Received: 12 August 2020 | Revised: 5 March 2021 | Accepted: 5 May 2021

DOI: 10.1111/tbed.14148

ORIGINAL ARTICLE

Transboundary and Emerging Diseases WILEY

Tracing genetic signatures of bat-to-human coronaviruses and early transmission of North American SARS-CoV-2

Xumin Ou[1,2,3] | Zhishuang Yang[1,3,4] | Dekang Zhu[1,3] | Sai Mao[1,3,4] | Mingshu Wang[1,3,4] | Renyong Jia[1,3,4] | Shun Chen[1,3,4] | Mafeng Liu[1,3,4] | Qiao Yang[1,3,4] | Ying Wu[1,3,4] | Xinxin Zhao[1,3,4] | Shaqiu Zhang[1,3,4] | Juan Huang[1,3,4] | Qun Gao[1,3,4] | Yunya Liu[1,3,4] | Ling Zhang[1,3,4] | Maikel Peppelenbosch[2] | Qiuwei Pan[2,5] | Anchun Cheng[1,3,4]

[1]Institute of Preventive Veterinary Medicine, Sichuan Agricultural University, Chengdu, China

[2]Department of Gastroenterology and Hepatology, Erasmus MC - University Medical Center Rotterdam, Rotterdam, The Netherlands

[3]Key Laboratory of Animal Disease and Human Health of Sichuan Province, Sichuan Agricultural University, Chengdu, China

[4]Research Center of Avian Diseases, College of Veterinary Medicine, Sichuan Agricultural University, Chengdu, China

[5]Biomedical Research Center, Northwest Minzu University, Lanzhou, China

Correspondence
Anchun Cheng, Institute of Preventive Veterinary Medicine, Sichuan Agricultural University, Chengdu city, Sichuan, China.
Email: chenganchun@vip.163.com

Funding information
China Agricultural Research System, Grant/Award Number: CARS-42-17; Integration and Demonstration of Key Technologies for Goose Industrial Chain in Sichuan Province, Grant/Award Number: 2018NZ0005; Science and Technology Program of Sichuan Province, Grant/Award Number: 2020YJ0396; Sichuan Veterinary Medicine and Drug Innovation Group of China Agricultural Research System, Grant/Award Number: SCCXTD-2020-18

Abstract

Highly pathogenic coronaviruses, including SARS-CoV-2, SARS-CoV and MERS-CoV, are thought to be transmitted from bats to humans, but the viral genetic signatures that contribute to bat-to-human transmission remain largely obscure. In this study, we identified an identical ribosomal frameshift motif among the three bat–human pairs of viruses and strong purifying selection after jumping from bats to humans. This represents genetic signatures of coronaviruses that are related to bat-to-human transmission. To further trace the early human-to-human transmission of SARS-CoV-2 in North America, a geographically stratified genome-wide association study (North American isolates and the remaining isolates) and a retrospective study were conducted. We determined that the single nucleotide polymorphisms (SNPs) 1,059.C > T and 25,563.G > T were significantly associated with approximately half of the North American SARS-CoV-2 isolates that accumulated largely during March 2020. Retrospectively tracing isolates with these two SNPs was used to reconstruct the early, reliable transmission history of North American SARS-CoV-2, and European isolates (February 26, 2020) showed transmission 3 days earlier than North American isolates and 17 days earlier than Asian isolates. Collectively, we identified the genetic signatures of the three pairs of coronaviruses and reconstructed an early transmission history of North American SARS-CoV-2. We envision that these genetic signatures are possibly diagnosable and predic markers for public health surveillance.

KEYWORDS

genetic signatures, geographic transmission, GWAS, SARS-CoV-2

Xumin Ou, Zhishuang Yang and Anchun Cheng contributed equally to this work.

1 | INTRODUCTION

Since 2003, coronaviruses (CoVs), specifically, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), severe acute respiratory syndrome-related coronavirus (SARS-CoV, 2003) and Middle East respiratory syndrome-related coronavirus (MERS-CoV, 2012), have caused three epidemics in human populations worldwide, including the ongoing COVID-19 pandemic caused by SARS-CoV-2 (Perlman, 2020; P. Zhou et al., 2020). Interestingly, these CoVs are thought to be associated with bat CoVs; for instance, human SARS-CoV-2 shares 96% nucleotide identity with bat CoVs (W. Li et al., 2005; P. Zhou et al., 2020). Human-to-human transmission of SARS-CoV-2 was confirmed on 14 January 2020 (Wu, Leung, & Leung, 2020). On 11 March 2020, the World Health Organization (WHO) officially declared that human COVID-19 caused by SARS-CoV-2 was a global pandemic. As of May 31, over 5.93 million human cases have been confirmed globally, of which over 2.74 million cases were from America, particularly North America (Wu et al., 2020). Thus, it is very important to understand the cross-species transmission of SARS-CoV-2 and its human-to-human transmission in America to inform public health measures (Ji et al., 2020).

It is well known that viral genetic variations are associated with many aspects of virology, most notably viral infectivity and zoonotic transfer (Ou et al., 2019; J. H. Zhou et al., 2019). Identifying distinctive genetic signatures of CoVs common to those viruses found in different host species (human and bat) as well as different geographic regions (North America and the rest) may provide differential markers to support epidemiological surveillance. However, these distinctive signatures are currently unknown. SARS-CoV-2 massively mutates, which hampers its transmission tracing (Tang et al., 2020). Herein, we aimed to identify the genetic signatures of three pairs of bat-to-human CoVs as well as those of the North American SARS-CoV-2 isolates associated with the early human-to-human transmission history of COVID-19. The whole genomic sequences of three bat–human pairs of SARS-CoV-2 were analysed as well as bat–human pairs of SARS-CoV and MERS-CoV; the latter also included a MERS-CoV strain isolated from camel (Azhar et al., 2014). After a virus jumps to a human host, new mutations are fixed in the viral genome that may be geographically different. Identification of these fixed and common mutations remains a great challenge because of the complexity of these mutable viruses. In human population genetics studies, this type of complexity can be particularly addressed by genome-wide association studies (GWASs) (Power et al., 2017). Therefore, we aimed to use a geographically stratified GWAS to address the complexity of global SARS-CoV-2 isolates.

We primarily found that the genomic organization of the three human CoVs was similar to that of their paired bat CoVs within lineages and underwent strong purifying selection after jumping to the human host. For the early human-to-human transmission of North American SARS-CoV-2, we identified that two SNPs of complete linkage disequilibrium were exclusively present in more than half of the North American SARS-CoV-2-dominated lineage B.1. By retrospectively tracing isolates with these two SNPs, an early transmission history of North American SARS-CoV-2 isolates was reconstructed.

2 | MATERIALS AND METHODS

2.1 | Data acquisition

Human SARS-CoV-2 isolated from Wuhan, China, was obtained from the NCBI database (GenBank No.: MN908947.3). To identify the phylogenetically closed bat CoVs associated with human SARS-CoV-2, the BLAST searching tool of the NCBI viral database was used. Based on the constructed phylogenetic tree, we identified two bat SARS-like CoVs (referred to here as bat SARS-CoV-2) (GenBank No: MG772933.1 and MG772934.1) that were evolutionarily close to human SARS-CoV-2 (Table S1) (Fig. S1). For human SARS-CoVs and MERS-CoVs, similar approaches were used to identify their bat CoV pairs; the camel MERS-CoV-related literature was also reviewed (Madani et al., 2014).

For the GWAS, full-genome sequences of global SARS-CoV-2 collected from 12 December 2019 to 24 April 2020 (8:00 GMT +8) were archived from the database of the GISAID Initiative EpiCoV platform (GISAID; https://www.epicov.org). A total of 8,480 sequences were archived and filtered by criteria, including high coverage only (> 29,000 bp, 1X coverage of genome), exclusion of low coverage and sequences with unconfident bases (N) inside. The PANGOLIN isolate (EPI_ISL_410539) and bat CoVRaTG13 isolate were also used for phylogenetic analysis. The identical sequences were further removed by CD-HIT software (version 4.8.1, parameters: -aL 1 -aS 1 -c 1 -s 1) (Huang et al., 2010). A final 2,599 sequences were used in this study (data file S1).

2.2 | Sequence alignment

A codon-based Cluster W method was used for the multiple sequence alignment and identification of the ribosomal frameshift motifs among the three CoV bat–human pairs. The genomic organization of annotated CoVs was visualized by SnapGene (Version 4.2.4).

2.3 | Phylogenetic analysis

For the accessory and necessary protein-coding genes, phylogenetic trees were constructed by the maximum likelihood method. The evolutionary distances were calculated by the differences between amino acid substitutions or nucleotide substitutions per site. Confidence probability was estimated by the bootstrap test (100 replicates).

The 2,599 full-genome sequences were aligned by MAFFT software (version 7.407, algorithm: FFT-NS-2) (Nakamura et al., 2018). The phylogenetic tree was constructed by IQ-TREE 2 (version 2.1.2, parameter: -nt -gtr -gamma) using the GTR+Γ model of nucleotide

TABLE 1 List of top 21 hits of causative SNPs

#	SNP	Gene	Effects	*p*-value	SNP Frequencies (%)				
					North America	Lineage B*	Lineage B.1*	Europe	Asia
1	**25,563.G > T**	**ORF3a**	**Missense(Gln57His)**	**2.98E−261**	**574/1063(54%)**	**574/818(70%)**	**573/691(83%)**	**119/951(13%)**	**19/448(4%)**
2	**1,059.C > T**	**ORF1ab**	**Missense(Thr265Ile)**	**2.44E−212**	**479/1063(45%)**	**479/818(59%)**	**479/691(69%)**	**94/951(10%)**	**0/448(0%)**
3	17,858.A > G	ORF1ab	Missense(Met5865Val)	3.57E−117	194/1063(18%)	0/818(0%)	0/691(0%)	0/951(0%)	0/448(0%)
4	17,747.C > T	ORF1ab	Synonymous	1.43E−116	193/1063(18%)	0/818(0%)	0/691(0%)	0/951(0%)	4/448(1%)
5	18,060.C > T	ORF1ab	Missense(Ser5932Phe)	2.42E−113	196/1063(18%)	1/818(0%)	0/691(0%)	0/951(0%)	0/448(0%)
6	29,553.G > A	ORF10	Upstream	7.26E−51	80/1063(8%)	80/818(10%)	80/691(12%)	0/951(0%)	30/448(7%)
7	28,882.G > A	*N*	Synonymous	6.15E−39	53/1063(5%)	53/818(6%)	53/691(8%)	207/951(22%)	30/448(7%)
8	28,883.G > C	*N*	Missense(Gly204Arg)	6.15E−39	53/1063(5%)	53/818(6%)	53/691(8%)	207/951(22%)	30/448(7%)
9	28,881.G > A	*N*	Missense(Arg203Lys)	2.77E−38	54/1063(5%)	54/818(7%)	54/691(8%)	207/951(22%)	87/448(19%)
10	28,144.T > C	ORF8	Missense(Leu84Ser)	9.73E−33	245/1063(23%)	0/818(0%)	0/691(0%)	49/951(5%)	0/448(0%)
11	27,964.C > T	ORF8	Missense(Ser24Leu)	2.74E−29	47/1063(4%)	47/818(6%)	47/691(7%)	0/951(0%)	2/448(0%)
12	11,916.C > T	ORF1ab	Missense(Ser3884Leu)	4.87E−29	51/1063(5%)	51/818(6%)	51/691(7%)	0/951(0%)	96/448(21%)
13	8,782.C > T	ORF1ab	Synonymous	4.03E−28	245/1063(23%)	0/818(0%)	0/691(0%)	53/951(6%)	9/448(2%)
14	15,324.C > T	ORF1ab	Missense(Thr5020Ile)	5.83E−27	2/1063(0.2%)	2/818(0%)	2/691(0%)	80/951(8%)	130/448(29%)
15	11,083.G > T	ORF1ab	Missense(Leu3606Phe)	3.99E−25	71/1063(7%)	65/818(8%)	5/691(1%)	92/951(10%)	2/448(0%)
16	18,998.C > T	ORF1ab	Missense(His6245Tyr)	1.10E−22	41/1063(4%)	41/818(5%)	41/691(6%)	0/951(0%)	2/448(0%)
17	29,540.G > A	ORF10	Upstream	1.10E−22	41/1063(4%)	41/818(5%)	41/691(6%)	0/951(0%)	8/448(2%)
18	18,877.C > T	ORF1ab	Synonymous(His6245Tyr)	2.31E−19	65/1063(6%)	63/818(8%)	62/691(9%)	10/951(1%)	0/448(0%)
19	29,711.G > T	5'UTR	Downstream	4.30E−19	31/1063(3%)	31/818(4%)	1/691(0%)	0/951(0%)	1/448(0%)
20	1604.AATG>A	ORF1ab	Deletion(delTGA)	2.87E−17	4/1063(0.4%)	4/818(0%)	0/691(0%)	69/951(7%)	2/448(0%)
21	27,046.C > T	M	Missense(Thr175Met)	3.48E−17	2/1063(0.2%)	2/818(0%)	2/691(0%)	60/951(6%)	60/951(13%)

The 'bold values' refers to the two SNPs used for the reconstruction of early transmission history.

*North America lineage B or B.1.

substitution (Minh et al., 2020). The phylogenetic tree was visualized by FigTree (version 1.4.4, http://tree.bio.ed.ac.uk/software/figtree/). Each descendant lineage was annotated according to criteria from recent publications (Rambaut et al., 2020).

2.4 | Mutation analysis

Synonymous and non-synonymous differences per sequence between human and bat CoVs were estimated using the Nei-Gojobori model by MEGA-X software (Table S2–11). The dN/dS ratio is an indicator of directional selection: a ratio above 1 implies positive selection (nature), a ratio less than 1 implies negative selection (purifying), and a ratio equal to 1 indicates no selection (neutral). The dN/dS ratio (Table S12–16) is calculated by the following equations:

$$d_N = -\frac{3}{4}\ln\left(1 - 4\frac{p_N}{3}\right)$$

$$d_S = -\frac{3}{4}\ln\left(1 - 4\frac{p_S}{3}\right)$$

$$d_N/d_S = \frac{d_N}{d_S}$$

2.5 | Codon usage bias analysis

Relative synonymous codon usage (RSCU) of CoV necessary protein-coding genes (i.e. ORF1ab-S-E-M-N) was analysed by CODONW software (http://www.molbiol.ox.ac.uk/cu, version 1.4.2) using standard genetic codes. The linear regression of RSCU between bat–human CoV pairs was analysed by GraphPad Prism 8.0.

2.6 | SNP calling

SNPs and INDEL polymorphisms were detected by MUMmer software (version 3.0, nucmer, show-snps) (Kurtz et al., 2004) using the Wuhan-Hu-1 strain (GISAID: EPI_ISL_402125, GenBank: NC_045512.2) as a reference genome. To validate the identity of the resulting polymorphisms, raw reads (40 out of 2,599 strains, NCBI SRA database) were analysed by the bwa program (version 0.7.16a) (H. Li & Durbin, 2010) and the mpileup program of the SAMtools software (version 1.10) (H. Li, 2011). The validation was consistent with the polymorphisms detected by MUMmer software.

2.7 | GWAS and linkage disequilibrium (LD) analysis

To identify causative SNPs in the population of North American SARS-CoV-2 (cases = 1,063, controls = 1536), a geographically stratified genome-wide association study of 5,312 mutations was performed using PLINK software (version 1.90) (Purcell et al., 2007). The empirical threshold of the p-value was suggested to be 9.41×10^{-6} ($0.05/5312 = 9.41 \times 10^{-6}$) calculated by the (Benjamini & Hochberg, 1995), but we further increased the threshold of the p-value to 1.00×10^{-15} to detect the most causative SNPs. The top 21 significant SNPs were listed (Table 1), and the LD of pairing SNPs was estimated and visualized by Haploview software (version 4.1) (Barrett et al., 2005).

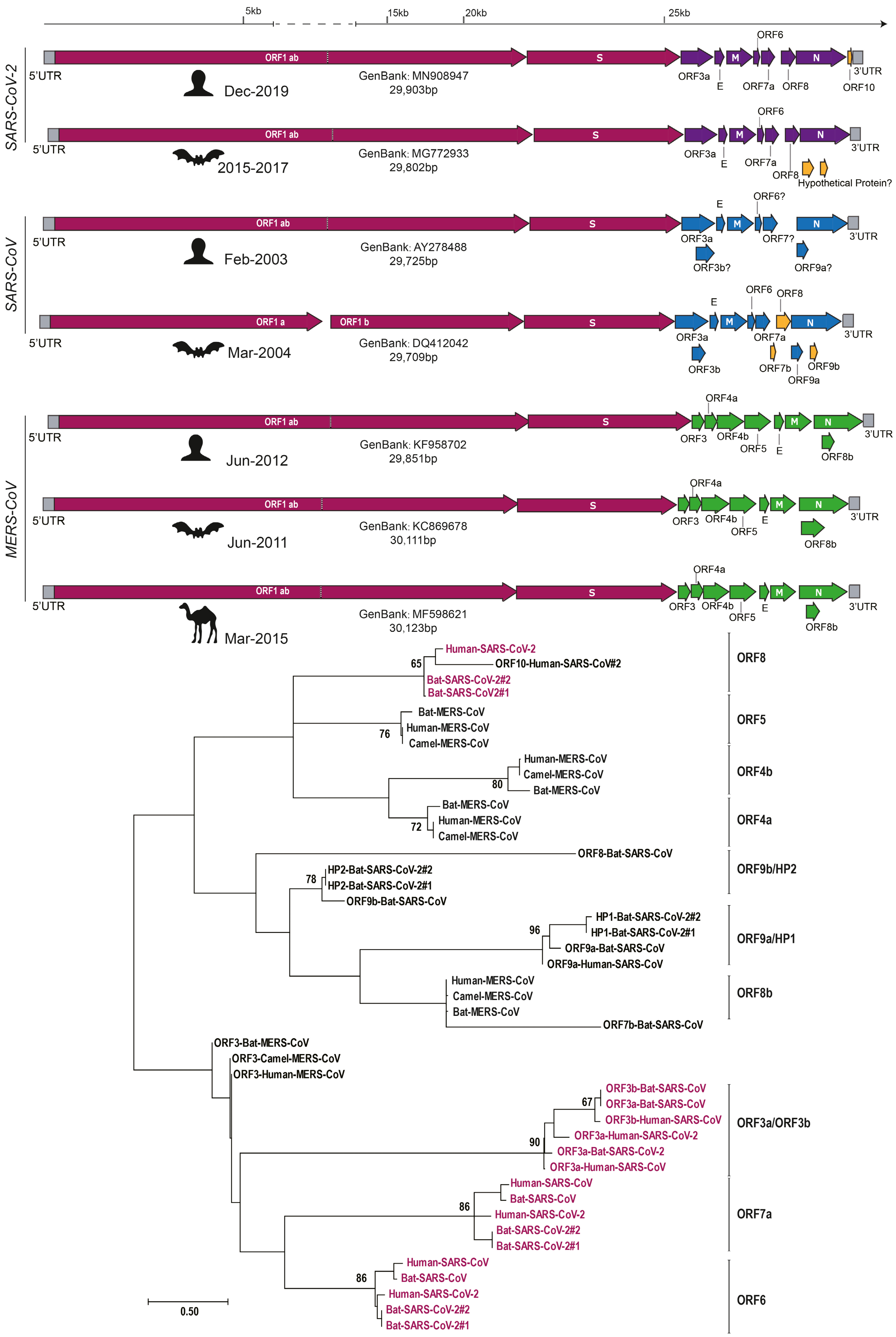
2.8 | SNP accumulating analysis

To analyse the trend of SNP accumulation during March 2020, the frequencies of average SNP accumulation per day were counted. This same trend in 1,059.C > T and 25,563.G > T in North American SARS-CoV-2 and that of the other continents was traced by the date of occurrence of the two SNPs. The same analysis of these two SNPs in North American lineage B.1 was also conducted. These analyses were performed by Microsoft® Excel 2016 (data file S2).

2.9 | Statistical analysis

The probability of rejecting the null hypothesis of strict neutrality (dN = dS) in favour of the alternative hypothesis (dN < dS) was calculated by the codon-based Z test of purifying selection. Data from the SNP accumulation analysis were plotted by GraphPad (Version 8.2.1). The mean differences of all types of SNP accumulation per day per strain were determined by the Mann–Whitney U test (interval = 10 days) (R version 3.6.2). *p* values less than .05 were considered significant.

FIGURE 1 Genomic signatures of bat–human SARS-CoV-2, SARS-CoV and MERS-CoV pairs. For the three CoV pairs, the necessary proteins are encoded in the same order in the genome (i.e. ORF1ab-S-E-M-N). The accessory protein-encoding genes vary by locus. SARS-CoV-2 and SARS-CoV have gene insertions between the M protein and N protein-coding genes, such as ORF6, ORF7a(b) and ORF8, while the same locus has no insertion in MERS-CoV. Of note, a novel ORF10 (yellow box) of human SARS-CoV-2 is less related to any human or CoV gene. Importantly, ORF6 and ORF7a of human SARS-CoV-2 are related to the equivalents of SARS-CoV isolated from both bats and humans (lower panel). Two new hypothetical proteins (i.e. HP1 and HP2) (yellow box) of bat SARS-CoV-2 are evolutionarily close to ORF9a and ORF9b of SARS-CoV. Bat-SARS-CoV-2 #1 and #2 represent two related bat CoVs. The phylogenetic tree was constructed by the maximum likelihood method. The evolutionary distances are calculated by base differences per site. Confidence probability was estimated using the bootstrap test (100 replicates)


5kb
15kb
20kb
25kb
SARS-CoV-2
5'UTR
ORF1 ab
S
ORF3a
E
M
ORF6
ORF7a
ORF8
N
ORF10
3'UTR
Dec-2019
GenBank: MN908947
29,903bp
2015-2017
GenBank: MG772933
29,802bp
Hypothetical Protein?
SARS-CoV
ORF6?
ORF7?
ORF3b?
ORF9a?
Feb-2003
GenBank: AY278488
29,725bp
ORF1 a
ORF1 b
ORF3b
ORF7b
ORF9b
ORF9a
Mar-2004
GenBank: DQ412042
29,709bp
MERS-CoV
ORF3
ORF4a
ORF4b
ORF5
ORF8b
Jun-2012
GenBank: KF958702
29,851bp
Jun-2011
GenBank: KC869678
30,111bp
Mar-2015
GenBank: MF598621
30,123bp
65
Human-SARS-CoV-2
ORF10-Human-SARS-CoV#2
Bat-SARS-CoV-2#2
Bat-SARS-CoV2#1
ORF8
76
Bat-MERS-CoV
Human-MERS-CoV
Camel-MERS-CoV
ORF5
80
Human-MERS-CoV
Camel-MERS-CoV
Bat-MERS-CoV
ORF4b
72
Bat-MERS-CoV
Human-MERS-CoV
Camel-MERS-CoV
ORF4a
ORF8-Bat-SARS-CoV
78
HP2-Bat-SARS-CoV-2#2
HP2-Bat-SARS-CoV-2#1
ORF9b-Bat-SARS-CoV
ORF9b/HP2
96
HP1-Bat-SARS-CoV-2#2
HP1-Bat-SARS-CoV-2#1
ORF9a-Bat-SARS-CoV
ORF9a-Human-SARS-CoV
ORF9a/HP1
Human-MERS-CoV
Camel-MERS-CoV
Bat-MERS-CoV
ORF8b
ORF7b-Bat-SARS-CoV
ORF3-Bat-MERS-CoV
ORF3-Camel-MERS-CoV
ORF3-Human-MERS-CoV
67
ORF3b-Bat-SARS-CoV
ORF3a-Bat-SARS-CoV
ORF3b-Human-SARS-CoV
90
ORF3a-Human-SARS-CoV-2
ORF3a-Bat-SARS-CoV-2
ORF3a-Human-SARS-CoV
ORF3a/ORF3b
86
Human-SARS-CoV
Bat-SARS-CoV
Human-SARS-CoV-2
Bat-SARS-CoV-2#2
Bat-SARS-CoV-2#1
ORF7a
86
Human-SARS-CoV
Bat-SARS-CoV
Human-SARS-CoV-2
Bat-SARS-CoV-2#2
Bat-SARS-CoV-2#1
ORF6
0.50

3 | RESULTS

3.1 | Genomic organization of bat–human CoV pairs

CoVs are positive single-stranded RNA viruses with a non-segmented genome. The genome encodes a fixed array of necessary proteins (NPs), in the order ORF1ab, spike (S) protein, envelope (E) protein, membrane glycoprotein (M) protein and nucleocapsid (N), as well as accessory proteins (APs) that differ by number and order among closely related CoVs (Figure 1). We found that the genomic organization of each bat–human CoV pair was similar , as well as that of MERS-CoV between humans, bats and camels. For the NPs, the genomic organization among all three CoVs followed the same order (i.e. ORF1ab-S-E-M-N) (Figure 1). The loci of the APs were largely different, which was likely caused by gene recombination (Figure 1). Specifically, for SARS-CoV-2 and SARS-CoV, the ORF6, ORF7 and ORF8 genes are equally inserted between the M gene and the N gene. However, for MERS-CoV, this location has no gene insertion.

For the APs, the phylogenetic analysis suggested that the APs of human SARS-CoV-2 are evolutionally close to those of bat SARS-CoV-2 (i.e. ORF3a, ORF6, ORF7a and ORF8), in which ORF6 and ORF7a are also close to those of bat or human SARS-CoV (Figure 1, Fig. S2 and S3). Strikingly, the human SARS-CoV-2 genome contains a novel ORF10 that is less related to any other human or bat CoV ORF.

3.2 | Identical ribosomal frameshift motif between bat–human CoV pairs

For all CoVs, a programmed −1 ribosomal frameshift signal is essential, as it controls viral translation. The slippage signal is characterized by an X_3Y_3Z motif (X_3, any three identical nucleotides; Y_3, typically UUU or AAA; Z, A, C or U). We found that the slippage signal, U_UUA_AAC, was identical among SARS-CoV-2, SARS-CoV and MERS-CoV (Baranov et al., 2005) (Fig. S4). The slippage of ribosomal frameshifting from U_UUA_AAC to UUU_AAA_C does not change the growing peptides, as they both encode identical dipeptides because of the degeneracy of codon position 3. For the two flanking motifs, the 5' attenuator hairpin and 3' frameshift-stimulating three-stemmed pseudoknots are relatively conserved in both SARS-CoV-2 and SARS-CoV. However, inside the second flanking motif of MERS-CoV, an insertion of the AAT codon (encoding asparagine) was newly identified (Fig. S4). Prior research conducted by mutating this slippage signal to C_CUC_AAC shows thorough inhibition of the ribosomal frameshift (Kelly et al., 2020).

3.3 | Strong purifying selection of CoVs

Zoonotic transfer of CoVs involves mutagenesis and directional selection (Forni et al., 2017). During the 2002–2004 epidemic, SARS-CoV mutated extensively, which enhanced its virulence (Consortium, 2004). To measure which type of mutation is linked to CoV virulence, synonymous differences (dSs) and non-synonymous differences (dNs) were analysed between intro- and extra-branches (Figure 2). The number of dSs between CoVs of intro-branches is much higher than that of dNs, which is similar among all NPs (i.e. ORF1ab-S-E-M-N) (Table S2–11). For instance, for ORF1ab, the number of dSs between bat and human SARS-CoV-2 (dS = 1875.75) is approximately four times higher than the number of dNs (dN = 441.25). The trend is the same for SARS-CoV and MERS-CoV. The number of dSs and dNs within intro-branches of CoVs is smaller than the number within extra-branches. This is consistent with the evolutionary distances of phylogenetic trees (Figure 2).

Because massively synonymous mutations do not modify the protein sequence but change the overall codon usage bias, we performed linear regression of relative synonymous codon usage (RSCU) to see whether bat CoVs show codon usage bias to human CoVs. However, only a slight RSCU shift from bat-to-human CoVs was observed, as the slope of the linear regression was slightly smaller than 1 (Fig. S5). Collectively, the large synonymous mutations were augmented, which may enhance SARS-CoV-2 virulence, similar to SARS-CoV, 2003.

The dN/dS ratio is a classic indicator of directional selection: a ratio above 1 implies positive selection (nature), a ratio less than 1 implies negative selection (purifying), and a ratio equal to 1 indicates no selection (neutral) (Kryazhimskiy & Plotkin, 2008). In contrast, purifying selection involves more synonymous mutations than non-synonymous mutations. As discussed, this is true for the three bat–human CoV pairs (Table S2–11). Purifying selection primarily changes viral codon usage bias and thus can regulate viral virulence via optimization of a specific codon context (Coleman et al., 2008; Hanson & Coller, 2017). For ORF1ab, the dN/dS ratio between human and bat SARS-CoV-2 is 0.05, which is much lower than that of SARS-CoV-2 and MERS-CoV (Table S12–16). A similar trend was confirmed in the other NPs. This is supported by a viral culture experiment, as SARS-CoV-2 grows better than SARS-CoV and MERS-CoV in human cells (Perlman, 2020). Using the codon-based Z test of selection, the statistic shows that human SARS-CoV-2 undergoes significantly strong purifying selection (Table S12–16).

3.4 | Phylogenetics of global SARS-CoV-2 reveals that early North American isolates dominate lineage B.1

To understand the early human-to-human transmission of SARS-CoV-2 in North America, a phylogenetic analysis of the global SARS-CoV-2 population (2,599 strains with high confidence) was conducted. We found that global SARS-CoV-2 was rooted in two lineages, lineages A (*n* = 413) and B (*n* = 2,186), in which North American isolates dominated lineage B (*n* = 818) and lineage B.1 (*n* = 691) (Figure 3 and Table 1) (Rambaut et al., 2020). Importantly, the phylogenetic tree was inferred by producing mutations, and the identification of key mutations can provide clues for tracing the transmission route of SARS-CoV-2.

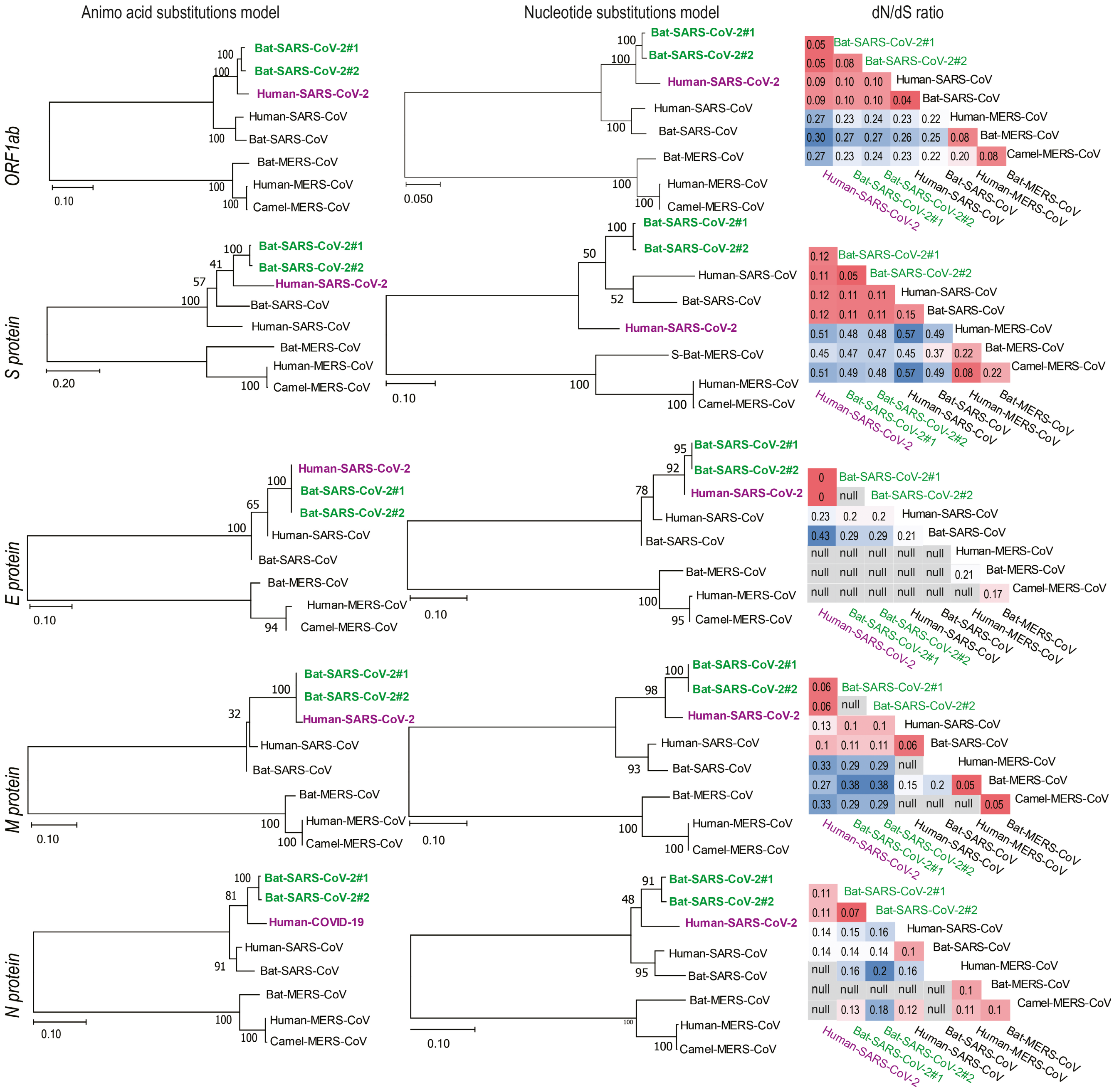


FIGURE 2 Evolutionary signatures of necessary protein-encoding genes of SARS-CoV-2. The differences between amino acid substitutions were used to construct phylogenetic trees (left panel). The same analysis was also performed by the differences in nucleotide substitutions (middle panel). dN/dS ratio matrixes are displayed (Tables S12–16) (right panel). The sequential analysis of necessary proteins was performed from top to bottom (i.e. ORF1ab-S-E-M-N). The phylogenetic tree was constructed by the maximum likelihood method. Bat-SARS-CoV-2 #1 and #2 represent two related bat CoVs. The evolutionary distances are calculated by the differences between amino acid substitutions or nucleotide substitutions per site. Confidence probability was estimated using the bootstrap test (100 replicates)

3.5 | Geographic GWAS reveals SNPs associated with North American isolates

Calling key SNPs from massive mutations of the SARS-CoV-2 population requires a GWAS that has been learned from a human GWAS (Power et al., 2017). Because of the complexity of the phylogenetic tree, a phylogenetically stratified GWAS may not be feasible. Therefore, a geographically stratified GWAS was carried out, as the geographic location of individual isolates was reliable. The mutation features of SARS-CoV-2 between continents may reflect the incidence of emergence of a given viral population in different human hosts (Rambaut et al., 2020). By using a geographically stratified GWAS comparing North American isolates (*n* = 1,063) with the remaining isolates (*n* = 1536), we found 21 significant

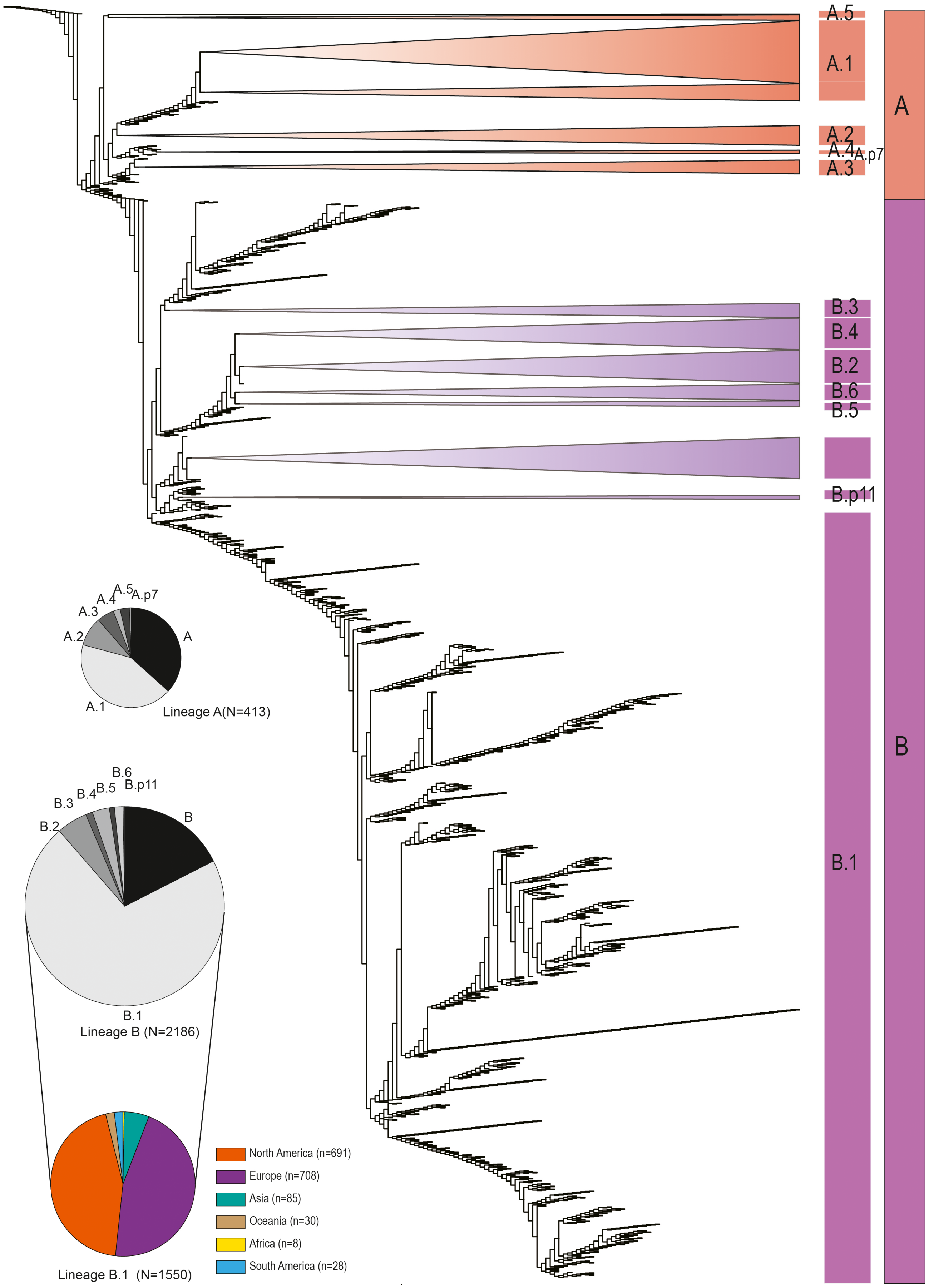

A.5
A.1
A.2
A.4
A.p7
A.3
A
B.3
B.4
B.2
B.6
B.5
B.p11
B.1
B
A.5
A.4
A.p7
A.3
A.2
A
A.1
Lineage A(N=413)
B.6
B.4
B.5
B.p11
B.3
B.2
B
B.1
Lineage B (N=2186)
Lineage B.1 (N=1550)
North America (n=691)
Europe (n=708)
Asia (n=85)
Oceania (n=30)
Africa (n=8)
South America (n=28)

FIGURE 3 Phylogenetic tree of early global SARS-CoV-2. The 2,599 full-genome sequences were used to construct the phylogenetic tree via the maximum likelihood (ML) method using IQ-TREE 2 software (version 2.1.2, model: GTR+Γ). Accordingly, the early SARS-CoV-2 isolates were rooted in two lineages, lineages A ($n = 413$) and B ($n = 2{,}186$), in which North American isolates dominated lineage B ($n = 818$) and sub-lineage B.1 ($n = 691$). The constituents of the main lineages A and B as well as lineage B.1 are displayed by three pie charts. Specifically, North American and European isolates dominate lineage B.1

SNPs or small insertion deletions (INDELs) out of 5,312 (threshold p-value =1.00×10^{-15}) (Figure 4a). Specifically, the top two SNPs (i.e. 1,059.C > T and 25,563.G > T) were present in approximately half of North American SARS-CoV-2 isolates (479/1063 = 45% and 574/1063 = 54%), particularly North American lineage B.1 (479/691 = 69% and 573/691 = 83%) (Table 1). Interestingly, the two SNPs were in complete linkage disequilibrium, suggesting that the two SNPs concurrently occurred in the North American dominating lineage B.1 (479/691, 69%) (Figure 4b). Importantly, the two SNPs resulted in two mutations (i.e. Thr265 Ile and Gln57 His) in ORF1ab and ORF3a, respectively.

Among these 21 SNPs, we also identified two previously reported SNPs, 8,782.C > T and 28,144.T > C (p-value = 4.03×10^{-28} and 9.73×10^{-33}), resulting in a synonymous mutation and a missense mutation (Leu 84 Ser) (Tang et al., 2020) (Table 1). Interestingly, three sequential SNP sites (28881–3.GGG>ACA) were fixed in 22% (207/951) of the European SARS-CoV-2 isolates, resulting in a synonymous mutation and two missense mutations (Arg 203 Lys and Gly 204 Arg) (Table 1). Tracing these three SNPs showed that the recent reemergence of COVID-19 in the Xinfadi market in Beijing, China, was associated with European isolates (Wenjie et al., 2020).

3.6 | SNP tracing reconstructed an early transmission history of North American isolates

In the North American SARS-CoV-2 population, 45% of strains have these two SNPs, and 69% have these SNPs for North American lineage B.1 (Table 1). Because of the high occurrence of the two SNPs, tracing the two SNPs may provide a reliable transmission route of SARS-CoV-2 in the major North American human population. We thus performed a retrospective tracing study in our high confidential data sets (2,599 flittered strains) to identify the time order of isolates occurring at the two SNPs on all continents and in lineage B.1. We found that the first isolate started in Europe (26 February 2020) 3 days earlier than the occurrence date of the North American isolates (29 February 2020) and 17 days earlier than the Asian isolates (Taiwan China dominated) (13 March 2020) (Figure 5a). By further tracing the accumulating frequencies per day of the two SNPs during mid to late March, we found that North American lineage B.1 highly accumulated these two SNPs (Figure 5b). In addition, the mean number of all accumulating SNPs during mid to late March was significantly lower than that before or after the same period (Fig. S6). This evidence indicated that the two SNPs were strongly selected in the North American SARS-CoV-2 isolates, in particular lineage B.1 from mid to late March. The accumulation of the two SNPs may explain the sharp increase in confirmed cases in North America before early April (WHO reported) (WHO, 2020).

SARS-CoV-2 is thought to be transmitted from wildlife to humans before human-to-human transmission occurs (Andersen et al., 2020; Shi et al., 2020; Tang et al., 2020). We carefully checked whether bat or pangolin CoVs evolved with the two mutations before they jumped to the human species. However, we did not find either of these SNPs in the bat- or pangolin-related CoVs (P. Zhou et al., 2020) (Fig. S7). Alternatively, in bat or pangolin CoVs, the 1,059 site has no or a C > A variant, and the 25,563 site has a G > A variant instead (Fig. S7). (P. Zhou et al., 2020).

4 | DISCUSSION

Herein, we identified the genetic signatures of bat-to-human CoVs and specified an early transmission history of North American SARS-CoV-2. Although human CoVs are highly similar to bat CoVs by sequence and genome organization (Perlman, 2020; P. Zhou et al., 2020), several specific genetic signatures were newly identified in this study, such as a unique ORF10 in human SARS-CoV-2, an identical ribosomal frameshift motif, and strong purifying selection after zoonotic transfer. In addition, we also found that the two causative SNPs that were present in approximately half of the North American SARS-CoV-2 isolates represented 69% of the isolates of North American lineage B.1 (Rambaut et al., 2020). The early transmission history of the major North American SARS-CoV-2 isolates was reconstructed by tracing the occurrence date of isolates with these two SNPs, and transmission started in Europe, North America, South America, and later Asia and Oceania.

The genetic signature and its extent determine the bat-to-human cross-species transmission of CoVs, which is still largely undocumented. However, distinctive genetic signatures can possibly predict and estimate the risk of zoonotic transmission. The unique ORF10 of human SARS-CoV-2 and the insertion of the AAT codon in the slippage signal of MERS-CoV could serve as novel targets for differential diagnosis. Importantly, human SARS-CoV-2 as well as SARS-CoV and MERS-CoV undergo strong purifying selection. Strong purifying selection involving large synonymous mutations may promote fitness in the human system by regulating viral translation efficiency (Ou et al., 2018). Monitoring the mutation rate in particular synonymous mutations and its possible impact would help to predict the risk of zoonotic transfer of bat CoVs.

Following zoonotic transfer, understanding the trend of SARS-CoV-2 geographic transmission is important for public measures (Centers for Disease Control and Prevention-USA, 2020). This study was published as a BioRvix preprint and gained the

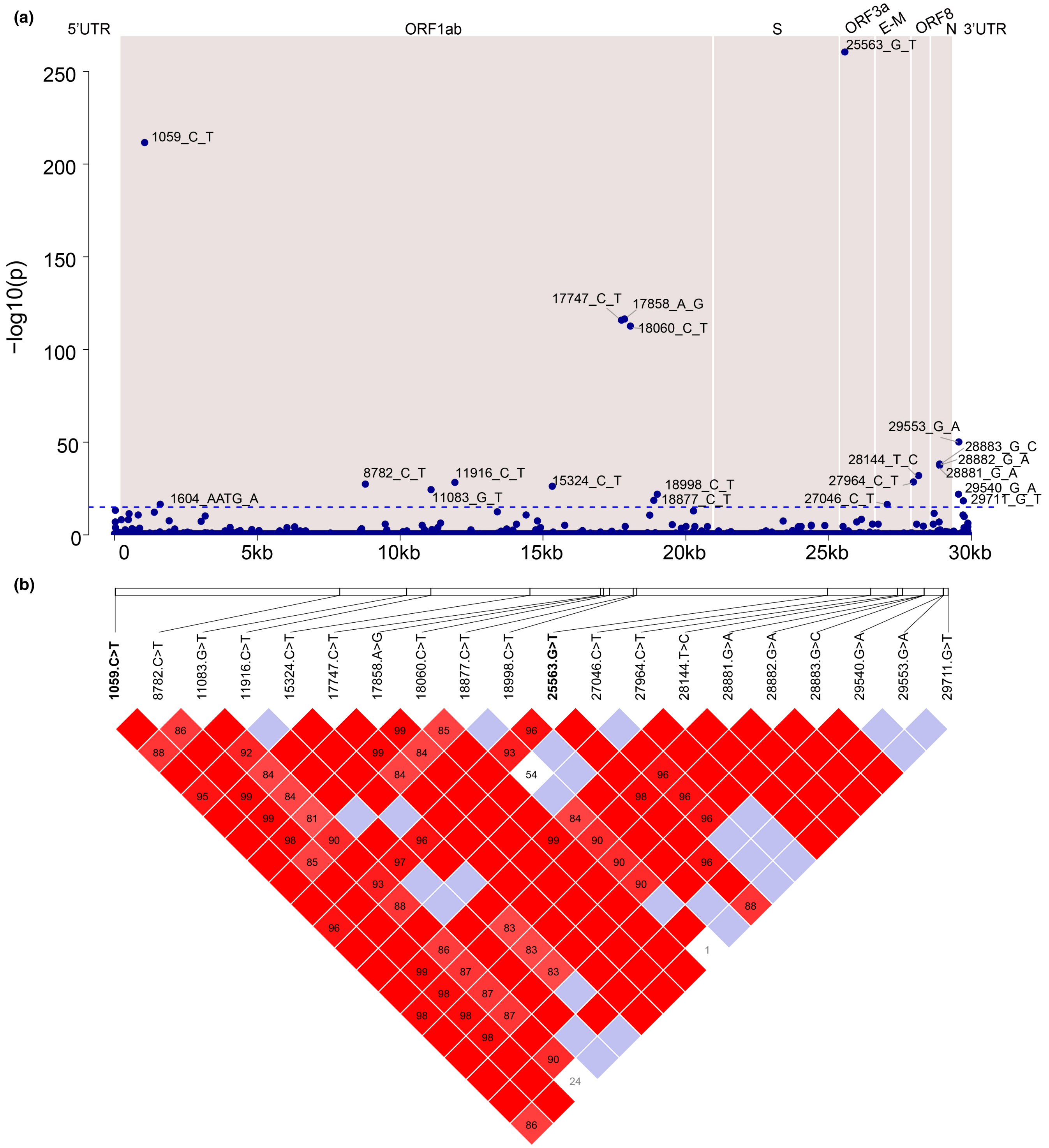


FIGURE 4 GWAS and linkage disequilibrium (LD) analysis. (a) Manhattan plot comparing the North American SARS-CoV-2 isolates (n = 1,063) to the isolates from the remaining continents (n = 1536). Genomic coordinates are displayed along the X-axis, and -log10 of the association p-value for each SNP is displayed on the Y-axis (threshold p-value =1.00 × 10^{-15}). Different blocks indicate the different protein-encoding regions. (b) Linkage disequilibrium between SNPs in SARS-CoV-2. LD plot of any two SNP pairs among the 21 sites. The number near slashes shows the genomic coordinates. The colour in the square is given by the standard (D'/LOD), and the number in the square is the r2 value

attention of certain medical communities, such as the Centers for Disease Control and Prevention of the USA and Washington State Department of Health (Centers for Disease Control and Prevention-USA, 2020; Ou et al., 2020; Washington State Department of Health, 2020). Regardless of the early transmission history, the genetic signatures identified may help methodology

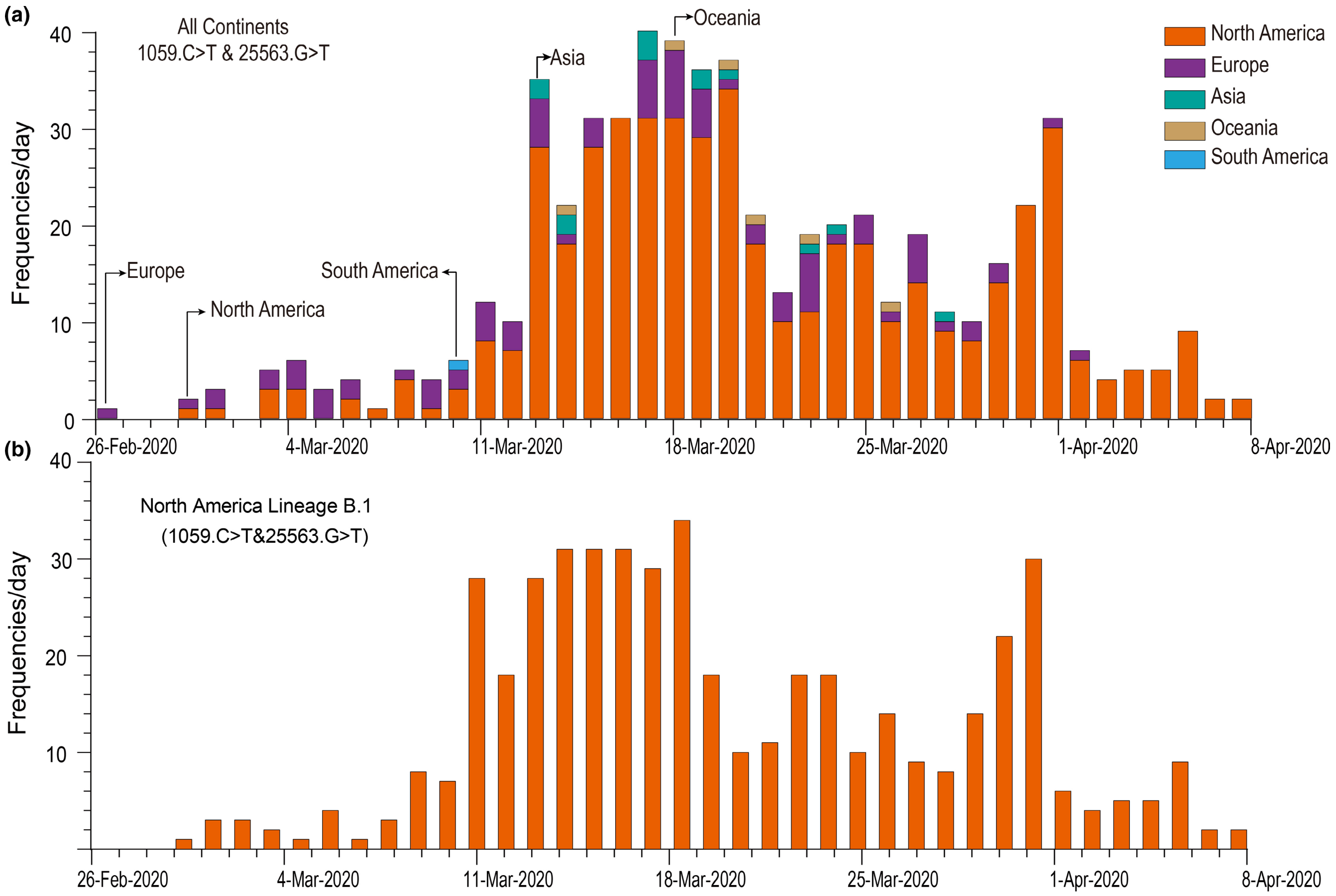


FIGURE 5 Retrospectively tracing the early SARS-CoV-2 isolates with SNPs (1,059.C > T & 25,563.G > T) of all continents and North American lineage B.1. (a) The time-dependent accumulating plot for frequencies of the two SNPs (1,059.C > T & 25,563.G > T) between continents. The continents are labelled by different colours, and the continents of the first occurrence of the two SNPs are indicated. (b) The time-dependent accumulation plot of North American lineage B.1. Of note, the two SNPs largely and concurrently accumulated during mid to late March and occurred most frequently in isolates of North American lineage B.1 (479/691, 69.31%) (Table 1)

development for precisely tracing viral transmission in real time, such as via the two causative SNPs. These two SNPs pose a great possibility for epidemiological surveillance in the North American population due to their high prevalence in the same population. In clinical diagnosis, these SNPs may improve methodology development to specifically detect North American isolates, such as via SNP-based allele-specific polymerase chain reaction (ASPCR) (Corman et al., 2020; Ugozzoli & Wallace, 1991). It has been reported that SARS-CoV-2 with the D614G mutation in the S protein increases infectivity in human lung cells (Yurkovetskiy et al., 2020). The two SNPs we identified are responsible for two missense mutations that probably change the protein structure and function to some extent. More mechanical investigations of the functional impact caused by these two SNPs would be enhanced in this aspect, as they are possibly druggable targets.

The hard lesson of the ongoing SARS-CoV-2 pandemic is its strain of the global public health system (Ji et al., 2020). Before the pandemic, researchers detected the proximal origin of SARS-CoV-2 in bats from 2015 to 2017 (Hu et al., 2018). However, its risks to public health were largely ignored. A platform for early surveillance and risk estimation of bat CoVs is very much needed. In the future, it is hoped that these exclusively genetic signatures may help public health surveillance and measures.

ACKNOWLEDGEMENTS

This work was supported by the China Agricultural Research System (CARS-42-17), Sichuan Veterinary Medicine and Drug Innovation Group of China Agricultural Research System (SCCXTD-2020-18), Science and Technology Program of Sichuan Province (2020YJ0369) and Integration and Demonstration of Key Technologies for Goose Industrial Chain in Sichuan Province (2018NZ0005).

ETHICAL APPROVAL

This study was approved by Sichuan Agricultural University Ethical Committee. No clinical data, animal and human material has been disclosed or used in this study.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

X.O., Q.P. and A.C. devised the project and the main conceptual ideas; X.O., Z.Y., D.Z., S.M. and M.W. acquired the data; X.O. and Z.Y. designed and performed the computations; X.O., Z.Y., D.Z., S.M., R.J., S.C., M.L. and Q.Y. analysed and interpreted the data; X.O. and Z.Y. drafted the manuscript; Y.W., X.Z., S.Z., J.H., Q.G., Y.L., L.Z., M.P. and Q.P. proofread the draft; and all authors approved the final version of the manuscript.

DATA AVAILABILITY STATEMENT

All data will be available in the manuscript.

ORCID

Xumin Ou https://orcid.org/0000-0003-0456-6362
Sai Mao https://orcid.org/0000-0001-5411-4706
Qiuwei Pan https://orcid.org/0000-0001-9982-6184
Anchun Cheng https://orcid.org/0000-0001-6093-353X

REFERENCES

Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., & Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nature Medicine*, *26*(4), 450–452. https://doi.org/10.1038/s41591-020-0820-9

Azhar, E. I., El-Kafrawy, S. A., Farraj, S. A., Hassan, A. M., Al-Saeed, M. S., Hashem, A. M., & Madani, T. A. (2014). Evidence for camel-to-human transmission of MERS coronavirus. *New England Journal of Medicine*, *370*(26), 2499–2505. https://doi.org/10.1056/NEJMoa1401505

Baranov, P. V., Henderson, C. M., Anderson, C. B., Gesteland, R. F., Atkins, J. F., & Howard, M. T. (2005). Programmed ribosomal frameshifting in decoding the SARS-CoV genome. *Virology*, *332*(2), 498–510. https://doi.org/10.1016/j.virol.2004.11.038

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*, *21*(2), 263–265. https://doi.org/10.1093/bioinformatics/bth457

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Centers for Disease Control and Prevention-USA (2020). *Hot topics of the day*. Retrieved from https://phgkb.cdc.gov/PHGKB/phgHome.action?action=archive&date=05/14/2020

Coleman, J. R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E., & Mueller, S. (2008). Virus attenuation by genome-scale changes in codon pair bias. *Science*, *320*(5884), 1784–1787. https://doi.org/10.1126/science.1155761

Consortium, & C. S. M. E (2004). Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science*, *303*(5664), 1666–1669. https://doi.org/10.1126/science.1092002

Corman, V. M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D. K. W., Bleicker, T., Brünink, S., Schneider, J., Schmidt, M. L., Mulders, D. G. J. C., Haagmans, B. L., van der Veer, B., van den Brink, S., Wijsman, L., Goderski, G., Romette, J.-L., Ellis, J., Zambon, M., … Drosten, C. (2020). Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance Weekly*, *25*(3), 2000045. https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045

Forni, D., Cagliani, R., Clerici, M., & Sironi, M. (2017). Molecular evolution of human coronavirus genomes. *Trends in Microbiology*, *25*(1), 35–48. https://doi.org/10.1016/j.tim.2016.09.001

Hanson, G., & Coller, J. (2017). Codon optimality, bias and usage in translation and mRNA decay. *Nature Reviews Molecular Cell Biology*, *19*, 20. https://doi.org/10.1038/nrm.2017.91

Hu, D., Zhu, C., Ai, L., He, T., Wang, Y. I., Ye, F., Yang, L. U., Ding, C., Zhu, X., Lv, R., Zhu, J., Hassan, B., Feng, Y., Tan, W., & Wang, C. (2018). Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. *Emerging Microbes & Infections*, *7*(1), 154. https://doi.org/10.1038/s41426-018-0155-5

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics*, *26*(5), 680–682. https://doi.org/10.1093/bioinformatics/btq003

Ji, Y., Ma, Z., Peppelenbosch, M. P., & Pan, Q. (2020). Potential association between COVID-19 mortality and health-care resource availability. *The Lancet. Global Health*, *8*(4), e480. https://doi.org/10.1016/S2214-109X(20)30068-1

Kelly, J. A., Olson, A. N., Neupane, K., Munshi, S., San Emeterio, J., Pollack, L., Woodside, M. T., & Dinman, J. D. (2020). Structural and functional conservation of the programmed −1 ribosomal frameshift signal of SARS coronavirus 2 (SARS-CoV-2). *Journal of Biological Chemistry*, https://doi.org/10.1074/jbc.AC120.013449

Kryazhimskiy, S., & Plotkin, J. B. (2008). The Population Genetics of dN/dS. *Plos Genetics*, *4*(12), e1000304. https://doi.org/10.1371/journal.pgen.1000304

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, *5*(2), R12. https://doi.org/10.1186/gb-2004-5-2-r12

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, W., Shi, Z., Yu, M., Ren, W., Smith, C., Epstein, J. H., & Wang, L.-F. (2005). Bats Are Natural Reservoirs of SARS-Like Coronaviruses. *Science*, *310*(5748), 676–679. https://doi.org/10.1126/science.1118391

Madani, T. A., Azhar, E. I., & Hashem, A. M. (2014). Evidence for camel-to-human transmission of MERS coronavirus. *New England Journal of Medicine*, *371*(14), 1360. https://doi.org/10.1056/NEJMc1409847

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Nakamura, T., Yamada, K. D., Tomii, K., & Katoh, K. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics*, *34*(14), 2490–2492. https://doi.org/10.1093/bioinformatics/bty121

Ou, X., Cao, J., Cheng, A., Peppelenbosch, M. P., & Pan, Q. (2019). Errors in translational decoding: tRNA wobbling or misincorporation? *Plos Genetics*, *15*(3), e1008017. https://doi.org/10.1371/journal.pgen.1008017

Ou, X., Wang, M., Mao, S., Cao, J., Cheng, A., Zhu, D., Chen, S., Jia, R., Liu, M., Yang, Q., Wu, Y., Zhao, X., Zhang, S., Liu, Y., Yu, Y., Zhang, L., Chen, X., Peppelenbosch, M. P., & Pan, Q. (2018). Incompatible Translation Drives a Convergent Evolution and Viral Attenuation During the Development of Live Attenuated Vaccine. *Frontiers in Cellular and Infection Microbiology*, *8*, https://doi.org/10.3389/fcimb.2018.00249

Ou, X., Yang, Z., Zhu, D., Mao, S., Wang, M., Jia, R., & Wu, Y. J., (2020). Tracing two causative SNPs reveals SARS-CoV-2 transmission in North America population. *BioRvix*, https://doi.org/10.1101/2020.05.12.092056

Perlman, S. (2020). Another decade, another coronavirus. *New England Journal of Medicine*, *382*(8), 760–762. https://doi.org/10.1056/NEJMe2001126

Power, R. A., Parkhill, J., & de Oliveira, T. (2017). Microbial genome-wide association studies: Lessons from human GWAS. *Nature Reviews Genetics*, *18*(1), 41–50. https://doi.org/10.1038/nrg.2016.132

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L., & Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*, *5*(11), 1403–1407. https://doi.org/10.1038/s41564-020-0770-5

Shi, J, Wen, Z, Zhong, G, Yang, H, Wang, C, Huang, B... Bu, Z. (2020). Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS–coronavirus 2. *Science*, *368*(6494), 1016–1020. https://doi.org/10.1126/science.abb7015

Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., Duan, Y., Zhang, H., Wang, Y., Qian, Z., Cui, J., & Lu, J. (2020). On the origin and continuing evolution of SARS-CoV-2. *National Science Review*, *26*(4), 450–452. https://doi.org/10.1093/nsr/nwaa036

Ugozzoli, L., & Wallace, R. B. (1991). Allele-specific polymerase chain reaction. *Methods*, *2*(1), 42–48. https://doi.org/10.1016/S1046-2023(05)80124-0

Washington State Department of Health (2020). *2019-nCoV Literature Situation Report* (pp. 1–8). Health. Retrieved from https://www.doh.wa.gov/Portals/1/Documents/1600/coronavirus/LitRep-20200514.pdf

Wenjie, T., Peihua, N., Xiang, Z., Yang, P., Yong, Z., Lijuan, C., & Guizhen, W. (2020). *Notes from the Field: Reemergent Cases of COVID-19 — Xinfadi Wholesales Market, Beijing Municipality, China, June 11, 2020*. China CDC Weekly, https://doi.org/10.46234/ccdcw2020.132

WHO (2020). *WHO report Situation Report – 160*. Retrieved from https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200513-covid-19-sitrep-114.pdf?sfvrsn=17ebbbe_4

Wu, J. T., Leung, K., & Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. *The Lancet*, *395*(10225), 689–697. https://doi.org/10.1016/s0140-6736(20)30260-9

Yurkovetskiy, L., Wang, X., Pascal, K. E., Tomkins-Tinch, C., Nyalile, T. P., Wang, Y., Baum, A., Diehl, W. E., Dauphin, A., Carbone, C., Veinotte, K., Egri, S. B., Schaffner, S. F., Lemieux, J. E., Munro, J. B., Rafique, A., Barve, A., Sabeti, P. C., Kyratsous, C. A., ... Luban, J. (2020). Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. *Cell*, *183*(3), 739–751 e738. https://doi.org/10.1016/j.cell.2020.09.032

Zhou, J. H., Li, X. R., Lan, X., Han, S. Y., Wang, Y. N., Hu, Y. H., & Pan, Q. W. (2019). The genetic divergences of codon usage shed new lights on transmission of hepatitis E virus from swine to human. *Infection Genetics and Evolution*, *68*, 23–29. https://doi.org/10.1016/j.meegid.2018.11.024

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X. I., ... Shi, Z.-L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, *579*(7798), 270–273. https://doi.org/10.1038/s41586-020-2012-7

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Ou X, Yang Z, Zhu D, et al. Tracing genetic signatures of bat-to-human coronaviruses and early transmission of North American SARS-CoV-2. *Transbound Emerg Dis*. 2022;69:1748–1760. https://doi.org/10.1111/tbed.14148