The Origins of SARS-CoV-2: A Critical Review

Supplementary Information

Edward C. Holmes[1], Stephen A. Goldstein[2], Angela L. Rasmussen[3], David L. Robertson[4], Alexander Crits-Christoph[5], Joel O. Wertheim[6], Simon J. Anthony[7], Wendy S. Barclay[8], Maciej F. Boni[9], Peter C. Doherty[10], Jeremy Farrar[11], Jemma L. Geoghegan[12], Xiaowei Jiang[13], Julian L. Leibowitz[14], Stuart J. D. Neil[15], Tim Skern[16], Susan R. Weiss[17], Michael Worobey[18], Kristian G. Andersen[19], Robert F. Garry[20,21], Andrew Rambaut[22].

[1]School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia.

[2]Department of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA.

[3]Vaccine and Infectious Disease Organization, University of Saskatchewan, Saskatoon, SK, S7N 5E3, Canada.

[4]MRC-University of Glasgow Centre for Virus Research, Glasgow, G61 1QH, UK.

[5]Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA 94704, USA.

[6]Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA.

[7]Department of Pathology, Microbiology, and Immunology, University of California Davis School of Veterinary Medicine, Davis, CA 95616, USA.

[8]Department of Infectious Disease, Imperial College London, W2 1PG, UK.

[9]Center for Infectious Disease Dynamics, Department of Biology, The Pennsylvania State University, University Park, PA 16802, USA.

[10]Department of Microbiology and Immunology, The University of Melbourne at the Doherty Institute, 792 Elizabeth St, Melbourne, VIC 3000, Australia.

[11]The Wellcome Trust, London, NW1 2BE, UK.

[12]Department of Microbiology and Immunology, University of Otago, Dunedin, New Zealand. Institute of Environmental Science and Research, Wellington, New Zealand.

[13]Department of Biological Sciences, Xi'an Jiaotong-Liverpool University (XJTLU), Suzhou, China.

[14]Department of Microbial Pathogenesis and Immunology, Texas A&M University, College Station, TX 77807, USA.

[15]Department of Infectious Diseases, King's College London, Guy's Hospital, London SE1 9RT, UK.

[16]Max Perutz Labs, Medical University of Vienna, Vienna Biocenter, Dr. Bohr-Gasse 9/3, A-1030 Vienna, Austria.

[17]Perelman School of Medicine, University of Pennsylvania. Philadelphia, PA 19104, USA.

[18]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA.

[19]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA.

[20]Department of Microbiology and Immunology, Tulane University School of Medicine, New Orleans, LA 70112, USA.

[21]Zalgen Labs, Germantown, MD 20876, USA.

[22]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, EH9 FL, UK.

Figure 1 Phylogenetic tree

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' MN908947 using Minimap2[1] and the 5' and 3' untranslated regions masked to avoid areas of low sequencing coverage. A maximum likelihood phylogenetic tree was reconstructed using IQTREE2[2] under the Jukes-Cantor model of molecular evolution. The tree was rooted at the midpoint between lineage A and lineage B.

Three genomes from late January 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, '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 the WHO report Tables 6 and 7 [3]. 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 Ref 3, only one representative was included (Table S2).

Figure 1 Map Locations

Xiao, X., Newman, C., Buesching, C.D. et al. Animal sales from Wuhan wet markets immediately prior to the COVID-19 pandemic. Sci Rep 11, 11898 (2021) [2]

- Baishazhou market, Wanfulin, Wuchang District, Wuhan, Hubei, China: 30°31'17.7"N 114°17'47.7"E

- Qiyimen Shengxian farmer's market, 588 Zhongshan Rd, Wuchang District, Wuhan, Hubei, China: 30°31'30.1"N 114°18'35.0"E

"These shops selling live, often wild, animals included two at 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) laboratories: 30°22'35.0"N 114°15'45.0"E

Panels b-d: Map data was manually extracted from Fig 17 (Page 157) of the Annexes of reference 10 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 reference 10 is indicated for selected dates.

Fig. 2 (Methods).

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.

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	1820709	0.9986784	1813363	0.99454945
CGT	R	2120	0.00116273	2457	0.00134756
CGC	R	201	0.00011024	25	1.3711E-05
CGA	R	172	9.4334E-05	381	0.00020896
AGG	R	27	1.4808E-05	350	0.00019196
TGG	W	34	1.8647E-05	148	8.1171E-05
CAG	Q	23	1.2614E-05	68	3.7295E-05
CTG	L	15	8.2268E-06	95	5.2103E-05
CCG	P	0	0	6	3.2907E-06
CAT	H	0	0	1	5.4846E-07
GGG	G	0	0	1	5.4846E-07
	Total	1823301	0.00138211*	1816895	0.00176219*
			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	
2019-12-16	2019-12-30	41M	Wuhan/IPBCAMS-WH-03/2019	EPI_ISL_403930	none	[4]	Note 1
2019-12-15	2019-12-24	65M	Wuhan/IPBCAMS-WH-01/2019	EPI_ISL_402123	vendor	[4]	
2019-12-17	2019-12-26	44M	Wuhan/WH01/2019	EPI_ISL_406798	purchaser	[3]	
2019-12-19	2019-12-30	32M	Wuhan/HBCDC-HB-02/2019	EPI_ISL_412898	vendor	[3]	Note 2
2019-12-20	2019-12-30	61M	Wuhan/IPBCAMS-WH-05/2020	EPI_ISL_403928	purchaser	[4]	Note 1
2019-12-20	2019-12-26	41M	Wuhan/Hu-1/2019	EPI_ISL_402125	worker	[5]	
2019-12-20	2020-01-02	39M	Wuhan/WHU01/2020	EPI_ISL_406716	vendor	[6]	
2019-12-20	2019-12-30	56M	Wuhan/IME-WH04/2019	EPI_ISL_529216	vendor	[3]	Note 3
2019-12-22	2020-01-02	21F	Wuhan/WHU02/2020	EPI_ISL_406717	Contact with Huanan Market staff	[6]	
2019-12-23	2019-12-30	49F	Wuhan/IPBCAMS-WH-02/2019	EPI_ISL_403931	vendor	[4]	Note 1
2019-12-23	2019-12-30	52F	Wuhan/IPBCAMS-WH-04/2019	EPI_ISL_403929	vendor	[4]	Note 1
2019-12-23	2019-12-30	40M	Wuhan/WIV06/2019	EPI_ISL_402129	vendor	[3]	
2019-12-26	2019-12-30		Wuhan/IME-WH01/2019	EPI_ISL_529213	visitor to another market	[3]	

- Note 1: Patient 1,2,3 & 5 from Ref 4 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 Ref 3 to be the same case.
- Note 3: Age/sex taken from EPI_ISL_402130 - WIV07 - Linked in Table 6 of Ref 3 to be the same case.

References

1. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34. doi:10.1093/bioinformatics/bty191

2. Xiao X, Newman C, Buesching CD, Macdonald DW, Zhou Z-M. Animal sales from Wuhan wet markets immediately prior to the COVID-19 pandemic. Sci Rep. 2021;11: 1–7.

3. WHO. WHO-convened global study of origins of SARS-CoV-2: China Part. World Health Organization; 30 Mar 2021 [cited 28 Jun 2021]. Available: https://www.who.int/publications/i/item/who-convened-global-study-of-origins-of-sars-cov-2-china-part

4. Ren L-L, Wang Y-M, Wu Z-Q, Xiang Z-C, Guo L, Xu T, et al. Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. Chin Med J . 2020;133: 1015.

5. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579: 265–269.

6. Chen L, Liu W, Zhang Q, Xu K, Ye G, Wu W, et al. RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. Emerg Microbes Infect. 2020;9: 313–319.