

1 December 14th, 2021

2 **Evidence for a mouse origin of the SARS-CoV-2 Omicron variant**

3 Changshuo Wei,^{1,2,3} Ke-Jia Shan,^{1,2,3} Weiguang Wang,^{1,2,3} Shuya Zhang,^{1,2} Qing Huan,^{1,*}
4 and Wenfeng Qian^{1,2,*}

5

6 ¹ State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental
7 Biology, Innovation Academy for Seed Design, Chinese Academy of Sciences, Beijing
8 100101, China

9 ² University of Chinese Academy of Sciences, Beijing 100049, China

10 ³ These authors contributed equally to this work

11 * Correspondence: qhuan@genetics.ac.cn (Q.H.) or wfqian@genetics.ac.cn (W.Q.)

12

13 **ABSTRACT**

14 The rapid accumulation of mutations in the SARS-CoV-2 Omicron variant that enabled
15 its outbreak raises questions as to whether its proximal origin occurred in humans or
16 another mammalian host. Here, we identified 45 point mutations that Omicron acquired
17 since divergence from the B.1.1 lineage. We found that the Omicron spike protein
18 sequence was subjected to stronger positive selection than that of any reported SARS-
19 CoV-2 variants known to evolve persistently in human hosts, suggesting the possibility of
20 host-jumping. The molecular spectrum (*i.e.*, the relative frequency of the twelve types of
21 base substitutions) of mutations acquired by the progenitor of Omicron was significantly
22 different from the spectrum for viruses that evolved in human patients, but was highly
23 consistent with spectra associated with evolution in a mouse cellular environment.
24 Furthermore, mutations in the Omicron spike protein significantly overlapped with
25 SARS-CoV-2 mutations known to promote adaptation to mouse hosts, particularly
26 through enhanced spike protein binding affinity for the mouse cell entry receptor.
27 Collectively, our results suggest that the progenitor of Omicron jumped from humans to
28 mice, rapidly accumulated mutations conducive to infecting that host, then jumped back
29 into humans, indicating an inter-species evolutionary trajectory for the Omicron outbreak.

30 **Keywords:** SARS-CoV-2; Omicron; evolutionary origin; molecular spectrum of
31 mutations; spike protein; ACE2

32 INTRODUCTION

33 The coronavirus disease 2019 (COVID-19) pandemic, caused by the SARS-CoV-2 RNA
34 virus, has led to significant illness and death worldwide. The SARS-CoV-2 Omicron
35 variant was first reported in South Africa on November 24th, 2021, and was designated as
36 a variant of concern (VOC) within two days by the World Health Organization (WHO)
37 based on the increase in infections attributable to this variant in South Africa (*i.e.*,
38 Omicron outbreak). In addition, the open reading frame encoding the spike protein (ORF
39 S) of Omicron harbors an exceptionally high number of mutations. These mutations are
40 particularly relevant to SARS-CoV-2 infection characteristics because the spike protein is
41 well-known to mediate viral entry into the host cell by interacting with angiotensin-
42 converting enzyme 2 (ACE2) on the cell surface (Zhou et al., 2020). In addition, the
43 spike protein is also a target for vaccine development and antibody-blocking therapy
44 (Huang et al., 2020; Martinez-Flores et al., 2021).

45 The proximal origins of Omicron have quickly become a controversial topic of heated
46 debate in the scientific and public health communities (Callaway, 2021; Kupferschmidt,
47 2021). Many mutations detected in Omicron were rarely reported among previously
48 sequenced SARS-CoV-2 variants (Shu and McCauley, 2017; Hadfield et al., 2018),
49 leading to three prevalent hypotheses regarding its evolutionary history. The first
50 hypothesis is that Omicron could have “cryptically spread” and circulated in a population
51 with insufficient viral surveillance and sequencing. Second, Omicron could have evolved
52 in a chronically infected COVID-19 patient, such as an immunocompromised individual
53 who provided a suitable host environment conducive to long-term virus adaptation. The
54 third possibility is that Omicron could have accumulated mutations in a nonhuman host
55 and then jumped into humans. Currently, the second scenario represents the most popular
56 hypothesis regarding the proximal origins of Omicron (Callaway, 2021; Kupferschmidt,
57 2021).

58 The first two hypotheses assume that Omicron acquired these mutations in humans
59 (collectively to as “human origin hypothesis” hereafter), while the third assumes that
60 Omicron acquired mutations in a nonhuman species. Based on our previous work in viral

61 evolution, we hypothesized that the host species in which Omicron acquired its particular
62 set of mutations could be determined by analyzing the molecular spectra of mutations
63 (*i.e.*, the relative frequency of the twelve types of base substitutions). In previous work,
64 we showed that *de novo* mutations in RNA virus genomes are generated in a replication-
65 independent manner and are highly dependent on mutagenic mechanisms specific to the
66 host cellular environment, resulting in overrepresentation with specific mutation types.
67 For example, reactive oxygen species (ROS) can oxidize guanine to 8-oxoguanine and
68 thereby induce the G>U transversion (Li et al., 2006; Kong and Lin, 2010), while
69 cytidine deaminases can induce RNA editing such as C>U transitions (Blanc and
70 Davidson, 2010; Harris and Dudley, 2015). Consistent with this phenomenon, viruses
71 belonging to different orders (*e.g.*, poliovirus, Ebola virus, and SARS-CoV-2) were found
72 to exhibit similar molecular spectra of mutations when evolving in the same host species,
73 while members of the same virus species exhibit divergent molecular spectra when
74 evolving in different host species (Shan et al., 2021). Since *de novo* mutations can thus
75 strongly influence the molecular spectrum of mutations that accumulate during virus
76 evolution in a host-specific manner, the host species in which Omicron acquired its
77 mutations could be determined by analyzing information carried by the mutations
78 themselves.

79 In this study, we identified mutations acquired by Omicron before its outbreak, and tested
80 whether the molecular spectrum of these mutations was consistent with the cellular
81 environment of human hosts. Prominent dissimilarities were observed between the
82 molecular spectrum of Omicron and a relatively comprehensive set of molecular spectra
83 from variants known to have evolved in humans, including those of three isolates from
84 chronic SARS-CoV-2 patients. Therefore, we next examined the molecular spectra of
85 mutations obtained from a wide range of host mammals for comparison with that of
86 Omicron. Finally, we used molecular docking-based analyses to investigate whether the
87 mutations in the Omicron spike protein could be associated with adaptation to the host
88 species inferred from molecular spectrum analysis. Our study provides insight into the
89 evolutionary trajectory and proximal origins of Omicron through careful scrutiny of its
90 mutations, and suggests strategies for avoiding future outbreaks caused by potentially
91 dangerous SARS-CoV-2 variants.

92

93 **RESULTS**

94 **Over-representation of nonsynonymous mutations in Omicron ORF *S* suggests** 95 **strong positive selection**

96 To first identify mutations that accumulated in the SARS-CoV-2 genome prior to the
97 Omicron outbreak, we constructed a phylogenetic tree that included the genomic
98 sequences of the reference SARS-CoV-2, two variants in the B.1.1 lineage which were
99 genetically close to Omicron, and 48 Omicron variants sampled before November 15th,
100 2021 (**Fig. 1A**). These two B.1.1 variants were sampled during April 22nd–May 5th, 2020,
101 which suggested that the progenitor of Omicron diverged from the B.1.1 lineage roughly
102 in mid-2020. Intermediate versions have gone largely undetected, thus resulting in an
103 exceptionally long branch leading to the most recent common ancestor (MRCA) of
104 Omicron in the phylogenetic tree (**Fig. 1A**). We hereafter refer to this long branch as
105 Branch O.

106 We identified 45 point mutations that were introduced in Branch O (hereafter referred to
107 as “pre-outbreak Omicron mutations”; **Fig. 1A**). Visual assessment suggested that the
108 pre-outbreak Omicron mutations were over-represented in ORF *S* (**Fig. 1B**). To test if the
109 rate at which mutations accumulated in ORF *S* was accelerated in Branch O, we
110 randomly sampled one SARS-CoV-2 variant per day since December 24th, 2019 from the
111 Global Initiative on Sharing All Influenza Data (GISAID) (Shu and McCauley, 2017) to
112 compare mutation accumulation rates among different variants. We found that mutations
113 accumulated in ORF *S* at a rate of ~0.45 mutations per month on average. In sharp
114 contrast, 27 mutations accumulated in ORF *S* in Branch O during the 18 months spanning
115 May 2020–November 2021, equivalent to ~1.5 mutations per month, or ~3.3 times faster
116 than the average rate of other variants (**Fig. 1C**).

117 Counting mutations across the whole SARS-CoV-2 genome indicated that Omicron
118 acquired mutations in the genome at a similar rate to other variants (**Fig. 1D**), suggesting
119 that the accelerated evolutionary rate of ORF *S* could not be explained by an overall

120 elevated mutation rate in Omicron progenitors. In light of these findings, we
121 hypothesized that positive selection could have helped accelerate the evolutionary rate of
122 ORF *S*. To test this hypothesis, we sought to infer the strength of positive selection by
123 estimating the ratio of nonsynonymous to synonymous mutations. Twenty-six of the 27
124 pre-outbreak mutations in the ORF *S* of Omicron were nonsynonymous (**Fig. 1E**),
125 resulting in a d_N/d_S ratio of 6.64, significantly greater than a d_N/d_S of 1.00 ($P = 0.03$,
126 Fisher's exact test). These results indicated that positive selection contributed to
127 increasing the mutation rate in ORF *S* in Branch O.

128 To test if such a level of positive selection is common among SARS-CoV-2 variants, we
129 counted the number of nonsynonymous and synonymous mutations in ORF *S* in other
130 SARS-CoV-2 VOC lineages (*i.e.*, Alpha, Beta, Gamma, and Delta) as well as in the
131 genomes of SARS-CoV-2 variants isolated from three chronically infected patients
132 (Kemp et al., 2021; Truong et al., 2021). None of these other VOCs or isolates exhibited
133 comparable numbers of nonsynonymous mutations as that of mutations in Branch O (**Fig.**
134 **1E**). These observations strongly suggested that the Omicron variant had undergone a
135 strong positive selection for the spike protein that no other known SARS-CoV-2 variants
136 evolved in humans had been subjected to. Considering that the spike protein determines
137 the host range of a coronavirus (*i.e.*, which organisms it can infect), we therefore
138 hypothesized that the progenitor of Omicron might host-jump from humans to a
139 nonhuman species because this process would require substantial mutations in the spike
140 protein for rapid adaptation to a new host.

141

142 **The molecular spectrum of pre-outbreak Omicron mutations is inconsistent with an** 143 **evolutionary history in humans**

144 Previous studies showed that the molecular spectrum of mutations that accumulate in a
145 viral genome reflect a host-specific cellular environment (Deng et al., 2021; Shan et al.,
146 2021). To test the human origin hypothesis of Omicron, we compared the molecular
147 spectrum of the 45 pre-outbreak Omicron mutations with the “standard” molecular
148 spectrum for SARS-CoV-2 variants known to have evolved strictly in humans (hereafter

149 referred to as “the hSCV2 spectrum”, **Fig. 2A**). The hSCV2 spectrum included 6,986
150 point mutations that were compiled from 34,853 high-quality sequences of SARS-CoV-2
151 variants isolated from patients worldwide (Shan et al., 2021). We found that the
152 molecular spectrum of the pre-outbreak Omicron mutations was significantly different
153 from the hSCV2 spectrum ($P = 0.004$, G -test, **Fig. 2B**). In particular, transitions were
154 more abundant than transversions and C>U mutation was more abundant than its
155 complementary mutation G>A, as in the hSCV2 spectrum; However, a hallmark of RNA
156 virus mutations when evolving in humans—a higher rate of G>U mutation than its
157 complementary mutation C>A (Panchin and Panchin, 2020; De Maio et al., 2021; Deng
158 et al., 2021; Shan et al., 2021), likely caused by cellular ROS—was absent in the pre-
159 outbreak Omicron mutations.

160 To exclude the possibility that this apparent difference in the molecular spectrum was
161 caused by the relatively small number of pre-outbreak Omicron mutations, we generated
162 100 “pseudo” variants *in silico* by randomly down sampling 45 mutations from the
163 hSCV2 spectrum. None of the pseudo variants showed smaller P values (based on G -
164 tests) than that obtained using the pre-outbreak Omicron mutations (**Fig. 2C**), nor did the
165 SARS-CoV-2 isolates with mutations known to be acquired in the three chronically
166 infected patients (# of mutations are 30, 47, and 81, **Fig. 2C**). These observations
167 indicated that the difference between the molecular spectrum of pre-outbreak Omicron
168 mutations and the hSCV2 spectrum could not be strictly attributed to statistical
169 randomness.

170 To exclude the possibility that some mutations which occurred early in the evolution of
171 Omicron distorted the molecular spectrum of mutations that accumulated afterward, we
172 identified 120 point mutations on top of the MRCA of Omicron, by screening 695
173 Omicron variants collected spanning November 8th–December 7th, 2021 (hereafter
174 referred to as “post-outbreak Omicron mutations”). The molecular spectrum of these
175 post-outbreak Omicron mutations was not significantly different from the hSCV2
176 spectrum (**Fig. 2B–C**). This finding indicated that Omicron acquired mutations following
177 the same molecular spectrum as other SARS-CoV-2 variants during its evolution in

178 human hosts. Collectively, these molecular spectrum analyses revealed that pre-outbreak
179 Omicron mutations were unlikely to have been acquired in humans.

180

181 **The molecular spectrum of pre-outbreak Omicron mutations is consistent with an** 182 **evolutionary history in mice**

183 In light of our findings that Omicron may have evolved in another host before its
184 outbreak, we next sought to determine the nonhuman host species in which these
185 mutations accumulated. To this end, we first characterized the molecular spectra of
186 coronaviruses that evolved in different host species for comparison with that of Omicron.
187 Specifically, we retrieved 17 sequences of murine hepatitis viruses, 13 canine
188 coronaviruses, 54 feline coronaviruses, 23 bovine coronaviruses, and 110 porcine
189 deltacoronaviruses (**Table S1**), constructed the phylogenetic tree for the coronaviruses
190 isolated from each host species (canine coronavirus as an example shown in **Fig. 3A** and
191 the rest are shown in **Fig. S1**), and identified the mutations that accumulated in each
192 branch (**Fig. 3A**). We also included some previously reported molecular spectra (Shan et
193 al., 2021), including 17 spectra of mutations acquired by SARS-CoV-, SARS-CoV-2-,
194 and MERS-CoV-related coronaviruses during their evolution in bats, two spectra of
195 camel MERS-CoV, one spectrum estimated from 807 MERS-CoV mutations
196 accumulated in human (the hMERS spectrum), as well as the hSCV2 spectrum.
197 Furthermore, we also included the molecular spectrum of mutations identified in an early
198 variant of each of the other four VOCs.

199 We performed principal component analysis to reduce the dimensionality of the
200 molecular spectrum of mutations, and subsequently visualized the data using the first two
201 principal components (**Fig. 3B**). Consistent with the results of our previous study (Shan
202 et al., 2021), drawing 95% confidence ellipses for each host species showed that the
203 molecular spectra clustered according to their respective hosts (**Fig. 3B**), likely because
204 viruses evolving in the same host species share the mutagens specific to that host's
205 cellular environment. Supporting this point, the molecular spectrum of post-outbreak
206 Omicron mutations (which are known to have accumulated in humans) was located

207 within the human 95% confidence ellipse. In contrast, the molecular spectrum of pre-
208 outbreak Omicron mutations was within the mouse ellipse, suggesting that the pre-
209 outbreak mutations accumulated in a rodent (in particular mouse) host.

210

211 **Pre-outbreak Omicron mutations in the spike protein significantly overlap with** 212 **mutations in mouse-adapted SARS-CoV-2**

213 Mice were previously reported to serve as poor hosts for SARS-CoV-2 because the spike
214 protein of early SARS-CoV-2 variants exhibit low-affinity interactions with mouse ACE2
215 (Lam et al., 2020; Zhou et al., 2020; Ren et al., 2021; Wong et al., 2021). However, over
216 the course of the pandemic SARS-CoV-2 variants emerged that could infect mice. For
217 example, variants harboring the spike mutation N501Y, which are relatively common in
218 human patients (24.7%), could infect mice (Gu et al., 2020; Leist et al., 2020; Sun et al.,
219 2021). If the progenitor of Omicron indeed evolved in a mouse species before the
220 Omicron outbreak, we postulated that its spike protein likely adapted through increased
221 binding affinity for mouse ACE2. To test this possibility, we projected the pre-outbreak
222 Omicron mutations in the spike protein onto a three-dimensional structural model of the
223 spike:ACE2 complex (Lan et al., 2020). Seven mutations (*i.e.*, K417N, G446S, E484A,
224 Q493R, G496S, Q498R, and N501Y) were located at the interface of ACE2 and the
225 receptor-binding domain (RBD) of the spike protein and could potentially affect their
226 interactions (**Fig. 4A**).

227 Previous studies of SARS-CoV-2 variants isolated from mice reported specific amino
228 acid mutations in the spike protein that could promote its interactions with mouse ACE2
229 (Leist et al., 2020; Wu et al., 2020b; Huang et al., 2021; Montagutelli et al., 2021; Sun et
230 al., 2021; Wong et al., 2021; Zhang et al., 2021). In addition, previous studies have
231 described some reverse zoonotic events (*e.g.*, from humans to other mammals such as
232 mink and white-tailed deer) for SARS-CoV-2 (Chandler et al., 2021; Oude Munnink et
233 al., 2021), and the variants isolated from these mammalian hosts presumably harbored
234 amino acid mutations that could potentially participate in their adaptation to these hosts.
235 Thus, if the progenitor of Omicron evolved in mice and adapted to mouse ACE2, we

236 predicted that the pre-outbreak Omicron mutations should share considerable overlap
237 with the mutations identified in these mouse-adapted SARS-CoV-2 variants, but not
238 those of other mammalian species.

239 To test this prediction, we identified the mutations in ORF *S* of SARS-CoV-2 variants
240 isolated from 15 mammalian species (*e.g.*, mice, cats, dogs, minks, and deer, **Table S2**)
241 and found that pre-outbreak Omicron mutations tended to share the same positions as the
242 ORF *S* mutations identified in mice (odds ratio = 231.4, $P = 1.6 \times 10^{-11}$, Fisher's exact
243 test, **Fig. 4B–C**). In contrast, same statistical test showed much lower odds ratios and
244 significance levels for overlap in these mutations with other species (**Fig. 4C**). Pre-
245 outbreak Omicron mutations also overlapped with some mutations detected in isolates
246 from chronically infected patients (Kemp et al., 2021; Truong et al., 2021), although they
247 too showed substantially lower odds ratios and significance levels (**Fig. 4C**). These
248 observations implied that the pre-outbreak Omicron mutations in ORF *S* promoted its
249 adaptation to a mouse host.

250 We then conducted enrichment analysis for each of the seven mouse-adapted SARS-
251 CoV-2 variants and observed statistical significance for all these variants (**Fig. 4D**). In
252 particular, we observed amino acid mutations at residues 493 and 498 in five and six of
253 the seven mouse-adapted SARS-CoV-2 variants, respectively (**Fig. 4D**). Identical amino
254 acid mutations (*i.e.*, Q493R and Q498R) were both observed in two variants
255 (Montagutelli et al., 2021; Wong et al., 2021), and considering that these two amino acid
256 mutations are uncommon in human patients infected by non-Omicron SARS-CoV-2
257 variants (0.005% and 0.002%, respectively) we concluded that the progenitor of Omicron
258 evolved in mouse species (or at least rodent species).

259

260 **Pre-outbreak Omicron mutations in the spike protein significantly enhance binding** 261 **affinity with mouse ACE2**

262 To investigate the mechanisms by which the pre-outbreak Omicron mutations in the spike
263 protein could have contributed to its adaptation to a mouse host, we examined their

264 interaction through molecular docking analysis of the spike protein RBD and mouse
265 ACE2 (**Fig. 5A**). Following previous studies (Lam et al., 2020; Rodrigues et al., 2020),
266 we estimated the HADDOCK score (van Zundert et al., 2016), which is positively
267 associated with the dissociation constant (K_D , with smaller K_D indicating stronger
268 binding) of protein interactions (Kastritis and Bonvin, 2010), and can be used to predict
269 the susceptibility of a mammalian species to infection with SARS-CoV-2 (Rodrigues et
270 al., 2020).

271 To confirm the accuracy of inferences regarding the binding affinity between spike
272 protein RBD and ACE2 based on the HADDOCK score, we calculated the HADDOCK
273 score for eight experimentally determined K_D values between four RBD variants and
274 human (or mouse) ACE2 (Sun et al., 2021). The HADDOCK scores were positively
275 correlated with the K_D values in the analysis (Pearson's correlation coefficient $r = 0.93$, P
276 $= 0.002$, **Fig. S2A–C**), thus supporting the validity of molecular docking-based
277 predictions of ACE2-binding affinity for other RBD variants.

278 The molecular docking-based predictions suggested that the RBD of Omicron exhibited
279 higher binding affinity for mouse ACE2 than that of RBD encoded in the reference
280 SARS-CoV-2 genome, further suggesting an evolutionary history in mice (**Fig. 5B**). And
281 as expected, the mutations detected in the RBD of the other four VOCs of SARS-CoV-2
282 as well as those of variants isolated from chronically infected human patients showed no
283 apparent changes in their binding affinity for mouse ACE2 compared with the reference
284 RBD (**Fig. 5B**).

285 Since five amino acid mutations were shared between Omicron and mouse-adapted
286 SARS-CoV-2 variants in RBD (*i.e.*, K417N, E484A, Q493R, Q498R, and N501Y; **Fig.**
287 **4B**), and that they together enhanced RBD binding affinity for mouse ACE2 (**Fig. 5B**),
288 we next determined the individual effects of each of these five mutations. Notably, only
289 Q493R and Q498R significantly increased the binding affinity with mouse ACE2, which
290 was consistent with their repeated detection in mouse-adapted SARS-CoV-2 variants
291 (Montagutelli et al., 2021; Wong et al., 2021). Indeed, docking analysis showed that
292 Q493R/Q498R double mutation could further increase the RBD binding affinity for

293 mouse ACE2 (**Fig. 5B**). By contrast, the other three mutations showed no significant
294 effects on the binding affinity between RBD and mouse ACE2, neither in the reference
295 RBD nor in the Q493R/Q498R double mutant (**Fig. 5B**), suggesting that they did not
296 contribute to the enhanced interaction between Omicron RBD and mouse ACE2. Indeed,
297 previous studies showed that K417N, E484K, and N501Y were related to escape from
298 neutralizing antibodies (Li et al., 2021; Nelson et al., 2021).

299 **Mice as the most likely rodent species in which the progenitor of Omicron evolved**

300 While the observations regarding both the molecular spectrum of mutations and the
301 RBD-ACE2 interaction suggested that mice were candidate host species in which the
302 progenitor of Omicron evolved, it remained plausible that Omicron evolved in some
303 other rodent species with similar cellular mutagen environment and ACE2 structure to
304 mice. We postulated that if Omicron evolved in another rodent species, the amino acid
305 mutations in Omicron RBD should elevate its interaction with the ACE2 of this host. To
306 test this prediction, we applied molecular docking analysis to four additional rodent
307 species representing different lineages of rodents (Kumar et al., 2017)—brown rats
308 (*Rattus norvegicus*), guinea pigs (*Cavia porcellus*), golden hamsters (*Mesocricetus*
309 *auratus*), and Daurian ground squirrels (*Spermophilus dauricus*)—as well as a close
310 relative of rodents, European rabbits (*Oryctolagus cuniculus*). Omicron RBD showed
311 higher ACE2-binding affinity (compared with the reference RBD) only to the mouse
312 ACE2 (**Fig. 5C**), suggesting that mice are the most likely host species in which the
313 progenitor of Omicron evolved.

314 **DISCUSSION**

315 In this study, we used the molecular spectrum of mutations of the SARS-CoV-2 Omicron
316 variant to trace its proximal host origins. We found that the molecular spectrum of pre-
317 outbreak Omicron mutations was inconsistent with the rapid accumulation of mutations
318 in humans, but rather suggested a trajectory in which the progenitor of Omicron
319 experienced a reverse zoonotic event from humans to mice sometime during the
320 pandemic (most likely in mid-2020) and accumulated mutations in a rodent host (most
321 likely mouse) for more than one year before jumping back to humans in late-2021. While
322 evolving in mice, the progenitor of Omicron adapted to the mouse host by acquiring
323 amino acid mutations in the spike protein that increased its binding affinity with mouse
324 ACE2. In addition, mutations associated with immune escape also accumulated, which
325 may also be a contributing factor in its rapid spread.

326 While we show a phylogenetically long branch leading to the MRCA of current Omicron
327 variants (*i.e.*, Branch O), it is worth noting that intermediate versions of Omicron were
328 occasionally reported. For example, a SARS-CoV-2 variant (EPI_ISL_7136300) was
329 collected by the Utah Public Health Laboratory on December 1st, 2021 which harbored 32
330 of the 45 pre-outbreak Omicron mutations. However, the 13 mutations absent in this
331 variant clustered within residues 371–501 of the spike protein (**Fig. S3**). The absence of
332 these spike protein mutations thus suggested that this variant was a product of
333 recombination between an Omicron variant and another SARS-CoV-2 variant, rather than
334 a direct progenitor of Omicron. Considering the large number of pre-outbreak Omicron
335 mutations (45) combined with the sparsity of intermediate versions identified to date, this
336 long branch leading to Omicron in our phylogenetic reconstruction remains valid.

337 Although we primarily focused on point mutations because the molecular spectrum of
338 these mutations can reflect the host cellular environment (Deng et al., 2021; Shan et al.,
339 2021), we also used the information of deletions and insertions to infer the evolutionary
340 trajectory of Omicron. For example, a B.1.1 variant (EPI_ISL_493480) shared the same
341 deletion (Δ 105–107 in non-structural protein 6) as the Omicron variants, which was used
342 to infer that B.1.1 is a close relative of Omicron. In addition, spike Δ 69–70 deletion is

343 shared by Omicron and many non-Omicron variants isolated from patients (Meng et al.,
344 2021), but is absent in the early samples of SARS-CoV-2 (Wu et al., 2020a), strongly
345 suggesting that the progenitor of Omicron was jumped from humans to mice during the
346 pandemic.

347

348 In addition, we noted that Omicron harbored a nine nucleotide insertion (GAGCCAGAA,
349 encoding the peptide EPE) after residue 214 in the spike protein. This insertion is
350 identical to the sequence of *TMEM245* in the human genome or that of ORF *S* in the
351 human coronavirus hCoV-229E, which was used as evidence to support a human origin
352 for Omicron (Venkatakrisnan et al., 2021). However, we provide a simpler explanation
353 for this insertion, namely that it was derived from an RNA fragment of ORF *N* in the
354 SARS-CoV-2 genome (**Fig. S4**). We believe that the insertion of an ORF *N* fragment is
355 more likely because the RNA abundance of ORF *N* is much higher than that of mRNA
356 encoded by the human genome (Wei et al., 2021). That is, in SARS-CoV-2-infected cells,
357 a substantial proportion of RNAs are viral, and especially so for ORF *N* due to the nested
358 nature of the coronavirus genome and subgenomes (Kim et al., 2020).

359 It also warrants mention that all of the mouse-adapted SARS-CoV-2 variants were
360 amplified/purified in Vero cells (a cell line originally isolated from the kidney of green
361 monkey) at some stage of experimentation, which could impose an additional selection
362 pressure to enhance the spike protein binding affinity towards primate ACE2 (Leist et al.,
363 2020; Wu et al., 2020b; Huang et al., 2021; Montagutelli et al., 2021; Sun et al., 2021;
364 Wong et al., 2021; Zhang et al., 2021). Consistent with this experimental process, the
365 amino acid mutations acquired by Omicron and mouse-adapted viruses were not always
366 identical, even if mutations occurred at the same residue. For example, Q493H and
367 Q493K were also detected in the mouse-adapted SARS-CoV-2 at residue 493 of the spike
368 protein, in addition to mutations observed in Omicron (Q493R). Different from the
369 effects of Q493R, these two mutations increased the binding affinity toward both mouse
370 and human ACE2 (**Fig. S2C**), indicating that SARS-CoV-2 could potentially evolve
371 remarkably high diversity in its adaptation to ACE2 from various host species. Consistent

372 with this possibility, numerous mutations were also identified in the spike protein of
373 SARS-CoV-2 RNA fragment amplified from wastewater samples (Smyth et al., 2021).

374 Humans represent the largest known reservoir of SARS-CoV-2. Our study suggests that
375 SARS-CoV-2 could have spilled over from humans to wild animals, and that the variants
376 which successfully infected animal hosts could then accumulate new mutations before
377 jumping back into humans as a variant of concern. Given the ability of SARS-CoV-2 to
378 jump across various species, it appears likely that global populations will face additional
379 animal-derived variants until the pandemic is well under control. Viral surveillance and
380 sequencing in wild animals will likely help to prevent future outbreaks of dangerous
381 SARS-CoV-2 variants.

382 **METHODS**

383 **Identification of pre-outbreak and post-outbreak Omicron mutations**

384 Genomic sequences of 695 SARS-CoV-2 Omicron variants were downloaded from
385 GISAID (<https://www.gisaid.org/>) on December 7th, 2021. The reference genome of
386 SARS-CoV-2 (EPI_ISL_402125) and two variants in the B.1.1 lineage (EPI_ISL_698296
387 and EPI_ISL_493480) were also downloaded from GISAID.

388 The genomes of SARS-CoV-2 variants were aligned by MUSCLE v3.8.1551 (Edgar,
389 2004). The phylogenetic tree and ancestral sequences were reconstructed using FastML
390 v3.11 (Ashkenazy et al., 2012) with default parameters. The single-nucleotide
391 substitutions obtained by the most recent common ancestor (MRCA) of Omicron variants
392 after its divergence from the B.1.1 lineage were defined as pre-outbreak Omicron
393 mutations. To detect the post-outbreak Omicron mutations, the sequences of 695
394 Omicron variants were aligned to the Omicron's MRCA sequence, and sequences with
395 >10 single-nucleotide substitutions were discarded. The single-nucleotide substitutions
396 detected in at least two variants were defined as the post-outbreak Omicron mutations.

397 The numbers of synonymous and nonsynonymous sites in ORF *S* of SARS-CoV-2 were
398 estimated by PAML in a previous study (Wei et al., 2021). Briefly, d_N was calculated as
399 the ratio between the number of nonsynonymous mutations and the number of
400 nonsynonymous sites, while d_S was calculated as the ratio between the number of
401 synonymous mutations and the number of synonymous sites.

402 The frequencies of three mutations (Q493R, Q498R, and N501Y) among patients were
403 retrieved from CoV-GLUE-Viz (<http://cov-glue-viz.cvr.gla.ac.uk/>) updated at November
404 23th, 2021.

405 **Comparison between the sequence evolutionary rate of Omicron and other SARS-** 406 **CoV-2 variants**

407 A total of 764 variant sequences were randomly sampled from the SARS-CoV-2 genomic
408 sequences deposited at GISAID, one variant each day since COVID-19 outbreak. The
409 progenitors of other four VOCs (Alpha, Beta, Gamma, and Delta) were retrieved from

410 Nextstrain (<https://nextstrain.org/>) (Hadfield et al., 2018). Single-nucleotide substitutions
411 (relative to the reference genome) of each variant were defined as the mutations acquired
412 by the SARS-CoV-2 variant. The single-nucleotide base substitutions of three chronically
413 infected patients were retrieved from two previous studies (Kemp et al., 2021; Truong et
414 al., 2021). The mutations with allele frequency >50% on the final monitored day were
415 used to count mutations that accumulated in a chronically infected patient.

416 We performed resampling test to estimate the statistical significance. Specifically, we
417 randomly sampled 45 mutations from the 6,986 point mutations identified in a previous
418 study from the 34,853 high-quality sequences of SARS-CoV-2 variants isolated from
419 patients worldwide (Shan et al., 2021). This operation was repeated 100 times *in silico*.

420 **Characterization of molecular spectra of mutations**

421 Complete genomic sequences of 23 bovine coronavirus (*Betacoronavirus 1*), 13 canine
422 coronavirus (*Alphacoronavirus 1*), 54 feline coronavirus (*Alphacoronavirus 1*), 17
423 murine hepatitis virus (*Murine coronavirus*), and 110 porcine deltacoronavirus
424 (*Coronavirus HKU15*) were downloaded from National Center for Biotechnology
425 Information (NCBI) Virus database (<https://www.ncbi.nlm.nih.gov/labs/virus/vssi/>)
426 (Hatcher et al., 2017), querying the hosts as *Bos taurus* (cattle), *Canis lupus familiaris*
427 (dogs), *Felis catus* (cats), *Mus musculus* (mice), and *Sus scrofa* (pigs), respectively
428 (**Table S1**).

429 The virus genome sequences were aligned by MUSCLE, and the phylogenetic trees and
430 ancestral sequences were reconstructed using FastML. Since the roots of these
431 phylogenetic trees were not readily identified, we kept only external branches to ensure
432 the correction direction of base substitutions (*e.g.*, C>U vs. U>C). For the sake of clarity,
433 we showed the molecular spectra for five branches with the largest number of mutations
434 for each coronavirus species in the main text. The full data set is available in **Table S1**.

435 We characterized the molecular spectra of mutations accumulated in chronically infected
436 patients, in which single-nucleotide base substitutions that ever occurred during the
437 monitored period were counted. We downloaded the genomic sequences of four variants
438 (EPI_ISL_5803018, EPI_ISL_3730369, EPI_ISL_4003132, and EPI_ISL_6260720),

439 each from one of the other four VOCs (Alpha, Beta, Gamma, and Delta, respectively), to
440 estimate the molecular spectra of mutations accumulated in VOCs.

441 **Principal component analyses**

442 We performed principal component analysis (prcomp function in *R*) with the proportions
443 of the 12 base-substitution types as the input, and then projected molecular spectra into a
444 two-dimensional space according to the first two principal components. To define the
445 borderlines of molecular spectra for each host species (*i.e.*, cattle, bats, dogs, cats, mice,
446 pigs, or humans), we estimated the 95% confidence ellipses (stat_ellipse option in *R*)
447 from the molecular spectra of these host species. The spectra of pre- and post-outbreak
448 Omicron mutations were further projected into the same two-dimensional space.

449 **Comparison of pre-outbreak Omicron mutations with mutations detected in SARS-** 450 **COV-2 variants isolated from various mammalian hosts**

451 We downloaded from GISAID the genomic sequences of SARS-CoV-2 variants isolated
452 from 18 mammalian hosts (**Table S2**): *Aonyx cinereus* (Asian small-clawed otter);
453 *Arctictis binturong* (binturong); *Canis lupus familiaris* (dog); *Crocuta crocuta* (spotted
454 hyena); *Felis catus* (cat); *Gorilla gorilla* (western gorilla); *Mus musculus* (mouse);
455 *Mustela furo* (ferret); *Neovison vison* (American mink); *Odocoileus virginianus* (white-
456 tailed deer); *Panthera leo* (lion), *Panthera tigris* (tiger); *Panthera uncia* (snow leopard);
457 *Prionailurus bengalensis* (leopard cat); *Prionailurus viverrinus* (fishing cat);
458 *Mesocricetus auratus* (golden hamster); *Chlorocebus sabaeus* (green monkey) and *Puma*
459 *concolor* (puma). BLASTx was performed to identify ORF *S* in each variant, and
460 mutations were identified at the same time. Three species (*Mesocricetus auratus*,
461 *Chlorocebus sabaeus*, and *Puma concolor*) were discarded because they harbored less
462 than three single amino acid mutations. Amino acid mutation data from three additional
463 viruses isolated from mice were retrieved from three studies (Leist et al., 2020;
464 Montagutelli et al., 2021; Sun et al., 2021).

465 **Estimation of the binding affinity of RBD-ACE2 interaction by molecular docking**

466 We extracted three-dimensional structures of the spike RBD and human ACE2 from the
467 crystal structure (PDB: 6M0J) reported in a previous study (Lan et al., 2020), and those
468 of rodent ACE2 from the predicted models reported in a previous study (Lam et al.,
469 2020). The structure models of the Omicron RBD and the RBD with five mutations
470 (K417N, E484A, Q493R, Q498R, and N501Y) were generated using SWISS-MODEL
471 (Waterhouse et al., 2018), and those of other RBD variants were generated using PyMOL
472 “mutagenesis” (<https://pymol.org/>). The structure models of the RBD:ACE2 complex
473 were generated by aligning against the reported complex structure of the corresponding
474 species using PyMOL (Lam et al., 2020; Lan et al., 2020).

475 We performed molecular docking following previous studies (Lam et al., 2020;
476 Rodrigues et al., 2020). Briefly, we refined the three-dimensional models using default
477 refinement protocols, and then estimated the HADDOCK scores for each RBD:ACE2
478 complex using HADDOCKv2.4 web server (van Zundert et al., 2016). Docking results of
479 each RBD-ACE2 variant pair were clustered, and the average HADDOCK score of the
480 top cluster was reported for the RBD:ACE2 complex.

481

482 **ACKNOWLEDGMENTS**

483 We thank Dr. Xionglei He from Sun Yat-sen University and Dr. Mingkun Li from
484 Beijing Institute of Genomics CAS for discussion. We acknowledge the authors and
485 laboratories for generating and submitting the sequences to GISAID Database on which
486 this research is based. The list is detailed in Table S2. This work was supported by grants
487 from the National Natural Science Foundation of China (31922014).

488

489 **SUPPLEMENTARY MATERIAL**

490 Supplementary material includes Supplemental Figures S1–4 and Supplemental Tables
491 S1–2.

492

493 **AUTHOR CONTRIBUTIONS**

494 W.Q. designed the study; C.W., K.-J.S., W.W., and S.Z. performed data analyses; C.W.,
495 K.-J.S., Q.H., and W.Q. wrote the manuscript.

496

497 **DECLARATION OF INTERESTS**

498 The authors declare that they have no competing interests.

499

500 **DATA AVAILABILITY**

501 All scripts used to analyze the data and to generate the figures are available at github
502 (https://github.com/ChangshuoWei/Omicron_origin) and Zenodo (DOI:
503 10.5281/zenodo.5778199). All data that were used to support the findings of this study
504 are available in the public databases.

505 REFERENCES

- 506 Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O., Pupko, T.,
507 2012. Fastml: A web server for probabilistic reconstruction of ancestral sequences.
508 Nucleic Acids Res 40, W580-584.
- 509 Blanc, V., Davidson, N.O., 2010. Apobec-1-mediated rna editing. Wiley Interdiscip Rev Syst Biol
510 Med 2, 594-602.
- 511 Callaway, E., 2021. Heavily mutated omicron variant puts scientists on alert. Nature.
- 512 Chandler, J.C., Bevins, S.N., Ellis, J.W., Linder, T.J., Tell, R.M., Jenkins-Moore, M., Root, J.J.,
513 Lench, J.B., Robbe-Austerman, S., DeLiberto, T.J., *et al.*, 2021. Sars-cov-2 exposure in
514 wild white-tailed deer (*odocoileus virginianus*). Proc Natl Acad Sci U S A 118.
- 515 De Maio, N., Walker, C.R., Turakhia, Y., Lanfear, R., Corbett-Detig, R., Goldman, N., 2021.
516 Mutation rates and selection on synonymous mutations in sars-cov-2. Genome Biol Evol
517 13.
- 518 Deng, S., Xing, K., He, X., 2021. Mutation signatures inform the natural host of sars-cov-2.
519 National Science Review.
- 520 Edgar, R.C., 2004. Muscle: A multiple sequence alignment method with reduced time and space
521 complexity. BMC Bioinformatics 5, 113.
- 522 Gu, H., Chen, Q., Yang, G., He, L., Fan, H., Deng, Y.Q., Wang, Y., Teng, Y., Zhao, Z., Cui, Y., *et al.*,
523 2020. Adaptation of sars-cov-2 in balb/c mice for testing vaccine efficacy. Science 369,
524 1603-1607.
- 525 Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford,
526 T., Neher, R.A., 2018. Nextstrain: Real-time tracking of pathogen evolution.
527 Bioinformatics 34, 4121-4123.
- 528 Harris, R.S., Dudley, J.P., 2015. Apobec and virus restriction. Virology 479-480, 131-145.
- 529 Hatcher, E.L., Zhdanov, S.A., Bao, Y., Blinkova, O., Nawrocki, E.P., Ostapchuck, Y., Schaffer,
530 A.A., Brister, J.R., 2017. Virus variation resource - improved response to emergent viral
531 outbreaks. Nucleic Acids Res 45, D482-D490.
- 532 Huang, K., Zhang, Y., Hui, X., Zhao, Y., Gong, W., Wang, T., Zhang, S., Yang, Y., Deng, F., Zhang,
533 Q., *et al.*, 2021. Q493k and q498h substitutions in spike promote adaptation of sars-cov-
534 2 in mice. EBioMedicine 67, 103381.
- 535 Huang, Y., Yang, C., Xu, X.F., Xu, W., Liu, S.W., 2020. Structural and functional properties of sars-
536 cov-2 spike protein: Potential antiviral drug development for covid-19. Acta Pharmacol
537 Sin 41, 1141-1149.
- 538 Kastritis, P.L., Bonvin, A.M., 2010. Are scoring functions in protein-protein docking ready to
539 predict interactomes? Clues from a novel binding affinity benchmark. J Proteome Res 9,
540 2216-2225.
- 541 Kemp, S.A., Collier, D.A., Datir, R.P., Ferreira, I., Gayed, S., Jahun, A., Hosmillo, M., Rees-Spear,
542 C., Mlcochova, P., Lumb, I.U., *et al.*, 2021. Sars-cov-2 evolution during treatment of
543 chronic infection. Nature 592, 277-282.
- 544 Kim, D., Lee, J.Y., Yang, J.S., Kim, J.W., Kim, V.N., Chang, H., 2020. The architecture of sars-cov-2
545 transcriptome. Cell 181, 914-921 e910.
- 546 Kong, Q., Lin, C.L., 2010. Oxidative damage to rna: Mechanisms, consequences, and diseases. Cell
547 Mol Life Sci 67, 1817-1829.
- 548 Kumar, S., Stecher, G., Suleski, M., Hedges, S.B., 2017. Timetree: A resource for timelines,
549 timetrees, and divergence times. Mol Biol Evol 34, 1812-1819.
- 550 Kupferschmidt, K., 2021. Where did 'weird' omicron come from? Science 374, 1179.

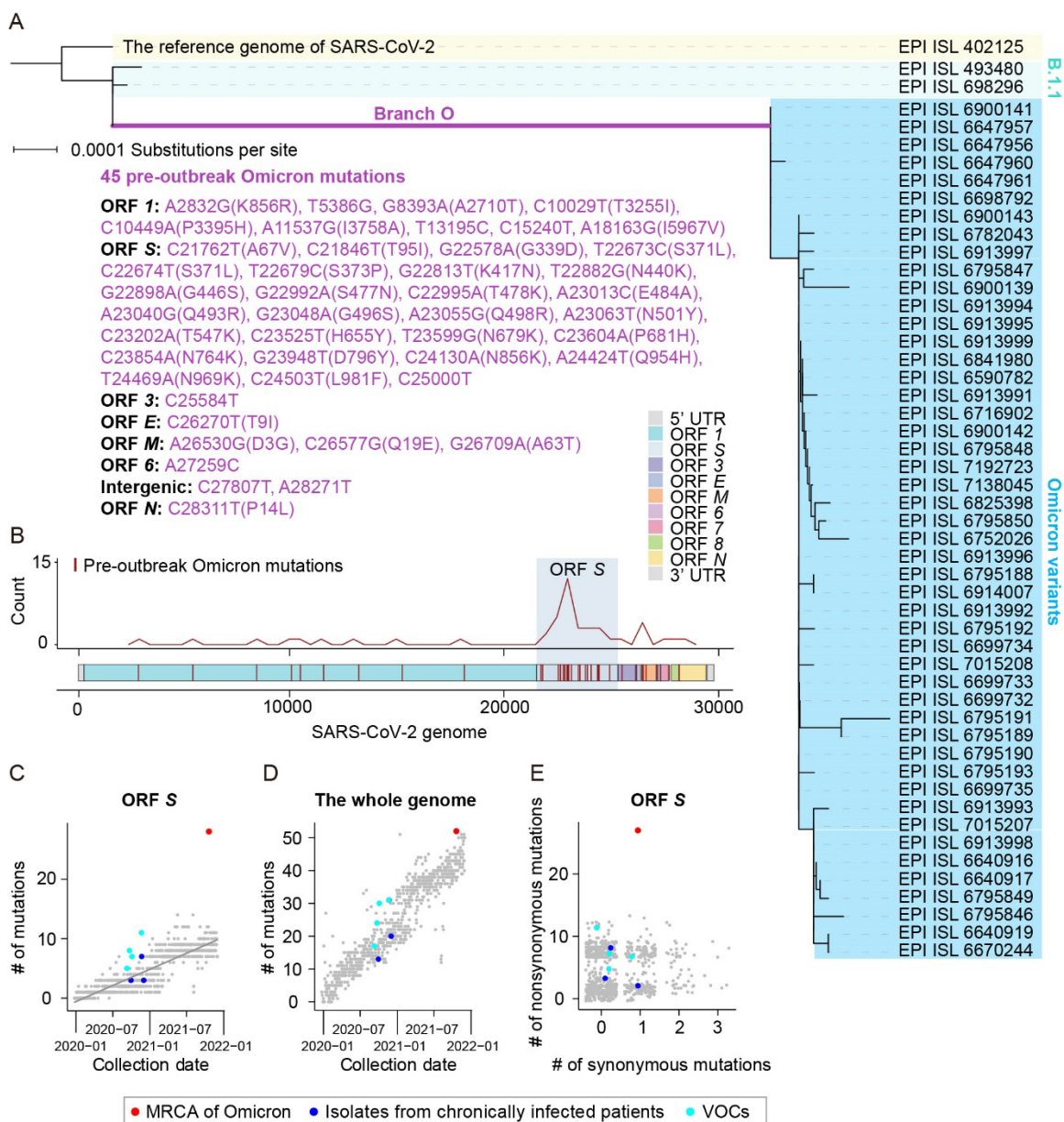
- 551 Lam, S.D., Bordin, N., Waman, V.P., Scholes, H.M., Ashford, P., Sen, N., van Dorp, L., Rauer, C.,
552 Dawson, N.L., Pang, C.S.M., *et al.*, 2020. Sars-cov-2 spike protein predicted to form
553 complexes with host receptor protein orthologues from a broad range of mammals. *Sci*
554 *Rep* 10, 16471.
- 555 Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., *et al.*, 2020.
556 Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor.
557 *Nature* 581, 215-220.
- 558 Leist, S.R., Dinno, K.H., 3rd, Schafer, A., Tse, L.V., Okuda, K., Hou, Y.J., West, A., Edwards, C.E.,
559 Sanders, W., Fritch, E.J., *et al.*, 2020. A mouse-adapted sars-cov-2 induces acute lung
560 injury and mortality in standard laboratory mice. *Cell* 183, 1070-1085 e1012.
- 561 Li, Q., Nie, J., Wu, J., Zhang, L., Ding, R., Wang, H., Zhang, Y., Li, T., Liu, S., Zhang, M., *et al.*, 2021.
562 Sars-cov-2 501y.V2 variants lack higher infectivity but do have immune escape. *Cell* 184,
563 2362-2371 e2369.
- 564 Li, Z., Wu, J., Deleo, C.J., 2006. Rna damage and surveillance under oxidative stress. *IUBMB Life*
565 58, 581-588.
- 566 Martinez-Flores, D., Zepeda-Cervantes, J., Cruz-Resendiz, A., Aguirre-Sampieri, S., Sampieri,
567 A., Vaca, L., 2021. Sars-cov-2 vaccines based on the spike glycoprotein and implications
568 of new viral variants. *Front Immunol* 12, 701501.
- 569 Meng, B., Kemp, S.A., Papa, G., Datir, R., Ferreira, I., Marelli, S., Harvey, W.T., Lytras, S.,
570 Mohamed, A., Gallo, G., *et al.*, 2021. Recurrent emergence of sars-cov-2 spike deletion
571 h69/v70 and its role in the alpha variant b.1.1.7. *Cell Rep* 35, 109292.
- 572 Montagutelli, X., Prot, M., Jouvion, G., Levillayer, L., Conquet, L., Reyes-Gomez, E., Donati, F.,
573 Albert, M., van der Werf, S., Jaubert, J., *et al.*, 2021. A mouse-adapted sars-cov-2 strain
574 replicating in standard laboratory mice. *bioRxiv*, 2021.2007.2010.451880.
- 575 Nelson, G., Buzko, O., Spilman, P., Niazi, K., Rabizadeh, S., Soon-Shiong, P., 2021. Molecular
576 dynamic simulation reveals e484k mutation enhances spike rbd-ace2 affinity and the
577 combination of e484k, k417n and n501y mutations (501y.V2 variant) induces
578 conformational change greater than n501y mutant alone, potentially resulting in an
579 escape mutant. *bioRxiv*, 2021.2001.2013.426558.
- 580 Oude Munnink, B.B., Sikkema, R.S., Nieuwenhuijse, D.F., Molenaar, R.J., Munger, E.,
581 Molenkamp, R., van der Spek, A., Tolsma, P., Rietveld, A., Brouwer, M., *et al.*, 2021.
582 Transmission of sars-cov-2 on mink farms between humans and mink and back to
583 humans. *Science* 371, 172-177.
- 584 Panchin, A.Y., Panchin, Y.V., 2020. Excessive g-u transversions in novel allele variants in sars-cov-
585 2 genomes. *PeerJ* 8, e9648.
- 586 Ren, W., Zhu, Y., Wang, Y., Shi, H., Yu, Y., Hu, G., Feng, F., Zhao, X., Lan, J., Wu, J., *et al.*, 2021.
587 Comparative analysis reveals the species-specific genetic determinants of ace2 required
588 for sars-cov-2 entry. *PLoS Pathog* 17, e1009392.
- 589 Rodrigues, J., Barrera-Vilarmau, S., J, M.C.T., Sorokina, M., Seckel, E., Kastiris, P.L., Levitt, M.,
590 2020. Insights on cross-species transmission of sars-cov-2 from structural modeling.
591 *PLoS Comput Biol* 16, e1008449.
- 592 Shan, K.J., Wei, C., Wang, Y., Huan, Q., Qian, W., 2021. Host-specific asymmetric accumulation of
593 mutation types reveals that the origin of sars-cov-2 is consistent with a natural process.
594 *The Innovation* 2, 100159.
- 595 Shu, Y., McCauley, J., 2017. Gisaid: Global initiative on sharing all influenza data - from vision to
596 reality. *Euro Surveill* 22, 30494.

- 597 Smyth, D.S., Trujillo, M., Gregory, D.A., Cheung, K., Gao, A., Graham, M., Guan, Y.,
598 Guldenpfennig, C., Hoxie, I., Kannyo, S., *et al.*, 2021. Tracking cryptic sars-cov-2 lineages
599 detected in nyc wastewater. medRxiv, 2021.2007.2026.21261142.
- 600 Sun, S., Gu, H., Cao, L., Chen, Q., Ye, Q., Yang, G., Li, R.T., Fan, H., Deng, Y.Q., Song, X., *et al.*,
601 2021. Characterization and structural basis of a lethal mouse-adapted sars-cov-2. Nat
602 Commun 12, 5654.
- 603 Truong, T.T., Ryutov, A., Pandey, U., Yee, R., Goldberg, L., Bhojwani, D., Aguayo-Hiraldo, P.,
604 Pinsky, B.A., Pekosz, A., Shen, L., *et al.*, 2021. Increased viral variants in children and
605 young adults with impaired humoral immunity and persistent sars-cov-2 infection: A
606 consecutive case series. EBioMedicine 67, 103355.
- 607 van Zundert, G.C.P., Rodrigues, J., Trellet, M., Schmitz, C., Kastiris, P.L., Karaca, E., Melquiond,
608 A.S.J., van Dijk, M., de Vries, S.J., Bonvin, A., 2016. The haddock2.2 web server: User-
609 friendly integrative modeling of biomolecular complexes. J Mol Biol 428, 720-725.
- 610 Venkatakrisnan, A., Anand, P., Lenehan, P., Suratekar, R., Raghunathan, B., Niesen,
611 M.J., Soundararajan, V., 2021. Omicron variant of sars-cov-2 harbors a unique insertion
612 mutation of putative viral or human genomic origin. OSF Preprints.
- 613 Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de
614 Beer, T.A.P., Rempfer, C., Bordoli, L., *et al.*, 2018. Swiss-model: Homology modelling of
615 protein structures and complexes. Nucleic Acids Res 46, W296-W303.
- 616 Wei, C., Chen, Y.M., Chen, Y., Qian, W., 2021. The missing expression level-evolutionary rate
617 anticorrelation in viruses does not support protein function as a main constraint on
618 sequence evolution. Genome Biol Evol 13, evab049.
- 619 Wong, L.Y.R., Zheng, J., Wilhelmsen, K., Li, K., Ortiz, M.E., Schnicker, N.J., Pezzulo, A.A.,
620 Szachowicz, P.J., Klumpp, K., Aswad, F., *et al.*, 2021. Eicosanoid signaling as a therapeutic
621 target in middle-aged mice with severe covid-19. bioRxiv.
- 622 Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., *et*
623 *al.*, 2020a. A new coronavirus associated with human respiratory disease in china.
624 Nature 579, 265-269.
- 625 Wu, S., Zhong, G., Zhang, J., Shuai, L., Zhang, Z., Wen, Z., Wang, B., Zhao, Z., Song, X., Chen, Y., *et*
626 *al.*, 2020b. A single dose of an adenovirus-vectored vaccine provides protection against
627 sars-cov-2 challenge. Nat Commun 11, 4081.
- 628 Zhang, Y., Huang, K., Wang, T., Deng, F., Gong, W., Hui, X., Zhao, Y., He, X., Li, C., Zhang, Q., *et al.*,
629 2021. Sars-cov-2 rapidly adapts in aged balb/c mice and induces typical pneumonia. J
630 Virol.
- 631 Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L.,
632 *et al.*, 2020. A pneumonia outbreak associated with a new coronavirus of probable bat
633 origin. Nature 579, 270-273.

634

635

636 **FIGURES**



637

638 **Fig. 1. The characterization of pre-outbreak Omicron mutations.**

639 **A.** The phylogenetic tree of Omicron variants, including the reference genome of SARS-
 640 CoV-2 (EPI_ISL_402125), two B.1.1 variants, and 48 Omicron variants. A total of 45
 641 pre-outbreak Omicron point mutations in the long branch leading to the MRCA of
 642 Omicron (Branch O, labeled in purple) in the phylogenetic tree are grouped according to
 643 ORFs.

644 **B.** The distribution of pre-outbreak Omicron mutations across the SARS-CoV-2 genome.

645 The curve indicates the density of mutations. UTR stands for the untranslated region.

646 **C.** Number of mutations that accumulated in ORF *S* of the MRCA of current Omicron

647 variants (red), the other four VOCs (*i.e.*, Alpha, Beta, Gamma, and Delta; cyan), and

648 three SARS-CoV-2 isolates from chronically infected patients (blue), against the date of

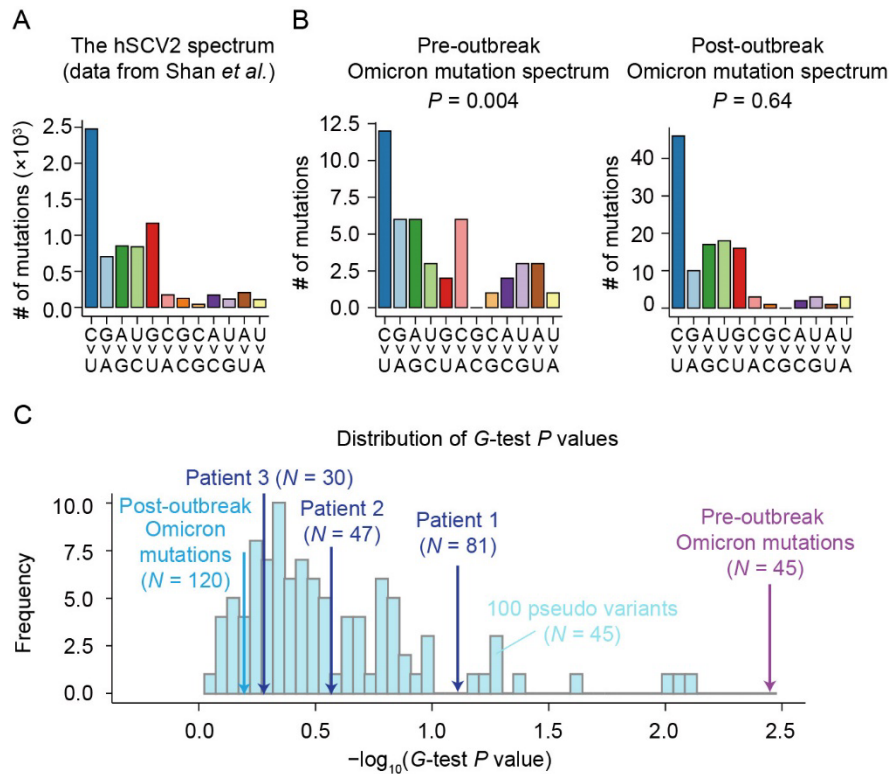
649 sample collection. SARS-CoV-2 variants randomly sampled (one variant per day) are

650 shown in grey, and the grey line represents their linear regression.

651 **D.** Similar to (C), for the whole genome.

652 **E.** A scatterplot shows the numbers of synonymous and nonsynonymous mutations in

653 ORF *S* (jittered in order to reduce overplotting).



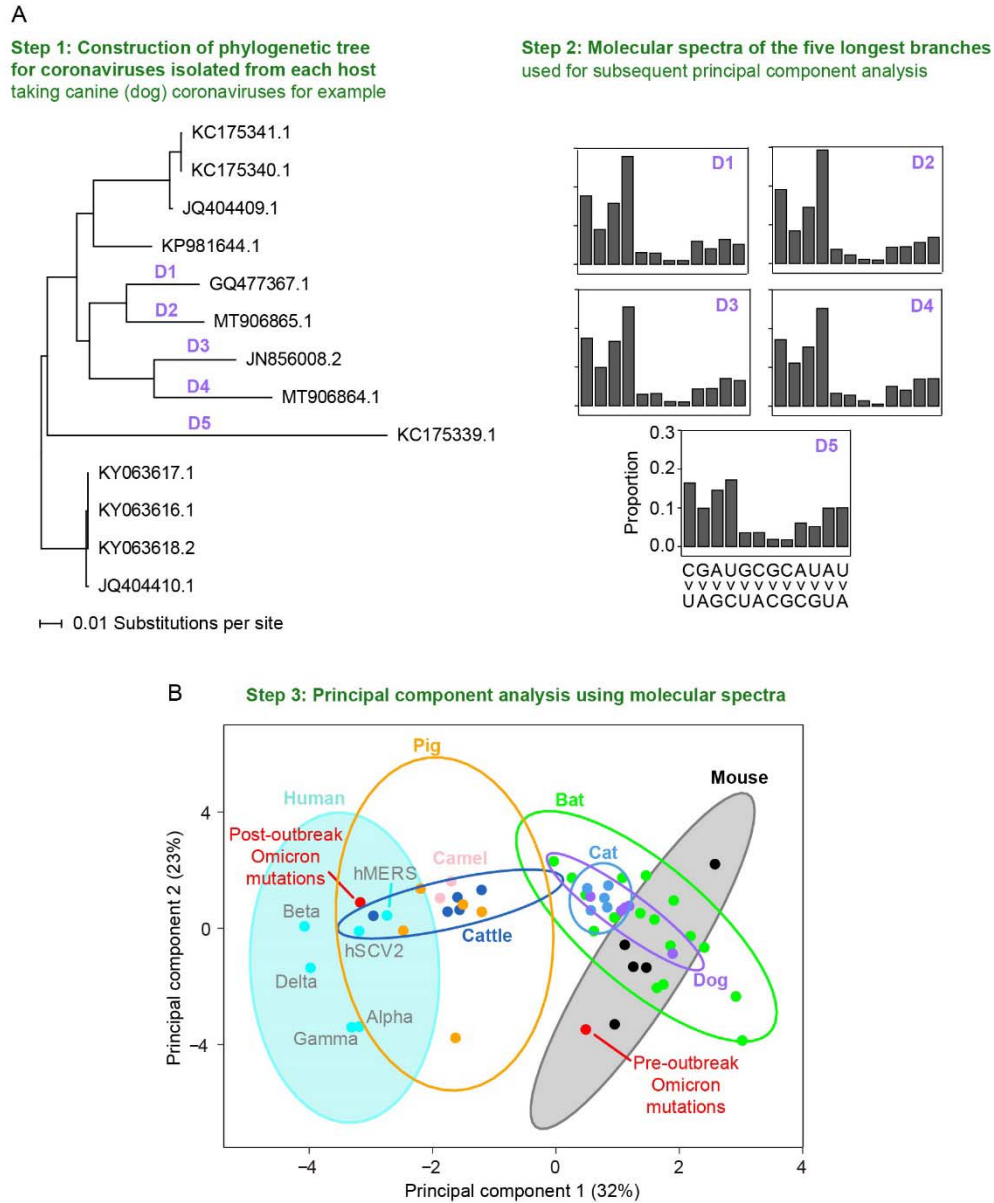
654

655 **Fig. 2. Comparison of the molecular spectrum of pre-outbreak Omicron mutations**
656 **and spectra of mutations known to accumulate in humans.**

657 **A.** The molecular spectrum of viral mutations that accumulated in humans (the hSCV2
658 spectrum).

659 **B.** The molecular spectra of pre- and post-outbreak Omicron mutations. P values were
660 given by G -test to test whether a molecular spectrum was significantly different from the
661 hSCV2 spectrum.

662 **C.** The distribution of P values (given by G -test) of 100 pseudo samples that were down
663 sampled from the hSCV2 spectrum. The number of mutations (N) of each pseudo sample
664 was equal to 45. The molecular spectra of SARS-CoV-2 isolates from three chronically
665 infected patients were also labeled. SARS-CoV-2 data of patient 1 were retrieved from
666 Kemp *et al.* (2021) and those of patients 2 and 3 were retrieved from Truong *et al.*
667 (2021).

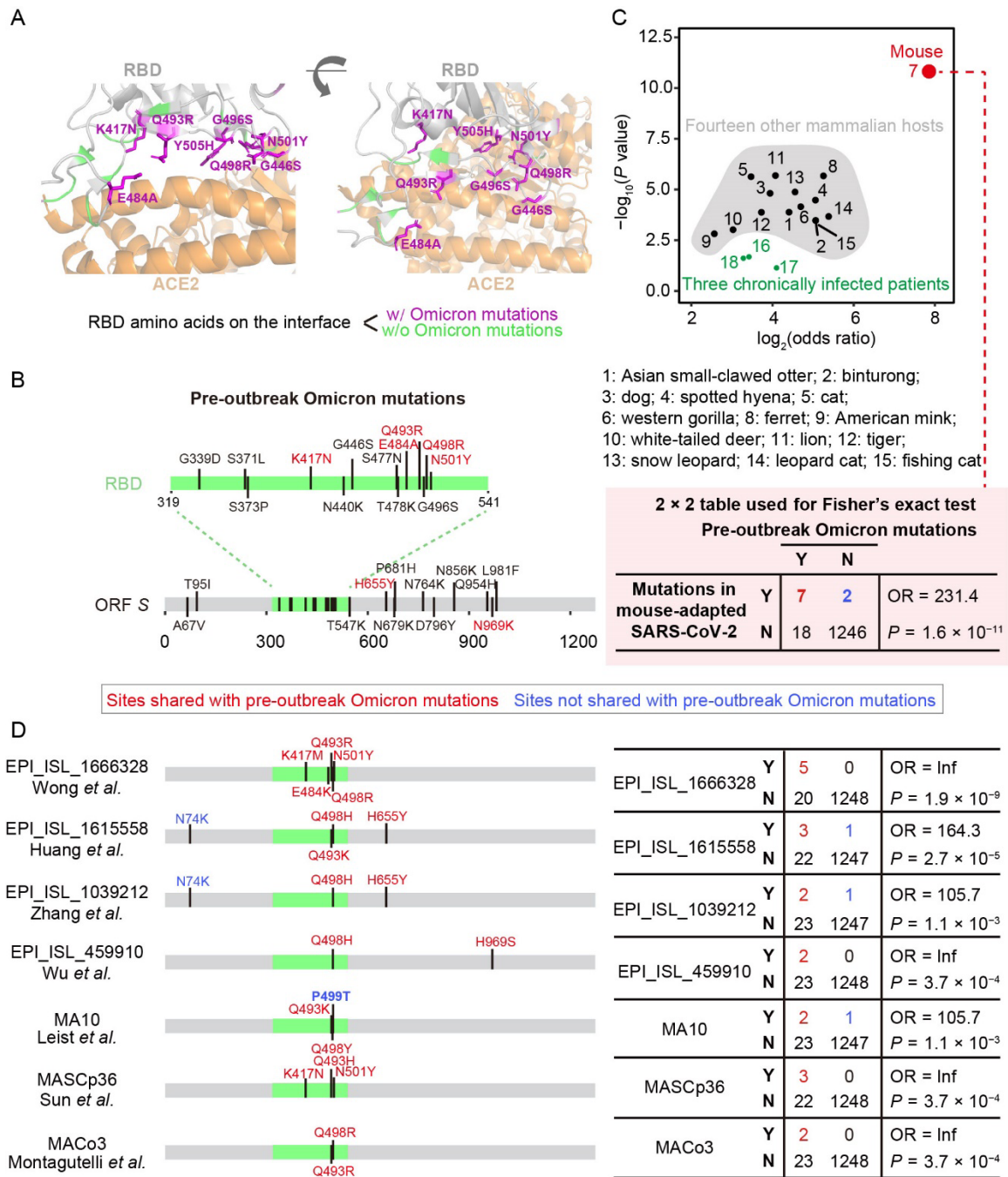


668

669 **Fig. 3. The similarity in molecular spectra between Omicron and coronaviruses**
 670 **isolated from various mammalian species.**

671 **A.** A schematic shows the workflow for analyzing the similarity in molecular spectra
 672 across various hosts, taking variants in dogs as an example.

673 **B.** The principal component analysis plot depicts the molecular spectra of virus mutations
 674 that accumulated in humans and various host species. Dots were colored according to the
 675 corresponding host species. The 95% confidence ellipses are shown for each host species.



676

677 **Fig. 4. The similarity in the spike protein sequence between Omicron and SARS-**
678 **CoV-2 variants isolated from various hosts.**

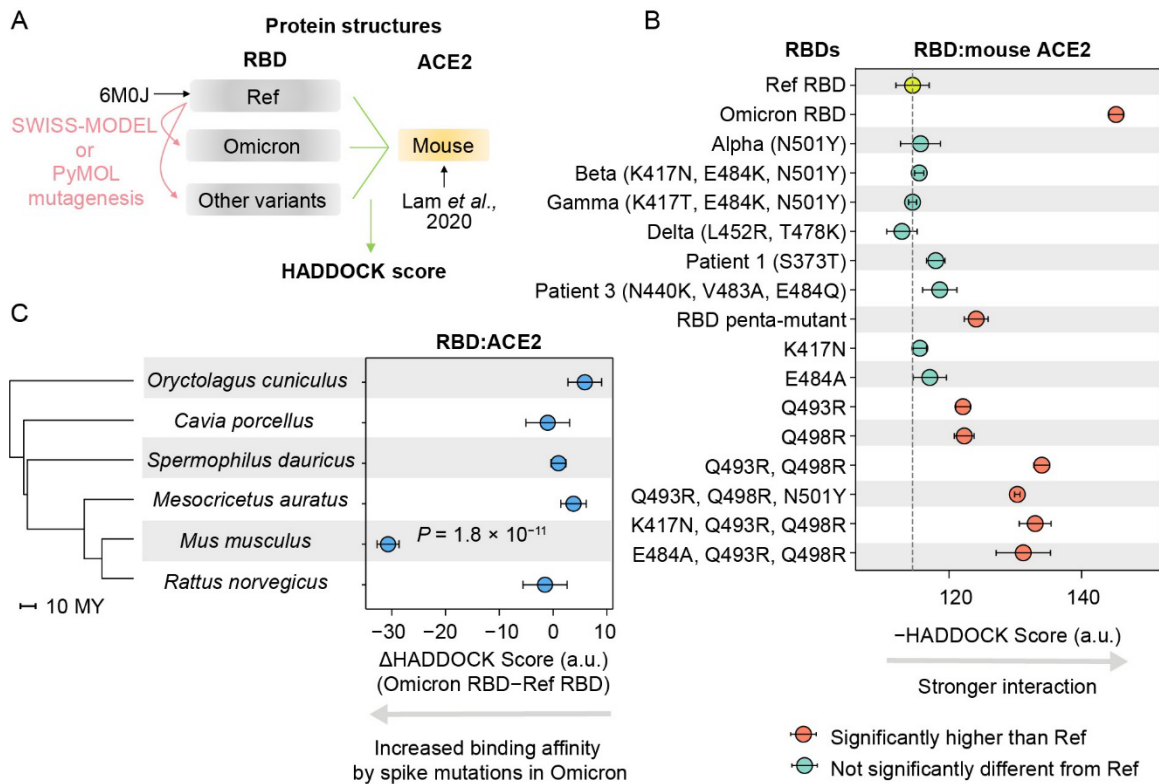
679 **A.** The structure of the interface between SARS-CoV-2 spike protein and human ACE2,
680 from the crystal structure of the spike:ACE2 (human) complex (PDB: 6M0J). RBD
681 residues on the interface (defined within 5 Å distance) were colored.

682 **B.** The amino acid mutations in the spike protein in the MRCA of Omicron variants.

683 **C.** The statistical assessment on the overlapping in mutated positions between Omicron
684 and SARS-CoV-2 variants using Fisher's exact test. The 2×2 contingency table for mice
685 is shown. OR stands for the odds ratio.

686 **D.** Comparison between pre-outbreak Omicron mutations and mutations detected in
687 seven SARS-COV-2 variants isolated from mice, in the spike protein. RBD of spike
688 protein was colored green.

689



690

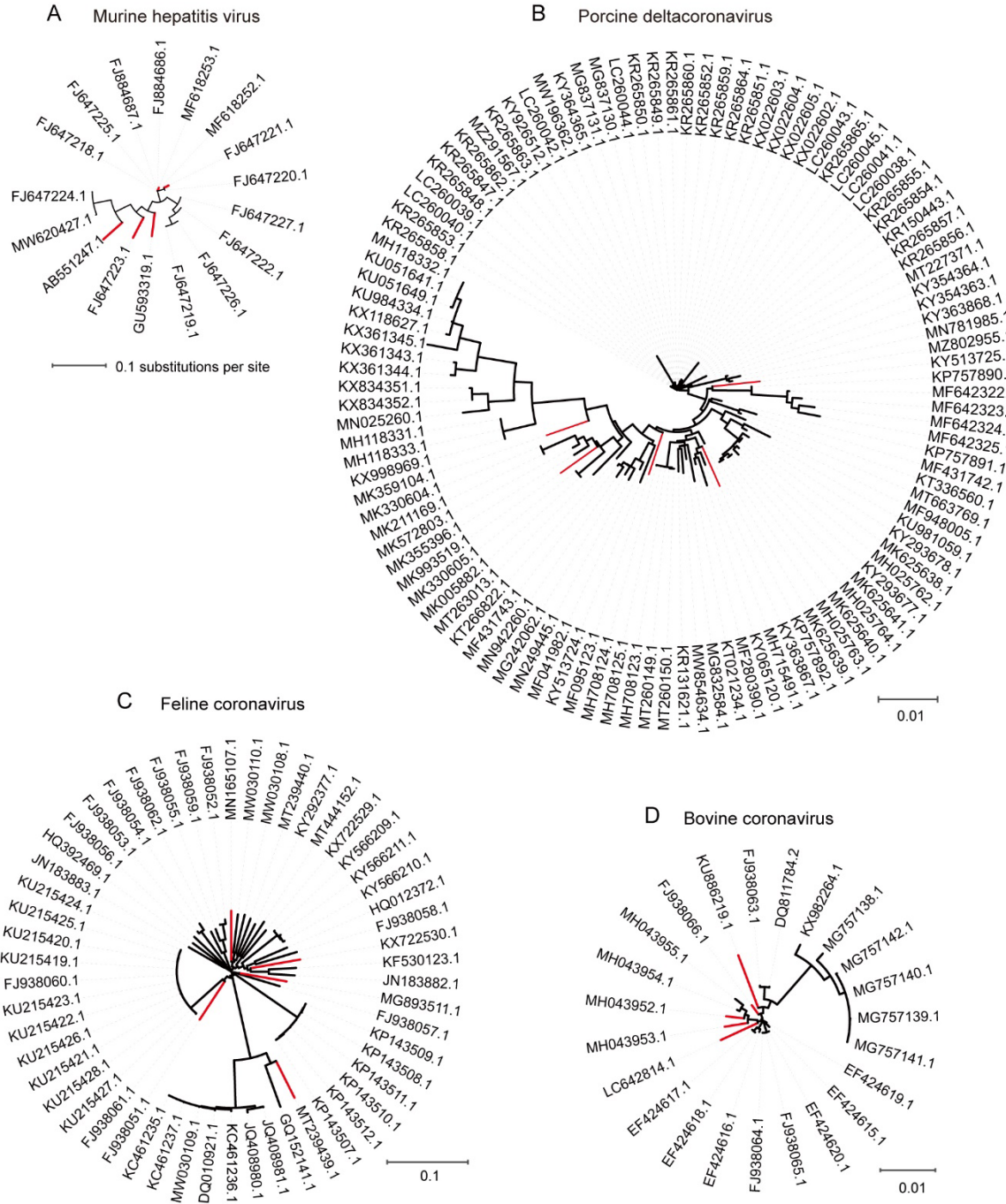
691 **Fig. 5. Predicted binding affinities between RBD variants and ACE2.**

692 **A.** A schematic shows the workflow to estimate the HADDOCK scores between RBD
693 variants and mouse ACE2.

694 **B.** The HADDOCK scores for various RBD variants and mouse ACE2. The error bars
695 represent standard errors. Penta-mutant of RBD harbored five mutations (K417N, E484A,
696 Q493R, Q498R, and N501Y). Patient 2 who did not harbor any amino acid mutations in
697 RBD was not shown.

698 **C.** The Δ HADDOCK scores for five rodent species and the European rabbit. *P* values
699 were given by two-tailed *t*-tests. Only *P* values <0.05 are labeled. The phylogenetic tree
700 was constructed using TIMETREE, in the unit of million years (MY).

701



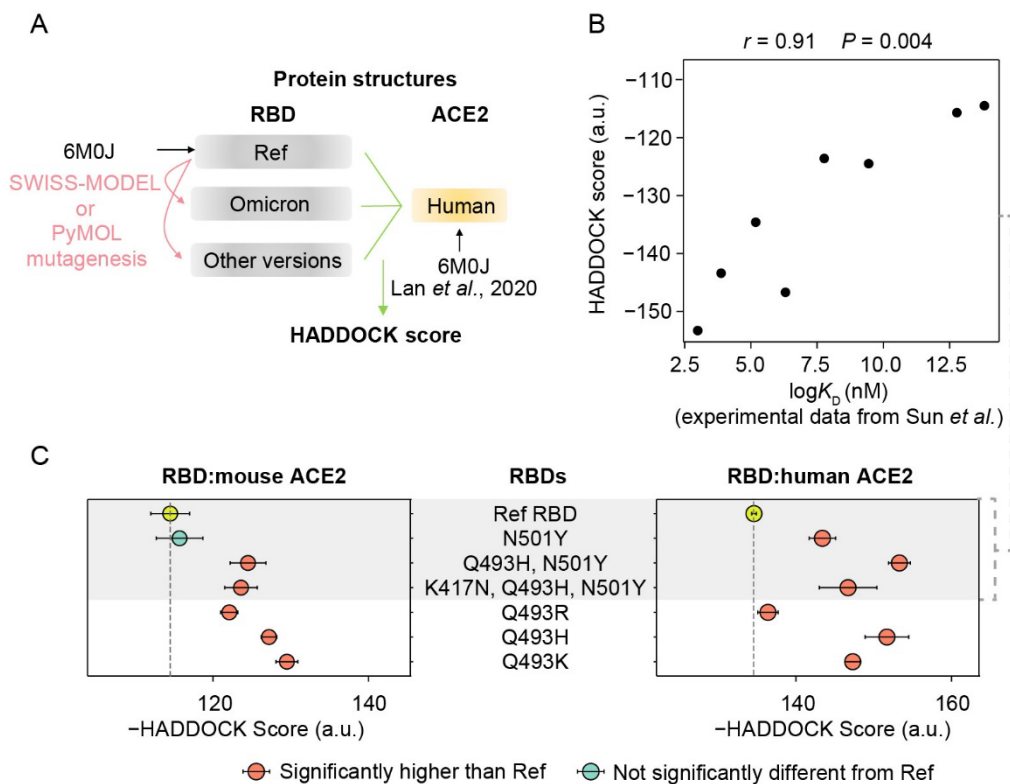
702

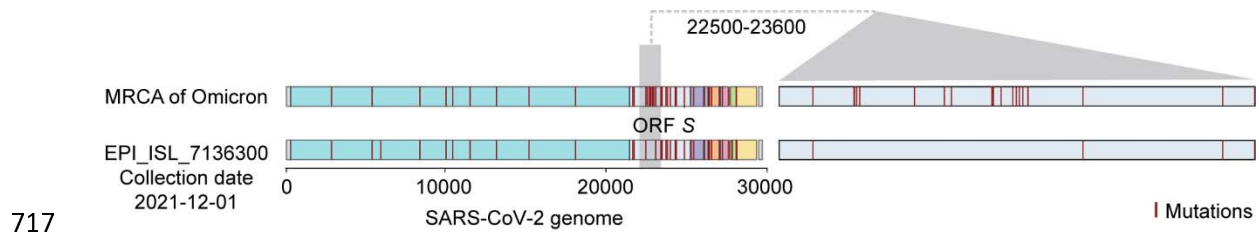
703 **Fig. S1. The phylogenetic trees of four coronaviruses species.**

704 The five external branches with the largest number of accumulated mutations were

705 colored red, and were used in the principal component analysis.

706

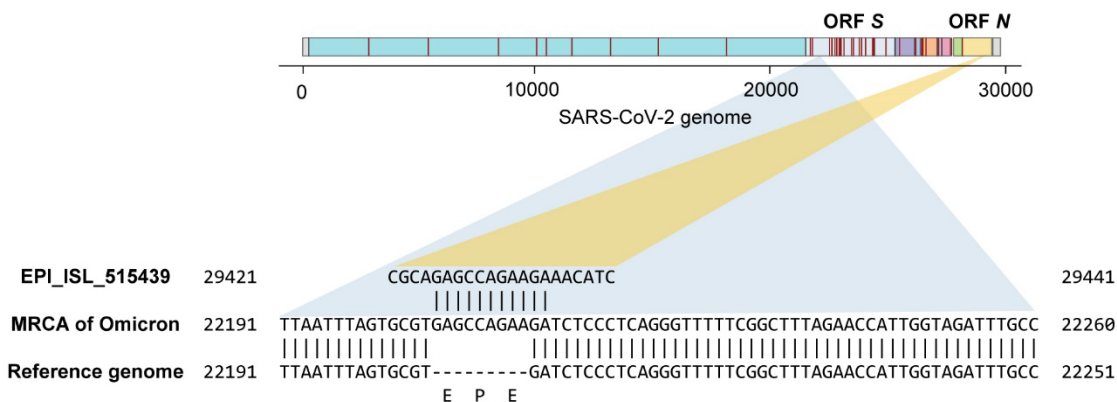




718 **Fig. S3. The distribution of mutations in a potential intermediate of the current**
719 **Omicron variants.**

720 Red lines represent the positions of point mutations (relative to the B.1.1 lineage). The
721 genomic regions were colored according to ORFs as in **Fig. 1B**.

722



724 **Fig. S4. A potential explanation for the “EPE insertion” in the spike protein of**
 725 **Omicron**

726 Red lines represent the positions of point mutations (relative to the B.1.1 lineage). The
 727 genomic regions were colored according to ORFs as in **Fig. 1B**. The “EPE insertion”
 728 related regions were shown at the nucleotide resolution. The SARS-CoV-2 variant
 729 (EPI_ISL_515439) collected by the Nevada State Public Health Laboratory on May 19th,
 730 2020 was used as an example to illustrate a potential source sequence of the insertion.