Cell, Volume *184*

Supplemental information

The origins of SARS-CoV-2: A critical review

Edward C. Holmes, Stephen A. Goldstein, Angela L. Rasmussen, David L. Robertson, Alexander Crits-Christoph, Joel O. Wertheim, Simon J. Anthony, Wendy S. Barclay, Maciej F. Boni, Peter C. Doherty, Jeremy Farrar, Jemma L. Geoghegan, Xiaowei Jiang, Julian L. Leibowitz, Stuart J.D. Neil, Tim Skern, Susan R. Weiss, Michael Worobey, Kristian G. Andersen, Robert F. Garry, and Andrew Rambaut

Supplementary Information

Supplementary information to main text Figure 1.

Phylogenetic tree. SARS-CoV-2 genome sequences were downloaded from the GISAID EpiCoV database (http://gisaid.org). All complete and high coverage genomes from Wuhan, China with collection dates from December 2019 to January 2020 were downloaded. Genomes were pairwise aligned to the reference genome, 'Wuhan/Hu-1/2019' (NCBI accession MN908947) using Minimap2 (Li, 2018) and the 5' and 3' untranslated regions were masked to avoid areas of low sequencing coverage. A maximum likelihood phylogenetic tree was estimated using IQ-TREE2 (Minh et al., 2021) under the Jukes-Cantor model of nucleotide substitution. The tree was rooted at the midpoint between lineage A and lineage B.

Three genomes from late January 2020 were removed ('Wuhan/0126-C94/2020', 'Wuhan/0126-C100/2020', 'Wuhan/0126-C93/2020' – GISAID accessions EPI_ISL_493180, EPI_ISL_493182, EPI_ISL_493179, respectively) because although they had the mutation 8782T indicative of lineage A, they did not have the corresponding 28144C mutation. One of these sequences, 'Wuhan/0126-C93/2020', shares a mutation (13402G) with a lineage A genome from the same collection date and laboratory ('Wuhan/0126-C77/2020'). It is likely that the nucleotide at 28144 has been called as the reference allele (28144T - using the Wuhan-Hu-1 reference genome).

Information about 13 early cases linked to genomes was collected from published work and Tables 6 and 7 from the WHO report (WHO, 2021). Where there were discrepancies, the published reports were given priority. In particular, the case in Tables 6 and 7 with the earliest onset date (2019-12-08) seems to have been mistakenly linked to a genome (see Table S2, note 1). Where multiple genomes were linked to the same case in Table 6 of the WHO report (WHO, 2021), only one representative was included (Table S2).

Map Locations. Information taken from Xiao et al. (2021).

- Baishazhou market, Li Shui Lu, Hongshan Qu, Wuhan Shi, Hubei Sheng, China: 30.4626°N 114.2565°E

- Qiyimen Shengxian farmer's market, 588 Zhongshan Rd, Wuchang District, Wuhan, Hubei, China: 30.5232°N 114.3096°E

- Huanan Seafood Wholesale Market, Fazhan Ave, Jianghan District, Wuhan, Hubei, China: 30.6196°N 114.2576°E

"These shops selling live, often wild, animals included two at the Baishazhou market (a large market comprising c. 400 other types of shop), seven at Huanan seafood market (c. 120 other shops), four at Dijiao outdoor pet market (c. 100 other shops), and four at Qiyimen live animal market (c. 40 other shops)."

The Wuhan Institute of Virology (WIV):

- The Wuhan National Biosafety Laboratory at the Zhengdian Scientific Park of the Wuhan Institute of Virology: 30.376389°N, 114.262500°E

Panels b-d: Map data was manually extracted from Fig. 17 (Page 157) of the Annexes of the WHO report (WHO, 2021) using Adobe Illustrator. Because of multiple overlapping points there will be errors in the extraction process. Peripheral districts are: DXH: Dongxihu, CD: Caidian, JX: Jiangxia, HP: Huangpi, XZ: Xinzhou and HN: Hannan.

Panels e-f. Excess mortality from pneumonia by district/governmental areas from Fig. 21 (p. 40) of the WHO report (WHO, 2021) is indicated for selected dates.

Map data and polygons from OpenStreetMap (http://openstreetmap.org) and copyright © OpenStreetMap contributors – see https://www.openstreetmap.org/copyright for details.

Supplementary information to main text Figure 2.

Panel A. Alignment of the nucleotide sequences encoding the S1/S2 cleavage sites of the spike proteins of SARS-CoV-2 (YP_009724390 and bat Coronavirus RaTG13 (QHR63300.2). The reading frame for the amino acids can be inferred from the variation in the third base of several codons (yellow). Two possible insertions are indicated by capital letters, both of which are out-of-frame (-1 or -2). Numbers represent amino acids of the Spike proteins and nucleotides of the entire genomes.

Panel B. Amino acid alignment of the S1/S2 cleavage sites of selected beta spike proteins. Accession numbers: SARS-CoV-2 YP_009724390, SARS-CoV AAP13441.1, RaTG13 QHR63300.2, RmYN02 EPI_ISL_412977, MERS-CoV AGG22542.1, HKU4 MH002339.1 HKU5 AGP04943.1, HKU5 AGP04943.1, HKU1a ABD75561_1, HKU1b ABD96196_1, OC43 AIX10760.1, Bovine CoV CCE89341.1, HKU24 YP_009113025.1, Chinese *Hipposideros pratti* Bat-betacoronavirus/Zhejiang2013 (HpZJ13) and Nigerian *Hipposideros commersoni* Zaria bat coronavirus (HcNG08). To facilitate the identification of insertions we aligned a conserved cysteine residue (green) and included spikes from viruses that appear to be ancestral to the subgenuses where known. O-linked glycosylation sites were predicted by Net-O-Glyc v. 4.0.

Data files, map data and other supplementary materials are available from http://github.com/sars-cov-2-origins/critical-review/

Supplementary Table S1. Codons in the spike furin cleavage site of SARS-CoV-2.

		Residue 682		Residue 683	
Codon	Amino acid	count	Proportion	count	Proportion
CGG	R	2317072	0.99851	2308542	0.99484
CGT	R	2894	0.00125	3181	0.00137
CGC	R	215	9.26516E-05	52	2.24088E-05
CGA	R	200	8.61876E-05	449	0.000193491
TGG	W	50	2.15469E-05	218	9.39445E-05
AGG	R	34	1.46519E-05	402	0.000173237
CAG	Q	26	1.12044E-05	239	0.000102994
CTG	L	26	1.12044E-05	156	6.72263E-05
CCG	P	2	8.61876E-07	11	4.74032E-06
CAT	H	1	4.30938E-07	1	4.30938E-07
GGG	G	0	0	1	4.30938E-07
AAG	K	0	0	1	4.30938E-07
		2320520	0.00144*	2313252	0.00177*
			99.86%**		99.82%**

* Proportion of non-CGG arginine codons

** Percentage CGG relative to all arginine codons

Supplementary Table S2. Early cases linked to genome sequences.

Onset date	Collection date	Age/Sex	Sequence name	GISAID ID	Relation to the Huanan market	Reference	
2019-12-15	2019-12-24	65M	Wuhan/IPBCAMS-WH-01/2019	EPI_ISL_402123	Vendor	Ren et al., 2020	Note 1
2019-12-16	2019-12-30	41M	Wuhan/IPBCAMS-WH-03/2019	EPI_ISL_403930	None	Ren et al., 2020	Note 1
2019-12-17	2019-12-26	44M	Wuhan/WH01/2019	EPI_ISL_406798	Purchaser	WHO, 2021	
2019-12-19	2019-12-30	32M	Wuhan/HBCDC-HB-02/2019	EPI_ISL_412898	Vendor	WHO, 2021	Note 2
2019-12-20	2019-12-30	61M	Wuhan/IPBCAMS-WH-05/2020	EPI_ISL_403928	Purchaser	Ren et al., 2020	Note 1
2019-12-20	2019-12-26	41M	Wuhan/Hu-1/2019	EPI_ISL_402125	Worker	Wu et al., 2020	
2019-12-20	2020-01-02	39M	Wuhan/WHU01/2020	EPI_ISL_406716	Vendor	Chen et al., 2020	
2019-12-20	2019-12-30	56M	Wuhan/IME-WH04/2019	EPI_ISL_529216	Vendor	WHO, 2021	Note 3
2019-12-22	2020-01-02	21F	Wuhan/WHU02/2020	EPI_ISL_406717	Contact with Staff	Chen et al., 2020	
2019-12-23	2019-12-30	49F	Wuhan/IPBCAMS-WH-02/2019	EPI_ISL_403931	Vendor	Ren et al., 2020	Note 1
2019-12-23	2019-12-30	52F	Wuhan/IPBCAMS-WH-04/2019	EPI_ISL_403929	Vendor	Ren et al., 2020	Note 1
2019-12-23	2019-12-30	40M	Wuhan/WIV06/2019	EPI_ISL_402129	Vendor	WHO, 2021	
2019-12-26	2019-12-30		Wuhan/IME-WH01/2019	EPI_ISL_529213	Visitor to another market	WHO, 2021	

- Note 1: Patient 1, 2, 3 and 5 from Ren et al., (2021) were matched by age/sex and collection date in GISAID entry. Patient 4 was matched by elimination.
- Note 2: Age/sex taken from EPI_ISL_402127 - WIV02 - Linked in Table 6 of the WHO report (WHO, 2021) to be the same case.
- Note 3: Age/sex taken from EPI_ISL_402130 - WIV07 - Linked in Table 6 of the WHO report (WHO, 2021) to be the same case.