The evolutionary history of ACE2 usage within the coronavirus subgenus *Sarbecovirus*

Wells, H.L[1,2]*; Letko, M[3,4]; Lasso, G[5]; Ssebide, B[6]; Nziza, J[6]; Byarugaba, D.K[7,8]; Navarrete-Macias[1], I; Liang, E[1]; Cranfield, M[9,10]; Han, B.A[11]; Tingley, M.W[12]; Diuk-Wasser, M[2]; Goldstein, T[9]; Johnson, C.K[9]; Mazet, J[9]; Chandran, K[5]; Munster, V.J[3]; Gilardi, K[6,9]; Anthony, S.J[1,2,13]*

1. Center for Infection and Immunity, Mailman School of Public Health, Columbia University, New York, NY, USA
2. Department of Ecology, Evolution, and Environmental Biology, Columbia University, New York, NY, USA
3. Laboratory of Virology, Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, MT, USA
4. Paul G. Allen School for Global Animal Health, Washington State University, Pullman, WA, USA
5. Department of Microbiology and Immunology, Albert Einstein College of Medicine, New York, NY 10461, USA
6. Gorilla Doctors, c/o MGVP, Inc., Davis, California, USA
7. Makerere University Walter Reed Project, Kampala, Uganda
8. Makerere University, College of Veterinary Medicine, Kampala, Uganda
9. One Health Institute and Karen C. Drayer Wildlife Health Center, School of Veterinary Medicine, University of California, Davis, California, USA
10. Department of Microbiology and Immunology, University of North Carolina School of Medicine, Chapel Hill, North Carolina, USA
11. Cary Institute of Ecosystem Studies, Millbrook, New York, USA
12. Department of Ecology and Evolutionary Biology, University of California Los Angeles, Los Angeles, CA, USA
13. Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA

* Co-corresponding authors. Email: hlw2124@cumc.columbia.edu, sja2127@cumc.columbia.edu

Abstract

SARS-CoV-1 and SARS-CoV-2 are not phylogenetically closely related; however, both use the ACE2 receptor in humans for cell entry. This is not a universal sarbecovirus trait; for example, many known sarbecoviruses related to SARS-CoV-1 have two deletions in the receptor binding domain of the spike protein that render them incapable of using human ACE2. Here, we report three novel sarbecoviruses from Rwanda and Uganda which are phylogenetically intermediate to SARS-CoV-1 and SARS-CoV-2 and demonstrate via in vitro studies that they are also unable to utilize human ACE2. Furthermore, we show that the observed pattern of ACE2 usage among sarbecoviruses is most likely due to recombination. We show that the lineage that includes SARS-CoV-2 is most likely the ancestral ACE2-using lineage, and that recombination with at least one virus from this group conferred ACE2 usage to the progenitor of SARS-CoV-1 at some time in the past. We argue that alternative scenarios such as convergent evolution are much less parsimonious; we show that biogeography and patterns of host tropism support the plausibility of a recombination scenario; and we propose a competitive release hypothesis to explain how this recombination event could have occurred and why it is evolutionarily advantageous. The findings provide important insights into the natural history of ACE2 usage for both SARS-CoV-1 and SARS-CoV-2, and a greater understanding of the evolutionary mechanisms that shape zoonotic potential of coronaviruses. This study also underscores the need for increased surveillance for sarbecoviruses in southwestern China, where most ACE2-using viruses have been found to date, as well as other regions including Africa, where these viruses have only recently been discovered.

47 **Introduction**

48 The recent emergence of *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2) in China and
49 its rapid spread around the globe demonstrates that coronaviruses (CoVs) from wildlife remain an urgent
50 threat to global public health and economic stability. In particular, coronaviruses from the subgenus
51 *Sarbecovirus* (which includes SARS-CoV-2, SARS-CoV-1, numerous bat viruses, and a small number of
52 pangolin viruses) [1] are considered to be a high-risk group for potential emergence. As both
53 sarbecoviruses that have caused human disease (SARS-CoV-1 and -2) use angiotensin-converting enzyme
54 2 (ACE2) as their cellular receptor [2,3], the evolution of this trait is of particular importance for
55 understanding the emergence pathway for sarbecoviruses. Bat SARS-like coronavirus Rp3 (GenBank
56 accession DQ071615), a phylogenetically close relative of SARS-CoV-1, is unable to bind human ACE2
57 (hACE2) in vitro [4]. In contrast, another close relative of SARS-CoV-1, bat SARS-like coronavirus
58 WIV1 (GenBank accession KF367457) was shown to have the capacity to bind hACE2 [5]. A number of
59 other SARS-CoV-1-like viruses have also been tested for the ability to utilize hACE2 [6,7], and
60 comparison of their spike gene sequences shows that viruses that are unable to utilize hACE2
61 unanimously have one or two deletions in their RBDs that make them structurally very different than
62 those that do use hACE2 [7]. As all of these viruses are otherwise closely phylogenetically related, the
63 evolutionary mechanism explaining this variation in ACE2 usage among sarbecoviruses has thus far been
64 unclear.

65

66 Because of the diversity of sarbecoviruses that have been sequenced to date from Chinese horseshoe bats
67 (*Rhinolophidae*), this family of bats is thought to be the primary natural reservoir [5,6,8–10]. Rhinolophid
68 bats are also considered to be the source of the progenitor virus to SARS-CoV-1, as several viruses with
69 genetic regions with high identity to SARS-CoV-1 have been sequenced from rhinolophid bat samples,
70 although none of these bat viruses are completely identical to SARS-CoV-1 across the entire genome
71 [6,11]. It is hypothesized that SARS-CoV-1 obtained genomic regions from different strains of bat SARS-
72 1-like CoVs in or near Yunnan Province through recombination before spilling over into humans

73 [6,11,12]. In particular, one region of SARS-CoV-1 that is known to have a recombinant origin is the
74 spike gene, as a breakpoint has been detected at the junction of ORF1b and the spike [11,13]. The SARS-
75 1-CoV spike is genomically very different from other viruses in the same clade that have large deletions
76 in the receptor binding domain (RBD) and are unable to use hACE2. The exact minor parent that
77 contributed the recombinant region is still unknown, but it was previously hypothesized that the
78 recombination occurred with a yet undiscovered lineage of sarbecoviruses and that this event contributed
79 strongly to its potential for emergence [11,14]. Recombination has also been shown within the spike
80 genes of other CoVs that have spilled over into humans and domestic animals and is potentially an
81 important driver of emergence for all coronaviruses [15–20].
82
83 In order for CoVs to recombine, they must first have the opportunity to do so by sharing overlapping
84 geographic ranges, host species tropism, and cell and tissue tropism. Sarbecoviruses in bats tend to
85 phylogenetically cluster according to the geographic region in which they were found [6,21]. Yu et. al
86 showed that there are three lineages of SARS-CoV-1-like viruses: lineage 1 from southwestern China
87 (Yunnan, Guizhou, and Guangxi, and including SARS-CoV-1), lineage 2 from other southern regions
88 (Guangdong, Hubei, Hong Kong, and Zhejiang), and lineage 3 from central and northern regions (Hubei,
89 Henan, Shanxi, Shaanxi, Hebei, and Jilin) [21]. Studies in Europe and Africa have shown that there are
90 distinct sarbecovirus clades in each of these regions as well [22–27]. Sarbecoviruses appear to switch
91 easily among bat hosts of co-occurring *Rhinolophus* species [28,29] but appear to rarely occupy more
92 than one geographic area, despite the fact that some of these bat species have widespread distributions
93 across China.
94
95 Shortly after the emergence of SARS-CoV-2, Zhou et al. showed a high degree of homology across the
96 genome between a bat virus (RaTG13) sampled from Yunnan Province in 2013 and SARS-CoV-2 [3].
97 RaTG13 has also been shown to bind hACE2, although with decreased affinity compared to SARS-CoV-
98 2 [30]. Subsequently, seven full- or near full-length SARS-CoV-2-like viruses were published that had

been sampled from Malayin pangolins (*Manis javanica*) in 2017 and 2019 [31,32]. In addition, the most recently described bat virus (RmYN02) is even more closely related to SARS-CoV-2 than RaTG13 and was also found in Yunnan Province [33]. These viruses together with SARS-CoV-2 form a fourth phylogenetic lineage ("lineage 4") that is distinct from all other lineages of sarbecoviruses despite having been detected in the same geographic regions as lineage 1. This is inconsistent with the previously observed pattern of geographic concordance with phylogeny. SARS-CoV-2 was isolated first from people in Hubei Province and most of the pangolin viruses were isolated from animals sampled in Guangdong, neither of which are lineage 1 provinces. However, the true geographic origins of these viruses are unknown as it is possible they were anthropogenically transported to the regions in which they were detected. For example, the Malayan pangolin (*Manis javanica*) has a natural range that reaches southwestern China (Yunnan Province) at its northermost edge and extends further south into Myanmar, Lao PDR, Thailand, and Vietnam [34]. So, if they were naturally infected (as opposed to infection via wildlife trade), the infection was potentially not acquired from Guangdong Province. Similarly, SARS-CoV-2 cannot be guaranteed to have emerged from bats in Hubei Province, as humans are highly mobile and the exact spillover event was not observed. If the clade containing SARS-CoV-2 and its close relatives is indeed endemic in animals in Yunnan and the nearby Southeast Asian regions as suggested by the presence of RaTG13, RmYN02, and the natural range of the Malayan pangolin, whatever mechanism is facilitating the biogeographical concordance of lineages 1, 2 and 3 within China appears to no longer apply for the biogeography of lineage 4, since they all appear to overlap in and around Yunnan Province.

Here, we report a series of observations that together suggest that SARS-CoV-1 and its close relatives gained the ability to utilize ACE2 through a recombination event that happened between the progenitor of SARS-CoV-1 and a lineage 4 virus closely related to SARS-CoV-2, which could only have occurred with the lineages occupying the same geographic and host space. Viruses related to SARS-CoV-1 that cannot use hACE2 appear to have evolutionarily lost this trait after deletions that arose in the RBD. We also report three full-length genomes of sarbecoviruses from Rwanda and Uganda and demonstrate that the

125 RBDs of these viruses are genetically intermediate between viruses that use hACE2 and those that do not.
126 Accordingly, we also investigate the potential for these viruses to utilize hACE2 in vitro. Together, these
127 findings help illuminate the evolutionary history of ACE2 usage within sarbecoviruses and provide
128 insight into identifying their risk of emergence in the future. We also propose a mechanism that could
129 explain the pattern of phylogeography across lineages 1, 2, and 3, and why lineage 4 viruses (including
130 SARS-CoV-2 and its relatives) do not adhere to this pattern.
131
132 **Results**
133 Screening of bat samples from Africa as part of the USAID-PREDICT project resulted in the
134 identification of three novel bat sarbecoviruses with RBDs that appear to be intermediate to those of
135 ACE2-using viruses and non-ACE2-using viruses, as they have one of the two major deletions at the
136 ACE2 interface. All three are closely related to each other, as two viruses sequenced from bats in Uganda
137 (PREDICT_PDF-2370 and PREDICT_PDF-2386) are nearly identical to each other and the third from
138 Rwanda (PREDICT_PRD-0038) shares 99% overall nucleotide identity with the viruses from Uganda.
139 Each was most closely related to the previously reported SARS-related coronavirus BtKY72 (GenBank
140 accession KY352407) found in bats in Kenya [27]. Each shared ~76% overall identity with SARS-CoV-1
141 and ~74% overall identity with SARS-CoV-2. Phylogenetically, they lie within a clade that is between
142 SARS-CoV-1 and SARS-CoV-2 and cluster with BtKY72 and bat coronavirus BM48_21 from Bulgaria
143 (GenBank accession GU190215) [24]. The host species of all three viruses could not be definitively
144 identified in the field or in the lab but are all genetically identical. They may represent a cryptic species,
145 as the mitochondrial sequences are ~94% identical with *Rhinolophus ferrumequinum* in the cytochrome
146 oxidase I gene (COI) and ~96% identical with *Rhinolophus clivosus* in the cytochrome b (cytb) gene.
147 Sequences of COI and and cytb have been deposited in GenBank.
148
149 In order to understand the evolutionary history of these viruses within the context of previously
150 sequenced sarbecoviruses from elsewhere in the world, we first constructed a phylogenetic tree of the

151 RNA-dependent RNA polymerase (RdRp) gene (also known as nsp12). Additionally, we accounted for
152 the geographic origin of each virus as well as the host species in which each was identified by illustrating
153 these traits on the phylogeny (Figure 1). Using this tree, we observe the same geographic pattern of
154 concordance reported by Yu et al [21], where viruses in each lineage show a striking pattern of fidelity
155 with particular geographic regions within China. The viruses from Africa and Europe form a distinct
156 phylogenetic cluster, as would be expected with this pattern of phylogeography. Notably, SARS-CoV-2
157 does not lie within the clade of bat sarbecorviruses that have been detected in bats in China but rather lies
158 outside an additional clade that contains viruses in bats found in Africa and eastern Europe. The discovery
159 of the "lineage 4" clade containing SARS-CoV-2 and related viruses in pangolins and bats is a deviation
160 from the geographic patterns observed for other sarbecoviruses. RaTG13 and RmYN02 were found in the
161 same geographic region as lineage 1 viruses (including SARS-CoV-1), yet forms a completely distinct
162 phylogenetic branch.
163
164 To more specifically investigate the evolutionary history of ACE2 usage, we built a second phylogenetic
165 tree using only the RBD of the spike gene (Figure 2). This region was selected because the spike protein
166 mediates cell entry and because previous reports showed that SARS-CoV-1 and SARS-CoV-2 both use
167 hACE2, despite being distantly related in the RdRp (Figure 1) [2,3]. Within the RBD region of the
168 genome, SARS-CoV-1 and all ACE2-using viruses are much more closely related to SARS-CoV-2 than
169 to other lineage 1 viruses (Figure 2). Interestingly, bat virus RmYN02 is no longer associated with SARS-
170 CoV-2 in the RBD and is instead within the clade of non-ACE2-using viruses. We also found that within
171 the RBD, ACE2-using viruses and non-ACE2-using viruses are perfectly phylogenetically separated. The
172 viruses from Africa sequenced here form a distinct clade that is intermediate between the ACE2-using
173 and non-ACE2-using groups, but appears more closely related to the ACE2-using group.
174
175 While these viruses from Africa are slightly more similar to the ACE2-using group, they differ somewhat
176 in amino acid sequence from the ACE2-users at the binding interface, including a small deletion in the 5'

177 end (Figure 3). In order to assess the ability for these sarbecoviruses to use hACE2, we performed *in vitro*
178 experiments in which we replaced the RBD of SARS-CoV-1 with the RBD from the Uganda (PDF-2370,
179 PDF-2386) and Rwanda viruses (PRD-0038), as done previously [7]. Single-cycle Vesicular Stomatitis
180 Virus (VSV) reporter particles containing the recombinant SARS-Uganda and SARS-Rwanda spike
181 proteins were then used to infect BHK cells expressing hACE2. While VSV-SARS-CoV-1 showed
182 efficient usage of hACE2, VSV-Uganda and VSV-Rwanda did not (Figure 4), indicating that they are not
183 able to use hACE2 as a cellular receptor.
184
185 In order to elucidate the potential amino acid variations that hamper hACE2 binding for the viruses from
186 Africa, we modeled the RBD domain of SARS-CoV found in Uganda (PREDICT_PDF-2370,
187 PREDICT_PDF-2386) and Rwanda (PREDICT_PRD-0038). Unlike other non-hACE2 binders,
188 homology modeling suggests that the RBDs of these viruses from Africa are structurally similar to SARS-
189 CoV-1 and SARS-CoV-2 (Figure 5A). However, modeling the interaction with hACE2 reveals amino
190 acid differences at key interfacial positions that can help explain the lack of interaction observed for
191 rVSV-Uganda and rVSV-Rwanda viruses (Figure 5B-C). There are four regions of the RBD that lie
192 within 10Å of the interface with hACE2, one of which is the receptor binding ridge (SARS-CoV-1
193 residues 459-477) which is critical for hACE2 binding [30,35]. For the purpose of this study, we have
194 designated the remaining segments as regions 1 (residues 390-408), 2 (residues 426-443), and 3 (residues
195 478-491) (Figure 4).
196
197 Sarbecoviruses from Africa studied here have a 2-3 amino acid deletion (SARS-CoV-1 residues 434-436)
198 in region 2 (Figure 3). As many of the residues in this region make close contact with hACE2 (<5Å), it is
199 possible that this contributes to the disruption of hACE2 binding. One of these residues, Y436, establishes
200 hydrogen bonds with hACE residues D38 and Q42 in both SARS-CoV-1 and SARS-CoV2 (Figure 5C).
201 Notably, all other non-hACE2 binders also have deletions in residues 432-436.
202

203 Moreover, ACE2 contains two hotspots (K31 and K353) that are crucial targets for binding by SARS-
204 RBDs and amino acid variations in the RBD sequence enclosing these ACE2 hotspots have been shown
205 to shape viral infectivity, pathogenesis, and determine the host range of SARS-CoV-1 [36–38]. All
206 sarbecoviruses from Africa contain a Lys (K) at SARS-CoV-1 position 479 within region 3 (positions 481
207 and 482 for Uganda and Rwanda, respectively), which makes contact with these ACE2 hotspots (as
208 compared to N479 or Q493 in SARS-Cov-1 and 2 respectively; Figure 4). K479 decreases binding
209 affinity by more than 20-fold in SARS-CoV-1 [39]. The negative contribution of K479 in region 3 is
210 likely due to unfavorable electrostatic contributions with ACE2 hostspot K31 (Figure 5C) [37,40]. On the
211 other hand, SARS-CoV-1 residue T487 (N501 in SARS-CoV-2) is located near hotspot K353 and has a
212 valine in the viruses from Africa (residues 489 and 490) (Figure 3). As with residue 479, the amino acid
213 identity at position 487 contributes to the enhanced hACE2 binding observed in SARS-CoV-2 [37,38,40].
214 The presence of a hydrophobic residue at position 487, not previously observed in any ACE2 binding
215 sarbecovirus (Figure 3), might lead to a local rearrangement at the K353 hotspot that hinders hACE2
216 binding. Indeed, most non-ACE2 binders have a Val (V) in SARS-CoV-1 position 487 (Figure 3).
217
218 Finally, the receptor binding ridge, which is conspicuously absent from all non-ACE2 binders, is present
219 in the sarbecoviruses from Africa but has amino acid variations that differ significantly from both SARS-
220 CoV-1 and SARS-CoV-2 (Figure 3). Changes in the structure of this ridge contribute to increased binding
221 affinity of SARS-CoV-2, as a Pro-Pro-Ala (PPA) motif in SARS-CoV-1 (residues 469-471) replaced with
222 Gly-Val-Glu-Gly (GVEG) in SARS-CoV-2 results in a more compact loop and better binding with
223 hACE2 [30]. Changes within this ridge may be negatively contributing to hACE2 binding of viruses from
224 Africa, which have Ser-Thr-Ser-Gln (STSQ) or Ser-Iso-Ser-Gln (SISQ) in this position (Figure 3 and 5C).
225 Altogether, our results suggest that these viruses from Africa do not utilize hACE2 and provide additional
226 structural evidence that aids in distinguishing viruses which bind hACE2 from those that do not. This also
227 demonstrates that hACE2 usage within sarbecoviruses is restricted to those viruses within the SARS-
228 CoV-1 and SARS-CoV-2 clade in the RBD (Figure 2).

229

230 In order to evaluate an alternative scenario where hACE2 usage arose in SARS-CoV-1 and SARS-CoV-2
231 clades indepedently via convergent evolution, we compared the topology of the RdRp phylogeny with the
232 amino acid sequences of the interfacial residues within the RBD (Figure 3). In theory, if there was no
233 recombination between these two regions, we would expect both regions to follow the same evolutionary
234 trajectory. From this perspective, the deletions in region 2 and the receptor binding ridge of the RBD
235 appear to have been lost in a stepwise fashion. The small deletion in region 2 arose first, before the
236 diversification of clades in Africa and Europe (Figure 3). The larger deletion in the receptor binding ridge,
237 not present in known sequences from Africa and Europe, likely arose second, but before the
238 diversification of lineages 1, 2, and 3 (Figure 3). If this were the case, since SARS-CoV-1 and all ACE2-
239 using lineage 1 viruses are nested within a clade of non-ACE2-using viruses with deletions in region 2
240 and the receptor binding ridge, reinsertion of these deleted regions would be required to regain ACE2
241 usage. Further, RmYN02 is within the lineage 4 clade of ACE2-using viruses in RdRp but its RBD
242 sequence contains both deletions (Figure 3). Without recombination, these deletions would have had to be
243 independently lost from exactly the same regions as in the lineage 1 viruses.
244

245 Finally, to look for evidence of selective pressure in the RBD, we also performed selection analysis on the
246 spike protein of ACE2-using and non-ACE2-using sarbecoviruses independently using the codeml
247 package in PAML [41]. We found strong evidence that many residues in S1, specifically those that are
248 interfacial to ACE2 in the RBD, are undergoing positive selection in the ACE2-using group, including a
249 total of 16 residues in the interfacial region (Figure 6, Table 2). Notably, all five residues that were
250 implicated in the increased binding affinity of SARS-CoV-2 to hACE2 over SARS-CoV-1 show evidence
251 of positive selection (Table 2) [42]. In the non-ACE2-using group, however, most of these same residues
252 are not positively selected and appear to be under strong purifying selection instead (Figure 6). Only two
253 sites in this same region appear to be under positive selection for non-ACE2-users, neither of which are in
254 homologous positions with positively selected sites in the ACE2-using group (Figure 3). Almost all

255 signatures of positive selection are restricted to S1, as the vast majority of sites in the S2 region are not
256 under positive selection in either group.
257
258 **Discussion**
259 *ACE2 usage in lineage 1 viruses was acquired via recombination*
260 At first glance, hACE2 usage does not appear to be phylogenetically conserved among sarbecoviruses,
261 especially since many phylogenies are built using RdRp. This naturally leads to the hypothesis that
262 hACE2 usage arose independently in SARS-CoV-1 and SARS-CoV-2 via convergent evolution. This has
263 been suggested previously for another ACE2-using human coronavirus, NL63 [43]. However, a
264 phylogeny constructed using the RBD perfectly separates viruses that have been shown to utilize ACE2
265 from those that do not (Figure 2). Viruses that cannot utilize ACE2 have significant differences in their
266 RBDs, including large deletions in critical interfacial residues and low amino acid identity with viruses
267 that do use ACE2 (Figure 3). Notably, in addition to the large deletions, viruses that cannot use ACE2
268 deviate considerably at the interacting suface, including positions that play fundamental roles dictating
269 binding and cross-species transmission (Figure 3) [30,36,39,42]. It is unknown whether viruses that
270 cannot use hACE2 are utilizing bat ACE2 or an entirely different receptor altogether.
271
272 The difference in topology, specifically in the positioning of ACE2-using lineage 1 viruses, between
273 RdRp and RBD trees is most parsimoniously explained by a recombination event between a recent
274 ancestor of the ACE2-using lineage 1 viruses (including SARS-CoV-1) and a relative of SARS-CoV-2.
275 As the rest of the lineages of viruses remain in the same orientation between the two trees and only the
276 SARS-CoV-1 and ACE2-using relatives change positions, an ancestor of all ACE2-using lineage 1
277 viruses is most likely the recombinant virus. There are two groups of closely related RBD sequences in
278 the viruses that change positions (Figure 3), suggesting there may have been at least two separate
279 recombination events in this clade. Alternatively, it is possible there was a single recombination event that
280 was long enough ago such that there has been noticeable sequence divergence between the two groups.

281 The lack of monophyly among these recombinant viruses in RdRp can be explained by incomplete
282 lineage sorting, where the population of the common ancestor of these viruses continued to evolve
283 alongside the recombinant viruses.
284
285 This finding is also supported by numerous studies that have provided evidence that SARS-CoV-1 is
286 recombinant and SARS-CoV-2 is not [3,11,13,44]. We also demonstrate that recombination is possible
287 given that viruses related to SARS-CoV-1 and -2 appear to share both geographic and host space in
288 southwestern China and in *R. sinicus* and *R. affinis* bats. Highlighting that this previously known
289 recombination event occurred with a previously unknown group of viruses that are related to SARS-CoV-
290 2 is an important finding of this study and demonstrates that recombination is an important driver of
291 spillover for sarbecoviruses.
292
293 *A series of deletion events most likely resulted in the loss of ACE2 usage*
294 Assuming the RdRp tree represents the true evolutionary history of these viruses, sequences without the
295 deletions in the RBD most likely represent the ancestral state, as the SARS-CoV-2 lineage 4 viruses at the
296 base of the tree do not show this trait. Alternatively, it is possible that the deletion state is the ancestral
297 state and insertions were acquired during the evolution of the SARS-CoV-2 lineage; however, the only
298 virus in the subgenus *Hibecovirus*, the next closest virus phylogenetically to *Sarbecovirus*, does not show
299 the deletion trait (GenBank accesstion KF636752, Bat Hp-betacoronavirus/Zhejiang2013, not shown).
300 Further, the viruses from bats in Africa and Europe have one of the two deletions, which may indicate that
301 these are descendant from an evolutionary intermediate and support a stepwise deletion hypothesis. Since
302 ACE2-using lineage 1 viruses including SARS-CoV-1 are nested within a clade of viruses that all have
303 both deletions, this implies that both deletions arose before the diversification of lineage 1, 2, and 3
304 viruses (Figure 3). The smaller deletion in the receptor binding ridge was likely acquired earliest, before
305 the diversification of the clades into Africa and Europe, since it is shared by all clades with the exception
306 of SARS-CoV-2 lineage 4 at the base of the tree (Figure 3). These large deletions in the RBD-hACE2

307 interface also suggest that non-ACE2-using viruses, including lineages 1, 2, 3, and viruses from Africa
308 and Europe, are using at least one receptor other than ACE2 [7].
309
310 *ACE2 usage is not well explained by convergent evolution*
311 Under a hypothetical convergent evolution scenario, we would expect identical (or similar) residues at
312 key sites with structural or functional roles but not significant overall pairwise identity. We would also
313 expect to see evidence of selective pressure at these key sites that promotes the convergence to a
314 particular phenotype. Our results indicate that there is strong evidence of positive selection within the
315 receptor binding domain of ACE2-using viruses, but that this pressure is much less strong for non-ACE2-
316 using viruses. Although there is evidence of positive selection in this region for non-ACE2-users, many of
317 the residues in between these positively selected sites are strongly negatively selected, including critical
318 residues in the receptor binding ridge. We also highlight that the bat virus RmYN02, which is highly
319 similar to SARS-CoV-2 within the RdRp, actually has a RBD with the deletion trait associated with the
320 inability to use hACE2. If selective pressure was converging toward ACE2 usage, it would not have been
321 advantageous for a lineage 4 virus (and potentially ACE2-using, given its relation to SARS-CoV-2) to
322 lose this trait.
323
324 It is also unclear how this pressure for convergence would have arisen without these viruses switching to
325 a novel host [14]. Both the ACE2-using viruses and non-ACE2-using viruses in lineage 1 in the RdRp
326 tree are found within the same hosts, mostly bats belonging to the species *R. sinicus* and *R. affinis*,
327 leaving the source of the positive selection pressure driving convergent evolution a mystery. While
328 switching hosts and subsequently evolving via adaptation to the new host would be a potential case for
329 the convergent evolution hypothesis, the fact that these viruses a) do not seem to be specifically adapted
330 to any one particular species of bat given that we see nearly identical viruses in *R. affinis* and *R sinicus*
331 bat hosts, and b) have probably not undergone novel host shifts outside these bat species in recent history,
332 would suggest that this hypothesis is at best unsupported. Host shifts to civets and pangolins have

occurred; however, they are currently not known to be the progenitor of either SARS-CoV-1 or SARS-CoV-2 [12,32]. The most current knowledge suggests that SARS-CoV-1 in particular could have emerged directly from bats, and that civets may not have been a required step in the emergence pathway [5,12].

Not only would this selective pressure on a multitude of negatively selected sites be required in order for ACE2 usage to be regained within lineage 1, but also large insertions would have had to be re-acquired in precisely the same regions from which they were lost within the RBD. The most parsimonious argument is that ACE2-using lineage 1 viruses are descendent from at least one recombinant virus and that this recombination event explains the non-monophyletic pattern of ACE2 usage within the *Sarbecovirus* subgenus. In contrast, human coronavirus NL63 is an alphacoronavirus that is also an hACE2 user but most likely represents a true case of convergent evolution. The RBD of SARS-CoV-1 and SARS-CoV-2 are structurally identical, while NL63 has a different structural fold, suggesting that they are not evolutionarily homologous [43]. Nonetheless, NL63 also binds to hACE2 in the same region – suggesting all of the ACE2-using viruses have converged towards this interaction mode [43].

Differences in receptor usage within sarbecoviruses would explain observed phylogeographic patterns

Lineage 1 and SARS-CoV-2-like bat viruses appear to occupy the same geographic space but are quite distantly phylogenetically related, which is a notable deviation from previous observations that show sarbecovirus phylogeny mirrors geography. One hypothesis for this geographic specificity is that immune cross-reactivity between closely related viruses within hosts results in indirect competitive exclusion and priority effects. Antibodies against the spike protein are critical components of the immune response against CoVs [45–47]. Hosts that have been infected by one sarbecovirus may be immunologically resistant to infection from a related sarbecovirus, leading to geographic exclusion of closely related strains and a pattern of evolution that is concordant with geography (Figure 1). It is unlikely that this pattern is caused by differing competencies amongst *Rhinolophus* bats, as host-switching of these viruses appears to be common and many of these bat species are not geographically limited to the areas in which

359 the different virus lineages are found. The finding of an additional clade of SARS-CoV-2-like viruses
360 circulating in the same species and the same geographic location may suggest a release in the competitive
361 interactions maintaining geographic specificity. This would preclude recognition by cross-reactive
362 antibodies, such as those produced against the spike protein, and may be evolutionarily advantageous for
363 the recombinant virus. Furthermore, if these two groups of viruses utilize different receptors, antibodies
364 against one would be ineffective at excluding the other, potentially allowing both viral groups to infect
365 the same hosts. If competitive release has indeed occurred among these viruses, it is likely that the SARS-
366 CoV-2 clade is potentially much more diverse and geographically widespread than currently understood.
367
368 *Implications for future research*
369 Here, we highlight the critical need for further surveillance specifically in southernwestern China and
370 surrounding regions in southeast Asia given that all hACE2-using bat viruses discovered to date were
371 isolated from bats in Yunnan Province. Southeast Asia and parts of Europe and Africa have been
372 previously identified as hotspots for sarbecoviruses [48], but increased surveillance will help characterize
373 the true range of ACE2-using sarbecoviruses in particular. The receptors for viruses from northern China
374 and other regions such as Europe and Africa remain unknown, and may not pose a threat to human health
375 if they cannot utilize hACE2. It is unclear whether the lack of hACE2 binding for sarbecoviruses from
376 Uganda and Rwanda is due to the small deletion in region 2 or to the numerous amino acid changes in
377 other interfacial residues. It is possible that sarbecoviruses in Africa with different residues in these
378 interfacial regions could potentially still use hACE2. It is also unknown whether the sarbecoviruses from
379 Africa in particular use a divergent form of bat ACE2 or a different receptor altogether, or whether
380 sarbecoviruses with the potential to utilize hACE2 without the region 2 deletion have also diversified into
381 Africa or Europe. If competitive release between groups of viruses utilizing different receptors has indeed
382 occurred, further surveillance is needed to determine the true extent of lineage 4 viruses. In addition,
383 experimental evidence examining viral competition in vitro to support a competitive release hypothesis
384 should be prioritized.

385

386 This study highlights that hACE2 usage is unpredictable using phylogenetic proximity to SARS-CoV-1 or
387 SARS-CoV-2 in the RdRp gene. This is due to vastly different evolutionary histories in different parts of
388 the viral genome due to recombination. Phylogenetic relatedness in the RdRp gene is not an appropriate
389 proxy for pandemic potential among CoVs (the 'nearest neighbor' hypothesis). By extension, the
390 consensus PCR assays most commonly used for surveillance and discovery, which mostly generate a
391 small fragment of sequence from within this gene [49–51], are insufficient to predict hACE2 usage. Using
392 phylogenetic distance in RdRp as a quantitative metric to predict the potential for emergence is tempting
393 because of the large amount of data available, but this approach is unlikely to capture the biological
394 underpinnings of emergence potential compared to more robust data sources such as full viral genome
395 sequences. The current collection of full-length sarbecovirus genomes is heavily weighted toward China
396 and *Rhinolophus* hosts, despite evidence of sarbecoviruses prevalent outside of China (such as in Africa)
397 and in other mammalian hosts (such as pangolins). Further, investigations into determinants of
398 pathogenicity and transmission for CoVs and the genomic signatures of such features will be an important
399 step towards the prediction of viruses with spillover potential, and distinguishing those with pandemic
400 potential.

401

402 Finally, these findings reiterate the importance of recombination as a driver of spillover and emergence,
403 particularly in the spike gene. If SARS-CoV-1 gained the ability to use hACE2 through recombination,
404 other non-ACE2-using viruses could become human health threats through recombination as well. We
405 know that recombination occurs much more frequently than just this single event with SARS-CoV-1, as
406 the RdRp phylogeny does not mirror host phylogeny and the RBD tree has significantly different
407 topology across all geographic lineages. In addition, the bat virus RmYN02 appears to be recombinant in
408 the opposite direction (lineage 4 backbone with lineage 1 RBD) [33], again supporting the hypothesis that
409 recombination occurs between these lineages. Our analyses support two hypotheses: first, that
410 sarbecoviruses frequently undergo recombination in this region of the genome, resulting in this pattern,

411 and second, that sarbecoviruses are commonly shared amongst multiple host species, resulting in a lack of
412 concordance with host species phylogeny and a reasonable opportunity for coinfection and
413 recombination. Bats within the family *Rhinolophidae* have also repeatedly shown evidence of
414 introgression between species, supporting the hypothesis that many species in this family have close
415 contact with one another [52–57]. Given that we have shown that hACE2-using viruses are co-occuring
416 with a large diversity of non-ACE2-using viruses in Yunnan Province and in a similar host landscape,
417 recombination poses a significant threat to the emergence of novel sarbecoviruses [6].
418
419 With recombination constituting such an important variable in the emergence of novel CoVs,
420 understanding the genetic and ecological determinants of this process is a critical avenue for future
421 research. Here we have shown not only that recombination was involved in the emergence of SARS-CoV-
422 1, but also demonstrated how knowledge of the evolutionary history of these viruses can be used to infer
423 the potential for other viruses to spillover and emerge. Understanding this evolutionary process is highly
424 dependent on factors influencing viral co-occurrence and recombination, such as the geographic range of
425 these viruses and their bat hosts, competitive interactions with co-circulating viruses within the same
426 hosts, and the range of host species these viruses are able to infect. Our understanding depends on the data
427 we have available - the importance of generating more data for such investigations cannot be understated.
428 Investing effort now into further sequencing these viruses and describing the mechanisms that underpin
429 their circulation and capacity for spillover will have important payoffs for predicting and preventing
430 sarbecovirus pandemics in the future.
431
432 **Methods**
433 *Consensus PCR and sequencing of sarbecoviruses from Africa*
434 Oral swabs, rectal swabs, whole blood, and urine samples collected from bats sampled and released in
435 Uganda and Rwanda were assayed for CoVs using consensus PCR as previously described [20]. All
436 sampling was conducted under UC Davis IACUC Protocol No. 16048. Bands of the expected size were

437 purified and confirmed positive by Sanger sequencing and the PCR fragments were deposited to GenBank
438 (accessions). Samples were subsequently deep sequenced using the Illumina HiSeq platform and reads
439 were bioinformatically de novo assembled using MEGAHIT v1.2.8 [58] after quality control steps and
440 subtraction of host reads using Bowtie2 v2.3.5. Contigs were aligned to a reference sequence and any
441 overlaps or gaps were confirmed with iterative local alignment using Bowtie2. The full genome
442 sequences are deposited in GenBank. Cytochrome b and cytochrome oxidase I host sequences were also
443 extracted bioinformatically by mapping reads to *Rhinolophus ferrumequinum* reference genes using
444 Bowtie2 and deposited in GenBank.
445
446 *Phylogenetic reconstruction*
447 All publicly available full genome sarbecovirus sequences were collected from GenBank and SARS-
448 CoV-2, pangolin virus genomes, RaTG13, and RmYN02 were downloaded from GISAID (Table 1). All
449 relevant metadata (geographic origin, host species, date of collection) was retrieved from GenBank or the
450 corresponding publications. The RdRp gene (nucleotides 13,431 to 16,222 based on SARS-CoV-2
451 sequence EPI_ISL_402125 from GISAID) and RBD region (nucleotides 22,506 to 23,174 based on the
452 same SARS-CoV-2 reference genome) were extracted and aligned using ClustalW. Trees were built and
453 statistically supported using BEAST v2.6.2 with a Yule process prior, relaxed molecular clock and 10
454 million MCMC samples.
455
456 *Cell culture and transfection*
457 BHK and 293T cells were obtained from the American Type Culture Collection and maintained in
458 Dulbecco's modified Eagle's medium (DMEM; Sigma–Aldrich) supplemented with 10% fetal bovine
459 serum (FBS), penicillin/streptomycin and L-glutamine. BHK cells were seeded and transfected the next
460 day with 100ng of plasmid encoding hACE2 or an empty vector using polyethylenimine (Polysciences).
461 VSV plasmids were generated and transfected onto 293T cells to produce seed particles as previously
462 described [7]. CoV spike pseudotypes were generated as described in [59] and transfected onto 293T

463 cells. After 24h, cells were infected with VSV particles as described in [60], and after 1h of incubating at
464 37 °C, cells were washed three times and incubated in 2 ml DMEM supplemented with 2% FBS,
465 penicillin/streptomycin and L-glutamine for 48 h. Supernatants were collected and centrifuged at 500*g* for
466 5 min, then aliquoted and stored at −80 °C.
467
468 *Western blots*
469 293T cells transfected with CoV spike pseudotypes (producer cells) were lysed in 1% sodium dodecyl
470 sulfate, 150mM NaCl, 50 mM Tris-HCl and 5 mM EDTA and centrifuged at 14,000*g* for 20 minutes.
471 Pseudotyped particles were concentrated from producer cell supernatants that were overlaid on a 10%
472 OptiPrep cushion in PBS (Sigma–Aldrich) and centrifuged at 20,000*g* for 2 h at 4 °C. Lysates and
473 concentrated particles were analysed for FLAG (Sigma–Aldrich; A8592; 1:10,000), GAPDH (Sigma–
474 Aldrich; G8795; 1:10,000) and/or VSV-M (Kerafast; 23H12; 1:5,000) expression on 10% Bis-Tris PAGE
475 gel (Thermo Fisher Scientific).
476
477 *Cell entry assays*
478 Luciferase-based cell entry assays were performed as described in [7]. For each experiment, the relative
479 light unit for spike pseudotypes was normalized to the plate relative light unit average for the no-spike
480 control, and relative entry was calculated as the fold-entry over the negative control. Three replicates
481 were performed for each CoV pseudotype.
482
483 *Structural modeling*
484 RBDs were modeled using Modweb [61]. Modeled RBDs were docked to hACE2 by structural
485 superposition to the experimentally determined interaction complex between SARS-CoV-1 RBD and
486 hACE2 (PDB 2ajf) [36] using Chimera [62].
487
488 *Positive selection analysis*

489 The spike gene (SARS-CoV-2 nucleotides 21,525 to 25,373) was extracted and aligned from full genome
490 sequences separately for ACE2-using and non-ACE2-using groups of viruses (Table 1). Each alignment
491 was analyzed for positive selection using the codeml package within PAML using the M8 model. The
492 Bayes empirical Bayes posterior estimates of the selection coefficient ω with standard errors generated by
493 the M8 model were extracted from the output and plotted. Each estimate is statistically supported with a
494 posterior probability that the selection coefficient ω for the site is greater than 1. If the posterior
495 probability associated with the site was greater than 0.5, the estimate was considered borderline positively
496 selected; greater than 0.9, positively selected; and greater than 0.99, strongly positively selected (Table 2).

497 **Tables**

498 *Table 1. Full list of sequences and accession numbers used in this study.* All accession numbers are from
499 GenBank with the exception of those beginning with EPI_ISL, which are from GISAID. Metadata
500 includes equencing year, geographic origin, and host species. Group refers to which group, if either, the
501 sequence was included in for positive selection analysis. Group 1 viruses are those with the structure of
502 ACE2-using viruses and group 2 are those with the interfacial residue deletions which do not have the
503 structure for ACE2 binding.

Accession	Name	Date	Country	Host	Group
AY304486	SARS coronavirus SZ3	2003	Guangdong, China	Paguma larvata (civet)	1
AY304488	SARS coronavirus SZ16	2003	Hong Kong, China	Paguma larvata (civet)	1
AY572034	SARS coronavirus civet007	2004	Guangdong, China	Paguma larvata (civet)	1
DQ022305	Bat SARS coronavirus HKU3 1	2005	Hong Kong, China	Rhinolophus sinicus	2
DQ071615	Bat SARS coronavirus Rp3	2004	Guangxi, China	Rhinolophus pearsonii	2
DQ084199	Bat SARS coronavirus HKU3 2	2005	Hong Kong, China	Rhinolophus sinicus	2
DQ084200	Bat SARS coronavirus HKU3 3	2005	Hong Kong, China	Rhinolophus sinicus	2
DQ412042	Bat SARS coronavirus Rf1	2004	Hubei, China	Rhinolophus ferrumequinum	2
DQ412043	Bat SARS coronavirus Rm1	2004	Hubei, China	Rhinolophus macrotis	2
DQ648856	Bat coronavirus BtCoV/273/2005	2004	Hubei, China	Rhinolophus ferrumequinum	2
DQ648857	Bat coronavirus BtCoV/279/2005	2004	Hubei, China	Rhinolophus macrotis	2
EPI_ISL_402125	BetaCoV/Wuhan Hu 1	2019	Hubei, China	human	1
EPI_ISL_402131	BetaCoV/RaTG13	2013	Yunnan, China	Rhinolophus affinis	1
EPI_ISL_412977	BetaCoV/RmYN02	2019	Yunnan, China	Rhinolophus malayanus	NA
EPI_ISL_410538	BetaCoV/P4L	2017	Guangxi, China	Manis javanica (pangolin)	1
EPI_ISL_410539	BetaCoV/P1E	2017	Guangxi, China	Manis javanica (pangolin)	1
EPI_ISL_410540	BetaCoV/P5L	2017	Guangxi, China	Manis javanica (pangolin)	1

EPI_ISL_410541	BetaCoV/P5E	2017	Guangxi, China	Manis javanica (pangolin)	NA (incomplete sequence)
EPI_ISL_410542	BetaCoV/P2V	2017	Guangxi, China	Manis javanica (pangolin)	1
EPI_ISL_410543	BetaCoV/P3B	2017	Guangxi, China	Manis javanica (pangolin)	1
EPI_ISL_410544	BetaCoV/P2S	2019	Guangdong, China	Manis javanica (pangolin)	NA (incomplete sequence)
FJ588686	Bat SARS coronavirus Rs672/2006	2006	Guizhou, China	Rhinolophus sinicus	2
GQ153539	Bat SARS coronavirus HKU3 4	2005	Hong Kong, China	Rhinolophus sinicus	2
GQ153540	Bat SARS coronavirus HKU3 5	2005	Hong Kong, China	Rhinolophus sinicus	2
GQ153541	Bat SARS coronavirus HKU3 6	2005	Hong Kong, China	Rhinolophus sinicus	2
GQ153542	Bat SARS coronavirus HKU3 7	2006	Guangdong, China	Rhinolophus sinicus	2
GQ153543	Bat SARS coronavirus HKU3 8	2006	Guangdong, China	Rhinolophus sinicus	2
GQ153544	Bat SARS coronavirus HKU3 9	2006	Hong Kong, China	Rhinolophus sinicus	2
GQ153545	Bat SARS coronavirus HKU3 10	2006	Hong Kong, China	Rhinolophus sinicus	2
GQ153546	Bat SARS coronavirus HKU3 11	2007	Hong Kong, China	Rhinolophus sinicus	2
GQ153547	Bat SARS coronavirus HKU3 12	2007	Hong Kong, China	Rhinolophus sinicus	2
GQ153548	Bat SARS coronavirus HKU3 13	2007	Hong Kong, China	Rhinolophus sinicus	2
GU190215	Bat coronavirus BM48 21/BGR/2008	2008	Bulgaria	Rhinolophus blasii	NA
JX993987	Bat coronavirus Rp/Shaanxi2011	2011	Shaanxi, China	Rhinolophus pusillus	2
JX993988	Bat coronavirus Cp/Yunnan	2011	Yunnan, China	Chaerephon plicatus	2
KC881005	Bat SARS-like coronavirus RsSHC014	2012	Yunnan, China	Rhinolophus sinicus	1
KC881006	Bat SARS-like coronavirus Rs3367	2012	Yunnan, China	Rhinolophus sinicus	1
KF294457	SARS related bat coronavirus Longquan 140	2012	Guizhou, China	Rhinolophus monoceros	2
KF367457	Bat SARS-like coronavirus WIV1	2012	Yunnan, China	Rhinolophus sinicus	1
KF569996	Rhinolophus affinis coronavirus LYRa11	2011	Yunnan, China	Rhinolophus affinis	1
KF636752	Bat Hp betacoronavirus/Zhejiang2013	2013	Zhejiang, China	Hipposideros pratti	NA
KJ473811	Bat coronavirus BtRf BetaCoV/JL2012	2012	Jilin, China	Rhinolophus ferrumequinum	2

KJ473812	Bat coronavirus BtRf BetaCoV/HeB2013	2013	Hebei, China	Rhinolophus ferrumequinum	2
KJ473813	Bat coronavirus BtRf BetaCoV/SX2013	2013	Shanxi, China	Rhinolophus ferrumequinum	2
KJ473814	Bat coronavirus BtRs BetaCoV/HuB2013	2013	Hubei, China	Rhinolophus sinicus	2
KJ473815	Bat coronavirus BtRs BetaCoV/GX2013	2013	Guangxi, China	Rhinolophus sinicus	2
KJ473816	Bat coronavirus BtRs BetaCoV/YN2013	2013	Yunnan, China	Rhinolophus sinicus	2
KP886808	Bat SARS-like coronavirus YNLF 31C	2013	Yunnan, China	Rhinolophus sinicus	2
KP886809	Bat SARS-like coronavirus YNLF 34C	2013	Yunnan, China	Rhinolophus sinicus	2
KT444582	SARS-like coronavirus WIV16	2013	Yunnan, China	Rhinolophus sinicus	1
KU182964	Bat coronavirus JTMC15	2013	Yunnan, China	Rhinolophus sinicus	2
KU182963	Bat coronavirus MLHJC35	2012	Jilin, China	Rhinolophus sinicus	2
KU973692	SARS related coronavirus F46	2012	Yunnan, China	Rhinolophus pusillus	2
KY352407	SARS related coronavirus BtKY72	2007	Kenya	Rhinolophus sp.	NA
KY417142	Bat SARS-like coronavirus As6526	2014	Yunnan, China	Aselliscus stoliczkanus	2
KY417143	Bat SARS-like coronavirus Rs4081	2012	Yunnan, China	Rhinolophus sinicus	2
KY417144	Bat SARS-like coronavirus Rs4084	2012	Yunnan, China	Rhinolophus sinicus	1
KY417145	Bat SARS-like coronavirus Rf4092	2012	Yunnan, China	Rhinolophus ferrumequinum	2
KY417146	Bat SARS-like coronavirus Rs4231	2013	Yunnan, China	Rhinolophus sinicus	2
KY417147	Bat SARS-like coronavirus Rs4237	2013	Yunnan, China	Rhinolophus sinicus	2
KY417148	Bat SARS-like coronavirus Rs4247	2013	Yunnan, China	Rhinolophus sinicus	2
KY417149	Bat SARS-like coronavirus Rs4255	2013	Yunnan, China	Rhinolophus sinicus	2
KY417150	Bat SARS-like coronavirus Rs4874	2013	Yunnan, China	Rhinolophus sinicus	1
KY417151	Bat SARS-like coronavirus Rs7327	2014	Yunnan, China	Rhinolophus sinicus	1
KY417152	Bat SARS-like coronavirus Rs9401	2015	Yunnan, China	Rhinolophus sinicus	1
KY770858	Bat coronavirus Anlong 103	2013	Guizhou, China	Rhinolophus sinicus	2
KY770859	Bat coronavirus Anlong 112	2013	Guizhou, China	Rhinolophus sinicus	2

KY770860	Bat coronavirus Jiyuan 84	2012	Henan, China	Rhinolophus ferrumequinum	2
KY938558	Bat coronavirus 16BO133	2016	South Korea	Rhinolophus ferrumequinum	2
MG772933	Bat SARS-like coronavirus SL CoVZC45	2017	Zhejiang, China	Rhinolophus sinicus	2
MG772934	Bat SARS-like coronavirus SL CoVZXC21	2015	Zhejiang, China	Rhinolophus sinicus	2
MK211374	Bat coronavirus BtRl BetaCoV/SC2018	2018	Sichuan, China	Rhinolophus	2
MK211375	Bat coronavirus BtRs BetaCoV/YN2018A	2018	Yunnan, China	Rhinolophus affinis	2
MK211376	Bat coronavirus BtRs BetaCoV/YN2018B	2018	Yunnan, China	Rhinolophus affinis	1
MK211377	Bat coronavirus BtRs BetaCoV/YN2018C	2018	Yunnan, China	Rhinolophus affinis	2
MK211378	Bat coronavirus BtRs BetaCoV/YN2018D	2018	Yunnan, China	Rhinolophus affinis	2
NC_004718	SARS coronavirus	2003	Canada	human	1
TBA	PREDICT PDF-2370	2013	Uganda	Rhinolophus sp.	NA
TBA	PREDICT PDF-2386	2013	Uganda	Rhinolophus sp.	NA
TBA	PREDICT PRD-0038	2010	Rwanda	Rhinolophus sp.	NA

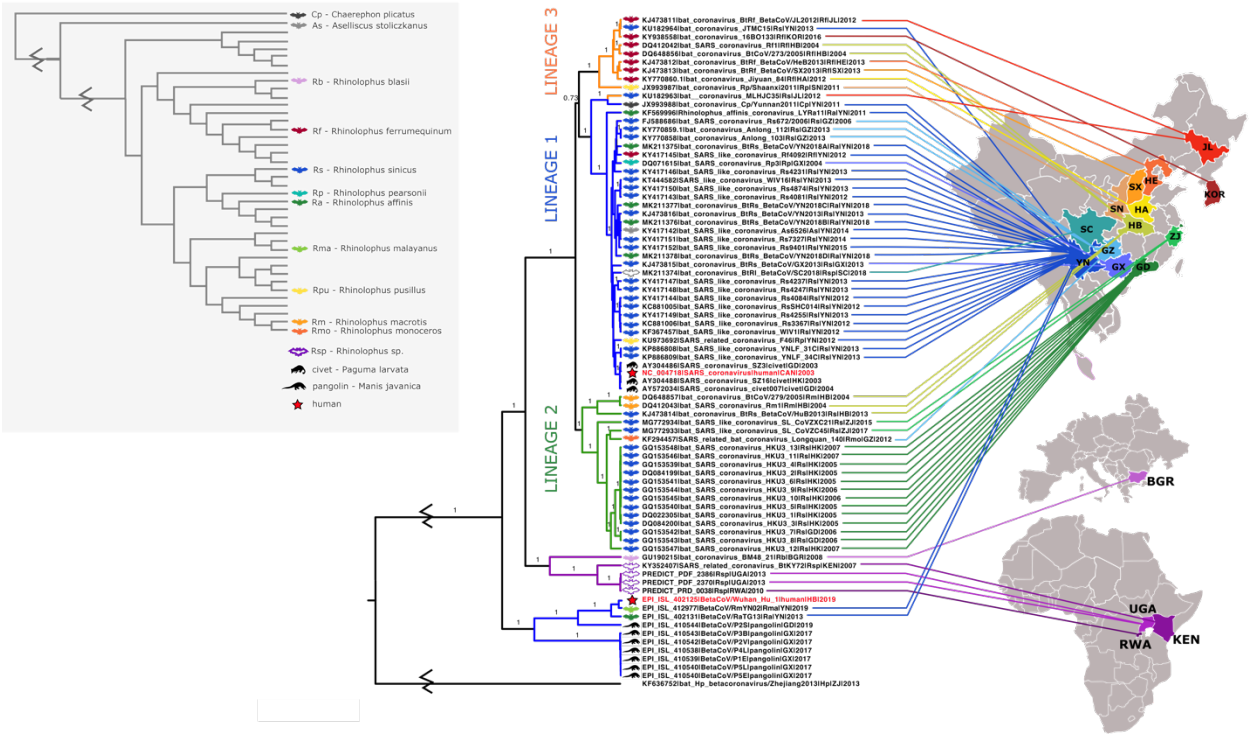
504

505 *Table 2. Residues identified as being under positive selection in the spike gene*. Group 1 viruses are those
506 with the structure of ACE2-using viruses and group 2 are those with the interfacial residue deletions
507 which do not have the structure for ACE2 binding. Group 1 residue numbering is relative to SARS-CoV-
508 2 (GISAID accession EPI_ISL_402125) and group 2 residue numbering is relative to bat coronavirus
509 16BO133 (GenBank accession KY938558). Residues that were also shown to be responsible for
510 increased binding affinity to hACE2 in SARS-CoV-2 are indicated with asterisks in column 2 [42]. The
511 posterior estimate of ω ± the standard error from the M8 model in the codeml analysis from PAML is
512 given in column 3. If the posterior probability in column 4 associated with the estimate at that site was
513 greater than 0.5, the estimate was considered borderline positively selected; greater than 0.9, positively
514 selected (indicated with one asterisk); and greater than 0.99, strongly positively selected (indicated with
515 two asterisks).

Group	Residue	Posterior estimate of ω ± s.e.	Posterior probability of ω>1
1	18L	1.371 ± 0.321	0.851
1	19T	1.430 ± 0.242	0.915 (*)
1	27A	1.408 ± 0.272	0.888
1	53T	1.294 ± 0.398	0.776
1	66H	1.058 ± 0.501	0.544
1	152W	1.375 ± 0.319	0.857
1	218Q	1.398 ± 0.289	0.881
1	250T	1.248 ± 0.435	0.735
1	252G	1.267 ± 0.406	0.740
1	255S	1.425 ± 0.249	0.908 (*)
1	256S	1.402 ± 0.282	0.883
1	344E	1.303 ± 0.385	0.780
1	459N	1.480 ± 0.135	0.971 (*)
1	440N	1.360 ± 0.337	0.841
1	441L	1.345 ± 0.352	0.826
1	444K	1.032 ± 0.506	0.520
1	445V	1.499 ± 0.055	0.997 (**)
1	452L	1.361 ± 0.328	0.838
1	455L*	1.387 ± 0.310	0.873
1	458K	1.337 ± 0.349	0.811
1	459S	1.419 ± 0.259	0.902 (*)
1	484E	1.466 ± 0.170	0.955 (*)
1	486F*	1.440 ± 0.223	0.925 (*)
1	490F	1.492 ± 0.094	0.987 (*)

1	493Q*	1.495 ± 0.081	0.990 (**)
1	494S*	1.484 ± 0.121	0.978 (*)
1	498Q	1.328 ± 0.362	0.805
1	501N*	1.467 ± 0.169	0.957 (*)
1	554E	1.022 ± 0.500	0.502
1	642V	1.231 ± 0.430	0.0.705
1	679N	1.353 ± 0.346	0.836
1	688A	1.044 ± 0.513	0.538
2	143G	1.465 ± 0.154	0.950 (*)
2	147T	1.455 ± 0.177	0.938 (*)
2	153I	1.197 ± 0.403	0.624
2	214A	1.444 ± 0.196	0.922 (*)
2	429V	1.500 ± 0.013	1.000 (**)
2	431S	1.463 ± 0.170	0.941 (*)
2	499P	1.494 ± 0.065	0.990 (**)
2	521D	1.168 ± 0.409	0.587
2	594P	1.407 ± 0.249	0.873
2	710L	1.325 ± 0.342	0.783

517 **Figures**



518

519 *Figure 1: Phylogenetic tree of the RNA dependent RNA polymerase (RdRp) gene (nsp12) and associated*

520 *geographic origin and host species for sarbecoviruses.* Colors of clade bars represent the different

521 geographic lineages. Lineage 1 is shown in blue, lineage 2 in green, and lineage 3 in orange. The clade of

522 viruses from Africa and Europe is shown in purple. The putative "lineage 4" containing SARS-CoV-2 is

523 also shown in blue at the bottom of the tree since the sequences are from the same regions as lineage 1

524 viruses. The geographic origin of each virus is indicated by the lines that terminate in the respective

525 country or province with the same color code. The full province and country names for all two- and three-

526 letter codes can be found in Table 1. As human, civet, and pangolin viruses cannot be certain to have

527 naturally originated in the province in which they were first found, their locations are not illustrated.

528 Hosts are also shown with colored symbols according to the key on the left. The host phylogeny in the

529 key was adapted from [63].

530

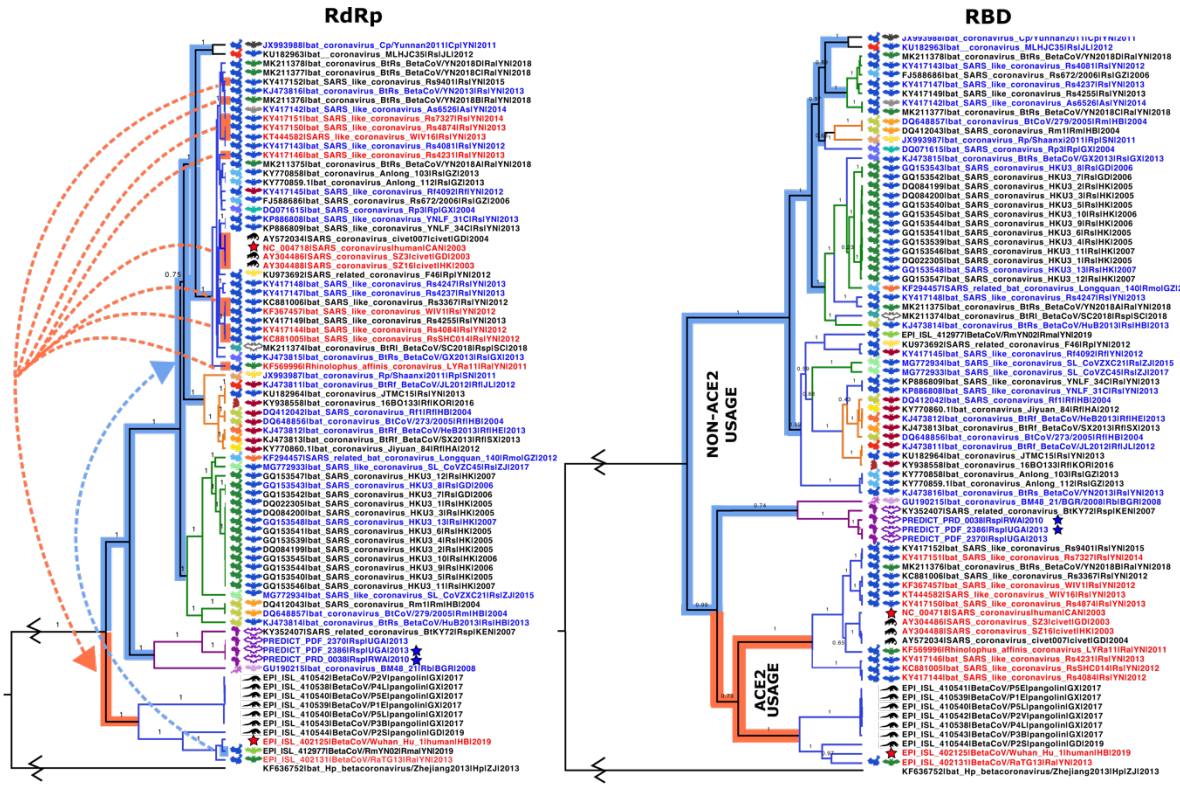


Figure 2: Phylogenetic trees of RdRp (left) and the RBD (right) demonstrating recombination events between ACE2-users and non-ACE2-users. Names of viruses that have been confirmed to use hACE2 are shown in red font, and those that have been shown to not use hACE2 are shown in blue font. Viruses in black font have not yet been tested, and viruses indicated with blue stars are tested in the present study. The red and blue highlighted clade bars separate viruses with the structure associated with ACE2 usage (highly similar to viruses confirmed to use hACE2 specifically) and the structure with deletions that cannot use ACE2, respectively. Geography and host are reiterated here in the same fashion as from Figure 1. Dashed lines indicate recombination events that resulted in a gain of ACE2 usage (red) or a loss of ACE2 usage (blue). The arrows only represent the direction of the recombination event and while they originate at the recombinant sequence, they should not be interpreted as pointing to any putative parental sequences or as pointing to the evolutionary distance of the clade bars.

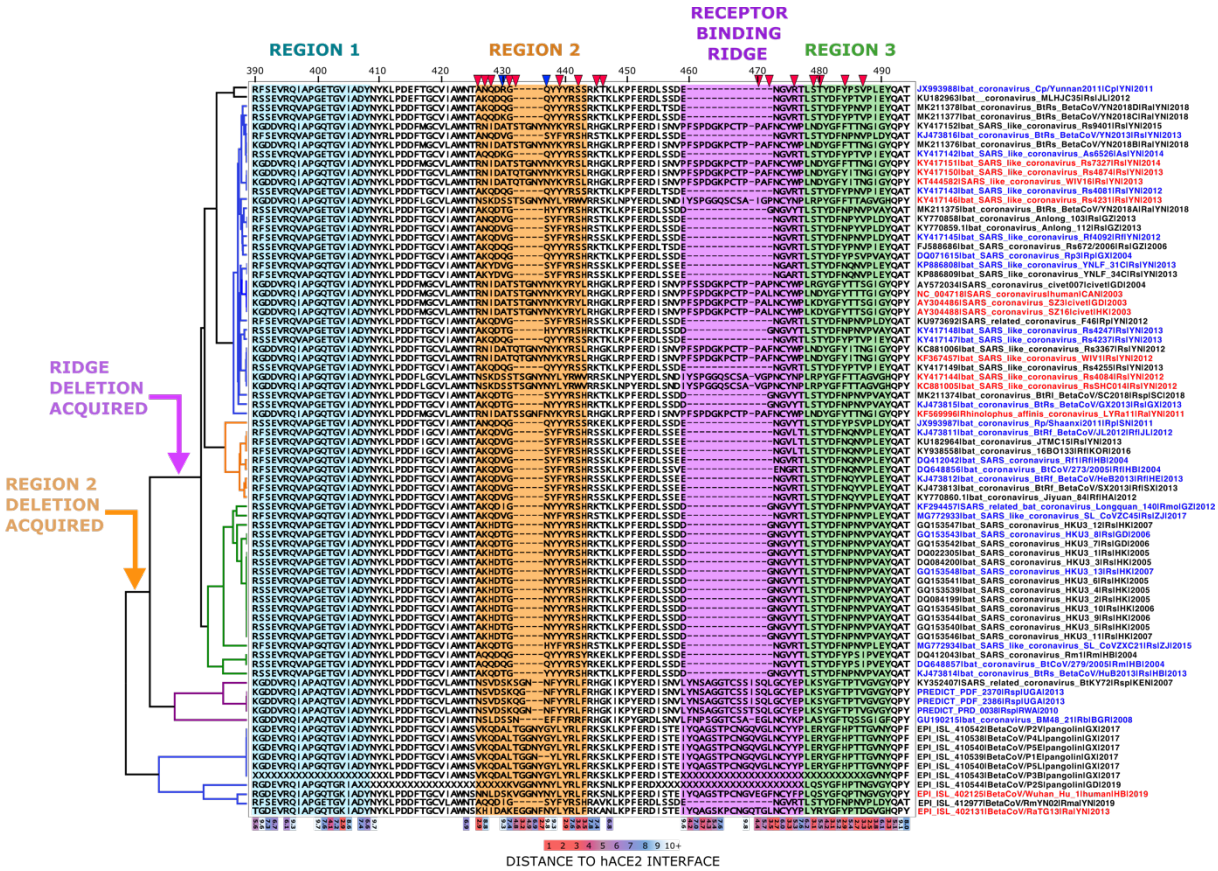
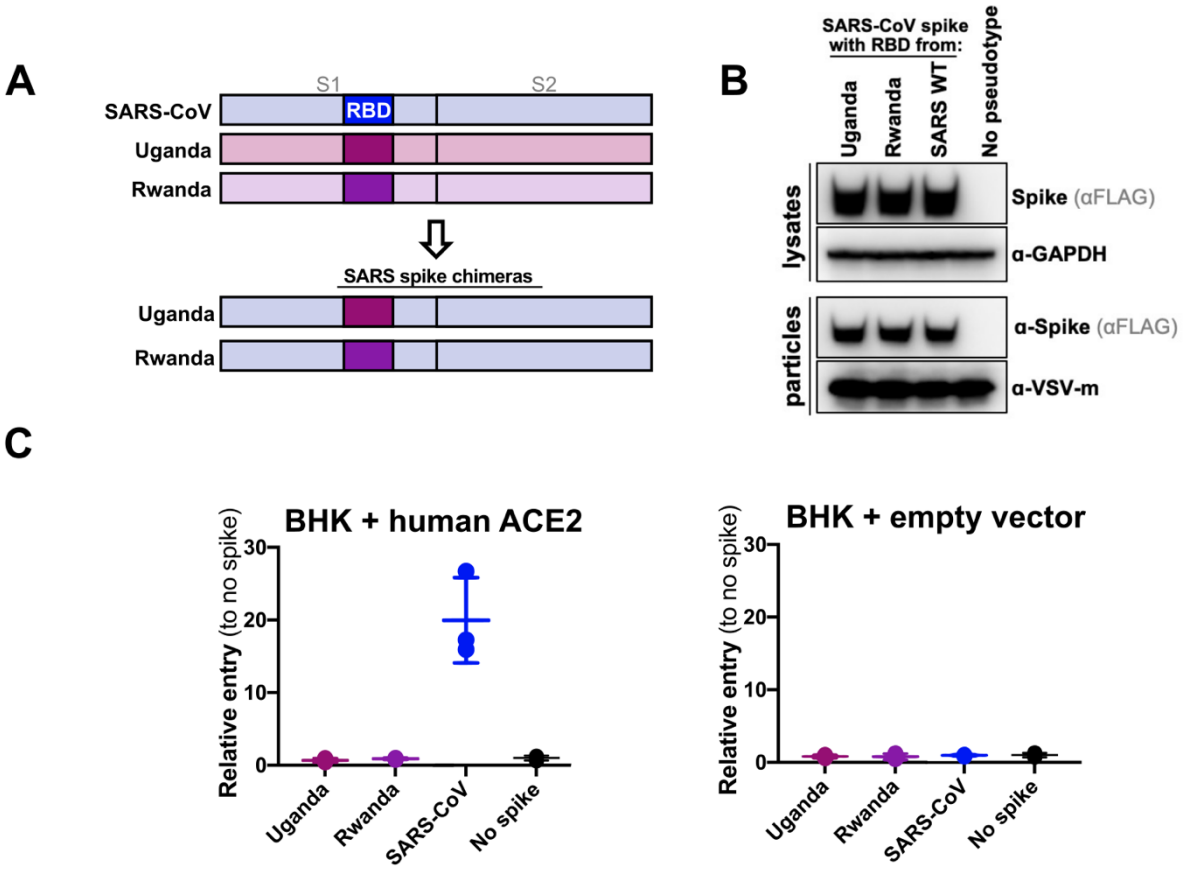


Figure 3: The phylogenetic backbone of the RdRp gene alongside the amino acid sequences of the RBM. Amino acid numbering is relative to SARS-CoV-1. Virus names in red font are known hACE2 users, those in blue are known non-users, and those in black have not been tested. Residues that show evidence of positive selection within the ACE2-using group are designated with red arrows at the top, and those that show positive selection within the non-ACE2-using group are designated with blue arrows. Residues within 10Å of the interface with ACE2 are considered interfacial, and exact distances between each interfacial residue and the closest ACE2 residue (based on structural modeling of SARS-CoV-1 bound with hACE2) are shown along the bottom. Residues that are closer to the interface (3Å or less) and thus make strong interactions with ACE2 are shown in red, and as distance increases this color transitions to purple, blue, and finally to white. The receptor bindng ridge sequences are highlighted in purple and the remaining interfacial segments have been numbered regions 1, 2, and 3 for clarity within the main text. The colors of these regions correspond with the colors in the structural models of Figure 5. Arrows over

556 the clade bars indicate potential points where deletions were acquired in a stepwise fashion in region2 and
557 the receptor binding ridge. True gaps in sequences are represented with hyphens, and residues that are
558 expected to be present but where sequence was not recovered are marked with 'X'.



559

560 *Figure 4: hACE2 useage of bat sarbecoviruses investigated using a surrogate VSV-psuedotyping system.*
561 (A) Schematic showing the structure of chimeric spike proteins. The SARS-CoV-1 spike backbone is
562 used in conjunction with the RBD from the Uganda and Rwanda strains. (B) Incorporation of chimeric
563 SARS-CoV-1 spike proteins into VSV. Western blots show successful expression of chimeric spikes
564 (lysates) and their incorporation into VSV (particles). (C) hACE2 entry assay. Wildtype SARS-CoV spike
565 protein is able to mediate entry into BHK cells expressing hACE2. In contrast, recombinant spike proteins
566 containing either the Uganda or Rwanda RBD were unable to mediate entry. Entry is expressed relative to
567 VSV particles with no spike protein. (D) Control experiment for entry assay. BHK cells do not express
568 hACE2 and therefore do not permit entry of hACE2-dependent VSV pseudotypes.

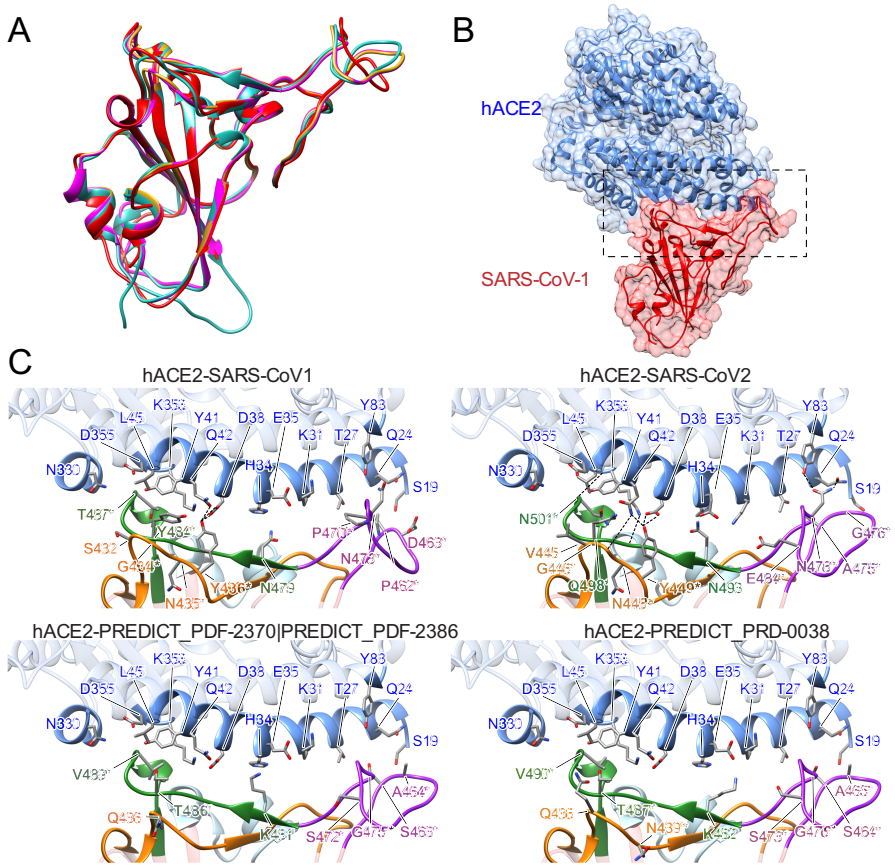


Figure 5. Structural modeling of sarbecovirus RBDs found in Uganda and Rwanda. **(**A) Structural superposition of the X-ray structures for the RBDs in SARS-CoV-1 (PDB 2ajf, red) [36] and SARS-CoV-2 (PDB 6m0j, cyan) [64] and homology models for SARS-CoV found in Uganda (PREDICT_PDF-2370 and PREDICT_PDF-2386, purple) and Rwanda (PREDICT_PRD-0038, yellow). (B) Overview of the X-ray structure of SAR-CoV-1 RBD (red) bound to hACE2 (blue) (PDB 2ajf, red) [36]. (C) Close-up view of the interface between hACE2 (blue) and RBDs in SARS-CoV-1 (PDB 2ajf, red) [36] and SARS-CoV-2 (PDB 6m0j, cyan) [64] and homology models for viruses found in Uganda (PREDICT_PDF-2370 and PREDICT_PDF-2386, purple) and Rwanda (PREDICT_PRD-0038, yellow). Labeled RBD residues correspond to residues whose identity is not shared by SARS-CoV-1 and/or SARS-CoV-2 (asterisks denote residues whose identity is not shared by any ACE-2 binding SARS-CoV as dictated by Figure 3). Labeled hACE2 residues correspond to residues within 5Å of RBD residues depicted.

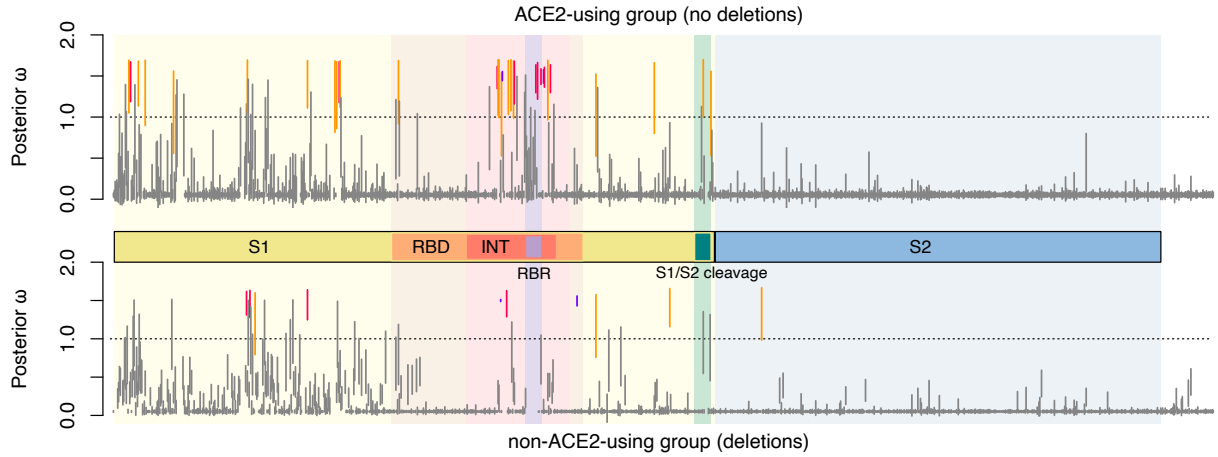


Figure 6: Positive selection analysis of the spike protein. Bayes empirical Bayes posterior estimates of the selection coefficient ω ± standard error generated using the M8 model within the codeml package in PAML were extracted and plotted. Viruses with RBD sequences with the structure for ACE2 binding (those without deletions) are shown above, and viruses with RBD sequences that do not have the structure for ACE2 binding (with deletions in interfacial residues) are shown below. Residues with estimates of ω greater than 1 were considered positively selected. Uncertainty assessed by the posterior probability that ω>1 for each site is shown by the color of each bar: grey bars are not positively selected (posterior probability less than 0.5), orange bars are borderline positively selected (posterior probability between 0.5 and 0.9), red bars are positively selected (posterior probability between 0.9 and 0.99), and purple bars are strongly positively selected (posterior probability >0.99). The domains of the spike gene are shown in the center: spike subunit 1 (S1), receptor binding domain (RBD), interfacial residues (INT), receptor binding ridge (RBR), the S1/S2 cleavage site unique to SARS-CoV-2 [65], and spike subunit 2 (S2).

Acknowledgements

The research reported in this publication was supported by the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health under Award Number R01AI149693 (PI Anthony). ML and VJM are supported by the Intramural Research Program of the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH). GL and KC are supported by National Institutes of Health (NIH) grant U19AI142777. This study was also made possible by the support of the American people through the United States Agency for International Development (USAID) Emerging Pandemic Threats PREDICT project, GHN-A-OO-09-00010-00 (PI Mazet) and AID-OAA-A-14-00102 (PI Mazet). The content is solely the responsibility of the authors and does not necessarily represent the official views of the U.S. Government.

Statement of Data Availability

All sequences have been submitted to GenBank and accession numbers will be provided prior to acceptance/publication.

References

1 International Committee on Taxonomy of Viruses (ICTV) Virus Taxonomy: 2019 Release. . [Online]. Available: https://talk.ictvonline.org/taxonomy/. [Accessed: 15-May-2020]

2 Li, W. *et al.* (2003) Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* 426, 450–454

3 Zhou, P. *et al.* (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* DOI: 10.1038/s41586-020-2012-7

4 Ren, W. *et al.* (2008) Difference in Receptor Usage between Severe Acute Respiratory Syndrome (SARS) Coronavirus and SARS-Like Coronavirus of Bat Origin. *J. Virol.* 82, 1899–1907

5 Ge, X.Y. *et al.* (2013) Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* 503, 535–538

6 Hu, B. *et al.* (2017) Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog.* DOI: 10.1371/journal.ppat.1006698

7 Letko, M. *et al.* (2020) Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol.* 5, 562–569

8 Lau, S.K.P. *et al.* (2005) Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc. Natl. Acad. Sci. U. S. A.* DOI: 10.1073/pnas.0506735102

9 Li, W. *et al.* (2005) Bats are natural reservoirs of SARS-like coronaviruses. *Science (80-.).* DOI: 10.1126/science.1118391

10 He, B. *et al.* (2014) Identification of Diverse Alphacoronaviruses and Genomic Characterization of a Novel Severe Acute Respiratory Syndrome-Like Coronavirus from Bats in China. *J. Virol.* DOI: 10.1128/jvi.00631-14

11 Hon, C.-C. *et al.* (2008) Evidence of the Recombinant Origin of a Bat Severe Acute Respiratory Syndrome (SARS)-Like Coronavirus and Its Implications on the Direct Ancestor of SARS Coronavirus. *J. Virol.* DOI: 10.1128/jvi.01926-07

12 Luk, H.K.H. *et al.* Molecular epidemiology, evolution and phylogeny of SARS coronavirus. , *Infection, Genetics and Evolution*. (2019)

13 Lau, S.K.P. *et al.* (2010) 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.* DOI: 10.1128/jvi.02219-09

14 Yuan, J. *et al.* (2010) Intraspecies diversity of SARS-like coronaviruses in Rhinolophus sinicus and its implications for the origin of SARS coronaviruses in humans. *J. Gen. Virol.* 91, 1058–1062

15 Graham, R.L. and Baric, R.S. (2010) Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission. *J. Virol.* 84, 3134–3146

16 Su, S. *et al.* Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. , *Trends in Microbiology*, 24. 01-Jun-(2016) , Elsevier Ltd, 490–502

17 Menachery, V.D. *et al.* Jumping species—a mechanism for coronavirus persistence and survival. , *Current Opinion in Virology*. (2017)

18 Woo, P.C.Y. *et al.* Coronavirus diversity, phylogeny and interspecies jumping. , *Experimental Biology and Medicine*. (2009)

19 Lu, G. *et al.* Bat-to-human: Spike features determining “host jump” of coronaviruses SARS-CoV, MERS-CoV, and beyond. , *Trends in Microbiology*. (2015)

20 Anthony, S.J. *et al.* (2017) Further evidence for bats as the evolutionary source of middle east respiratory syndrome coronavirus. *MBio* DOI: 10.1128/mBio.00373-17

21 Yu, P. *et al.* Geographical structure of bat SARS-related coronaviruses. , *Infection, Genetics and Evolution*, 69. 01-Apr-(2019) , Elsevier B.V., 224–229

22 Lecis, R. *et al.* (2019) Molecular identification of Betacoronavirus in bats from Sardinia (Italy): first detection and phylogeny. *Virus Genes* 55, 60–67

23 Ar Gouilh, M. *et al.* (2018) SARS-CoV related Betacoronavirus and diverse Alphacoronavirus

659 members found in western old-world. *Virology* 517, 88–97
660 24 Drexler, J.F. *et al.* (2010) Genomic Characterization of Severe Acute Respiratory Syndrome-
661 Related Coronavirus in European Bats and Classification of Coronaviruses Based on Partial RNA-
662 Dependent RNA Polymerase Gene Sequences. *J. Virol.* 84, 11336–11349
663 25 Rihtarič, D. *et al.* (2010) Identification of SARS-like coronaviruses in horseshoe bats
664 (Rhinolophus hipposideros) in Slovenia. *Arch. Virol.* 155, 507–514
665 26 Lelli, D. *et al.* (2013) Detection of Coronaviruses in Bats of Various Species in Italy. *Viruses* 5,
666 2679–2689
667 27 Tao, Y. and Tong, S. (2019) Complete Genome Sequence of a Severe Acute Respiratory
668 Syndrome-Related Coronavirus from Kenyan Bats. *Microbiol. Resour. Announc.* 8,
669 28 Cui, J. *et al.* (2007) Evolutionary relationships between bat coronaviruses and their hosts. *Emerg.*
670 *Infect. Dis.* 13, 1526–1532
671 29 Leopardi, S. *et al.* (2018) Interplay between co-divergence and cross-species transmission in the
672 evolutionary history of bat coronaviruses. *Infect. Genet. Evol.* DOI: 10.1016/j.meegid.2018.01.012
673 30 Shang, J. *et al.* (2020) Structural basis of receptor recognition by SARS-CoV-2. *Nature* 581, 221–
674 224
675 31 Liu, P. *et al.* (2019) Viral metagenomics revealed sendai virus and coronavirus infection of
676 malayan pangolins (manis javanica). *Viruses* DOI: 10.3390/v11110979
677 32 Lam, T.T.Y. *et al.* (2020) Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins.
678 *Nature* DOI: 10.1038/s41586-020-2169-0
679 33 Zhou, H. *et al.* (2020) A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains
680 Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. *Curr. Biol.* DOI:
681 10.1016/j.cub.2020.05.023
682 34 Challender, D. *et al.* (2014) Manis javanica. *IUCN Red List Threat. Species 2014* DOI:
683 http://dx.doi.org/10.2305/IUCN.UK.2014-2.RLTS.T12763A45222303.en.
684 35 Prabakaran, P. *et al.* (2004) A model of the ACE2 structure and function as a SARS-CoV receptor.
685 *Biochem. Biophys. Res. Commun.* 314, 235–241
686 36 Li, F. *et al.* (2005) Structural biology: Structure of SARS coronavirus spike receptor-binding
687 domain complexed with receptor. *Science (80-.).* 309, 1864–1868
688 37 Li, F. (2008) Structural Analysis of Major Species Barriers between Humans and Palm Civets for
689 Severe Acute Respiratory Syndrome Coronavirus Infections. *J. Virol.* DOI: 10.1128/jvi.00442-08
690 38 Wu, K. *et al.* (2012) Mechanisms of host receptor adaptation by severe acute respiratory syndrome
691 coronavirus. *J. Biol. Chem.* DOI: 10.1074/jbc.M111.325803
692 39 Li, W. *et al.* (2005) Receptor and viral determinants of SARS-coronavirus adaptation to human
693 ACE2. *EMBO J.* 24, 1634–1643
694 40 Wan, Y. *et al.* (2020) Receptor recognition by novel coronavirus from Wuhan: 2 An analysis
695 based on decade-long structural studies of SARS 3 4 Downloaded from. DOI: 10.1128/JVI.00127-
696 20
697 41 Yang, Z. (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* DOI:
698 10.1093/molbev/msm088
699 42 Wan, Y. *et al.* (2020) Receptor recognition by novel coronavirus from Wuhan: An analysis based
700 on decade-long structural studies of SARS. *J. Virol.* DOI: 10.1128/jvi.00127-20
701 43 Chen, Y. *et al.* (2020) Structure analysis of the receptor binding of 2019-nCoV. *Biochem. Biophys.*
702 *Res. Commun.* 525, 135–140
703 44 Wu, F. *et al.* (2020) A new coronavirus associated with human respiratory disease in China.
704 *Nature* DOI: 10.1038/s41586-020-2008-3
705 45 Buchholz, U.J. *et al.* (2004) Contributions of the structural proteins of severe respiratory syndrome
706 coronavirus to protective immunity. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9804–9809
707 46 Lu, L. *et al.* (2004) Immunological Characterization of the Spike Protein of the Severe Acute
708 Respiratory Syndrome Coronavirus. *J. Clin. Microbiol.* 42, 1570–1576
709 47 Prabakaran, P. *et al.* (2006) Structure of severe acute respiratory syndrome coronavirus receptor-

710 binding domain complexed with neutralizing antibody. *J. Biol. Chem.* 281, 15829–15836
711 48 Anthony, S.J. *et al.* (2017) Global patterns in coronavirus diversity. *Virus Evol.* DOI:
712 10.1093/ve/vex012
713 49 Quan, P.L. *et al.* (2010) Identification of a severe acute respiratory syndrome coronavirus-like
714 virus in a leaf-nosed bat in Nigeria. *MBio* DOI: 10.1128/mBio.00208-10
715 50 Watanabe, S. *et al.* (2010) Bat coronaviruses and experimental infection of bats, the Philippines.
716 *Emerg. Infect. Dis.* DOI: 10.3201/eid1608.100208
717 51 De Souza Luna, L.K. *et al.* (2007) Generic detection of coronaviruses and differentiation at the
718 prototype strain level by reverse transcription-PCR and nonfluorescent low-density microarray. *J.*
719 *Clin. Microbiol.* DOI: 10.1128/JCM.02426-06
720 52 Mao, X. *et al.* (2016) Differential introgression suggests candidate beneficial and barrier
721 locibetween two parapatric subspecies of Pearson′s horseshoe bat Rhinolophuspearsoni. *Curr.*
722 *Zool.* 62, 405
723 53 Mao, X. *et al.* (2013) Lineage Divergence and Historical Gene Flow in the Chinese Horseshoe Bat
724 (Rhinolophus sinicus). *PLoS One* 8,
725 54 Mao, X. *et al.* (2014) Differential introgression among loci across a hybrid zone of the
726 intermediate horseshoe bat (Rhinolophus affinis). *BMC Evol. Biol.* 14, 154
727 55 Mao, X. *et al.* (2013) Multiple cases of asymmetric introgression among horseshoe bats detected
728 by phylogenetic conflicts across loci. *Biol. J. Linn. Soc.* 110, 346–361
729 56 MAO, X. *et al.* (2010) Historical male-mediated introgression in horseshoe bats revealed by
730 multilocus DNA sequence data. *Mol. Ecol.* 19, 1352–1366
731 57 Dool, S.E. *et al.* (2016) Nuclear introns outperform mitochondrial DNA in inter-specific
732 phylogenetic reconstruction: Lessons from horseshoe bats (Rhinolophidae: Chiroptera). *Mol.*
733 *Phylogenet. Evol.* 97, 196–212
734 58 Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced
735 methodologies and community practices. , *Methods*, 102. 01-Jun-(2016) , Academic Press Inc., 3–
736 11
737 59 Letko, M. *et al.* (2018) Adaptive Evolution of MERS-CoV to Species Variation in DPP4. *Cell*
738 *Rep.* 24, 1730–1737
739 60 Takada, A. *et al.* (1997) A system for functional analysis of Ebola virus glycoprotein. *Proc. Natl.*
740 *Acad. Sci. U. S. A.* 94, 14764–14769
741 61 Pieper, U. *et al.* (2011) ModBase,a database of annotated comparative protein structure
742 models,and associated resources. *Nucleic Acids Res.* DOI: 10.1093/nar/gkq1091
743 62 Pettersen, E.F. *et al.* (2004) UCSF Chimera - A visualization system for exploratory research and
744 analysis. *J. Comput. Chem.* DOI: 10.1002/jcc.20084
745 63 Agnarsson, I. *et al.* (2011) A time-calibrated species-level phylogeny of bats (chiroptera,
746 mammalia). *PLoS Curr.* DOI: 10.1371/currents.RRN1212
747 64 Lan, J. *et al.* (2020) Structure of the SARS-CoV-2 spike receptor-binding domain bound to the
748 ACE2 receptor. *Nature* DOI: 10.1038/s41586-020-2180-5
749 65 Hoffmann, M. *et al.* (2020) A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is
750 Essential for Infection of Human Lung Cells. *Mol. Cell* 78, 779-784.e5
751