Peng et al. *Zool. Res.* 2021, 42(6): 833−843
https://doi.org/10.24272/j.issn.2095-8137.2021.334

ZR | Zoological Research

Article Open Access

The high diversity of SARS-CoV-2-related coronaviruses in pangolins alerts potential ecological risks

Min-Sheng Peng[1,2,3,#,*], Jian-Bo Li[1,2,#], Zheng-Fei Cai[4,#], Hang Liu[1,2,#], Xiaolu Tang[5,#], Ruochen Ying[5], Jia-Nan Zhang[6], Jia-Jun Tao[6], Ting-Ting Yin[1], Tao Zhang[4], Jing-Yang Hu[4], Ru-Nian Wu[1], Zhong-Yin Zhou[1], Zhi-Gang Zhang[4], Li Yu[4], Yong-Gang Yao[2,3,7], Zheng-Li Shi[8], Xue-Mei Lu[1,2,9], Jian Lu[5,*], Ya-Ping Zhang[1,2,3,4,9,*]

[1] *State Key Laboratory of Genetic Resources and Evolution*, *Yunnan Laboratory of Molecular Biology of Domestic Animals*, *Kunming Institute of Zoology*, *Chinese Academy of Sciences*, *Kunming*, *Yunnan* 650201, *China*

[2] *Kunming College of Life Science*, *University of Chinese Academy of Sciences*, *Kunming*, *Yunnan* 650204, *China*

[3] *KIZ/CUHK Joint Laboratory of Bioresources and Molecular Research in Common Diseases*, *Kunming Institute of Zoology*, *Chinese Academy of Sciences*, *Kunming*, *Yunnan* 650201, *China*

[4] *State Key Laboratory for Conservation and Utilization of Bio-resources in Yunnan*, *Yunnan University*, *Kunming*, *Yunnan* 650091, *China*

[5] *State Key Laboratory of Protein and Plant Gene Research*, *Center for Bioinformatics*, *School of Life Sciences*, *Peking University*, *Beijing* 100871, *China*

[6] *Molbreeding Biotechnology Co., Ltd.*, *Shijiazhuang*, *Hebei* 050035, *China*

[7] *Key Laboratory of Animal Models and Human Disease Mechanisms of the Chinese Academy of Sciences & Yunnan Province*, *Kunming Institute of Zoology*, *Kunming*, *Yunnan* 650201, *China*

[8] *CAS Key Laboratory of Special Pathogens and Biosafety*, *Wuhan Institute of Virology*, *Chinese Academy of Sciences*, *Wuhan*, *Hubei* 430071, *China*

[9] *Center for Excellence in Animal Evolution and Genetics*, *Chinese Academy of Sciences*, *Kunming*, *Yunnan* 650201, *China*

ABSTRACT

Understanding the zoonotic origin and evolution history of SARS-CoV-2 will provide critical insights for alerting and preventing future outbreaks. A significant gap remains for the possible role of pangolins as a reservoir of SARS-CoV-2 related coronaviruses (SC2r-CoVs). Here, we screened SC2r-CoVs in 172 samples from 163 pangolin individuals of four species, and detected positive signals in muscles of four *Manis javanica* and, for the first time, one *M. pentadactyla*. Phylogeographic analysis of pangolin mitochondrial DNA traced their origins from Southeast Asia. Using in-solution hybridization capture sequencing, we assembled a partial pangolin SC2r-CoV (pangolin-CoV) genome sequence of 22 895 bp (MP20) from the *M.*

Received: 22 September 2021; Accepted: 09 November 2021; Online: 10 November 2021

Foundation items: This work was supported by the National Key Research and Development Projects of the Ministry of Science and Technology of China, National Key Research and Development Program of China (2021YFC0863300), Ministry of Agriculture of China (2016ZX08009003-006), Key Program of Chinese Academy of Sciences (KJZD-SW-L11), and Animal Branch of the Germplasm Bank of Wild Species, Chinese Academy of Sciences (the Large Research Infrastructure Funding)

#Authors contributed equally to this work

*Corresponding authors, E-mail: pengminsheng@mail.kiz.ac.cn; luj@pku.edu.cn; zhangyp@mail.kiz.ac.cn

pentadactyla sample. Phylogenetic analyses revealed MP20 was very closely related to pangolin-CoVs that were identified in *M. javanica* seized by Guangxi Customs. A genetic contribution of bat coronavirus to pangolin-CoVs via recombination was indicated. Our analysis revealed that the genetic diversity of pangolin-CoVs is substantially higher than previously anticipated. Given the potential infectivity of pangolin-CoVs, the high genetic diversity of pangolin-CoVs alerts the ecological risk of zoonotic evolution and transmission of pathogenic SC2r-CoVs.

Keywords: SARS-CoV-2; Pangolin; Recombination; Diversity; Sequencing; mtDNA

INTRODUCTION

The COVID-19 pandemic, caused by the global spread of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has had significant impacts on the global economy and public health (Hu et al., 2021). Efforts to trace the zoonotic origin of SARS-CoV-2 and investigate its evolution and transmission have attracted worldwide attention (Banerjee et al., 2021; Hu et al., 2021; WHO, 2021; Wong et al., 2020; Zhang & Holmes 2020). Recent studies have identified coronaviruses closely related to SARS-CoV-2 (genomic identity >90%) in various bat species, including *Rhinolophus affinis* (Zhou et al., 2020b), *R. malayanus* (Zhou et al., 2020a), *R. cornutus* (Murakami et al., 2020), *R. shameli* (Hul et al., 2021), *R. acuminatus* (Wacharapluesadee et al., 2021), and *R. pusillus* (Zhou et al., 2021). Among these bat-derived SARS-CoV-2-related coronaviruses (SC2r-CoVs), RaTG13 from *R. affinis* is the closest to SARS-CoV-2, with the sequence identity of 96.2% between the two viral genomes (Zhou et al., 2020b).

Besides bats, SC2r-CoVs have also been detected in the Malayan pangolins (*Manis javanica*) (Lam et al., 2020; Liu et al., 2020; Xiao et al., 2020; Zhang et al., 2020). The pangolin-derived SC2r-CoVs (pangolin-CoVs) can be classified into two sublineages. The first sublineage, defined by a pangolin-CoV strain detected in Malayan pangolins confiscated during anti-smuggling operations by Guangdong Customs (pangolin-CoV-GDC), shares a genomic sequence similarity of 92.4% to SARS-CoV-2 (Liu et al., 2020; Xiao et al., 2020; Zhang et al., 2020). The second sublineage was defined with six strains of pangolin-CoVs that were identified in Malayan pangolins obtained during anti-smuggling operations performed by Guangxi Customs (pangolin-CoV-GXC) (Lam et al., 2020). The genome sequences of the six pangolin-CoV-GXC strains were highly similar to each other (>99%), and they overall show a similarity of 85.5% to that of SARS-CoV-2 (Lam et al., 2020). The two pangolin-CoV sublineages share a genomic sequence similarly of 84%, presenting substantial genetic divergence. Currently, *M. javanica* is the only known pangolin species that harbors coronaviruses closely related to SARS-CoV-2. Although the pangolin-CoVs tend to be more distantly related to SARS-CoV-2 than the aforementioned bat-derived SC2r-CoVs, recent studies show that the receptor-binding domain (RBD) of both pangolin-CoV sublineages can potentially bind the human ACE2 receptor (Dicken et al., 2021; Guo et al., 2021; Nie et al., 2021; Niu et al., 2021; Zhang et al., 2021), raising concerns about the potential risks of spillover of these pangolin-CoVs to humans.

Different hypotheses have been proposed to account for the possible role of pangolins as a source of potential zoonosis of SC2r-CoVs (see Supplementary Table S1 for a summary). Based on the studies that the two divergent pangolin-CoV sublineages were detected in the samples of *M. javanica* collected before the COIVD-19 pandemic, one hypothesis is that pangolins are reservoirs or natural hosts of SC2r-CoVs (the natural host hypothesis). According to this hypothesis, one or both sublineages of pangolin-CoVs might have a long history of circulating in pangolins in the wild (Lam et al., 2020; Liu et al., 2020; Zhang et al., 2020). However, a recent study in which 334 live *M. javanica* individuals confiscated in Peninsular Malaysia and Borneo were screened for pangolin-CoVs yielded no positive results, suggesting that pangolins are only incidental hosts of SC2r-CoVs in smuggling networks, and that SC2r-CoVs were recently transmitted from bats or other unknown animals to pangolins (the incidental host hypothesis) (Lee et al., 2020). Indeed, unlike bats, which typically carry coronaviruses asymptomatically, some pangolin individuals infected with SC2r-CoVs have shown clinical signs of severe respiratory disease (Lam et al., 2020; Liu et al., 2019, 2020; Xiao et al., 2020); this is in line with the incidental host hypothesis. If this hypothesis is correct, the presence of SC2r-CoVs with high similarity to pangolin-CoVs would be expected in wildlife across the same region in which pangolins are distributed or along pangolin smuggling routes. However, no evidence of such SC2r-CoVs has yet been found despite extensive surveys of wildlife and livestock in China and Southeast Asia (Hul et al., 2021; Wacharapluesadee et al., 2021; WHO, 2021). Collectively, there is yet no consensus view on the two competing hypotheses, and more studies are required for a better understanding of whether pangolins act as natural, incidental, or intermediate hosts.

Here, we first screened the SC2r-CoVs in 172 samples from 163 pangolin individuals belonging to four species that were confiscated during anti-smuggling operations in Yunnan Province of southwestern China. Then we analyzed the genome sequence of a pangolin-CoV that was isolated and sequenced from a *M. pentadactyla* sample (MP20). Our results, together with the re-analyses of pangolin-CoVs sequenced in a previous study (Lam et al., 2020), reveal that the genetic diversity of pangolin-CoVs are substantially higher than previously anticipated, and provide more evidence that favors the natural host over the incidental host hypothesis.

MATERIALS AND METHODS

Sample collection

This study was approved by the Zoology Animal Care and Ethics Committee of the Kunming Institute (SMKX-20200210-01). The pangolin samples were collected by the Animal Branch of the Germplasm Bank of Wild Species of the Chinese Academy of Sciences from seizures of illegal wildlife

trade and smuggling in the international border regions of Yunnan Province, southwestern China, between 1990 and 2018. All the pangolin individuals had been dead when the samples were collected, with the postmortem intervals unknown. The samples were collected and preserved in 95% alcohol at –80 °C immediately after sample collection.

Viral RNA extraction

The samples were disrupted in 200 μL of PBS by ultrasonication for 5 min. Viral RNA was extracted using the High Pure Viral RNA Kit (Roche, Switzerland) according to the manufacturer-recommended procedure. After elution with 50 μL of RNase-free water, the RNA purity (OD260/280) and concentration and the absorption peak were measured with a NanoDrop 2000 (Thermo Scientific, USA).

RT-PCR and Sanger sequencing

We followed a published RT-PCR strategy (Wu et al., 2020) to screen the viral RNA extracts. PCR primers and a 5'FAM-labeled probe (F1: 5'-GGTCATGTGTGGCGGCTC-3'; R1: 5'-GCTGTAACAGCTTGACAAATGTTAAAG-3'; Probe: 5'-CTATATGTTAAACCAGGTGGAAC-3') were designed to target the *S* gene of SARS-CoV-2. AgPath-ID™ One-Step RT-PCR reagents (Thermo Scientific, USA) were used for amplification in a QuantStudio 12K Flex Real-Time PCR System (ABI, USA) as follows: 50 ℃ for 10 min, 95 ℃ for 10 min, and 40 cycles consisting of 95 ℃ for 15 s and 55 ℃ for 45 s, followed by a default melting curve step. Each run was repeated at least three times. The PCR products from RT-PCR-positive samples were TA cloned with the vector pMD®18-T (Takara, Japan) and subsequently Sanger sequenced. We then aligned the sequences using BLAST with the 2019 Novel Coronavirus Resource (2019nCoVR) (Gong et al., 2020).

Library preparation and screening

Viral RNA sequencing library preparation and depletion of ribosomal RNA (rRNA) were performed using the SMARTer Stranded Total RNA-seq Kit V2-Pico Input Mammalian (Takara, Japan) according to the manufacturer's protocol, with 400 ng of input RNA. The five libraries were PE150 sequenced on the NovaSeq 6000 platform (Illumina, USA) by Novogene Co., Ltd., China.

In-solution capture and next-generation sequencing

We used an in-solution capture kit and corresponding protocol that was originally developed for laboratory diagnosis of SARS-CoV-2 by Molbreeding Biotechnology Co., Ltd., China. A total of 502 GenoBaits DNA probes with an average length of 120 nucleotides were designed based on 50 different human SARS-CoV-2 genome sequences. A similar system was applied to capture and enrich SARS-CoV-2 viral sequences from viral particles at low concentrations (Wen et al., 2020). The coverage depth for the regions with mutations across diverse viral lineages ranged from 2 to 3 X to enhance the capture capacity (Johnson et al., 2019). A total of 500 ng of RNAs was used for in-solution capture. The commercial kit, together with the protocol, is available from Molbreeding Biotechnology Co., Ltd, China. Finally, the GenoBaits DNA probe capture library was PE150 sequenced on the HiSeq X Ten platform (Illumina, USA) by Berry Genomics Co., Ltd., China.

Analysis of NGS data

Reads that did not include one of the tags/indices used in library construction were discarded to avoid potential contamination (Briggs et al., 2007; Green et al., 2009). We removed the tag/index and adaptor sequences with Cutadapt v2.10 (Martin, 2011). Reads with a length less than 25 bp were excluded. We used BWA-MEM 0.7.15-r1140 (Li, 2014) to map the reads to the 21 coronaviral genomes (Supplementary Table S2) with the parameter -M -k 15. SAMtools v1.9 was used to filter unmapped reads with the parameter -F 4 (Li et al., 2009). Duplicates were marked by using the MarkDuplicates module of GATK v4.1.3.0 (McKenna et al., 2010) with default parameters. Duplicate reads were removed by SAMtools with the parameter -F 1024.

Genome assembly

We downloaded 21 representative coronavirus genomes (Drexler et al., 2010; Ge et al., 2013; He et al., 2014; Hu et al., 2018; Hul et al., 2021; Lam et al., 2020; Lau et al., 2005; Li et al., 2005, 2021; Lin et al., 2017; Liu et al., 2020; Murakami et al., 2020; Qin et al., 2003; Tao & Tong, 2019; Wacharapluesadee et al., 2021; Wu et al., 2020; Zhou et al., 2020a, 2020b) presenting relatively high similarities to SARS-CoV-2 especially when we conducted the analysis. We used MP789 and GXP5L as representatives of the pangolin-CoV-GDC (MP789) and pangolin-CoV-GXC (GXP5L) substrains, respectively. We performed *de novo* assembly of the 21 coronavirus-mapped reads with SPAdes v3.15.2 (Bankevich et al., 2012) and MegaHit v1.2.9 (Li et al., 2015). We queried the assembled contigs by using BLAST to search NCBI nucleotide collection standard databases to exclude contigs unrelated to coronaviruses. Reference-guided scaffolding was conducted to align the contigs and reads mapped to the 21 coronaviruses that were not assembled into contigs to the reference genome of pangolin-CoV-GXC (GXP5L) (Lam et al., 2020) with blastn v2.10.0+. Reads that mapped to the 21 coronavirus sequences were remapped to the consensus sequence by using BWA-MEM (Li, 2014) and checked by IGV (Robinson et al., 2011). The gaps in the assembled consensus sequence were filled in with "N" to facilitate subsequent multiple genome sequence alignment. The sequencing coverage depth was estimated with SAMtools (Li et al., 2009).

Phylogenetic and evolutionary analysis

We conducted a phylogenetic analysis of MP20 in the context of 21 coronavirus genomes (Supplementary Table S2). A total of 21 sequences were aligned with the assembled consensus sequence with the online service MAFFT v7 (Katoh et al., 2019). The alignment was checked manually with reference to the genome annotation of MN908947. The substitution models and related parameters were determined by the Bayesian information criterion (Schwarz, 1978) in ModelTest-NG v0.1.6 (Darriba et al., 2020). Nucleotide substitution saturation was tested using DAMBE7 (Xia, 2018). MatGAT v2.01 was used to analyze sequence similarity and identity (Campanella et al., 2003). The results were visualized with SimPlot v3.5.1 (Lole et

al., 1999). All gaps and indels were discarded. The maximum likelihood phylogenetic trees for nucleotide alignments were constructed on the basis of 1 000 bootstrap replications by using MEGA X v10.1 (Kumar et al., 2018).

The pairwise nucleotide distance based on transitions and transversions was calculated with the Jukes-Cantor model with gamma distribution (G) +4 in MEGA X v10.1 (Kumar et al., 2018). We calculated the nonsynonymous/synonymous rate ratio (dN/dS) between (1) pangolin-CoV-GDC (MP789) and pangolin-CoV-GXC (GXP5L) and between (2) pangolin-CoV-GXC (GXP5L) and MP20 with the program YN00 implemented in PAML v4.9a (Yang, 2007).

Intra-host variant analysis

Six pangolin-CoV sample sequences were downloaded from the Sequence Read Archive (accession numbers: SRR11093266, SRR11093267, SRR11093268, SRR11093269, SRR11093270, and SRR11093271). The reads were mapped to the pangolin-CoV-GXC reference genome (GXP5L) using BWA-MEM (Li, 2014). SAMtools v1.9 was used to filter out reads with a mapping quality less than 30. Freebayes v9.9.2-27-g5d5b8ac (Garrison & Marth, 2012) was used to call the iHVs (--min-mapping-quality 20 --min-base-quality 20). The GXP3B library (SRR11093270) was excluded due to low coverage depth.

Recombination analysis

We used RDP, GENECONV, BootScan, MaxChi, Chimaera, SiScan and 3Seq implemented in RDP v.4.101 to detect putative recombination events (Martin et al., 2015). SimPlot v.3.5.1 was employed to visualize the recombination events and breakpoints in a sliding window (Lole et al., 1999). The candidate recombination regions were further checked by constructing neighbor-joining trees with the Kimura 2-parameter model and 1 000 bootstrap replications by using MEGA X v10.1 (Kumar et al., 2018).

Pangolin mitochondrial DNA analysis

The DNA barcoding classification of four pangolin samples (MP20, MJ20, MJ57, and MJ54) was published previously (Hu et al., 2020). An 829 bp region of the mitochondrial *COI* gene and a 790 bp region of the cyt *b* gene were sequenced for MJ18007 according to published protocols (Hu et al., 2020; Nash et al., 2018). The *COI* sequence was queried using BLAST searches of GenBank to identify the pangolin species. Mitochondrial DNA clustering based on the *COI* and cyt *b* genes was performed according to the protocol used in a reported phylogeographic analysis of pangolins (Hu et al., 2020). *M. pentadactyla* MP20 was classified into the MPB^{mt} cluster, indicating that it likely originated in Myanmar (Hu et al., 2020). *M. javanica* MJ20 was assigned to the MJA^{mt} cluster, restricted it to northern Mainland Southeast Asia (e.g., Myanmar), while MJ54, MJ57, and MJ18007 were classified into the MJB^{mt} cluster, which is prevalent in Island Southeast Asia (Hu et al., 2020). In particular, MJ18007 shared its mtDNA haplotype with the sample MZBR_1085, recorded as seized illegal trade from Borneo in GenBank (*COI*: MG825608; cyt *b* : MG825549) (Nash et al., 2018).

RESULTS

Screening SC2r-CoVs in samples of four pangolin species

To gain a more comprehensive view on the prevalence of SC2r-CoVs in pangolins, we performed reverse-transcription PCR (RT-PCR) of viral RNA extracted from muscles of 172 samples collected from 163 pangolin individuals that were confiscated during anti-smuggling operations in Yunnan Province, southwestern China, between 1990 and 2018. To infer the species and potential geographic sources of the pangolins, we analyzed mitochondrial DNA using a pangolin phylogeographic framework as described previously (Hu et al., 2020). Altogether, in this study, we tested 172 samples (163 muscles, two blood samples, three kidneys, two small intestines, one heart, and one liver) from 163 pangolin individuals of 4 species (*M. javanica*, 106; *M. pentadactyla*, 53; *Phataginus tricuspis*, 3; and *P. tetradactyla*, 1).

We followed a published RT-PCR strategy (Wu et al., 2020) that target a 121-nucleotide fragment of the *S* gene of SARS-CoV-2 to screen the viral RNA extracts. In total, we detected positive RT-PCR signals in muscles of four *M. javanica* samples and, for the first time, one *M. pentadactyla* sample (Figure 1A). Sanger sequencing revealed that the RT-PCR fragment was 81.01%–83.54% similar to the SARS-CoV-2 sequence (MN908947). Phylogeographic analysis of mitochondrial DNA showed that the five pangolin individuals were obtained from a wide area across Southeast Asia (Figure 1A).

Low amounts of viral RNA in pangolin samples

To determine the genome sequences of the SC2r-CoVs from the five muscle samples that yielded positive RT-PCR results, we performed viral RNA sequencing for each sample with the Illumina platform (each mRNA-Seq library generated 16.6–18.4 Gb of data). The majority (93.44%–97.25%) of the mRNA-Seq reads were mapped to the pangolin reference genomes (Hu et al., 2020). Only twelve reads from sample MP20 (*M. pentadactyla*) and eight reads from sample MJ57 (*M. javanica*) could be mapped to one of the 21 known SC2r-CoVs and SARS-related CoVs (SARSr-CoV) genomes that were discovered in bats and pangolins in previous studies (Tables S2, S3), and we failed to detect reads that can be reliably mapped on these SC2r-CoV and SARSr-CoV genomes in the remaining three libraries. The scarcity of the mapped reads was not unexpected since the pangolin muscle samples were collected after the animals were dead (postmortem intervals unknown), and the samples had been preserved in 95% alcohol, mainly for DNA studies, likely leading to extremely low abundance and high degradation of the RNAs.

A novel partial pangolin-CoV genome sequence (MP20) from *M. pentadactyla*

The development of targeted enrichment technologies for samples that are degraded has considerably facilitated the sequencing of ancient pathogen genomes (Spyrou et al., 2019). Here, we applied a similar strategy to RNAs extracted from the muscle of the *M. pentadactyla* individual (sample MP20), as it showed the highest abundance of SC2r-CoV nucleotides among the five pangolins for which mRNA-Seq libraries were available. Specifically, we carried out in-solution

hybridization capture with dense probes that cover the whole genome of SARS-CoV-2 to enrich SC2r-CoV sequences. Targeted enrichment sequencing generated ~74 Gb of data, with 395 424 769 PE150 reads. After a series of quality control checks and data filtering, we obtained 901 clean reads that could be mapped to the 21 known SC2r-CoV and SARSr-CoV genomes (Supplementary Table S2; aligned length >100 nucleotides and blast E<10^{-5} were required). We *de novo* assembled these mapped reads using both SPAdes (Bankevich et al., 2012) and MegaHit (Li et al., 2015). The SPAdes analysis yielded 36 contigs that ranged from 230 to 1 994 nucleotides in length, with a mean length of 525.6 nucleotides. The MegaHit analysis similarly yielded 36 contigs, ranging from 201 to 1 695 nucleotides in length with a mean length of 475.25 nucleotides. Since these contigs were more similar to the pangolin-CoV-GXC sequence than to the other SC2r-CoV sequences we analyzed, we pooled the contigs assembled by SPAdes and MegaHit together, and scaffolded the contigs and some mapped reads (without assembly into contigs), using the GXP5L strain of pangolin-CoV-GXC as a reference. In addition, two unique reads from MJ57 were used to fill two gaps (23 266−23 301 and 25 577−25 588). We finally obtained a consensus genome from sample MP20 that contained sequenced nucleotides at 22 895 sites, and this viral genome (named MP20) covers ~76.6% of the SARS-CoV-2 genome and has a coverage depth from 1 to 64 X and a mean depth of 8.45 X (Figure 1B).

Divergence between MP20 and other pangolin-CoVs

The MP20 genome exhibited 86.3% nucleotide sequence identity and 93.9% amino acid identity to the reference genome of SARS-CoV-2 (MN908947) and the corresponding translated polypeptide, respectively (Figure 1C). The alignment for MP20 (red bar) to SARS-CoV-2 genes indicated 79.47% coverage of *ORF1ab*, 66.09% of *S*, 78.26% of *ORF3a*, 7.89% of *E*, 85.80% of *M*, 100% of *ORF6*, 88.80% of *ORF7a*, 0.82% of *ORF8*, 99.52% of *N* and 100% of *ORF10*. MP20 was more similar to pangolin-CoV-GXC (96% sequence identity) than to pangolin-CoV-GDC (86.4% sequence identity). Indeed, the construction of a phylogenetic tree based on the whole-genome sequences of MP20 and the 21 known SC2r-CoV and SARSr-CoVs clustered MP20 with GXP5L (the reference sequence of pangolin-CoV-GXC) in the pangolin-CoV-GXC sublineage (Figure 1D).

Although we did not retrieve the full-length sequence of the *S* gene of MP20, which encodes the spike protein and plays a crucial role in the binding to the host's cellular receptor and cellular entry, we achieved an alignment with 66.09% coverage (2 526 bp), mainly covering the S2 subunit (Figure 1C). MP20 and GXP5L showed 95% nucleotide sequence identity and 98.8% amino acid identity at the *S* gene and spike protein, respectively. There are six amino acids in the RBD region of the S protein that are crucial for the attachment of SARS-CoV-2 to the ACE2 receptor (Lu et al., 2020; Ou et al., 2020; Qu et al., 2005; Ren et al., 2008; Wan et al., 2020; Wrapp et al., 2020), but unfortunately, the MP20 sequence only covers 60.99% of the RBD region and does not cover any of those crucial sites, which hampered further investigation of the potential infection ability of MP20 (Supplementary Figure S1). However, our sequence comparisons revealed there were five amino acid changes in the RBD regions between GXP5L and MP20 (I324T, N504D, G520A, T532Q, and N536D; Supplementary Figure S1). Further studies are required to investigate whether those differences might be associated with the difference in the cellular entry between MP20 and GXP5L.

The average genome-level nucleotide divergence between MP20 and GXP5L was 0.041, which is very similar to those between human SARS-CoV-2 and the bat SC2r-CoVs RaTG13 (0.036) and RmYN02 (0.045), suggesting the genetic diversity within the pangolin-CoV-GXC sublineage should not be neglected. Previous evolutionary analysis suggested that strong purifying selection underlay the protein changes between SARS-CoV-2 and RaTG13 (Li et al., 2020; Tang et al., 2020). Likewise, we identified strong signals indicating purifying selection between MP20 and GXP5L (genomic average dN/dS=0.011/0.181=0.058; Supplementary Table S4) and between GXP5L and pangolin-CoV-GDC (dN/dS=0.060/0.718=0.083; Supplementary Table S5), suggesting functional constraints have operated the protein evolution of the pangolin-CoV strains. By assuming a molecular clock and a neutral substitutional rate of 1.69×10^{-3} substitutions/site/year (Boni et al., 2020), we estimated the divergence time between pangolin-CoV-GDC and pangolin-CoV-GXC based on synonymous substitutions (dS) as 212.4 ($0.718/(2\times1.69\times10^{-3})\approx212.4$) years ago and that between GXP5L and MP20 as 53.6 ($0.181/(2\times1.69\times10^{-3})\approx53.6$) years ago. Overall, our analyses suggest the divergence of MP20 from other pangolin-CoV strains was not very recent and support the scenarios that these pangolin-CoVs have had a long history of circulating in different species of pangolins in the wild.

Pervasive intra-host variants in the pangolin-CoVs

Intra-host variants (iHV) provide key insights into the genetic diversity and evolutionary dynamics of SARS-CoV-2 (Armero et al., 2021; Wang et al., 2021). Despite the low sequencing coverage of the MP20 coronavirus, we still observed a modest number of intra-host variants (iHVs) at six sites (three of them were nonsynonymous: G1299T and G6105T in *ORF1ab* and A26530C in *M*) that had at least 5X coverage depth in the MP20 sequencing results (Supplementary Table S6). To further examine the occurrence of iHVs in pangolin-CoVs, we analyzed the libraries from six previously published pangolin-CoV-GXC viruses that were detected in various tissues of *M. javanica* (Lam et al., 2020), using GXP5L as the reference genome. To avoid sequencing errors, we considered only those sites for both single nucleotide variants and small indels that had coverage of ≥150X and ≥10 reads covering the alternative alleles in each library. The number of identified iHVs per library ranged from 71 to 432 (the GXP3B library SRR11093270 was excluded because it had a median coverage of 5X, and no iHV site met our criteria; Table 1).

In total, we identified 852 iHVs in at least one of five libraries, including 217 synonymous, 336 nonsynonymous, 264 frameshift, and 3 stop-gain variants. 566 (66.4%) of the iHVs were restricted to a single library (Figure 2A). Consistent with previous observations that intra-host variants of SARS-

A

Information from five pangolin samples with positive RT-PCR results for SARS-CoV-2

Sample ID	Species	Date	Blast of RT-PCR sequence	mtDNA cluster	Inferred geographic source
MP20	*Manis pentadactyla*	2017	81.01% to MN908947	MPB^{mt}	Myanmar
MJ20	*Manis javanica*	2017	83.54% to MN908947	MJA^{mt}	Northern Mainland Southeast Asia
MJ57	*Manis javanica*	2017	83.54% to MN908947	MJB^{mt}	Island Southeast Asia
MJ54	*Manis javanica*	2017	83.54% to MN908947	MJB^{mt}	Island Southeast Asia
MJ18007	*Manis javanica*	2018	83.54% to MN908947	MJB^{mt}	Borneo

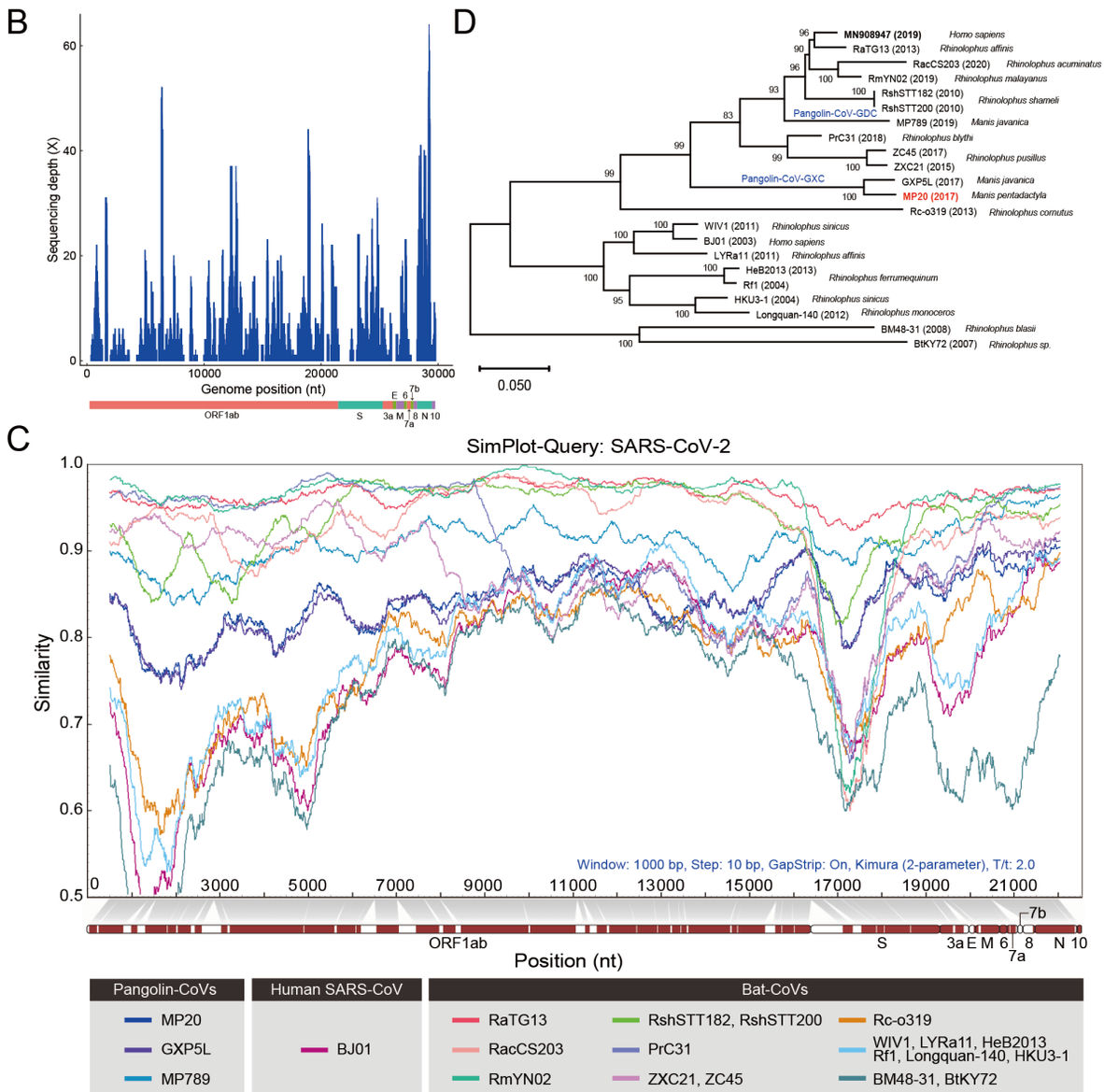

Figure 1 Genetic characterization of pangolin-CoVs in pangolin muscles

A: The five samples that yielded positive RT-PCR signals of SC2r-CoVs. The nomenclature of mitochondrial DNA (mtDNA) clusters used for inferring the possible geographic origins of the samples is given in reference (Hu et al., 2020). B: Sequencing depth of clean reads remapped to MP20 using nucleotide extraction and enrichment technologies. The genome organization of the SARS-CoV-2 reference strain MN908947 is shown. C: Sliding window analysis of nucleotide sequence similarity between SARS-CoV-2 and SC2r-CoVs/SARSr-CoVs from bats and pangolins. D: Maximum likelihood tree based on the alignment of the 22 genomes with the substitution model GTR+I+G4. Bootstrap values calculated from 1000 replicates are shown.

CoV-2 could be shared among samples collected from different patients (Armero et al., 2021; Wang et al., 2021), here, we found a total of 286 (33.6%) of these iHVs were detected in at least two of the five libraries (Figure 2). A total of 158 iHVs were observed in the *S* gene (Table 1), and 58 (36.7%) of these *S* gene iHVs were shared by at least two libraries (Figure 2B). Collectively, the large numbers of iHVs we observed here adds another layer of complexity for

Table 1 The number of putative iHVs identified in the pangolin-CoV-GXC sublineage

Library	GXP1E	GXP5L	GXP5E	GXP4L	GXP2V	Total iHVs(union)
Median coverage(25%, 75%)	420(56, 2558)	341(37, 1547)	829(130, 3662)	247(126, 443)	133(88, 254)	
Genome-wide						
Synonymous	93	61	139	15	1	217
Nonsynonymous	162	107	181	31	8	336
Frameshift	105	85	108	22	114	264
Stop-gain	3	1	0	0	0	3
Intergenic	7	9	2	3	20	29
Other	3	2	2	0	0	3
Total	373	265	432	71	143	852
***S* gene**						
Synonymous	31	8	13	1	0	36
Nonsynonymous	63	20	28	6	0	73
Frameshift	23	20	16	6	18	46
Stop-gain	1	0	0	0	0	1
Other	2	1	1	0	0	2
Total	120	49	58	13	18	158

The SRA accession Nos. are as follows: SRR11093266 (GXP1E), SRR11093267 (GXP5L), SRR11093268 (GXP5E), SRR11093269 (GXP4L), and SRR11093271 (GXP2V). Sites with coverage of ≥150X and ≥10 reads covering the alternative alleles in each library.

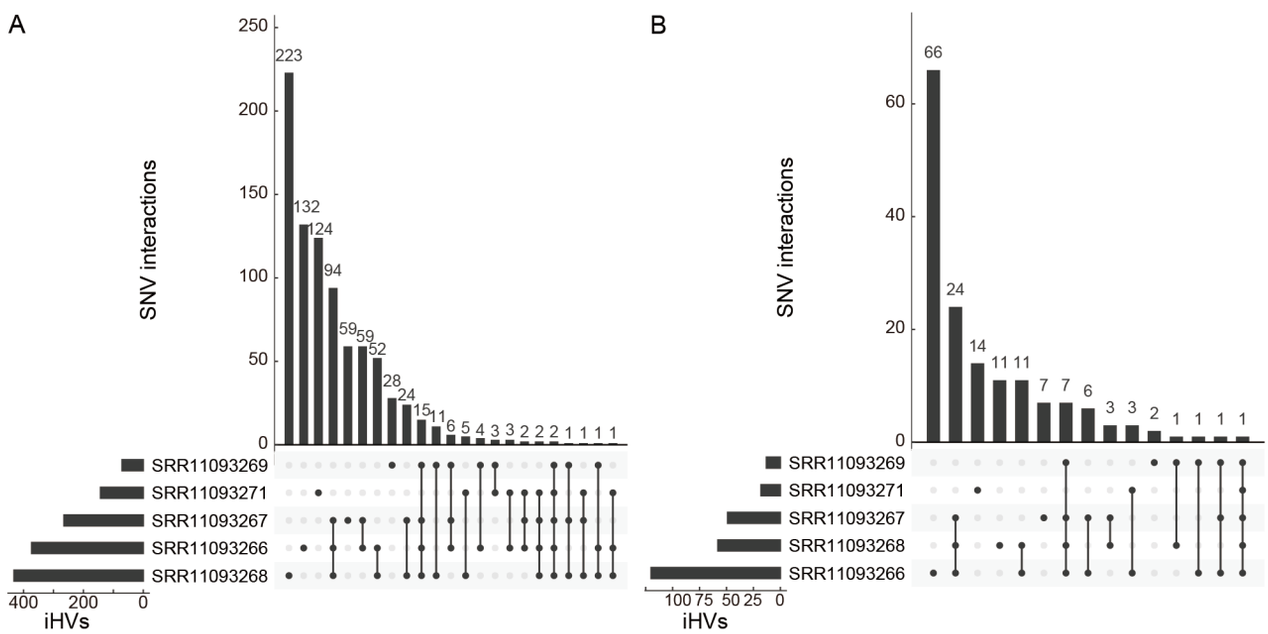

Figure 2 The numbers of overlapping iHVs between pangolin viral sequencing libraries at the genome-wide level (A) and at the level of the *S* gene (B)

interrogating the genetic diversity of the pangolin-CoVs.

Recombination events during Pangolin-CoV evolution

Since recombination is commonly observed in coronaviruses and is important for their evolution (Li et al., 2021; Sun et al., 2020; Wu et al., 2020), we asked whether different pangolin-CoV sublineages or strains have undergone recombination. We followed a previously reported strategy (Paraskevis et al., 2020) to screen putative recombination signals in the pangolin-CoV genomes and other SC2r-CoV and SARSr-CoV genomes. A putative recombination event involving a 346-nucleotide fragment (position 13 666–14 011 on the MP20 genome) in the *ORF1ab* gene was detected with the recombination detection program RDP4 (Martin et al., 2015) (Figure 3). The 346-nucleotide fragment, shared between MP20 and GXP5L, was likely introduced into the common ancestor of the pangolin-CoV-GXC sublineage from bat-derived SARSr-CoVs. Given the potential contribution of the spike protein RBD from the pangolin-CoV-GDC sublineage to SARS-CoV-2 via recombination (Liu et al., 2020; Xiao et al., 2020; Zhang et al., 2020), these results reveal that the occurrence of different recombination events in different

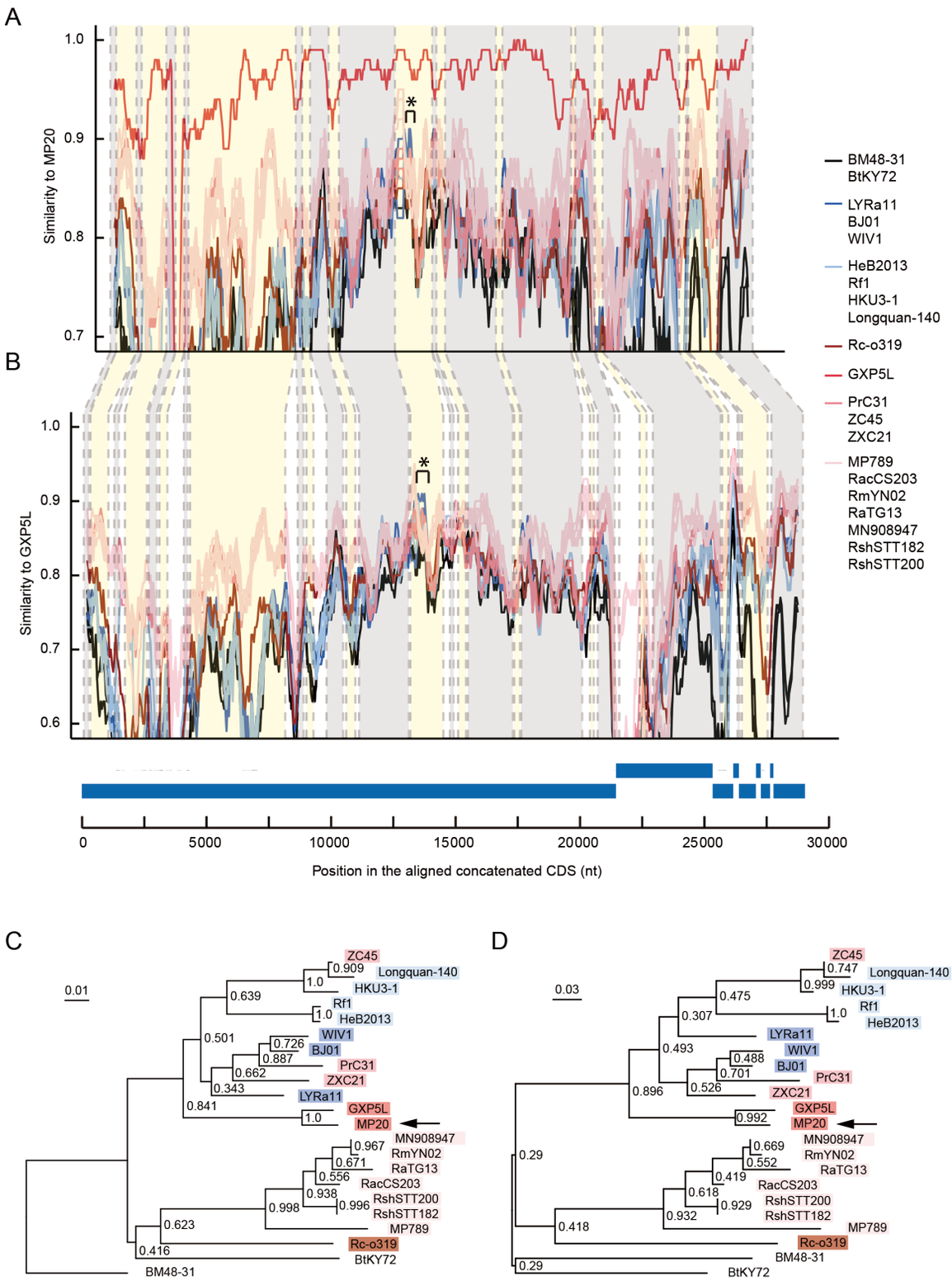

Figure 3 Recombination analysis of the pangolin-CoV-GXC sublineage

A: Similarity plot comparing 21 viruses related to MP20 from successfully captured coding sequences (CDSs). *: region for which a recombination signal was detected. SC2r-CoVs are displayed in warm colors, and SARSr-CoVs are displayed in cool colors. B: Similarity plot comparing 20 viruses related to the pangolin-CoV GXP5L. *: region for which a recombination signal was detected. C, D: Neighboring-joining phylogenetic trees for regions with a recombination signal based on all positions (C) and based on only the third positions (D) in the codons. The bootstrap values calculated from 1000 replicates and branch scale bars are shown.

pangolin-CoV sublineages further contributes to the diversity of pangolin-CoVs.

DISCUSSION

Due to the COVID-19 pandemic, coronaviruses as a research topic have garnered a tremendous amount of recent interest, and the literature contains a number of studies in which the roles of different animals in coronavirus transmission to humans are examined. Numerous studies have demonstrated that bats are rich reservoirs of the SC2r-CoVs (Hul et al., 2021; Murakami et al., 2020; Wacharapluesadee et al., 2021; Zhou et al., 2020a, 2020b, 2021). Nevertheless, a significant gap remains for the possible role of pangolins as a source of potential zoonosis of SC2r-CoVs, presumably due to the limited numbers of pangolins used to screen the SC2r-CoVs in previous studies or due to long postmortem intervals when the pangolins were confiscated in the illegal animal trades which typically lead to RNA degradation. Here, we applied the in-solution hybridization capture sequencing to retrieve the genomic sequence of an SC2r-CoV from a highly degraded biomaterial. Technologies developed for the extraction and enrichment of nucleotides from ancient samples have the potential to build more complete viral genomes (Spyrou et al., 2019). Applying this strategy to massive collections in biobanks or museums will afford promising insights into the evolutionary history of SARS-CoV-2 and other SC2r-CoVs.

This study reveals the high level of genetic diversity of SC2r-CoVs hosted in pangolins and highlights the possibility that pangolins are an important natural host or reservoir of SC2r-CoVs. Both the pangolin-CoV-GDC and pangolin-CoV-GXC viruses were isolated from *M. javanica* (Malayan pangolin) in previous studies, and our results demonstrate for the first time that pangolin-CoV is also present in *M. pentadactyla*, suggesting that pangolin-CoVs have a relatively complex evolutionary history that deserves further study. It should be noted that both pangolin-CoV-GXC and pangolin-CoV-GDC had entry efficiencies into cells expressing the human ACE2 receptor comparable to those of SARS-CoV-2 pseudoviruses (Dicken et al., 2021; Guo et al., 2021; Nie et al., 2021; Niu et al., 2021; Zhang et al., 2021). Together, the observation of multiple recombination events during the evolution of SARS-CoV-2 (Boni et al., 2020; Gryseels et al., 2021; Zhou & Shi 2021), the high genetic diversity of pangolin-CoVs in Southeast Asia and neighboring regions, and the overlapping distributions of the pangolin-CoVs, bat-CoVs, and other coronaviruses (Lacroix et al., 2017; Latinne et al., 2020), highlight the ecological risk of zoonotic transmission of novel pathogenic SC2r-CoVs. The accumulated evidence demonstrates that pangolins should be included in the search for possible natural hosts or intermediate hosts of novel coronaviruses.

DATA AVAILABILITY

The raw sequencing reads generated in this study are available in the National Genomics Data Center (NGDC) Genome Sequence Archive (GSA) database (https://ngdc.cncb.ac.cn/gsa/) under BioProject accession No. PRJCA003816. The consensus sequence of pangolin-CoV MP20 has been deposited in the NGDC Genome Warehouse (GWH) (https://ngdc.cncb.ac.cn/gwh/) under the accession No. GWHBEBQ00000000. The Sanger sequencing data were deposited in GenBank with accession Nos. MW173323, MW173324, and MW960369.

SUPPLEMENTARY DATA

Supplementary data to this article can be found online.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

Y.P.Z., M.S.P. and J.L. conceived the project and designed the research. J.B.L. and R.N.W. conducted the wet lab experiments. J.N.Z. and J.J.T. provided an in-solution capture kit and protocol. H.L., Z.F.C., X.T., M.S.P., J.B.L., R.Y., J.Y.H., Z.Y.Z., T.Z., and J.L. performed data analyses. T.T.Y. and R.N.W. prepared the samples. M.S.P., J.L., J.B.L., H.L., Z.F.C., X.T., R.Y., and T.T.Y. drafted the manuscript with input from all authors. Y.P.Z., Z.G.Z., Z.L.S., Y.G.Y., L.Y., and X.M.L. revised the manuscript. All authors read and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

We thank Shi-Fang Wu and volunteers for the sampling work. We appreciate the suggestion for in-solution capture from Yun-Gui Yang, Ming-Kun Li, and Qiao-Mei Fu. We thank Wei Chen for comments on the manuscript.

REFERENCES

Armero A, Berthet N, Avarre JC. 2021. Intra-host diversity of SARS-Cov-2 should not be neglected: case of the state of Victoria, Australia. *Viruses*, **13**(1): 133.

Banerjee A, Doxey AC, Mossman K, Irving AT. 2021. Unraveling the zoonotic origin and transmission of SARS-CoV-2. *Trends in Ecology & Evolution*, **36**(3): 180−184.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, **19**(5): 455−477.

Boni MF, Lemey P, Jiang XW, Lam TTY, Perry BW, Castoe TA, et al. 2020. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nature Microbiology*, **5**(11): 1408−1417.

Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, et al. 2007. Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(37): 14616−14621.

Campanella JJ, Bitincka L, Smalley J. 2003. MatGAT: an application that generates similarity/identity matrices using protein or DNA sequences. *BMC Bioinformatics*, **4**: 29.

Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. 2020. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Molecular Biology and Evolution*, **37**(1): 291−294.

Dicken SJ, Murray MJ, Thorne LG, Reuschl AK, Forrest C, Ganeshalingham M, et al. 2021. Characterisation of B. 1.1. 7 and pangolin coronavirus spike provides insights on the evolutionary trajectory of SARS-CoV-2. *bioRxiv*,doi: 10.1101/2021.03.22.436468.

Drexler JF, Gloza-Rausch F, Glende J, Corman VM, Muth D, Goettsche M, et al. 2010. Genomic characterization of severe acute respiratory syndrome-related coronavirus in European bats and classification of

coronaviruses based on partial RNA-dependent RNA polymerase gene sequences. *Journal of Virology*, **84**(21): 11336−11349.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv preprint*, **arXiv**: 1207.3907.

Ge XY, Li JL, Yang XL, Chmura AA, Zhu GJ, Epstein JH, et al. 2013. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature*, **503**(7477): 535−538.

Gong Z, Zhu JW, Li CP, Jiang S, Ma LN, Tang BX, et al. 2020. An online coronavirus analysis platform from the National Genomics Data Center. *Zoological Research*, **41**(6): 705−708.

Green RE, Briggs AW, Krause J, Prüfer K, Burbano HA, Siebauer M, et al. 2009. The neandertal genome and ancient DNA authenticity. *The EMBO Journal*, **28**(17): 2494−2502.

Gryseels S, De Bruyn L, Gyselings R, Calvignac-Spencer S, Leendertz FH, Leirs H. 2021. Risk of human-to-wildlife transmission of SARS-CoV-2. *Mammal Review*, **51**(2): 272−292.

Guo H, Hu B, Si HR, Zhu Y, Zhang W, Li B, et al. 2021. Identification of a novel lineage bat SARS-related coronaviruses that use bat ACE2 receptor. *Emerging Microbes & Infections*, **10**(1): 1507−1514.

He B, Zhang Y, Xu L, Yang W, Yang F, Feng Y, et al. 2014. Identification of diverse alphacoronaviruses and genomic characterization of a novel severe acute respiratory syndrome-like coronavirus from bats in China. *Journal of Virology*, **88**(12): 7070−7082.

Hu B, Guo H, Zhou P, Shi ZL. 2021. Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology*, **19**(3): 141−154.

Hu D, Zhu C, Ai L, He T, Wang Y, Ye F, et al. 2018. Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. *Emerging Microbes & Infections*, **7**(1): 1−10.

Hu JY, Hao ZQ, Frantz L, Wu SF, Chen W, Jiang YF, et al. 2020. Genomic consequences of population decline in critically endangered pangolins and their demographic histories. *National Science Review*, **7**(4): 798−814.

Hul V, Delaune D, Karlsson EA, Hassanin A, Tey PO, Baidaliuk A, et al. 2021. A novel SARS-CoV-2 related coronavirus in bats from Cambodia. *bioRxiv*,doi: 10.1101/2021.01.26.428212.

Johnson MG, Pokorny L, Dodsworth S, Botigué LR, Cowan RS, Devault A, et al. 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology*, **68**(4): 594−606.

Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, **20**(4): 1160−1166.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, **35**(6): 1547−1549.

Lacroix A, Duong V, Hul V, San S, Davun H, Omaliss K, et al. 2017. Genetic diversity of coronaviruses in bats in Lao PDR and Cambodia. *Infection, Genetics and Evolution*, **48**: 10−18.

Lam TTY, Jia N, Zhang YW, Shum MHH, Jiang JF, Zhu HC, et al. 2020. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature*, **583**(7815): 282−285.

Latinne A, Hu B, Olival KJ, Zhu G, Zhang L, Li H, et al. 2020. Origin and cross-species transmission of bat coronaviruses in China. *Nature Communications*, **11**(1): 4235.

Lau SKP, Woo PCY, Li KSM, Huang Y, Tsoi HW, Wong BHL, et al. 2005. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(39): 14040−14045.

Lee J, Hughes T, Lee MH, Field H, Rovie-Ryan JJ, Sitam FT, et al. 2020. No evidence of coronaviruses or other potentially zoonotic viruses in Sunda pangolins (*Manis javanica*) entering the wildlife trade via Malaysia. *EcoHealth*, **17**(3): 406−418.

Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics*, **31**(10): 1674−1676.

Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, **30**(20): 2843−2851.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**(16): 2078−2079.

Li LL, Wang JL, Ma XH, Sun XM, Li JS, Yang XF, et al. 2021. A novel SARS-CoV-2 related coronavirus with complex recombination isolated from bats in Yunnan province, China. *Emerging Microbes & Infections*, **10**(1): 1683−1690.

Li T, Tang X, Wu C, Yao X, Wang Y, Lu X, et al. 2020. The use of SARS-CoV-2-related coronaviruses from bats and pangolins to polarize mutations in SARS-Cov-2. *Science China Life Sciences*, **63**(10): 1608−1611.

Li W, Shi Z, Yu M, Ren W, Smith C, Epstein JH, et al. 2005. Bats are natural reservoirs of SARS-like coronaviruses. *Science*, **310**(5748): 676−679.

Lin XD, Wang W, Hao ZY, Wang ZX, Guo WP, Guan XQ, et al. 2017. Extensive diversity of coronaviruses in bats from China. *Virology*, **507**: 1−10.

Liu P, Chen W, Chen JP. 2019. Viral metagenomics revealed sendai virus and coronavirus infection of Malayan pangolins (*Manis javanica*). *Viruses*, **11**(11): 979.

Liu P, Jiang JZ, Wan XF, Hua Y, Li L, Zhou J, et al. 2020. Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)?. *PLoS Pathogens*, **16**(5): e1008421.

Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, et al. 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *Journal of Virology*, **73**(1): 152−160.

Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet*, **395**(10224): 565−574.

Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. 2015. RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, **1**(1): vev003.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, **17**(1): 10−12.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**(9): 1297−1303.

Murakami S, Kitamura T, Suzuki J, Sato R, Aoi T, Fujii M, et al. 2020. Detection and characterization of bat sarbecovirus phylogenetically related to SARS-CoV-2, Japan. *Emerging Infectious Diseases*, **26**(12): 3025−3029.

Nash HC, Wirdateti, Low GW, Choo SW, Chong JL, Semiadi G, et al. 2018. Conservation genomics reveals possible illegal trade routes and admixture across pangolin lineages in Southeast Asia. *Conservation Genetics*, **19**(5): 1083−1095.

Nie J, Li Q, Zhang L, Cao Y, Zhang Y, Li T, et al. 2021. Functional

comparison of SARS-CoV-2 with closely related pangolin and bat coronaviruses. *Cell Discovery*, **7**(1): 21.

Niu S, Wang J, Bai B, Wu L, Zheng A, Chen Q, et al. 2021. Molecular basis of cross-species ACE2 interactions with SARS-CoV-2-like viruses of pangolin origin. *The EMBO Journal*, **40**(16): e107786.

Ou X, Liu Y, Lei X, Li P, Mi D, Ren L, et al. 2020. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nature Communications*, **11**(1): 1620.

Paraskevis D, Kostaki EG, Magiorkinis G, Panayiotakopoulos G, Sourvinos G, Tsiodras S. 2020. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. *Infection, Genetics and Evolution*, **79**: 104212.

Qin E, Zhu Q, Yu M, Fan B, Chang G, Si B, et al. 2003. A complete sequence and comparative analysis of a SARS-associated virus (Isolate BJ01). *Chinese Science Bulletin*, **48**(10): 941−948.

Qu XX, Hao P, Song XJ, Jiang SM, Liu YX, Wang PG, et al. 2005. Identification of two critical amino acid residues of the severe acute respiratory syndrome coronavirus spike protein for its variation in zoonotic tropism transition via a double substitution strategy. *Journal of Biological Chemistry*, **280**(33): 29588−29595.

Ren W, Qu X, Li W, Han Z, Yu M, Zhou P, et al. 2008. Difference in receptor usage between severe acute respiratory syndrome (SARS) coronavirus and SARS-like coronavirus of bat origin. *Journal of Virology*, **82**(4): 1899−1907.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. 2011. Integrative genomics viewer. *Nature Biotechnology*, **29**(1): 24−26.

Schwarz G. 1978. Estimating the dimension of a model. *The Annals of Statistics*, **6**(2): 461−464.

Spyrou MA, Bos KI, Herbig A, Krause J. 2019. Ancient pathogen genomics as an emerging tool for infectious disease research. *Nature Reviews Genetics*, **20**(6): 323−340.

Sun J, He WT, Wang L, Lai A, Ji X, Zhai X, et al. 2020. COVID-19: epidemiology, evolution, and cross-disciplinary perspectives. *Trends in Molecular Medicine*, **26**(5): 483−495.

Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. , 2020. On the origin and continuing evolution of SARS-CoV-2. *National Science Review*, **7**(6): 1012–1023.

Tao Y, Tong S. 2019. Complete genome sequence of a severe acute respiratory syndrome-related coronavirus from Kenyan bats. *Microbiology Resource Announcements*, **8**(28): e00548−19.

Wacharapluesadee S, Tan CW, Maneeorn P, Duengkae P, Zhu F, Joyjinda Y, et al. 2021. Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia. *Nature Communications*, **12**(1): 972.

Wan Y, Shang J, Graham R, Baric RS, Li F. 2020. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. *Journal of Virology*, **94**(7): e00127−20.

Wang Y, Wang D, Zhang L, Sun W, Zhang Z, Chen W, et al. 2021. Intra-host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19 patients. *Genome Medicine*, **13**(1): 30.

Wen S, Sun C, Zheng H, Wang L, Zhang H, Zou L, et al. 2020. High-coverage SARS-CoV-2 genome sequences acquired by target capture sequencing. *Journal of Medical Virology*, **92**(10): 2221−2226.

WHO. 2021. WHO-convened global study of origins of SARS-CoV-2: China part. Geneva: WHO, 120.

Wong G, Bi YH, Wang QH, Chen XW, Zhang ZG, Yao YG. 2020. Zoonotic origins of human coronavirus 2019 (HCoV-19 / SARS-CoV-2): why is this work important?. *Zoological Research*, **41**(3): 213−219.

Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, et al. 2020. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, **367**(6483): 1260−1263.

Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. 2020. A new coronavirus associated with human respiratory disease in China. *Nature*, **579**(7798): 265−269.

Xia X. 2018. DAMBE7: new and improved tools for data analysis in molecular biology and evolution. *Molecular Biology and Evolution*, **35**(6): 1550−1552.

Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou JJ, et al. 2020. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature*, **583**(7815): 286−289.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**(8): 1586−1591.

Zhang S, Qiao S, Yu J, Zeng J, Shan S, Tian L, et al. 2021. Bat and pangolin coronavirus spike glycoprotein structures provide insights into SARS-CoV-2 evolution. *Nature Communications*, **12**(1): 1607.

Zhang T, Wu Q, Zhang Z. 2020. Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. *Current Biology*, **30**(7): 1346−1351.e2.

Zhang YZ, Holmes EC. 2020. A genomic perspective on the origin and emergence of SARS-CoV-2. *Cell*, **181**(2): 223−227.

Zhou H, Chen X, Hu T, Li J, Song H, Liu Y, et al. 2020a. A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the spike protein. *Current Biology*, **30**(11): 2196−2203.e3.

Zhou H, Ji J, Chen X, Bi Y, Li J, Wang Q, et al. 2021. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. *Cell*, **184**(17): 4380−4391.e14.

Zhou P, Shi ZL. 2021. SARS-CoV-2 spillover events. *Science*, **371**(6525): 120−122.

Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. 2020b. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, **579**(7798): 270−273.