

1 Title: Exceptional diversity and selection pressure on SARS-CoV and SARS-CoV-2 host
2 receptor in bats compared to other mammals

3

4 Authors: Hannah K. Frank¹, David Enard² and Scott D. Boyd¹

5

6 Affiliations:

7 1. Department of Pathology, Stanford University School of Medicine, Stanford, CA

8 2. Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ

9 **Abstract:**

10 Pandemics originating from pathogen transmission between animals and humans
11 highlight the broader need to understand how natural hosts have evolved in response to
12 emerging human pathogens and which groups may be susceptible to infection. Here, we
13 investigate angiotensin-converting enzyme 2 (ACE2), the host protein bound by SARS-CoV and
14 SARS-CoV-2. We find that the ACE2 gene is under strong selection pressure in bats, the group
15 in which the progenitors of SARS-CoV and SARS-CoV-2 are hypothesized to have evolved,
16 particularly in residues that contact SARS-CoV and SARS-CoV-2. We detect positive selection
17 in non-bat mammals in ACE2 but in a smaller proportion of branches than in bats, without
18 enrichment of selection in residues that contact SARS-CoV or SARS-CoV-2. Additionally, we
19 evaluate similarity between humans and other species in residues that contact SARS-CoV or
20 SARS-CoV-2, revealing potential susceptible species but also highlighting the difficulties of
21 predicting spillover events. This work increases our understanding of the relationship between
22 mammals, particularly bats, and coronaviruses, and provides data that can be used in functional
23 studies of how host proteins are bound by SARS-CoV and SARS-CoV-2 strains.

24

25 **Main:**

26 The recent coronavirus pandemic has highlighted the disastrous impacts of zoonotic
27 spillovers and underscores the need to understand how pathogens and hosts evolve in
28 response to one another. Evolutionary analyses of host proteins targeted by infections reveal
29 the pressures that hosts have faced from pathogens and how they have evolved to resist
30 disease, informing predictions about spread of infections and how to counter them. The virus,
31 SARS-CoV-2, the causative virus of COVID-19, like its close relative SARS-CoV, is thought to
32 have its progenitor origins in bats¹⁻³. Bats have been suggested to be “special” reservoirs of
33 emerging infectious viruses⁴ and of coronaviruses in particular⁵. However, often this species-
34 rich, ecologically diverse clade is treated as a homogenous group, represented by one or two

35 species, particularly when considering the interaction of SARS-CoV and SARS-CoV-2 with host
36 proteins (but see Hou et al.⁶ and Demogines et al.⁷ which consider multiple bat species).

37 Examination of host proteins bound by potential zoonoses can be used not only to infer
38 past and current evolutionary pressure but to inform the likelihood of cross-species
39 transmission. One major barrier to cross-species transmission is the ability of the virus, adapted
40 to one host protein, to bind another species' protein^{8,9}. Accordingly, many studies have
41 examined the ACE2 sequence of suspected disease reservoirs to understand how different viral
42 strains bind different species' ACE2 and where zoonotic spillover may have originated^{6,8,10-12}.
43 These studies, especially ones involving functional assays or in-depth modeling of virus-host
44 contacts, are usually limited in their comparisons to a small subset of domestic animals or
45 suspected reservoirs, e.g. rhinolophid bats or civets – a suspected intermediate host for SARS-
46 CoV¹³. Often similarity, or lack thereof, between humans and other species in key ACE2
47 residues are used to predict the species that may have transmitted viruses to humans but
48 because studies only examine a small subset of the existing diversity, it is hard to determine
49 whether the selected species are more or less similar to humans than a random sample of
50 animals.

51 Here, we investigate how angiotensin converting enzyme 2 (ACE2), the host protein
52 bound by SARS-CoV and SARS-CoV-2^{3,14,15}, has evolved in bats compared to other mammals.
53 We analyze sequences drawn from 90 bat species, including 55 sequences generated for this
54 study, an eight-fold increase over prior studies, and 108 other mammal species. Finally, we use
55 our dataset of ACE2 sequences to highlight the potential for transmission of COVID-19 from
56 humans into wildlife and the difficulty of predicting intermediate and amplifying hosts in spillover
57 based on receptor similarity alone.

58

59

60 Results and Discussion

61 *Mammals, particularly bats, are diverse at ACE2 contact residues for SARS-CoV and SARS-*
62 *CoV-2*

63 We analyzed a total of 207 ACE2 sequences from 198 species (90 bat species; 108
64 non-bat species) representing 18 mammalian orders (Table S1). There are 24 amino acid sites
65 on ACE2 that are important for stabilizing the binding of ACE2 with the receptor binding domain
66 of SARS-CoV (22 sites; Table S2) and/or SARS-CoV-2 (21 sites; Table S2)^{6,8,10,11,14–16}. Across
67 these 24 sites, which we refer to by their position in the human ACE2, we found a minimum of
68 132 unique amino acid combinations in the 207 sequences; across a subset of 7 amino acids
69 identified as virus-contacting residues or important for the maintenance of salt bridges by most
70 studies^{6,8,15,16} we found a minimum of 82 unique amino acid sequences (Figure 1). In bats
71 (N=96), we found a minimum of 64 unique amino acid sequences across the 24 amino acids
72 and 49 across the 7 amino acids, while in other mammals (N=111), we found a minimum of 68
73 unique amino acid sequences across the 24 amino acids and 38 across the 7 amino acids.
74 Within species for which we were able to observe multiple individuals, we observed differing
75 levels of diversity in the 24 sites with *Bos indicus*, *Rousettus leschenaultii*, *Camelus*
76 *dromedarius* and *Rhinolophus ferrumequinum* identical within species across individuals and
77 sites; *Canis lupus* showed one amino acid difference across two individuals; in contrast all four
78 individuals of *Rhinolophus sinicus* were different from one another.

79 Across all sequences, the 24 amino acid sites varied from monomorphic across all
80 examined sequences (e.g. Phe²⁸ and Arg³⁵⁷) to having 10 or more possible amino acids (e.g.
81 24, 27, 31, 34, 79, 82, 329). The most diverse sites (as measured by Shannon's diversity index)
82 were amino acid positions 24, 34, 79, 82, 329 and 354, while the most even sites were positions
83 24, 30, 34, 41, 82 and 329 (Table S2). Of the 22 sites with more than one amino acid, bats were
84 more diverse than other mammals at 13 and were more even at 15. That bats demonstrate a
85 similar diversity in their ACE2 across these loci and greater diversity in some sites than that

86 observed across the rest of mammals suggests they may be particularly diverse in their ACE2,
87 and supports the idea that bats are more diverse than other suspected SARS-CoV and SARS-
88 like CoV hosts⁶.

89

90 *Bats drive the signal of mammalian selection and adaptation to SARS-CoV and SARS-CoV-2*

91 We also conducted a series of selection analyses each on 5 phylogenetic trees drawn
92 from Upham et al. ¹⁷. Across all mammals, the 20 variable sites in ACE2 that contact SARS-
93 CoV were not more likely to be under positive selection than other residues in the gene (MEME
94 $p < 0.05$, Fisher's exact test, $p_{\text{all trees}} > 0.35$; Table S3), though there was some marginal
95 evidence for increased selection in these residues (MEME $p < 0.1$, Fisher's exact test, $p_{\text{two trees}} <$
96 0.06 ; Table S3). Similarly, the 19 variable sites that contact SARS-CoV-2 were not more likely to
97 be under positive selection than other residues in the gene when considering sites under
98 selection at $p < 0.05$ (MEME $p < 0.05$, Fisher's exact test, $p_{\text{all trees}} > 0.1$; Table S3). However,
99 when considering sites under selection at $p < 0.1$, residues that contact SARS-CoV-2 do indeed
100 appear to be more likely to be under selection than other residues in the gene, likely due to the
101 reduction in statistical power loss (MEME $p < 0.1$, Fisher's exact test, $p_{\text{all trees}} < 0.02$). Therefore
102 there is some evidence that the locus is evolving in response to coronaviruses; this is similar to
103 the finding of strong selection in aminopeptidase N (ANPEP) in response to coronaviruses in
104 mammals¹⁸. However, this pattern is driven by and strengthened in bats; in bats a greater
105 proportion of residues that contact SARS-CoV (MEME $p < 0.05$, Fisher's exact test, $p_{\text{all trees}} <$
106 0.03 ; MEME $p < 0.1$, Fisher's exact test, $p_{\text{all trees}} < 0.02$; Table S3) and SARS-CoV-2 (MEME $p <$
107 0.05 , Fisher's exact test, $p_{\text{all trees}} < 0.02$; MEME $p < 0.1$, Fisher's exact test, $p_{\text{all trees}} < 0.0004$;
108 Table S3) were under selection than other residues in the gene, whereas residues that contact
109 SARS-CoV (MEME $p < 0.05$, Fisher's exact test, $p_{\text{all trees}} > 0.4$; MEME $p < 0.1$, Fisher's exact test,
110 $p_{\text{all trees}} > 0.5$; Table S3) or SARS-CoV-2 (MEME $p < 0.05$, Fisher's exact test, $p_{\text{all trees}} > 0.4$;
111 MEME $p < 0.1$, Fisher's exact test, $p_{\text{all trees}} > 0.5$; Table S3) were not more likely to be under

112 selection in non-bat mammals. Increased sampling can improve the ability of MEME to detect
113 selection at individual sites¹⁹. Because our dataset of bat sequences is smaller than our
114 mammalian dataset, it further strengthens our conclusion that bats are under positive selection
115 in contact residues. Across all mammals, positions 24 and 42 were under selection (Table S2; 5
116 trees, MEME, $p < 0.05$), but in bats positions 27, 31, 35 and 354 (Table S2, 5 trees, MEME, $p <$
117 0.05) and 30, 38, 329 and 393 (Table S2, 5 trees, MEME, $p < 0.1$) were additionally under
118 positive selection while positions 45 (Table S2, 5 trees, MEME, $p < 0.05$) and 353 (Table S2, 5
119 trees, MEME, $p < 0.1$) were under selection in non-bat mammals but not bats.

120 Using aBSREL we tested two *a priori* hypotheses, the first that bats are under positive
121 selection in ACE2 and the second that the family Rhinolophidae, the bat family in which the
122 progenitors of SARS-CoV and SARS-CoV-2 are hypothesized to have evolved^{3,20}, specifically,
123 is under positive selection. Both bats ($p_{\text{all trees}} < 0.002$) and Rhinolophidae ($p_{\text{all trees}} < 0.00007$) are
124 under positive selection in ACE2 (Table S4). When we conducted an adaptive branch-site test
125 of positive selection on all branches without specifying a foreground branch, branches in the bat
126 clade were more likely to be selected than branches in other parts of the mammalian phylogeny
127 (Fisher's exact test, $p_{\text{all trees}} < 0.05$; Table S4) and the branch at the base of Rhinolophidae was
128 under positive selection ($p_{\text{all trees, Holm-Bonferroni correction}} < 0.04$, Table S4). We found that bat
129 branches are more likely to be under positive selection than other branches despite the fact that
130 these branches are a subset of the total phylogeny and branch-site tests of positive selection
131 have reduced power to detect selection on shorter branches, making our test conservative^{18,21}.
132 It is possible that the sequences we generated through target capture and genomic sequence
133 were of poorer sequencing quality than the reference genomes (though the number of residues
134 covered by sequences we generated and publicly available reference sequences was similar;
135 two-tailed t-test, $t = -0.49$, $p = 0.63$). When we removed the bat sequences we generated and
136 examined the remaining terminal branches, a greater proportion of bat branches were under
137 selection than non-bat branches, but statistical significance was lost, likely due to reduced

138 power (Table S4). Increased positive selection in bats in ACE2 compared with other mammals
139 is consistent with their status as rich hosts of coronaviruses⁵. Host diversity of bats in a region is
140 associated with higher richness of coronaviruses⁵; the diversity of bat ACE2 is consistent with
141 the idea that a diversity of bats and their ACE2 sequences are coevolving with a diversity of
142 viruses.

143 Two bat families, Rhinolophidae and Hipposideridae, have been associated with SARS-
144 related betacoronaviruses⁵, which use the ACE2 molecule as a viral receptor⁹. Interestingly,
145 while we found evidence that Rhinolophidae are under selection in ACE2, we found widespread
146 selection across bats. Branches in the rhinolophid/ hipposiderid clade were not more likely to be
147 under selection than other branches within bats (Fisher's exact test, $p_{\text{all trees}} > 0.7$; Table S4) and
148 bat lineages that live outside the predicted range of these viruses (e.g. in the Americas⁵) are
149 also under positive selection. Therefore, there are still aspects of the bat-coronavirus
150 relationship that we do not fully understand. At least one other coronavirus uses ACE2 to gain
151 entry into the host cell, HCoV-NL63, which may have its origin in bats²²; we found some
152 evidence for increased selection in the residues that contact this virus in bats (MEME $p < 0.05$,
153 Fisher's exact test, $p_{\text{all trees}} < 0.07$; MEME $p < 0.1$, Fisher's exact test, $p_{\text{all trees}} < 0.08$; Table S3),
154 but not in non-bat mammals (MEME $p < 0.05$, Fisher's exact test, $p_{\text{all trees}} > 0.6$; MEME $p < 0.1$,
155 Fisher's exact test, $p_{\text{all trees}} > 0.4$; Table S3). Many ACE2 residues that interact with HCoV-NL63
156 also interact with one or both of SARS-CoV and SARS-CoV-2^{23,24}, which may be driving the
157 evidence of selection in these residues. However, we did find evidence of selection in residues
158 321 and 326 in both bats and non-bat mammals (Table S2, 5 trees, MEME, $p < 0.05$), as well as
159 selection in bats in residue 322 (Table S2, 5 trees, MEME, $p < 0.05$); these three residues
160 contact HCoV-NL63 but not SARS-CoV or SARS-CoV-2. Our finding of selection in residues
161 that contact HCoV-NL63 but not SARS-CoV or SARS-CoV-2 contrasts with the findings of a
162 smaller dataset of bats mostly from Europe, Asia and Africa⁷ and may result from our greater
163 power to detect signal or signal originating from bats in different regions than previously tested.

164 ACE2 regulation is known to impact survival in some influenza A infections²⁵; New World bats
165 are known to host many influenza A viruses²⁶, so it is possible the selection we detect could
166 result from infection from non-coronavirus infections. Still, our results raise questions about
167 whether there are or were SARS-related coronaviruses in these regions that have not been
168 detected?

169

170 *Similarity of ACE2 yields predictions of susceptible hosts but cannot completely determine host*
171 *range of SARS-CoV and SARS-CoV-2*

172 To determine how similar bats, civet, pangolin (a suspected source of SARS-CoV-2²⁷)
173 and other groups are to humans in 24 ACE2 residues that bind SARS-CoV (22 residues) and/or
174 SARS-CoV-2 (21 residues), in all 207 sequences we quantified how many of the residues were
175 identical or very similar to humans, likely maintaining current binding properties, versus how
176 many were likely to disrupt binding. All of the apes and most of the Old World monkeys we
177 examined were identical to humans across all amino acid sites; those that were not identical
178 differed by only 1 or 2 amino acids (Figure 2; Table S1). However, amino acid similarity in these
179 sites across different species often diverged from what we would have predicted from phylogeny
180 alone. Notably, two rodents (*Mesocricetus auratus* and *Peromyscus leucopus*) had identical or
181 very similar amino acids to humans in all but 2 sites for each virus, and many artiodactyls (e.g.
182 cows, deer, sheep, goats), cetaceans, cats, and pangolin were as similar or more similar to
183 humans than New World monkeys both in residues that contact SARS-CoV and in residues that
184 contact SARS-CoV-2. The civet fell in the middle of mammals in its similarity to humans in
185 residues that contact either or both viruses. In general, bats were not very similar to humans at
186 these 24 amino acid sites, some with as many as five changes that would likely reduce virus
187 binding, the most observed across mammals. Additionally, most bat sequences (56 of 91)
188 showed that at least one of the two salt bridges (Lys³¹-Glu³⁵; Asp³⁸-Lys³⁵³ in humans) within
189 ACE2 would be disrupted by changing a charged amino acid to an uncharged amino acid or to

190 an amino acid with a clashing charge (Table S1). In Rhinolophidae, only one sequence of the
191 ten examined did not have a change in position 31 or 35 that would result in a clash between
192 two positively charged amino acids. Because of the large overlap in residues that contact
193 SARS-CoV and SARS-CoV-2 (19 residues) generally species were roughly as similar to
194 humans in residues that contact SARS-CoV and in residues that contact SARS-CoV-2 (Figure
195 2). However, bats (two-sided Wilcoxon signed rank test, $p < 0.0001$) and carnivores (two-sided
196 Wilcoxon signed rank test, $p < 0.0004$), particularly mustelids including ferrets, were more
197 similar to humans in residues that contact SARS-CoV-2 than residues that contact SARS-CoV
198 (Figure 2).

199 Examination of the diversity of ACE2 sequences across mammals and the similarity
200 between distantly related groups at key residues for interaction with SARS-CoV and SARS-
201 CoV-2 allows one to make predictions about potential spillover hosts or other susceptible
202 species. In some cases, similarity of host residues seems to predict infection ability well. Old
203 World primates were identical to humans across all 24 residues and, consistent with the idea
204 that identical residues would confer susceptibility, experimental infections have demonstrated
205 that SARS-CoV replicates in multiple macaque species²⁸. Additionally, in our analysis, domestic
206 cats were among the species most similar humans in residues that contact SARS-CoV and
207 SARS-CoV-2. Notably, cats can become infected and can shed both SARS-CoV and SARS-
208 CoV-2^{29,30}. Pangolins were as similar in their ACE2 residues to humans as cats, lending some
209 support for the idea that a virus that can bind pangolin ACE2 might be able to transmit to
210 humans. Accordingly, it seems prudent to exercise precautions when interacting with species
211 whose ACE2 is similar to humans in the contact residues for SARS-CoV and SARS-CoV-2,
212 especially domestic animals such as cats, cows, goats and sheep. Care should also be taken
213 with wild animals; for example, interactions between people with macaques or visitation of
214 mountain gorillas by tourists could lead to cross-species transmission, endangering the health of
215 humans and wildlife.

216 In other cases, it can be hard to predict susceptibility to SARS-CoV or SARS-CoV-2
217 infection based on overall similarity of ACE2 residues. A single amino acid change can impact
218 the binding of a virus to ACE2. In position 24, a diverse, even and selected site across all
219 mammals that contacts both viruses, mutation from Gln (human) to Lys (16 bat species and
220 *Rattus norvegicus*) reduced association between the SARS-CoV spike protein and ACE2¹².
221 Position 27, a selected site in bats with many amino acid variants, is a Thr in humans; when
222 mutated to a Lys (as in some bats), the interaction disfavored SARS-CoV binding by disrupting
223 hydrophobic interactions with the SARS-CoV virus, while mutation to Ile, found in other bats,
224 increased the ability of the virus to infect cells⁶. Some rodents, including *Mesocricetus auratus*
225 and *Peromyscus leucopus*, which were among the most similar species to humans, have a
226 glycosylated Asn in position 82 that disrupts the hydrophobic contact with Leu⁴⁷² on SARS-CoV,
227 reducing association between the spike protein and ACE2^{12,15}; in conjunction with other
228 mutations this glycosylation can disrupt binding¹¹. We predict this same glycosylated Asn is also
229 present in some rhinolophid bats (*R. ferrumequinum* and some *R. sinicus*), though not all (*R.*
230 *pusillus*, *R. macrotis* and some *R. sinicus*). Additionally, intraspecies variation could be an
231 important component of reservoir competency that we are unable to assess. It is likely that not
232 all individuals in a species are equally susceptible to infection, complicating the identification of
233 reservoirs.

234 Additionally, spillover potential is not regulated solely by ACE2 sequence and sometimes
235 SARS-CoV or SARS-CoV-2 are able to replicate in hosts with divergent ACE2. Compatibility of
236 the host protease with the virus is important for determining host range⁹ and viral strains vary in
237 their binding properties to different species¹⁵, with some SARS-like coronaviruses able to bind
238 human, civet and rhinolophid ACE2, despite major ACE2 sequence differences between the
239 species²⁰. Additionally, SARS-CoV can enter cells expressing the ACE2 of *Myotis daubentonii*
240 and *Rhinolophus sinicus* with limited efficiency⁶, even though these species only share 13-16
241 amino acids with humans that contact either virus and each contain some mutations that should

242 interfere with binding. Similarly, both SARS-CoV and SARS-CoV-2 replicated well in ferrets,
243 whose ACE2 ranked among the least similar to humans in their contact residues for SARS-CoV,
244 though more similar for SARS-CoV-2^{29,30}. And species whose ACE2 sequence is not very
245 similar to humans can be experimentally infected with SARS-CoV^{11,31}.

246

247 *Conclusions*

248 Taken together, our data suggest that mammals, particularly bats, are evolving in
249 response to coronaviruses with a diverse suite of ACE2 sequences that likely confer differing
250 susceptibility to various coronavirus strains. Predicting which species will expose humans to
251 potential zoonoses is difficult. Data about viruses circulating in wildlife can help trace the origins
252 of zoonotic disease outbreaks but are of minimal use if people continue to expose themselves to
253 the reservoirs. Growing evidence suggests that some viruses are capable of evolving to infect
254 different hosts, even when there might be sizable barriers such as divergent host receptors. The
255 best solution is to prevent people and domestic animals from contacting wildlife to minimize
256 opportunities for disease transmission and host switching.

257

258 **Acknowledgments:** We thank Ellie Armstrong for assistance with bioinformatic analysis and
259 Sandra Nielsen for insightful comments. We thank CONAGEBIO and the Organization for
260 Tropical Studies for assistance and access to Costa Rican genetic resources. We thank the
261 following museums for grants of tissue: Field Museum, Museum of Southwestern Biology,
262 University of Alaska Museum, Museum of Vertebrate Zoology, University of Kansas Museum,
263 and Texas Tech Museum. Additionally, we are grateful to the following organizations for funding
264 the work: National Science Foundation Doctoral Dissertation Improvement Grant (1404521;
265 HKF), Life Sciences Research Foundation Fellowship (HKF), Open Philanthropy Project,
266 Stanford Woods Environmental Venture Program, Bing-Mooney Fellowship in Environmental
267 Science, Stanford Center for Computational, Evolutionary and Human Genomics Postdoctoral

268 Fellowship (HKF), National Institutes of Health grants “Molecular and Cellular Immunobiology”
269 (5 T32 AI07290), Stanford School of Medicine Dean’s Postdoctoral Fellowship (HKF), and
270 endowment from the Crown Family foundation (SDB).

271

272 **Author contributions:** HKF, DE and SB planned the project. HKF collected and analyzed the
273 data; DE contributed bioinformatic tools. HKF wrote the manuscript; all authors edited the
274 manuscript.

275

276 **Data availability:** The DNA sequences generated in this study are available from GenBank with
277 the primary accession codes MT333480-MT333534.

278

279 **Code availability:** No custom code was created for this analysis.

280

281 **Experimental methods:**

282 *Alignment of mammalian ACE2 sequences:*

283 Sequences for ACE2 were obtained either through Genbank or, in the case of several
284 bat species, sequenced for this study. On 21 February 2020, ACE2 orthologs for all jawed
285 vertebrates were downloaded from Genbank³². In addition, we sought out all bat sequences of
286 ACE2, adding sequences from Hou et al.⁶, as well as the palm civet ACE2 sequences because
287 of their putative role as reservoirs of SARS-CoV. Only sequences from mammalian species
288 found in Upham et al.¹⁷ were retained and are listed in Table S1. Two Costa Rican bat species
289 (*Artibeus [Dermanura] watsoni* and *Artibeus [Dermanura] phaeotis*) were not included in the
290 phylogenetic hypotheses¹⁷ and were therefore excluded from the molecular evolution analyses
291 but were included in analyses of amino acid identity and diversity. Multiple individuals were
292 available for a few species (four individuals of *Rhinolophus sinicus*; three individuals of
293 *Rhinolophus ferrumequinum*; and two each for: *Bos indicus*, *Camelus dromedarius*, *Canis lupus*

294 and *Rousettus leschenaultii*), however only one sequence per species was used in molecular
295 evolution analyses, noted in Table S1.

296 We sought additional data on the diversity of the ACE2 gene across bats using a
297 combination of samples collected in the field in Costa Rica and granted from museums (63
298 species; summarized in table S1). For samples collected in the field, bats were captured in mist
299 nets and a wing biopsy sample was collected. Bats were released immediately after sampling.
300 Research was approved by the Stanford Institutional Animal Care and Use Committee
301 (protocols 26920 and 29978) and conducted under the appropriate Costa Rican permits. From
302 these species, we isolated DNA from tissue using the Qiagen DNeasy Blood and Tissue kit
303 (Valencia, CA, USA) and created genomic libraries using the NEBNext Ultra II kit (New England
304 BioLabs; Ipswich, MA, USA), according to manufacturer's instructions. For some species, ACE2
305 was captured as part of a targeted capture using genomic libraries and a custom target
306 enrichment kit (Arbor Biosciences; Ann Arbor, MI, USA) according to the manufacturer's
307 instructions with modifications³³, while in other cases ACE2 was isolated from total genomic
308 sequence. Briefly, genomic reads were mapped to the nearest bat genome of *Rousettus*
309 *aegyptiacus*, *Pteropus alecto*, *Desmodus rotundus*, *Myotis lucifugus* or *Eptesicus fuscus* using
310 LAST³⁴ to generate a consensus sequence and the ACE2 coding regions were extracted using
311 a translated DNA search in BLAT³⁵ and the ACE2 coding sequence from *Myotis lucifugus*³⁶.
312 Sequences are available from Genbank (MT333480-MT333534; Table S1).

313 All sequences for ACE2 were aligned in Geneious³⁷. Sequences were corrected by hand
314 to remove sequences outside the coding region and adjust gaps to be in frame with the coding
315 region using the human mRNA as a guide. Missing sequence, gaps and premature stop codons
316 were converted to Ns for downstream analyses and comparison of residues with the human
317 coding region.

318

319

320 *Investigation of important residues for CoV binding*

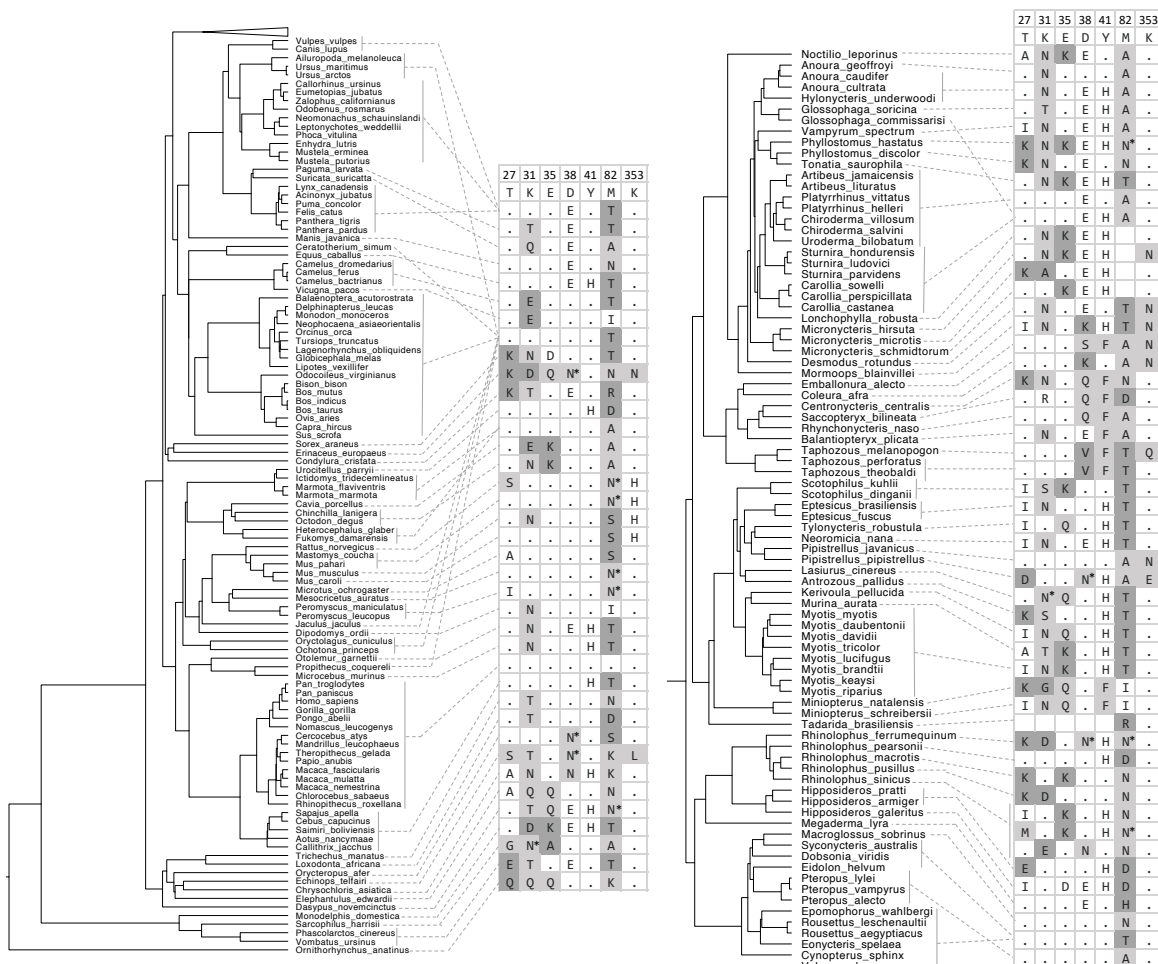
321 We sought to understand how the residues important for coronavirus binding are
322 conserved (or not) across mammals to determine probable host range of SARS-CoV and
323 SARS-CoV-2. We compared amino acid sequences across 24 positions known to be important
324 for binding of SARS-CoV and/or SARS-CoV-2 as determined by others^{6,8,10,11,14-16}. To determine
325 which amino acid positions were the most variable, we calculated Shannon's diversity index
326 (which accounts for the number and evenness of amino acids), number of unique amino acids
327 and evenness for each of the 24 amino positions using the vegan³⁸ package (version 2.5-6) in
328 R³⁹ (version 3.6.2). We also calculated how "human-like" a species was across these 24 amino
329 acids, as well as separately for residues contacting SARS-CoV and residues contacting SARS-
330 CoV-2 by giving a score to each amino acid in each position. Residues that were identical or
331 relatively equivalent to humans were given a score of 1; relative equivalency was inferred when
332 amino acids retained similar properties and abilities to participate in hydrogen bonds, Van der
333 Waals forces or salt bridges. Residues that would likely be worse at binding were given scores
334 of -1; reduced binding was inferred when amino acid properties were dramatically altered from
335 that of the human amino acid motif (e.g. replacement of a positively charged amino acid with a
336 negatively charged amino acid in a salt bridge). In general, asparagine and glutamine were
337 considered similar enough not to disrupt binding, as were amino acids with the same charge
338 and amino acids with small hydrophobic side chains (valine, leucine, isoleucine and methionine).
339 Amino acids whose effect was hard to determine were given scores of zero. Exact
340 determinations of the impact can be found in Table S2. The human-like score was calculated as
341 a sum of each amino acid score divided by the total amino acids observed across all 24 sites or
342 all sites that contacted a given virus (since some species had missing data). We predicted the
343 N-linked glycosylation of Asn when Asn was found in the following motif N-X-S/T where X is not
344 a proline⁴⁰. Glycosylation was not taken into account when calculating the human-like score.
345

346 *Molecular evolution analyses*

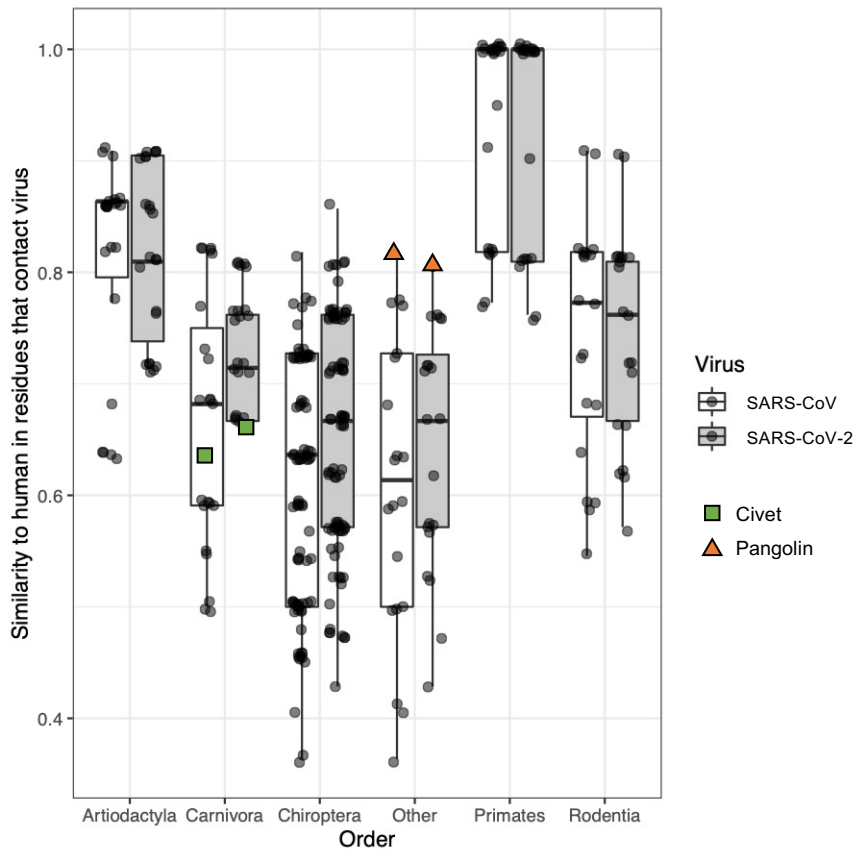
347 To determine whether it was likely to be interactions with coronavirus driving the
348 evolution of ACE2 we used MEME¹⁹ to infer the residues under selection across the mammal
349 phylogeny, in just bats and in non-bat mammals and used a Fisher's exact test to determine
350 whether residues that interact with SARS-CoV, SARS-CoV-2 or HCoV-NL63²³ were more likely
351 to be under selection than other residues in ACE2. Only codons that showed variation (e.g.
352 more than one amino acid across all 198 species) and that were present in humans were
353 considered in the Fisher's exact test. We used a $p < 0.05$ cutoff for inferring selection at each
354 site via MEME but some results were shaper when using a $p < 0.1$ cutoff, likely due to the
355 reduction in loss of statistical power (Table S3).

356 Additionally, to determine whether bats, and specifically the family Rhinolophidae, are
357 under strong selection to adapt to viruses we used the adaptive branch-site random effects
358 model test of positive selection, aBSREL²¹, as implemented in HyPhy, version 2.5.8⁴¹, using a
359 pruned subset of five phylogenetic hypotheses chosen from the pseudo-posterior distribution of
360 Upham et al.¹⁷ to account for phylogenetic uncertainty. We tested three conditions: 1) in which
361 the branch leading to Rhinolophidae was considered the foreground branch; 2) in which the
362 branch leading to the common ancestor of all bats was considered the foreground branch; and
363 3) in which all branches were tested without *a priori* specification of background and foreground
364 branches. In determining whether bats are more likely to be under selection than other
365 mammals, we used Fisher's exact tests to test whether an excess of branches in the bat lineage
366 were under selection compared to the rest of the phylogeny. We used $p < 0.05$ as our cutoff for
367 a branch being under selection without any Holm-Bonferroni correction because it seemed
368 unlikely that bat branches were more susceptible to false positives than any other branch and all
369 our comparisons were between branches within the same aBSREL analysis. If one only accepts
370 significance at a $p < 0.05$ with Holm-Bonferroni correction, a very stringent requirement, the
371 general trends remain but the results lose statistical significance (Table S4). As described in the

372 results, to guard against bias due to potentially lower sequence quality in the sequences we
373 generated, we repeated our Fisher's exact test using only terminal branches and removing
374 sequences we generated; the trend of a larger proportion of bat branches being under selection
375 was maintained but the results lose statistical significance (Table S4).
376



377
 378
 379 **Figure 1: ACE2 is diverse across mammalian phylogeny at 7 residues responsible for**
 380 **contact with SARS-CoV and SARS-CoV-2.** Dashed lines connect species to their ACE2
 381 sequence. Numbers at the top of the table indicate the amino acid position in the human ACE2;
 382 human residues are listed as the reference. A period indicates identity with the human amino
 383 acid; unshaded boxes are amino acids that are identical or similar to humans; light gray
 384 indicates no or unknown impact on the interaction of ACE2 and SARS-CoV/ SARS-CoV-2 and
 385 dark gray indicates an amino acid that would likely disrupt the virus-host interaction. Asterisks
 386 on N indicate predicted presence of N-linked glycosylation. Blank boxes indicate a lack of data
 387 for that species and residue. The bat clade is collapsed on the left (top) and expanded in the
 388 right column.



389

390 **Figure 2: Similarity of ACE2 residues contacting SARS-CoV or SARS-CoV-2 to human.**

391 Similarity of residues was calculated based on the number of residues that were identical or

392 highly similar in binding properties to those found in human ACE2 with penalties for residues

393 that would likely disrupt binding (see methods). Scores of 1 indicate residues that contact the

394 virus are identical (or highly similar to humans). Boxes cover the interquartile range with a line

395 indicating the median and whiskers extending to the largest value less than 1.5 times the

396 interquartile range. Each point indicates a single sequence; only sequences with data for at

397 least 21 of 24 residues are shown. Sequences are grouped by mammalian order; "other"

398 includes all orders with fewer than 4 sequences. The green squares in Carnivora indicate the

399 civet (*Paguma larvata*) and the orange triangles at the top of the "Other" distribution indicate the

400 pangolin (*Manis javanica*).

401

402 REFERENCES

403

404 1. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The Proximal
405 Origin of SARS-CoV-2. *Nat. Med.* (2020) doi:<https://doi.org/10.1038/s41591-020-0820-9>.

406 2. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus:
407 implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).

408 3. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat
409 origin. *Nature* **579**, 270–273 (2020).

410 4. Brook, C. E. & Dobson, A. P. Bats as ‘special’ reservoirs for emerging zoonotic
411 pathogens. *Trends Microbiol.* **23**, 172–180 (2015).

412 5. Anthony, S. J. *et al.* Global patterns in coronavirus diversity. *Virus Evol.* **3**, vex012 (2017).

413 6. Hou, Y. *et al.* Angiotensin-converting enzyme 2 (ACE2) proteins of different bat species
414 confer variable susceptibility to SARS-CoV entry. *Arch. Virol.* **155**, 1563–1569 (2010).

415 7. Demogines, A., Farzan, M. & Sawyer, S. L. Evidence for ACE2-Utilizing Coronaviruses
416 (CoVs) Related to Severe Acute Respiratory Syndrome CoV in Bats. *J. Virol.* (2012)
417 doi:10.1128/jvi.00311-12.

418 8. Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. Receptor recognition by novel
419 coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS.
420 *J. Virol.* (2020) doi:10.1128/jvi.00127-20.

421 9. Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor
422 usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol.* **5**, 562–
423 569 (2020).

424 10. Liu, Z. *et al.* Composition and divergence of coronavirus spike proteins and host ACE2
425 receptors predict potential intermediate hosts of SARS-CoV-2. *J. Med. Virol.* 1–7 (2020)
426 doi:10.1002/jmv.25726.

427 11. Lu, G., Wang, Q. & Gao, G. F. Bat-to-human: Spike features determining ‘host jump’ of
428 coronaviruses SARS-CoV, MERS-CoV, and beyond. *Trends Microbiol.* **23**, 468–478
429 (2015).

430 12. Li, W. *et al.* Receptor and viral determinants of SARS-coronavirus adaptation to human
431 ACE2. *EMBO J.* **24**, 1634–1643 (2005).

432 13. Guan, Y. *et al.* Isolation and characterization of viruses related to the SARS coronavirus
433 from animals in Southern China. *Science (80-)*. **302**, 276–8 (2003).

434 14. Yan, R. *et al.* Structural basis for the recognition of the SARS-CoV-2 by full-length human
435 ACE2. *Science* (2020) doi:10.1126/science.abb2762.

436 15. Li, F., Li, W., Farzan, M. & Harrison, S. C. Structure of SARS coronavirus spike receptor-
437 binding domain complexed with receptor. *Science (80-)*. **309**, 1864–1868 (2005).

438 16. Lan, J. *et al.* Crystal structure of the 2019-nCoV spike receptor-binding domain bound
439 with the ACE2 receptor. *bioRxiv* (2020) doi:10.1101/2020.02.19.956235.

440 17. Upham, N. S., Esselstyn, J. A. & Jetz, W. Inferring the mammal tree: Species-level sets
441 of phylogenies for questions in ecology, evolution, and conservation. *PLoS Biol.* **17**,
442 e3000494 (2019).

443 18. Enard, D., Cai, L., Gwennap, C. & Petrov, D. A. Viruses are a dominant driver of protein

- 444 adaptation in mammals. *Elife* **5**, e12469 (2016).
- 445 19. Murrell, B. *et al.* Detecting individual sites subject to episodic diversifying selection. *PLoS*
446 *Genet.* **8**, e1002764 (2012).
- 447 20. Ge, X. Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses
448 the ACE2 receptor. *Nature* **503**, 535–538 (2013).
- 449 21. Smith, M. D. *et al.* Less is more: An adaptive branch-site random effects model for
450 efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* **32**, 1342–1353
451 (2015).
- 452 22. Banerjee, A., Kulcsar, K., Misra, V., Frieman, M. & Mossman, K. Bats and coronaviruses.
453 *Viruses* **11**, 41 (2019).
- 454 23. Wu, K., Li, W., Peng, G. & Li, F. Crystal structure of NL63 respiratory coronavirus
455 receptor-binding domain complexed with its human receptor. *Proc. Natl. Acad. Sci. U. S.*
456 *A.* **106**, 19970–19974 (2009).
- 457 24. Li, W. *et al.* The S proteins of human coronavirus NL63 and severe acute respiratory
458 syndrome coronavirus bind overlapping regions of ACE2. *Virology* **367**, 367–374 (2007).
- 459 25. Zou, Z. *et al.* Angiotensin-converting enzyme 2 protects from lethal avian influenza A
460 H5N1 infections. *Nat. Commun.* **5**, 3594 (2014).
- 461 26. Tong, S. *et al.* New World Bats Harbor Diverse Influenza A Viruses. *PLoS Pathog.* **9**,
462 e1003657 (2013).
- 463 27. Zhang, T., Wu, Q. & Zhang, Z. Probable Pangolin Origin of SARS-CoV-2 Associated with
464 the COVID-19 Outbreak. *Curr. Biol.* **30**, 1346–1351 (2020).
- 465 28. McAuliffe, J. *et al.* Replication of SARS coronavirus administered into the respiratory tract
466 of African Green, rhesus and cynomolgus monkeys. *Virology* **330**, 8–15 (2004).
- 467 29. Shi, J. *et al.* Susceptibility of ferrets, cats, dogs, and different domestic animals to SARS-
468 coronavirus-2. *bioRxiv* doi:<https://doi.org/10.1101/2020.03.30.015347>.
- 469 30. Martina, B. E. E. *et al.* SARS virus infection of cats and ferrets. *Nature* **425**, 915 (2003).
- 470 31. Shi, Z. & Hu, Z. A review of studies on animal reservoirs of the SARS coronavirus. *Virus*
471 *Res.* **133**, 74–87 (2008).
- 472 32. ACE2 - angiotensin I converting enzyme 2. *Bethesda (MD): National Library of Medicine*
473 *(US), National Center for Biotechnology Information*
474 <https://www.ncbi.nlm.nih.gov/gene/59272/ortholog/?scope=7776> (2004).
- 475 33. Portik, D. M., Smith, L. L. & Bi, K. An evaluation of transcriptome-based exon capture for
476 frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order:
477 Anura). *Mol. Ecol. Resour.* **16**, 1069–1083 (2016).
- 478 34. Kielbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame
479 genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).
- 480 35. Kent, W. J. BLAT - The BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
- 481 36. Smedley, D. *et al.* The BioMart community portal: An innovative alternative to large,
482 centralized data repositories. *Nucleic Acids Res.* **43**, W589–W598 (2015).
- 483 37. Kears, M. *et al.* Geneious Basic: an integrated and extendable desktop software
484 platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–9
485 (2012).

- 486 38. Oksanen, J. *et al.* vegan: Community Ecology Package. (2013).
487 39. R Core Team. R: A language and environment for statistical computing. (2019).
488 40. UniProt Consortium, . Glycosylation. <https://www.uniprot.org/help/carbohydr> (2018).
489 41. Kosakovsky Pond, S. L. *et al.* HyPhy 2.5 - A Customizable Platform for Evolutionary
490 Hypothesis Testing Using Phylogenies. *Mol. Biol. Evol.* **37**, 295–299 (2020).
491

492 Supplementary tables: Available as excel spreadsheets

493

494 **Table S1:** Data for each sequence on accession number, whether sequence is in selection
495 analyses, preservation of salt bridges, identity of residues contacting SARS-CoV or SARS-CoV-
496 2, combination of residues, and scores for similarity to humans in residues contacting SARS-
497 CoV and SARS-CoV-2

498

499 **Table S2:** Summary data about residues that contact SARS-CoV, SARS-CoV-2 and HCoV-
500 NL63 including diversity metrics, number of trees in which residues are inferred to be under
501 selection, which virus is contacted by each residue and the identity of amino acids that lead to a
502 positive or negative score in terms of similarity to humans.

503

504 **Table S3:** Results of Fisher's exact tests on the number of selected and non-selected residues
505 contacting SARS-CoV, SARS-CoV-2 and HCoV-NL63 as determined by MEME at $p < 0.05$ and
506 $p < 0.1$. Results are summarized for tests on all 5 phylogenetic trees in all mammals, just bats
507 or just non-bat mammals.

508

509 **Table S4:** Results of aBSREL analyses. Includes p-values from analyses in which the branch at
510 the base of all bats is specified as the foreground branch, the branch at the base of
511 Rhinolophidae is specified as the foreground branch or no foreground branches were specified.
512 Also includes results of Fisher's exact tests on the number of branches under selection in the
513 bat clade compared to other branches; number of branches in the rhinolophid/ hipposiderid
514 clade under selection compared to other branches in the bat clade; number of terminal bat
515 branches (including or excluding sequences generated as part of the study) under selection
516 compared to terminal branches in the rest of the phylogeny. Results using uncorrected p-values
517 and Holm-Bonferroni corrected p-values are reported.