A critical reexamination of recovered SARS-CoV-2 sequencing data

F. Débarre[1] & Z. Hensel[2]

[1] Institute of Ecology and Environmental Sciences, CNRS UMR 7618, Sorbonne Université, UPEC, IRD, INRAE, Paris, France
https://orcid.org/0000-0003-2497-833X
[2] ITQB NOVA, Universidade NOVA de Lisboa, Lisbon, Av. da República, 2780-157, Oeiras, Portugal
Corresponding author: florence.debarre@normalesup.org

Abstract

SARS-CoV-2 genomes collected at the onset of the Covid-19 pandemic are valuable because they could help understand how the virus entered the human population. In 2021, Jesse Bloom reported on the recovery of a dataset of raw sequencing reads that had been removed from the NCBI SRA database at the request of the data generators, a scientific team at Wuhan University (Wang *et al.*, 2020b). Bloom suggested that the data may have been removed in order to obfuscate the origin of SARS-CoV-2, and he questioned the generating authors' statements that the samples had been collected on and after January 30, 2020. Here, we show that sample collection dates were published in 2020 by Wang *et al.* together with the sequencing reads, and match the dates given by the authors in 2021. We examine mutations in these sequences and confirm that they are entirely consistent with the previously known genetic diversity of SARS-CoV-2 of late January 2020. Finally, we explain how an apparent phylogenetic rooting paradox described by Bloom was resolved by subsequent analysis. Our reanalysis demonstrates that allegations of cover-up or of metadata manipulation were unwarranted.

Note for bioRxiv readers

The automatically generated Full Text version of our manuscript is missing footnotes; they are available in the PDF version.

1. Introduction

In June 2021, Jesse Bloom described the recovery of SARS-CoV-2 sequencing read data that had been deleted from the Sequence Read Archive (SRA) at the request of the data generators based at Wuhan University (Bloom, 2021a). Bloom claimed that the recovered data shed light on the early days—and thereby the origin—of the Covid-19 pandemic. His results, initially presented in a bioRxiv preprint and accompanied by a Twitter thread, reverberated in popular media[1,2] and were addressed at a press conference by a vice minister of China's National Health Commission.[3] Bloom's study was later published in *Molecular Biology and Evolution* (MBE; Bloom, 2021b).

The study for which the sequencing data had been generated, suddenly under international public scrutiny, presented a diagnostic technique based on amplifying fragments from a portion of the SARS-CoV-2 genome using nanopore technology (Wang *et al.*, 2020b). The article, published in the journal *Small*, had initially been shared as a preprint on the medRxiv server (Wang *et al.*, 2020a) (see Table 1 for a timeline). After the preprint was posted, raw sequence read data were submitted by Wang *et al.* to SRA as Bio-Project PRJNA612766 in mid-March

[1] References to non-academic work are presented as footnotes.

[2] e.g., C. Zimmer, Scientist Finds Early Virus Sequences That Had Been Mysteriously Deleted, `https://www.nytimes.com/2021/06/23/science/coronavirus-sequences.html`; see `https://medrxiv.altmetric.com/details/108029569/news/page:3` for other examples.

[3] Press conference recording, 22 July 2021: `https://www.youtube.com/watch?v=UA2P8hlurlQ&t=4606s`; Transcript: `https://www.pekingnology.com/p/why-did-wuhan-university-researchers`.

2020; these data were removed in mid-June 2020. Neither the preprint nor the published article mentioned the public availability of raw sequencing data.

Central to Bloom's claim was the argument that the removal of these data by Chinese scientists was carried out in secret, and with the intent to obstruct investigations of pandemic origins. This claim was promoted directly by Bloom (2021a) using phrases including: "*the sequences were deleted to obscure their existence*", "*surreptitiously delete the partial sequences*", and "*trusting structures of science have been abused to obscure sequences relevant to the early spread of SARS-CoV-2*". The argument is part of a narrative, presented in Bloom's introduction, which claims that Chinese researchers were "gagged" by China's government, and had to retract previously released data related to cases prior to mid-December 2019 to comply with one government order or another. [4]

Although the raw sequencing data had been removed from the SRA by the National Center for Biotechnology Information (NCBI) after a request by the authors, the published, peer-reviewed article (Wang *et al.*, 2020b) included as its Table 1 all of the mutations identified in the raw reads, the same ones later found in Bloom's reanalysis. The preprint (Wang *et al.*, 2020a) contained a less complete version of the table that nevertheless identified the mutation central in Bloom's analysis: mutation C29095T, in sample C2. Bloom (2021b)'s reanalysis indicated that Wang *et al.*'s results were consistent with data they submitted to SRA. In other words, the information contained in the raw data that Bloom recovered from SRA was available and described in documents published in 2020. The fact that Wang *et al.*'s description of the mutations identified in their samples is *even more complete* in the published article than in the preprint (Wang *et al.*, 2020a,b, Table 1) directly refutes the hypothesis that the raw sequence reads were removed from the SRA to obfuscate the origin of SARS-CoV-2.

Bloom initially discovered the existence of the Wang *et al.* sequences via a paper which had referenced sequencing data published on the SRA at the end of March 2020 (Farkas *et al.*, 2020). Sequencing data relating to Wang *et al.* were listed in Supplementary Table 1 of Farkas *et al.* (2020), but the data were no longer available and not findable on SRA when Bloom looked for them in 2021. However, the data had been backed up to the cloud, and Bloom recovered sequencing data from the backup.

In reply to Bloom's preprint, in 2021, Wang *et al.* responded that the samples from which sequences had been obtained had been collected on 30 January 2020 at the earliest.[5,6] According to the authors, the partial sequences were therefore not relevant to the origin of SARS-CoV-2. A co-author of Wang *et al.* also described the rationale for withdrawal of sequencing data from the SRA in an interview:[7] raw sequence data had been submitted to the SRA to accompany the submitted manuscript. After their article was accepted by the journal *Small*, the proofs received by the authors did not include a data availability statement. The authors thought it was appropriate to request deletion because journal editors did not retain their data availability statement and because SRA data would not be referenced in the manuscript.[8,9] This explana-

[4] There were multiple such official notices at different dates in early 2020; see timeline in Table 1. Different censorship narratives inconsistently refer to one or the other.

[5] Press conference – Zeng Yixin, vice minister of China's National Health Commission: "*According to our understanding, the earliest sampling time of this batch of samples was January 30 - some time has passed since the COVID outbreak began. In fact, it is not an early sample. These sequences provide limited information and value for COVID-19 origin tracing.*"; Zichen Wang, 22 July 2021, Why did Wuhan University researchers delete Covid-19 data at NIH?: `https://www.pekingnology.com/p/why-did-wuhan-university-researchers`.

[6] "*According to the researcher, a total of two batches of samples were taken. In the first batch, a total of 45 samples were taken randomly from patients that sought treatment in Wuhan on Jan. 30th, 2020. The second batch of samples was taken from a group of patients in mid-February, 2020.*" Zichen Wang, 24 July 2021, The Chinese side of the COVID data withdrawal controversy: `https://www.pekingnology.com/p/the-chinese-side-of-the-covid-data` (Interview conducted by Yang Liu).

[7] Zichen Wang, 24 July 2021, *ibid.*

[8] "*When we saw that the journal had deleted the paragraph, we believed that then the paragraph was unnecessary.*", *ibid.*

[9] "*Because the paper no longer included this descriptive paragraph (of the link to the database), the data that was stored in the database was like a headless fly. Nobody would know the data's association, maybe after some time, even we wouldn't be able to find the data, since there was no link. So we asked for the data to be deleted. This took place in*

Date (UTC)	Event description	Source
2020-01-29	China Ministry of Science and Technology notice encouraging scientists to fight the epidemic and publish in Chinese journals	https://m.sohu.com/a/369721616_120059213/
2020-02-25	China CDC notice on Covid-19 publications and data sharing	https://www.documentcloud.org/documents/7340336-China-CDC-Sup-Regs.html
2020-03-02	Wuhan University press release on the nanopore paper	https://web.archive.org/web/20211203030758/https://news.whu.edu.cn/info/1002/57753.htm
2020-03-03	Notice by the Chinese Minister of Science and Technology on Covid-19 scientific research	https://www.documentcloud.org/documents/7340337-State-Research-regulations.html
2020-03-04	Wang *et al.* nanopore paper sent to medRxiv	https://www.medrxiv.org/content/10.1101/2020.03.04.20029538v1.article-info
2020-03-06	Wang *et al.* (2020a) nanopore paper posted on medRxiv	https://www.medrxiv.org/content/10.1101/2020.03.04.20029538v1.article-info
2020-03-16	PRJNA612766 submitted to SRA, SUB7147304	https://justthenews.com/sites/default/files/2022-03/nih-foia-request-56712_redacted.pdf
∼2020-03-31	Farkas *et al.* download of SRA metadata	https://peerj.com/articles/9255/
2020-04-03	Wang *et al.* nanopore paper received by *Small*	https://onlinelibrary.wiley.com/doi/epdf/10.1002/smll.202002169
2020-04-17	Application filed by Wuhan Zhenxi Medical Laboratory Co Ltd for patent related to the nanopore paper	https://patents.google.com/patent/CN111662958A/zh
2020-05-27	Wang *et al.* nanopore paper revision received by *Small*	https://onlinelibrary.wiley.com/doi/epdf/10.1002/smll.202002169
2020-06-01	Wang *et al.* nanopore paper accepted by *Small*	Feb 2024 email from Wiley's Integrity Assurance & Case Resolution team to FD
2020-06-09/12	Proofs of the Wang *et al.* nanopore paper sent to the authors	Feb 2024 email from Wiley's Integrity Assurance & Case Resolution team to FD
2020-06-16	Authors request withdrawal of SUB7147304 (PRJNA612766)	https://justthenews.com/sites/default/files/2022-03/nih-foia-request-56712_redacted.pdf
2020-06-17	PRJNA612766 withdrawn	https://justthenews.com/sites/default/files/2022-03/nih-foia-request-56712_redacted.pdf
2020-06-24	Wang *et al.* (2020b) nanopore paper published online at *Small*	https://onlinelibrary.wiley.com/doi/epdf/10.1002/smll.202002169
2020-06-28	Wuhan University tweets about the publication of Wang *et al.* (2020b) in *Small*	https://x.com/WHU_1893/status/1277218113642086402
2020-09-15	Publication of patent CN111662958A	https://patents.google.com/patent/CN111662958A/zh

Table 1: Timeline of 2020 events related to Wang *et al.*'s study and sequencing data. The dates are written in the YYYY-MM-DD format.

tion is consistent with the timeline of events (see Table 1). It was also confirmed to us by Wiley's Integrity Assurance & Case Resolution team, who conducted an investigation on the case, that the data availability statement was removed by the journal during copy-editing.[10]

In his study, Jesse Bloom explored possible roots of the early SARS-CoV-2 phylogeny in the context of the data from Wang *et al.*. Assuming that the sequence of the most recent common ancestor of SARS-CoV-2 should resemble the most closely related viruses sampled in bats, and neglecting sampling dates, Bloom considered three roots: all were of lineage A (defined by sub-

June 2020.", ibid.

[10]Email to FD, 9 February 2024.

stitutions C8782T and T28144C compared to the reference sequence Wuhan-Hu-1), with each proposed root haplotype containing one additional mutation towards a bat-virus outgroup (either T3171C, C18060T, or C29095T) compared to Wuhan-Hu-1 (which is of lineage B; the positions and names are summarized in Table 2). This analysis also did not account for the fact that C→T is the most frequent type of single-nucleotide substitution in SARS-CoV-2 genomes (Azgari *et al.*, 2021; De Maio *et al.*, 2021; Bloom *et al.*, 2023; Ruis *et al.*, 2023).

Position	3171	8782	18060	28144*	29095*
Lineage B (Wuhan-Hu-1)	T	C	C	T	C
Lineage A	T	**T**	C	**C**	C
Bloom 1: A + C18060T	T	**T**	**T**	**C**	C
Bloom 2: A + C29095T	T	**T**	C	**C**	**T**
Bloom 3: A + T3171C	**C**	**T**	C	**C**	C

Table 2: Substitutions in the different lineages and Bloom's proposed roots. We use Wuhan-Hu-1 as reference. The positions highlighted with a star (*) are covered in the "recovered" sequences. The lineage defined by A + C18060T is referred to as "proCoV2" by Bloom (2021b), following previous analysis (Kumar *et al.*, 2021). Since the original "proCoV2" had three additional substitutions in the Kumar *et al.* preprint, we avoid this nomenclature.

Rooting the SARS-CoV-2 tree had long been identified as a difficult problem (Pipes *et al.*, 2021), for which different methods give different answers (Pekar *et al.*, 2021). In particular, an early sequence with three spurious mutations caused rooting issues (J. Wertheim, personal communication; Pekar *et al.*, 2022) until these errors were corrected in the China-WHO joint mission report (World Health Organization, 2021, Table 6, ID: S02, `IPBCAMS-WH-01`). Pekar *et al.* (2022) later showed that the root almost certainly lies along one branch including lineages B and A. However, uncertainty remains regarding whether the ancestral state is lineage A, lineage B, or an intermediate between them.

Bloom's rooting methodology led to a known conundrum (Rambaut *et al.*, 2020): all three of the roots that Bloom considered plausible roots did not resemble the sequences with earliest collection dates. To explain this discrepancy, Bloom suggested that critical early (meta)data may be missing or altered, identifying the "recovered sequences" as examples for which true collection dates were potentially earlier than reported: "*The press conference and blog posts also stated that the sequences were all collected on or after January 30, 2020, rather than "early in the epidemic" as originally described in Wang et al. (2020).*" (Bloom, 2021b). This suggestion was also expressed on Twitter, when a news article reporting the story behind Bloom's preprint was published:[11] "*Dr. Zeng Yixin* [vice-minister of China National Health Commission] *also said earliest collection time for deleted sequences was Jan-30-2020 & so they were "not early-stage samples." In contrast, Chinese authors originally said samples were from "early in the epidemic." I lack data to reconcile these differing descriptions (18/n)*".[12] In other words, Bloom suggested that the 30 January 2020 collection date was incorrect.

Here we provide multiple lines of evidence showing that the 30 January 2020 date put forward by the Chinese scientists in July 2021 was correct, including the crucial fact that collection dates were available in the dataset analyzed by Bloom. Speculation that scientists may have been lying about the collection dates of these samples was, and remains, unsubstantiated and unwarranted.

[11] Katherine Eban, 31 March 2022, `https://www.vanityfair.com/news/2022/03/the-virus-hunting-nonprofit-at-the-center-of-the-lab-leak-controversy`.

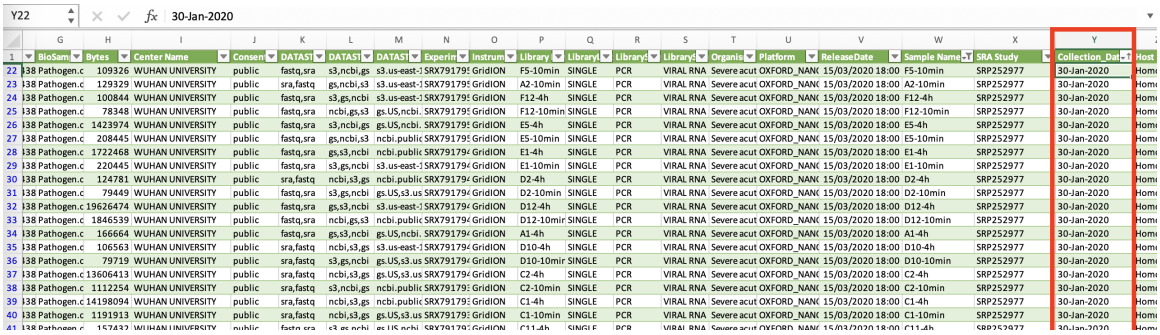
[12] `https://twitter.com/jbloom_lab/status/1509598923588993027`, Mar 31, 2022.

2. Results

2.1 The 30 January 2020 collection date was present in the data from Wang *et al.* (2020)

In his article, Bloom questioned the veracity of 30 January 2020 collection dates reported by authors of Wang *et al.* (2020b) in 2021. However, there is contemporaneous evidence confirming the dates. The collection dates were indeed present in the SRA metadata and remain visible today.[13] The collection date visible today, 30 January 2020, matches the date reported by authors of Wang *et al.* in 2021. According to the SRA team, the collection dates were the same in 2020. This is further confirmed by Supplementary Table 1 of Farkas *et al.* (2020), which was compiled at the end of March 2020 (see Figure 1). The same supplementary table was used by Bloom (2021a,b) to originally identify raw sequencing data from Wang *et al.* (2020b). This table consists of sequencing metadata downloaded after publication of the Wang *et al.* (2020a) preprint, but before its submission to *Small*; see Table 1 for a chronology. In summary, there is zero evidence that co-authors of Wang *et al.* ever fabricated or altered sample collection dates, and ample evidence that they did not.

Collection dates on and after 30 January 2020 are significant, as a small number of partial sequences from samples collected at this time are unlikely to substantially shift likelihoods of proposed SARS-CoV-2 progenitor genomes. Full genome sequences from samples collected on or before 30 January are not rare: there are 507 such sequences in data considered by Bloom (2021b), and there are 430 such sequences the dataset considered by Pekar *et al.* (2022), with more extensive quality control.



	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	BioSam	Bytes	Center Name	Consen	DATAST	DATAST	DATAST	Experin	Instrum	Library	Library	Library	Library	Organis	Platform	ReleaseDate	Sample Name	SRA Study	Collection_Da
22	138 Pathogen.c	109326	WUHAN UNIVERSITY	public	fastq,sra	s3,ncbi,gs	s3.us-east-1	SRX79179	GridION	F5-10min	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	F5-10min	SRP252977	30-Jan-2020
23	138 Pathogen.c	129329	WUHAN UNIVERSITY	public	sra,fastq	gs,ncbi,s3	s3.us-east-1	SRX79179	GridION	A2-10min	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	A2-10min	SRP252977	30-Jan-2020
24	138 Pathogen.c	100844	WUHAN UNIVERSITY	public	fastq,sra	s3,gs,ncbi	s3.us-east-1	SRX79179	GridION	F12-4h	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	F12-4h	SRP252977	30-Jan-2020
25	138 Pathogen.c	78348	WUHAN UNIVERSITY	public	fastq,sra	ncbi,gs,s3	gs.US,ncbi.	SRX79179	GridION	F12-10min	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	F12-10min	SRP252977	30-Jan-2020
26	138 Pathogen.c	1423974	WUHAN UNIVERSITY	public	fastq,sra	s3,ncbi,gs	gs.US,ncbi.	SRX79179	GridION	E5-4h	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	E5-4h	SRP252977	30-Jan-2020
27	138 Pathogen.c	208445	WUHAN UNIVERSITY	public	fastq,sra	gs,ncbi,s3	ncbi.public	SRX79179	GridION	E5-10min	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	E5-10min	SRP252977	30-Jan-2020
28	138 Pathogen.c	1722468	WUHAN UNIVERSITY	public	fastq,sra	gs,s3,ncbi	ncbi.public	SRX79179	GridION	E1-4h	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	E1-4h	SRP252977	30-Jan-2020
29	138 Pathogen.c	220445	WUHAN UNIVERSITY	public	fastq,sra	s3,gs,ncbi	s3.us-east-1	SRX79179	GridION	E1-10min	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	E1-10min	SRP252977	30-Jan-2020
30	138 Pathogen.c	124781	WUHAN UNIVERSITY	public	sra,fastq	ncbi,s3,gs	ncbi.public	SRX79179	GridION	D2-4h	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	D2-4h	SRP252977	30-Jan-2020
31	138 Pathogen.c	79449	WUHAN UNIVERSITY	public	fastq,sra	s3,gs,ncbi	gs.US,s3.us	SRX79179	GridION	D2-10min	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	D2-10min	SRP252977	30-Jan-2020
32	138 Pathogen.c	19626474	WUHAN UNIVERSITY	public	fastq,sra	gs,s3,ncbi	s3.us-east-1	SRX79179	GridION	D12-4h	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	D12-4h	SRP252977	30-Jan-2020
33	138 Pathogen.c	1846539	WUHAN UNIVERSITY	public	fastq,sra	ncbi,gs,s3	ncbi.public	SRX79179	GridION	D12-10min	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	D12-10min	SRP252977	30-Jan-2020
34	138 Pathogen.c	166664	WUHAN UNIVERSITY	public	fastq,sra	gs,s3,ncbi	gs.US,ncbi.	SRX79179	GridION	A1-4h	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	A1-4h	SRP252977	30-Jan-2020
35	138 Pathogen.c	106563	WUHAN UNIVERSITY	public	sra,fastq	ncbi,s3,gs	s3.us-east-1	SRX79179	GridION	D10-4h	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	D10-4h	SRP252977	30-Jan-2020
36	138 Pathogen.c	79719	WUHAN UNIVERSITY	public	sra,fastq	s3,gs,ncbi	gs.US,s3.us	SRX79179	GridION	D10-10min	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	D10-10min	SRP252977	30-Jan-2020
37	138 Pathogen.c	13606413	WUHAN UNIVERSITY	public	sra,fastq	ncbi,s3,gs	gs.US,s3.us	SRX79179	GridION	C2-4h	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	C2-4h	SRP252977	30-Jan-2020
38	138 Pathogen.c	1112254	WUHAN UNIVERSITY	public	sra,fastq	s3,ncbi,gs	ncbi.public	SRX79179	GridION	C2-10min	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	C2-10min	SRP252977	30-Jan-2020
39	138 Pathogen.c	14198094	WUHAN UNIVERSITY	public	sra,fastq	ncbi,s3,gs	ncbi.public	SRX79179	GridION	C1-4h	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	C1-4h	SRP252977	30-Jan-2020
40	138 Pathogen.c	1191913	WUHAN UNIVERSITY	public	sra,fastq	ncbi,s3,gs	gs.US,s3.us	SRX79179	GridION	C1-10min	SINGLE	PCR	VIRAL RNA	Severe acut	OXFORD_NAN	15/03/2020 18:00	C1-10min	SRP252977	30-Jan-2020

Figure 1: Screenshot of Supplementary Table 1 from Farkas *et al.* (2020). The highlighted cell is a collection date (`30-Jan-2020`). Source: `https://dfzljdn9uc3pi.cloudfront.net/2020/9255/1/Supplementary_Table_1.xlsx`; `https://peerj.com/articles/9255/#supp-2`. The file is also available in Jesse Bloom's Github repository (`https://github.com/jbloom/SARS-CoV-2_PRJNA612766/blob/main/manual_analyses/PRJNA612766/Supplementary_Table_1.xlsx`).

Finally, we note that there is no discrepancy to reconcile between the 30 January 2020 date and the phrase "*early in epidemic*". Wang *et al.* were developing a new test in a first-line hospital, and January 30 was indeed early in the context of the response to the epidemic at Renmin Hospital of Wuhan University.

2.2 The study timeline is incompatible with the alleged censorship

Bloom speculated about a scenario in which co-authors of Wang *et al.* were compelled to remove their data from SRA under the pressure from China's government, which published notices regarding scientific publication during the Covid-19 pandemic. The actual timeline of events contradicts this narrative: the preprint itself was posted to medRxiv *after* the publication of the second notice referenced by Bloom, as were the raw sequencing data on the SRA (see Table 1).

[13] e.g., `https://www.ncbi.nlm.nih.gov/biosample/?term=SAMN14381071`.

2.3 The "recovered" sequences are compatible with a late January collection date

To further test the veracity of the 30 January 2020 sampling date announced by Wang *et al.*, we compare partial sequences from Wang *et al.* (2020b) to the corresponding region of other early sequences, following the approach in Figure 4 of Bloom (2021b).

As a first comparison, we turn to sequencing data generated via a similar nanopore-based technology, obtained from samples collected in Wuhan, with similar collection dates to those reported for the earliest samples in Wang *et al.*. These data were reported in the context of an article by Yan *et al.* (2021); the samples were collected from "*various Wuhan health care facilities*" on 25 and 26 January 2020, and consensus sequences were deposited on GISAID. Two sequences from the Yan *et al.* (2021) dataset are present in proposed progenitor nodes in Bloom (2021b).[14] Figure 2 shows that the distribution of substitutions in sequences from Yan *et al.* (2021) is similar to that of Wang *et al.* (2020b) ("recovered" sequences). The two distributions remain similar when the outgroup comparator is changed (Figure S1) or when a partial dataset of Yan *et al.* (2021) is used, removing sequences with potential sequencing errors (Figure S2).

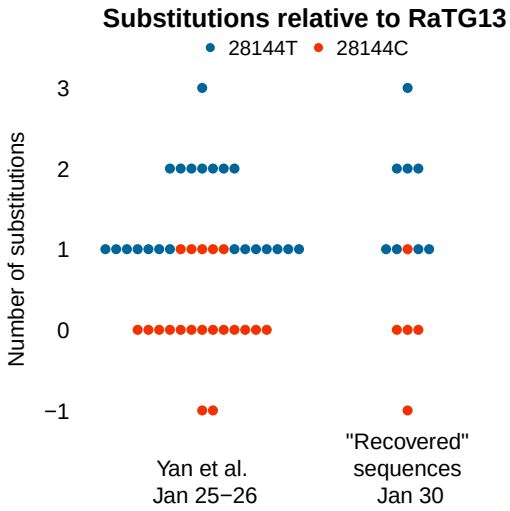


Figure 2: Number of substitutions from bat SARS-like coronavirus RaTG13 (relative to lineage A, or, equivalently, lineage A + C18060T, because the C18060T mutation distinguishing them is not included in the region considered, between nucleotides 21,570–29,550 as in Figure 4 in Bloom (2021b)). Sequences from Yan *et al.* (2021) are compared to those from Wang *et al.* (2020b) ("Recovered" sequences). Substitutions are counted such that 0 corresponds to the same distance as between RaTG13 and lineage A; negative values (−1) correspond to additional substitutions towards RaTG13 (C29095T for a "recovered" sequence and for one of the Yan *et al.* (2021) sequences, and C22747T for the other Yan *et al.* sequence). Substitution T28144C is characteristic of lineage A and is highlighted in red. (NB: We use RaTG13 only for the sake of comparison with Bloom's analysis.)

Figure 2 also illustrates that substitutions towards the chosen outgroup are not necessarily signs of their ancestral nature. The −1 positions of three sequences in Figure 2 are due to C29095T (one "recovered sequence" and one sequence from Yan *et al.* (2021)) and to C22747T (the other Yan *et al.* (2021) sequence). Both substitutions have subsequently reappeared in other SARS-CoV-2 lineages (see Figure S3). Outside of the region covered in sequences from Wang *et al.*, the Yan *et al.* sequence with C22747T also contains T4402C and G5062T, identifying C22747T as a reversion subsequent to mutations that characterize a common early epidemic genome in lineage A sampled in China (Beijing), South Korea, and Japan (i.e., not an

[14] `hCoV-19/Wuhan/0126-C13/2020` in the A + C18060T root, and `hCoV-19/Wuhan/0126-C31/2020` in the A + C29095T root. The `C31` sequence has two additional mutations, but they are unique mutations in Bloom (2021b)'s dataset and were therefore discarded in his workflow.

early ancestral genome).

The comparison can be extended to a broader set of early sequences. Figure 3 shows that the number of substitutions in the Wang *et al.* (2020b) dataset is consistent with those observed in other sequences with similar collection dates. The pattern holds when changing the comparator (Figure S4; all data points instead of averages are shown in Figure S5).

By implicitly assuming that positions that are not covered are not mutated, Bloom's methodology will underestimate divergence for sequences with low coverage. Bloom (2021b) highlighted "*a sequence (Guangdong/FS-[S]30-P00502/2020 reportedly collected in late February that is actually two mutations more similar to RaTG13 than lineage A + C18060T*" (corresponding to a point at "−2" in the "Other China" panel of his Fig. 2). We doubt it is a coincidence that the most striking outlier in this figure is also a sequence with one of the lowest levels of coverage in his dataset (84%, ranking 5th of 1886 sequences).

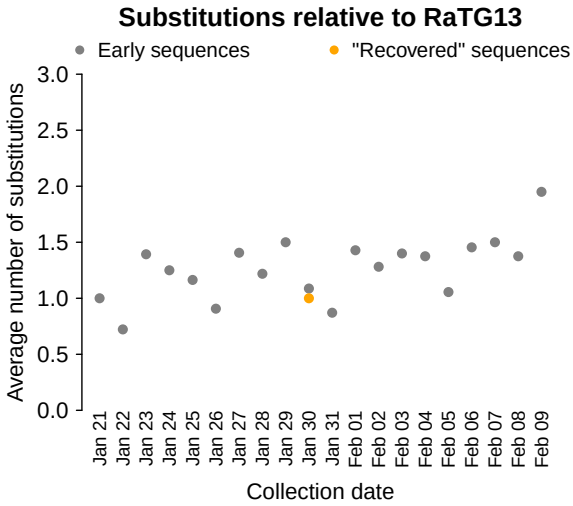


Figure 3: Average number of substitutions relative to bat SARS-like coronavirus RaTG13 (from lineage A, or, equivalently, lineage A + C18060T), over SARS-CoV-2 nucleotides 21,570–29,550, comparing all available sequences (after de-duplications and curation; Pekar *et al.* (2022); gray points) to Wang *et al.* (2020b)'s "recovered" sequences (orange).

2.4 No evidence of a widespread undetected circulation of virus with C29095T in Wuhan

Analysis in Bloom (2021b) highlighted sequences collected in the Guangdong province that were related to what he considers a very plausible progenitor SARS-CoV-2 genome (lineage A + C29095T). The presence of C29095T brought them closer to the bat virus outgroup, and positioned them at a striking "−1" in Figure 4 of Bloom (2021b) (orange dots in his figure). Initially described as belonging to "*two different clusters of patients who traveled to Wuhan in late December of 2019*", these sequences could be interpreted as evidence of a widespread but so far undetected circulation of similar viruses in Wuhan in late 2019.

Examination of the included sequences, however, indicated that there was only one cluster rather than two; Bloom recently corrected the article after we and others pointed this out (Bloom, 2023). All the patients were from the same family group, and therefore the sequences were not independent. In addition, we found that multiple sequences collected from the same patients were included in Bloom's dataset, sometimes labeled as "Other China". Briefly, at least seven sequences belong to the same family cluster detected in Guangdong (Chan *et al.*, 2020; Kang *et al.*, 2020), corresponding to four patients, two of which were sampled multiple times; two of the sequences were labeled as "Other China" by Bloom (see supplementary Tables S1 and S2 for details). Further, following Bloom (2021b)'s advice to "*[go] beyond the annotations in GISAID to carefully trace the location of patient infection and sample sequencing*", we note that

plausible index patients in the Guangdong cluster did not just travel to Wuhan in late December 2019, but had visited a relative hospitalized in Wuhan for febrile pneumonia (Chan *et al.*, 2020). In other words, they had been to one of the few places other than the Huanan market where one was most likely to encounter other people infected by SARS-CoV-2 in Wuhan at that early date.

Bloom argued that a A + C29095T root is "*now more consistent with the evidence that the pandemic originated in Wuhan, as half*[15] *its progenitor node is derived from early Wuhan infections, which is more than any other equivalently large node.*". There was in fact a single "recovered" sequence in the node (sample C2); the rest of the increase in support was due to the relabeling of Guangdong cluster sequences. Moreover, in addition to the documented epidemiological link between annotated cases discussed above, there is no reason to expect that sample C2 necessarily lacks additional mutations outside of the region covered in "recovered" sequences[16]. Furthermore, epidemiological links to Wuhan are very common in case reports from January 2020, and not only for A + C29095T. For example, all eight sequences in Bloom's proposed A + T3171C root have a documented epidemiological link to Wuhan (Jiang *et al.*, 2020), as does the first Covid-19 case detected in the United States with A + C18060 (Holshue *et al.*, 2020). But this remark should not be interpreted as support for those haplotypes as roots, because lineage A and lineage B exports were for instance found in Australia (Eden *et al.*, 2020) and in the first two cases in Thailand (Okada *et al.*, 2020). Lastly, a complete annotation of exposure history for cases outside of Wuhan should note the case with symptom onset predating any case in the Guangdong cluster by almost two weeks: this is a lineage B case identified in Beijing with a link not only to Wuhan, but to the Huanan market specifically (Liu, 2020).

The history of the Guangdong cluster indicated that the C29095T substitution was present in Wuhan in late December 2019; it is therefore unsurprising that it was sampled in late January 2020 as well. While C29095T is a mutation towards the most closely related viruses sampled in bats, the A + C29095T haplotype is rejected as the root of the SARS-CoV-2 phylogenetic tree in humans by Pekar *et al.* (2022). Methods in Bloom (2021b) neglect that C→T mutations are by far the most frequent type of mutation during the pandemic (Azgari *et al.*, 2021; De Maio *et al.*, 2021; Bloom *et al.*, 2023; Ruis *et al.*, 2023), and that the C29095T mutation, specifically, occurs much more frequently than expected for a typical C→T mutation (Bloom and Neher, 2023, supplementary data `nt_fitness.csv`). This mutation regularly reappeared in multiple lineages during the last four years of SARS-CoV-2 evolution in humans (Figure S3). For example, it is a defining mutation in the HP.1.1 lineage that emerged in North America in mid-2023. In fact, C29095T is even recurrent in Bloom (2021b)'s phylogenetic trees (Figs 3 and 5), where this position mutates three times.

3. Discussion

The facts we present do not support the conclusion that recovered sequences "shed more light on the early Wuhan SARS-CoV-2 epidemic" as promised by Bloom (2021b). First, Bloom's phylogenetic analyses did not require any of the recovered raw data, as all the data utilized were publicly available in a peer-reviewed article (Wang *et al.*, 2020b). Second, the "recovered" sequences are partial: only a fraction of the whole genome was sequenced, by design, seriously limiting the usefulness of these "recovered" sequences to infer ancestral states. Finally, the samples were collected in late January 2020, and as such were unlikely to provide useful information on the genome of the proximal ancestor of SARS-CoV-2 prior to the outbreak in Wuhan. Mutations identified in these samples, including C29095T, are unsurprising to find in Wuhan in late January.

[15] We note that "half its progenitor node" was true in Bloom (2021a), but is not in Bloom (2021b), owing to a shift in methods from suppressing rare haplotypes to suppressing rare mutations.

[16] Considering full sequences in Bloom's data set collected on Jan 30, 2020 ±5 days, those with C29095T have, on average, 1.1 substitutions (±0.2; $n = 27$) outside of the region covered by the "recovered" sequences.

The haplotypes proposed by Bloom as roots of the SARS-CoV-2 tree led to a conundrum: they were of lineage A, while the earliest known sequences were of lineage B, as were all the Huanan market sequences known at the time. The issue of the absence of lineage-A sequences directly linked to Huanan market was resolved a few months after the publication of Bloom's article. In early 2022, Liu *et al.* (2022) revealed that a lineage-A genome had been detected in an environmental sample collected in the Huanan market on the 1st of January 2020. Raw sequencing data from this sample, shared a year later (Liu *et al.*, 2023), confirmed the lineage assignment.

The question of the precise identity of SARS-CoV-2's root remains unresolved. Pekar *et al.* (2022) proposed a SARS-CoV-2 origin scenario resolving the rooting conundrum: the root may never have been in humans, but only in the animals from which SARS-CoV-2 spilled over more than once. Under this scenario, the two early lineages, A and B, would have been the products of two spillovers close in time and space, possibly from the same group of animals. The more "bat-like" lineage A likely spilled over after B, resulting in most early sequences being derived from lineage B. Among the several examples of animal-to-human SARS-CoV-2 transmission documented later in the pandemic (reviewed in EFSA Panel on Animal Health and Welfare (AHAW) *et al.*, 2023), such a scenario of multiple transmissions close in time and space, from a group of animals to humans, occurred notably with pet hamsters in Hong Kong, for which a genomic investigation indicated that different patients had been independently infected by hamsters at a pet shop (Yen *et al.*, 2022). Low diversity in viral genomes identified in samples from bats at the same time and place is also common; for example, RshSTT182 and RshSTT200 genomes differ by only 3 nucleotides (Delaune *et al.*, 2021).

Facing the same conundrum, Bloom proposed a different explanation: the record of published SARS-CoV-2 genomic sequences would have been intentionally altered by China authorities through selective suppression of sequences of some of the earliest samples. We demonstrated that there is no evidence supporting this speculation.

A divergence between data and expectations from a theory can be due to issues with the data, or to issues with the theory. Checking the reliability of data, especially from diverse sources, is an essential step in any scientific endeavor. There are for instance some aberrant collection dates in public databases, and data need to be curated to avoid absurd conclusions due to issues in input data. When the data are consistent, however, it is also important to challenge one's theory and to reconsider methodological choices. Hybrid methods taking into account both the bat virus relatives and collection dates find support for A or B roots, but confidently reject both the A + C29095T and A + C18060T roots considered by Bloom (2021b) (Pekar *et al.*, 2022, Table 1).

In an attempt to reconcile the proposed roots with the lack of supporting data, Bloom (2021b) suggested that the collection date indicated by Wang *et al.* in 2021 could be incorrect. Our investigation invalidates this suggestion. We demonstrate that collection dates have been available since March 2020, and that the data are also fully compatible with reported collection dates. The unavailability of the Wang *et al.* raw sequencing data on the SRA after June 2020 was shown to be the product of a human error (Berman *et al.*, 2022). According to policies of the International Nucleotide Sequence Database Collaboration (INSDC; Brunak *et al.*, 2002), of which the SRA is a member, data once made public on this repository are supposed to belong to the scientific record, and to remain accessible. Mechanisms exist to remove data from indexing, but keep them available by accession ("suppress" command) (Berman *et al.*, 2022). Due to human error however, the data were instead made unavailable ("kill" command), i.e. made unavailable on the SRA (Berman *et al.*, 2022). The avoidable deletion of raw data motivated speculation that caused harm to scientists who had submitted them.

In Bloom (2021b), the removal of the raw data was presented as part of a larger narrative, set up in his Introduction, in which Chinese authorities would have gagged researchers and made them retract or falsely amend previous statements, in particular on early Covid-19 cases. In this narrative, changes in the inclusion of early cases are seen as censorship rather than the simple correction of errors. The censorship narrative comes from the fanciful interpretation of

a news article published on a blog by a lab leak activist.[17] The source news article,[18] however does not support this narrative, when the quotes are read in full (emphasis added):

> *As of February 25, our entire database has about 47,000 cases. The database has some data on patients who developed the disease before December 8 last year,* ***but we cannot be sure of the authenticity of these data and further verification is needed.****.*
> *Professor Yu Chuanhua explains,*
>
> > *"For example, there is data on a patient who developed the disease on September 29, the data shows that the patient did not undergo nucleic acid testing, the clinical diagnosis (CT diagnosis) is a suspected case and the patient has died, this data has no confirmed diagnosis and no time of death,* ***it could also be wrong data****."*

These quotes in the original news article make it clear that the retrospective search of Covid-19 cases was work in progress, and that the results could change. There is no evidence that the researcher was forced by Chinese authorities to walk back earlier comments because of a gag order; instead, the researcher later gave an updated report of an ongoing analysis. Likewise, it is important to emphasize that the order to destroy samples, mentioned in Bloom's Introduction, was not specific to Covid-19: it followed from a biosafety regulation published long before Covid-19.[19]

The notion that Wang *et al.*'s data withdrawal was linked to something nefarious was pervasive throughout Bloom's article. In the final version of his article, Bloom (2021b) added a note suggesting that Wang *et al.* might have wanted to retract their preprint to cover their tracks.[20] There is no evidence that Wang *et al.* wanted to delete their paper. On the contrary, their work was featured in official channels, both before and after the removal of data on SRA (see Table 1). First, the preprint did not contain a link to the data—the data were submitted to the SRA only after the preprint was posted. Second, a press release about the work was posted before the preprint was submitted to medRxiv and was still online at the end of 2021. Third, the peer-reviewed paper itself was also advertised: Wuhan University tweeted about the paper when it came out. Finally, the work appears to have been part of a patent application.

Beyond the data from Wang *et al.* (2020b), but still in the context of the Bloom (2021a,b) study, Bloom also investigated other sequence datasets that were either removed or corrected to answer questions about whether sequencing (meta)data relevant to the origin of SARS-CoV-2 had been suppressed. We show in the Appendix that in all cases investigated, the answer is "no." While we reject unfounded speculation of this sort, we recognize that some datasets relevant to the origin and early spread of SARS-CoV-2 are known to exist and remain unpublished (Holmes, 2024). We hope that these datasets will be published to help resolve unanswered questions.

Bloom's article illustrates how prejudices can influence scientific conclusions. Data and analysis were presented through the lens of Chinese censorship and the implication that research in China is inherently untrustworthy. We conclude by noting that we initially took it for granted when we read Bloom's claim that "*the sequences were deleted to obscure their existence*" (ZH), or were initially captivated by the feat of recovering deleted data (FD). The fact that this narrative captured so much attention despite *a complete lack of supporting evidence* prompts us to reflect on how our biases shape our interpretation of data, and how extreme differences in believing people based on where they work can lead to incorrect and harmful conclusions. Here, we are reflecting on our experiences, and we invite readers to do the same.

---

[17] `https://github.com/jbloom/SARS-CoV-2_PRJNA612766/blob/main/literature_notes/README.md`, citing "Rushed data collection of suspected early Covid-19 cases in Wuhan".

[18] Health Times, 2020, `https://www.guancha.cn/politics/2020_02_27_538822.shtml`.

[19] Law text: `https://www.gov.cn/gongbao/content/2019/content_5468882.htm`.

[20] "*Notably, it is not possible to delete preprints from bioRxiv and medRxiv, so once Wang* et al. *(2020) had posted their preprint, it was permanently committed to the public record (withdrawn preprints are still accessible, for instance see Yang et al. 2020).*", Bloom (2021b).

Methods

We followed the same methods as Bloom (2021b) to compare sequences to outgroups. We used data shared by Bloom on Github, and outputs of a dataset curated by Pekar *et al.* (2022). We gratefully acknowledge the authors from the originating laboratories and the submitting laboratories, who generated and shared through GISAID the viral genomic sequences and metadata on which this research is based. Accessions used are the same as Pekar *et al.* (2022) data S1. The Yan *et al.* (2021) data correspond to EPI_ISL_493149 to EPI_ISL_493190.

Data and code are available on Zenodo `https://zenodo.org/doi/10.5281/zenodo.10665464`.

Acknowledgements

FD thanks the SRA team for their answers to her questions. We thank Alex Crits-Christoph, Joel Wertheim and Mike Worobey for comments and discussions. Zhihua Chen first spotted the two Guangdong clusters error in Bloom (2021b). FD thanks Dake Kang for providing details on the Chinese law behind the destruction of samples. FD thanks Wiley for sharing details on the Wang *et al.* (2020b) publication process and on their investigation. We thank Jonathan Pekar and Alex Crits-Christoph for sharing their files of substitutions. Finally, we thank all data producers for sharing their sequencing data on GISAID and on open platforms (the GISAID accessions are those from Pekar *et al.* (2022), listed in their data S1). ZH is supported by Fundação para a Ciência e a Tecnologia (FCT) through MOSTMICRO-ITQB (UIDB/04612/2020, UIDP/04612/2020) and LS4FUTURE (LA/P/0087/2020).

Appendix

Here we provide details or other datasets related to Bloom's study. These datasets were cited in different versions of the study or in accompanying communication. They correspond to data that were removed at some point from public repositories, or associated with metadata that changed. In all three cases that we present, the (meta)data do not shed light on the origin of SARS-CoV-2, and the explanations for their removal or modification are mundane.

SRR11119760 and SRR11119761, PRJNA607174

Bloom's original preprint (Bloom, 2021a, v1; Figure 2) contained the screenshot of an email showing another group of Chinese scientists asking SRA for the removal of their data. The screenshot had been obtained through a FOIA request by an activist group pursuing the hypothesis that papers on pangolin viruses were part of a concerted diversion.[21] By pure happenstance, the data were put back online on June 16, 2021,[22] that is, two days before Bloom posted his preprint to bioRxiv and shared it with NIH leadership, and six days before the preprint was published on bioRxiv (June 22, 2021).

The SRA team indicated that the data had be released "*at the request of a user*". Whether this is related or not, we can note that a Zenodo document was updated on June 21, 2021 with an analysis of that dataset (Daoyu Zhang, 2020, version 14), i.e., before Bloom's preprint was even published on bioRxiv and therefore before attention was drawn to those data.

PRJNA637497

In the revised version of his study, Bloom included an email by SRA confirming that two Wang *et al.* datasets had been removed (Bloom, 2021b, Figure 6). Although the accessions were

[21] `https://usrtk.org/wp-content/uploads/2020/12/NCBI-Emails.pdf`.

[22] See "Published date", `https://www.ncbi.nlm.nih.gov/sra/?term=SRR11119760` and `https://www.ncbi.nlm.nih.gov/sra/?term=SRR11119761`.

redacted in Bloom's article, Bloom shared the second accession on social media,[23] and the email is available in documents posted on the Internet. The metadata of this dataset are back online on SRA (under SAMN15143806[24]/SRR11931188[25]), and indicate that it contained a single sample collected on 23 March 2020, i.e. too late to be relevant to the origin of SARS-CoV-2. The SRA team confirmed that the collection date has not been modified since the initial submission in 2020.

PRJNA605907

A separate study on early cases (Shen *et al.*, 2020) was discussed by Bloom in Twitter threads[26] related to his 2021 MBE article. The main text of the Shen *et al.* article initially was not consistent with sequence metadata; a correction was published after Bloom's initial tweets (Shen *et al.*, 2021). An in-depth examination of the data indicates that the samples were collected as announced in the sequence metadata and as later corrected. The samples were collected from known patients from 30 December 2019, and sent to separate groups for analysis. The patients are known, the timeline is clear, and there is zero evidence that the samples were collected earlier than the stated date.[27]

[23] Jesse Bloom, March 31, 2022: "*Finally, e-mails show Wuhan University deleted *two* projects, only one of which (SUB7147304=PRJNA612766) was published in journal Small & described in my paper. Initial email focused on deleting another previously unknown project (SUB7554642=PRJNA637497). (23/n)*" `https://x.com/jbloom_lab/status/1509598938772361218`

[24] `https://www.ncbi.nlm.nih.gov/biosample/?term=SAMN15143806`.

[25] `https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR11931188&display=data-access`.

[26] `https://x.com/jbloom_lab/status/1432903935312818178` on September 1st, 2021 and `https://x.com/jbloom_lab/status/1509599601753395210` on March 31, 2022.

[27] see `https://github.com/flodebarre/Shen-etal_2020/tree/main` for an analysis.

References

Azgari C, Kilinc Z, Turhan B, Circi D, Adebali O. 2021. The Mutation Profile of SARS-CoV-2 Is Primarily Shaped by the Host Antiviral Defense. Viruses. 13:394. https://www.mdpi.com/1999-4915/13/3/394.

Berman A, Boykin L, Ceasar M, Sowa A, Twigger S. 2022. NIH/NLM: Root Cause Analysis: Removal of SRA Sequence Data Records. Technical report. BioTeam, Inc.. https://ftp.ncbi.nlm.nih.gov/sra/doc/BioTeam-RCA-RedactedReport.pdf.

Bloom JD. 2021a. Recovery of deleted deep sequencing data sheds more light on the early Wuhan SARS-CoV-2 epidemic. Preprint. Evolutionary Biology. doi:10.1101/2021.06.18.449051.

Bloom JD. 2021b. Recovery of Deleted Deep Sequencing Data Sheds More Light on the Early Wuhan SARS-CoV-2 Epidemic. Molecular Biology and Evolution. 38:5211–5224. doi:10.1093/molbev/msab246.

Bloom JD. 2023. Correction to: Recovery of Deleted Deep Sequencing Data Sheds More Light on the Early Wuhan SARS-CoV-2 Epidemic. Molecular Biology and Evolution. 40:msad201. doi:10.1093/molbev/msad201.

Bloom JD, Beichman AC, Neher RA, Harris K. 2023. Evolution of the SARS-CoV-2 Mutational Spectrum. Molecular Biology and Evolution. 40:msad085. doi:10.1093/molbev/msad085.

Bloom JD, Neher RA. 2023. Fitness effects of mutations to SARS-CoV-2 proteins. Virus Evolution. 9:vead055. https://doi.org/10.1093/ve/vead055.

Brunak S, Danchin A, Hattori M, Nakamura H, Shinozaki K, Matise T, Preuss D. 2002. Nucleotide Sequence Database Policies. Science. 298:1333–1333. https://www.science.org/doi/abs/10.1126/science.298.5597.1333b.

Chan JFW, Yuan S, Kok KH, To KKW, Chu H, Yang J, Xing F, Liu J, Yip CCY, Poon RWS *et al.* 2020. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster. The Lancet. 395:514–523. doi:10.1016/S0140-6736(20)30154-9.

Chen C, Nadeau S, Yared M, Voinov P, Xie N, Roemer C, Stadler T. 2022. CoV-Spectrum: Analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. Bioinformatics. 38:1735–1737. doi:10.1093/bioinformatics/btab856.

Daoyu Zhang. 2020. The Pan-SL-CoV/GD sequences may be from contamination. Technical report. Zenodo. doi:10.5281/ZENODO.5004213.

De Maio N, Walker CR, Turakhia Y, Lanfear R, Corbett-Detig R, Goldman N. 2021. Mutation Rates and Selection on Synonymous Mutations in SARS-CoV-2. Genome Biology and Evolution. 13:evab087. https://academic.oup.com/gbe/article/13/5/evab087/6251359.

Delaune D, Hul V, Karlsson EA, Hassanin A, Ou TP, Baidaliuk A, Gámbaro F, Prot M, Tu VT, Chea S *et al.* 2021. A novel sars-cov-2 related coronavirus in bats from cambodia. Nature communications. 12:6563. https://www.nature.com/articles/s41467-021-26809-4.

Eden JS, Rockett R, Carter I, Rahman H, De Ligt J, Hadfield J, Storey M, Ren X, Tulloch R, Basile K *et al.* 2020. An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. Virus Evolution. 6:veaa027.

EFSA Panel on Animal Health and Welfare (AHAW), Nielsen SS, Alvarez J, Bicout DJ, Calistri P, Canali E, Drewe JA, Garin-Bastuji B, Gonzales Rojas JL, Gortázar C *et al.* 2023. SARS-CoV-2 in animals: Susceptibility of animal species, risk for animal and public health, monitoring, prevention and control. EFSA Journal. 21. `https://efsa.onlinelibrary.wiley.com/doi/full/10.2903/j.efsa.2023.7822`.

Farkas C, Fuentes-Villalobos F, Garrido JL, Haigh J, Barría MI. 2020. Insights on early mutational events in SARS-CoV-2 virus reveal founder effects across geographical regions. PeerJ. 8:e9255. doi:10.7717/peerj.9255.

Holmes EC. 2024. The emergence and evolution of SARS-CoV-2. Annu. Rev.Virol.. p. *In Press.*

Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, Spitters C, Ericson K, Wilkerson S, Tural A *et al.* 2020. First Case of 2019 Novel Coronavirus in the United States. New England Journal of Medicine. 382:929–936.

Jiang XL, Zhang XL, Zhao XN, Li CB, Lei J, Kou ZQ, Sun WK, Hang Y, Gao F, Ji SX *et al.* 2020. Transmission Potential of Asymptomatic and Paucisymptomatic Severe Acute Respiratory Syndrome Coronavirus 2 Infections: A 3-Family Cluster Study in China. The Journal of Infectious Diseases. 221:1948–1952.

Kang M, Wu J, Ma W, He J, Lu J, Liu T, Li B, Mei S, Ruan F, Lin L *et al.* 2020. Evidence and characteristics of human-to-human transmission of SARS-CoV-2. Preprint. Epidemiology. doi:10.1101/2020.02.03.20019141.

Kumar S, Tao Q, Weaver S, Sanderford M, Caraballo-Ortiz MA, Sharma S, Pond SLK, Miura S. 2021. An Evolutionary Portrait of the Progenitor SARS-CoV-2 and Its Dominant Offshoots in COVID-19 Pandemic. Molecular Biology and Evolution. 38:3046–3059. doi:10.1093/molbev/msab118.

Liu J. 2020. *Epidemiological, Clinical and Viral Gene Evolution Characteristics of Important Emerging Infectious Diseases (SFTS and COVID-19)*. Ph.D. thesis.

Liu W, Liu P, Lei W, Jia Z, He X, Liu LL, Shi W, Tan Y, Zou S, Zhao X *et al.* 2022. Surveillance of SARS-CoV-2 in the environment and animal samples of the Huanan Seafood Market. Preprint. Research Square. doi:10.21203/rs.3.rs-1370392/v1.

Liu WJ, Liu P, Lei W, Jia Z, He X, Shi W, Tan Y, Zou S, Wong G, Wang J *et al.* 2023. Surveillance of SARS-CoV-2 at the Huanan Seafood Market. Nature. doi:10.1038/s41586-023-06043-2.

Okada P, Buathong R, Phuygun S, Thanadachakul T, Parnmen S, Wongboot W, Waicharoen S, Wacharapluesadee S, Uttayamakul S, Vachiraphan A *et al.* 2020. Early transmission patterns of coronavirus disease 2019 (COVID-19) in travellers from Wuhan to Thailand, January 2020. Eurosurveillance. 25.

Pekar J, Worobey M, Moshiri N, Scheffler K, Wertheim JO. 2021. Timing the SARS-CoV-2 index case in Hubei province. Science. 372:412–417.

Pekar JE, Magee A, Parker E, Moshiri N, Izhikevich K, Havens JL, Gangavarapu K, Malpica Serrano LM, Crits-Christoph A, Matteson NL *et al.* 2022. The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2. Science. 377:960–966. doi:10.1126/science.abp8337.

Pipes L, Wang H, Huelsenbeck JP, Nielsen R. 2021. Assessing Uncertainty in the Rooting of the SARS-CoV-2 Phylogeny. Molecular Biology and Evolution. 38:1537–1543. doi:10.1093/molbev/msaa316.

Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, Du Plessis L, Pybus OG. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature Microbiology. 5:1403–1407. doi:10.1038/s41564-020-0770-5.

Ruis C, Peacock TP, Polo LM, Masone D, Alvarez MS, Hinrichs AS, Turakhia Y, Cheng Y, McBroome J, Corbett-Detig R *et al.* 2023. A lung-specific mutational signature enables inference of viral and bacterial respiratory niche. Microbial Genomics. 9. https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.001018.

Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, Zhou Z, Yang J, Zhong J, Yang D *et al.* 2020. Genomic Diversity of Severe Acute Respiratory Syndrome–Coronavirus 2 in Patients With Coronavirus Disease 2019. Clinical Infectious Diseases. 71:713–720. doi:10.1093/cid/ciaa203.

Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, Zhou Z, Yang J, Zhong J, Yang D *et al.* 2021. Corrigendum to: Genomic Diversity of Severe Acute Respiratory Syndrome–Coronavirus 2 in Patients With Coronavirus Disease 2019. Clinical Infectious Diseases. 73:2374–2374. doi:10.1093/cid/ciab900.

Wang M, Fu A, Hu B, Tong Y, Liu R, Gu J, Liu J, Jiang W, Shen G, Zhao W *et al.* 2020a. Nanopore target sequencing for accurate and comprehensive detection of SARS-CoV-2 and other respiratory viruses. Preprint. Infectious Diseases (except HIV/AIDS). doi:10.1101/2020.03.04.20029538.

Wang M, Fu A, Hu B, Tong Y, Liu R, Liu Z, Gu J, Xiang B, Liu J, Jiang W *et al.* 2020b. Nanopore Targeted Sequencing for the Accurate and Comprehensive Detection of SARS-CoV-2 and Other Respiratory Viruses. Small. 16:2002169. doi:10.1002/smll.202002169.

World Health Organization. 2021. *WHO-convened Global Study of Origins of SARS-CoV-2: China Part: Joint WHO-China Study, 14 January-10 February 2021 : Joint Report*. WHO. https://www.who.int/publications/i/item/who-convened-global-study-of-origins-of-sars-cov-2-china-part.

Yan Y, Wu K, Chen J, Liu H, Huang Y, Zhang Y, Xiong J, Quan W, Wu X, Liang Y *et al.* 2021. Rapid Acquisition of High-Quality SARS-CoV-2 Genome via Amplicon-Oxford Nanopore Sequencing. Virologica Sinica. 36:901–912. doi:10.1007/s12250-021-00378-8.

Yen HL, Sit THC, Brackman CJ, Chuk SSY, Gu H, Tam KWS, Law PYT, Leung GM, Peiris M, Poon LLM *et al.* 2022. Transmission of SARS-CoV-2 delta variant (AY.127) from pet hamsters to humans, leading to onward human-to-human transmission: A case study. The Lancet. 399:1070–1078. doi:10.1016/S0140-6736(22)00326-9.

Supplementary figures and tables

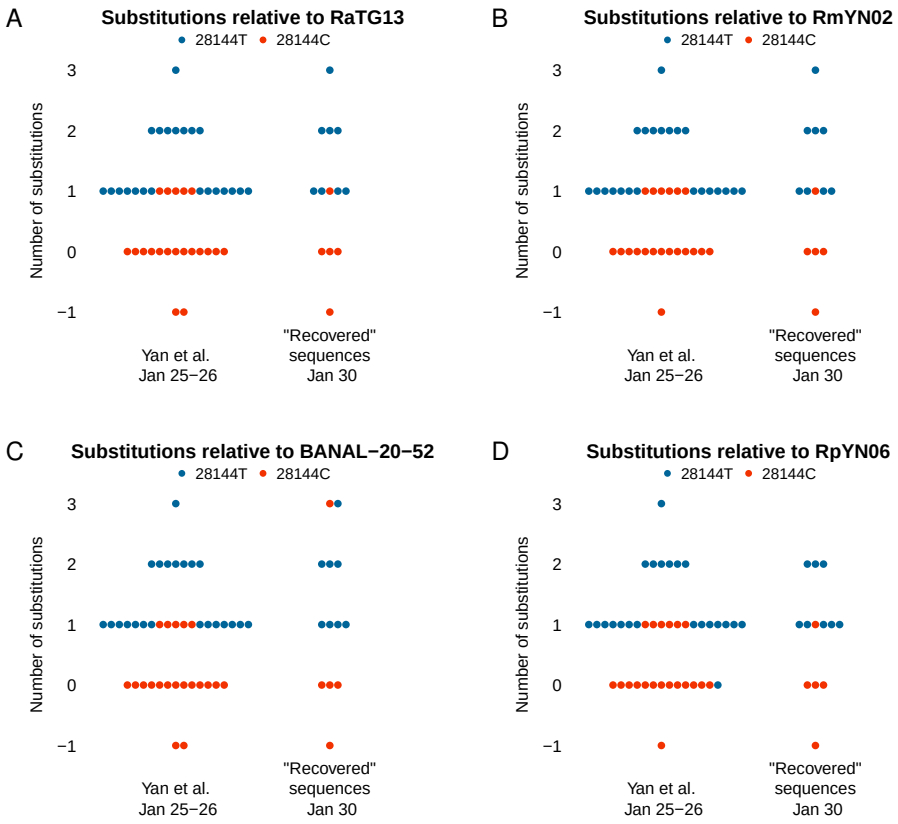


Figure S1: Equivalent of Figure 2, changing the outgroup comparator, shown as title of each panel.

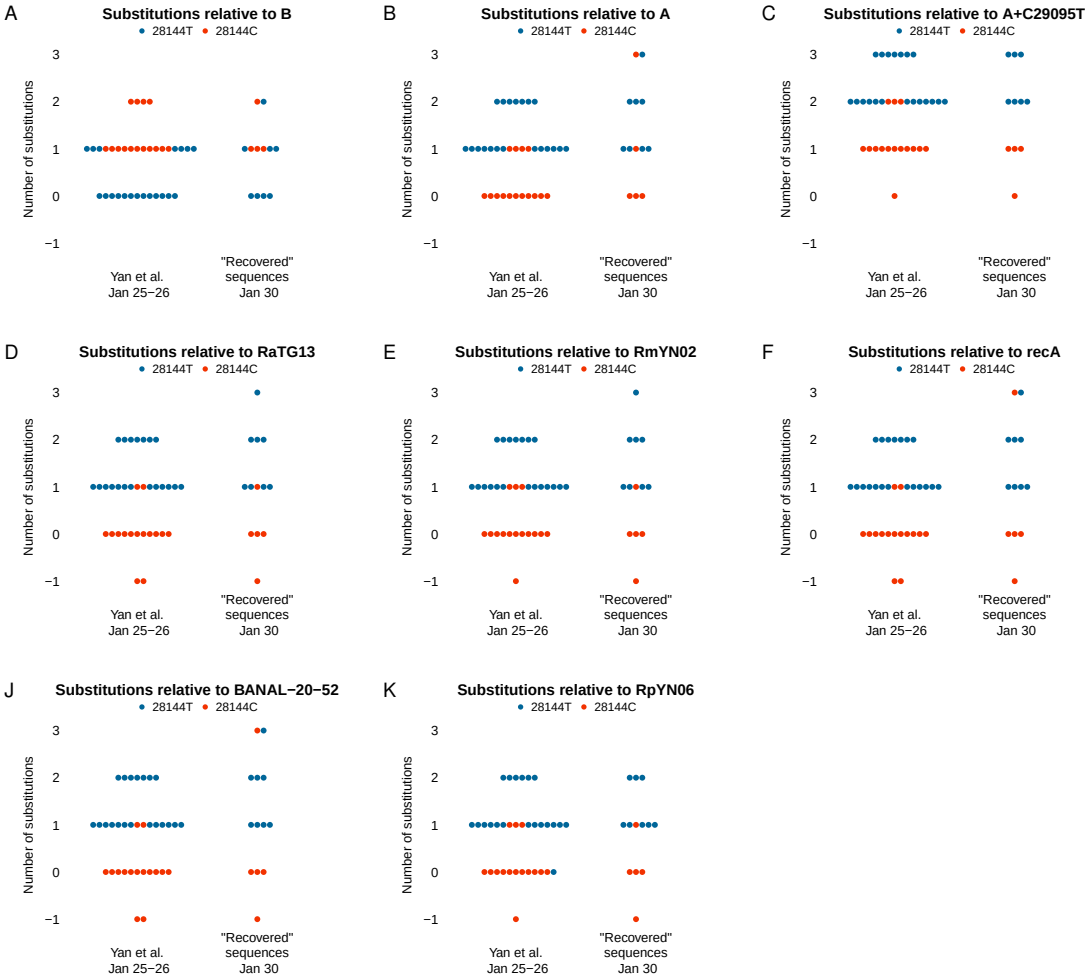


Figure S2: Equivalent of Figure 2, removing sequences with potential sequencing errors (Pekar *et al.* (2022) dataset), and changing the outgroup comparator, shown as title of each panel.

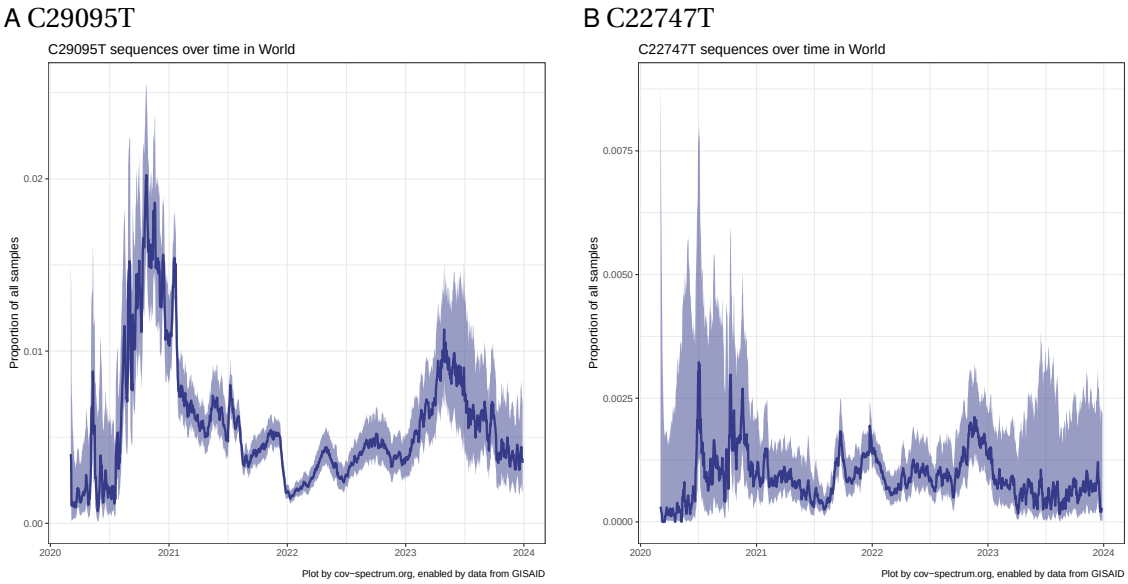


Figure S3: Proportion of sequences with C29095T and with C22747T among all sequences available on GISAID from 1 March 2020 through the end of 2023. Plots generated by CoV-Spectrum (Chen *et al.*, 2022), from `https://cov-spectrum.org/explore/World/AllSamples/from%3D2020-03-01%26to%3D2024-01-01/variants?nucMutations=C29095T&` and `https://cov-spectrum.org/explore/World/AllSamples/from%3D2020-03-01%26to%3D2024-01-01/variants?nucMutations=C22747T&` . Note the different vertical axis scales.

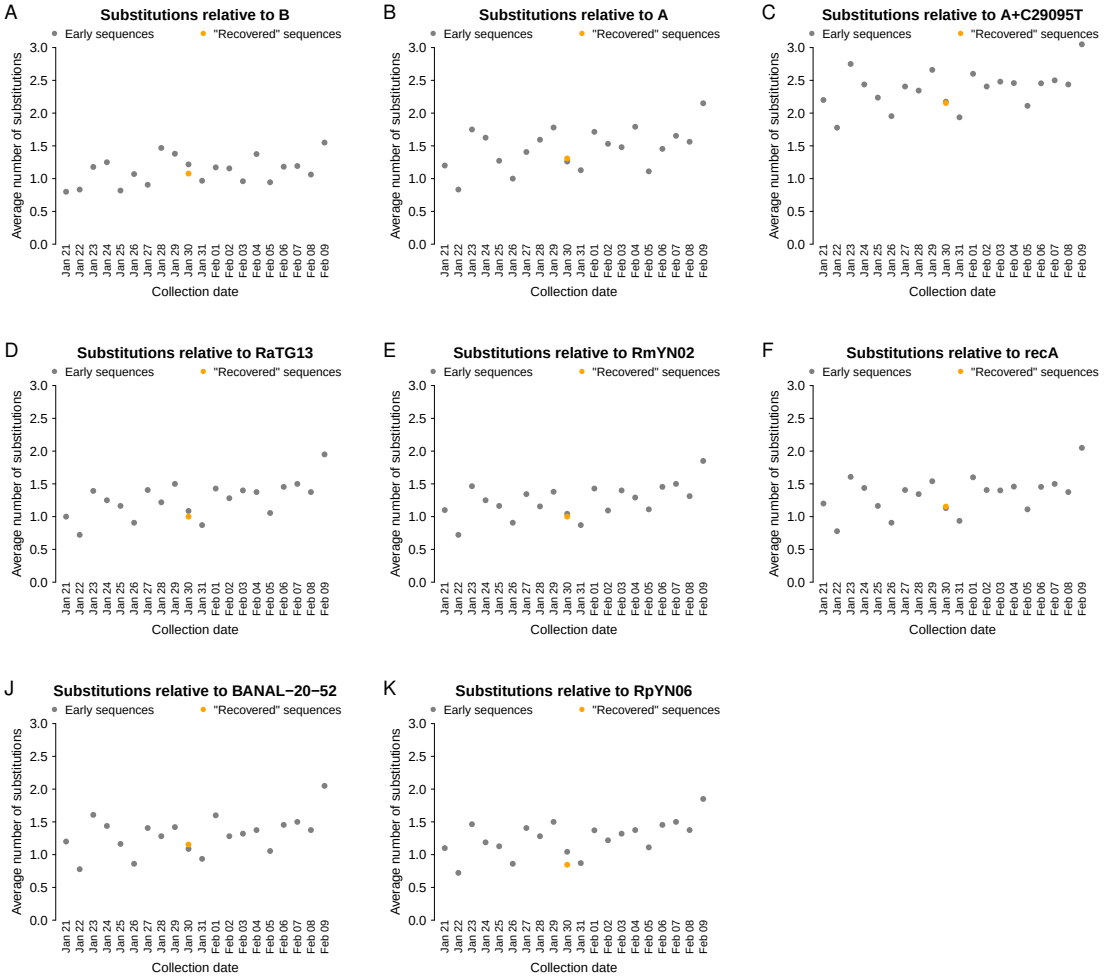


Figure S4: Equivalent of Figure 3, changing the outgroup comparator (shown as panel title).

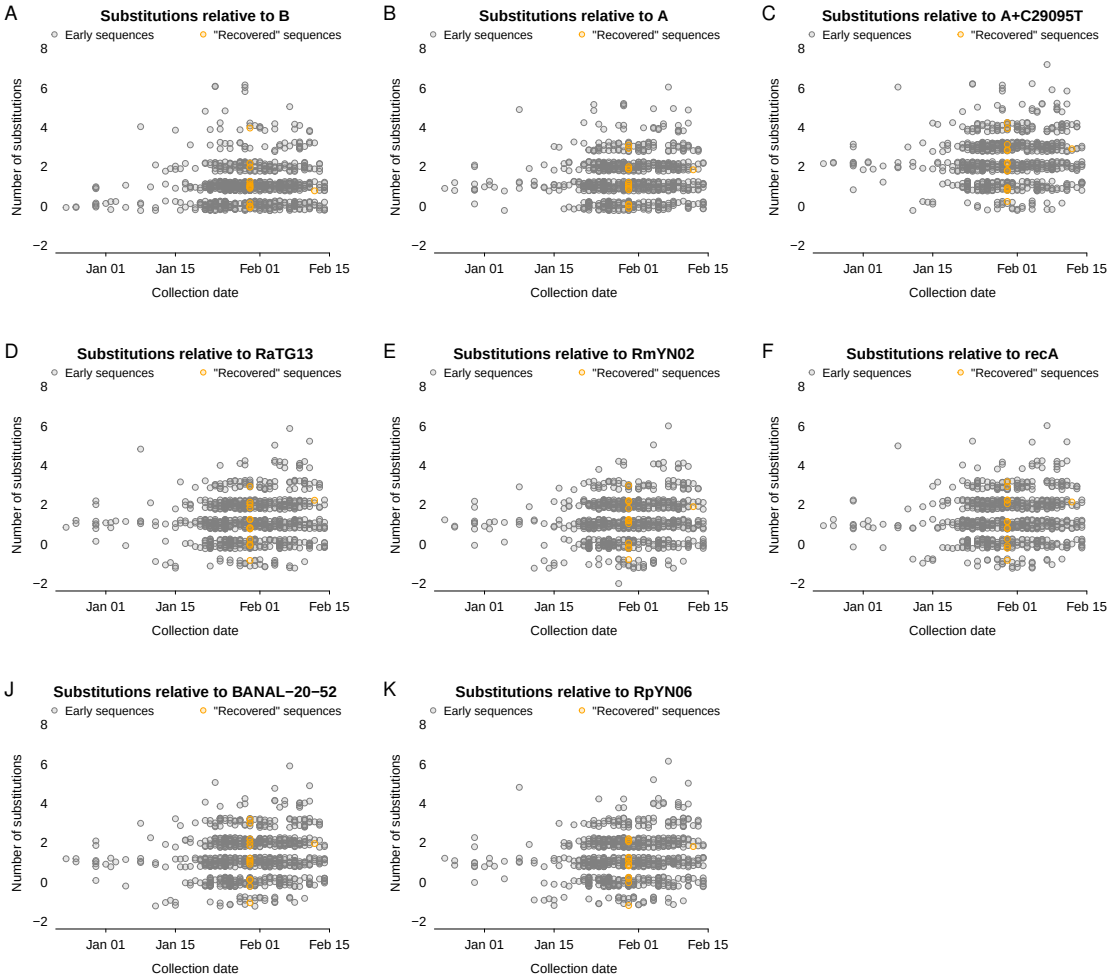


Figure S5: Equivalent of Figure 3, showing all points instead of averages and over a wider time window. The outgroup comparators are shown as panel titles.

ID	Age	Sex	Link	Onset	Hospitalization
SZ01	65	F	Mother of SH03	2020-01-03	2020-01-10
SZ02	66	M	Father of SH03	2020-01-04	2020-01-10
SZ03	37	F	Daughter of SH01 and SH02	2020-01-09	2020-01-11
SZ04	36	M	Son in law of SH01 and SH02	2020-01-05	2020-01-11
SZ05	10	M	Grandson of SH01 and SH02		2020-01-11
SZ06	63	F	Mother of SH04		2020-01-11

Table S1: Patients in the early January 2020 Shenzhen cluster. Hospitalization refers to the date of admission at HKU-SZH. Metadata from Chan *et al.* (2020).

Accession GISAID	ID	Collection	Source	Bloom label
`EPI_ISL_406592`	SZ01	2020-01-13	Yang *et al.*	
`EPI_ISL_403933`	SZ01	2020-01-15	Kang *et al.* (2020)	Guangdong patients
`EPI_ISL_406030`	SZ02	2020-01-10	Chan *et al.* (2020)	Guangdong patients
`EPI_ISL_406593`	SZ02	2020-01-13	Yang *et al.*	other China
`EPI_ISL_403932`	SZ02	2020-01-14	Kang *et al.* (2020)	Guangdong patients
`EPI_ISL_405839`	SZ05	2020-01-11	Chan *et al.* (2020)	other China
`EPI_ISL_403935`	SZ06	2020-01-15	Kang *et al.* (2020)	Guangdong patients

Table S2: Sequences in the early January 2020 Shenzhen cluster. Yang *et al.* are Yang Yang, Chenguang Shen, Li Xing, Zhixiang Xu, Haixia Zheng, Yingxia Liu, as listed on GISAID; we have not found a specific article presenting the sequences. The ID column corresponds to patient IDs introduced in Table S1.