Contrasting Patterns in the Early Stage of SARS-CoV-2 Evolution between Humans and Minks

Jui-Hung Tai,[1,2] Hsiao-Yu Sun,[3] Yi-Cheng Tseng,[4] Guanghao Li,[5] Sui-Yuan Chang,[6] Shiou-Hwei Yeh,[7] Pei-Jer Chen,[1,7,8,9,10] Shu-Miaw Chaw,*[,2,11] and Hurng-Yi Wang*[,1,4,12]

[1]Graduate Institute of Clinical Medicine, College of Medicine, National Taiwan University, Taipei, Taiwan

[2]Genome and Systems Biology Degree Program, National Taiwan University and Academia Sinica, Taipei, Taiwan

[3]Taipei Municipal Zhongshan Girls High School, Taipei, Taiwan

[4]Institute of Ecology and Evolutionary Biology, National Taiwan University, Taipei, Taiwan

[5]CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

[6]Department of Clinical Laboratory Sciences and Medical Biotechnology, College of Medicine, National Taiwan University, Taipei, Taiwan

[7]Department of Microbiology, College of Medicine, National Taiwan University, Taipei, Taiwan

[8]Hepatitis Research Center, National Taiwan University College of Medicine and National Taiwan University Hospital, Taipei, Taiwan

[9]Department of Internal Medicine, National Taiwan University College of Medicine and National Taiwan University Hospital, Taipei, Taiwan

[10]Department of Medical Research, National Taiwan University College of Medicine and National Taiwan University Hospital, Taipei, Taiwan

[11]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan

[12]Graduate Institute of Medical Genomics and Proteomics, National Taiwan University College of Medicine, Taipei, Taiwan

***Corresponding authors:** E-mails: smchaw@sinica.edu.tw; hurngyi@ntu.edu.tw.

Associate editor: Meredith Yeager

Abstract

One of the unique features of SARS-CoV-2 is its apparent neutral evolution during the early pandemic (before February 2020). This contrasts with the preceding SARS-CoV epidemics, where viruses evolved adaptively. SARS-CoV-2 may exhibit a unique or adaptive feature which deviates from other coronaviruses. Alternatively, the virus may have been cryptically circulating in humans for a sufficient time to have acquired adaptive changes before the onset of the current pandemic. To test the scenarios above, we analyzed the SARS-CoV-2 sequences from minks (*Neovision vision*) and parental humans. In the early phase of the mink epidemic (April to May 2020), nonsynonymous to synonymous mutation ratio per site in the spike protein is 2.93, indicating a selection process favoring adaptive amino acid changes. Mutations in the spike protein were concentrated within its receptor-binding domain and receptor-binding motif. An excess of high-frequency derived variants produced by genetic hitchhiking was found during the middle (June to July 2020) and late phase I (August to September 2020) of the mink epidemic. In contrast, the site frequency spectra of early SARS-CoV-2 in humans only show an excess of low-frequency mutations, consistent with the recent outbreak of the virus. Strong positive selection in the mink SARS-CoV-2 implies that the virus may not be preadapted to a wide range of hosts and illustrates how a virus evolves to establish a continuous infection in a new host. Therefore, the lack of positive selection signal during the early pandemic in humans deserves further investigation.

Key words: **positive selection, site frequency spectrum, genetic hitchhiking, Ka/Ks.**

Article

Introduction

The pandemic coronavirus disease 2019 (COVID-19), first recorded in the city of Wuhan, China, is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Wu et al. 2020). SARS-CoV-2 is the seventh coronavirus found to infect humans. Among the other six, SARS-CoV and MERS-CoV can cause severe respiratory illness, whereas seasonal 229E, HKU1, NL63, and OC43 produce mild symptoms (Corman et al. 2018). SARS-CoV-2 exhibits 96% similarity to a coronavirus collected in Yunnan Province, China, from a bat, *Rhinolophus affinis*. It is, therefore, possible that the virus may have a zoonotic origin from bats (Andersen et al. 2020; Zhou et al. 2020).

It is generally believed that for a cross-species transmitted virus to achieve high infectiousness in a new host, multiple changes, each of conferring a selective advantage, are necessary (Parrish et al. 2008; Plowright et al. 2017; Ruan, Wen, He, et al. 2021). For example, a series of small incremental adaptations appear to underlie the emergence of SARS-CoV that infect humans (Chinese SARS Molecular Epidemiology Consortium 2004). Both SARS-CoV and SARS-CoV-2 use their spike protein to mediate entry into host cells. This protein first binds to its host receptor, angiotensin converting enzyme 2 (ACE2) and subsequently mediates fusion of viral and host membranes. This receptor binding is the first and one of the most important steps in viral infection of host cells. During the short epidemic in 2002–2003, several rounds of adaptive changes have been documented, especially in the spike protein, in SARS-CoV genomes (Yeh et al. 2004; Wu et al. 2012). As a result, SARS-CoV samples isolated from humans during the late phase of the outbreak in 2002–2003 exhibited higher affinity for human ACE2 than their early porgenitors (Cui et al. 2019).

Nevertheless, in the first several months of the current pandemic, the evidence of positive selection in SARS-CoV-2 was scarce (Chaw et al. 2020; Chiara et al. 2021; MacLean et al. 2021; Martin et al. 2021). The only probable exception is the D614G mutation in the spike protein that increases viral transmissibility (Korber et al. 2020; Plante et al. 2021). Although some ORFs such as orf3a and orf8 show Ka/Ks > 1 in the early pandemic (Chaw et al. 2020), it was due to co-segregation of both ancestral and derived alleles, such as G215V (orf3a) and L84S (orf8), at the same time. Since these derived alleles finally went extinct, it is unclear if they were, in fact, adaptive.

The lack of a positive selection signature in the early SARS-CoV-2 pandemic is in contrast to its predecessors, that is, SARS-CoV. It is possible that SARS-CoV-2 exhibits a unique feature that distinguishes it from other coronaviruses (MacLean et al. 2021) and enables its efficient cross-transmission to humans and other species without altering its genome. Alternatively, there may be a SARS-CoV-2 progenitor or prototype which has been cryptically circulating in humans for some time before being noticed (Kumar et al. 2021). During that period, the progenitor virus may have acquired adaptive changes to become the present SARS-CoV-2 and efficiently transmit among humans. After adapting to the new host, most RNA viruses exhibit strong negative selection (Lin et al. 2019). Therefore, the signature of positive selection may have become obscured by the time the pandemic took off.

These two scenarios can be tested by examining evolutionary patterns of the virus causing epidemics in other species. If SARS-CoV-2 is preadapted for cross-species transmission, positive selection should not be expected. Otherwise, if a SARS-CoV-2 progenitor has experienced adaptive evolution to cause the pandemic in humans, signatures of accelerated adaptation should be revealed when the current virus jumps to other species. The transmission of SARS-CoV-2 from humans to minks (*Neovision vision*), thus, provides an excellent opportunity to test these scenarios.

The first SARS-CoV-2 infection in minks was reported in the Netherlands in April 2020. Three of five initial introductions of SARS-CoV-2 led to subsequent spread between mink farms until November 2020. Although the modes and mechanisms of most farm-to-farm transmissions remain unknown, a study has found that movement of people and distance between farms were statistically significant predictors of virus dispersal between farms (Lu et al. 2021). In this study, we analyzed the sequences from SARS-CoV-2 viruses that infected minks. Our results show a strong signature of positive selection during the early epidemic, with the signal rapidly diminishing later in the outbreak. We also discuss how the virus can have circulated within human populations without being noticed while accumulating adaptive changes.

Results

Adaptive Evolution of SARS-CoV-2 in Minks

The phylogeny of SARS-CoV-2 derived from minks and their parental human strains is shown in figure 1*A*. Human to mink transmission clearly occurred multiple times, the majority of these events failing to trigger an epidemic. Because selection has been continuously operating in the human hosts, it is reasonable to expect higher infectiousness in humans than in minks (Ruan, Wen, He, et al. 2021). Among all inter-species transmission events, we observed three clusters of infections (mink-1 to 3, all from the Netherlands), suggesting that the emergence of new SARS-CoV-2 strains can efficiently infect minks (colored clade in fig. 1*A*). One of the clusters (mink-1, Cluster A of Lu et al. (2021)) lasted for more than 6 months (fig. 1*B*), implying that the strain may have acquired new mutations to sustain its infection in minks.

Consistent with the above scenario, strong evidence of positive selection was detected in the mink-1 clade: the ratio of nonsynonymous (Ka) to synonymous (Ks) changes per site at the spike locus is 5.33 (table 1 and supplementary table S1, Supplementary Material online). In addition to overall Ka/Ks > 1 in the spike, these mutations are concentrated in the domains critical for infection. Four of the seven amino acid changes within the spike protein are in the receptor-binding domain (RBD) ($P = 0.013$; Fisher exact test), and three of these four mutations are in the receptor-binding motif (RBM) ($P = 0.004$) (table 2 and supplementary table S2, Supplementary Material online).

To further search for evidence of positive selection, we used the fixed effects likelihood (FEL; Kosakovsky Pond and Frost 2005) and adaptive branch-site random effects likelihood mixed effects models (aBSREL; Smith et al. 2015) implemented in HyPhy (Pond et al. 2005). We only included nonredundant high-quality sequences for these analyses (supplementary fig. S1, Supplementary Material online) (see Materials and Methods). The FEL method identified eight codons putatively under positive selection ($P < 0.05$) within the mink-1 clade (fig. 2*A*). Based on epidemiological data, the course of the outbreak in minks was

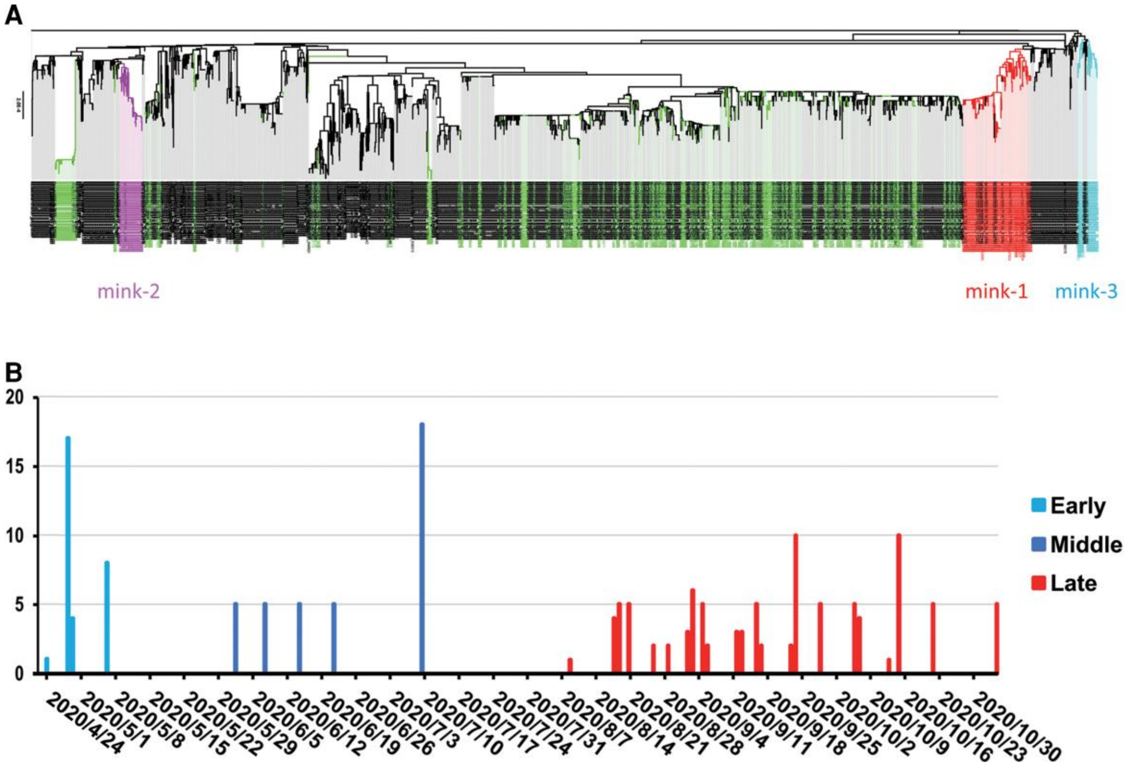

Fig. 1. Phylogeny and epidemiology of SARS-CoV-2 that infected minks and humans. (*A*) The maximum likelihood phylogeny of SARS-CoV-2 derived from minks (colored sequences) and associated humans as of December 31, 2020. (*B*) Distribution of case numbers in mink-1.

divided into early (April to May), middle (June and July), and late (August to November) phases (fig. 1*B*). The positively selected lineage identified by the aBSREL method is the lineage leading to middle and late phases (fig. 2*B*). In addition, seven putatively selected sites gradually increased in frequency from early on and were finally fixed in the late phase (fig. 2*B*). The above observations demonstrate that SARS-CoV-2 gradually acquired adaptive changes to effectively transmit among minks.

When Ka/Ks was calculated separately, the whole genome Ka/Ks was highest in the early phase (0.75) and gradually decreased to 0.61 and 0.57 at middle and late phases, respectively (table 1). Focusing on individual genes, we found that only orf1a and spike in the early phase have Ka/Ks > 1. Some short open-reading frames (ORFs) occasionally have Ka/Ks > 1 but that is because their Ks = 0, thus the evidence of positive selection on these ORFs is in doubt (supplementary table S1, Supplementary Material online). Within the spike, three amino acid mutations in the early phase are all in the RBD ($P < 0.01$) and two are in the RBM ($P < 0.01$) (table 2 and supplementary table S3, Supplementary Material online). This concentration of mutations within RBD/RBM was not seen in the middle or late phases.

In addition to mink-1, signatures of positive selection were also detected in two other clusters, mink-2 and -3. The Ka/Ks of the spike protein in mink-2 was 2.37 and 0.51 in mink-3 (table 1 and supplementary table S1, Supplementary Material online). Consistently, both lineages have mutations concentrated in the RBM of the spike protein ($P < 0.01$; table 2 and supplementary table S4, Supplementary Material online). It is noteworthy that two mutations within the RBM, Y453F and F486L, probably optimize spike binding affinity to the mink ACE2 receptor (fig. 2*C*; Ren et al. 2021; Welkers et al. 2021) and repeatedly occurred on different mink lineages. The convergence of identical mutations from different lineages provides strong evidence of positive selection and implies adaptation of the virus to its new mink hosts.

Hitchhiking under Positive Selection in Mink SARS-CoV-2

We found a strong signature of positive selection on the mink-1 lineage. When different phases of the epidemic were analyzed separately, evidence of adaptive evolution was most prominent during the early phase, but weak in the middle and late phases (fig. 2*B* and table 1). Nevertheless, the effect of positive selection can leave a trace on linked neutral variation, that is, the linked variation hitchhikes to either low or high frequencies. Although the frequency distribution of variation can be influenced by several evolutionary processes, an excess of derived variants at high frequency is a unique pattern produced by genetic hitchhiking due to positive selection (Fay and Wu 2000). We thus constructed site frequency spectra (SFSs) of both synonymous and nonsynonymous changes to look for evidence of positive selection.

SFSs of both synonymous and nonsynonymous changes in mink-1 were skewed toward high-frequency mutations

Table 1. Ka, Ks, and Ka/Ks of Different Open-Reading Frames of SARS-CoV-2 in Different Lineages.

	All			*orf1a*			*orf1b*			*Spike*		
	Ka × 10^4	Ks × 10^4	Ka/Ks	Ka × 10^4	Ks × 10^4	Ka/Ks	Ka × 10^4	Ks × 10^4	Ka/Ks	Ka × 10^4	Ks × 10^4	Ka/Ks
mink-1 (*n* = 163)	2.71	5.00	0.54	2.27	2.19	1.04	0.97	4.84	0.20	5.75	1.08	5.33
mink-1 Early Phase (*n* = 30)	1.86	2.49	0.75	2.39	1.00	2.39	0.10	2.35	0.04	2.55	0.87	2.93
mink-1 Middle Phase (*n* = 38)	2.33	3.79	0.61	2.07	3.57	0.58	1.07	2.76	0.39	2.44	0	0.64
mink-1 Late Phase (*n* = 95)	1.11	1.95	0.57	1.31	1.83	0.72	0.17	3.18	0.05	0.35	1.57	0.22
mink-2 (*n* = 59)	2.14	4.69	0.46	0.93	3.64	0.26	0.51	4.59	0.11	3.71	1.57	2.37
mink-3 (*n* = 41)	1.92	4.29	0.45	1.63	3.01	0.54	1.32	2.74	0.48	2.86	5.66	0.51
Early-Stage Human[a] (*n* = 1,476)	1.76	4.07	0.43	1.35	4.66	0.29	1.17	3.55	0.33	1.93	2.13	0.91

[a]Early-stage SARS-CoV-2 are sequences from December 2019–February 2020.

(fig. 3*A*). Assuming that the SARS-CoV-2 population has grown exponentially (MacLean et al. 2021; Martin et al. 2021), the observed SFSs significantly deviated from neutral expectation under this population growth model (Durrett 2013), further supporting the idea that the SARS-CoV-2 that transferred from human to mink hosts experienced strong positive selection.

Analyzing each phase separately, we find no excess of high-frequency mutations at either nonsynonymous ($P = 0.23$) or synonymous sites ($P = 0.27$) (fig. 3*B*) in the early phase. That is because the power to detect genetic hitchhiking is compromised when the frequency of advantageous mutations is low (Stephan et al. 2006; Zeng et al. 2006; Cutter and Payseur 2013). As shown in figure 2*B*, putative sites under selection were in low frequency, resulting in a lack of detectable deviation from neutrality.

When advantageous mutations reach high frequency in a population, the SFS reflects a deviation from neutral expectation (Zeng et al. 2006). Consequently, it is reasonable to expect the excess of high-frequency mutations in the middle and late phases of the outbreak as shown in figure 3*C*,*D*. If we further divide the late phase into late phase I (August and September) and late phase II (October and November), the hitchhiking effect is most prominent in late phase I (supplementary fig. S2*A*, Supplementary Material online) but less so in late phase 2 (supplementary fig. S2*B*, Supplementary Material online), demonstrating a rapid decay in the signature of positive selection.

Evolution of SARS-CoV-2 during the Early Epidemic in Humans

Weak signs of adaptive evolution during the early phase of the epidemic of SARS-CoV-2 in humans have been observed in many studies (Chaw et al. 2020; Tang et al. 2020; MacLean et al. 2021; Martin et al. 2021). The Ka/Ks in the whole genome is 0.43 and the spike protein 0.91 before February 29, 2020 (table 1). In addition, very few mutations occurred in the RBD or RBM of the spike protein (table 2 and supplementary table S5, Supplementary Material online).

We next constructed SFSs of both synonymous and nonsynonymous changes and found they were skewed toward high frequency, which may suggest a signature of positive selection (fig. 4*A*). However, the pattern should be interpreted with caution. The results shown in figure 4*A* were based on an outgroup comparison. The divergence at synonymous sites between SARS-CoV-2 and RaTG13 was 17%, approximately three-fold greater than between

Table 2. The Distribution of Mutation Within Spike in Different Lineages of SARS-CoV-2.

	Outside RBD[a] (1090 a.a)	Outside RBM[b] (1216 a.a)	Within RBD[c] (126 a.a)	Within RBM (69 a.a)	Fisher exact *P*-value (RBD/RBM)
mink-1	3	4	V367F	Y453F, F486L, N501T	0.01/<0.01
mink-1 early phase	0	1	V367F	F486L, N501T	<0.01/<0.01
mink-1 middle phase	2	2	–	Y453F	0.39/0.15
mink-1 late phase	2	2	–	F486L	0.39/0.15
mink-2	1	1	–	L452M, F486L	0.06/<0.01
mink-3	1	1	–	Y453F, F486L	0.06/<0.01
Early stage SARS-CoV-2	29	30	V367F	V483A	0.30/1.00

[a]RBD, Receptor-binding domain.
[b]RBM, Receptor-binding motif.
[c]The RBM is included within RBD. Thus, the mutations listed within RBM are also in RBD.

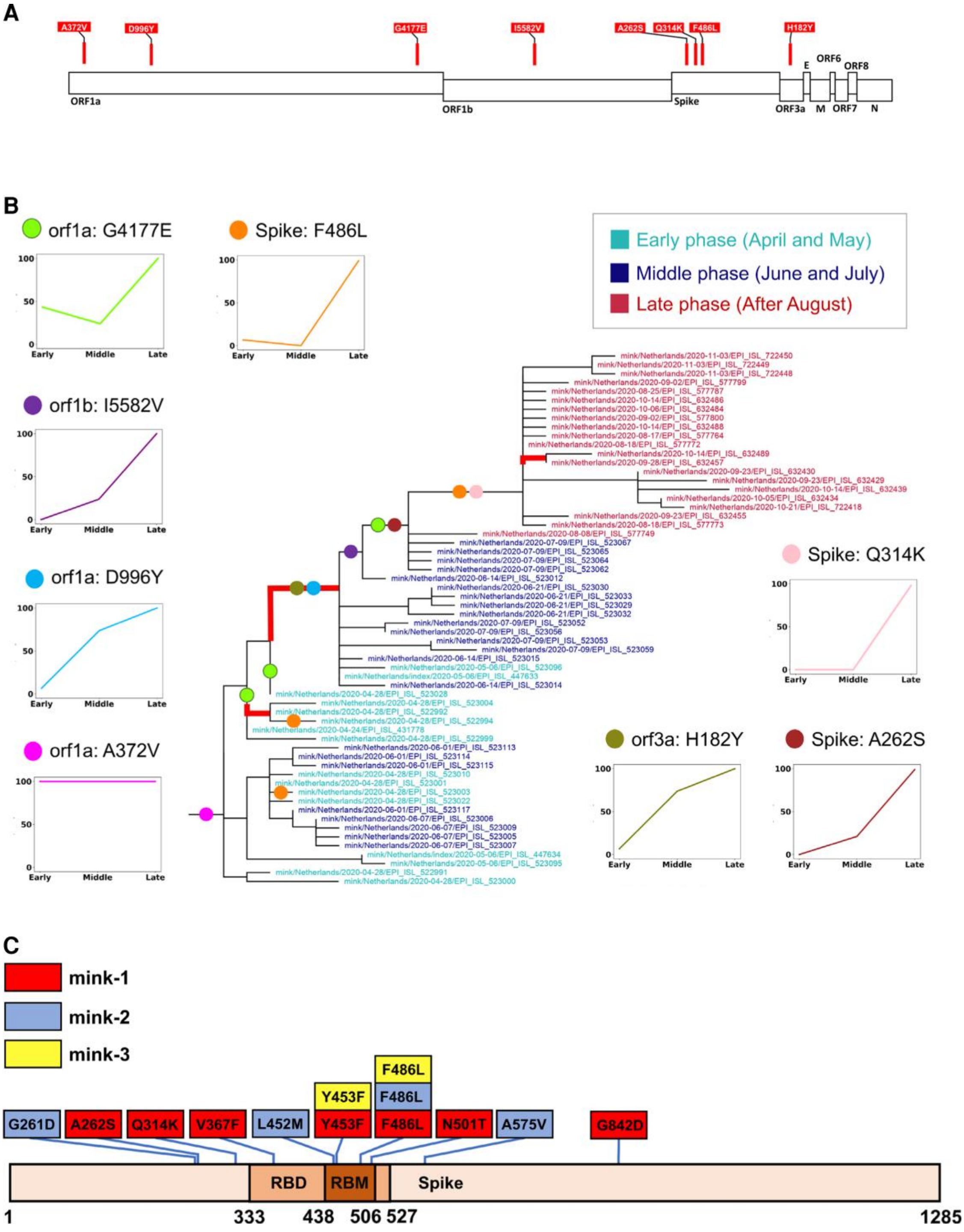
A
A372V
D996Y
G4177E
I5582V
A262S
Q314K
F486L
H182Y
ORF1a
ORF1b
Spike
E
ORF6
ORF8
ORF3a
M
ORF7
N
B
orf1a: G4177E
Spike: F486L
orf1b: I5582V
orf1a: D996Y
orf1a: A372V
Spike: Q314K
orf3a: H182Y
Spike: A262S
100
50
0
Early
Middle
Late
Early phase (April and May)
Middle phase (June and July)
Late phase (After August)
mink/Netherlands/2020-11-03/EPI_ISL_722450
mink/Netherlands/2020-11-03/EPI_ISL_722449
mink/Netherlands/2020-11-03/EPI_ISL_722448
mink/Netherlands/2020-09-02/EPI_ISL_577799
mink/Netherlands/2020-08-25/EPI_ISL_577787
mink/Netherlands/2020-10-14/EPI_ISL_632486
mink/Netherlands/2020-10-06/EPI_ISL_632484
mink/Netherlands/2020-09-02/EPI_ISL_577800
mink/Netherlands/2020-10-14/EPI_ISL_632488
mink/Netherlands/2020-08-17/EPI_ISL_577764
mink/Netherlands/2020-08-18/EPI_ISL_577772
mink/Netherlands/2020-10-14/EPI_ISL_632489
mink/Netherlands/2020-09-28/EPI_ISL_632457
mink/Netherlands/2020-09-23/EPI_ISL_632430
mink/Netherlands/2020-09-23/EPI_ISL_632429
mink/Netherlands/2020-10-14/EPI_ISL_632439
mink/Netherlands/2020-10-05/EPI_ISL_632434
mink/Netherlands/2020-10-21/EPI_ISL_722418
mink/Netherlands/2020-09-23/EPI_ISL_632455
mink/Netherlands/2020-08-18/EPI_ISL_577773
mink/Netherlands/2020-08-08/EPI_ISL_577749
mink/Netherlands/2020-07-09/EPI_ISL_523067
mink/Netherlands/2020-07-09/EPI_ISL_523065
mink/Netherlands/2020-07-09/EPI_ISL_523064
mink/Netherlands/2020-07-09/EPI_ISL_523062
mink/Netherlands/2020-06-14/EPI_ISL_523012
mink/Netherlands/2020-06-21/EPI_ISL_523030
mink/Netherlands/2020-06-21/EPI_ISL_523033
mink/Netherlands/2020-06-21/EPI_ISL_523029
mink/Netherlands/2020-06-21/EPI_ISL_523032
mink/Netherlands/2020-07-09/EPI_ISL_523052
mink/Netherlands/2020-07-09/EPI_ISL_523056
mink/Netherlands/2020-07-09/EPI_ISL_523053
mink/Netherlands/2020-07-09/EPI_ISL_523059
mink/Netherlands/2020-06-14/EPI_ISL_523015
mink/Netherlands/2020-05-06/EPI_ISL_523096
mink/Netherlands/index/2020-05-06/EPI_ISL_447633
mink/Netherlands/2020-06-14/EPI_ISL_523014
mink/Netherlands/2020-04-28/EPI_ISL_523028
mink/Netherlands/2020-04-28/EPI_ISL_523004
mink/Netherlands/2020-04-28/EPI_ISL_522992
mink/Netherlands/2020-04-28/EPI_ISL_522994
mink/Netherlands/2020-04-24/EPI_ISL_431778
mink/Netherlands/2020-04-28/EPI_ISL_522999
mink/Netherlands/2020-06-01/EPI_ISL_523113
mink/Netherlands/2020-06-01/EPI_ISL_523114
mink/Netherlands/2020-06-01/EPI_ISL_523115
mink/Netherlands/2020-04-28/EPI_ISL_523010
mink/Netherlands/2020-04-28/EPI_ISL_523001
mink/Netherlands/2020-04-28/EPI_ISL_523003
mink/Netherlands/2020-04-28/EPI_ISL_523022
mink/Netherlands/2020-06-01/EPI_ISL_523117
mink/Netherlands/2020-06-07/EPI_ISL_523006
mink/Netherlands/2020-06-07/EPI_ISL_523009
mink/Netherlands/2020-06-07/EPI_ISL_523005
mink/Netherlands/2020-06-07/EPI_ISL_523007
mink/Netherlands/index/2020-05-06/EPI_ISL_447634
mink/Netherlands/2020-05-06/EPI_ISL_523095
mink/Netherlands/2020-04-28/EPI_ISL_522991
mink/Netherlands/2020-04-28/EPI_ISL_523000
C
mink-1
mink-2
mink-3
F486L
Y453F
F486L
G261D
A262S
Q314K
V367F
L452M
Y453F
F486L
N501T
A575V
G842D
RBD
RBM
Spike
1
333
438
506
527
1285

Fig. 2. Signature of positive selection in the mink-1 lineage. (*A*) Amino acid sites putatively under positive selection identified by the fixed effects likelihood (FEL) method implemented in HyPhy. (*B*) Phylogeny of sequences from the mink-1 lineage. Only nonredundant sequences are included. Red branches are the positively selected lineage identified by the adaptive branch-site random effects likelihood method implemented in HyPhy. Sites putatively under positive selection identified by the FEL method are indicated by colored circles along the branches with their frequencies in different phases shown. (*C*) Mutations within the spike protein across clusters of minks.

humans and rhesus macaques (Wang et al. 2007). Indeed, inference of change directionality via the bat outgroup does not appear to be credible (Morel et al. 2020). With such high level of divergence, the possibility of multiple substitutions cannot be ignored (van Dorp, Acman, et al. 2020), especially since substitutions in coronavirus genomes are strongly biased toward transitions (Matyášek and Kovařík 2020; van Dorp, Richard, et al. 2020).

Among all mutations in figure 4A, 32.88% of the changes were C to T transitions (table 3A), two-fold higher than T to C transitions (15.02%), as shown in the previous study (Simmonds and Schwemmle 2020). The bias toward

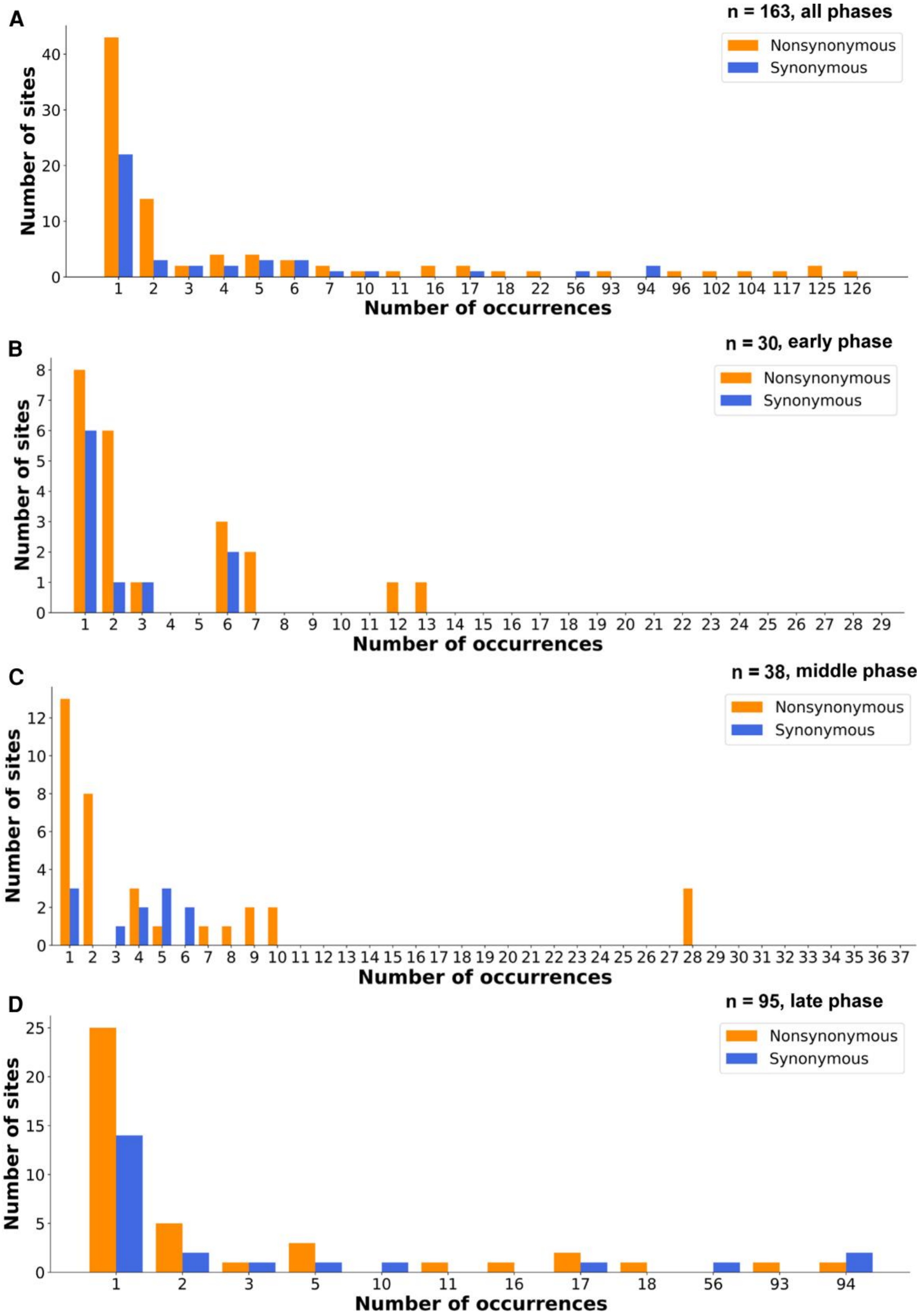

Fig. 3. Site frequency spectra of SARS-CoV-2 in the mink-1 lineage in each epidemic phase. (*A*) Site frequency spectra (SFSs) of the mink-1 lineage. Significant deviation from neutral expectation under exponential population growth was found in both synonymous ($P < 10^{-3}$) and nonsynonymous mutations ($P < 10^{-5}$). (*B*) SFSs of the mink-1 lineage in the early phase. Neither synonymous nor nonsynonymous mutations deviate from the neutral expectation. (*C*) SFSs of mink-1 in middle phase. Only nonsynonymous mutations are deviated from neutral expectation ($P < 10^{-4}$). (*D*) SFSs of the mink-1 lineage in the late phase. Significant deviation from neutral expectation is seen in both synonymous ($P < 10^{-5}$) and nonsynonymous ($P < 10^{-3}$) mutations.

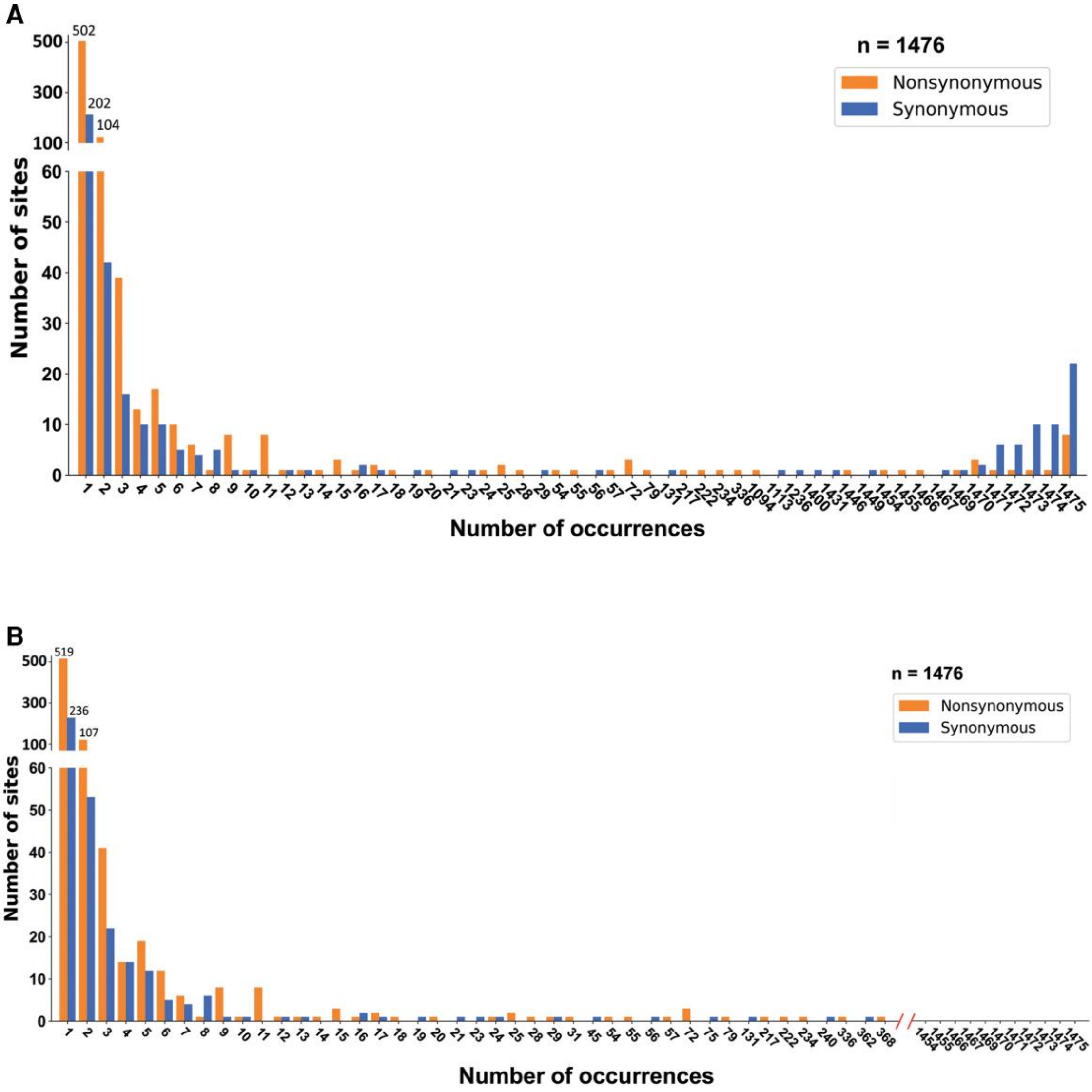

FIG. 4. Site frequency spectra of SARS-CoV-2 during the early epidemic (December 2019–February 2020) in humans. (*A*) Site frequency spectra (SFSs) inferred using RaTG13 as the outgroup. Significant deviation from neutral expectation under exponential population growth was found in both synonymous ($P < 10^{-5}$) and nonsynonymous mutations ($P < 10^{-5}$). (*B*) SFSs cross-referenced by the phylogeny and date of sampling (see main text for details). Neither synonymous nor nonsynonymous mutations deviate from the neutral expectation.

C to T changes was probably mediated either through selective pressures by a CpG-targeting mechanism involving the Zinc finger Antiviral Protein (ZAP), C to U hypermutation by APOBEC3 cytidine deaminases, or escape from the host immune system (Bishop et al. 2004; Greenbaum et al. 2008; Takata et al. 2017; Kmiec et al. 2020; Kustin and Stern 2020; Pollock et al. 2020).

Strikingly, the changes from the common ancestor to SARS-CoV-2 and RaTG-13 were strongly biased toward T to C, 50% higher than C to T (table 3B). Contrasting patterns between divergence and polymorphism imply that many nucleotide sites may have changed back and forth during evolution, further increasing the complexity of inferring directionality of changes using outgroups. For example, the earliest available SARS-CoV-2 genome (December 24, 2019) has the characteristic motif C:C:T at the nucleotide residues 8,782, 18,060, and 28,144, different from the RaTG13 **T:T:C** sequence (supplementary table S6, Supplementary Material online). The first mutant strain carrying the **T:C:C** motif, with the bold SNVs matching RaTG13, was found on December 30, 2019. At the same time, another 20 genome sequences carrying the C:C:T motif were recovered in Wuhan, China. The first viral strain carrying the **T:T:C** motif was found in the USA on January 19, 2020 and Guangdong on January 20, 2020. Therefore, the **T:T:C** motif, although matching RaTG13, is composed of derived instead of ancestral nucleotides.

To get round the potential problem of multiple mutations, we cross-referenced phylogeny and date of sampling (Chaw et al. 2020). Many T to C changes based on outgroup comparison were inferred as C to T changes (table 3A), because those T were first recorded outside Wuhan and appeared after mid-January 2020 (supplementary table S7, Supplementary Material online). As a result, all

Table 3. Directionality of Nucleotide Changes of SARS-CoV-2 in (*A*) Polymorphism (Early Stage (December 2019–February 2020)) and (*B*) Divergence.

(*A*) Polymorphism

Changes	Outgroup comparison		Adjusted	
	#	(%)	#	(%)
C->T	381	32.88	420	36.24
T->C	174	15.02	132	11.39
A->G	147	12.69	134	11.57
G->T	144	12.43	149	12.86
G->A	99	8.55	107	9.24
A->T	46	3.97	43	3.72
T->A	44	3.8	49	4.23
C->A	32	2.77	32	2.77
A->C	31	2.68	30	2.59
T->G	29	2.51	29	2.51
G->C	18	1.56	20	1.73
C->G	14	1.21	14	1.21

(*B*) Divergence

Changes	Ancestor to SARS2		Ancestor to RaTG13	
	#	(%)	#	(%)
T->C	185	38.07	212	37.53
C->T	112	23.05	137	24.25
A->G	75	15.44	83	14.7
G->A	40	8.24	51	9.03
T->A	26	5.35	26	4.61
A->T	14	2.89	25	4.43
C->A	10	2.06	8	1.42
T->G	7	1.45	8	1.42
A->C	7	1.45	7	1.24
G->T	6	1.24	4	0.71
C->G	3	0.62	3	0.54
G->C	1	0.21	1	0.18

mutations listed as high frequency in figure 4A were reassigned to the other side of the frequency spectra (fig. 4B). The re-estimated SFSs only show an excess of low-frequency mutations, consistent with a recent origin of SARS-CoV-2 and suggesting that population expansion is the major force shaping site frequency spectra during the evolution of this virus. However, we did not observe a significant deviation from neutral expectation of growing populations. Thus, in good agreement with many previous studies, we did not detect positive selection during the early phase of the SARS-CoV-2 pandemic (Chaw et al. 2020; MacLean et al. 2021).

Discussion

Adaptation of SARS-CoV-2 in Minks

Using several approaches, we identified signs of adaptive evolution in lineages leading to mink-1. Transmission between humans and minks illustrates how a virus evolves to establish a continuous infection in a new host. In the beginning, a virus may invade a new host multiple times but fail to trigger an epidemic. Because natural selection is working in the original hosts, higher infectiousness of a virus in the original than in new hosts is expected (Ruan, Wen, He, et al. 2021).

In the case of SARS-CoV-2, while many studies documented the susceptibility of different animal species to the virus, including cats, dogs, tigers, lions, ferrets, and rhesus macaques (Salajegheh Tazerji et al. 2020; Shan et al. 2020; Shi et al. 2020; Sit et al. 2020), the virus only caused outbreaks in minks (Oude Munnink et al. 2021). It is possible that high population density of farm minks facilitates virus spread. However, as shown in figure 1, transmission from humans to minks occurred multiple times, the majority of them failing to trigger an epidemic. Similar phenomenon was also found in free-ranging white-tailed deer (*Odocoileus virginianus*) (Hale et al. 2021). Deer in six locations of northeast Ohio (USA) were infected by different SARS-CoV-2 lineages/variants originated from humans. Evidence of deer-to-deer transmission was confirmed only in one location. The above observations support previous suggestion that many attempts of SARS-CoV-2 to jump cross-species boundaries appeared as “spill-over” infections (Pekar et al. 2021).

Occasionally, after many failed attempts, the emergence of a new strain can sustain the infection long enough to acquire new mutations for further enhancement of infectiousness in the new hosts. A series of putative adaptive changes identified in the lineages leading to late phase of mink-1 outbreak supports the adaptation of SARS-CoV-2 in the new host (fig. 2).

Meanwhile, because the virus may travel back and forth between old and new hosts (Oude Munnink et al. 2021) (supplementary fig. S3, Supplementary Material online), evolution may not only be in the recipients, but also in the donors. We observed that multiple putatively adaptive mutations (Y453F, F486L, and N501T) repeatedly occurred in different mink lineages (supplementary tables S3 and S4, Supplementary Material online). The introduction of these mutations into human populations from minks was documented (Welkers et al. 2021). Y453F enhances binding to the mink ACE2 and other orthologs of *Mustela* species without compromising, and even enhancing, its ability to utilize human ACE2 as a receptor for entry (Ren et al. 2021). Interestingly, in a case report involving a patient receiving the Regeneron treatment, escape mutations were identified in several positions of the S protein, including 486 and 501 (Choi et al. 2020). Thus, while adaptation of SARS-CoV-2 occurred in minks, it is possible that these mutants may affect evolution of the virus in humans as well.

On the Origin of SARS-CoV-2 in Humans

Similar to previous studies, we did not detect evidence of positive selection in the early episode of the SARS-CoV-2 pandemic (Chaw et al. 2020; MacLean et al. 2021; Martin et al. 2021). Virus adaptation to new hosts that is sufficient to produce a pandemic often entails a significant adaptive challenge and requires the acquisition of the ability to (1) bind and enter hosts’ cells, (2) evade host restriction factors and immune responses, and (3) transmit effectively among hosts. Such adaptation is not likely to have emerged suddenly but, instead, may have evolved step by step with each step favored by natural selection (Parrish et al. 2008; Pepin et al. 2010; Plowright et al. 2017; Cui et al. 2019; Long et al. 2019; Kang et al. 2021; Ruan, Wen, He, et al. 2021). Some viruses may possess the baseline abilities needed to ensure its onward transmission in the new host (Geoghegan and Holmes 2017). For instance, EBOV has crossed the species barrier from its reservoir hosts to humans and caused several localized epidemics since 1976. Nonetheless, these outbreaks were resolved after at most a few hundred cases (Jacob et al. 2020). The 2013–2016 Western African EBOV disease outbreak (the largest in history), which caused 28,625 infections and 11,325 deaths, was linked to a series of active adaptive events of EBOV to humans (Diehl et al. 2016; Urbanowicz et al. 2016).

While the history of early adaptation is unknown, the RBD of the SARS-CoV-2 spike protein appears to be highly specialized to human ACE2 (Delgado Blanco et al. 2020). Substitution of eight SARS-CoV-2 RBD residues proximal to the ACE2-binding surface with those found in RaTG13 is almost universally detrimental to human ACE2 receptor usage (Conceicao et al. 2020). Both SARS-CoV-2 and RaTG13 bind poorly to *R. sinicus* ACE2 (Li et al. 2020). These findings indicate that SARS-CoV-2 is well adapted to humans. In this study, we observe strong signatures of positive selection in the viral strains that successfully established continuous infections among minks. By analogy, it appears unlikely that a nonhuman progenitor of SARS-CoV-2 would require little or no novel adaptation to successfully infect humans.

We thus hypothesize that the progenitor of SARS-CoV-2 may have been cryptically circulating among humans before the current outbreak. During the period of unawareness, the virus had gradually accumulated adaptive changes that enabled it to effectively infect humans and finally cause the pandemic (Kumar et al. 2021). For example, Kang et al. (2021) showed that a putative adaptive nonsynonymous change (A1114G; T372A) within the RBD of the spike protein likely contributed to SARS-CoV-2 emergence from animal reservoirs or enabled sustained human-to-human transmission.

Although the exact time the prototype SARS-CoV-2 jumped to humans is difficult to estimate without further information, we may refer to the timing from minks and SARS-CoV. After adapting to the new hosts, most RNA viruses exhibit strong negative selection (Lin et al. 2019). While the signature of positive selection diminished quickly, it may still leave a trace on linked neutral variation which can be revealed by SFSs if the selection event is relatively recent, as shown in mink-1 (fig. 3). During the short episode of the SARS-CoV outbreak in 2002–2003, despite the Ka/Ks ratio of its spike gene being reduced from 1.248 in the early phase to 0.219 in the late phase (Chinese SARS Molecular Epidemiology Consortium 2004), the SFS still showed evidence of genetic hitchhiking due to positive selection in later stages (supplementary fig. S4, Supplementary Material online). However, SFSs of SARS-CoV-2 before February 29, 2020 only show recent population expansion with no sign of genetic hitchhiking. Therefore, it is reasonable to presume that the SARS-CoV-2 progenitor may have associated with humans unnoticed for longer than the SARS-CoV-2 mink infection and the SARS-CoV epidemic episodes (Kang et al. 2021; Kumar et al. 2021; Ruan, Wen, Hou, et al. 2021), perhaps before June 2019.

Alternatively, we cannot rule out the possibility that, unlike its precedent SARS-CoV and counterpart in the minks, the SARS-CoV-2 exhibits a unique property which facilitates successful infection in humans (MacLean et al. 2021). Therefore, adaptive change in the SARS-CoV-2 genome may not be necessary and the origin of current pandemic may be very close to the first case of SARS-CoV-2 emerged in Hubei province, China (Pekar et al. 2021). To verify these hypotheses, more extensive analysis and experimental confirmation are required. It is essential to collect archive samples from environments and pneumonia patients in the Wuhan area for analysis. These data are needed to trace its evolutionary path and/or to reveal critical steps required for effective spread.

Why and How the Current Pandemic Occurred

It is not uncommon that the origin of virus infection dates back before the awareness of the epidemic. For example, molecular clock dating suggests the onset of HIV-M and -O epidemics occurred at the beginning of the 20th century (Korber et al. 2000; Lemey et al. 2004; Worobey et al. 2004). The earliest documented HIV-1 infection was discovered in a preserved blood sample taken in 1959 from a man living in what was then Belgian Congo (Nahmias et al. 1986; Zhu et al. 1998). However, it was not until 1980 that the virus was finally confirmed as the causal agent of AIDS (Barre-Sinoussi et al. 1983; Gallo et al. 1984; Popovic et al. 1984). After the last reported cases of rabies in a human in 1959 and a nonhuman animal in 1961, Taiwan was considered free from rabies. However, a rabies outbreak occurred among ferret badgers in Taiwan in 2012 and 2013. Further field survey confirmed that the ferret badger (*Melogale moschata*) is the sole reservoirs species of rabies in Taiwan (Lan et al. 2017), indicating that the rabies may have associate with ferret badgers for many years without being noticed. Furthermore, phylogeographic analyses suggest that the virus has been in Taiwan for more than 100 years, demonstrating that even the rabies virus can circulate cryptically in the environment (Chiou et al. 2014).

Even viruses that appear to be well adapted to humans may fail to induce an outbreak (Geoghegan and Holmes 2017). That is because herd immunity can develop in local populations and impede epidemic spread while the virus is building up its ability to infect humans. Such viruses would then be more infectious outside the enzootic area as outside populations are immunologically naïve (Ruan, Wen, He, et al. 2021). That is why the place of origin is not necessarily the same as the outbreak location, as can be seen in the cases of HIV and influenza (Crosby 2003; Barry 2004;

Sharp and Hahn 2011). It is also possible for a changing ecological environment to impact virus spread. In the case of the canine influenza virus (CIV), which jumped to dogs in the late 1990s from an equine influenza strain prevalent in horses (Crawford et al. 2005), Dalziel et al., found that CIV is largely confined to dog shelters in the US, where most dogs are infected soon after they arrive. But the virus cannot be maintained for long in smaller facilities or in the companion dog population without input from the larger shelters (Dalziel et al. 2014). These hotspot dynamics give a clear picture of what can happen in the time between the beginning of a host range shift and the onset of a possible pandemic.

Materials and Methods

Data Collection

All sequences were downloaded from the Global Initiative on Sharing Avian Influenza Data (GISAID, https://www.gisaid.org/) on or before February 5, 2021. Only complete and high coverage genomes were used. All 796 SARS-CoV-2 genomes labeled as *Neovision vision* (minks) from Denmark, the Netherlands, the USA, Poland, and Canada were included. We also retrieved all human SARS-CoV-2 sequences from the Netherlands (6,625), Poland (406), and Canada (7,102). For Denmark and the USA, due to the sheer amount of data available, only sequences collected between dates that are 7 days before the first mink sequence and 7 days after the last mink sequence were included. As a result, 27,971 SARS-CoV-2 genomic sequences were used.

For early-stage data, a collection of 1,476 complete and high coverage genomic sequences, with the collection starting from the earliest sequence to February (December 24, 2019–February 29, 2020), were retrieved.

Sequence Analyses and Phylogenetic Reconstruction

All sequences were aligned against the reference genome (EPI_ISL_402125) using the default settings in ClustalW (Thompson et al. 1994). Phylogenies were constructed using IQ-TREE 2.1.2 (Minh et al. 2020). Numbers of nonsynonymous changes per nonsynonymous site (Ka) and synonymous changes per synonymous site (Ks) among genomes were estimated based on Li-Wu-Luo's method (Li et al. 1985) implemented in MEGA-X (Kumar et al. 2018). Kimura's two-parameter model was used for estimating genetic distances between sequences.

For site frequency spectrum (SFS) construction of mink-1, sequence sets that were immediate sister groups to target groups were used to infer directionality of changes (supplementary fig. S3, Supplementary Material online). We first used EPI_ISL_422678 to infer changes. Several other closely related sequences were also applied to identify mutations unique to EPI_ISL_422678. We also used different sequences as shown in supplementary figure S3, Supplementary Material online to construct SFSs and the results were essentially the same. Therefore, our SFS construction should be authentic.

For the early-stage SARS-CoV-2 sequences, we first used RaTG13 as an outgroup to construct the SFS. We also cross-referenced the directionality of changes based on phylogeny and date of collection. To test whether the observed SFS deviates from neutral expectation under exponential population growth, a custom R script was utilized based on the theorem of Durrett (2013). The method was developed to analyze the SFS of cancer genomes, which are exponentially growth and largely nonrecombining.

The ancestor sequences of SARS-CoV-2 and RaTG13 were reconstructed using *codeml* implemented in PAML 4 (Yang 2007) under the free ratio model. The sequences used in this analysis included Rf1 (DQ412042.1), HKU3-1 (DQ022305.2), BM48-31 (NC_014470.1), ZC45 (MG772933.1), and ZXC21 (MG772934.1). The reconstructed ancestral sequence was used to infer nucleotide changes after the divergence of SARS-CoV-2 and RaTG13.

Detection of Positive Selection in SARS-CoV-2

To examine signatures of positive selection in the SARS-CoV-2 isolates derived from minks, we included all sequences from the Netherlands. In order to facilitate our analyses, we only retained sequences derived from humans with less than 99.9% nucleotide identify. For SARS-CoV-2 from minks, sequences with ambiguous nucleotides were removed. The resulting dataset contains 92 sequences (32 humans and 60 minks). A maximum likelihood tree was constructed using MEGA-X.

An array of selection detection methods implemented in HyPhy was applied to detect whether the lineage leading to minks has experienced adaptive evolution (Pond et al. 2005). We employed the FEL method (Kosakovsky Pond and Frost 2005) to infer amino acid sites under positive selection within minks. We also searched for evidence of positive selection on specific branches using the aBSREL method (Smith et al. 2015). Because identical or essentially identical sequences do not increase power for codon-based methods to detect selection, we set genetic distance of 0.001 for the above analyses.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We would like to thank Anthony Greenberg for his comments and all those who had contributed sequences to the Global Initiative on Sharing Avian Influenza Data (GISAID, https://www.gisaid.org/). This work was supported by grants from the Ministry of Science and Technology (MOST), Taiwan (109-2311-B-002-023-MY3, 109-2327-B-002-009, 109-2634-F-002-043, 110-2621-B-001-003, 111-2321-B-002-017, and 111-2634-F-002-017), Biodiversity Research Center, Academia Sinica (2020–2021 to SMC), the "Center of Precision Medicine" from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE), and National Taiwan University, College of Medicine (NSC-131-5).

Data Availability

All sequences used in this study are listed in https://github.com/ala98412/Sequence-List#sequence-list.

References

Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. 2020. The proximal origin of SARS-CoV-2. *Nat Med*. **26**(4):450–452.

Barre-Sinoussi F, Chermann J, Rey F, Nugeyre M, Chamaret S, Gruest J, Dauguet C, Axler-Blin C, Vezinet-Brun F, Rouzioux C, *et al*. 1983. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220**-(4599):868–871.

Barry JM. 2004. The site of origin of the 1918 influenza pandemic and its public health implications. *J Transl Med*. **2**(1):3.

Bishop KN, Holmes RK, Sheehy AM, Malim MH. 2004. APOBEC-mediated editing of viral RNA. *Science* **305**(5684):645.

Chaw S-M, Tai J-H, Chen S-L, Hsieh C-H, Chang S-Y, Yeh S-H, Yang W-S, Chen P-J, Wang H-Y. 2020. The origin and underlying driving forces of the SARS-CoV-2 outbreak. *J Biomed Sci*. **27**(1):73.

Chiara M, Horner DS, Gissi C, Pesole G. 2021. Comparative genomics reveals early emergence and biased spatiotemporal distribution of SARS-CoV-2. *Mol Biol Evol*. **38**(6):2547–2565.

Chinese SARS Molecular Epidemiology Consortium. 2004. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science* **303**(5664):1666–1669.

Chiou H-Y, Hsieh C-H, Jeng C-R, Chan F-T, Wang H-Y, Pang V-F. 2014. Molecular characterization of cryptically circulating rabies virus from ferret badgers, Taiwan. *Emerg Infect Dis*. **20**(5):790–798.

Choi B, Choudhary MC, Regan J, Sparks JA, Padera RF, Qiu X, Solomon IH, Kuo HH, Boucau J, Bowman K, *et al*. 2020. Persistence and evolution of SARS-CoV-2 in an immunocompromised host. *New Engl J Med*. 383(23):2291–2293.

Conceicao C, Thakur N, Human S, Kelly JT, Logan L, Bialy D, Bhat S, Stevenson-Leggett P, Zagrajek AK, Hollinghurst P, *et al*. 2020. The SARS-CoV-2 Spike protein has a broad tropism for mammalian ACE2 proteins. *PLoS Biol*. **18**(12):e3001016.

Corman VM, Muth D, Niemeyer D, Drosten C. 2018. Hosts and sources of endemic human coronaviruses. *Adv Virus Res*. **100**: 163–188.

Crawford PC, Dubovi EJ, Castleman WL, Stephenson I, Gibbs EPJ, Chen L, Smith C, Hill RC, Ferro P, Pompey J, *et al*. 2005. Transmission of equine influenza virus to dogs. *Science* **310**-(5747):482–485.

Crosby AW. 2003. *America's forgotten pandemic: the influenza of 1918*. Cambridge: Cambridge University Press.

Cui J, Li F, Shi Z-L. 2019. Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol*. **17**(3):181–192.

Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat Rev Genet*. **14**(4):262–274.

Dalziel BD, Huang K, Geoghegan JL, Arinaminpathy N, Dubovi EJ, Grenfell BT, Ellner SP, Holmes EC, Parrish CR. 2014. Contact heterogeneity, rather than transmission efficiency, limits the emergence and spread of canine influenza virus. *PLoS Pathog*. **10**(10): e1004455.

Delgado Blanco J, Hernandez-Alias X, Cianferoni D, Serrano L. 2020. In silico mutagenesis of human ACE2 with S protein and translational efficiency explain SARS-CoV-2 infectivity in different species. *PLoS Comp Biol*. **16**(12):e1008450.

Diehl WE, Lin AE, Grubaugh ND, Carvalho LM, Kim K, Kyawe PP, McCauley SM, Donnard E, Kucukural A, McDonel P, *et al*. 2016. Ebola virus glycoprotein with increased infectivity dominated the 2013–2016 epidemic. *Cell* **167**(4):1088–1098.e6.

Durrett R. 2013. Population genetics of neutral mutations in exponentially growing cancer cell populations. *Ann Appl Probab*. **23**(1):230–250.

Fay JC, Wu C-I. 2000. Hitchhiking under positive Darwinian selection. *Genetics* **155**(3):1405–1413.

Gallo R, Salahuddin S, Popovic M, Shearer G, Kaplan M, Haynes B, Palker T, Redfield R, Oleske J, Safai B, *et al*. 1984. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science* **224**-(4648):500–503.

Geoghegan JL, Holmes EC. 2017. Predicting virus emergence amid evolutionary noise. *Open Biol*. **7**(10):170189.

Greenbaum BD, Levine AJ, Bhanot G, Rabadan R. 2008. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. *PLoS Pathog*. **4**(6):e1000079.

Hale VL, Dennis PM, McBride DS, Nolting JM, Madden C, Huey D, Ehrlich M, Grieser J, Winston J, Lombardi D, *et al*. 2021. SARS-CoV-2 infection in free-ranging white-tailed deer. *Nature* **602**:481–486.

Jacob ST, Crozier I, Fischer WA, Hewlett A, Kraft CS, de La Vega MA, Soka MJ, Wahl V, Griffiths A, Bollinger L, *et al*. 2020. Ebola virus disease. *Nat Rev Dis Primers* **6**(1):13.

Kang L, He G, Sharp AK, Wang X, Brown AM, Michalak P, Weger-Lucarelli J. 2021. A selective sweep in the Spike gene has driven SARS-CoV-2 human adaptation. *Cell* **184**(17): 4392–4400.e4.

Kmiec D, Nchioua R, Sherrill-Mix S, Stürzel CM, Heusinger E, Braun E, Gondim MVP, Hotter D, Sparrer KMJ, Hahn BH, *et al*. 2020. CpG frequency in the 5′ third of the *env* gene determines sensitivity of primary HIV-1 strains to the zinc-finger antiviral protein. *mBio* **11**(1):e02903-19.

Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, *et al*. 2020. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**(4): 812–827.e19.

Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, Bhattacharya T. 2000. Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**(5472):1789–1796.

Kosakovsky Pond SL, Frost SDW. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol*. **22**(5):1208–1222.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol Biol Evol*. **35**(6):1547–1549.

Kumar S, Tao Q, Weaver S, Sanderford M, Caraballo-Ortiz MA, Sharma S, Pond SLK, Miura S. 2021. An evolutionary portrait of the progenitor SARS-CoV-2 and its dominant offshoots in COVID-19 pandemic. *Mol Biol Evol*. **38**(8):3046–3059.

Kustin T, Stern A. 2020. Biased mutation and selection in RNA viruses. *Mol Biol Evol*. **38**(2):575–588.

Lan Y-C, Wen T-H, Chang C-C, Liu H-F, Lee P-F, Huang C-Y, Chomel BB, Chen Y-MA. 2017. Indigenous wildlife rabies in Taiwan: Ferret Badgers, a long term terrestrial reservoir. *Biomed Res Int* **2017**: 5491640.

Lemey P, Pybus OG, Rambaut A, Drummond AJ, Robertson DL, Roques P, Worobey M, Vandamme A-M. 2004. The molecular population genetics of HIV-1 group O. *Genetics* **167**(3): 1059–1068.

Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol*. **2**(2):150–174.

Li Y, Wang H, Tang X, Fang S, Ma D, Du C, Wang Y, Pan H, Yao W, Zhang R, *et al*. 2020. SARS-CoV-2 and three related coronaviruses utilize multiple ACE2 orthologs and are potently blocked by an improved ACE2-Ig. *J Virol*. **94**(22):e01283-20.

Lin J-J, Bhattacharjee MJ, Yu C-P, Tseng YY, Li W-H. 2019. Many human RNA viruses show extraordinarily stringent selective constraints on protein evolution. *Proc Natl Acad Sci USA* **116**(38): 19009–19018.

Long JS, Mistry B, Haslam SM, Barclay WS. 2019. Host and viral determinants of influenza A virus species specificity. *Nat Rev Microbiol*. **17**(2):67–81.

Lu L, Sikkema RS, Velkers FC, Nieuwenhuijse DF, Fischer EAJ, Meijer PA, Bouwmeester-Vincken N, Rietveld A, Wegdam-Blans MCA, Tolsma P, *et al.* 2021. Adaptation, spread and transmission of SARS-CoV-2 in farmed minks and associated humans in the Netherlands. *Nat Commun*. **12**(1):6802.

MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, Kosakovsky Pond SL, Robertson DL. 2021. Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. *PLoS Biol*. **19**(3): e3001115.

Martin DP, Weaver S, Tegally H, San JE, Shank SD, Wilkinson E, Lucaci AG, Giandhari J, Naidoo S, Pillay Y, *et al.* 2021. The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell* **184**(20):5189–5200.e7.

Matyášek R, Kovařík A. 2020. Mutation patterns of human SARS-CoV-2 and bat RaTG13 coronavirus genomes are strongly biased towards C>U transitions, indicating rapid evolution in their hosts. *Genes* **11**(7):761.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. **37**(5):1530–1534.

Morel B, Barbera P, Czech L, Bettisworth B, Hübner L, Lutteropp S, Serdari D, Kostaki E-G, Mamais I, Kozlov AM, *et al.* 2020. Phylogenetic analysis of SARS-CoV-2 data is difficult. *Mol Biol Evol*. **38**(5):1777–1791.

Nahmias AJ, Weiss J, Yao X, Lee F, Kodsi R, Schanfield M, Matthews T, Bolognesi D, Durack D, Motulsky A, *et al.* 1986. Evidence for human infection with an HTLV-III LAV-like virus in Central-Africa, 1959. *Lancet* **1**(8492):1279–1280.

Oude Munnink BB, Sikkema RS, Nieuwenhuijse DF, Molenaar RJ, Munger E, Molenkamp R, van der Spek A, Tolsma P, Rietveld A, Brouwer M, *et al.* 2021. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* **371**(6525):172–177.

Parrish CR, Holmes EC, Morens DM, Park E-C, Burke DS, Calisher CH, Laughlin CA, Saif LJ, Daszak P. 2008. Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiol Mol Biol Rev*. **72**(3):457–470.

Pekar J, Worobey M, Moshiri N, Scheffler K, Wertheim JO. 2021. Timing the SARS-CoV-2 index case in Hubei province. *Science* **372**(6540):412–417.

Pepin KM, Lass S, Pulliam JRC, Read AF, Lloyd-Smith JO. 2010. Identifying genetic markers of adaptation for surveillance of viral host jumps. *Nat Rev Microbiol*. **8**(11):802–813.

Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, Zhang X, Muruato AE, Zou J, Fontes-Garfias CR, *et al.* 2021. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* **592**(7852):116–121.

Plowright RK, Parrish CR, McCallum H, Hudson PJ, Ko AI, Graham AL, Lloyd-Smith JO. 2017. Pathways to zoonotic spillover. *Nat Rev Microbiol*. **15**(8):502–510.

Pollock DD, Castoe TA, Perry BW, Lytras S, Wade KJ, Robertson DL, Holmes EC, Boni MF, Kosakovsky Pond SL, Parry R, *et al.* 2020. Viral CpG deficiency provides no evidence that dogs were intermediate hosts for SARS-CoV-2. *Mol Biol Evol*. **37**(9): 2706–2710.

Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**(5):676–679.

Popovic M, Sarngadharan M, Read E, Gallo R. 1984. Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. *Science* **224**-(4648):497–500.

Ren W, Lan J, Ju X, Gong M, Long Q, Zhu Z, Yu Y, Wu J, Zhong J, Zhang R, *et al.* 2021. Mutation Y453F in the spike protein of SARS-CoV-2 enhances interaction with the mink ACE2 receptor for host adaption. *PLoS Pathog*. **17**(11):e1010053.

Ruan Y, Wen H, He X, Wu C-I. 2021. A theoretical exploration of the origin and early evolution of a pandemic. *Sci Bull*. **66**(10): 1022–1029.

Ruan Y, Wen H, Hou M, He Z, Lu X, Xue Y, He X, Zhang Y-P, Wu C-I. 2021. The twin-beginnings of COVID-19 in Asia and Europe – one prevails quickly. *Natl Sci Rev*. **9**:nwab223.

Salajegheh Tazerji S, Magalhães Duarte P, Rahimi P, Shahabinejad F, Dhakal S, Singh Malik Y, Shehata AA, Lama J, Klein J, Safdar M, *et al.* 2020. Transmission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) to animals: an updated review. *J Transl Med*. **18**(1):358.

Shan C, Yao Y-F, Yang X-L, Zhou Y-W, Gao G, Peng Y, Yang L, Hu X, Xiong J, Jiang R-D, *et al.* 2020. Infection with novel coronavirus (SARS-CoV-2) causes pneumonia in *Rhesus macaques*. *Cell Res*. **30**(8):670–677.

Sharp PM, Hahn BH. 2011. Origins of HIV and the AIDS pandemic. *Cold Spring Harbor Perspect Med*. **1**:1.

Shi J, Wen Z, Zhong G, Yang H, Wang C, Huang B, Liu R, He X, Shuai L, Sun Z, *et al.* 2020. Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS-coronavirus 2. *Science* **368**-(6494):1016–1020.

Simmonds P, Schwemmle M. 2020. Rampant C→U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses: causes and consequences for their short- and long-term evolutionary trajectories. *mSphere* **5**(3):e00408-20.

Sit THC, Brackman CJ, Ip SM, Tam KWS, Law PYT, To EMW, Yu VYT, Sims LD, Tsang DNC, Chu DKW, *et al.* 2020. Infection of dogs with SARS-CoV-2. *Nature* **586**(7831):776–778.

Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. 2015. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol Biol Evol*. **32**(5):1342–1353.

Stephan W, Song YS, Langley CH. 2006. The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics* **172**(4):2647–2663.

Takata MA, Gonçalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, Bieniasz PD. 2017. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature* **550**-(7674):124–127.

Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, *et al.* 2020. On the origin and continuing evolution of SARS-CoV-2. *Natl Sci Rev*. **7**(6):1012–1023.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. **22**(22): 4673–4680.

Urbanowicz RA, McClure CP, Sakuntabhai A, Sall AA, Kobinger G, Müller MA, Holmes EC, Rey FA, Simon-Loriere E, Ball JK. 2016. Human adaptation of Ebola virus during the West African Outbreak. *Cell* **167**(4):1079–1087.e5.

van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CCS, Boshier FAT, *et al.* 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol*. **83**:104351.

van Dorp L, Richard D, Tan CCS, Shaw LP, Acman M, Balloux F. 2020. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nat Commun*. **11**(1): 5986.

Wang H-Y, Chien H-C, Osada N, Hashimoto K, Sugano S, Gojobori T, Chou C-K, Tsai S-F, Wu C-I, Shen C-K. 2007. Rate of evolution in brain-expressed genes in humans and other primates. *PLoS Biol*. **5**(2):e13.

Welkers MRA, Han AX, Reusken CBEM, Eggink D. 2021. Possible host-adaptation of SARS-CoV-2 due to improved ACE2 receptor binding in mink. *Virus Evol*. **7**(1):veaa094.

Worobey M, Santiago ML, Keele BF, Ndjango J-BN, Joy JB, Labama BL, Dhed'a BD, Rambaut A, Sharp PM, Shaw GM, *et al*. 2004. Contaminated polio vaccine theory refuted. *Nature* **428**(6985): 820.

Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y, *et al*. 2020. A new coronavirus associated with human respiratory disease in China. *Nature* **579**(7798):265–269.

Wu K, Peng G, Wilken M, Geraghty RJ, Li F. 2012. Mechanisms of host receptor adaptation by severe acute respiratory syndrome coronavirus. *J Biol Chem*. **287**(12):8904–8911.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. **24**(8):1586–1591.

Yeh S-H, Wang H-Y, Tsai C-Y, Kao C-L, Yang J-Y, Liu H-W, Su I-J, Tsai S-F, Chen D-S, Chen P-J, *et al*. 2004. Characterization of severe acute respiratory syndrome coronavirus genomes in Taiwan: molecular epidemiology and genome evolution. *Proc Natl Acad Sci USA* **101**(8):2542–2547.

Zeng K, Fu Y-X, Shi S, Wu C-I. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* **174**(3):1431–1439.

Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, *et al*. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**(7798): 270–273.

Zhu T, Korber BT, Nahmias AJ, Hooper E, Sharp PM, Ho DD. 1998. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature* **391**(6667):594–597.