

1 A novel bat coronavirus reveals natural insertions at the S1/S2  
2 cleavage site of the Spike protein and a possible recombinant  
3 origin of HCoV-19

4  
5 Hong Zhou<sup>1,8</sup>, Xing Chen<sup>2,8</sup>, Tao Hu<sup>1,8</sup>, Juan Li<sup>1,8</sup>, Hao Song<sup>3</sup>, Yanran Liu<sup>1</sup>, Peihan Wang<sup>1</sup>, Di  
6 Liu<sup>4</sup>, Jing Yang<sup>5</sup>, Edward C. Holmes<sup>6</sup>, Alice C. Hughes<sup>2,\*</sup>, Yuhai Bi<sup>5,\*</sup>, Weifeng Shi<sup>1,7,\*</sup>

7  
8 <sup>1</sup>Key Laboratory of Etiology and Epidemiology of Emerging Infectious Diseases in Universities  
9 of Shandong, Shandong First Medical University & Shandong Academy of Medical Sciences,  
10 Taian 271000, China

11 <sup>2</sup>Landscape Ecology Group, Center for Integrative Conservation, Xishuangbanna Tropical  
12 Botanical Garden, Chinese Academy of Sciences, Menglun, Mengla, Yunnan 666303, China

13 <sup>3</sup>Research Network of Immunity and Health (RNIH), Beijing Institutes of Life Science, Chinese  
14 Academy of Sciences, Beijing 100101, China

15 <sup>4</sup>Computational Virology Group, Center for Bacteria and Virus Resources and Bioinformatics,  
16 Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan 430071, China

17 <sup>5</sup>CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology,  
18 CAS Center for Influenza Research and Early-warning (CASCIRE), CAS-TWAS Center of  
19 Excellence for Emerging Infectious Diseases (CEEID), Chinese Academy of Sciences, Beijing  
20 100101, China

21 <sup>6</sup>Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and  
22 Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney,  
23 New South Wales 2006, Australia

24 <sup>7</sup>The First Affiliated Hospital of Shandong First Medical University (Shandong Provincial  
25 Qianfoshan Hospital), Ji'nan 250014, China

26 <sup>8</sup>These authors contributed equally

27 \*Correspondence: shiwf@ioz.ac.cn (W.S.), beeyh@im.ac.cn (Y.B.),  
28 ach\_conservation2@hotmail.com (A.C.H.)

## 29 **Summary**

30 The unprecedented epidemic of pneumonia caused by a novel coronavirus, HCoV-19, in  
31 China and beyond has caused public health concern at a global scale. Although bats are  
32 regarded as the most likely natural hosts for HCoV-19<sup>1,2</sup>, the origins of the virus remain unclear.  
33 Here, we report a novel bat-derived coronavirus, denoted RmYN02, identified from a  
34 metagenomics analysis of samples from 227 bats collected from Yunnan Province in China  
35 between May and October, 2019. RmYN02 shared 93.3% nucleotide identity with HCoV-19 at  
36 the scale of the complete virus genome and 97.2% identity in the 1ab gene in which it was the  
37 closest relative of HCoV-19. In contrast, RmYN02 showed low sequence identity (61.3%) to  
38 HCoV-19 in the receptor binding domain (RBD) and might not bind to angiotensin-  
39 converting enzyme 2 (ACE2). Critically, however, and in a similar manner to HCoV-19,  
40 RmYN02 was characterized by the insertion of multiple amino acids at the junction site of the

41 S1 and S2 subunits of the Spike (S) protein. This provides strong evidence that such insertion  
42 events can occur in nature. Together, these data suggest that HCoV-19 originated from  
43 multiple naturally occurring recombination events among those viruses present in bats and  
44 other wildlife species.

45

## 46 **Text**

47 Coronaviruses (CoVs) are common viral respiratory pathogens that primarily cause symptoms  
48 in the upper respiratory and gastrointestinal tracts. In 1960s, two CoVs, 229E and OC43, were  
49 identified in clinical samples from patients experiencing the common cold<sup>3</sup>. More recently,  
50 four additional human CoVs have been successively identified: severe acute respiratory  
51 syndrome coronavirus (SARS-CoV) in 2002, NL63 in late 2004, HKU1 in January 2005, and  
52 Middle East respiratory syndrome coronavirus (MERS-CoV) in 2012. However, only two  
53 betacoronaviruses (beta-CoVs), SARS-CoV and MERS-CoV, are able to cause severe and fatal  
54 infections, leading to 774 and 858 deaths, respectively, suggesting that beta-CoVs may be of  
55 particular concern to human health. In December 2019, viral pneumonia caused by an  
56 unidentified microbial agent was reported, which was soon identified to be a novel  
57 coronavirus<sup>4</sup>, now termed SARS-CoV-2 by the International Committee for the Taxonomy of  
58 Viruses<sup>5</sup> and HCoV-19 by a group of Chinese scientists<sup>6</sup>. The number of patients infected with  
59 HCoV-19 has increased sharply since January 21, 2020, and as of March 2rd, 2020, more than  
60 80,000 confirmed HCoV-19 cases have been reported, with >11,000 severe cases and >2900  
61 deaths in China. By the end of January confirmed HCoV-19 cases were present in all the

62 Chinese provinces and municipalities and at the time of writing the virus has been detected  
63 in over 60 countries.

64

65 An epidemiological survey of several HCoV-19 cases at an early stage of the outbreak  
66 revealed that most had visited the Huanan seafood market in Wuhan city prior to illness,  
67 where various wild animals were on sale before it was closed on January 1, 2020 due to the  
68 outbreak. Phylogenetic analysis has revealed that HCoV-19 is a novel beta-CoV distinct from  
69 SARS-CoV and MERS-CoV<sup>1,2,4</sup>. To date, the most closely related virus to HCoV-2019 is RaTG13,  
70 identified from a *Rhinolophus affinis* bat sampled in Yunnan province in 2013<sup>2</sup>. This virus  
71 shared 96.1% nucleotide identity and 92.9% identity in the S gene, again suggesting that bats  
72 play a key role as coronavirus reservoirs<sup>2</sup>. Notably, however, two research groups recently  
73 reported several novel beta-CoVs related to HCoV-19 in Malayan pangolins (*Manis javanica*)  
74 that were illegally imported into Guangxi (GX) and Guangdong (GD) provinces, southern  
75 China<sup>7,8</sup>. Although these pangolins CoVs are more distant to HCoV-19 than RaTG13 across  
76 the virus genome as a whole, they are very similar to HCoV-19 in the receptor binding domain  
77 (RBD) of the S protein, including at the amino acid residues thought to mediate binding to  
78 ACE2<sup>8</sup>. It is therefore possible that pangolins play an important role in the ecology and  
79 evolution of CoVs, although whether they act as intermediate hosts for HCoV-19 is currently  
80 unclear. Indeed, the discovery of viruses in pangolins suggests that there is a wide diversity  
81 of CoVs still to be sampled in wildlife, some of which may be directly involved in the  
82 emergence of HCoV-19.

83

84 Between May and October, 2019, we collected a total of 302 samples from 227 bats from  
85 Mengla County, Yunnan Province in southern China (Extended Data Table 1). These bats  
86 belonged to 20 different species, with the majority of samples from *Rhinolophus malayanus*  
87 (n=48, 21.1%), *Hipposideros larvatus* (n=41, 18.1%) and *Rhinolophus steno* (n=39, 17.2%). The  
88 samples comprised multiple tissues, including patagium (n=219), lung (n=2) and liver (n=3),  
89 and feces (n=78). All but three bats were sampled alive and subsequently released. Based on  
90 the bat species primarily identified according to morphological criteria and confirmed  
91 through DNA barcoding, the 224 tissues and 78 feces were merged into 38 and 18 pools,  
92 respectively, with each pool including 1 to 11 samples of the same type (Extended Data Table  
93 1). These pooled samples were then used for next generation sequencing (NGS).

94  
95 Using next-generation metagenomic sequencing we successfully obtained 11954 and 64224  
96 reads in pool No. 39 (from a total of 78,477,464 clean reads) that mapped to a SARS-like bat  
97 coronavirus, Cp/Yunnan2011<sup>9</sup> (JX993988), and to HCoV-19. From this, we generated two  
98 preliminary consensus sequences. Pool 39 comprised 11 feces from *Rhinolophus malayanus*  
99 collected between May 6 and July 30, 2019. After a series of verification steps, including re-  
100 mapping and Sanger sequencing (Extended Data Table 2 and Figures 1-3), one partial (23395  
101 bp) and one complete (29671 bp) beta-CoV genome sequences were obtained and termed  
102 BetaCoV/Rm/Yunnan/YN01/2019 (RmYN01) and BetaCoV/Rm/Yunnan/YN02/2019  
103 (RmYN02), respectively. Notably, 20 positions in the RmYN02 genome displayed nucleotide  
104 polymorphisms in the NGS data, although these did not include the S1/S2 cleavage site  
105 (Extended Data Figure 3). Only a few reads in the remaining 55 pools could be mapped to

106 the reference CoV genomes. The sequence identity between RmYN01 and Cp/Yunnan2011  
107 across the aligned regions was 96.9%, whereas that between RmYN01 and HCoV-19 was only  
108 79.7% across the aligned regions and 70.4% in the spike gene.

109

110 In contrast, RmYN02 was closely related to HCoV-19, exhibiting 93.3% nucleotide sequence  
111 identity, although it was less similar to HCoV-19 than RaTG13 (96.1%) across the genome as  
112 a whole (Fig. 1a). RmYN02 and HCoV-19 were extremely similar (>96% sequence identity) in  
113 most genomic regions (e.g. 1ab, 3a, E, 6, 7a, N and 10) (Fig. 1a). In particular, RmYN02 was  
114 97.2% identical to HCoV-19 in the longest encoding gene region, 1ab (n=21285). However,  
115 RmYN02 exhibited far lower sequence identity to HCoV-19 in the S gene (nucleotide 71.8%,  
116 amino acid 72.9%), compared to 97.4% amino acid identity between RaTG13 and HCoV-19  
117 (Fig. 1a). Strikingly, RmYN02 only possessed 62.4% amino acid identity to HCoV-19 in the RBD,  
118 whereas the pangolin beta-CoV from Guangdong had amino acid identity of 97.4%<sup>7</sup>, and was  
119 the closest relative of HCoV-19 in this region. A similarity plot estimated using Simplot<sup>10</sup> also  
120 revealed that RmYN02 was more similar to HCoV-19 than RaTG13 in most genome regions  
121 (Fig. 1b). Again, in the RBD, the pangolin/MP789/2019 virus shared the highest sequence  
122 identity to HCoV-19 (Fig. 1c).

123

124 Results from both homology modelling<sup>1</sup>, *in vitro* assays<sup>2</sup> and resolved three-dimensional  
125 structure of the S protein<sup>11</sup> have revealed that like SARS-CoV, HCoV-19 could also use ACE2  
126 as a cell receptor. We analyzed the RBD of RmYN02, RaTG13, and the two pangolin beta-  
127 CoVs using homology modelling (Fig. 2a-2f and Extended Data Figure 4 for sequence

128 alignment). The amino acid deletions in RmYN02 RBD made two loops near the receptor  
129 binding site that are shorter than those in HCoV-19 RBD (Fig. 2a and 2f). Importantly, the  
130 conserved disulfide bond in the external subdomain of SARS-CoV (PDB: 2DD8)<sup>12</sup>, HCoV-19  
131 (PDB: 6LZG), RaTG13 (Fig. 2b), pangolin/MP789/2019 (Fig. 2c) and pangolin/GX/P5L/2017  
132 (Fig. 2d) was missing in RmYN02 (Fig. 2f). We speculate that these deletions may cause  
133 conformational variations and consequently reduce the binding of RmYN02 RBD with ACE2  
134 or even cause non-binding. It is possible that the bat SARS-related CoVs with loop deletions,  
135 including RmYN02, ZXC21 and ZC45, use a currently unknown receptor. In contrast, RaTG13  
136 (Fig. 2b), pangolin/MP789/2019 (Fig. 2c) and pangolin/P5L/2017 (Fig. 2d) did not have the  
137 deletions, and had similar conformations at their external domains, indicating that they may  
138 also use ACE2 as cell receptor although, with the exception of pangolin/MP789/2019 (see  
139 below), all exhibited amino acid variation to HCoV-19. Indeed, the pangolin/MP789/2019  
140 virus showed highly structural homology with HCoV-19 (Fig. 2e).

141

142 Six amino acid residues at the RBD (L455, F486, Q493, S494, N501 and Y505) have been  
143 reported to be major determinants of efficient receptor binding of HCoV-19 to ACE2<sup>13</sup>. As  
144 noted above, and consistent with the homology modelling, pangolin/MP789/2019 possessed  
145 the identical amino acid residues to HCoV-19 at all six positions<sup>7</sup>. In contrast, both RaTG13,  
146 RmYN02 and RmYN01 possessed the same amino acid residue as HCoV-19 at only one of  
147 the six positions each (RaTG13, L455; RmYN02, Y505; RmYN01, Y505) (Fig. 2g), despite  
148 RaTG13 being the closest relative in the spike protein. Such an evolutionary pattern is  
149 indicative of a complex combination of recombination and natural selection<sup>7,14</sup>.

150

151 The S protein of CoVs is functionally cleaved into two subunits, S1 and S2<sup>15</sup> in a similar manner  
152 to the haemagglutinin (HA) protein of avian influenza viruses (AIVs). The insertion of polybasic  
153 amino acids at the cleavage site in the HAs of some AIV subtypes is associated with enhanced  
154 pathogenicity<sup>16,17</sup>. Notably, HCoV-19 is characterized by a four-amino-acid-insertion at the  
155 junction of S1 and S2, not observed in other lineage B beta-CoVs<sup>18</sup>. This insertion, which  
156 represents a poly-basic (furin) cleavage site, is unique to HCoV-19 and is present in all HCoV-  
157 19 sequenced so far. The insertion of three residues, PAA, at the junction of S1 and S2 in  
158 RmYN02 (Fig. 2h and Extended Data Figure 2) is therefore of major importance. Although the  
159 inserted residues (and hence nucleotides) are not the same as those in RmYN02, and hence  
160 are indicative of an independent insertion event, that they are presented in wildlife (bats)  
161 strongly suggests that they are of natural origin and have likely acquired by recombination.  
162 As such, these data are strongly suggestive of a natural zoonotic origin of HCoV-19.

163

164 We next performed a phylogenetic analysis of RmYN02, RaTG13, HCoV-19 and the pangolin  
165 beta-CoVs. Consistent with a previous research<sup>7</sup>, the pangolin beta-CoVs formed two well-  
166 supported sub-lineages, representing animal seized by anti-smuggling authorities in Guangxi  
167 (Pangolin-CoV/GX) and Guangdong (Pangolin-CoV/GD) provinces (Fig. 3a and Extended  
168 Data Figure 5). However, whether pangolins are natural reservoirs for these viruses, or they  
169 acquired these viruses independently from bats or other wildlife, requires further sampling<sup>7</sup>.  
170 More notable was that RmYN02 was the closest relative of HCoV-19 in most of the virus  
171 genome, although these two viruses were still separated from each other by a relatively long



172 branch length (Fig. 3a and Extended Data Figure 5). In the spike gene tree, HCoV-19 clustered  
173 with RaTG13 and was distant from RmYN02, suggesting that the latter virus has experienced  
174 recombination in this gene (Fig. 3b and Extended Data Figure 6). In phylogeny of the RBD,  
175 HCoV-19 was most closely related to pangolin-CoV/GD, with the bat viruses falling in more  
176 divergent positions, again indicative of recombination (Fig. 3c and Extended Data Figure 7).  
177 Finally, phylogenetic analysis of the complete RNA dependent RNA polymerase (RdRp) gene,  
178 which is often used in the phylogenetic analysis of RNA viruses, revealed that RmYN02,  
179 RaTG13 and HCoV-19 formed a well-supported sub-cluster distinct from the pangolin viruses  
180 (Fig. 3d and Extended Data Figure 8).

181

182 We confirmed the bat host of RmYN02, *Rhinolophus malayanus*, by analyzing the sequence  
183 of the cytochrome b (*Cytb*) gene from the next generation sequencing data; this revealed 100%  
184 sequence identity to a *Rhinolophus malayanus* isolate (GenBank accession MK900703). Both  
185 *Rhinolophus malayanus* and *Rhinolophus affinis* are widely distributed in southwest China  
186 and southeast Asia. Generally, they do not migrate over long distances and are highly  
187 gregarious such that they are likely to live in the same caves, which might facilitate the  
188 exchange of viruses between them and the occurrence of recombination. Notably, RaTG13  
189 was identified from anal swabs and RmYN02 was identified from feces, which is a simple, but  
190 feasible way for bats to spread the virus to other animals, especially species that can utilize  
191 cave environments.

192

193 Based on the currently available data we propose that HCoV-19 likely originates from multiple

194 naturally occurring recombination events in wildlife. A virus from bats likely provides the  
195 genetic backbone of HCoV-19, with further recombination events with bats and perhaps  
196 other wildlife species resulting in the acquisition of the Spike protein, RBD and the polybasic  
197 cleavage site. Similar recombination events have been also implicated in the origin of SARS-  
198 CoV<sup>19</sup>, although it is clear that a far wider sampling of wildlife will be required to reveal the  
199 exact species involved and the exact series of recombination events.

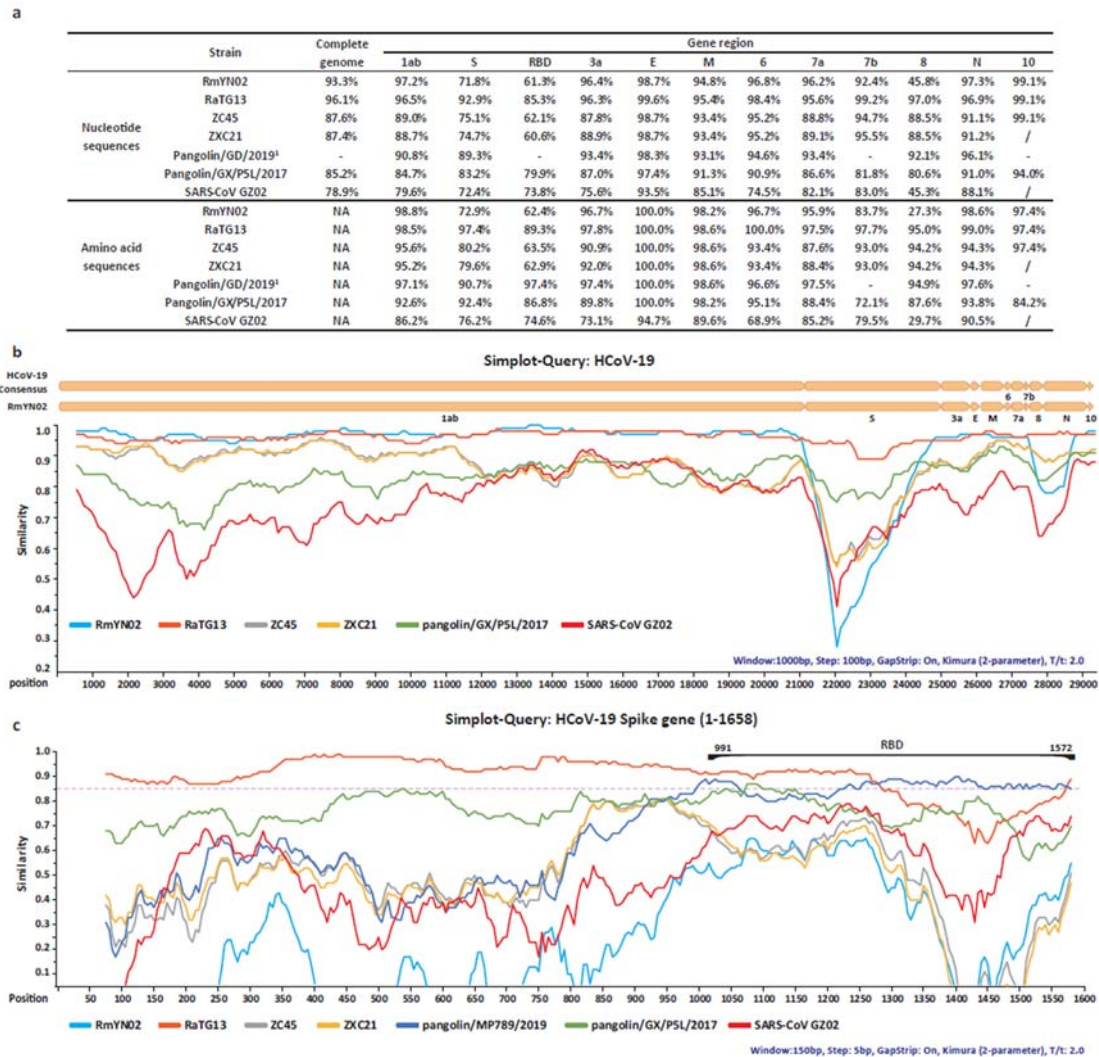
200

## 201 **References**

- 202 1 Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus:  
203 implications for virus origins and receptor binding. *Lancet* **395**, 565-574,  
204 doi:10.1016/S0140-6736(20)30251-8 (2020).
- 205 2 Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat  
206 origin. *Nature*, doi:10.1038/s41586-020-2012-7 (2020).
- 207 3 Su, S. *et al.* Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses.  
208 *Trends Microbiol* **24**, 490-502, doi:10.1016/j.tim.2016.03.003 (2016).
- 209 4 Gorbalenya, A. E. *et al.* Severe acute respiratory syndrome-related coronavirus: The  
210 species and its viruses – a statement of the Coronavirus Study Group. *bioRxiv*,  
211 2020.2002.2007.937862, doi:10.1101/2020.02.07.937862 (2020).
- 212 5 Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New*  
213 *England Journal of Medicine* **382**, doi:10.1056/NEJMoa2001017 (2020).
- 214 6 Jiang, S. *et al.* A distinct name is needed for the new coronavirus. *The Lancet*,

- 215           doi:[https://doi.org/10.1016/S0140-6736\(20\)30419-0](https://doi.org/10.1016/S0140-6736(20)30419-0) (2020).
- 216    7       Lam, T. T.-Y. *et al.* Identification of 2019-nCoV related coronaviruses in Malayan pangolins  
217       in southern China. *bioRxiv*, 2020.2002.2013.945485, doi:10.1101/2020.02.13.945485  
218       (2020).
- 219    8       Xiao, K. *et al.* Isolation and Characterization of 2019-nCoV-like Coronavirus from Malayan  
220       Pangolins. *bioRxiv*, 2020.2002.2017.951335, doi:10.1101/2020.02.17.951335 (2020).
- 221    9       Wu, Z. *et al.* Deciphering the bat virome catalog to better understand the ecological  
222       diversity of bat viruses and the bat origin of emerging infectious diseases. *The ISME*  
223       *journal* **10**, 609-620, doi:10.1038/ismej.2015.138 (2016).
- 224    10      Lole, K. S. *et al.* Full-length human immunodeficiency virus type 1 genomes from subtype  
225       C-infected seroconverters in India, with evidence of intersubtype recombination. *J Virol*  
226       **73**, 152-160 (1999).
- 227    11      Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation.  
228       *Science*, doi:10.1126/science.abb2507 (2020).
- 229    12      Prabakaran, P. *et al.* Structure of severe acute respiratory syndrome coronavirus receptor-  
230       binding domain complexed with neutralizing antibody. *The Journal of biological chemistry*  
231       **281**, 15829-15836, doi:10.1074/jbc.M600697200 (2006).
- 232    13      Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. Receptor recognition by novel  
233       coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. *J*  
234       *Virology*, doi:10.1128/JVI.00127-20 (2020).
- 235    14      Wong, M. C., Javornik Cregeen, S. J., Ajami, N. J. & Petrosino, J. F. Evidence of  
236       recombination in coronaviruses implicating pangolin origins of nCoV-2019. *bioRxiv*,

- 237 2020.2002.2007.939207, doi:10.1101/2020.02.07.939207 (2020).
- 238 15 He, Y. *et al.* Receptor-binding domain of SARS-CoV spike protein induces highly potent  
239 neutralizing antibodies: implication for developing subunit vaccine. *Biochemical and*  
240 *biophysical research communications* **324**, 773-781, doi:10.1016/j.bbrc.2004.09.106  
241 (2004).
- 242 16 Monne, I. *et al.* Emergence of a highly pathogenic avian influenza virus from a low-  
243 pathogenic progenitor. *J Virol* **88**, 4375-4388, doi:10.1128/JVI.03181-13 (2014).
- 244 17 Zhang, F. *et al.* Human infections with recently-emerging highly pathogenic H7N9 avian  
245 influenza virus in China. *J Infect* **75**, 71-75, doi:10.1016/j.jinf.2017.04.001 (2017).
- 246 18 Kristian G., A., Andrew Rambaut, W. Ian Lipkin, Holmes, E. C. & Garry, R. F. The Proximal  
247 Origin of SARS-CoV-2. (2020).
- 248 19 Hu, B. *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides new  
249 insights into the origin of SARS coronavirus. *PLoS pathogens* **13**, e1006698 (2017).  
250 <<http://europepmc.org/abstract/MED/29190287>>.
- 251 20 Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and  
252 complexes. *Nucleic Acids Res* **46**, W296-W303, doi:10.1093/nar/gky427 (2018).
- 253 21 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large  
254 phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).
- 255
- 256
- 257
- 258



259

260 **Fig. 1. Patterns of sequence identity between the consensus sequences of HCoV-19 and**  
 261 **representative beta-CoVs.**

262 (a) Sequence identities for HCoV-19 compared to representative beta-CoVs, including

263 RmYN02, RaTG13 (EPI\_ISL\_402131), ZC45 (MG772933), ZXC21 (MG772934),

264 pangolin/GX/P5L/2017 (EPI\_ISL\_410540) and SARS-CoV GZ02 (AY390556).

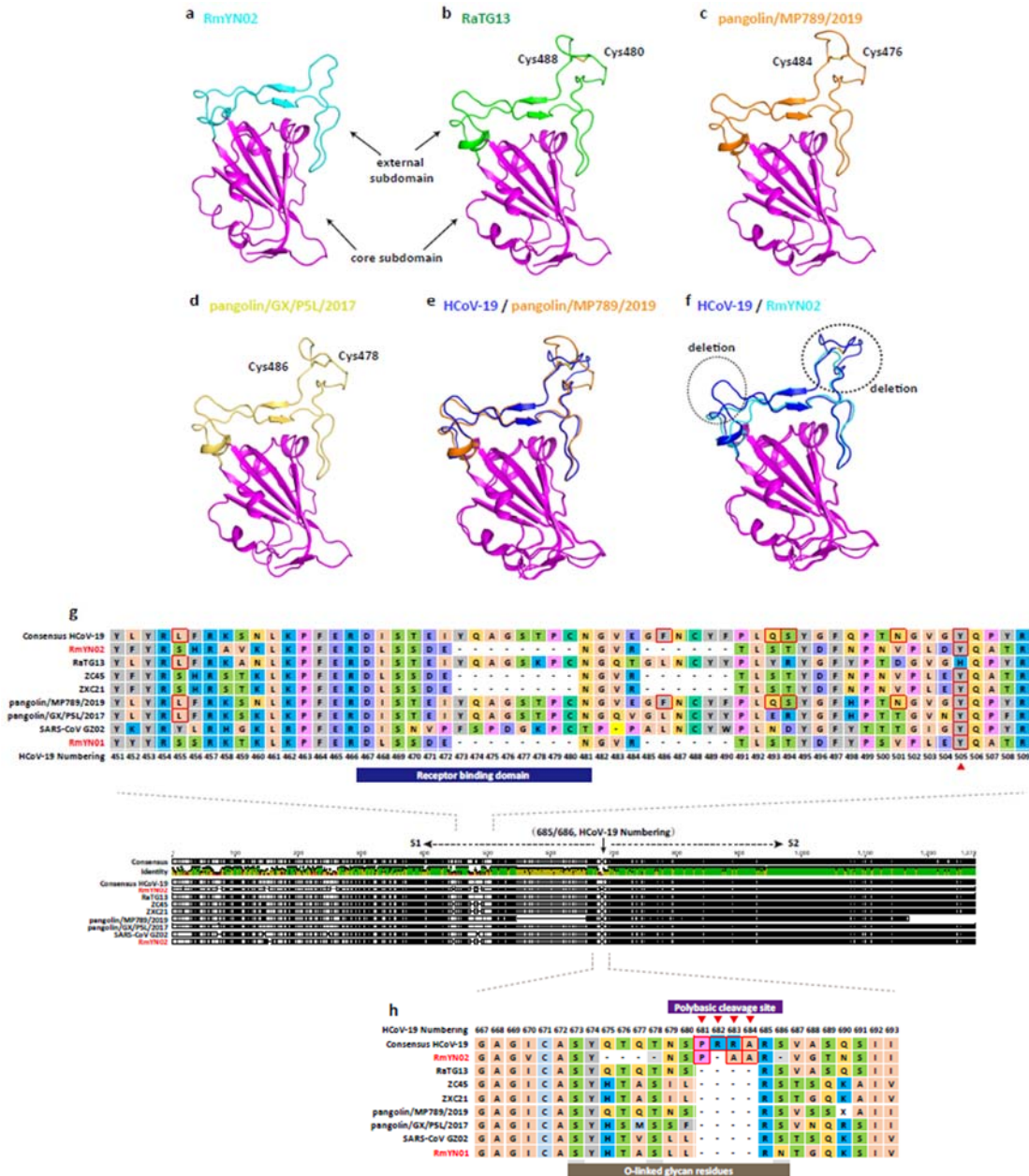
265 <sup>1</sup>Pangolin/GD/2019 represents a merger of GD/P1L and GD/P2S, and these values were

266 adapted from the reference<sup>7</sup>. “-” : No corresponding values in reference<sup>7</sup>. “/” : This orf is

267 not found. (b) Whole genome similarity plot between HCoV-19 and representative viruses

268 listed in panel (a). The analysis was performed using Simplot, with a window size of 1000bp

269 and a step size of 100bp. (c) Similarity plot in the spike gene (positions 1-1658) between  
 270 HCoV-19 and representative viruses listed in panel (a). The analysis was performed using  
 271 Simplot, with a window size of 150bp and a step size of 5bp.



272  
 273 **Fig. 2. Homology modelling of the RBD structures and molecular characterizations of**  
 274 **the S1/S2 cleavage site of RmYN02 and representative beta-CoVs.**

275 (a-d) Homology modelling and structural comparison of the RBD structures of RmYN02 and  
 276 representative beta-CoVs, including (a) RmYN02, (b) RaTG13, (c) pangolin/MP789/2019 and



277 (d) pangolin/GX/P5L/2017. The three-dimensional structures of the RBD from Bat-SL-CoV  
278 RmYN02, RaTG13, pangolin/MP789/2019 and pangolin/GX/P5L/2017 were modeled using  
279 the Swiss-Model program<sup>20</sup> employing the RBD of SARS-CoV (PDB: 2DD8) as a template. All  
280 the core subdomains are colored magenta, and the external subdomains of RmYN02, RaTG13,  
281 pangolin/MP789/2019 and pangolin/GX/P5L/2017 are colored cyan, green, orange and  
282 yellow, respectively. The conserved disulfide bond in RaTG13, pangolin/GD and pangolin/GX  
283 is highlighted, while it is missing in RmYN02 due to a sequence deletion.

284 (e-f) Superimposition of the RBD structure of pangolin/MP789/2019 (e) and RmYN02 (f) with  
285 that of HCoV-19. The two deletions located in respective loops in RmYN02 are highlighted  
286 using dotted cycles.

287 (g) Molecular characterizations of the RBD of RmYN02 and the representative beta-CoVs.

288 (h) Molecular characterizations of the cleavage site of RmYN02 and the representative beta-  
289 CoVs.

290

291

292

293

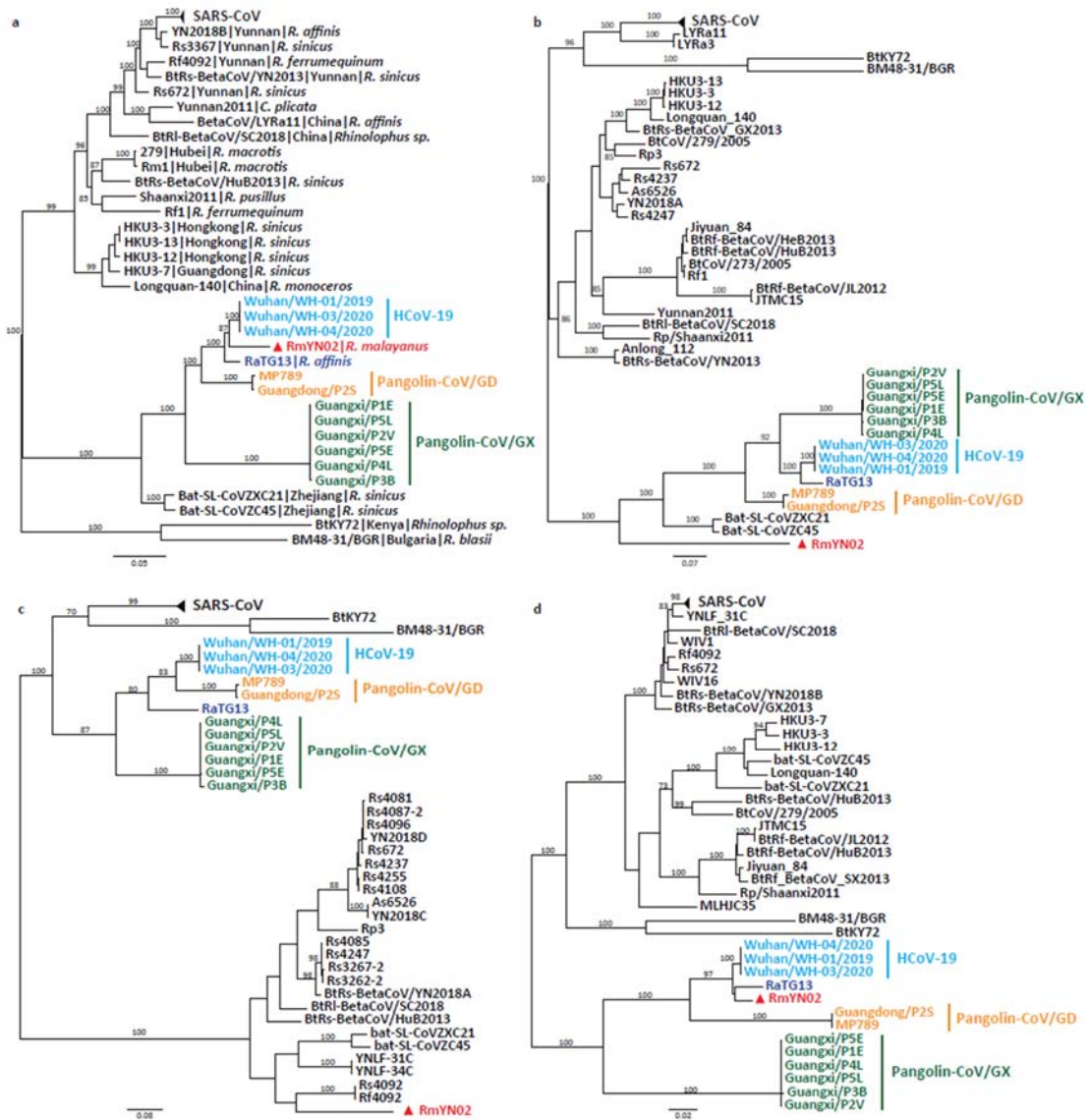
294

295

296

297

298



299

300 **Fig. 3. Phylogenetic analysis of HCoV-19 and representative viruses from the subgenus**

301 ***Sarbecoronavirus*.**

302 (a) Phylogenetic tree of the full-length virus genome. (b) the S gene. (c) the RBD. (d) the RdRp.

303 Phylogenetic analysis was performed using RAxML<sup>21</sup> with 1000 bootstrap replicates,

304 employing the GTR nucleotide substitution model. RBD is delimited as the gene region 991-

305 1572 of the spike gene according to the reference<sup>7</sup>. All the trees are midpoint rooted for

306 clarity.

307