Briefings in Bioinformatics, 00(00), 2021, 1–9

doi: 10.1093/bib/bbab144
Case Study

Comparative phylogenetic analysis of SARS-CoV-2 spike protein—possibility effect on virus spillover

Abozar Ghorbani, Samira Samarfard, Neda Eskandarzade, Alireza Afsharifar, Mohammad Hadi Eskandari, Ali Niazi, Keramatollah Izadpanah and Thomas P. Karbanowicz

Corresponding author: Abozar Ghorbani, Plant Virology Research Centre, College of Agriculture, Shiraz University, Shiraz, Iran. Tel.: +98 7132286154, +98 9166632176; Fax: +98 7132286154; E-mail: Abozar.ghorbani@shirazu.ac.ir

Abstract

Coronavirus disease 2019 has developed into a dramatic pandemic with tremendous global impact. The receptor-binding motif (RBM) region of the causative virus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), binds to host angiotensin-converting enzyme 2 (ACE2) receptors for infection. As ACE2 receptors are highly conserved within vertebrate species, SARS-CoV-2 can infect significant animal species as well as human populations. An analysis of SARS-CoV-2 genotypes isolated from human and significant animal species was conducted to compare and identify mutation and adaptation patterns across different animal species. The phylogenetic data revealed seven distinct phylogenetic clades with no significant relationship between the clades and geographical locations. A high rate of variation within SARS-CoV-2 mink isolates implies that mink populations were infected before human populations. Positions of most single-nucleotide polymorphisms (SNPs) within the spike (S) protein of SARS-CoV-2 genotypes from the different hosts are mostly accumulated in the RBM region and highlight the pronounced accumulation of variants with mutations in the RBM region in comparison with other variants. These SNPs play a crucial role in viral transmission and pathogenicity and are keys in identifying other animal species as potential intermediate hosts of SARS-CoV-2. The possible roles in the emergence of new viral strains and the possible implications of these changes, in compromising vaccine effectiveness, deserve urgent considerations.

Key words: SARS-CoV-2; coronavirus; selective pressure; SNP

Introduction

The coronavirus disease 2019 (COVID-19) rapidly developed into a dramatic pandemic with a tremendous global impact. First diagnosed in Wuhan, China, in December 2019, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has spread globally. The causal agent of COVID-19, SARS-CoV-2, infects the lower respiratory tract, leading to severe acute respiratory syndrome and multiple organ failure [1, 2]. As of 10 January 2021, there have been 88 387 352 cumulative cases globally, with 1 919 204 cumulative deaths [3]. As SARS-CoV-2 continued to spread globally, the scientific community mobilized to urgently conduct a comprehensive examination of the SARS-CoV-2 virus to develop effective vaccines and therapeutic monoclonal antibodies (mAbs) [4].

Abozar Ghorbani is a Postdoctoral Researcher at Shiraz University and member of SARS-CoV-2 vaccine production team in Shiraz University.
Samira Samarfard is a PhD in virology from The University of Queensland, Australia.
Neda Eskandarzade is a Researcher at School of Veterinary Medicine, Shahid Bahonar University of Kerman.
Alireza Afsharifar is a Professor of plant virology and member of SARS-CoV-2 vaccine production team in Shiraz University.
Mohammad Hadi Eskandari is a Professor of food science and technology and head of SARS-CoV-2 vaccine production team in Shiraz University.
Ali Niazi is a Professor of biotechnology and member of SARS-CoV-2 vaccine production team in Shiraz University.
Keramatollah Izadpanah is a Professor of virology and member of SARS-CoV-2 vaccine production team in Shiraz University.
Thomas P. Karbanowicz is a PhD in microbiology at The University of Queensland.
Submitted: 19 February 2021; **Received (in revised form):** 18 March 2021

SARS-CoV-2 genome is a ~30 kb none-segmented, single-stranded, positive-sense ribonucleic acid comprising of 14 open reading frames (ORFs). The 5' end of the genome possesses a leader sequence and a noncoding/untranslated region, followed by two overlapping ORFs (ORF1/ab) spinning most of the genome and encode nonstructural proteins that facilitate viral replication and transcription [5–7]. The remaining downstream (~10 kb) segment includes genes associated with the viral envelope encoding canonical structural proteins (SPs) such as spike (S), nucleocapsid, envelope and membrane proteins, and accessory proteins [8–10].

Of all the SPs, the trimeric S proteins that protrude from the virus envelope play a key role in facilitating viral entry into the host cell [11, 12]. The S protein of SARS-CoV-2 is a type-I transmembrane (TM) protein that comprises a large ectodomain, a single-pass TM and an intracellular tail. The ectodomain part of S protein includes the S1 subunit with a receptor-binding domain (RBD) and the membrane-fusion subunit (S2) [12]. To initiate viral infection and host cell entry, the receptor-binding motif (RBM) of the S protein, which is present in the S1 subunit, binds to the host angiotensin-converting enzyme 2 (ACE2) receptor. Once attached, the host membrane protein TM protease serine 2 (TMPRSS2) transfers the virus into the target cell through cleaving S1 and S2 subunits [13, 14]. As the S protein is critical to SARS-CoV-2 viral entry, it has been the primary target of antibody therapies and the identification of vaccine candidates. Therefore, a comprehensive understanding of the structural, molecular and mutational traits of the S protein and its interaction with cognate receptors is important in designing vaccines and antiviral drugs against SARS-CoV-2.

The RBM region (residues 319–541) is important in binding the host ACE2 receptor for infection; thus, mutations in this region can facilitate virus transmission and increase the diversity of SARS-CoV-2-prone hosts [15]. As ACE2 is highly conserved in a broad spectrum of vertebrates, a wide array of animals have been predicted to be intermediate hosts for SARS-CoV-2 [16, 17]. This has been supported by *in vivo* studies into SARS-CoV-2 infection and transmission in dogs, cats, ferrets, hamsters, *Rhesus macaques*, etc. [18–20]. Phylogenetic studies have exhibited that SARS-CoV-2 strains isolated from infected dogs, cats, tigers and lions have diverged from human strains [21–23]. However, animals such as American mink, cows, rabbits and sheep are also at risk of SARS-CoV-2 infection [24]. Malayan pangolins (*Manis javanica*) have been proposed as the first potential intermediate host for SARS-CoV-2 due to 97.4% similarity between amino acid sequences of RBM in pangolin-CoVs and SARS-CoV-2; a higher genetic similarity (>95%) between SARS-CoV-2 and horseshoe bat SARS-CoV (RaTG13) suggests that SARS-CoV-2 is supposedly diverged from RaTG13 many years ago [25–28].

SARS-CoV-2 infections in mink farms have been reported in the Netherlands, Denmark, Spain, France, Sweden, Italy, United States and Greece [29, 30]. Munnink *et al*. [31] identified human to mink, mink to mink and mink to human transmission of SARS-CoV-2, with the presence of human sequence clusters reported from mink isolates from the Netherlands outbreak [16]. Some farmers, workers and their close contacts in Europe were infected with SARS-CoV-2 strains containing the animal sequence signature, suggesting a SARS-CoV-2 spillover between animal and human populations [31]. However, SARS-CoV-2 was reported in mink farm workers both before and after the summer outbreak, indicating that vertebrate species such as mink may act as reservoir hosts, leading to circular reinfection [16, 32]. The potential of circular reinfection, and the presence of human clusters within mink isolates, emphasize the importance to monitor mutations of SARS-CoV-2 strains. As they are transmissible between human and animal populations, mutations occurring within viral genome regions targeted by current diagnostic tests, antiviral drugs and vaccine development may affect their efficacy. Currently, there is no evidence indicating if mutations within mink strains of SARS-CoV-2 affect the recognition and neutralization of antibodies designed to target dominant human strains.

Analysis of viral sequences from the Netherlands and Denmark revealed multiple substitutions including Y453F, F486L and N501T in the RBD of S protein, which supposedly mediates increased binding affinity for mink ACE2 compared with human ACE2 [31, 33, 34]. Furthermore, the Danish variant of the virus termed 'cluster 5' revealed three additional mutations in the S protein (del69_70, I692V and M1229I). Transmission of the SARS-CoV-2 variant from animal to human suggests that the variant is not fully compatible with the human host [28, 35], indicating potential selective pressure for the virus to adapt to the new host, by optimizing its S protein-binding affinity with the host's ACE2 receptor [33]. Virus adaptation in human or susceptible animal hosts leads to mutations and genetic diversity in the virus in the frame of 'evolution', which can allow the production of different clades or a new lineage with higher transmission power [36].

A comparative analysis of the mutation patterns in different animal hosts can inform on SARS-CoV-2 evolution and adaptation within animal species. This information could provide a foundation to further control the pandemic by producing host-specific mAbs or vaccines. To address this, our study aims to analyze the phylogenetic relationship between SARS-CoV-2 genotypes from human and animal hosts to predict the evolutionary forces behind this pandemic, and the potential risks of selective pressures that may affect the pathogenicity of SARS-CoV-2 or the efficacy of vaccines and antiviral therapies. Screening genomic changes in the virus specifically at the S protein level is essential and plays a pivotal role in all the above biological concepts of the virus, as the emergence of genetic variants could affect the efficacy of currently developed vaccines. Therefore, we performed single-nucleotide polymorphism (SNP) analysis in S protein sequences of viral variants isolated from human and animal hosts by determining nucleotides that differed from the reference sequence using CLC Genomics Workbench software (version 20, QIAGEN, Venlo, the Netherlands).

Materials and methods

Data retrieval and arrangements

Whole-genome sequences (WGS) of 504 SARS-CoV-2 genotypes isolated from human and eight animal species (pangolin, bat, dogs, mouse, mink, cat, tiger and lion), including their geographical annotations, were downloaded from Global Initiative on Sharing All Influenza Data (GISAID) databases [37]. Representative sequences were selected from Asia, Europe, North America, Africa, South America and Oceania.

The raw sequences were aligned to the corresponding sequence (reference sequence: NC_045512.2) of type strain using ClustalW (version 2.1) in Geneious Prime 2019 (Biomatters, New Zealand). Low-quality sequences with ambiguous nucleotides were manually trimmed to obtain sequences of the approximate size.

Genetic variation of S protein of viral genotypes at interhost and intrahost levels

To assess the genetic diversity of S protein in the selected viral population, the consensus sequence of the S protein region was retrieved from the entire genome sequences of SARS-CoV-2 strains (Supplementary Table S1). To confirm the sequences of the S protein region, spike sequences were reviewed to ensure they were in the correct frame for translation. Then, the evolutionary distance analysis among the S protein sequences from different host groups and estimation of genetic variability of the S protein consensus sequences within individual species were assessed by the maximum likelihood (ML) algorithm via MEGA 7 [38]. Gamma-distributed evolutionary rates between selected groups were further calculated via the maximum parsimony method by the MEGA 7 package [38].

Phylogenetic analysis of S proteins of SARS-CoV-2 genotypes

A distance-based phylogenetic tree based on S protein consensus sequences was constructed to infer the phylogenetic relationship of S protein sequences from the virus genotypes (see Materials and Methods). Individual S protein consensus sequences were first constructed using multiple sequence alignment via the ClustalW approach implemented in MEGA 7. The distance-based neighbor-joining (NJ) phylogenetic tree was then constructed based on nucleotide sequences in MEGA 7 using bootstrap values of 1000 replicates and a 70% threshold score [38].

SNP prediction and imposed selective pressure at S protein sequence level

Variant discovery/SNP prediction of S protein sequences was performed through detection of nucleotides that differed from the S protein of reference sequence (NC_045512.2) using CLC Genomics Workbench software. The detected SNPs were then compared with the original S protein structure with common mutations from GISAID databases [37]. The direction and magnitude of natural selection acting on the S protein gene interpreted via computing the confidence estimation for the nonsynonymous and synonymous nucleotide substitution rates ($dN/dS = \omega$) and degree of selective constraints imposed on S protein via the bootstrap method (1000 replicates) and Tamura–Nei model using MEGA version 7 [38]. *P*-value < 0.05 was significant. The dN/dS nucleotide changes of $dN/dS < 1$, $dN/dS = 1$ and $dN/dS > 1$ were regarded as purifying (negative), neutrality and positive selection, respectively.

Results and discussion

GISAID clades of SARS-CoV-2 complete genomes from different host species

As RNA viruses replicate through error-prone RNA polymerase, mutations occur within each copying cycle. These mutations may affect some biological traits of the virus including transmission and virulence [39]. In comparison with DNA viruses, RNA viruses such as SARS-CoV-2 have a higher mutation rate, and thus a greater opportunity for genetic diversity [40]. However, in comparison with other RNA viruses, coronaviruses mutate less due to corrective enzymes that rectify some replication errors, with newly arising mutants determined by natural selection and selective pressures [41]. In addition to mutation by replication errors, variation in the virus can be caused by recombination or host immune system RNA editing [35]. To date, hundreds of homoplastic sites have been detected among the nucleotides of the SARS-CoV-2 genome, most of which are related to cytosine depletion and uracil conversion [35]. Since these mutations are mostly found in positions that are the site of action of RNA-editing enzymes, it seems that alteration of the virus genome by host deaminase enzymes is the primary cause of mutations in SARS-CoV-2 [35]. The role of host immune enzymes in ongoing mutations within the virus reveals the importance of monitoring hotspot points in the viral genome. Therefore, comparing the pattern of mutations in different hosts can determine how the virus evolves or adapts to a particular animal. The analysis of such data is paramount in informing the development of therapeutics, such as host-specific mAbs or the identification of vaccine candidates.

In our study, representative sequences of human and animal hosts from all clades, isolated from various continents, were selected to investigate and compare phylogenetic relationships. The complete genome-based NJ tree was constructed via aligned SARS-CoV-2 sequences collected globally and subjected to the GISAID database [37]. Seven distinct phylogenetic clades (S, L, V, G, GH, GR and GV) of the virus were observed in the NJ tree, with no significant relationship between the clade types and geographical locations (Figure 1), and individual clades could be observed in each continent. Based on SNP and mutation on WGS of the virus, the clades were separated. Each clad represented the same mutations. The distribution of viral variant sequences corresponding to different clades, present in several countries, highlights the uniform distribution of prevalent mutation types, the potential genetic variability of SARS-CoV-2 in different geographic areas as well as the ability of new virus strains to easily spread between different countries via infected humans and/or animals. The clade 'G' of the NJ tree represented the most divergent profiles of the virus, whereas the clades 'GV' and 'GR' appeared to be less diverged and principally contained within European strains (Figure 1). Novel genome sequences of SARS-CoV-2 variants from Africa, Oceania and Asia were scattered across the NJ-based phylogenetic divergence tree without a specific pattern of mass dissemination (Figure 1). The phylogenetic congruences could be concluded between our findings and the phylogenetic tree that was previously constructed by Alouane *et al.* [42], who declared that the distribution of mutants is uniform in all geographical areas. Overall, American and European variants appear to be responsible for most of the virus spread. For instance, variants from Oceania revealed a close evolutionary relationship with the variant genomes from American strains. Moreover, intragenomic clustering among all assessed countries with no clear pattern of geographic distributions represents the impact of migration and globalization on SARS-CoV-2 dissemination during pandemics [42]. Early introduction and rapid spread of genotypes genetically close to the original strain in continents with high infection rates, such as Europe and North America, can be suggested based on the NJ tree that has been constructed in this study using the GISAID database [37].

Phylogenetic and variation analysis of SARS-CoV-2 genotypes at S protein sequence level

The viral origin and mechanisms of the first cross-species spillover event of SARS-CoV-2 remain unclear; however, some researchers have attempted to identify susceptible hosts by analyzing the genome of the virus isolates, presuming that bats were the primary natural reservoir of the virus [43]. Further

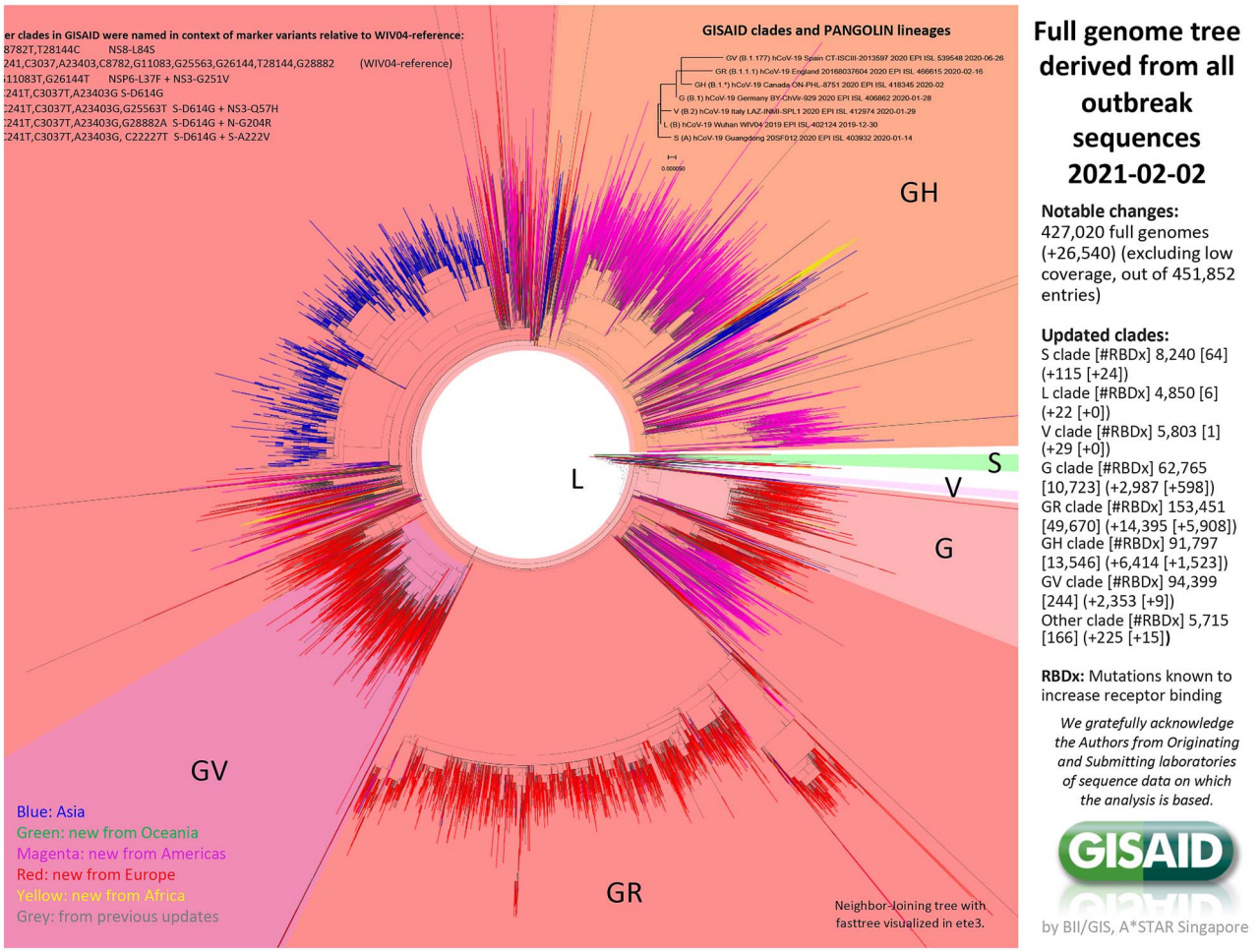

Figure 1. A full reference phylogenetic tree is drawn from 243.779 high-coverage SARS-CoV-2 sequences that have been registered in the GISAID database from all over the world. The tree was constructed using a neighbor-joining (NJ) method. It shows that the SARS-CoV-2 is generally divided into seven different clades (S, L, V, G, GH, GR and GV). The presence of different clades of the virus was possible in all geographical areas. Color codes show different continents.

analysis into the lineage related to pangolin-CoVs sequences has suggested that pangolins facilitated the transmission of the virus by acting as an intermediary host between bats and humans [43]. Considering the key role of S protein infusing the virus to the host ACE2 receptor as well as the presence of homolog of this receptor in different animals, understanding the genetic diversity of S protein may provide pivotal clues to identify possible host sources of SARS-CoV-2, and how this virus has been transmitted between species [16, 17]. Here, the evolutionary distance matrix of the S protein between- and within-host groups using the ML approach was estimated and applied as an input for the reconstruction of a phylogenetic tree. The distance of the S protein variants as a measure of the dissimilarity of virus sequences between- and within-host groups has been represented (Tables 1 and 2). However, the numbers of delineated sequences related to bat and mouse submitted to National Center for Biotechnology Information were lower than the other analyzed species. Thus, the software was not able to calculate the dissimilarity values of viral genotypes within individual groups associated with these two host species (Table 2). The highest evolutionary distances between different host groups were related to pangolin, followed by mink and bat (Table 1). Higher evolutionary distances between pangolins and other species compared with estimated distance values between bats and other hosts species have shed light on some possible clues, revealing that pangolins may not be a simple intermediate host and represent a stronger evolutionary origin than previously thought [43] (Table 1). Evolutionary divergences between humans and companion animals such as cats, monkey and dogs were however negligible, and it might be considered that the virus may have been transmitted from humans to these animals (Table 1). In addition, pangolins and minks had the highest average evolutionary divergence within their own groups (Table 2). The fact that the evolutionary divergence in the mink group was greater than in humans also suggests that the virus could not have been transmitted from humans to minks, and we suggest that the virus-infected minks a long time before humans were infected.

The colored-based phylogenetic tree of S protein from different genotypes, constructed via the NJ method, showed that the pangolin sequences formed a distinct group evolutionarily farther away from others but closer to the mink group (Figure 2). Even though the clades related to the S protein sequences of the viral genotypes associated with humans from different geographical areas showed a close evolutionary relationship with each other, the S protein variants of animal hosts were still scattered within the human groups (Figure 2). This highlights the idea that the virus has been transmitted from humans to

Table 1. Matrix of estimated evolutionary distance between different host groups calculated with maximum likelihood (ML) approach

	Human	Bat	Dog	Cat	Lion	Mink	Mouse	Pangolin
Bat	1.009							
Dog	0.000	1.011						
Cat	0.000	1.010	0.000					
Lion	0.001	1.011	0.001	0.001				
Mink	2.193	3.045	2.196	2.194	2.193			
Mouse	0.001	1.006	0.001	0.001	0.001	2.187		
Pangolin	8.819	8.075	8.834	8.813	8.819	8.427	8.789	
Tiger	0.000	1.010	0.000	0.000	0.001	2.194	0.001	8.814

Table 2. Average evolutionary divergence and gamma distribution between coronavirus populations within each species based on maximum likelihood (ML) algorithm

Animal host	Transition/transversion bias (r)	Gamma distribution	Average evolutionary divergence over sequence pairs within groups
Human	3.56	200	0.000408
Dog	1	11.2382	0.000263
Pangolin	1.24	1.3765	3.5921
Cat	1	37.1736	0.000105
Lion	2 765 874.42	0.05	0.000526
Mink	1.28	0.05	3.2147
Tiger	2	66.0911	0.000175

other species. Consistent with these findings, several studies have shown the presence of human sequences among viral genome isolates from different animals such as dogs, cats and tigers from the same geographical regions, and some infected animals had owners who tested positive for COVID-19 [21–23, 32]. The detection of human sequences in the SARS-CoV-2 genome isolated from other animals such as minks in the Netherland reinforced the theory of virus exchange between humans and other intermediate hosts [16, 34]. This was manifested by the emergence of new cases of SARS-CoV-2-infected farm workers after the COVID-19 mink outbreak [16, 34]. Despite several studies confirming the capability of companion animals to act as hosts, there is yet to be a reported case of pet to human transmission [16, 18]. Therefore, SARS-CoV-2 acts as a multiple species' viral pathogen, with the potential of spillover between human and various animal species, including domestic, exotics or wild animals, and like other zoonotic diseases, vaccination of animals should be considered to mitigate circular transmission of the virus between susceptible hosts.

SNP discovery

SARS-CoV-2 has been evolved in a nondeterministic process with limited selective pressure exerted on the viral genome that has been imposed by viral transmission among the human populations [44, 45]. Random genetic drift plays a significant role in the frequency of SNPs of RNA viruses. Hotspot mutations are driven by positive selection leading to the substitution of particular amino acids and offering an adaptive advantage in specific conditions [46]. The frequency and position of SNPs estimated for the genome structure of SARS-CoV-2 S protein in all sets of virus samples in this study revealed a total of 105 SNPs: 37 in pangolin, 31 in bat, 14 in human, 1 in monkey, 9 in mink, 4 in tiger, 3 in cat and lion, and 2 in dog and mouse groups (Figure 3). Our study showed that more than half of the mutations detected in S protein of genotypes from bats (58%) and human (57%) were related to the RBM region; however, this value was only 8% for pangolins, and most of the mutations in the S protein of the virus strain from this animal were associated with the N-terminal domain region (NTD, 14–305 residues). Similarly, most of the mutations in the mink population were detected in the S1 subunit, especially in the RBM region of viral genotypes (Figure 3). Furthermore, about half of the mutations detected within the S protein sequence of viral genotypes from humans were also detected at the same location of the S protein sequence of bats-related genotypes. The SNPs profile of S protein within human-related genotypes of the virus was more similar to those of bats and minks than pangolins; therefore, taken together with the phylogenetic section, it could be presumed that the virus was first circulating within pangolin populations, was then transmitted to bats and/or minks, and a spillover event led to dissemination within the humans. All of the SNPs in cats were also observed in humans (Figure 3); this supports the previously proposed concept of human to the feline transmission of SARS-CoV-2 [22, 23, 47]. The most common mutation type found in all assessed hosts was the nonsynonymous substitution of the aspartate to glycine (D614G), which was detectable in viral genotypes from minks, humans, monkeys, cats, dogs and tigers, but not in pangolins or bats (Figure 3). The D614G mutant was formerly reported in viral strains from all geographical areas with high frequency, and this mutant replaced the ancestral virus genotype globally and selectively increased the viral loading and infection rate [41, 48, 49]. Three-dimensional modeling of the S protein structure indicated that the aspartate substitution does not alter the conformation of S protein, but reduces the number of hydrogen bonds between S1/S2 subunits among spike promoters, increasing the likelihood of cleavage between the two subunits [41, 50]. Since the cleavage between these two subunits is essential for the virus's entrance to the host cell, any mutation in this area changes the infectivity potential of the virus [41]. Another mutation that alters the furin cleavage site is P681H, which has recently been observed with many other spike mutations in the new English variant, B.1.1.7 lineage [36]. Despite having the same mutation rate as other mutants, this

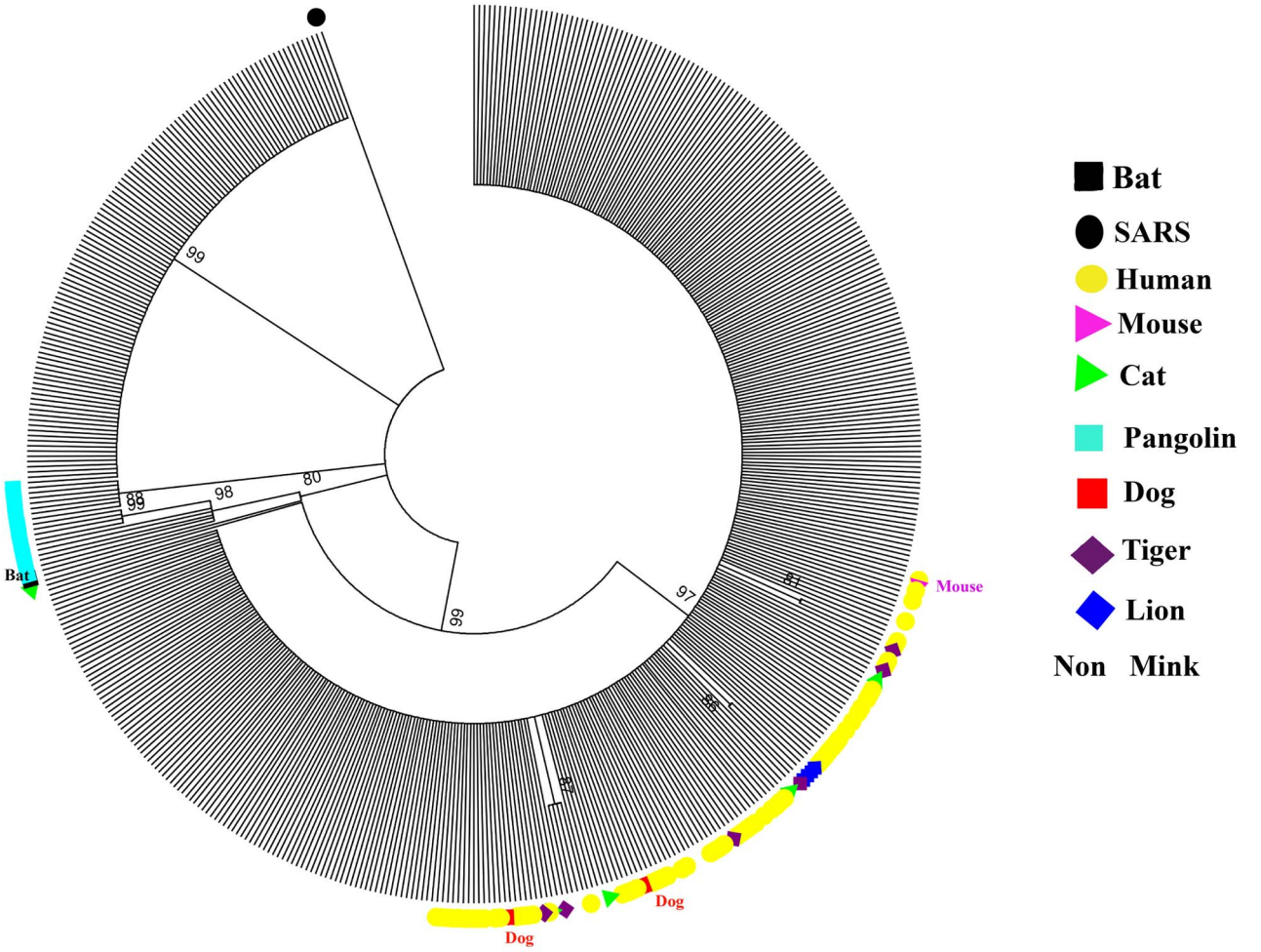

Figure 2. Phylogenetic analysis of SARS CoV-2 S protein sequence. The radial phylogenetic tree was constructed using the neighbor-joining (NJ) algorithm in the MEGA 7 package to illustrate the evolutionary relationship between different host sets of the virus. The colors used in the tree indicate different hosts from which the sequences were sampled.

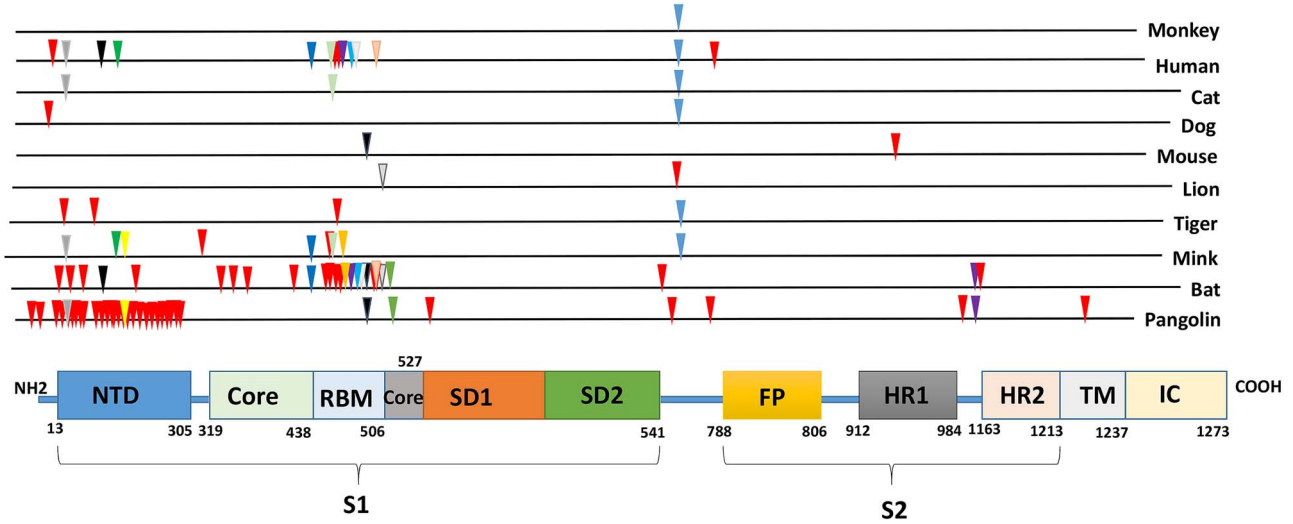

Figure 3. Abundance and position of SNP in SARS-CoV-2 S protein in different hosts. S1 subunit (14–685 residues) comprised NTD (N-terminal domain, 14–305 residues), RBM (receptor-binding motif, 319–541 residues). S2 subunit (686–1273 residues) comprised FP (fusion peptide, 788–806 residues), HR1 (heptapeptide repeat sequence 1, 912–984 residues), HR2 (heptapeptide repeat sequence 2, 1163–1213 residues), TM (transmembrane domain, 1213–1237 residues) and IC (cytoplasmic domain, 1237–1273 residues). Symbols of the same color lined up in a column indicate the same nucleotide position.

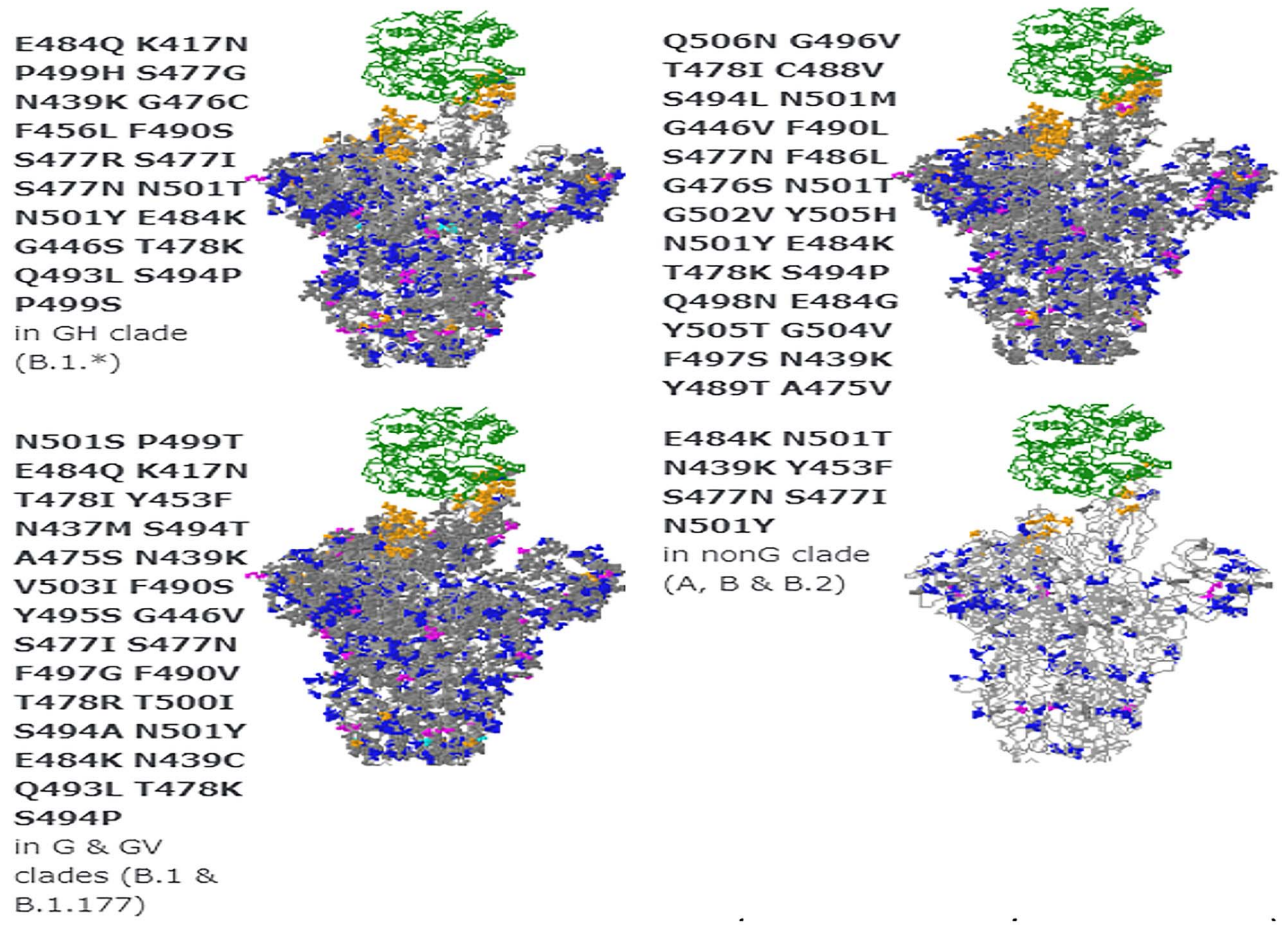

Figure 4. Frequency and location of SNPs on the 3D structure of SARS-CoV-2 S protein among all virus isolates from different clades. Green line: ACE2 human host receptor, gray line: spike glycoprotein trimer.

: Spike glycoprotein variation occurring >100 times.

: Spike glycoprotein variation occurring 100 times or less.

: Spike glycoprotein variation near host receptor with effect history.

: Spike glycoprotein variation near host receptor or other functional annotation.

: Insertion/deletion.

: Spike glycoprotein variation altering potential N-glycosylation sites.

novel lineage has a faster molecular evolutionary rate and a high ratio of nonsynonymous to synonymous changes within the RBD and furin cleavage site [36]. Other significant mutations within the B.1.1.7 lineage are the removal of two amino acids at the 69–70 position in the NTD and N501Y within RBM in the S protein [51–53]. N501Y and P681H mutations in the new English variant are biologically significant in increasing spike-binding affinity to the host receptor in N501Y mutant or induced changing spike conformation from prefusion to fusion state in P681H mutant [54]. These mutations had been previously reported individually, but their accumulation together with other sets of mutations such as ΔH69/ΔV70 in the S protein produced synergistic effects on virus infectivity, leading to the emergence of a novel lineage [51, 54]. It was shown that N501Y appeared after transmission of the virus to mice for passaging in vaccine production, so it occurs as part of the virus adaptation process in the host [54]. In addition, N501Y had been detected in humans related to ΔH69/ΔV70 in the mink outbreak in the Netherlands before the emergence of the new lineage in England. In our study, deletion at position 69 (Δ69) was detected in pangolins, minks and cats, whereas in humans, this mutation was observed in combination with N501Y (Figure 3). Antibody production against the NTD of the spike has been demonstrated; therefore, mutants with genetic changes in the NTD such as ΔH69/ΔV70 were not recognized by neutralizing antibodies specific to this region due to changes in the orientation of the protruding loops [51, 52, 55]. Selective pressure of the host immune system can cause mutations in NTD. Therefore, mutations in this area should not be underestimated, and animals with ΔH69 such as pangolins, minks and cats should be constantly monitored (Figure 3). This study and previous reports confirm that ΔH69/V70 is associated with other mutations in the RBM region such as N439K, which is observed in viral genotypes from bats, minks and humans (Figure 3) [56]. *In vivo* and *in vitro* evidence show that the mutants with N439K in the RBM region have a higher ability to escape from antibody-mediated immunity [51, 56]. In minks, ΔH69/ΔV70 in the NTD has been associated with some other mutations in the RBM region (Y453F and F486L) in the outbreak in the Netherlands. Evidence indicates that the RBM region with Y453F and F486L mutations enabled the virus to elude

neutralizing antibodies [51, 56]. In this study, Y453F was observed in viral genotypes from a cat, mink and human, and F486L mutant was also detected in viral genotypes from bats and minks (Figure 3). Data recorded on the GISAID site regarding the location of the amino acid changes in the schematic map of SARS-CoV-2 S protein has been presented in Figure 4. The most observed SNP mutations in SARS-CoV-2 S protein accumulated in the RBM region (Figure 4). This was consistent with our SNP analysis, revealing that the majority of the SNPs in S protein of SARS-CoV-2 genotypes isolated from humans were mostly accumulated in the RBM area of this protein (Figure 3). Therefore, the evolutionary analysis shows that variants arising from the mutations in the RBM region have been more preserved compared with the other variants during the viral evolution and have an important role in the transmission and pathogenicity of the virus [15, 56].

Conclusion

Apart from the question of whether adaptive selections will further develop in the viral populations in different hosts, we can now conclude that the viral strains currently circulating in humans consist of a rather homogeneous population. We can therefore be hopeful that the current genetic variability of SARS-CoV-2 will not affect the efficacy of S protein-based universal vaccine candidates. However, possible circulation of the virus among different host species may result in a buildup of variants that can evade current vaccines. On the other hand, the adaptation of SARS-CoV-2 to different hosts causes mutations in different parts of the spike protein. The accumulation of such mutations can increase the risk of the emergence of dangerous variants that may be resistant to general vaccines or mAbs. Therefore, there must be an urgent drive to vaccinate and monitor not only human beings but also other susceptible host species, particularly companion and commercial animals. Constant monitoring of genetic changes in the S protein of the virus in different hosts would help us to design host-specific vaccines or antibodies as an important part of a multi-faceted strategy for containment of the virus pandemic [57].

Key Points

- This study investigated SNP variation of SARS-CoV-2 spike protein in human and animal hosts.
- We showed that the circulation of the virus among different host species may result in a buildup of new variants that can evade current vaccines.
- We investigated the virus population in their hosts and showed the high rate of variation in mink populations implies a prior evolution and the virus-infected minks before the human infection.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

AG was supported by a fellowship from Shiraz University and Iran National Foundation of Elites.

References

1. Lv M, Luo X, Estill J, *et al*. Coronavirus disease (COVID-19): a scoping review. *Eurosurveillance* 2020;**25**:2000125.
2. Yang X, Yu Y, Xu J, *et al*. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *Lancet Respir Med* 2020;**8**:475–81.
3. WHO. Weekly Epidemiological Update 19 – 23 March 2021. *World Health Organization* [Online], 2021. https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---23-march-2021.
4. Pardi N, Weissman D. Development of vaccines and antivirals for combating viral pandemics. *Nat Biomed Eng* 2020;**4**:1128–33.
5. Gao Y, Yan L, Huang Y, *et al*. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* 2020;**368**:779–82.
6. Wu C, Liu Y, Yang Y, *et al*. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharm Sin B* 2020;**10**:766–88.
7. Snijder EJ, Decroly E, Ziebuhr J. The nonstructural proteins directing coronavirus RNA synthesis and processing. *Adv Virus Res* 2016;**96**:59–126.
8. Lu R, Zhao X, Li J, *et al*. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 2020;**395**:565–74.
9. Chan JF-W, Kok K-H, Zhu Z, *et al*. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg Microbes Infect* 2020;**9**:221–36.
10. Michel CJ, Mayer C, Poch O, *et al*. Characterization of accessory genes in coronavirus genomes. 2020;**17**:1–13.
11. Li F. Structure, function, and evolution of coronavirus spike proteins. *Annu Rev Virol* 2016;**3**:237–61.
12. Wrapp D, Wang N, Corbett KS, *et al*. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 2020;**367**:1260–3.
13. Huang Y, Yang C, Xu X, *et al*. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol Sin* 2020;**41**:1141–9.
14. Fehr AR, Perlman S. Coronaviruses: an overview of their replication and pathogenesis. In: Maier HJ, Bickerton E, Britton P, (eds). *Coronaviruses: Methods and Protocols, Methods in Molecular Biology*. New York: Springer, 2015; **1282**:1–23.
15. Becerra-Flores M, Cardozo T. SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. *Int J Clin Pract* 2020;**74**:e13525.
16. Munnink BBO, Sikkema RS, Nieuwenhuijse DF, *et al*. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* 2020;**371**:172–77.
17. Lam TT, Jia N, Zhang Y-W, *et al*. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* 2020;**583**:282–5.
18. Shi J, Wen Z, Zhong G, *et al*. Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS–coronavirus 2. *Science* 2020;**368**:1016–20.
19. Halfmann PJ, Hatta M, Chiba S, *et al*. Transmission of SARS-CoV-2 in domestic cats. *N Engl J Med* 2020;**383**:592–4.
20. Munster VJ, Feldmann F, Williamson BN, *et al*. Respiratory disease in Rhesus macaques inoculated with SARS-CoV-2. *Nature* 2020;**585**:268–72.

21. Sit THC, Brackman CJ, Ip SM, *et al*. Infection of dogs with SARS-CoV-2. *Nature* 2020;**586**:776–8.
22. Sailleau C, Dumarest M, Vanhomwegen J, *et al*. First detection and genome sequencing of SARS-CoV-2 in an infected cat in France. *Transbound Emerg Dis* 2020;**67**:2324–8.
23. Gollakner R, Capua I. Is COVID-19 the first pandemic that evolves into a panzootic? *Vet Ital* 2020;**56**:11–2.
24. Lam SD, Bordin N, Waman VP, *et al*. SARS-CoV-2 spike protein predicted to form stable complexes with host receptor protein orthologues from mammals. *Scientific Reports* 2020;**10**:1–14.
25. Zhou P, Yang X-L, Wang X-G, *et al*. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020;**579**:270–3.
26. Boni MF, Lemey P, Jiang X, *et al*. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol* 2020;**5**:1408–1717.
27. Liu P, Chen W, Chen J-P. Viral metagenomics revealed Sendai virus and coronavirus infection of Malayan pangolins (Manis javanica). *Viruses* 2019;**11**:979.
28. Lau SKP, Luk HKH, Wong ACP, *et al*. Possible bat origin of severe acute respiratory syndrome coronavirus 2. *Emerg Infect Dis* 2020;**26**:1542.
29. OIE. COVID-19 Portal: Events in Animals. *World Organisation for Animal Health* [Online], 2020. www.oie.int/en/scientific-expertise/specific-information-and-recommendations/questionsand-answers-on-2019novel-coronavirus/events-inanimals/
30. ECDC. Detection of New SARS-CoV-2 Variants Related to Mink. *European Centre for Disease Prevention and Control* [Online], 2020. https://www.ecdc.europa.eu/sites/default/files/documents/RRA-SARS-CoV-2-in-mink-12-nov-2020.pdf
31. Munnink BBO, Sikkema RS, Nieuwenhuijse DF, *et al*. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* 2021;**371**:172–7.
32. Hossain MG, Akter S, Saha S. SARS-CoV-2 host diversity: an update of natural infections and experimental evidences. *J Microbiol Immunol Infect* 2020. doi: https://doi.org/10.1016/j.jmii.2020.06.006.
33. Welkers MRA, Han AX, Reusken CBEM, *et al*. Possible host-adaptation of SARS-CoV-2 due to improved ACE2 receptor binding in mink. *Virus Evol* 2021;**7**:veaa094.
34. Oreshkova N, Molenaar RJ, Vreman S, *et al*. SARS-CoV-2 infection in farmed minks, the Netherlands, April and May 2020. *Eurosurveillance* 2020;**25**:2001005.
35. Van Dorp L, Richard D, Tan CCS, *et al*. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nat. Commun* 2020;**11**:1–8.
36. Rambaut A, Loman N, Pybus O, *et al*. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *Genom Epidemiol* 2020. https://virological.org/t/preliminary-genomiccharacterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-ofspike-mutations/563 (accessed Dec 21, 2020).
37. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* 2017;**1**:33–46.
38. Kumar S, Stecher G, Tamura K, *et al*. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* 2016;**33**:1870–4.
39. Holmes EC. *The Evolution and Emergence of RNA Viruses*. New York, NY, USA: Oxford University Press, 2009; E01031–17.
40. Peck KM, Lauring AS. Complexities of viral mutation rates. *J Virol* 2018;**92**:e01031–17.
41. Lauring AS, Hodcroft EB. Genetic variants of SARS-CoV-2—what do they mean? *JAMA* 2021;**325**:529–31.
42. Alouane T, Laamarti M, Essabbar A, *et al*. Genomic diversity and hotspot mutations in 30,983 SARS-CoV-2 genomes: moving toward a universal vaccine for the "confined virus"? *Pathogens* 2020;**9**:829.
43. Dong R, Pei S, Yin C, *et al*. Analysis of the hosts and transmission paths of SARS-CoV-2 in the COVID-19 outbreak. *Genes (Basel)* 2020;**11**:637.
44. Koyama T, Weeraratne D, Snowdon JL, *et al*. Emergence of drift variants that may affect COVID-19 vaccine development and antibody treatment. *Pathogens* 2020;**9**:324.
45. Lai A, Bergna A, Caucci S, *et al*. Molecular tracing of SARS-CoV-2 in Italy in the first three months of the epidemic. *Viruses* 2020;**12**:798.
46. Chattopadhyay S, Weissman SJ, Minin VN, *et al*. High frequency of hotspot mutations in core genes of Escherichia coli due to short-term positive selection. *Proc Natl Acad Sci* 2009;**106**:12412–7.
47. Garigliany M, Van Laere A-S, Clercx C, *et al*. SARS-CoV-2 natural transmission from human to cat, Belgium, March 2020. *Emerg Infect Dis* 2020;**26**:3069.
48. Callaway E. The coronavirus is mutating-does it matter? *Nature* 2020;**585**:174–7.
49. Korber B, Fischer WM, Gnanakaran S, *et al*. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* 2020;**182**:812–27.
50. Yurkovetskiy L, Wang X, Pascal KE, *et al*. Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell* 2020;**183**:739–51.
51. Kemp SA, Datir RP, Collier DA, *et al*. Recurrent emergence and transmission of a SARS-CoV-2 spike deletion ΔH69/V70. *bioRxiv* 2020.
52. McCarthy KR, Rennick LJ, Nambulli S, *et al*. Natural deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *bioRxiv* 2020.
53. Starr TN, Greaney AJ, Hilton SK, *et al*. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* 2020;**182**:1295–310.
54. Gu H, Chen Q, Yang G, *et al*. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science* 2020;**369**:1603–7.
55. Chi X, Yan R, Zhang J, *et al*. A neutralizing human antibody binds to the N-terminal domain of the spike protein of SARS-CoV-2. *Science* 2020;**369**:650–5.
56. Thomson EC, Rosen LE, Shepherd JG, *et al*. The circulating SARS-CoV-2 spike variant N439K maintains fitness while evading antibody-mediated immunity. *bioRxiv* 2020.
57. Uludağ H, Parent K, Aliabadi HM, *et al*. Prospects for RNAi therapy of COVID-19. *Front Bioeng Biotechnol* 2020;**8**:916.