Note: This paper has been revised after peer review, so that it can be considered technically correct.

Complete genome characterisation of a novel coronavirus associated with severe human respiratory disease in Wuhan, China

Fan Wu[1,6], Su Zhao[2,6], Bin Yu[3,6], Yan-Mei Chen[1,6], Wen Wang[4,6], Yi Hu[2,6], Zhi-Gang Song[1,6], Zhao-Wu Tao[2], Jun-Hua Tian[3], Yuan-Yuan Pei[1], Ming-Li Yuan[2], Yu-Ling Zhang[1], Fa-Hui Dai[1], Yi Liu[1], Qi-Min Wang[1], Jiao-Jiao Zheng[1], Lin Xu[1], Edward C. Holmes[5], Yong-Zhen Zhang[1,4]*

[1]Shanghai Public Health Clinical Center & School of Public Health, Fudan University, Shanghai, China.

[2]Department of Pulmonary and Critical Care Medicine, The Central Hospital of Wuhan, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430014, China.

[3]Wuhan Center for Disease Control and Prevention, Wuhan, Hubei, China

[4]Department of Zoonosis, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Changping, Beijing, China.

[5]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, Australia.

[6]These authors contributed equally: Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Yi Hu, Zhi-Gang Song. *e-mail: zhangyongzhen@shphc.org.cn

Emerging and re-emerging infectious diseases, such as SARS, MERS, Zika and highly pathogenic influenza present a major threat to public health[1-3]. Despite intense research effort, how, when and where novel diseases appear are still the source of considerable uncertainly. A severe respiratory disease was recently reported in the city of Wuhan, Hubei province, China. At the time of writing, at least 62 suspected cases have been reported since the first patient was hospitalized on December 12nd 2019. Epidemiological investigation by the local Center for Disease Control and Prevention (CDC) suggested that the outbreak was associated with a sea food market in Wuhan. We studied seven patients who were workers at the market, and collected bronchoalveolar lavage fluid (BALF) from one patient who exhibited a severe respiratory syndrome including fever, dizziness and cough, and who was admitted to Wuhan Central Hospital on December 26th 2019. Next generation metagenomic RNA sequencing[4] identified a novel RNA virus from the family *Coronaviridae* designed WH-Human-1 coronavirus (WHCV). Phylogenetic analysis of the complete viral genome (29,903 nucleotides) revealed that WHCV was most closely related (89.1% nucleotide similarity similarity) to a group of Severe Acute Respiratory Syndrome (SARS)-like coronaviruses (genus *Betacoronavirus*, subgenus *Sarbecovirus*) previously sampled from bats in China and that have a history of genomic recombination. This outbreak highlights the ongoing capacity of viral spill-over from animals to cause severe disease in humans.

Seven patients, comprising five men and two women, were hospitalized at the Central Hospital of Wuhan from December 14 through December 28, 2019. The median age of the

patients was 43, ranging from 31 to 70 years old. The clinical characteristics of the patients are shown in Table 1. Fever and cough were the most common symptoms. All patients had fever with body temperatures ranging from 37.2°C to 40°C. Patients 1, 2, 5, 6 and 7 had cough, while patients 1, 2 and 7 presented with severe cough with phlegm at onset of illness. Patients 4 and 5 also complained of chest tightness and dyspnea. Patients 1, 3, 4 and 6 experienced dizziness and patient 3 felt weakness. No neurological symptoms were observed in any of the patients. Bacterial culture revealed the presence of *Streptococcus* bacteria in throat swabs from patients 3, 4 and 7. Combination antibiotic, antiviral and glucocorticoid therapy were administered. Unfortunately, patient 1 and 4 showed respiratory failure: patient 1 was given high flow noninvasive ventilation, while patient 4 was provided with nasal/face mask ventilation (Table 1).

Epidemiological investigation by the Wuhan CDC revealed that all the suspected cases were linked to individuals working in a local indoor seafood market. Notably, in addition to fish and shell fish, a variety of live wild animals including hedgehogs, badgers, snakes, and birds (turtledoves) were available for sale in the market before the outbreak began, as well as animal carcasses and animal meat. No bats were available for sale. While the patients might have had contact with wild animals in the market, none recalled exposure to live poultry.

Patient 1 was a 41-year-old man with no history of hepatitis, tuberculosis or diabetes. He was admitted and hospitalized in Wuhan Central Hospital 6 days after the onset of illness. The patient reported fever, chest tightness, unproductive cough, pain and weakness for one week on presentation. Physical examination of cardiovascular, abdominal and neurologic examination was normal. Mild lymphopenia (less than 900 cells per cubic milli-meter) was

observed, but white blood cell and blood platelet count was normal in a complete blood count (CBC) test. Elevated levels of C-reactive protein (CRP, 41.4 mg/L of blood, reference range 0-6 mg/L) was observed and levels of aspartate aminotransferase, lactic dehydrogenase, and creatine kinase were slightly elevated in blood chemistry tests. The patient had mild hypoxemia with oxygen levels of 67mmHg by the Arterial Blood Gas (ABG) Test. On the first day of admission (day 6 after the onset of illness), chest radiographs were abnormal with air-space shadowing such a ground-glass opacities, focal consolidation and patchy consolidation in both lungs (Figure 1). Chest computed tomographic (CT) scans revealed bilateral focal consolidation, lobar consolidation and patchy consolidation, especially in the lower lung. A chest radiograph revealed a bilateral diffuse patchy and fuzzy shadow on day 5 after admission (day 11 after the onset of illness). Preliminary aetiological investigation excluded the presence of influenza virus, *Chlamydia pneumoniae* and *Mycoplasma pneumoniae* by commercial pathogen antigen detection kits and confirmed by PCR. Other common respiratory pathogens, including adenovirus, were also negative by qPCR (Figure S1). The condition of the patient did not improve after three days of treatment with combined antiviral and antibiotic therapy. He was admitted to the intensive care unit (ICU) and treatment with a high flow non-invasive ventilator was initiated. The patient was transferred to another hospital in Wuhan for further treatment 6 days after admission.

To investigate the possible aetiologic agents associated this disease, we collected bronchoalveolar lavage fluid (BALF) from patient 1 and performed deep meta-transcriptomic sequencing. All the clinical specimens were handled in a biosafety level 3 laboratory at the Shanghai Public Health Clinical Center. Total RNA was extracted from 200μl BAL fluid and

93 a meta-transcriptomic library was constructed for pair-end (150 bp) sequencing using an
94 Illumina MiniSeq as previously described[4-7]. In total, we generated 56,565,928 sequence
95 reads that were *de novo* assembled and screened for potential aetiologic agents. Of the
96 384,096 contigs assembled by Megahit[8], the longest (30,474 nucleotides [nt]) had high
97 abundance and was closely related to a bat SARS-like coronavirus isolate - bat-SL-CoVZC45
98 (GenBank Accession MG772933) - previously sampled in China, with a nt identity of 89.1%
99 (Table S1 and S2). The genome sequence of this novel virus, as well as its termini, were
100 determined and confirmed by RT-PCR[9] and 5'/3' RACE kits (TaKaRa), respectively. This
101 new virus was designated as WH-Human 1 coronavirus (WHCV) (and has also been referred
102 to as '2019-nCoV') and its whole genome sequence (29,903 nt) has been assigned GenBank
103 accession number MN908947. Remapping the RNA-seq data against the complete genome of
104 WHCV resulted in an assembly of 123,613 reads, providing 99.99% genome coverage at a
105 mean depth of 6.04X (range: 0.01X -78.84X) (Figure S2). The viral load in the BALF sample
106 was estimated by quantitative PCR (qPCR) to be 3.95×10^{8} copies/mL (Figure S3).

107 The viral genome organization of WHCV was characterized by sequence alignment
108 against two representative members of the genus *Betacoronavirus*: a human-origin
109 coronavirus (SARS-CoV Tor2, AY274119) and a bat-origin coronavirus (Bat-SL-CoVZC45,
110 MG772933) (Figure 2). The un-translational regions (UTR) and open reading frame (ORF) of
111 WHCV were mapped based on this sequence alignment and ORF prediction. The WHCV
112 viral genome was similar to these two coronaviruses (Figure 2 and Table S3), with a gene
113 order 5'-replicase ORF1ab-S-envelope(E)-membrane(M)-N-3'. WHCV has 5' and 3' terminal
114 sequences typical of the betacoronaviruses, with 265 nt at the 5' terminal and 229 nt at the 3'

terminal region. The predicted replicase ORF1ab gene of WHCV is 21,291 nt in length and contained 16 predicted non-structural proteins (Table S4), followed by (at least) 13 downstream ORFs. Additionally, WHCV shares a highly conserved domain (LLRKNGNKG: amino acids 122-130) with SARS-CoV in nsp1. The predicted S, ORF3a, E, M and N genes of WHCV are 3,822, 828, 228, 669 and 1,260 nt in length, respectively. In addition to these ORFs regions that are shared by all members of the subgenus *Sarbecovirus*, WHCV is similar to SARS-CoV in that it carries a predicted ORF8 gene (366 nt in length) located between the M and N ORF genes. The functions of WHCV ORFs were predicted based on those of known coronaviruses and given in Table S5. In a manner similar to SARS CoV Tor2, a leader transcription regulatory sequence (TRS) and nine putative body TRSs could be readily identified upstream of the 5' end of ORF, with the putative conserved TRS core sequence appeared in two forms – the ACGAAC or CUAAAC (Table S6).

To determine the evolutionary relationships between WHCV and previously identified coronaviruses, we estimated phylogenetic trees based on the nucleotide sequences of the whole genome sequence, non-structural protein genes ORF1a and 1b, and the main structural proteins encoded by the S, E, M and N genes (Figures 3 and S4). In all phylogenies WHCV clustered with members of the subgenus *Sarbecovirus*, including the SARS-CoV responsible for the global SARS pandemic of 2002-2003[1,2], as well as a number of SARS-like coronaviruses sampled from bats. However, WHCV changed topological position within the subgenus *Sarbecovirus* depending on which gene was used, suggestive of a past history of recombination in this group of viruses (Figures 3 and S4). Specifically, in the S gene tree (Figure S4), WHCV was most closely related to the bat coronavirus bat-SL-CoVZC45 with

137 82.3% amino acid (aa) identity (and ~77.2% aa identity to SARS CoV; Table S3), while in the
138 ORF1b phylogeny WHCV fell in a basal position within the subgenus *Sarbecovirus* (Figure
139 3). This topological division was also observed in the phylogenetic trees estimated for
140 conserved domains in the replicase polyprotein pp1ab (Figure S5).

141 To better understand the potential of WHCV to infect humans, the receptor-binding
142 domain (RBD) of its spike protein was compared to those in SARS-CoVs and bat SARS-like
143 CoVs. The RBD sequences of WHCV were more closely related to those of SARS-CoVs
144 (73.8%-74.9% aa identity) and SARS-like CoVs including strains Rs4874, Rs7327 and
145 Rs4231 (75.9%-76.9% aa identity) that are able to use the human ACE2 receptor for cell entry
146 (Table S7)[10]. In addition, the WHCV RBD was only one amino acid longer than the SARS-
147 CoV RBD (Figure 4a). In contrast, other bat SARS-like CoVs including the Rp3 strain that
148 cannot use human ACE2[11], had amino acid deletions at positions 473-477 and 460-472
149 compared to the SARS-CoVs (Figure 4a). The previously determined[12] crystal structure of
150 SARS-CoV RBD complexed with human ACE2 (PDB 2AJF) revealed that regions 473-477
151 and 460-472 directly interact with human ACE2 and hence may be important in determining
152 species specificity (Figure 4b). We predicted the three-dimension protein structures of WHCV,
153 Rs4874 and Rp3 RBD domains by protein homology modelling using the SWISS-MODEL
154 server and compared them to the crystal structure of SARS-CoV RBD domains (PDB 2GHV)
155 (Figure 4, c-f). In accord with the sequence alignment, the predicted protein structures of
156 WHCV and Rs4874 RBD domains were closely related to that of SARS-CoVs and different
157 from the predicted structure of the RBD domain from Rp3. In addition, the N-terminus of
158 WHCV S protein is more similar to that of SARS-CoV rather than other human coronaviruses

159 (HKU1 and OC43) (Figure S6) that can bind to sialic acid[13]. In sum, the high similarities of
160 amino acid sequences and predicted protein structure between WHCV and SARS-CoV RBD
161 domains suggest that WHCV may efficiently use human ACE2 as a cellular entry receptor,
162 perhaps facilitating human-to-human transmission[10, 14-15].

163 To further characterize putative recombination events in the evolutionary history of the
164 sarbecoviruses the whole genome sequence of WHCV and four representative coronaviruses -
165 Bat SARS-like CoV Rp3, CoVZC45, CoVZXC21 and SARS-CoV Tor2 - were analysed using
166 the Recombination Detection Program v4 (RDP4)[16]. Although the similarity plots suggested
167 possible recombination events between WHCV and SARS CoVs or SARS-like CoVs (Figure
168 S7), there was no significant evidence for recombination across the genome as a whole.
169 However, some evidence for past recombination was detected in the S gene of WHCV and
170 SARS CoV and bat SARS-like CoVs (WIV1 and RsSHC014) ($p<3.147\times10^{-3}$ to $p<9.198\times10^{-}$
171 ^{9}), with similarity plots suggesting the presence of recombination break points at nucleotides
172 1,029 and 1,652 that separated the WHCV S gene into three regions (Figure 5). In
173 phylogenies of the fragment nt 1 to 1029 and nt 1652 to the end of the sequence, WHCV was
174 most closely related to Bat-SL-CoVZC45 and Bat-SL-CoVZXC21, whereas in the region nt
175 1030 to 1651 (the RBD region) WHCV grouped with SARS CoV and bat SARS-like CoVs
176 (WIV1 and RsSHC014) that are capable of direct human transmission[14,17].

177 Coronaviruses are associated with a number of infectious disease outbreaks in humans,
178 including SARS in 2002/3 and MERS in 2012[1,18]. Four other coronaviruses - human
179 coronaviruses HKU1, OC43, NL63 and 229E - are also associated with respiratory disease[19-
180 22]. Although SARS-like coronaviruses have been widely identified in mammals including bats

181 since 2005 in China[9,23-25], the exact origin of human-infected coronaviruses remains unclear.
182 Herein, we describe a novel coronavirus - WHCV - in BALF from a patient experiencing
183 severe respiratory disease in Wuhan, China. Phylogenetic analysis suggested that WHCV
184 represents a novel virus within genus *Betacoronavirus* (subgenus *Sarbecovirus*) and hence
185 that exhibits some genomic and phylogenetic similarity to SARS-CoV[1], particularly in the
186 RBD. These genomic and clinical similarities to SARS, as well as its high abundance in
187 clinical samples, provides evidence for an association between WHCV and the ongoing
188 outbreak of respiratory disease in Wuhan.

189 The identification of multiple SARS-like-CoVs in bats led to the idea that these animals
190 act as the natural reservoir hosts of these viruses[19,20]. Although SARS-like viruses have been
191 identified widely in bats in China, viruses identical to SARS-CoV have not yet been
192 documented. Notably, WHCV is most closely related to bat coronaviruses, even exhibiting
193 100% aa similarity to Bat-SL-CoVZC45 in the nsp7 and E proteins. Hence, these data suggest
194 that bats are a possible reservoir host of WHCV. However, as a variety of animal species were
195 for sale in the market when the disease was first reported, more work is needed to determine
196 the natural reservoir and any intermediate hosts of WHCV.

197

198 **Acknowledgements** This study was supported by the Special National Project on
199 investigation of basic resources of China (Grant SQ2019FY010009) and the National Natural
200 Science Foundation of China (Grants 81861138003 and 31930001). ECH is supported by an
201 ARC Australian Laureate Fellowship (FL170100022).

202

Author Contributions Y.-Z.Z. conceived and designed the study. S.Z, Y.H, Z.-W.T. and M.-L.Y. performed the clinical work and sample collection. B.Y and J.-H.T. performed epidemiological investigation and sample collection. F.W, Z.-G.S., L.X., Y.-Y.P., Y.-L.Z., F.-H.D., Y.L., J.-J.Z. and Q.-M.W. performed the experiments. Y.-M.C., W.W., F.W., E.C.H. and Y.-Z.Z. analysed the data. Y.-Z.Z. E.C.H. and F.W. wrote the paper with input from all authors. Y.-Z.Z. led the study.

Correspondence and requests for materials should be addressed to Y.-Z.Z. (zhangyongzhen@shphc.org.cn or zhangyongzhen@icdc.cn).

Competing interests The authors declare no competing interests.

References

1. Drosten, C., et al., Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N. Engl. J. Med.* **348**,1967–1976 (2003).
2. Wolfe, N.D., Dunavan, C.P., Diamond, J. Origins of major human infectious diseases. *Nature*. **447**, 279-283 (2007).
3. Ventura, C.V., Maia, M., Bravo-Filho, V., Góis, A.L. & Belfort, R. Jr. Zika virus in Brazil and macular atrophy in a child with microcephaly. *Lancet.* **387**, 228 (2016).
4. Shi, M. et al. Redefining the invertebrate RNA virosphere. *Nature.* **540**, 539-543 (2016).
5. Shi, M., et al. The evolutionary history of vertebrate RNA viruses. *Nature*. **556**,197-202 (2018).
6. Yadav, P.D., et al. Nipah virus sequences from humans and bats during Nipah outbreak, Kerala, India, 2018. *Emerg. Infect. Dis.* **25**, 1003-1006 (2019).
7. McMullan, L.K., et al. Characterisation of infectious Ebola virus from the ongoing outbreak to guide response activities in the Democratic Republic of the Congo: a phylogenetic and in vitro analysis. *Lancet. Infect. Dis.* **19**, 1023-1032 (2019).
8. Li, D., Liu, C.M., Luo, R., Sadakane, K. & Lam, T.W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674-1676 (2015).
9. Wang, W., et al. Discovery, diversity and evolution of novel coronaviruses sampled from rodents in China. *Virology.* **474**, 19-27 (2015).
10. Hu, B. et al. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog.* **13**: e1006698 (2017).

236 11. Ren, W. et al. Difference in receptor usage between severe acute respiratory syndrome
237 (SARS) coronavirus and SARS-like coronavirus of bat origin. *J Virol.* **82**:1899-1907
238 (2008).
239 12. Li, F., Li, W., Farzan, M., Harrison, S.C. Structure of SARS coronavirus spike receptor-
240 binding domain complexed with receptor. *Science.* **309**, 1864-1868 (2005).
241 13. Hulswit, R.J.G., et al. Human coronaviruses OC43 and HKU1 bind to 9-O-acetylated
242 sialic acids via a conserved receptor-binding site in spike protein domain A. *Proc Natl*
243 *Acad Sci USA*., **116**, 2681-2690 (2019).
244 14. Ge, X.Y. et al. Isolation and characterization of a bat SARS-like coronavirus that uses the
245 ACE2 receptor. *Nature.* **503**: 535-538 (2013).
246 15. Yang, X.L., et al. Isolation and characterization of a novel bat coronavirus closely related
247 to the direct progenitor of severe acute respiratory syndrome coronavirus. *J Virol.* **90**:
248 3253-3256 (2016).
249 16. Martin, D.P., Lemey, P., Lott, M., Moulton, V., Posada, D., Lefeuvre, P. RDP3: a flexible
250 and fast computer program for analyzing recombination. *Bioinformatics* 26:2462–2463
251 (2010).
252 17. Menachery, V.D., et al. A SARS-like cluster of circulating bat coronaviruses shows
253 potential for human emergence. *Nat Med.* **21**:1508-1513 (2015).
254 18. Bermingham, A., et al. Severe respiratory illness caused by a novel coronavirus, in a
255 patient transferred to the United Kingdom from the Middle East, September 2012. *Euro.*
256 *Surveill.* **17**, 20290 (2012).
257 19. Hamre, D. & Procknow, J.J. A new virus isolated from the human respiratory tract. *Proc.*

Soc. Exp. Biol. Med. **121**, 190–193 (1966).

20. McIntosh, K., Becker, W.B., Chanock, R.M. Growth in suckling-mouse brain of "IBV-like" viruses from patients with upper respiratory tract disease. *Proc Natl Acad Sci USA.* **58**, 2268-73(1967).
21. van der Hoek, L., et al. Identification of a new human coronavirus. *Nat. Med.***10**, 368–373 (2004).
22. Woo, P.C., et al. Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. *J. Virol.***79**,884–895 (2005).
23. Li, W., et al. Bats are natural reservoirs of SARS-like coronaviruses. *Science* **310**, 676–679 (2005).
24. Lau S.K., et al. Severe acute respiratory syndrome coronavirus- like virus in Chinese horseshoe bats. *Proc.Natl.Acad.Sci.U.S.A.***102**, 14040–14045 (2005).
25. Wang, W., et al. Discovery of a highly divergent coronavirus in the Asian house shrew from China illuminates the origin of the Alphacoronaviruses. *J. Virol*. **91**, e00764-17 (2017).

274 **Figure legends**

275

276 **Figure 1.** Chest radiographs of patient 1. **a. b. c. d.** Chest computed tomographic scans of
277 Patient 1 were obtained on the day of admission (day 6 after the onset of illness). Bilateral
278 focal consolidation, lobar consolidation, and patchy consolidation were clearly observed,
279 especially in the lower lung. **e**. Chest radiograph of patient 1 was obtained on day 5 after
280 admission (day 11 after the onset of illness). Bilateral diffuse patchy and fuzzy shadow were
281 observed.

282

283 **Figure 2.** Genome organization of SARS and SARS-like CoVs including Tor2, CoVZC45
284 and WHCV determined here.

285

286 **Figure 3.** Maximum likelihood phylogenetic trees of nucleotide sequences of the ORF1a,
287 ORF1b, E and M genes of WHCV and related coronaviruses. Numbers (>70) above or below
288 branches indicate percentage bootstrap values for the associated nodes. The trees were mid-
289 point rooted for clarity only. The scale bar represents the number of substitutions per site.

290

291 **Figure 4.** Analysis of receptor-binding domain (RBD) of the spike (S) protein of WHCV
292 coronavirus. **(a)** Amino acid sequence alignment of SARS-like CoV RBD sequences. Three
293 bat SARS-like CoVs, which could efficiently utilize the human ACE2 as receptor, had an
294 RBD sequence of similar size to SARS-CoV, and WHCV contains a single Val 470 insertion.
295 The key amino acid residues involved in the interaction with human ACE2 are marked with a

brown box. In contrast, five bat SARS-like CoVs had amino acid deletions at two motifs (amino acids 473-477 and 460-472) compared with those of SARS-CoV, and Rp3 has been reported not to use ACE2.[11] **(b)** The two motifs (aa 473-477 and aa 460-472) are shown in red on the crystal structure of the SARS-CoV spike RBD complexed with receptor human ACE2 (PDB 2AJF). Human ACE2 is shown in blue and the SARS-CoV spike RBD is shown in green. Important residues in human ACE2 that interact with SARS-CoV spike RBD are marked. **(c)** Predicted protein structures of RBD of WHCV spike protein based on target-template alignment using ProMod3 on the SWISS-MODEL server. The most reliable models were selected based on GMQE and QMEAN Scores. Template: 2ghw.1.A, GMQE: 0.83; QMEAN:-2.67. Motifs resembling amino acids 473-477 and 460-472 of the SARS-CoV spike protein are shown in red. **(d)** Predicted structure of RBD of SARS-like CoV Rs4874. Template: 2ghw.1.A, GMQE:0.99; QMEAN:-0.72. Motifs resembling amino acids 473-477 and 460-472 of the SARS-CoV spike protein are shown in red. (**e**) Predicted structure of the RBD of SARS-like CoV Rp3. Template: 2ghw.1.A, GMQE:0.81, QMEAN:-1.50. **(f)** Crystal structure of RBD of SARS-CoV spike protein (green) (PDB 2GHV). Motifs of amino acids 473-477 and 460-472 are shown in red.

Figure 5. Possible recombination events in the S gene of sarbecoviruses. A sequence similarity plot (upper panel) reveals two putative recombination break-points shown by black dashed lines, with their locations indicated at the bottom. The plot shows S gene similarity comparisons of the WHCV (query) against SARS-CoV Tor2 and bat SARS-like CoVs WIV1, Rf1 and CoVZC45. Phylogenies of major parental region (1-1028 and 1653-3804) and minor

318 parental region (1029-1652) are shown below the similarity plot. Phylogenies were estimated
319 using a ML method and mid-point rooted for clarity only. Numbers above or below branches
320 indicate percentage bootstrap values.
321
322 **Methods**
323 **Cases and collection of clinical data and samples**
324 Patients presenting with acute onset of fever (>37.5°C), cough, and chest tightness, and who
325 were admitted to Wuhan Central Hospital in Wuhan city, China, were considered as suspected
326 cases. During admission, bronchoalveolar lavage fluid (BALF) was collected and stored at -
327 80°C until further processing. Demographic, clinical and laboratory data were retrieved from
328 the clinical records of the confirmed patients. The study was reviewed and approved by the
329 ethics committee of the National Institute for Communicable Disease Control and Prevention,
330 Chinese Center for Disease Control and Prevention (CDC).
331
332 **RNA library construction and sequencing**
333 Total RNA was extracted from the BALF sample of patient 1 using the RNeasy Plus
334 Universal Mini Kit (Qiagen) following the manufacturer's instructions. The quantity and
335 quality of the RNA solution was assessed using a Qbit machine and an Agilent 2100
336 Bioanalyzer (Agilent Technologies) before library construction and sequencing. An RNA
337 library was then constructed using the SMARTer Stranded Total RNA-Seq Kit v2 (TaKaRa,
338 Dalian, China). Ribosomal RNA (rRNA) depletion was performed during library construction
339 following the manufacturer's instructions. Paired-end (150 bp) sequencing of the RNA library

340 was performed on the MiniSeq platform (Illumina). Library preparation and sequencing were
341 carried out at the Shanghai Public Health Clinical Center, Fudan University, Shanghai, China.

342 **Data processing and viral agent identification**

343 Sequencing reads were first adaptor- and quality-trimmed using the Trimmomatic program[26].
344 The remaining reads (56,565,928 reads) were assembled *de novo* using both the Megahit
345 (version 1.1.3)[8] and Trinity program (version 2.5.1)[27] with default parameter settings. Megahit
346 generated a total of 384,096 assembled contigs (size range: 200-30,474 nt), while Trinity
347 generated 1,329,960 contigs with a size range of 201 to 11,760 nt. All of these assembled
348 contigs were compared (using blastn and Diamond blastx) against the entire non-redundant
349 nucleotide (Nt) and protein (Nr) database, with *e*-values set to 1×10^{-10} and 1×10^{-5},
350 respectively. To identify possible aetiologic agents present in the sequence data, the
351 abundance of the assembled contigs was first evaluated as the expected counts using the
352 RSEM program[28] implemented in Trinity. Non-human reads (23,712,657 reads), generated by
353 filtering host reads using the human genome (human release 32, GRCh38.p13, downloaded
354 from Gencode) by Bowtie2[29], were used for the RSEM abundance assessment.

355 As the longest contigs generated by Megahit (30,474 nt) and Trinity (11,760 nt) both
356 had high similarity to the bat SARS-like coronavirus isolate bat-SL-CoVZC45 and were at
357 high abundance (Table S1 and S2), the longer one (30,474 nt) that covered almost the whole
358 virus genome was used for primer design for PCR confirmation and genome termini
359 determination. Primers used in PCR, qPCR and RACE experiments are listed in Table S8. The
360 PCR assay was conducted as described previously[9] and the complete genome termini was
361 determined using the Takara SMARTer RACE 5'/3' kit (TaKaRa) following the

362 manufacturer's instructions. Subsequently, the genome coverage and sequencing depth were
363 determined by remapping all of the adaptor- and quality-trimmed reads to the whole genome
364 of WHCV using Bowtie2[29] and Samtools[30].

365 The viral loads of WHCV in BALF of patient 1 were determined by quantitative real-
366 time RT-PCR with Takara One Step PrimeScript™ RT-PCR Kit (Takara RR064A) following
367 the manufacturer's instructions. Real-time RT-PCR was performed using 2.5μl RNA with
368 8pmol of each primer and 4pmol probe under the following conditions: reverse transcription
369 at 42°C for 10 minutes, and 95°C for 1 minute, followed by 40 cycles of 95°C for 15 seconds
370 and 60°C for 1 minute. The reactions were performed and detected by ABI 7500 Real-Time
371 PCR Systems. PCR product covering the Taqman primers and probe region was cloned into
372 pLB vector using the Lethal Based Simple Fast Cloning Kit (TIAGEN) as standards for
373 quantitative viral load test.

374 **Virus genome characterization and phylogenetic analysis**

375 For the newly identified virus genome, the potential open reading frames (ORFs) were
376 predicted and annotated using the conserved signatures of the cleavage sites recognized by
377 coronavirus proteinases, and were processed in the Lasergene software package (version 7.1,
378 DNAstar). The viral genes were aligned using the L-INS-i algorithm implemented in MAFFT
379 (version 7.407)[31].

380 Phylogenetic analyses were then performed using the nucleotide sequences of various
381 CoV gene data sets: (i) Whole genome, (ii) ORF1a, (iii) ORF1b, (iv) nsp5 (3CLpro), (v)
382 RdRp (nsp12), (vi) nsp13 (Hel), (vii) nsp14 (ExoN), (viii) nsp15 (NendoU), (ix) nsp16 (O-

383 MT), (x) spike (S), and the (xi) nucleocapsid (N). Phylogenetic trees were inferred using the
384 Maximum likelihood (ML) method implemented in the PhyML program (version 3.0)[32], using
385 the Generalised Time Reversible substitution (GTR) model and Subtree Pruning and
386 Regrafting (SPR) branch-swapping. Bootstrap support values were calculated from 1,000
387 pseudo-replicate trees. The best-fit model of nucleotide substitution was determined using
388 MEGA (version 5)[33]. Amino acid identities among sequences were calculated using the
389 MegAlign program implemented in the Lasergene software package (version 7.1, DNAstar).

390 **Genome recombination analysis**

391 Potential recombination events in the history of the sarbecoviruses were assessed using both
392 the Recombination Detection Program v4 (RDP4)[16] and Simplot (version 3.5.1)[34]. The RDP4
393 analysis was conducted based on the complete genome (nucleotide) sequence, employing the
394 RDP, GENECONV, BootScan, maximum chi square, Chimera, SISCAN, and 3SEQ methods.
395 Putative recombination events were identified with a Bonferroni corrected p-value cut-off of
396 0.01. Similarity plots were inferred using Simplot to further characterize potential
397 recombination events, including the location of breakpoints.

398 **Analysis of RBD domain of WHCV spike protein**

399 An amino acid sequence alignment of WHCV, SARS-CoVs, bat SARS-like CoVs RBD
400 sequences was performed using MUSCLE[35]. The predicted protein structures of the spike
401 protein RBD were estimated based on target-template alignment using ProMod3 on SWISS-
402 MODEL server (https://swissmodel.expasy.org/). The sequences of the spike RBD domains of
403 WHCV, Rs4874 and Rp3 were searched by BLAST against the primary amino acid sequence
404 contained in the SWISS-MODEL template library (SMTL, last update: 2020-01-09, last

405 included PDB release: 2020-01-03). Models were built based on the target-template alignment
406 using ProMod3. The global and per-residue model quality were assessed using the QMEAN
407 scoring function[36]. The PDB files of the predicted protein structures were displayed and
408 compared with the crystal structures of SARS-CoV spike RBD (PDB 2GHV)[37] and the crystal
409 of structure of SARS-CoV spike RBD complexed with human ACE2 (PDB 2AJF)[12].
410

411 **References**

412 26. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
413 sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
414 27. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a
415 reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
416 28. Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A. & Dewey, C.N. RNA-Seq gene
417 expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493-500 (2010).
418 29. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods.*
419 **9**, 357–359 (2012).
420 30. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **15**,
421 2978-2079 (2009).
422 31. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7:
423 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).
424 32. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood
425 phylogenies: assessing the performance of PhyML 3.0. *Syst.Biol.* **59**, 307-321 (2010).
426 33. Tamura, K. et al. MEGA5: molecular evolutionary genetics analysis using maximum
427 likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**,
428 2731–2739 (2011).
429 34. Lole, K.S. et al. Full-length human immunodeficiency virus type 1 genomes from
430 subtype C-infected seroconverters in India, with evidence of intersubtype recombination.
431 *J. Virol*. **73**, 152–160 (1999).
432 35. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high

433 throughput. *Nucleic Acids Res.* **32**, 1792-1797 (2004).

434 36. Waterhouse, A. et al. SWISS-MODEL: homology modelling of protein structures and

435 complexes. *Nucleic Acids Res.* **46**, W296-W303 (2018).

436 37. Hwang, W.C. et al. Structural basis of neutralization by a human anti-severe acute

437 respiratory syndrome spike protein antibody, 80R. *J Biol Chem.* **281**, 34610-34616

438 (2006).

439

440 **Supplementary legends**

441

442 **Supplementary Tables**

443 **Table S1.** The top 50 abundant assembled contigs generated using the Megahit program.

444 **Table S2.** The top 80 abundant assembled contigs generated using the Trinity program.

445 **Table S3.** Amino acid identities of the selected predicted gene products between the novel
446 coronavirus (WHCV) and known betacoronaviruses.

447 **Table S4.** Cleavage products of the replicase polyproteins of WHCV.

448 **Table S5.** Predicted gene functions of WHCV ORFs.

449 **Table S6**. Coding of potential and putative transcription regulatory sequences of the genome
450 sequence of WHCV.

451 **Table S7.** Amino acid identities of the RBD sequence between SARS- and bat SARS-like
452 CoVs.

453 **Table S8.** PCR primers used in this study.

454

455 **Supplementary Figures**

456 **Figure S1.** Detection of other respiratory pathogens by qPCR.

457 **Figure S2.** Mapped read count plot showing the coverage depth per base of the WHCV
458 genome.

459 **Figure S3.** Detection of WHCV in clinical samples by RT-qPCR. **(a)** Specificity of the
460 WHCV primers used in RT-qPCR. Test samples comprised clinical samples that are positive
461 for at least one of the following viruses: Influenza A virus (09H1N1 and H3N2), Influenza B
462 virus, Human adenovirus, Respiratory syncytial virus, Rhinovirus, Parainfluenza virus type 1-
463 4, Human bocavirus, Human metapneumovirus, Coronavirus OC43, Coronavirus NL63,
464 Coronavirus 229E and Coronavirus HKU1. **(b-c)** Standard curve. **(d)** Amplification curve of
465 WHCV.

466 **Figure S4**. Maximum likelihood phylogenetic trees of the nucleotide sequences of the whole
467 genome, S and N genes of WHCV and related coronaviruses. Numbers (>70) above or below
468 branches indicate percentage bootstrap values. The trees were mid-point rooted for clarity
469 only. The scale bar represents the number of substitutions per site.

470 **Figure S5**. Maximum likelihood phylogenetic trees of the nucleotide sequences of the 3CL,
471 RdRp, Hel, ExoN, NendoU, and O-MT genes of WHCV and related coronaviruses. Numbers
472 (>70) above or below branches indicate percentage bootstrap values. The trees were mid-point
473 rooted for clarity only. The scale bar represents the number of substitutions per site.

474 **Figure S6.** Amino acid sequence comparison of the N-terminal domain (NTD) of spike
475 protein of WHCV, bovine coronavirus (BCoV), mouse hepatitis virus (MHV) and human

476 coronavirus (HCoV OC43 and HKU1) that can bind to sialic acid and the SARS-CoVs that

477 cannot. The key residues[13] for sialic acid binding on BCoV, MHV, HCoV OC43 and HKU1

478 were marked with a brown box.

479 **Figure S7.** A sequence similarity plot of WHCV, SARS- and bat SARS-like CoVs revealing

480 putative recombination events.

481 **Table 1.** Clinical symptoms and patient data

Characteristic	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Patient 7
Age (Year)	41	44	42	70	31	51	43
Sex	M	F	M	F	M	M	M
Date of illness onset	Dec 20,2019	Dec 22,2019	Dec 24,2019	Dec 24,2019	Dec 21,2019	Dec 16,2109	Dec 14,2019
Date of admission	Dec 26,2019	Dec 22,2019	Dec 28,2019	Dec 28,2019	Dec 28,2019	Dec 27,2019	Dec 14,2019
Fever	+	+	+	+	+	+	+
Body Temperature (ºC)	38.4	37.3	39	37.9	38.7	37.2	38
Cough	+	+	+	+	+	+	+
Sputum Production	+	+	-	-	-	-	+
Dizzy	+	-	+	+	-	+	-
Weakness	+	-	+	-	-	-	-
Chest tightness	+	-	-	+	+	-	-
Dyspnea	+	-	-	+	+	+	-
Bacterial culture	-	-	streptococcus pneumoniae	streptococcus pneumoniae	-	-	streptococcus pneumoniae
Glucocorticoid therapy	No	No	Yes	Yes	Yes	Yes	No
Antibiotic therapy	Cefoselis	Ceftazidime, Levofloxacin	Cefminox	Cefminox, moxifloxacin	Cefminox	No	No
Antiviral therapy	Oseltamivir	No	Oseltamivir, ganciclovir	Oseltamivir, ganciclovir	Oseltamivir, ganciclovir	Oseltamivir, ganciclovir	No
Oxygen therapy	mechanical ventilation	No	No	Mask	No	No	No

482
483

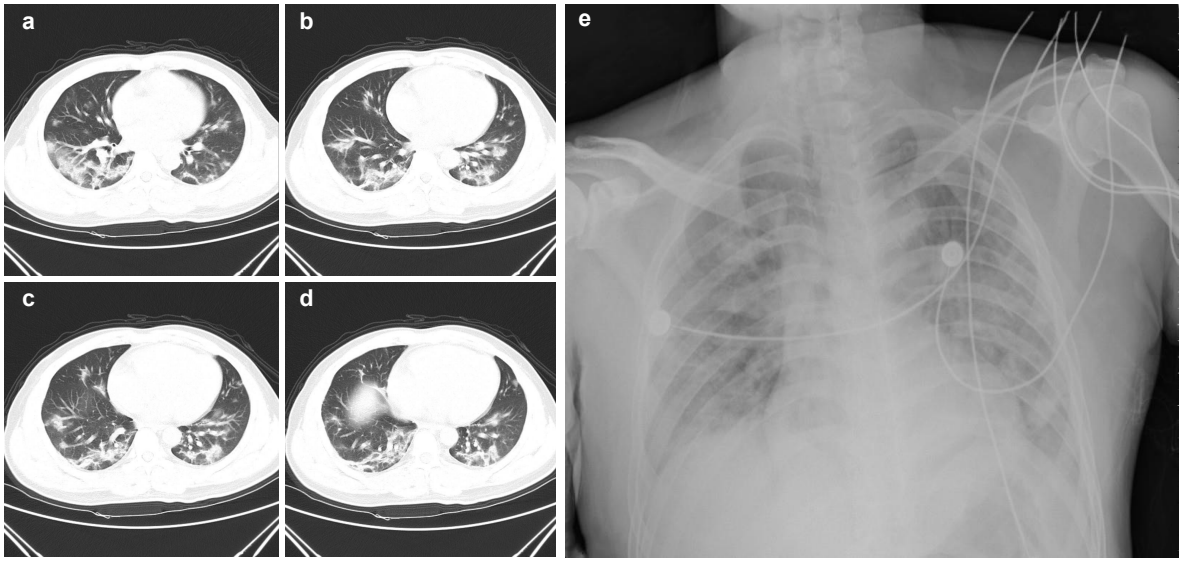


Figure 1. Chest radiographs of patient 1. **a. b. c. d.** Chest computed tomographic scans of Patient 1 were obtained on the day of admission (day 6 after the onset of illness). Bilateral focal consolidation, lobar consolidation, and patchy consolidation were clearly observed, especially in the lower lung. **e**. Chest radiograph of patient 1 was obtained on day 5 after admission (day 11 after the onset of illness). Bilateral diffuse patchy and fuzzy shadow were observed.

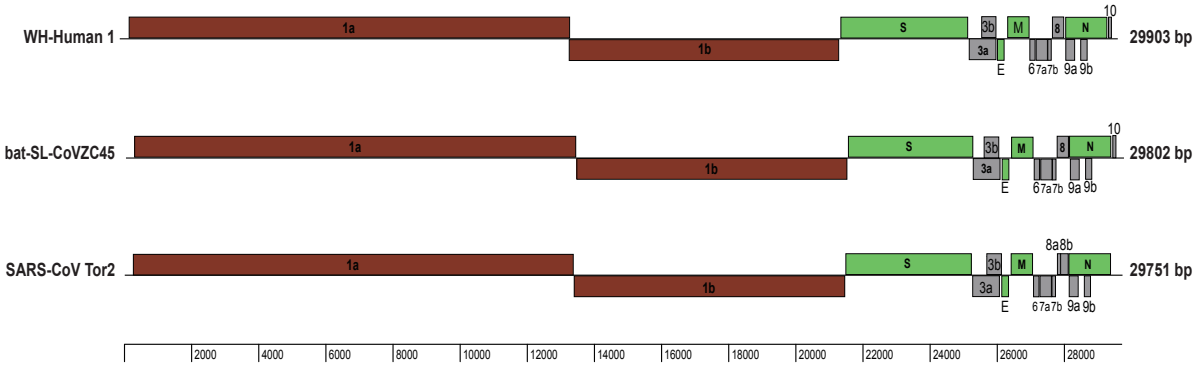


Figure 2. Genome organization of SARS and SARS-like CoVs including Tor2, CoVZC45 and WHCV determined here.

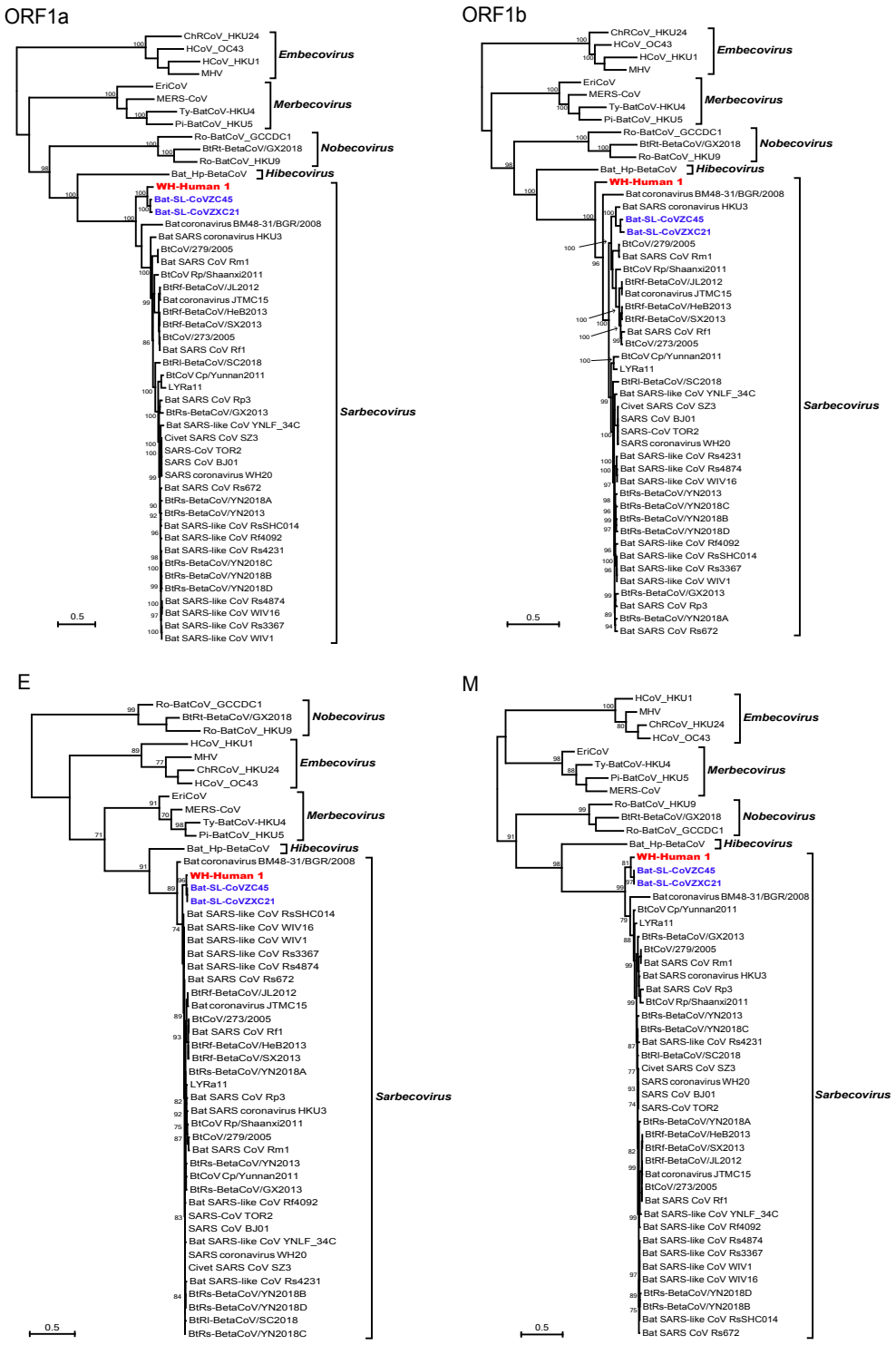
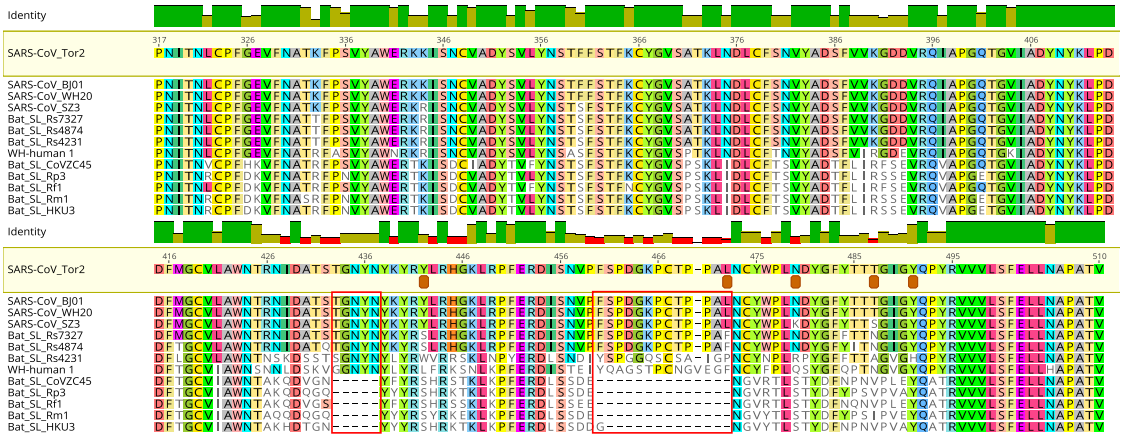
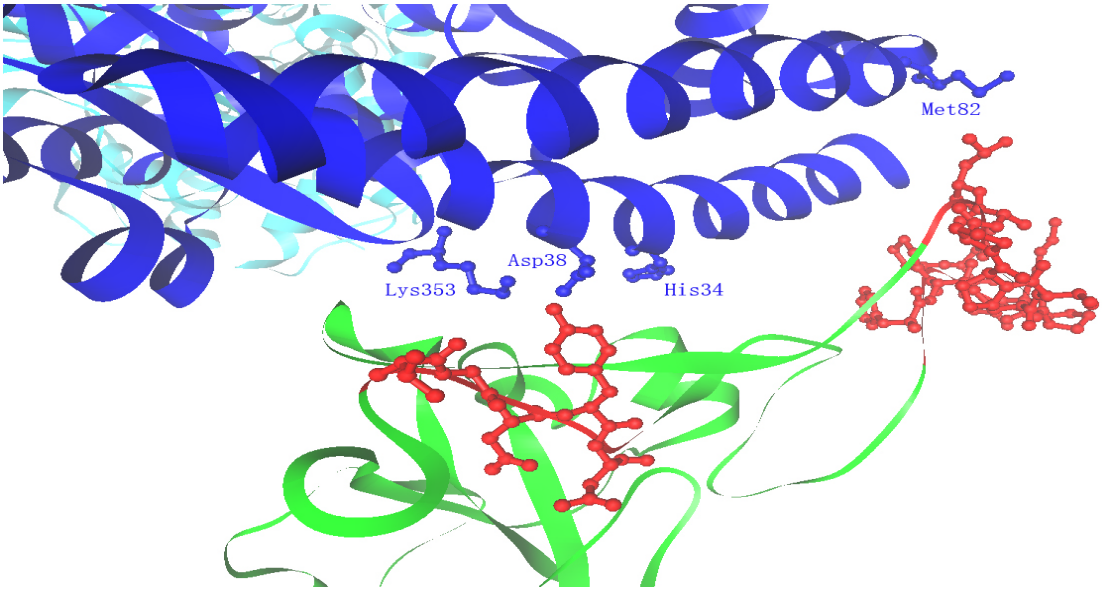


Figure 3. Maximum likelihood phylogenetic trees of nucleotide sequences of the ORF1a, ORF1b, E and M genes of WHCV and related coronaviruses. Numbers (>70) above or below branches indicate percentage bootstrap values for the associated nodes. The trees were mid-point rooted for clarity only. The scale bar represents the number of substitutions per site.

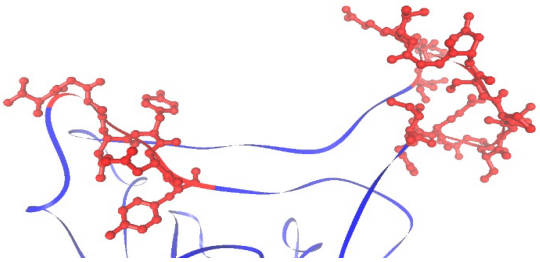
(a)


Identity
SARS-CoV_Tor2
SARS-CoV_BJ01
SARS-CoV_WH20
SARS-CoV_SZ3
Bat_SL_Rs7327
Bat_SL_Rs4874
Bat_SL_Rs4231
WH-human 1
Bat_SL_CoVZC45
Bat_SL_Rp3
Bat_SL_Rf1
Bat_SL_Rm1
Bat_SL_HKU3

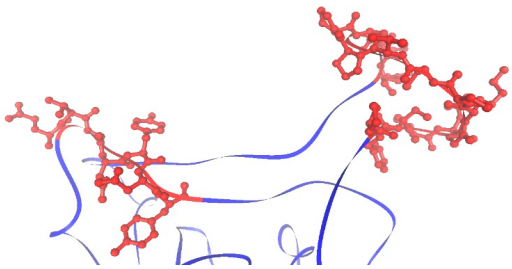
(b)


Met82
Asp38
Lys353
His34

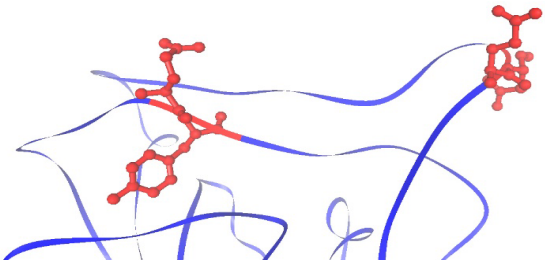
(c)



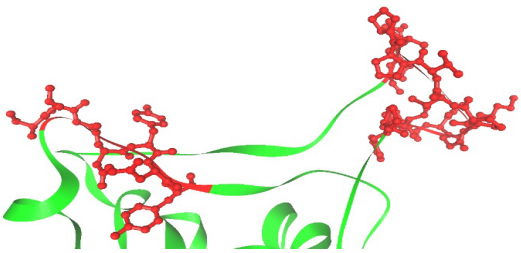
(d)



(e)



(f)



502

503 **Figure 4.** Analysis of receptor-binding domain (RBD) of the spike (S) protein of WHCV
504 coronavirus. **(a)** Amino acid sequence alignment of SARS-like CoV RBD sequences. Three
505 bat SARS-like CoVs, which could efficiently utilize the human ACE2 as receptor, had an
506 RBD sequence of similar size to SARS-CoV, and WHCV contains a single Val 470 insertion.
507 The key amino acid residues involved in the interaction with human ACE2 are marked with a
508 brown box. In contrast, five bat SARS-like CoVs had amino acid deletions at two motifs
509 (amino acids 473-477 and 460-472) compared with those of SARS-CoV, and Rp3 has been
510 reported not to use ACE2.[11] **(b)** The two motifs (aa 473-477 and aa 460-472) are shown in red
511 on the crystal structure of the SARS-CoV spike RBD complexed with receptor human ACE2
512 (PDB 2AJF). Human ACE2 is shown in blue and the SARS-CoV spike RBD is shown in
513 green. Important residues in human ACE2 that interact with SARS-CoV spike RBD are
514 marked. **(c)** Predicted protein structures of RBD of WHCV spike protein based on target-
515 template alignment using ProMod3 on the SWISS-MODEL server. The most reliable models
516 were selected based on GMQE and QMEAN Scores. Template: 2ghw.1.A, GMQE: 0.83;
517 QMEAN:-2.67. Motifs resembling amino acids 473-477 and 460-472 of the SARS-CoV spike
518 protein are shown in red. **(d)** Predicted structure of RBD of SARS-like CoV Rs4874.
519 Template: 2ghw.1.A, GMQE:0.99; QMEAN:-0.72. Motifs resembling amino acids 473-477
520 and 460-472 of the SARS-CoV spike protein are shown in red. (**e**) Predicted structure of the
521 RBD of SARS-like CoV Rp3. Template: 2ghw.1.A, GMQE:0.81, QMEAN:-1.50. **(f)** Crystal
522 structure of RBD of SARS-CoV spike protein (green) (PDB 2GHV). Motifs of amino acids
523 473-477 and 460-472 are shown in red.
524

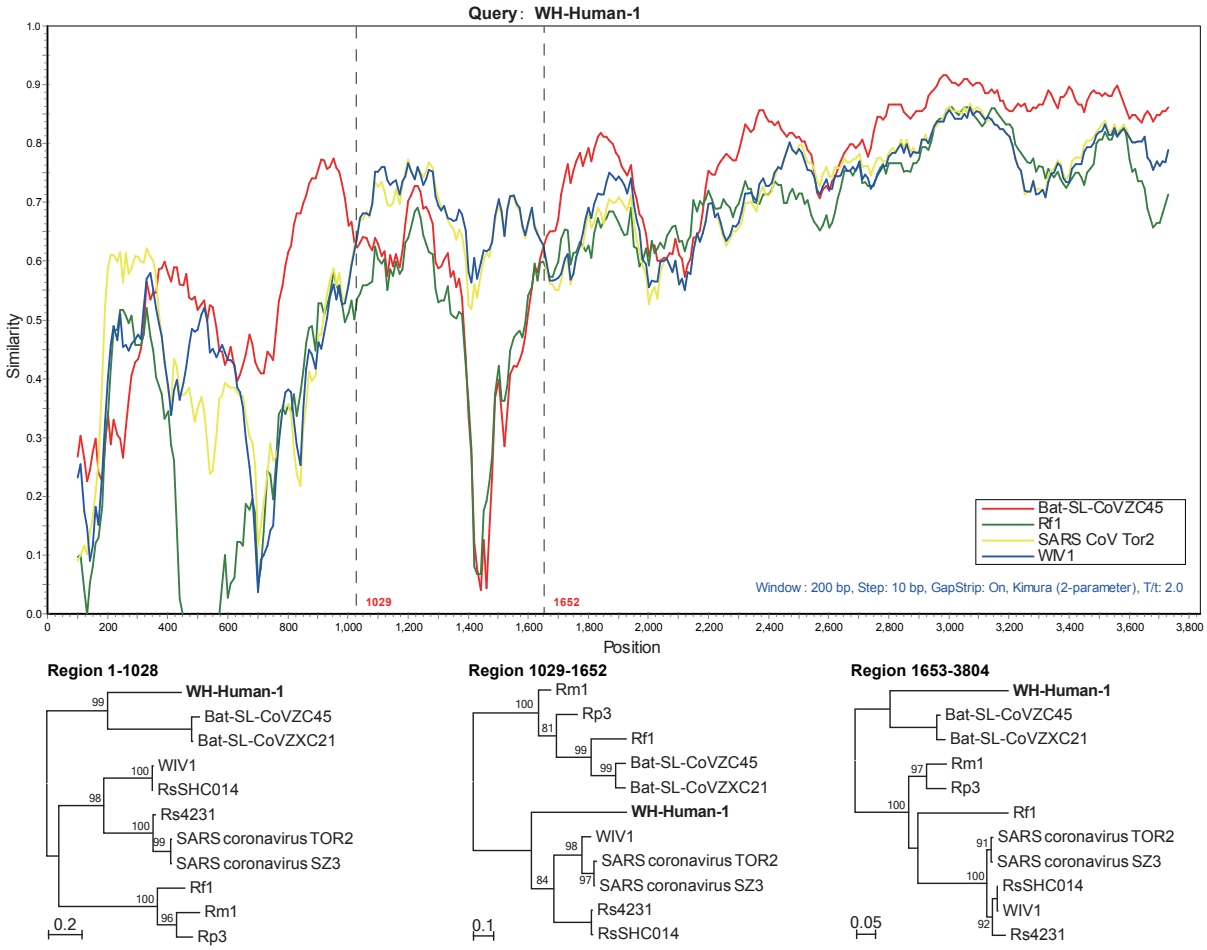


Figure 5. Possible recombination events in the S gene of sarbecoviruses. A sequence similarity plot (upper panel) reveals two putative recombination break-points shown by black dashed lines, with their locations indicated at the bottom. The plot shows S gene similarity comparisons of the WHCV (query) against SARS-CoV Tor2 and bat SARS-like CoVs WIV1, Rf1 and CoVZC45. Phylogenies of major parental region (1-1028 and 1653-3804) and minor parental region (1029-1652) are shown below the similarity plot. Phylogenies were estimated using a ML method and mid-point rooted for clarity only. Numbers above or below branches indicate percentage bootstrap values.