SARS-CoV-2 spike protein predicted to form stable complexes with host receptor protein orthologues from mammals

Lam SD[1,2,†], Bordin N[2,†], Waman VP[2], Scholes HM[2], Ashford P[2], Sen N[2,3], van Dorp L[4], Rauer C[2], Dawson NL[2], Pang CSM[2], Abbasian M[2], Sillitoe I[2], Edwards SJL[5], Fraternali F[6], Lees JG[7], Santini JM[2], Orengo CA[2,*]

1 Department of Applied Physics, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, 43600, Malaysia
2 Institute of Structural and Molecular Biology, University College London, London, WC1E 6BT, UK
3 Indian Institute of Science Education and Research, Pune, 411008, India
4 UCL Genetics Institute, University College London, London, WC1E 6BT, UK
5 Department of Science and Technology Studies, University College London, London, WC1E 6BT, UK
6 Randall Division of Cell and Molecular Biophysics, Guy's Campus, New Hunt's House, King's College London, London, SE1 1UL, UK
7 Department of Biological and Medical Sciences, Faculty of Health and Life Sciences, Oxford Brookes University, Oxford, OX3 OBP, UK
† These authors contributed equally
* To whom correspondence should be addressed: c.orengo@ucl.ac.uk

Abstract

SARS-CoV-2 has a zoonotic origin and was transmitted to humans via an undetermined intermediate host, leading to infections in humans and other mammals. To enter host cells, the viral spike protein (S-protein) binds to its receptor, ACE2, and is then processed by TMPRSS2. Whilst receptor binding contributes to the viral host range, S-protein:ACE2 complexes from other animals have not been investigated widely. To predict infection risks, we modelled S-protein:ACE2 complexes from 215 vertebrate species, calculated their relative energies, correlated these energies to COVID-19 infection data, and analysed structural interactions. We predict that known mutations are more detrimental in ACE2 than TMPRSS2. Finally, we demonstrate phylogenetically that human SARS-CoV-2 strains have been isolated in animals. Our results suggest that SARS-CoV-2 can infect a broad range of mammals, but not fish, birds or reptiles. Susceptible animals could serve as reservoirs of the virus, necessitating careful ongoing animal management and surveillance.

Keywords: SARS-CoV-2, COVID-19, spike protein, ACE2, TMPRSS2, structural bioinformatics

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a novel coronavirus that emerged towards the end of 2019 and is responsible for the coronavirus disease 2019 (COVID-19) global pandemic. Available data suggests that SARS-CoV-2 has a zoonotic source[1], with the closest sequence currently available deriving from the horseshoe bat[2]. As yet, the transmission route to humans, including the intermediate host, is unknown. So far, little work has been done to assess the animal reservoirs of SARS-CoV-2, or the potential for the virus to spread to other species living with, or in close proximity to, humans in domestic, rural, agricultural or zoological settings.

Coronaviruses, including SARS-CoV-2, are major multi-host pathogens and can infect a wide range of non-human animals[3–5]. SARS-CoV-2 is in the Betacoronavirus genus, which includes viruses that infect economically important livestock, including cows[6], pigs[7], mice[8], rats[9], rabbits[10], and wildlife, such as antelope and giraffe[11]. Severe acute respiratory syndrome coronavirus (SARS-CoV), the betacoronavirus that caused the 2002-2004 SARS outbreak[12], likely jumped to humans from its original bat host via civets. Viruses genetically similar to human SARS-CoV have been isolated from animals as diverse as racoon dogs, ferret-badgers[3] and pigs[13], suggesting the existence of a large host reservoir. It is therefore probable that SARS-CoV-2 can also infect a wide range of species.

Real-world SARS-CoV-2 infections have been reported in cats[14], tigers[15], dogs[16,17] and minks[16,17]. Animal infection studies have also identified cats[18] and dogs[18] as hosts, as well as ferrets[18], macaques[19] and marmosets[19]. Recent *in vitro* studies have also suggested an even broader set of animals may be infected[20–22]. To understand the potential host range of SARS-CoV-2, the plausible extent of zoonotic and anthroponotic transmission, and to guide surveillance efforts, it is vital to know which species are susceptible to SARS-CoV-2 infection.

The receptor binding domain (RBD) of the SARS-CoV-2 spike protein (S-protein) binds to the extracellular peptidase domain of angiotensin I converting enzyme 2 (ACE2) mediating cell entry[23]. The sequence of ACE2 is highly conserved across vertebrates, suggesting that SARS-CoV-2 could use orthologues of ACE2 for cell entry. The structure of the SARS-CoV-2 S-protein RBD has been solved in complex with human ACE2[24] and identification of critical binding residues in this structure have provided valuable insights into viral recognition of the host receptor[24–28]. Compared with SARS-CoV, the SARS-CoV-2 S-protein has a 10–22-fold higher affinity for human ACE2[24,25,29], which is thought to be caused by three classes of mutations in the S-protein[30]. Similarly, variations in human ACE2 have also been found to increase affinity for S-protein receptor binding [31]. These factors may contribute to the host range and infectivity of SARS-CoV-2.

Both SARS-CoV-2 and SARS-CoV additionally require the transmembrane serine protease (TMPRSS2) to mediate cell entry. Together, ACE2 and TMPRSS2 confer specificity of host cell types that the virus can enter[32,33]. Upon binding to ACE2, the S-protein is cleaved by TMPRSS2 at two cleavage sites on separate loops, which primes the S-protein for cell entry[34]. TMPRSS2 has been docked against the SARS-CoV-2 S-protein, which revealed its binding site to be adjacent to these two cleavage sites[33]. An approved TMPRSS2 protease inhibitor drug is able to block SARS-CoV-2 cell entry[35], which demonstrates the key role of TMPRSS2 alongside ACE2. [36]. As such, both ACE2 and TMPRSS2 represent attractive therapeutic targets against SARS-CoV-2[37].

Recent work has predicted possible hosts for SARS-CoV-2 using the structural interplay between the

S-protein and ACE2. These studies proposed a broad range of hosts, covering hundreds of mammalian species, including tens of bat[27] and primate[38] species, and more comprehensive studies analysing all classes of vertebrates[38,39], including agricultural species of cow, sheep, goat, bison and water buffalo. In addition, sites in ACE2 have been identified as under positive selection in bats, particularly in regions involved in binding the S-protein[27]. The impacts of mutations in ACE2 orthologues have also been tested, for example structural modelling of ACE2 from 27 primate species[38] demonstrated that apes and African and Asian monkeys may also be susceptible to SARS-CoV-2. However, whilst cell entry is necessary for viral infection, it may not be sufficient alone to cause disease. For example, variations in other proteins may prevent downstream events that are required for viral replication in a new host. Hence, examples of real-world infections[14–17] and experimental data from animal infection studies[18–22] are required to validate hosts that are predicted to be susceptible.

Here, we analysed the effect of known mutations in orthologues of ACE2 and TMPRSS2 from a broad range of 215 vertebrate species, including primates, rodents and other placental mammals; birds; reptiles; and fish. For each species, we generated a 3-dimensional model of the ACE2 protein structure from its protein sequence and calculated the impacts of known mutations in ACE2 on the stability of the S-protein:ACE2 complex. We correlated changes in the energy of the complex with changes in the structure of ACE2, chemical properties of residues in the binding interface, and experimental COVID-19 infection phenotypes from *in vivo* and *in vitro* animal studies. To further test our predictions, we performed detailed manual structural analyses, presented as a variety of case studies for different species. Unlike other studies that analyse interactions that the S-protein makes with the host, we also analyse the impact of mutations in vertebrate orthologues of TMPRSS2. Our results suggest that SARS-CoV-2 can infect a broad range of vertebrates, which could serve as reservoirs of the virus, supporting future anthroponotic and zoonotic transmission.

Results

Conservation of ACE2 in vertebrates

We aligned protein sequences of 247 vertebrate orthologues of ACE2. Most orthologues have more than 60% sequence identity with human ACE2 (Supplementary Fig. 1A). For each orthologue, we generated a 3-dimensional model of the protein structure from its protein sequence using FunMod[40,41]. We were able to build high-quality models for 236 vertebrate orthologues, with nDOPE scores > -1 (Supplementary Table 5). 11 low-quality models were removed from the analysis. After this, we removed a further 21 models that were missing > 10 DCEX residues, leaving 215 models to take forward for further analysis. We observed high sequence (> 60% identity) and structure similarity (> 90) between ACE2 proteins for all species (Supplementary Results 1).

Identification of critical S-protein:ACE2 interface residues

ACE2 residues directly contacting the S-protein (DC residues) were identified in a structure of the complex (PDB ID 6M0J; Fig. 1a, Supplementary Results 2, Supplementary Fig. 3). We also identified a more extended set of both DC residues and residues within 8Å of DC residues likely to be influencing binding (DCEX residues).

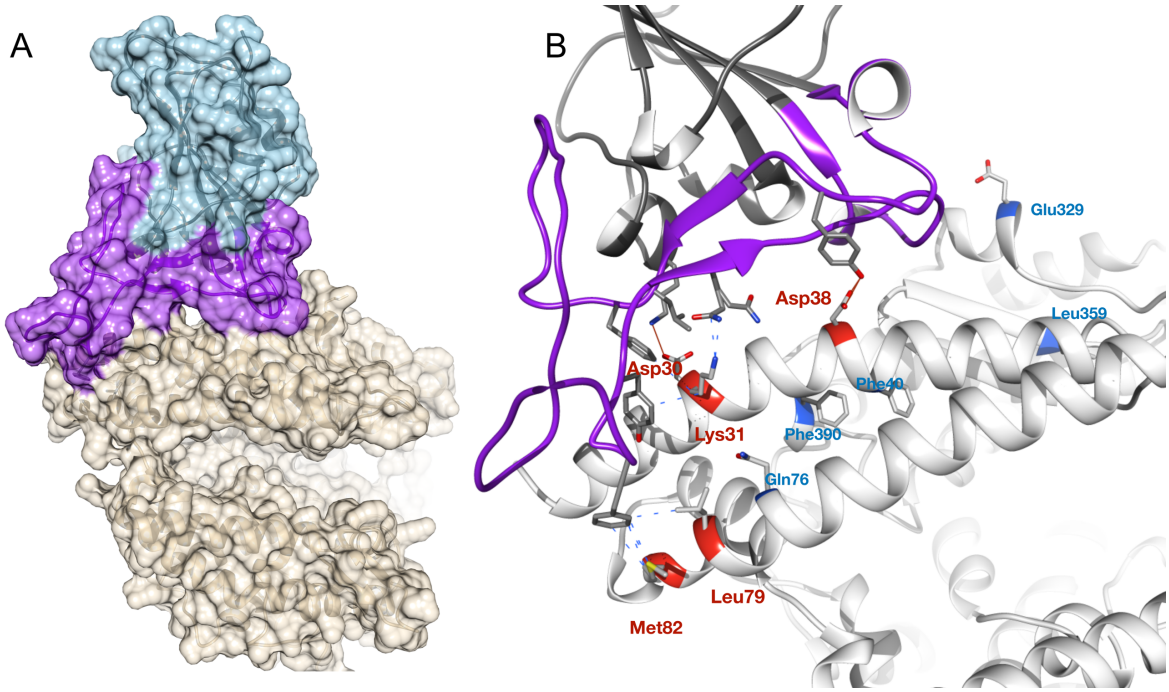

Figure 1. Overview of SARS-CoV-2 S-protein:human ACE2 complex and interface. (a) SARS-CoV-2 S-protein RBD (light blue, purple) showing the receptor binding motif (purple) at the interface with ACE2 (tan). (b) DC residues in the SARS-CoV-2 S-protein:human ACE2 complex interface. A subset of contact residues that are not conserved in vertebrates are highlighted for DC (red) and DCEX residues (blue) (PDB ID 6M0J).

Changes in the energy of the S-protein:ACE2 complex in vertebrates

We used three methods to assess the relative change in binding energy (ΔΔG) of the SARS-CoV-2 S-protein:ACE2 complex following mutations in DC residues and DCEX residues, that are likely to influence binding. We found that protocol 2 employing mCSM-PPI2 (henceforth referred to as P(2)-PPI2), calculated over the DCEX residues, correlated best with the phenotype data (Supplementary Results 3, Supplementary Fig. 4, Supplementary Table 6), justifying the use of animal models to calculate ΔΔG values in this context. Since this protocol considers mutations from animal to human, lower ΔΔG values correspond to stabilisation of the animal complex relative to the human complex, and therefore higher risk of infection. To consider ΔΔG values in an evolutionary context, we annotated phylogenetic trees for all 215 vertebrate species analysed (Supplementary Fig. 8) and for a subset of animals that humans come into close contact with in domestic, agricultural or zoological settings (Fig. 2). We show the residues that P(2)-PPI2 reports as stabilising or destabilising for the SARS-CoV-2 S-protein:ACE2 animal complex for DC (Supplementary Fig. 7) and DCEX (Supplementary Fig. 8) residues.

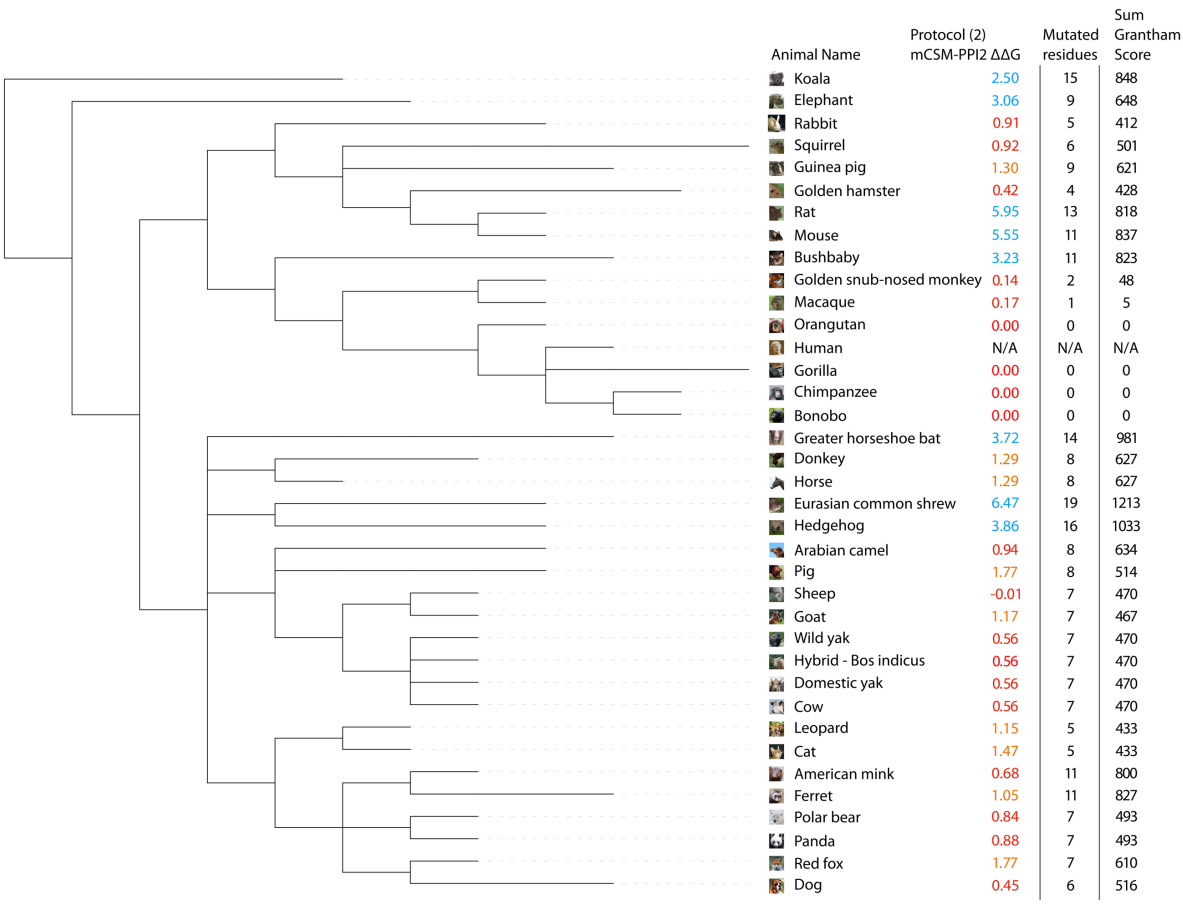

Figure 2 Phylogenetic tree of species that humans come into close contact with in domestic, agricultural or zoological settings. Leaves are annotated by the change in energy of the complex (ΔΔG), as measured by protocol 2 - PPI2. Animals are categorised according to risk of infection by SARS-CoV-2, with ΔΔG ≤ 1 high risk (red), 1 < ΔΔG < 2 medium risk (orange) and ΔΔG ≥ 2 low risk (blue). These thresholds were chosen as they agree well with the available experimental data. For each animal, the number of residue changes compared to human ACE2 sequence and the total chemical shift across the DCEX residues are also shown. This tree contains a subset of animals from Supplementary Fig. 6.

In general we see a high infection risk for most mammals, with a notable exception for all non-placental mammals. ΔΔG values measured by P(2)-PPI2 correlate well with the infection phenotypes (Supplementary Table 6). As shown in previous studies, many primates are at high risk[19,38,39]. Exceptions include New World monkeys, for which the capuchin and the squirrel monkey all show no infection risk in experimental studies, in agreement with our predicted energies for the complex[20]. Zoological animals that come into contact with humans, such as pandas, leopards and bears, are also at risk of infection. In agricultural settings, camels, cows, sheep, goats and horses also have relatively low ΔΔG values, suggesting comparable binding affinities to humans, in agreement with experimental data(20) (Supplementary Table 6). Whilst in domestic settings, dogs, cats, hamsters, and rabbits are also at risk from infection. Importantly, mice and rats are not susceptible, so hamsters and ferrets are being used as models of human COVID-19, instead of mice. Of the 35 birds tested only the blue tit shows an infection risk. Similarly, the Nile tilapia is the only fish out of the 72 in this study which shows a low change in energy of the complex, suggesting susceptibility to infection. Also, all 14 reptiles and amphibians we investigated do not show any risk.

Our predictions do not always agree with the experimental data. For some cases, we predict that some animals are at medium risk of infection, in conflict with experimental data. For example, we predict that guinea pigs and donkeys have a medium risk of infection, but no infections were observed *in vitro* for these animals[21]. However, infection has been observed *in vitro* for horse[21] and horse and donkey have identical DCEX residues and the same ΔΔG. On the other hand, we predict that some animals are at low risk of infection, despite experimental evidence to the contrary. For example, *in vivo* studies have shown that horseshoe bats[2] and marmosets[19] can be infected by SARS-CoV-2, but we predict that both animals have a low risk of infection, in agreement with *in vitro* data on marmosets[20]. We considered these discrepancies further using detailed structural analyses (Supplementary Results 4). Additionally, we compared changes in energy of the S-protein:ACE2 complex in SARS-CoV-2 and SARS-CoV and found similar changes suggesting that the range of animals susceptible to the virus is likely to be similar for SARS-CoV-2 and SARS-CoV (Supplementary Results 5).

Conservation of TMPRSS2 and its role in SARS-CoV-2 infection

ACE2 and TMPRSS2 are key factors in the SARS-CoV-2 infection process. Both are highly co-expressed in susceptible cell types, such as type II pneumocytes in the lungs, ileal absorptive enterocytes in the gut, and nasal goblet secretory cells[42]. Since both proteins are required for infection of host cells, and since our analyses clearly support suggestions of conserved binding of S-protein:ACE2 across animal species, we decided to analyse whether the TMPRSS2 was similarly conserved. There is no known structure of TMPRSS2, so we built a high-quality model from a template structure (PDB ID 5I25). Since TMPRSS2 is a serine protease, and the key catalytic residues are known, we used FunFams to identify highly conserved residues in the active site and the cleavage site that are likely to be involved in substrate binding. This resulted in two sets of residues that we analysed: the active site and cleavage site residues (ASCS), and the active site and cleavage site residues plus residues within 8Å of catalytic residues that are highly conserved in the FunFam (ASCSEX). The sum of Grantham scores for mutations in the active site and cleavage site for TMPRSS2 is zero or consistently lower than ACE2 in all organisms under consideration, for both ASCS and ASCSEX residues (Fig. 3). This means that the mutations in TMPRSS2 involve more conservative changes.

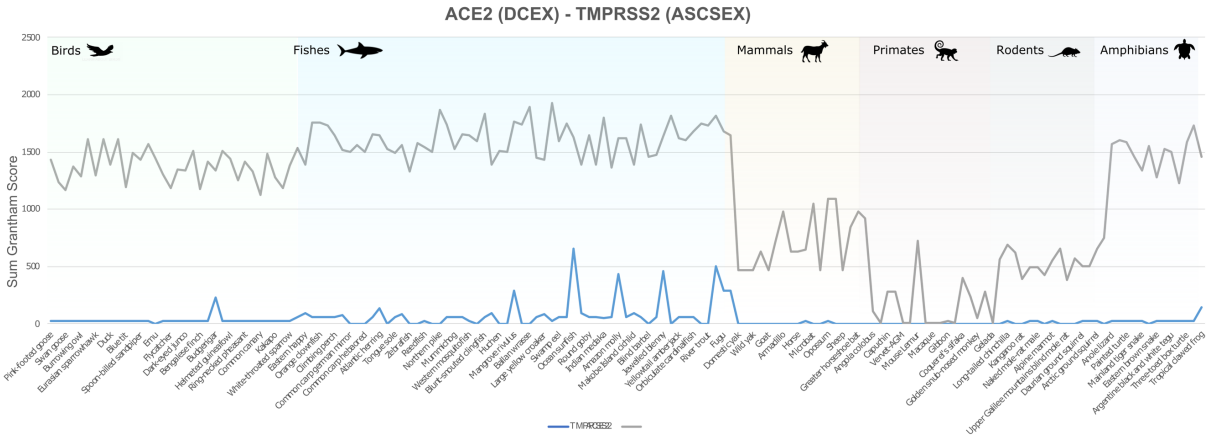

Figure 3: Comparison of Grantham score sums for ASCSEX residues in ACE2 and TMPRSS2.

Mutations DCEX residues seem to have a more disruptive effect in ACE2 than in TMPRSS2. Whilst we expect orthologues from organisms that are close to humans to be conserved and have lower Grantham scores, we observed some residue substitutions that have high Grantham scores for

primates, such as capuchin, marmoset and mouse lemur. In addition, primates, such as the coquerel sifaka, greater bamboo lemur and Bolivian squirrel monkey, have mutations in DCEX residues with high Grantham scores. Mutations in TMPRSS2 may render these animals less susceptible to infection by SARS-CoV-2.

Phylogenetic Analysis of SARS-like strains in different animal species

A small-scale phylogenetic analysis was performed on a subset of SARS-CoV-2 assemblies in conjunction with a broader range of SARS-like betacoronaviruses (Supplementary Table 3), including SARS-CoV isolated from humans and civets. The phylogeny is consistent with previous work[2] which identified the virus isolated from horseshoe bats (RaTG13, EPI_ISL_402131) as the closest genome to SARS-CoV-2 strains currently available (Fig. 4). Aided by a large community effort, thousands of human-associated SARS-CoV-2 genome assemblies are now accessible on GISAID[16,17]. To date, these also include one assembly generated from a virus infecting a domestic dog (EPI_ISL_414518), one obtained from a zoo tiger (EPI_ISL_420923) and one obtained from a mink (EPI_ISL_431778). SARS-CoV-2 strains from animal infections all fall within the phylogenetic diversity observed in human lineages (Fig. 4a). The receptor binding domain is completely conserved (Fig. 4b-c) across both human and animal SARS-CoV-2, with replacements in the spike protein of dog (S-protein V8L), tiger (S-protein D614G) and mink (S-protein D614G) strains relative to Wuhan-Hu-1 also observed in human-associated lineages[43], consistent with circulation in non-human hosts. Of note, whilst genome-wide data indicates a closer phylogenetic relationship between SARS-CoV-2 strains and RaTG13, the receptor binding domain alignment instead supports a closer relationship with a virus isolated from pangolins[44] (EPI_ISL_410721; Fig. 4c), in line with previous reports[45]. This highlights the importance of considering variations in structures of proteins that may determine the host range.

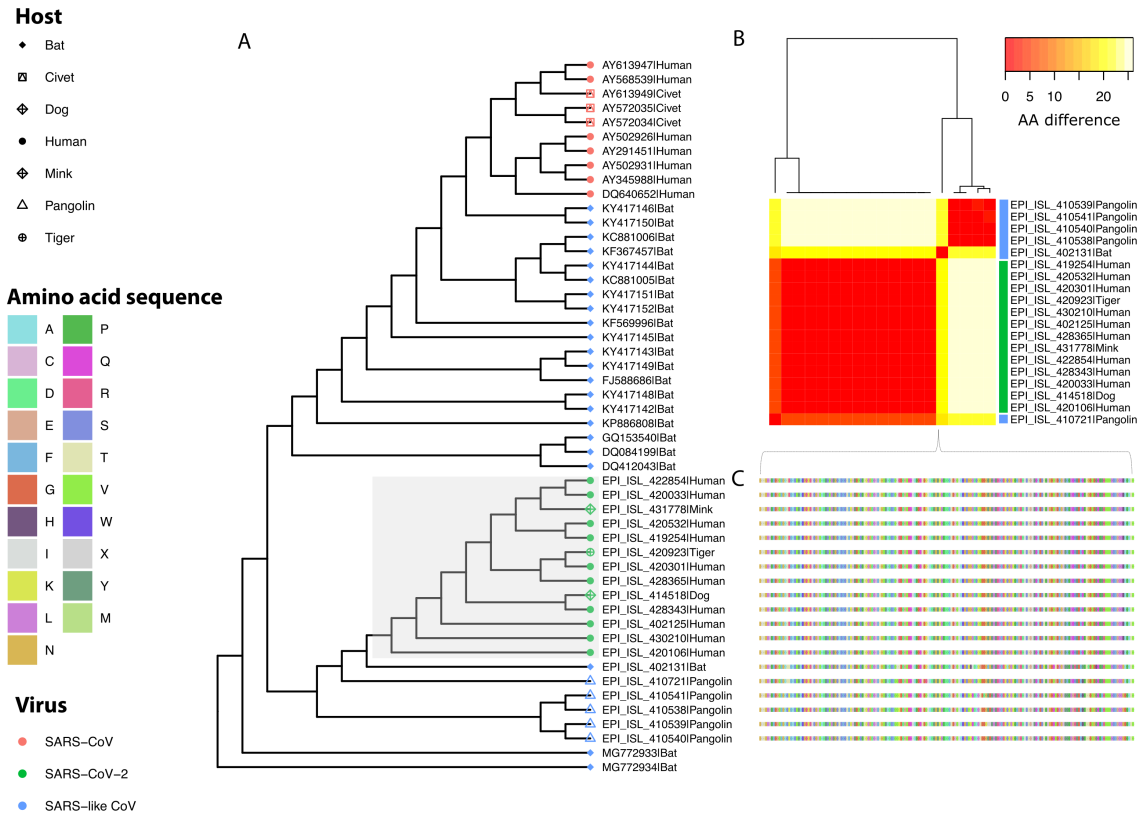

Figure 4: Phylogeny of SARS-like viruses. (a) Genome-wide maximum likelihood phylogenetic tree of SARS-like betacoronavirus strains sampled from diverse hosts (coloured tip symbols; Supplementary Table 3). Genome EPI_ISL_402131 is sample RaTG13 from a horseshoe bat. (b) Pairwise amino acid differences at the S-protein RBD between human and animal strains of SARS-CoV-2, relative to closely related-SARS-like viruses in animal hosts. (c) Sequence alignment for the spike protein receptor binding domain.

Discussion

The ongoing COVID-19 global pandemic has a zoonotic origin, necessitating investigations into how SARS-CoV-2 infects animals, and how the virus can be transmitted across species. Given the role that the stability of the complex, formed between the S-protein and its receptors, could contribute to the viral host range, zoonosis and anthroponosis, there is a clear need to study these interactions. However, relative changes in the energies of the S-protein:ACE2 complex have not been explored experimentally or *in silico*. A number of recent studies[20,21,39] have suggested that, due to high conservation of ACE2, some animals are vulnerable to infection by SARS-CoV-2. Concerningly, these animals could, in theory, serve as reservoirs of the virus, increasing the risk of future transmission across species, but transmission rates across species are not known. Therefore, it is important to try to predict which other animals could potentially be infected by SARS-CoV-2, so that the plausible extent of zoonotic transmission can be estimated, and surveillance efforts can be guided appropriately.

Animal susceptibility to infection by SARS-CoV-2 has been studied *in vivo*[18,19,46–48] and *in vitro*[20–22]

during the course of the pandemic. Parallel *in silico* work has made use of the protein structure of the S-protein:ACE2 complex to computationally predict the breadth of possible viral hosts. Most studies simply considered the number of residues mutated relative to human ACE2[29,49,50], although some also analyse the effect that these mutations have on the interface stability[31,38,51]. The most comprehensive of these studies analysed the number, and locations, of mutated residues in ACE2 orthologues from 410 species[39], but did not perform detailed energy calculations as we have done. Also, no assessment of TMPRSS2 was made. Furthermore, our work is the only study that has so far explored changes in the energy of the S-protein:ACE2 complex on a large scale.

In this study, we performed a comprehensive analysis of the major proteins that SARS-CoV-2 uses for cell entry. We predicted structures of ACE2 and TMPRSS2 orthologues from 215 vertebrate species and modelled S-protein:ACE2 complexes. We calculated relative changes in energy (ΔΔG) of S-protein:ACE2 complexes, *in silico,* following mutations from animal residues to those in human. Our predictions suggest that, whilst many mammals are susceptible to infection by SARS-CoV-2, birds, fish and reptiles are not likely to be. We manually analysed residues in the S-protein:ACE2 interface, including DC residues that directly contacted the other protein, and DCEX residues that also included residues within 8Å of the binding residues, that may affect binding. We clearly showed the advantage of performing more sophisticated studies of the changes in energy of the complex, over more simple measures—such as the number or chemical nature of mutated residues—used in other studies. Furthermore, the wider set of DCEX residues that we identified near the binding interface had a higher correlation to the phenotype data than the DC residues. In addition to ACE2, we also analysed how mutations in TMPRSS2 impact binding to the S-protein. We found that mutations in TMPRSS2 are less disruptive than mutations in ACE2, indicating that binding interactions in the S-protein:TMPRSS2 complex in different species will not be affected.

To increase our confidence in assessing changes in the energy of the complex, we developed multiple protocols using different, established methods. We correlated these stability measures with experimental infection phenotypes in the literature, from *in vivo*[18,19,46–48] and *in vitro*[20–22] studies of animals. Protocol 2 using mCSM-PPI2 (P(2)-PPI2) correlated best with the number of mutations, chemical changes induced by mutations and infection phenotypes, so we chose to focus our analysis employing this protocol.

Humans are likely to come into contact with 26 of these species in domestic, agricultural or zoological settings (Fig. 5). Of particular concern are sheep, that have no change in energy of the S-protein:ACE2 complex, as these animals are farmed and come into close contact with humans. We also provide phylogenetic evidence that human SARS-CoV-2 is also present in tigers[15], dogs[16,17] and minks[16,17] (Fig. 4), consistent with reports of human-to-animal transmission. Our measurements of the change in energy of the complex for the SARS-CoV S-protein were highly correlated with SARS-CoV-2, so our findings are also applicable to SARS-CoV.

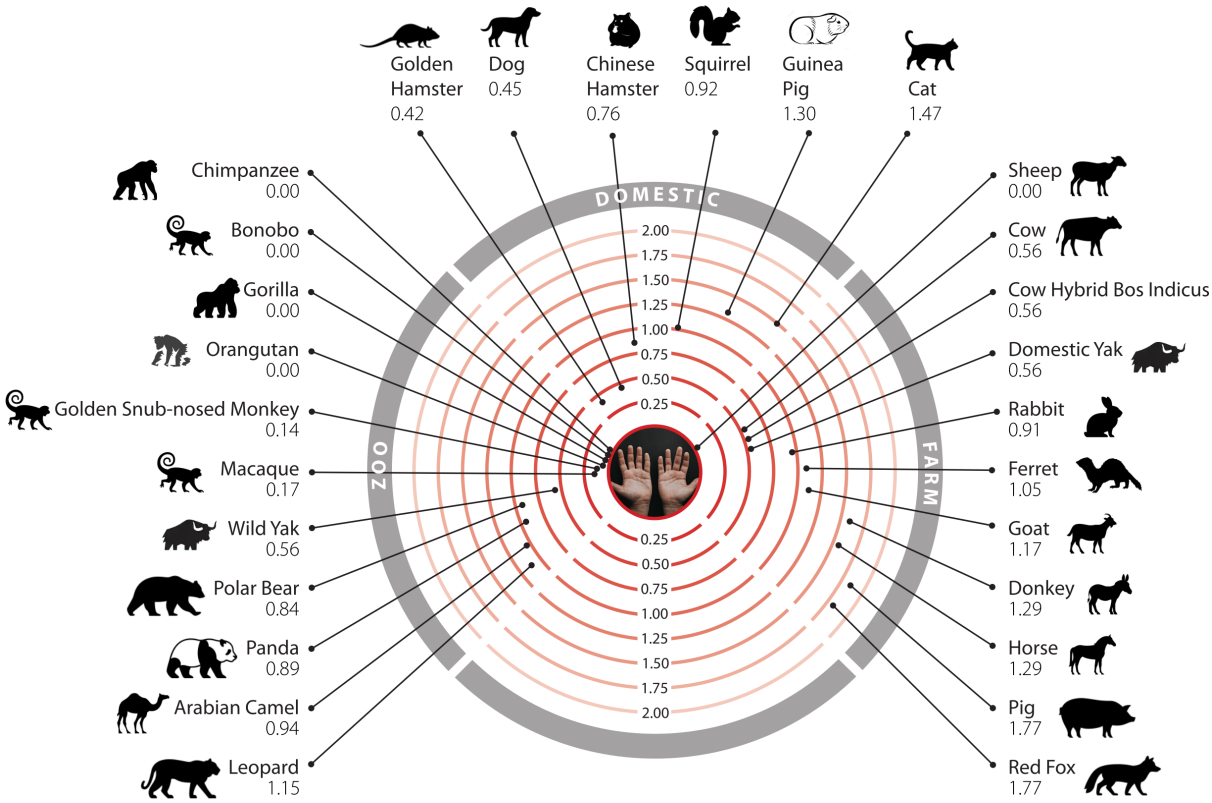

Figure 5: Mammals that humans come into contact with that are at risk of infection by SARS-CoV-2. Twenty-six mammals are categorised into* domestic, agricultural or zoological settings*. Numbers represent the change in binding energy (ΔΔG) of the S-protein:ACE2.

To gain a better understanding of the nature of the S-protein:ACE2 interface, we performed more detailed structural analyses for a subset of species. In some cases, we had found discrepancies between our energy calculations and experimental phenotypes. To test our predictions, we manually analysed how the shape or chemistry of residues may impact complex stability for all DC residues and a selection of DCEX residues. In agreement with other studies[25], we identified a number of locations in ACE2 that are important for binding the S-protein. These locations, namely the hydrophobic cluster near the N-terminus and two hotspot locations near residues 31 and 353, stabilise the binding interface. Five DC residues have species-specific variants and influence how well S-protein can bind utilising these key interface regions. In agreement with our calculations in changes in energy of the S-protein:ACE2 complex, our structural studies do not support *(in vivo)* infection of horseshoe bat, which has variants at three out of five variant DC residues, one of which (D38N) causes the loss of a salt bridge and H-bonding interactions between ACE2 and S-protein at hotspot 353. These detailed structural analyses are supported by the high Grantham score and calculated total ΔΔG for the change in energy of the complex. Both dog and cat have a physico-chemically similar variant at this hotspot (D38E), which although disrupting the salt bridge still permits alternative H-bonding interactions between the spike RBD and ACE2.

SARS-CoV-2 is better able to exploit the hydrophobic pocket than SARS-CoV by increased flexibility of its RBD loop and by mutation of L486F[25]. Our structures show how SARS-CoV-2 can utilise this pocket for binding at the interface in a wide range of species. Of species with DCEX variants at this pocket, only guinea pig maintains an entirely hydrophobic environment (M82A), whilst also conserving three

out of the five variant DC residues. This helps explain why we predict moderate risk of infection, in contrast to the *in vitro* experimental data that reports no infection. For marmoset, there is contradictory *in vivo* and *in vitro* experimental data. Our energy calculations suggest no risk and the structural analyses support this by identifying a large structural difference caused by a 39 residue insert. This alters the overall structural superposition of the marmoset and human structures, which could affect the energy of the complex. Finally, some DCEX residues were predicted to be allosteric sites, which may be promising drug targets[52].

We applied protocols that enabled a comprehensive study of host range, within a reasonable time, for identifying species at risk of infection by SARS-CoV-2, or of becoming reservoirs of the virus. Although we felt that these faster methods were justified by the need for timely answers to these questions, there are clearly caveats to our work that should be taken into account. Whilst we use a state of the art modelling tool[53] and an endorsed method for calculating changes in energy of the complex[54], molecular dynamics may give a more accurate picture of energy changes by sampling rotamer space more comprehensively. However, such an approach would have been prohibitively expensive at a time when it is clearly important to identify animals at risk as quickly as possible. Furthermore, although the animals we highlight at risk from our changes in binding energy calculations correlate well with the experimental data, there is only a small amount of such data currently available, and many of the experimental papers reporting these data are yet to be peer reviewed. Finally, we restricted our analyses to one strain of SARS-CoV-2, but other strains may have evolved with mutations that give more complementary interfaces. Recent work reporting a new SARS-CoV-2 strain that can infect mice(107) suggests that this could be the case.

The ability of SARS-CoV-2 to infect host cells and cause COVID-19, sometimes resulting in severe disease, ultimately depends on a multitude of other host-virus protein interactions[37]. While we do not investigate them all in this study, our results suggest that SARS-CoV-2 can indeed infect a broad range of mammals. As there is a possibility of creating new reservoirs of the virus, we should now consider how to identify such transmission early and to mitigate against such risks. Animals living in close contact with humans should be monitored and farm animals should be protected where possible and managed accordingly[55].

Methods

Sequence Data

ACE2 protein sequences for 239 vertebrates, including humans, were obtained from ENSEMBL[56] version 99 and eight sequences from UniProt release 2020_1 (Supplementary Table 1). TMPRSS2 protein sequences for 278 vertebrate sequences, including the human sequence, were obtained from ENSEMBL (Supplementary Table 2).

A phylogenetic tree of species, to indicate the evolutionary relationships between animals, was downloaded from ENSEMBL[56].

Structural Data

The structure[24] of the SARS-CoV-2 S-protein bound to human ACE2 at 2.45Å was used throughout (PDB ID 6M0J).

Sequence analysis

We used standard methods to analyse the sequence similarity between human ACE2 and other vertebrate species (Supplementary Methods 1). We also mapped ACE2 and TMPRSS2 sequences to our CATH functional families to detect residues highly conserved across species (Supplementary Methods 1).

Structure analysis

Identifying residues in ACE2

In addition to residues in ACE2 that contact the S-protein directly, various other studies have also considered residues that are in the second shell, or are buried, and could influence binding[57]. Therefore, in our analyses we built on these approaches and extended them to compile the following sets for our study:

1. *Direct contact (DC)* residues. This includes a total of 20 residues that are involved in direct contact with the S-protein[24] identified by PDBe[58] and PDBSum[59].
2. *Direct Contact Extended (DCEX) residues.* This dataset includes residues within 8Å of DC residues, that are likely to be important for binding. These were selected by detailed manual inspection of the complex, and also considering the following criteria: (i) reported evidence from deep mutagenesis[57], (ii) *in silico* alanine scanning (using mCSM-PPI[60]), (iii) residues with high evolutionary conservation patterns identified by the FunFams-based protocol described above, i.e. residues identified with DOPS ≥ 70 and ScoreCons score ≥ 0.7, (iv) allosteric site prediction (Supplementary Methods 2), and (v) sites under positive selection (Supplementary Methods 2). Selected residues are shown in Supplementary Fig. 1 and residues very close to DC residues (i.e. within 5Å) are annotated.

We also included residues identified by other related structural analyses, reported in the literature (Supplementary Methods 2).

Generating 3-dimensional structure models

Using the ACE2 protein sequence from each species, structural models were generated for the S-protein:ACE2 complex for 247 animals using the FunMod modelling pipeline[40,41], based on MODELLER[53] (Supplementary Methods 3). Models were refined by MODELLER to optimise the geometry of the complex and the interface. Only high-quality models were used in this analysis, with nDOPE[61] score < -1 and with < 10 DCEX residues missing. This gave a final dataset of 215 animals for further analysis.

The modelled structures of ACE2 were compared against the human structure (PDB ID 6M0J) and pairwise, against each other, using SSAP[62]. SSAP returns a score in the range 0-100, with identical structures scoring 100.

We also built models for TMPRSS2 proteins in all available species and identified the residues likely to be involved in the protein function (Supplementary Methods 3).

Measuring changes in the energy of the S-protein:ACE2 complex in SARS-CoV-2 and SARS-CoV

We calculated the changes in binding energy of the SARS-CoV-2 S-protein:ACE2 complex and the SARS-CoV S-protein:ACE2 complex of different species, compared to human, following three different

protocols:

1. *Protocol 1:* Using the human complex and mutating the residues for the ACE2 interface to those found in the given animal sequence and then calculating the ΔΔG of the complex using both mCSM-PPI1[60] and mCSM-PPI2[54] (Supplementary Methods 4). This gave a measure of the destabilisation of the complex in the given ***animal relative to the human*** complex. ΔΔG values < 0 are associated with destabilising mutations, whilst values ≥ 0 are associated with stabilising mutations.
2. *Protocol 2:* We repeated the analysis with both mCSM-PPI1 and mCSM-PPI2 as in protocol 1, but using the animal 3-dimensional models, instead of the human ACE2 structure, and calculating the ΔΔG of the complex by mutating the animal ACE2 interface residue to the appropriate residue in the human ACE2 structure. This gave a measure of the destabilisation of the complex in the ***human complex relative to the given animal***. Values ≤ 0 are associated with destabilisation of the human complex (i.e. animal complexes more stable), whilst values > 0 are associated with stabilisation of the human complex (i.e. animal complexes less stable).
3. *Protocol 3:* We used the PRODIGY server[63] to calculate the binding energy for the human complex and for the 3-dimensional models of the other 215 animal complexes. As this method delivers an absolute binding energy, we calculated the change in binding energy from the human complex to the animal complex as $\Delta\Delta G = \Delta G_{human} - \Delta G_{animal}$, at 298 and 310 Kelvin.

We subsequently correlated ΔΔG values with available *in vivo* and *in vitro* experimental data on COVID-19 infection data for mammals. Protocol 2, mCSM-PPI2, correlated best with these data. This allowed us to assign thresholds for risk of infection by SARS-CoV-2 on the ΔΔG values with ΔΔG ≤ 1 high risk, 1 < ΔΔG < 2 medium risk and ΔΔG ≥ 2 low risk.

Change in residue chemistry for mutations

To measure the degree of chemical change associated with mutations occurring in DC and DCEX residues, we computed the Grantham score[64] for each vertebrate compared to the human sequence (Supplementary Methods 5).

Phylogeny of SARS-like betacoronaviruses

We performed phylogenetic analyses for a subset of SARS-CoV (n = 10), SARS-like (n = 27) and SARS-CoV-2 (n = 13) viruses from publicly available data in NCBI[65–72] and GISAID[16,17] (Supplementary Methods 6).

Funding

HS is funded by Wellcome [203780/Z/16/A]. LvD acknowledges financial support from the Newton Fund UK-China NSFC initiative [MR/P007597/1] and a BBSRC equipment grant [BB/R01356X/1]. ND is funded by Wellcome [104960/Z/14/Z]. The following people acknowledge BBSRC for their funding: NB [BB/R009597/1], PA [BB/S016007/1], NS [BB/S020144/1], CR [BB/T002735/1], IS [BB/R014892/1], VW [BB/S020039/1]. SE is funded by EDCTP PANDORA-ID NET, UCLH/UCL Biomedical Research Centre, and the Medical Research Council.

Author contributions

SL conceived the idea of analysing structures and effects of mutations in the S-protein:human ACE2 complex. JL conceived the idea of extending the analyses to animal complexes, for animals reported to be infected. JS conceived the idea of extending to a larger set of animals to explore host range. CO conceived the idea of contrasting multiple protocols to validate predictions. SL, CO, JL, JS, LvD designed the experiments. SL, NB, VW, PA, LvD performed the experiments. SL, NB, VW, PA, NS, JS, LvD, CR, IS, JL, CO analysed data. SL, NB, VW, HS, PA, NS, JS, LvD, CR, ND, IS, JL, CO interpreted the results. SL, NB, VW, HS, PA, NS, JS, LvD, CR, ND, CSMP, MA, IS, JL, CO contributed to the manuscript and figures. HS, VW, PA, SL, CO wrote the manuscript. JL, SE, FF, JS, CO revised the manuscript.

References

1. World Health Organisation. COVID-19 situation report - 94.
2. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
3. Guan, Y. *et al.* Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* **302**, 276–278 (2003).
4. Wang, L. F. & Eaton, B. T. Bats, civets and the emergence of SARS. *Curr. Top. Microbiol. Immunol.* **315**, 325–344 (2007).
5. Shaw, L. P. *et al. The phylogenetic range of bacterial and viral pathogens of vertebrates*. http://biorxiv.org/lookup/doi/10.1101/670315 (2019) doi:10.1101/670315.
6. Saif, L. J. Bovine Respiratory Coronavirus. *Vet. Clin. North Am. Food Anim. Pract.* **26**, 349–364 (2010).
7. Vlasova, A. N. *et al.* Porcine Coronaviruses. *Emerg. Transbound. Anim. Viruses* 79–110 (2020) doi:10.1007/978-981-15-0402-0_4.
8. Wang, W. *et al.* Discovery, diversity and evolution of novel coronaviruses sampled from rodents in China. *Virology* **474**, 19–27 (2015).
9. Lau, S. K. P. *et al.* Discovery of a Novel Coronavirus, China Rattus Coronavirus HKU24, from Norway Rats Supports the Murine Origin of Betacoronavirus 1 and Has Implications for the Ancestor of Betacoronavirus Lineage A. *J. Virol.* **89**, 3076–3092 (2015).

10. Lau, S. K. P. *et al.* Isolation and Characterization of a Novel Betacoronavirus Subgroup A Coronavirus, Rabbit Coronavirus HKU14, from Domestic Rabbits. *J. Virol.* **86**, 5481–5496 (2012).
11. Hasoksuz, M. *et al.* Biologic, antigenic, and full-length genomic characterization of a bovine-like coronavirus isolated from a giraffe. *J. Virol.* **81**, 4981–4990 (2007).
12. WHO | SARS (Severe Acute Respiratory Syndrome). *WHO* https://www.who.int/ith/diseases/sars/en/.
13. Chen, W. *et al.* SARS-associated Coronavirus Transmitted from Human to Pig. *Emerg. Infect. Dis.* **11**, 446–448 (2005).
14. Zhang, Q. *et al. SARS-CoV-2 neutralizing serum antibodies in cats: a serological investigation*. http://biorxiv.org/lookup/doi/10.1101/2020.04.01.021196 (2020) doi:10.1101/2020.04.01.021196.
15. USDA APHIS | USDA Statement on the Confirmation of COVID-19 in a Tiger in New York. https://www.aphis.usda.gov/aphis/newsroom/news/sa_by_date/sa-2020/ny-zoo-covid-19.
16. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health: Data, Disease and Diplomacy. *Glob. Chall.* **1**, 33–46 (2017).
17. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull.* **22**, (2017).
18. Shi, J. *et al. Susceptibility of ferrets, cats, dogs, and different domestic animals to SARS-coronavirus-2*. http://biorxiv.org/lookup/doi/10.1101/2020.03.30.015347 (2020) doi:10.1101/2020.03.30.015347.
19. Lu, S. *et al. Comparison of SARS-CoV-2 infections among 3 species of non-human primates*. http://biorxiv.org/lookup/doi/10.1101/2020.04.08.031807 (2020) doi:10.1101/2020.04.08.031807.
20. Liu, Y. *et al.* Functional and Genetic Analysis of Viral Receptor ACE2 Orthologs Reveals Broad Potential Host Range of SARS-CoV-2. *bioRxiv* 2020.04.22.046565 (2020) doi:10.1101/2020.04.22.046565.

21. Li, Y. *et al. Potential host range of multiple SARS-like coronaviruses and an improved ACE2-Fc variant that is potent against both SARS-CoV-2 and SARS-CoV-1*. http://biorxiv.org/lookup/doi/10.1101/2020.04.10.032342 (2020) doi:10.1101/2020.04.10.032342.
22. Zhao, X. *et al.* Broad and differential animal ACE2 receptor usage by SARS-CoV-2. *bioRxiv* 2020.04.19.048710 (2020) doi:10.1101/2020.04.19.048710.
23. Li, W. *et al.* Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* **426**, 450–454 (2003).
24. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* 1–6 (2020) doi:10.1038/s41586-020-2180-5.
25. Shang, J. *et al.* Structural basis of receptor recognition by SARS-CoV-2. *Nature* (2020) doi:10.1038/s41586-020-2179-y.
26. Li, F., Li, W., Farzan, M. & Harrison, S. C. Structure of SARS Coronavirus Spike Receptor-Binding Domain Complexed with Receptor. *Science* **309**, 1864–1868 (2005).
27. Frank, H. K., Enard, D. & Boyd, S. D. Exceptional diversity and selection pressure on SARS-CoV and SARS-CoV-2 host receptor in bats compared to other mammals. *bioRxiv* 2020.04.20.051656 (2020) doi:10.1101/2020.04.20.051656.
28. Sun, J. *et al.* COVID-19: Epidemiology, Evolution, and Cross-Disciplinary Perspectives. *Trends Mol. Med.* **0**, (2020).
29. Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020).
30. Ou, J. *et al. Emergence of RBD mutations in circulating SARS-CoV-2 strains enhancing the structural stability and human ACE2 receptor affinity of the spike protein*. http://biorxiv.org/lookup/doi/10.1101/2020.03.15.991844 (2020) doi:10.1101/2020.03.15.991844.
31. Stawiski, E. W. *et al. Human ACE2 receptor polymorphisms predict SARS-CoV-2 susceptibility*.

http://biorxiv.org/lookup/doi/10.1101/2020.04.07.024752 (2020) doi:10.1101/2020.04.07.024752.

32. Ziegler, C. *et al. SARS-CoV-2 Receptor ACE2 is an Interferon-Stimulated Gene in Human Airway Epithelial Cells and Is Enriched in Specific Cell Subsets Across Tissues*. https://papers.ssrn.com/abstract=3555145 (2020) doi:10.2139/ssrn.3555145.
33. Hussain, M. *et al.* Structural Basis of SARS-CoV-2 Spike Protein Priming by TMPRSS2. *bioRxiv* 2020.04.21.052639 (2020) doi:10.1101/2020.04.21.052639.
34. Heurich, A. *et al.* TMPRSS2 and ADAM17 Cleave ACE2 Differentially and Only Proteolysis by TMPRSS2 Augments Entry Driven by the Severe Acute Respiratory Syndrome Coronavirus Spike Protein. *J. Virol.* **88**, 1293–1307 (2014).
35. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).
36. Asselta, R., Paraboschi, E. M., Mantovani, A. & Duga, S. ACE2 and TMPRSS2 variants and expression as candidates to sex and country differences in COVID-19 severity in Italy. *medRxiv* 2020.03.30.20047878 (2020) doi:10.1101/2020.03.30.20047878.
37. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 1–13 (2020) doi:10.1038/s41586-020-2286-9.
38. Melin, A. D., Janiak, M. C., Marrone, F., Arora, P. S. & Higham, J. P. *Comparative ACE2 variation and primate COVID-19 risk*. http://biorxiv.org/lookup/doi/10.1101/2020.04.09.034967 (2020) doi:10.1101/2020.04.09.034967.
39. Damas, J. *et al. Broad Host Range of SARS-CoV-2 Predicted by Comparative and Structural Analysis of ACE2 in Vertebrates*. http://biorxiv.org/lookup/doi/10.1101/2020.04.16.045302 (2020) doi:10.1101/2020.04.16.045302.
40. Lam, S. D. *et al.* Gene3D: expanding the utility of domain assignments. *Nucleic Acids Res.* **44**, D404–D409 (2016).
41. Lam, S. D., Das, S., Sillitoe, I. & Orengo, C. An overview of comparative modelling and resources

dedicated to large-scale modelling of genome sequences. *Acta Crystallogr. Sect. Struct. Biol.* **73**, 628–640 (2017).

42. Hamming, I. *et al.* Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis. *J. Pathol.* **203**, 631–637 (2004).
43. CVR Bioinformatics. CoV-GLUE: Virus genome analysis for the SARS-CoV-2 pandemic. http://cov-glue.cvr.gla.ac.uk/#/home.
44. Xiao, K. *et al.* Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature* (2020) doi:10.1038/s41586-020-2313-x.
45. Zhang, T., Wu, Q. & Zhang, Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Curr. Biol.* **30**, 1346-1351.e2 (2020).
46. Kim, Y.-I. *et al.* Infection and Rapid Transmission of SARS-CoV-2 in Ferrets. *Cell Host Microbe* S1931312820301876 (2020) doi:10.1016/j.chom.2020.03.023.
47. Temmam, S. *et al. Absence of SARS-CoV-2 infection in cats and dogs in close contact with a cluster of COVID-19 patients in a veterinary campus*. http://biorxiv.org/lookup/doi/10.1101/2020.04.07.029090 (2020) doi:10.1101/2020.04.07.029090.
48. Shan, C. *et al. Infection with Novel Coronavirus (SARS-CoV-2) Causes Pneumonia in the Rhesus Macaques*. https://www.researchsquare.com/article/rs-15756/v1 (2020) doi:10.21203/rs.2.25200/v1.
49. Liu, Z. *et al.* Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2. *J. Med. Virol.* **92**, 595–601 (2020).
50. Luan, J., Jin, X., Lu, Y. & Zhang, L. SARS-CoV-2 spike protein favors ACE2 from Bovidae and Cricetidae. *J. Med. Virol.* **n/a**,.
51. Jia, Y. *et al. Analysis of the mutation dynamics of SARS-CoV-2 reveals the spread history and emergence of RBD mutant with lower ACE2 binding affinity*. http://biorxiv.org/lookup/doi/10.1101/2020.04.09.034942 (2020)

doi:10.1101/2020.04.09.034942.

52. Han, Y. & Král, P. Computational Design of ACE2-Based Peptide Inhibitors of SARS-CoV-2. *ACS Nano* acsnano.0c02857 (2020) doi:10.1021/acsnano.0c02857.
53. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinforma.* **54**, 5.6.1-5.6.37 (2016).
54. Rodrigues, C. H. M., Myung, Y., Pires, D. E. V. & Ascher, D. B. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic Acids Res.* **47**, W338–W344 (2019).
55. Santini, J. & Edwards, S. J. Host-range of human SARS-CoV-2 and implications for public health. Under Review. *Rev.*
56. Herrero, J. *et al.* Ensembl comparative genomics resources. *Database* **2016**, bav096 (2016).
57. Procko, E. The sequence of human ACE2 is suboptimal for binding the S spike protein of SARS coronavirus 2. *bioRxiv* 2020.03.16.994236 (2020) doi:10.1101/2020.03.16.994236.
58. Armstrong, D. R. *et al.* PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Res.* gkz990 (2019) doi:10.1093/nar/gkz990.
59. Laskowski, R. A. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.* **29**, 221–222 (2001).
60. Pires, D. E. V., Ascher, D. B. & Blundell, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinforma. Oxf. Engl.* **30**, 335–342 (2014).
61. Shen, M.-Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci. Publ. Protein Soc.* **15**, 2507–2524 (2006).
62. Orengo, C. A. & Taylor, W. R. SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.* **266**, 617–635 (1996).
63. Xue, L. C., Rodrigues, J. P., Kastritis, P. L., Bonvin, A. M. & Vangone, A. PRODIGY: a web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics* btw514 (2016) doi:10.1093/bioinformatics/btw514.
64. Grantham, R. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* **185**,

862–864 (1974).

65. Yeh, S.-H. *et al.* Characterization of severe acute respiratory syndrome coronavirus genomes in Taiwan: Molecular epidemiology and genome evolution. *Proc. Natl. Acad. Sci.* **101**, 2542–2547 (2004).
66. Chim, S. S. C. *et al.* Genomic characterisation of the severe acute respiratory syndrome coronavirus of Amoy Gardens outbreak in Hong Kong. *Lancet Lond. Engl.* **362**, 1807–1808 (2003).
67. Hu, B. *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLOS Pathog.* **13**, e1006698 (2017).
68. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).
69. Li, W. *et al.* Bats are natural reservoirs of SARS-like coronaviruses. *Science* **310**, 676–679 (2005).
70. Lau, S. K. P. *et al.* Ecoepidemiology and Complete Genome Comparison of Different Strains of Severe Acute Respiratory Syndrome-Related Rhinolophus Bat Coronavirus in China Reveal Bats as a Reservoir for Acute, Self-Limiting Infection That Allows Recombination Events. *J. Virol.* **84**, 2808–2819 (2010).
71. Lau, S. K. P. *et al.* Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 14040–14045 (2005).
72. Hu, D. *et al.* Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. *Emerg. Microbes Infect.* **7**, 154 (2018).