RESEARCH ARTICLE

Evolutionary Tracking of SARS-CoV-2 Genetic Variants Highlights an Intricate Balance of Stabilizing and Destabilizing Mutations

Jobin John Jacob,[a] Karthick Vasudevan,[a,d] Agila Kumari Pragasam,[a] Karthik Gunasekaran,[b] Balaji Veeraraghavan,[a] Ankur Mutreja[c]

[a]Department of Clinical Microbiology, Christian Medical College, Vellore, India

[b]Department of General Medicine (Unit-V), Christian Medical College, Vellore, India

[c]Cambridge Institute of Therapeutic Immunology & Infectious Disease (CITIID) Department of Medicine, University of Cambridge, Cambridge, United Kingdom

[d]Department of Biotechnology, School of Applied Sciences, REVA University, Bangalore, India

Jobin John Jacob and Karthick Vasudevan contributed equally to the manuscript. Author order was determined alphabetically.

ABSTRACT The currently ongoing COVID-19 pandemic caused by SARS-CoV-2 has accounted for millions of infections and deaths across the globe. Genome sequences of SARS-CoV-2 are being published daily in public databases and the availability of these genome data sets has allowed unprecedented access to the mutational patterns of SARS-CoV-2 evolution. We made use of the same genomic information for conducting phylogenetic analysis and identifying lineage-specific mutations. The catalogued lineage-defining mutations were analyzed for their stabilizing or destabilizing impact on viral proteins. We recorded persistence of D614G, S477N, A222V, and V1176F variants and a global expansion of the PANGOLIN variant B.1. In addition, a retention of Q57H (B.1.X), R203K/G204R (B.1.1.X), T85I (B.1.2-B.1.3), G15S+T428I (C.X), and I120F (D.X) variations was observed. Overall, we recorded a striking balance between stabilizing and destabilizing mutations, therefore leading to well-maintained protein structures. With selection pressures in the form of newly developed vaccines and therapeutics to mount in the coming months, the task of mapping viral mutations and recording their impact on key viral proteins should be crucial to preemptively catch any escape mechanism for which SARS-CoV-2 may evolve.

IMPORTANCE Since its initial isolation in Wuhan, China, large numbers of SARS-CoV-2 genome sequences have been shared in publicly accessible repositories, thus enabling scientists to do detailed evolutionary analysis. We investigated the evolutionarily associated mutational diversity overlaid on the major phylogenetic lineages circulating globally, using 513 representative genomes. We detailed the phylogenetic persistence of key variants facilitating global expansion of the PANGOLIN variant B.1, including the recent, fast-expanding, B.1.1.7 lineage. The stabilizing or destabilizing impact of the catalogued lineage-defining mutations on viral proteins indicates their possible involvement in balancing the protein function and structure. A clear understanding of this mutational profile is of high clinical significance to catch any vaccine escape mechanism, as the same proteins make crucial components of vaccines that have recently been approved or are in development. In this vein, our study provides an imperative framework and baseline data upon which further analysis could be built as newer variants of SARS-CoV-2 continue to appear.

KEYWORDS SARS-CoV-2, mutation, evolution, stability, vaccine

The emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Wuhan, China and the subsequent global spread has brought the world to a

Citation Jacob JJ, Vasudevan K, Pragasam AK, Gunasekaran K, Veeraraghavan B, Mutreja A. 2021. Evolutionary tracking of SARS-CoV-2 genetic variants highlights an intricate balance of stabilizing and destabilizing mutations. mBio 12:e01188-21. https://doi.org/10.1128/mBio.01188-21.

Invited Editor Maria Kalamvoki, University of Kansas Medical Center

Editor Indranil Biswas, KUMC

Address correspondence to Balaji Veeraraghavan, vbalaji@cmcvellore.ac.in, or Ankur Mutreja, am872@medschl.cam.ac.uk.

Received 6 May 2021
Accepted 14 June 2021
Published 20 July 2021

standstill (1). During the course of 11 months, the coronavirus disease 19 (COVID-19) pandemic has caused more than 81 million confirmed cases in 220 countries, with close to 1,770,000 fatalities. (2). Initially, and rightly, the efforts were focused on minimizing the number of cases and deaths due to COVID-19 (3). This included fast tracking the search and development of novel treatment and prevention options (4). Today, however, as vaccine candidates have started showing promising results, there is a cautious shift towards assessing the efficacy of vaccine candidates with respect to the circulating diversity of SARS-CoV-2 and its continuously evolving genetic variants (5).

Functional mutations that help the virus to adapt to the recent host-shift events are hypothesized to drive the evolution of transmissibility and virulence in SARS-CoV-2 (6). Shortly after the first isolated SARS-CoV-2 genome from China was published, >30,500 distinct mutations were catalogued in the CoV-GLUE database (http://cov-glue.cvr.gla.ac.uk/) among globally circulating strains of this virus (7). Variations in the genetic makeup are key determinants in measuring the evolutionary distance and stability of SARS-CoV-2 from the first sequenced isolate (8). Moreover, tracking the evolution of SARS-CoV-2 since its introduction in humans is a high-priority undertaking to prevent future waves of this pandemic from escaping the global preparedness (9). Since many vaccine candidates currently under development are derived from the first available SARS-CoV-2 sequences, recurrent genetic changes may have an unforeseen impact on their sustained effectiveness in the longer term (10).

The availability of whole-genome sequences of SARS-CoV-2 in public repositories such as Global Initiative on Sharing All Influenza Data (GISAID) and real-time data visualization pipeline NextStrain (https://nextstrain.org) offers a great opportunity for scientists to track the evolutionary path of this virus (11, 12). Phylogenetic Assignment of Named Global Outbreak LINeages tool (PANGOLIN) has been the most widely used tool for lineage assignment to newly emerging variants. PANGOLIN (https://cov-lineages.org/pangolin.html) has also been deployed in establishing the transmission patterns of various clones of this virus (13). Since coronaviruses frequently recombine, tracking the evolution and assigning lineages has been challenging (13, 14). As a result, multiple studies that tracked the evolution of SARS-CoV-2 have been hugely controversial. For example, doubts have been cast on the claim of finding more aggressive L type strains emerging from S type strains (14). Similarly, the hypothesis that rapid spread of the D614G variant of SARS-CoV-2 indicates a possible fitness advantage has been questioned (15–17). Therefore, in the current and highly sensitive global circumstances due to this pandemic, having a detailed map of mutations highlighting their prospective role in therapeutics and vaccine development can prepare us better for the future waves of continuously evolving SARS-CoV-2. In this study, we present a catalogue of the most important genomic mutations recorded between December 2019 and November 2020 in SARS-Cov-2 and their possible impact on the stability of protein candidates that form the most crucial part of vaccines and also constitute the most common therapeutic targets.

RESULTS

Diversity of SARS-CoV-2 genomes. Of the 7,000 SARS-CoV-2 genomes screened, we constructed a robust phylogenetic tree of 513 genomes strategically selected to reflect the most complete diversity among the isolates by covering all the PANGOLIN lineages. Lineage assignment based on the PANGOLIN tool indicated the circulation of seven distinct lineages and/or sublineages, such as A, B.1, B.1.1, B.1.1.1, B.2, B.3, and B.6. This is in line with the phylogenetic groupings by GISAID (S, L, V, O, G, GH, and GR) (Fig. 1). As the epidemic has progressed and mutations have accumulated, further subdivision of major lineages into sublineages has been observed. Overall, a total of 61 lineages and sublineages have been found to be circulating concurrently in multiple countries around the world. In general, numerous introductions of different variants were observed across the globe with a few sublineages (C.2, D.2) being restricted to

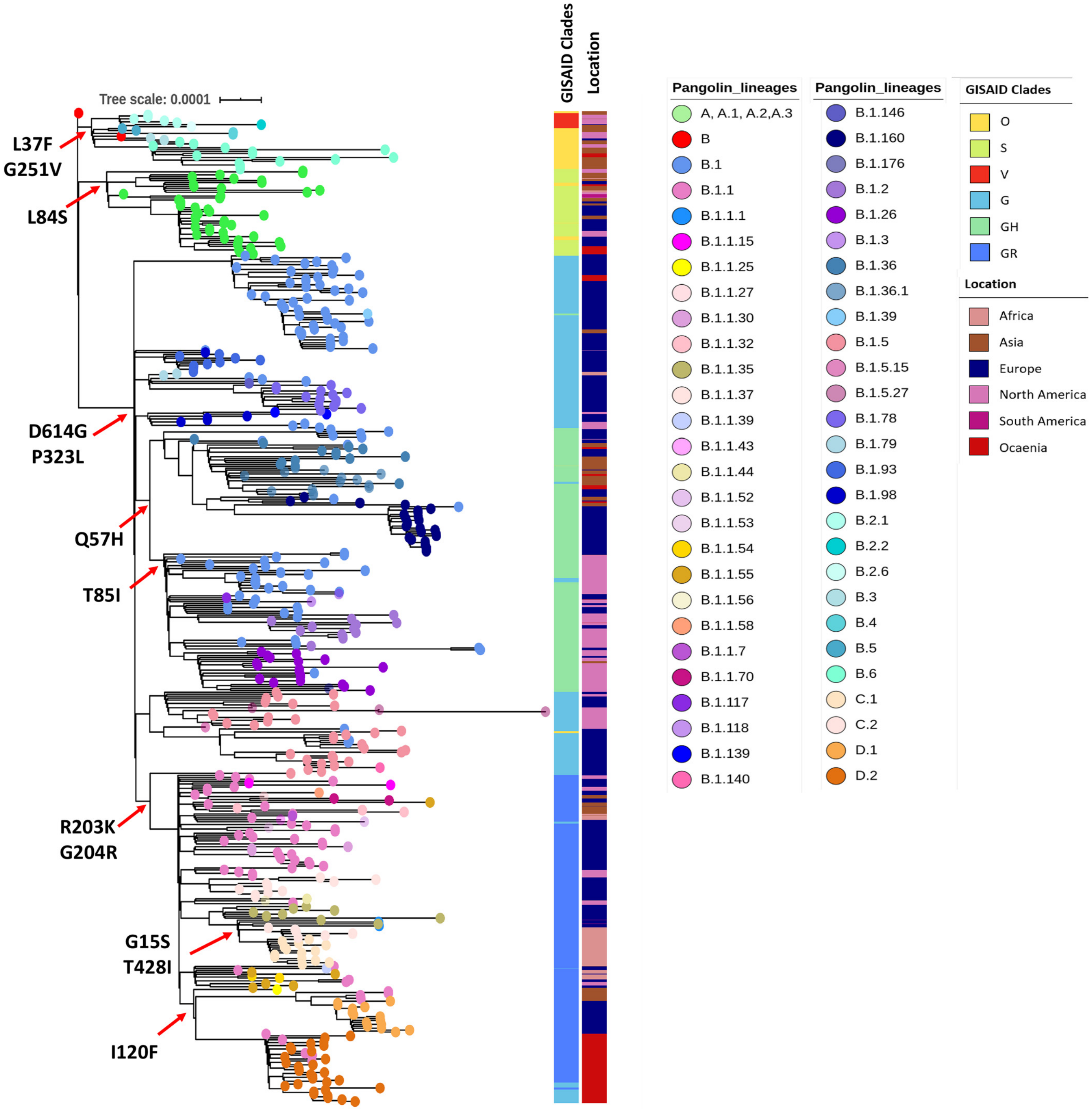

FIG 1 Maximum likelihood phylogenetic tree inferred from 513 SARS-CoV-2 genomes. The tree was constructed using multiple genome sequence alignment (MAFFT) by mapping against the Wuhan-Hu-1 strain (accession NC_045512). Tips are colored with the major lineages assigned by PANGOLIN. Respective lineages assigned by GISAID and origin of sequence are labeled as color strips. The scale bar indicates the distance corresponding to substitution per site.

certain regions. While the B.1.113 lineage, for example, has been exclusively reported from India, lineages C.2 and D.2 have been geographically confined to South Africa and Australia, respectively.

Major amino acid substitutions. Mutation mapping showed a total of 106 amino acid substitutions (missense mutations in >5 genomes) from a representative set of 513 genomes. The analysis also revealed 36 mutations that were found in >5% of genome sequences, while 12 major substitutions were lineage-defining mutations (Fig. 1). The first major mutation to appear was L84S in ORF8 (present in 8.6% of the

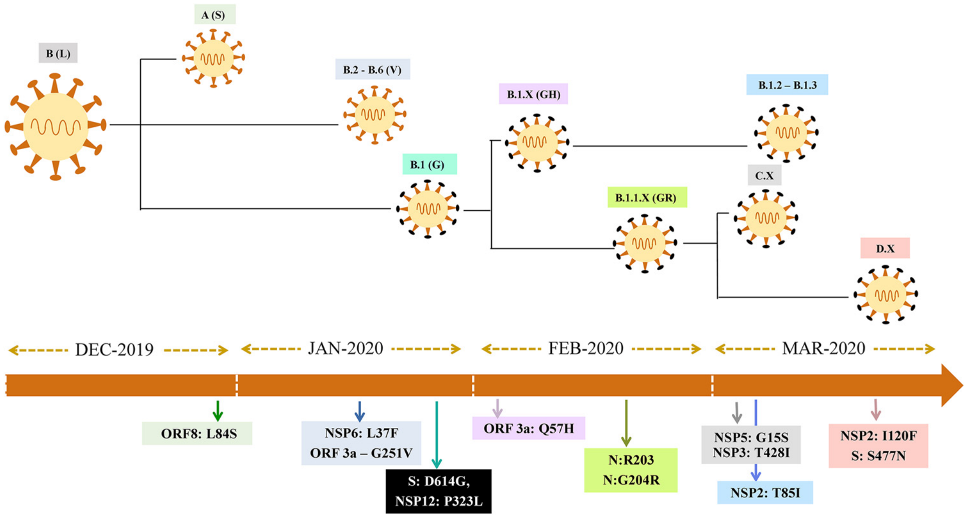

FIG 2 Schematic representation of the major evolutionary events/amino acid substitutions that gave rise to SARS-CoV-2 variants in sequential order.

genomes) that has defined the A lineage (i.e., clade S in the GISAID classification). The subsequent amino acid substitutions L37F in ORF3a and G251V in nsp6 were found to be present in 13.3% and 1.4% of genomes, respectively. The combination of G251V and L37F, which was initially considered a defining mutation pattern for the B.2 to B.6 lineage (clade V in GISAID classification), has shown under more detailed analysis that isolates carrying the G251V mutation are distributed in other lineages too. The predominant lineage-defining mutations in the whole data set were D614G (85.5%) and P323L (85.5%), after originally appearing in late January 2020 (Fig. 2). Other major mutations noted are Q57H (26.5%), R203K/G204R (33%), G15S (12%), I120F (11.5%), and T85I (14%).

Dominance of the D614G variant. Two mutations have become consensus: D614G in S (nucleotide 23,403, A to G) and P323L (also known as P4715L) in nsp12 (nucleotide 14,143, C to T). These mutations were present in 80.5% of the sequences and have defined the B.1 lineage (G in GISAID classification). The widely discussed D614G variant is speculated to have been introduced in Europe at the end of January (EPI_ISl_422424) before becoming globally dominant. Genomes with D614G mutations were assigned as B.1 by PANGOLIN or GH/GR by GISAID. Notably, founder lineage B.1 and its sublineages B.1.X, B.1.1.X, D.X, and C.X that carry both D614G and P323L mutations have become the dominant variants across the world (87% of global collection per CoV-GLUE as of 30 November 2020).

As the pandemic has progressed, several other major substitutions affecting the protein structure have appeared. These are Q57H (nucleotide 25,563, G to T) in ORF3a, the R203K + G204R combination (nucleotide 28,881, GGG to AAC) in nucleocapsid, and T85I (nucleotide 1,059, C to T) in ORF1a. The region-specific sub lineages C.1, C.2, D.1, and D.2 were found to cumulatively harbor multiple mutations. Amino acid substitutions such as T428I and G15S in ORF1a were reported in sublineages C.1 and C.2, and the S477N substitution in the spike (S) protein along with I120F in nsp2 specifically established the sublineage D.2 (Fig. 1).

Structural analysis of SARS-CoV-2 mutants. The possible structural consequences of 11 lineage-defining missense mutations identified in this study were investigated. Among the mutations, three were considered stabilizing to the respective protein structures, while six mutations were destabilizing (Table 1). The significance of these mutations in evolutionary selection cannot be solely predicted by ΔΔG, or change in free energy. Hence for a precise interpretation, correlation of ΔΔG, ΔΔS, and N-H S^2 (Table S2 in the supplemental material) order parameter values of the proteins have been taken into account based on fine local alterations in structures. All lineage-

TABLE 1 Lineage-defining SNPs and their impact on protein structures

Protein	Lineage-defining mutation	ΔΔS in kcal mol^{-1} K^{-1}	Change in dynamics	ΔΔG in kcal mol^{-1} (DUET)	ΔΔG in kcal mol^{-1} (SDM)	Stability
Nsp12	P323L	−0.33	Decreasing flexibility	0.43 (stabilizing)	1.57 (stabilizing)	Stabilizing
Spike	D614G	−0.01	Decreasing flexibility	0.46 (stabilizing)	2.33 (stabilizing)	Stabilizing
Orf3a	G251V	−0.39	Decreasing flexibility	−0.6 (destabilizing)	−2.19 (destabilizing)	Destabilizing
	Q57H	0.44	Increasing flexibility	−1.25 (destabilizing)	0.87 (stabilizing)	Inconclusive
Orf8	L84S	0.30	Increasing flexibility	−1.41 (destabilizing)	−1.41 (destabilizing)	Destabilizing
Nsp2	T85I	0.07	Increasing flexibility	0.54 (stabilizing)	1.93 (stabilizing)	Stabilizing
	I120F	−1.30	Decreasing flexibility	−1.04 (destabilizing)	−0.21 (destabilizing)	Destabilizing
Nsp6	L37F	−0.29	Decreasing flexibility	−0.72 (Destabilizing)	−0.04 (neutral)	Inconclusive
Nucleocapsid protein (N)	R203K	−0.98	Decreasing flexibility	−1.57 (destabilizing)	−0.48 (destabilizing)	Destabilizing
	G204R	−0.16	Decreasing flexibility	−1.06 (destabilizing)	−1.95 (destabilizing)	Destabilizing
Nsp5	G15S	−0.31	Decreasing flexibility	−0.98 (destabilizing)	−0.79 (destabilizing)	Destabilizing

defining mutations except two have reduced the vibrational entropies of the proteins, thereby decreasing the flexibility in the structures (Table 1).

Additionally, the impact of mutations in key structural proteins that potentially allows any pathogen to escape available treatment and prevention regime was investigated. Among the 59 major missense mutations, our analysis using both the SDM and DUET servers predicted 16 missense mutations as stabilizing and 23 missense mutations as destabilizing the protein structure. Twenty major mutations were predicted to be neither stabilizing nor destabilizing, as the ΔΔG values provided by the SDM and DUET servers were contradictory (Table 2).

Balance of stabilizing and destabilizing mutations. Overall, from both the data sets, 70 amino acid substitutions in SARS-CoV-2 were tested for stability, of which 19 were stabilizing, 29 were destabilizing, and 22 showed inconclusive results. Computational prediction to understand the effect of amino acid substitutions in SARS-CoV-2 revealed a balance of stabilization and destabilization of the proteins.

When checked for amino acid substitutions, the stabilizing mutation in spike (S) protein predicted an increase in the rigidity of its structure (Fig. 3; Fig. S1). The increased rigidities of the structure may provide a stable conformation to the protein that may positively influence the binding of spike protein to the ACE2 receptor (18). The major mutations D614G and S477N were located at potential epitope regions (codons 469 to 882), with S477N particularly positioned in the receptor-binding domain (RBD) of the S protein (319 to 541).

The most frequent amino acid substitutions were observed in the nucleocapsid (N) protein, in which the variants S194L, D103Y, P13L, S197L, M234I, and S188L were predicted to be stabilizing according to both the analytical servers (Table 2). In contrast, membrane (M) and envelope (E) proteins accounted for the least number of amino acid substitutions. The amino acid changes in M (T175M) indicated a stabilizing effect, while E does not account for any stabilizing variant. Structural analysis of double (D614G + S477N; D614G + A222V) and triple (D614G + S477N + A222V) mutation patterns in the S protein indicated ΔΔG values of 0.228, 0.195 and 0.129, respectively (Table 3). This signifies that accumulation of spike mutation in D614G-bearing lineages could potentially be affecting the stability of the spike and therefore may influence the binding affinity toward the ACE2 receptor.

DISCUSSION

Since the beginning of the COVID-19 pandemic, whole-genome, sequence-based phylogenetic inference has been heavily utilized in tracing viral origins and transmission chains (19). However, as the virus has evolved with time, genomic data are being increasingly used in guiding infection risk and control strategies. Several genomic mutations have been mapped that seem to be of advantage to the virus (20). In parallel, numerous vaccine candidates have been designed using genomic data from the original SARS-CoV-2 strain of Wuhan and many are now approved for use or at late-stage trials (21, 22). Based on immunological data obtained from infected and

TABLE 2 Predicted effect of protein stability in the presence of amino acid mutations in the SARS-COV-2 genomes

Protein	Mutation	ΔΔG SDM (kcal/mol)	ΔΔG DUET (kcal/mol)	Stability
Spike	A222V	0.95	0.91	Stabilizing
	S477N	0.31	0.02	Stabilizing
	L18F	−0.801	−0.46	Destabilizing
	N439K	−0.29	0.3	Inconclusive
	L5F	−0.801	−0.1	Destabilizing
	W1214G	−1.913	−0.28	Destabilizing
	R21I	−0.856	0.46	Inconclusive
	A262S	−2.13	−1.66	Destabilizing
	S98F	1.23	−0.58	Inconclusive
	D1163Y	0.21	0.26	Stabilizing
	G1167V	−0.58	−2.25	Destabilizing
	D936Y	−0.13	−0.32	Destabilizing
	P272L	2.12	0.36	Stabilizing
	D80Y	0.77	−3.08	Inconclusive
	E583D	−0.69	−0.86	Destabilizing
	P1263L	−0.231	1.29	Inconclusive
	K1073N	−0.48	−0.45	Destabilizing
	D253G	−0.19	0.04	Inconclusive
	T723I	0.64	0.21	Stabilizing
	A688V	−0.18	0.03	Inconclusive
	A626S	−2.6	−1.66	Destabilizing
	L54F	−0.61	−1.33	Destabilizing
	H655Y	0.6	1.44	Stabilizing
	G769V	0.6	0.14	Stabilizing
	L176F	0.12	−0.95	Inconclusive
	G1124V	−1.52	−0.14	Destabilizing
	V622F	−0.2	−0.67	Destabilizing
	S255F	0.94	−0.8	Inconclusive
	H49Y	0.4	1.14	Stabilizing
	D839Y	−0.389	−1.08	Destabilizing
	V1176F	−0.92	−0.55	Destabilizing
	D215H	0.8	1.35	Stabilizing
	H146Y	1.139	−0.29	Inconclusive
	A879S	−2.57	−1.69	Destabilizing
	Q677H	0.98	−0.48	Inconclusive
	D1084Y	−0.43	−0.03	Destabilizing
	V1068F	−1.05	−1.15	Destabilizing
	P25S	−0.392	0.93	Inconclusive
	A520S	−1.23	−0.15	Destabilizing
	G261V	0.16	0.16	Stabilizing
	D574Y	−0.56	−0.45	Destabilizing
	T29I	0.48	0.51	Stabilizing
	Y453F	−0.17	−0.48	Destabilizing
	N501Y	0.41	−0.42	Inconclusive
	S939F	0.76	−0.71	Inconclusive
	T95I	1.91	0.37	Stabilizing
	Q675H	0.8	−0.4	Inconclusive
Nucleocapsid	A220V	−0.51	−1.13	Destabilizing
	S194L	1.15	−0.02	Inconclusive
	D103Y	1.45	0.55	Stabilizing
	P13L	0.84	0.23	Stabilizing
	S197L	1.27	0.26	Stabilizing
	A398V	−0.98	−1.03	Destabilizing
	P199L	1.27	−0.21	Inconclusive
	M234I	0.69	0.41	Stabilizing
	S188L	1.21	−0.06	Inconclusive
	S183Y	0.05	−0.71	Inconclusive
Membrane	T175M	0.69	−0.26	Inconclusive
Envelope	P71S	−0.03	−2.35	Destabilizing

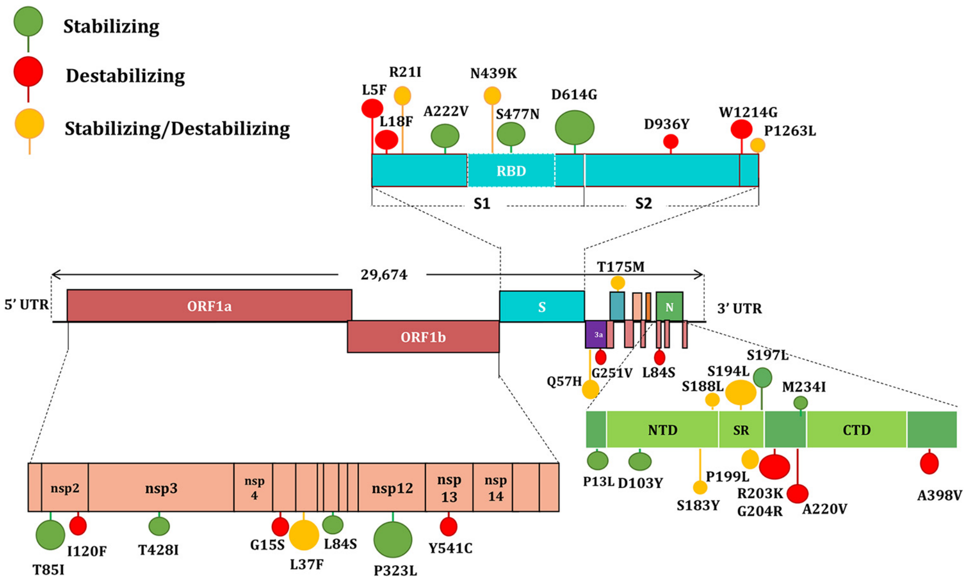

FIG 3 Schematic representation of SARS-CoV-2 genome organization, the major amino acid substitutions, and stability of amino acid changes. Stabilizing mutations are colored green, destabilizing mutations are colored red, and mutations that neither stabilize nor destabilize are colored yellow.

recovered patients, almost all COVID-19 vaccine candidates of today are based on the original SARS-CoV-2 spike protein or its RBD domain (23–25). However, as vaccines are introduced and successful treatment options become available, it is vital that we carefully monitor the mutations in the immunogenic region of SARS-CoV-2 genome (26). Mapping these changes to protein structure will allow preemptive forecasting of the direction of change in vaccine effectiveness and guide future preparedness efforts. We analyzed the impact of recurrent amino acid replacements in the genomic evolution and proteome stability of SARS-CoV-2 from its introduction in December 2019 through to November 2020. Our analysis found an intriguing balance of stabilizing and destabilizing mutations, which may have allowed SARS-CoV-2 to evolve and persist without losing pathogenicity.

SARS-CoV-2 is considered a slowly evolving virus, as it possesses an inherent proofreading mechanism to repair the mismatches during its replication. This is believed to have a crucial role in maintaining the stability and integrity of the viral genome (27, 28). Our analysis confirmed previously recorded positive natural selection of the D614G, S477N (29), A222V, and V1176F (30) variants and a global expansion of the PANGOLIN variant B.1 (11). In addition, we also observed a positive natural selection of Q57H (B.1.X), R203K/G204R (B.1.1.X), T85I (B.1.2-B.1.3), G15S+T428I (C.X), and I120F (D.X) variants (Fig. 2).

Apart from the 11 clade-defining mutations, some of the major missense mutations were in the four structural proteins (E, M, N, and S). When analyzed for their impact ($n = 59$) in the respective protein structure, the spike glycoprotein, more specifically its RBD domain, was found to be most vulnerable to frequent mutations. This may be due to the immunological observation that most neutralizing anti-SARS-CoV-2 antibodies have been found to target the RBD domain of the S protein (31, 32). Consistent with

TABLE 3 Impact of independent, double and triple mutations in the spike protein

Protein	Combinations	Mutations	ΔΔG (pred)	C (pred)
Spike	Independent	D614G	0.422	0.892
	Double	D614G+S477N	0.228	0.896
		D614G+A222V	0.195	0.889
	Triple	D614G+S477N+A222V	0.129	0.129

this finding, a total of 4,170 missense mutations have been reported in the spike protein, with 683 on the RBD domain alone (CoV-Glue, accessed 12 December 2020). Computational prediction to understand the effect of amino acid substitutions in E, M, N, and S proteins revealed a balance of stabilization and destabilization of the proteins. While viral populations carrying mutations with higher stabilizing effects (positive $\Delta\Delta G$ values) would be expected to become dominant variants, it is interesting to note that destabilization mutations in the major protein targets of SARS-CoV-2 have also generated variants that have been hugely successful. For example, many of the favorably selected variants, such as L18F, L5F (spike); R203K, G204R, and A220V (nucleocapsid), were found to be destabilizing the respective protein structure (Table 1). As destabilizing mutations are known for their crucial functional roles, a trade-off between stabilizing and destabilizing mutations may balance the protein function and structure in ways that are not yet fully understood (33, 34).

In our study, the effect of mutations on respective proteins was primarily estimated based on the physical change in free energy for a single “native” protein conformation. To allow the most robust correlation of mutations with molecular evolution, the mutational effects for the protein in an unfolded state, and the possibility of structural adjustment of the folded state in response to the mutation, needs to be explored in future studies when more structural dynamic information becomes available (35). While our study highlights the impact of $\Delta\Delta G$ analyses as a reference frame for evolutionary evaluation, molecular evolution is likely a consequence of complex amalgamation of changes in free energy, entropy, solvent accessibilities, etc. (36). As the data on these unchecked parameters becomes available, predicting evolutionary selection of mutation with respect to the phylogeny would become confirmatory. Our study highlighting preliminary data linking free energy and phylogeny would help streamline the scope of future studies by providing a baseline matrix.

The currently circulating spike variants or RBD variants need to be taken into account while evaluating vaccine candidates or neutralizing monoclonal antibodies against SARS-CoV-2 (37). Mapping the viral mutations that escape antibody binding is essential for accessing the efficacy of therapeutic and prophylactic anti-SARS-CoV-2 agents (29, 38). Recently generated experimental evidence suggests that leading vaccines (mRNA-1273, BNT162b1, and ChAdOx1 nCoV-19) and two potent neutralizing antibodies (REGN10987 and REGN10933) are unlikely to be affected by the dominant variant D614G (23, 24, 39–41). As all three candidate vaccines encode RBD or the part of spike protein as antigens, the viral population is expected to try and escape by altering the positioning of the respective antigens (42) under vaccine-induced selection pressure. Notably, complete escape mutation maps of 3,804 of the 3,819 possible RBD amino acid mutations against 10 human monoclonal antibodies are already in place (29, 42). The antigenic effect of key RBD mutations against the REGN-COV2 cocktail (REGN10933 and REGN10987) showed N439K and K444R variants escaped neutralization only by REGN10987, while E406W escaped both individual REGN-COV2 antibodies and the cocktail (38). Similar strategies should be adopted to map all antibody resistance mutations against neutralizing antibodies elicited after vaccination. Once mutation escape maps are available for all successful vaccine candidates, vaccine roll-out strategies should be carefully planned to counter geographically confined escape mutants.

In conclusion, our study highlights the importance of continued genomic surveillance, mutation mapping, stability analysis, and potential escape mutation cataloguing both in the pre- and postvaccination period of SARS-CoV-2 so as to design the epidemiologically best vaccination programs. The currently observed mutation pattern and subsequent phylogenetic diversification of SARS-CoV-2 seem to be strongly influenced by the negative and positive selection pressures. The overall variation in SARS-CoV-2 sequences is currently low compared to many other RNA viruses. One of the possible reasons for the low rate of mutations can be attributed to the widespread absence of neutralizing antibodies or the selective pressure. Once the virus population is challenged with the vaccine candidates or therapeutic monoclonal antibodies, the currently known epitopes on surfaces of SARS-CoV-2 proteins are likely to undergo rapid forced change for survival. Thus, the prevalence of such possible escape mutations needs to be monitored even more carefully after vaccination if we are to remain ahead of this rapidly shifting pandemic curve.

MATERIALS AND METHODS

Data acquisition and curation. In total, we have retrieved 7,000 genomes from GISAID EpiCoV database (https://www.gisaid.org/). Data sets that were flagged as complete (>28,000 bp) were screened and subsequently manually curated for excluding low quality/coverage sequences and duplicates. Sequence metadata was retrieved and only genomes containing sampling time and location were chosen for the study. Lineages were assigned from alignment file using the Phylogenetic Assignment of Named Global Outbreak LINeages tool PANGOLIN v1.07 (https://github.com/hCoV-2019/pangolin). We selected a subset of 513 genomes (Table S1 in the supplemental material) that belongs to all major PANGOLIN lineages and common mutations for the optimal output of the phylogenetic tree.

Phylogenetic analysis. Genome sequences were aligned against the original Wuhan-Hu-1 genome (accession: NC_045512) using multiple genome sequence alignment tool MAFFT (v6.240) (43). Subsequently, the error prone 5′-UTR and 3′-UTR regions were masked and the genome size was adjusted without losing key sites. A maximum likelihood (ML) tree was generated using IQTREE v.1.6.1 (http://www.iqtree.org/) under the GTR nucleotide substitution model with 1,000 bootstrap replicates (44). The ML tree was visualized and labeled using the interactive tree of life software iTOL v.3 (45).

Mutation profiling. In order to identify the genetic variants, assembled genomes were mapped against the reference (Wuhan-Hu-1: accession: NC_045512) using Snippy mapping and variant calling pipeline (https://github.com/tseemann/snippy) (46). Among the SNPs, missense SNPs (nonsynonymous) were extracted using custom-written bash scripts and manually curated as per the CoV-GLUE database (http://cov-glue.cvr.gla.ac.uk/). Specifically, we considered 11 lineage-defining mutations and 59 major missense mutations in four major structural proteins: envelope protein (E), membrane glycoprotein (M), nucleocapsid phosphoprotein (N), and spike protein (S). Structural analysis of 70 amino acid substitutions in SARS CoV-2 mutants were analyzed to examine the potential impact of these mutations on protein stability.

Structural analysis. The structural impact of mutations has been assessed from the COVID-3D server (http://biosig.unimelb.edu.au/covid3d), which has integrated analytics regarding mutation-based structural changes in a protein. Vibrational entropy (VE) ($\Delta\Delta S$) and unfolding Gibbs free energy (FE) ($\Delta\Delta G$) were considered markers to ascertain the stability of the variants. Gibbs free energy (FE) ($\Delta\Delta G$) values from the site directed mutator (SDM), DUET, and DynaMut tools available in COVID-3D server were considered (47, 48). The change in vibrational entropy energy ($\Delta\Delta$SVib ENCoM) between wild-type and mutant protein was calculated using DynaMut (49). VE explains the occupation probabilities of protein residues in an energy landscape based on average configurational entropies. Considerable decrease in VE increases the rigidity of the proteins (50). FE, on the other hand, describes the free energy alterations while unfolding a kinetically stable protein (49). The positive and negative values of $\Delta\Delta G$ indicate the stabilizing and destabilizing mutations. DynaMine (http://dynamine.ibsquare.be/) was employed to validate the stability profiles through residue level (sequence-based) dynamics. Backbone N-H S^2 order parameter values (atomic bond vector's movement restrictions) were generated according to the molecular reference frame. These N-H S^2 order parameter values are evaluated from experimentally determined NMR chemical shifts. A value above 0.8 is considered highly stable, values between 0.6 and 0.8 can be considered to be functionally contextual, and values >0.6 are highly flexible (51).

Data availability. The genome sequences used in this study are available in the Global Initiative on Sharing All Influenza Data (GISAID) with accession IDs (see Table S1 in the supplemental material).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, TIF file, 0.1 MB.

TABLE S1, XLSX file, 0.03 MB.

TABLE S2, XLSX file, 0.1 MB.

ACKNOWLEDGMENTS

We thank Soumya Basu (ICMR, Senior research Fellow) for his contribution and helpful advice in the structural analysis. We gratefully acknowledge the Department of Clinical Microbiology, Christian Medical College, Vellore, Tamil Nadu, India, for providing all the necessary computational facilities for this work. We are grateful to the staff of Christian Medical College for their assistance with data curation.

This work received no specific external funding and the work was carried out depending on the resources of the host institute.

We 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.

REFERENCES

1. World Health Organization. 2020. Coronavirus disease 2019 (COVID-19): situation report. World Health Organization, Geneva, Switzerland.
2. Worldometer. 2020. COVID-19 Coronavirus. https://www.worldometers.info/coronavirus/. Accessed 27 December 2020.
3. World Health Organization. 2020. Rolling updates on coronavirus disease (COVID19). https://www.who.int/emergencies/diseases/novel-coronavirus2019/events-as-they-happen. Accessed 18 May 2020.
4. Li G, De Clercq E. 2020. Therapeutic options for the 2019 novel coronavirus (2019-nCoV). Nat Rev Drug Discov 19:149–150. https://doi.org/10.1038/d41573-020-00016-0.
5. Hodgson SH, Mansatta K, Mallett G, Harris V, Emary KR, Pollard AJ. 2020. What defines an efficacious COVID-19 vaccine? A review of the challenges assessing the clinical efficacy of vaccines against SARS-CoV-2. Lancet Infect Dis 21:e26–e35. https://doi.org/10.1016/S1473-3099(20)30773-8.
6. Sironi M, Hasnain SE, Rosenthal B, Phan T, Luciani F, Shaw M-A, Sallum MA, Mirhashemi ME, Morand S, González-Candelas F, Editors of Infection, Genetics and Evolution. 2020. SARS-CoV-2 and COVID-19: a genetic, epidemiological, and evolutionary perspective. Infect Genet Evol 84:104384. https://doi.org/10.1016/j.meegid.2020.104384.
7. Singer J, Gifford R, Cotten M, Robertson D. Accessed 12 December 2020. CoV-GLUE: a web application for tracking SARS-CoV-2 genomic variation. Preprints 2020060225. https://doi.org/10.20944/preprints202006.0225.v1.
8. Hu B, Guo H, Zhou P, Shi ZL. 2020. Characteristics of SARS-CoV-2 and COVID-19. Nature Rev Microbiol 19:141–154. https://doi.org/10.1038/s41579-020-00459-7.
9. 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.
10. Dearlove B, Lewitus E, Bai H, Li Y, Reeves DB, Joyce MG, Scott PT, Amare MF, Vasan S, Michael NL, Modjarrad K, Rolland M. 2020. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. Proc Natl Acad Sci U S A 117:23652–23662. https://doi.org/10.1073/pnas.2008281117.
11. Shu Y, McCauley J. 2017. GISAID: global initiative on sharing all influenza data–from vision to reality. Eurosurveillance 22:30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.
12. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34:4121–4123. https://doi.org/10.1093/bioinformatics/bty407.
13. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, Du Plessis L, Pybus OG. 2020. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. Nat Microbiol 5:1403–1407. https://doi.org/10.1038/s41564-020-0770-5.
14. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui J, Lu J. 2020. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev 7:1012–1023. https://doi.org/10.1093/nsr/nwaa036.
15. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, Zhang X, Muruato AE, Zou J, Fontes-Garfias CR, Mirchandani D, Scharton D, Bilello JP, Ku Z, An Z, Kalveram B, Freiberg AN, Menachery VD, Xie X, Plante KS, Weaver SC, Shi P-Y. 2021. Spike mutation D614G alters SARS-CoV-2 fitness. Nature 592:116–116. https://doi.org/10.1038/s41586-020-2895-3.
16. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, Hastie KM, Parker MD, Partridge DG, Evans CM, Freeman TM, de Silva TI, McDanal C, Perez LG, Tang H, Moon-Walker A, Whelan SP, LaBranche CC, Saphire EO, Montefiori DC, Sheffield COVID-19 Genomics Group. 2020. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 182:812–827. https://doi.org/10.1016/j.cell.2020.06.043.
17. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, Schiergens TS, Herrler G, Wu N-H, Nitsche A, Müller MA, Drosten C, Pöhlmann S. 2020. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 181:271–280.e8. https://doi.org/10.1016/j.cell.2020.02.052.
18. Ou J, Zhou Z, Dai R, Zhao S, Wu X, Zhang J, Lan W, Cui L, Wu J, Seto D, Chodosh J. 2021. V367F mutation in SARS-CoV-2 spike RBD emerging during the early transmission phase enhances viral infectivity through increased human ACE2 receptor binding affinity. J Virol e00617-21. https://doi.org/10.1128/JVI.00617-21.
19. Oude Munnink BB, Nieuwenhuijse DF, Stein M, O'Toole Á, Haverkate M, Mollers M, Kamga SK, Schapendonk C, Pronk M, Lexmond P, van der Linden A, Bestebroer T, Chestakova I, Overmars RJ, van Nieuwkoop S, Molenkamp R, van der Eijk AA, GeurtsvanKessel C, Vennema H, Meijer A, Rambaut A, van Dissel J, Sikkema RS, Timen A, Koopmans M, Dutch-Covid-19 response team. 2020. Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. Nat Med 26:1802. https://doi.org/10.1038/s41591-020-1128-5.
20. Gómez-Carballa A, Bello X, Pardo-Seco J, Martinón-Torres F, Salas A. 2020. Mapping genome variation of SARS-CoV-2 worldwide highlights the impact of COVID-19 super-spreaders. Genome Res 30:1434–1448. https://doi.org/10.1101/gr.266221.120.
21. Dong Y, Dai T, Wei Y, Zhang L, Zheng M, Zhou F. 2020. A systematic review of SARS-CoV-2 vaccine candidates. Signal Transduct Target Ther 5:237. https://doi.org/10.1038/s41392-020-00352-y.
22. Alturki SO, Alturki SO, Connors J, Cusimano G, Kutzler MA, Izmirly AM, Haddad EK. 2020. The 2020 pandemic: current SARS-CoV-2 vaccine development. Front Immunol 11:1880. https://doi.org/10.3389/fimmu.2020.01880.
23. Corbett KS, Edwards D, Leist SR, Abiona OM, Boyoglu-Barnum S, Gillespie RA, Himansu S, Schafer A, Ziwawo CT, DiPiazza AT, Dinnon KH, et al. 2020. SARS-CoV-2 mRNA vaccine development enabled by prototype pathogen preparedness. bioRxiv https://doi.org/10.1101/2020.06.11.145920.
24. Sahin U, Muik A, Derhovanessian E, Vogler I, Kranz LM, Vormehr M, Baum A, Pascal K, Quandt J, Maurus D, Brachtendorf S, Lörks V, Sikorski J, Hilker R, Becker D, Eller A-K, Grützner J, Boesler C, Rosenbaum C, Kühnle M-C, Luxemburger U, Kemmer-Brück A, Langer D, Bexon M, Bolte S, Karikó K, Palanche T, Fischer B, Schultz A, Shi P-Y, Fontes-Garfias C, Perez JL, Swanson KA, Loschko J, Scully IL, Cutler M, Kalina W, Kyratsous CA, Cooper D, Dormitzer PR, Jansen KU, Türeci Ö. 2020. COVID-19 vaccine BNT162b1 elicits human antibody and TH 1 T cell responses. Nature 586:594–599. https://doi.org/10.1038/s41586-020-2814-7.
25. Poland GA, Ovsyannikova IG, Crooke SN, Kennedy RB. 2020. SARS-CoV-2 vaccine development: current status. Mayo Clin Proc 95:2172–2188. https://doi.org/10.1016/j.mayocp.2020.07.021.
26. Li Q, Wu J, Nie J, Zhang L, Hao H, Liu S, Zhao C, Zhang Q, Liu H, Nie L, Qin H, Wang M, Lu Q, Li X, Sun Q, Liu J, Zhang L, Li X, Huang W, Wang Y. 2020. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. Cell 182:1284–1294. https://doi.org/10.1016/j.cell.2020.07.012.
27. Robson F, Khan KS, Le TK, Paris C, Demirbag S, Barfuss P, Rocchi P, Ng WL. 2020. Coronavirus RNA proofreading: molecular basis and therapeutic targeting. Mol Cell 79:710–727. https://doi.org/10.1016/j.molcel.2020.07.027.
28. Bar-On YM, Flamholz A, Phillips R, Milo R. 2020. SARS-CoV-2 (COVID-19) by the numbers. Elife 9:e57309. https://doi.org/10.7554/eLife.57309.
29. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, Navarro MJ, Bowen JE, Tortorici MA, Walls AC, King NP, Veesler D, Bloom JD. 2020. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell 182:1295–1310. https://doi.org/10.1016/j.cell.2020.08.012.
30. Farkas C, Mella A, Haigh JJ. 2020. Large-scale population analysis of SARS-CoV2 whole genome sequences reveals host-mediated viral evolution with emergence of mutations in the viral Spike protein associated with elevated mortality rates. medRxiv https://doi.org/10.1101/2020.10.23.20218511.
31. Ju B, Zhang Q, Ge J, Wang R, Sun J, Ge X, Yu J, Shan S, Zhou B, Song S, Tang X, Yu J, Lan J, Yuan J, Wang H, Zhao J, Zhang S, Wang Y, Shi X, Liu L, Zhao J, Wang X, Zhang Z, Zhang L. 2020. Human neutralizing antibodies elicited by SARS-CoV-2 infection. Nature 584:115–119. https://doi.org/10.1038/s41586-020-2380-z.
32. Liu L, Wang P, Nair MS, Yu J, Rapp M, Wang Q, Luo Y, Chan JF-W, Sahi V, Figueroa A, Guo XV, Cerutti G, Bimela J, Gorman J, Zhou T, Chen Z, Yuen K-Y, Kwong PD, Sodroski JG, Yin MT, Sheng Z, Huang Y, Shapiro L, Ho DD. 2020. Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. Nature 584:450–456. https://doi.org/10.1038/s41586-020-2571-7.
33. Laha S, Chakraborty J, Das S, Manna SK, Biswas S, Chatterjee R. 2020. Characterizations of SARS-CoV-2 mutational profile, spike protein stability and viral transmission. Infect Genet Evol 85:104445. https://doi.org/10.1016/j.meegid.2020.104445.
34. Teng S, Sobitian A, Rhoades R, Liu D, Tang Q. 2020. Systemic effects of missense mutations on SARS-CoV-2 spike glycoprotein stability and receptor binding affinity. Brief Bioinform 22:1239–1253. https://doi.org/10.1093/bib/bbaa233.
35. Sikosek T, Chan HS. 2014. Biophysics of protein evolution and evolutionary protein biophysics. J R Soc Interface 11:20140419. https://doi.org/10.1098/rsif.2014.0419.
36. Echave J, Wilke CO. 2017. Biophysical models of protein evolution: understanding the patterns of evolutionary sequence divergence. Annu Rev Biophys 46:85–103. https://doi.org/10.1146/annurev-biophys-070816-033819.

37. Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, Lorenzi JCC, Muecksch F, Rutkowska M, Hoffmann H-H, Michailidis E, Gaebler C, Agudelo M, Cho A, Wang Z, Gazumyan A, Cipolla M, Luchsinger L, Hillyer CD, Caskey M, Robbiani DF, Rice CM, Nussenzweig MC, Hatziioannou T, Bieniasz PD. 2020. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. Elife 9:e61312. https://doi.org/10.7554/eLife.61312.
38. Starr TN, Greaney AJ, Addetia A, Hannon WH, Choudhary MC, Dingens AS, Li JZ, Bloom JD. 2020. Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. Science 371:850–854. https://doi.org/10.1126/science.abf9302.
39. Voysey M, Clemens SAC, Madhi SA, Weckx LY, Folegatti PM, Aley PK, Angus B, Baillie VL, Barnabas SL, Bhorat QE, Bibi S. 2020. 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. https://doi.org/10.1016/S0140-6736(20)32661-1.
40. Baum A, Fulton BO, Wloga E, Copin R, Pascal KE, Russo V, Giordano S, Lanza K, Negron N, Ni M, Wei Y, Atwal GS, Murphy AJ, Stahl N, Yancopoulos GD, Kyratsous CA. 2020. Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. Science 369:1014–1018. https://doi.org/10.1126/science.abd0831.
41. McAuley AJ, Kuiper MJ, Durr PA, Bruce MP, Barr J, Todd S, Au GG, Blasdell K, Tachedjian M, Lowther S, Marsh GA, Edwards S, Poole T, Layton R, Riddell S-J, Drew TW, Druce JD, Smith TRF, Broderick KE, Vasan SS. 2020. Experimental and in silico evidence suggests vaccines are unlikely to be affected by D614G mutation in SARS-CoV-2 spike protein. NPJ Vaccines 5:1–5. https://doi.org/10.1038/s41541-020-00246-8.
42. Greaney AJ, Starr TN, Gilchuk P, Zost SJ, Binshtein E, Loes AN, Hilton SK, Huddleston J, Eguia R, Crawford KH, Dingens AS. 2020. Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. Cell Host Microbe 29:44–57. https://doi.org/10.1016/j.chom.2020.11.007.
43. Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform 20:1160–1166. https://doi.org/10.1093/bib/bbx108.
44. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274. https://doi.org/10.1093/molbev/msu300.
45. Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:W256–W259. https://doi.org/10.1093/nar/gkz239.
46. Seemann T. 2015. Snippy: rapid haploid variant calling and core SNP phylogeny. GitHub. github.com/tseemann/snippy.
47. Pandurangan AP, Ochoa-Montaño B, Ascher DB, Blundell TL. 2017. SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res 45:W229–W235. https://doi.org/10.1093/nar/gkx439.
48. Pires DE, Ascher DB, Blundell TL. 2014. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 42:W314–W319. https://doi.org/10.1093/nar/gku411.
49. Rodrigues CH, Pires DE, Ascher DB. 2018. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res 46:W350–W355. https://doi.org/10.1093/nar/gky300.
50. Goethe M, Fita I, Rubi JM. 2015. Vibrational entropy of a protein: large differences between distinct conformations. J Chem Theory Comput 11:351–359. https://doi.org/10.1021/ct500696p.
51. Cilia E, Pancsa R, Tompa P, Lenaerts T, Vranken WF. 2014. The DynaMine webserver: predicting protein dynamics from sequence. Nucleic Acids Res 42:264–270. https://doi.org/10.1093/nar/gku270.