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

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

[1]Department of Biochemistry, Microbiology & Immunology, Faculty of Medicine, University of Ottawa, Ontario, Canada K1H 8M5.

[2]Department of Cellular & Molecular Medicine, Faculty of Medicine, University of Ottawa, Ontario, Canada K1H 8M5.

[3]Human Health Therapeutics Research Centre, National Research Council Canada, Ottawa, Canada K1A 0R6

[4]uOttawa Center for Infection, Immunity and Inflammation (CI3).

Running title: **Genome Evolution of SARS-CoV-2.**

*Correspondence should be addressed to: langlois@uottawa.ca

ABSTRACT

The novel Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) emerged in late December 2019 in Wuhan, China, and is the causative agent for the worldwide COVID-19 pandemic. SARS-CoV-2 is a 29,811 nucleotides positive-sense single-stranded RNA virus belonging to the betacoronavirus genus. Due to inefficient proofreading ability of the viral RNA-dependent polymerase complex, coronaviruses are known to acquire new mutations following replication, which constitutes one of the main factors driving the evolution of its genome and the emergence of new genetic variants. In the last few months, the identification of new B.1.1.7 (UK), B.1.351 (South Africa) and P.1 (Brazil) variants of concern (VOC) highlighted the importance of tracking the emergence of mutations in the SARS-CoV-2 genome and their impact on transmissibility, infectivity, and neutralizing antibody escape capabilities. These VOC demonstrate increased transmissibility and antibody escape, and reduce current vaccine efficacy. Here we analyzed the appearance and prevalence trajectory of mutations that appeared in all SARS-CoV-2 genes from December 2019 to January 2021. Our goals were to identify which modifications are the most frequent, study the dynamics of their spread, their incorporation into the consensus sequence, and their impact on virus biology. We also analyzed the structural properties of the spike glycoprotein of the B.1.1.7, B.1.351 and P.1 variants. This study offers an integrative view of the emergence, disappearance, and consensus sequence integration of successful mutations that constitute new SARS-CoV-2 variants and their impact on neutralizing antibody therapeutics and vaccines.

IMPORTANCE

SARS-CoV-2 is the etiological agent of COVID-19, which has caused > 2 million deaths worldwide as of January, 2021. Mutations occur in the genome of SARS-CoV-2 during viral replication and affect viral infectivity, transmissibility and virulence. In early March 2020, the D614G mutation in the spike protein emerged, which increased the viral transmissibility and is now found in >90% of all SARS-CoV-2 genomic sequences in GISAID database. Between October and December 2020, B.1.1.7 (UK), B.1.351 (South Africa) and P.1 (Brazil) variants of concern (VOCs) emerged, which have increased neutralizing antibody escape capabilities because of mutations in the receptor binding domain of the spike protein. Characterizing mutations in these variants is crucial because of their effect on adaptive immune response, neutralizing antibody therapy, and their impact on vaccine efficacy. Here we tracked and analyzed mutations in SARS-CoV-2 genes over a twelve-month period and investigated functional alterations in the spike of VOCs.

INTRODUCTION

In late December 2019, a new betacoronavirus known as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV-2) emerged in Wuhan, the province of Hubei, China (1). SARS-CoV-2 is the etiological agent for the worldwide COVID-19 pandemic resulting in more than 90 million infected and 2 million death worldwide as of Dec. 2020 (2,3). SARS-CoV-2 is an enveloped, positive-sense single-stranded RNA (+ssRNA) virus with a genome length of 26,000 to 32,000 bps (4). The mutation rates of RNA viruses are generally higher than that of DNA viruses because of the low fidelity of their viral RNA polymerases (5,6). Mutations occur when viral replication enzymes introduce errors in the viral genome resulting in the creation of premature termination codons, deletions and insertions of nucleotides that can change open reading frames, and changes in the nucleotide sequence that can result in amino acid substitutions in the viral proteins. These mutations combined with the selective pressure of the human immune system leads to the selection and evolution of viral genomes (6,7). However, coronaviruses are one of the few members of the RNA virus family that possess limited but measurable proofreading ability via the 3' to 5' exoribonuclease activity of the non-structural viral protein 14 (nsp14) (8,9). Coronaviruses are therefore expected to evolve through genetic drift much slower than other RNA viruses that do not have this ability, such as influenza and HIV (8,10). Additionally, SARS-CoV-2 and other coronaviruses have low known occurrences of recombination between family members (i.e., genetic shift), and therefore are mostly susceptible to genetic drift (11).

SARS-CoV-2 has reached pandemic status due to its presence on every continent and has since maintained a high level of transmissibility across hosts of varied ethnical and genetic backgrounds (2, 12). Moreover, SARS-CoV-2 infections have been reported to naturally infect minks, ferrets, cats, and dogs, which allows the virus to replicate in completely new hosts and mutate to produce new variants and possibly new strains (13,14). In March 2020, the now dominant D614G mutation first emerged in the spike protein (S) of SARS-CoV-2 . The S protein is present as a trimer at the surface of the viral envelope and is responsible for attachment of the virus to the human angiotensin converting enzyme 2 (hACE2), the entry receptor for SARS-CoV-2 (15). Published evidence has now shown that D614G increases viral fitness, transmissibility and viral load but does not directly affect COVID-19 pathogenicity (16,17,18,19). Additionally, emerging evidence indicates that D614G may have epistatic interactions and exacerbate the impact of several other independent mutations (19). Mutations in the S protein, and particularly in receptor binding domain (RBD), are of very high concern given that they can direct influence viral infectivity, transmissibility and resistance to neutralizing antibodies and T cell responses.

New mutations are frequently and regularly detected in the genome of SARS-CoV-2 through whole genome sequencing, however very few of these mutations make it into the viral consensus sequence. The

consensus sequence or reference strain is generally regarded as the dominant transmitted strain at a given time. This sequence is determined by aligning large numbers of recently sequenced genomes and establishing a consensus of the highest frequency nucleotide for each position in the viral genome. A genetic variant is a version of the reference strain that has acquire one or several mutations and acts as the founder for further genetic diversification. Mutations arise regularly in the reference strain, but few are longitudinally conserved. Genetic variants are therefore successful offshoots of the reference strain.

Some variants rise rapidly in frequency and then collapse and disappear, others will rise and overtake the frequency of the reference strain and become the new reference, while others acquire additional mutations and continue their upward prevalence trajectory and evolution. There are three main genetic variants that have emerged in the past few months with a sustained upward frequency trajectory. The UK and South African variants have been reported in September and October 2020, respectively (20,21,22), while the P.1/501Y.N3 (P.1) variant is a branch off of the B.1.1.28 lineage, was first detected in travelers from Brazil that landed in Japan in January 2021 (23). These variants are associated with increased resistance to neutralizing antibodies (Nabs) (24). The UK variant known as B.1.1.7/501Y.V1 (B1.1.7) is present worldwide with local transmission in Europe, China, Oceania, North and South America (20,21). The South African variant also known as B.1.352/501Y.V2 (B.1.351) has reported local transmission in the South African region, Europe, and North America (21,22). The P.1. variant has since been identified in 42% of specimens in the Amazonian city of Manaus and was detected for the first time in the U.S. at the end of January 2021 (landed in Japan in January 2021) (23). All these variants possess the N501Y mutation, which is a mutation in the RBD that is critical for the spike to interact with hACE2 (25). This mutation is reported to cause increased resistance to Nabs, increased infectivity, and virulence in animal models (26). In addition to the N501Y mutation, both the South African and Brazil variants possesses RBD mutations K417N and E484K, which are also associated with increased Nabs escape capabilities (24).

Here we present a retrospective metadata analysis study of mutations reaching higher than 1% global frequency occurring throughout the SARS-CoV-2 genome over the past year and we specifically investigate their frequency trajectory over time and their fixation into the reference sequencing using the Global Initiative on Sharing Avian Influenza Data (GISAID) (27). Additionally, we analyzed mutations in the S protein of the B.1.1.7, B.1.351 and P.1 variants and illustrated their impact on molecular interactions between the S protein and hACE2 and their potential impact on Nabs.

MATERIALS AND METHODS

Data collection and mutational analysis

Genomes uploaded to the GISAID EpiCoV^{TM} server database were analyzed from December 1st, 2019, to December 31st, 2020, with collection of viral sequences from December 1st, 2019 to January 6, 2021. The mutations were selected by being in more than 500 reported genomes in August 2020, and another selection was made in January 2021 to capture mutations with more than 4000 reported genomes. Thus, allowing us to study mutations reaching at least 1% in global frequency. We filtered through 309,962 genomes for the analysis of selected mutations. The variants hCoV-19/Wuhan, hCoV-19/D614G, B.1.1.7, B.1.351, and P.1 were analyzed from December 1st, 2019 to February 17th, 2021. The collection dates were the same dates as the analysis. For the analysis of the mutations in B.1.1.7, B.1.351, and P.1 variants, we used the GISAID EpiCoV^{TM} server database and analyzed the mutations in the variants from December 1st, 2019, to February 17th, 2021. The same timeline was used for the collection dates of the viral sequences. We filtered through 429,514 genomes for the analysis of the variants. Only complete SARS-CoV-2 genomes (28 to 30 Kbps) isolated from human hosts were analyzed. MUSCLE alignment tool on UGene, and SnapGene was used to determine the nucleotide mutations and codon changes of the non-synonymous and synonymous mutations by sequence alignments with NCBI SARS-CoV-2 reference genome (NC_045512). We acknowledge all genomes uploaded to the GISAID database, and aligned genomes used for **Tables 1, 2, 3 and 4** are presented in supplementary material 1. Graphs of mutations and variants were performed on RStudio with timelines, and genomes illustrations were produced in Biorender.

Structural modeling

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

RESULTS

Identification of emerging mutations in various SARS-CoV-2 genes. Emerging mutations in the SARS-CoV-2 genome were investigated to determine the fluctuations of these mutations during a period of twelve months. We compiled mutations reported in 500> genomes in August, and 4000> genomes in early January from the GISAID database, and followed their appearance in reported SARS-CoV-2 genomes until December 31th , 2020. Genes NSP8, NSP10, NS6, NS7a, and E are not illustrated in **Figure 1** given that did not display mutation frequencies sufficiently high to meet our inclusion conditions during the periods of data collection of our study. This is also indicative that these appear to be the most conserved sequences of the SARS-CoV-2 genome. Our analysis focused on the emergence, fixation, and fading of the mutations analyzed that met our inclusion criteria. Our analysis highlights the fixation of the D614G mutation in the S protein and the P323L in the RdRp (**Fig.1**). Both are the only mutations to have been successful to reach reference strain status until now. They appeared to have emerged simultaneously in January, 2020 and became present in >90% of all sequenced genomes by June, 2020. However, some other mutations emerged rapidly and then stabilized or faded out. For example, Q57H (NS3), R203K (N), G204R (N) are mutations that emerged rapidly and appeared to have stabilized at a frequency of 15% to 40%. Others like I120F (NSP2), L37F (NSP6), S477N (S), and L84S (NS8) illustrate mutations that emerged rapidly and then faded-out. We also demonstrate that most genes in SARS-CoV-2 have mutations with frequencies lower than 10% (**Fig.1)**. Also, these mutations are summarized into **Table 1**, which illustrates nucleotide substitution producing the amino acid change, the frequency at the end of 2020 and their respective effects. (**Table 1**). These results indicate that only D614G (S) and P323L (NSP12) were fixed in the viral consensus sequence, many of the mutations either faded out or emerged and then stabilized at a frequency lower than 50%.

Geographic localization and timeline of the viral genes with mutations higher than 50% frequency. To further study the mutations that have frequencies higher than 50%, we took the graphs of NSP12, S, and N genes from **Figure 1** and added worldwide geographic locations of the mutations provided from GISAID. These maps are useful to track down if a mutation is a localized and regional event or found worldwide. In the S protein, D614G is found worldwide with higher reported cases in the US, UK, and Australia, probably due to more large-scale testing, while A222V is mostly reported in the UK, but has not been reported in South America, Central and East African regions (**Fig. 2A)**. Like D614G, P323L in the RdRp is also found worldwide, with higher reported cases in the US, UK, and Australia (**Fig. 2B)**. The N gene doesn't have successful mutations that have attained reference strain status, but R203K, G204R and A220V all had a frequency of 50% or higher during the past twelve months. Even if R203K and G204R frequencies have decreasing since July of 2020 and stabilized in November of 2020, both

mutations are currently found worldwide. A220V emerged in August of 2020 and reached a frequency higher than 50% in October of 2020. Mutations G204R, R203K and A220V are reported at high frequency in the UK, but have not been detected in South America, Central, East, and South African regions (**Fig. 2C)**. The analyses of these data illustrate the localization of the most prevalent mutations to date, which appear to be mostly present in Western and developed countries. This, however, is undoubtable attributable to overall more testing.

Localization and molecular interactions of recurrent S protein mutations. Here we illustrated the molecular interactions and spatial localization of the mutations in the S protein. PyMOL was used to model the structures of S protein and analyze the possible effects of specific mutations at given positions in the protein. **Figure 3A** presents an overview of the localization of the S protein mutations and their interactions with their environment. The A222V mutation has no reported effects on protein stability, neutralizing antibody escape, and affinity for hACE2 **(Table 1)**. The substitution from A to V results in a low steric clash between neighboring residues (**Fig. 3B)**. The S477N substitution in the RBD enables increased stability during hACE2-RBD interactions (**Table 1, Fig. 3C)**. In the N-terminal domain (NTD), mutagenesis of L18F leads to a steric clash between neighboring residues (**Fig. 3D)**. However, this does not appear to impact the stability of the protein given that no such effects have been reported for L18F so far (**Table 1)**. In the closed conformation, D614 makes an ionic bond with K854 in the S2 subunit of another S protein monomer (**Fig. 3F)** (28). In the open conformation, D614 (or G614) doesn't make interactions or display steric clashes with neighboring residues (**Fig. 3E)**.

The emergence of the B.1.1.7 variant in the UK. A new variant was discovered in late 2020 in the UK that displayed increased affinity to hACE2 and Nabs escape capabilities (24,29,30). Here we attempted to further investigate the B.1.1.7 variant by looking at S protein mutations of this variant in complex with Nabs and hACE2. We mapped the localization of the mutations with available Cryo-EM structures of the S protein and assessed the frequency of the variant by interrogating the GISAID database. There are nine mutations in the S protein out of the total 24 mutations in the B.1.1.7 SARS-CoV-2 genome (**Fig. 5A & 5E)**. Deletions and mutations in the S protein of the B.1.1.7 variant, apart from D614G, emerged in October of 2020 and reached a frequency of 68% to 72% mid February (**Fig. 4, Fig. 5C, Table 2)**. N501Y is found in the RBD and can interact with a lysine residue in hACE2 (**Fig. 5B & 5D, Fig. 8C)**. The N501Y mutation is associated with an increased affinity to hACE2, along with an increase in infectivity and virulence (**Table 2)**. **Figure 5E** illustrates the whole genome of SARS-CoV-2 with all nucleotide substitutions of the B.1.1.7 variant. The C913T, C5986T, C14676T, C15279T, C16176T in ORF1ab, and T26801C in M protein are synonymous mutations. Also, the C27972T mutation has a frequency of 71% and produces a premature stop codon in NS8 that inactivates the protein (Q27stop)

without obvious consequences (**Fig. 5E, Table 2)** (31). These results allow us to better understand the frequencies, localization, and interactions of mutations in the S protein of the B.1.1.7 variant. Importantly, for viruses, not only synonymous mutations are of important. Non-synonymous mutations can exercise very important roles at various stages of the viral infection cycle, such as replication and creating functional RNA loops that serve as docking points for ribonucleoproteins and primers.

The emergence of the B.1.351 variant in South Africa. During the emergence of the B.1.1.7 variant in the UK, another variant was emerging in South Africa, known as B.1.351 (21,22). Similar to **Figure 5**, we illustrate the mutations in S protein and their respective frequencies. The GISAID database was used to track down mutations and Cryo-EM structures of the S protein to model the effects of point mutations. Most of the mutations in the S protein of the B.1.351 variant are localized in the S1 subunit, with only A701V in the S2 subunit. Additionally, three mutations reside in the RBD, among which two of them are not found in the B.1.1.7 variant (K417N, E484K) (**Fig. 6A & 6B**). This variant contains the D614G and N501Y, which are also seen in the B.1.1.7 variant (**Fig. 5A)**. Furthermore, many of the mutations found in the B.1.351 variant have global frequencies lower than 2%. The exception is D614G, L18F and N501Y (**Fig. 6C, Table 3)**. In comparison to B.1.1.7 and P.1 variants, B.1.351 variant has not reached a frequency higher than 1% as of mid February 2021 (**Fig. 4**). By using the mutagenesis tool of PyMOL, we modeled the S protein in complex with hACE2, and with C103 and C121 Nabs, which are human recombinant class I & II neutralizing antibodies, respectively(32). Our *in silico* mutagenesis predicts that mutations in the RBD induce a loss of interactions with C102 Nab and C121 Nab. At the position 417 in the S protein, a loss of interaction is predicted between the RBD and C102 Nab when the K417 is mutated to Asn producing Nabs escape capability. (**Fig. 6D & 8A).** Another loss of interaction is predicted with the E484K mutation and the C121 Nab, which could lead to Nabs escape capability (**Fig. 6E & 8B)**. As previously mentioned, N501Y is also found in the B.1.351 strain and our modeling predicts that it will have similar effects as those observed with the B.1.1.7 variant (**Fig. 6F**). **Figure 6G** illustrates the nucleotide substitutions of the B.1.351 variant.

The emergence of the P.1 variant in Brazil. Similar to the B.1351 variant, the P.1 variant harbours the N501Y and E484K mutations, but position 417 of the S protein displays a threonine instead of a lysine residue (19,23). Similar to the UK and South African variants, we demonstrate mutations in the S protein producing the P.1 variant and their frequencies from December 2019 to 17th of February 2021 (**Fig. 7**). The P.1 variant harbours substitutions L18F, T20N, P26S, D138Y, and R190S in the NTD of the S protein. H655Y and T1027I are in the subdomain 2 (SD2) and S2 subunit, respectively (**Fig. 7A & 7B, Table 4).** Overall, P.1-specific mutations have worldwide frequencies less than 2% (**Fig. 7C, Table 4)**. D614G, L18F and N501Y are not specific to P1. Similar to B.1.351, P.1 variant has low global frequency

with less than 1% of global variant frequency as of mid February 2021 (**Fig. 4**). We then modeled mutations to investigate interaction alterations with known recombinant neutralizing antibodies (32). The K417T mutation reduces interactions with neighboring residues in the C102 Nab and therefore we predict, as with K417N in B.1.1.7, an increased ability to escape neutralization (**Fig. 7D**). Also, the P.1 variant has E484K, and N501Y mutations in the RBD. We predict they will have the same effect reported for the B.1.1.7, and B.1.351 variants (**Fig. 7E & 7F)**. In **Figure 7G**, we illustrate the synonymous, non-synonymous, and deletions in SARS-CoV-2 P.1 genome (**Table 4**).

DISCUSSION

Research on the effect of mutations in SARS-CoV-2 has been carried out since the appearance of the virus, and the emergence of new genetic variants that are more transmissible and resistant to antibody neutralization have highlighted the importance of studying these mutations. The number of sequenced viral genomes uploaded to the GISAID database grew fast from 131,417 at the end of September to 451,913 by January 30th, 2021 (27). GISAID is a formidable tool for tracking the emergence of mutations, identifying the region where it emerged, and track its spread around the globe. Given the risk mutations and new variants pose for neutralizing antibody therapy and vaccines for SARS-CoV-2, it is crucial to continuously monitor the susceptibility of these variants to neutralization by humoral and cellular immune responses either induced through natural exposure to the reference strain or induced by vaccination (24). Recent reports on the vaccine efficacy of the Moderna, Pfizer-BioNTech, and Oxford-AstraZeneca vaccines against the B.1.1.7 and B.1.351 variant is variable. All of these vaccines remain efficacious against the B.1.1.7 variant (30,33,34). However, the Oxford-AstraZeneca vaccine has displayed compromised efficacy against the B.1.351 variant with 21.4% (35). Preliminary data with the Pfizer-BioNTech and Moderna mRNA vaccines also show reduction of efficacy against B.1.351 (34,36). Furthermore, antibodies induced by the Pfizer-BioNTech vaccine appear to display a 15.1-fold decrease in neutralization efficacy against the P.1 variant (37) (**Table 5**). Nevertheless, humoral responses are only one component of the adaptive immune response. T and B cell responses have not been probed in detail against these variants at this time and may still provide robust protection.

Furthermore, there is likely epistatic mutations in the S protein. Epistasis is the combinatory effect of two or more mutations in a genome (41). Epistasis has previously been studied in the surface protein hemagglutinin (HA) of the influenza viruses, and have illustrated positive epistasis in 11 regions of the HA receptor-binding domain (42). In relation to the S protein of SARS-CoV-2, it could, for instance, allow the S protein to adopt a specific conformation when all the mutations are present, thereby producing

a unique folding of the protein. A recent study demonstrated the impact of antigenicity and infectivity of the D614G SARS-CoV-2 variants with a combination of different mutations occurring in the S protein (19). The study shows that D614G alone increases infectivity, but in combination with different other mutations in the S protein, these can either increase or decrease viral infectivity. Similar findings have been reported regarding sensitivity to Nabs. D614G alone has undetectable effects on Nabs escape. However, the combination of D614G with other mutations in S can enable Nabs escape. This data suggests that the continuous emergence of epistatic mutations in SARS-CoV-2 will likely be involved in further altering properties of the virus, including transmissibility, pathogenicity, stability and Nab resistance.

Our analyses have highlighted that several of the successful mutations analyzed had frequency trajectories that eventually plunged or stabilized at low frequencies. Only the D614G in the S protein and the P323L in the RdRp maintained their presence in the consensus sequence (**Fig. 1 & 2**). Analyses of the GISAID database also reveal which countries upload the most sequences to the database and are therefore carrying out the most testing and sequencing. The global frequency of mutations and variants in the database is therefore biased to represent the genetic landscape of the countries doing the most testing. Emergence of new variants may therefore go undetected until they leave their point of origin and enter countries with high testing and sequencing rates. This delayed notification constitutes a major obstacle in preventing the spread of nefarious variants that are potentially resistant to current vaccines and neutralizing antibody therapy.

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

ACKNOWLEDGEMENTS

The authors wish to thank Sean Li for helpful comments on our manuscript. M.-A.L. holds a Canada Research Chair in Molecular Virology and Intrinsic Immunity. This study was supported by a COVID-19 Rapid Response grant to M-A Langlois by the Canadian Institute of Health Research (CIHR) and by a grant supplement by the Canadian Immunity Task Force (CITF).

CONFLICTS OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W. 2020. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* **382**:727–733.

2. Dong E, Du H, Gardner L. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* **20**:533–534.

3. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y, Yuan M-L, Zhang Y-L, Dai F-H, Liu Y, Wang Q-M, Zheng J-J, Xu L, Holmes EC, Zhang Y-Z. 2020. A new coronavirus associated with human respiratory disease in China. *Nature* **579**:265–269.

4. Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, Chang H. 2020. The Architecture of SARS-CoV-2 Transcriptome. *Cell 181*:914-921.e10.

5. Duffy S. 2018. Why are RNA virus mutation rates so damn high? *PLOS Biology* **16**:e3000003.

6. Mandary MB, Masomian M, Poh CL. 2019. Impact of RNA Virus Evolution on Quasispecies Formation and Virulence. *Int J Mol Sci* 20:4657.

7. Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics* 9:267–276.

8. Eckerle LD, Becker MM, Halpin RA, Li K, Venter E, Lu X, Scherbakova S, Graham RL, Baric RS, Stockwell TB, Spiro DJ, Denison MR. 2010. Infidelity of SARS-CoV Nsp14-Exonuclease Mutant Virus Replication Is Revealed by Complete Genome Sequencing. *PLOS Pathogens* 6:e1000896.

9. Minskaia E, Hertzig T, Gorbalenya AE, Campanacci V, Cambillau C, Canard B, Ziebuhr J. 2006. Discovery of an RNA virus 3′→5′ exoribonuclease that is critically involved in coronavirus RNA synthesis. *Proc Natl Acad Sci U S A* 103:5108.

10. Ma Y, Wu L, Shaw N, Gao Y, Wang J, Sun Y, Lou Z, Yan L, Zhang R, Rao Z. 2015.

Structural basis and functional analysis of the SARS coronavirus nsp14–nsp10 complex. *Proc Natl Acad Sci USA* 112:9436.

11. Rausch JW, Capoferri AA, Katusiime MG, Patro SC, Kearney MF. 2020. Low genetic diversity may be an Achilles heel of SARS-CoV-2. *Proc Natl Acad Sci USA* 117:24614.

12. Andreano E, Piccini G, Licastro D, Casalino L, Johnson NV, Paciello I, Monego SD, Pantano E, Manganaro N, Manenti A, Manna R, Casa E, Hyseni I, Benincasa L, Montomoli E, Amaro RE, McLellan JS, Rappuoli R. 2020. SARS-CoV-2 escape *in vitro* from a highly neutralizing COVID-19 convalescent plasma. *bioRxiv* 2020.12.28.424451.

13. Shi J, Wen Z, Zhong G, Yang H, Wang C, Huang B, Liu R, He X, Shuai L, Sun Z, Zhao Y, Liu P, Liang L, Cui P, Wang J, Zhang X, Guan Y, Tan W, Wu G, Chen H, Bu Z. 2020. Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS–coronavirus 2. *Science* 368:1016.

14. Li X, Giorgi EE, Marichannegowda MH, Foley B, Xiao C, Kong X-P, Chen Y, Gnanakaran S, Korber B, Gao F. 2020. Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci Adv* 6:eabb9153.

15. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. 2020. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* 181:281-292.e6.

16. 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. 2020. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* https://doi.org/10.1038/s41586-020-2895-3.

17. Zhang L, Jackson CB, Mou H, Ojha A, Peng H, Quinlan BD, Rangarajan ES, Pan A, Vanderheiden A, Suthar MS, Li W, Izard T, Rader C, Farzan M, Choe H. 2020. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nature Communications* 11:6013.

18. 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, Angyal A, Brown RL, Carrilero L, Green LR, Groves DC, Johnson KJ, Keeley AJ, Lindsey BB, Parsons PJ, Raza M, Rowland-Jones S, Smith N, Tucker RM, Wang D, Wyles MD, McDanal C, Perez LG, Tang H, Moon-Walker A, Whelan SP, LaBranche CC, Saphire EO, Montefiori DC. 2020. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* 182:812-827.e19.

19. 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*, 2020/07/17 ed. 182:1284-1294.e9.

20. Rambaut A, Loman N, Pybus O, Barclay W, Barrett J, Carabelli A, Connor T, Peacock T, Robertson D, Volz E, on behalf of CoG-UK. 2020. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. 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

21. O'Toole, A, Hill V, et al. (2021) Tracking the international spread of SARS-CoV-2 lineage B.1.1.7 and B.1.351/501Y-V2. https://virological.org/t/tracking-the-international-spread-of-sars-cov-2-lineages-b-1-1-7-and-b-1-351-501y-v2/592

22. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, Doolabh D, Pillay S, San EJ, Msomi N, Mlisana K, von Gottberg A, Walaza S, Allam M, Ismail A, Mohale T, Glass AJ, Engelbrecht S, Van Zyl G, Preiser W, Petruccione F, Sigal A, Hardie D, Marais G, Hsiao M, Korsman S, Davies M-A, Tyers L, Mudau I, York D, Maslo C, Goedhals D, Abrahams S, Laguda-Akingba O, Alisoltani-Dehkordi A, Godzik A, Wibmer CK, Sewell BT, Lourenço J, Alcantara LCJ, Pond SLK, Weaver S, Martin D, Lessells RJ, Bhiman JN, Williamson C, de Oliveira T. 2020. Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv* 2020.12.21.20248640.

23. Faria, Nuno R.; et al. (12 January 2021). Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. *Virological.* Retrieved 23 January 2021. https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586

24. Wang P, Liu L, Iketani S, Luo Y, Guo Y, Wang M, Yu J, Zhang B, Kwong PD, Graham BS, Mascola JR, Chang JY, Yin MT, Sobieszczyk M, Kyratsous CA, Shapiro L, Sheng Z, Nair MS, Huang Y, Ho DD. 2021. Increased Resistance of SARS-CoV-2 Variants B.1.351 and B.1.1.7 to Antibody Neutralization. *bioRxiv* 2021.01.25.428137.

25. Yi C, Sun X, Ye J, Ding L, Liu M, Yang Z, Lu X, Zhang Y, Ma L, Gu W, Qu A, Xu J, Shi Z, Ling Z, Sun B. 2020. Key residues of the receptor binding motif in the spike protein of SARS-CoV-2 that interact with ACE2 and neutralizing antibodies. *Cellular & Molecular Immunology* 17:621–630.

26. Gu H, Chen Q, Yang G, He L, Fan H, Deng Y-Q, Wang Y, Teng Y, Zhao Z, Cui Y, Li Y, Li X-F, Li J, Zhang N-N, Yang X, Chen S, Guo Y, Zhao G, Wang X, Luo D-Y, Wang H, Yang X, Li Y, Han G, He Y, Zhou X, Geng S, Sheng X, Jiang S, Sun S, Qin C-F, Zhou Y. 2020. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science* 369:1603.

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

28. Benton DJ, Wrobel AG, Xu P, Roustan C, Martin SR, Rosenthal PB, Skehel JJ, Gamblin SJ. 2020. Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. *Nature* 588:327–330.

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.e20.

30. Muik A, Wallisch A-K, Sänger B, Swanson KA, Mühl J, Chen W, Cai H, Maurus D, Sarkar

R, Türeci Ö, Dormitzer PR, Şahin U. 2021. Neutralization of SARS-CoV-2 lineage B.1.1.7 pseudovirus by BNT162b2 vaccine–elicited human sera. *Science* eabg6105.

31. Pereira F. 2020. Evolutionary dynamics of the SARS-CoV-2 ORF8 accessory gene. *Infect Genet Evol*, 2020/09/02 ed. 85:104525–104525.

32. Barnes CO, Jette CA, Abernathy ME, Dam K-MA, Esswein SR, Gristick HB, Malyutin AG, Sharaf NG, Huey-Tubman KE, Lee YE, Robbiani DF, Nussenzweig MC, West AP, Bjorkman PJ. 2020. SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature* 588:682–687.

33. Emary KR, Golubchik T, Aley PK, Ariani CV, Angus BJ, Bibi S, Blane B, Bonsall D, Cicconi P, Charlton S. 2021. Efficacy of ChAdOx1 nCoV-19 (AZD1222) vaccine against SARS-CoV-2 VOC 202012/01 (B. 1.1. 7).

34. Wu K, Werner AP, Koch M, Choi A, Narayanan E, Stewart-Jones GBE, Colpitts T, Bennett H, Boyoglu-Barnum S, Shi W, Moliva JI, Sullivan NJ, Graham BS, Carfi A, Corbett KS, Seder RA, Edwards DK. 2021. Serum Neutralizing Activity Elicited by mRNA-1273 Vaccine — *Preliminary Report. N Engl J Med* https://doi.org/10.1056/NEJMc2102179

35. Madhi SA, Baillie V, Cutland CL, Voysey M, Koen AL, Fairlie L, Padayachee SD, Dheda K, Barnabas SL, Bhorat QE, Briner C, Kwatra G, Ahmed K, Aley P, Bhikha S, Bhiman JN, Bhorat AE, Plessis J du, Esmail A, Groenewald M, Horne E, Hwa S-H, Jose A, Lambe T, Laubscher M, Malahleha M, Masenya M, Masilela M, McKenzie S, Molapo K, Moultrie A, Oelofse S, Patel F, Pillay S, Rhead S, Rodel H, Rossouw L, Taoushanis C, Tegally H, Thombrayil A, Eck S van, Wibmer CK, Durham NM, Kelly EJ, Villafana TL, Gilbert S, Pollard AJ, de Oliveira T, Moore PL, Sigal A, Izu A. 2021. Safety and efficacy of the ChAdOx1 nCoV-19 (AZD1222) Covid-19 vaccine against the B.1.351 variant in South Africa. *medRxiv* 2021.02.10.21251247.

36. Liu Y, Liu J, Xia H, Zhang X, Fontes-Garfias CR, Swanson KA, Cai H, Sarkar R, Chen W, Cutler M, Cooper D, Weaver SC, Muik A, Sahin U, Jansen KU, Xie X, Dormitzer PR, Shi P-Y. 2021. Neutralizing Activity of BNT162b2-Elicited Serum — *Preliminary Report. N Engl J Med* https://doi.org/10.1056/NEJMc2102017.

37. Garcia-Beltran WF, Lam EC, Denis KSt, Nitido AD, Garcia ZH, Hauser BM, Feldman J, Pavlovic MN, Gregory DJ, Poznansky MC, Sigal A, Schmidt AG, Iafrate AJ, Naranbhai V, Balazs AB. 2021. Circulating SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *medRxiv* 2021.02.14.21251704.

38. Polack FP, Thomas SJ, Kitchin N, Absalon J, Gurtman A, Lockhart S, Perez JL, Pérez Marc G, Moreira ED, Zerbini C, Bailey R, Swanson KA, Roychoudhury S, Koury K, Li P, Kalina WV, Cooper D, Frenck RW, Hammitt LL, Türeci Ö, Nell H, Schaefer A, Ünal S, Tresnan DB, Mather S, Dormitzer PR, Şahin U, Jansen KU, Gruber WC. 2020. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *N Engl J Med* 383:2603–2615.

39. Baden LR, El Sahly HM, Essink B, Kotloff K, Frey S, Novak R, Diemert D, Spector SA, Rouphael N, Creech CB, McGettigan J, Khetan S, Segall N, Solis J, Brosz A, Fierro C, Schwartz H, Neuzil K, Corey L, Gilbert P, Janes H, Follmann D, Marovich M, Mascola J, Polakowski L, Ledgerwood J, Graham BS, Bennett H, Pajon R, Knightly C, Leav B, Deng W, Zhou H, Han S, Ivarsson M, Miller J, Zaks T. 2020. Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N Engl J Med* 384:403–416.

40. Voysey M, Clemens SAC, Madhi SA, et al. 2021. Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *The Lancet* 397:99–111.

41. Phillips PC. 2008. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics* 9:855–867.

42. Wu NC, Xie J, Zheng T, Nycholat CM, Grande G, Paulson JC, Lerner RA, Wilson IA. 2017. Diversity of Functionally Permissive Sequences in the Receptor-Binding Site of Influenza Hemagglutinin. *Cell Host & Microbe* 21:742-753.e8.

43. COVID-19 GENOMIC UK CONSORTIUM. 2021. COG-UK report on SARS-CoV-2 Spike mutations of interest in the UK. https://www.cogconsortium.uk/wp-content/uploads/2021/01/Report-2_COG-UK_SARS-CoV-2-Mutations.pdf

44. Zahradník J, Marciano S, Shemesh M, Zoler E, Chiaravalli J, Meyer B, Rudich Y, Dym O,

Elad N, Schreiber G. 2021. SARS-CoV-2 RBD *in vitro* evolution follows contagious mutation spread, yet generates an able infection inhibitor. *bioRxiv* 2021.01.06.425392.

45. Kannan SR, Spratt AN, Quinn TP, Heng X, Lorson CL, Sönnerborg A, Byrareddy SN, Singh K. 2020. Infectivity of SARS-CoV-2: there Is Something More than D614G? *J Neuroimmune Pharmacol*, 2020/09/15 ed. 15:574–577.

46. Ugurel OM, Mutlu O, Sariyer E, Kocer S, Ugurel E, Inci TG, Ata O, Turgut-Balik D. 2020. Evaluation of the potency of FDA-approved drugs on wild type and mutant SARS-CoV-2 helicase (Nsp13). *Int J Biol Macromol*, 2020/09/24 ed. 163:1687–1696.

47. Wu S, Tian C, Liu P, Guo D, Zheng W, Huang X, Zhang Y, Liu L. 2020. Effects of SARS-CoV-2 mutations on protein structures and intraviral protein–protein interactions. *Journal of Medical Virology* n/a.

TABLES

Table 1: Non-synonymous mutations in SARS-CoV-2 genes with a worldwide frequency >0.01%.

Gene	Genome Nucleotide Mutation	Codon	AA Mutation	Frequency (%)*	Effect
S	A23403G	G**A**T to G**G**T	D614G	99.70	Moderate increase in Transmissibility[43]
	C22227T	G**C**T to G**T**T	A222V	42.79	No mutational effect[43]
	C21615T	**C**TT to **T**TT	L18F	19.86	N/A
	G22992A	A**G**C to AA**C**	S477N	5.36	Spike-hACE2 complex stability and Nab interference[42]
	C22879A	AA**C** to AA**A**	N439K	4.09	Antibody Escape ability[19]
	C21575T	**C**TT to **T**TT	L5F	1.44	N/A
	C21855T	T**C**T to T**T**T	S98F	2.74	N/A
	G22346T	**G**CT to **T**CT	A262S	1.64	N/A
NSP1	C335T	**C**GC to **T**GC	R24C	0.70	N/A
NSP2	A1163T	**A**TT to **T**TT	I120F	0.74	N/A
	G1210T	AT**G** to AT**T**	M135I	0.124	N/A
	C1059T	A**C**C to A**T**C	T85I	5.55	N/A
	T1947C	G**T**T to G**C**T	V381A	0.55	N/A
	C2453T	**C**TC to **T**TC	L550F	2.73	N/A
NSP3	C7926T	G**C**A to G**T**A	A1736V	1.23	N/A
	G2891A	**G**CA to **A**CA	A58T	0.03	N/A
	T7767C	A**T**C to A**C**C	I1683T	4.05	N/A
	C5622T	C**C**T to C**T**T	P968L	0.63	N/A
	G8371T	CA**G** to CA**T**	Q1884H	0.05	N/A
	C4002T	A**C**T to A**T**T	T428I	0.38	N/A
	C5388A	G**C**T to G**A**T	A890D	21.37	N/A
	G8083A	AT**G** to AT**A**	M1788I	1.23	N/A
	C3602T	**C**AC to **T**AC	H295Y	2.51	N/A
NSP4	C9246T	G**C**T to G**T**T	A231V	0.12	N/A
	G9526T	AT**G** to AT**T**	M324I	5.18	N/A
	A10323G	A**A**G to A**G**G	K90R	1.60	N/A
NSP5	G10097A	**G**GT to **A**GT	G15S	0.36	N/A
	G10265A	**G**GT to **A**GT	G71S	0.69	N/A
	C10319T	**C**TT to **T**TT	L89F	3.21	N/A
NSP6	C11109T	G**C**T to G**T**T	A46V	0.08	N/A
	C11396T	**C**TT to **T**TT	L142F	2.64	N/A
	G11083T	TTG to TT**T**	L37F	2.73	N/A
	C11195T	**C**TT to **T**TT	L75F	0.05	N/A
	G11230T	AT**G** to AT**T**	M86I	0.52	N/A
	G11801A	**G**GT to **A**GT	G277S	0.01	N/A
	G11132T	**G**CT to **T**CT	A54S	1.83	N/A
	11288-11296 del	N/A	ΔS106	22.50	N/A
NSP7	G12067T	AT**G** to AT**T**	M75I	0.171	N/A
	C11916T	T**C**A to T**T**A	S25L	0.06	N/A
NSP9	G12988T	AT**G** to AT**T**	M101I	3.20	N/A
NSP12	C14408T	C**C**T to C**T**T	P323L	99.53	Improve processivity, by interaction with NSP8[45]
	G15766T	**G**TG to **T**TG	V776L	5.27	N/A
	G13993T	**G**CT to **T**CT	A185S	5.22	N/A
	G15598A	**G**TC to **A**TC	V720I	3.21	N/A
	G14202T	GA**G** to GA**T**	E254D	0.68	N/A
	C13730T	G**C**T to G**T**T	A97V	0.04	N/A
NSP13	C16289T	G**C**T to G**T**T	A18V	0.05	N/A
	G17019T	GA**G** to GA**T**	E261D	5.20	N/A
	C17104T	**C**AT to **T**AT	H290Y	4.20	N/A
	C17747T	C**C**T to C**T**T	P504L	0.10	Increase hydrophobicity of 2A domain[46]

	C17639T	T**C**A to T**T**A	S468L	0.02	N/A
	A17615G	A**A**G to A**G**G	K460R	10.20	N/A
	A16889G	A**A**A to A**G**A	K218R	5.18	N/A
	G18028T	**G**CA to **T**CA	A598S	3.21	N/A
NSP14	C18998T	G**C**A to G**T**A	A320V	0.02	N/A
	C18568T	**C**TC to **T**TC	L177F	0.16	N/A
	G19542T	AT**G** to AT**T**	M501I	0.71	N/A
	A18424G	**A**AT to **G**AT	N129D	3.00	N/A
NSP15	C19718T	A**C**A to A**T**A	T33I	1.78	N/A
NSP16	A21137G	A**A**G to A**G**G	K160R	0.39	N/A
	A21390G	TT**A** to TT**G**	R216C	2.77	N/A
N	C28932T	G**C**T to G**T**T	A220V	46.41	N/A
	G28580T	**G**AT to **T**AT	D103Y	0.13	N/A
	G28883C	**G**GA to **C**GA	G204R	29.55	Destabilizing N protein[47]
	C28311T	C**C**C to C**T**C	P13L	0.27	N/A
	C28869T	C**C**A to C**T**A	P199L	5.55	N/A
	GC28882AA	A**GC** to A**AA**	R203K	29.74	Destabilizing N protein[47]
	C28854T	T**C**A to T**T**A	S194L	4.65	Destabilizing N protein[47]
	C28977T	T**C**T to T**T**T	S235F	22.01	N/A
	GAT28282CTA	GAT to CTA	D3L	22.01	N/A
	G28975C	AT**G** to AT**C**	M234I	5.58	N/A
	C28472T	**C**CT to **T**CT	P67S	3.09	N/A
	G29399A	**G**CT to **A**CT	A376T	5.14	N/A
	C29366T	**C**CA to **T**CA	P365S	5.64	N/A
	G29402T	**G**AT to **T**AT	D377Y	1.29	N/A
	C28887T	A**C**T to A**T**T	T205I	1.47	N/A
	C29466T	G**C**A to G**T**A	A398V	0.78	N/A
M	A26530G	G**A**T to G**G**T	D3G	0.42	N/A
	C27046T	A**C**G to A**T**G	T175M	0.02	N/A
NS3	G25907T	G**G**T to G**T**T	G172V	2.98	N/A
	G25617T	AA**G** to AA**T**	K75N	0.39	N/A
	G25563T	CA**G** to CA**T**	Q57H	13.00	N/A
	C26060T	A**C**T to A**T**T	T223I	1.87	N/A
	G25429T	**G**TA to **T**TA	V13L	0.10	N/A
	G25996T	**G**TA to **T**TA	V202L	2.64	N/A
	A25505G	C**A**A to C**G**A	Q38R	2.46	N/A
	G25906C	**G**GT to **C**GT	G172R	2.32	N/A
NS7b	AT27866TA	C**AT** to C**TA**	H37L	0.058	N/A
	C27769T	T**C**A to T**T**A	S5L	1.68	N/A
NS8	C28087T	G**C**T to G**T**T	A65V	2.68	N/A
	T28144C	T**T**A to T**C**A	L84S	0.12	N/A
	C27964T	T**C**A to T**T**A	S24L	3.58	N/A
	G28077T	**G**TG to **T**TG	V62L	0.27	N/A
	C27972T	**C**AA to **T**AA	Q27stop	21.17	Inactivation of NS8[20]
	G28048T	A**G**A to A**T**A	R52I	21.14	N/A

*Frequency of the mutation as of December 31, 2020.

Table 2: Non-synonymous, synonymous and deletion mutations in the B.1.1.7 variant

Variant	Gene	Genome Nucleotide Mutation[17]	S or NS	AA mutation[17]	Domain	Frequency* (%)	Effect
B.1.1.7	*ORF1ab*	C3267T	NS	T1001I		70.7	N/A
		C5388A	NS	A1708D		71.5	N/A
		T6954C	NS	I2230T		70	N/A
		11288-11296 del	NS	ΔS3675/ ΔG3676/ΔF3677		69.2	N/A
		C913T	S			N/A	N/A
		C5986T	S			N/A	N/A
		C14676T	S			N/A	N/A
		C15279T	S			N/A	N/A
		C16176T	S			N/A	N/A
	Spike	21765- 21770	NS	ΔH69/ΔV70	NTD	70.8	Antibody escape[43]
		21991-21993	NS	ΔY144	NTD	0.01	Decrease infectivity, Antibody escape[19]
		A23063T	NS	N501Y	RBD	71.9	Increase infectivity, virulence and affinity for hACE2[43]
		C23271A	NS	A570D	SD1	71.6	
		A23403G	NS	D614G	SD2	99.4	Moderate increase transmissibility[43]
		C23709A	NS	P681H	SD2	72.7	N/A
		C23709T	NS	T716I		72.1	N/A
		T24506G	NS	S982A	HR1	71.3	N/A
		G24914C	NS	D1118H		71.2	N/A
	ORF8	C27972T	NS	Q27stop		71.2	Inactivation of NS8[20]
		G28048T	NS	R52I		71	N/A
		A28111G	NS	Y73C		71.3	N/A
	M	T26801C	S			N/A	N/A
	N	GAT28280CTA	NS	D3L		70.5	N/A
	N	C28977T	NS	S235F		71.4	N/A

*Frequency of the mutation as of February 17th 2021.

Table 3: Non-synonymous mutations and deletions in the B.1.351 variant

Variant	Gene	Genome Nucleotide Mutation[19]	A.A mutation[19]	Domain	Frequency * (%)	Effect
B.1.351	*ORF1ab*	C1059T	T265I		7.3	N/A
		G5230T	K1655N		0.57	N/A
		A10323G	K3353R		1.7	N/A
	Spike	C21614T	L18F	NTD	4.5	N/A
		A21801C	D80A	NTD	0.47	N/A
		A22206G	D215G	NTD	0.51	N/A
		22286-22294	ΔL242/ΔA243/ΔL244	NTD	0.311	N/A
		G22299T	R246I	NTD	0	N/A
		G22813T	K417N	RBD	0.50	Antibody escape[24]
		G23012A	E484K	RBD	1.74	Antibody escape[43]
		A23063T	N501Y	RBD	71.9	Increase affinity for hACE2[43]
		A23403G	D614G	SD2	99.4	Moderate increase in transmissbility[43]
		G23664T	A701V	S1/S2 – S2'	1.40	N/A
	ORF3a	G25563T	Q57H		10.2	N/A
		C25904T	S171L		1.29	N/A
	E	C26456T	P71L		0.55	N/A
	N	C28887T	T205I		2.5	N/A

*Frequency of the mutation as of February 17th 2021.

Table 4: Non-synonymous, synonymous and deletions in the P.1 variant.

Variant	Gene	Genome Nucleotide Mutation	S or NS	A.A mutation	Domain	Frequency* (%)	Effect
P.1	*ORF1ab*	T733C	S			N/A	N/A
		C2749T	S			N/A	N/A
		C3828T	NS	S1188L		0.22	N/A
		A5648C	NS	K1795Q		0.19	N/A
		11288-11296 del	NS	ΔS3675/ ΔG3676/ΔF3677		71.1	N/A
		C12778T	S			N/A	N/A
		C13860T	S			N/A	N/A
		G17259T	NS	E5665D		0.24	N/A
	Spike	C21614T	NS	L18F	NTD	4.5	N/A
		C21621A	NS	T20N	NTD	0.30	N/A
		C21638T	NS	P26S	NTD	0.44	N/A
		G21974T	NS	D138Y	NTD	0.28	N/A
		G22132T	NS	R190S	NTD	0.73	N/A
		A22812C	NS	K417T	RBD	0.11	N/A
		G23012A	NS	E484K	RBD	1.7	Antibody escape[43]
		A23063T	NS	N501Y	RBD	71.9	Increase affinity for hACE2[43]
		A23403G	NS	D614G	SD2	99.4	Moderate increase in transmissbility[43]
		C23525T	NS	H655Y	SD2	0.25	N/A
		C24642T	NS	T1027I	S2	0.20	N/A
	ORF8	G28167A	NS	E92K		0.30	N/A
		Ins28269-28273	S			N/A	N/A
	N	C28512G	NS	P80R		0.18	N/A

*Frequency of the mutation as of February 17th 2021.

Table 5: Efficacy of vaccines against SARS-CoV-2 variants.

Vaccine	**hCoV-19/Wuhan/WIV-4/2019**	**B.1.1.7**	**B.1.351**	**P.1**
Pfizer-BioNTech	95.0%[38]	~ 95.0%[30]	2/3 reduction in neutralization[36]	15.1-fold decrease[37]
Moderna	94.1%[39]	~ 94.1%[34]	6.4 reduction in[34] titers	N/A
Oxford-AstraZeneca	70.4%[40]	74.6%[33]	21.9%[35]	N/A

FIGURE LEGENDS

Figure 1: Variation of mutation frequency in SARS-CoV-2 genes. The occurrence and frequency of mutations in various SARS-CoV-2 genes are presented for the time period between December 2019 to December 2020. SARS-CoV-2 genes are represented with non-structural proteins (NSPs) and the function on the genes in parentheses. Graphs were generated using RStudio.

Figure 2: Geographic location and timeline of dominant mutations in NSP12, S, and N genes. A) Frequency of S protein mutations with corresponding geographic maps. B) Frequency of RdRp mutations with corresponding geographic maps. C) Frequency of Nucleoprotein mutations with corresponding geographic maps. D) Mutations reaching a frequency higher than 50% between December 2019 and December 2020. A low frequency of reported cases of the mutations is represented in white, while higher frequencies are represented in red. All maps were taken from GISAID. Graphs were generated using RStudio and Biorender.

Figure 3: Structural rendering of most frequent mutation sin the S protein. A) Surface representation of hACE2 (yellow) in complex with S protein trimers illustrated in grey, blue, and magenta. Recurrent mutations are represented in green. Cartoon representation of B) A222V, C) S477N, D) L18F, E) D614G in open conformation and F) D614G in closed conformation. Reference sequence residues are illustrated in green, and the mutant amino acid is represented in purple. The red circles illustrate the steric clash when the mutations are inserted into the structure. Graphs were generated using PyMOL.

Figure 4: Frequency of B.1.1.7, B.1351, P.1, D614G, and reference variants. Database variants frequency were analyzed from December 2019, to February 17th, 2021. The hCoV-19/Wuhan is the reference strain, and hCoV-19/D614G is representing the D614G mutation in the S protein. B.1.1.7, B.1.3531, and P.1 represent the UK, South African, and Brazilians variants. *(overlapping curves).

Figure 5: Frequency of B.1.1.7 spike protein mutations with structural confirmation and genome map. A & B) Colour representation of S protein subdomains with mutations of the B.1.1.7 variant in red. NTD (green), RBD (blue), SD1 (purple), SD2 (light blue), and S2 (magenta) are illustrated. The other S protein monomers are displayed in grey and white. C) Frequency of the mutations in the S protein,

B.1.1.7 variant from December 2019 to February 17th 2021. D) Interaction of the N501Y (red) mutation in the RBD (blue) of S protein with hACE2 (yellow). The dash lines indicate interactions with adjacent residues. E) Genome of the SARS-CoV-2 B.1.1.7 variant with identified nucleotide substitutions and deletions. Graphs were generated using Biorender, PyMOL, and RStudio. *(overlapping curves).

Figure 6: Frequency of B.1.351 spike protein mutations with structural confirmation and genome map. A & B) Colour representation of S protein subdomains with mutations of the B.1.351 variant in orange. NTD (green), RBD (blue), SD1 (purple), SD2 (light blue), and S2 (magenta) are illustrated. The other S protein monomers are illustrated in grey and white. C) Frequency of the mutations in the S protein B.1.351 variant from December 2019 to February 17th 2021. Interaction of D) 417N with C102 Nab (green), E) 484K with C121 Nab (light pink), and F) 501Y with hACE2 (yellow). The mutant residues are illustrated in orange, and the dashed lines represent interactions with adjacent residues. G) Genome of the SARS-CoV-2 B.1.351 variant with identified nucleotide substitutions or deletions. Graphs were generated using Biorender, PyMOL, and RStudio. *(overlapping curves).

Figure 7: Frequency of P.1 spike protein mutations with structural confirmation and genome map. A& B) Colour representation of S protein subdomains with mutations of the P.1 variant in black. NTD (green), RBD (blue), SD1 (purple), SD2 (light blue), and S2 (magenta) are illustrated. The other S protein monomers are illustrated in grey and white. C) Frequency of P.1 variant S protein mutations from December 2019 to February 17th 2021. Interaction of D) 417T with C102 Nab (green), E) 484K with C121 Nab (light pink), and F) 501Y with hACE2 (yellow). The mutations are coloured in black and interaction with adjacent residues are demonstrated by dashed lines. G) Genome of the SARS-CoV-2 P.1 variant with identified nucleotides substitution, deletions, and insertions. Figures were generated using Biorender, PyMOL and RStudio. *(overlapping curves).

Figure 8: Interactions of K417, E484, and N501 of the S protein with neutralizing antibodies and hACE2. A) Interaction of K417 (blue) with C102 Nab (green) residues. B) Interaction of E484 (blue) with C121 Nab (light pink). C) Interaction of N501 (blue) with hACE2 (yellow) Dashes lines indicate interactions between residues. The graphs were generated using PyMOL.

Figure 1

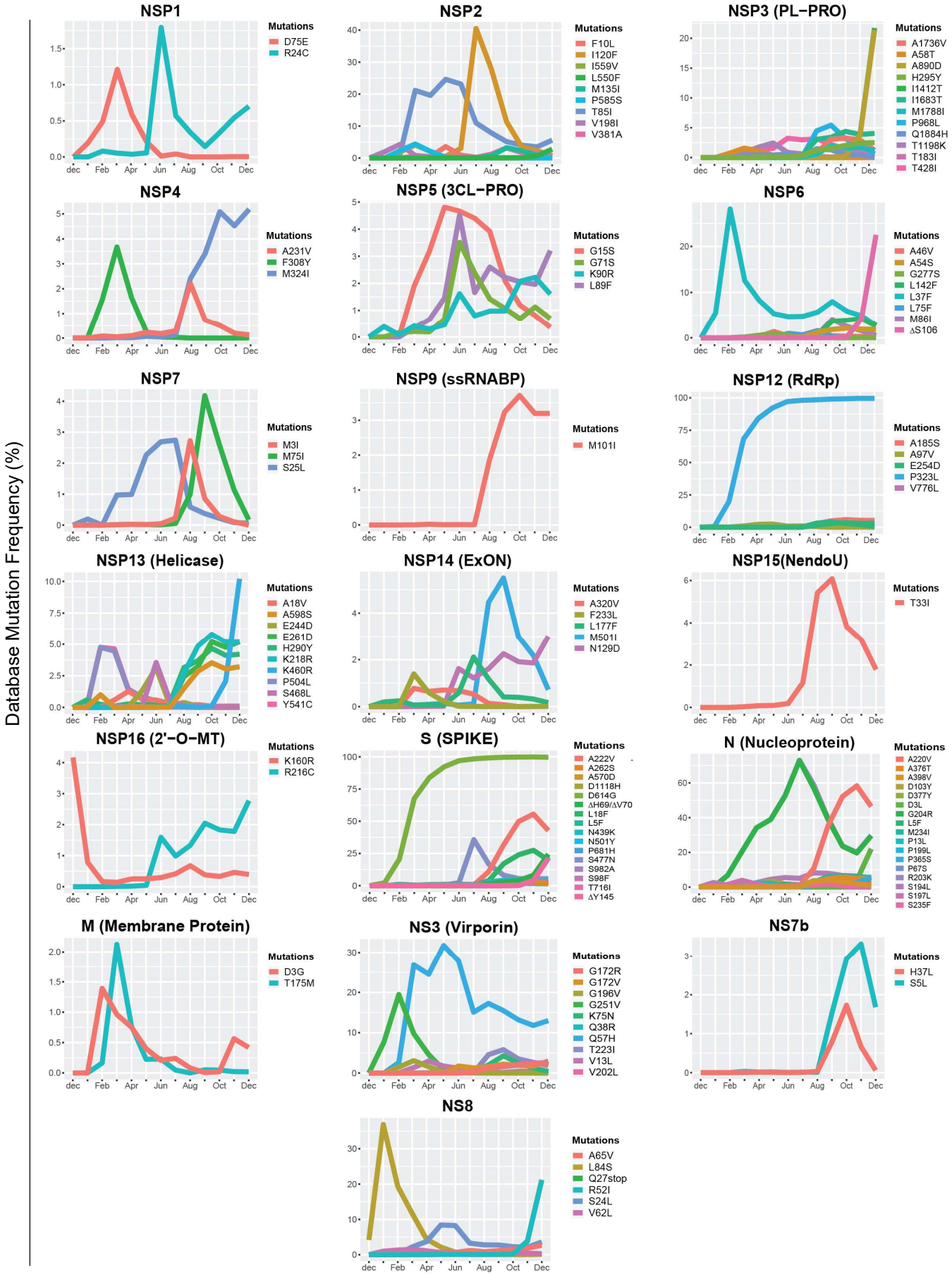
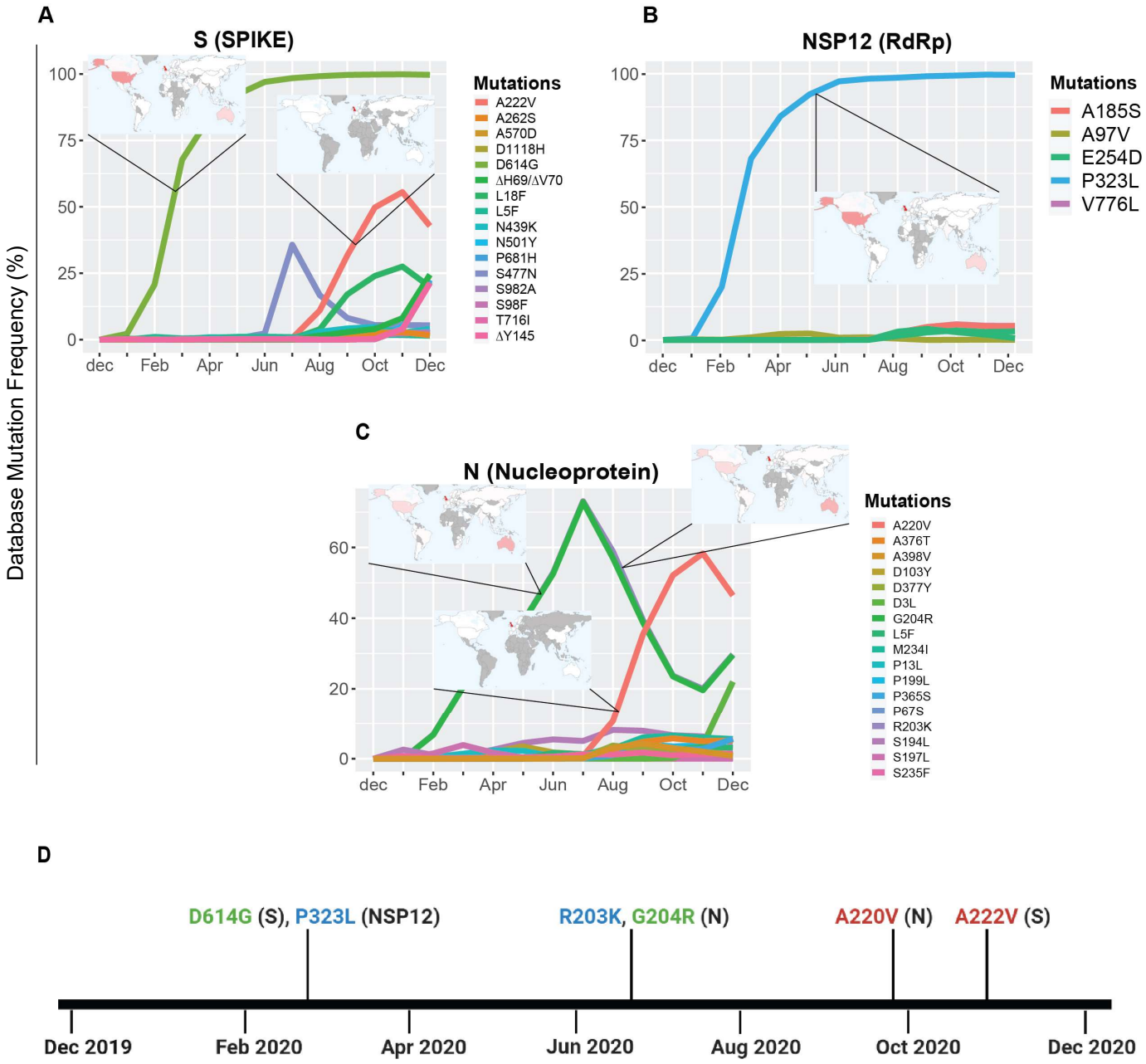

Database Mutation Frequency (%)
NSP1
Mutations
D75E
R24C
NSP2
Mutations
F10L
I120F
I559V
L550F
M135I
P585S
T85I
V198I
V381A
NSP3 (PL−PRO)
Mutations
A1736V
A58T
A890D
H295Y
I1412T
I1683T
M1788I
P968L
Q1884H
T1198K
T183I
T428I
NSP4
Mutations
A231V
F308Y
M324I
NSP5 (3CL−PRO)
Mutations
G15S
G71S
K90R
L89F
NSP6
Mutations
A46V
A54S
G277S
L142F
L37F
L75F
M86I
ΔS106
NSP7
Mutations
M3I
M75I
S25L
NSP9 (ssRNABP)
Mutations
M101I
NSP12 (RdRp)
Mutations
A185S
A97V
E254D
P323L
V776L
NSP13 (Helicase)
Mutations
A18V
A598S
E244D
E261D
H290Y
K218R
K460R
P504L
S468L
Y541C
NSP14 (ExON)
Mutations
A320V
F233L
L177F
M501I
N129D
NSP15(NendoU)
Mutations
T33I
NSP16 (2'−O−MT)
Mutations
K160R
R216C
S (SPIKE)
Mutations
A222V
A262S
A570D
D1118H
D614G
ΔH69/ΔV70
L18F
L5F
N439K
N501Y
P681H
S477N
S982A
S98F
T716I
ΔY145
N (Nucleoprotein)
Mutations
A220V
A376T
A398V
D103Y
D377Y
D3L
G204R
L5F
M234I
P13L
P199L
P365S
P67S
R203K
S194L
S197L
S235F
M (Membrane Protein)
Mutations
D3G
T175M
NS3 (Virporin)
Mutations
G172R
G172V
G196V
G251V
K75N
Q38R
Q57H
T223I
V13L
V202L
NS7b
Mutations
H37L
S5L
NS8
Mutations
A65V
L84S
Q27stop
R52I
S24L
V62L
dec
Feb
Apr
Jun
Aug
Oct
Dec

Figure 2


A
S (SPIKE)
Database Mutation Frequency (%)
100
75
50
25
0
dec
Feb
Apr
Jun
Aug
Oct
Dec
Mutations
A222V
A262S
A570D
D1118H
D614G
ΔH69/ΔV70
L18F
L5F
N439K
N501Y
P681H
S477N
S982A
S98F
T716I
ΔY145
B
NSP12 (RdRp)
Mutations
A185S
A97V
E254D
P323L
V776L
C
N (Nucleoprotein)
60
40
20
Mutations
A220V
A376T
A398V
D103Y
D377Y
D3L
G204R
L5F
M234I
P13L
P199L
P365S
P67S
R203K
S194L
S197L
S235F

D
D614G (S), P323L (NSP12)
R203K, G204R (N)
A220V (N)
A222V (S)
Dec 2019
Feb 2020
Apr 2020
Jun 2020
Aug 2020
Oct 2020
Dec 2020

Figure 3

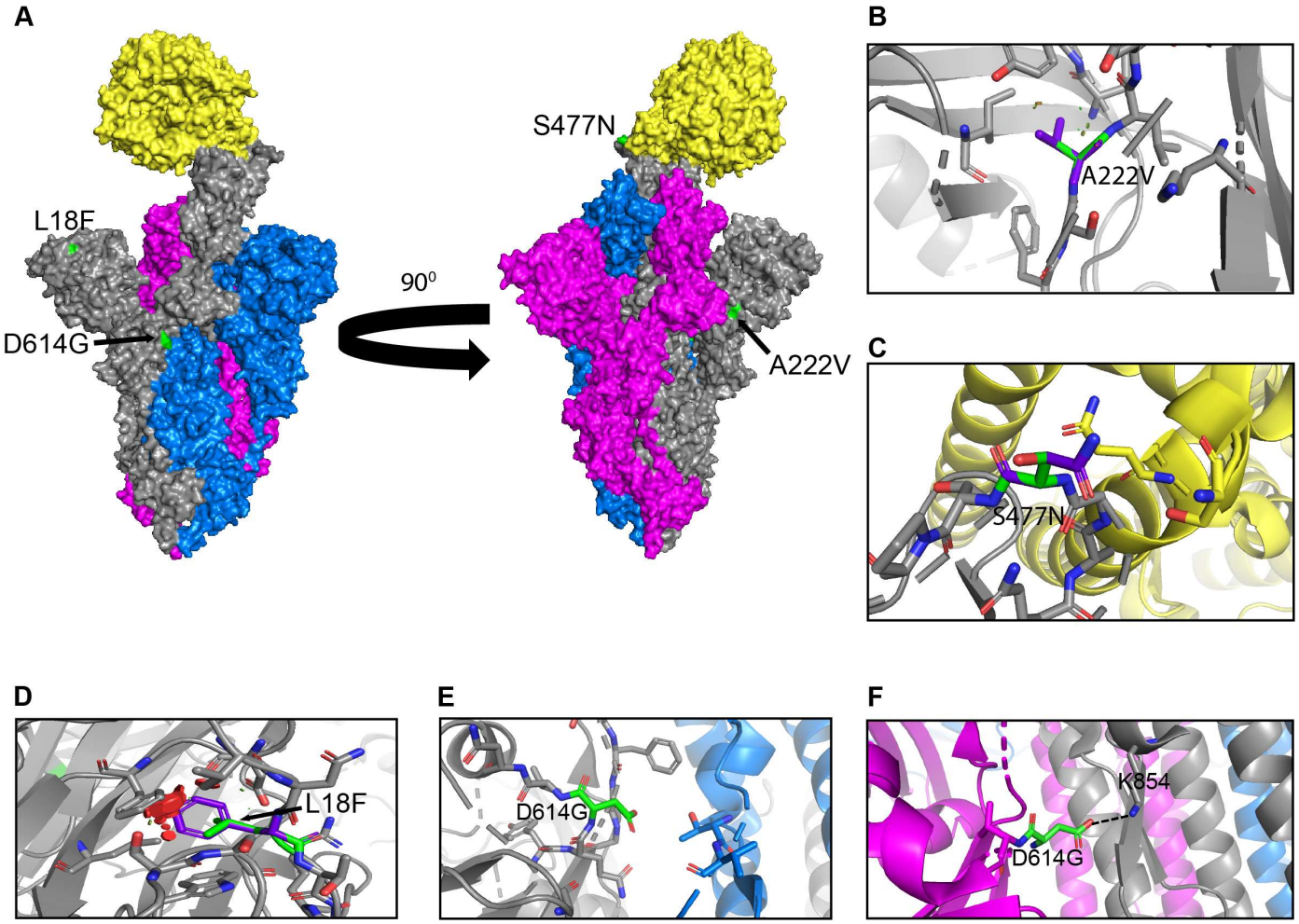

A
L18F
D614G
90⁰
S477N
A222V
B
A222V
C
S477N
D
L18F
E
D614G
F
K854
D614G

Figure 4

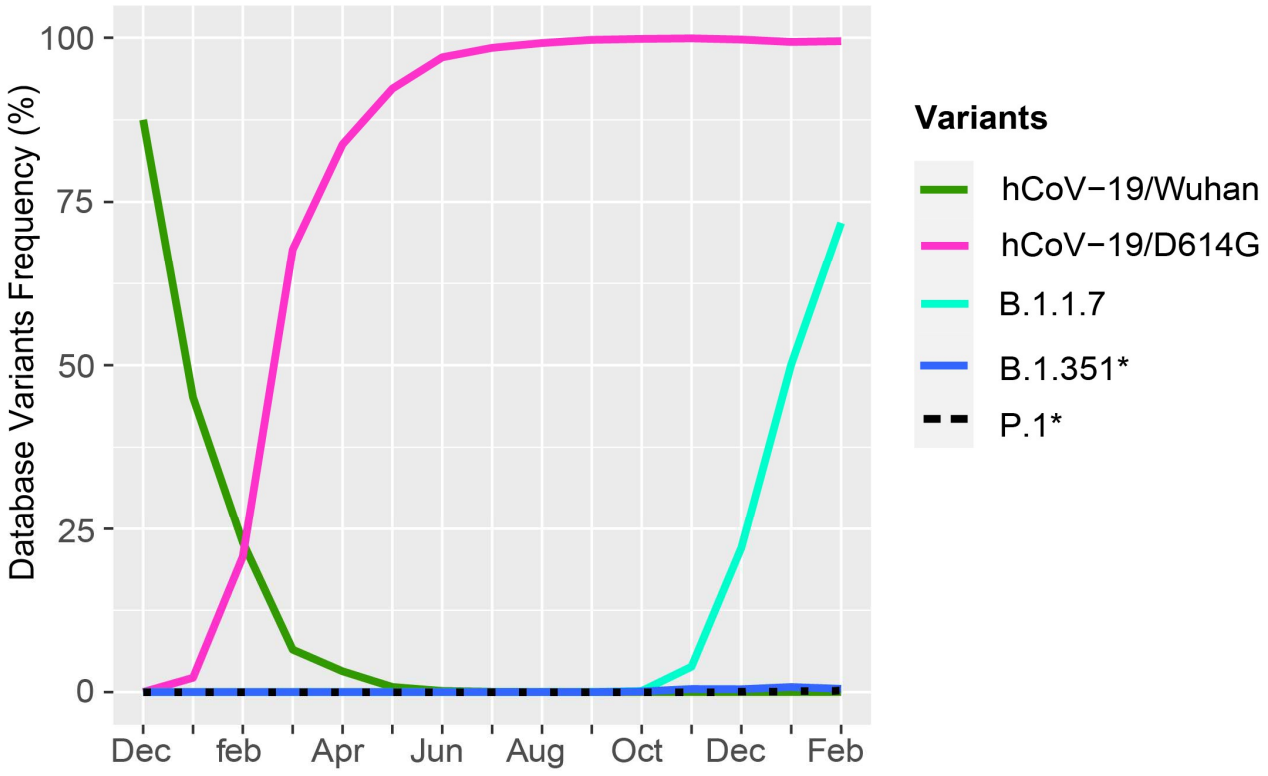

Database Variants Frequency (%)
100
75
50
25
0
Dec
feb
Apr
Jun
Aug
Oct
Dec
Feb
Variants
hCoV−19/Wuhan
hCoV−19/D614G
B.1.1.7
B.1.351*
P.1*

Figure 5

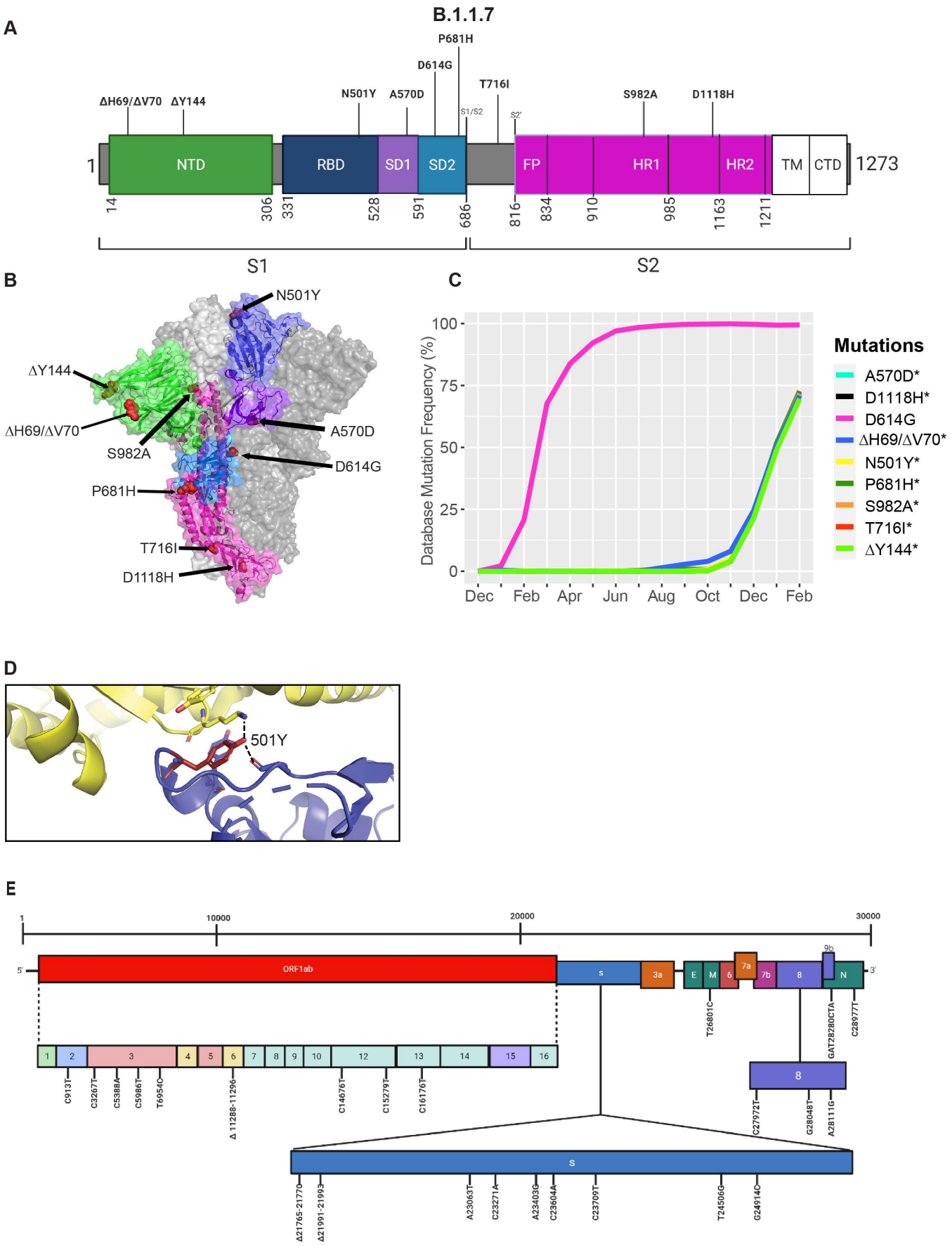

A
B.1.1.7
P681H
D614G
T716I
ΔH69/ΔV70
ΔY144
N501Y
A570D
S1/S2
S2'
S982A
D1118H
1
NTD
RBD
SD1
SD2
FP
HR1
HR2
TM
CTD
1273
14
306
331
528
591
686
816
834
910
985
1163
1211
S1
S2
B
N501Y
ΔY144
ΔH69/ΔV70
S982A
A570D
D614G
P681H
T716I
D1118H
C
Database Mutation Frequency (%)
100
75
50
25
0
Dec
Feb
Apr
Jun
Aug
Oct
Dec
Feb
Mutations
A570D*
D1118H*
D614G
ΔH69/ΔV70*
N501Y*
P681H*
S982A*
T716I*
ΔY144*
D
501Y
E
1
10000
20000
30000
5'
3'
ORF1ab
S
3a
E
M
6
7a
7b
8
9b
N
T26801C
GAT28280CTA
C28977T
1
2
3
4
5
6
7
8
9
10
12
13
14
15
16
C913T
C3267T
C5388A
C5986T
T6954C
Δ 11288-11296
C14676T
C15279T
C16176T
8
C27972T
G28048T
A28111G
S
Δ21765-21770
Δ21991-21993
A23063T
C23271A
A23403G
C23604A
C23709T
T24506G
G24914C

Figure 6

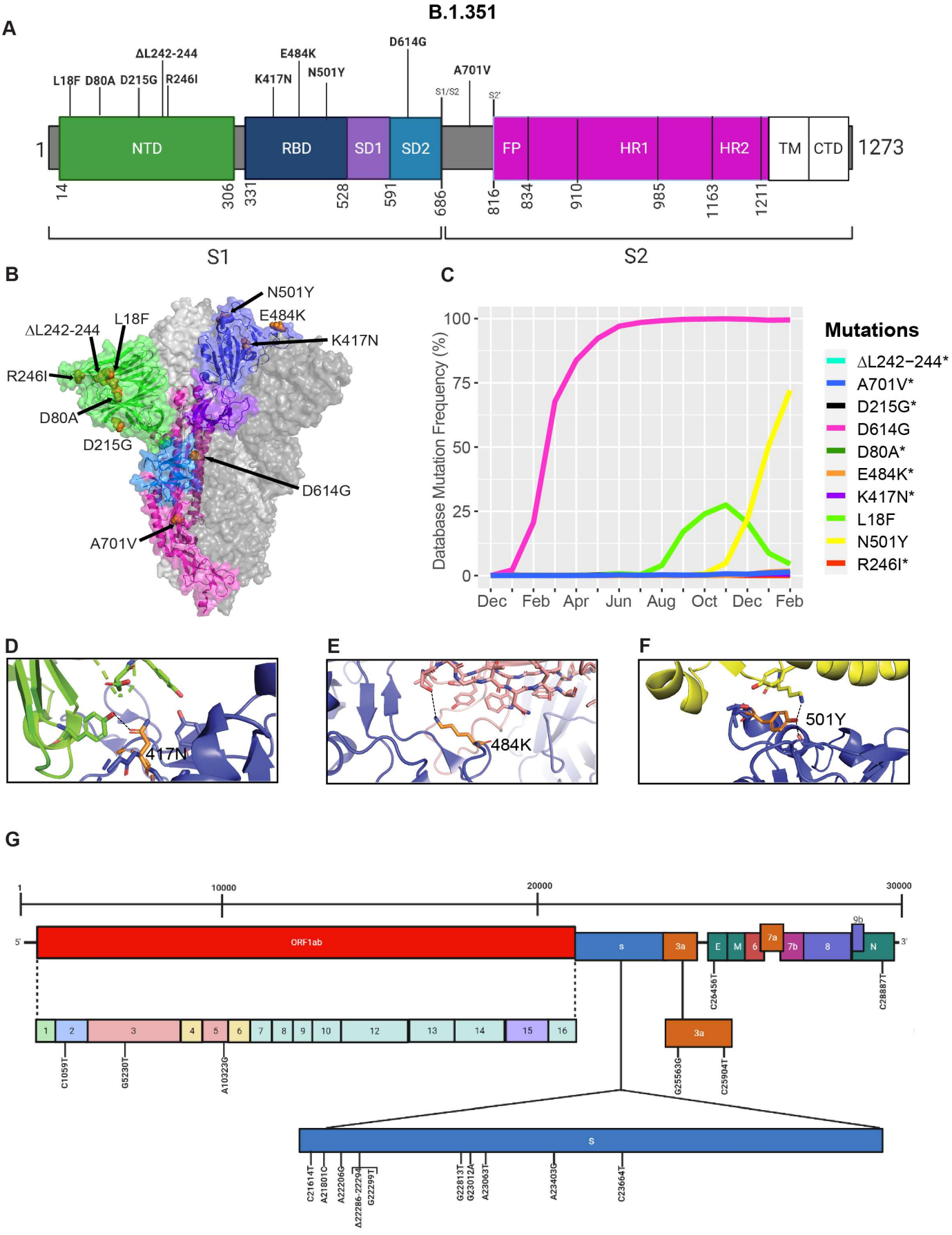
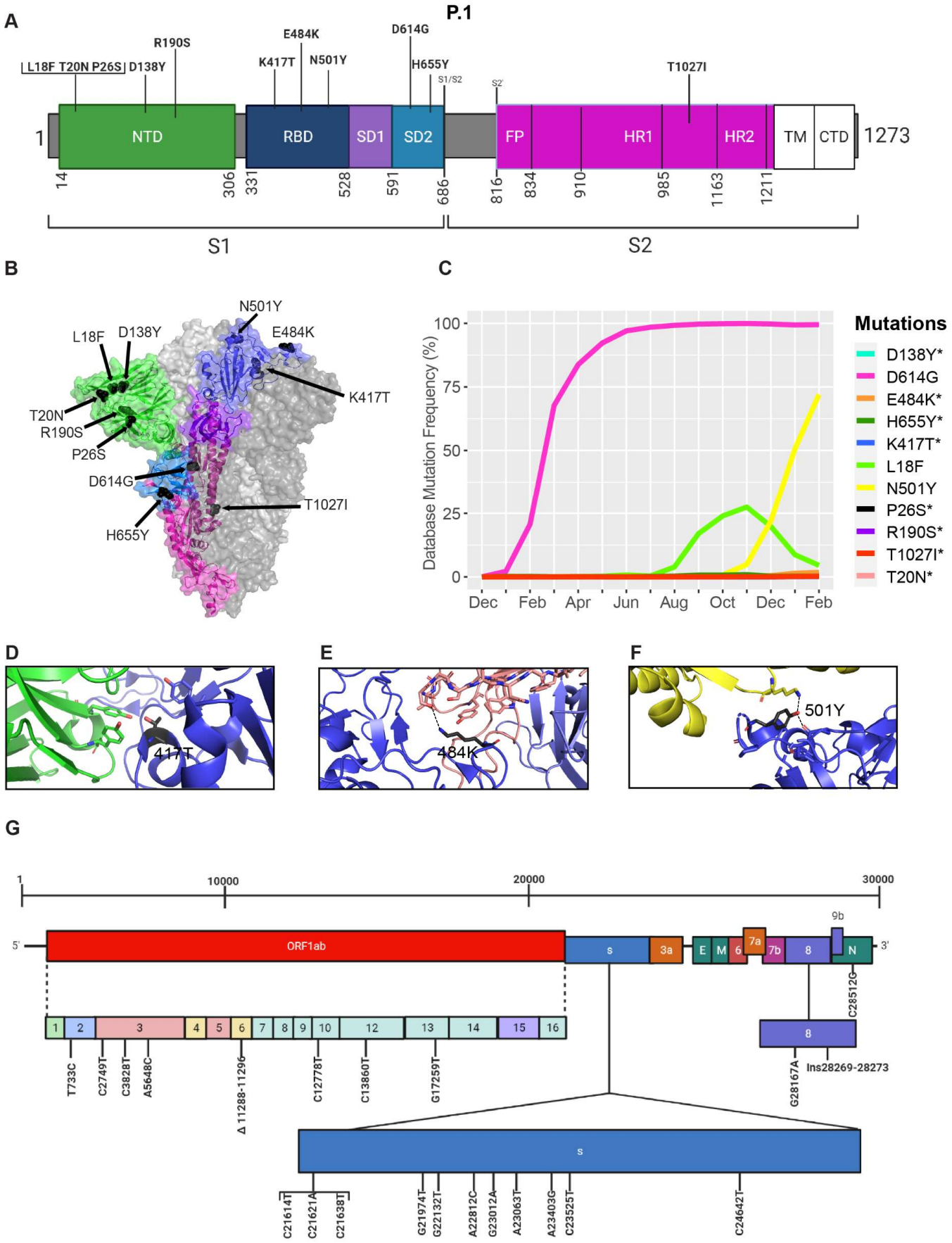

A
B.1.351
ΔL242-244
L18F D80A D215G R246I
E484K
K417N N501Y
D614G
A701V
S1/S2
S2'
1
NTD
RBD
SD1
SD2
FP
HR1
HR2
TM
CTD
1273
14
306
331
528
591
686
816
834
910
985
1163
1211
S1
S2
B
N501Y
E484K
ΔL242-244
L18F
K417N
R246I
D80A
D215G
D614G
A701V
C
Database Mutation Frequency (%)
100
75
50
25
0
Dec
Feb
Apr
Jun
Aug
Oct
Dec
Feb
Mutations
ΔL242−244*
A701V*
D215G*
D614G
D80A*
E484K*
K417N*
L18F
N501Y
R246I*
D
417N
E
484K
F
501Y
G
1
10000
20000
30000
5'
3'
ORF1ab
s
3a
E
M
6
7a
7b
8
9b
N
C26456T
C28887T
1
2
3
4
5
6
7
8
9
10
12
13
14
15
16
C1059T
G5230T
A10323G
3a
G25563G
C25904T
S
C21614T
A21801C
A22206G
Δ22286-22294
G22299T
G22813T
G23012A
A23063T
A23403G
C23664T

Figure 7


A
P.1
L18F T20N P26S
D138Y
R190S
K417T
E484K
N501Y
D614G
H655Y
S1/S2
S2'
T1027I
1
NTD
RBD
SD1
SD2
FP
HR1
HR2
TM
CTD
1273
14
306
331
528
591
686
816
834
910
985
1163
1211
S1
S2
B
N501Y
L18F
D138Y
E484K
K417T
T20N
R190S
P26S
D614G
T1027I
H655Y
C
Database Mutation Frequency (%)
100
75
50
25
0
Dec
Feb
Apr
Jun
Aug
Oct
Dec
Feb
Mutations
D138Y*
D614G
E484K*
H655Y*
K417T*
L18F
N501Y
P26S*
R190S*
T1027I*
T20N*
D
417T
E
484K
F
501Y
G
1
10000
20000
30000
5'
3'
ORF1ab
s
3a
E
M
6
7a
7b
8
9b
N
C28512G
1
2
3
4
5
6
7
8
9
10
12
13
14
15
16
T733C
C2749T
C3828T
A5648C
Δ 11288-11296
C12778T
C13860T
G17259T
G28167A
Ins28269-28273
s
C21614T
C21621A
C21638T
G21974T
G22132T
A22812C
G23012A
A23063T
A23403G
C23525T
C24642T

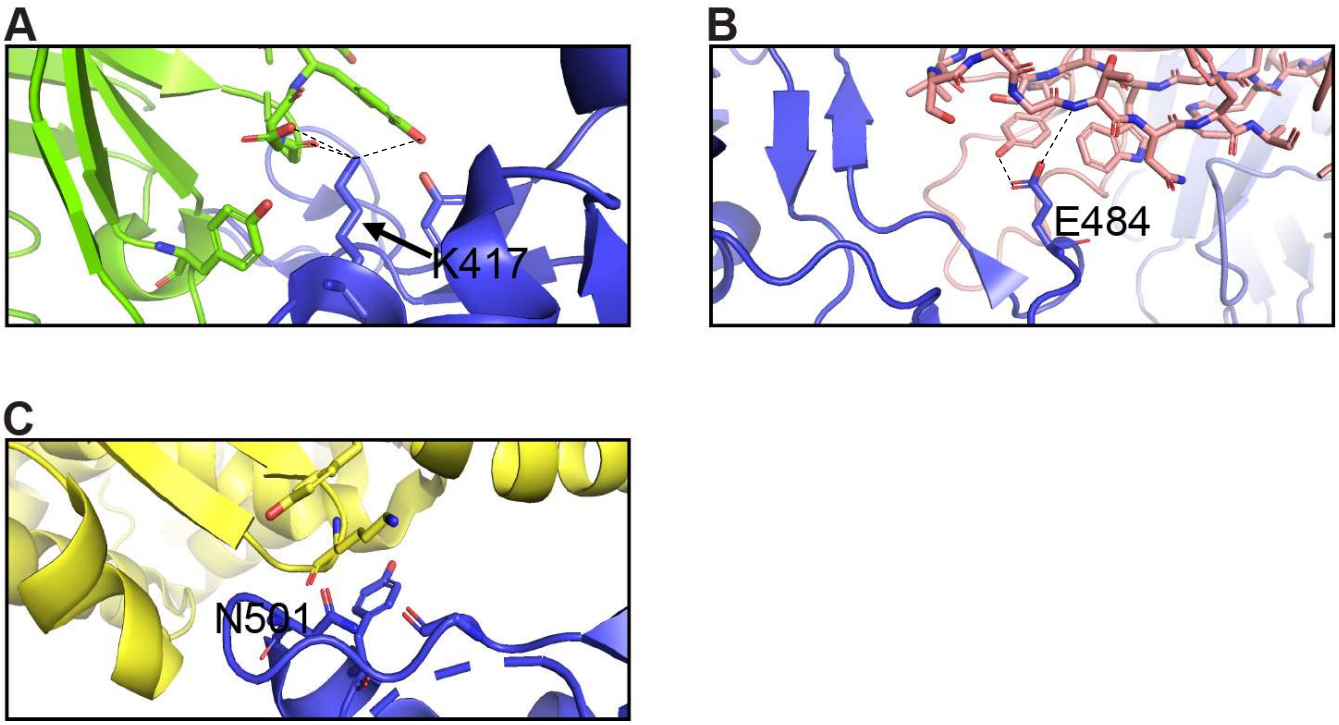



Figure 8