-19 Genomic UK Consortium Available online at: https://www.cogconsortium.uk/wp-content/uploads/2021/01/Report-2_COG-UK_SARS-CoV-2-Mutations.pdf (accessed May 13, 2021).

Dejnirattisai, W., Zhou, D., Supasa, P., Liu, C., Mentzer, A. J., Ginn, H. M., et al. (2021). Antibody evasion by the Brazilian P.1 strain of SARS-CoV-2. *bioRxiv* [preprint] doi: 10.1101/2021.03.12.435194

Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* 20, 533–534. doi: 10.1016/s1473-3099(20)30120-1

Drake, J. W. (1993). Rates of spontaneous mutation among RNA viruses. *Proc. Natl. Acad. Sci. U.S.A.* 90, 4171. doi: 10.1073/pnas.90.9.4171

Duffy, S. (2018). Why are RNA virus mutation rates so damn high? *PLoS Biol.* 16:e3000003. doi: 10.1371/journal.pbio.3000003

Eckerle, L. D., Becker, M. M., Halpin, R. A., Li, K., Venter, E., Lu, X., et al. (2010). Infidelity of SARS-CoV Nsp14-Exonuclease mutant virus replication is revealed by complete genome sequencing. *PLoS Pathog.* 6:e1000896. doi: 10.1371/journal.ppat.1000896

Emary, K. R. W., Golubchik, T., Aley, P. K., Ariani, C. V., Angus, B., Bibi, S., et al. (2021). Efficacy of ChAdOx1 nCoV-19 (AZD1222) vaccine against SARS-CoV-2 variant of concern 202012/01 (B.1.1.7): an exploratory analysis of a randomised controlled trial. *Lancet* 397, 1351–1362.

Faria, N. R., Claro, I. M., Candido, D., Moyses Franco, L. A., Andrade, P. S., Coletti, T. M., et al. (2021). *Genomic Characterisation of an Emergent SARS-CoV-2 Lineage in Manaus: Preliminary Findings. Virological*. Available online at: https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586 (accessed January 23, 2021).

Garcia-Beltran, W. F., Lam, E. C., St. Denis, K., Nitido, A. D., Garcia, Z. H., Hauser, B. M., et al. (2021). Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell* 184, 2372–2383.e9. doi: 10.1016/j.cell.2021.03.013

Greaney, A. J., Starr, T. N., Gilchuk, P., Zost, S. J., Binshtein, E., Loes, A. N., et al. (2021). Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. *Cell Host Microbe* 29, 44–57.e9.

Gu, H., Chen, Q., Yang, G., He, L., Fan, H., Deng, Y.-Q., et al. (2020). Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science* 369:1603.

Haas, E. J., Angulo, F. J., McLaughlin, J. M., Anis, E., Singer, S. R., Khan, F., et al. (2021). Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. *Lancet* 397, 1819–1829. doi: 10.1016/S0140-6736(21)00947-8

Kannan, S. R., Spratt, A. N., Quinn, T. P., Heng, X., Lorson, C. L., Sönnerborg, A., et al. (2020). Infectivity of SARS-CoV-2: there is something more than D614G? *J. Neuroimmune Pharmacol.* 15, 574–577. doi: 10.1007/s11481-020-09954-3

Kim, D., Lee, J.-Y., Yang, J.-S., Kim, J. W., Kim, V. N., and Chang, H. (2020). The architecture of SARS-CoV-2 transcriptome. *Cell* 181, 914–921.e10.

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., et al. (2020). Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* 182, 812–827.e19.

Li, Q., Wu, J., Nie, J., Zhang, L., Hao, H., Liu, S., et al. (2020). The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell* 182, 1284–1294.e9.

Li, X., Giorgi, E. E., Marichannegowda, M. H., Foley, B., Xiao, C., Kong, X.-P., et al. (2020). Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci. Adv.* 6:eabb9153. doi: 10.1126/sciadv.abb9153

Liu, Y., Liu, J., Xia, H., Zhang, X., Fontes-Garfias, C. R., Swanson, K. A., et al. (2021). Neutralizing activity of BNT162b2-elicited serum — preliminary report. *N. Engl. J. Med.* 384, 1466–1468. doi: 10.1056/NEJMc2102017

Ma, Y., Wu, L., Shaw, N., Gao, Y., Wang, J., Sun, Y., et al. (2015). Structural basis and functional analysis of the SARS coronavirus nsp14–nsp10 complex. *Proc. Natl. Acad. Sci. U.S.A.* 112:9436. doi: 10.1073/pnas.1508686112

Madhi, S. A., Baillie, V., Cutland, C. L., Voysey, M., Koen, A. L., Fairlie, L., et al. (2021). Efficacy of the ChAdOx1 nCoV-19 Covid-19 vaccine against the B.1.351 variant. *N. Engl. J. Med.* 384, 1885-1898. doi: 10.1056/NEJMoa2102214

Minskaia, E., Hertzig, T., Gorbalenya, A. E., Campanacci, V., Cambillau, C., Canard, B., et al. (2006). Discovery of an RNA virus 3′→5′ exoribonuclease that is critically involved in coronavirus RNA synthesis. *Proc. Natl. Acad. Sci. U.S.A.* 103:5108. doi: 10.1073/pnas.0508200103

O'Toole, A., Hill, V., Pybus, O. G., Watts, A., Bogoch, I. I., Khan, K., et al. (2021). *Tracking the international spread of SARS-CoV-2 lineage B.1.1.7 and B.1.351/501Y-V2*. Available online at: https://virological.org/t/tracking-the-international-spread-of-sars-cov-2-lineages-b-1-1-7-and-b-1-351-501y-v2/592 (accessed May 13, 2021).

Peck, K. M., and Lauring, A. S. (2018). Complexities of viral mutation rates. *J. Virol.* 92:e01031-17.

Pereira, F. (2020). Evolutionary dynamics of the SARS-CoV-2 ORF8 accessory gene. *Infect. Genet. Evol.* 85, 104525–104525. doi: 10.1016/j.meegid.2020.104525

Phillips, P. C. (2008). Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* 9, 855–867. doi: 10.1038/nrg2452

Plante, J. A., Liu, Y., Liu, J., Xia, H., Johnson, B. A., Lokugamage, K. G., et al. (2020). Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* 592, 116–121. doi: 10.1038/s41586-020-2895-3

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., et al. (2020). Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N. Engl. J. Med.* 383, 2603–2615.

Rambaut, A., Loman, N., Pybus, O., Barclay, W., Barrett, J., Carabelli, A., et al. (2020). *Preliminary Genomic Characterisation of an Emergent SARS-CoV-2 Lineage in the UK Defined by a Novel Set of Spike Mutations*. Available online at: https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563 (accessed May 13, 2021).

Rausch, J. W., Capoferri, A. A., Katusiime, M. G., Patro, S. C., and Kearney, M. F. (2020). Low genetic diversity may be an Achilles heel of SARS-CoV-2. *Proc. Natl. Acad. Sci. U.S.A.* 117:24614. doi: 10.1073/pnas.2017726117

Rocheleau, L., Laroche, G., Fu, K., Stewart, C. M., Mohamud, A. O., Côté, M., et al. (2021). Identification of a high-frequency intra-host SARS-CoV-2 spike variant with enhanced cytopathic and fusogenic effect. *bioRxiv* [preprint] doi: 10.1101/2020.12.03.409714

Sadoff, J., Gray, G., Vandebosch, A., Cárdenas, V., Shukarev, G., Grinsztejn, B., et al. (2021). Safety and efficacy of single-dose Ad26.COV2.S vaccine against Covid-19. *N. Engl. J. Med.* 384, 2187–2201. doi: 10.1056/NEJMoa2101544

Shen, X., Tang, H., McDanal, C., Wagh, K., Fischer, W., Theiler, J., et al. (2021a). SARS-CoV-2 variant B.1.1.7 is susceptible to neutralizing antibodies elicited by ancestral spike vaccines. *Cell Host Microbe* 29, 529–539.e3.

Shen, X., Tang, H., Pajon, R., Smith, G., Glenn, G. M., Shi, W., et al. (2021b). Neutralization of SARS-CoV-2 Variants B.1.429 and B.1.351. *N. Engl. J. Med.*

Shi, J., Wen, Z., Zhong, G., Yang, H., Wang, C., Huang, B., et al. (2020). Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS–coronavirus 2. *Science* 368:1016. doi: 10.1126/science.abb7015

Shinde, V., Bhikha, S., Hoosain, Z., Archary, M., Bhorat, Q., Fairlie, L., et al. (2021). Efficacy of NVX-CoV2373 Covid-19 vaccine against the B.1.351 variant. *N. Engl. J. Med.* 384, 1899-1909.

Shu, Y., and McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* 22:30494.

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H. D., Dingens, A. S., et al. (2020). Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* 182, 1295–1310.e20.

Supasa, P., Zhou, D., Dejnirattisai, W., Liu, C., Mentzer, A. J., Ginn, H. M., et al. (2021). Reduced neutralization of SARS-CoV-2 B.1.1.7 variant by convalescent and vaccine sera. *Cell* 184, 2201–2211.e7.

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., et al. (2021). Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* 592, 438–443.

Ugurel, O. M., Mutlu, O., Sariyer, E., Kocer, S., Ugurel, E., Inci, T. G., et al. (2020). Evaluation of the potency of FDA-approved drugs on wild type and mutant SARS-CoV-2 helicase (Nsp13). *Int. J. Biol. Macromol.* 163, 1687–1696. doi: 10.1016/j.ijbiomac.2020.09.138

Voysey, M., Clemens, S. A. C., Madhi, S. A., Weckx, L. Y., Folegatti, P. M., Aley, P. K., et al. (2021). Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *Lancet* 397, 99–111.

Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., McGuire, A. T., and Veesler, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* 181, 281–292.e6.

Wang, P., Liu, L., Iketani, S., Luo, Y., Guo, Y., Wang, M., et al. (2021). Increased resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7 to antibody neutralization. *bioRxiv* [preprint] doi: 10.1101/2021.01.25.428137

Weisblum, Y., Schmidt, F., Zhang, F., DaSilva, J., Poston, D., Lorenzi, J. C., et al. (2020). Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *ELife* 9:e61312.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269.

Wu, K., Werner, A. P., Koch, M., Choi, A., Narayanan, E., Stewart-Jones, G. B. E., et al. (2021). Serum neutralizing activity elicited by mRNA-1273 vaccine — Preliminary Report. *N. Engl. J. Med.* 384, 1468–1470. doi: 10.1056/NEJMc2102179

Wu, N. C., Xie, J., Zheng, T., Nycholat, C. M., Grande, G., Paulson, J. C., et al. (2017). Diversity of functionally permissive sequences in the receptor-binding site of influenza hemagglutinin. *Cell Host Microbe* 21, 742–753.e8.

Wu, S., Tian, C., Liu, P., Guo, D., Zheng, W., Huang, X., et al. (2020). Effects of SARS-CoV-2 mutations on protein structures and intraviral protein–protein interactions. *J. Med. Virol.* 93, 2132–2140. doi: 10.1002/jmv.26597

Yi, C., Sun, X., Ye, J., Ding, L., Liu, M., Yang, Z., et al. (2020). Key residues of the receptor binding motif in the spike protein of SARS-CoV-2 that interact with ACE2 and neutralizing antibodies. *Cell. Mol. Immunol.* 17, 621–630. doi: 10.1038/s41423-020-0458-z

Zahradník, J., Marciano, S., Shemesh, M., Zoler, E., Chiaravalli, J., Meyer, B., et al. (2021). SARS-CoV-2 RBD in vitro evolution follows contagious mutation spread, yet generates an able infection inhibitor. *bioRxiv* [preprint] doi: 10.1101/2021.01.06.425392

Zhang, L., Jackson, C. B., Mou, H., Ojha, A., Peng, H., Quinlan, B. D., et al. (2020). SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat. Commun.* 11:6013.

Zhou, D., Dejnirattisai, W., Supasa, P., Liu, C., Mentzer, A. J., Ginn, H. M., et al. (2021). Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. *Cell* 184, 2348–2361.e6.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* 382, 727–733.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.