Computational and Structural Biotechnology Journal 19 (2021) 3799–3809

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Evolution, correlation, structural impact and dynamics of emerging SARS-CoV-2 variants

Austin N. Spratt [a,1], Saathvik R. Kannan [a,1], Lucas T. Woods [a,b], Gary A. Weisman [a,b], Thomas P. Quinn [b], Christian L. Lorson [a,c], Anders Sönnerborg [d,e,*], Siddappa N. Byrareddy [e,f,*], Kamal Singh [a,c,e,g,*]

[a] Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA
[b] Department of Biochemistry, University of Missouri, Columbia, MO 65211, USA
[c] Department of Veterinary Pathobiology, University of Missouri, Columbia, MO 65211, USA
[d] Division of Infectious Diseases, Department of Medicine, Karolinska Institute, Huddinge 14186, Stockholm, Sweden
[e] Division of Clinical Microbiology, Department of Laboratory Medicine, Karolinska Institute, Huddinge 14186, Stockholm, Sweden
[f] Department of Pharmacology and Experimental Neuroscience, University of Nebraska Medical Center, Omaha, NE 68198, USA
[g] Sanctum Therapeutics Corporation, Sunnyvale, CA 94087, USA

ARTICLE INFO

Article history:
Received 8 March 2021
Received in revised form 22 June 2021
Accepted 22 June 2021
Available online 24 June 2021

Keywords:
SARS-CoV-2
SARS-CoV
COVID-19
B.1.1.7
B.1.351, P.1, CAL.20C

ABSTRACT

Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) infections remain unmanageable in some parts of the world. As with other RNA viruses, mutations in the SARS-CoV-2 gene have been continuously evolving. Recently, four variants have been identified, B.1.1.7, B.1.351, P.1 and CAL.20C. These variants appear to be more infectious and transmissible than the original Wuhan-Hu-1 virus. Using a combination of bioinformatics and structural analyses, we show that the new SARS-CoV-2 variants emerged in the background of an already known Spike protein mutation D614G together with another mutation P323L in the RNA polymerase of SARS-CoV-2. The phylogenetic analysis showed that the CAL.20C and B.1.351 shared one common ancestor, whereas the B.1.1.7 and P.1 shared a different ancestor. Structural comparisons did not show any significant difference between the wild-type and mutant ACE2/Spike complexes. Structural analysis indicated that the N501Y mutation may increase hydrophobic interactions at the ACE2/Spike interface. However, reported greater binding affinity of N501Y Spike with ACE2 does not seem to be entirely due to increased hydrophobic interactions, given that Spike mutation R417T in P.1 or K417N in B.1.351 results in the loss of a salt-bridge interaction between ACE2 and S-RBD. The calculated change in free energy did not provide a clear trend of S protein stability of mutations in the variants. As expected, we show that the CAL.20C generally migrated from the west coast to the east coast of the USA. Taken together, the analyses suggest that the evolution of variants and their infectivity is complex and may depend upon many factors.

1. Introduction

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), the virus responsible for Coronavirus Disease 2019 (COVID-19), is the seventh human coronavirus (hCoV) discovered to date. The other six are hCoV-229E, hCoV-NL63, hCoV-HKU1, hCoV-OC43, SARS-CoV and Middle East Respiratory Syndrome (MERS–CoV). Four hCoVs (hCoV-229E, hCoV-NL63, hCoV-HKU1 and hCoV-OC43) typically cause mild and self-limiting cold-like diseases. However, three hCoVs (SARS-CoV, MERS-CoV, and SARS-CoV-2) cause fatal respiratory illness. CoVs are single-stranded, non-segmented positive (+)-sense RNA viruses of ~ 30 kb genome length. They encode up to 16 nonstructural proteins (nsps) and 10–12 structural/accessory proteins [1,2]. Several nsps form a replication/transcription complex (RTC) that synthesizes the nascent (-) strand RNA to be used as a template for the (+) sense genome and a set of subgenomic RNAs (sgRNA) with a common 5′-leader sequence and a 3′-poly-A tail, whereas other nsps partici-

[*] Corresponding authors at: Division of Clinical Microbiology, Department of Laboratory Medicine, Karolinska Institute, Huddinge 14186, Stockholm, Sweden (A. Sönnerborg, Siddappa N. Byrareddy, K. Singh), Department of Pharmacology and Experimental Neuroscience, University of Nebraska Medical Center, Omaha, NE 68198, USA (Siddappa N. Byrareddy) and Department of Veterinary Pathobiology, University of Missouri, Columbia, MO 65211, USA (K. Singh).

E-mail addresses: anders.sonnerborg@ki.se (A. Sönnerborg), sid.byrareddy@unmc.edu (S.N. Byrareddy), singhka@missouri.edu, kamlendra.singh@ki.se (K. Singh).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.csbj.2021.06.037
2001-0370/© 2021 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

Table 1
Major new SARS-CoV-2 variants and mutations in the variants.

Variant Name	Mutations	Known Mutations (in Addition to Those in S Protein) in Different Variants
B.1.1.7 (UK)	Δ H69, Δ V70, Δ Y144, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H	N protein: S235F, D3L, Y73C, R521; ORF8: Q27 stop (changes 27th amino acid from Q to STOP codon and leaves a 26 amino acid-long stump; 2 other mutations appear after the stop codon, but since the protein is cut short, they may have no function); ORF1a: SGF 3675–3677 deletion, I2230T, A1708D, T1001I; there are six additional silent mutations
B.1.351(South Africa)	L18F, D80A, D215G, Δ L242, Δ A243, Δ L244, R246I, K417N, E484K, N501Y, D614G, A701V	N protein: T205I, M/E, P71L; ORF1a: SGF 3675–3677 deletion, K1655N
P.1 (Brazil/Japan)	L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, D614G, H655Y, T1027I	N protein: P80R, 28269–73 insertion, E92K; ORF1b: E5665D; ORF1a: SGF 3675–3677 deletion, K1795Q, S1188L
B.1.429 (California)	S13I, W152C, L452R, D614G	ORF1a: I4205V; ORF1b: D1183Y

pate in the processing of translated polyprotein into functional nsps. Of 10–12 structural proteins, there are four major proteins: Spike (S), nucleocapsid (N), membrane (M) and envelope (E) that participate in the infection and assembly of infectious progeny virus [3,4].

In the first step of viral infection, the Spike protein (S protein) of hCoVs directly interacts with distinct host cell receptors. Angiotensin-converting enzyme 2 (ACE2) serves as the cell-surface receptor for SARS-CoV and SARS-CoV-2. The action of host cell serine protease TMPRSS2 on the ACE2/Spike complex cleaves S protein into S1 and S2 subunits and facilitates endocytosis of viral particles [5]. Due to its primary role in hCoV infections, S protein is a crucial target for developing vaccines, therapeutic antibodies and diagnostics [6]. Structural data show that the Spike receptor-binding domain (S-RBD) within the S1 subunits of both SARS-CoV and SARS-CoV-2 has largely conserved interactions with ACE2, despite being the domain of lowest homology between the two viruses [7–10]. Genetics and chemical perturbation studies have demonstrated that the ACE2-mediated entry of SARS-CoV and CoV-2 requires the cell surface heparan sulfate (HS) as a cofactor by directly binding to Spike [11]. Many structures of the ACE2/Spike or ACE2/S-RBD complex have been solved [9,10,12–17]. These structures showed that the S-RBDs of SARS-CoV and SARS-CoV-2 share structural homology, and the S-RBDs of the two CoVs bind to overlapping sites on ACE2 [12].

RNA viruses are known to mutate for adaptation and evolution. Since the first report of the D614G mutation [18], other mutations throughout the SARS-CoV-2 genome have been reported [18–20]. Additional prominent mutations in nsp12 (RNA-dependent RNA polymerase), nsp13 (helicase) and the 5′-UTR (C241U) have been reported [19]. Subsequently, it was shown that the D614G mutation increases viral fitness and infectivity [21–23]. The impact of the P323L mutation in nsp12 on viral function has yet to be established. However, the P323L mutation has been proposed to enhance the processivity of nsp12 [19]. Similarly, the effect of C241U in the 5′-UTR is not known, but C241U may change the structure of the SL5 loop or affect protein binding for easy ribosomal scanning, which is favored in translation initiation [19,24]. A recently reported comprehensive analysis of mutation dynamics revealed the putative original status of SARS-CoV-2 and the early-stage spread history [25], and patient-derived mutations that alter the pathogenicity of SARS-CoV-2 virus [26].

Some mutations result in the transmittable strains giving rise to new variants. While new variants are continuously emerging, four main variants have appeared in the past few months that have caused a new wave of infections. These variants are B.1.1.7/501Y.V1 (B.1.1.7 or the UK variant or the α-variant), B.1.351/501Y.V2 (B.1.351 or the South Africa variant or the β-variant), P.1/501Y.V3 (P.1 or the Brazil/Japan variant or the γ-variant) and CAL.20C (20C/S:452R/B1.429 or the California variant or the ε-variant). Variant B.1.1.7 was identified in September 2020 [27]. Variant B.1.351 was reported in October 2020 [28], variant P.1 was first detected among Brazilians who traveled to Japan in January 2021 [29] and variant CAL.20C was reported in February 2021 [30]. Three of the four variants (B.1.1.7, B.1.351 and P.1) have the N501Y mutation in the Spike protein, which most likely results in the increased resistance to neutralizing antibodies [31–33] and an enhanced binding affinity of Spike with ACE2 [16,17]. N501Y also has been shown to cause increased virulence in animal models [34]. N501Y is not a signature mutation of the CAL.20C variant and the reason for this variant's higher transmissibility or fitness remains unknown. Additionally, the mutations in a given variant are not limited to Spike protein, but they are found in other parts of the genome. For example, CAL.20C contains three signature mutations in the Spike protein and one mutation each in nsp9 and nsp13 (Table 1).

The importance of ACE2 in SARS-CoV-2 infection may not be limited to just binding to S protein at the beginning of the infection. ACE2 is a zinc-dependent metalloprotease with four distinct regions/domains: an N-terminal signaling sequence, a catalytic domain containing the zinc-binding motif (HEMGH), a transmembrane region and a C-terminal cytosolic domain [35]. ACE2 cleaves angiotensin I (Ang I) and angiotensin II (Ang II). Ang II, which is derived from Ang I [36], is a vasoactive peptide hormone, and its cleavage by ACE2 results in a truncated heptapeptide Ang (1–7), which opposes the harmful effects of Ang II-dependent AT$_1$R signaling through its actions on the Mas receptor. Thus, Ang (1–7) exerts anti-inflammatory and anti-fibrotic effects, but also indirectly regulates AT$_1$R functions [37]. Hence, the importance of a link between SARS-CoV-2 and Ang II signaling can be underscored by inappropriately enhancing Ang II signaling predominately through AT$_1$R-induced inflammation, as seen in disease states such as obesity, diabetes, hypertension and aging [38,39].

Due to the critical role of the S-RBD in SARS-CoV-2 infection, almost all vaccines target the S-RBD. Since currently approved vaccines are based on the RBD sequence of the first isolated virus (Wuhan-Hu-1; GenBank accession no. NC_045512), newly emerged SARS-CoV-2 variants containing mutations within the S-RBD may present a challenge for available vaccines. Here, we provide genetic and structural analyses suggesting that SARS-CoV-2 variants B.1.1.7, B.1.351, P.1 and CAL.20C have evolved from two different common ancestors. The variants evolved in the background of already reported D614G mutation together with P323L mutation in the RNA dependent RNA polymerase. A lack of significant difference in the available structures of wild-type and mutation S protein receptor binding domain (S-RBD) suggests that structural changes are not involved in the higher infectivity of the variants. Additionally, calculated free energy of change upon mutation did not provide a definitive answer regarding the stability of the S protein. Initially available data suggested that the CAL.20C variant emerged before B.1.351 followed by B.1.1.7 and P.1. However, as more and more sequences are becoming available, a change in the temporal evolution of variant may be observed. The enhanced binding affinity of the N501Y-containing variants may be due to increased hydrophobicity due to the stacking π-π interaction of Y501 with Y41 of ACE2, thereby increasing the infectivity and/or transmissibility of new variants. Additionally, the propensity of SARS-CoV-2 interaction with specific integrins, presumably alternate receptors of SARS-CoV-2, may also contribute to the increased infectivity of the original Wuhan-Hu-1 virus and new variants. Taken together, the greater infectivity and/or transmutability of the variants appears complex requiring additional structural and biochemical data to clarify.

2. Results

2.1. Binding of two S protein trimers to the ACE2 dimer

A cryoEM structure of SARS-CoV-2 S-RBD, ACE2 and the sodium-dependent neutral amino acid transporter B^0AT1 (SLC6A19) showed that ACE2 in its dimeric form interacts with SLC6A19, and two SARS-CoV-2 S-RBDs bind to the ACE2 dimer [15]. For functional expression on the cell surface, SLC6A19 requires an obligatory protein, collectrin (TMEM27) in the kidney and ACE2 in the intestine [40–43]. The binding interface for the two S-RBDs on ACE2 is the same as reported for the crystal structures. However, this report suggests that two S protein trimers can bind to the ACE2 dimer (Fig. 1) when bound to B^0AT1, thereby increasing the infectivity of SARS-CoV-2. However, the SARS-CoV S-RBD/B^0AT1/ACE2 complex structure has not been resolved, and it remains to be seen whether S protein trimers enhance the infectivity of SARS-CoV-2.

2.2. Geo-prevalence and evolutionary relationship of re-emerging SARS-CoV-2

Several new SARS-CoV-2 variants have emerged over the past few months and new variants are continuously emerging. We first analyzed the mutation prevalence in variants based on their geo-prevalence. We then included the variant sequences from different countries. A list of the S protein and other SARS-CoV-2 mutations in the four major variants is given in Table 1. We analyzed sequences (n = 7,232) from different regions: Brazil (n = 119), California (n = 1,683), South Africa (n = 129) and the United Kingdom (n = 5,301) for the prevalence of signature Spike mutations. The results presented in Fig. 2 show that the S protein mutations are limited almost exclusively to specific variants and geographic locations, except for N501Y, which is present in Brazil, South Africa and UK at 56, 99, and 83% frequencies, respectively (Fig. 2). N501Y is not a signature mutation in CAL.20C. However, our analysis showed that ~ 6% of California sequences also have N501Y. The low prevalence of common mutations among different regions also suggested that these variants have evolved independently. All regions also had other mutations in addition to their respective variant's signature mutations. For example, sequences from California also had the mutations P26S, N501Y, A570D, P681H, T716I, S982A and D1118H with significant prevalence (up to 9%) in addition to the four mutations (S13I, W152C, L452R, D614G) that characterize the CAL.20C variant. Additionally, the V1176F mutation, which has not been included as a signature mutation of the P.1 variant, was present at a very high frequency (~81%) in Brazil. Thus, considering that P681H is a signature mutation of B.1.1.7, nearly 9% of infections in the USA (*i.e.,* California) may be related to B.1.1.7 (Fig. 2).

To further examine if the four variants evolved independently, we included variants from the associated countries and different parts of the world. We analyzed a significantly larger number of sequences (n = 225,368) that included B.1.1.7 (207,088), B.1.351 (3,273), P.1 (2,541) and CAL.20C (12,466). Of these, 47,078 (~23%) sequences from the UK are B.1.1.7, 590 (~18%) in South Africa are B.1.351, 61 (~2%) in Brazil are P.1, and 1,212 (~10%) in the United States are CAL.20C (as of Feb. 6, 2021). While most sequences belonged to the associated regions, variants B.1.1.7, B.1.351, P.1 and CAL.20C included 4,402, 4,068, 368 and 2,947 sequences from regions other than the UK, SA, Brazil, and the USA, respectively. These sequences were downloaded from GISAID; however, the regions were based on the three highest countries of prevalence reported on the Web Portal outbreak.info. As expected, all sequences included in this study have the D614G mutation in the S protein. A phylogenetic analysis of the first 10 high-quality, reported and dated sequences (from n = 225,368 sequences) of four variants is presented in Fig. 3**a**. We did not show the phylogenetic analysis of all (n = 225,368) sequences. Additionally, the phylogenetic analysis of the generated consensus sequences of each variant is shown in Fig. 3**b**. The results showed that: (i) the variants share common ancestors, thus B.1.1.7 and P.1 evolved from one common ancestor, and B.1.351 and CAL.20C evolved from another common ancestor (Fig. 3**b**); and (ii) the CAL.20C variant was first followed by B.1.351, then B.1.1.7 and then P.1 (Fig. 3**a and b**). Additionally, we used available Nextstrain global SARS-CoV-2 analyses web portal [44] to infer the evolutionary relationship among the variants as well as the first date of emergence, which confirmed our results (**Supplementary Fig. S1**). Supplementary **Fig. S1b** shows the zoom-in phylogenetic tree of CAL.20C and B.1.351, whereas Supplementary **Fig. S1c** shows the zoom-in region of the B.1.1.7 and P.1 variants. It is clear from this supplementary figure that CAL.20C emerged slightly before B.1.351 followed by B.1.1.7 and then P.1, as we observed in our phylogenetic analysis (Fig. 3).

We also used Nextstrain's NextClade [44] tool to further analyze and validate the evolutionary relationships between variants from variant sequences (n = 6,052); B.1.1.7 (n = 1,637), B.1.351 (n = 1,069), P.1 (n = 1,926), and Cal.20C (n = 4,182). The results (**Supplementary Figure S2**) confirmed that B.1.351 and CAL.20C shared a common ancestor, whereas B.1.1.7 and P.1 shared a different common ancestor. However, additional sequences of variants are becoming available, the temporal evolution of variants may be changing. As seen in **Supplementary Fig. S2**, it appears that B.1.351 and CAL.20C may have evolved around the same time spanning between late March 2020 and Early April 2020.

2.3. Correlation among variant-specific mutations

To determine the correlation among variant-specific mutations, we conducted a correlation analysis that included mutations P323L and C241U in addition to the mutations in the S protein, since these two mutations co-existed in the USA sequences [19]. The results presented in Fig. 4**a-d** show that D614G and P323L are highly correlated independent of the geographical locations of the variants. However, a strong correlation among the three mutations (D614G, P323L and C241U) was seen only in the USA infections, suggesting that the SARS-CoV-2 variant in the U. S. is different from the ones in the UK, Brazil and South Africa.

To determine the correlation among variant-specific mutations in S protein, we excluded the P323L and C241U mutations. Thus, B.1.1.7 mutations are strongly correlated with D614G except for T716I (Fig. 4**e**), which is negatively correlated (-0.3), suggesting that T716I most likely evolved separately in the background of (D614G, P681H, N501Y, S982A, A570D and D1118H). We also included mutation A626T in our analysis, as it is significantly prevalent in the UK viruses (75%) (**Supplementary Fig. S3**) that also contained P323L (Fig. 4**a**).

The CAL.20C (Fig. 4**b**) variant has two distinct groups of correlated mutations. The first group includes D614G, P323L and C241U (Fig. 4**b**), confirming our previous results [19]. The second

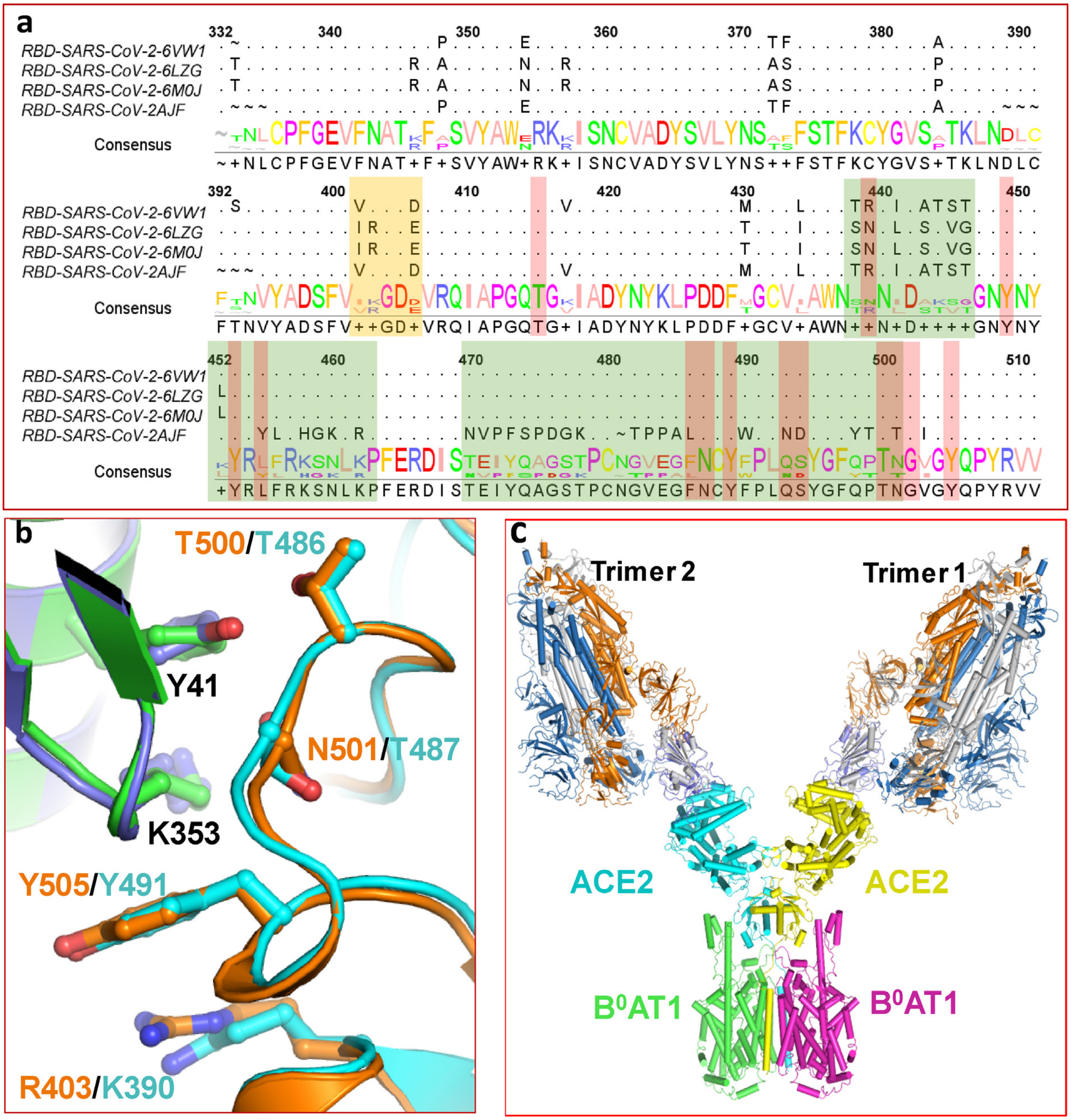
a
RBD-SARS-CoV-2-6VW1
RBD-SARS-CoV-2-6LZG
RBD-SARS-CoV-2-6M0J
RBD-SARS-CoV-2-2AJF
Consensus
~+NLCPFGEVFNAT+F+SVYAW+RK+ISNCVADYSVLYNS++FSTFKCYGVS+TKLNDLC
FTNVYADSFV++GD+VRQIAPGQTG+IADYNYKLPDDF+GCV+AWN++N+D++++GNYNY
+YRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVV
b
T500/T486
Y41
N501/T487
K353
Y505/Y491
R403/K390
c
Trimer 2
Trimer 1
ACE2
ACE2
B0AT1
B0AT1

Fig. 1. Details of genetics and structural components likely to contribute to SARS-CoV-2 binding, entry and infectivity in host cells. **a**. Structure-based sequence alignment of S-RBDs from SARS-CoV (PDB entry 2AJF) and SARS-CoV-2 (Protein Data Bank entries 6WV1, 6LZG and 6M0J). Green-shaded regions represent the receptor-binding motif (RBM), the orange-shaded region represents the RGD motif and the red/pink-shaded residues in the RBD directly interact with ACE2. **b**. Location of RGD and KGD motifs in S-RBDs of SARS-CoV-2 (orange ribbon) and SARS-CoV (cyan ribbon) and the KGD motif in ACE2 (violet and green ribbons). Residues are shown in ball-and-stick representation with carbons in orange for SARS-CoV-2, cyan for SARS-CoV, and green/violet for ACE2. Oxygen and nitrogen atoms are shown in red and blue colors, respectively. Y41 and K353 belong to ACE2. **c**. A proposed model for binding two S trimers with ACE2 based on the SARS-CoV-2 S-RBD/B^{0}AT1/ACE2 complex structure. The S-RBDs of the 2 trimers were superposed on the S-RBD in the cryoEM structure of the S-RBD/B^{0}AT1/ACE2 complex. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

group of correlated mutations includes S13I, W152C and L452R (Fig. 4**b**). The CAL.20C variant has four signature mutations (S13I, W152C, L452R and D614G) in the S protein (Table 1 **and** Fig. 2). To determine the correlation among the S protein mutations of CAL.20C, we excluded P323L and C241U mutations from the correlation analyses. The results in Fig. 4**f** clearly show that all four S protein mutations are strongly correlated. We also included additional S protein mutations in our analyses based on data presented in Fig. 2. The results show a positive correlation of N501Y with D614G, suggesting that these mutations coexist in some viruses in the United States. Another variant in the U. S. may emerge that includes D614G and N501Y mutations.

As expected, all the mutations in B.1.351 (SA variant) are correlated except D614G, which is clustered with P323L (Fig. 4**c**). We re-analyzed the SA sequences after excluding P323L and C241U mutations to gain insight into the correlation among S protein mutations. The correlation analysis presented in Fig. 4**g** reveals that mutations in the S protein of the B.1.351 variant are positively correlated except for D614G and R246I, which are clustered together. These results indicate a complex evolution of the mutations in B.1.351 S protein. Considering that a majority of viruses

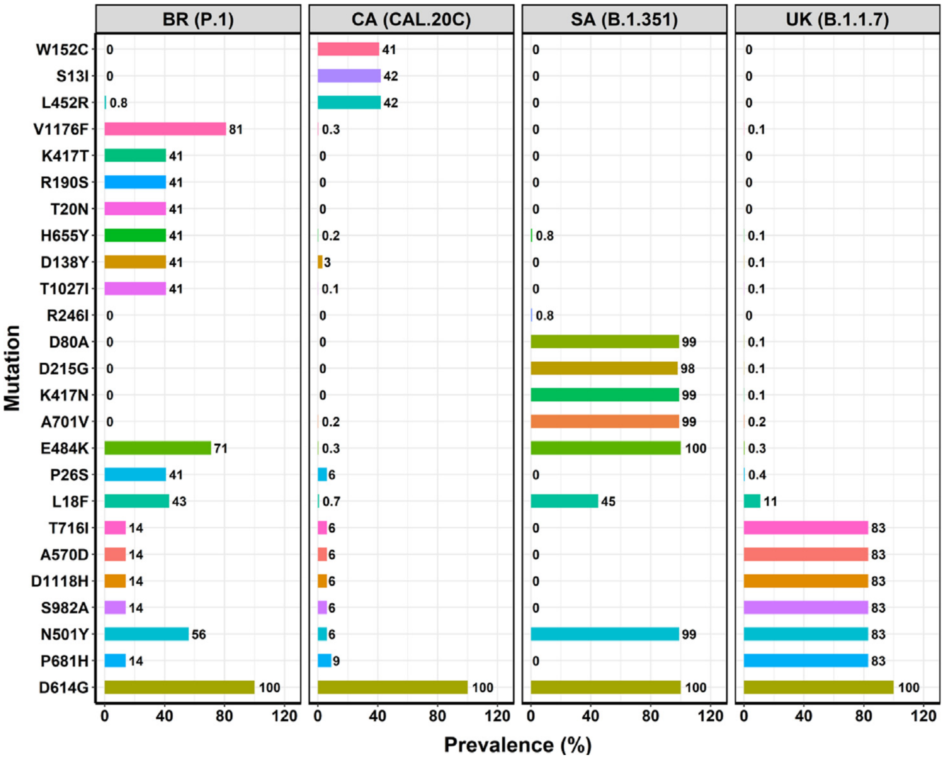

Fig. 2. Geo-prevalence of mutations within S protein of four variants. Additional mutations were added to this figure after we noted that these mutations were not reported previously for these variants. For example, the CAL.20C variant has been characterized by S13I, W152C, L452R and D614G mutations. However, we noted additional mutations in this region (*e. g.* P681H ~ 9%). Also, previous characterization of the P.1 variant did not include V1176F. Our results showed this mutation was at high prevalence (~81%). Therefore, this mutation is included in the figure.

contain D614G, it appears that R246I emerged separately from other B.1.351 mutations.

The correlation analysis of the P.1 (Brazil) variant mutations is extremely complex (Fig. 4**d**). There are at least 6 groups of correlated mutations. While D614G and P323L are positively correlated, C241U does not cluster with D614G and P323L. Instead, C241U shows a positive correlation with N501Y. To better understand the correlation of mutations in the S protein of P.1, we excluded P323L and C241U and re-analyzed the correlation of mutations (Fig. 4**h**). There are two major groups of positively correlated mutations. The first group includes D614G, N501Y, L18F, D138Y, and K417T, whereas the second group of correlated mutations includes T20N, P26S, R190S and T1027I. Additionally, the V1176F mutation is positively correlated with R190S. Considering that V1176F is not a signature mutation in the P.1 variant, a different and probably more complex variant may emerge from P.1. While this correlation analysis shows other degrees of correlation in S protein mutations, this analysis does not provide an evolutionary relationship among variants, indicating that specific variants may not be solely due to mutations in the S protein.

We also determined relative abundance (RA) of S protein mutations within a given variant for three previously reported co-existing mutations (D614G, P323L and C241U) in the United States [19]. The results of this analysis shown in **Supplementary Fig. S3** confirmed that these three mutations co-existed primarily in the United States [19] as the three mutations did not exist at the same frequencies in the B.1.1.7, B.1.351 and P.1 variants (**Supplementary Fig. S3**). Our sequence analysis also revealed that the four variants had additional mutations than the reported Spike mutations in these variants. For example, the prevalence of P681H was 14% in P.1 and 9% in CAL.20C (Fig. 2). In addition, the V1176F mutation at a prevalence of 81% co-exists with the signature mutation E484K in P.1 (Fig. 2).

2.4. Prevalence of CAL.20C in different states of the United States

We used signature mutations of CAL.20C to map the prevalence of this variant in the United States. As of June. 5, 2021, all states in the USA had CAL.20C infections. A map of the CAL.20C distribution in the USA is shown in Fig. 5. A time sampling of CAL.20C sequences deposited to GISAID showed that California had the maximum prevalence of CAL.20C (~32%) followed by Oregon (~25%) > South Carolina (~19%) > North Dakota = District of Columbia (~17%) > Arizona = Mississippi (15%) > New Mexico (~12%) > W ashington (~10%) until Feb. 24, 2021. However, this trend changed by June 5, 2021. Both California and Hawaii had comparable prevalence of CAL.20C variant (~31%) followed by Nevada (~25%) > Was hington (~15%) > Montana (~14%) > Colorado = Arizona = Oregon (~12%) > North Dakota (~10%). A high prevalence of CAL.20C in these states is not surprising as this variant can migrate easily to nearby locations except Montana and North Dakota, which are somewhat far from California. Our analysis also showed that, in general, the prevalence of L452R was greater than S13I and W152C in many states. It is possible that another variant containing L452R is evolving.

Four S protein mutations of CAL.20C were present in California as early as April 11, 2020 (GISAID sequence # EPI_ISL_2304313) [45]. To evaluate how CAL.20C spread across the United States, we aligned first collected and dated CAL.20C sequence from each state. We then grouped the sequences based upon the date collected and percent homology cut-offs of 96, 94, 93, 92, 91 and 90 with the previous group, except the first CAL.20C sequence from California, which has a homology of 73% with Wuhan-Hu-1. Thus, the sequence homologies of the first sequence of CAL.20C variants from UT, NM, NV and TX with the CAL.20C sequence were 96.7, 96.3, 97.3 and 96.1, respectively. Using this data, we generated a Sankey diagram showing the dynamics of the the CAL.20C variant

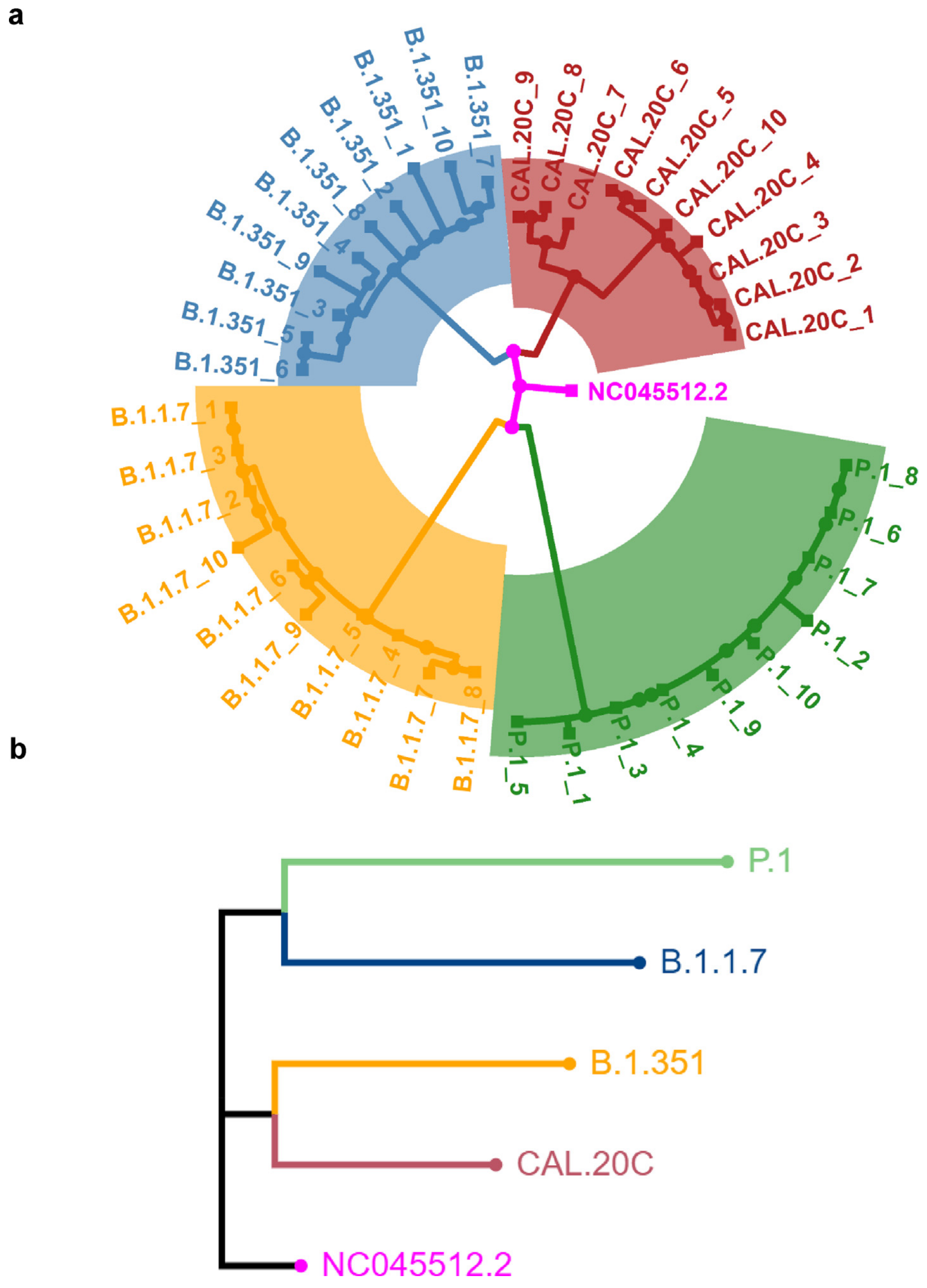

Fig. 3. Phylogenetic analysis of variant sequences. Panel a shows the phylogenetic analysis of the first 10 sequences obtained from the four variants B.1.1.7, B.1.351, CAL.20C and P.1. These variants B.1.1.7, B.1.351, CAL.20C and P.1 are shaded and labeled in orange, light blue, dark red and forest green, respectively. Since the entry names in GISAID are too long, we have renamed them sequences 1 through 10 for each variant. The list of sequences and GISAID identification numbers of the sequences are provided in Supplementary Table 1. The sequence alignment was generated using MAFFT (version 7) [49] and MEGA X [50]. The Nearest Neighbor End joining method was used to generate the intial tree. The final tree was generated using R package ggtree. Panel b shows a phylogenetic tree of the consensus sequences generated from the four variants. The tree was generated using the methods used for Panel 3a. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

in the United States (Fig. 6). This figure shows that, in general, CAL.20C variant migrated from the west coast to east coast over a period of 10 months. The data also showed that all but two states had CAL.20C within 6 months (Sept. 2020 – Feb. 2021) indicating a high transmissibility of this variant.

2.5. Structural implications of variant-specific mutations

As of May 27, 2021, 400 structures of S protein either in the apo form or in complex with different proteins (*e.g.*, antibody, nanobody, ACE2 or vaccine candidates) have been deposited in the Protein Data Bank (RCSB.org). These 116 structures have been solved by X-ray crystallography, whereas Cryo-Electron Microscopy (CryoEM) determined the rest (284). Four of the 400 deposited structures contain the N501Y mutation (PDB files 7NXC [16], 7MJN, 7MJM and 7MJG [17]), and one structure solved by X-ray crystallography (PDB file 7NXC [16]) contains three mutations (K417T, E484K and N501Y) in the S-RBD of the P.1 variant. The three structures with only the N501Y mutation (PDB files 7MJN, 7MJM and 7MJG [17]) were solved by CryoEM and represent an S-RBD/ACE2 complex (PDB file 7MJN), an S protein/ACE2 complex (PDB file 7MJM) and an apo S protein (PDB file 7MJG).

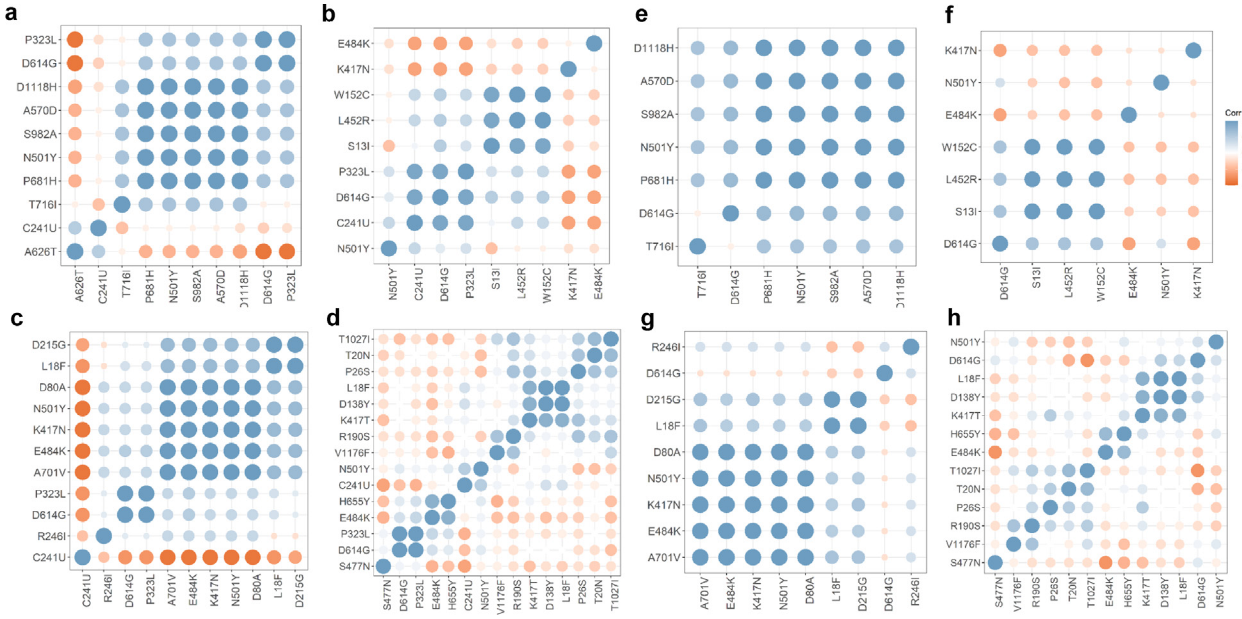

Fig. 4. Correlation analyses of mutations in four variants. Panels **a, b, c and d** show the analyses of mutations in S protein together with D614G, P323L and C241U(T) from B.1.1.7, CAL.20C, B.1.351 and P.1 variants, respectively. Panels **e, f, g and h** show the correlation analyses of mutations in S protein from B.1.1.7, CAL.20C, B.1.351 and P.1 variants, respectively. The additional mutations, V1176F in P.1 and N501Y, E484K and K417N in CAL.20C, were included in our analyses as these mutations were significantly prevalent in respective variants.

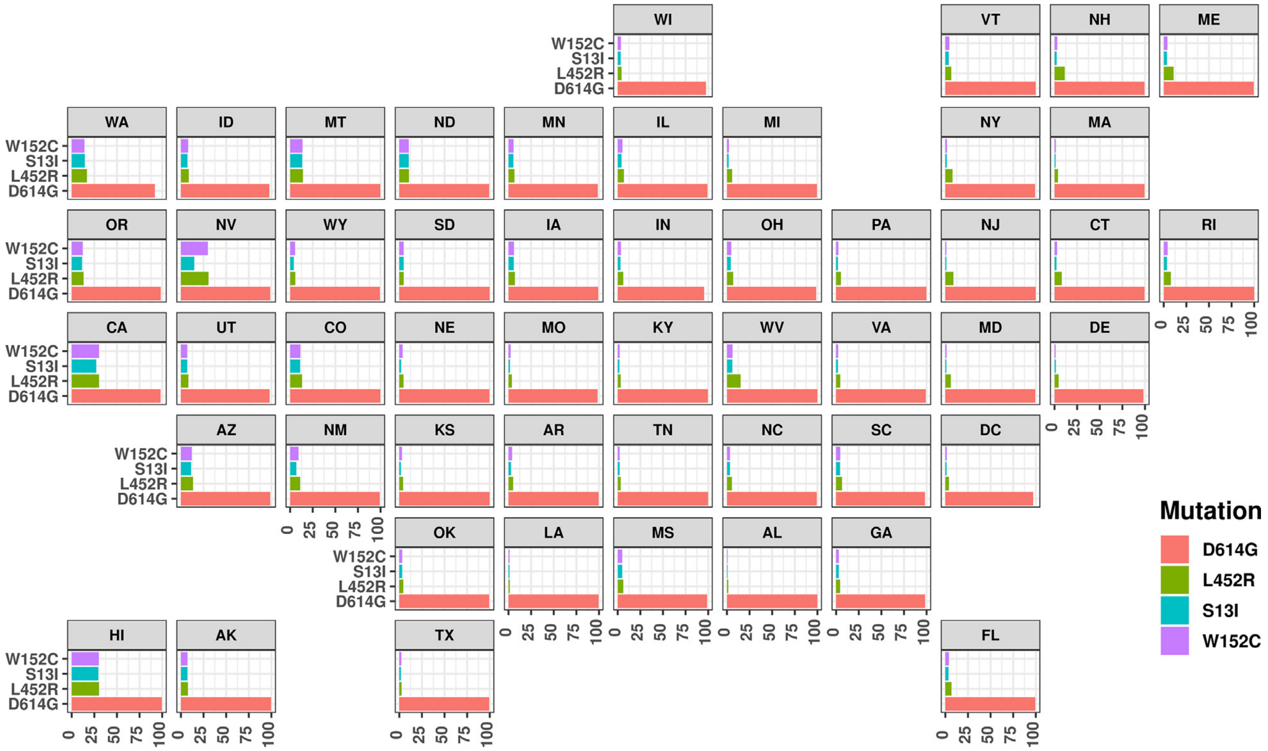

Fig. 5. Prevalence of CAL.20C in different states of the United States. Four signature mutations, D614G, S13I, W152C and L452R, were used to generate this map. This figure was generated by an R code that utilizes 'geofacet' and 'ggplot2' libraries.

The S-RBD derived from the P.1 variant structure containing three mutations showed an ~ 19-fold enhanced binding affinity to ACE2 compared to WT S-RBD binding to ACE2 [16]. We first compared the crystal structures of the WT S-RBD/ACE2 complex (PDB file 6M0J [9]) with the P.1 S-RBD/ACE2 complex (PDB file 7NXC) [16]. This comparison showed an overall root-mean-square-deviation of 0.56 Å of Cα of the common structure, suggesting that the P.1 S-RBD mutations do not cause significant structural changes in the S protein. This comparison showed that Y501 is inserted into a cavity at the binding interface and forms a stacking interaction with Y41 of ACE2. The same observation was noted in the CryoEM structures of ACE2/S-RBD or ACE2/S protein complexes (PDB files 7MJN and 7MJM, respectively). Additionally, the mutations K417N and E484K did not generate any new interactions with

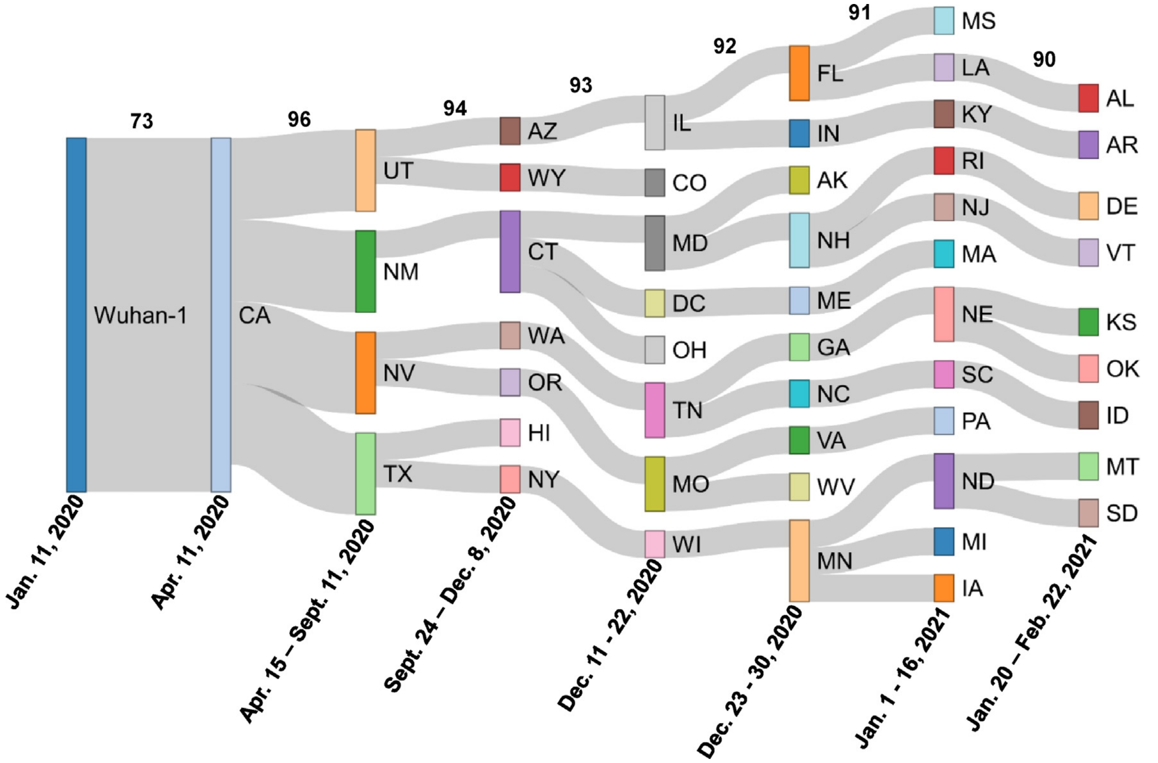

Fig. 6. Sankey diagram showing the dynamics of CAL.20C in the United States. To generate the Sankey diagram we aligned the first collected and dated CAL.20C sequence from each state. We then grouped the sequences based upon the date collected and percent homology cut-offs of 96, 94, 93, 92, 91 and 90 with the previous group. Thus, the sequence homologies of the first sequence of CAL.20C variants from UT, NM, NV and TX with the CAL.20C sequence were 96.7, 96.3, 97.3 and 96.1, respectively. This figure shows that, in general, CAL.20C variant migrated from the west coast to east coast over a period of 10 months.

ACE2. On the contrary, a salt-bridge between D30 (ACE2) and K417 (WT S-RBD) was lost with the mutation K417T (as in P.1) or K417N (as in B.1.351). Therefore, it appears that the increase in binding affinity of P.1 S-RBD to ACE2 may be due to the newly generated hydrophobic interactions caused by the N501Y mutation (Fig. 7), which overcomes the loss of a salt-bridge.

Recently, it was reported that a pseudovirus mimicking B.1.351 did not confer increased infectivity in multiple cell types except for murine ACE2-overexpressing cells [46]. However, another report showed that a mouse-adapted strain (MASCp6) had increased infectivity in mouse lung and led to interstitial pneumonia and inflammatory responses in both young and aged mice after intranasal inoculation [34]. Therefore, it appears that other factors, such as the stability of the S protein, may contribute to the transmission of the variants. To gain insight into the stability of the S protein, we calculated the change in free energy ($\Delta\ \Delta$ G) between the wild-type and mutant protein, as described by Pandurangan *et al.* [47]. The calculated $\Delta\ \Delta$ G values for a majority of the signature mutations are shown in Fig. 8. A positive value of $\Delta\ \Delta$ G indicates an increase in protein stability, whereas a negative $\Delta\ \Delta$ G suggests a decrease in protein stability. D614G (the globally dominant virus variant) has the highest $\Delta\ \Delta$ G, suggesting that G614 S protein is more stable than D614 S protein. K417N, part of the B.1.351 variant, has the lowest $\Delta\ \Delta$ G, which indicates a decrease in protein stability. While the computation of $\Delta\ \Delta$ G does not provide a definitive answer regarding the overall stability of S protein, it does demonstrate an increased stability of S protein with specific mutations.

3. Summary and conclusion

Here, we present an analysis of SARS-CoV-2 variant sequences to shed light on the evolution of the variants. Based on the analyses of available sequences of new SARS-CoV-2 variants, it is evident that these variants emerged from common ancestors. It is interesting to note that B.1.1.7 and P.1 are genetically related. It is possible that the ancestor virus emerged at a third location, followed by its travel to U.K. and Brazil. Similarly, the CAL.20C and B.1.351 are genetically related. A similar scenario of evolution of these two variants can also be imagined. However, given short period, volume of variants and limited travel around the world, it is safe to postulate that new variants have greater infectivity than the original (Wuhan-Hu-1) virus. A recent deadly surge caused by variant B.1.617.2 or the Delta variant (although not included in this study) in India provides additional justification for the increased infectivity of new variants.

Although our correlation analysis indicates that variant-specific mutations in the S protein of a given variant do not reflect evolution of the variant, but may increase variant infectivity, more experimental data are needed to validate these observations. Nonetheless, the mutations in S protein may play a role in the greater infectivity of SARS-CoV-2 variants compared to the original Wuhan-Hu-1 virus (GenBank accession # NC_045512).

As with all RNA viruses, the resultant evolved virus goes through quasispecies [48] in which the viral populations consist of mutant spectra (or mutant clouds) rather than genomes with the same nucleotide sequence. It is possible that SARS-CoV-2 genomic sequences rapidly expand in sequence space and lose biological information that enables the elimination of fitness-compromised genomes. Subsequent mutations or mutant spectra in the quasispecies of SARS-CoV-2 are most likely generated to circumvent reduced viral fitness for adaptability. This is because the quasispecies mutations constitute dynamic (continuously changing) repositories of genotypic and phenotypic viral variants [48]. Thus, the new variants may have evolved through these strategies and may still be evolving. In the end, we would like

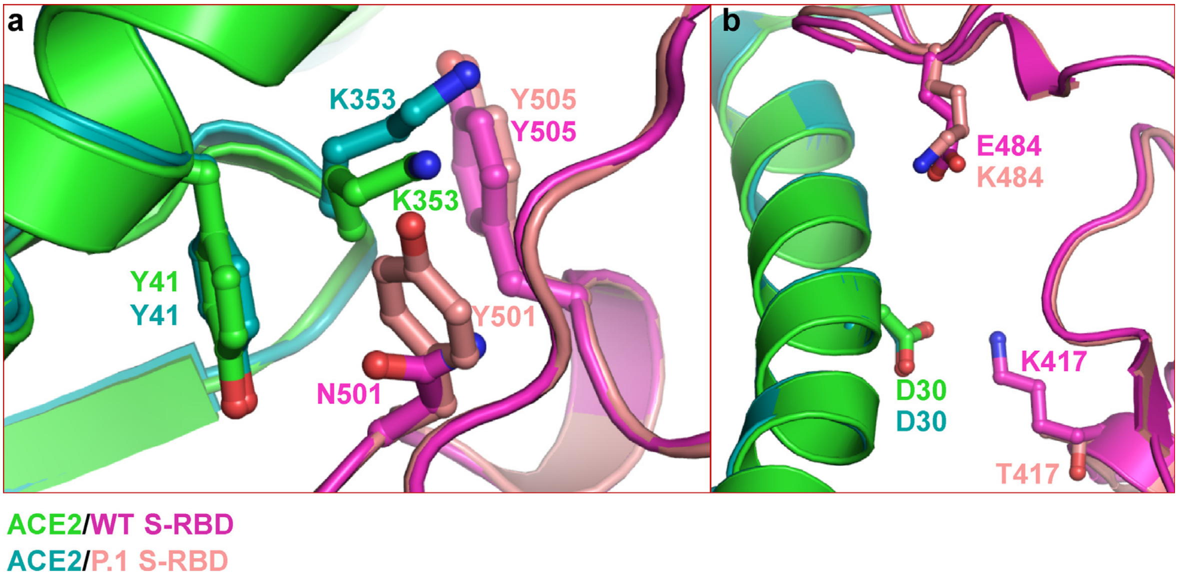

Fig. 7. Structural implications of mutations in different variants. This figure depicts two examples of mutations that may enhance the binding of S protein and ACE2 (N501Y; panel a) or have no effect (K417T; panel b). Additionally, the solved structure does not have P.1 signature mutation L452R. Therefore, it is not shown in this figure.

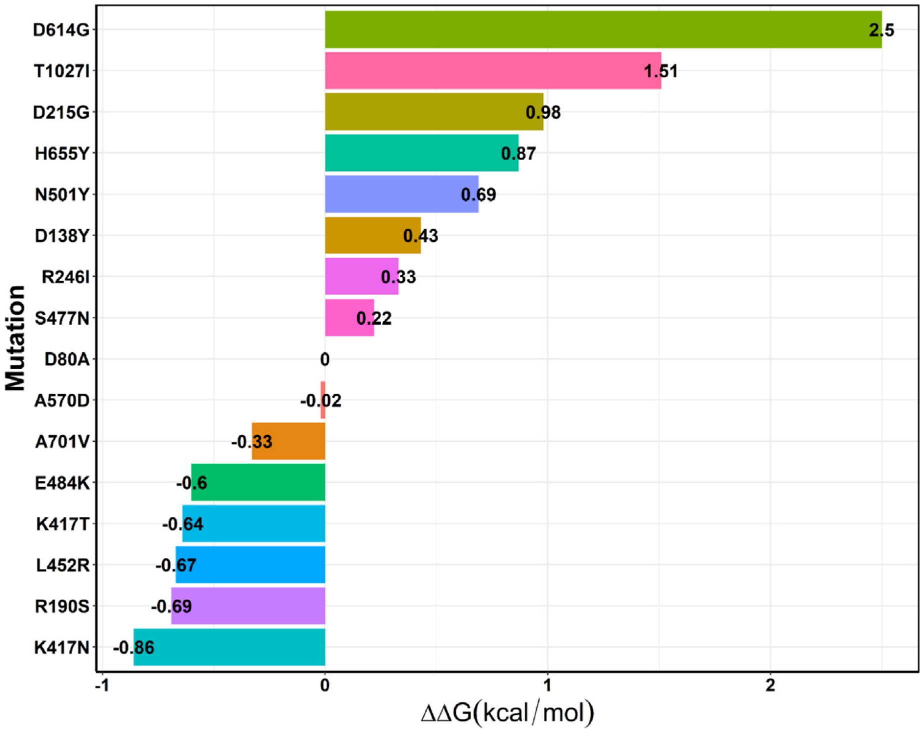

Fig. 8. Computed free energy change (Δ Δ G) upon mutation. PDB file 6XM4 was used for Δ Δ G computation, except for N501Y, S477N and E484K, for which PDB file 6M0J [9] was used.

to caution that the sequences of viruses are being deposited at an unprecedented rate. It is possible that some conclusions may change as more sequences of variants become available. For example, the GISAID repository accessed late Feb. 2021 showed that the first available sequence of CAL.20C was collected on July 7, 2020 (GISAID ID hCoV-19/USA/CA-LACPHL-AE00055/2020|EPI_ISL_765994|2020-07-07), whereas the current GISAID data (as of June 21, 2021) show that the first sequence of CAL.20C was collected on April 11, 2020.

4. Material and Methods.

4.1. Sequence acquisition, alignment and analysis

The prevalence of each mutation in B.1.1.7, B.1.351, CAL.20C and P.1 variants was obtained from the GISAID repository [45]. These sequences were aligned using the MAFFT [49], MEGA X [50] or JalView [51] sequence alignment programs. The phylogenetic tree was constructed using the maximum likelihood tree construction method through MEGA X [52] with 500 Bootstrap steps. The final phylogenetic trees shown in Fig. 3a and 3b were generated by the 'ggtree' package of the R programming language [53]. Relative abundance (RA) of mutations in S protein from different geographic regions was determined by an in-house Python script using the scikit-learn (Python) library [54] and plotted with R (codes available upon request). Final values were multiplied by 100 and expressed as percentages. We also used an R package 'corrplot' to create an in-house R script for plotting purpose. The sequence IDs corresponding to the tip labels on the phylogenetic tree are given in **Supplementary Table T1**. All scripts are available upon request. Additional phylogenetic analysis was conducted using sequences (n = 6,052) from each variant: B.1.1.7 sequences (n = 1,637) collected from January 1, 2020 to Nov 25, 2020, B.1.351 sequences (n = 1,069) collected from January 1, 2020 to March 1, 2021, P.1 sequences (n = 1,926) collected from January 1, 2021 to March 1, 2021, and CAL.20C sequences (n = 4,182) collected from January 1, 2020 to January 15, 2021. These sequences were aligned using the online MAFFT alignment server along with the reference sequence (NCBI: NC_045512.2). The resultant alignment was submitted into NextStrain's NextClade [44], which generated a Phylogeny using the sequences. The resulting phylogeny was filtered to highlight the four variants (B.1.1.7, CAL.20C, B.1.351, and P.1) (**Supplementary Fig. S2**).

4.2. Prevalence of CAL.20C and time-sampled spread of this variant.

An in-house Python script was used to compute the prevalence of CAL.20C within the United States. The visualization of prevalence of the mutations in different states was generated using R package 'geofacet'. The Sankey diagram was generated by first assembling the first sequence of CAL.20C variant from each state of the U.S. based on date collected. The pairwise sequence homology was calculated using the EMBL-EBI search and sequence analysis tools [55]. For the Sankey diagram, the sequences were grouped according to date of collection (shown at the bottom) and the percent sequence homology (nodes) of percent homologies of 96, 94, 93, 92, 91 and 90 as shown at the top of Fig. 6. The figure was generated using R. The Metadata associated with Sankey diagram is provided in **Supplementary Table T3**.

4.3. Structural analysis

An in-house R program was written to retrieve sequences from the Protein Data Bank (www.rcsb.org). Specific structures were extracted using the 'grepl' function of R package 'dplyr'. The structures were then downloaded and analyzed using either the Schrodinger Suite (Schrodinger LLC, NY) or PyMol [56]. Fig. 7 was generated using PyMol. The free energy change upon mutations in S protein ($\Delta\ \Delta$ G) was computed through the SDM server as detailed by Pandurangan *et al.* [47].

4.4. Metadata summary

All sequences used in this study were extracted from GISAID except Wuhan-Hu-1, which was obtained from National Center for Biotechnology Information (NCBI, accession number NC_045512.2). Sequences were aligned and analyzed through MEGA X or MAFFT multiple sequence analysis programs. The quality of the sequences for gaps and unidentified nucleotides (N's in place of nucleotides) was verified by an in-house Python script. Any sequence that contained more than 0.1% N's was excluded from the analysis. The overall sequence of S protein gene was 3,821 nucleotides and the average sequence length of variant genomes used for phylogenetic analysis was 29,907 nucleotides. Apart from the analyses for Fig. 2., Fig. 5., and Fig. 6, the sequences were not restricted to specific geological region. Instead, they were collected irrespective of the country where the variant emerged.

Author contributions

Conceptualization: KS, ANS and SRK the study; Data curation: KS, ANS, SRK; Formal analysis: KS, AS, GAW, and LTW; Investigation: KS, ANS, SRK and LTW; Methodology: KS, ANS and SRK; Project administration: KS, SNB, and AS; Resources: KS, SNB, AS; Software: KS, ANS and SRK; Supervision: KS; Validation: KS and SNB and GAW; Visualization: KS; Writing - original draft: KS and SRK; KS, TPQ, AS, SNB, CLL and GAW.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. CLL is cofounder of Shift Pharmaceuticals but that has not influenced the work reported in this paper.

Acknowledgments

KS acknowledges support from the Office of Research, University of Missouri (Bond Life Sciences Center, Early Concept Grant). AS acknowledges funding from the Swedish Research Council (2016-01675). LTW, GAW and KS acknowledge support from the National Institute of Dental and Craniofacial Research Grant DE007389 and the American Lung Association. KS and TPQ acknowledge the computation facilities of the Molecular Interaction Core at the University of Missouri, Columbia, MO 65212. We also thank the laboratories that have submitted the virus sequences to the GISAID.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.06.037.

References

[1] Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The architecture of SARS-CoV-2 transcriptome. Cell 2020;13:914–21. https://doi.org/10.1016/j.cell.2020.04.011.
[2] Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, et al. Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. J Mol Biol 2003;331:991–1004.
[3] Masters PS. The molecular biology of coronaviruses. Adv Virus Res 2006;66:193–292.
[4] Ziebuhr J. Molecular biology of severe acute respiratory syndrome coronavirus. Curr Opin Microbiol 2004;7:412–9.
[5] Yan L, Ge J, Zheng L, Zhang Y, Gao Y, Wang T, et al. Cryo-EM structure of an extended SARS-CoV-2 replication and transcription complex reveals an intermediate state in cap synthesis. Cell 2021;184(184–193):e110.
[6] Dai L, Gao GF. Viral targets for vaccines against COVID-19. Nat Rev Immunol 2021;21:73–82.
[7] Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 2020;181(281–292):e286.
[8] Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 2020;367:1260–3.
[9] Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature 2020;581:215–20.
[10] Li F, Li W, Farzan M, Harrison SC. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. Science 2005;309:1864–8.
[11] Zhang Q, Chen CZ, Swaroop M, Xu M, Wang L, Lee J, et al. Heparan sulfate assists SARS-CoV-2 in cell entry and can be targeted by approved drugs in vitro. Cell Discov 2020;6:80.

[12] Wu K, Li W, Peng G, Li F. Crystal structure of NL63 respiratory coronavirus receptor-binding domain complexed with its human receptor. Proc Natl Acad Sci U S A 2009;106:19970–4.
[13] Wang Q, Zhang Y, Wu L, Niu S, Song C, Zhang Z, et al. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. Cell 2020;181(894–904): e899.
[14] Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, et al. Structural basis of receptor recognition by SARS-CoV-2. Nature 2020;581:221–4.
[15] Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science 2020;367:1444–8.
[16] Dejnirattisai W, Zhou D, Supasa P, Liu C, Mentzer AJ, Ginn HM, et al. Antibody evasion by the p. 1 strain of SARS-CoV-2. Cell 2021.
[17] Zhu X, Mannar D, Srivastava SS, Berezuk AM, Demers JP, Saville JW, et al. Cryo-electron microscopy structures of the n501y SARS-CoV-2 spike protein in complex with ACE2 and 2 potent neutralizing antibodies. PLoS Biol 2021;19: e3001237.
[18] Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. Cell 2020;182(812–827):e819.
[19] Kannan SR, Spratt AN, Quinn TP, Heng X, Lorson CL, Sonnerborg A, et al. Infectivity of SARS-CoV-2: There is something more than D614G?. J Neuroimmune Pharmacol 2020;15:574–7.
[20] Wang R, Chen J, Gao K, Hozumi Y, Yin C, Wei GW. Analysis of SARS-CoV-2 mutations in the united states suggests presence of four substrains and novel variants. Commun Biol 2021;4:228.
[21] Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, et al. Spike mutation D614G alters SARS-CoV-2 fitness. Nature 2021;592:116–21.
[22] Daniloski Z, Jordan TX, Ilmain JK, Guo X, Bhabha G, tenOever BR, et al. The spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types. Elife 2021;10.
[23] Li Q, Wu J, Nie J, Zhang L, Hao H, Liu S, et al. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. Cell 2020;182(1284–1294): e1289.
[24] Sun L, Li P, Ju X, Rao J, Huang W, Ren L, et al. In vivo structural characterization of the SARS-CoV-2 RNA genome identifies host proteins vulnerable to repurposed drugs. Cell 2021;184(1865–1883):e1820.
[25] Jia, Y., G.S. Nguyen, S., Zhang, Y., Huang, K.-S., Ho, H.-Y., Hor, W.-S., Yang, C.-H., Bruning, J.B., Li, C., Wang, W.-L. Analysis of the mutation dynamics of SARS-CoV-2 reveals the spread history and emergence of rbd mutant with lower ACE2 binding affinity. BioRxiv; 2021.
[26] Yao H, Lu X, Chen Q, Xu K, Chen Y, Cheng M, et al. Patient-derived SARS-CoV-2 mutations impact viral replication dynamics and infectivity in vitro and with clinical implications in vivo. Cell Discov 2020;6:76.
[27] Rambaut, A., Loman, N., Pybus, O., Barclay, W., Barrett, J., Carabelli, A., Connor, T., Peacock, T., Robertson, D.L., Volz, E., CoG-UK, o.b.o. Preliminary genomic characterisation of an emergent SARS CoV-2 lineage in the uk defined by a novel set 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 2020, Accessed on May 26, 2021.
[28] Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. Nature 2021;592:438–43.
[29] Faria, N.R., Mellan, T.A., Whittaker, C., Claro, I.M., Candido, D.D.S., Mishra, S., et al., Genomics and epidemiology of the p.1 SARS-CoV-2 lineage in manaus, brazil. Science; 2021, 372, 815-821.
[30] Zhang W, Davis BD, Chen SS, Sincuir Martinez JM, Plummer JT, Vail E. Emergence of a novel SARS-CoV-2 variant in southern california. JAMA 2021;325:1324–6.
[31] Wang, P., Nair, M.S., Liu, L., Iketani, S., Luo, Y., Guo, Y., et al., Increased resistance of SARS-CoV-2 variants b.1.351 and b.1.1.7 to antibody neutralization. bioRxiv 2021.
[32] Greaney AJ, Starr TN, Gilchuk P, Zost SJ, Binshtein E, Loes AN, et al. Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. Cell Host Microbe 2021;29(44–57):e49.
[33] Yi C, Sun X, Ye J, Ding L, Liu M, Yang Z, et al. 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 2020;17:621–30.
[34] Gu H, Chen Q, Yang G, He L, Fan H, Deng YQ, et al. Adaptation of SARS-CoV-2 in balb/c mice for testing vaccine efficacy. Science 2020;369:1603–7.
[35] Turner AJ, Tipnis SR, Guy JL, Rice G, Hooper NM. Aceh/ACE2 is a novel mammalian metallocarboxypeptidase and a homologue of angiotensin-converting enzyme insensitive to ace inhibitors. Can J Physiol Pharmacol 2002;80:346–53.
[36] Bellomo R, Wunderink RG, Szerlip H, English SW, Busse LW, Deane AM, et al. Angiotensin i and angiotensin ii concentrations and their ratio in catecholamine-resistant vasodilatory shock. Crit Care 2020;24:43.
[37] Gheblawi M, Wang K, Viveiros A, Nguyen Q, Zhong JC, Turner AJ, et al. Angiotensin-converting enzyme 2: SARS-CoV-2 receptor and regulator of the renin-angiotensin system: Celebrating the 20th anniversary of the discovery of ACE2. Circ Res 2020;126:1456–74.
[38] Aroor AR, Demarco VG, Jia G, Sun Z, Nistala R, Meininger GA, et al. The role of tissue renin-angiotensin-aldosterone system in the development of endothelial dysfunction and arterial stiffness. Front Endocrinol (Lausanne) 2013;4:161.
[39] Aroor AR, McKarns S, Demarco VG, Jia G, Sowers JR. Maladaptive immune and inflammatory pathways lead to cardiovascular insulin resistance. Metabolism 2013;62:1543–52.
[40] Danilczyk U, Sarao R, Remy C, Benabbas C, Stange G, Richter A, et al. Essential role for collectrin in renal amino acid transport. Nature 2006;444:1088–91.
[41] Malakauskas SM, Quan H, Fields TA, McCall SJ, Yu MJ, Kourany WM, et al. Aminoaciduria and altered renal expression of luminal amino acid transporters in mice lacking novel gene collectrin. Am J Physiol Renal Physiol 2007;292:F533–544.
[42] Kowalczuk S, Broer A, Tietze N, Vanslambrouck JM, Rasko JE, Broer S. A protein complex in the brush-border membrane explains a hartnup disorder allele. FASEB J 2008;22:2880–7.
[43] Fairweather SJ, Broer A, O'Mara ML, Broer S. Intestinal peptidases form functional complexes with the neutral amino acid transporter b(0)at1. Biochem J 2012;446:135–48.
[44] Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: Real-time tracking of pathogen evolution. Bioinformatics 2018;34:4121–3.
[45] Elbe S, Buckland-Merrett G. Data, disease and diplomacy: Gisaid's innovative contribution to global health. Glob Chall 2017;1:33–46.
[46] Li Q, Nie J, Wu J, Zhang L, Ding R, Wang H, et al. SARS-CoV-2 501y.V2 variants lack higher infectivity but do have immune escape. Cell 2021;184(2362–2371):e2369.
[47] Pandurangan AP, Ochoa-Montano B, Ascher DB, Blundell TL. SDM: A server for predicting effects of mutations on protein stability. Nucleic Acids Res 2017;45: W229–35.
[48] Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. Microbiol Mol Biol Rev 2012;76:159–216.
[49] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 2013;30:772–80.
[50] Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA x: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 2018;35:1547–9.
[51] Troshin PV, Procter JB, Barton GJ. Java bioinformatics analysis web services for multiple sequence alignment–jabaws:MSA. Bioinformatics 2011;27:2001–2.
[52] Hall BG. Building phylogenetic trees from molecular data with MEGA. Mol Biol Evol 2013;30:1229–35.
[53] Yu G. Using ggtree to visualize data on tree-like structures. Curr Protoc Bioinformat 2020;69:e96.
[54] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in python. J Mach Learn Res 2011;12:2825–30.
[55] Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The embl-ebi search and sequence analysis tools apis in 2019. Nucleic Acids Res 2019;47:W636–41.
[56] DeLano, W.L. An open-source molecular graphics tool. CCP4 newsletter on protein crystallography; 2002, 40, 82-92.