
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Infection, Genetics and Evolution 93 (2021) 104933

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

# Exploration of hosts and transmission traits for SARS-CoV-2 based on the k-mer natural vector

Yuyan Zhang [a,1], Jia Wen [a,b,1,*], Xin Li [c], Guizhi Li [d]

[a] School of Information Engineering, Suihua University, Suihua 152061, China
[b] Warshel Institute for Computational Biology, The Chinese University of Hong Kong (Shenzhen), Shenzhen 518172, China
[c] Tianjin International Joint Research Center for Neural Engineering, Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin 30072, China
[d] Yingkou Institute of Technology, Yingkou 115014, China

ARTICLE INFO

*Keywords:*
Phylogenetic analysis
Bayesian
K-mer model
Intermediate host
Cross-species

ABSTRACT

A severe respiratory pneumonia COVID-19 has raged all over the world, and a coronavirus named SARS-CoV-2 is blamed for this global pandemic. Despite intensive research into the origins of the COVID-19 pandemic, the evolutionary history of its agent SARS-CoV-2 remains unclear, which is vital to control the pandemic and prevent another round of outbreak. Coronaviruses are highly recombinogenic, which are not well handled with alignment-based method. In addition, deletions have been found in the genomes of several SARS-CoV-2, which cannot be resolved with current phylogenetic methods. Therefore, the k-mer natural vector is proposed to explore hosts and transmission traits for SARS-CoV-2 using strict phylogenetic reconstruction. SARS-CoV-2 clustering with bat-origin coronaviruses strongly suggests bats to be the natural reservoir of SARS-CoV-2. By building bat-to-human transmission route, pangolin is identified as an intermediate host, and civet is predicted as a possible candidate. We speculate that SARS-CoV-2 undergoes cross-species recombination between bat and pangolin coronaviruses. This study also demonstrates transmission mode and features of SARS-CoV-2 in the COVID-19 pandemic when it broke out early around the world.

## 1. Introduction

Emerging and re-emerging of virulent infection disease presents a great threat to the public health (Gao, 2018). The outbreak of COVID-19, a severe respiratory pneumonia, in Wuhan, China, has captured the attention of the world. A novel coronavirus named SARS-CoV-2 is thought as the culprit of this epidemic, which is the seventh pathogenic coronavirus to human (Su et al., 2016; Lu et al., 2020). Four coronaviruses of 229E, OC43, NL63, and HKU1 are mild and typically cause cold symptoms in immunocompetent individuals (Drosten et al., 2003), whereas severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV) are highly pathogenic and linked to high mortality (Cui et al., 2019). Depending on a high transmissibility, the COVID-19 has spread throughout the world and upgraded to a global pandemic.

Initial analysis indicates that SARS-CoV-2 belongs to the genus Betacoronavirus (BetaCoV), containing six major open-reading frames (ORFs) in virus genome and some accessory genes (Wu et al., 2020b; Zhou et al., 2020a). The first ORF (denoted as Orf1ab) occupying nearly half of entire virus genome encodes 16 non-structure proteins, while remaining ORFs encode structural proteins and accessory proteins, of which four main structural proteins are spike surface glycoprotein (S), small envelop protein (E), matrix protein (M), and nucleocapsid protein (N). Of note, the S protein mediates receptor binding and membrane fusion, and determines host tropism and transmission capacity (Jaimes et al., 2020).

Coronaviruses are zoonotic pathogens that are naturally hosted by bats (Guan et al., 2003; Lau et al., 2020). Phylogenetic analysis has shown SARS-CoV-2 clustering with bat-derived SARS related coronaviruses (SARSr-CoVs) within the genus BetaCoV, of which RaTG13 is observed the highest degree of sequence identity to SARS-CoV-2 (Zhou et al., 2020a). It is also confirmed that SARS-CoV-2 has close similarity to SARS-CoV, particular in the receptor-binding domain (RBD) of the S protein. Since human infections of bat-origin viruses typically occur through intermediate hosts, the Malayan pangolin has been suggested as

* Corresponding author at: Warshel Institute for Computational Biology, The Chinese University of Hong Kong (Shenzhen), Shenzhen 518172, China
*E-mail address:* wenjia@cuhk.edu.cn (J. Wen).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.meegid.2021.104933
Received 18 April 2021; Received in revised form 8 May 2021; Accepted 19 May 2021
Available online 20 May 2021

**Abbreviations**

| | |
|---|---|
| SARS-CoV | severe acute respiratory syndrome coronavirus |
| MERS-CoV | Middle East respiratory syndrome coronavirus |
| BetaCoV | Betacoronavirus |
| ORFs | open-reading frames |
| SARSr-CoVs | SARS related coronaviruses |
| RBD | receptor-binding domain |
| Pan-CoV | Pangolin coronavirus |
| ML | Maximum-likelihood |
| MSA | multiple sequence alignment |
| SL | SARS-like |
| Env-CoVs | environmental coronaviruses |
| RMSD | root-mean-square deviation |
| ACE2 | angiotensin-converting enzyme 2 |
| NJ | Neighbor-Joining |

an intermediate host of SARS-CoV-2 (Lam et al., 2020; Xiao et al., 2020). Although the RBDs in the S protein from Malayan pangolin are well conserved to SARS-CoV-2 (Highest at 97%), whole-genome analysis reveals 85.5%–92.0% sequence identity, which are less than what is observed from RaTG13 (over 96%). Thus, the phylogenies between pangolin coronavirus (Pan-CoV) and SARS-CoV-2 could not be served as the direct evidence of pangolin being an intermediate host of SARS-CoV-2.

Coronaviruses are highly recombinogenic that are not well handled with alignment-based method (Zielezinski, 2017), so (Boni et al., 2020) had to remove the effects of recombination and used putative non-recombinant regions to predict the origin of SARS-CoV-2. In addition, deletions have been found in the genomes of several SARS-CoV-2, which indicates human adaptation after transmission and could not be accurately reflected by current phylogenetic methods (Young et al., 2020). It was demonstrated that the k-mer model method could capture recombination events and deal with the cases with deletions efficiently (Bauer et al., 2020). However, the k-mer approach is not suggested to tract potential transmission route for its non-uniqueness. To this end, the k-mer natural vector is proposed to characterize the compositions and distributions of k-mers occurance in a virus genome, and construct one-to-one relationship between a virus genome and its k-mer natural vector. Based on this, we determine the classification of SARS-CoV-2, identify its origin and intermediate hosts, and tract transmission mode and features of SARS-CoV-2 in the COVID-19 epidemic, which deepen our understanding of the recombination for viruses among cross-species transmission.

## 2. Results

### 2.1. Classification of SARS-CoV-2

To validate the efficiency of the k-mer natural vector, all viruses from the family Coronaviridae in NCBI's RefSeq database are applied to determine the classification of SARS-CoV-2, in which one sequence designated as Wuhan-Hu-1 is the reference strain for SARS-CoV-2. Phylogenetic tree for coronaviruses in RefSeq database is shown in Fig. 1(a), in which different colors represent different virus types. As a comparison, results obtained by multiple sequence alignment (MSA) with ClustalW are shown in Fig. 1(b). Comparing Figs. 1(a) and (b), both results are consistent with each other, which cannot precisely depicted by common k-mer model methods. Wuhan-Hu-1 and SARS-CoV are clustered together, grouping with bat viruses of BM48-31/BGR/2008 and Hp_BetaCoV. It is indicated that Wuhan-Hu-1 is closely related to SARS-CoV in phylogeny and suggested as a sister clade to SARS-CoV, which was, therefore, named SARS-CoV-2 by the International Committee on Taxonomy of Viruses.

Furthermore, viruses in the genus BetaCoV are used to determine the classification of SARS-CoV-2 at Genus level. Phylogenies for whole-genome sequence and genes encoding non-structural protein Orf1ab and structural proteins of S, E, M, and N are shown with similar structures (Figs. 2(a)–2(f)), in which viruses are classified into subgenera of Sarbecovirus, Hibecovirus, Merbecovirus, Nobecovirus, and Embevovirus. In special, Wuhan-Hu-1 always falls in basal position within the subgenus Sarbecovirus, and tends to cluster with bat SARS-like (SL) viruses of CoVZC45 and CoVZXC21, which is in line with results obtained by alignment-based methods (Wu et al., 2020a; Zhu et al., 2020).

### 2.2. Origin and intermediate hosts of SARS-CoV-2

Identification of the origin and intermediate hosts is current urgent task to be done, which is vital to control virus spread and prevent another round of epidemic outbreak. It has been shown that SARS-CoV-2 clusters with bat-derived SL-CoVs, indicating that SARS-CoV-2 might originate from bats. In Fig. 3(a), closely related coronaviruses are utilized to identify the origin of SARS-CoV-2. Wuhan-Hu-1 clusters with viruses of SARS-CoV-2 (WIV02, WIV04-07) with high sequence identity, plus environmental coronaviruses (Env-CoVs) sampled from the seafood market (IVDC-HB-envF13-20, 21) with distances less than 0.0010 (data are not shown). Bat-CoV RaTG13 shows the highest homology to SARS-CoV-2 among all current known SARSr-CoVs. In addition, Bat/Yunnan/RmYN02 is closely related to SARS-CoV-2, especially a peptide insertion at S1/S2 cleavage site (Zhou et al., 2020a). RaTG13 and RmYN02 are both obtained from Rhinolophus bats, and SARS-CoV-2 cluster with bat-origin CoVs, so bats are identified as natural reservoir of SARS-CoV-2.

Bats' ecological separation from human makes it probable that other animals act as intermediate hosts that transmit viruses to human (Cui et al., 2019; Lam et al., 2020; Xiao et al., 2020; Shi, 2020; Sit, 2020). For example, SARS-CoV and MERS-CoV are originated from bat, then transmitted to civet (Song et al., 2005) or camel (Wang et al., 2016), and finally to human. It is reported that the RBD of the S gene from Guangdong Pan-CoV is conserved to SARS-CoV-2 (Lam et al., 2020; Xiao et al., 2020). Besides pangolin, mink, snake, turtle, cat, and dog have been proposed as intermediate hosts (Li et al., 2020; Xia, 2020; Shi et al., 2020; Oreshkova et al., 2020; Sit et al., 2020; Zhang et al., 2020). Since there is no possible way to get sufficient sampling to determine intermediate hosts of SARS-CoV-2, it is necessary to build transmission route from the origin to intermediate hosts. Since human is thought as the terminal host of SARS-CoV-2, an inference of bat-to-human transmissiion route looks more effective. Based on the transmission modes of animal origins of human coronaviruses (Cui et al., 2019), coronavirus groups are chosen from all possible animal hosts, and distance for each pair of virus groups is depicted the similarity between animal hosts (Tables S1–S4), in which both whole-genome and S gene sequences are considered. Moreover, Mean distance and Center distance are applied. In Fig. 3(b), two bat-to-human transmission routes are inferred (see Text S1 for more detail). The only difference between two transmission routes is whether civet has taken part in the genetic recombination of SARS-CoV-2; however, pangolin is always adjacent to human, and identified as an intermediate host of SARS-CoV-2. Meanwhile, civet is predicted as a possible candidate.

The S protein is a significant driver in virus evolution through binding with receptor protein (Wrobel et al., 2021). To validate pangolin as an intermediate host of SARS-CoV-2, the crystal structure of the S protein for representatives from SARS-CoV-2, Pan-CoV, SARS-CoV, and Bat-CoV are built by homology modelling using SWISS-MODEL server, and pairwise values of root-mean-square deviation (RMSD) to the 3D structure of Wuhan-Hu-1 are 2.34 (M789), 3.09 (Civet007), and 5.78 (HKU2), respectively. In Fig. 3(c), the structure of the S protein from Pan-CoV (M789) is the most similar to that of SARS-CoV-2 (Wuhan-Hu-1), which coincides with results from RMSD values. In addition, the similarity analysis for the S genes from close related coronaviruses is

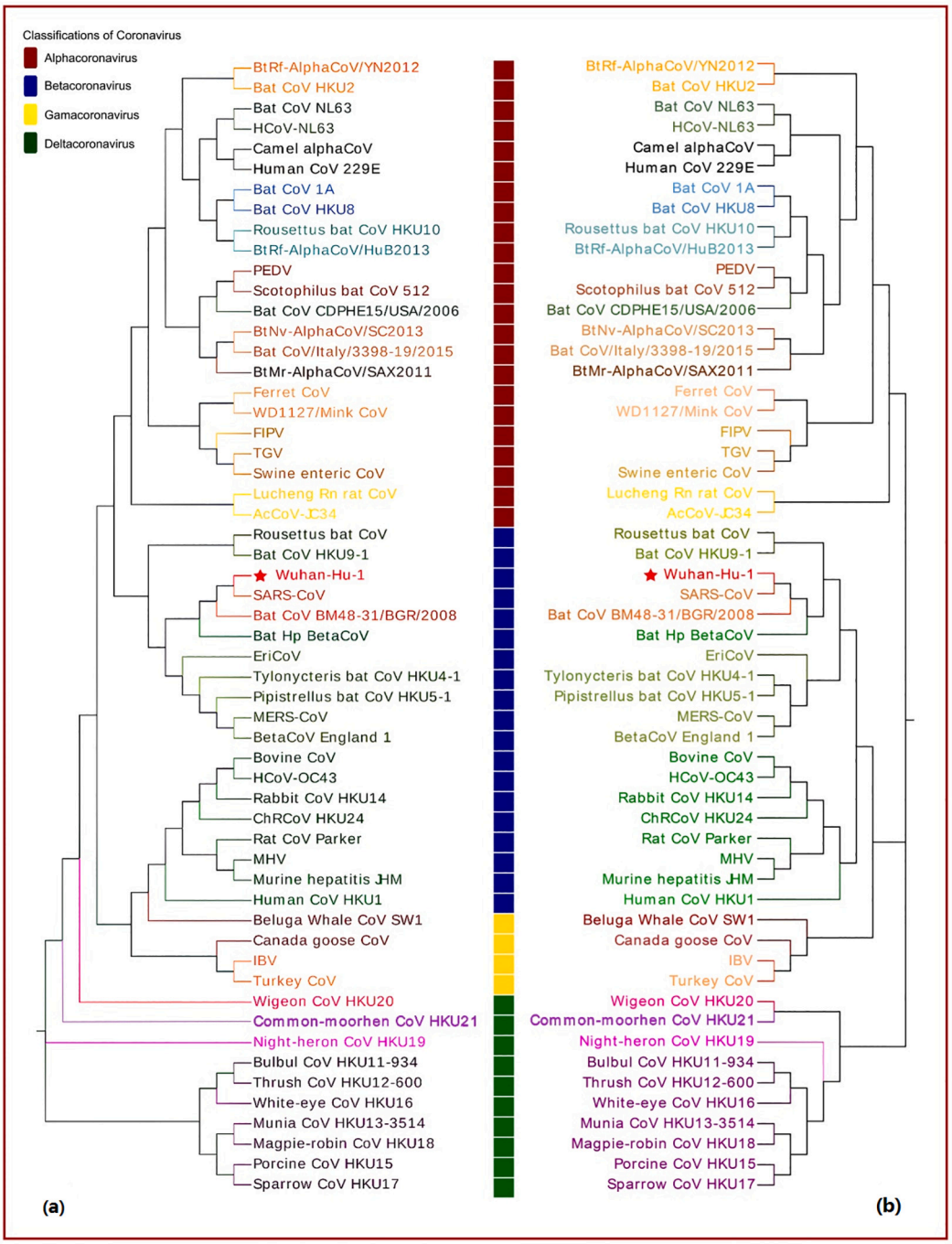


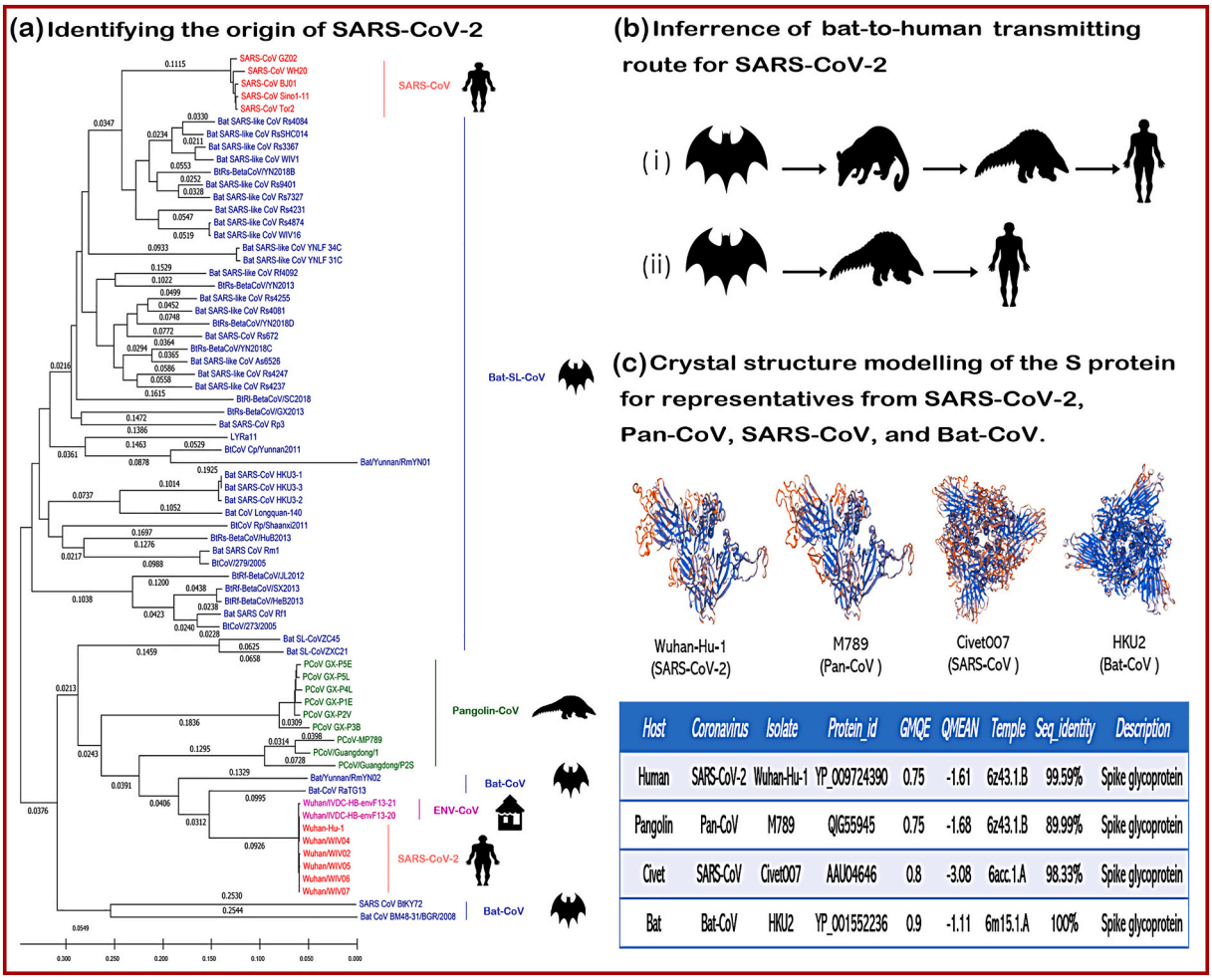

**Fig. 1.** Phylogenetic trees of all viruses from the family Coronaviridae in NCBI's RefSeq database is shown the classification of SARS-CoV-2, which are classified into four clades, including Alphacoronavirus, Betacoronavirus, Gammacoronavirus, and Deltacoronavirus. Phylogenetic tree constructing with the k-mer natural vector is shown in (a), in which different colors represent different virus types. As a comparison, the results gotten by MSA with ClustalW are shown in (b).

performed. As shown in Fig. 4(a), it is confirmed again that RaTG13 is the closest to SARS-CoV-2, and the recombination in SARS-CoV-2 is noted, which suggests cross-species recombination between bat and pangolin CoVs exists in the evolution of SARS-CoV-2. Furthermore, the RBDs in the S protein are compared, in which the ACE2 critical contact sites are highlighted with arrows in Fig. 4(b). It is obvious that all critical contact sites in the Pan-CoV Guangdong/1 are consistent with that of SARS-CoV-2, which proves that pangolin should be an intermediate host in the emergence of SARS-CoV-2.

### *2.3. Transmitting mode and features of SARS-CoV-2 at the beginning of the COVID-19 pandemic*

It has been more than one year since the outbreak of COVID-19 in Wuhan, China, but transmission mode and features are still unclear. Because of many asymptomatic infections, it is likely that virus emerged earlier in human than envisaged (Chinazzi et al., 2020). SARS-CoV-2 sampled at early stage of the epidemic is closely related to Env-CoVs sampled from the seafood market (Fig. 3(a)). It is indicated that there existed plenty of viruses at the seafood market when the epidemic broke out, and this "clammy" market should play an important role in virus


**Fig. 2.** Phylogenies of viruses in the genus BetaCoV is shown the classification of SARS-CoV-2 at Genus level. Phylogenies of whole-genome sequence (a), non-structural protein gene Orf1ab (b), genes encoding structural proteins of S (c), E (d), M (e), and N (f) are shown with the k-mer natural vector, in which Beta-CoVs are classified into subgenera of Sarbecovirus, Hibecovirus, Merbecovirus, Nobecovirus, and Embevovirus.




| Host | Coronavirus | Isolate | Protein_id | GMQE | QMEAN | Temple | Seq_identity | Description |
|---|---|---|---|---|---|---|---|---|
| Human | SARS-CoV-2 | Wuhan-Hu-1 | YP_009724390 | 0.75 | -1.61 | 6z43.1.B | 99.59% | Spike glycoprotein |
| Pangolin | Pan-CoV | M789 | QIG55945 | 0.75 | -1.68 | 6z43.1.B | 89.99% | Spike glycoprotein |
| Civet | SARS-CoV | Civet007 | AAU04646 | 0.8 | -3.08 | 6acc.1.A | 98.33% | Spike glycoprotein |
| Bat | Bat-CoV | HKU2 | YP_001552236 | 0.9 | -1.11 | 6m15.1.A | 100% | Spike glycoprotein |

**Fig. 3.** Origin and intermediate hosts of SARS-CoV-2. (a) Closely related coronaviruses from SARS-CoV, Bat-SL-CoV, Pan-CoV, Bat-CoV, and Env-CoV sampled from the seafood market are utilized to identify the origin of SARS-CoV-2 using k-mer natural vector. (b) Bat-to-human transmission route is inferred to ascertain intermediate hosts of SARS-CoV-2. (c) Crystal structure modelling of the S protein for representatives from SARS-CoV-2 (Wuhan-Hu-1), Pan-CoV (M789), SARS-CoV (Civet007), and Bat-CoV (HKU2).

transmission to human. In addition, human-to-human transmission has been confirmed in family clustering and hospital personnel (Lu et al., 2020; Zhou et al., 2020b; Chan et al., 2020).

SARS-CoV-2 is shown with location-linkage: viruses from neighboring locations commonly clustering together (see Fig. 5(a)–(c)). Besides different out-groups utilized, 141 virus genomes from human SARS-CoV-2 were downloaded from GISAID database with submission date on or before February 29, 2020, when the COVID-19 had escalated to a global pandemic. To crack transmitting features of SARS-CoV-2 at the beginning of the COVID-19 pandemic, the root of viruses was carefully tested by introducing out-groups of Bat-CoVs, Pan-CoVs, and HIVs, respectively. Since phylogenetic trees are shown with similar topologies and viruses are hypothesized spread from the root region, it is indicated that SARS-CoV-2 might have existed in several regions of the world when it broke out in Wuhan, China (Deslandes et al., 2020). It is also noted that most viruses near the root region are from Australia and the USA, which is consistent with results from phylogenetic network analysis of SARS-CoV-2 (Forster et al., 2020).

## 3. Discussion

The COVID-19 caused by SARS-CoV-2 had terrible influences on human lives, so it is urgent to identify the origin and intermediate hosts, which is the main objective of this study. The k-mer natural vector is proposed to fulfil this tough task. SARS-CoV-2 clustering with bat-origin coronaviruses strongly suggests bats serving as the natural reservoir for SARS-CoV-2. Although Malayan pangolin was thought as a possible intermediate host, the result from phylogenies does not support this induction.

To ascertain intermediate hosts of SARS-CoV-2, bat-to-human transmission route is built based on the similarities of coronavirus groups chosen from all possible animal hosts. It is identified that pangolin is an intermediate host in SARS-CoV-2 transmission, which coincides with results from the modelled structure comparisons of the S proteins, as well as the high sequence and structural similarities among RBDs. In addition, civet is predicted as a possible candidate, because SARS-CoV-2 is closely related to SARS-CoV in phylogeny, especially the peptide insertion at S1/S2 cleavage site in the S protein. It is strongly suggested that SARS-CoV-2 has a history of cross-species recombination

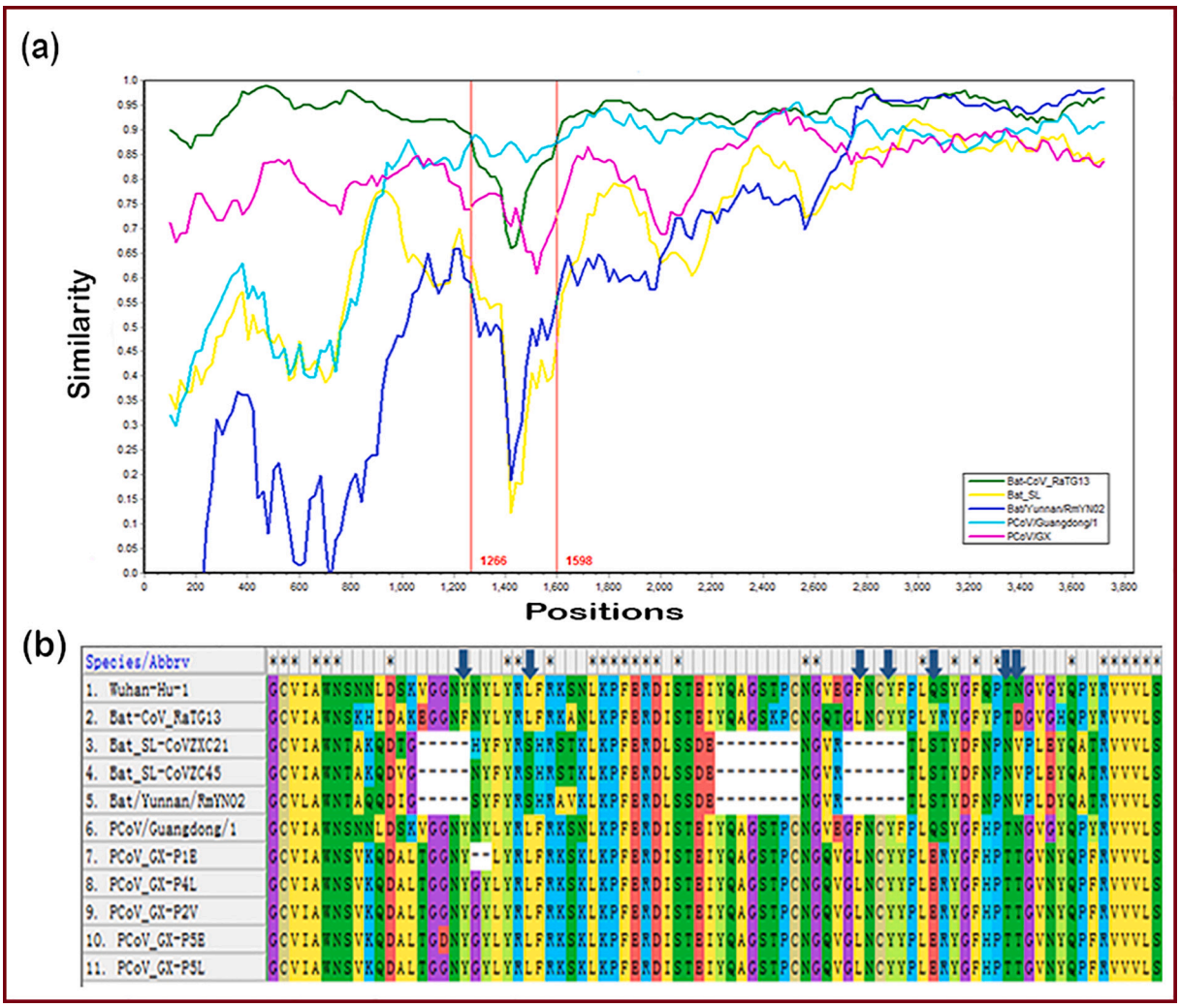



**Fig. 4.** Similarity analysis for the S genes between SARS-CoV-2 and several close related coronaviruses. (a) The similarity plot for the S gene comparisons between Human-Hu-1 and CoVs from bat and pangolin indicates the recombination of SARS-CoV-2 among bat and pangolin CoVs using Simplot. (b) The RBDs in the S proteins are compared with ClustalW, in which the ACE2 critical contact sites are highlighted with arrows. Here, Human-Hu-1 is the reference sequence for SARS-CoV-2.

between bat and pangolin CoVs. Moreover, pangolin and civet are both wild mammals sold at the seafood market when the epidemic broke out, which coincides with rules for intermediate host (Zhang and Holmes, 2020).

It is predicted that the virus might have spread when it broke out in Wuhan, China, for many asymptomatic infections. In addition, several evidences have shown that the cold and wet circumstance is good for virus transmission, as well as human-to-human transmission. To depict transmission features of SARS-CoV-2, the root of viruses has been carefully tested by introducing different out-groups. It is obvious that viruses in neighboring locations often cluster together showing with strong location-linkage. Combining virus location with the timeline, it is suggested the virus having existed in several regions of the world when it broke out in Wuhan, China, which needs to be verified with more evidences from different research areas.

In this study, the k-mer natural vector is proposed to explore hosts and transmitting traits for SARS-CoV-2 using strict phylogenetic reconstruction, in which the k-mer natural vector is well kept the ability to deal with recombination and deletions often existing in virus genome, and overcomes the deficiencies of previous k-mer models. Although the k-mer model methods have been proposed for several years, and some methods based on the k-mer models have been optimized, but all of these methods lose many important biological information, namely there is no way to recover the original genome sequence. One significant novelty of our k-mer natural vector is that each virus genome can be rigorously recovered by its corresponding k-mer natural vector. Compared with alignment-based method, our k-mer natural vector concerns global similarities of genomes, such as the changes averaged across whole genome rather than at specific locations (shared mutations) and require no evolutionary model or human intervention. The k-mer natural vector is a good choice in virus research that precisely describes the phylogenetic relationships and greatly enhances computational efficiency (see Table S5), especially facing volumes of data extremely increasing.

## 4. Material and methods

### 4.1. Dataset

Virus genomes used in this study are collected from datasets of GenBank and GISAID with basic sequence information (see Dataset.xls).

Dataset 1: all viruses from the family Coronaviridae in NCBI's RefSeq database are collected to determine the classification of SARS-CoV-2, in which Wuhan-Hu-1 is the reference sequence for SARS-CoV-2.

Dataset 2: viruses in the genus BetaCoV are used to determine the classification of SARS-CoV-2 at Genus level, in which whole-genome sequence, non-structural protein gene Orf1ab, and genes encoding structural proteins of S, E, M, and N are utilized.

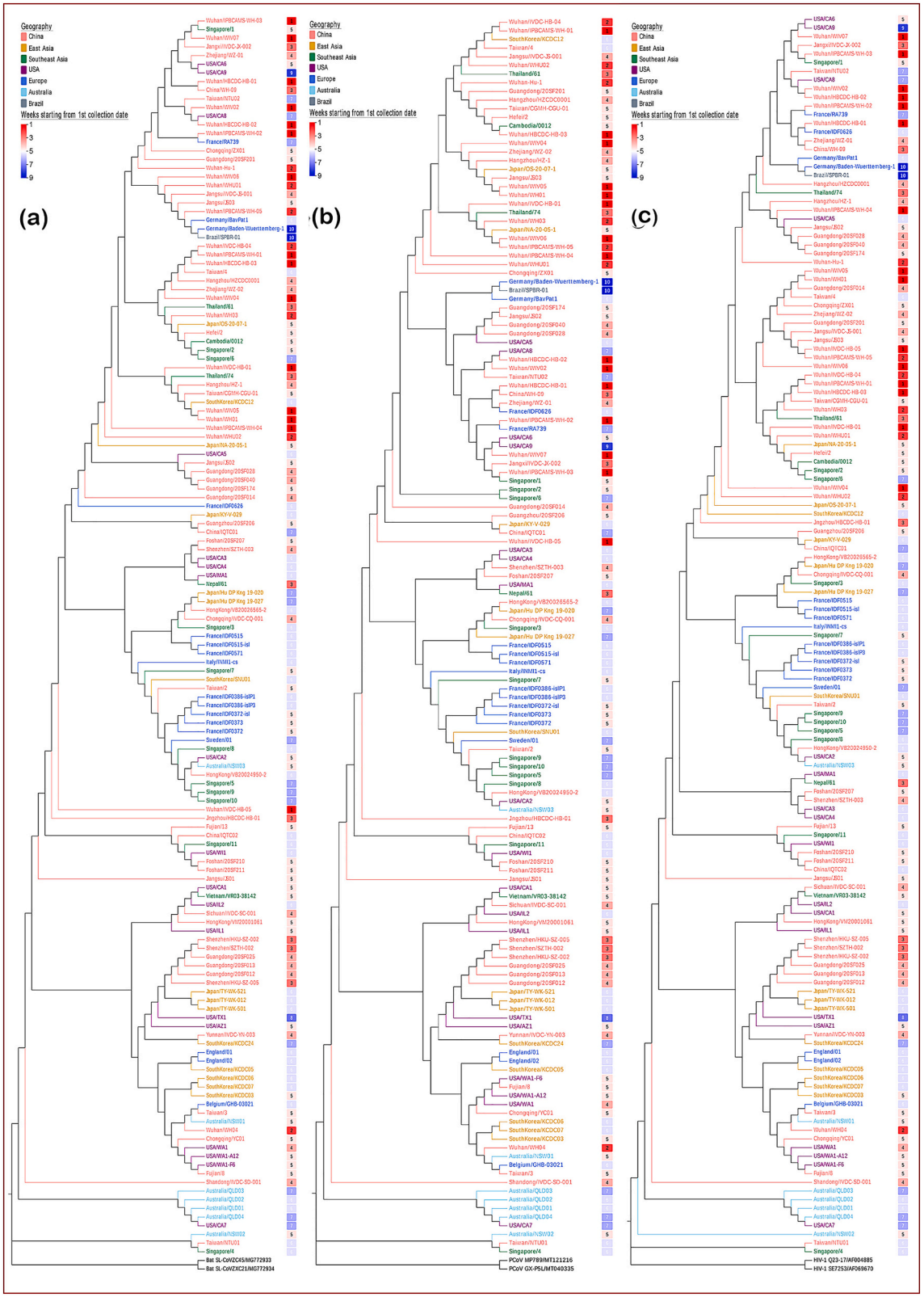



**Fig. 5.** Phylogenies for 141 viruses from Human SARS-CoV-2 are applied to crack transmission features of SARS-CoV-2 at the beginning of the COVID-19 pandemic based on the k-mer natural vector, by introducing out-groups of Bat-CoVs (a), Pan-CoVs (b), and HIVs (c), respectively.

Dataset 3: closely related coronaviruses from SARS-CoV, Bat-SL-CoV, Pan-CoV, Bat-CoV, and Env-CoV sampled from the seafood market are applied to identify the origin of SARS-CoV-2.

Dataset 4: according to transmission mode of animal origins of human coronaviruses (Cui et al., 2019), coronavirus groups are chosen for possible animals to ascertain the intermediate hosts of SARS-CoV-2 (Lam, 2020; Shi, 2020; Sit, 2020; Xia, 2020; Xiao, 2020; Zhang, 2020).

Dataset 5: except different out-groups, a total of 141 virus genomes from human SARS-CoV-2 viruses from GISAID with submission date on or before February 29, 2020, are applied to crack transmitting features. Any sequence with Ns is discarded. Sequences of Shenzhen/SZTH-001 (EPI_ISL_406592), Shenzhen/SZTH-004 (EPI_ISL_406595), TaiWan/NTU01 (EPI_ISL_408489), and Singapore/4 (EPI_ISL_410535) are also excluded;.

### 4.2. K-mer natural vector for virus genome

Let s = $'N_1N_2\cdots N_L'$ be a virus genome with slength $L$, Lwhere $N_l \in \{A, C, G, T\}$, $l = 1, 2, \cdots, L$, and s[$j$][$i$] be the location of the $i$-th occurrence of a k-mer s[$j$] in s, $j = 1, 2, \cdots, 4^k$. For each given $k$, the distributions of a k-mer s[$j$] can be described by three quantities:

$n_{s[j]}$: Number of s[$j$] occurrences in $s$;

$\mu_{s[j]}$: Mean distance of s[$j$] from the first position of $s$;

$D_m^{s[j]}$: Central moment of s[$j$], that is,

$$D_m^{s[j]} = \sum_{i=1}^{n_{s[j]}} \frac{\left(s[j][i] - \mu_{s[j]}\right)^m}{n_{s[j]}^{m-1}(L-k+1)^{m-1}}, m = 1, 2, \cdots, n_{s[j]}$$

Thus, the k-mer natural vector for virus genome $s$ is defined by.

($n_{s[j]}$, $\mu_{s[j]}$, $D_m^{s[j]}$), $j = 1, 2, \cdots, 4^k$

By the definition above, the k-mer natural vector concatenates the numbers of occurrence and mean distance for k-mer with its central moments, it therefore contains the information of k-mers and avoids the deficiencies of previous k-mer models. Moreover, the relationship between a virus genome and its k-mer natural vector is one-to-one for each given $k$, which has been mathematically proved in the Test S1. In addition, it has been verified that a k-mer natural vector with order two central moment is enough to represent a virus genome, so ($n_{s[j]}$, $\mu_{s[j]}$, $D_2^{s[j]}$) is effectively depict a virus genome, and still satisfies one-to-one mapping.

### 4.3. Selection of the k-value and distance metric for k-mer natural vector

Parameter k has a great influence on obtaining result and computational complexity for k-mer model methods. Following our former work, we choose optimal k value for k-mer natural vector is within a range of $[\text{ceil}(\log_4 \min(L)), \text{ceil}(\log_4 \max(L)) + 1]$, where $L$ is the set of lengths of genome sequences considered (Wen et al., 2014). In this study, values of k chosen for whole-genome sequence, non-structural protein gene Orf1ab, and genes encoding the structural proteins of S, E, M, and N are 8, 8, 7, 6, 6, and 7, respectively.

Once each virus genome is uniquely represented by a k-mer natural vector, the Cosine distance metric is used to calculate pairwise distance of virus genomes, which eliminates the effects of high dimensionality and thus widely used in k-mer models (Zhang et al., 2019). Then, Neighbor-Joining (NJ) tree is constructed to show the phylogenies of virus genomes, which can be drawn by MEGA (version 7.0) with default parameters (Kumar, 2016).

### 4.4. Mean distance and Center distance

Mean distance and Center distance are proposed to quantify the distance between two point sets. Mean distance is defined as the average of distances between two point sets. Although Center distance is similar to Mean distance, they are different, in that, Center distance is proposed based on convex hulls (Lin and Kwan, 2016; Dong et al., 2020).

Let $A = \{V_1, V_2, \cdots, V_n\}$ represents a point set of $V_s$ of $n$ points. Then the convex hull of A is defined as

$$C(A) = \left\{ p | p = \sum_{i=1}^{n} \alpha_i V_i, \sum_{i=1}^{n} \alpha_i = 1, \alpha_i \geq 0, 0 \leq i \leq n \right\}$$

A convex hull is the smallest convex set containing a given point set. For two point sets, each point set can be described by its convex hull, and the barycenter of each hull is considered as the representative of the hull. Thus, the distance between two barycenter represents the average distance of two point sets as well.

## Declarations of interest

None.

## Acknowledgments

We sincerely thank the authors of the coronavirus related data from GenBank and GISAID. This work was supported by Natural Scientific Research Funding of Heilongjiang (LH2019A031), and Scientific Research Funding of Suihua University (2017-XKYYWF-017).

## Appendix A. Supplementary data

Supplementary material to this article can be found online. All datasets and matlab code used in this paper are available at https://github.com/wenjia198021/Hosts-and-transmission-traits-for-SARS-CoV-2. Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2021.104933.

## References

Bauer, D.C., et al., 2020. Supporting pandemic response using genomics and bioinformatics: a case study on the emergent SARS-CoV-2 outbreak. Transbound. Emerg. Dis. 67, 1453–1462.
Boni, M.F., et al., 2020. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat. Microbiol. 5, 1408–1417.
Chan, J.F., et al., 2020. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg Microbes Infect. 9, 221–236.
Chinazzi, M., et al., 2020. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science 368, 395–400.
Cui, J., et al., 2019. Origin and evolution of pathogenic coronaviruses. Nat. Rev. Microbiol. 17, 181–192.
Deslandes, A., et al., 2020. SARS-CoV-2 was already spreading in France in late December 2019. Int. J. Antimicrob. Agents 55, 106006.
Dong, R., et al., 2020. Analysis of the hosts and transmission paths of SARS-CoV-2 in the COVID-19 outbreak. Genes (Basel) 11, 637.
Drosten, C., et al., 2003. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N. Engl. J. Med. 348, 1967–1976.
Forster, P., et al., 2020. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc. Natl. Acad. Sci. U. S. A. 117, 9241–9243.
Gao, G.F., 2018. From "a" IV to "Z" IKV: attacks from emerging and re-emerging pathogens. Cell 172, 1157–1159.
Guan, Y., et al., 2003. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science 302, 276–278.
Jaimes, J.A., et al., 2020. Phylogenetic analysis and structural modeling of SARS-CoV-2 spike protein reveals an evolutionary distinct and proteolytically sensitive activation loop. J. Mol. Biol. 432, 3309–3325.
Kumar, S., et al., 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 33, 1870–1874.
Lam, T.T., et al., 2020. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature 583, 282–285.
Lau, S.K.P., et al., 2020. Possible bat origin of severe acute respiratory syndrome coronavirus 2. Emerg. Infect. Dis. 26, 1542–1547.
Li, C., et al., 2020. Genetic evolution analysis of 2019 novel coronavirus and coronavirus from other species. Infect. Genet. Evol. 82, 104285.
Lin, Z., Kwan, R., 2016. Local convex hulls for a special class of integer multicommodity flow problems. Comput. Optim. Appl. 64, 881–919.
Lu, R., et al., 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395, 565–574.
Oreshkova, N., et al., 2020. SARS-CoV-2 infection in farmed minks, the Netherlands, April and May 2020. Euro Surveill. 25, 2001005.
Shi, J., et al., 2020. Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS–coronavirus 2. Science. 368, 1016–1020.

Sit, T.H.C., et al., 2020. Infection of dogs with SARS-CoV-2. Nature 586, 776–778.
Song, H.D., et al., 2005. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. Proc. Natl. Acad. Sci. U. S. A. 102, 2430–2435.
Su, S., et al., 2016. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol. 24, 490–502.
Wang, Q., et al., 2016. MERS-CoV spike protein: targets for vaccines and therapeutics. Antivir. Res. 133, 165–177.
Wen, J., et al., 2014. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. Gene 546, 25–34.
Wrobel, A.G., et al., 2021. Structure and binding properties of Pangolin-CoV spike glycoprotein inform the evolution of SARS-CoV-2. Nat. Commun. 12, 837.
Wu, A., et al., 2020a. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. Cell Host Microbe 27, 325–328.
Wu, F., et al., 2020b. A new coronavirus associated with human respiratory disease in China. Nature 579, 265–269.
Xia, X., 2020. Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense. Mol. Biol. Evol. 37, 2699–2705.
Xiao, K.P., et al., 2020. Isolation and characterization of 2019-nCoV-like coronavirus from Malayan pangolins. Nature 583, 286–289.
Young, B.E., et al., 2020. Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. Lancet 396, 603–611.
Zhang, Y.Z., Holmes, E.C., 2020. A genomic perspective on the origin and emergence of SARS-CoV-2. Cell 181, 223–227.
Zhang, Y., et al., 2019. Phylogenetic analysis of protein sequences based on a novel k-mer natural vector method. Genomics 111, 1298–1305.
Zhang, Q., et al., 2020. A serological survey of SARS-CoV-2 in cat in Wuhan. Emerg Microbes Infect. 9, 2013–2019.
Zhou, H. Chen, 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. Curr. Biol. 30, 3896.
Zhou, P., et al., 2020b. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273.
Zhu, N., et al., 2020. A novel coronavirus from patients with pneumonia in China, 2019. N. Engl. J. Med. 382, 727–733.
Zielezinski, et al., 2017. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol. 18, 186.