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

Wells, H.L[1]*; Letko, M[2,3]; Lasso, G[4]; Ssebide, B[5]; Nziza, J[5]; Byarugaba, D.K[6,7]; Navarrete-Macias[8], I; Liang, E[8]; 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[1,13]*

1. Department of Ecology, Evolution, and Environmental Biology, Columbia University, New York, NY, USA
2. Laboratory of Virology, Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, MT, USA
3. Paul G. Allen School for Global Animal Health, Washington State University, Pullman, WA, USA
4. Department of Microbiology and Immunology, Albert Einstein College of Medicine, New York, NY 10461, USA
5. Gorilla Doctors, c/o MGVP, Inc., Davis, California, USA
6. Makerere University Walter Reed Project, Kampala, Uganda
7. Makerere University, College of Veterinary Medicine, Kampala, Uganda
8. Center for Infection and Immunity, Mailman School of Public Health, Columbia University, New York, NY, USA
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 Pathology, Microbiology, and Immunology, School of Veterinary Medicine, University of California Davis, California, USA

* Co-corresponding authors. Email: hlw2124@columbia.edu, sjanthony@ucdavis.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 sequences of a novel sarbecovirus 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 best explained by recombination not of SARS-CoV-2, but of SARS-CoV-1 and its relatives. 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 lineage including 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 such as Africa, where these viruses have only recently been discovered.

48 **Introduction**

49 The recent emergence of *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2) in China and
50 its rapid spread around the world demonstrates that coronaviruses (CoVs) from wildlife remain an urgent
51 threat to global public health and economic stability. In particular, coronaviruses from the subgenus
52 *Sarbecovirus* (which includes SARS-CoV-2, SARS-CoV-1, numerous bat viruses, and a small number of
53 pangolin viruses) [1] are considered to be a high-risk group for potential emergence. As both
54 sarbecoviruses that have caused human disease (SARS-CoV-1 and -2) use angiotensin-converting enzyme
55 2 (ACE2) as their cellular receptor [2,3], the evolution of this trait is of particular importance for
56 understanding the emergence pathway for sarbecoviruses. Bat SARS-like coronavirus Rp3 is a
57 phylogenetically close relative of SARS-CoV-1 but is unable to bind human ACE2 (hACE2) *in vitro* [4].
58 In contrast, other close relatives of SARS-CoV-1, including bat SARS-like coronavirus WIV1 and
59 WIV16, do have the capacity to bind hACE2 [5,6]. A number of other SARS-CoV-1-like viruses have
60 also been tested for their ability to utilize hACE2 [7–9] and comparison of their spike protein sequences
61 shows that viruses that are unable to utilize hACE2 unanimously have one or two deletions in their RBDs
62 that make them structurally very different than those that do use hACE2 [8]. As SARS-CoV-1, Rp3,
63 WIV1, and WIV16 viruses are closely phylogenetically related, the evolutionary mechanism explaining
64 the variation in their ability to utilize hACE2 (and likely also bat ACE2) as a cellular receptor has thus far
65 been unclear.

66

67 Chinese horseshoe bats (*Rhinolophidae*) are thought to be the primary natural reservoir of sarbecoviruses
68 [5,7,10–12]. Bats within this family are also considered to be the source of the progenitor virus to SARS-
69 CoV-1, as related viruses with high sequence identity to SARS-CoV-1 have been sequenced from
70 Rhinolophid bats, although none have high sequence similarity to SARS-CoV-1 across the entire genome
71 [7,13]. 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 by recombination before spilling over into humans [7,13,14]. In
73 particular, one region of SARS-CoV-1 that is known to have a recombinant origin is the spike gene, as a

74 breakpoint has been detected at the junction of ORF1b and the spike [13,15]. The SARS-1-CoV spike is
75 genomically very different from other viruses in the same clade that have large deletions in the receptor
76 binding domain (RBD) and are unable to use hACE2. The exact minor parent that contributed the
77 recombinant region is still unknown, but it was previously hypothesized that the recombination occurred
78 with a yet undiscovered lineage of sarbecoviruses and that this event contributed strongly to its potential
79 for emergence [13,16]. Recombination has also been shown within the spike genes of other CoVs that
80 have spilled over into humans and domestic animals and is potentially an important driver of emergence
81 for all coronaviruses [17–22].
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 [7,23]. 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) [23]. Studies in Europe and Africa have shown that there are
90 distinct sarbecovirus clades in each of these regions as well, herein named "Lineage 4" [24–29].
91 Sarbecoviruses appear to switch easily among co-occurring *Rhinolophus* species [30,31]; however, they
92 appear to rarely occupy more than one geographic area, despite the fact that some of these bat species
93 have widespread distributions 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 [32]. Subsequently, seven full- or near full-length SARS-CoV-2-like viruses were published that had
99 been sampled from Malayan pangolins (*Manis javanica*) in 2017 and 2019 [33,34], one of which has also

100 been tested and found to bind hACE2 [35]. Neither SARS-CoV-2, RaTG13, nor the pangolin CoVs have
101 deletions in their RBDs. In contrast, the most recently described bat virus (RmYN02) is even more
102 closely related to SARS-CoV-2 than RaTG13 in the polymerase gene and was also found in Yunnan
103 Province; however, this sequence has deletions in the RBD and homology modeling suggests it likely
104 does not use hACE2 [36]. Together, these viruses form a fifth phylogenetic lineage ("Lineage 5") that is
105 distinct from all other lineages of sarbecoviruses despite having been detected in Yunnan, where all
106 viruses found until this point had belonged to Lineage 1.
107
108 This finding of overlapping Lineage 1 and Lineage 5 viruses in geographic space is inconsistent with the
109 previously observed pattern of biogeography for sarbecoviruses. SARS-CoV-2 was isolated first from
110 people in Hubei Province and one of the pangolin viruses was isolated from an animal sampled in
111 Guangdong, neither of which are Lineage 1 provinces. However, the true geographic origins of these
112 viruses are unknown as it is possible they were anthropogenically transported to the regions in which they
113 were detected. For example, the Malayan pangolin (*Manis javanica*) has a natural range that reaches
114 southwestern China (Yunnan Province) at its northernmost edge and extends further south into Myanmar,
115 Lao PDR, Thailand, and Vietnam [37]. So, if they were naturally infected (as opposed to infection via
116 wildlife trade), the infection was potentially not acquired from Guangdong Province. Similarly, SARS-
117 CoV-2 cannot be guaranteed to have emerged from bats in Hubei Province, as humans are highly mobile
118 and the exact spillover event was not observed. If the clade containing SARS-CoV-2 and its close
119 relatives is indeed endemic in animals in Yunnan and the nearby Southeast Asian regions as suggested by
120 the presence of RaTG13, RmYN02, and the natural range of the Malayan pangolin, whatever mechanism
121 is facilitating the biogeographical concordance of Lineages 1, 2 and 3 within China appears to no longer
122 apply for the biogeography of Lineage 5, since they all appear to overlap in and around Yunnan Province.
123
124 Here, we report a series of observations that together suggest that SARS-CoV-1 and its close relatives
125 gained the ability to utilize ACE2 through a recombination event that happened between an ancestor of

126 SARS-CoV-1 and a Lineage 5 virus phylogenetically related to SARS-CoV-2, which could only have
127 occurred with the lineages occupying the same geographic and host space. We also report three full-
128 length genomes of sarbecoviruses from Rwanda and Uganda and demonstrate that the RBDs of these
129 viruses are genetically intermediate between viruses that use ACE2 and those that do not. Accordingly,
130 we also investigate the potential for these viruses to utilize hACE2 *in vitro*. Together, our findings help
131 illuminate the evolutionary history of ACE2 usage within sarbecoviruses and provide insight into
132 identifying their risk of emergence in the future. We also propose a mechanism that could explain the
133 pattern of phylogeography across Lineages 1, 2, and 3, and why Lineage 5 viruses (including SARS-
134 CoV-2 and its relatives) represent an inconsistency to this pattern.
135
136 **Results**
137 To better understand the evolutionary history of sarbecoviruses we first constructed a phylogenetic tree of
138 the RNA-dependent RNA polymerase (RdRp) gene, also known as nsp12 (Figure 1). The tree was
139 constructed using sequences from GenBank as well as three sequences of a novel sarbecovirus detected in
140 bats from Uganda and Rwanda as part of the USAID-PREDICT project. The three novel sequences share
141 >99% nucleotide identity to each other and ~76% and ~74% nucleotide identity with SARS-CoV-1 and
142 SARS-CoV-2, respectively. Phylogenetically, they lie within Lineage 4, clustering with previously
143 reported SARS-related coronavirus BtKY72 found in bats in Kenya [29] and bat coronavirus BM48-31
144 from Bulgaria [26]. The topology of the sarbecovirus phylogeny is uncertain with respect to the
145 placement of the Lineage 4 viruses, with some models placing them between Lineage 5 and Lineages 1, 2,
146 and 3, and others placing them at the base of the tree, depending on the methodology and alignment used
147 [3,38,39] (Supplementary Figure S1). Our results place Lineage 4 in the former position with high
148 posterior support for the RdRp gene, though the variability in this placement must be recognized. Figure 1
149 also demonstrates the same geographic pattern of concordance reported by Yu et al [23], where viruses in
150 each lineage show a clear pattern of fidelity with particular geographic regions. However, SARS-CoV-2
151 does not lie within the clade of bat sarbecoviruses that have been detected in bats in China to date but

152 rather forms a much deeper, separate lineage. The discovery of the "Lineage 5" clade containing SARS-
153 CoV-2 and related viruses in pangolins and bats is a deviation from the geographic patterns observed for
154 other sarbecoviruses.
155
156 To investigate the evolutionary history of ACE2 usage, we built a second phylogenetic tree using only the
157 RBD of the spike gene and compared it to the phylogeny of RdRp (Figure 2). This region was selected
158 because the spike protein mediates cell entry and because previous reports showed that SARS-CoV-1 and
159 SARS-CoV-2 both use hACE2, despite being distantly related in the RdRp [2,3]. Within the RBD region
160 of the genome, SARS-CoV-1 and all ACE2-using viruses are much more closely related to SARS-CoV-2
161 than to other Lineage 1 viruses (Figure 2). Interestingly, bat virus RmYN02 is no longer associated with
162 SARS-CoV-2 in the RBD and is instead within the clade of non-ACE2-using viruses. We also found that
163 within the RBD, ACE2-using viruses and non-ACE2-using viruses are perfectly phylogenetically
164 separated. The viruses from Africa and Europe form a distinct clade that is intermediate between the
165 ACE2-using and non-ACE2-using groups, but appears more closely related to the ACE2-using group.
166
167 While these viruses from Africa and Europe are slightly more similar to the ACE2-using group, they
168 differ somewhat in amino acid sequence from the ACE2-users at the binding interface, including a small
169 deletion in the middle of the sequence (Figure 5, region 2). Thus, to determine the ability of these
170 sarbecoviruses to use hACE2 and better delineate the boundaries of ACE2 usage, we performed *in vitro*
171 experiments in which we replaced the RBD of SARS-CoV-1 with the RBD from the Uganda (PDF-2370,
172 PDF-2386) and Rwanda viruses (PRD-0038) [8]. Single-cycle Vesicular Stomatitis Virus (VSV) reporter
173 particles containing the recombinant SARS-Uganda and SARS-Rwanda spike proteins were then used to
174 infect BHK cells expressing hACE2. While VSV-SARS-CoV-1 showed efficient usage of hACE2, VSV-
175 Uganda and VSV-Rwanda did not (Figure 3).
176

177 To try and explain why the African sarbecoviruses are unable to use hACE2, we modeled the RBD
178 domain of the sequences from Uganda (PDF-2370, PDF-2386) and Rwanda (PRD-0038). Unlike other
179 non-ACE2 binders, homology modeling suggests that the RBDs of these viruses from Africa are
180 structurally similar to SARS-CoV-1 and SARS-CoV-2 (Figure 4A). However, modeling the interaction
181 with hACE2 reveals amino acid differences at key interfacial positions that can help explain the lack of
182 interaction observed for the rVSV-Uganda and rVSV-Rwanda viruses (Figure 4B-C). There are four
183 regions of the RBD that lie within 10Å of the interface with hACE2, one of which is the receptor binding
184 ridge (SARS-CoV-1 residues 459-477) that is critical for hACE2 binding [32,40]. We have designated the
185 remaining regions as regions 1 (residues 390-408), 2 (residues 426-443), and 3 (residues 478-491) (Figure
186 5).
187
188 The sarbecoviruses from Africa evaluated here have a 2-3 amino acid deletion (SARS-CoV-1 residues
189 434-436) in region 2 (Figure 5). As many of the residues in this region make close contact with hACE2
190 (<5Å), it is possible that this contributes to the disruption of hACE2 binding. One of these residues,
191 Y436, establishes hydrogen bonds with human ACE residues D38 and Q42 in both SARS-CoV-1 and
192 SARS-CoV2 (Figure 4C). Notably, all other non-ACE2 binders also have deletions in residues 432-436.
193 While this deletion is thought to interfere or reduce binding, restoring a similar deletion (SARS-CoV-1
194 residues 432-437) in the S protein of a European CoV (BM48-31) with the corresponding consensus
195 segment obtained from Lineage 1 ACE2-binding viruses did not restore hACE2-mediated entry; only
196 replacing the receptor-binding motif (RBM) increased hACE2-mediated entry [8].
197
198 Moreover, sarbecoviruses from Africa contain additional amino acid changes at the interface that can also
199 contribute to hACE2 binding disruption (Figure 4C). hACE2 contains two hotspots (K31 and K353) that
200 are crucial targets for binding by SARS-RBDs and amino acid variations in the RBD sequence enclosing
201 these ACE2 hotspots have been shown to shape viral infectivity, pathogenesis, and determine the host
202 range of SARS-CoV-1 [41–43]. All sarbecoviruses from Africa contain a Lys (K) at SARS-CoV-1

203 position 479 within region 3 (positions 481 and 482 for Uganda and Rwanda, respectively), which makes
204 contact with these ACE2 hotspots (as compared to N479 or Q493 in SARS-CoV-1 and 2 respectively;
205 Figure 4C). K479 decreases binding affinity by more than 20-fold in SARS-CoV-1 [44]. The negative
206 contribution of K479 in region 3 is likely due to unfavorable electrostatic contributions with ACE2
207 hotspot K31 (Figure 4C) [42,45]. On the other hand, SARS-CoV-1 residue T487 (N501 in SARS-CoV-2)
208 interacts with ACE2 hotspot K353 and has a Val (V) in the viruses from Africa (residues 489 and 490)
209 (Figure 5). As with residue 479, the amino acid identity at position 487 contributes to the enhanced
210 hACE2 binding observed in SARS-CoV-2 [42,43,45]. The presence of a hydrophobic residue at position
211 487, not previously observed in any ACE2 binding sarbecovirus, might lead to a local rearrangement at
212 the K353 hotspot that hinders hACE2 binding. Indeed, most non-ACE2 binders have a Val (V) in SARS-
213 CoV-1 position 487 (Figure 5).
214
215 Finally, the receptor binding ridge, which is conspicuously absent from all non-ACE2 binders, is present
216 in the sarbecoviruses from Africa but has amino acid variations that differ significantly from both SARS-
217 CoV-1 and SARS-CoV-2 (Figure 5). Changes in the structure of this ridge contribute to increased binding
218 affinity of SARS-CoV-2, as a Pro-Pro-Ala (PPA) motif in SARS-CoV-1 (residues 469-471) replaced with
219 Gly-Val-Glu-Gly (GVEG) in SARS-CoV-2 results in a more compact loop and better binding with
220 hACE2 [32]. Changes within this ridge may be negatively contributing to hACE2 binding of viruses from
221 Africa, which have Ser-Thr-Ser-Gln (STSQ) or Ser-Iso-Ser-Gln (SISQ) in this position (Figure 4C and 5).
222
223 While our studies suggest that these viruses from Africa do not utilize hACE2, it is not clear whether they
224 are still ACE2-users but are adapted to divergent forms of bat ACE2 in their natural hosts. The specific
225 bat host species for the Uganda and Rwanda viruses reported here could not be definitively identified in
226 the field or in the lab, but are all genetically identical. They may represent a cryptic species, as the
227 mitochondrial sequences are ~94% identical with *Rhinolophus ferrumequinum* in the cytochrome oxidase
228 I gene (COI) and ~96% identical with *Rhinolophus clivosus* in the cytochrome b (cytb) gene, each of

229 which have been deposited in GenBank (accessions MT738926-MT738928, MT732776). We were also
230 able to extract ACE2 sequences from the deep sequencing reads of PDF-2370 (GenBank accession
231 MW183243) to compare it to ACE2 sequences from species that are known to host ACE2 binders
232 (human, civet, pangolin), non-ACE2 binders (*R. macrotis, pearsonii, pusillus, ferrumequinum*), and both
233 (*R. sinicus*). Comparison of the ACE2 sequences shows that they are highly similar, with only a few
234 amino acids that are changed in hosts of viruses that utilize ACE2 compared to the host of our African bat
235 sample (Supplementary File 1). *R. sinicus* in particular is a known host of viruses that utilize ACE2 as
236 well as viruses with the deletions that do not, suggesting that adaptation to divergent bat ACE2 is not a
237 likely explanation for the deviation in sequence and structure of the RBD of viruses with deletions,
238 including the novel sarbecoviruses from Uganda and Rwanda. These findings provide additional
239 structural evidence that aids in distinguishing viruses which bind ACE2 from those that do not. They also
240 demonstrate that ACE2 usage within sarbecoviruses is restricted to those viruses within the SARS-CoV-1
241 and SARS-CoV-2 clade in the RBD (Lineages 1 and 5, Figure 2).
242
243 The finding of discordant evolutionary trees for RdRp and the RBD in Figure 2 more strongly supports a
244 recombination scenario; however, to consider an alternate scenario where ACE2 usage arose in Lineages
245 1 and 5 independently through convergent evolution, we compared the RdRp phylogeny with the amino
246 acid sequences of the interfacial residues in the RBD (Figure 5). When mapped to the RdRp tree, the
247 'extra' RBD sequence present in the ACE2-using viruses is conspicuous within the Lineage 1 clade of
248 otherwise non-ACE2-using viruses that have large deletions. We also note that there are two distinct
249 groups of RBD sequences within ACE2-using Lineage 1 viruses: Type 1, containing SARS-CoV-1,
250 SARS-SZ3 (civet), Rs3367, WIV1, Rs7327, YN2018B, Rs9401, WIV16, Rs4874, and LYRa11, and
251 Type 2, containing Rs4231, Rs4084, and RsSHC014. Further, RmYN02 is within the Lineage 5 clade of
252 ACE2-using viruses in RdRp but its RBD sequence contains both deletions (Figure 5). Without
253 recombination, the viruses with deletions in region 2 and in the receptor binding ridge would have had to
254 be gained and lost in precisely the same positions for ACE2-using Lineage 1 viruses and RmYN02,

255 respectively, which is not a parsimonious explanation. The phylogeny and sequence in Figure 5 also
256 illustrate that ACE2-usage appears to be an ancestral trait conserved in Lineage 5 [39] and a derived trait
257 in each of the 13 Lineage 1 viruses with ACE2-using structure.
258
259 Finally, we further investigated support for the recombination scenario by examining the region of
260 sequence between RdRp and the RBD for possible breakpoints. Only the 13 Lineage 1 viruses with
261 ACE2-using structure were targets of this analysis as we were primarily interested in explaining the
262 discordant phylogeny and variation in ACE2 usage (Figure 2), not in fully describing the recombination
263 history of every sarbecovirus. Using 3SEQ, we show that all of the ACE2-using Lineage 1 sequences
264 show extensive evidence of recombination within S1 and the RBD specifically (Table 2, Figure 6A).
265 Further, the assignment of the parental sequence that donated the recombinant region (the minor parent)
266 always resulted in the identification of one of the other recombinant sequences. This would not have been
267 possible, as the recombinant region would have had to come from somewhere other than these 13
268 sequences, indicating that the true minor parent does not exist in our alignment. Using these breakpoints,
269 we designated six subregions that were relatively free of recombination within these 13 sequences,
270 mirroring the approach of Boni et al. 2020 [39], and built phylogenetic trees for each region. We show
271 that in orf1ab (region A) and S2 (region F) these 13 sequences fall within Lineage 1, but within S1 and
272 particularly the RBD (B through E) they switch phylogenetic positions and cluster with Lineage 5 (Figure
273 6B), supporting the recombination scenario.
274
275 Despite only investigating the Lineage 1 recombinants for the locations of sequence breakpoints, the
276 phylogenetic trees provide evidence that recombination has occurred frequently in other sarbecoviruses in
277 this genomic region as well (Figure 6B). Of note, Rs4084 and RsSHC014 cluster with Type 1 RBDs in
278 regions B, C, and D, but with swap to cluster with Rs4231 (Type 2) in Region E, even though Rs4084,
279 RsSHC014, WIV1, and Rs3367 are all nearly identical in every other region. This suggests that a
280 WIV1/Rs3367-like Type 1 virus which had already undergone recombination in regions B through E

281 underwent a second recombination event with a Type 2 virus on top of the first in region E. A number of
282 other viruses also appear to have recombinant history in regions B, C, and D (SL-CoVZC45 and SL-
283 CoVZXC21, YN2013, Anlong-103, and Anlong 112), but these viruses do not show evidence of
284 recombination that spans the RBD in region E, which contains the amino acid deletions in region 2 and
285 the receptor binding ridge and appears to primarily determine ACE2-using potential. The frequency of
286 recombination in this region among Lineage 1 viruses strongly supports the hypothesis that after ACE2-
287 usage was acquired in Lineage 1, it subsequently spread throughout the clade via additional
288 recombination events with other Lineage 1 viruses.
289
290 As all of our evidence supports a recombination scenario over convergent evolution, we sought to
291 construct a possible timeline of events that could explain our observations. Using tip dating in BEAST2,
292 we constructed a time-calibrated phylogeny for RdRp using a substitution rate prior inferred from Boni et
293 al. 2020 [39]. Using the RdRp tree as an evolutionary backbone, the deletions in region 2 and the receptor
294 binding ridge of the RBD appear to have been lost in a stepwise fashion (Figure 5). The small deletion in
295 region 2 likely arose first, before the diversification of Lineage 4 in Africa and Europe (Figure 5) and was
296 dated using the MRCA of Lineages 1, 2, 3 and 4 (Figure 8). Alternatively, as the boundaries of the
297 deletion in region 2 in Lineage 4 and Lineages 1, 2, and 3 do not align perfectly and there is uncertainty in
298 the position of this branch in the phylogeny, it is equally possible that this deletion was lost independently
299 in Lineage 4. The larger deletion in the receptor binding ridge, not present in known sequences from
300 Lineage 4, likely arose second, but before the diversification of Lineages 1, 2, and 3 (Figure 5) and was
301 dated with the MRCA of these three lineages (Figure 8). Because no ACE2-using viruses have been
302 discovered in Lineage 2 or 3 to date, we propose that the re-appearance of this trait arose after the MRCA
303 of Lineage 1 on the tree (Figure 8). As SARS-CoV-1 was the earliest Lineage 1 virus sequenced with
304 ACE2-using structure, the emergence of ACE2 usage in Lineage 1 must have occurred in the time
305 between the MRCA of Lineage 1 (1852, 95% HPD 1804-1901) and the emergence of SARS-CoV-1 in
306 2003.

307

308 Next, we constructed a time-calibrated phylogeny for RBD with a strict MRCA age prior informed by the
309 estimation of the tree height in RdRp (see *Methods*), such that the timescale would be comparable even
310 though the evolutionary rates between these two regions likely are not the same (Figure 7). To account for
311 variability in lineage-specific substitution rates, we also generated a time-calibrated model using a relaxed
312 lognormal clock (Figure 7). Comparing the time-calibrated RBD tree to the time-calibrated RdRp tree, the
313 divergence dates for the two types of RBD sequence observed in the recombinant Lineage 1 sequences
314 are incompatible, suggesting that more than one recombination event donating ACE2 usage from Lineage
315 5 to Lineage 1 must have occurred. The 13 Lineage 1 recombinants (both Type 1 and Type 2) coalesce
316 between 119-216 years ago in RdRp and between 259-490 years ago in the RBD (Figure 7). If these time
317 estimates reflect true rates of diversification, a single introduction of the ACE2-using phenotype via
318 recombination would not allow enough time for the sequence divergence between Type 1 and Type 2
319 RBDs to accumulate, even when accounting for the substitution rate in RBD being estimated as an order
320 of magnitude higher than that of RdRp (5.248e-4 in RdRp, 2.181e-3 in RBD). Further, the substitution
321 rate that would be needed for the observed sequence divergence in the RBD of the 13 recombinants to
322 have accumulated since their MRCA in RdRp (1852) is more than double the estimated rate of our time-
323 calibrated tree (5.899e-3). Even with a relaxed clock assumption, the maximum value of the posterior
324 distribution of the mean rate is only 4.733e-3. From this, we conclude that two independent
325 recombination events occurred between Lineage 5 and Lineage 1 resulting in two distinct RBD types.
326

327 We propose two main hypotheses for the acquisition and spread of the two distinct RBD types donating
328 ACE2 usage from Lineage 5 to Lineage 1. The recombination hypothesis posits that two recombination
329 events donated Type 1 and Type 2 RBD sequence from Lineage 5 to Lineage 1; however, these two
330 events are insufficient to explain the non-monophyletic pattern of ACE2 usage in Lineage 1. We further
331 hypothesize that whichever Lineage 1 virus first gained Type 1 and Type 2 ACE2 usage in each group
332 then donated the trait to other Lineage 1 viruses through subsequent recombination events (Figure 8). It is

333 difficult to approximate a date for such an event, but the MRCA of the Type 1 recombinants in the RBD
334 may be a close estimation (between 42 and 77 years ago) (Figure 7). The events must have been recent
335 enough that the observed diversity of Type 2 RBD sequences is quite low, yet not so recent such that
336 there would not have been time for recombination to have occurred twice in region E for sequences
337 Rs4084 and RsSHC014 (Figure 6B).
338
339 The second hypothesis and only remaining possibility for ACE2 usage in Lineage 1 (besides
340 convergence) is that perhaps the trait persisted in this Lineage from the ancestral state (Figure 8). Because
341 no viruses demonstrating ACE2 usage have been discovered in Lineages 2, 3, and 4, this would mean that
342 the ACE2 usage trait would have been lost via deletion in these lineages. Further, because of the non-
343 monophyletic branching order of these lineages, this would require multiple independent and identical
344 losses of the region 2 and receptor binding ridge deletions in all three of these lineages. If this did indeed
345 occur, in order to then observe the pattern of ACE2 usage in Lineage 1 where some viruses, but not all,
346 have the ACE2 usage trait, further independent losses would be required in individual viruses. In much
347 the same manner as convergence would require multiple independent and identical events, persistence of
348 ACE2 usage with multiple independent deletions for the entire clades of Lineages 2, 3, and 4 and only
349 some of the viruses in Lineage 1 is also highly non-parsimonious. Persistence is also a poor explanation
350 for the pattern of the two RBD types observed, particularly for Type 2, where the RBD sequences are
351 highly similar but the RdRp sequences are quite divergent. If both genes were vertically inherited via
352 persistence, we would expect these genes to have approximately equal MRCA ages. Instead, we observe
353 that the MRCA age for Type 2 RBDs in region E are much younger than for RdRp.
354
355 **Discussion**
356 *ACE2 usage in Lineage 1 viruses was acquired via recombination*
357 At first glance, ACE2 usage does not appear to be phylogenetically conserved among sarbecoviruses,
358 especially since many phylogenies are built using RdRp. This naturally leads to the hypothesis that ACE2

359 usage arose independently in SARS-CoV-1 and SARS-CoV-2 via convergent evolution. This has been
360 suggested previously for another ACE2-using human coronavirus, NL63 [46]. However, a phylogeny
361 constructed using the RBD perfectly separates viruses that have been shown to utilize ACE2 from those
362 that do not (Figure 2). Viruses that cannot utilize ACE2 have significant differences in their RBDs,
363 including large deletions in critical interfacial residues and low amino acid identity with viruses that do
364 use ACE2 (Figure 5). Notably, in addition to the large deletions, viruses that cannot use ACE2 deviate
365 considerably at the interacting surface, including positions that play fundamental roles dictating binding
366 and cross-species transmission [32,41,44,47]. It is unknown whether viruses that cannot use hACE2 are
367 utilizing bat ACE2 or an entirely different receptor altogether, but since mammalian ACE2 is so
368 conserved [48,49] and ACE2-using viruses demonstrate broad host tropism [42,50–52], we hypothesize
369 that there is likely a different receptor involved for the non-ACE2 users (see Supplementary File 1).
370
371 The difference in topology, specifically in the positioning of ACE2-using Lineage 1 viruses, between
372 RdRp and RBD trees suggests that the ability to use ACE2 was introduced into Lineage 1 by
373 recombination between a recent ancestor of the ACE2-using Lineage 1 viruses (including SARS-CoV-1)
374 and an undiscovered Lineage 5 virus in the RBD. As there are two types of closely related RBD
375 sequences in the recombinant Lineage 1 viruses (Figure 2) with incompatible divergence dates (Figure 7),
376 we suggest that two such recombination events occurred between Lineage 1 and Lineage 5 (Figure 8)
377 independently introducing ACE2-usage into Lineage 1. The non-monophyletic nature of ACE2 usage
378 within Lineage 1 can then be most parsimoniously explained by secondary intra-lineage recombination
379 events (Figure 8). It is possible that both hypotheses are partially true and that both intra-lineage
380 recombination as well as the persistence of this trait alongside sister Lineage 1 viruses without the trait
381 gave rise to the observed patterns of Type 1 and Type 2 ACE2 usage within Lineage 1. It is also very
382 possible that further sampling may illuminate that some of the events proposed here have been distorted
383 by sampling bias. We have estimated that these events may have occurred roughly within the last two
384 centuries, though this estimate will likely change with further sampling as well. Our intention is not

385 necessarily to date these events exactly, but rather to infer their order relative to each other and to make
386 hypotheses based on this order of events. Confidence intervals for many node dates overlap, but high
387 posterior probabilities on internal nodes indicate that events most likely occurred in a certain order.
388
389 Our conclusion that ACE2 usage originated in Lineage 5 and was introduced into Lineage 1 by
390 recombination is based on phylogenetics; however, studies of recombination using phylogenetics are
391 often limited in their ability to definitively determine the direction of recombination. Nonetheless, there
392 are several lines of evidence that support the direction having occurred from Lineage 5 to Lineage 1.
393 First, recombination is notoriously more frequent in spike compared to orf1ab [39,53,54]. Second,
394 Lineage 5 constitutes the base of the tree and has the oldest MRCA, meaning it likely shares more
395 ancestral traits with the MRCA of all sarbecoviruses. Third, phylogenetic topology in orf1ab before the
396 recombinant region of the genome mirrors that of S2 after the recombinant region (Figure 6A), orienting
397 orf1ab/S2 as sequence from the major parent of the recombination event. And finally, that spike is the
398 recombinant region as opposed to RdRp is also supported by numerous studies that have provided
399 evidence that SARS-CoV-1 is recombinant and SARS-CoV-2 is not [3,13,15,55].
400
401 In order for recombination to have occurred between Lineage 1 and Lineage 5, these viruses must have
402 had the opportunity to coinfect the same host cell. We demonstrate that recombination is possible given
403 that viruses related to SARS-CoV-1 and -2 appear to share both geographic and host space in
404 southwestern China and in *R. sinicus* and *R. affinis* bats. Highlighting that this previously known
405 recombination event (i.e. SARS-CoV-1) occurred with a previously unknown group of viruses that are
406 related to SARS-CoV-2 is an important finding of this study and demonstrates that recombination is an
407 important driver of spillover for sarbecoviruses.
408
409 *A series of deletion events most likely resulted in the ancestral loss of ACE2 usage in Lineages 1-4*

410 Using the RdRp tree as the evolutionary history to which to compare because of its stability and relative
411 lack of recombination, sequences without the deletions in the RBD most likely represent the ancestral
412 state, as the SARS-CoV-2 Lineage 5 viruses at the base of the tree do not show this trait (Figure 2). This
413 is in accordance with the findings of Boni et al. [39]. Alternatively, it is possible that the deletion state is
414 the ancestral state, and that this ancestral deletion state was conserved in Lineages 1, 2, and 3; however,
415 insertions acquired during the evolution of Lineages 4 and 5 would have had to have occurred
416 independently, which is less parsimonious. Persistence of the ACE2 usage trait from the MRCA of
417 Lineage 5 all the way to Lineage 1 is also not parsimonious, as the RBD deletions would have had to have
418 been lost many times independently (Figure 8).
419
420 Further, the viruses from bats in Africa and Europe have one of the two deletions, which may indicate that
421 these are descendant from an evolutionary intermediate and support a stepwise deletion hypothesis;
422 however, this hypothesis hinges completely on the uncertain positioning of Lineage 4 on the phylogeny,
423 which may support independent deletion within region 2 in Lineage 4 instead. Since ACE2-using Lineage
424 1 viruses including SARS-CoV-1 are nested within a clade of viruses that all have both deletions, this
425 implies that both deletions arose before the diversification of Lineages 1, 2, and 3 viruses (Figures 5 and
426 8). According to the branching order shown here, the smaller deletion in region 2 was likely acquired
427 earliest, before the diversification of the clades into Africa and Europe, since it is shared by all clades
428 with the exception of SARS-CoV-2 Lineage 5 at the base of the tree (Figure 5). These large deletions in
429 the RBD-ACE2 interface and the similarity of Rhinolophid and hACE2 also suggest that non-ACE2-
430 using viruses, including Lineages 1, 2, 3, and 4, are using at least one receptor other than ACE2 [8,36].
431
432 *ACE2 usage is not well explained by convergent evolution*
433 Under a hypothetical convergent evolution scenario, large insertions would have had to be reacquired in
434 precisely the same regions from which they were lost within the RBD independently in ACE2-using
435 Lineage 1 viruses. The most parsimonious argument is that ACE2-using Lineage 1 viruses are descendent

436 from at least two recombinant viruses (containing Types 1 and 2 RBDs) and that recombination best
437 explains the non-monophyletic pattern of ACE2 usage within the *Sarbecovirus* subgenus. In contrast,
438 human coronavirus NL63 is an alphacoronavirus that is also a hACE2 user but most likely represents a
439 true case of convergent evolution. The RBD of SARS-CoV-1 and SARS-CoV-2 are structurally identical,
440 while NL63 has a different structural fold, suggesting that they are not evolutionarily homologous [46].
441 Nonetheless, NL63 also binds to hACE2 in the same region – suggesting all of the ACE2-using viruses
442 have converged towards this interaction mode [46].
443
444 Additional evidence supports a recombination scenario over convergent evolution, including (i) the
445 detection of statistically supported recombination breakpoints in all ACE2-using Lineage 1 viruses
446 between RdRp and the RBD, and (ii) a growing number of reports identifying recombination in the spike
447 gene of other CoVs [22,56–59]. We also highlight an additional unreported recombination event between
448 Lineage 5 and Lineage 1 giving rise to RmYN02 that further demonstrates the importance of this
449 evolutionary mechanism. We observed that the Lineage 5 bat virus RmYN02, which is highly similar to
450 SARS-CoV-2 within the RdRp, actually has a RBD with the Lineage 1 deletion trait associated with the
451 inability to use ACE2. This indicates a recombination in the opposite direction, From Lineage 1 to
452 Lineage 5, and is again consistent with their overlapping host and geographic ranges. The RmYN02 virus
453 was sequenced from a pooled sample that also contained a second strain, RmYN01, so the possibility that
454 the assembled RmYN02 sequence is chimeric cannot be ruled out. However, both RmYN01 and
455 RmYN02 have deletions in the RBD, so whether or not the sequence is chimeric, it is most likely still
456 recombinant. Again, recombination is a much more parsimonious explanation for the loss of ACE2 usage
457 in RmYN02 rather than convergence, which would require independent and identical deletions in the
458 interfacial residues of the RBD.
459
460 *Differences in receptor usage within sarbecoviruses would explain observed phylogeographic patterns*

461 Lineage 1 and Lineage 5 viruses appear to occupy the same geographic space, which is necessary for the
462 opportunity to recombine to exist. However, the co-circulation of these distantly phylogenetically related
463 viruses is a notable deviation from previous observations that show sarbecovirus phylogeny mirrors
464 geography. It is unknown why Lineages 1-4 show strong phylogeographic clustering. Isolation by
465 distance (IBD) is one ecological mechanism that could explain concordance between phylogeny and
466 geography; however, this would not explain why Lineage 5 deviates from this pattern and overlaps
467 geographically with Lineage 1. Instead, we hypothesize that immune cross-reactivity between closely
468 related viruses within hosts results in indirect competitive exclusion and priority effects, and that this
469 explains the phylogeographic signal of Lineages 1-3. Antibodies against the spike protein are critical
470 components of the immune response against CoVs [60–62]. Hosts that have been infected by one
471 sarbecovirus may be immunologically resistant to infection from a related sarbecovirus, leading to
472 geographic exclusion of closely related strains and a pattern of evolution that is concordant with
473 geography despite the fact that species and individuals are not strictly confined (Figure 1). It is unlikely
474 that this pattern is caused by differing competencies amongst *Rhinolophus* bats, as host-switching of these
475 viruses appears to be common. The co-circulation of Lineage 5 viruses (including SARS-CoV-2 and
476 related viruses) in the same species and the same geographic location as Lineage 1 viruses may suggest a
477 release in the competitive interactions maintaining geographic specificity. This would preclude
478 recognition by cross-reactive antibodies, such as those produced against the spike protein, and may be
479 evolutionarily advantageous for the recombinant virus. Furthermore, if these two groups of viruses utilize
480 different receptors, antibodies against one would be ineffective at excluding the other, potentially
481 allowing both viral groups to infect the same hosts. If competitive release has indeed occurred among
482 these viruses, it is likely that the SARS-CoV-2 clade is potentially much more diverse and geographically
483 widespread than currently understood.
484
485 *Implications for future research*

486 Here, we highlight the critical need for further surveillance specifically in southwestern China and
487 surrounding regions in southeast Asia given that all ACE2-using bat viruses discovered to date were
488 isolated from bats in Yunnan Province. If this holds true, it would support the hypothesis that SARS-
489 CoV-2 originated in Yunnan or the surrounding regions of southwest China before the initial epidemic
490 then amplified in Wuhan. Southeast Asia and parts of Europe and Africa have been previously identified
491 as hotspots for sarbecoviruses [63], but increased surveillance will help characterize the true range of
492 ACE2-using sarbecoviruses in particular. The receptors for viruses from northern China and other regions
493 such as Europe and Africa remain unknown, and may not pose a threat to human health if they cannot
494 utilize hACE2, though their potential to acquire hACE2-usage by recombination should be considered
495 along with the potential for their existing spike proteins to use other human receptors for cell entry. It is
496 unclear whether the lack of hACE2 binding for sarbecoviruses from Uganda and Rwanda is due to the
497 small deletion in region 2 or to the numerous amino acid changes in other interfacial residues. It is
498 possible that sarbecoviruses in Africa with different residues in these interfacial regions could potentially
499 still use hACE2. It is also unknown whether the sarbecoviruses from Africa in particular use a different
500 receptor altogether, or whether sarbecoviruses with the potential to utilize hACE2 without the region 2
501 deletion have also diversified into Africa or Europe. If competitive release between groups of viruses
502 utilizing different receptors has indeed occurred, further surveillance is needed to determine the true
503 extent of Lineage 5 viruses. In addition, experimental evidence to support or refute a competitive release
504 hypothesis should be prioritized.
505
506 This study highlights that hACE2 usage is unpredictable using phylogenetic proximity to SARS-CoV-1 or
507 SARS-CoV-2 in the RdRp gene. This is due to vastly different evolutionary histories in different parts of
508 the viral genome due to recombination. Phylogenetic relatedness in the RdRp gene is not an appropriate
509 proxy for pandemic potential among CoVs (the 'nearest neighbor' hypothesis). By extension, the
510 consensus PCR assays most commonly used for surveillance and discovery, which mostly generate a
511 small fragment of sequence from within this gene [64–66], are insufficient to predict hACE2 usage. Using

512 phylogenetic distance in RdRp as a quantitative metric to predict the potential for emergence is tempting
513 because of the large amount of data available, but this approach is unlikely to capture the biological
514 underpinnings of emergence potential compared to more robust data sources such as full viral genome
515 sequences. The current collection of full-length sarbecovirus genomes is heavily weighted toward China
516 and *Rhinolophus* hosts, despite evidence of sarbecoviruses prevalent outside of China (such as in Africa)
517 and in other mammalian hosts (such as pangolins). Further, investigations into determinants of
518 pathogenicity and transmission for CoVs and the genomic signatures of such features will be an important
519 step towards the prediction of viruses with spillover potential, and distinguishing those with pandemic
520 potential.
521
522 Finally, these findings reiterate the importance of recombination as a driver of spillover and emergence,
523 particularly in the spike gene. If SARS-CoV-1 gained the ability to use hACE2 through recombination,
524 other non-ACE2-using viruses could become human health threats through recombination as well. We
525 know that recombination occurs much more frequently than just this single event with SARS-CoV-1, as
526 the RdRp phylogeny does not mirror host phylogeny and the RBD tree has significantly different
527 topology across all geographic lineages. In addition, the bat virus RmYN02 appears to be recombinant in
528 the opposite direction (Lineage 5 backbone with Lineage 1 RBD) [36], again supporting the hypothesis
529 that recombination occurs between these lineages. Our analyses support two hypotheses: first, that
530 sarbecoviruses frequently undergo recombination in this region of the genome, resulting in this pattern,
531 and second, that sarbecoviruses are commonly shared amongst multiple host species, resulting in a lack of
532 concordance with host species phylogeny and a reasonable opportunity for coinfection and
533 recombination. Bats within the family *Rhinolophidae* have also repeatedly shown evidence of
534 introgression between species [67–72], supporting the hypothesis that many species in this family have
535 close contact with one another which may facilitate viral host switching. Given that we have shown that
536 ACE2-using viruses are co-occurring with a large diversity of non-ACE2-using viruses in Yunnan

537 Province and in a similar host landscape, recombination poses a significant threat to the emergence of
538 novel sarbecoviruses [7].
539
540 With recombination constituting such an important variable in the emergence of novel CoVs,
541 understanding the genetic and ecological determinants of this process is a critical avenue for future
542 research. Here we have shown not only that recombination was involved in the emergence of SARS-CoV-
543 1, but also demonstrated how knowledge of the evolutionary history of these viruses can be used to infer
544 the potential for other viruses to spillover and emerge. Understanding this evolutionary process is highly
545 dependent on factors influencing viral co-occurrence and recombination, such as the geographic range of
546 these viruses and their bat hosts, competitive interactions with co-circulating viruses within the same
547 hosts, and the range of host species these viruses are able to infect. Our understanding depends on the data
548 we have available - the importance of generating more data for such investigations cannot be understated.
549 Investing effort now into further sequencing these viruses and describing the mechanisms that underpin
550 their circulation and capacity for spillover will have important payoffs for predicting and preventing
551 sarbecovirus pandemics in the future.
552
553 **Methods**
554 *Consensus PCR and sequencing of sarbecoviruses from Africa*
555 Oral swabs, rectal swabs, whole blood, and urine samples collected from bats sampled and released in
556 Uganda and Rwanda were assayed for CoVs using consensus PCR as previously described [22]. All
557 sampling was conducted under UC Davis IACUC Protocol No. 16048. Bands of the expected size were
558 purified and confirmed positive by Sanger sequencing and the PCR fragments were deposited to GenBank
559 (accessions MT738926-MT738928, MT732776). Samples were subsequently deep sequenced using the
560 Illumina HiSeq platform and reads were bioinformatically de novo assembled using MEGAHIT v1.2.8
561 [73] after quality control steps and subtraction of host reads using Bowtie2 v2.3.5. Contigs were aligned
562 to a reference sequence and any overlaps or gaps were confirmed with iterative local alignment using

563 Bowtie2. The full genome sequences are deposited in GenBank. Cytochrome b, cytochrome oxidase I.
564 and ACE22 host sequences were also extracted bioinformatically where possible by mapping reads to
565 *Rhinolophus ferrumequinum* reference genes using Bowtie2 and deposited in GenBank.
566
567 *Phylogenetic reconstruction*
568 All publicly available full genome sarbecovirus sequences were collected from GenBank and SARS-
569 CoV-2, pangolin virus genomes, RaTG13, and RmYN01/RmYN02 were downloaded from GISAID
570 (Table 1). All relevant metadata (geographic origin, host species, date of collection) was retrieved from
571 GenBank or the corresponding publications. The RdRp gene (nucleotides 13,431 to 16,222 based on
572 SARS-CoV-2 sequence EPI_ISL_402125 from GISAID) and RBD region (nucleotides 22,506 to 23,174
573 based on the same SARS-CoV-2 reference genome) were extracted and aligned using Muscle v10.2.6.
574 We chose RdRp as a backbone to which to compare because of the strong evolutionary constraints
575 imposed by its fundamental biological role in viral replication [53]. Indeed, the RdRp is generally
576 considered to be a primary genetic trait in viral taxonomy [1,38] and most viruses exhibit strong purifying
577 selection in this gene [74]. Further, the orf1ab region of coronaviruses (which contains the RdRp) also
578 tends to be more recombination-free as compared to the recombination-frequent latter half of the genome
579 [39,54]. Since many of our conclusions are based around phylogenetic topology, we confirmed the
580 robustness of the topology of our nucleotide trees by also building identical trees with alignments of other
581 relatively stable genes in orf1ab frequently used for taxonomic classification [38] (Supplementary Figure
582 S1). Phylogenetic reconstruction was performed using BEAST v2.6.3 [75] with partitioned codon
583 positions, a GTR+Γ substitution model for each of the three codon positions, a constant size coalescent
584 process prior, and a strict molecular clock model. Log files were examined using Tracer v1.7.1 to confirm
585 that the model converged and that the effective sample size (ESS) for each parameter was at least 100.
586 Chains were run until these convergence criteria were met (~2-10 million samples) and multiple chains

587 were run independently to ensure convergence to the same estimates. Use of Beagle 2.1.2 was chosen to
588 increase computational speed.
589
590 Maximum clade credibility trees were built using TreeAnnotator and visualized with FigTree with
591 branches scaled by distance. Posterior probabilities are shown on the preceding branch for each node and
592 probabilities for nodes near the tips of the tree were removed for visual clarity as the exact reconstruction
593 of the most recent divergence events are not within the scope of this study and bear no impact on the
594 interpretation of evolutionary events deeper within the tree.
595
596 Finally, for time-calibrated phylogenies, we minimized the effect of recombination on our estimates by
597 using regions of the genome that were free of recombination for the 13 Lineage 1 sequences of interest
598 (further detailed below). In place of RdRp we used Region A, and in place of RBD we used Region E.
599 These regions were determined to be completely breakpoint free for all sequences using 3SEQ. We
600 started by adding tip dates to Region A and used a strict molecular clock with a normally distributed prior
601 informed from estimates derived in Boni et al. (mean 5.5e-4, sd 5.5e-5) [39]. The prior distribution for the
602 coalescent population size was set to lognormal with mean 1 and standard deviation 10 to help with
603 convergence, as the default of 1/X is an improper prior. Our phylogenetics and time estimates are in
604 accordance with those proposed by Boni et al [39]. As the substitution rate in the spike gene is
605 undoubtedly higher than in RdRp, the same clock rate prior could not be used for the Region E time-
606 calibrated phylogeny because the divergence dates would not be comparable. Instead, we assumed the age
607 of the root of this tree should be approximately the same as the age of the Region A tree and fixed the tree
608 height to match the posterior estimate of the tree height for Region A (770 years before present, 1250
609 AD). This was done by adding a monophyletic MRCA prior to all taxa with a Laplace distribution with
610 mu 1250 and scale 0.1. To account for lineage-specific substitution rates, we also tested a relaxed
611 lognormal clock model.
612

613 *Screening for recombination using detection algorithms*

614 We restricted our search for recombination breakpoints to the region of sequence beginning 750 base
615 pairs upstream from RdRp (SARS-CoV-2 nucleotide 12,681) through the end of S2 (through SARS-CoV-
616 2 nucleotide 25,176). There are undoubtedly other breakpoints outside of this region, but since our
617 analysis focuses primarily on RdRp and the spike, the recombination events elsewhere in the genome are
618 outside the scope of this study. We used the program 3SEQ [76] to test the 13 putative recombinants
619 within Lineage 1 (SARS-CoV-1, SARS-SZ3, LYRa11, Rs3367, WIV1, RsSHC014, Rs4084, YN2018B,
620 Rs7327, Rs9401, Rs4231, WIV16, Rs4874) and RmYN02 individually. If breakpoints were found, each
621 subregion on either side of the breakpoint was assessed separately to fine-tune our assessments until no
622 further breakpoints were identified. We did not test any of the remaining sequences for recombination.
623 We were able to identify six regions across all 13 recombinants that appear to be free of recombination
624 and chose these for further phylogenetic analysis (above). The topologies of regions A and E are not
625 significantly different from the topologies of RdRp and the RBD, respectively, suggesting that our use of
626 RdRp and RBD phylogenies in Figures 1, 2, and 5 is a sufficient representation despite some minor
627 evidence of recombination (*e.g.*, LYRa11).

628

629 *Cell culture and transfection*

630 BHK and 293T cells were obtained from the American Type Culture Collection and maintained in
631 Dulbecco's modified Eagle's medium (DMEM; Sigma–Aldrich) supplemented with 10% fetal bovine
632 serum (FBS), penicillin/streptomycin and L-glutamine. BHK cells were seeded and transfected the next
633 day with 100ng of plasmid encoding hACE2 or an empty vector using polyethylenimine (Polysciences).
634 VSV plasmids were generated and transfected onto 293T cells to produce seed particles as previously
635 described [8]. CoV spike pseudotypes were generated as described in [77] and transfected onto 293T
636 cells. After 24h, cells were infected with VSV particles as described in [78], and after 1h of incubating at
637 37 °C, cells were washed three times and incubated in 2 ml DMEM supplemented with 2% FBS,

638 penicillin/streptomycin and L-glutamine for 48 h. Supernatants were collected and centrifuged at 500*g* for
639 5 min, then aliquoted and stored at −80 °C.
640
641 *Western blots*
642 293T cells transfected with CoV spike pseudotypes (producer cells) were lysed in 1% sodium dodecyl
643 sulfate, 150mM NaCl, 50 mM Tris-HCl and 5 mM EDTA and centrifuged at 14,000*g* for 20 minutes.
644 Pseudotyped particles were concentrated from producer cell supernatants that were overlaid on a 10%
645 OptiPrep cushion in PBS (Sigma–Aldrich) and centrifuged at 20,000*g* for 2h at 4 °C. Lysates and
646 concentrated particles were analyzed for FLAG (Sigma–Aldrich; A8592; 1:10,000), GAPDH (Sigma–
647 Aldrich; G8795; 1:10,000) and/or VSV-M (Kerafast; 23H12; 1:5,000) expression on 10% Bis-Tris PAGE
648 gel (Thermo Fisher Scientific).
649
650 *Cell entry assays*
651 Luciferase-based cell entry assays were performed as described in [8]. For each experiment, the relative
652 light unit for spike pseudotypes was normalized to the plate relative light unit average for the no-spike
653 control, and relative entry was calculated as the fold-entry over the negative control. Three replicates
654 were performed for each CoV pseudotype.
655
656 *Structural modeling*
657 RBDs were modeled using Modweb [79]. Modeled RBDs were docked to hACE2 by structural
658 superposition to the experimentally determined interaction complex between SARS-CoV-1 RBD and
659 hACE2 (PDB 2ajf) [41] using Chimera [80].

660 **Tables**

661 *Table 1. Full list of sequences and accession numbers used in this study.* All accession numbers are from
662 GenBank with the exception of those beginning with EPI_ISL, which are from GISAID. Metadata
663 includes sequencing year, geographic origin, and host species. Sequence names marked with an asterisk
664 (*) indicate those that were not included in the final phylogenetic reconstruction due to high genetic
665 identity with another sequence in the alignment. Citations used to determine hACE2 binding capability
666 are also included.

Accession	Name	Date	Country	Host	ACE2 usage
AY304486	SARS coronavirus SZ3	2003	Guangdong, China	*Paguma larvata* (civet)	[44]†
AY304488	SARS coronavirus SZ16*	2003	Hong Kong, China	*Paguma larvata* (civet)	
AY572034	SARS coronavirus civet007*	2004	Guangdong, China	*Paguma larvata* (civet)	
DQ022305	Bat SARS coronavirus HKU3 1	2005	Hong Kong, China	*Rhinolophus sinicus*	
DQ071615	Bat SARS coronavirus Rp3	2004	Guangxi, China	*Rhinolophus pearsonii*	[4]† [7]† [8]†
DQ084199	Bat SARS coronavirus HKU3 2*	2005	Hong Kong, China	*Rhinolophus sinicus*	
DQ084200	Bat SARS coronavirus HKU3 3*	2005	Hong Kong, China	*Rhinolophus sinicus*	
DQ412042	Bat SARS coronavirus Rf1	2004	Hubei, China	*Rhinolophus ferrumequinum*	[8]†
DQ412043	Bat SARS coronavirus Rm1	2004	Hubei, China	*Rhinolophus macrotis*	
DQ648856	Bat coronavirus BtCoV/273/2005	2004	Hubei, China	*Rhinolophus ferrumequinum*	[8]†
DQ648857	Bat coronavirus BtCoV/279/2005	2004	Hubei, China	*Rhinolophus macrotis*	[8]†
EPI_ISL_ 402125	BetaCoV/Wuhan Hu 1	2019	Hubei, China	human	[3]
EPI_ISL_ 402131	BetaCoV/RaTG13	2013	Yunnan, China	*Rhinolophus affinis*	[32]†
EPI_ISL_ 412976	BetaCoV/RmYN01	2019	Yunnan, China	*Rhinolophus malayanus*	
EPI_ISL_ 412977	BetaCoV/RmYN02	2019	Yunnan, China	*Rhinolophus malayanus*	
EPI_ISL_ 410538	BetaCoV/P4L*	2017	Guangxi, China	*Manis javanica* (pangolin)	
EPI_ISL_ 410539	BetaCoV/P1E*	2017	Guangxi, China	*Manis javanica* (pangolin)	

EPI_ISL_ 410540	BetaCoV/P5L*	2017	Guangxi, China	*Manis javanica* (pangolin)	
EPI_ISL_ 410541	BetaCoV/P5E*	2017	Guangxi, China	*Manis javanica* (pangolin)	
EPI_ISL_ 410542	BetaCoV/P2V	2017	Guangxi, China	*Manis javanica* (pangolin)	
EPI_ISL_ 410543	BetaCoV/P3B*	2017	Guangxi, China	*Manis javanica* (pangolin)	
EPI_ISL_ 410544	BetaCoV/P2S	2019	Guangdong, China	*Manis javanica* (pangolin)	[35]†
FJ588686	Bat SARS coronavirus Rs672/2006	2006	Guizhou, China	*Rhinolophus sinicus*	
GQ153539	Bat SARS coronavirus HKU3 4*	2005	Hong Kong, China	*Rhinolophus sinicus*	
GQ153540	Bat SARS coronavirus HKU3 5*	2005	Hong Kong, China	*Rhinolophus sinicus*	
GQ153541	Bat SARS coronavirus HKU3 6*	2005	Hong Kong, China	*Rhinolophus sinicus*	
GQ153542	Bat SARS coronavirus HKU3 7*	2006	Guangdong, China	*Rhinolophus sinicus*	
GQ153543	Bat SARS coronavirus HKU3 8	2006	Guangdong, China	*Rhinolophus sinicus*	[8]†
GQ153544	Bat SARS coronavirus HKU3 9*	2006	Hong Kong, China	*Rhinolophus sinicus*	
GQ153545	Bat SARS coronavirus HKU3 10*	2006	Hong Kong, China	*Rhinolophus sinicus*	
GQ153546	Bat SARS coronavirus HKU3 11*	2007	Hong Kong, China	*Rhinolophus sinicus*	
GQ153547	Bat SARS coronavirus HKU3 12	2007	Hong Kong, China	*Rhinolophus sinicus*	
GQ153548	Bat SARS coronavirus HKU3 13*	2007	Hong Kong, China	*Rhinolophus sinicus*	[8]†
GU190215	Bat coronavirus BM48-31/BGR/2008	2008	Bulgaria	*Rhinolophus blasii*	[8]†
JX993987	Bat coronavirus Rp/Shaanxi2011	2011	Shaanxi, China	*Rhinolophus pusillus*	[8]†
JX993988	Bat coronavirus Cp/Yunnan2011	2011	Yunnan, China	*Chaerephon plicatus*	[8]†
KC881005	Bat SARS-like coronavirus RsSHC014	2012	Yunnan, China	*Rhinolophus sinicus*	[8,]† [9]†
KC881006	Bat SARS-like coronavirus Rs3367	2012	Yunnan, China	*Rhinolophus sinicus*	
KF294457	SARS related bat coronavirus Longquan 140	2012	Guizhou, China	*Rhinolophus monoceros*	[8]†
KF367457	Bat SARS-like coronavirus WIV1	2012	Yunnan, China	*Rhinolophus sinicus*	[5] [8]†
KF569996	Rhinolophus affinis coronavirus LYRa11	2011	Yunnan, China	*Rhinolophus affinis*	[8]†
KF636752	Bat Hp betacoronavirus/Zhejiang2013	2013	Zhejiang, China	*Hipposideros pratti*	

KJ473811	Bat coronavirus BtRf BetaCoV/JL2012	2012	Jilin, China	*Rhinolophus ferrumequinum*	[8]†
KJ473812	Bat coronavirus BtRf BetaCoV/HeB2013	2013	Hebei, China	*Rhinolophus ferrumequinum*	[8]†
KJ473813	Bat coronavirus BtRf BetaCoV/SX2013	2013	Shanxi, China	*Rhinolophus ferrumequinum*	
KJ473814	Bat coronavirus BtRs BetaCoV/HuB2013	2013	Hubei, China	*Rhinolophus sinicus*	[8]†
KJ473815	Bat coronavirus BtRs BetaCoV/GX2013	2013	Guangxi, China	*Rhinolophus sinicus*	[8]†
KJ473816	Bat coronavirus BtRs BetaCoV/YN2013	2013	Yunnan, China	*Rhinolophus sinicus*	[8]†
KP886808	Bat SARS-like coronavirus YNLF 31C	2013	Yunnan, China	*Rhinolophus sinicus*	
KP886809	Bat SARS-like coronavirus YNLF 34C	2013	Yunnan, China	*Rhinolophus sinicus*	
KT444582	SARS-like coronavirus WIV16	2013	Yunnan, China	*Rhinolophus sinicus*	[6]
KU182964	Bat coronavirus JTMC15	2013	Yunnan, China	*Rhinolophus sinicus*	
KU182963	Bat coronavirus MLHJC35	2012	Jilin, China	*Rhinolophus sinicus*	
KU973692	SARS related coronavirus F46	2012	Yunnan, China	*Rhinolophus pusillus*	
KY352407	SARS related coronavirus BtKY72	2007	Kenya	*Rhinolophus sp.*	
KY417142	Bat SARS-like coronavirus As6526	2014	Yunnan, China	*Aselliscus stoliczkanus*	[7]† [8]†
KY417143	Bat SARS-like coronavirus Rs4081	2012	Yunnan, China	*Rhinolophus sinicus*	[7]† [8]†
KY417144	Bat SARS-like coronavirus Rs4084	2012	Yunnan, China	*Rhinolophus sinicus*	[8]†
KY417145	Bat SARS-like coronavirus Rf4092	2012	Yunnan, China	*Rhinolophus ferrumequinum*	[8]†
KY417146	Bat SARS-like coronavirus Rs4231	2013	Yunnan, China	*Rhinolophus sinicus*	[7]† [8]†
KY417147	Bat SARS-like coronavirus Rs4237	2013	Yunnan, China	*Rhinolophus sinicus*	[8]†
KY417148	Bat SARS-like coronavirus Rs4247	2013	Yunnan, China	*Rhinolophus sinicus*	[8]†
KY417149	Bat SARS-like coronavirus Rs4255	2013	Yunnan, China	*Rhinolophus sinicus*	
KY417150	Bat SARS-like coronavirus Rs4874	2013	Yunnan, China	*Rhinolophus sinicus*	[7]
KY417151	Bat SARS-like coronavirus Rs7327	2014	Yunnan, China	*Rhinolophus sinicus*	[7]† [8]†
KY417152	Bat SARS-like coronavirus Rs9401	2015	Yunnan, China	*Rhinolophus sinicus*	
KY770858	Bat coronavirus Anlong 103	2013	Guizhou, China	*Rhinolophus sinicus*	

KY770859	Bat coronavirus Anlong 112	2013	Guizhou, China	*Rhinolophus sinicus*	
KY770860	Bat coronavirus Jiyuan 84	2012	Henan, China	*Rhinolophus ferrumequinum*	
KY938558	Bat coronavirus 16BO133	2016	South Korea	*Rhinolophus ferrumequinum*	
MG772933	Bat SARS-like coronavirus SL CoVZC45	2017	Zhejiang, China	*Rhinolophus sinicus*	[8]†
MG772934	Bat SARS-like coronavirus SL CoVZXC21	2015	Zhejiang, China	*Rhinolophus sinicus*	[8]†
MK211374	Bat coronavirus BtRl BetaCoV/SC2018	2018	Sichuan, China	*Rhinolophus sp.*	
MK211375	Bat coronavirus BtRs BetaCoV/YN2018A	2018	Yunnan, China	*Rhinolophus affinis*	
MK211376	Bat coronavirus BtRs BetaCoV/YN2018B	2018	Yunnan, China	*Rhinolophus affinis*	
MK211377	Bat coronavirus BtRs BetaCoV/YN2018C	2018	Yunnan, China	*Rhinolophus affinis*	
MK211378	Bat coronavirus BtRs BetaCoV/YN2018D	2018	Yunnan, China	*Rhinolophus affinis*	
NC_004718	SARS coronavirus	2003	Canada	human	[2]
MT726044	PREDICT PDF-2370	2013	Uganda	*Rhinolophus sp.*	
MT726043	PREDICT PDF-2386	2013	Uganda	*Rhinolophus sp.*	
MT726045	PREDICT PRD-0038	2010	Rwanda	*Rhinolophus sp.*	

† Indicates viruses that were not cultured but their spike was shown to enable (or not) hACE2-mediated entry using pseudotyped or recombinant viruses

669 *Table 2. Recombination breakpoints detected in ACE2-using Lineage 1 viruses by the program 3SEQ.*

670 Each recombinant Lineage 1 virus was set as the child sequence, and the parental sequences between the

671 breakpoints identified (minor parent) and on either side (major parent) are listed. The *p*-value indicates

672 the level of significance indicated by 3SEQ. Breakpoint estimates are given as ranges, and the minimum

673 length of the recombinant region between these breakpoints is given. Numbering is relative to the

674 alignment, which begins at SARS-CoV-2 nucleotide 12,681. When 3SEQ identified more than one set of

675 breakpoint estimates, all were included in the table. Each recombinant region was further analyzed

676 separately for more breakpoints within, since 3SEQ identifies only one at a time.

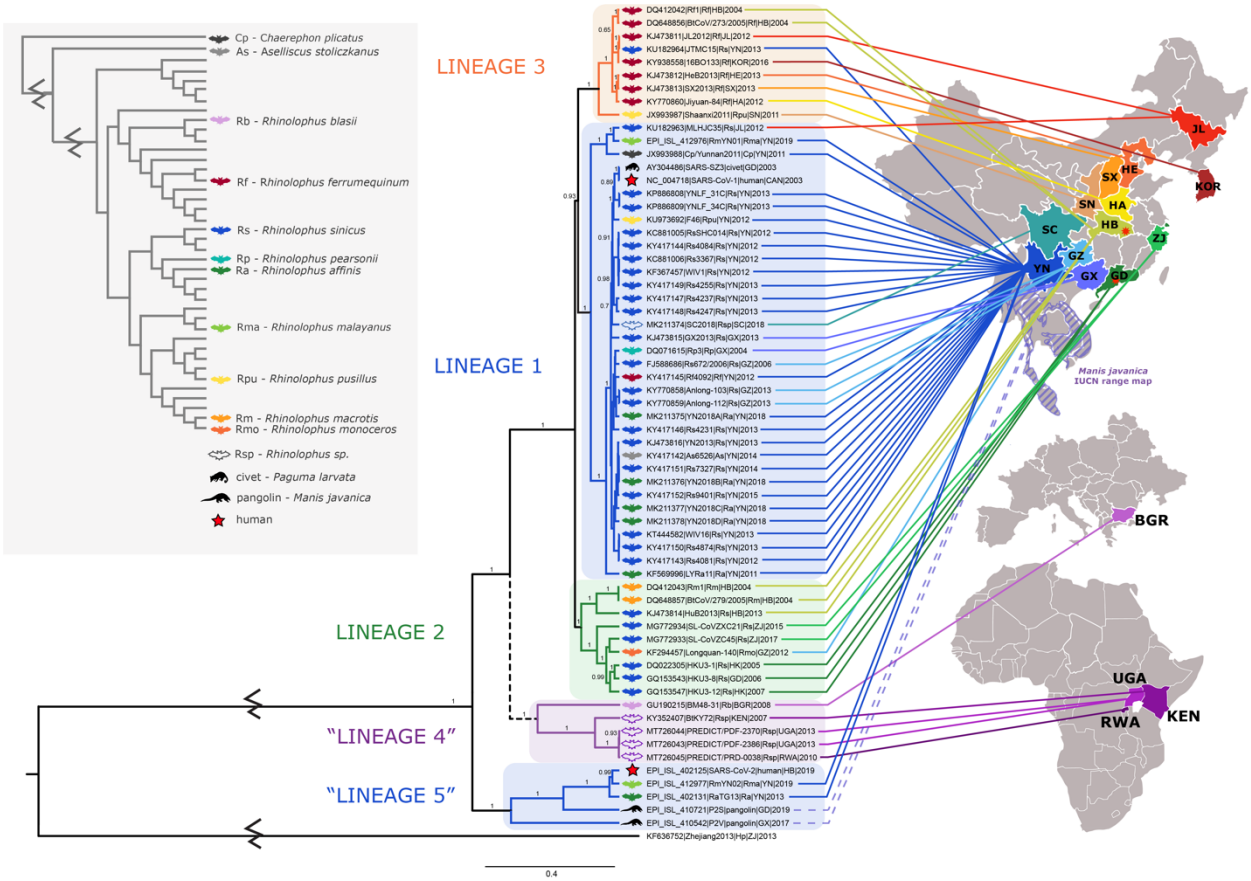
Major Parent	**Minor Parent**	**Child**	***p***	**Length**	**Breakpoint Estimates**
KU973692 F46	EPI_ISL_402131 RaTG13	NC_004718 SARS-CoV-1	0	952	8836-8837 & 10510-10542 8836-8837 & 10726-10752
MK211374 SC2018	EPI_ISL_412976 RmYN01	NC_004718 SARS-CoV-1	0	1290	6497-6519 & 8363-8365 6401-6406 & 8363-8365 6440-6472 & 8363-8365
KY417146 Rs4231	KY417151 Rs7327	NC_004718 SARS-CoV-1	0	573	9760-9772 & 10702-10704
MG772933 SL-CoVZC45	KY770860 Jiyuan-84	NC_004718 SARS-CoV-1	1.4775E-07	1072	11035-11037 & 12610-12624
KY770859 Anlong-112	KY352407 BtKY72	AY304486 SARS-SZ3	0	993	8620-8681 & 10732-10771
MK211374 SC2018	KJ473814 HuB2013	AY304486 SARS-SZ3	1.1774E-07	1077	6755-6784 & 8397-8431
KY417146 Rs4231	MK211376 YN2018B	AY304486 SARS-SZ3	0	558	9760-9772 & 10702-10704
MG772933 SL-CoVZC45	KP886808 YNLF_31C	AY304486 SARS-SZ3	1.592E-07	791	11260-11273 & 12543-12558
EPI_ISL_412976 RmYN01	NC_004718 SARS-CoV-1	KF569996 LYRa11	0	921	9107-9113 & 10700-10701 9027-9043 & 10865-10869 9077-9095 & 10865-10869 9107-9113 & 10865-10869 9027-9043 & 10840-10842 9077-9095 & 10840-10842 9107-9113 & 10840-10842 9027-9043 & 10700-10701 9077-9095 & 10700-10701
JX993988 Cp/Yunnan2011	KY770859 Anlong-112	KF569996 LYRa11	0	1627	1658-1714 & 4151-4199 1368-1428 & 4229-4240 1487-1498 & 4229-4240 1658-1714 & 4229-4240 1368-1428 & 4151-4199 1487-1498 & 4151-4199

NC_004718 SARS-CoV-1	KY417142 As6526	KC881006 Rs3367	0	2117	0-11 & 9245-9251
KC881005 RsSHC014	KF569996 LYRa11	KC881006 Rs3367	0	168	10201-10233 & 10549-10565
KY417151 Rs7327	KY417142 As6526	KC881006 Rs3367	0	3036	1853-3932 & 8288-8374
NC_004718 SARS-CoV-1	KY417142 As6526	KF367457 WIV1	0	2116	0-11 & 9245-9251
KC881005 RsSHC014	KF569996 LYRa11	KF367457 WIV1	0	168	10201-10233 & 10549-10565
KY417151 Rs7327	KY417142 As6526	KF367457 WIV1	0	3036	1853-3932 & 8288-8374
KF367457 WIV1	KY417146 Rs4231	KC881005 RsSHC014	0	378	9841-9915 & 10549-10572
KY417151 Rs7327	KY417142 As6526	KC881005 RsSHC014	0	3037	1853-3932 & 8288-8374
KF367457 WIV1	KY417146 Rs4231	KY417144 Rs4084	0	378	9841-9915 & 10549-10572
KY417151 Rs7327	KY417142 As6526	KY417144 Rs4084	0	3034	1853-3932 & 8288-8374
NC_004718 SARS-CoV-1	MK211377 YN2018C	MK211376 YN2018B	0	2417	411-551 & 9245-9251
KC881005 RsSHC014	KF569996 LYRa11	MK211376 YN2018B	0	122	10201-10233 & 10469-10497
KY417151 Rs7327	MK211378 YN2018D	MK211376 YN2018B	0	2205	4541-5578 & 8766-8789
NC_004718 SARS-CoV-1	KY417142 As6526	KY417151 Rs7327	0	2112	0-11 & 9245-9251
KC881005 RsSHC014	KF569996 LYRa11	KY417151 Rs7327	0	122	10201-10233 & 10469-10497
KY417144 Rs4084	MK211377 YN2018C	KY417151 Rs7327	0	3260	924-1939 & 8186-8374
NC_004718 SARS-CoV-1	KY417142 As6526	KY417152 Rs9401	0	2112	0-11 & 9245-9251
KC881005 RsSHC014	KF569996 LYRa11	KY417152 Rs9401	0	122	10201-10233 & 10469-10497
KY417144 Rs4084	MK211377 YN2018C	KY417152 Rs9401	0	3260	924-1939 & 8186-8374
NC_004718 SARS-CoV-1	KY417149 Rs4255	KY417146 Rs4231	0	2296	0-11 & 8838-8840
NC_004718 SARS-CoV-1	KC881005 RsSHC014	KY417146 Rs4231	0	1788	9769-9780 & 12448-12793
NC_004718 SARS-CoV-1	KY417143 Rs4081	KT444582 WIV16	0	2293	0-32 & 8838-8840
KF367457 WIV1	KY417146 Rs4231	KT444582 WIV16	0	541	0-8891 & 9973-10233
KC881005 RsSHC014	NC_004718 SARS-CoV-1	KT444582 WIV16	0	403	0-8891 & 9769-9780
KY417143 Rs4081	KY417146 Rs4231	KT444582 WIV16	4E-12	1781	5975-6133 & 8727-12793 3536-5782 & 8727-12793

NC_004718 SARS-CoV-1	KY417143 Rs4081	KY417150 Rs4874	0	2294	0-32 & 8838-8840
KF367457 WIV1	KY417146 Rs4231	KY417150 Rs4874	0	541	0-8891 & 9973-10233
KC881005 RsSHC014	NC_004718 SARS-CoV-1	KY417150 Rs4874	0	403	0-8891 & 9769-9780
KY417143 Rs4081	KY417146 Rs4231	KY417150 Rs4874	4E-12	1782	5975-6133 & 8727-12793 3536-5782 & 8727-12793
EPI_ISL_402125 SARS-CoV-2	KU182964 JTMC15	EPI_ISL_412977 RmYN02	0	1111	8957-8957 & 10827-10828 8938-8941 & 10831-10845 8957-8957 & 10831-10845 8938-8941 & 10827-10828
EPI_ISL_410542 P2V	KY770859 Anlong-112	EPI_ISL_412977 RmYN02	0	3218	1904-1907 & 5126-5128 1862-1879 & 5126-5128 1883-1885 & 5126-5128

677

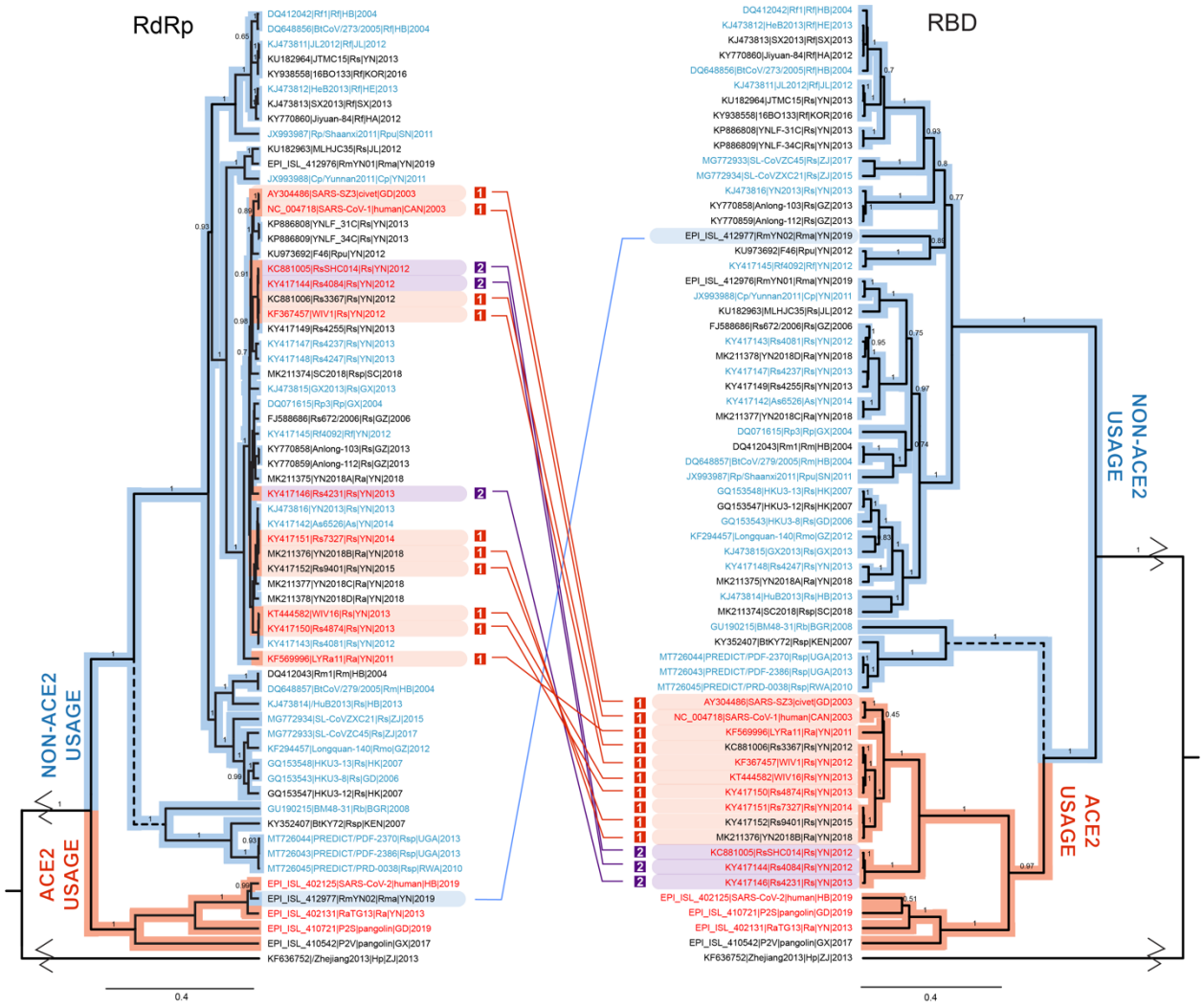
678 **Figures**



679

680 *Figure 1: Phylogenetic tree of the RNA dependent RNA polymerase (RdRp) gene (nsp12) and associated*
681 *geographic origin and host species.* Colors of clade bars represent the different geographic lineages.
682 Lineage 1 is shown in blue, Lineage 2 in green, and Lineage 3 in orange. The clade of viruses from Africa
683 and Europe is putatively named "Lineage 4" and is shown in purple. The phylogeny shows strong
684 posterior support for the branching order presented; however, different models or genes have produced
685 trees with different branching orders placing Lineage 4 outside Lineage 5, so the branch to Lineage 4 is
686 dashed to represent this uncertainty (Supplementary Figure S1). The putative "Lineage 5" containing
687 SARS-CoV-2 is also shown in blue at the bottom of the tree to demonstrate that the sequences are from
688 the same regions as Lineage 1 viruses. The geographic origin of each virus is indicated by the lines that
689 terminate in the respective country or province with the same color code. The full province and country
690 names for all two- and three-letter codes can be found in Table 1. As human, civet, and pangolin viruses

691 cannot be certain to have naturally originated in the province in which they were first found, their
692 locations are not illustrated, but the natural range of the pangolin (*Manis javanica*) is denoted with dashed
693 shading and the origins of the SARS-CoV-1 and SARS-CoV-2 human outbreaks are designated with red
694 stars in Guangdong and Hubei, respectively. Hosts are also shown with colored symbols according to the
695 key on the left. The host phylogeny in the key was adapted from [81]. The root of the tree was shortened
696 for clarity.



697

698 *Figure 2: Phylogenetic trees of RdRp (left) and the RBD (right) demonstrating recombination events*

699 *between ACE2-users and non-ACE2-users*. Names of viruses that have been confirmed to use hACE2 are

700 shown in red font, and those that have been shown to not use hACE2 are shown in blue font (citations can

701 be found in Table 1). Viruses in black font have not yet been tested. The red and blue highlighted clade

702 bars separate viruses with the structure associated with ACE2 usage (highly similar to viruses confirmed

703 to use hACE2 specifically) and the structure with deletions that cannot use ACE2, respectively.

704 Connecting lines indicate recombination events that resulted in a gain of ACE2 usage (red) or a loss of

705 ACE2 usage (blue). The two different groups of RBD sequence within the Lineage 1 recombinants that

706 gained ACE2 usage are distinguished in red (Type 1) and purple (Type 2) highlighting. The distances of

707 the roots have been shortened for clarity. The branch leading to Lineage 4 is dashed to demonstrate

708 uncertainty in its positioning.

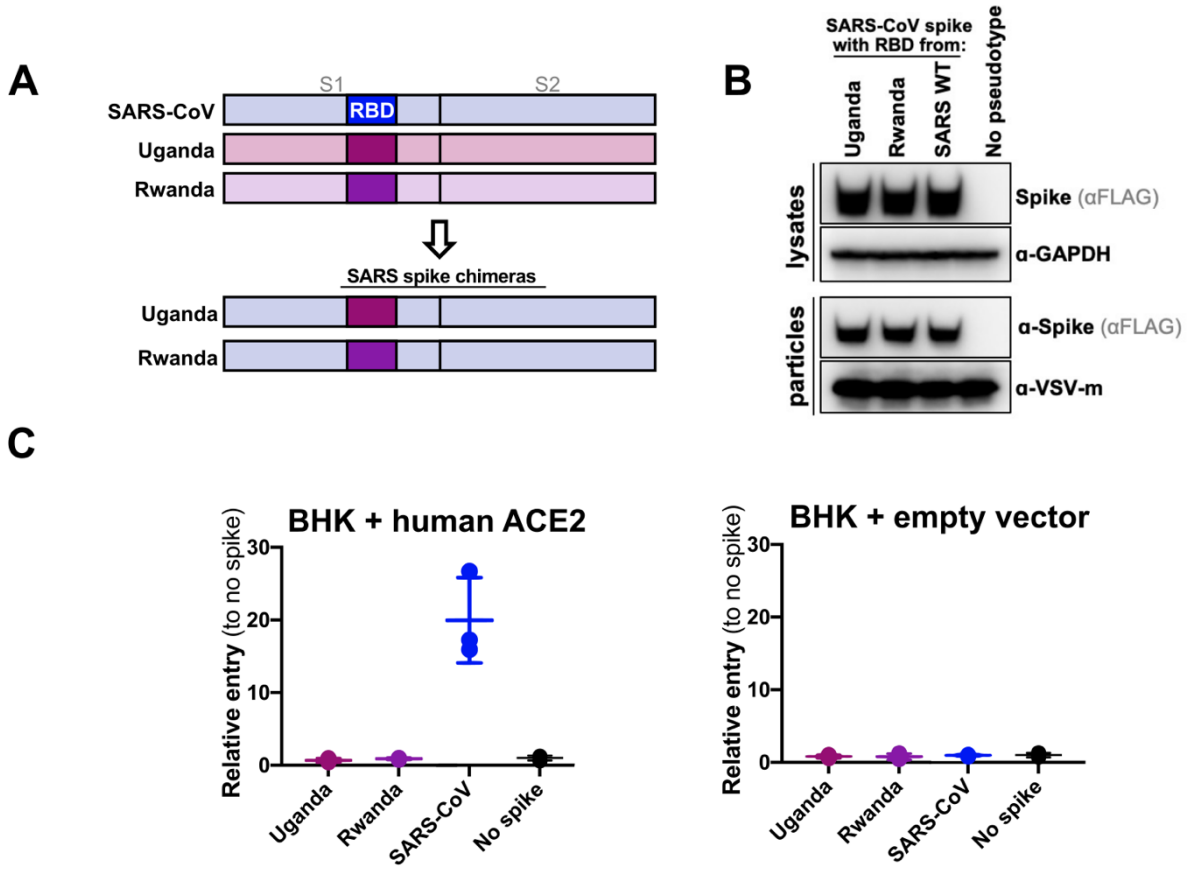


Figure 3: hACE2 usage of bat sarbecoviruses investigated using a surrogate VSV-psuedotyping system. (A) Schematic showing the structure of chimeric spike proteins. The SARS-CoV-1 spike backbone is used in conjunction with the RBD from the Uganda and Rwanda strains. (B) Incorporation of chimeric SARS-CoV-1 spike proteins into VSV. Western blots show successful expression of chimeric spikes (lysates) and their incorporation into VSV (particles). (C) hACE2 entry assays. Left, wildtype SARS-CoV spike protein is able to mediate entry into BHK cells expressing hACE2. In contrast, recombinant spike proteins containing either the Uganda or Rwanda RBD were unable to mediate entry. Entry is expressed relative to VSV particles with no spike protein. Right, control experiment for entry assay. BHK cells do not express hACE2 and therefore do not permit entry of hACE2-dependent VSV pseudotypes.

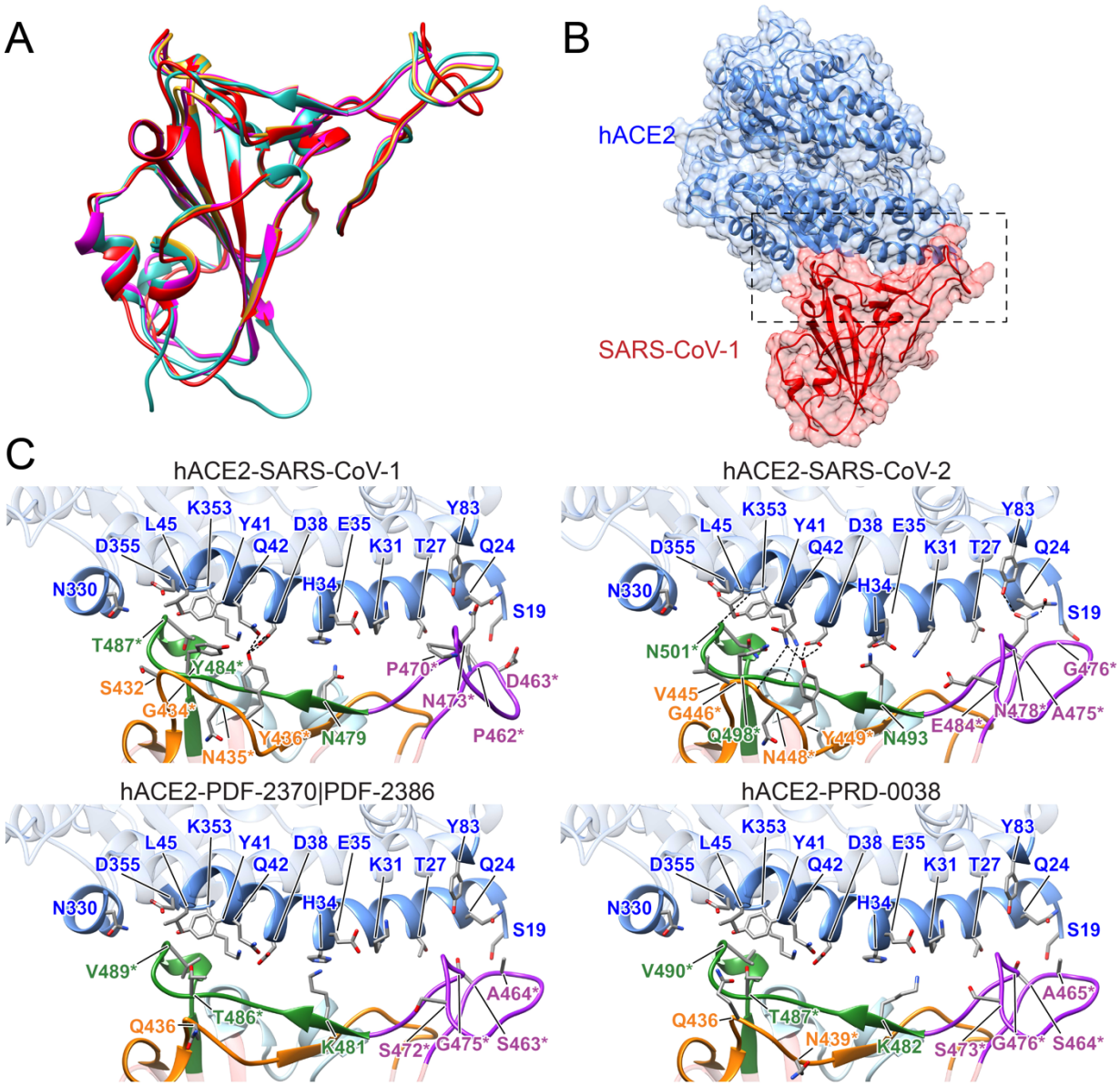


Figure 4. Structural modeling of sarbecovirus RBDs found in Uganda and Rwanda. **(**A) Structural superposition of the X-ray structures for the RBDs in SARS-CoV-1 (PDB 2ajf, red) [41] and SARS-CoV-2 (PDB 6m0j, cyan) [82] and homology models for SARS-CoV found in Uganda (PDF-2370 and PDF-2386, magenta) and Rwanda (PRD-0038, yellow). (B) Overview of the X-ray structure of SAR-CoV-1 RBD (red) bound to hACE2 (blue) (PDB 2ajf, red) [41]. (C) Close-up view of the interface between hACE2 (blue) and RBDs in SARS-CoV-1 (PDB 2ajf, top left) [41] and SARS-CoV-2 (PDB 6m0j, top right) [82] and homology models for viruses found in Uganda (PDF-2370 and PDF-2386, bottom, left)

727 and Rwanda (PRD-0038, bottom, right). The color of the RBD loops corresponds to the colors of the
728 labeled sequence regions in Figure 5: region 1 in cyan, region 2 in orange, the receptor binding ridge in
729 purple, and region 3 in green. Labeled RBD residues correspond to interfacial residues whose identity
730 differ in African sarbecoviruses and SARS-CoV-1 or SARS-CoV-2 (labels are included in all four panels
731 to facilitate the identification of counterpart residues in each virus). Asterisks denote residues whose
732 identity is not shared by any ACE-2 binding SARS-CoV as dictated by Figure 5. Labeled hACE2 residues
733 correspond to residues within 5Å of RBD residues depicted.

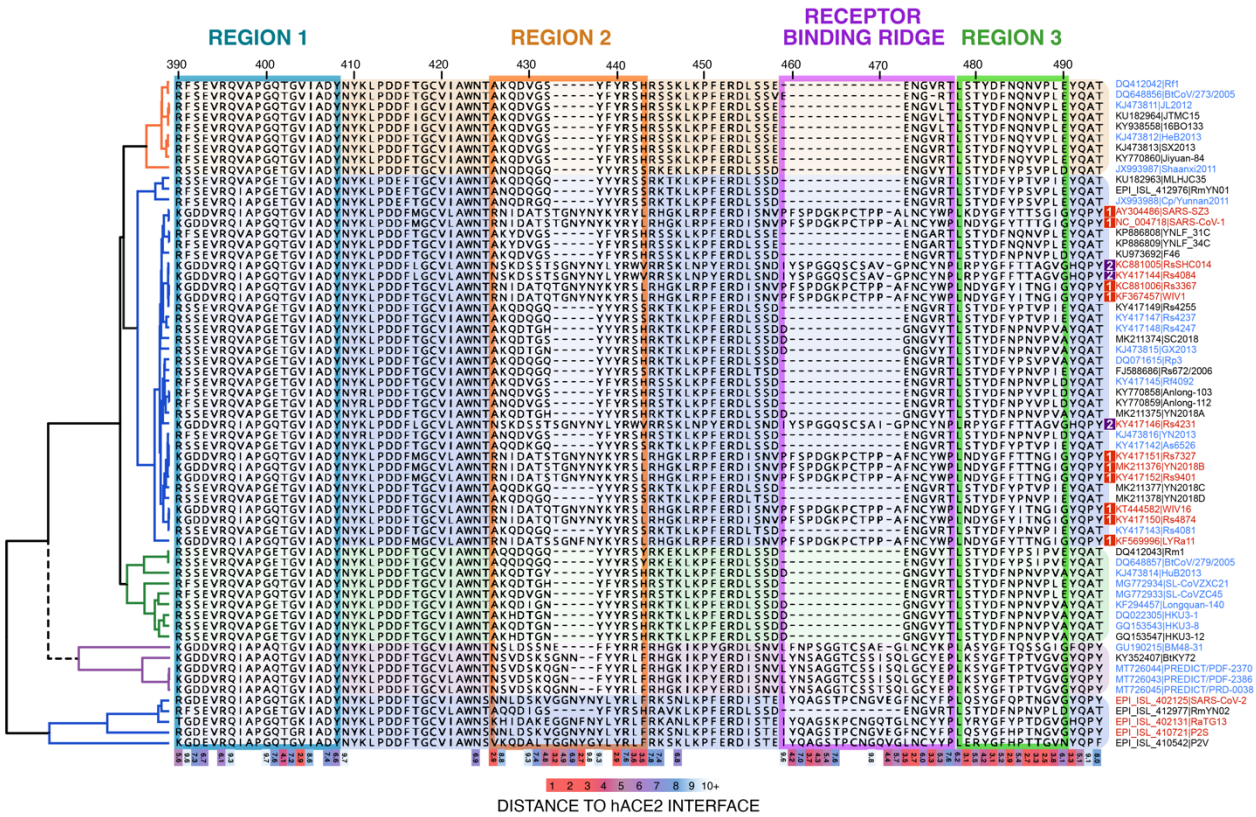


Figure 5: The phylogenetic backbone of the RdRp gene alongside the amino acid sequences of the RBM. Amino acid numbering is relative to SARS-CoV-1. Virus names in red font are known hACE2 users, those in blue are known non-users, and those in black have not been tested. Residues within 10Å of the interface with hACE2 are considered interfacial, and exact distances between each interfacial residue and the closest hACE2 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 hACE2 are shown in red, and as distance increases this color transitions to purple, blue, and finally to white. The receptor binding 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 4. The branch leading to Lineage 4 is dashed to demonstrate uncertainty in its positioning.

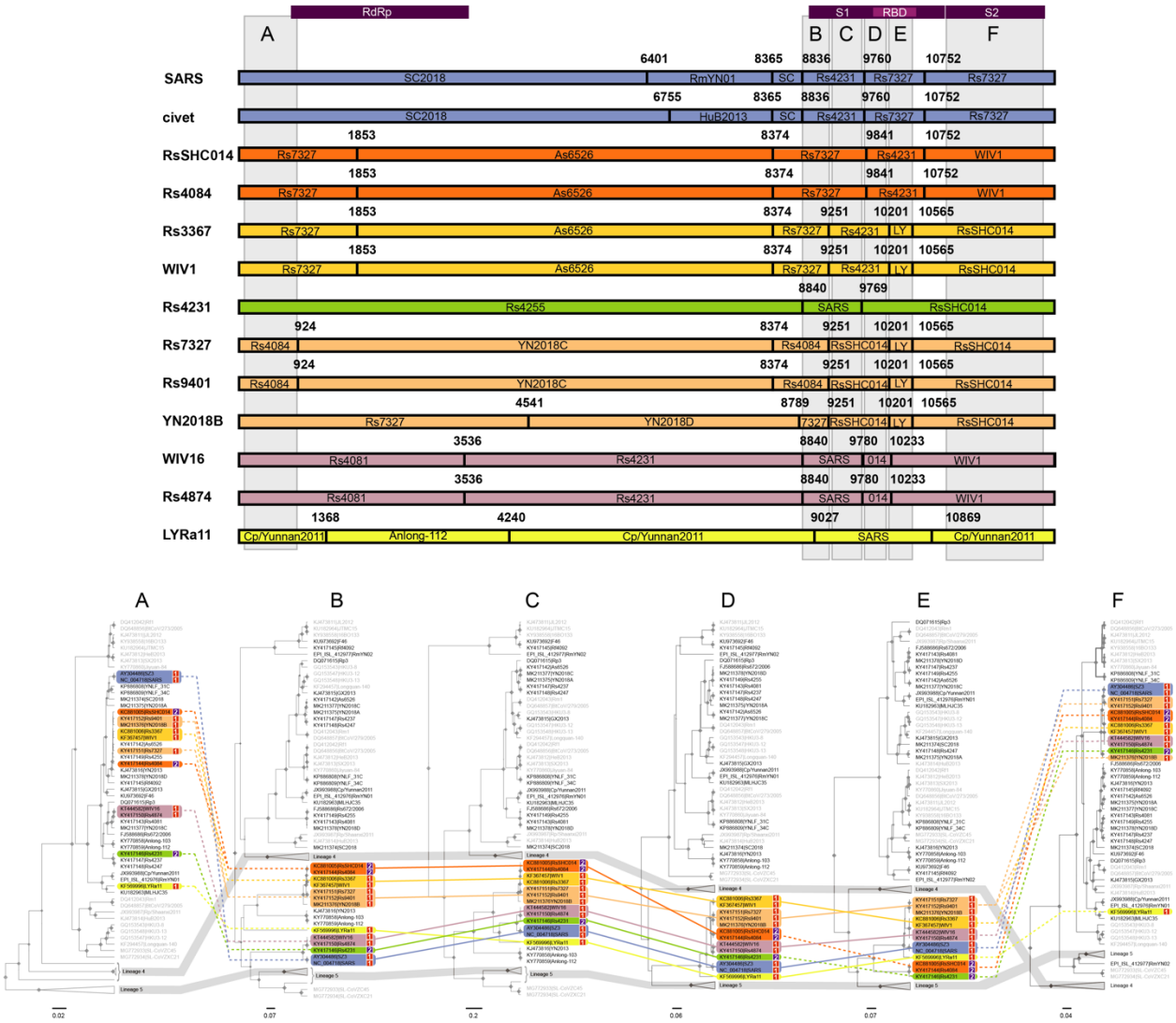
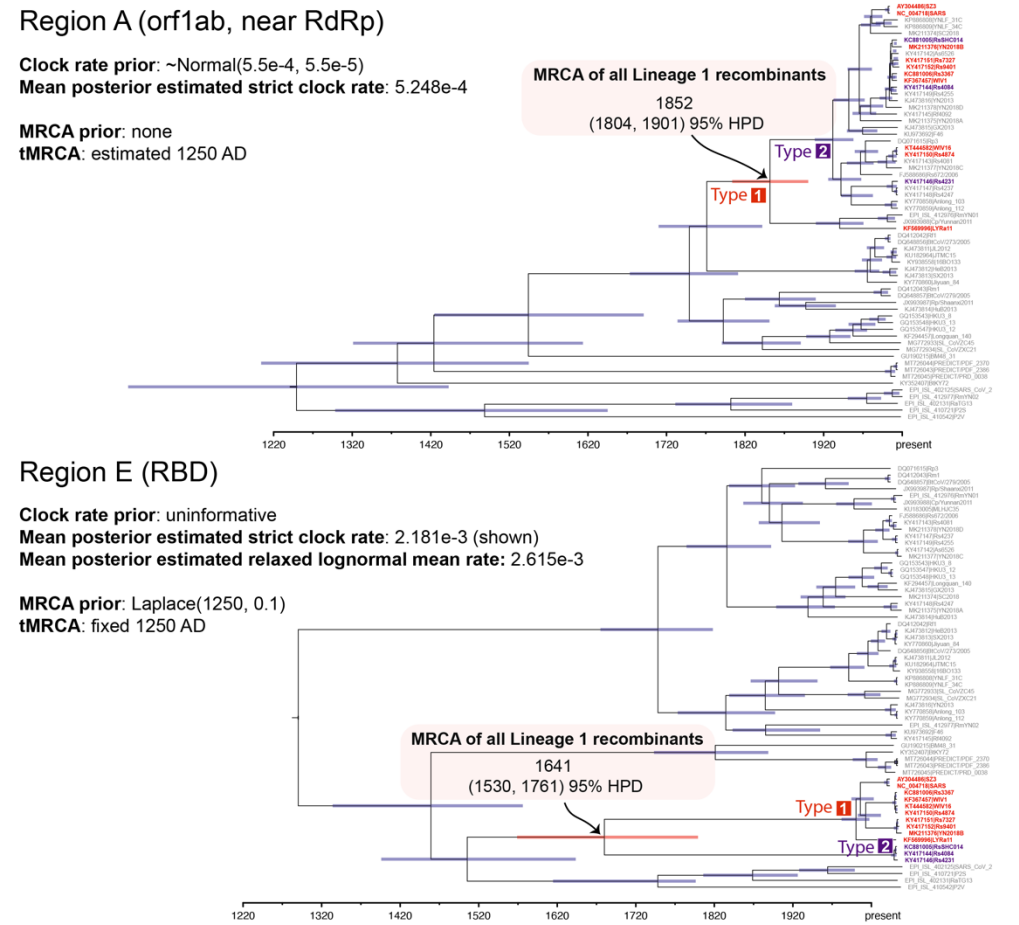


Figure 6. Recombination breakpoints detected in Lineage 1 ACE2-using sequences. The top of this figure illustrates that the recombination suggested by the change in topology in Figure 2 for 13 Lineage 1 viruses is supported by formal breakpoint analysis. The breakpoints detected for each of the 13 recombinant Lineage 1 sequences with ACE2-using structure (no deletions) are shown. Sequences that are nearly identical are colored the same for simplicity. The bars represent the sequence of genome beginning 750 bp before RdRp spanning through the end of S2 (SARS-CoV-2 nucleotides 12,681 through 25,176) and each box within represents a recombinant section within the sequence. The breakpoints correspond to those identified in Table 2. Numbering is relative to the alignment. The parental sequence is shown within each box. Sequences identified as the minor parent by 3SEQ were labeled within the

756 breakpoint margins and the major parent outside. Six regions where these sequences appear to be free of
757 recombination are labeled A-F and a corresponding phylogeny for each region is shown below. Regions
758 A and E were further tested for recombination breakpoints in all sequences, not just the 13 Lineage 1
759 viruses, and were found to be breakpoint-free. The topology of regions A and E is not different enough
760 from Figure 2 to suggest that recombination within RdRp or RBD significantly changed the interpretation
761 of our results. For each region, sequences were tracked with connecting lines of corresponding color to
762 identify where recombination may have occurred between Lineage 1 and Lineage 5 and hypothesized
763 events are specifically marked with dotted lines. This highlights the secondary recombination of Rs4084
764 and RsSHC014 in region E on top of the primary recombination in regions B through E. Sequence names
765 of Lineage 2 and 3 viruses are greyed out and Lineages 4 and 5 are collapsed and highlighted in darker
766 grey to make the changes in topology between the trees more visible.
767



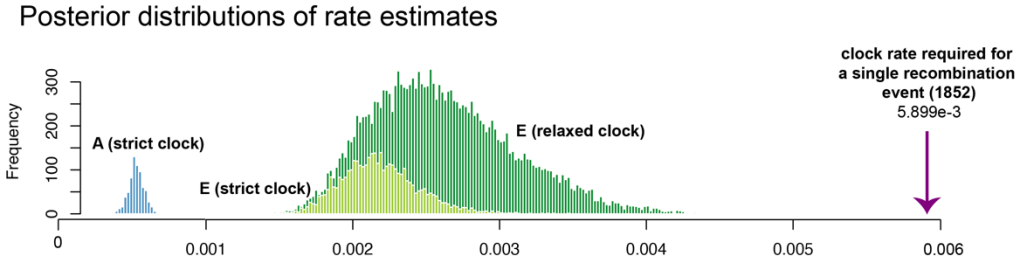


Figure 7. Time-calibrated phylogenies for recombination-free regions of the genome. Breakpoint-free regions A and E from Figure 6 were chosen for time calibration since evidence of recombination was found in both RdRp and RBD. Both regions A and E were free of recombination for all sequences included in the tree, ensuring the best possible dating estimates. The MRCA of all Lineage 1 recombinants and its corresponding divergence date are labeled on each tree, demonstrating that the MRCA in region E (within the RBD) is much older than the MRCA in region A (proxy for RdRp, see Figure 6). This suggests that there would not have been enough time for the RBDs of the recombinants to

776 diversify to the extent shown here if only a single recombination event occurred between Lineage 5 and
777 Lineage 1. The MRCAs of each type are labeled in red (Type 1) and purple (Type 2). Posterior
778 distributions of rate estimates are also shown for each model as well as for a relaxed clock model of
779 region E. For the observed sequence divergence in region E to have accumulated since the MRCA of the
780 13 recombinants in region A (1852), a clock rate of 5.899e-3 would be required, which is well outside the
781 posterior distributions estimated by both our strict and relaxed clock models.

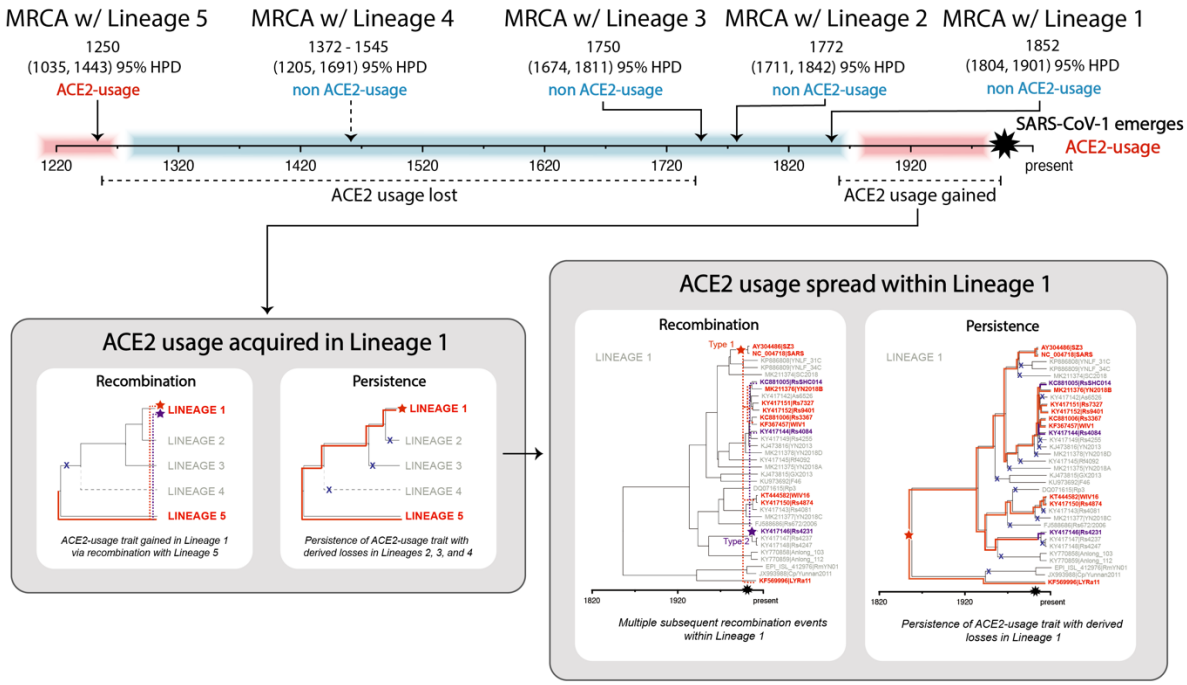


Figure 8. Proposed timeline of deletion and recombination events. The timeline demonstrates the sequence of events that led to loss of ACE2 usage in Lineages 2, 3, and 4 and gain of ACE2 usage within Lineage 1, leading to the emergence of SARS-CoV-1. Events are dated with MRCA age estimates; however, the exact intention is less to provide exact dates and more to suggest a particular order of events, which is strongly supported by the posterior probabilities of the time-calibrated phylogenies. The arrow for the Lineage 4 event is again dashed to demonstrate uncertainty in its positioning. We illustrate two hypotheses for the acquisition and subsequent spread of ACE2 usage in Lineage 1: recombination and persistence. The recombination hypothesis is much more parsimonious, as persistence would require multiple independent deletion events to generate the observed pattern of ACE2 usage.

Acknowledgements

We also thank three anonymous reviewers who provided thoughtful and robust suggestions that substantially improved this manuscript. 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 alignments used for phylogenetics are included as supplementary materials.

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 Yang, X.-L. *et al.* (2016) Isolation and Characterization of a Novel Bat Coronavirus Closely Related to the Direct Progenitor of Severe Acute Respiratory Syndrome Coronavirus. *J. Virol.* 90, 3253–3256

7 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

8 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

9 Menachery, V.D. *et al.* (2015) A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat. Med.* 21, 1508–1513

10 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

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

12 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

13 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

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

15 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

16 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

17 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

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

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

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

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

22 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

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

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

25 Ar Gouilh, M. *et al.* (2018) SARS-CoV related Betacoronavirus and diverse Alphacoronavirus members found in western old-world. *Virology* 517, 88–97

26 Drexler, J.F. *et al.* (2010) Genomic Characterization of Severe Acute Respiratory Syndrome-Related Coronavirus in European Bats and Classification of Coronaviruses Based on Partial RNA-Dependent RNA Polymerase Gene Sequences. *J. Virol.* 84, 11336–11349

27 Rihtarič, D. *et al.* (2010) Identification of SARS-like coronaviruses in horseshoe bats (Rhinolophus hipposideros) in Slovenia. *Arch. Virol.* 155, 507–514

28 Lelli, D. *et al.* (2013) Detection of Coronaviruses in Bats of Various Species in Italy. *Viruses* 5, 2679–2689

29 Tao, Y. and Tong, S. (2019) Complete Genome Sequence of a Severe Acute Respiratory Syndrome-Related Coronavirus from Kenyan Bats. *Microbiol. Resour. Announc.* 8,

30 Cui, J. *et al.* (2007) Evolutionary relationships between bat coronaviruses and their hosts. *Emerg. Infect. Dis.* 13, 1526–1532

31 Leopardi, S. *et al.* (2018) Interplay between co-divergence and cross-species transmission in the evolutionary history of bat coronaviruses. *Infect. Genet. Evol.* DOI: 10.1016/j.meegid.2018.01.012

32 Shang, J. *et al.* (2020) Structural basis of receptor recognition by SARS-CoV-2. *Nature* 581, 221–224

33 Liu, P. *et al.* (2019) Viral metagenomics revealed sendai virus and coronavirus infection of malayan pangolins (manis javanica). *Viruses* DOI: 10.3390/v11110979

34 Lam, T.T.Y. *et al.* (2020) Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature* DOI: 10.1038/s41586-020-2169-0

35 Antoni G. Wrobel, D.J.B.P.X.A.B.C.R.S.R.M.P.B.R.J.J.S.S.J.G. (2020) Structure and binding properties of Pangolin-CoV Spike glycoprotein inform the evolution of SARS-CoV-2. *Res. Sq. [preprint]* DOI: 10.21203/RS.3.RS-83072/V1

36 Zhou, H. *et al.* (2020) A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. *Curr. Biol.* DOI: 10.1016/j.cub.2020.05.023

37 Challender, D. *et al.* (2014) Manis javanica. *IUCN Red List Threat. Species 2014* DOI: http://dx.doi.org/10.2305/IUCN.UK.2014-2.RLTS.T12763A45222303.en.

38 Gorbalenya, A.E. *et al.* The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. , *Nature Microbiology*. (2020)

39 Boni, M.F. *et al.* (2020) Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.* DOI: 10.1038/s41564-020-0771-4

40 Prabakaran, P. *et al.* (2004) A model of the ACE2 structure and function as a SARS-CoV receptor. *Biochem. Biophys. Res. Commun.* 314, 235–241

41 Li, F. *et al.* (2005) Structural biology: Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science (80-.).* 309, 1864–1868

42 Li, F. (2008) Structural Analysis of Major Species Barriers between Humans and Palm Civets for Severe Acute Respiratory Syndrome Coronavirus Infections. *J. Virol.* DOI: 10.1128/jvi.00442-08

43 Wu, K. *et al.* (2012) Mechanisms of host receptor adaptation by severe acute respiratory syndrome coronavirus. *J. Biol. Chem.* DOI: 10.1074/jbc.M111.325803

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

45 Wan, Y. *et al.* (2020) Receptor recognition by novel coronavirus from Wuhan: 2 An analysis based on decade-long structural studies of SARS 3 4 Downloaded from. DOI: 10.1128/JVI.00127-20

46 Chen, Y. *et al.* (2020) Structure analysis of the receptor binding of 2019-nCoV. *Biochem. Biophys.*

Res. Commun. 525, 135–140

47 Wan, Y. *et al.* (2020) Receptor recognition by novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. *J. Virol.* DOI: 10.1128/jvi.00127-20

48 Damas, J. *et al.* (2020) Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates. *Proc. Natl. Acad. Sci. U. S. A.* 117, 22311–22322

49 Lam, S.D. *et al.* (2020) SARS-CoV-2 spike protein predicted to form complexes with host receptor protein orthologues from a broad range of mammals. *Sci. Rep.* 10, 16471

50 Hou, Y. *et al.* (2010) Angiotensin-converting enzyme 2 (ACE2) proteins of different bat species confer variable susceptibility to SARS-CoV entry. *Arch. Virol.* DOI: 10.1007/s00705-010-0729-6

51 Zheng, M. *et al.* (2020) Bat SARS-Like WIV1 coronavirus uses the ACE2 of multiple animal species as receptor and evades IFITM3 restriction via TMPRSS2 activation of membrane fusion. *Emerg. Microbes Infect.* 9, 1567–1579

52 Zhao, X. *et al.* (2020) Broad and Differential Animal Angiotensin-Converting Enzyme 2 Receptor Usage by SARS-CoV-2. *J. Virol.* 94,

53 Ulferts, R. *et al.* (2010) Expression and functions of SARS coronavirus replicative proteins. In *Molecular Biology of the SARS-Coronavirus* pp. 75–98, Springer Berlin Heidelberg

54 Fu, K. and Baric, R.S. (1994) Map locations of mouse hepatitis virus temperature-sensitive mutants: confirmation of variable rates of recombination. *J. Virol.* 68, 7458–7466

55 Wu, F. *et al.* (2020) A new coronavirus associated with human respiratory disease in China. *Nature* DOI: 10.1038/s41586-020-2008-3

56 Regan, A.D. *et al.* (2012) Characterization of a recombinant canine coronavirus with a distinct receptor-binding (S1) domain. *Virology* DOI: 10.1016/j.virol.2012.04.013

57 Terada, Y. *et al.* (2014) Emergence of pathogenic coronaviruses in cats by homologous recombination between feline and canine coronaviruses. *PLoS One* DOI: 10.1371/journal.pone.0106534

58 Tao, Y. *et al.* (2017) Surveillance of Bat Coronaviruses in Kenya Identifies Relatives of Human Coronaviruses NL63 and 229E and Their Recombination History. *J. Virol.* DOI: 10.1128/jvi.01953-16

59 Boniotti, M.B. *et al.* (2016) Porcine epidemic diarrhea virus and discovery of a recombinant swine enteric coronavirus, Italy. *Emerg. Infect. Dis.* DOI: 10.3201/eid2201.150544

60 Buchholz, U.J. *et al.* (2004) Contributions of the structural proteins of severe respiratory syndrome coronavirus to protective immunity. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9804–9809

61 Lu, L. *et al.* (2004) Immunological Characterization of the Spike Protein of the Severe Acute Respiratory Syndrome Coronavirus. *J. Clin. Microbiol.* 42, 1570–1576

62 Prabakaran, P. *et al.* (2006) Structure of severe acute respiratory syndrome coronavirus receptor-binding domain complexed with neutralizing antibody. *J. Biol. Chem.* 281, 15829–15836

63 Anthony, S.J. *et al.* (2017) Global patterns in coronavirus diversity. *Virus Evol.* DOI: 10.1093/ve/vex012

64 Quan, P.L. *et al.* (2010) Identification of a severe acute respiratory syndrome coronavirus-like virus in a leaf-nosed bat in Nigeria. *MBio* DOI: 10.1128/mBio.00208-10

65 Watanabe, S. *et al.* (2010) Bat coronaviruses and experimental infection of bats, the Philippines. *Emerg. Infect. Dis.* DOI: 10.3201/eid1608.100208

66 De Souza Luna, L.K. *et al.* (2007) Generic detection of coronaviruses and differentiation at the prototype strain level by reverse transcription-PCR and nonfluorescent low-density microarray. *J. Clin. Microbiol.* DOI: 10.1128/JCM.02426-06

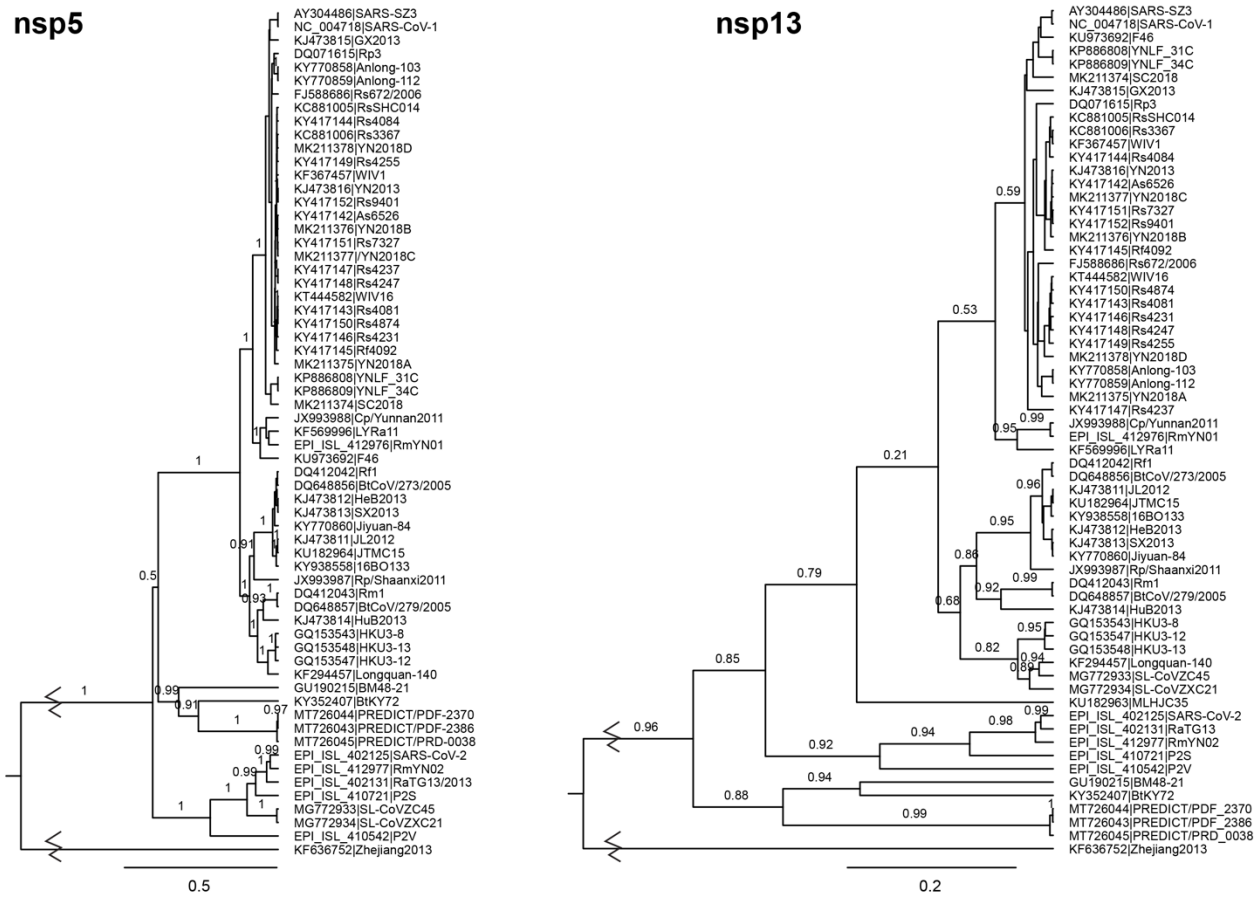
67 Mao, X. *et al.* (2016) Differential introgression suggests candidate beneficial and barrier locibetween two parapatric subspecies of Pearson′s horseshoe bat Rhinolophuspearsoni. *Curr. Zool.* 62, 405

68 Mao, X. *et al.* (2013) Lineage Divergence and Historical Gene Flow in the Chinese Horseshoe Bat (Rhinolophus sinicus). *PLoS One* 8,

69 Mao, X. *et al.* (2014) Differential introgression among loci across a hybrid zone of the

960 intermediate horseshoe bat (Rhinolophus affinis). *BMC Evol. Biol.* 14, 154
961 70 Mao, X. *et al.* (2013) Multiple cases of asymmetric introgression among horseshoe bats detected
962 by phylogenetic conflicts across loci. *Biol. J. Linn. Soc.* 110, 346–361
963 71 MAO, X. *et al.* (2010) Historical male-mediated introgression in horseshoe bats revealed by
964 multilocus DNA sequence data. *Mol. Ecol.* 19, 1352–1366
965 72 Dool, S.E. *et al.* (2016) Nuclear introns outperform mitochondrial DNA in inter-specific
966 phylogenetic reconstruction: Lessons from horseshoe bats (Rhinolophidae: Chiroptera). *Mol.*
967 *Phylogenet. Evol.* 97, 196–212
968 73 Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced
969 methodologies and community practices. , *Methods*, 102. 01-Jun-(2016) , Academic Press Inc., 3–
970 11
971 74 Tang, X. *et al.* (2009) Differential stepwise evolution of SARS coronavirus functional proteins in
972 different host species. *BMC Evol. Biol.* 9, 52
973 75 Bouckaert, R. *et al.* (2019) BEAST 2.5: An advanced software platform for Bayesian evolutionary
974 analysis. *PLoS Comput. Biol.* 15, e1006650
975 76 Lam, H.M. *et al.* (2018) Improved Algorithmic Complexity for the 3SEQ Recombination
976 Detection Algorithm. *Mol. Biol. Evol.* 35, 247–251
977 77 Letko, M. *et al.* (2018) Adaptive Evolution of MERS-CoV to Species Variation in DPP4. *Cell*
978 *Rep.* 24, 1730–1737
979 78 Takada, A. *et al.* (1997) A system for functional analysis of Ebola virus glycoprotein. *Proc. Natl.*
980 *Acad. Sci. U. S. A.* 94, 14764–14769
981 79 Pieper, U. *et al.* (2011) ModBase,a database of annotated comparative protein structure
982 models,and associated resources. *Nucleic Acids Res.* DOI: 10.1093/nar/gkq1091
983 80 Pettersen, E.F. *et al.* (2004) UCSF Chimera - A visualization system for exploratory research and
984 analysis. *J. Comput. Chem.* DOI: 10.1002/jcc.20084
985 81 Agnarsson, I. *et al.* (2011) A time-calibrated species-level phylogeny of bats (chiroptera,
986 mammalia). *PLoS Curr.* DOI: 10.1371/currents.RRN1212
987 82 Lan, J. *et al.* (2020) Structure of the SARS-CoV-2 spike receptor-binding domain bound to the
988 ACE2 receptor. *Nature* DOI: 10.1038/s41586-020-2180-5
989

990 **Supplementary Materials**



991

992 *Supplementary Figure S1. Phylogenetic trees of additional orf1ab genes used for taxonomic*

993 *classification*. To investigate the robustness of the position of Lineage 4 in the RdRp phylogeny, we also

994 constructed phylogenies of nsp5 (3CLpro) and nsp13 (HEL1 core) using identical methods to those used

995 to generate Figure 1. While nsp5 supports the topology we observed for RdRp (nsp12), nsp13 supports

996 the positioning of Lineage 4 at the base of the tree instead. Because of the deep time scale and relatively

997 few sequences used to construct these trees, we must interpret hypotheses that depend on the branching

998 order with caution. The topology is robust to the inclusion or exclusion of the *Hibecovirus* sequence root

999 (data not shown). This pattern of inconsistency was also found for nsp14 and nsp15, with nsp14 matching

1000 the topology with Lineage 4 in an intermediate position and nsp15 matching the topology with Lineage 4

1001 at the base (data not shown). The roots of the trees were shortened for clarity.

1002 *Supplementary File 1. Excel spreadsheet of ACE2 amino acid alignment for host species of ACE2-using*
1003 *and non-ACE2-using viruses.* Host ACE2 sequences involved in interfacial interactions with the RBD of
1004 SARS-CoV-1 and SARS-CoV-2 are shown for human, civet (*Paguma larvata*), pangolin (*Manis*
1005 *javanica*), and species of bats that are known to both harbor ACE2-binding and non-ACE2-binding
1006 viruses (*Rhinolophus sinicus*) or only non-ACE2-binding viruses (*Rhinolophus macrotis, pearsonii,*
1007 *pusillus, ferrumequinum*). The ACE2 sequence from the African bat species from which the PDF-2370
1008 sample was taken is unidentified and also shown. At the time of publication, the ACE2 sequence of
1009 *Rhinolophus affinis* was not available. GenBank accession numbers for each sequence are provided.
1010 Distance in angstroms to the nearest SARS-CoV-1 (row 14) or SARS-CoV-2 (row 15) residues are shown
1011 and color coded according to the legend in row 18. Residues in hosts of non-ACE2-binders that differ
1012 from hosts of ACE2-binders (human, civet, pangolin, and *R. sinicus*) are outlined with black boxes.
1013
1014 *Supplementary File 2. Alignments used for building all phylogenetic trees included in this study.*
1015 Alignment files are provided in FASTA format and are named according to the Figure containing the
1016 phylogeny constructed from each one.