AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology®

GENETIC DIVERSITY AND EVOLUTION

Genome-Wide Variation in Betacoronaviruses

Katherine LaTourrette,[a,b,c] Natalie M. Holste,[a,b] Rosalba Rodriguez-Peña,[a,b] Raquel Arruda Leme,[a] Hernan Garcia-Ruiz[a,b]

[a]Nebraska Center for Virology, University of Nebraska-Lincoln, Lincoln, Nebraska, USA
[b]Department of Plant Pathology, University of Nebraska-Lincoln, Lincoln, Nebraska, USA
[c]Complex Biosystems Interdisciplinary Life Sciences Program, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

ABSTRACT The *Severe acute respiratory syndrome coronavirus* (SARS-CoV) and SARS-CoV-2 originated in bats and adapted to infect humans. Several SARS-CoV-2 strains have been identified. Genetic variation is fundamental to virus evolution and, in response to selection pressure, is manifested as the emergence of new strains and species adapted to different hosts or with novel pathogenicity. The combination of variation and selection forms a genetic footprint on the genome, consisting of the preferential accumulation of mutations in particular areas. Properties of betacoronaviruses contributing to variation and the emergence of new strains and species are beginning to be elucidated. To better understand their variation, we profiled the accumulation of mutations in all species in the genus *Betacoronavirus*, including SARS-CoV-2 and two other species that infect humans: SARS-CoV and *Middle East respiratory syndrome coronavirus* (MERS-CoV). Variation profiles identified both genetically stable and variable areas at homologous locations across species within the genus *Betacoronavirus*. The S glycoprotein is the most variable part of the genome and is structurally disordered. Other variable parts include proteins 3 and 7 and ORF8, which participate in replication and suppression of antiviral defense. In contrast, replication proteins in ORF1b are the least variable. Collectively, our results show that variation and structural disorder in the S glycoprotein is a general feature of all members of the genus *Betacoronavirus*, including SARS-CoV-2. These findings highlight the potential for the continual emergence of new species and strains with novel biological properties and indicate that the S glycoprotein has a critical role in host adaptation.

IMPORTANCE Natural infection with SARS-CoV-2 and vaccines triggers the formation of antibodies against the S glycoprotein, which are detected by antibody-based diagnostic tests. Our analysis showed that variation in the S glycoprotein is a general feature of all species in the genus *Betacoronavirus*, including three species that infect humans: SARS-CoV, SARS-CoV-2, and MERS-CoV. The variable nature of the S glycoprotein provides an explanation for the emergence of SARS-CoV-2, the differentiation of SARS-CoV-2 into strains, and the probability of SARS-CoV-2 repeated infections in people. Variation of the S glycoprotein also has important implications for the reliability of SARS-CoV-2 antibody-based diagnostic tests and the design and deployment of vaccines and antiviral drugs. These findings indicate that adjustments to vaccine design and deployment and to antibody-based diagnostic tests are necessary to account for S glycoprotein variation.

KEYWORDS COVID-19, MERS-CoV, S protein, SARS-CoV, SARS-CoV-2, coronavirus, genomic variation, glycoprotein S, protein S, vaccine

Coronaviruses cause respiratory and intestinal infections in animals, including humans. Three species are highly pathogenic to humans: *Severe acute respiratory syndrome coronavirus* (SARS-CoV), first described in China in 2002; *Middle East respiratory syndrome coronavirus* (MERS-CoV), first described in South Arabia in 2012 (1); and

Citation LaTourrette K, Holste NM, Rodriguez-Peña R, Leme RA, Garcia-Ruiz H. 2021. Genome-wide variation in betacoronaviruses. J Virol 95: e00496-21. https://doi.org/10.1128/JVI.00496-21.

Editor Colin R. Parrish, Cornell University

Address correspondence to Hernan Garcia-Ruiz, hgarciaruiz2@unl.edu.

Received 20 March 2021
Accepted 4 May 2021
Accepted manuscript posted online 12 May 2021
Published 12 July 2021

TABLE 1 Betacoronavirus (order *Nidovirales*, family *Coronaviridae*, subfamily *Coronavirinae*) nucleotide accessions used in this study[a]

Subgenus species	No. of accessions	No. of complete genomes	Reference accession no.	Length (nt)
Embecovirus				
Bovine coronavirus	1148	111	AF220295.1	31,100
Camel coronavirus HKU23	22	9	MN514966.1	31,075
Canine respiratory coronavirus	60	3	LR721664.1	31,190
Equine coronavirus	37	4	EF446615.1	30,992
Human coronavirus HKU1	416	48	DQ415901.1	30,097
Human coronavirus OC43	1386	178	MN306053.1	30,818
Murine coronavirus	258	38	AC_000192.1	31,526
Porcine hemagglutinating encephalomyelitis	92	13	KY994645.1	30,684
Rattus coronavirus HKU24	4	4	NC_026011.1	31,249
Hibecovirus				
Bat Hp-betacoronavirus Zhejiang2013	1	1	NC_025217.1	31,491
Merbecovirus				
Betacoronavirus erinaceus	10	6	KC545386.1	30,175
Bat coronavirus HKU4	87	11	EF065508.1	30,316
Bat coronavirus HKU5	66	10	MH002342.1	30,529
MERS-CoV	1351	572	MG987420.1	30,484
Nobecovirus				
Rousettus bat coronavirus HKU9-1	116	10	EF065516.1	29,155
Rousettus bat coronavirus GCCDC1	27	3	NC_030886.1	30,161
Sarbecovirus				
Bat coronavirus	2	2	GU190215	29,276
Bat SARS coronavirus	45	28	MN996532.1	29,855
Bat SARS-like coronavirus	147	19	KY417150.1	30,311
SARS-CoV	637	254	MK062183.1	29,874
SARS-CoV-2	2,379	2,315	NC_045512.2	29,903

[a]For each virus species, one annotated accession describing the full genome was used as a reference. Only accessions describing complete genomes were used for nucleotide variation analyses. Accessions were downloaded from NCBI on 6 April 2020. For SARS-CoV-2, additional accessions were downloaded on 13 May 2020.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), first detected in December 2019 in Wuhan, China (2, 3).

Coronaviruses consist of a group of four genera (*Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus*, and *Deltacoronavirus*) in the order *Nidovirales*, family *Coronaviridae*, subfamily *Coronavirinae* (2, 3). Gammacoronaviruses and deltacoronaviruses infect birds, and some species infect mammals. Alphacoronaviruses and betacoronaviruses (β-CoVs) infect mammals, primarily bats and humans (1). The genus *Betacoronavirus* contains five subgenera (Table 1): *Embevorirus*, *Merbecovirus*, *Nobecovirus*, *Hibecovirus*, and *Sarbecovirus* (2, 3). The subgenus *Sarbecovirus* contains species that infect bats or humans. Among species that infect humans, the closest relatives to SARS-CoV-2 are SARS-CoV and MERS-CoV (Table 1) (3, 4).

β-CoVs have a monopartite, linear, positive-strand RNA genome of approximately 30,000 nucleotides (nt). The virion is spherical, enveloped, and about 120 nm in diameter (1, 5). Genomic RNA associates with the nucleoprotein (N) to form a nucleocapsid. The membrane (M) protein forms part of the envelope, which also contains the small membrane E protein. The virion surface displays spikes formed by the S glycoprotein (6–8) that mediate cell entry by interacting with cellular receptors and entry cofactors (9–12). The S glycoprotein is divided into S1 and S2 subunits that are separated by proteolytic cleavage via cellular proteases and cofactors (12–15). The virion contains large and small spikes formed by S1 and S2 together and by S2 subunits, respectively (8). In subunit S1, the carboxyl-terminal domain contains a core and the receptor binding subdomains (1, 7–11).

Several lines of evidence show that β-CoVs, including SARS-CoV-2, are evolving and accumulate mutations in their genome (1, 16). For RNA viruses, important sources of

genetic variation are RNA recombination and nucleotide insertions, deletions, and substitutions introduced during RNA replication (17). Genetic variation combined with selection pressure imposed by genetically diverse hosts favors the accumulation of mutations that support the emergence of new virus strains and species (18–20). These general principles of virus evolution explain several features of β-CoVs, including SARS-CoV-2. Both SARS-CoV and SARS-CoV-2 likely emerged through RNA recombination of two species infecting bats. The recombinant progeny then adapted to infect humans (3). Within a year of the initial description, several SARS-CoV-2 strains have been detected that mainly accumulate mutations in the S glycoprotein (16, 21). Temporal and spatial genetic relationships show accumulation of mutations in the genome, allowing SARS-CoV-2 isolates to be differentiated (22, 23). In the United States, a variant carrying the Q677P substitution in the S glycoprotein is now abundant in the southwest, and new mutations within this variant have created sublineages (24). Furthermore, variants infecting the same individual have been detected (20, 25–27).

The S glycoprotein is the common target for neutralizing antibodies developed in response to natural infection or vaccines (28–31). Neutralizing antibodies are formed against the prefusion conformation of the entire S glycoprotein. In contrast, nonneutralizing antibodies are formed against the S2 subunit (8, 32). Antibodies against the S glycoprotein are used as markers in diagnostic assays (28, 29, 32). Accordingly, the emergence of new variants with mutations in the S glycoprotein has the potential to compromise the efficacy of vaccines and the immunity mediated by natural infection (20, 27, 33–36). Conversely, antibodies developed through natural infection or vaccines may impose selection pressure on β-CoVs (20, 36). For these and other reasons, it is imperative to understand the biological properties of β-CoVs that contribute to the emergence of new strains and species.

Virus variation, evolution, and adaptation to diverse hosts are mediated by genetic determinants in the viral genome and selection pressure imposed by the host (37–40). Accordingly, characterization of genomic variation is fundamental to our understanding of β-CoV evolution and host adaptation. We hypothesized that genetic determinants of variation in β-CoVs are conserved across species, including SARS-CoV-2. In this study, we profiled the genomic variation in all species in the genus *Betacoronavirus*. Genome-wide nucleotide variation analyses combined with amino acid variation analyses revealed that variation patterns are conserved across β-CoVs, including the presence of variable areas at homologous locations. The most variable parts are the S glycoprotein, followed by open reading frame 8, accessory proteins 3 and 7, and the N protein. Genome-wide distribution of mutations of all β-CoVs provides an explanation for the emergence of new β-CoV species, such as SARS-CoV-2, and for the emergence of strains. These findings and published results (16, 20, 24, 41, 42) predict how and where SARS-CoV-2 will accumulate mutations and differentiate into new biological strains. SARS-CoV-2 will likely evolve as it adapts to genetically diverse human populations (20, 43) and possibly to selection constraints imposed by vaccines, antiviral drugs, and antibodies developed against natural infections (20, 30, 36). Our results underscore the potential for the continual emergence of new β-CoV species and strains with novel biological properties.

RESULTS

S glycoprotein is more variable than the rest of the genome. SARS-CoV-2 strains identified to date (16, 21) differ in the accumulation of nonsynonymous substitutions in the entire genome (Fig. 1A). However, nonsynonymous substitutions preferentially accumulate in the S glycoprotein and in the N protein. In all other parts of the genome, mutations accumulate to a frequency that is equal to or less than that expected randomly (Fig. 1B). New mutations continue to arise and diversify SARS-CoV-2 into variants (20, 24). However, recurrent mutations mapped along the SARS-CoV-2 genome preferentially accumulate in the S glycoprotein (22, 23).

To further test preferential accumulation of mutations in the S glycoprotein, we measured single-nucleotide polymorphisms (SNPs) in the entire SARS-CoV-2 genome

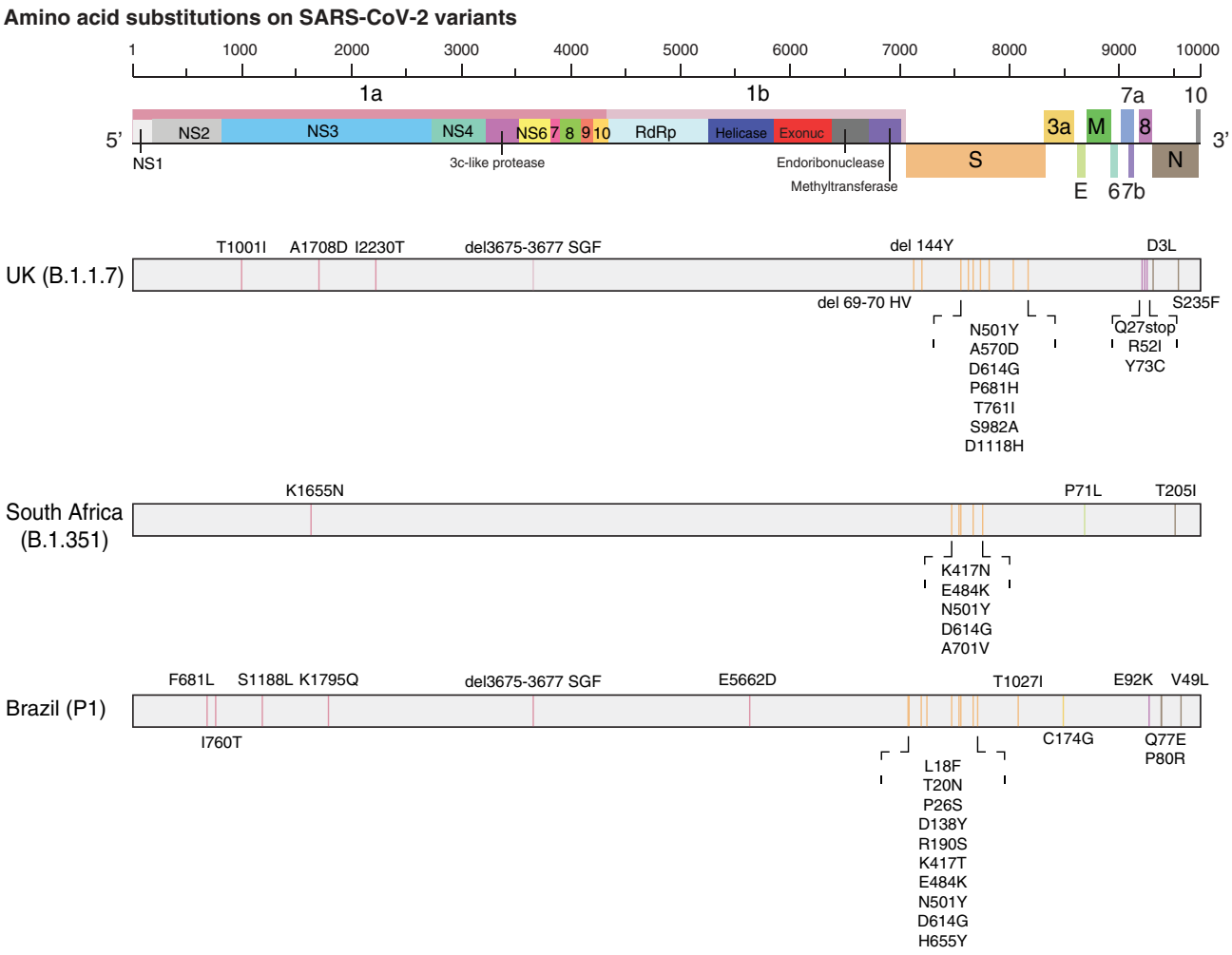

FIG 1 Amino acid substitutions in UK, South Africa, and Brazil SARS-CoV-2 strains. (A) Location of amino acid mutations on the SARS-CoV-2 genome. Coordinates are based on accession no. NC_045512.2. (B) Proportion of amino acid substitutions observed and randomly expected based on the length of each protein or open reading frame. Strains are color coded.

and separately in the S glycoprotein, normalized to their respective length. SARS-CoV was included in the analysis as it was closely related to SARS-CoV-2. Nucleotide accessions were added in increments of 50 or 10 for SARS-CoV-2 and SARS-CoV, respectively. In both species, the S glycoprotein accumulated more polymorphic sites than the rest of the genome proportionately (Fig. 2A). Using a chronological approach, SARS-CoV-2 accessions were analyzed by month from December 2019 to April 2020. For SARS-CoV, accessions were analyzed by year from 2003 to 2017. The number of polymorphic sites

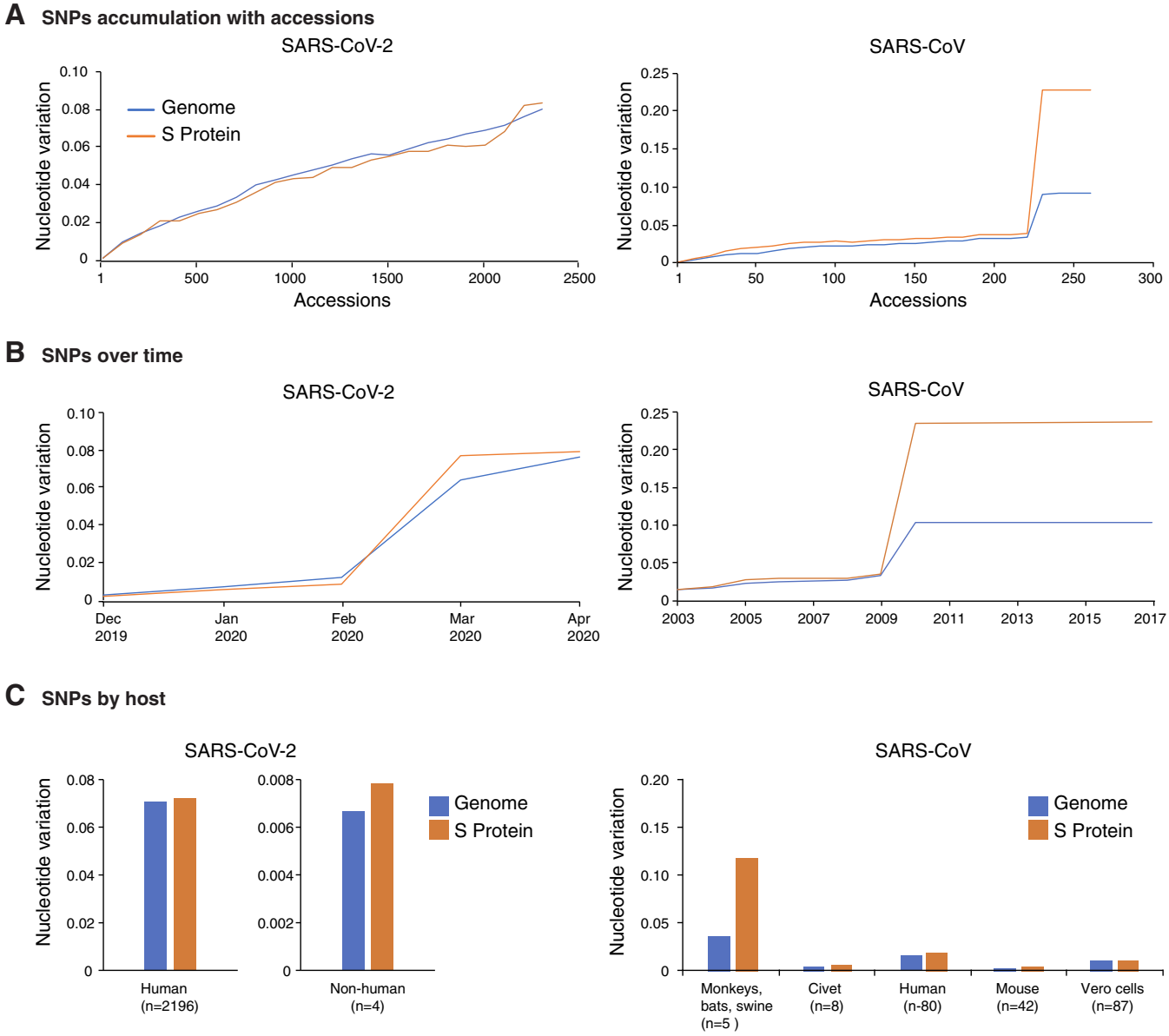

FIG 2 Nucleotide variation in SARS-CoV-2 and SARS-CoV. (A) Nucleotide variation over accessions. (B) Nucleotide variation over time. (C) Nucleotide variation by host. The number of accessions in the analysis is indicated in parenthesis.

in the S glycoprotein was similar to or higher than that of the rest of the genome in SARS-CoV and SARS-CoV-2 over time (Fig. 2B) and in all hosts (civets, humans, mice, and Vero cells) (Fig. 2C). In SARS-CoV-2, the S glycoprotein represents approximately 17% of the genome (Fig. 1B). However, it accumulated at least 50% of the mutations (Fig. 1B). Accordingly, the S glycoprotein accumulated mutations at a frequency that is at least 3-fold higher than would be expected randomly. The frequency was even higher in accessions derived from nonhuman hosts and in SARS-CoV (Fig. 2C). These results show that the SARS-CoV-2 S glycoprotein is the most variable part of the genome.

Differentiation of SARS-CoV-2 based on the S glycoprotein. The variable nature of the S glycoprotein suggests it is a major contributor to SARS-CoV-2 evolution and diversification. To further test this model, using accessions from the United States, we generated a phylogenetic tree using the S glycoprotein and two time periods. Based on nucleotide (Fig. 3A) or amino acid sequence (Fig. 3B), accessions from January 2020 clustered separately from accessions from July and August 2020. With respect to the

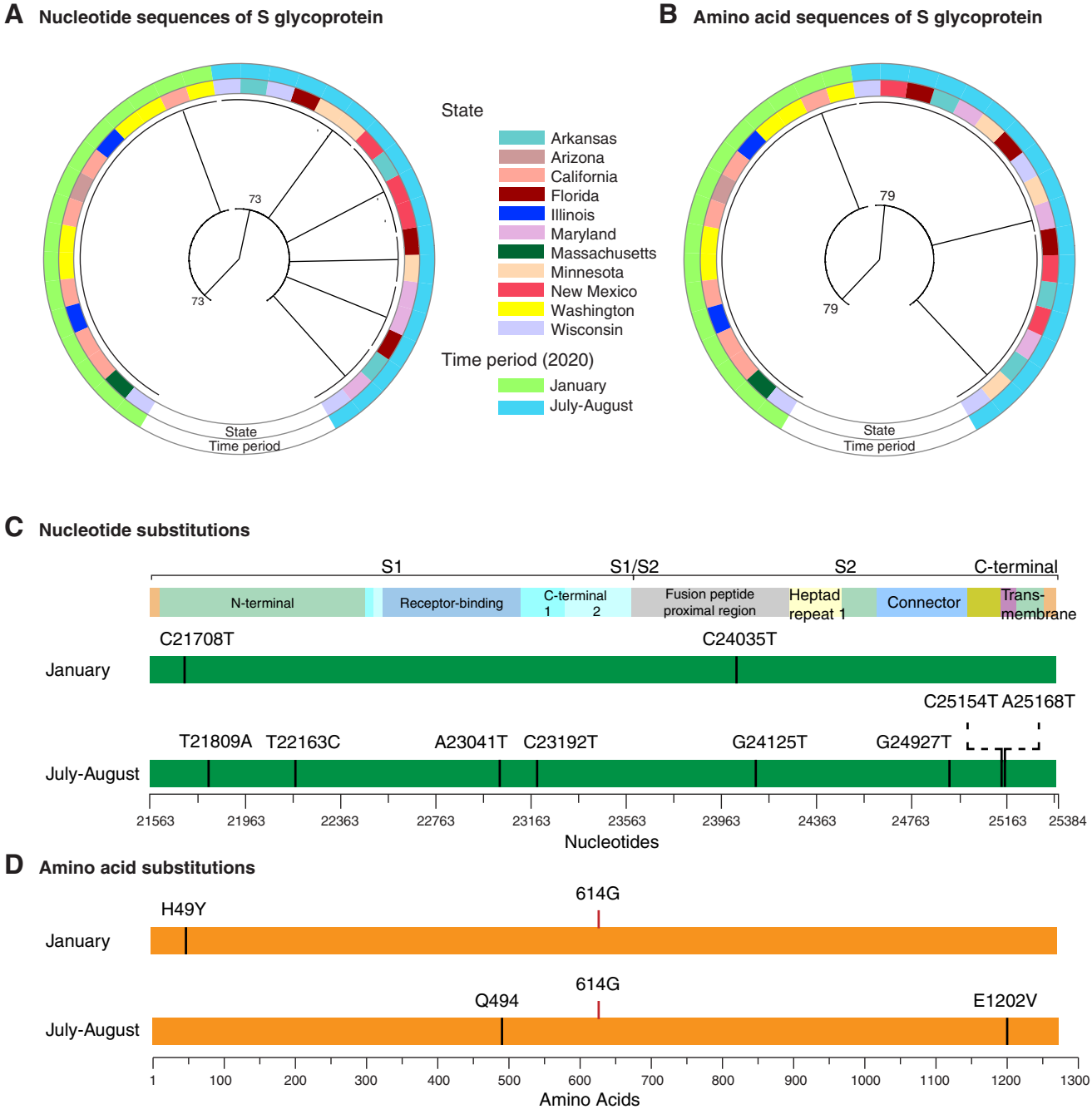

FIG 3 Phylogram and mutations of U.S. SARS-CoV-2 accessions based on the S glycoprotein. (A) Phylogenetic tree of the coding sequence for S protein based on nucleotide accessions from early (January) and middle (July and August) 2020. Color-coded rings indicate the time period and state of origin. (B) Phylogenetic tree of S protein generated using amino acid sequences from early and late time periods. (C) Mutations found in nucleotide accessions from early and late time periods. (D) Mutations found in protein accessions from early and late time periods. Amino acid 614 (G) is indicated.

reference sequence Wuhan-Hu-1 (NC_045512.2), U.S. accessions from January accumulated two nucleotide substitutions (Fig. 3C) and one amino acid (H49Y) substitution that maps to the N-terminal domain of the S1 subunit (Fig. 3D). These mutations were not present on accessions from July and August 2020. Instead, eight new nucleotide (Fig. 3C) and two new amino acid mutations were detected, a Q494L substitution in the receptor binding motif and an E1202V substitution in the transmembrane motif (Fig. 3D). In both time periods, all accessions contained the D614G mutation (Fig. 3D). Consistent with recent observations (16, 24), these results support the model that

mutations in the S glycoprotein mediate the emergence of new strains and that the S glycoprotein is a major contributor to SARS-CoV-2 evolution.

Genome-wide variation in SARS-CoV-2. Although nonsynonymous substitutions preferentially accumulate in the S glycoprotein, other areas of the genome also accumulate mutations (Fig. 1). To characterize SARS-CoV-2 genome variation, we compared accessions early and later in the pandemic. Using a 50-nt window, variation in SARS-CoV-2 was measured for the 106 nucleotide accessions available on NCBI on 23 March 2020, the 2,315 nucleotide accessions available on 13 May 2020, and the 2,299 amino acid accessions available on 14 May 2020. No variable areas in the genome were detected in accessions representing the early part of the pandemic (Fig. 4A). However, by May 2020, variation was detected in several areas (Fig. 4B). SNPs were higher than the average of the genome in the S glycoprotein, ORF8, and at the N-terminal part of ORF 1a (Fig. 4B). A single-amino-acid polymorphism (SAP) analysis showed variation in the S glycoprotein maps to the N-terminal domain, the receptor binding domain, the fusion peptide-proximal region, the heptad repeat 2, and the transmembrane domain (Fig. 4C). These results are consistent with the detection of recurrent deletions that map to the N-terminal domain in immunocompromised patients (20). An order/disorder analysis showed that, in SARS-CoV-2, the S glycoprotein has intrinsically disordered areas in the receptor binding domain, C-terminal domain 2, and the fusion peptide proximal region (Fig. 4D). Intrinsically disordered regions often interact with multiple molecular partners, are highly plastic, and show high evolutionary rates (44). Consistent with these results, major mutations (D614G and Q677P) that render the virus more transmissible and pathogenic to humans (24, 34, 45) map near the hypervariable region in the disordered C-terminal domain 2 in S1 (Fig. 4C). The S1/S2 cleavage site is located within a variable region (Fig. 4D). In the U.S. accession subset analyzed here (Fig. 3), no variation was detected in the S1/S2 cleavage site. However, in the larger data set, mutations were detected at the S1/S2 cleavage site (Fig. 4C). Because variation in the S1/S2 cleavage site contributes to cellular tropisms and pathogenesis (13), these results suggest that the S1/S2 cleavage site tolerates mutations and can contribute to SARS-CoV-2 diversification.

Collectively, these observations support the model that the S glycoprotein is variable and mutationally robust and contains intrinsically disordered areas.

Genome-wide variation in betacoronaviruses. The variation pattern described above could be a property exclusive to SARS-CoV-2, to a subset of species, or a general property of β-CoVs. To distinguish the difference, we profiled the genome variation in all members of the genus *Betacoronavirus*. To ensure statistical power (46), the analyses described here were based on species with three or more accessions. At least three accessions describing complete genomes were available for 19 of the 21 species in the genus *Betacoronavirus*. The length of the genomes ranged from 29,855 to 31,190 nt (Table 1).

We measured nucleotide variation in all members of the genus *Betacoronavirus* (Table 1). SNPs and nucleotide diversity, estimated in a 50-nt window, showed that β-CoVs in general and species in the subgenus *Sarbecovirus* in particular are highly variable (Fig. 5A). Due to its high variability (47), and as a point of comparison, we estimated HIV-1 variation using the same method. The most diversity was observed in *Rousettus bat coronavirus* HKU9, other species infecting bats, and MERS-CoV (Fig. 5A). In these species, more than 25% of the nucleotides in the genome were polymorphic. Genomic variation is not a function of the number of accessions, because similar results were observed using the nucleotide diversity index (Pi), which normalizes for the number of accessions (48) (Fig. 5A).

Variation in HIV-1 (Fig. 6) was higher than that for all β-CoVs, with one exception (Fig. 5A). Based on nucleotide diversity (Pi), the genome of *Rousettus bat coronavirus* accumulated more variation than HIV-1. The next three most variable species included bat SARS coronaviruses, and their nucleotide variation varied from 57% to 86% of that observed for HIV-1 (Fig. 5A). In contrast, nucleotide diversity in SARS-CoV and SARS-CoV-2 was approximately 10% of that observed for HIV (Fig. 5A). Variation estimated for all β-CoVs was higher than that of polioviruses (0.1%), known for being genetically

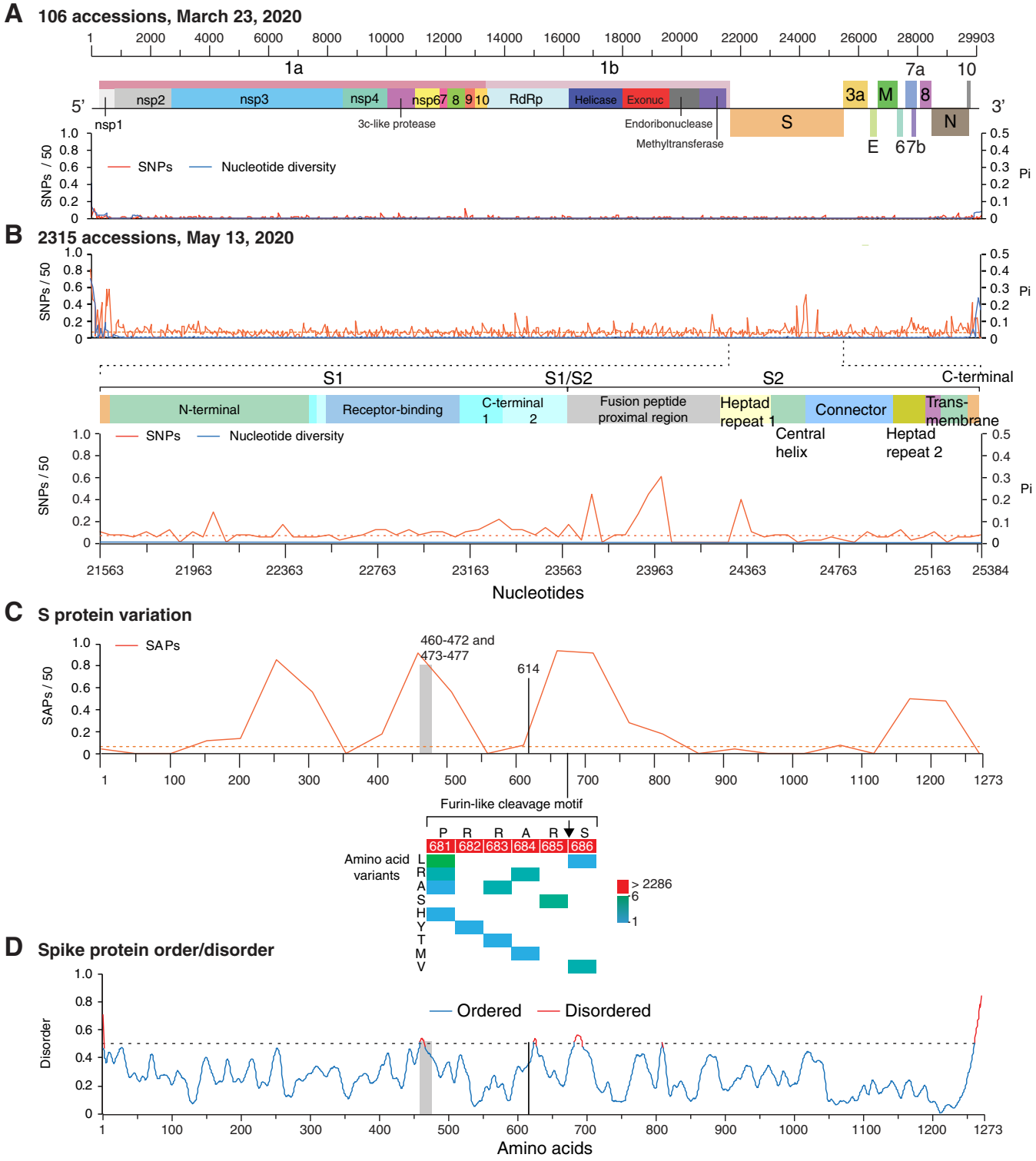

FIG 4 Variation in SARS-CoV-2 genome and S protein. (A) Single-nucleotide polymorphism (SNP) and nucleotide diversity (Pi) plotted with respect to the SARS-CoV-2 genome based on accessions available on 23 March 2020. A 99% confidence interval ($P < 0.01$) is indicated as a horizontal line for each parameter. Nucleotide and amino acid coordinates are based on accession no. NC_045512.2. (B) Variation in SARS-CoV-2 downloaded on 13 May 2020 with 2,315 accessions. (C) Single-amino-acid polymorphism (SAP) plotted with respect to SARS-CoV-2 S protein based on accessions available 14 May 2020. A vertical gray bar represents the ACE2 binding domain. Amino acid 614 is indicated. Mutations in the Furin-like cleavage motif and their frequency are indicated with a heat map. (D) Order and disorder in S protein.

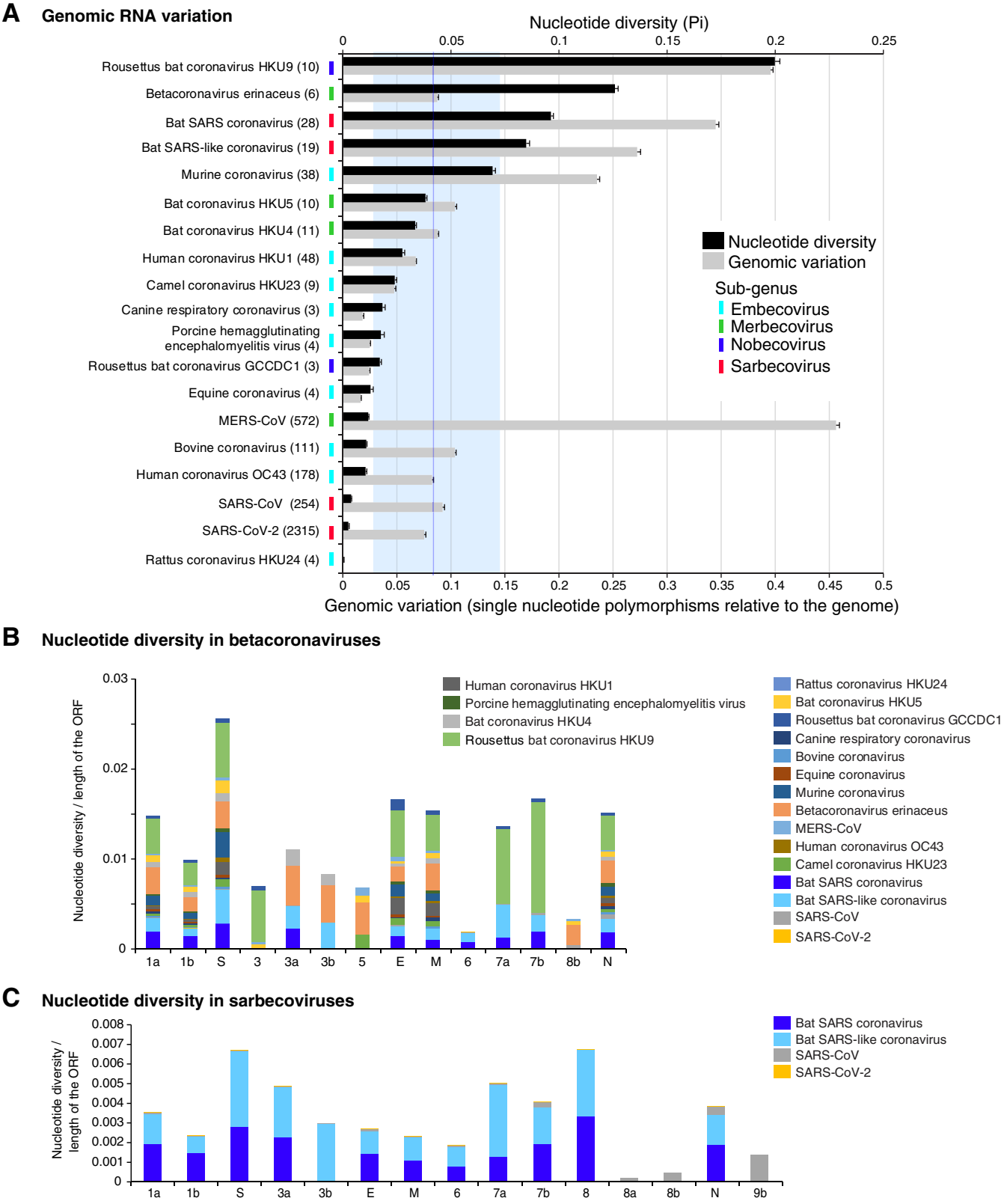

FIG 5 Genomic variation in β-CoVs. (A) Nucleotide variation as determined by nucleotide diversity (Pi) and genomic variation index (proportion of single-nucleotide polymorphisms with respect to the length of the genome). Bars represent the averages and standard errors for each species. The vertical line represents the mean Pi and 99% confidence interval ($P < 0.01$). Virus names are color-coded by genera. (B) Cumulative nucleotide diversity by ORF in the genome of all β-CoVs. (C) Cumulative nucleotide diversity by ORF in the genome of all sarbecoviruses.

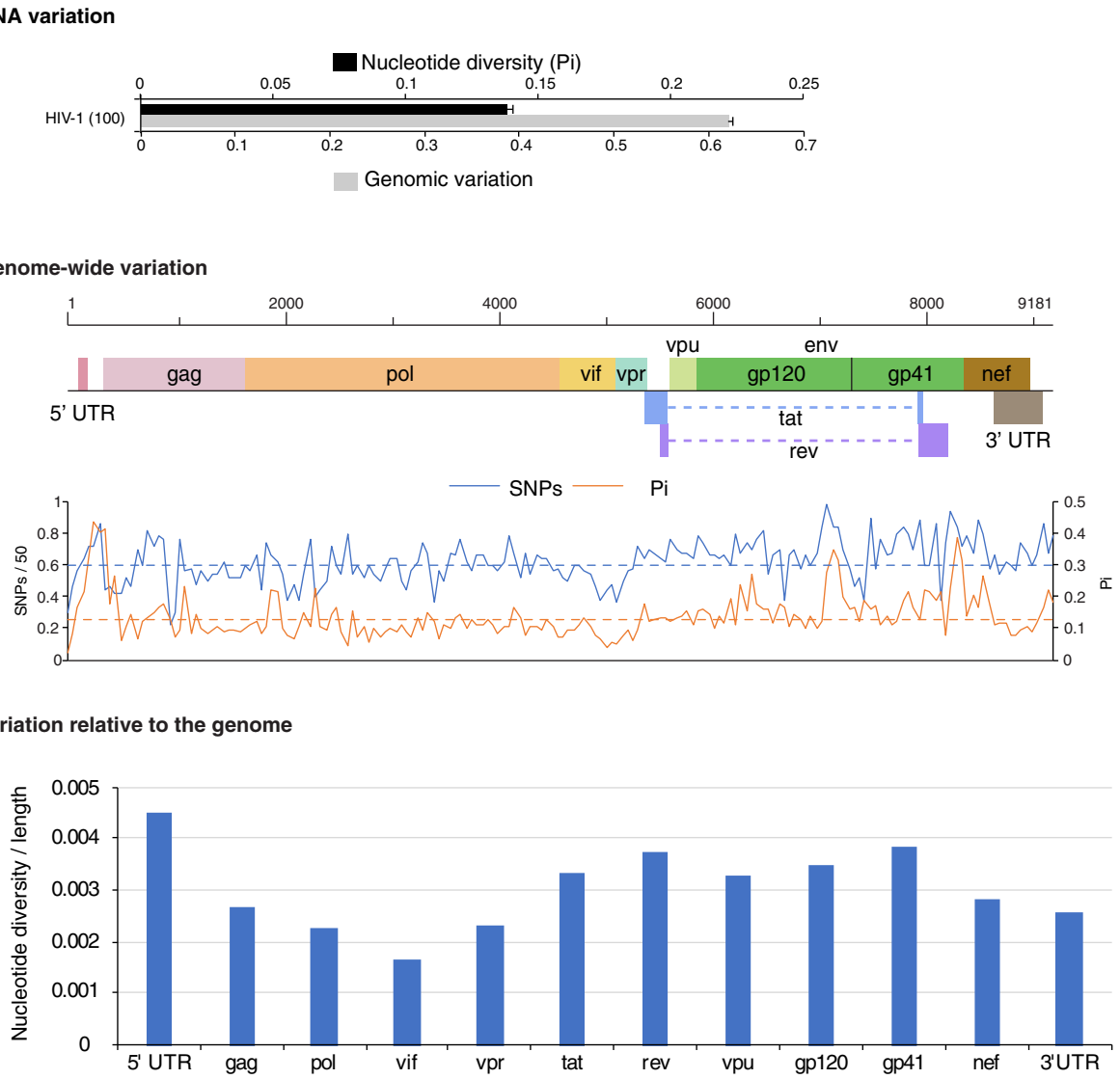

FIG 6 Genomic variation in HIV-1. (A) Nucleotide diversity and genomic variation index (single-nucleotide polymorphisms relative to the genome) estimated using 100 accessions. Bars represent the averages and standard errors. (B) Genome-wide nucleotide variation. Single-nucleotide polymorphism (SNP) and nucleotide diversity (Pi) are plotted with respect to the genome. The average and 99% confidence interval ($P < 0.01$) are indicated as a horizontal line for each parameter. Coordinates are based on accession no. NC_001802.1. (C) Nucleotide diversity normalized to the length of the ORF or UTR.

stable (49). The wide range of genomic variation across β-CoV species may reflect a bias in the source of the accessions, such as being obtained from genetically diverse hosts. However, these results show that β-CoVs have the potential to be highly variable.

The genome of β-CoVs consists of 11 to 14 open reading frames (ORFs). Coding regions for accessory and structural proteins located 3′ of the S proteins are not in synteny (1, 50). After normalizing to their length, the most variable parts of the genome were the S glycoprotein, followed by the E protein, protein 7, the M protein, and the N protein (Fig. 5B). The lowest variation was detected in open reading frame 1b (Fig. 5B), which codes for nonstructural proteins that mediate virus replication: RNA-dependent RNA polymerase, RNA helicase, exonuclease, endoribonuclease, and methyltransferase. Within the subgenus *Sarbecovirus*, the most variable part of the genome was the S glycoprotein, followed by ORF8, accessory proteins 3a, 7a, and 7b, and N protein. The lowest variation was detected in open reading frame 1b (Fig. 5C).

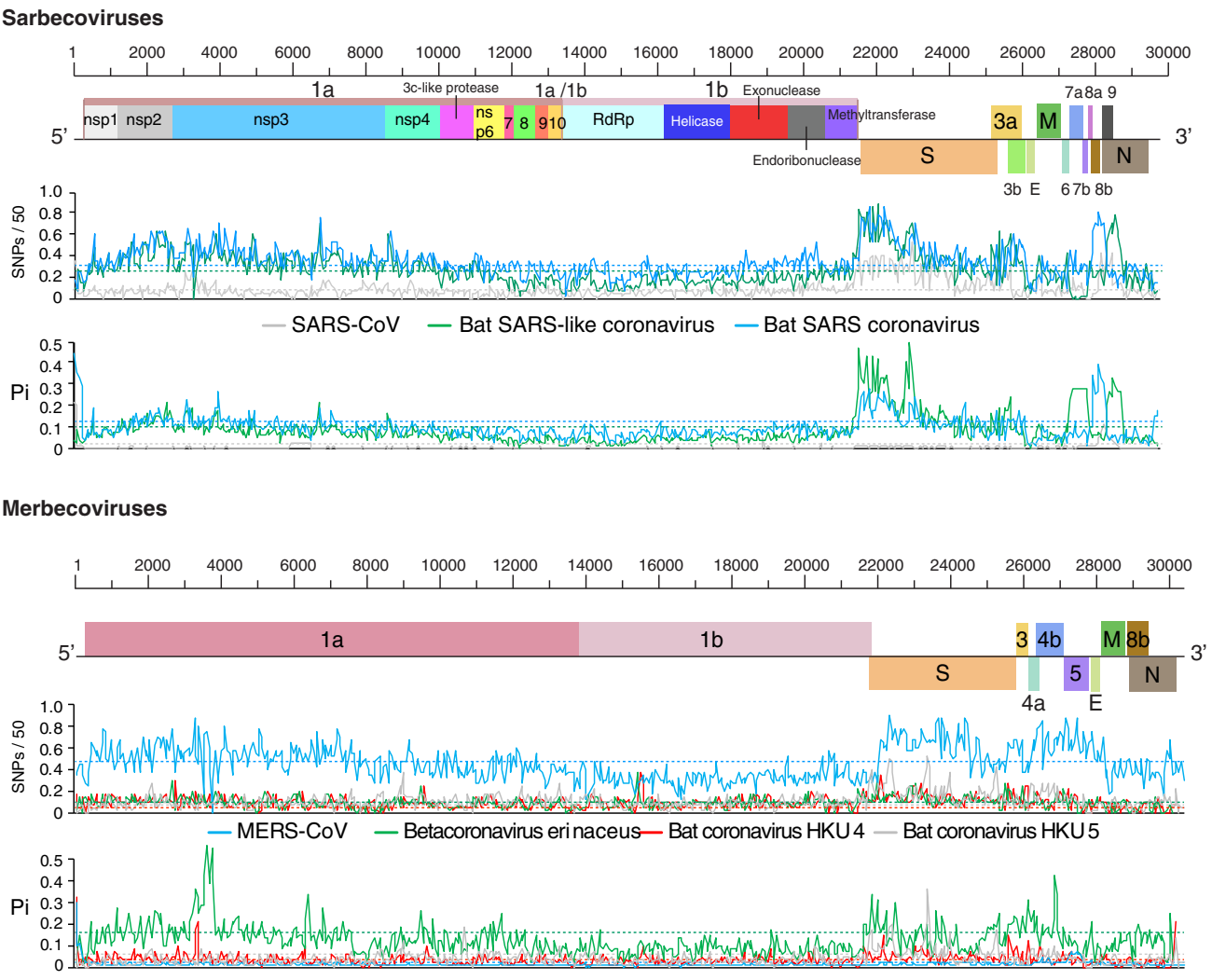

FIG 7 Genome-wide nucleotide variation in sarbecoviruses and merbecoviruses. Single-nucleotide polymorphism (SNP) and nucleotide diversity (Pi) were plotted with respect to the genome. The average and 99% confidence interval ($P < 0.01$) are indicated as a horizontal line for each parameter. Species are represented by colored horizontal lines, and ORFs are color coordinated. (A) Sarbecovirus variation. Coordinates are based on SARS-CoV accession no. MK062183.1. (B) Merbecovirus variation. Coordinates are based on MERS-CoV accession no. MG987420.1.

Genome-wide maps illustrate the distribution of nucleotide variation (Fig. 7). In the S glycoprotein, ORF8, and 3a, nucleotide polymorphisms were higher than the average of the genome. In the two species infecting bats, variation was also detected in nsp1 (inhibits host gene expression) (51), nsp2 (inhibits cell signaling) (52), and nsp3 (papain-like protease) (53). The lowest variation was detected in the 3′ half of ORF1a and in ORF1b (Fig. 7A). This pattern was observed in all species in the subgenus *Sarbecovirus*, including SARS-CoV (Fig. 7A).

In the subgenera *Merbecoviruses* (Fig. 7B), *Nobecovirus* (Fig. 8A), and *Embecovirus* (Fig. 8B), the S glycoprotein is the most variable part of the genome. Other areas of hypervariation include nsp1, nsp2, and the nsp3 protease, and the lowest variation was detected in the 3′ half of ORF1a and in ORF1b (Fig. 8).

Collectively, genome-wide variation described above show that, in all members of the genus *Betacoronavirus*, the S glycoprotein is the most variable part of the genome, and replication proteins in ORF1b are the least variable (Fig. 5, 7, and 8).

Betacoronavirus differentiation into strains. The S glycoprotein mediates receptor recognition and membrane fusion during viral entry into the cells (8–11). Consistent with this role, ACE2 receptor binding is a determinant of host range in sarbecoviruses (11, 54–57). For viruses in the subfamily *Torovirinae* within the family *Coronaviridae*,

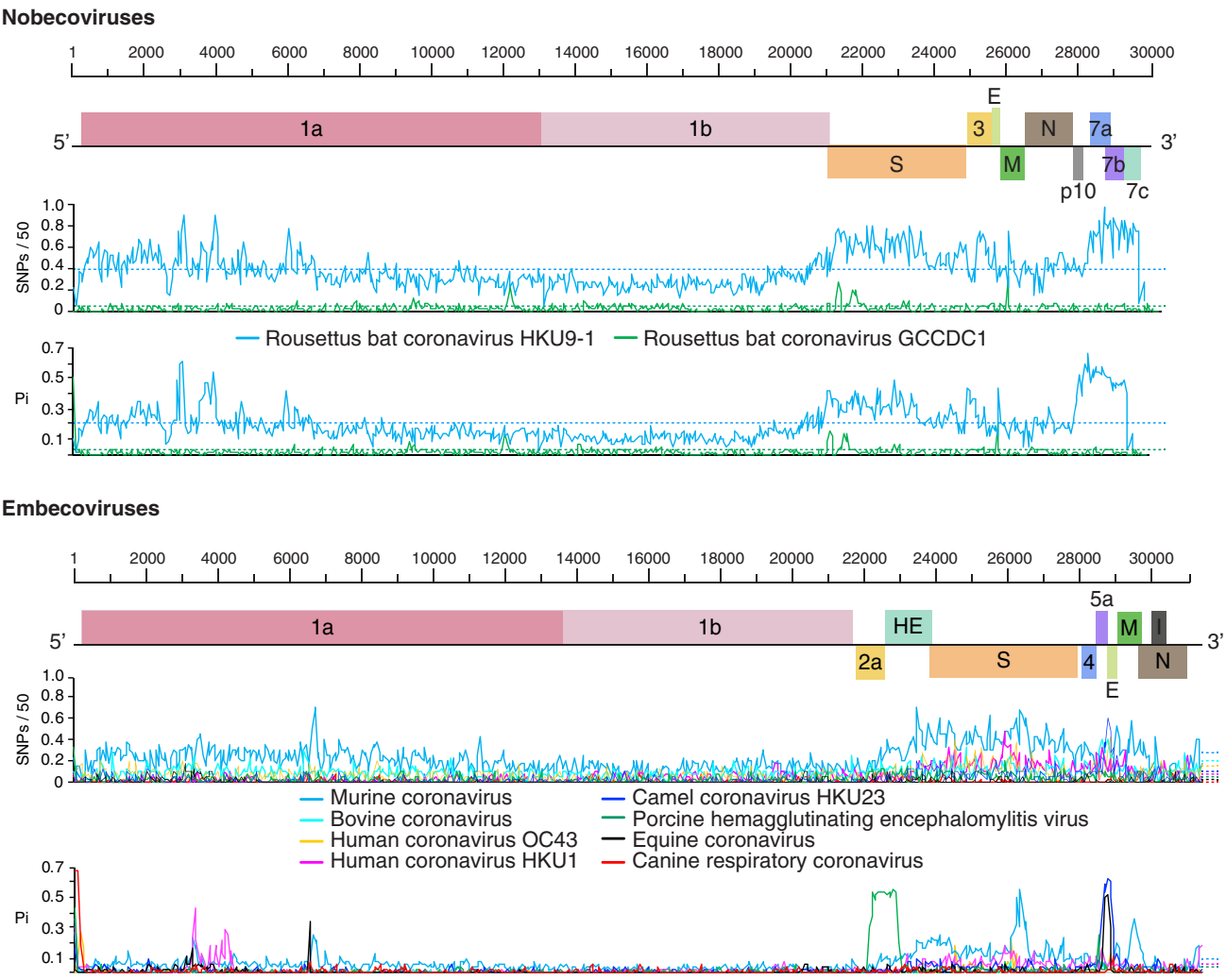

FIG 8 Genome-wide nucleotide variation in nobecoviruses and embecoviruses. Labels are as described for Fig. 7. (A) Nobecovirus variation. Coordinates are based on Rousettus bat coronavirus HKU9 accession no. EF065516.1. (B) Embecovirus variation. Coordinates are based on Bovine coronavirus accession no. AF220295.1.

receptor binding is a determinant of host range (58). In the β-CoV genome, as described above, the S glycoprotein is the most variable (Fig. 5B). These observations predict that coronaviruses differentiate into strains based on selection pressure from the host. To test this hypothesis, we generated a phylogeny based on the S glycoprotein for SARS-CoV, MERS-CoV, and closely related species infecting bats. In SARS-CoV (Fig. 9A) and MERS-CoV (Fig. 9B), accessions formed clusters that correlated with the country of origin and host species. Accessions representing bat SARS coronavirus (Fig. 9C) and bat SARS-like coronavirus (Fig. 9D) originated exclusively from China, and accessions clustered according to the host.

These results support the model that, in β-CoVs, variation in the S glycoprotein separates isolates into strains and may reflect the effect of selection imposed by the host. Consistent with this model, reinfection in humans has been confirmed (25, 27). In Brazil, a case of reinfection occurred with a new strain containing an E484K mutation in the S protein (27).

Amino acid variation in the S glycoprotein. Amino acid sequence in the S glycoprotein is variable in all species of the subgenus *Sarbecovirus* (Fig. 10). Variation localizes to subunit S1, particularly to the receptor binding domain, which is predicted to be intrinsically disordered for bat-SARS and bat-SARS-like coronaviruses (Fig. 10). Intrinsically disordered proteins mediate functional diversity and interactions with

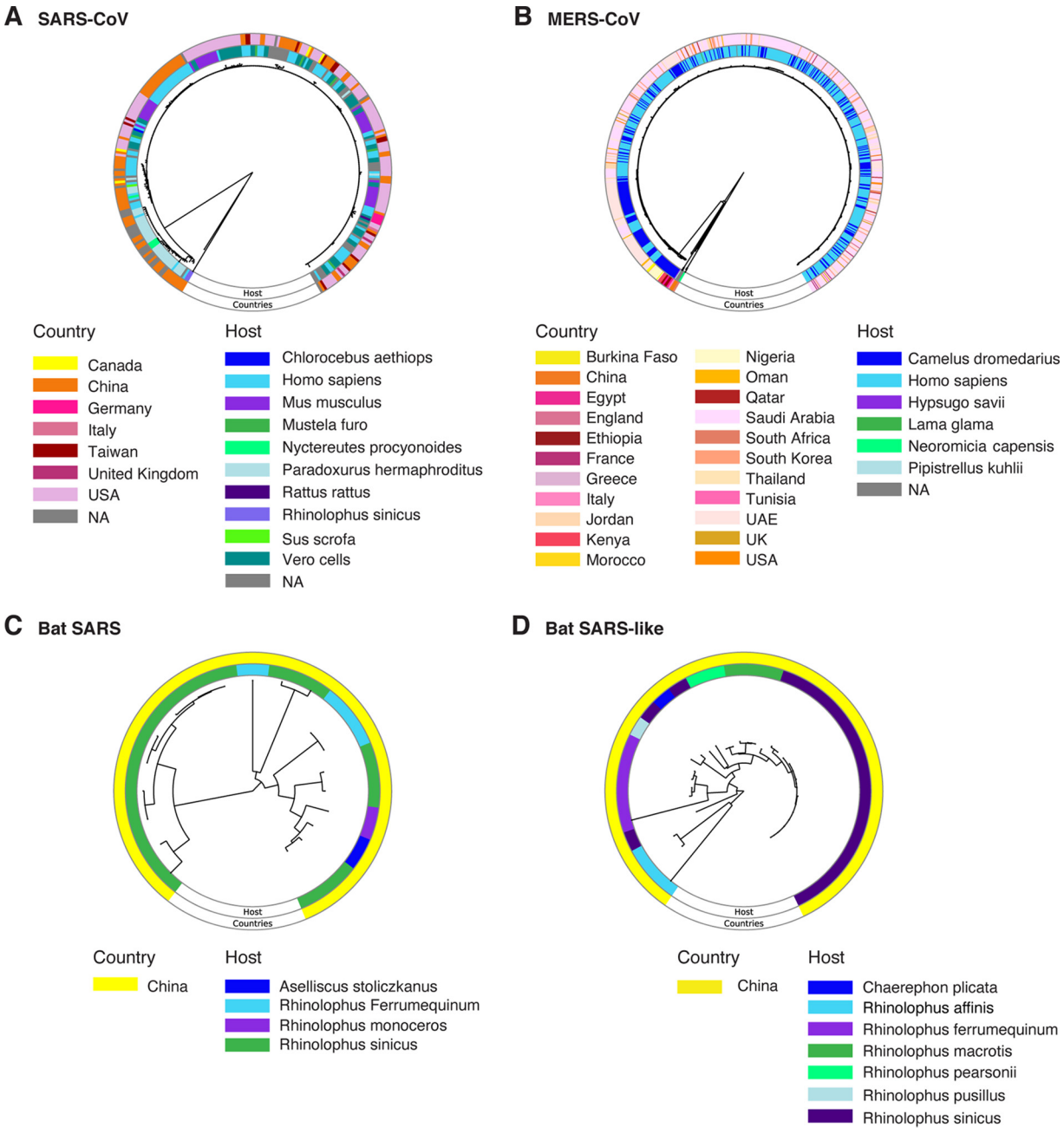

FIG 9 Phylogram based on the coding region for S protein in species closely related to SARS-CoV-2. Color-coded rings indicate hosts and country of origin. (A) SARS-CoV phylogenetic tree based on S protein amino acid sequences. (B) MERS-CoV. (C) Bat SARS coronavirus. (D) Bat SARS-like coronavirus.

multiple partners (59, 60). These observations support the model that, in β-CoVs, the S glycoprotein is mutationally robust and contains disordered areas.

DISCUSSION

Host and viral factors contribute to virus evolution (38, 61). The starting material is the introduction of mutations (nucleotide substitutions, insertions, or deletions) in the genome through RNA-dependent RNA polymerase errors during replication, RNA recombination, and reassortment (in segmented viruses) (39, 61). While they may occur randomly, selection separates beneficial from detrimental and neutral mutations. Selection is imposed by the host, the environment, and their interaction. Mutations

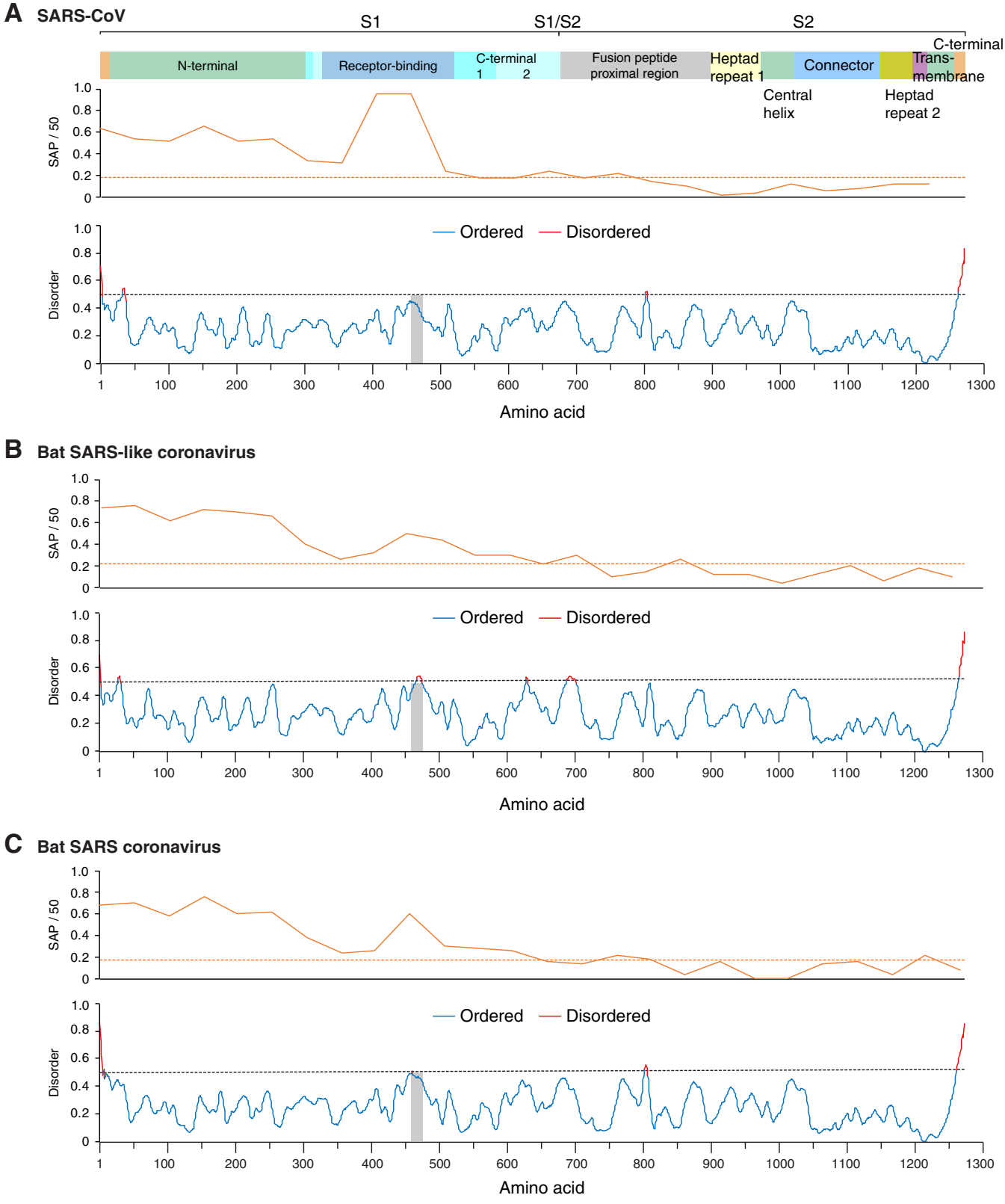

FIG 10 S protein amino acid variation and disorder in sarbecoviruses. Single-amino-acid polymorphism (SAP) and disorder/order of S protein. The average and 99% confidence interval are represented by the dotted line. The vertical gray line marks the location of the ACE2 receptor binding domain. (A) SARS-CoV-2. (B) Bat SARS-like coronavirus. (C) Bat SARS coronavirus.

that provide a beneficial advantage are more likely to be fixed in the genome (37). Under this scenario, the distributions of mutations in the viral genome are not random. Instead, mutations accumulate to higher than random frequencies in areas of the genome that contribute to fitness by enhancing stability, transmission, replication efficiency, escape from immunity, suppression of immunity responses, or a combination (37–39). The collective results of these effects may be evident through biological properties such as host adaptation, pathogenicity, or others.

The emergence of new species, such as SARS-CoV-2, and the rapid emergence of new SARS-CoV-2 strains are an indication that β-CoVs evolve quickly and have a high capacity to switch hosts and to adapt to new hosts. Results described here show that β-CoVs are more variable than polioviruses (49). In some species, variation is close to that observed for HIV-1 (Fig. 5 and 6). The wide host range observed across species suggests that, in β-CoVs, genomic variation is related to the genetic diversity of the host.

While mutations accumulate in the entire genome, they are not randomly or equally distributed. Instead, preferential accumulation of mutations in the S glycoprotein is a general feature of all members of the genus *Betacoronavirus* (Fig. 5B). This was particularly evident in SARS-CoV-2. Strains identified to date (16, 21) and isolates from early and middle 2020 can be distinguished based on the S glycoprotein sequence alone (Fig. 3).

In HIV-1, glycoproteins gp120 and gp41 are the most variable in the genome (Fig. 6). Both β-CoV S glycoprotein and HIV gp120 and gp41 are envelope proteins and mediate viral entry. The S glycoprotein binds to the ACE2 receptor (55), while gp120 binds to the CD4 receptor (62). Both the S glycoprotein and gp120 induced the formation of neutralizing antibodies. These features suggest several mechanisms driving diversifying selection in envelope glycoproteins: cellular receptors, entry cofactors, and antibodies.

Within the variety of coronavirus hosts, the cellular receptors, entry cofactors, and cellular proteases that process the S1/S2 cleavage site and immunity responses are likely diverse (12, 63). Results described here show that, for β-CoVs, the S glycoprotein is variable and mutationally robust and contains intrinsically disordered areas (Fig. 4, 5B, and 9). Disordered proteins allow functionality with a diverse set of interaction partners (44). These observations are consistent with a model in which host diversity pushes diversifying selection in the S glycoprotein. Mutational and structural robustness in the S glycoprotein provide a selection advantage, are major contributors to β-CoV evolution, and may lead to the emergence of new strains and species.

In the reference sequence Wuhan-Hu-1 (NC_045512.2), the S glycoprotein contains residues that are compatible, but not optimal, for binding human receptor ACE2 (55). Accordingly, SARS-CoV-2 has the potential to accumulate mutations for more efficient entry into human cells and to escape from neutralizing antibodies. Consistent with this model, SARS-CoV-2 strains detected to date mainly differ in the S glycoprotein (Fig. 1A) (16, 21). The D614G and Q677P mutations make the virus more transmissible and more pathogenic to humans (24, 45) and have been detected in several parts of the world (33, 64). Amino acids 614 and 677 are near hypervariable C-terminal domain 2 in subunit S1 (Fig. 4C). Furthermore, the D614G mutation and others in the receptor binding domain reduce affinity to monoclonal antibody CR3022 (64). In Mexico (65) and Wisconsin (Fig. 3), the H49Y mutation in the S protein was the most frequent and appears to have independent origins, suggesting convergent evolution.

In humans, the strength of the immune responses is not uniform, as indicated by immunocompromised patients (2, 20), and immunity-driven selection in SARS-CoV-2 has been documented (66). The human population is genetically diverse enough to select for variants in SARS-CoV and MERS-CoV (Fig. 9), and there is genetic variability in human leukocyte antigen genes that affect susceptibility to SARS-CoV-2 and the severity of the disease (67). Thus, it is likely that SARS-CoV-2 will continue to accumulate mutations for efficient transmission and genome replication and differentiate into

biological strains as the virus faces selection pressure from genetically distinct human populations or immunocompromised individuals. For example, a 27-amino-acid deletion was detected in protein 7 in Arizona (68), new mutations formed sublineages in the southwest (24), and SARS-CoV-2 populations were grouped into clades in Mexico (65).

Our analysis identified proteins 3 and 7 and ORF8 as variable in SARS-CoV and bat SARS (Fig. 5C). The 3b protein is an inhibitor of the interferon response, and a variant with a longer 3b protein induces more severe symptoms and has enhanced ability to suppress induction of type I interferon (69). The 7a protein antagonizes antiviral factor BST-2 to enhance virion release (70). ORF8 is a cofactor of the RNA-dependent RNA polymerase and an inhibitor of the type I interferon response (71–73). Further, a 29-nt deletion in ORF8 attenuated SARS-CoV, and mutations or deletions in SARS-CoV-2 ORF8 caused attenuation. Although no difference was detected *in vivo*, higher replication was detected *in vitro* than in the wild-type virus (66, 74). ORF8 is different between SARS-CoV and SARS-CoV-2 and does not include functional motifs (50). Because ORF8 induces a robust antibody response, deletions may reflect immunity-driven selection (66). Due to their biological role in replication, suppression of antiviral defense, and hypervariable nature (Fig. 5C), proteins 3 and 7 and ORF8 are likely contributors to pathogenicity and host adaptation in sarbecoviruses.

Vaccines against SARS-CoV-2 induce neutralizing antibodies against the prefusion conformation of the S glycoprotein (8, 75, 76). However, nonneutralizing antibodies against subunit S2 are also developed (8). Variation within the S2 subunit is among the highest in the genome (Fig. 4). Nonneutralizing antibodies may provide a mechanism for the virus to escape from the immune response (77). Thus, variation in the S glycoprotein provides β-CoVs a mechanism to escape the immune response and an important selection advantage.

Vaccines and antiviral drugs might function as selection agents (36). In an infected individual, new variants are generated (26) and may be selected to escape neutralizing antibodies, which were developed against natural infection or triggered by a vaccine. Indeed, in immunocompromised patients, recurrent deletions in the N-terminal domain were detected, and these deletions mediate escape from neutralizing antibodies (20).

Factors contributing to β-CoV evolution include intrinsic properties of the S glycoprotein (mutationally robust and intrinsically disordered), natural genetic diversity in their hosts, and diversity in the strength of the immune response. Similarly, several factors contribute to SARS-CoV-2 differentiation into strains, including natural genetic diversity in the human population, diversity in the strength of the immune response, and, possibly, selection imposed by vaccines. This represents a challenge for vaccine development and deployment, because vaccines may only be efficient against closely related strains, ineffective against diverse strains, and fail to prevent reinfection.

MATERIALS AND METHODS

All computational analyses were done on high-performance computing nodes. Custom scripts are available upon request.

Genomic RNA sequences. All available genomic sequences for the genus *Betacoronavirus* were obtained from NCBI (http://www.ncbi.nlm.nih.gov/) on 6 April 2020 using customized scripts based on Entrez Programming Utilities (E-utilities; https://eutils.ncbi.nlm.nih.gov/entrez/eutils/). For SARS-CoV-2, nucleotide accessions were redownloaded on 13 May 2020. For HIV-1, 100 random full-length sequences were obtained from NCBI to provide a representative sample of HIV-1 variation on 19 February 2021. Only accessions with at least 95% of the reference genome length were retained. For each species, the reference accession describing a complete genome was identified (Table 1).

Genomic and amino acid variation. For each species, nucleotide and amino acid variation analyses were conducted either on the entire genome or the spike S protein. Both were estimated in a 50-nt window. Nucleotide substitutions on the genome were measured based on nucleotide diversity (48) and genomic variation (40). Nucleotide diversity was calculated using TASSEL (https://www.maizegenetics.net/tassel) (78). Amino acid substitutions were measured based on SAPs (40). SNPs or SAPs were identified and mapped using SNP-sites version 2.4.1 (https://github.com/sanger-pathogens/snp-sites) (79) and VCFtools (80). The average and 99% confidence interval ($P < 0.01$) were estimated and plotted for each species. For variation per ORF, only ORFs present in at least 25% of the β-CoVs were counted.

Annotated phylogram for the S protein. Phylograms (40) were made using GraPhlAn (http://segatalab.cibio.unitn.it/tools/graphlan/) to illustrate the geographical location, host, and variation in the S protein (81).

Disorder of the S protein. Order/disorder was estimated using the Multilayered Fusion-based Disorder predictor (MFDp) with a false-positive rate of 5% (82). For each species, the amino acid sequence of the reference accession was used. Ordered and disordered areas are below and above the 0.5 threshold, respectively.

Annotated phylogram of U.S. species. All U.S. January sequences from the 23 March 2020 download were included in the early (January) time period. Three random sequences were chosen from each state with accessions from the late (July-August) time period to ensure a representative sample. Neighbor-joining phylogenetic trees were created using MAFFT version 7.4 (https://mafft.cbrc.jp/alignment/software/) with bootstrap values of 100. Accessions were aligned and mutations were identified using Geneious version 8.0 (https://www.geneious.com).

Data availability. All accession numbers used in this study were downloaded from GenBank (Table 1).

ACKNOWLEDGMENTS

This research was supported by NIH grant R01GM120108 to H.G.-R. and by the Nebraska Agricultural Experiment Station with funding from the Hatch Act (accession number 1007272) through the USDA National Institute of Food and Agriculture. The same grant provided open access costs.

We thank the Holland Computing Center (https://hcc.unl.edu/) for technical assistance and Jim Van Etten, David D. Dunnigan, Tom Petro, and Qingsheng Li for critical readings of the manuscript.

H.G.-R. conceived the study. K.L. performed the analysis. K.L., N.M.H., R.R.P., R.A.L., and H.G.-R. made the figures. H.G.-R. and K.L. wrote the paper.

REFERENCES

1. Cui J, Li F, Shi ZL. 2019. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol 17:181–192. https://doi.org/10.1038/s41579-018-0118-9.
2. 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, China Novel Coronavirus Investigative Research Team. 2020. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 382:727–733. https://doi.org/10.1056/NEJMoa2001017.
3. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, Ma X, Zhan F, Wang L, Hu T, Zhou H, Hu Z, Zhou W, Zhao L, Chen J, Meng Y, Wang J, Lin Y, Yuan J, Xie Z, Ma J, Liu WJ, Wang D, Xu W, Holmes EC, Gao GF, Wu G, Chen W, Shi W, Tan W. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395:565–574. https://doi.org/10.1016/S0140-6736(20)30251-8.
4. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270–273. https://doi.org/10.1038/s41586-020-2012-7.
5. Brian DA, Baric RS. 2005. Coronavirus genome structure and replication. Curr Top Microbiol Immunol 287:1–30. https://doi.org/10.1007/3-540-26765-4_1.
6. Neuman BW, Kiss G, Kunding AH, Bhella D, Baksh MF, Connelly S, Droese B, Klaus JP, Makino S, Sawicki SG, Siddell SG, Stamou DG, Wilson IA, Kuhn P, Buchmeier MJ. 2011. A structural analysis of M protein in coronavirus assembly and morphology. J Struct Biol 174:11–22. https://doi.org/10.1016/j.jsb.2010.11.021.
7. Li F. 2016. Structure, function, and evolution of coronavirus spike proteins. Annu Rev Virol 3:237–261. https://doi.org/10.1146/annurev-virology-110615-042301.
8. Cai Y, Zhang J, Xiao T, Peng H, Sterling SM, Walsh RM, Jr, Rawson S, Rits-Volloch S, Chen B. 2020. Distinct conformational states of SARS-CoV-2 spike protein. Science 369:1586–1592. https://doi.org/10.1126/science.abd4251.
9. Gallagher TM, Buchmeier MJ. 2001. Coronavirus spike proteins in viral entry and pathogenesis. Virology 279:371–374. https://doi.org/10.1006/viro.2000.0757.
10. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. 2020. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science 367:1444–1448. https://doi.org/10.1126/science.abb2762.
11. Li F. 2013. Receptor recognition and cross-species infections of SARS coronavirus. Antiviral Res 100:246–254. https://doi.org/10.1016/j.antiviral.2013.08.014.
12. Cantuti-Castelvetri L, Ojha R, Pedro LD, Djannatian M, Franz J, Kuivanen S, van der Meer F, Kallio K, Kaya T, Anastasina M, Smura T, Levanov L, Szirovicza L, Tobi A, Kallio-Kokko H, Osterlund P, Joensuu M, Meunier FA, Butcher SJ, Winkler MS, Mollenhauer B, Helenius A, Gokce O, Teesalu T, Hepojoki J, Vapalahti O, Stadelmann C, Balistreri G, Simons M. 2020. Neuropilin-1 facilitates SARS-CoV-2 cell entry and infectivity. Science 370:856–860. https://doi.org/10.1126/science.abd2985.
13. Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG, Decroly E. 2020. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Res 176:104742. https://doi.org/10.1016/j.antiviral.2020.104742.
14. Xia S, Liu M, Wang C, Xu W, Lan Q, Feng S, Qi F, Bao L, Du L, Liu S, Qin C, Sun F, Shi Z, Zhu Y, Jiang S, Lu L. 2020. Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. Cell Res 30:343–355. https://doi.org/10.1038/s41422-020-0305-x.
15. Millet JK, Whittaker GR. 2015. Host cell proteases: critical determinants of coronavirus tropism and pathogenesis. Virus Res 202:120–134. https://doi.org/10.1016/j.virusres.2014.11.021.
16. Mercatelli D, Giorgi FM. 2020. Geographic and genomic distribution of SARS-CoV-2 mutations. Front Microbiol 11:1800. https://doi.org/10.3389/fmicb.2020.01800.
17. Sanjuan R, Domingo-Calap P. 2016. Mechanisms of viral mutation. Cell Mol Life Sci 73:4433–4448. https://doi.org/10.1007/s00018-016-2299-6.
18. Lauring AS, Andino R. 2010. Quasispecies theory and the behavior of RNA viruses. PLoS Pathog 6:e1001005. https://doi.org/10.1371/journal.ppat.1001005.
19. Domingo E, Perales C. 2019. Viral quasispecies. PLoS Genet 15:e1008271. https://doi.org/10.1371/journal.pgen.1008271.
20. McCarthy KR, Rennick LJ, Nambulli S, Robinson-McCarthy LR, Bain WG, Haidar G, Duprex WP. 2021. Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. Science 371:1139–1142. https://doi.org/10.1126/science.abf6950.
21. CDC. 2021. Emerging SARS-CoV-2 variants. Centers for Disease Control and Prevention, Atlanta, GA.
22. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of

pathogen evolution. Bioinformatics 34:4121–4123. https://doi.org/10.1093/bioinformatics/bty407.
23. Fernandes JD, Hinrichs AS, Clawson H, Gonzalez JN, Lee BT, Nassar LR, Raney BJ, Rosenbloom KR, Nerli S, Rao AA, Schmelter D, Fyfe A, Maulding N, Zweig AS, Lowe TM, Ares M, Jr, Corbet-Detig R, Kent WJ, Haussler D, Haeussler M. 2020. The UCSC SARS-CoV-2 genome browser. Nat Genet 52:991–998. https://doi.org/10.1038/s41588-020-0700-8.
24. Hodcroft EB, Domman DB, Snyder DJ, Oguntuyo K, Van Diest M, Densmore KH, Schwalm KC, Femling J, Carroll JL, Scott RS, Whyte MM, Edwards MD, Hull NC, Kevil CG, Vanchiere JA, Lee B, Dinwiddie DL, Cooper VS, Kamil JP. 2021. Emergence in late 2020 of multiple lineages of SARS-CoV-2 spike protein variants affecting amino acid position 677. medRxiv https://doi.org/10.1101/2021.02.12.21251658:2021.02.12.21251658.
25. Tillett RL, Sevinsky JR, Hartley PD, Kerwin H, Crawford N, Gorzalski A, Laverdure C, Verma SC, Rossetto CC, Jackson D, Farrell MJ, Van Hooser S, Pandori M. 2021. Genomic evidence for reinfection with SARS-CoV-2: a case study. Lancet Infect Dis 21:52–58. https://doi.org/10.1016/S1473-3099(20)30764-7.
26. Jary A, Leducq V, Malet I, Marot S, Klement-Frutos E, Teyssou E, Soulie C, Abdi B, Wirden M, Pourcher V, Caumes E, Calvez V, Burrel S, Marcelin AG, Boutolleau D. 2020. Evolution of viral quasispecies during SARS-CoV-2 infection. Clin Microbiol Infect 26:1560. https://doi.org/10.1016/j.cmi.2020.07.032.
27. Vasques Nonaka CK, Miranda FM, Gräf T, Almeida MA, Santana DAR, Giovanetti M, Solano de Freitas Souza B. 27 January 2021. Genomic evidence of a Sars-Cov-2 reinfection case with E484K spike mutation in Brazil. Preprints https://doi.org/10.20944/preprints202101.0132.v1.
28. Zhu Z, Chakraborti S, He Y, Roberts A, Sheahan T, Xiao X, Hensley LE, Prabakaran P, Rockx B, Sidorov IA, Corti D, Vogel L, Feng Y, Kim JO, Wang LF, Baric R, Lanzavecchia A, Curtis KM, Nabel GJ, Subbarao K, Jiang S, Dimitrov DS. 2007. Potent cross-reactive neutralization of SARS coronavirus isolates by human monoclonal antibodies. Proc Natl Acad Sci U S A 104:12123–12128. https://doi.org/10.1073/pnas.0701000104.
29. Li CK, Wu H, Yan H, Ma S, Wang L, Zhang M, Tang X, Temperton NJ, Weiss RA, Brenchley JM, Douek DC, Mongkolsapaya J, Tran BH, Lin CL, Screaton GR, Hou JL, McMichael AJ, Xu XN. 2008. T cell responses to whole SARS coronavirus in humans. J Immunol 181:5490–5500. https://doi.org/10.4049/jimmunol.181.8.5490.
30. Baum A, Fulton BO, Wloga E, Copin R, Pascal KE, Russo V, Giordano S, Lanza K, Negron N, Ni M, Wei Y, Atwal GS, Murphy AJ, Stahl N, Yancopoulos GD, Kyratsous CA. 2020. Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. Science 369:1014–1018. https://doi.org/10.1126/science.abd0831.
31. Pinto D, Park YJ, Beltramello M, Walls AC, Tortorici MA, Bianchi S, Jaconi S, Culap K, Zatta F, De Marco A, Peter A, Guarino B, Spreafico R, Cameroni E, Case JB, Chen RE, Havenar-Daughton C, Snell G, Telenti A, Virgin HW, Lanzavecchia A, Diamond MS, Fink K, Veesler D, Corti D. 2020. Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. Nature 583:290–295. https://doi.org/10.1038/s41586-020-2349-y.
32. Brochot E, Demey B, Touze A, Belouzard S, Dubuisson J, Schmit JL, Duverlie G, Francois C, Castelain S, Helle F. 2020. Anti-spike, anti-nucleocapsid and neutralizing antibodies in SARS-CoV-2 inpatients and asymptomatic individuals. Front Microbiol 11:584251. https://doi.org/10.3389/fmicb.2020.584251.
33. Volz E, Hill V, McCrone JT, Price A, Jorgensen D, O'Toole A, Southgate J, Johnson R, Jackson B, Nascimento FF, Rey SM, Nicholls SM, Colquhoun RM, da Silva Filipe A, Shepherd J, Pascall DJ, Shah R, Jesudason N, Li K, Jarrett R, Pacchiarini N, Bull M, Geidelberg L, Siveroni I, Consortium C-U, Goodfellow I, Loman NJ, Pybus OG, Robertson DL, Thomson EC, Rambaut A, Connor TR. 2021. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. Cell 184:64–75. https://doi.org/10.1016/j.cell.2020.11.020.
34. 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 PY. 2021. Spike mutation D614G alters SARS-CoV-2 fitness. Nature 592:116–121. https://doi.org/10.1038/s41586-020-2895-3.
35. 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, Sheffield C-GG, 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. https://doi.org/10.1016/j.cell.2020.06.043.
36. Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, Lorenzi JC, Muecksch F, Rutkowska M, Hoffmann HH, Michailidis E, Gaebler C, Agudelo M, Cho A, Wang Z, Gazumyan A, Cipolla M, Luchsinger L, Hillyer CD, Caskey M, Robbiani DF, Rice CM, Nussenzweig MC, Hatziioannou T, Bieniasz PD. 2020. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. Elife 9:e61312. https://doi.org/10.7554/eLife.61312.
37. Duffy S. 2018. Why are RNA virus mutation rates so damn high? PLoS Biol 16:e3000003. https://doi.org/10.1371/journal.pbio.3000003.
38. Obenauer JC, Denson J, Mehta PK, Su X, Mukatira S, Finkelstein DB, Xu X, Wang J, Ma J, Fan Y, Rakestraw KM, Webster RG, Hoffmann E, Krauss S, Zheng J, Zhang Z, Naeve CW. 2006. Large-scale sequence analysis of avian influenza isolates. Science 311:1576–1580. https://doi.org/10.1126/science.1121586.
39. Smith EC, Sexton NR, Denison MR. 2014. Thinking outside the triangle: replication fidelity of the largest RNA viruses. Annu Rev Virol 1:111–132. https://doi.org/10.1146/annurev-virology-031413-085507.
40. Nigam D, LaTourrette K, Souza PFN, Garcia-Ruiz H. 2019. Genome-wide variation in potyviruses. Front Plant Sci 10:1439. https://doi.org/10.3389/fpls.2019.01439.
41. Phan T. 2020. Genetic diversity and evolution of SARS-CoV-2. Infect Genet Evol 81:104260. https://doi.org/10.1016/j.meegid.2020.104260.
42. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CCS, Boshier FAT, Ortiz AT, Balloux F. 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol 83:104351. https://doi.org/10.1016/j.meegid.2020.104351.
43. Forster P, Forster L, Renfrew C, Forster M. 2020. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci U S A 117:9241–9243. https://doi.org/10.1073/pnas.2004999117.
44. Charon J, Barra A, Walter J, Millot P, Hebrard E, Moury B, Michon T. 2018. First experimental assessment of protein intrinsic disorder involvement in an RNA virus natural adaptive process. Mol Biol Evol 35:38–49. https://doi.org/10.1093/molbev/msx249.
45. Becerra-Flores M, Cardozo T. 2020. SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. Int J Clin Pract 74:e13525. https://doi.org/10.1111/ijcp.13525.
46. Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, Liu Y, Weinstock GM, Wheeler DA, Gibbs RA, Yu F. 2010. A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res 20:273–280. https://doi.org/10.1101/gr.096388.109.
47. Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R. 2010. Viral mutation rates. J Virol 84:9733–9748. https://doi.org/10.1128/JVI.00694-10.
48. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497. https://doi.org/10.1093/bioinformatics/btg359.
49. Fitzsimmons WJ, Woods RJ, McCrone JT, Woodman A, Arnold JJ, Yennawar M, Evans R, Cameron CE, Lauring AS. 2018. A speed-fidelity trade-off determines the mutation rate and virulence of an RNA virus. PLoS Biol 16:e2006459. https://doi.org/10.1371/journal.pbio.2006459.
50. Chan JF, Kok KH, Zhu Z, Chu H, To KK, Yuan S, Yuen KY. 2020. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg Microbes Infect 9:221–236. https://doi.org/10.1080/22221751.2020.1719902.
51. Jauregui AR, Savalia D, Lowry VK, Farrell CM, Wathelet MG. 2013. Identification of residues of SARS-CoV nsp1 that differentially affect inhibition of gene expression and antiviral signaling. PLoS One 8:e62416. https://doi.org/10.1371/journal.pone.0062416.
52. Cornillez-Ty CT, Liao L, Yates JR, Kuhn P, Buchmeier MJ. 2009. Severe acute respiratory syndrome coronavirus nonstructural protein 2 interacts with a host protein complex involved in mitochondrial biogenesis and intracellular signaling. J Virol 83:10314. https://doi.org/10.1128/JVI.00842-09.
53. Báez-Santos YM, St John SE, Mesecar AD. 2015. The SARS-coronavirus papain-like protease: structure, function and inhibition by designed antiviral compounds. Antiviral Res 115:21–38. https://doi.org/10.1016/j.antiviral.2014.12.015.
54. Li W, Wong SK, Li F, Kuhn JH, Huang IC, Choe H, Farzan M. 2006. Animal origins of the severe acute respiratory syndrome coronavirus: insight from ACE2-S-protein interactions. J Virol 80:4211–4219. https://doi.org/10.1128/JVI.80.9.4211-4219.2006.
55. Wan Y, Shang J, Graham R, Baric RS, Li F. 2020. Receptor recognition by the novel coronavirus from Wuhan: an Analysis based on decade-long structural studies of SARS coronavirus. J Virol 94:e00127-20. https://doi.org/10.1128/JVI.00127-20.
56. Li W, Zhang C, Sui J, Kuhn JH, Moore MJ, Luo S, Wong SK, Huang IC, Xu K, Vasilieva N, Murakami A, He Y, Marasco WA, Guan Y, Choe H, Farzan M.

2005. Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. EMBO J 24:1634–1643. https://doi.org/10.1038/sj.emboj.7600640.
57. Damas J, Hughes GM, Keough KC, Painter CA, Persky NS, Corbo M, Hiller M, Koepfli KP, Pfenning AR, Zhao H, Genereux DP, Swofford R, Pollard KS, Ryder OA, Nweeia MT, Lindblad-Toh K, Teeling EC, Karlsson EK, Lewin HA. 2020. Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates. Proc Natl Acad Sci U S A 117:22311–22322. https://doi.org/10.1073/pnas.2010146117.
58. Edwards CE, Yount BL, Graham RL, Leist SR, Hou YJ, Dinnon KH, III, Sims AC, Swanstrom J, Gully K, Scobey TD, Cooley MR, Currie CG, Randell SH, Baric RS. 2020. Swine acute diarrhea syndrome coronavirus replication in primary human cells reveals potential susceptibility to infection. Proc Natl Acad Sci U S A 117:26915–26925. https://doi.org/10.1073/pnas.2001046117.
59. Hebrard E, Bessin Y, Michon T, Longhi S, Uversky VN, Delalande F, Van Dorsselaer A, Romero P, Walter J, Declerck N, Fargette D. 2009. Intrinsic disorder in viral proteins genome-linked: experimental and predictive analyses. Virol J 6:23. https://doi.org/10.1186/1743-422X-6-23.
60. Rantalainen KI, Eskelin K, Tompa P, Makinen K. 2011. Structural flexibility allows the functional diversity of potyvirus genome-linked protein VPg. J Virol 85:2449–2457. https://doi.org/10.1128/JVI.02051-10.
61. Elena SF, Fraile A, Garcia-Arenal F. 2014. Evolution and emergence of plant viruses. Adv Virus Res 88:161–191. https://doi.org/10.1016/B978-0-12-800098-4.00003-9.
62. Woodham AW, Skeate JG, Sanna AM, Taylor JR, Da Silva DM, Cannon PM, Kast WM. 2016. Human immunodeficiency virus immune cell receptors, coreceptors, and cofactors: implications for prevention and treatment. AIDS Patient Care STDS 30:291–306. https://doi.org/10.1089/apc.2016.0100.
63. Kuo L, Godeke GJ, Raamsman MJ, Masters PS, Rottier PJ. 2000. Retargeting of coronavirus by substitution of the spike glycoprotein ectodomain: crossing the host cell species barrier. J Virol 74:1393–1406. https://doi.org/10.1128/JVI.74.3.1393-1406.2000.
64. Long SW, Olsen RJ, Christensen PA, Bernard DW, Davis JJ, Shukla M, Nguyen M, Saavedra MO, Yerramilli P, Pruitt L, Subedi S, Kuo HC, Hendrickson H, Eskandari G, Nguyen HAT, Long JH, Kumaraswami M, Goike J, Boutz D, Gollihar J, McLellan JS, Chou CW, Javanmardi K, Finkelstein IJ, Musser JM. 2020. Molecular architecture of early dissemination and massive second wave of the SARS-CoV-2 virus in a major metropolitan area. mBio 11:e02707-20. https://doi.org/10.1128/mBio.02707-20.
65. Taboada B, Vazquez-Perez JA, Munoz-Medina JE, Ramos-Cervantes P, Escalera-Zamudio M, Boukadida C, Sanchez-Flores A, Isa P, Mendieta-Condado E, Martinez-Orozco JA, Becerril-Vargas E, Salas-Hernandez J, Grande R, Gonzalez-Torres C, Gaytan-Cervantes FJ, Vazquez G, Pulido F, Araiza-Rodriguez A, Garces-Ayala F, Gonzalez-Bonilla CR, Grajales-Muniz C, Borja-Aburto VH, Barrera-Badillo G, Lopez S, Hernandez-Rivas L, Perez-Padilla R, Lopez-Martinez I, Avila-Rios S, Ruiz-Palacios G, Ramirez-Gonzalez JE, Arias CF. 2020. Genomic analysis of early SARS-CoV-2 variants introduced in Mexico. J Virol 94:e01056-20. https://doi.org/10.1128/JVI.01056-20.
66. Su YCF, Anderson DE, Young BE, Linster M, Zhu F, Jayakumar J, Zhuang Y, Kalimuddin S, Low JGH, Tan CW, Chia WN, Mak TM, Octavia S, Chavatte JM, Lee RTC, Pada S, Tan SY, Sun L, Yan GZ, Maurer-Stroh S, Mendenhall IH, Leo YS, Lye DC, Wang LF, Smith GJD. 2020. Discovery and genomic characterization of a 382-nucleotide deletion in ORF7b and ORF8 during the early evolution of SARS-CoV-2. mBio 11:e01610-20. https://doi.org/10.1128/mBio.01610-20.
67. Nguyen A, David JK, Maden SK, Wood MA, Weeder BR, Nellore A, Thompson RF. 2020. Human leukocyte antigen susceptibility map for severe acute respiratory syndrome coronavirus 2. J Virol 94:e00510-20. https://doi.org/10.1128/JVI.00510-20.
68. Holland LA, Kaelin EA, Maqsood R, Estifanos B, Wu LI, Varsani A, Halden RU, Hogue BG, Scotch M, Lim ES. 2020. An 81-nucleotide deletion in SARS-CoV-2 ORF7a identified from sentinel surveillance in Arizona (January to March 2020). J Virol 94:e00711-20. https://doi.org/10.1128/JVI.00711-20.
69. Konno Y, Kimura I, Uriu K, Fukushi M, Irie T, Koyanagi Y, Sauter D, Gifford RJ, Consortium U-C, Nakagawa S, Sato K. 2020. SARS-CoV-2 ORF3b is a potent interferon antagonist whose activity is increased by a naturally occurring elongation variant. Cell Rep 32:108185. https://doi.org/10.1016/j.celrep.2020.108185.
70. Taylor JK, Coleman CM, Postel S, Sisk JM, Bernbaum JG, Venkataraman T, Sundberg EJ, Frieman MB. 2015. Severe acute respiratory syndrome coronavirus ORF7a inhibits bone marrow stromal antigen 2 virion tethering through a novel mechanism of glycosylation interference. J Virol 89:11820–11833. https://doi.org/10.1128/JVI.02274-15.
71. Gao Y, Yan L, Huang Y, Liu F, Zhao Y, Cao L, Wang T, Sun Q, Ming Z, Zhang L, Ge J, Zheng L, Zhang Y, Wang H, Zhu Y, Zhu C, Hu T, Hua T, Zhang B, Yang X, Li J, Yang H, Liu Z, Xu W, Guddat LW, Wang Q, Lou Z, Rao Z. 2020. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. Science 368:779–782. https://doi.org/10.1126/science.abb7498.
72. Subissi L, Posthuma CC, Collet A, Zevenhoven-Dobbe JC, Gorbalenya AE, Decroly E, Snijder EJ, Canard B, Imbert I. 2014. One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities. Proc Natl Acad Sci U S A 111:E3900–E3909. https://doi.org/10.1073/pnas.1323705111.
73. Li JY, Liao CH, Wang Q, Tan YJ, Luo R, Qiu Y, Ge XY. 2020. The ORF6, ORF8 and nucleocapsid proteins of SARS-CoV-2 inhibit type I interferon signaling pathway. Virus Res 286:198074. https://doi.org/10.1016/j.virusres.2020.198074.
74. Muth D, Corman VM, Roth H, Binger T, Dijkman R, Gottula LT, Gloza-Rausch F, Balboni A, Battilani M, Rihtaric D, Toplak I, Ameneiros RS, Pfeifer A, Thiel V, Drexler JF, Muller MA, Drosten C. 2018. Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. Sci Rep 8:15177. https://doi.org/10.1038/s41598-018-33487-8.
75. Yuan M, Wu NC, Zhu X, Lee CD, So RTY, Lv H, Mok CKP, Wilson IA. 2020. A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. Science 368:630–633. https://doi.org/10.1126/science.abb7269.
76. Wrapp D, De Vlieger D, Corbett KS, Torres GM, Wang N, Van Breedam W, Roose K, van Schie L, Team V-C-R, Hoffmann M, Pohlmann S, Graham BS, Callewaert N, Schepens B, Saelens X, McLellan JS. 2020. Structural basis for potent neutralization of betacoronaviruses by single-domain camelid antibodies. Cell 181:1004–1015. https://doi.org/10.1016/j.cell.2020.04.031.
77. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, Graham BS, McLellan JS. 2020. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 367:1260–1263. https://doi.org/10.1126/science.abb2507.
78. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635. https://doi.org/10.1093/bioinformatics/btm308.
79. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom 2:e000056. https://doi.org/10.1099/mgen.0.000056.
80. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, Genomes Project Analysis Group. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158. https://doi.org/10.1093/bioinformatics/btr330.
81. Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. 2015. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ 3:e1029. https://doi.org/10.7717/peerj.1029.
82. Mizianty MJ, Stach W, Chen K, Kedarisetti KD, Disfani FM, Kurgan L. 2010. Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics 26:i489–i496. https://doi.org/10.1093/bioinformatics/btq373.