



Review

Zhaohui Qian*, Pei Li, Xiaolu Tang and Jian Lu*

Evolutionary dynamics of the severe acute respiratory syndrome coronavirus 2 genomes

<https://doi.org/10.1515/mr-2021-0035>

Received December 19, 2021; accepted January 23, 2022;
published online March 1, 2022

Abstract: The coronavirus disease 2019 (COVID-19) pandemic has caused immense losses in human lives and the global economy and posed significant challenges for global public health. As severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of COVID-19, has evolved, thousands of single nucleotide variants (SNVs) have been identified across the viral genome. The roles of individual SNVs in the zoonotic origin, evolution, and transmission of SARS-CoV-2 have become the focus of many studies. This review summarizes recent comparative genomic analyses of SARS-CoV-2 and related coronaviruses (SC2r-CoVs) found in non-human animals, including delineation of SARS-CoV-2 lineages based on characteristic SNVs. We also discuss the current understanding of receptor-binding domain (RBD) evolution and characteristic mutations in variants of concern (VOCs) of SARS-CoV-2, as well as possible co-evolution between RBD and its receptor, angiotensin-converting enzyme 2 (ACE2). We propose that the interplay between SARS-CoV-2 and host RNA editing mechanisms might have partially resulted in the bias in nucleotide changes during SARS-CoV-2 evolution. Finally, we outline some current challenges, including difficulty in deciphering the complicated relationship between viral pathogenicity and infectivity of different variants, and monitoring transmission of SARS-CoV-2 between humans and animals as the pandemic progresses.

Keywords: molecular evolution; population genetics; severe acute respiratory syndrome coronavirus 2; variant; virus.

The structure, composition, replication and transcription of the SARS-CoV-2 genome

The coronavirus disease 2019 (COVID-19) pandemic has caused immense losses in human lives and the global economy and posed significant challenges for global public health. As of December 14th, 2021, there were over 270 million confirmed cases and 5.32 million reported deaths [1]. The SARS-CoV-2, a newly identified sarbecovirus in the genus Betacoronavirus (β -CoVs or Beta-CoVs), is the causative agent of COVID-19 pandemic [2–5].

SARS-CoV-2 is a single-stranded, positive-sense RNA virus. The genome of SARS-CoV-2 is approximately 29.9 kb with a cap structure at the 5' end and a poly-A tail at the 3' end [5], similar to host cellular mRNA; and it has 13–15 open reading frames (ORFs) flanked by 5' and 3' untranslated regions (UTRs) [6, 7], which contain cis-elements essential for RNA synthesis (Figure 1A). ORF1ab occupies two-thirds of the viral genome and is synthesized as a single poly-protein (Figure 1A) and then cleaved into 16 nonstructural proteins (nsps) by viral proteases encoded in nsp3 and nsp5. Most nsps are essential for the formation of the viral replication and transcription complex (RTC) [8]. The remaining one-third of the viral genome encodes four viral structural proteins: the spike protein (S), an envelope protein (E), membrane protein (M), and nucleoprotein (N), and several viral accessory proteins, including ORF3a, ORF3b, ORF6, ORF7a, ORF7b, ORF8, ORF9b, and ORF10 [6, 7]. The S protein is essential for binding its entry receptor, angiotensin-converting enzyme 2 (ACE2), and contains two subunits, S1 and S2, separated by a furin cleavage site (Figure 1B). S1 can be further divided into two domains, N-terminal domain (NTD) and receptor-binding domain (RBD). After binding to their receptor, S proteins may mediate membrane fusion either at the cell plasma membrane directly or at lysosomal

***Corresponding authors: Zhaohui Qian**, NHC Key Laboratory of Systems Biology of Pathogens, Institute of Pathogen Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100176, China, E-mail: zqian2013@sina.com; and **Jian Lu**, State Key Laboratory of Protein and Plant Gene Research, Center for Bioinformatics, School of Life Sciences, Peking University, Beijing, 100871, China. E-mail: LUJ@pku.edu.cn. <https://orcid.org/0000-0002-4409-1667> (J. Lu)

Pei Li, NHC Key Laboratory of Systems Biology of Pathogens, Institute of Pathogen Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100871, China

Xiaolu Tang, State Key Laboratory of Protein and Plant Gene Research, Center for Bioinformatics, School of Life Sciences, Peking University, Beijing, 100176, China

membranes after internalization through endocytosis, depending on the availability of appropriate host proteases [9, 10]. The E and M proteins are required for effective virus assembly and budding, and the N protein binds to the viral genome and forms a helical ribonucleocapsid (RNP) that is essential for virus assembly [11]. Viral accessory proteins are not required for virus replication in cell culture, but they are suspected of playing important roles in viral pathogenesis in the natural host. Not all ORFs listed here have been experimentally verified, and the exact number of accessory proteins encoded in the SARS-CoV-2 genome remains to be determined [6, 7, 12, 13]. In addition, the SARS-CoV-2 genome might encode other unknown ORFs involved in the regulation of viral replication or host immune responses [6].

Once SARS-CoV-2 enters a cell, viral RNA replication and transcription, which are controlled by RTC, begin. Like many other positive-sense RNA viruses, SARS-CoV-2 RNA synthesis likely occurs inside the endoplasmic reticulum (ER)-derived double-membrane vesicles (DMVs) [14]. DMVs may not only protect viral RNA replication intermediates from host cytosolic innate immune sensors but also provide a place with adequate concentrations of substrates required for RNA synthesis. Viral nsp3, nsp4, and nsp6 have been implicated in the formation of DMVs [14]. In the RTC, nsp12 serves as an RNA-dependent RNA polymerase (RdRp), catalyzing viral RNA synthesis with the help of two viral cofactors, nsp7 and nsp8 [15]. The nsp8 protein is a primase, and nsp7, nsp8, and nsp12 together form the core of the RTC (Figure 1C). The nsp9 protein forms a dimer and regulates the replication process. Viral nsp13 [16] and nsp14 [17] also play important roles in regulating viral RNA synthesis during elongation: nsp13 is a viral helicase [18], and nsp14 provides 3′–5′ exonuclease activity with a proofreading function [19]. Nsp13 and nsp14 also contribute to the 5′ capping of viral RNAs [18, 20]. The coronavirus capping machinery includes nsp10, nsp13, nsp14, and nsp16. Nsp13 provides RNA 5′-triphosphatase activity [18], nsp14 has N7-methyltransferase activity [20], nsp16 is a 2′-O-methyltransferase [21], and nsp10 acts as a cofactor for nsp14 and nsp16 [22, 23].

During viral replication, the positive-sense viral genome is used as the template to synthesize full-length negative-sense genome copies (Figure 1D). In return, negative-sense genomes serve as the templates for the generation of progeny viral RNA genomes, which can be translated to produce more nsps and RTCs. Viral structural proteins and accessory proteins are generated from individual viral subgenomic RNAs (sgRNAs), which result from a unique discontinuous transcription process during negative-strand synthesis, a hallmark feature of CoV replication and transcription [8, 24]. There are transcription regulatory sequences (TRS) located

upstream of most ORFs in the coronavirus genome [8]. The TRS adjacent to the leader sequence in the 5′ UTR of the viral genome is named “TRS-L”, whereas all other TRSs are called TRS-“body” or TRS-B. In the case of SARS-CoV-2, the TRS sequence is “ACGAAC” [5–7]. During negative-strand RNA synthesis, the RTC likely pauses on specific sequences containing TRS-B and reinitiates synthesis at TRS-L (Figure 1D). The nascent negative-sense sgRNAs are then used as templates to synthesize positive-sense sgRNAs for the expression of structural and accessory proteins. The discontinuous transcription process likely involves interactions between complementary TRSs of the nascent negative-strand RNA (negative-sense TRS-B) and the positive-strand genomic RNA (positive-sense TRS-L) [8]. The exact molecular mechanism underlying discontinuous transcription remains elusive.

Comparative genomics of SARS-CoV-2 and related CoVs

Identification of SARS-CoV-2-related CoVs

Great efforts have been undertaken worldwide to trace the origin of SARS-CoV-2, but it remains elusive when and where SARS-CoV-2 originated. The current consensus is that it is extremely unlikely that a lab leak was the source of the pandemic virus [25]. Instead, many studies have supported the view that SARS-CoV-2 had a zoonotic origin and evolved in nature [26–29]. Because the place of virus origin is usually different from the place of the first recognized outbreak [30] and investigating the origin of a virus can take tremendous time and effort [31, 32], further studies are needed to better understand the origin of SARS-CoV-2 [33].

Despite the zoonotic signatures observed in the SARS-CoV-2 genome, it remains unclear how this virus was transmitted from animals to human populations [28]. Nevertheless, recent studies have identified various CoVs in bats closely related to SARS-CoV-2 (termed SC2r-CoVs), including RaTG13 from *Rhinolophus affinis* [3], BANAL-20-236 from *R. marshalli* [34], BANAL-20-52, BANAL-20-116, BANAL-20-247, and RmYN02 from *Rhinolophus malayanus* [34, 35], Rc-o319 from *Rhinolophus cornutus* [36], RshSTT182 and RshSTT200 from *Rhinolophus shameli* [37], RacCs203 from *Rhinolophus acuminatus* [38], and BANAL-20-103 and RpYN06 from *Rhinolophus pusillus* [34, 39]. Bats are common natural hosts for CoVs [40–43], supporting that SARS-CoV-2 likely had a bat origin.

As shown previously [38, 44], the currently known CoVs in the sarbecovirus lineage of the β -CoV genus can

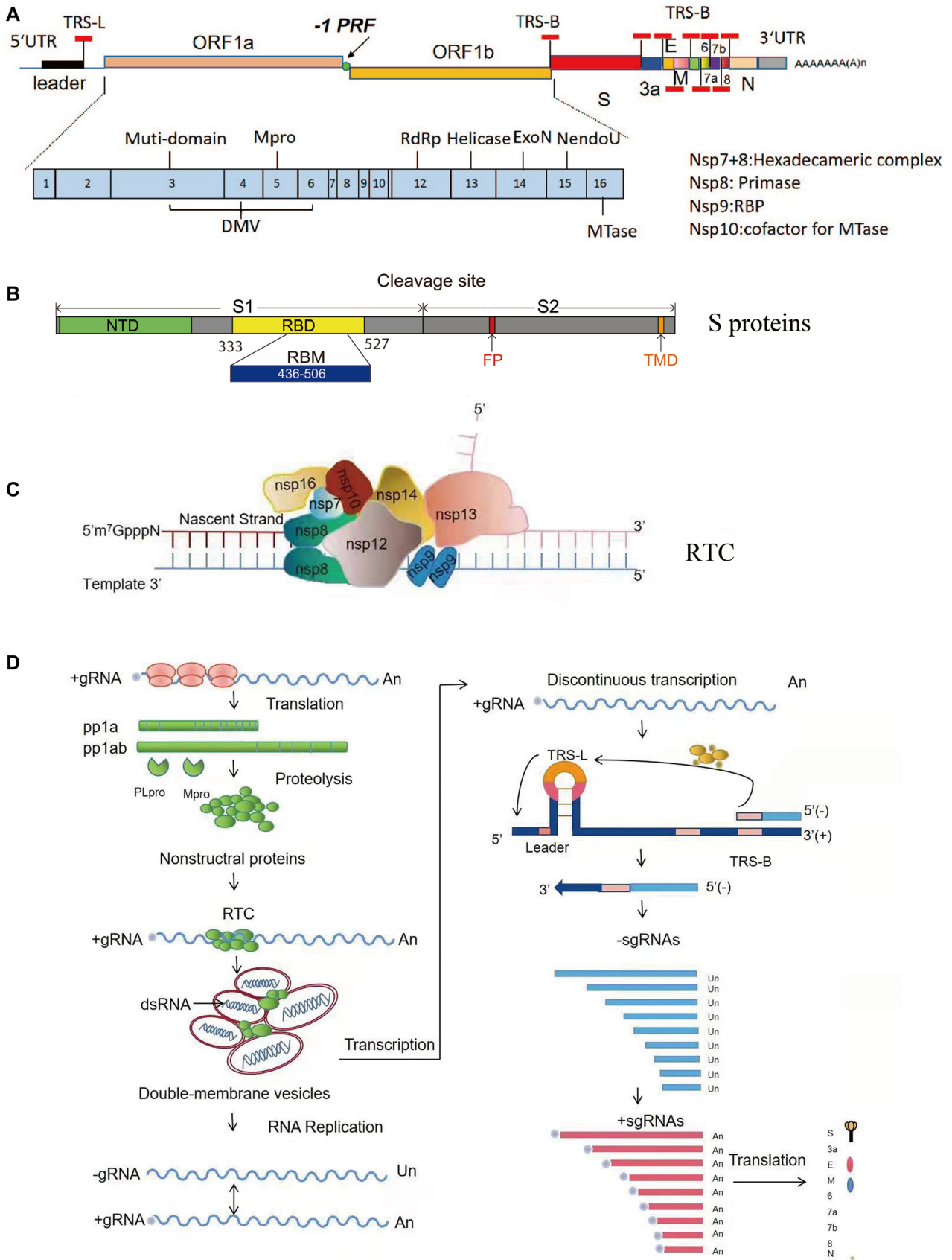


Figure 1: The scheme of genome organization and replication/transcription of SARS-CoV-2.

be categorized into two major clades, with one clade clustering with SARS-CoV-2 (termed SC2r-CoVs) and the other clade grouping with SARS-CoV (SC1r-CoVs) (Figure 2A). Notably, several bat CoVs, including RaTG13, BANAL-20-52, BANAL-20-103, and BANAL-20-236, share ~96% nucleotide sequence identity with SARS-CoV-2 [3, 34]. Divergence of SARS-CoV-2 and RaTG13 was inferred to have occurred roughly 50 years ago [45, 46], although the exact divergence time depends on the substitution rate assumed in various analyses. Of note, a recent large-scale survey of CoV in bats in China failed to find any SC2r-CoV sequences, indicating that active circulation of SC2r-CoV might be limited in China [47–52].

Besides bats, SC2r-CoVs have also been detected in Malayan pangolins (*Manis javanica*) [53–56]. Pangolin-derived SC2r-CoVs (pangolin-CoVs) can be further classified into two sublineages, pangolin-CoV-GDC [53–55] and pangolin-CoV-GXC, which were found during anti-smuggling operations by Guangdong and Guangxi customs, respectively. While pangolin-CoV-GDC and SARS-CoV-2 share genomic sequence similarity of 92.4% [53–55], pangolin-CoV-GXC and SARS-CoV-2 show 85.5% nucleotide identity [56]. Both sublineages of pangolin-CoVs were isolated from *M. javanica*. However, a recent study identified an SC2r-CoV, MP20, from *M. pentadactyla* that likely originated from Southeast Asia [57]. MP20 is closely related to pangolin-CoV-GXC. An intra-host variant analysis revealed that the genetic diversity of pangolin-CoVs was substantially higher than expected, suggesting that pangolins might be the natural hosts of SC2r-CoVs [57]. The S proteins of viruses in both pangolin-CoV sublineages can bind hACE2 [44, 58–60],

pointing to the potential risk of zoonotic transmission of pathogenic SC2r-CoVs from pangolins to humans.

Roles of natural selection in SARS-CoV-2 and SC2r-CoVs divergence

The substitution rate at synonymous substitution sites in protein-coding regions (nucleotide changes that do not alter protein sequences) is routinely used as a proxy for the rate of neutral evolution. The synonymous substitution rate of SARS-CoV has been estimated to range from 1.67 to 4.67×10^{-3} substitutions/site/year [61], whereas the evolutionary rate of MERS-CoV was estimated at about 1.12×10^{-3} substitutions/site/year [62]. In comparison, the substitution rate in SARS-CoV-2 was roughly on the order of 10^{-3} substitutions/site/year [63–68], although the exact rate has varied slightly across studies. These results indicate that SARS-CoV-2 has an overall evolutionary rate similar to SARS-CoV and MERS-CoV. Nevertheless, as observed in other CoVs [69], the evolutionary rate is heterogeneous across SARS-CoV-2 genes. For example, in comparing SARS-CoV-2 with SC2r-CoVs, we found considerable differences in synonymous substitution rates across genes, with the S gene showing a much higher evolutionary rate than other genes [70].

Comparative genomics has revealed how natural selection shaped the genome-wide differences between SARS-CoV-2 and SC2r-CoVs. Comparison between dN (nonsynonymous substitutions per nonsynonymous site) vs. dS (synonymous substitutions per synonymous site) values in coding regions provides a measure of the selective pressure on protein evolution, with a dN/dS (ω)

A. The genome organization of SARS-CoV-2. The schematic diagram of the complete SARS-CoV-2 genome is shown at the top. ORF1a and ORF1b encode polyproteins pp1a and pp1ab, respectively, which are further processed into 10 and 16 nsps by viral proteases. The expression of ORF1b is regulated by a ribosomal frameshifting mechanism. –1PRF: –1 programmed ribosome frameshifting element. The leader and body copies of transcription-regulating sequences (TRS-L and TRS-B, respectively) are indicated by short thick red lines. The functions of important nsps are indicated in the scheme. **UTR**, untranslated regions; **Mpro**, main protease; **RBP**, RNA binding protein; **RdRp**, RNA-dependent RNA polymerase; **S**, spike protein; **E**, envelop protein; **M**, membrane protein; **N**, nucleocapsid protein. B. **The schematic diagram of S protein of the SARS-CoV-2 S protein.** **NTD**, N-terminal domain; **RBD**, receptor-binding domain; **FP**, fusion peptide; **TMD**, transmembrane domain; **cleavage site**, furin cleavage site. C. A hypothetical model of the replicase and transcriptase complex of SARS-CoV-2. The diagram above shows how the replication-related proteins form an RTC. Nascent RNA is synthesized at the nsp12 RdRp domain. The nsp7 and nsp8 form the primase complex, nsp9 is a single-stranded binding protein and forms a dimer in the complex. Formation of the 5' cap is catalyzed by nsp13, nsp14, nsp10, and nsp16. The locations of these proteins in the model are based on structural and functional analyses. D. A schematic of SARS-CoV-2 replication and transcription. Viral RNA replicates in the cytoplasm. ORF1a and ORF1ab are translated from the genomic RNA to produce pp1a and pp1ab polyproteins, which are then cleaved by viral papain-like protease (PLpro) and Mpro. Nsp 3, 4, and 6 are responsible for remodeling cellular membranes to form double-membrane vesicles (DMVs) where viral replication and transcription occur. The positive-sense genome is used as the template to produce full-length (–) RNA copies, which are used as templates for making full-length (+) RNA genomes. Negative-stranded sub-genomic RNAs (–sgRNAs) are synthesized through a unique discontinuous transcription mechanism in which fusion and transfer of a leader RNA sequence to body RNAs occur at transcription-regulating sequences (TRSs) with the help of the viral N protein and host proteins. The –sgRNAs serve as templates for sub-genomic RNAs (+sgRNAs) that are capped and polyA-tailed, and, despite many ORFs, only the closest ORF is typically translated.

sites, analysis of sequence differences without separating these two classes of sites may underestimate the extent of molecular divergence several-fold. For instance, between SARS-CoV-2 and RaTG13, nucleotides genome-wide differ by -3.8% ; however, the average dN is 0.78% , and dS is 16.8% ($\omega=0.0465$), which means that, on average, 95.35% of nonsynonymous mutations that change protein sequences were removed by natural selection as SARS-CoV-2 and RaTG13 diverged. Although phylogenetic reconstruction using concatenated protein sequences indicates that some CoVs collected from bats in North Laos (BANAL-20-103 and BANAL-20-236) are more distantly related to SARS-CoV-2 than RaTG13 [34], the dS values from a comparison of SARS-CoV-2 and these two CoVs (0.1524 , and 0.1577 for BANAL-20-103 and 236, respectively) tend to be slightly lower than those from a comparison of SARS-CoV-2 and RaTG13 (0.1682). Teasing apart the effect of natural selection can yield a better understanding of the phylogenetic relationships of CoVs (Table 1).

Although the purifying selection is the predominant force governing the evolution of SARS-CoV-2 and SC2r-CoVs, signals of positive selection were also detected in nonsynonymous sites. By carrying out a CODEML analysis, we previously identified 10 nonsynonymous sites that showed strong signals of positive selection during the evolution of SARS-CoV-2 and other SC2r-CoVs [70]. Interestingly, five of these putative positively selected sites are located in the S protein (sites 46, 183, 439, 483, and 493), and three of them are located in the RBD of the S protein (439, 483 and 493). Using a similar analysis, Damas et al. identified three putative positively selected sites (455, 483, and 494) [73], and Cagliani et al. [72] found strong evidence of positive selection at seven sites, including six in the S protein (483, 484, 486, 490, 493, and 494). Sites 493 and 494 of the S protein were inferred to be positively selected in two of these studies, and site 483 was inferred to be positively selected in all three studies. Some of these inferences might have led to false positives, as the assumptions of CODEML were violated

in analysis [74]. Therefore, functional studies are needed to investigate the consequences of these amino acid changes.

Evolution of RBD and possible co-evolution with ACE2

Deletions and possible recombination in the RBD

Compared with the phylogenetic tree based on protein alignments of all the conserved genes (Figure 2A), a considerably different tree was obtained when only S gene sequences were used for phylogenetic reconstruction, as some CoVs in the SC2r-CoV clade (e.g. RmYN02 and RacCS203) grouped with viruses in the SC1r-CoV clade (e.g. Rf1 and HeB2013) (Figure 2B). This discrepancy might result from differences in genealogies of the S gene from those of other parts of the genome, as evidence of recombination is commonly observed in CoVs [5, 75, 76]. This pattern is manifested in the RBM (sites 436–506 of the S protein) of the RBD (Figure 2C).

There are two deletions in the RBMs of the RBD: deletions 1 (sites 445–449) and 2 (473–486). These deletions commonly coexist in coronavirus lineages such as RmYN02 and RacCS203. Previous analyses demonstrated that deletions in this sequence abolish the capacity of RBDs of RmYN02 and RacCS203 to bind to hACE2 [38, 44]. However, deletion 2 seems to be more important, because S proteins with deletion of sites 445–449 alone, such as that in RsYN04, retain some to bind to hACE2 [39]. The deletions are interspersed in SC1r-CoVs and SC2r-CoVs; intriguingly, however, viruses with both deletions also have highly similar sequences flanking these deletions (Figure 2C), suggesting that these deletions might have one single origin rather than multiple independent origins. Recombination may have shaped this discontinuity in the distribution of deletions within the phylogeny.

Amino acid changes in RBDs of bat SC2r-CoVs

There are at least 17 amino acid residues in the RBD of the SARS-CoV-2 S protein that interact with hACE2 [51, 52] (Figure 2C). Eight (Y449, Y453, N487, Y489, G496, T500, G502, and Y505) of these 17 residues are conserved between S proteins of SARS-CoV and SARS-CoV-2, and 11 of the 17 residues (K417, G446, Y453, L455, F456, A475, N487, Y489,

Table 1: The molecular divergence between SARS-CoV-2 and SC2r-CoVs.

	dN	dS	dN/dS
Bat RaTG13	0.0078	0.1682	0.0465
Bat BANAL-20-52	0.0059	0.1322	0.0447
Bat BANAL-20-103	0.013	0.1524	0.0855
Bat BANAL-20-116	0.0347	0.1967	0.1762
Bat BANAL-20-236	0.0133	0.1577	0.0841
Bat BANAL-20-247	0.0352	0.1988	0.177

G496, T500, and G502) are identical between the S proteins of SARS-CoV-2 and RaTG13 (Figure 2C). Nevertheless, the S protein of RaTG13 can still use hACE2 as its entry receptor, although the entry efficiency is lower than that of SARS-CoV-2 [77–79]. These results highlight the plasticity of RBD/hACE2 interactions. Recently, CoVs isolated from bats in North Laos were found to show very high homologies with SARS-CoV-2 [34]. The S proteins of BANAL-20-52 and BANAL-20-236 share amino acid sequence identities of over 98.4 and 90.6%, respectively, with that of SARS-CoV-2. Among the 17 ACE2-contacting residues in the S protein, there is only H498 in BANAL-20-52 and K493 and H498 in BANAL-20-236 that differ from those residues in SARS-CoV-2, indicating that both S proteins likely use hACE2 as the entry receptor. Interestingly, although deletion 1 is found in S proteins of bat RSHSTT182 and RSHSTT200 CoVs found in Cambodia, several critical ACE2 contact residues are preserved, including Q493, Q498, N501, and Y505 [37]. While the S protein of RSHSTT200 CoVs failed to bind to hACE2, it could bind *R. shameli* bat ACE2 for virus entry [37]. Similar to bat RSHSTT182 and RSHSTT200 CoVs, the S protein of RaTG15, a coronavirus isolated from *R. affinis*, has a short deletion 1 (Figure 2C). However, there are 10 residues (sites 417, 449, 475, 486, 487, 493, 498, 500, 501, and 502) and one deletion (site 446) in the 17 ACE2-contacting sites that differ between SARS-CoV-2 and RaTG15. Experimental results reveal that these differences, possibly combined with the effect of deletion 1, might be responsible for the loss of binding affinity to hACE2 [44]. In summary, there are substantial differences in the critical functional sites in RBDs across bat CoVs, but only some changes may be associated with differences in the entry of human cells.

Amino acid changes in RBDs of pangolin-CoVs

Neither of the two currently known pangolin-CoV sublineages has any deletions in the RBM (Figure 2C). Although the pangolin-CoVs are more distantly related to SARS-CoV-2 than RaTG13, previous studies have revealed almost identical amino acid sequences in the RBD region between pangolin-GDC-CoV and SARS-CoV-2 [55, 80]. As shown in Figure 2C, only two (417 and 498) out of the 17 ACE2-contacting residues differ between the RBDs of pangolin-GDC-CoV and SARS-CoV-2. It seems likely that the identical residues in SARS-CoV-2 and pangolin-GDC-CoV resulted from convergent evolution or recombination [56, 70, 80]. Notably, 12 of these 17 residues (G446, Y449, Y453, L455, F456, A475, N487, Y489, G496, T500, G502, and Y505) are identical between pangolin-GXC-CoV

and SARS-CoV-2 (Lam et al. [56]). Despite differences in key functional residues between the two sublineages of pangolin-CoVs (Figure 2C), the RBDs of both pangolin-CoV sublineages bind efficiently to hACE2 [44, 81]. In addition, the RBDs of pangolin-CoVs seem to indicate a broader host range than those of SARS-CoV-2 [81]. Of note, S protein residue 498 differs across SARS-CoV-2 (Q), RaTG13 (Y), pangolin-CoV (H), BANAL-20-52 (H), and BANAL-20-236 (H), and introducing a Q498H substitution in the SARS-CoV-2 RBD expands its binding capacity to ACE2 of mice, rats, and European hedgehogs [28, 81].

Evolution of ACE2 in animals and possible co-evolution with RBD

In addition to differences in the RBD region across SC2r-CoVs, sequence changes in the ACE2 receptor can influence RBD-ACE2 binding affinity. There are about 20 residues in ACE2 that interact with viral S proteins. Comparative genomic analyses revealed multiple amino acid changes in ACE2 across animals that putatively affect binding of the SARS-CoV-2 RBD to ACE2 [73, 82]. Bat *Rhinolophus macrotis* ACE2 (bACE2-Rm) exhibits a substantially lower affinity to the RBD of SARS-CoV-2 than hACE2 does [83]. A detailed analysis has revealed that residues 41 and 42 in bACE2-Rm play important roles in interactions of the receptor with SARS-CoV-2 RBD, with the Y41-Q42 combination yielding a high binding affinity and the H41-E42 combination resulting in a much weaker binding affinity [83]. While the S proteins of both SARS-CoV-2 and RaTG13 can bind to ACE2 of bat *R. affinis* (RaACE2), the binding affinity of SARS-CoV-2 RBD to RaACE2 is much weaker than that to hACE2 [28]. The RBD of RaTG15 shows clear discrepancies in binding to ACE2s from different species; it can bind to ACE2 of both *R. affinis* and Malayan pangolins, but fails to bind hACE2 [44]. More comprehensive studies are needed to dissect the complicated interactions between various RBDs of SC2r-CoVs and different ACE2 homologs, as well as to understand the mechanism of possible co-evolution between SC2r-CoVs and the animal host receptors.

Mutational bias in the genomes of SARS-CoV-2 and SC2r-CoVs

Besides replication errors, several confounding factors such as host antiviral proteins and spontaneous chemical reactions can lead to mutations in the genomes of RNA viruses [69]. RNA editing enzymes, including adenosine

deaminases acting on RNAs (ADARs) and apolipoprotein B mRNA editing enzyme, catalytic polypeptides (APOBECs), play important roles in the innate immune restriction to counter SARS-CoV-2 infection by inducing point mutations in SARS-CoV-2 genomes [6, 84, 85].

Several studies have shown excessive A-to-G and C-to-U mutations occurred during SARS-CoV-2 evolution [86–91]. For example, one study examined the nucleotide substitution matrix from the most recent common ancestor of SARS-CoV-2 and RaTG13 to that of SARS-CoV-2 and RaTG13, and observed a strong nucleotide substitution bias at synonymous sites [14]. Specifically, A-to-G, C-to-U, U-to-C, and G-to-A substitutions were the most abundant when the

ancestral nucleotide was A, C, U, and G, respectively (Figure 3A). It was proposed that the overabundance of C-to-U transitions in the SARS-CoV-2 genomes can be caused by the activity of APOBEC cytosine deaminases [92]. Here, we propose a model of possible RNA editing-induced mutational bias in SARS-CoV-2 evolution. Under this model, A-to-I editing events catalyzed by ADARs in the sense or antisense strand of SARS-CoV-2 cause A-to-G or U-to-C mutations; and C-to-U editing catalyzed by APOBECs in the sense or antisense strand of SARS-CoV-2 cause C-to-U or G-to-A mutations (Figure 3B). However, other mechanisms that might lead to similar observations could not be excluded.

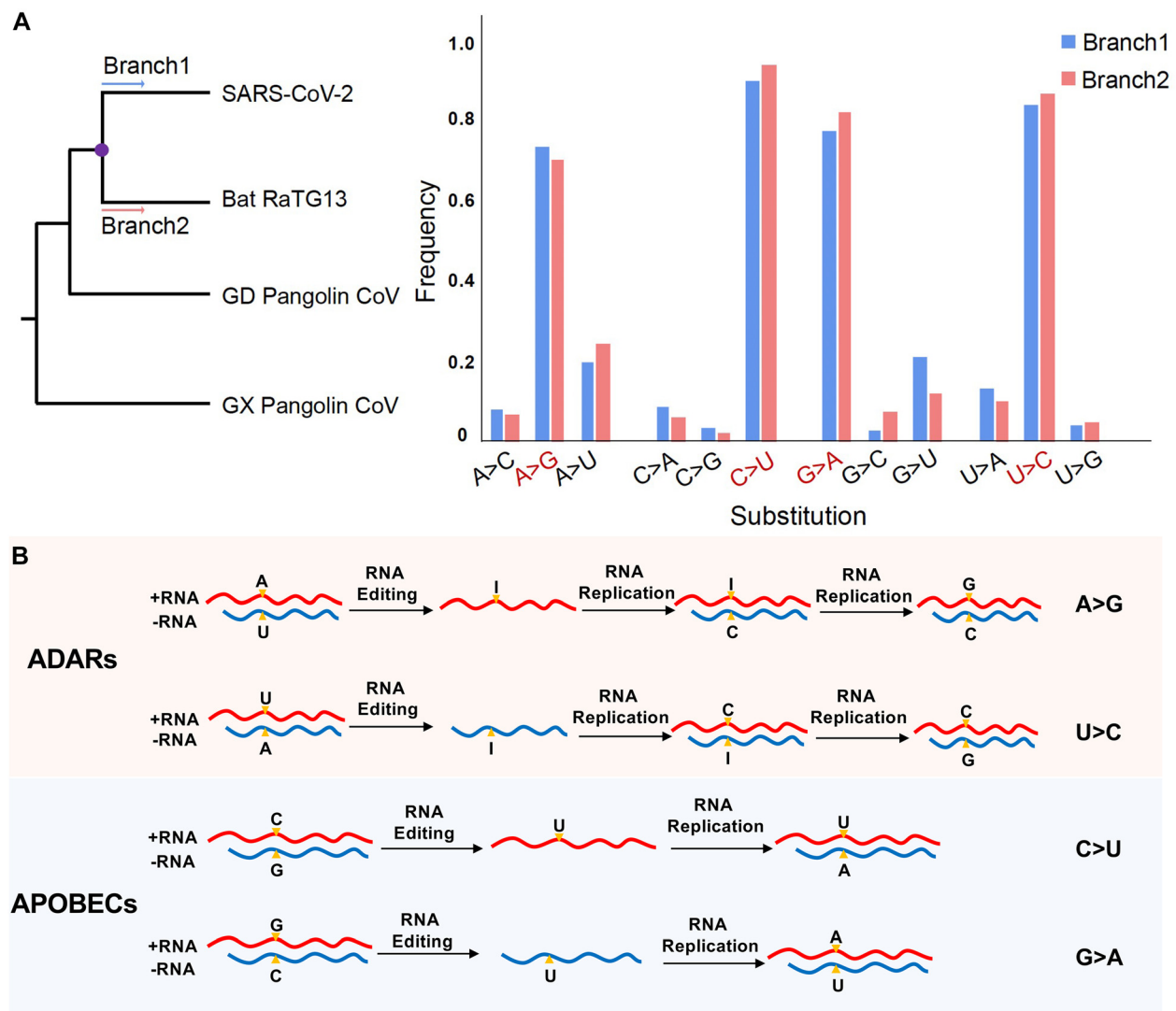


Figure 3: Model of possible RNA editing-induced mutational bias in SARS-CoV-2 evolution. A. Nucleotide substitution frequencies at synonymous sites in branches from the most recent common ancestor of SARS-CoV-2 and RaTG13 (the purple point in the phylogenetic tree on the left) to SARS-CoV-2 (B1) and RaTG13 (B2). B. A-to-I editing events catalyzed by ADARs in the sense or antisense strand of SARS-CoV-2 cause A-to-G or U-to-C mutations (upper panel); C-to-U editing catalyzed by APOBECs in the sense or antisense strand of SARS-CoV-2 cause C-to-U or G-to-A mutations (lower panel).

SARS-CoV-2 lineage analysis and the continuing evolution

Although SARS-CoV-2 has a proofreading mechanism, mutations remain inevitable during the replication of RNA viruses. By carrying out experimental evolution experiments with two circulating SARS-CoV-2 strains, a recent study estimates a genomic mutation rate of $2.9\text{--}3.7 \times 10^{-6}$ mutation/site/cycle for SARS-CoV-2 under cell culture condition [93], which yields roughly 0.1 mutations/genome/cycle. With millions of SARS-CoV-2 genome sequences deposited in databases, including the Global Initiative on Sharing All Influenza Data (GISAID; <https://www.epicov.org>) [94, 95] and National Genomic Data Center of China (<https://ngdc.cncb.ac.cn/>) databases, hundreds to thousands of single nucleotide variants (SNVs) have been identified [70, 96–100]. The roles of individual SNVs in zoonotic origin, evolution, and transmission of SARS-CoV-2 have become the focus of many studies [25, 64, 101–104]. Based on 103 available SARS-CoV-2 genomes, we found that SARS-CoV-2 could be divided into two major lineages, L and S, early in the COVID-19 pandemic [70]. The distinction between L and S lineages depends on two SNV pairs at sites 8,782 and 28,144 with nearly complete linkage: C8782/U28144 for L and U8782/C28144 for S, with the reference genome (NC_045512) belonging to the L lineage. Residue 8,782 is encoded in the nsp4 gene. The C-to-U change at position 8,782 has no effect on the resulting amino acid, whereas residue 28,144 is encoded in the accessory protein ORF8 and the U-to-C substitution at position 28,144 leads to a codon switch from leucine (L) to serine (S). The “L” and “S” lineages are named because of leucine and serine residues, respectively, at position 28,144. Of note, the S lineage is considered to be ancestral to the L lineage when a tree is rooted by bat and pangolin CoVs as the outgroup [99]. Using Forster’s nomenclature, SARS-CoV-2 variants are classified into three lineages: A, B, and C. “A” lineage is equivalent to our “S” lineage, “L” lineage is further divided into “B” and “C” lineages. Moreover, based on these two sites and other SNVs, GISAID (<http://gisaid.org>) divides SARS-CoV-2 genomes into four major groups (S, L, V, and G). In contrast, Nextstrain (<https://nextstrain.org>) [105] categorized SARS-CoV-2 variants into five major clades (19A, 19B, 20A, 20B, and 20C). Finally, in the popular Pango nomenclature of SARS-CoV-2 (<https://cov-lineages.org/index.html>), classification of “A” and “B” lineages is also based on SNVs at positions 8,782 and 28,144, with “A” equating to “S” and “B” equating to “L”. Despite the substantial expansion of the number of viral genomes analyzed, the distinction between SARS-CoV-2L and S lineages remains

robust. For instance, among the 127,119 high-quality SARS-CoV-2 genomes we previously analyzed, 120,958 (95.15%) belonged to the L lineage, 5,950 (4.68%) belonged to the S lineage, and only 211 (0.17%) could not be accurately assigned to either the L or S lineage [106].

Given the rapid accumulation of publicly available SARS-CoV-2 genome sequences, analyzing the relatedness of SARS-CoV-2 genomes using traditional phylogenetic methods is a significant challenge. The power of a phylogenetic analysis can also be limited for tracing genealogies when ancestral, and descendent sequences are pooled [107, 108]. Further, because viruses often evolve through multifurcation, especially when superspreaders play a role in transmission [109], the hierarchical bifurcating assumption in the traditional phylogenetic inference may be violated. As an alternative, we have proposed determination of the lineage of a SARS-CoV-2 genome combined with haplotype network analysis to trace genealogies [106]. Specifically, based on L/S delineation according to variants at sites 8,782 and 28,144, we further divided the L lineage into two major sublineages (L1 and L2) using three tightly linked variants at sites 3,037, 14,408, and 23,403, and further categorized SARS-CoV-2 strains into 130 sublineages with SNVs at 201 additional sites. Our lineage designation system is hierarchical and can be easily expanded with new variants that might arise and become prevalent. In Figure 4, we incorporated the characteristic mutations in representative variants of concern (VOCs) and variants of interest (VOIs) and updated the haplotype network of SARS-CoV-2 sublineages based on the previous L/S nomenclature system [106].

Important SARS-CoV-2 variants and their biological, immunological, and clinical characteristics

Important variants of concern (VOCs) and variants of interest (VOIs) and their biological, immunological, and transmissibility

Amino acid changes in the S protein affect virus infectivity and host immune responses against the virus [110]. For instance, an N234Q change promotes resistance to neutralizing antibodies, whereas an N165Q change makes the virus more sensitive. A D614G mutation in the S protein (A23403G) and C3037U and C14408U greatly enhances virus infectivity and transmission. Based on these mutations, L1 and L2 sublineages can be defined [106]. The L1

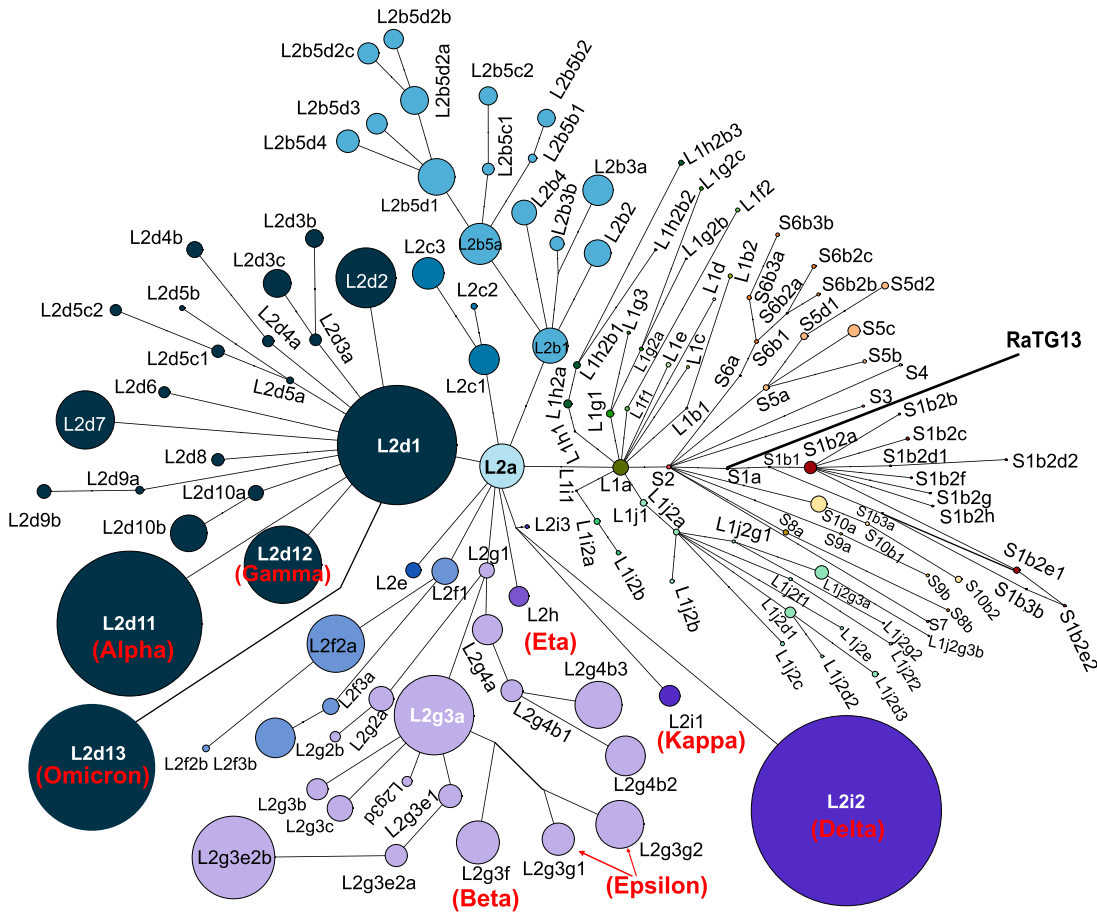


Figure 4: Updated haplotype network of SARS-CoV-2 sublineages based on the L/S nomenclature system [106]. Except for L2d11, L2i2, and L2d13 the size of each dot indicating a sublineage was scaled to the number of genomes in that sublineage (Sizes of L2d11, L2i2, and L2d13 have been reduced for better visualization). Some representative variants of concern (VOCs) and variants of interest (VOIs) are labeled in red in the network (Alpha: L2d11; Beta: L2g3f; Gamma: L2d12; Delta: L2i2; Omicron: L2d13; Epsilon: L2g3g1 + L2g3g2; Eta: L2h; Kappa: L2i1).

sublineage carries the ancestral D614 variant, whereas the L2 sublineage has the G614 mutation. As shown in Figure 5B, SARS-CoV-2 sublineages exhibit substantial differences in temporal distributions; currently, the L2 sublineage is dominant worldwide.

Population genetic analyses have revealed that the D614G change is driven by positive selection [111–113]. Precisely, selective coefficients of G614 over D614 were estimated to be 0.31–0.55 [112] and 0.06–0.56 [113] when considering SARS-CoV-2 strains sampled worldwide. Strains found in the initial outbreak mainly carry D614, whereas G614 strains became dominant during the pandemic [106, 111, 112]. Although residue 614 itself is not located on the surface of the RBM, the G614 S protein seems to adopt a more open conformation that allows the S protein to bind to hACE2 more efficiently [114]. Numerous experiments have shown D614G enhancement of viral replication in human lung epithelial cells and primary human airway tissues, leading to high virus titers in the

upper respiratory tract and greater transmissibility, that result from the greater infectivity and virion stability conferred by this change [115–121]. However, the open conformation of the S protein may also increase susceptibility to antibody neutralization [122]. Similar to D614G, some emerging variants alter receptor binding affinity, reduce antibody neutralization activity, and affect the T cell response, potentially impacting COVID-19 diagnosis, treatment, and vaccine effectiveness globally.

To mitigate the potential impacts of some important variants and communicate differences between variants more easily to the public, WHO developed a classification system that defines two classes of SARS-CoV-2 variants, VOIs and VOCs. A VOI is defined as a SARS-CoV-2 variant with specific genetic markers that are known or predicted to be associated with an increase in virus transmissibility, disease severity, or immune escape, or a decrease in the efficacy of treatments and diagnostic assays. A VOC is defined as a variant that shows evidence of an increase in transmissibility,

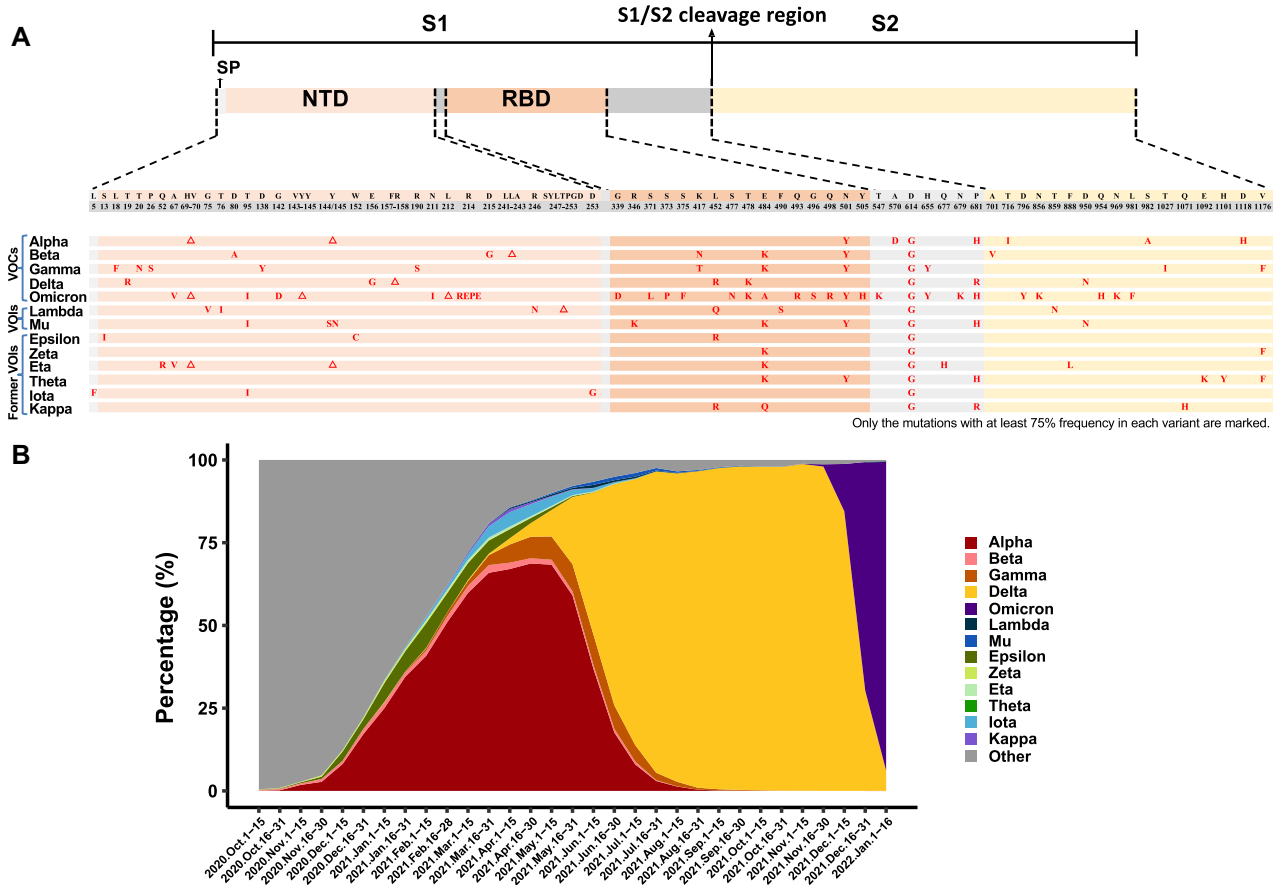


Figure 5: Prevalence and characteristic mutations in the S protein of VOCs and VOIs. A. Characteristic mutations (relative to the reference genome NC_045512) in the S proteins of VOCs and VOIs. In each variant lineage, the mutations that have a frequency of $\geq 75\%$ in the sequences of that lineage are shown in red (Data were taken from Outbreak.info, last accessed on 20 January 2022). Δ , deletion; A, alanine; R, arginine; N, asparagine; D, aspartic acid; C, cysteine; E, glutamic acid; Q, glutamine; G, glycine; H, histidine; I, isoleucine; L, leucine; K, lysine; F, phenylalanine; P, proline; S, serine; T, threonine; Y, tyrosine; V, valine. Note that a 3-amino-acid insertion (EPE) occurred after R214 of the Omicron Spike protein. B. Prevalence of VOCs and VOIs over time. SARS-CoV-2 genomes with collection date information in the GISAID database (6,977,884 in total, deposited between October 1st, 2020 and January 16th, 2022) were used in the analysis. The number of genomes was updated at two-week intervals.

disease severity, or immune escape. Both VOIs and VOCs are named with Greek alphabet letters, and the list is periodically adjusted as the pandemic progresses. As of January 18th, 2022, there are five VOCs, Alpha, Beta, Gamma, Delta and Omicron; two VOIs, Lambda and Mu; and six formerly circulating VOIs, Epsilon, Zeta, Eta, Theta, Iota, and Kappa. Figure 5A shows characteristic mutations in the S protein that define VOCs. In Figure 5B, we present the bi-weekly worldwide prevalence of these variants. In the following, we will briefly summarize the current understanding of the five VOC lineages, with a focus on the S protein.

The Alpha (B.1.1.7) variant has two deletions (sites 69–70 and site 144) and seven amino acid changes (N501Y, A570D, D614G, P681H, T716I, S982A, and D1118H) in the S protein. This lineage shows greater transmissibility than the SARS-CoV-2 variant circulating prior to its appearance

[113, 123]. The Alpha lineage showed a modest increase in resistance to neutralizing antibodies, but the E484K substitution in a small fraction of strains in the Alpha lineage ($\sim 0.3\%$) was found to facilitate immune escape [124–127].

The Beta (B.1.351) variant carries three mutations in the RBD (K417N, E484K, and N501Y), three in the N-terminal domain (D80A, D215G, and a deletion of sites 241–243), and one mutation in the S2 subunit (A701V). This variant was first identified in South Africa in October 2020, and it spreads rapidly in Africa due to a selective advantage putatively resulting from enhanced transmissibility [128] or immune escape [125, 129, 130].

The Gamma (P.1) variant has three mutations (K417N, E484K, and N501Y) in the RBD and nine other mutations (L18F, T20N, P26S, D138Y, R190S, D614G, H655Y, T1027I, and V1176F) in the S protein. The three mutations in the

RBD confer an increased binding affinity to hACE2 to strains in this lineage; strains in the gamma lineage may be 1.7- to 2.4-fold more transmissible than previously circulating non-Gamma strains [131]. In addition, virus strains in the gamma lineage have shown increased resistance to neutralizing antibodies [132, 133].

The Delta (B.1.617.2) lineage has one deletion (sites 157–158) and seven amino acid mutations in the S protein (T19R, E156G, L452R, T478K, D614G, P681R, and D950N). The infection rate with Delta strains is significantly higher than strains from other lineages. Delta is currently the most prevalent variant circulating worldwide. This variant is considerably less sensitive to serum neutralizing antibodies than pre-existing strains that only bear the D614G substitution [134]. The Delta lineage shows more efficient replication in airway organoid and human airway epithelial cells and spike-mediated entry than strains in the Alpha lineage [134]. The S protein of the Delta variant appears to mediate faster membrane fusion than other variants [135], resulting in a higher virus load and faster transmission rate [136]. The L452R mutation in the RBM confers increased infectivity and neutralizing antibody resistance to this variant [137–144]. Interestingly, the L452Q mutation in the Lambda variant might have similar effects as the L452R change in the Delta lineage [145–147]. Additionally, the P681R mutation in the Delta lineage may enhance furin cleavage of the spike protein into S1 and S2 subunits, facilitating more efficient cleavage by TMPRSS2 and increased virus infectivity [148].

The Omicron variant (B.1.1.529) was first detected in patients traveling from South Africa in November 2021, and has rapidly expanded globally. A substantial number of changes have occurred in the S protein of Omicron variant, including three deletions (sites 69–70, sites 143–145, and site 212), one insertion (insertion of EPE after site 214), and 26 point mutations (12 of them are located in the RBD: G339D, S371L, S373P, S375F, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, and Y505H). Despite the large number of mutations carried by the Omicron variant, it is not yet clear regarding the origin of this variant. It is possible that the Omicron variant might have evolved in human populations where the large-scale sequencing of SARS-CoV-2 genomes were not well carried out, or in immunocompromised people where many mutations in the SARS-CoV-2 genomes were allowed to accumulate, or resulted from cross-species transmission from animals that were infected with SARS-CoV-2 and accumulated adaptive mutations [149, 150]. Existing studies have shown that Omicron has a stronger immune evasion ability to neutralize antibodies than other mutant strains [151–154], but its pathogenicity might be significantly weakened [155–158].

In summary, the VOC lineages tend to be more infectious and have a greater capacity for immune escape.

Although numerous mutations have been found across the SARS-CoV-2 genome, changes in the S protein have received the most attention. These residues determine not only receptor usage and host range but also serve as major targets for host immune responses and, therefore, vaccines. Of note, as the pandemic developed, many mutations have occurred in the genomes of the VOCs which differentiated each VOC lineage into many sublineages. In Figures S1–5, we presented the characteristic mutations in the S protein of the sublineages and their prevalence in each of the five VOC lineages. Although new S protein variants have emerged and disappeared throughout the pandemic and most changes have had little to no impact on critical characteristics of the virus, some mutations such as D614G, L452R, E484, and N501Y significantly affect virus infectivity and/or sensitivity to neutralizing antibodies. In particular, the N501Y mutation in the S protein is present in four of the five VOC variants (Alpha, Beta, Gamma, and Omicron). N501 interacts with several residues in hACE2 [79], and the N501Y mutation increases RBD binding affinity to hACE2 [159, 160]. Further studies are required regarding whether the N501Y mutations in different VOCs were descendants from one single mutation event or resultant from multiple independent parallel mutations. E484K is present in Beta and Gamma lineage strains and a small fraction (0.3%) of Alpha strains. The E484K mutation confers immune escape [161, 162], and a combination of E484K and N501Y can further increase resistance to antibody neutralization [161–165]. In addition, many other mutations in S protein, such as N234Q, N165Q, L452R, A475V, V483A, F490L, and combinations of these, may also affect neutralization by antibodies [144, 166, 167].

Relationship between SARS-CoV-2 variants and pathogenicity

Despite the relatively well-understood relationships between a handful of variants and infectivity and immune escape of SARS-CoV-2 [110], how the variants affect pathogenicity and clinical manifestations of COVID-19 in patients is not yet well understood. Previously, among 271 patients (73 S- and 198 L-lineage patients) diagnosed with COVID-19 early during the COVID-19 outbreak in Wuhan, S-lineage patients exhibited significantly worse clinical outcomes than L-lineage patients, and this pattern held even after excluding other risk factors [168]. However, the underlying molecular mechanism, i.e. how changes at sites 8,782 and 28,144 affect the replication and transmission of SARS-CoV-2, is not yet clear. Although the D614G change in the S protein is associated with increased infectivity, D614 and

G614 SARS-CoV-2 variants do not appear to differ significantly in pathogenicity or clinical severity in patients [113] or in their pathogenicity in hamsters [169].

Mixed results were obtained regarding whether the COVID-19 patients infected with the Alpha variant [170–173] or Beta variant [128, 174] exhibit more severe disease than those infected with the prior SARS-CoV-2 strains, despite these two VOCs tending to show increased infectivity. Viral infection experiments of animal models yielded mixed results as well. Both the Alpha and Beta variants were shown to be 100-fold more lethal than the original SARS-CoV-2 bearing 614D in K18-hACE2 transgenic mice [175]. Similarly, in the hACE2-bearing mice, infection with the Beta variant resulted in more severe clinical symptoms and more weight loss than infection with prototype strain IME-BJ05 [176]. In contrast, very similar virus replication kinetics and amounts of virus shedding were observed in rhesus macaques infected with Alpha, Beta, and the variant containing only the D614G mutation; however, the Beta variant was found to be slightly less pathogenic than the other two variants in this host [177]. The Delta variant is more pathogenic than the prototypic SARS-CoV-2 strain in hamsters, and the P681R mutation in the Delta variant might be associated with this enhanced pathogenicity [178]. However, whether the Delta variant will induce more severe disease remains unclear. Altogether, it is challenging to decipher the relationship between SARS-CoV-2 variants and pathogenicity in COVID-19 patients, as multiple confounding factors such as age, gender, and underlying medical conditions affect symptoms and clinical severity of COVID-19 [111, 179–181]. Animal models have provided important insights into the phenotypical changes and pathogenesis of SARS-CoV-2 variants. However, outcomes in animal models affected by changes in ACE2 gene and protein sequences may not be recapitulated in humans.

Driving forces and possible trends in the evolution of SARS-CoV-2

Epistatic interactions between SARS-CoV-2 variants

One salient observation is that many genetic variants of SARS-CoV-2 are tightly linked together. For instance, by analyzing SNVs in 121,618 high-quality SARS-CoV-2 genomes, we identified 202 pairs of mutations (133 sites) that exhibit strong linkage [106]. Intriguingly, despite the tremendous accumulation of SARS-CoV-2 genome sequences to date, variants at sites 8,782 and 28,144 continue to show an

extremely high level of linkage, with C8782/U28144 variants specific to the L lineage and U8782/C28144 variants specific to the S lineage, and ~0.2% of genomes unable to be accurately assigned to the L or S lineage. In addition, within the L lineage, variants at sites 3,037, 14,408, and 23,404 are tightly linked, with C3037/C14408/A23403 variants belonging to the L1 lineage, and U3037/U14408/G23403 variants belonging to the L2 lineage, and <0.2% of L genomes belonging to neither the L1 nor L2 sublineage [106]. These observations support the notion of extensive epistasis and advantageous compensatory mutations between tightly linked variants, as illustrated in Figure 6. Nevertheless, the molecular mechanisms underlying these epistatic interactions in viral transmission or pathogenicity are largely unknown. For instance, in mouse-adapted SARS-CoV-2 strains, double mutations at N501Y/Q493H conferred higher binding affinity for hACE2 than N501Y alone; however, triple mutations at N501Y/Q493H/K417N substantially decreased binding affinity for hACE2 [182], highlighting the importance of epistasis between these variants. Overall, the impacts of individual variants and combined effects of tightly linked variants on the transmission and pathogenicity of SARS-CoV-2 require further studies.

Trends in SARS-CoV-2 evolution

One notable observation with SARS-CoV-2 is that one lineage tends to replace previously dominant lineages, with the Delta variant being predominant at present (e.g. Figure 5B). Nevertheless, the precise factors that shape this pattern are not clear. There are two competing theories on the evolution of virulence (i.e. pathogenicity) of a pathogen [183–186]. One long-standing view is that there is a trade-off between virulence and transmissibility, leading pathogens to evolve toward reduced virulence because weakening a host may reduce transmission. The alternative view is that a high level of virulence might be favored by natural selection if the more virulent strain compensates for the reduced transmission that can result from harming hosts. Continuous circulation of a SARS-CoV-2 variant is mainly driven by evolutionary forces that favor transmissibility and immune evasion rather than pathogenicity [177]. Further studies are needed to understand better the relationship between pathogenicity and infectivity caused by new SARS-CoV-2 variants that emerge.

Cross-species transmission between humans and animals

An early study suggested that residue 372 of the SARS-CoV-2 S protein plays a vital role in virus adaptation in humans. An

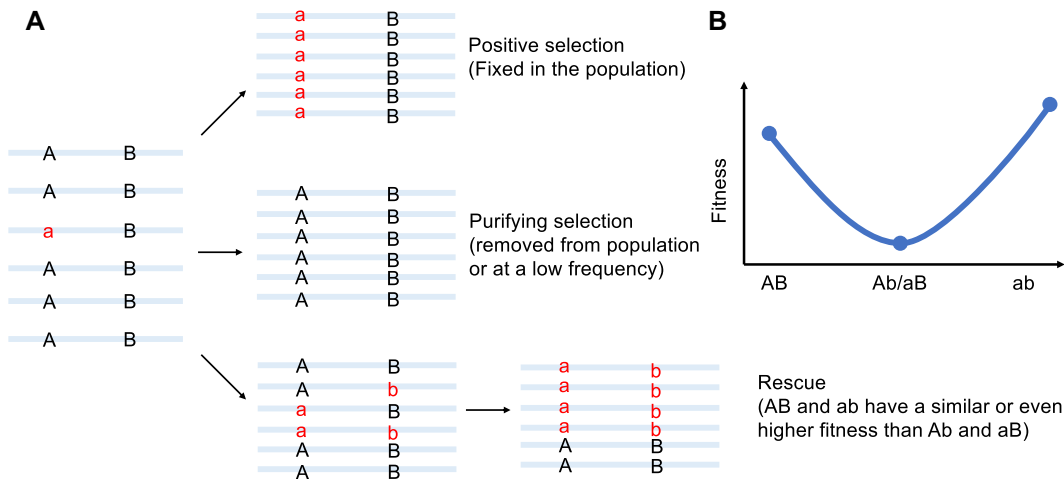


Figure 6: Schematic of epistatic interactions between two variants. A. A newly emerged variant ($A>a$) in the population might be harmful, neutral, or advantageous. Beneficial new variants are favored by natural selection and become fixed in the population very rapidly. In contrast, highly detrimental variants are removed by natural selection or persist in the population only at low frequencies. B. Under the epistatic interactions model, both $A>a$ and $B>b$ mutations are slightly deleterious. A virus with either an Ab or aB genotype has reduced fitness relative to the AB genotype. The virus with the ab genotype has a normal or even higher fitness. Thus, epistatic interactions can cause linkages between variants at the two sites to be maintained during viral evolution.

A372 (codon GCA at sites 22,676–22,678 of the genome) was found in nearly all 182,000 SARS-CoV-2 sequences surveyed [187], whereas T372 is found in the orthologous site (codon ACU or ACC) of all other SC2r-CoVs, indicating that the T372A change might have evolved during zoonotic transmission to humans. The T372A substitution not only abolishes N-linked glycosylation but increases affinity for hACE2, whereas A372 viruses replicate better than T372 viruses in human respiratory epithelial cells. Additionally, the T372A variant has shown evidence of positive selection upon examination of SARS-CoV-2 population data [187]. An N501Y substitution increases affinity of the S protein for hACE2, and viruses with Y at this site in the S protein show a broader host range, with mice susceptible to infection by N501Y strains [188]. In addition, mice can be infected with viruses carrying Q493K and Q498H changes in the S protein [189, 190].

Accumulating evidence indicates that SARS-CoV-2 can be transmitted from humans infected with COVID-19 to a wide range of mammals, including cats, minks, ferrets, lions, tigers, and white-tailed deer [191–193]. Notably, reverse zoonosis from humans to the animals can enable animals to become novel reservoirs for new SARS-CoV-2 variants that might be transmitted back to human populations with dramatic changes in infectivity or pathogenicity, as cross-species transmission can be accompanied by punctuated increases in mutations that may serve as raw materials for natural selection. For example, genome sequences prove that SARS-CoV-2 has been transmitted from humans to minks and then back to humans [194]. Thus, surveillance should be established

to monitor possible back-and-forth transmission of SARS-CoV-2 between humans and animals.

Concluding remarks and future perspectives

The COVID-19 pandemic has caused immense disruptions in the global economy and human health. Through comparative genomics, CoVs in animals that are closely related to SARS-CoV-2 has been identified. Evolutionary analysis of these CoV genomes has revealed strong signatures of positive and purifying selection as SARS-CoV-2, and SC2r-CoVs diverged. The interplay between SARS-CoV-2 and RNA editing mechanisms in hosts might have shaped the characteristic mutational bias in nucleotide changes during SARS-CoV-2 evolution. Genome sequencing has provided a powerful approach to identifying SARS-CoV-2 variants under putative positive selection, and defining lineages based on characteristic variants has facilitated studies of ongoing SARS-CoV-2 evolutionary dynamics. Although the molecular mechanisms and functional consequences of a few amino acid changes have been dissected, these changes have been mainly restricted to the S protein, with the causes and impacts of amino acid changes in other regions of SARS-CoV-2 genomes poorly understood. The epistatic relationship between amino acid changes and possible combined effects of such changes require further

exploration. One observation of SARS-CoV-2 variants is that one lineage replaces previously dominant lineages; however, the factors underlying these patterns are not yet clear. It is also a challenge to decipher the relationship between pathogenicity and infectivity of variants. Further studies are needed to monitor the possible cross-species transmission of SARS-CoV-2 between humans and other animals as the pandemic develops.

Acknowledgments: We thank Dr. Xiaowei Jiang for his helpful comments on this review. We apologize for any inadvertent omissions in the relevant literature.

Research funding: This was supported by National Key Research and Development Projects of the Ministry of Science and Technology of China, Grant Nos. 2021YFC0863300, 2020YFA0707600, 2020YFC0847000, 2021YFC2301300 and Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences, Grant No. 2021-12M-1-038.

Author contributions: J.L. and Z.Q. designed the outline of this review and wrote the manuscript; P.L. and X.T. drew the figures and tables. All the authors approved the version to be published, and ensured the accuracy and integrity of the work.

Competing interests: None.

Ethical approval: Not applicable.

Informed consent: Not applicable.

References

1. WHO. WHO coronavirus disease (COVID-19) dashboard; 2021. Available from: <https://covid19.who.int/>.
2. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 2020; 395:565–74.
3. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020;579:270–3.
4. Ren L-L, Wang YM, Wu ZQ, Xiang ZC, Guo L, Xu T, et al. Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. *Chinese Med J* 2020;133:1015–24.
5. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. *Nature* 2020;579:265–9.
6. Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, Chang H. The architecture of SARS-CoV-2 transcriptome. *Cell* 2020;181: 914–21 e910.
7. Parker MD, Lindsey BB, Leary S, Gaudieri S, Chopra A, Wyles M, et al. Subgenomic RNA identification in SARS-CoV-2 genomic sequencing data. *Genome Res* 2021;31:645–58.
8. Masters PS, Perlman S. In *fields virology*, Knipe DM, Howley PM, editors. Philadelphia, PA, USA: Lippincott Williams & Wilkins; 2013, vol. 1:825–58 pp. Ch. 28.
9. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* 2020;181:271–80.
10. Ou X, Liu Y, Lei X, Li P, Mi D, Ren L, et al. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nat Commun* 2020;11:1620.
11. Bracquemond D, Muriaux D. Betacoronavirus assembly: clues and perspectives for elucidating SARS-CoV-2 particle formation and egress. *mBio* 2021;12:e0237121.
12. Finkel Y, Mizrahi O, Nachshon A, Weingarten-Gabbay S, Morgenstern D, Yahalom-Ronen Y, et al. The coding capacity of SARS-CoV-2. *Nature* 2021;589:125–30.
13. Bojkova D, Klann K, Koch B, Widera M, Krause D, Ciesek S, et al. Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature* 2020;583:469–72.
14. Wolff G, Limpens RWAL, Zevenhoven-Dobbe JC, Laugks U, Zheng S, de Jong AWM, et al. A molecular pore spans the double membrane of the coronavirus replication organelle. *Science* 2020;369:1395–8.
15. Gao Y, Yan L, Huang Y, Liu F, Zhao Y, Cao L, et al. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* 2020;368:779–82.
16. Chen J, Malone B, Llewellyn E, Grasso M, Shelton PMM, Olinares PDB, et al. Structural basis for helicase-polymerase coupling in the SARS-CoV-2 replication-transcription complex. *Cell* 2020; 182:1560–73.
17. Ogando NS, Zevenhoven-Dobbe JC, van der Meer Y, Bredenbeek PJ, Posthuma CC, Snijder EJ. The enzymatic activity of the nsp14 exoribonuclease is critical for replication of MERS-CoV and SARS-CoV-2. *J Virol* 2020;94:e01246–20.
18. Ivanov KA, Thiel V, Dobbe JC, van der Meer Y, Snijder EJ, Ziebuhr J. Multiple enzymatic activities associated with severe acute respiratory syndrome coronavirus helicase. *J Virol* 2004; 78:5619–32.
19. Minskaia E, Hertzog T, Gorbalenya AE, Campanacci V, Cambillau C, Canard B, et al. Discovery of an RNA virus 3'-5'. *Proc Natl Acad Sci USA* 2006;103:5108–13.
20. Chen Y, Cai H, Pan J, Xiang N, Tien P, Ahola T, et al. Functional screen reveals SARS coronavirus nonstructural protein nsp14 as a novel cap N7 methyltransferase. *Proc Natl Acad Sci USA* 2009; 106:3484–9.
21. Decroly E, Imbert I, Coutard B, Bouvet M, Selisko B, Alvarez K, et al. Coronavirus nonstructural protein 16 is a cap-0 binding enzyme possessing (nucleoside-2'O)-methyltransferase activity. *J Virol* 2008;82:8071–84.
22. Bouvet M, Debarnot C, Imbert I, Selisko B, Snijder EJ, Canard B, et al. In vitro reconstitution of SARS-coronavirus mRNA cap methylation. *PLoS Pathog* 2010;6:e1000863.
23. Bouvet M, Imbert I, Subissi L, Gluais L, Canard B, Decroly E. RNA 3'-end mismatch excision by the severe acute respiratory syndrome coronavirus nonstructural protein nsp10/nsp14 exoribonuclease complex. *Proc Natl Acad Sci USA* 2012;109: 9372–7.
24. Sawicki SG, Sawicki DL. Coronaviruses use discontinuous extension for synthesis of subgenome-length negative strands. *Adv Exp Med Biol* 1995;380:499–506.
25. WHO. WHO-convened global study of origins of SARS-CoV-2: China part. WHO; 2021. Available from: <https://www.who.int/>

- emergencies/diseases/novel-coronavirus-2019/origins-of-the-virus.
26. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. *Nat Med* 2020;26:450–2.
 27. Wu C-I, Wen H, Lu J, Su X-d, Hughes AC, Zhai W, et al. On the origin of SARS-CoV-2—the blind watchmaker argument. *Sci China Life Sci* 2021;64:1560–3.
 28. Holmes EC, Goldstein SA, Rasmussen AL, Robertson DL, Crits-Christoph A, Wertheim JO, et al. The origins of SARS-CoV-2: a critical review. *Cell* 2021;184:4848–56.
 29. Li J, Lai S, Gao GF, Shi W. The emergence, genomic diversity and global spread of SARS-CoV-2. *Nature* 2021;600:408–18.
 30. Ruan Y, Wen H, He X, Wu C-I. A theoretical exploration of the origin and early evolution of a pandemic. *Sci Bull* 2021;66:1022–9.
 31. Tong Y, Liu W, Liu P, Liu WJ, Wang Q, Gao GF. The origins of viruses: discovery takes time, international resources, and cooperation. *Lancet* 2021;398:1401–2.
 32. Wang Q, Chen H, Shi Y, Hughes AC, Liu WJ, Jiang J, et al. Tracing the origins of SARS-CoV-2: lessons learned from the past. *Cell Res* 2021;31:1139–41.
 33. Wu Z, Jin Q, Wu G, Lu J, Li M, Guo D, et al. SARS-CoV-2's origin should be investigated worldwide for pandemic prevention. *Lancet* 2021;398:1299–303.
 34. Temmam S, Vongphayloth K, Salazar EB, Munier S, Bonomi M, Régnauld B, et al. Coronaviruses with a SARS-CoV-2-like receptor-binding domain allowing ACE2-mediated entry into human cells isolated from bats of Indochinese peninsula. *Research Square* 2021. <https://doi.org/10.21203/rs.3.rs-871965/v1>.
 35. Zhou H, Chen X, Hu T, Li J, Song H, Liu Y, et al. A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the spike protein. *Curr Biol* 2020;30:2196–203.e2193.
 36. Murakami S, Kitamura T, Suzuki J, Sato R, Aoi T, Fujii M, et al. Detection and characterization of bat sarbecovirus phylogenetically related to SARS-CoV-2, Japan. *Emerg Infect Dis* 2020;26:3025–9.
 37. Delaune D, Hul V, Karlsson EA, Hassanin A, Ou TP, Baidaliuk A, et al. A novel SARS-CoV-2 related coronavirus in bats from Cambodia. *Nat Commun* 2021;12:6563.
 38. Wacharapluesadee S, Tan CW, Maneeorn P, Duengkae P, Zhu F, Joyjinda Y, et al. Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia. *Nat Commun* 2021;12:972.
 39. Zhou H, Ji J, Chen X, Bi Y, Li J, Wang Q, et al. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. *Cell* 2021;184:4380–91.e4314.
 40. Li X, Song Y, Wong G, Cui J. Bat origin of a new human coronavirus: there and back again. *Sci China Life Sci* 2020;63:461–2.
 41. Li W, Shi Z, Yu M, Ren W, Smith C, Epstein JH, et al. Bats are natural reservoirs of SARS-like coronaviruses. *Science* 2005;310:676–9.
 42. Dominguez SR, O'Shea TJ, Oko LM, Holmes KV. Detection of group 1 coronaviruses in bats in North America. *Emerg Infect Dis* 2007;13:1295–300.
 43. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol* 2019;17:181–92.
 44. Guo H, Hu B, Si H-R, Zhu Y, Zhang W, Li B, et al. Identification of a novel lineage bat SARS-related coronaviruses that use bat ACE2 receptor. *Emerg Microb Infect* 2021;10:1507–14.
 45. Boni MF, Lemey P, Jiang X, Lam TT-Y, Perry BW, Castoe TA, et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat Microbiol* 2020;5:1408–17.
 46. Singh D, Yi SV. On the origin and evolution of SARS-CoV-2. *Exp Mol Med* 2021;53:537–47.
 47. Wu Z, Han Y, Wang Y, Liu B, Zhao L, Zhang J, et al. A comprehensive survey of bat sarbecoviruses across China for the origin tracing of SARS-CoV and SARS-CoV-2 [Preprint]. *Research Square* 2021. <https://doi.org/10.21203/rs.3.rs-885194/v1>.
 48. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* 2018;35:1547–9.
 49. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;4:406–25.
 50. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 1992;8:275–82.
 51. Wang Q, Zhang Y, Wu L, Niu S, Song C, Zhang Z, et al. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell* 2020;181:894–904.
 52. Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* 2020;581:215–20.
 53. Zhang T, Wu Q, Zhang Z. Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. *Curr Biol* 2020;30:1346–51.
 54. Liu P, Jiang J-Z, Wan X-F, Hua Y, Li L, Zhou J, et al. Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? *PLoS Pathog* 2020;16:e1008421.
 55. Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou J-J, et al. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature* 2020;583:286–9.
 56. Lam TT, Jia N, Zhang Y-W, Shum MH-H, Jiang J-F, Zhu H-C, et al. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* 2020;583:282–5.
 57. Peng M-S, Li JB, Cai ZF, Liu H, Tang X, Ying R, et al. The high diversity of SARS-CoV-2-related coronaviruses in pangolins alerts potential ecological risks. *Zool Res* 2021;42:833.
 58. Zhang S, Qiao S, Yu J, Zeng J, Shan S, Tian L, et al. Bat and pangolin coronavirus spike glycoprotein structures provide insights into SARS-CoV-2 evolution. *Nat Commun* 2021;12:1607.
 59. Nie J, Li Q, Zhang L, Cao Y, Zhang Y, Li T, et al. Functional comparison of SARS-CoV-2 with closely related pangolin and bat coronaviruses. *Cell Discov* 2021;7:21.
 60. Dicken SJ, Murray MJ, Thorne LG, Reuschl A-K, Forrest C, Ganeshalingham M, et al. Characterisation of B.1.1.7 and pangolin coronavirus spike provides insights on the evolutionary trajectory of SARS-CoV-2. *BioRxiv* 2021. <https://doi.org/10.1101/2021.03.22.436468>.
 61. Zhao Z, Li H, Wu X, Zhong Y, Zhang K, Zhang Y-P, et al. Moderate mutation rate in the SARS coronavirus genome and its implications. *BMC Evol Biol* 2004;4:21.

62. Cotten M, Watson SJ, Zumla AI, Makhdoom HQ, Palser AL, Ong SH, et al. Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus. *mBio* 2014;5:e01062–13.
63. Giovanetti M, Benvenuto D, Angeletti S, Ciccozzi M. The first two cases of 2019-nCoV in Italy: where they come from? *J Med Virol* 2020;92:518–21.
64. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol* 2020;83:104351.
65. Lai A, Bergna A, Acciarri C, Galli M, Zehender G. Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2. *J Med Virol* 2020;92:675–9.
66. Lycett S, Inward R. Phylogeography with whole genomes 24 Mar 2020. *Virological website*; 2020. Available from: <https://virological.org/t/phylogeography-with-whole-genomes-24-mar-2020/444>.
67. Li X, Zai J, Zhao Q, Nie Q, Li Y, Foley BT, et al. Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. *J Med Virol* 2020;92:602–11.
68. Chaw SM, Tai J-H, Chen S-L, Hsieh C-H, Chang S-Y, Yeh S-H, et al. The origin and underlying driving forces of the SARS-CoV-2 outbreak. *J Biomed Sci* 2020;27:73.
69. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* 2008;9:267–76.
70. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. *Natl Sci Rev* 2020;7:1012–23.
71. Li X, Giorgi EE, Marichannegowda MH, Foley B, Xiao C, Kong X-P, et al. Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci Adv* 2020;6:eabb9153.
72. Cagliani R, Forni D, Clerici M, Sironi M. Computational inference of selection underlying the evolution of the novel coronavirus, severe acute respiratory syndrome coronavirus 2. *J Virol* 2020;94:e00411–20.
73. Damas J, Hughes GM, Keough KC, Painter CA, Persky NS, Corbo M, et al. Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates. *Proc Natl Acad Sci USA* 2020;117:22311–22.
74. Nozawa M, Suzuki Y, Nei M. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci USA* 2009;106:6700.
75. Li L-L, Wang J-L, Ma X-H, Li J-S, Yang X-F, Shi W-F, et al. A novel SARS-CoV-2 related virus with complex recombination isolated from bats in Yunnan Province, China. *bioRxiv* 2021. <https://doi.org/10.1101/2021.03.17.435823>.
76. Sun J, He W-T, Wang L, Lai A, Ji X, Zhai X, et al. COVID-19: epidemiology, evolution, and cross-disciplinary perspectives. *Trends Mol Med* 2020;26:483–95.
77. Liu K, Pan X, Li L, Yu F, Zheng A, Du P, et al. Binding and molecular basis of the bat coronavirus RaTG13 virus to ACE2 in humans and other species. *Cell* 2021;184:3438–51.
78. Li P, Guo R, Liu Y, Zhang Y, Hu J, Ou X, et al. The *Rhinolophus affinis* bat ACE2 and multiple animal orthologs are functional receptors for bat coronavirus RaTG13 and SARS-CoV-2. *Sci Bull* 2021;66:1215–27.
79. Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, et al. Structural basis of receptor recognition by SARS-CoV-2. *Nature* 2020;581:221–4.
80. Wong MC, Cregeen SJJ, Ajami NJ, Petrosino JF. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019 [Preprint]. *bioRxiv* 2020. <https://doi.org/10.1101/2020.02.07.939207>.
81. Niu S, Wang J, Bai B, Wu L, Zheng A, Chen Q, et al. Molecular basis of cross-species ACE2 interactions with SARS-CoV-2-like viruses of pangolin origin. *EMBO J* 2021;40:e107786.
82. Bhattacharjee MJ, Lin J-J, Chang C-Y, Chiou Y-T, Li T-N, Tai C-W, et al. Identifying primate ACE2 variants that confer resistance to SARS-CoV-2. *Mol Biol Evol* 2021;38:2715–31.
83. Liu K, Tan S, Niu S, Wang J, Wu L, Sun H, et al. Cross-species recognition of SARS-CoV-2 to bat ACE2. *Proc Natl Acad Sci USA* 2021;118:e2020216118.
84. Di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci Adv* 2020;6:eabb5813.
85. Song Y, He X, Yang W, Tang T, Zhang R. ADAR mediated A-to-I RNA editing affects SARS-CoV-2 characteristics and fuels its evolution [Preprint]. *bioRxiv* 2021. <https://doi.org/10.1101/2021.07.22.453345>.
86. Kosuge M, Furusawa-Nishii E, Ito K, Saito Y, Ogasawara K. Point mutation bias in SARS-CoV-2 variants results in increased ability to stimulate inflammatory responses. *Sci Rep* 2020;10:17766.
87. Shan K-J, Wei C, Wang Y, Huan Q, Qian W. Host-specific asymmetric accumulation of mutation types reveals that the origin of SARS-CoV-2 is consistent with a natural process. *Innovation* 2021;2:100159.
88. Rice AM, Castillo Morales A, Ho AT, Mordstein C, Mühlhausen S, Watson S, et al. Evidence for strong mutation bias toward, and selection against, U content in SARS-CoV-2: implications for vaccine design. *Mol Biol Evol* 2021;38:67–83.
89. De Maio N, Walker CR, Turakhia Y, Lanfear R, Corbett-Detig R, Goldman N. Mutation rates and selection on synonymous mutations in SARS-CoV-2 [Preprint]. *bioRxiv* 2021. <https://doi.org/10.1101/2021.01.14.426705>.
90. Li T, Tang X, Wu C, Yao X, Wang Y, Lu X, et al. The use of SARS-CoV-2-related coronaviruses from bats and pangolins to polarize mutations in SARS-Cov-2. *Sci China Life Sci* 2020;63:1608–11.
91. Deng S, Xing K, He X. Mutation signatures inform the natural host of SARS-CoV-2 [Preprint]. *Natl Sci Rev* 2021. <https://doi.org/10.1093/nsr/nwab220>.
92. Ratcliff J, Simmonds P. Potential APOBEC-mediated RNA editing of the genomes of SARS-CoV-2 and other coronaviruses and its impact on their longer term evolution. *Virology* 2021;556:62–72.
93. Amicone M, Borges V, Alves MJ, Isidro J, Zé-Zé L, Duarte S, et al. Mutation rate of SARS-CoV-2 and emergence of mutators during experimental evolution [Preprint]. *bioRxiv* 2021. <https://doi.org/10.1101/2021.05.19.444774>.
94. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Chall* 2017;1:33–46.
95. Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data – from vision to reality. *Euro Surveill* 2017;22:30494.
96. Zhang L, Yang J-R, Zhang Z, Lin Z. Genomic variations of SARS-CoV-2 suggest multiple outbreak sources of transmission [Preprint]. *medRxiv* 2020. <https://doi.org/10.1101/2020.02.25.20027953>.

97. Yu W-B, Tang G-D, Zhang L, Corlett RT. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2/HCoV-19) using whole genomic data. *Zool Res* 2020;41:247–57.
98. Matsuda T, Suzuki H, Ogata N. Phylogenetic analyses of the severe acute respiratory syndrome coronavirus 2 reflected the several routes of introduction to Taiwan, the United States, and Japan. *arXiv* 2020. <https://arxiv.org/abs/2002.08802>.
99. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci USA* 2020; 117:9241–3.
100. Wu A, Niu P, Wang L, Zhou H, Zhao X, Wang W, et al. Mutations, recombination and insertion in the evolution of 2019-nCoV [Preprint]. *bioRxiv* 2020. <https://doi.org/10.1101/2020.02.29.971101>.
101. Hu B, Guo H, Zhou P, Shi ZL. Characteristics of SARS-CoV-2 and COVID-19. *Nat Rev Microbiol* 2020;19:1–14.
102. Zhang YZ, Holmes EC. A genomic perspective on the origin and emergence of SARS-CoV-2. *Cell* 2020;181:223–7.
103. Wong G, Bi YH, Wang QH, Chen XW, Zhang ZG, Yao YG. Zoonotic origins of human coronavirus 2019 (HCoV-19/SARS-CoV-2): why is this work important? *Zool Res* 2020;41:213–9.
104. Banerjee A, Doxey AC, Mossman K, Irving AT. Unraveling the zoonotic origin and transmission of SARS-CoV-2. *Trends Ecol Evol* 2021;36:180–4.
105. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 2018;34:4121–3.
106. Tang X, Ying R, Yao X, Li G, Wu C, Tang Y, et al. Evolutionary analysis and lineage designation of SARS-CoV-2 genomes. *Sci Bull* 2021;66:2297–311.
107. Jombart T, Eggo RM, Dodd PJ, Balloux F. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity* 2011; 106:383–90.
108. Paradis E. Analysis of haplotype networks: the randomized minimum spanning tree method. *Methods Ecol Evol* 2018;9: 1308–17.
109. Gomez-Carballa A, Bello X, Pardo-Seco J, Martinon-Torres F, Salas A. Mapping genome variation of SARS-CoV-2 worldwide highlights the impact of COVID-19 super-spreaders. *Genome Res* 2020;30:1434–48.
110. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* 2021;19:409–24.
111. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* 2020;182:812–27.
112. Trucchi E, Gratton P, Mafessoni F, Motta S, Cicconardi F, Mancina F, et al. Population dynamics and structural effects at short and long range support the hypothesis of the selective advantage of the G614 SARS-CoV-2 spike variant. *Mol Biol Evol* 2021;38: 1966–79.
113. Volz E, Hill V, McCrone JT, Price A, Jorgensen D, O’Toole Á, et al. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. *Cell* 2021;184:64–75.
114. Zhou B, Thao TTN, Hoffmann D, Taddeo A, Ebert N, Labrousseau F, et al. SARS-CoV-2 spike D614G change enhances replication and transmission. *Nature* 2021;592:122–7.
115. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, et al. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* 2020; 592:116–21.
116. Daniloski Z, Jordan TX, Ilmain JK, Guo X, Bhabha G, tenOever BR, et al. The spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types. *eLife* 2021;10:e65365.
117. Ozono S, Zhang Y, Ode H, Sano K, Tan TS, Imai K, et al. SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity. *Nat Commun* 2021;12:848.
118. Hu J, He C-L, Gao Q-Z, Zhang G-J, Cao X-X, Long Q-X, et al. The D614G mutation of SARS-CoV-2 spike protein enhances viral infectivity [Preprint]. *bioRxiv* 2020. <https://doi.org/10.1101/2020.06.20.161323>.
119. Zhang L, Jackson CB, Mou H, Ojha A, Peng H, Quinlan BD, et al. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat Commun* 2020;11:6013.
120. Yurkovetskiy L, Wang X, Pascal KE, Tomkins-Tinch C, Nyalile TP, Wang Y, et al. Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell* 2020;183: 739–51.
121. Hou YJ, Chiba S, Halfmann P, Ehre C, Kuroda M, Dinno KH, et al. SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* 2020;370:1464–8.
122. Weissman D, Alameh M-G, de Silva T, Collini P, Hornsby H, Brown R, et al. D614G spike mutation increases SARS CoV-2 susceptibility to neutralization. *Cell Host Microbe* 2021;29:23–31.
123. Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, Munday JD, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* 2021;372:eabg3055.
124. Collier DA, De Marco A, Ferreira IATM, Meng B, Datir RP, Walls AC, et al. Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. *Nature* 2021;593:136–41.
125. Wang P, Nair MS, Liu L, Iketani S, Luo Y, Guo Y, et al. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. *Nature* 2021;593:130–5.
126. Graham C, Seow J, Huettnner I, Khan H, Kouphou N, Acors S, et al. Neutralization potency of monoclonal antibodies recognizing dominant and subdominant epitopes on SARS-CoV-2 spike is impacted by the B.1.1.7 variant. *Immunity* 2021;54:1276–89.
127. Muik A, Wallisch A-K, Sanger B, Swanson KA, Muhl J, Chen W, et al. Neutralization of SARS-CoV-2 lineage B.1.1.7 pseudovirus by BNT162b2 vaccine-elicited human sera. *Science* 2021;371: 1152–3.
128. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* 2021;592:438–43.
129. Wibmer CK, Ayres F, Hermanus T, Madzivhandila M, Kgagudi P, Oosthuysen B, et al. SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *Nat Med* 2021;27: 622–5.
130. Cele S, Gazy I, Jackson L, Hwa S-H, Tegally H, Lustig G, et al. Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma. *Nature* 2021;593:142–6.
131. Faria NR, Mellan TA, Whittaker C, Claro IM, Candido DDS, Mishra S, et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* 2021;372:815–21.
132. Wang P, Casner RG, Nair MS, Wang M, Yu J, Cerutti G, et al. Increased resistance of SARS-CoV-2 variant P.1 to antibody neutralization. *Cell Host Microbe* 2021;29:747–51.

133. Garcia-Beltran WF, Lam EC, St Denis K, Nitido AD, Garcia ZH, Hauser BM, et al. Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell* 2021; 184:2372–83.
134. Mlcochova P, Kemp SA, Dhar MS, Papa G, Meng B, Ferreira IATM, et al. SARS-CoV-2 B.1.617.2 delta variant replication and immune evasion. *Nature* 2021;599:114–9.
135. Zhang J, Xiao T, Cai Y, Lavine CL, Peng H, Zhu H, et al. Membrane fusion and immune evasion by the spike protein of SARS-CoV-2 delta variant. *Science* 2021;374:1353–60.
136. Li B, Deng A, Li K, Hu Y, Li Z, Xiong Q, et al. Viral infection and transmission in a large, well-traced outbreak caused by the SARS-CoV-2 delta variant [Preprint]. *medRxiv* 2021. <https://doi.org/10.1101/2021.07.07.21260122>.
137. McCallum M, Bassi J, De Marco A, Chen A, Walls AC, Di Iulio J, et al. SARS-CoV-2 immune evasion by the B.1.427/B.1.429 variant of concern. *Science* 2021;373:648–54.
138. Motozono C, Toyoda M, Zahradnik J, Ikeda T, Saito A, Tan TS, et al. An emerging SARS-CoV-2 mutant evading cellular immunity and increasing viral infectivity [Preprint]. *bioRxiv* 2021. <https://doi.org/10.1101/2021.04.02.438288>.
139. Liu Z, VanBlargan LA, Bloyet L-M, Rothlauf PW, Chen RE, Stumpf S, et al. Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *Cell Host Microbe* 2021;29:477–88.
140. Greaney AJ, Starr TN, Gilchuk P, Zost SJ, Binshtein E, Loes AN, et al. Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. *Cell Host Microbe* 2021;29:44–57.
141. Deng X, Garcia-Knight MA, Khalid MM, Servellita V, Wang C, Morris MK, et al. Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant. *Cell* 2021;184: 3426–37.e3428.
142. Starr TN, Greaney AJ, Dingens AS, Bloom JD. Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016. *Cell Rep Med* 2021;2:100255.
143. Zhang W, Davis BD, Chen SS, Sincuir Martinez JM, Plummer JT, Vail E. Emergence of a novel SARS-CoV-2 variant in Southern California. *JAMA* 2021;325:1324–6.
144. Li QQ, Wu J, Nie J, Zhang L, Hao H, Liu S, et al. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell* 2020;182:1284–94.
145. Acevedo ML, Alonso-Palomares L, Bustamante A, Gaggero A, Paredes F, Cortés CP, et al. Infectivity and immune escape of the new SARS-CoV-2 variant of interest lambda [Preprint]. *medRxiv* 2021. <https://doi.org/10.1101/2021.06.28.21259673>.
146. Kimura I, Kosugi Y, Wu J, Yamasoba D, Butlertanaka EP, Tanaka YL, et al. SARS-CoV-2 lambda variant exhibits higher infectivity and immune resistance [Preprint]. *bioRxiv* 2021. <https://doi.org/10.1101/2021.07.28.454085>.
147. Tada T, Zhou H, Dcosta BM, Samanovic MI, Mulligan MJ, Landau NR. SARS-CoV-2 lambda variant remains susceptible to neutralization by mRNA vaccine-elicited antibodies and convalescent serum [Preprint]. *bioRxiv* 2021. <https://doi.org/10.1101/2021.07.02.450959>.
148. Liu Y, Liu J, Johnson BA, Xia H, Ku Z, Schindewolf C, et al. Delta spike P681R mutation enhances SARS-CoV-2 fitness over alpha variant [Preprint]. *bioRxiv* 2021. <https://doi.org/10.1101/2021.08.12.456173>.
149. Wei C, Shan K-J, Wang W, Zhang S, Huan Q, Qian W. Evidence for a mouse origin of the SARS-CoV-2 omicron variant. *J Genet Genom* 2021;48:1111–21.
150. Du P, Gao F, Wang Q. The mysterious origins of the omicron variant of SARS-CoV-2 [Preprint]. *Innovation* 2022. <https://doi.org/10.1016/j.xinn.2022.100206>.
151. Wang Y, Li Q, Liang Z, Li T, Liu S, Cui Q, et al. The significant immune escape of pseudotyped SARS-CoV-2 variant omicron. *Emerg Microb Infect* 2022;11:1–5.
152. Cao Y, Wang J, Jian F, Xiao T, Song W, Yisimayi A, et al. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies [Preprint]. *Nature* 2021. <https://doi.org/10.1038/s41586-021-04385-3>.
153. Cele S, Jackson L, Khoury DS, Khan K, Moyo-Gwete T, Tegally H, et al. Omicron extensively but incompletely escapes Pfizer BNT162b2 neutralization [Preprint]. *Nature* 2021. <https://doi.org/10.1038/s41586-021-04387-1>.
154. Rössler A, Riepler L, Bante D, von Laer D, Kimpel J. SARS-CoV-2 omicron variant neutralization in serum from vaccinated and convalescent persons [Preprint]. *N Engl J Med* 2022. <https://doi.org/10.1056/NEJMc2119236>.
155. Zhao H, Lu L, Peng Z, Chen L-L, Meng X, Zhang C, et al. SARS-CoV-2 omicron variant shows less efficient replication and fusion activity when compared with delta variant in TMPRSS2-expressed cells. *Emerg Microb Infect* 2021;11:1–18.
156. Meng B, Ferreira IATM, Abdullahi A, Goonawardane N, Saito A, Kimura I, et al. SARS-CoV-2 omicron spike mediated immune escape and tropism shift [Preprint]. *bioRxiv* 2022. <https://doi.org/10.1101/2021.12.17.473248>.
157. McMahan K, Giffin V, Tostanoski LH, Chung B, Siamatu M, Suthar MS, et al. Reduced pathogenicity of the SARS-CoV-2 omicron variant in hamsters [Preprint]. *bioRxiv* 2022. <https://doi.org/10.1101/2022.01.02.474743>.
158. Bentley EG, Kirby A, Sharma P, Kipar A, Mega DF, Bramwell C, et al. SARS-CoV-2 omicron-B.1.1.529 variant leads to less severe disease than pango B and delta variants strains in a mouse model of severe COVID-19 [Preprint]. *bioRxiv* 2021. <https://doi.org/10.1101/2021.12.26.474085>.
159. Zhu X, Mannar D, Srivastava SS, Berezuk AM, Demers J-P, Saville JW, et al. Cryo-electron microscopy structures of the N501Y SARS-CoV-2 spike protein in complex with ACE2 and 2 potent neutralizing antibodies. *PLoS Biol* 2021;19: e3001237.
160. Tian F, Tong B, Sun L, Shi S, Zheng B, Wang Z, et al. N501Y mutation of spike protein in SARS-CoV-2 strengthens its binding to receptor ACE2. *eLife* 2021;10:e69091.
161. Baum A, Fulton BO, Wloga E, Copin R, Pascal KE, Russo V, et al. Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science* 2020;369:1014–8.
162. Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, Lorenzi JC, et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *eLife* 2020;9:e61312.
163. Kuzmina A, Khalaila Y, Voloshin O, Keren-Naus A, Boehm-Cohen L, Raviv Y, et al. SARS-CoV-2 spike variants exhibit differential infectivity and neutralization resistance to convalescent or post-vaccination sera. *Cell Host Microbe* 2021;29:522–8.
164. Li Q, Nie J, Wu J, Zhang L, Ding R, Wang H, et al. SARS-CoV-2 501Y.V2 variants lack higher infectivity but do have immune escape. *Cell* 2021;184:2362–71.

165. Zhou D, Dejnirattisai W, Supasa P, Liu C, Mentzer AJ, Ginn HM, et al. Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. *Cell* 2021;184:2348–61.
166. Shi R, Shan C, Duan X, Chen Z, Liu P, Song J, et al. A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. *Nature* 2020;584:120–4.
167. Ju B, Zhang Q, Ge J, Wang R, Sun J, Ge X, et al. Human neutralizing antibodies elicited by SARS-CoV-2 infection. *Nature* 2020;584:115–9.
168. Hu B, Liu R, Tang X, Pan Y, Wang M, Tong Y, et al. The concordance between the evolutionary trend and the clinical manifestation of the two SARS-CoV-2 variants. *Natl Sci Rev* 2021; 8:nwab073.
169. Stauff CB, Lien CZ, Selvaraj P, Liu S, Wang TT. The G614 pandemic SARS-CoV-2 variant is not more pathogenic than the original D614 form in adult Syrian hamsters. *Virology* 2021;556:96–100.
170. Graham MS, Sudre CH, May A, Antonelli M, Murray B, Varsavsky T, et al. Changes in symptomatology, reinfection, and transmissibility associated with the SARS-CoV-2 variant B.1.1.7: an ecological study. *Lancet Public Health* 2021;6:E335–45.
171. Davies NG, Jarvis CI, CMMID COVID-19 Working Group, Edmunds WJ, Jewell NP, Diaz-Ordaz K, et al. Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. *Nature* 2021;593:270–4.
172. Frampton D, Rampling T, Cross A, Bailey H, Heaney J, Byott M, et al. Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B.1.1.7 lineage in London, UK: a whole-genome sequencing and hospital-based cohort study. *Lancet Infect Dis* 2021;21:1246–56.
173. Challen R, Brooks-Pollock E, Read JM, Dyson L, Tsaneva-Atanasova K, Danon L. Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study. *BMJ Br Med J* 2021;372:n579.
174. Jassat W, Mudara C, Ozougwu L, Tempia S, Blumberg L, Davies MA, et al. Difference in mortality among individuals admitted to hospital with COVID-19 during the first and second waves in South Africa: a cohort study. *Lancet Global Health* 2021;9: e1216–25.
175. Radvak P, Kwon H-J, Kosikova M, Ortega-Rodriguez U, Xiang R, Phue J-N, et al. SARS-CoV-2 B.1.1.7 (alpha) and B.1.351 (beta) variants induce pathogenic patterns in K18-hACE2 transgenic mice distinct from early strains. *Nat Commun* 2021;12:6559.
176. Chen Q, Huang X-Y, Tian Y, Fan C, Sun M, Zhou C, et al. The infection and pathogenicity of SARS-CoV-2 variant B.1.351 in hACE2 mice. *Virol Sin* 2021;36:1232–5.
177. Munster VJ, Flagg M, Singh M, Yinda CK, Williamson BN, Feldmann F, et al. Subtle differences in the pathogenicity of SARS-CoV-2 variants of concern B.1.1.7 and B.1.351 in rhesus macaques. *Sci Adv* 2021;7:eabj3627.
178. Saito A, Nasser H, Uriu K, Kosugi Y, Irie T, Shirakawa K, et al. SARS-CoV-2 spike P681R mutation, a hallmark of the delta variant, enhances viral fusogenicity and pathogenicity [Preprint]. *bioRxiv* 2021. <https://doi.org/10.1101/2021.06.17.448820>.
179. Teuwen L-A, Geldhof V, Pasut A, Carmeliet P. COVID-19: the vasculature unleashed. *Nat Rev Immunol* 2020;20:389–91.
180. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* 2020;395:507–13.
181. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA* 2020; 323:1061–9.
182. Sun S, Gu H, Cao L, Chen Q, Ye Q, Yang G, et al. Characterization and structural basis of a lethal mouse-adapted SARS-CoV-2. *Nat Commun* 2021;12:5654.
183. Alizon S, Hurford A, Mideo N, Van Baalen M. Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future. *J Evol Biol* 2009;22:245–59.
184. Cressler CE, McLeod DV, Rozins C, Van Den Hoogen J, Day T. The adaptive evolution of virulence: a review of theoretical predictions and empirical tests. *Parasitology* 2016;143: 915–30.
185. Bull JJ, Lauring AS. Theory and empiricism in virulence evolution. *PLoS Pathog* 2014;10:e1004387.
186. Lipsitch M, Moxon ER. Virulence and transmissibility of pathogens: what is the relationship? *Trends Microbiol* 1997;5: 31–7.
187. Kang L, He G, Sharp AK, Wang X, Brown AM, Michalak P, et al. A selective sweep in the spike gene has driven SARS-CoV-2 human adaptation. *Cell* 2021;184:4392–400.
188. Gu H, Chen Q, Yang G, He L, Fan H, Deng Y-Q, et al. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science* 2020;369:1603–7.
189. Huang K, Zhang Y, Hui X, Zhao Y, Gong W, Wang T, et al. Q493K and Q498H substitutions in spike promote adaptation of SARS-CoV-2 in mice. *EBioMedicine* 2021;67:103381.
190. Dinno KH, 3rd, Leist SR, Schäfer A, Edwards CE, Martinez DR, Montgomery SA, et al. A mouse-adapted model of SARS-CoV-2 to test COVID-19 countermeasures. *Nature* 2020; 586:560–6.
191. Gao GF, Wang L. COVID-19 expands its territories from humans to animals. *China CDC Wkly* 2021;3:855–8.
192. Shi J, Wen Z, Zhong G, Yang H, Wang C, Huang B, et al. Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS-coronavirus 2. *Science* 2020;368:1016–20.
193. Chandler JC, Bevins SN, Ellis JW, Linder TJ, Tell RM, Jenkins-Moore M, et al. SARS-CoV-2 exposure in wild white-tailed deer (*Odocoileus virginianus*) [Preprint]. *bioRxiv* 2021. <https://doi.org/10.1101/2021.07.29.454326>.
194. Oude Munnink BB, Sikkema RS, Nieuwenhuijse DF, Molenaar RJ, Munger E, Molenkamp R, et al. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* 2020;371:172–7.