scientific reports

OPEN

Study on sentinel hosts for surveillance of future COVID-19-like outbreaks

Yanjiao Li[1,2,7], Jingjing Hu[1,2,7], Jingjing Hou[3,7], Shuiping Lu[1,2,7], Jiasheng Xiong[4], Yuxi Wang[1,2], Zhong Sun[5], Weijie Chen[1,2], Yue Pan[1,2], Karuppiah Thilakavathy[5], Yi Feng[6], Qingwu Jiang[1,2], Weibing Wang[1,2]✉ & Chenglong Xiong[1,2]✉

The spread of SARS-CoV-2 to animals has the potential to evolve independently. In this study, we distinguished several sentinel animal species and genera for monitoring the re-emergence of COVID-19 or the new outbreak of COVID-19-like disease. We analyzed SARS-CoV-2 genomic data from human and nonhuman mammals in the taxonomic hierarchies of species, genus, family and order of their host. We find that SARS-CoV-2 carried by domestic dog (*Canis lupus familiaris*), domestic cat (*Felis catus*), mink (*Neovison vison*), and white-tailed deer (*Odocoileus virginianus*) cluster closely to human-origin viruses and show no differences in the majority of amino acids, but have the most positively selected sites and should be monitored to prevent the re-emergence of COVID-19 caused by novel variants of SARS-CoV-2. Viruses from the genera *Panthera* (especially lion (*Panthera leo*)), *Manis* and *Rhinolophus* differ significantly from human-origin viruses, and long-term surveillance should be undertaken to prevent the future COVID-19-like outbreaks. Investigation of the variation dynamics of sites 142, 501, 655, 681 and 950 within the S protein may be necessary to predict the novel animal SARS-CoV-2 variants.

Keywords SARS-CoV-2, Sentinel host, Spike protein, Phylogenetic tree, Amino acid polymorphism, Positively selected site

The COVID-19 pandemic, which began in 2019, has had a significant impact on the global economy and public health. As of 30 July 2023, more than 768 million confirmed cases and more than 6.9 million deaths had been reported worldwide[1]. The pathogen, SARS-CoV-2, is a single, positive-stranded, non-segmented RNA virus, and the viral genome can encode at least 29 proteins, of which the spike (S) protein is a critical structural protein involved in recognition and fusion with the host cell receptor angiotensin-converting enzyme 2 (ACE2)[2,3]. The S protein is composed of two functional regions, S1 and S2. The S1 region contains two crucial structural domains, the N-terminal domain (NTD) and the receptor-binding domain (RBD), which interfere with the immune response of the host cell and recognize and bind to host-associated receptors, respectively, while the S2 region mediates the fusion of the virus with the cell membrane[4–6]. It has been demonstrated that the high variability of the RBD and NTD relates to the virus' immune evasion and host selection[7].

Coronaviruses have a taxonomically and geographically broad host range and exhibit genetic diversity and frequent gene recombination. Mammals are important potential hosts for cross-species transmission of novel coronaviruses. SARS-CoV-2 is widely believed to have originated in bats[8–10], and previous studies have suggested that pangolins were likely to be its intermediate host due to their high amino acid sequence similarity[11–13]. However, later studies found that their nucleotide sequence was clearly different, so the conclusion that pangolins are intermediate hosts was not conclusive[14]. Since the outbreak of the COVID-19 pandemic, numerous infections have been observed in both domestic and wild mammals. Various mammals such as cats, dogs, white-tailed deer, lions and tigers can be infected with SARS-CoV-2 under natural conditions and become

[1]Department of Epidemiology, School of Public Health, Fudan University, Shanghai 200433, China. [2]Key Laboratory of Public Health Safety, Ministry of Education, School of Public Health, Fudan University, Shanghai 200433, China. [3]Department of Pulmonary and Critical Care Medicine, Tongji Hospital, School of Medicine, Tongji University, Shanghai 200065, China. [4]Division of Emergency Management, Shanghai Municipal Center for Disease Control and Prevention, Shanghai 200336, China. [5]Department of Biomedical Sciences, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang 43400, Malaysia. [6]Department of Integrative Medicine and Neurobiology, School of Basic Medical Sciences, Fudan University, Shanghai 200032, China. [7]Yanjiao Li, Jingjing Hu, Jingjing Hou and Shuiping Lu contributed equally to this work. ✉email: wwb@fudan.edu.cn; xiongchenglong@fudan.edu.cn

potential hosts for the virus[15–17]. Due to the need to adapt quickly to new hosts, transmission between different host species often leads to rapid variation or evolution[18]. Although the epidemic intensity of COVID-19 in humans has recently declined, the spread of SARS-CoV-2 in these domestic and wild animals remains a major threat and they may at any time cause a re-emerging outbreak, especially considering the possible presence of other SARS-like-CoVs in these animals as well as frequent genetic recombination[19], which has raised further concerns about the emergence and re-emergence of coronaviruses. Identifying sentinel hosts for surveillance is important for effective control of infection and transmission of SARS-CoV-2 or other SARS-like-CoVs. In this study, we classified animal hosts into four taxonomic hierarchies of species, genus, family and order, and compared the amino acid sequences of the human SARS-CoV-2 S protein with those of nonhuman mammals to identify critical polymorphic sites and perform phylogenetic and selective pressure analyses. We attempt to identify the sentinel animals that need to be closely monitored and the natural reservoir of SARS-CoV-2 in order to prevent an unexpected pandemic in the future.

Results

Taxonomic hierarchy of SARS-CoV-2's hosts

In addition to humans, a total of 6 orders, 14 families, 20 genera and 31 species of SARS-CoV-2 hosts were included in this study (Fig. 1).

Phylogenetic trees of the S gene of SARS-CoV-2

The phylogenetic tree of the S gene showed a strong tendency for SARS-CoV-2 to cluster according to epidemic area and host species, and a moderate tendency for temporal clustering. The viruses isolated from different hosts in the early phase can be clustered on the phylogenetic tree according to host species. Sequences from domestic dog (*Canis lupus familiaris*), domestic cat (*Felis catus*), mink (*Neovison vison*), and white-tailed deer (*Odocoileus virginianus*) showed a closer relationship to the reference, whereas sequences from the genera *Manis* and *Rhinolophus* were located at the root of the tree (Fig. 2A). Indeed, both before and after the pandemic, viruses from the genera *Manis* and *Rhinolophus* were distant from the reference (Fig. 2B, C, F, G), but for viruses isolated from other hosts, the trend of clustering based on host species was broken on the phylogenetic trees during the later phase of the pandemic, with sequences from each species dispersed and shifting from host species clustering to predominantly endemic area clustering (Fig. 2D, E). Sequences from the genus of *Panthera* were also distant from the reference (Fig. 2A, D, E). Unlike most other hosts, whose viruses tended to cluster separately during the early phase and then mix, the lion (*Panthera leo*) sequences consistently formed a distinct cluster both early and late phases of the pandemic (Fig. 2A, D, E). It is not easy to determine whether the lion-derived virus originated in humans during the pandemic or had circulated previously. It can therefore be confirmed that, with the exception of SARS-CoV-2 isolated from pangolins, bats and lions, the viruses isolated from other hosts all originated from human epidemic spillover infections.

Manis and *Rhinolophus* are two host genera worth investigating. On the pre-2020 phylogenetic trees, compared with the *Manis* genus sequences, the *Rhinolophus* genus sequences are closer to the human reference sequence, and the closest is the bat coronavirus RaTG13 from a Chinese horseshoe bat (*Rhinolophus affinis*) in Yunnan (Fig. 2F, G). Sequences from the *Manis* and *Rhinolophus* genera were both distant from the human sequences, suggesting that although it cannot be ruled out that the viruses they carry are of human-origin, they are more likely to serve as natural hosts and provide a persistent SARS-like-CoVs reservoir for infection. The identities of them are shown in Supplementary Table S1, S2.

Amino acid polymorphism of SARS-CoV-2's S protein

There were seven hosts in the species hierarchy, i.e., small-clawed otter (*Aonyx cinereus*), lion, tiger (*Panthera tigris*), Roborovski hamster (*Phodopus roborovskii*), golden hamster (*Mesocricetus auratus*), European mink (*Mustela lutreola*), ferret (*Mustela furo*), with significant difference sites, of which lion had the largest number of statistically significant differential majority of amino acids (9), with seven of them located in the key NTD or RBD (Fig. 3). In the hierarchy of genus, *Panthera*, *Mesocricetus*, *Mustela*, *Aonyx* and *Phodopus* had statistically significant differential majority of amino acids in the key NTD or RBD. A total of 15 statistically significant differential amino acid sites were identified across species, genus, family and order, of which eight were located in the NTD (19, 69, 95, 142, 144, 156, 157 and 158), four in the RBD (452, 478, 486, 501) and one in the furin protease cleavage site (Table 1). In the host taxonomic hierarchy of species, sites 452, 478, 486, and 501 of the S protein showed lower conservation (Fig. 4). In all host taxonomic hierarchies, RBD is more variable than NTD (Supplementary Fig. S1, S2). Detailed results of the differential majority of amino acids in four host taxonomic hierarchies are given in Supplementary Tables S3-S6.

Viruses isolated from domestic dog, domestic cat, mink, and white-tailed deer showed typical characteristics of those from humans in terms of amino acid polymorphism. The viral S protein from these animals showed no differences in the majority of amino acid compared with the full-length reference. The protein from the lion exhibited again much more divergent sites from the human-derived sequences, only next to those from the genera of *Manis* and *Rhinolophus* (Table 2).

Positively selected sites within S protein of SARS-CoV-2

FUBAR detected more positively selected sites than SLAC. In the species hierarchy, three host animals, mink, domestic dog, and green monkey (*Chlorocebus sabaeus*), had positively selected sites searched by SLAC, with the numbers of 3, 3, and 2, respectively, while FUBAR detected more positively selected sites, with a total of 14 species having positively selected sites. Combining the two methods, the top four species with the highest number of positively selected sites were domestic dog, domestic cat, mink and white-tailed deer with 42, 28,
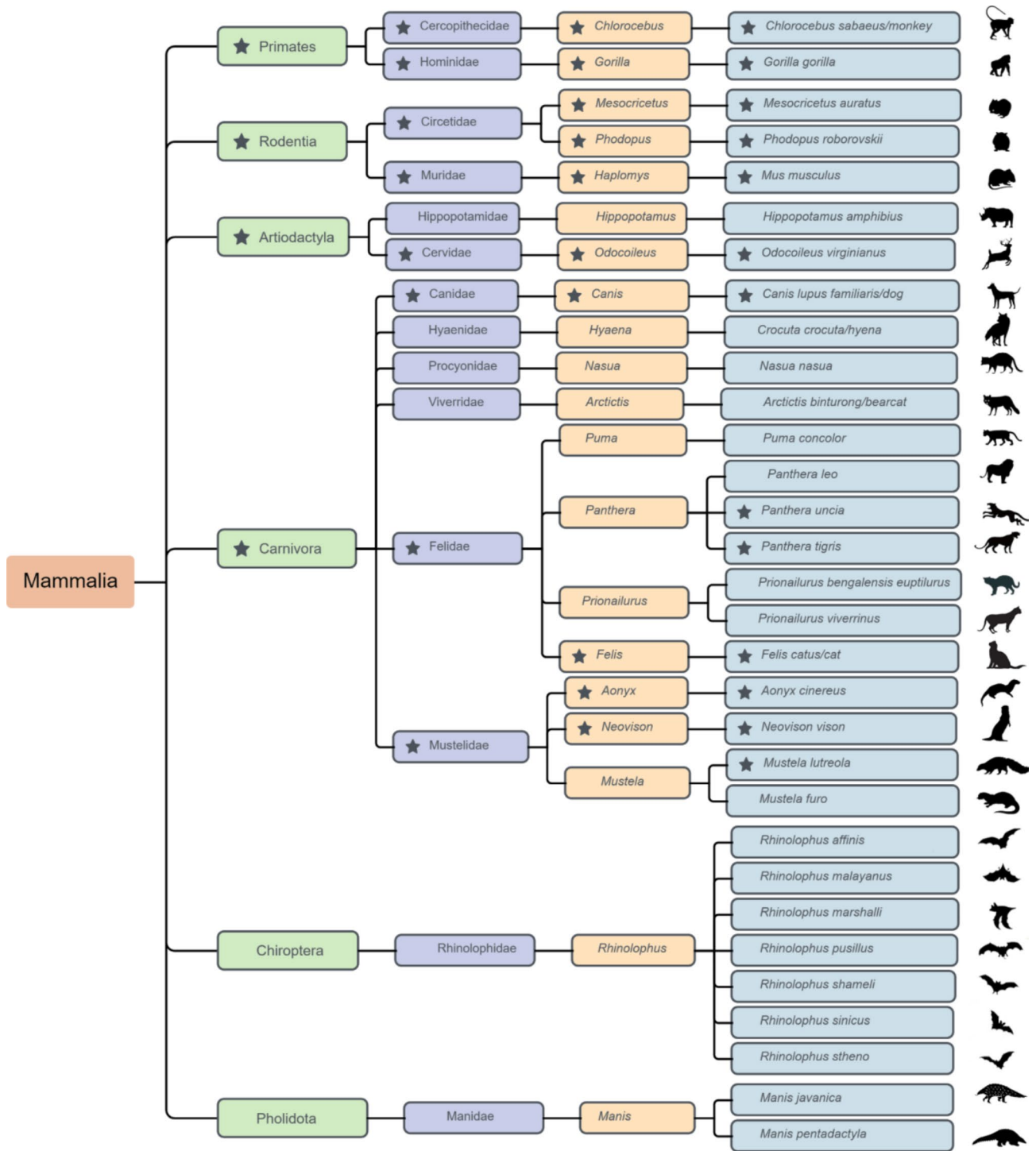
Fig. 1. Nonhuman mammalian hosts of SARS-CoV-2 and their taxonomic hierarchies. All hosts included in this study are mammals, and the taxonomic hierarchies of order, family, genus and species are shown in green, purple, yellow and blue respectively. In total, 6 orders, 14 families, 20 genera and 31 species were included. The star indicates mammalian hosts from which the majority of amino acids in the RBD of SARS-CoV-2 did not differ from those derived from humans.

21 and 18, respectively (Fig. 5). These results suggest that although SARS-CoV-2 in these animals originated from human spillover during the pandemic, it retains significant mutational potential during its circulation and may give rise to novel variants in the future. Lion, closely following the four species above, showed a total of 8 positively selected sites, 6 of which were in critical regions such as RBD and NTD using the FUBAR method. In the genus hierarchy, the top five genera with the highest number of positively selected sites were *Canis*, *Felis*, *Neovison*, *Odocoileus*, and *Panthera*. Among these, *Canis*, *Neovison*, and *Panthera* had significant positively
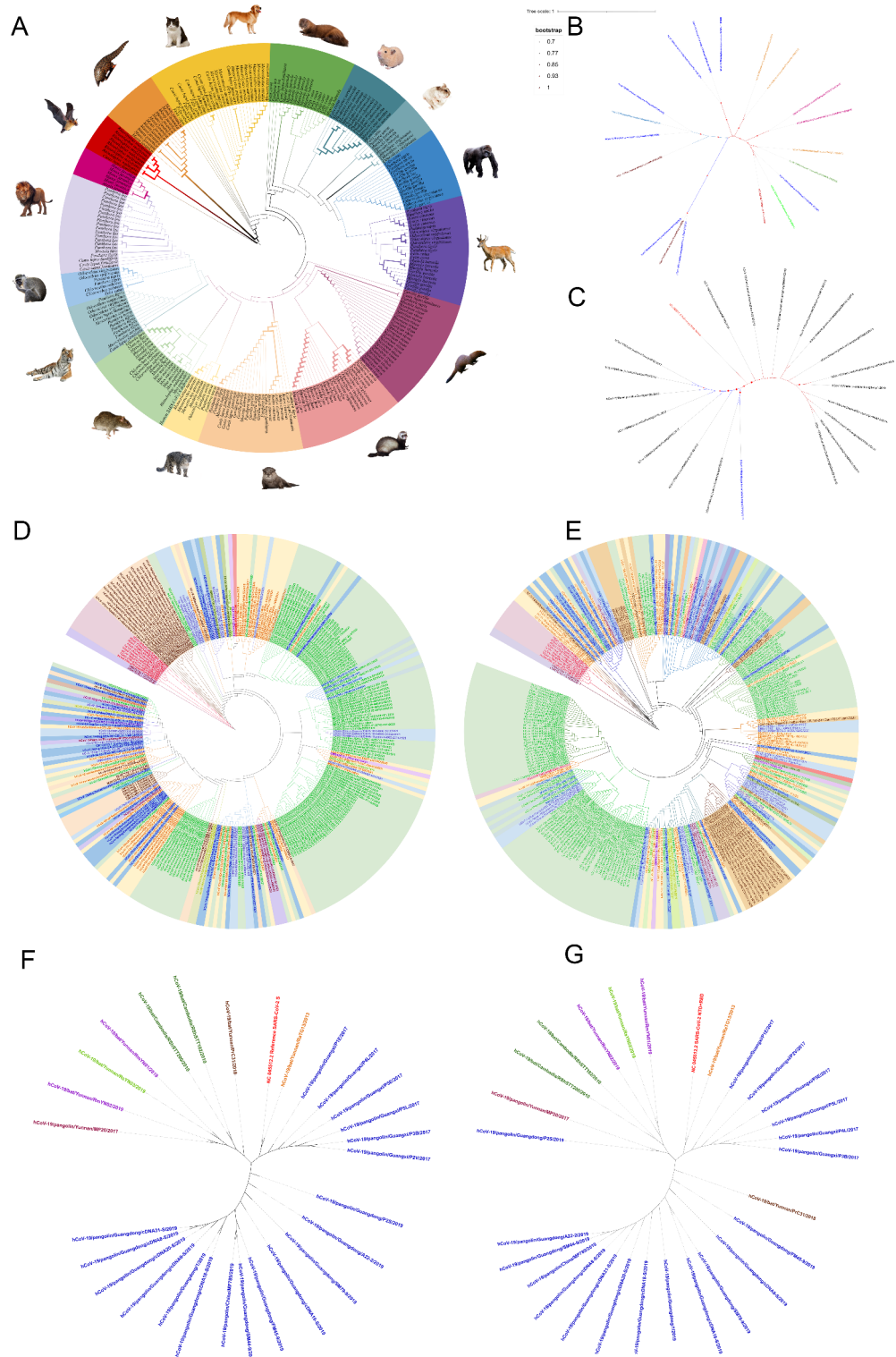
Fig. 2. Phylogenetic trees of SARS-CoV-2 spike (S) protein genes. No more than 20 earliest sequences from each species were retained to construct the overall phylogenetic tree (**A**). Due to the large gap between sequences from pangolin or bat and human sources, unrooted trees were constructed separately: pangolin-derived (**B**) and bat-derived (**C**). The phylogenetic trees of S protein sequences with similarity less than 0.9965 were constructed according to the time of sequence collection before 2020 and after 2020: phylogenetic tree of the complete S protein sequence after 2020 (**D**), phylogenetic tree of the NTD and RBD regions after 2020 (**E**), phylogenetic tree of the complete S protein sequence before 2020 (**F**), phylogenetic tree of the NTD and RBD regions before 2020 (**G**).
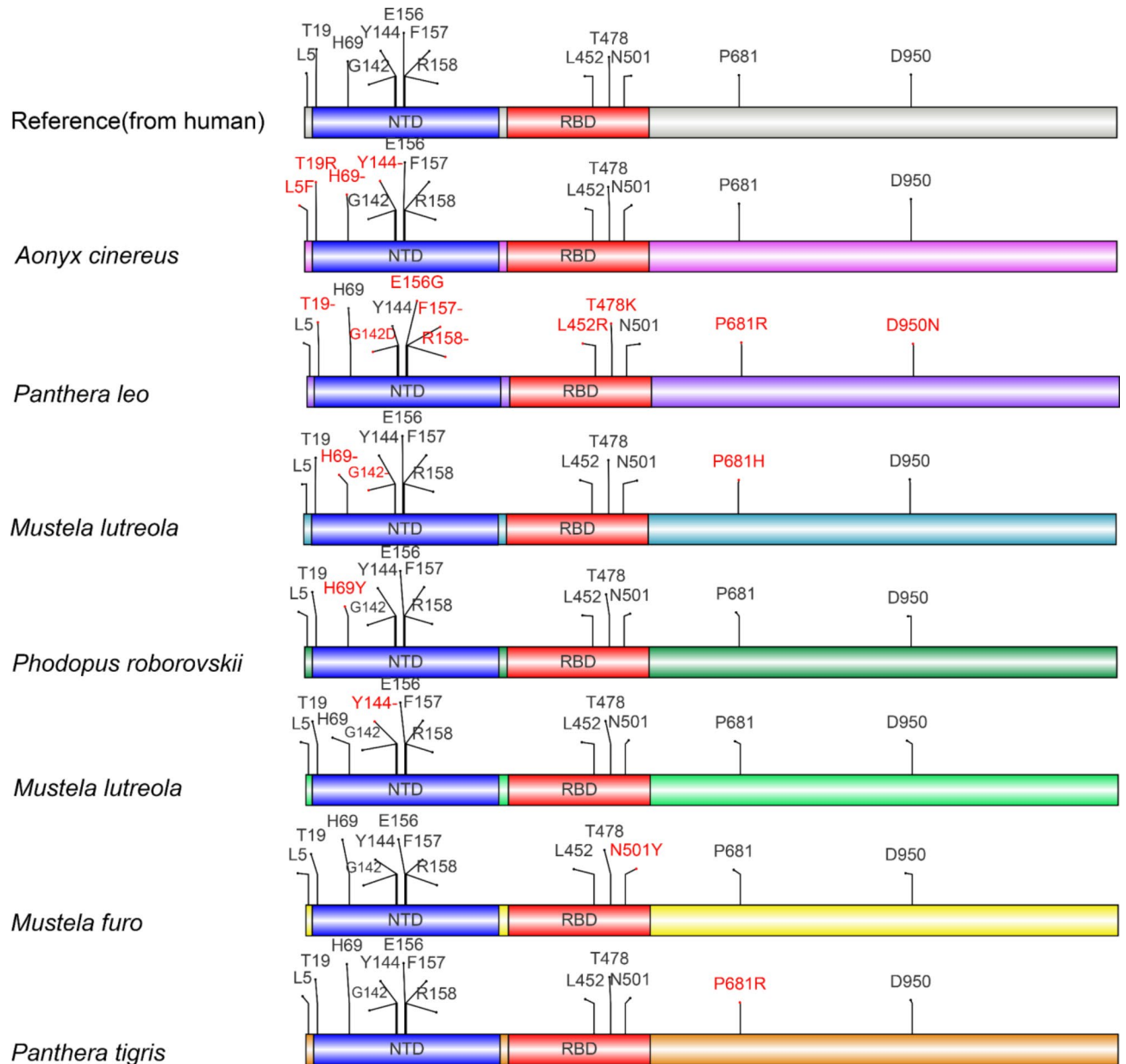
Fig. 3. Graph of SARS-CoV-2 S protein amino acid polymorphism in species of the host taxonomic hierarchy. The red font indicates that the majority of amino acid residues in the SARS-CoV-2 S protein from that species differed from those in humans at those positions, e.g. L5F indicates that the majority of amino acids in the SARS-CoV-2 S protein sequence from the small-clawed otter (*Aonyx cinereus*) differed at 5 positions from the human-derived SARS-CoV-2 sequence, which was varied from L (leucine) to F (phenylalanine). Lion (*Panthera leo*) had the largest number of statistically significant different majority amino acid residues (9), seven of which were in the NTD or RBD.

selected sites by both methods, including key sites such as 501 and 681. In particular, the genus *Panthera* not only had positively selected sites by both methods, but also had statistically significant differential majority amino acid residues in key regions such as the RBD, NTD and the furin protease cleavage site. In the taxonomic hierarchies of species, genus and order, and by the FUBAR method, the five positively selected sites with the highest frequency are 681, 501, 142, 222, 655; 681, 501, 142, 655, 950; and 681, 501, 655, 478, 18 respectively (Supplementary Table S7). When combined across three taxonomic hierarchies, the five positively selected sites with the highest frequency are sites 681, 501, 142, 655 and 950. Tracking the variation dynamics of these sites may be necessary to predict the novel animal-origin variants of SARS-CoV-2.

Discussion

Since the beginning of the COVID-19 pandemic, the range of infected animals has expanded to an unprecedented extent. There is a significant risk that they could transmit SARS-CoV-2 back to humans. Identification of sentinel

	Site	Variations*	Species	Genera	Families	Orders
other	5	L5F	*Aonyx cinereus*	*Aonyx*		
NTD	19	T19R	*Panthera leo*	*Panthera*		Primates
NTD	69	H69Y	*Phodopus roborovskii*	*Phodopus*		
		H69-	*Mesocricetus auratus; Aonyx cinereus*	*Mesocricetus; Aonyx*	Circetidae	Rodentia
NTD	95	T95I			Circetidae	Rodentia
NTD	142	G142D	*Panthera leo*	*Panthera; Mustela*	Circetidae	Primates
		G142-	*Mesocricetus auratus*	*Mesocricetus*		
NTD	144	Y144-	*Mustela lutreola; Aonyx cinereus*	*Mustela; Aonyx*		
NTD	156	E156G	*Panthera leo*	*Panthera*		
NTD	157	F157-	*Panthera leo*	*Panthera*	Circetidae	
NTD	158	R158-	*Panthera leo*	*Panthera*	Circetidae	
RBD	452	L452R	*Panthera leo*	*Panthera*		
RBD	478	T478K	*Panthera leo*	*Panthera*		
RBD	486	F486L		*Mustela*		
RBD	501	N501Y	*Mustela furo*			
S1/S2 cleavage region	681	P681R	*Panthera tigris; Panthera leo*	*Panthera*	Circetidae	
		P681H	*Mesocricetus auratus*	*Mesocricetus*		
HR1**	950	D950N	*Panthera leo*	*Panthera*		

Table 1. Amino acid polymorphism in different taxonomic hierarchies of the mammalian hosts. Note, * Compared with the reference Wuhan-Hu-1 (NC_045512). ** Heptad repeat-1 (HR1) is a fusogen that drives membrane fusion.

animals for epidemic surveillance is critical to control emerging or re-emerging outbreaks caused by novel SARS-CoV-2 variants or other SARS-like-CoVs. In this study, based on a large-scale bioinformatic analysis of viral genomes from human and nonhuman mammals, we identify domestic dog, domestic cat, mink, and white-tailed deer as well as their respective genera as the sentinel hosts that require special attention to prevent them from transmitting an evolved SARS-CoV-2 variant back to humans and causing a re-emerging outbreak. The genera *Panthera* (especially lion), *Manis* and *Rhinolophus* are proposed to be monitored for potential emerging outbreaks caused by other SARS-like-CoVs. In addition, it is recommended that the variation dynamics of sites 142, 501, 655, 681 and 950 within the S protein be studied to predict the novel variants of SARS-CoV-2 originating from animals.

Animals in close contact with humans, whether as pets or as a source of fur or food protein, have increasing opportunities to share and exchange pathogens with humans. In this study, in the species taxonomic hierarchy, the S protein of SARS-CoV-2 from four animal hosts, including domestic dog, domestic cat, mink and white-tailed deer, showed no or only minor differences from the human-derived reference, both in terms of phylogenetic trees and majority amino acids. It is more likely that they were infected with SARS-CoV-2 from humans. In previous studies, Hale VL et al. and Chandler JC et al. found that white-tailed deer can carry SARS-CoV-2[20,21]. Palmer MV et al. also found that white-tailed deer can exhibit subclinical symptoms after nasal inoculation with SARS-CoV-2 and that humans can be infected through contact with white-tailed deer[22]. Studies in Poland and Hong Kong, China, found SARS-CoV-2 infection in domestic dogs and cats, but it was not clear whether it could be transmitted to humans[15,23,24]. American mink on fur farms in the Netherlands, Denmark, Poland and Utah, USA, have been found to have widespread infections, posing a significant risk of transmission to farm workers and other local wildlife, making them worthy of attention as potential intermediate hosts[25–28]. Additionally, the Spanish study has identified SARS-CoV-2 infection in two wild minks within self-sustaining populations[29]. A plausible explanation for the infection in these animals is that they were sporadically exposed to viruses present in wastewater[30]. This highlights the potential significance of an indirect transmission route, possibly through wastewater, as a source of infection that warrants thorough investigation. Due to a relatively independent host ecosystem and microecosystem, the evolutionary dynamics of these viruses tend to be different from those that persist and circulate in the human population, which can lead to the emergence of novel variants that are more adapted to humans[18]. Selective pressure (usually from the host immune system) plays the dominant role. Under positively selective pressure, in order to evade the immune response of the animal host and establish a persistent infection, the virus exhibits high genetic variation, making it possible for new variation signatures to occur[31,32]. Our study confirmed the presence of most positively selected sites within the S protein of viruses isolated from these hosts as well as their respective genera. This suggests that, over time, human-derived SARS-CoV-2 infecting and circulating in these animals may sooner or later evolve into novel variants that may be better adapted to humans. If re-introduced into humans through cross-species transmission, it is possible to trigger re-emerging pandemics. In terms of SARS-CoV-2 alone, viruses currently circulating in domestic dog, domestic cat, mink and white-tailed deer as well as their respective genera may be more dangerous, as these viruses are subject to continuous selection in their animal hosts, potentially leading to novel variants of SARS-CoV-2. Therefore, monitoring of SARS-CoV-2 carried by these sentinel animals should be emphasized to prevent the emergence of novel variants and back transmission to humans.
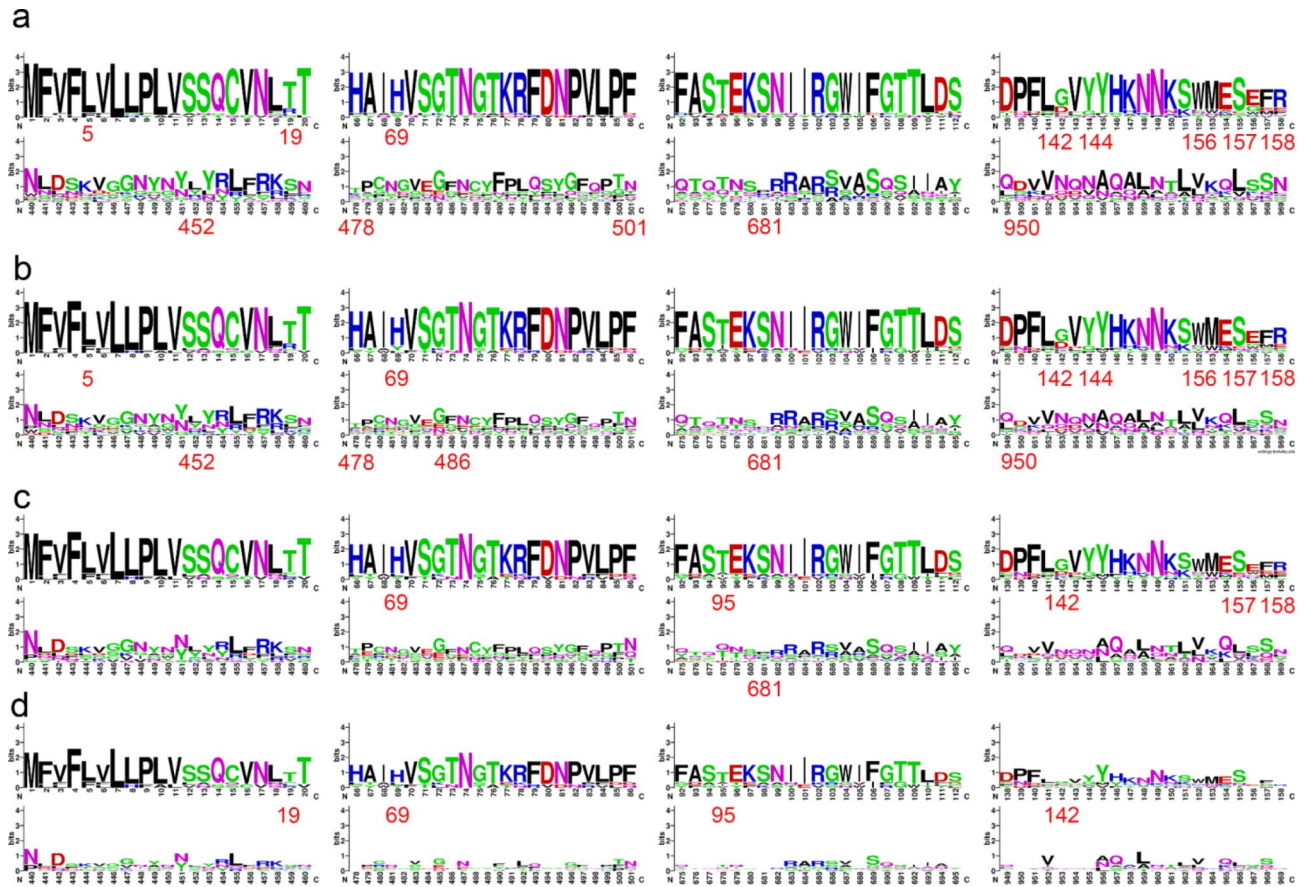
Fig. 4. SeqLogo of amino acid substitutions in SARS-CoV-2 S protein in different taxonomic hierarchies of the mammalian hosts. Sequence logos of conserved majority of amino acid from species (**a**), genera (**b**), families (**c**) and orders (**d**) are shown. Sites where the majority of amino acid of the S protein differ significantly from those of human-acquired viruses in that taxonomic hierarchy are marked in red. The height of each letter in the figure is proportional to the frequency of occurrence of the corresponding amino acid residue at that position, and the letters at each position are arranged in order of conservation from large to small, with the top letters representing conserved sequences.

	No differential site	No statistically significant differential site	Statistically significant differential sites, but not in RBD
13 Species	*Odocoileus virginianus* *Canis lupus familiaris* *Felis catus* *Neovison vison*	*Mus musculus* *Chlorocebus sabaeus* *Gorilla gorilla* *Panthera uncia*	*Panthera tigris* *Mesocricetus auratus* *Mustela lutreola* *Phodopus roborovskii* *Aonyx cinereus*
10 Genera	*Canis* *Felis* *Odocoileus* *Neovison*	*Chlorocebus* *Gorilla* *Haplomys*	*Mesocricetus* *Phodopus* *Aonyx*
8 Families	Cervidae Canidae Felidae Mustelidae	Cercopithecidae Hominidae Muridae	Circetidae
4 Orders	Artiodactyla carnivora		Primates Rodentia

Table 2. Differential majority of amino acid compared with human SARS-CoV-2's S protein in different host taxonomic hierarchies.

Viruses from the *Manis* and *Rhinolophus* genera showed significant differences from the human reference on the phylogenetic trees, suggesting that the possibility of direct transmission from them to humans and vice versa is not substantial. Previous and ongoing studies have shown that the beta-coronaviruses carried by bats and pangolins pose a potential threat to human public health, as they may evolve into new coronaviruses other than SARS-CoV-2 that can infect humans through continuous variation[11,12,33]. In order to monitor the possible
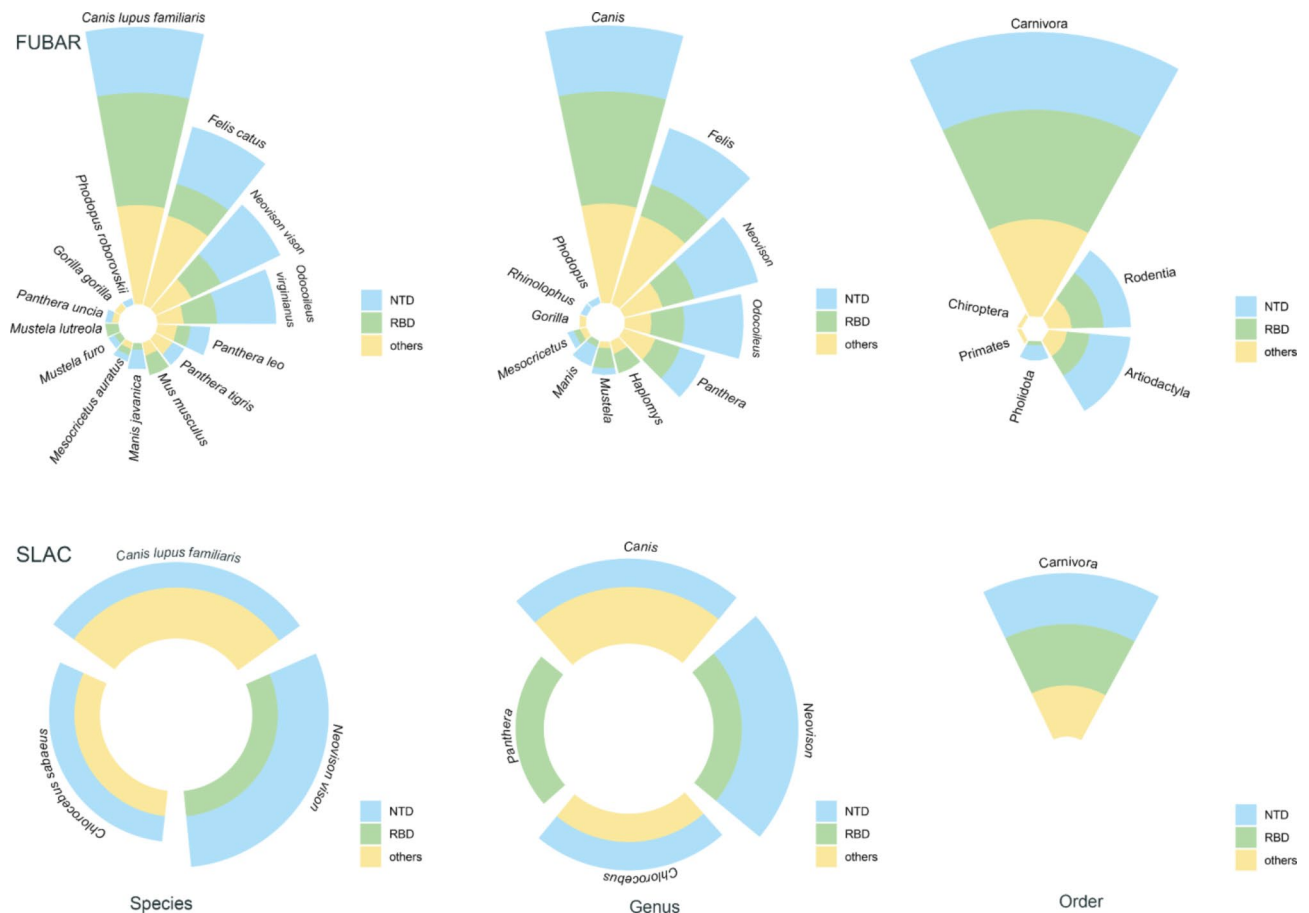
Fig. 5. Graph of the number of positively selected sites in the S protein of SARS-CoV-2 in different mammalian taxonomic hierarchies. The graph shows the number of positively selected sites of different species, genera and orders using the two methods of FUBAR and SLAC. The domestic dog (*Canis lupus familiaris*), domestic cat (*Felis catus*), mink (*Neovison vison*) and white-tailed deer (*Odocoileus virginianus*) have the most positively selected sites using the FUBAR method.

emergence of new SARS-like-CoVs outbreaks in the future, emphasis should be placed on real-time investigation of host animals belonging to the genera *Manis* and *Rhinolophus*. For the same purpose, the genus *Panthera* (especially lions) is particularly worthy of close monitoring, as it not only has the most statistically significant differential majority of amino acids (except for *Manis* and *Rhinolophus* genera), but also has the most positively selected sites (except for *Canis*, *Felis*, *Neovison* and *Odocoileus* genera). Our study showed that all host-derived viruses, except lion, tended to cluster separately during the early phase and mix together afterward. However, the lion-derived sequences consistently formed a distinct cluster both early and late phases of the pandemic (Fig. 2A, D, E). It is not easy to determine whether the lion-derived virus originated from humans during the pandemic or had been circulating previously. In addition, the lion, which has the highest number of statistically significant differential amino acids and a high number of positively selected sites (just after domestic dog, domestic cat, mink and white-tailed deer), clustered close to the genera of *Manis* and *Rhinolophus*. We cautiously raise the question of whether the lion-derived viruses are of human-origin, or whether there is a relatively independent lineage of SARS-CoV-2 within lion that triggered the pandemic. A recent review described positive contact tracing studies confirming possible lion-to-human transmission[34]. Although the lion-transmitted viruses exhibit complexity and uncertainty in both origin and evolutionary trends, the number of positively selected sites also warrants attention to prevent the emergence of new strains of beta-coronaviruses that could be transmitted to humans.

Specific laboratory animals such as mice (*Mus musculus*), green monkeys, western gorillas (*Gorilla gorilla*) and snow leopards (*Panthera uncia*) also need to be monitored. Nonhuman primates are often used as medical research animals to create disease models. According to the World Organization for Animal Health (WOAH) database (https://www.woah.org/en/document/sars-cov-2-in-animals-situation-report-20/), both natural infections with SARS-CoV-2 and spillover events between humans and gorillas have been reported. Once the virus carried by nonhuman primates has mutated into a novel variant, the species barrier that must be crossed to infect humans is minimal.

SARS-CoV-2 appears to be capable of infecting several mammalian species. Detections of this virus in pets, zoo animals, wildlife and livestock have raised questions about zoonotic (animal-to-human) reinfection of humans. To date, however, the COVID-19 pandemic has revealed a global paucity of publicly available

biological samples that are representative of space, time and taxonomic diversity. The shortage is particularly severe in biodiverse and low-income countries. Taxonomically, geographically, and temporally comprehensive surveillance in these countries may not be practical[35,36], but to extrapolate animals from known sentinel host species to the genus, family, and even order of the taxonomic hierarchies for monitoring purposes may be triable. Our study recommends the inclusion of animals belonging to the genera of *Canis*, *Felis*, *Neovison*, *Odocoileus*, *Panthera*, *Mustela*, and *Aonyx* etc. as potential sentinel hosts for surveillance, especially the first five genera. In the host taxonomic hierarchy of the genus, SARS-CoV-2 derived from them has been detected more positively selected sites within the S protein.

Combining the results of amino acid polymorphism and positively selected sites, five sites (142, 501, 655, 681 and 950) in the S protein require special attention, one of which is located in the NTD (142), one in the RBD (501) and one in the furin protease cleavage site (681). The amino acids at these five sites may have a tendency to vary, which could lead to the development of a novel SARS-CoV-2 variant that is better adapted to humans. This is consistent with much research on the relationship between these sites and the infectious properties of SARS-CoV-2[37–39]. In our study, for example, the amino acid residue at cleavage site 681 is differential in three taxonomic hierarchies: family (Circetidae), genera (*Panthera*, *Mesocricetus*) and species (lion, tiger, golden hamster); the residue at site 142 is differential in four taxonomic hierarchies: order (Primates), family (Circetidae), genera (*Panthera*, *Mustela*) and species (lion, golden hamster); the residue at site 950 is differential only in the Panthera genus and lion species; at site 501 is differential only in the ferret species. All of these sites are likely to be critical in determining host tropism[40]. Variations at these sites increase the binding affinity of the S protein to the ACE2 receptor or decrease its affinity to antibodies, leading to immune escape, making the mutant strains more transmissible and potentially increasing their host range[41–43].

Our study highlights the need for a robust One Health-based investigation to monitor susceptible animal hosts. Since the emergence of the COVID-19 epidemic, normal human life and economic activities have been severely disrupted worldwide. It is important to regularly monitor the evolution of SARS-CoV-2 in both human and animal populations and to adopt an approach that sustainably balances and optimizes human, animal and ecosystem health to build a community of human health.

Methods

Data sources and preprocessing

SARS-CoV-2 genomic data were obtained from two databases, the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/) and the Global Initiative of Sharing All Influenza Data (GISAID) EpiCoV (https://gisaid.org/). Nonhuman mammalian viral genomes were downloaded according to host species, while human viral genomes were downloaded separately according to PANGO lineages, with only the earliest 10 genomes downloaded for each PANGO lineage.

The viral genomes were aligned using MAFFT 7.503 with the complete S gene of isolate Wuhan-Hu-1 (NC_045512) as a reference. The aligned genomic sequences were trimmed using MEGA 6.0 software, retaining only the S genes. Due to the large number of human-derived viral genomes, a special quality control procedure was implemented. Any genomic sequence with a total number of 'n' greater than 40 or containing more than three consecutive 'n' was considered to be of sub-optimal sequencing quality and was therefore excluded from subsequent analyses.

Sequence stratifying based on host taxonomic hierarchy

The principle of sequence inclusion is that if the number of sequences in a given taxonomic hierarchy is less than six, it is considered insufficient for basic sampling needs and is therefore not analysed in that hierarchy, but is retained and merged up to the higher level of the taxonomic hierarchies, and so on. The taxonomic hierarchies used in this study were species, genus, family and order, all belonging to the class Mammalia.

Phylogenetic analysis

Phylogenetic trees were constructed using the neighbour-joining method with bootstrap analysis of 1,000 replicates in MEGA 6.0, using the p-distance nucleotide substitution model and selecting pairwise deletion for gaps data treatment.

For ease of presentation, no more than 20 of the earliest sequences from each species were retained to construct the overall phylogenetic tree. Sequences from the widespread genera Manis and Rhinolophus are quite different from those of human-origin, so additional phylogenetic analyses were performed for them.

To investigate potential SARS-like-CoVs (other than SARS-CoV-2) carried by these nonhuman mammals, phylogenetic analysis of the virus from each mammalian species was also performed separately according to time of collection (before and after 2020) and region (NTD and RBD and full-length S gene). Due to the large number of sequences from animal hosts after 2020, CD-HIT 4.8.1 Linux version was used with a threshold of 99.65% to exclude the redundancies.

Amino acid polymorphism analysis

The coding regions of human and nonhuman mammalian S sequences were translated into amino acids using the Megalign module of Lasergene software. The majority of amino acids at each site was counted in the taxonomic hierarchy levels of host species, genera, families and orders, and then compared one by one with that of the human-derived SARS-CoV-2. Chi-square tests were performed using SPSS 19.0 to determine whether there were statistically significant differences in the amino acid majority at each site.

Selective pressure analysis

The aligned S gene sequences were used to explore the selective pressure on the viruses in the taxonomic hierarchies of host species, genera, families and orders. The SLAC (Single-Likelihood Ancestor Counting) and FUBAR (Fast Unconstrained Bayesian Approximation) programs on the Datamonkey Web server (http://www.datamonkey.org/) were executed for this purpose[44].

The methodological flow chart for this study is shown in the Supplementary Fig. S3.

Data availability

Some or all data, models, or code generated or used during the study are available from the corresponding author by request.

References

1. WHO. COVID-19 Weekly Epidemiological Update. (2023).
2. Kadam, S. B., Sukhramani, G. S., Bishnoi, P., Pable, A. A. & Barvkar, V. T. SARS-CoV-2, the pandemic coronavirus: Molecular and structural insights. *J. Basic. Microbiol.* **61**, 180–202. https://doi.org/10.1002/jobm.202000537 (2021).
3. Wang, Q. et al. A unique protease cleavage site predicted in the spike protein of the Novel Pneumonia Coronavirus (2019-nCoV) potentially related to viral transmissibility. *Virol. Sin.* **35**, 337–339. https://doi.org/10.1007/s12250-020-00212-7 (2020).
4. Harvey, W. T. et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* **19**, 409–424. https://doi.org/10.1038/s41579-021-00573-0 (2021).
5. Xia, X. Domains and functions of spike protein in Sars-Cov-2 in the context of Vaccine Design. *Viruses.* **13** https://doi.org/10.3390/v13010109 (2021).
6. Zhang, J., Xiao, T., Cai, Y. & Chen, B. Structure of SARS-CoV-2 spike protein. *Curr. Opin. Virol.* **50**, 173–182. https://doi.org/10.1016/j.coviro.2021.08.010 (2021).
7. Chen, Y. et al. Broadly neutralizing antibodies to SARS-CoV-2 and other human coronaviruses. *Nat. Rev. Immunol.* **23**, 189–199. https://doi.org/10.1038/s41577-022-00784-3 (2023).
8. Xu, X. et al. Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. *Sci. China Life Sci.* **63**, 457–460. https://doi.org/10.1007/s11427-020-1637-5 (2020).
9. Paraskevis, D. et al. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. *Infect. Genet. Evol.* **79**, 104212. https://doi.org/10.1016/j.meegid.2020.104212 (2020).
10. Benvenuto, D. et al. The 2019-new coronavirus epidemic: evidence for virus evolution. *J. Med. Virol.* **92**, 455–459. https://doi.org/10.1002/jmv.25688 (2020).
11. Zhang, T., Wu, Q. & Zhang, Z. Probable pangolin origin of SARS-CoV-2 Associated with the COVID-19 outbreak. *Curr. Biol.* **30**, 1346–1351e1342. https://doi.org/10.1016/j.cub.2020.03.022 (2020).
12. Xiao, K. et al. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature.* **583**, 286–289. https://doi.org/10.1038/s41586-020-2313-x (2020).
13. Li, X. et al. Emergence of SARS-CoV-2 through recombination and strong purifying selection. *bioRxiv.* https://doi.org/10.1101/2020.03.20.000885 (2020).
14. Schindell, B. G., Allardice, M., McBride, J. A. M., Dennehy, B. & Kindrachuk, J. SARS-CoV-2 and the Missing Link of Intermediate hosts in viral emergence - what we can learn from other Betacoronaviruses. *Front. Virol.* **2** https://doi.org/10.3389/fviro.2022.875213 (2022).
15. Sit, T. H. C. et al. Infection of dogs with SARS-CoV-2. *Nature.* **586**, 776–778. https://doi.org/10.1038/s41586-020-2334-5 (2020).
16. McAloose, D. et al. From people to Panthera: natural SARS-CoV-2 infection in Tigers and Lions at the Bronx Zoo. *mBio.* **11** https://doi.org/10.1128/mBio.02220-20 (2020).
17. Gaudreault, N. N. et al. SARS-CoV-2 infection, disease and transmission in domestic cats. *Emerg. Microbes Infect.* **9**, 2322–2332. https://doi.org/10.1080/22221751.2020.1833687 (2020).
18. McBride, D. S. et al. Accelerated evolution of SARS-CoV-2 in free-ranging white-tailed deer. *Nat. Commun.* **14**, 5105. https://doi.org/10.1038/s41467-023-40706-y (2023).
19. Su, S. et al. Epidemiology, genetic recombination, and Pathogenesis of coronaviruses. *Trends Microbiol.* **24**, 490–502. https://doi.org/10.1016/j.tim.2016.03.003 (2016).
20. Hale, V. L. et al. SARS-CoV-2 infection in free-ranging white-tailed deer. *Nature.* **602**, 481–. https://doi.org/10.1038/s41586-021-04353-x (2022).
21. Chandler, J. C. et al. SARS-CoV-2 exposure in wild white-tailed deer (Odocoileus virginianus). *Proc. Natl. Acad. Sci. U S A.* **118** https://doi.org/10.1073/pnas.2114828118 (2021).
22. Palmer, M. V. et al. Susceptibility of white-tailed deer (Odocoileus virginianus) to SARS-CoV-2. *J. Virol.* **95** https://doi.org/10.1128/jvi.00083-21 (2021).
23. Bao, L. et al. Susceptibility and attenuated transmissibility of SARS-CoV-2 in domestic cats. *J. Infect. Dis.* **223**, 1313–1321. https://doi.org/10.1093/infdis/jiab104 (2021).
24. Sharun, K., Saied, A. A., Tiwari, R. & Dhama, K. SARS-CoV-2 infection in domestic and feral cats: current evidence and implications. *Vet. Q.* **41**, 228–231. https://doi.org/10.1080/01652176.2021.1962576 (2021).
25. Shriner, S. A. et al. SARS-CoV-2 exposure in escaped Mink, Utah, USA. *Emerg. Infect. Dis.* **27**, 988–990. https://doi.org/10.3201/eid2703.204444 (2021).
26. Khalid, M., Alshishani, A. & Al-Ebini, Y. Genome similarities between human-derived and mink-derived SARS-CoV-2 make Mink a potential Reservoir of the Virus. *Vaccines (Basel).* **10** https://doi.org/10.3390/vaccines10081352 (2022).
27. Rabalski, L. et al. Severe Acute Respiratory Syndrome Coronavirus 2 in Farmed Mink (Neovison vison), Poland. *Emerg. Infect. Dis.* **27**, 2333–2339. https://doi.org/10.3201/eid2709.210286 (2021).
28. Hammer, A. S. et al. SARS-CoV-2 transmission between Mink (Neovison vison) and humans, Denmark. *Emerg. Infect. Dis.* **27**, 547–551. https://doi.org/10.3201/eid2702.203794 (2021).
29. Aguiló-Gisbert, J. et al. First description of SARS-CoV-2 infection in two feral American Mink (Neovison vison) caught in the Wild. *Animals.* **11**, 1422. https://doi.org/10.3390/ani11051422 (2021).
30. Padilla-Blanco, M. et al. The Finding of the severe Acute Respiratory Syndrome Coronavirus (SARS-CoV-2) in a wild Eurasian River Otter (Lutra lutra) highlights the need for viral surveillance in wild mustelids. *Front. Veterinary Sci.* **9** https://doi.org/10.3389/fvets.2022.826991 (2022).
31. Nunes, B. et al. Heterogeneous selective pressure acting on influenza B Victoria- and Yamagata-like hemagglutinins. *J. Mol. Evol.* **67**, 427–435. https://doi.org/10.1007/s00239-008-9154-9 (2008).

32. Torresi, J., Johnson, D. & Wedemeyer, H. Progress in the development of preventive and therapeutic vaccines for hepatitis C virus. *J. Hepatol.* **54**, 1273–1285. https://doi.org/10.1016/j.jhep.2010.09.040 (2011).

33. Lau, S. K. P. et al. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proceedings of the National Academy of Sciences* 102, 14040–14045, doi: (2005). https://doi.org/10.1073/pnas.0506735102

34. Siegrist, A. A. et al. Probable transmission of SARS-CoV-2 from African Lion to Zoo employees, Indiana, USA, 2021. *Emerg. Infect. Dis.* **29**, 1102–1108. https://doi.org/10.3201/eid2906.230150 (2023).

35. Colella, J. P. et al. Leveraging natural history biorepositories as a global, decentralized, pathogen surveillance network. *PLoS Pathog.* **17**, e1009583. https://doi.org/10.1371/journal.ppat.1009583 (2021).

36. Goraichuk, I. V., Arefiev, V., Stegniy, B. T. & Gerilovych, A. P. Zoonotic and reverse zoonotic transmissibility of SARS-CoV-2. *Virus Res.* **302**, 198473. https://doi.org/10.1016/j.virusres.2021.198473 (2021).

37. Liu, Y. et al. Delta spike P681R mutation enhances SARS-CoV-2 fitness over alpha variant. *Cell. Rep.* **39**https://doi.org/10.1016/j.celrep.2022.110829 (2022).

38. Jeong, B. S. et al. Structural basis for the broad and potent cross-reactivity of an N501Y-centric antibody against sarbecoviruses. *Front. Immunol.* **13**, 1049867. https://doi.org/10.3389/fimmu.2022.1049867 (2022).

39. Liu, S., Huynh, T., Stauft, C. B., Wang, T. T. & Luan, B. Structure-function analysis of resistance to Bamlanivimab by SARS-CoV-2 variants Kappa, Delta, and Lambda. *J. Chem. Inf. Model.* **61**, 5133–5140. https://doi.org/10.1021/acs.jcim.1c01058 (2021).

40. Tan, C. C. S. et al. Transmission of SARS-CoV-2 from humans to animals and potential host adaptation. *Nat. Commun.* **13**, 2988. https://doi.org/10.1038/s41467-022-30698-6 (2022).

41. Wang, X. et al. Evaluating the effect of SARS-CoV-2 spike mutations with a linear doubly robust learner. *Front. Cell. Infect. Microbiol.* **13**, 1161445. https://doi.org/10.3389/fcimb.2023.1161445 (2023).

42. Chen, J., Gao, K., Wang, R. & Wei, G. W. Revealing the threat of emerging SARS-CoV-2 mutations to antibody therapies. *J. Mol. Biol.* **433**, 167155. https://doi.org/10.1016/j.jmb.2021.167155 (2021).

43. Kurhade, C. et al. Low neutralization of SARS-CoV-2 Omicron BA.2.75.2, BQ.1.1 and XBB.1 by parental mRNA vaccine or a BA.5 bivalent booster. *Nat. Med.* **29**, 344–347. https://doi.org/10.1038/s41591-022-02162-x (2023).

44. Weaver, S. et al. Datamonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. *Mol. Biol. Evol.* **35**, 773–777. https://doi.org/10.1093/molbev/msx335 (2018).

Acknowledgements

Author contributions

Y. L., S. L., J. X. and C. X. wrote the first draft of the paper; Y. L., J. Hou, Y. W. and W. C. performed the literature search; J. Hu, J. X., Z. S. and Y. P. performed the data processing; K. T., Y. F., Q. J., W. W. and C. X. performed the systems review. C.X. and W. W. supervised the study. All authors made substantial revisions and critical review, and all authors have seen and approved the final version.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-76506-7.

Correspondence and requests for materials should be addressed to W.W. or C.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.