



In Silico Infection Analysis (iSFA) Identified Coronavirus Infection and Potential Transmission Risk in Mammals

Yanyan Zou^{1,2†}, Xiaojian Cao^{1,3†}, Bing Yang^{1,3†}, Lulu Deng^{1,3}, Yangyang Xu^{1,3}, Shuang Dong⁴, Wentao Li³, Chengchao Wu^{1,3*} and Gang Cao^{1,3,5*}

¹State Key Laboratory of Agricultural Microbiology, Huazhong Agricultural University, Wuhan, China, ²College of Informatics, Huazhong Agricultural University, Wuhan, China, ³College of Veterinary Medicine, Huazhong Agricultural University, Wuhan, China, ⁴Department of Medical Oncology, Hubei Cancer Hospital, Huazhong University of Science and Technology, Wuhan, China, ⁵Bio-Medical Center, Huazhong Agricultural University, Wuhan, China

OPEN ACCESS

Edited by:

Toyotaka Ishibashi,
Hong Kong University of Science and
Technology, Hong Kong SAR, China

Reviewed by:

Taobo Hu,
Peking University People's Hospital,
China
Masashi Yukawa,
The Chinese University of Hong Kong,
China

*Correspondence:

Chengchao Wu
ccwu@mail.hzau.edu.cn
Gang Cao
gcao@mail.hzau.edu.cn

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Genome Organization and Dynamics,
a section of the journal
Frontiers in Molecular Biosciences

Received: 09 December 2021

Accepted: 04 January 2022

Published: 08 February 2022

Citation:

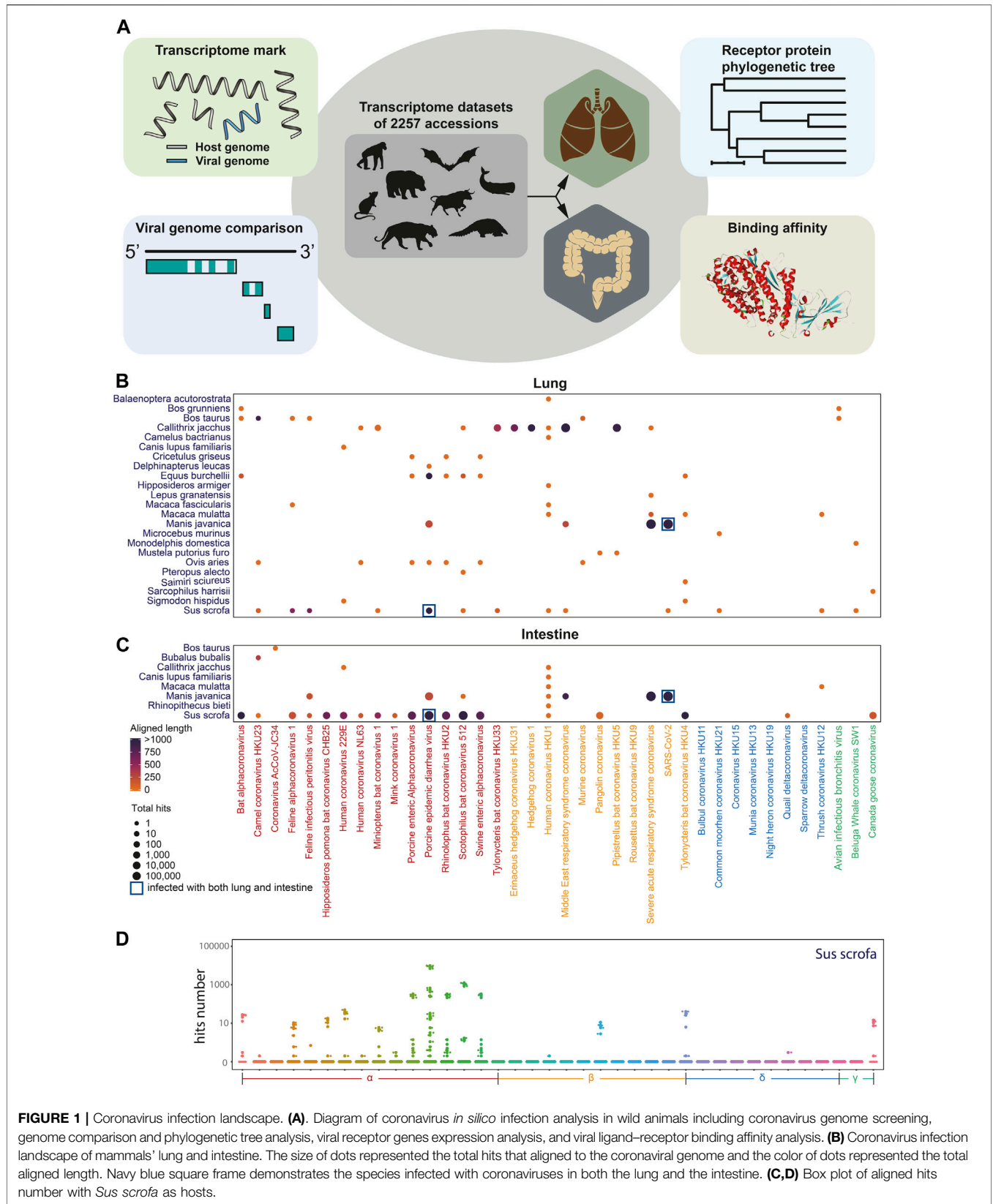
Zou Y, Cao X, Yang B, Deng L, Xu Y,
Dong S, Li W, Wu C and Cao G (2022)
In Silico Infection Analysis (iSFA)
Identified Coronavirus Infection and
Potential Transmission Risk
in Mammals.
Front. Mol. Biosci. 9:831876.
doi: 10.3389/fmolb.2022.831876

Coronaviruses are a great source of threat to public health which could infect various species and cause diverse diseases. However, the epidemic's spreading among different species remains elusive. This study proposed an *in silico* infection analysis (iSFA) system that includes pathogen genome or transcript mining in transcriptome data of the potential host and performed a comprehensive analysis about the infection of 38 coronaviruses in wild animals, based on 2,257 transcriptome datasets from 89 mammals' lung and intestine, and revealed multiple potential coronavirus infections including porcine epidemic diarrhea virus (PEDV) infection in *Equus burchellii*. Then, through our transmission network analysis, potential intermediate hosts of five coronaviruses were identified. Notably, iSFA results suggested that the expression of coronavirus receptor genes tended to be downregulated after infection by another virus. Finally, binding affinity and interactive interface analysis of S1 protein and ACE2 from different species demonstrated the potential inter-species transmission barrier and cross-species transmission of SARS-CoV-2. Meanwhile, the iSFA system developed in this study could be further applied to conduct the source tracing and host prediction of other pathogen-induced diseases, thus contributing to the epidemic prevention and control.

Keywords: coronaviruses, in silico infection analysis, in silico docking, data mining, COVID-19

INTRODUCTION

Coronaviruses are a great source of threat to humans which causes diverse diseases ranging from asymptomatic to severe respiratory deflection (Kahn and McIntosh, 2005). Generally, coronaviruses could be divided into four subtypes, including alpha-coronavirus, beta-coronavirus, delta-coronavirus, and gamma-coronavirus. Some of the known beta-coronavirus such as SARS-CoV, MERS-CoV, and recently emerged SARS-CoV-2 can lead to fatal repository health crisis in humans. Zoonotic transmission routes of coronaviruses usually include natural reservoir and several intermediate hosts, and end with infection of various human cells, causing distinct pathogenic effects in different hosts (Graham and Baric, 2010). The traditional epidemiological research methods which include statistics on the prevalence of large group, retrospective analysis have played an important role in clarifying the epidemic pattern, formulating prevention and control



countermeasures of emerging infectious diseases (Cho, 2020; Qiu et al., 2020). However, its traceability analysis is often restricted by tedious and limited sample collection and testing work which may expend much manpower and financial resources (Chen et al., 2020b; Hao et al., 2020).

The invasion of zoonotic coronaviruses to human cells is mediated by spike protein binding to various cell surface receptors, including ACE2, ANPEP, and DPP4. During the infection of SARS-CoV-2, the spike protein of SARS-CoV-2 binds to the ACE2 receptor of human cells, leading to virus invasion via endocytosis. Afterward, the virus-ACE2 complex is delivered into acidified phagolysosome. Then viral RNA is released, which consequently initiates viral replication and the assembly of the progeny viruses (Simmons et al., 2005). The risk level of coronavirus infection is associated with the binding affinity between the virus spike protein and host cell receptors. ANPEP is usually regarded as a receptor of a mild human coronavirus HCoV-229E, while it has been also suggested as one of the receptors for SARS-CoV (Yu et al., 2003). DPP4 has been reported to possess a highly conserved domain among Camelidae, primates, and Leporidae which is highly affinitive to MERS-CoV spike protein, making it one of the major mediators to MERS (Widagdo et al., 2019). However, the interaction of the diverse receptors and the ligands in different coronaviruses and the subsequent cross-species transmission remains elusive.

One of the most effective measures to control the coronavirus epidemic is cutting off the intermediate transmission route. Thus, it is urged to identify the intermediate host of the corresponding virus and the coronavirus infection among different species. Unfortunately, this effort confronts huge challenges due to the requirement of comprehensive screening. The intermediate host for SARS-CoV-2 remains undetermined more than a year after the burst of the epidemic. This study aims to explore the transmission route of coronaviruses in mammals and SARS-CoV-2 potential intermediate hosts. For this purpose, we nominated this integrated computational approach called *in silico* infection analysis to screen coronavirus infection in 89 non-human mammals and systematically analyzed the expression level of coronavirus receptors and binding affinity with the spike protein of coronaviruses.

RESULTS

In Silico Infection Analysis Identified Coronavirus Infection in Mammals

To explore the coronavirus infection in mammals, we performed *in silico* coronaviral genome screening in mammalian transcriptome datasets, except for widely studied *Homo sapiens*, *Mus musculus*, and *Rattus norvegicus* including lungs and intestines (Figure 1A). In this *in silico* infection analysis, the sequence of the genomic fragment of 38 coronaviruses was screened from 388 intestine datasets and 1,869 lung datasets of 89 mammal species from the NCBI Sequence Read Archive (SRA). Coronavirus genome fragments could not be detected in most of the selected transcriptomic datasets (Supplementary

Figures S1A,B). While some fragments of the coronavirus genome can be observed in certain species (Figures 1B,C) including porcine epidemic diarrhea virus (PEDV) identification in both the lung and the intestine of *Sus scrofa* (Figure 1D) and the lung of *Equus burchellii*. Importantly, numerous fragments of SARS-CoV-2-like virus existed in both the lung and intestine tissue of *Manis javanica*, which has been proposed as an intermediate host during the spread of SARS-CoV-2 (Lam et al., 2020). The virus fragments identified in *Equus burchellii* and *Manis javanica* were further analyzed for the exploration of the specific infection strain.

Potential Coronavirus Infection and Transmission Risk Analysis Among Different Mammals

All the nucleotide sequences of spike (S/GP2) protein were collected from 38 coronaviruses to construct the phylogenetic tree (Figure 2A). As shown in Figures 2B,C, the coronaviruses genome fragments were collected from lung and intestine transcriptome datasets of *Manis javanica* and *Equus burchellii*. The coronaviruses collected from *Manis javanica* and *Equus burchellii* showed the highest genome coverage rate, and most unique viral fragments can be mapped to the SARS-CoV-2 and PEDV genome, respectively. The result of identifying SARS-CoV-2 in *Manis javanica* is consistent with studies (Lam et al., 2020) reported earlier. As for the control group, we also collected the viral fragments isolated from the infection experiments of known strains, which exhibited a high degree of compliance (Supplementary Figures S2A,B). The S1 nucleotide and amino acid sequence of SARS-CoV-2 isolated from *Manis javanica*, *Chiroptera*, *Martes*, *Felinae*, *Panthera tigris*, *Canis lupus familiaris*, and *Homo sapiens* were further collected and subjected to nucleotide sequence phylogenetic tree and protein sequence alignment analysis, respectively (Figures 2D,E). These data demonstrated that GD/PL1 pangolin-CoV (Lam et al., 2020) and RaTG13 shared the highest nucleotide sequence similarity among species. Finally, we selected five coronaviruses including SARS-CoV-2, SARS-CoV, MERS-CoV, PEDV, and HCoV-229E for the infection network construction between 16 different mammals. Based on our transcriptome data screening and previously reported coronavirus host studies (Stadler et al., 2003; van der Hoek et al., 2004; Manzanillo et al., 2013; Mackay and Arden, 2015; Rahimi and Tehranchinia, 2020). Figure 2F demonstrated that coronaviruses have interlaced the transmission network and are widely spread across species. *Chiroptera* seems to be a central reservoir host of coronaviruses as most coronaviruses can use it as a host (Maxmen, 2017).

Expression Patterns of Coronavirus Receptors Among Species

To further evaluate the coronavirus transmission risk across species, the corresponding protein sequences of 24 putative respiratory virus receptors from 65 species (Supplementary Table S2), including *Primates*, *Rodentia*, *Carnivora*, *Artiodactyla*, *Perissodactyla*, *Cetacea*, *Chiroptera*, and *Pholidota* were collected and aligned (Figure 3A). Coronavirus receptors of non-human primates showed high sequence similarities with

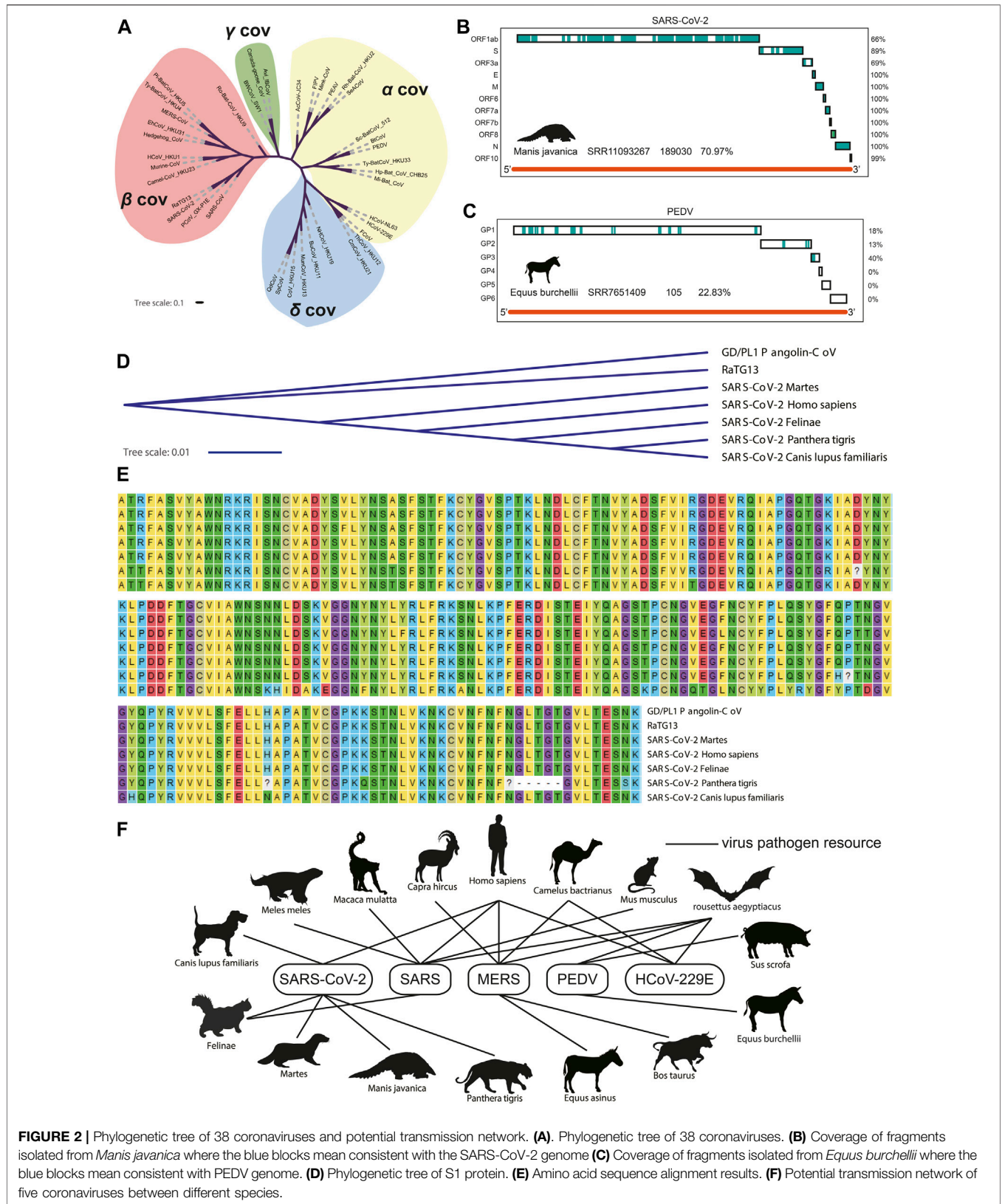
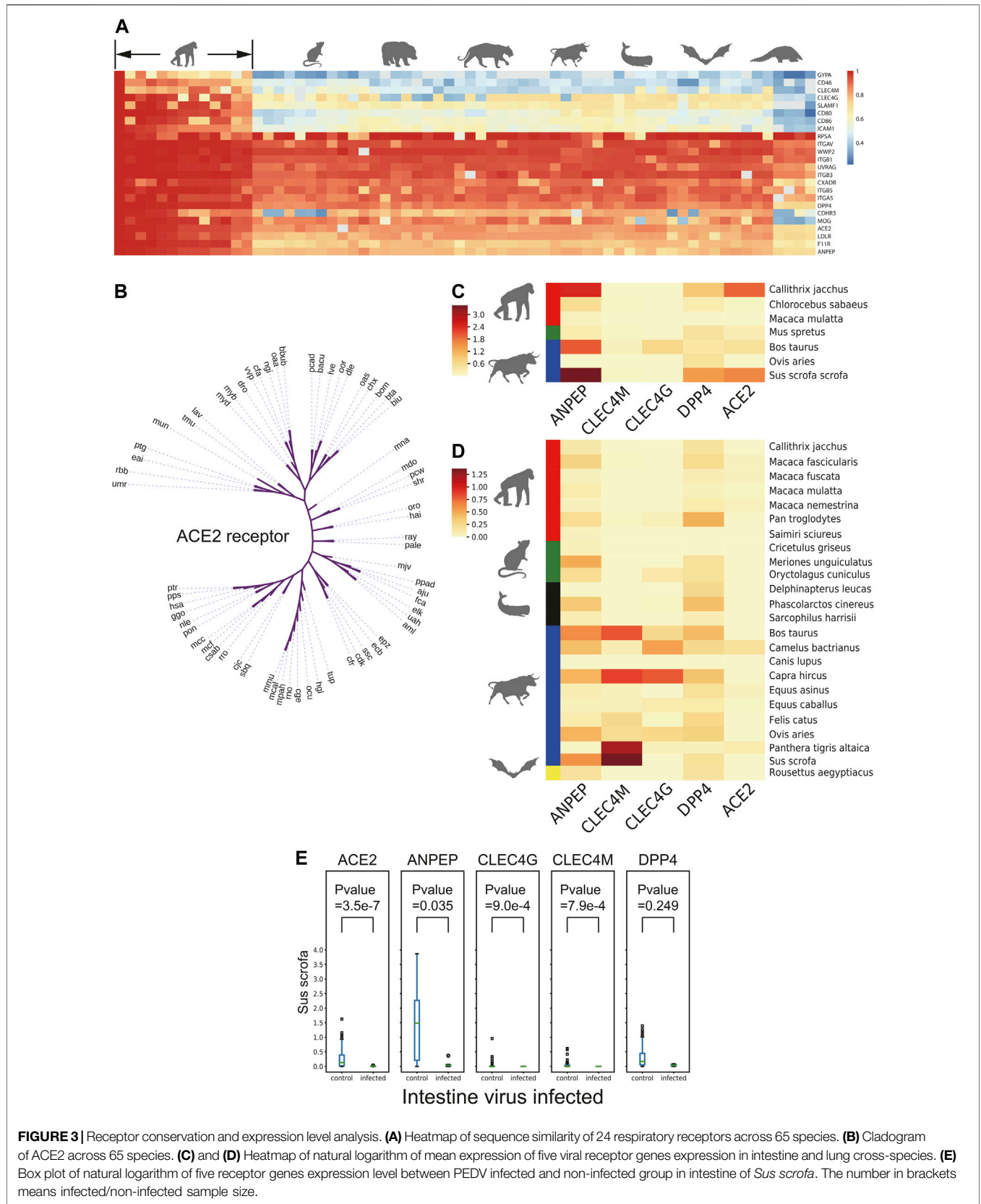


FIGURE 2 | Phylogenetic tree of 38 coronaviruses and potential transmission network. **(A)** Phylogenetic tree of 38 coronaviruses. **(B)** Coverage of fragments isolated from *Manis javanica* where the blue blocks mean consistent with the SARS-CoV-2 genome **(C)** Coverage of fragments isolated from *Equus burchellii* where the blue blocks mean consistent with PEDV genome. **(D)** Phylogenetic tree of S1 protein. **(E)** Amino acid sequence alignment results. **(F)** Potential transmission network of five coronaviruses between different species.



human receptors, which implied the similar binding affinity to coronavirus ligand. For example, gorilla and monkey, the nearest neighbors in the cladogram (Figure 3B), may be at similar potential risks of SARS-CoV-2 infection as they shared similar ACE2 receptor protein sequences, while some receptors of the rest species demonstrated distinct levels of protein sequence divergence, implying that the infection risk might differ among these species.

As the expression level of virus receptors may reflect the infection probability to some extent of certain viruses, the expression level of virus receptors was analyzed across the species from 1,869 lung transcriptomic datasets and 388 intestine transcriptomic datasets of different mammals (except for commonly studied *Homo sapiens*, *Mus musculus*, and *Rattus norvegicus*). After quality control steps, we kept the sequences of 26 mammalian species for further expression analysis. A total of 24 receptors including ACE2, a receptor of HCoV-NL63, SARS-CoV, and SARS-CoV-2; ANPEP, a receptor of several alpha-coronaviruses and delta-coronaviruses; CLEC4G and CLEC4M, putative receptors of SARS-CoV; and DPP4, a receptor of MERS-CoV and HKU4, were subjected to this expression analysis (MacLachlan and Dubovi, 2011) (Supplementary Figures S3A,B). As shown in Figures 3C,D, the receptor ANPEP presented relatively higher expression in the intestine, while the expression level of CLEC4M remained low in all tested species. Importantly, those five receptors were observed to be less expressed in *Macaca mulatta*. ANPEP has a similar expression pattern in the lung as in the intestine, while DPP4 generally showed higher expression than ACE2 across the species. As *Artiodactyla* mammals, the expression level of CLEC4M and CLEC4G were higher than other species.

Next, to study whether a virus infection would alter receptor gene expression of other viruses, the transcriptome data from *Sus scrofa* with or without PEDV infection in the intestine and lung transcriptome data with or without porcine reproductive and respiratory syndrome virus (PRRSV) infection were taken into further study. The expression level of different virus receptor genes including ACE2, ANPEP, CLEC4G, CLEC4M, and DPP4 were analyzed. Interestingly, we observed that after infection, CLEC4G, the putative receptor gene of SARS, was significantly downregulated in the lung of the PRRSV-infected pigs compared to the non-infected pigs (Supplementary Figure S3C). Notably, more receptor-encoding genes including ACE2, ANPEP, CLEC4G, and CLEC4M were significantly downregulated upon PEDV infection (Figure 3E) in the intestine. These data suggested that infection of a certain virus may suppress the expression of other virus receptors to avoid invasion of other viruses, which could maintain long-term persistence.

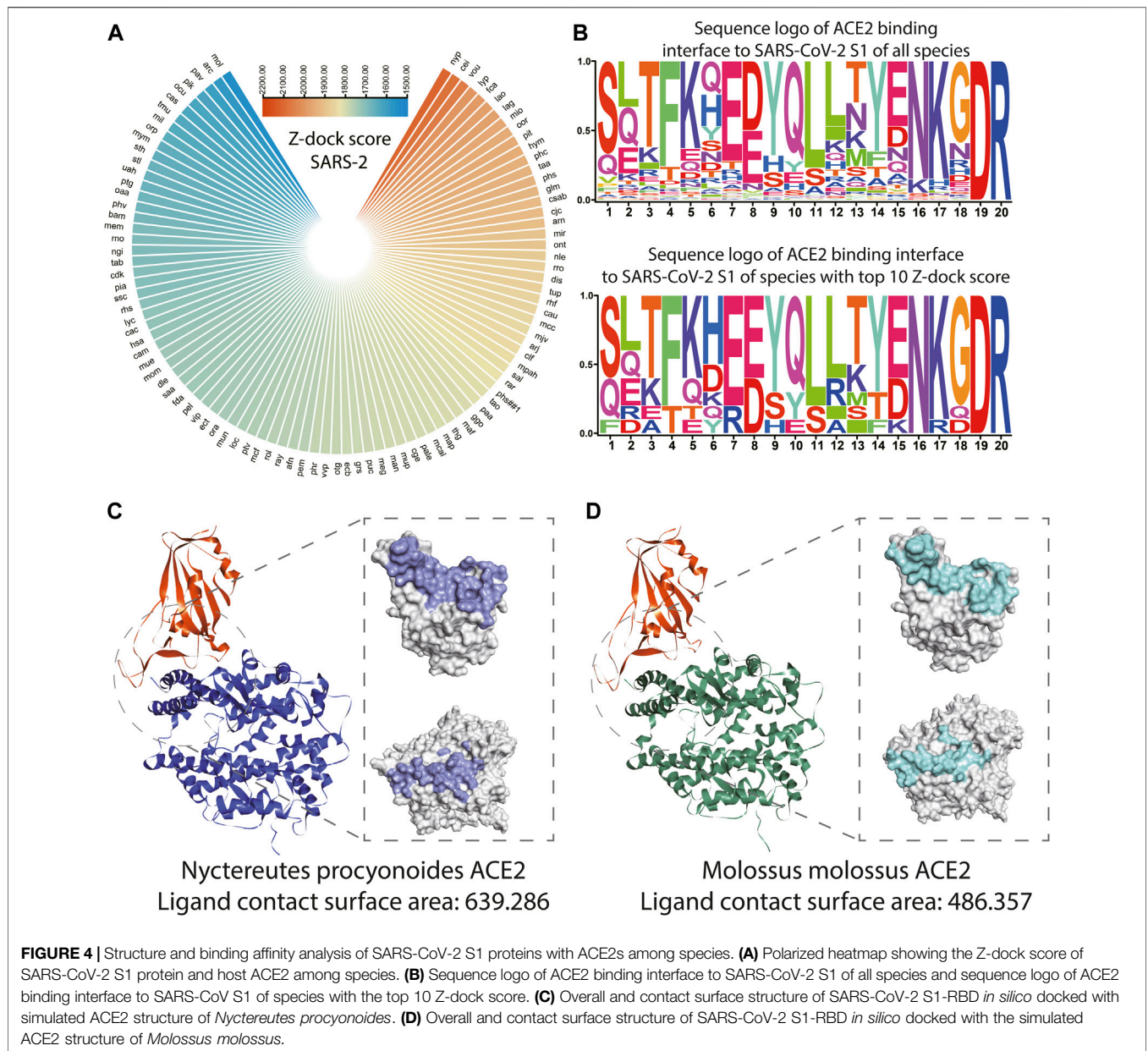
Binding Affinity Analysis of SARS-CoV-2 S1-RBD With ACE2 From Different Species

The protein sequences of ACE2 from 99 species were downloaded from the NCBI, then the homolog sequence was extracted for the structure simulation with the SARS-CoV S1-human ACE2

complex (PDB: 6acg) and the SARS-CoV-2 S1-human ACE2 complex (PDB: 7DF4) as the template with MODELLER. The strategy of *in silico* docking is shown in Supplementary Figure S4A.

As shown in Figure 4A and Supplementary Table S5, the top 10 species whose ACE2s are with the highest Z-dock score when docking with SARS-CoV-2 S1 were *Nyctereutes procyonoides*, *Cebus imitator*, *Vombatus ursinus*, *Lynx pardinus*, *Felis catus*, *Lagenorhynchus obliquidens*, *Lama glama*, *Microtus ochrogaster*, *Orcinus orca*, and *Ptilocolobus tephrosceles*. For docking with SARS-CoV S1, species with the top 10 Z-dock score were *Tachyglossus aculeatus*, *Talpa occidentalis*, *Macaca fascicularis*, *Artibeus jamaicensis*, *Felis catus*, *Ptilocolobus tephrosceles*, *Rhinopithecus roxellana*, *Meleagris gallopavo*, *Rattus rattus*, and *Rhinolophus ferrumequinum* (Supplementary Figure S4B).

The loci of the human binding site to S1 of SARS-CoV and SARS-CoV-2 (LBS) was extracted from the previous study (Shang et al., 2020), the un-gapped and unopened blast during homolog sequence extraction enabled direct extraction of corresponding loci from ACE2 of other species. The sequence logo was plotted with ACE2 LBS of SARS-CoV S1 and SARS-CoV-2 S1 of all species and species with the top 10 Z-dock score. It could be observed that ACE2 LBS to SARS-CoV-2 S1 of all the mammals involved in the analysis is conserved on the 19th and 20th amino acids, while the sixth, seventh, 13th, 15th, 16th, and 18th amino acids seem to be more conserved among the top 10 species. It should be noted that on the eighth loci, the most common amino acid is E, while D is the most common amino acid among all species (Figure 4B). In terms of ACE2 LBS to SARS-CoV S1, species with top 10 Z-dock scores showed higher conservatism on the 13th, 14th, and 15th amino acids. While it should be noted that all the mammals involved in the analysis are conserved on the 16th and 17th amino acids (Supplementary Figure S4C). The crystal structures and detailed interaction of ACE2 from *Nyctereutes procyonoides* or *Ptilocolobus tephrosceles* and SARS-CoV-2 S1-RBD were analyzed by the Dock and Analyze Protein Complex Module of Discovery Studio. Consistent with the results of binding affinity analysis, the complex simulation of *Nyctereutes procyonoides* ACE2-RBD displayed a higher Z-dock score than *Ptilocolobus tephrosceles* ACE2-RBD, which showed the highest and lowest Z-dock scores among the modeled structures. In addition, the contact surface area of SARS-CoV-2 S1-RBD in contact with ACE2 was calculated in our assays. We discovered that the contact surface area of *Nyctereutes procyonoides* ACE2 among the complex is 639.286, whereas that of SARS-CoV2 S1-RBD is 627.057 (Figure 4C). Notably, the area of *Ptilocolobus tephrosceles* ACE2 in contact with S1-RBD is 486.357, and the interactive surface acre in RBD is 478.622 (Figure 4D). The larger interactive area between the complex simulation of *Nyctereutes procyonoides* ACE2-RBD explains why *Nyctereutes procyonoides* ACE2 displayed the highest Z-dock score when docked with SARS-CoV2 S1-RBD, rather than other species. Comparison of PDB: 7DF4 (SARS-CoV-2 S1 and human ACE2 complex) and simulated human ACE2 *in silico* docked with SARS-CoV-2 S1 was also conducted, which showed significant overlap on the structure (Figure S4d),



indicating the reliability of our strategy of structuring modeling and *in silico* docking.

DISCUSSION

Detecting the general coronaviruses infection in the host can reduce the overall risk of infection and construct a coronavirus transmission network, which helps better prevention of coronavirus infection. In this study, we constructed a method, named *in silico* infection analysis (iSFA), which involves the transcriptome data mining from existing data available on the NCBI and could perform traceability and transmission risk evaluation of the coronaviruses and detect the coronavirus

infection in various mammals including PEDV in *Equus burchellii*. Instead of detecting the DNA fragments imprinted in the genome, Souilmi et al. (2021) identified infection traces of coronaviruses in human East Asian populations with adaptive events in coronavirus-related genes, which needs datasets of population size. The iSFA strategy was to focus on more hosts such as wild mammals and detect their infected situation with less cost. By integrating the previous related studies with these iSFA results, a transmission network for several coronaviruses was constructed, which could provide an important resource for cross-species transmission risk evaluation. Mammals play a critical role in the diffusion of coronaviruses and contribute to the evolution of some kind of viruses (Costagliola et al., 2020), which made it significant to amplify the coronavirus transmission

network. With the exponentially increasing number of deposited data from diverse tissues of vast species of animals, iSFA could be applied to detect virus genome fragments in multiple species to reveal the potential transmission route of a certain virus.

Rather than virus genome fragment detection, iSFA could be further utilized for expression analysis of the corresponding receptor that mediates the invasion and pathogenesis of the certain virus. In this study, receptor sequence similarity and expression patterns among 65 receptor genes and 26 non-human mammals were analyzed, respectively. Interestingly, PEDV infection was observed in *Sus scrofa* which can inhibit the expression of receptors of CoV_HKU15, SARS-CoV, and SARS-CoV-2 such as ANPEP, ACE2, CLEC4G, and CLEC4M. This result supported the hypothesis that the invasion of one virus may inhibit the invasion of another virus to ensure “the first invader” in an uncontested environment persist and proliferate (Hall et al., 1994), which may inhibit the expression of the receptor genes. With the rapidly increasing available amount of transcriptome data, especially the single-cell sequencing data from public databases, this analysis could shed more insights into the gene expression profile studies, such as the co-infection and co-expression of the receptor or co-receptor genes with a specific virus.

SARS-CoV and SARS-CoV-2 infection are initialized by the binding of S1 protein to the ACE2 receptor of the host. Thus, our *in silico* docking may benefit the construction of a landscape of the S1 binding to ACE2 of various species, which may reveal the infection risk of various mammals. It could be observed that *Nyctereutes procyonoides* (raccoon dog) ranks among the highest Z-dock score. And the detailed analysis for interactive surfaces among the modeled complexes also displayed a larger interface area on *Nyctereutes procyonoides* ACE2 and SARS-CoV2 S1-RBD, which agreed with the highest Z-dock score of complex simulation of *Nyctereutes procyonoides* ACE2-RBD. This consistency of contact area and the Z-dock score not only explained the variability of binding affinity among different species but also verified the reliability of our established *in silico* docking method. Previous studies also suggest the raccoon dog is the intermediate host of SARS-CoV and could also be infected by SARS-CoV-2 (Freuling et al., 2020). We also observed ACE2 of *Felis catus* (cat) showed a high Z-dock score when *in silico* docked with S1 of SARS-CoV and SARS-CoV-2, while the cat has also been reported as an intermediate host of SARS-CoV-2 (Stout et al., 2020). Reports of other species with high Z-dock scores infected with SARS-CoV-2 were failed to be found, but these species were geographically distributed distantly from human society, which might cause low chance of infection or their infection is rarely reported. Higher conservatism of LBS amino sequence could be observed in species with top 10 Z-dock scores, these loci might play an important role during ACE2 and S1 binding.

As for the conclusion, this study performed a comprehensive *in silico* analysis about the potential coronavirus mammalian host from different aspects and proposed an *in silico* infection analysis system that could be

extensively applied to explore the viral infection in different species. The infection of PEDV in *Equus burchellii* was revealed by using this approach. Moreover, this *in silico* infection analysis suggested that the expression of coronavirus receptor genes tended to be downregulated after infection from another virus. Finally, we provided and verified a modeling and docking strategy, which revealed conserved loci of the SARS-CoV and SARS-CoV-2 S1 binding site of ACE2, and the higher conservatism of the LBS sequence in the species with top 10 Z-dock scores. Our iSFA including transcriptome data mining and *in silico* modeling and docking could suggest potential intermediate hosts of SARS-CoV-2 and other coronaviruses and could be extensively applied to investigate other virus infections in various species in the future. Though it still needs further verification through animal experiments and clinical studies, iSFA of transcriptome data mining and *in silico* modeling and docking could provide a reference to the methodology of the intermediate host of pathogen anticipation and epidemic control.

MATERIALS AND METHODS

Transcriptomic Data

The mammalian classification information was obtained from The Catalogue of Life website. Transcriptome data of different mammalian from the lung and the intestine were obtained from the NCBI Sequence Read Archive (SRA), and those species including *Homo sapiens*, *Mus musculus*, and *Rattus norvegicus*, which have been frequently analyzed, were removed from further analysis. Also, those datasets from virus-infected experiments were removed for traceability analysis as well. After screening, 2,257 datasets were finally applied to further analyses, in which lung-associated 1,869 datasets belonged to 82 species (**Supplementary Table S3**) and intestine-associated 388 datasets belonged to 22 species (**Supplementary Table S4**).

Complete Genome of Coronaviruses

Complete genomes of 38 coronaviruses were downloaded from the Virus Pathogen Resource. These coronaviruses derived from four genera known as alpha-coronavirus, beta-coronavirus, delta-coronavirus, and gamma-coronavirus, which belong to Coronaviridae family and Orthocoronavirinae subfamily.

Receptor Protein Sequences

A total of 24 respiratory virus receptors were collected from the literature, and the corresponding gene sequences and protein sequences were downloaded from the KEGG resource with R.

Sequence Alignment

We took 38 coronaviruses except RaTG13 including AviCoV, Ro-Bat-CoV_HKU9, PEDV, HCoV-229E, HCoV-NL63, FIPV, Sc-BatCoV_512, SARS, MERS, SARS-CoV-2, BtCoV, BwCoV_

SW1, BuCoV_HKU11, Camel-CoV_HKU23, Canada-goose_CoV, CmCoV_HKU21, AcCoV-JC34, CoV_HKU15 (PDCoV), EhCoV_HKU31, FCoV, Hedgehog_CoV, Hp-Bat_CoV_CHB25, HCoV_HKU1, Mi-Bat_CoV, Mink-CoV, MunCoV_HKU13, Murine-CoV, NHCov_HKU19, PCoV_GX-P1E, RaTG13, Pi-BatCoV_HKU5, PEAV, QdCoV, Rh-Bat-CoV_HKU2, SpCoV, SeACoV, ThCoV_HKU12, Ty-BatCoV_HKU33, and Ty-BatCoV_HKU4 (Supplementary Table S1) into consideration.¹⁷ Because these data contains dozens of species, blastn (2.9.0+) (Shen et al., 2016) was chosen as aligner with complete virus genomes as reference. Those mapped reads whose E-value was less than 1e-5 and identity greater than 90% were counted for further downstream analysis. The virus infection traceability analysis also took genome coverage and unique viral genome fragments into consideration.

Multiple sequence alignments of spike glycoprotein nucleic acid sequences from 39 coronaviruses were displayed by MUSCLE (Edgar, 2004), while multiple sequence alignments of SARS-CoV-2 S1 protein fraction were carried out by ClustalW (Thompson et al., 2002) using sequences separated from animals including *Martes* (MT457393-MT457401), *Felinae* (MT457390-MT457392), *Panthera tigris* (MT365033), and *Canis lupus familiaris* (MT270814) released by GenBank, partial genomic sequence of the Guangdong pangolin coronavirus (GD/P1L) released by Lam et al. (2020), RaTG13 (MN996532.1) with genome highly similar to SARS-CoV-2 by GenBank and from *Homo sapiens* (MT676411.1) as reference.

Expression Measurement of Coronavirus Receptors

The differences among these protein sequences were measured by alignment using the Needleman–Wunsch algorithm. Blastn was selected to aligned transcriptomic data into receptor reference sequences for expression evaluation. We took steps of quality control, which first considered sequencing depth and genome size, then retained the sequences of 26 mammalian species for further analysis with sequencing depth greater than 2 and genome C-value around 3 (Supplementary Figures S3A,B). Normalization of receptors expression referred to the standardized method of TPM to balance the impact of gene length, and sequencing depth and was performed as follows:

$$\text{normalized expression} = \frac{\text{mapped reads num}}{\text{gene length}} \bigg/ \frac{\sum \text{reads length}}{\text{genome length}}$$

Genome length was estimated by C-value downloaded from the Animal Genome Size Database and $\sum \text{reads length}$ was statistics by seqkit (v0.12.1) (Shen et al., 2016).

Receptor Expression Level Among Infected and Non-Infected Samples

Considering the sample size, we chose datasets from *Sus scrofa* to compare the expression of receptor genes between the infected

and non-infected samples. We extracted non-infected and PEDV-infected expression data from intestines, and non-infected and PRRSV-infected data from the lung. Thereafter, the *t*-test was employed to examine whether virus infection will influence the expression of receptors significantly.

Phylogenetic Tree and Cladogram Construction

Phylogenetic tree was generated by FastMe in a distance-based manner (Lefort et al., 2015) and using iTol³¹ for advance annotation and management, while the cladogram was generated using the same procedure but with the branch lengths ignored during the visualization.

Protein Structure Simulation and Acquisition

The ACE2 protein sequences were acquired from the NCBI. The sequence and structure of ACE2 from the SARS-CoV S1-human ACE2 complex (PDB: 6acg) or the SARS-CoV-2 S1-human ACE2 complex (PDB: 7DF4) were extracted. The sequence was used as a template for the extraction of homolog sequence from the amino acid sequence of ACE2 of other species based on blast without gap and open. Extracted amino acid sequences were submitted to MODELLER (Webb and Sali, 2016) for the structure simulation with the extracted binding site structure mentioned earlier, while S1-RBD to ACE2 of SARS-CoV and SARS-CoV-2 was also extracted in the preparation of batch docking.

Z-Dock

Z-Dock software was downloaded to run the batch docking locally. Ligand (S1-RBD) and receptor (ACE2 from different species) structures were submitted to Z-dock, and the score was extracted from each docking output file and wrote into a text file with name of their corresponding organism. The Z-dock score represents the binding affinity of each ligand and receptor complexes.

Protein Docking and Complex Analysis

To understand the binding interaction of ACE2 receptors from different species with SARS-CoV-2 S1-RBD, the in-built protein docking module Z-dock in Discovery Studio 2018 software was employed to predict their binding poses. According to the protocol, first, the water molecules in receptors and ligands were removed before running the calculation to prevent the oxygen atom from being treated the same as an unknown atom in case if water is included. The proteins were prepared by Clean Protein tools to remove the alternative conformations, added the termini, and patched the missing side chain atoms. Thereafter, the 3D structures of ACE2 and SARS-CoV-2 S1-RBD were merged into a single Graphics View as a receptor and a ligand, respectively. The docking parameters were set as follows: angular step size was calculated with a defined value 15, RMSD cutoff was defined as 6.0, interface cutoff was defined as 9.0, maximum number of clusters was defined as 60, and other parameters were set as default. Finally, the docked protein complexes were optimized and re-scored using Refine Docked

Protein tools. The docked protein complexes were analyzed and visualized by the Analyze Protein Interface. Polarized heatmaps indicating the binding affinity of S1 and ACE2, and the sequence logo of S1 binding site were both generated by TBtools (Chen et al., 2020a).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding authors. The scripts for transcriptome data mining are available at <https://github.com/GangCaoLab/iSFA>. The scripts for protein modelling and in-silico docking are available at https://github.com/MSN-06s/Protein_modelling_and_docking.

AUTHOR CONTRIBUTIONS

GC and CW designed the study; YZ, CW, XC, BY, and LD conducted the analysis; YZ, CW, XC, BY, SD, WL, and GC wrote the manuscript; YZ, CW, XC, and BY contributed to the figure arrangement; and GC and CW were responsible for this work.

FUNDING

This work was supported by the National Key Research and Development Project of China (Grant 2017YFD0500300, to GC),

REFERENCES

- Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020a). TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol. Plant* 13 (8), 1194–1202. doi:10.1016/j.molp.2020.06.009
- Chen, C., Wang, H., Liang, Z., Peng, L., Zhao, F., Yang, L., et al. (2020b). Predicting Illness Severity and Short-Term Outcomes of COVID-19: A Retrospective Cohort Study in China. *The Innovation* 1 (1), 100007. doi:10.1016/j.xinn.2020.04.007
- Cho, S.-i. (2020). Academic Community's Efforts to Guide the Fight against Coronavirus Disease 2019 (COVID-19) Epidemic in Korea. *J. Prev. Med. Public Health* 53 (2), 65–66. doi:10.3961/jpmph.20.084
- Costagliola, A., Liguori, G., d'Angelo, D., Costa, C., Ciani, F., and Giordano, A. (2020). Do Animals Play a Role in the Transmission of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2)? A Commentary. *Animals* 11 (1), 16. doi:10.3390/ani11010016
- Edgar, R. C. (2004). MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Res.* 32 (5), 1792–1797. doi:10.1093/nar/gkh340
- Freuling, C. M., Breithaupt, A., Müller, T., Sehl, J., Balkema-Buschmann, A., Rissmann, M., et al. (2020). Susceptibility of Raccoon Dogs for Experimental SARS-CoV-2 Infection. *Emerg. Infect. Dis.* 26 (12), 2982–2985. doi:10.3201/eid2612.203733
- Graham, R. L., and Baric, R. S. (2010). Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission. *J. Virol.* 84 (7), 3134–3146. doi:10.1128/jvi.01394-09
- Hall, G. S., Pratt-Rippin, K., Meisler, D. M., Washington, J. A., Roussel, T. J., and Miller, D. (1994). Growth Curve for Propionibacterium Acnes. *Curr. Eye Res.* 13 (6), 465–466. doi:10.3109/02713689408999875
- Hao, X., Cheng, S., Wu, D., Wu, T., Lin, X., and Wang, C. (2020). Reconstruction of the Full Transmission Dynamics of COVID-19 in Wuhan. *Nature* 584 (7821), 420–424. doi:10.1038/s41586-020-2554-8
- Kahn, J. S., and McIntosh, K. (2005). History and Recent Advances in Coronavirus Discovery. *Pediatr. Infect. Dis. J.* 24 (11), S223–S227. doi:10.1097/01.inf.0000188166.17324.60
- Lam, T. T.-Y., Jia, N., Zhang, Y.-W., Shum, M. H.-H., Jiang, J.-F., Zhu, H.-C., et al. (2020). Identifying SARS-CoV-2-Related Coronaviruses in Malayan Pangolins. *Nature* 583 (7815), 282–285. doi:10.1038/s41586-020-2169-0
- Lefort, V., Desper, R., and Gascuel, O. (2015). FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program: Table 1. *Mol. Biol. Evol.* 32 (10), 2798–2800. doi:10.1093/molbev/msv150
- Mackay, I. M., and Arden, K. E. (2015). MERS Coronavirus: Diagnostics, Epidemiology and Transmission. *Virol. J.* 12, 222. doi:10.1186/s12985-015-0439-5
- MacLachlan, N. J., and Dubovi, E. J. (2011). "Chapter 24 - Coronaviridae," in *Fenner's Veterinary Virology*. Fourth Edition (San Diego: Academic Press), 393–413.
- Manzanillo, P. S., Ayres, J. S., Watson, R. O., Collins, A. C., Souza, G., Rae, C. S., et al. (2013). The Ubiquitin Ligase Parkin Mediates Resistance to Intracellular Pathogens. *Nature* 501 (7468), 512–516. doi:10.1038/nature12566
- Maxmen, A. (2017). Bats Are Global Reservoir for Deadly Coronaviruses. *Nature* 546 (7658), 340. doi:10.1038/nature.2017.22137
- Qiu, H., Wang, H. J., Chen, Q. L., and Yue, X. (2020). Safety protection of Forensic Examination during the Epidemic of COVID-19. *Fa Yi Xue Za Zhi* 36 (1), 24–28. doi:10.12116/j.issn.1004-5619.2020.01.006
- Rahimi, H., and Tehraninia, Z. (2020). A Comprehensive Review of Cutaneous Manifestations Associated with COVID-19. *Biomed. Res. Int.* 2020, 1–8. doi:10.1155/2020/1236520

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2022.831876/full#supplementary-material>

Supplementary Figure S1 | Details of different coronaviruses infected situation in *Manis javanica*. (A). Heatmap of sequence coverage on the viral genome. (B). Heatmap of total hits that aligned to the coronaviral genome.

Supplementary Figure S2 | Assembled genome of two coronaviruses. (A). Coverage of fragments isolated from virus infected *Callithrix jacchus* where the blue blocks mean consistent with MERS-CoV genome. (B). Coverage of fragments isolated from virus infected *Sus scrofa* where the blue blocks mean consistent with PEDV genome.

Supplementary Figure S3 | Receptor gene expression analysis. (A, B). Heatmap of natural logarithm of mean expression of 24 respiratory receptor genes in tissue intestine and lung of different species. (C). Box plot of natural logarithm of 5 receptor genes expression level between PRRSV infected and non-infected group in lung of *sus scrofa*.

Supplementary Figure S4 | Binding affinity analysis of SARS S1 proteins with ACE2s among species. (A). Schema of protein structure simulation and *in-silico* docking. (B). Polarized heatmap showing the z-dock score of SARS-CoV S1 protein and host ACE2 among species. (C). Sequence logo of ACE2 binding interface to SARS-CoV S1 of all species and sequence logo of ACE2 binding interface to SARS-CoV S1 of species with top 10 Z-dock score. (D). The overlay of simulated and docked human ACE2 + SARS-CoV-2 S1 structure on 7DF4.

- Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., et al. (2020). Structural Basis of Receptor Recognition by SARS-CoV-2. *Nature* 581 (7807), 221–224. doi:10.1038/s41586-020-2179-y
- Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* 11 (10), e0163962. doi:10.1371/journal.pone.0163962
- Simmons, G., Gosalia, D. N., Rennekamp, A. J., Reeves, J. D., Diamond, S. L., and Bates, P. (2005). Inhibitors of Cathepsin L Prevent Severe Acute Respiratory Syndrome Coronavirus Entry. *Proc. Natl. Acad. Sci.* 102 (33), 11876–11881. doi:10.1073/pnas.0505577102
- Souilmi, Y., Lauterbur, M. E., Tobler, R., Huber, C. D., Johar, A. S., Moradi, S. V., et al. (2021). An Ancient Viral Epidemic Involving Host Coronavirus Interacting Genes More Than 20,000 Years Ago in East Asia. *Curr. Biol.* 31 (16), 3504–3514. doi:10.1016/j.cub.2021.05.067
- Stadler, K., Masignani, V., Eickmann, M., Becker, S., Abrignani, S., Klenk, H.-D., et al. (2003). SARS - Beginning to Understand a New Virus. *Nat. Rev. Microbiol.* 1 (3), 209–218. doi:10.1038/nrmicro775
- Stout, A. E., André, N. M., Jaimes, J. A., Millet, J. K., and Whittaker, G. R. (2020). Coronaviruses in Cats and Other Companion Animals: Where Does SARS-CoV-2/covid-19 Fit. *Vet. Microbiol.* 247, 108777. doi:10.1016/j.vetmic.2020.108777
- Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2002). Multiple Sequence Alignment Using ClustalW and ClustalX. *Curr. Protoc. Bioinformatics* 0. doi:10.1002/0471250953.bi0203s00
- van der Hoek, L., Pyrc, K., Jebbink, M. F., Vermeulen-Oost, W., Berkhout, R. J. M., Wolthers, K. C., et al. (2004). Identification of a New Human Coronavirus. *Nat. Med.* 10 (4), 368–373. doi:10.1038/nm1024
- Webb, B., and Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinformatics* 54, 5–6. doi:10.1002/cpbi.3
- Widagdo, W., Sooksawasdi Na Ayudhya, S., Hundie, G. B., and Haagmans, B. L. (2019). Host Determinants of MERS-CoV Transmission and Pathogenesis. *Viruses* 11 (3), 280. doi:10.3390/v11030280
- Yu, X. J., Luo, C., Lin, J. C., Hao, P., He, Y. Y., Guo, Z. M., et al. (2003). Putative hAPN Receptor Binding Sites in SARS-CoV Spike Protein. *Acta Pharmacol. Sin* 24 (6), 481–488.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Zou, Cao, Yang, Deng, Xu, Dong, Li, Wu and Cao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.