nature

https://doi.org/10.1038/s41586-020-2008-3

Accelerated Article Preview

A new coronavirus associated with human respiratory disease in China

Received: 7 January 2020

Accepted: 28 January 2020

Accelerated Article Preview
Published online 3 February 2020

Cite this article as: Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* https://doi.org/10.1038/s41586-020-2008-3 (2020).

Open access

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

Article

A new coronavirus associated with human respiratory disease in China

https://doi.org/10.1038/s41586-020-2008-3

Received: 7 January 2020

Accepted: 28 January 2020

Published online: 3 February 2020

Open access

Fan Wu[1,6], Su Zhao[2,6], Bin Yu[3,6], Yan-Mei Chen[1,6], Wen Wang[4,6], Zhi-Gang Song[1,6], Yi Hu[2,6], 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]*

Emerging infectious diseases, such as SARS and Zika, present a major threat to public health[1–3]. Despite intense research efforts, how, when and where new diseases appear are still the source of considerable uncertainly. A severe respiratory disease was recently reported in the city Wuhan, Hubei province, China. Up to 25th of January 2020, at least 1,975 cases have been reported since the first patient was hospitalized on the 12th of December 2019. Epidemiological investigation suggested that the outbreak was associated with a seafood market in Wuhan. We studied one patient who was a worker at the market, and who was admitted to Wuhan Central Hospital on 26th of December 2019 experiencing a severe respiratory syndrome including fever, dizziness and cough. Metagenomic RNA sequencing[4] of a bronchoalveolar lavage fluid sample identified a novel RNA virus from the family *Coronaviridae*, designed here as WH-Human-1 coronavirus. Phylogenetic analysis of the complete viral genome (29,903 nucleotides) revealed that the virus was most closely related (89.1% nucleotide similarity) to a group of SARS-like coronaviruses (genus Betacoronavirus, subgenus Sarbecovirus) previously sampled from bats in China. This outbreak highlights the ongoing capacity of viral spill-over from animals to cause severe disease in humans.

The patient studied was a 41-year-old man with no history of hepatitis, tuberculosis or diabetes. He was admitted and hospitalized in Wuhan Central Hospital on December 26, 2019, 6 days after the onset of illness. The patient reported fever, chest tightness, unproductive cough, pain and weakness for one week on presentation (Table 1). Physical examination of cardiovascular, abdominal and neurologic examination was normal. Mild lymphopenia (less than 900 cells per cubic millimeter) was observed, but white blood cell and blood platelet count was normal in a complete blood count (CBC) test. Elevated levels of the 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 airspace shadowing such as ground-glass opacities, focal consolidation and patchy consolidation in both lungs (Extended Data Fig. 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 (Extended Data Fig. 2). Although combination antibiotic, antiviral and glucocorticoid therapy were administered, the patient exhibited respiratory failure and was given high flow noninvasive ventilation. The condition of the patient did not improve after three days of treatment and he was admitted to the intensive care unit (ICU). The patient was transferred to another hospital in Wuhan for further treatment 6 days after admission.

Epidemiological investigation by the Wuhan Center of Disease Control and Prevention (CDC) revealed that the patient worked at 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 patient might have had contact with wild animals in the market, he recalled no exposure to live poultry.

To investigate the possible aetiologic agents associated this disease, we collected bronchoalveolar lavage fluid (BALF) and performed deep meta-transcriptomic sequencing. The clinical specimen was handled in a biosafety level 3 laboratory at the Shanghai Public Health Clinical Center. Total RNA was extracted from 200μl BAL fluid and a meta-transcriptomic library was constructed for pair-end (150 bp)

[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, Zhi-Gang Song, Yi Hu. *e-mail: zhangyongzhen@shphc.org.cn

sequencing using an Illumina MiniSeq as previously described[4–7]. In total, we generated 56,565,928 sequence reads that were *de novo* assembled and screened for potential aetiologic agents. Of the 384,096 contigs assembled by Megahit[8], the longest (30,474 nucleotides [nt]) had high abundance and was closely related to a bat SARS-like coronavirus isolate - bat-SL-CoVZC45 (GenBank Accession MG772933) - previously sampled in China, with a nt identity of 89.1% (Supplementary Tables 1 and 2). The genome sequence of this novel virus, as well as its termini, were determined and confirmed by RT-PCR[9] and 5'/3' RACE kits (TaKaRa), respectively. This new virus was designated as WH-Human 1 coronavirus (WHCV) (and has also been referred to as '2019-nCoV') and its whole genome sequence (29,903 nt) has been assigned GenBank accession number MN908947. Remapping the RNA-seq data against the complete genome of WHCV resulted in an assembly of 123,613 reads, providing 99.99% genome coverage at a mean depth of 6.04X (range: 0.01X -78.84X) (Extended Data Fig. 3). The viral load in the BALF sample was estimated by quantitative PCR (qPCR) to be 3.95×10^8 copies/mL (Extended Data Fig. 4).

The viral genome organization of WHCV was characterized by sequence alignment against two representative members of the genus *Betacoronavirus*: a human-origin coronavirus (SARS-CoV Tor2, AY274119) and a bat-origin coronavirus (Bat-SL-CoVZC45, MG772933). The un-translational regions (UTR) and open reading frame (ORF) of WHCV were mapped based on this sequence alignment and ORF prediction. The WHCV viral genome was similar to these two coronaviruses (Fig. 1 and Supplementary Table 3), with a gene order 5'-replicase ORF1ab-S-envelope(E)-membrane(M)-N-3'. WHCV has 5' and 3' terminal 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 (Supplementary Table 4), 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 Supplementary Table 5. 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 (Supplementary Table 6).

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 (Fig. 2 and Extended Data Fig. 5). 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 (Fig. 2 and Extended Data Fig. 5). Specifically, in the S gene tree (Extended Data Fig. 5), WHCV was most closely related to the bat coronavirus bat-SL-CoVZC45 with 82.3% amino acid (aa) identity (and ~77.2% aa identity to SARS CoV; Supplementary Table 3), while in the ORF1b phylogeny WHCV fell in a basal position within the subgenus *Sarbecovirus* (Fig. 2). This topological division, likely reflecting recombination among the bat sarbecoviruses, was also observed in the phylogenetic trees estimated for conserved domains in the replicase polyprotein pp1ab (Extended Data Fig. 6).

To better understand the potential of WHCV to infect humans, the receptor-binding domain (RBD) of its spike protein was compared to those in SARS-CoVs and bat SARS-like CoVs. The RBD sequences of WHCV were more closely related to those of SARS-CoVs (73.8%-74.9% aa identity) and SARS-like CoVs including strains Rs4874, Rs7327 and Rs4231 (75.9%-76.9% aa identity) that are able to use the human ACE2 receptor for cell entry (Supplementary Table 7)[10]. In addition, the WHCV RBD was only one amino acid longer than the SARS-CoV RBD (Extended Data Fig. 7a). In contrast, other bat SARS-like CoVs including the Rp3 strain that cannot use human ACE2[11], had amino acid deletions at positions 473-477 and 460-472 compared to the SARS-CoVs (Extended Data Fig. 7a). The previously determined[12] crystal structure of SARS-CoV RBD complexed with human ACE2 (PDB 2AJF) revealed that regions 473-477 and 460-472 directly interact with human ACE2 and hence may be important in determining species specificity (Extended Data Fig. 7b). We predicted the three-dimension protein structures of WHCV, Rs4874 and Rp3 RBD domains by protein homology modelling using the SWISS-MODEL server and compared them to the crystal structure of SARS-CoV RBD domains (PDB 2GHV) (Extended Data Fig. 7c-f). In accord with the sequence alignment, the predicted protein structures of WHCV and Rs4874 RBD domains were closely related to that of SARS-CoVs and different from the predicted structure of the RBD domain from Rp3. In addition, the N-terminus of WHCV S protein is more similar to that of SARS-CoV rather than other human coronaviruses (HKU1 and OC43) (Extended Data Fig. 8) that can bind to sialic acid[13]. In sum, the high similarities of amino acid sequences and predicted protein structure between WHCV and SARS-CoV RBD domains suggest that WHCV may efficiently use human ACE2 as a cellular entry receptor, potentially facilitating human-to-human transmission[10,14–15].

To further characterize putative recombination events in the evolutionary history of the sarbecoviruses the whole genome sequence of WHCV and four representative coronaviruses - Bat SARS-like CoV Rp3, CoVZC45, CoVZXC21 and SARS-CoV Tor2 - were analysed using the Recombination Detection Program v4 (RDP4)[16]. Although the similarity plots suggested possible recombination events between WHCV and SARS CoVs or SARS-like CoVs (Extended Data Fig. 9), there was no significant evidence for recombination across the genome as a whole. However, some evidence for past recombination was detected in the S gene of WHCV and SARS CoV and bat SARS-like CoVs (WIV1 and RsSHC014) ($p<3.147\times10^{-3}$ to $p<9.198\times10^{-9}$), with similarity plots suggesting the presence of recombination break points at nucleotides 1,029 and 1,652 that separated the WHCV S gene into three regions (Fig. 3). In phylogenies of the fragment nt 1 to 1029 and nt 1652 to the end of the sequence, WHCV was most closely related to Bat-SL-CoVZC45 and Bat-SL-CoVZXC21, whereas in the region nt 1030 to 1651 (the RBD region) WHCV grouped with SARS CoV and bat SARS-like CoVs (WIV1 and RsSHC014) that are capable of direct human transmission[14,17]. Despite these recombination events, which seem relatively common among the sarbecoviruses, there is no evidence that recombination has facilitated the emergence of WHCV.

Coronaviruses are associated with a number of infectious disease outbreaks in humans, including SARS in 2002/3 and MERS in 2012[1,18]. Four other coronaviruses - human coronaviruses HKU1, OC43, NL63 and 229E - are also associated with respiratory disease[19–22]. Although SARS-like coronaviruses have been widely identified in mammals including bats since 2005 in China[9,23–25], the exact origin of human-infected coronaviruses remains unclear. Herein, we describe a novel coronavirus - WHCV (2019-nCoV) - in BALF from a patient experiencing severe respiratory disease in Wuhan, China. Phylogenetic analysis suggested that WHCV represents a novel virus within genus *Betacoronavirus* (subgenus *Sarbecovirus*) and hence that exhibits some genomic and phylogenetic similarity to SARS-CoV[1], particularly in the RBD. These genomic and clinical similarities to SARS, as well as its high abundance in clinical samples, provides evidence for an association between WHCV and the ongoing outbreak of respiratory disease in Wuhan. Although that

the isolation of the virus just from a single patient is not sufficient to conclude that it caused the respiratory symptoms, our findings have been independently corroborated in further patients[26].

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

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2008-3.

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).
11. Ren, W. et al. Difference in receptor usage between severe acute respiratory syndrome (SARS) coronavirus and SARS-like coronavirus of bat origin. *J Virol*. **82**:1899-1907 (2008).
12. Li, F., Li, W., Farzan, M., Harrison, S.C. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science*. **309**, 1864-1868 (2005).
13. Hulswit, R.J.G., et al. Human coronaviruses OC43 and HKU1 bind to 9-O-acetylated sialic acids via a conserved receptor-binding site in spike protein domain A. *Proc Natl Acad Sci USA*., **116**, 2681-2690 (2019).
14. Ge, X.Y. et al. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature*. **503**: 535-538 (2013).
15. Yang, X.L., et al. Isolation and characterization of a novel bat coronavirus closely related to the direct progenitor of severe acute respiratory syndrome coronavirus. *J Virol*. **90**: 3253-3256 (2016).
16. Martin, D.P., Lemey, P., Lott, M., Moulton, V., Posada, D., Lefeuvre, P. RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**:2462–2463 (2010).
17. Menachery, V.D., et al. A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat Med*. **21**:1508-1513 (2015).
18. Bermingham, A., et al. Severe respiratory illness caused by a novel coronavirus, in a patient transferred to the United Kingdom from the Middle East, September 2012. *Euro. Surveill*. **17**, 20290 (2012).
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).
26. Zhou, P., et al. Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin. https://www.biorxiv.org/content/10.1101/2020.01.22.914952v1. (2020)

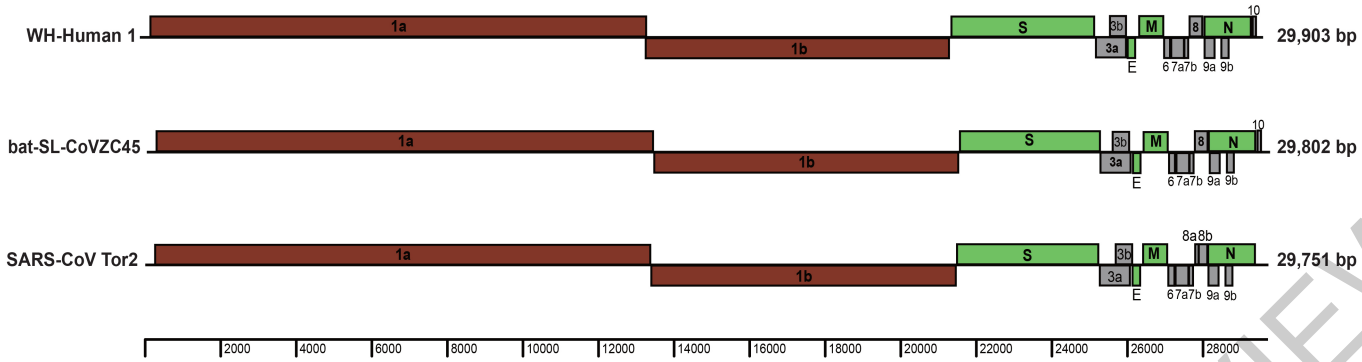


Fig. 1 | Genome organization of SARS and SARS-like CoVs including Tor2, CoVZC45 and WHCV determined here.

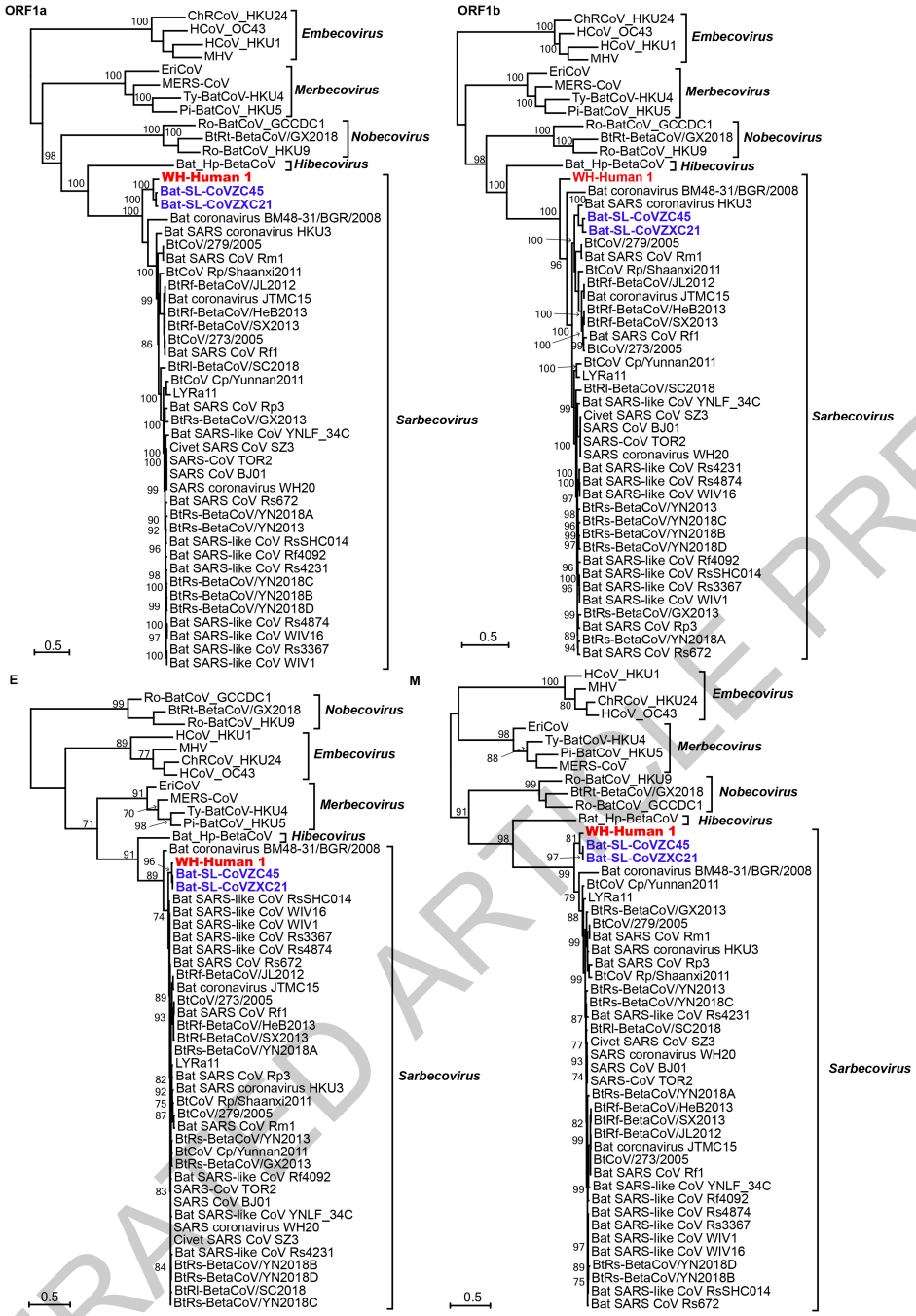


Fig. 2 | Maximum likelihood phylogenetic trees of nucleotide sequences of the ORF1a, ORF1b, E and M genes of WHCV and related coronaviruses. Numbers (>70) above or below branches indicate percentage bootstrap values for the associated nodes. The trees were mid-point rooted for clarity only. The scale bar represents the number of substitutions per site.

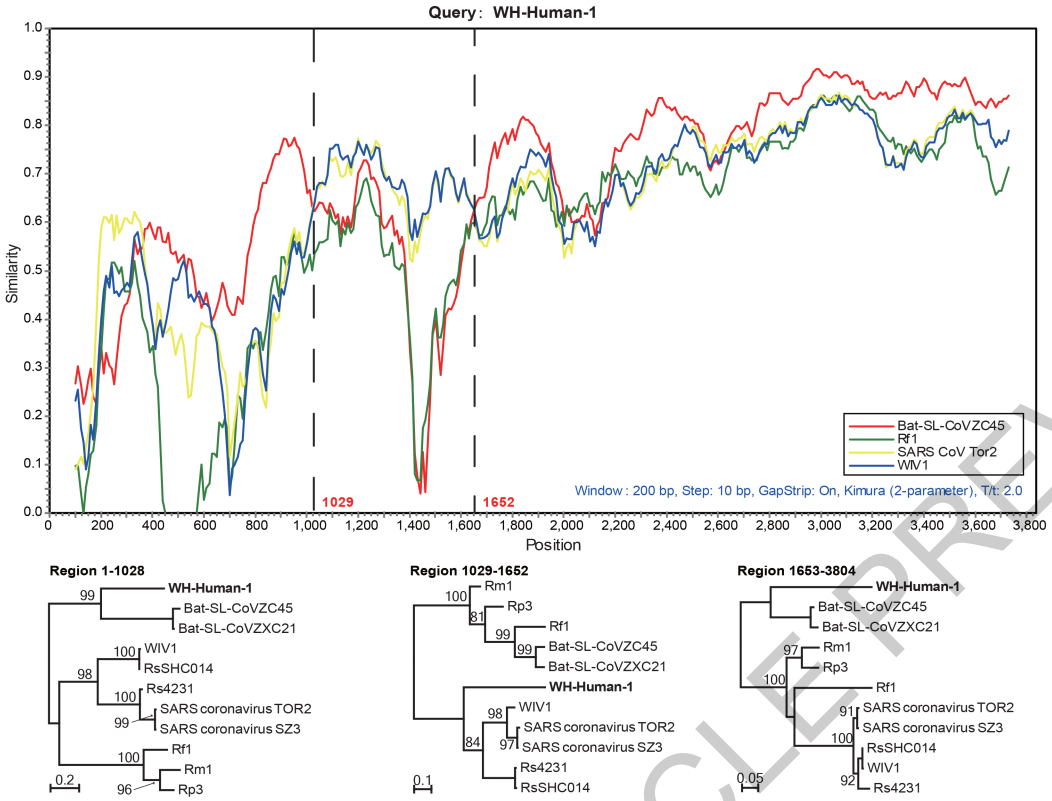


Fig. 3 | Possible recombination events in the S gene of sarbecoviruses. A sequence similarity plot (upper panel) reveals two putative recombination break-points shown by black dashed lines, with their locations indicated at the bottom. The plot shows S gene similarity comparisons of the WHCV (query) 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.

Table 1 | Clinical symptoms and patient data

Characteristic	Patient
Age (Year)	41
Sex	M
Date of illness onset	Dec 20,2019
Date of admission	Dec 26,2019
Signs and symptoms	
Fever	+
Body Temperature (°C)	38.4
Cough	+
Sputum Production	+
Dizzy	+
Weakness	+
Chest tightness	+
Dyspnea	+
Bacterial culture	-
Glucocorticoid therapy	No
Antibiotic therapy	Cefoselis
Antiviral therapy	Oseltamivir
Oxygen therapy	mechanical ventilation

Methods

Cases and collection of clinical data and samples

A patient presenting with acute onset of fever (>37.5 °C), cough, and chest tightness, and who were admitted to Wuhan Central Hospital in Wuhan city, China, was considered as a suspected case. During admission, bronchoalveolar lavage fluid (BALF) was collected and stored at -80 °C until further processing. Demographic, clinical and laboratory data was retrieved from the clinical records of the patient. The study was reviewed and approved by the ethics committee of the National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention (CDC). A signed written informed consent was obtained from the patient.

RNA library construction and sequencing

Total RNA was extracted from the BALF sample of the patient using the RNeasy Plus Universal Mini Kit (Qiagen) following the manufacturer's instructions. The quantity and quality of the RNA solution was assessed using a Qbit machine and an Agilent 2100 Bioanalyzer (Agilent Technologies) before library construction and sequencing. An RNA library was then constructed using the SMARTer Stranded Total RNA-Seq Kit v2 (TaKaRa, Dalian, China). Ribosomal RNA (rRNA) depletion was performed during library construction following the manufacturer's instructions. Paired-end (150 bp) sequencing of the RNA library was performed on the MiniSeq platform (Illumina). Library preparation and sequencing were carried out at the Shanghai Public Health Clinical Center, Fudan University, Shanghai, China.

Data processing and viral agent identification

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

As the longest contigs generated by Megahit (30,474 nt) and Trinity (11,760 nt) both had high similarity to the bat SARS-like coronavirus isolate bat-SL-CoVZC45 and were at high abundance (Supplementary Tables 1 and 2), the longer one (30,474 nt) that covered almost the whole virus genome was used for primer design for PCR confirmation and genome termini determination. Primers used in PCR, qPCR and RACE experiments are listed in Supplementary Table 8. The PCR assay was conducted as described previously[9] and the complete genome termini was determined using the Takara SMARTer RACE 5'/3' kit (TaKaRa) following the manufacturer's instructions. Subsequently, the genome coverage and sequencing depth were determined by remapping all of the adaptor- and quality-trimmed reads to the whole genome of WHCV using Bowtie2[30] and Samtools[31].

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

Virus genome characterization and phylogenetic analysis

For the newly identified virus genome, the potential open reading frames (ORFs) were predicted and annotated using the conserved signatures of the cleavage sites recognized by coronavirus proteinases, and were processed in the Lasergene software package (version 7.1, DNAstar). The viral genes were aligned using the L-INS-i algorithm implemented in MAFFT (version 7.407)[32].

Phylogenetic analyses were then performed using the nucleotide sequences of various CoV gene data sets: (i) Whole genome, (ii) ORF1a, (iii) ORF1b, (iv) nsp5 (3CLpro), (v) RdRp (nsp12), (vi) nsp13 (Hel), (vii) nsp14 (ExoN), (viii) nsp15 (NendoU), (ix) nsp16 (O-MT), (x) spike (S), and the (xi) nucleocapsid (N). Phylogenetic trees were inferred using the Maximum likelihood (ML) method implemented in the PhyML program (version 3.0)[33], using the Generalised Time Reversible substitution (GTR) model and Subtree Pruning and Regrafting (SPR) branch-swapping. Bootstrap support values were calculated from 1,000 pseudo-replicate trees. The best-fit model of nucleotide substitution was determined using MEGA (version 5)[34]. Amino acid identities among sequences were calculated using the MegAlign program implemented in the Lasergene software package (version 7.1, DNAstar).

Genome recombination analysis

Potential recombination events in the history of the sarbecoviruses were assessed using both the Recombination Detection Program v4 (RDP4)[16] and Simplot (version 3.5.1)[35]. The RDP4 analysis was conducted based on the complete genome (nucleotide) sequence, employing the RDP, GENECONV, BootScan, maximum chi square, Chimera, SISCAN, and 3SEQ methods. Putative recombination events were identified with a Bonferroni corrected p-value cut-off of 0.01. Similarity plots were inferred using Simplot to further characterize potential recombination events, including the location of possible breakpoints.

Analysis of RBD domain of WHCV spike protein

An amino acid sequence alignment of WHCV, SARS-CoVs, bat SARS-like CoVs RBD sequences was performed using MUSCLE[36]. The predicted protein structures of the spike protein RBD were estimated based on target-template alignment using ProMod3 on SWISS-MODEL server (https://swissmodel.expasy.org/). The sequences of the spike RBD domains of WHCV, Rs4874 and Rp3 were searched by BLAST against the primary amino acid sequence contained in the SWISS-MODEL template library (SMTL, last update: 2020-01-09, last included PDB release: 2020-01-03). Models were built based on the target-template alignment using ProMod3. The global and per-residue model quality were assessed using the QMEAN scoring function[37]. The PDB files of the predicted protein structures were displayed and compared with the crystal structures of SARS-CoV spike RBD (PDB 2GHV)[38] and the crystal of structure of SARS-CoV spike RBD complexed with human ACE2 (PDB 2AJF)[12].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Sequence reads generated in this study are available at the NCBI Sequence Read Archive (SRA) database under the BioProject accession PRJNA603194. The complete genome sequence of WHCV have been deposited in GenBank under the accession numbers MN908947.

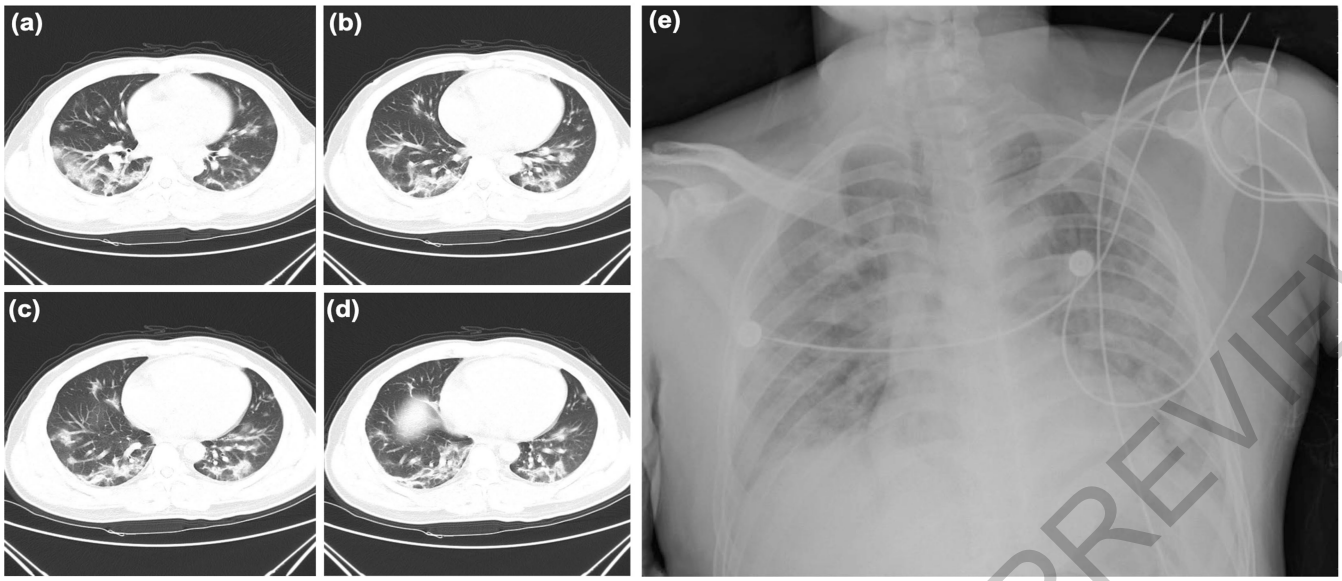
27. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
28. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
29. Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A. & Dewey, C.N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493-500 (2010).
30. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. **9**, 357–359 (2012).
31. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **15**, 2978-2079 (2009).
32. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).
33. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst.Biol*. **59**, 307-321 (2010).
34. Tamura, K. et al. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).
35. Lole, K.S. et al. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* **73**, 152–160 (1999).
36. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. **32**, 1792-1797 (2004).
37. Waterhouse, A. et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. **46**, W296-W303 (2018).
38. Hwang, W.C. et al. Structural basis of neutralization by a human anti-severe acute respiratory syndrome spike protein antibody, 80R. *J Biol Chem*. **281**, 34610-34616 (2006).

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

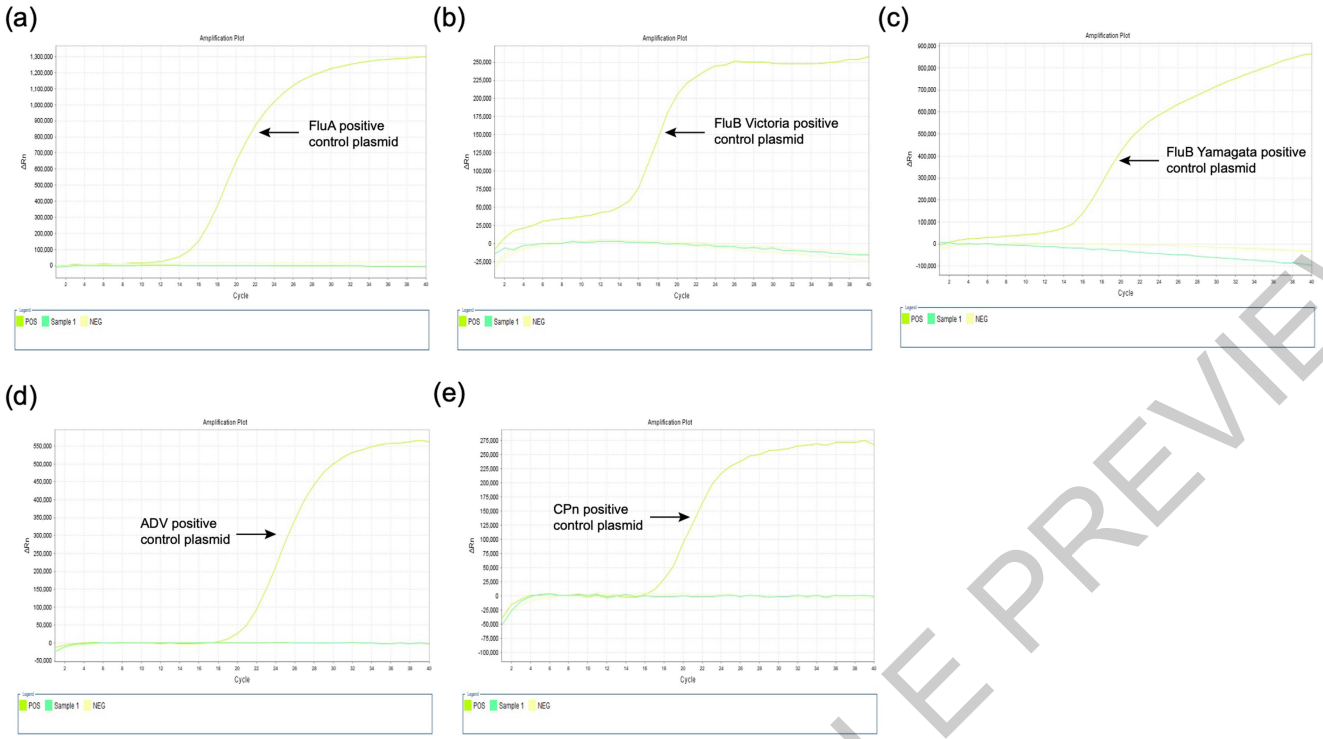
Author contributions Y.-Z.Z. conceived and designed the study. S.Z, Y.H, Z.-W.T. and M.-L.Y. performed the clinical work and sample collection. B.Y and J.-H.T. performed epidemiological investigation and sample collection. F.W, Z.-G.S., L.X., Y.-Y.P., Y.-L.Z., F.-H.D., Y.L., J.-J.Z. and Q.-M.W. performed 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.

Competing interests The authors declare no competing interests.

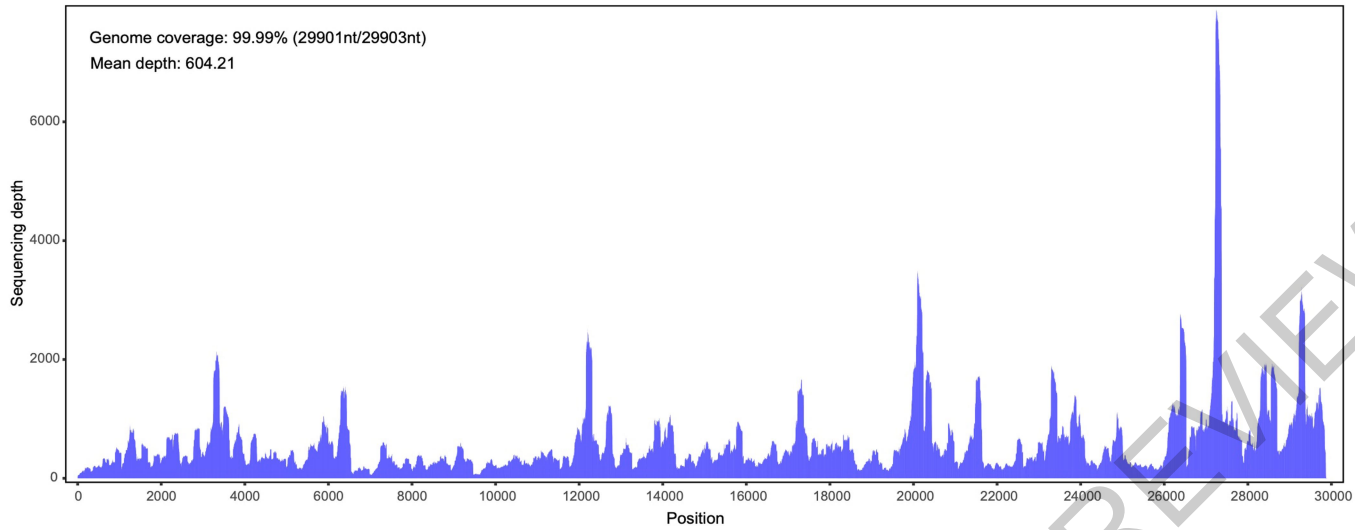
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2008-3.
Correspondence and requests for materials should be addressed to Y.-Z.Z.
Peer review information *Nature* thanks Nicholas Loman and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



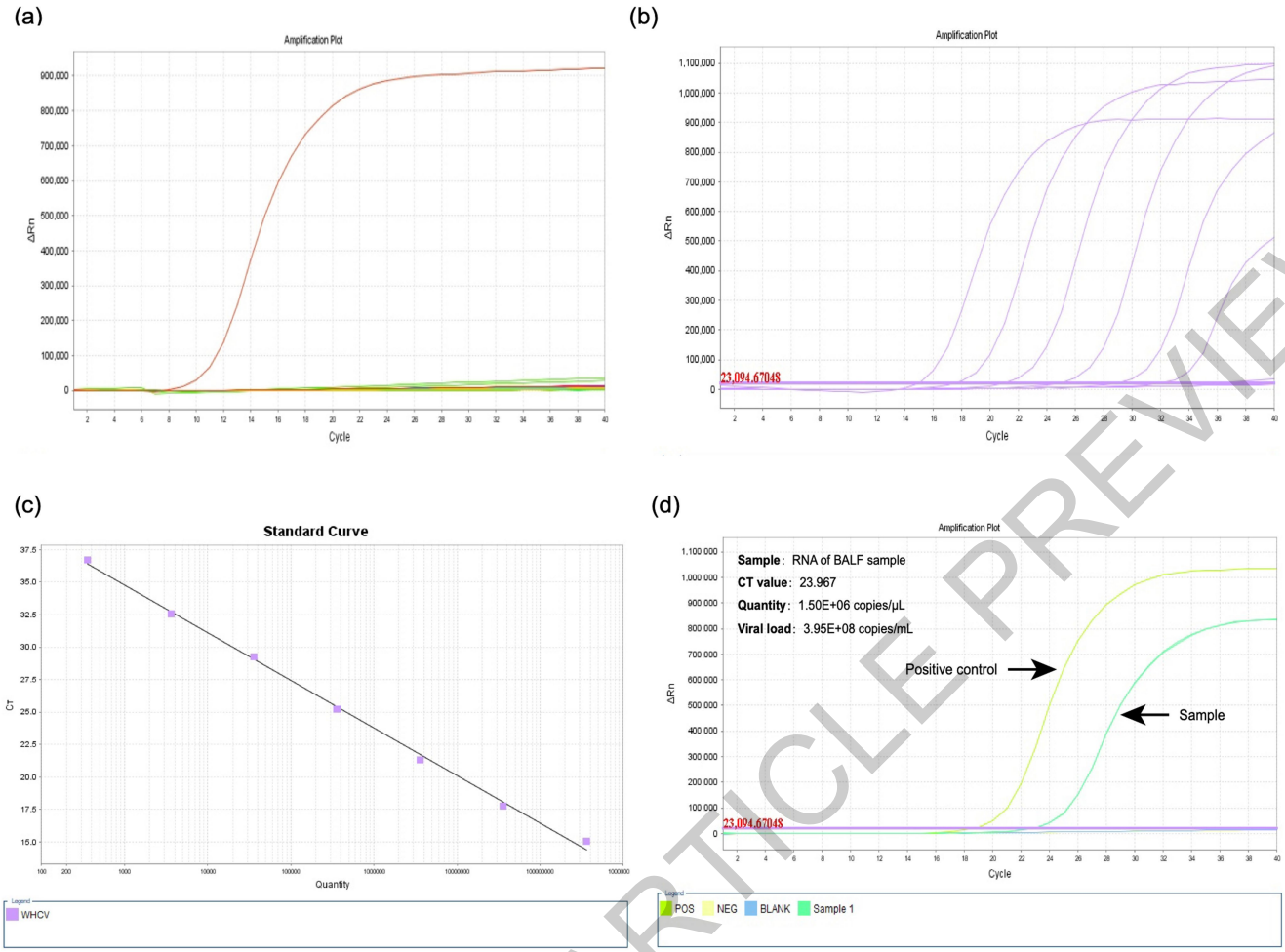
Extended Data Fig. 1 | Chest radiographs of the patient. (**a**–**d**), Chest computed tomographic scans were obtained on the day of admission (day 6 after the onset of illness). Bilateral focal consolidation, lobar consolidation, and patchy consolidation were clearly observed, especially in the lower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



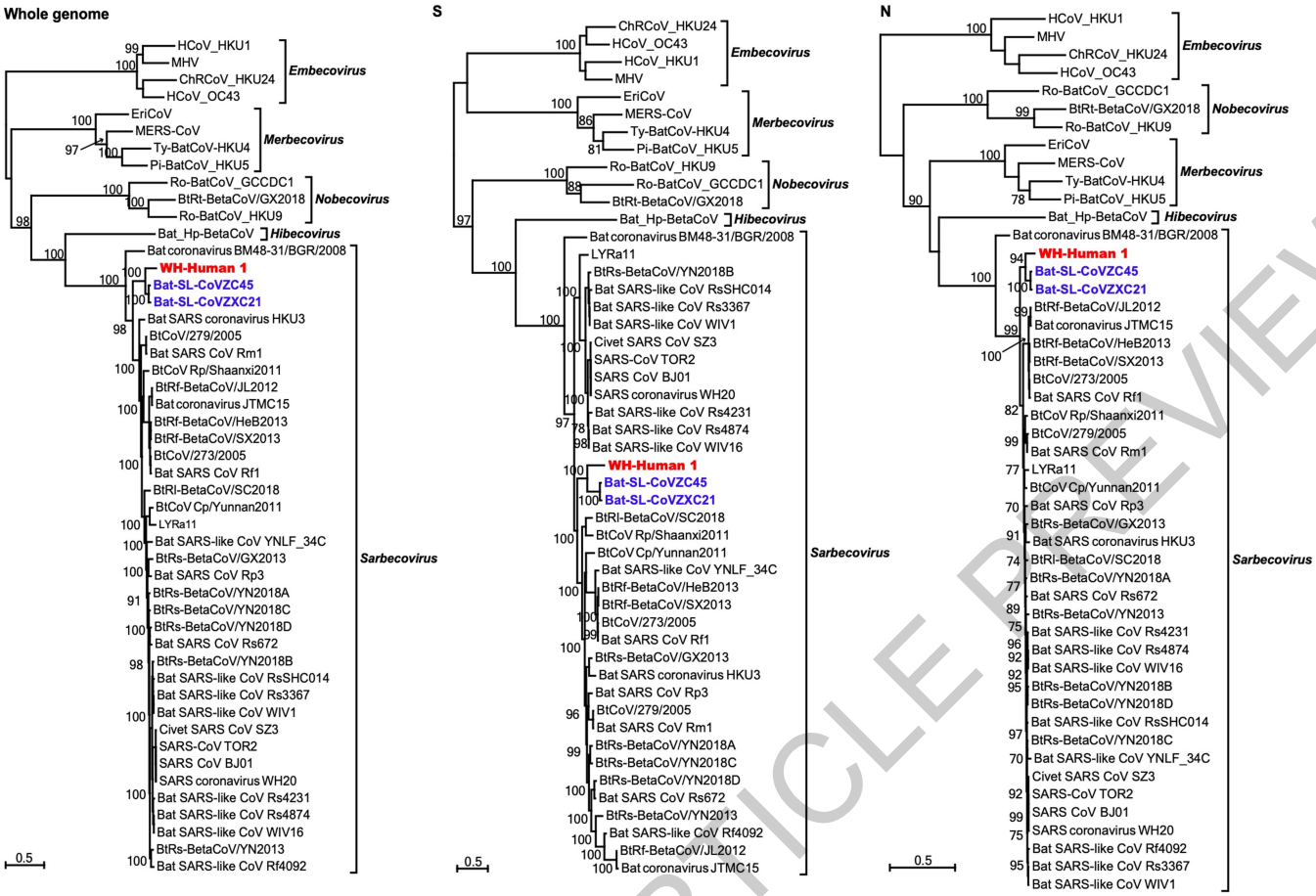
Extended Data Fig. 2 | Detection of other respiratory pathogens in patient in BALF sample by real-time RT-PCR. The presence of influenza A virus **(a)**, influenza B viruses victoria lineage **(b)**, influenza B virus yamagata lineage **(c)**, human adenovirus **(d)** and chlamydia pneumoniae **(e)** was test. Sample 1 was the patient's BALF sample, water was used as negative control (NEG), POS samples were plasmids covering the Taqman primers and probe regions of influenza A, influenza B Victoria lineage, influenza B virus yamagata lineage, human adenovirus and chlamydia pneumoniae as positive controls respectively.



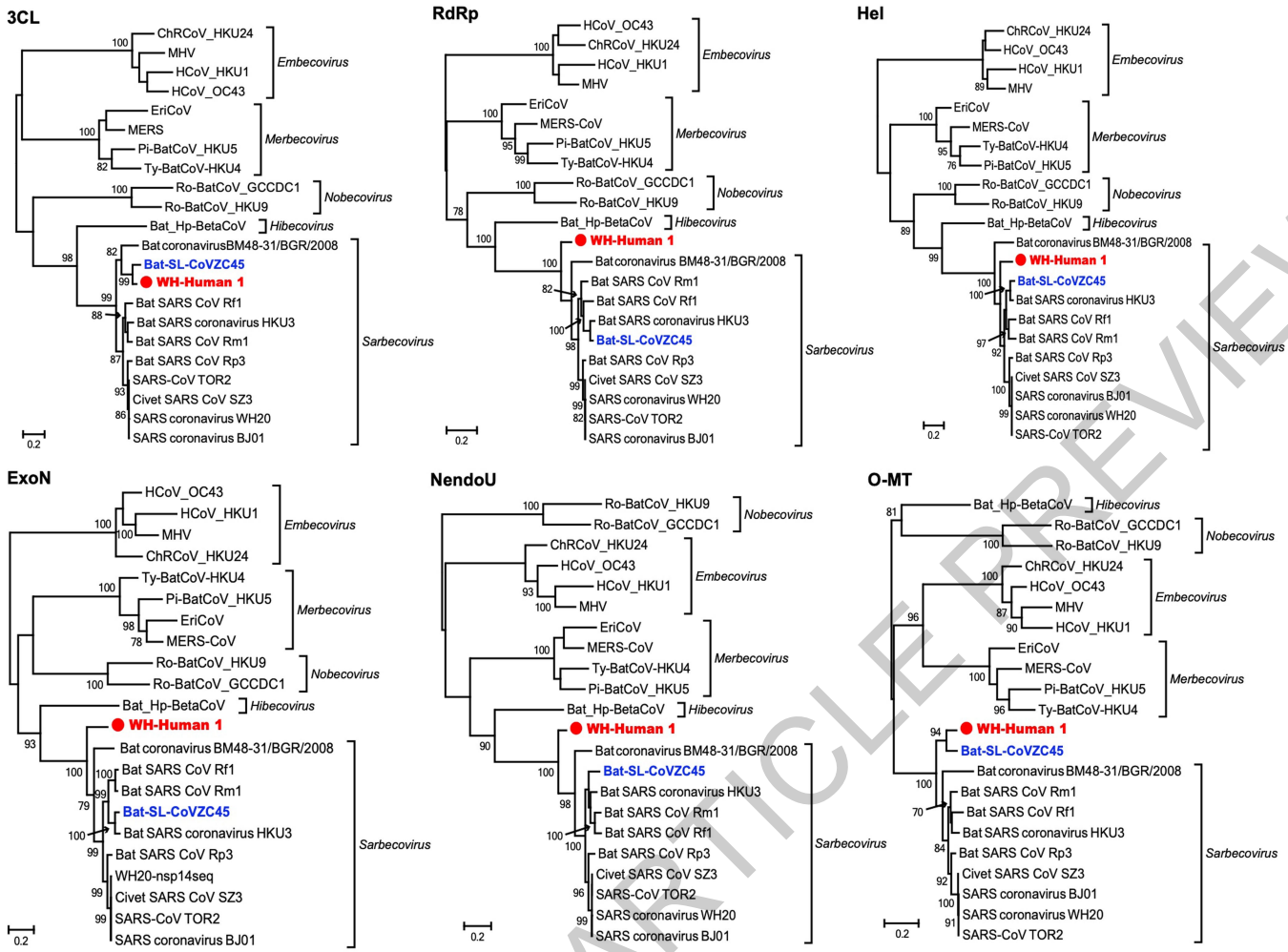
Extended Data Fig. 3 | Mapped read count plot showing the coverage depth per base of the WHCV genome.



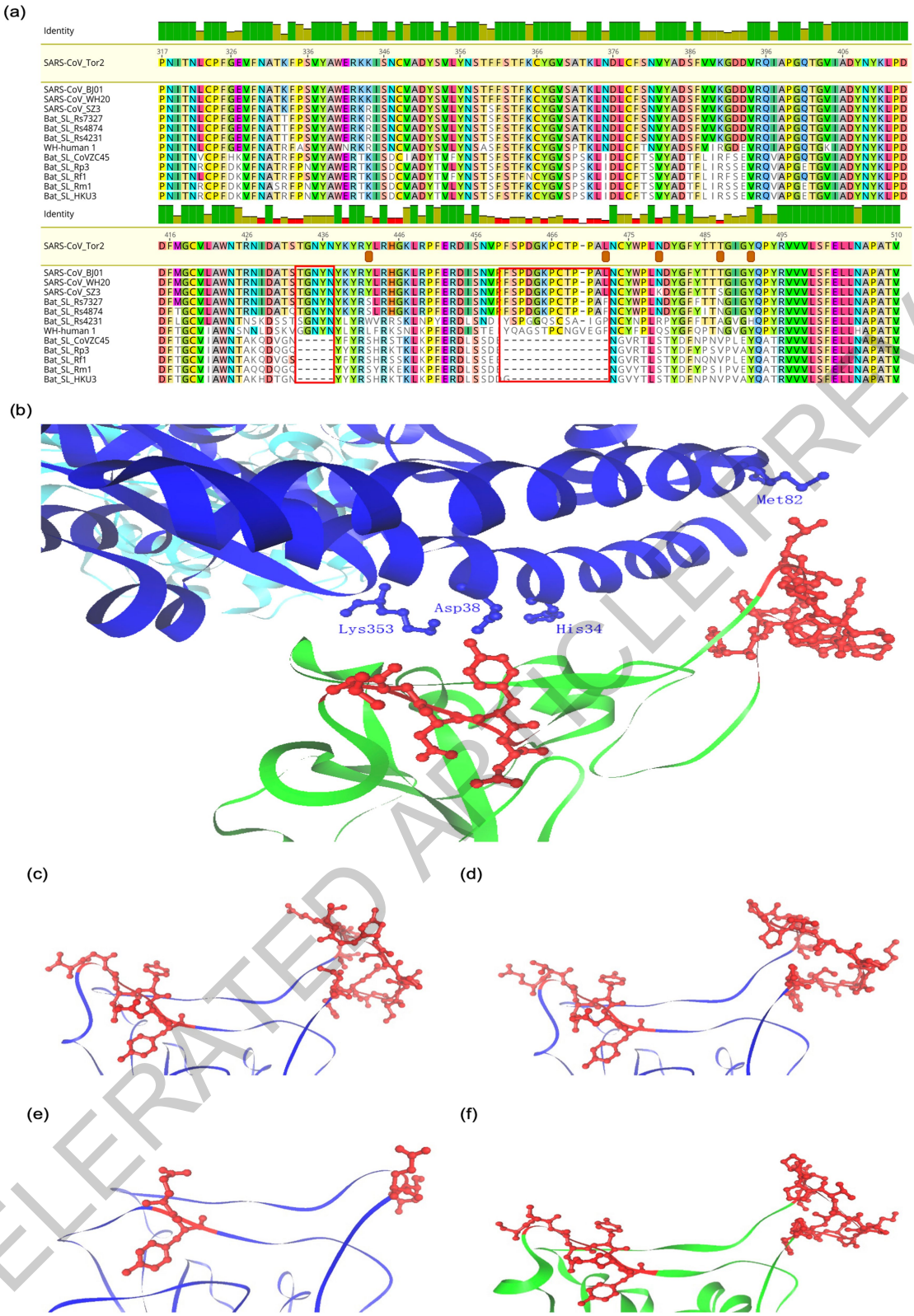
Extended Data Fig. 4 | Quantification of WHCV in clinical samples by real-time RT-PCR. (a) Specificity evaluation of the WHCV primers. Test samples comprised clinical samples that are positive for at least one of the following viruses: Influenza A virus (09H1N1 and H3N2), Influenza B virus, Human adenovirus, Respiratory syncytial virus, Rhinovirus, Parainfluenza virus type 1-4, Human bocavirus, Human metapneumovirus, Coronavirus OC43, Coronavirus NL63, Coronavirus 229E and Coronavirus HKU1. Only WHCV standard plasmid (WHCV 15704bp-16846 bp in pLB vector) generated positive amplification (brown curve). **(b)** Amplification curve of WHCV standard DNA. From left to right, the DNA concentrations were 1.8×10^8, 1.8×10^7, 1.8×10^6, 1.8×10^5, 1.8×10^4, 1.8×10^3, respectively. **(c)** Linear fit curve of CT values to WHCV standard DNA concentrations. **(d)** Quantification of WHCV in patient's BALF sample by real-time RT-PCR. WHCV standard DNA was used as positive control (POS), water (NEG) and blank were used as negative controls. The amplification curve of BALF sample was in green.



Extended Data Fig. 5 | Maximum likelihood phylogenetic trees of the nucleotide sequences of the whole genome, S and N genes of WHCV and related coronaviruses. Numbers (>70) above or below branches indicate percentage bootstrap values. The trees were mid-point rooted for clarity only. The scale bar represents the number of substitutions per site.



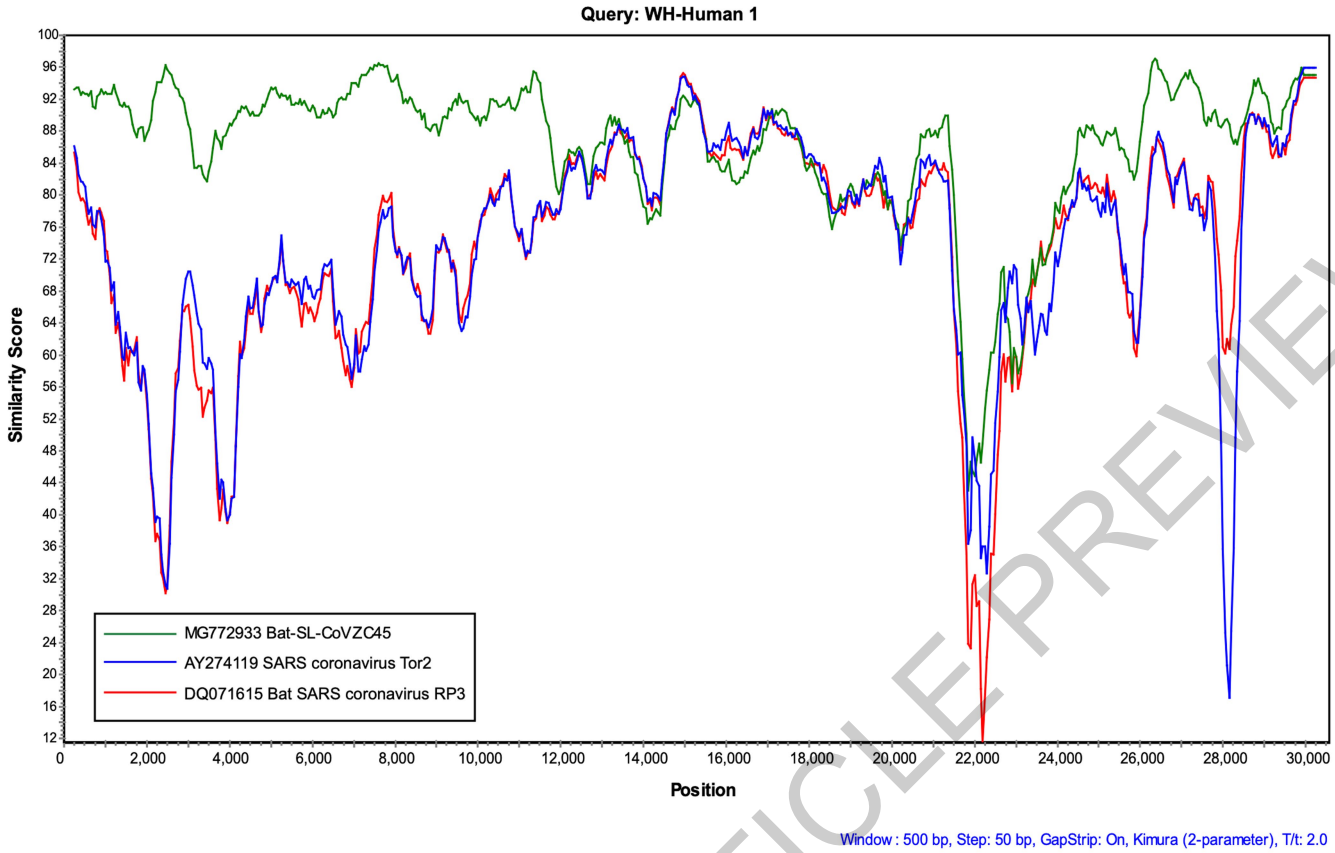
Extended Data Fig. 6 | Maximum likelihood phylogenetic trees of the nucleotide sequences of the 3CL, RdRp, Hel, ExoN, NendoU, and O-MT genes of WHCV and related coronaviruses. Numbers (>70) above or below branches indicate percentage bootstrap values. The trees were mid-point rooted for clarity only. The scale bar represents the number of substitutions per site.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Analysis of receptor-binding domain (RBD) of the spike (S) protein of WHCV coronavirus. (a) Amino acid sequence alignment of SARS-like CoV RBD sequences. Three bat SARS-like CoVs, which could efficiently utilize the human ACE2 as receptor, had an RBD sequence of similar size to SARS-CoV, and WHCV contains a single Val 470 insertion. 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 that had been reported not to use ACE2, had amino acid deletions at two motifs (amino acids 473-477 and 460-472) compared with those of SARS-CoV.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. **(d)** Predicted structure of RBD of SARS-like CoV Rs4874. **(e)** Predicted structure of the RBD of SARS-like CoV Rp3. **(f)** Crystal structure of RBD of SARS-CoV spike protein (green) (PDB 2GHV). Motifs resembling amino acids 473-477 and 460-472 of the SARS-CoV spike protein are shown in red.

Extended Data Fig. 8 | Amino acid sequence comparison of the N-terminal domain (NTD) of spike protein of WHCV, bovine coronavirus (BCoV), mouse hepatitis virus (MHV) and human coronavirus (HCoV OC43 and HKU1) that can bind to sialic acid and the SARS-CoVs that cannot. The key residues[13] for sialic acid binding on BCoV, MHV, HCoV OC43 and HKU1 were marked with a brown box.



Extended Data Fig. 9 | A sequence similarity plot of WHCV, SARS- and bat SARS-like CoVs revealing putative recombination events.

nature research

Corresponding author(s): Yong-Zhen Zhang

Last updated by author(s): Jan 25, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	Trimmotic (v0.39): adaptor- and quality-trimming of sequencing reads Megahit (v1.1.3) and Trinity (v2.5.1): de novo assembly of reads Blastn (v2.7.1), Diamond blastx (v0.9.21): homology based annotation of sequencing reads and contigs Bowtie2 (v2.3.4.1) and samtools (v 0.1.19-44428cd): read mapping and result analysis MAFFT (v7.407) and MUSCLE(v3.8.425): sequence alignment PhyML (v3.0): Phylogenetic tree estimation MEGA (v5): Best-fit model of nucleotide substitution determination and trees generation Lasergene software package (v7.1): ORF prediciion and annotation Geneious prime (v2019): Visualization of alignment Recombination Detection Program (v4, RDP4) and Simplot (v3.5.1): recombination analysis and similarity plot visualization SWISS-MODEL server (https://swissmodel.expasy.org/): spike protein RBD structure prediction.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The whole genome sequence obtained in this study was submitted to GenBank with the accession number MN908947.
Fig. 1-3, Extended Data Fig. 3, Extended Data Fig. 5-9 have associated raw data.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The goal of this study was to find out the possible aetiologic agents associated with the severe respiratory disease occurred recently in the city of Wuhan, Hubei province, China. We studied one patient, and collected bronchoalveolar lavage fluid (BALF) from him who exhibited severe respiratory syndrome including fever, dizzy and cough. Since it is a discovery study, the number of individuals is irrelevant to the conclusions drawn in the paper.
Data exclusions	No data was excluded from the analyses.
Replication	The de novo assembly of reads was performed using two programs. The whole genome viral sequence obtained from read assembly was confirmed by PCR assays. The results from phylogenetic and recombination analyses were confirmed by multiple runs.
Randomization	Not applicable. The goal of this study was to find out the possible aetiologic agent associated with the severe respiratory disease occurred recently in the city of Wuhan, Hubei province, China. Since we could obtain the BALF sample from only one patient who exhibited severe respiratory syndrome including fever, dizzy and cough, hence, randomization was not applicable to this study.
Blinding	Not applicable. Only one RNA library was generated in this study and thus no group allocation was performed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Recently, a severe respiratory disease emerged in the city of Wuhan, Hubei province of China. The aim of this study is to find out the etiologic agent. Although clinic records from seven patients were available in this study, BAFL sample was only obtained from one patient. Herein, only one patient was described in the text based on the comments by Referees.
Recruitment	The patient who exhibited clinic signs of respiratory disease including fever and cough was recruited.

Ethics oversight	This study was reviewed and approved by the ethics committees of the National Institute for Communicable Disease Control and Prevention of the China CDC. In addition, a signed individual written informed consent was obtained from the patient.

Note that full information on the approval of the study protocol must also be provided in the manuscript.