Molecular Biology and Evolution, 2025, **42**, 1–7
https://doi.org/10.1093/molbev/msaf109
Advance access publication 9 June 2025
Brief Communications

MBE OXFORD

A Critical Reexamination of Recovered SARS-CoV-2 Sequencing Data

Florence Débarre [1,*] **Zach Hensel** [2]

[1]Institute of Ecology and Environmental Sciences, CNRS UMR 7618, Sorbonne Université, UPEC, IRD, INRAE, Paris, France
[2]Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Av. da República, 2780-157 Oeiras, Portugal

***Corresponding author**: E-mail: florence.debarre@normalesup.org.
Associate editor: Brandon Gaut

Abstract

In 2021, Jesse Bloom published a study addressing why the earliest SARS-CoV-2 sequences in Wuhan from late December 2019 were not those most similar to viruses sampled in bats. The study concluded that recovered partial sequences from Wuhan and annotation of Wuhan links for other sequences increased support for one genotype as the progenitor of the SARS-CoV-2 pandemic. However, we show that the collection date for the recovered sequences was January 30, 2020, later than that of hundreds of other SARS-CoV-2 sequences. Mutations in these sequences also exhibit diversity consistent with SARS-CoV-2 sequences collected in late January 2020. Furthermore, we found that Wuhan exposure history was common for early samples, so Bloom's annotation for a single familial cluster does not support that an early genotype was undersampled in Wuhan. Both the recovered partial sequences and additional annotation align with contemporaneous data rather than increase support for a progenitor. Our findings clarify the significance of the recovered sequences and are supported by additional data and analysis published since mid-2021.

Keywords: SARS-CoV-2, COVID-19, sequence read archive, phylogenetics, data recovery, metadata, epidemiology

Introduction

In 2021, Jesse Bloom published an analysis of early SARS-CoV-2 sequences (Bloom 2021) that included data generated from samples collected at Renmin Hospital of Wuhan University (Wang et al. 2020b). Bloom concluded that these data increased the plausibility of one genotype as the progenitor of SARS-CoV-2. The data were generated during the development of a diagnostic method based on sequencing partial SARS-CoV-2 genomes using nanopore technology (Wang et al. 2020b). Primarily aimed at detecting infections by SARS-CoV-2, the technique could also be used for genotyping. Wang et al. first shared their results in a preprint in March 2020 (Wang et al. 2020a). They submitted sequencing data a few days later to the sequence read archive (SRA; see supplementary table S1, Supplementary Material online for a timeline). The work was submitted to the journal *Small* in April, revised in May, and published in June 2020 (Wang et al. 2020b). The article included in its Table 1 a list of mutations identified in samples, and grouped samples in what was later described as lineage A and lineage B (Rambaut et al. 2020; Tang et al. 2020). Following *Small*'s format at the time, the article did not include a data availability statement (Wang et al. 2021; Zimmer 2021). The data depositor requested that sequencing data be withdrawn from the SRA. Withdrawal should have just excluded the data from search results, but instead the data were deleted (Berman et al. 2022; Brunak et al. 2002). Some relevant details about this particular issue can be found in supplementary table S1, Supplementary Material online, and in a preprint version of our work (Débarre and Hensel 2024). However, the deletion did not impact Bloom's data or analyses, on which we focus here.

In 2021, Jesse Bloom discovered Wang et al.'s sequencing data via another study on early pandemic sequencing data (Bloom 2021; Farkas et al. 2020). Bloom recovered the sequencing data from NCBI Google Cloud Storage and from a mirror of SRA data (Bloom 2021; Lifebit 2020). In his study, Jesse Bloom explored possible roots of the early SARS-CoV-2 phylogeny on a set of sequences collected before February 2020. He addressed a well-known conundrum (Rambaut et al. 2020), illustrated in supplementary fig. S1A, Supplementary Material online, that he summarized as: "the earliest reported sequences from Wuhan are *not* the sequences most similar to SARS-CoV-2's bat coronavirus relatives" (Bloom 2021). Using the principle of outgroup rooting (rather than molecular clock rooting that accounts for collection dates), Bloom manually rooted phylogenetic trees of early SARS-CoV-2 sequences at three nodes most similar to a bat-virus outgroup (see supplementary table S2, Supplementary Material online). Next, Bloom qualitatively inferred the relative degree of support for each root by considering the locations and dates at which samples were collected, identifying "two plausible progenitor sequences." The third proposed progenitor (A+C3171T) was considered less plausible because "it has almost no weight from Wuhan and the first sequence identical to its progenitor was not collected until January 24[, 2020]." The A+C18060T progenitor was kept because one sequence was sampled in Wuhan, although it was collected on January 26, 2020. For A+C29095T, Bloom noted the presence of an A+C29095T sequence among the Wang et al. data (positions

Received: February 23, 2024. **Revised:** December 16, 2024. **Accepted:** January 7, 2025

3,171 and 18,060 were not covered in these sequences, so Wang et al.'s data cannot inform on the plausibility of the corresponding proposed roots; see supplementary table S2, Supplementary Material online), and he annotated four sequences collected on January 10 to 15 from patients in Guangdong province with Wuhan travel history (Chan et al. 2020; Kang et al. 2020). Bloom distinguished the recovered sequences and the Guangdong patient sequences from the broader pre-February dataset by annotating them in figures as "deleted early Wuhan" and "Guangdong patient infected in Wuhan before January 5" (respectively; see supplementary fig. S1B, Supplementary Material online). The reported collection dates of recovered sequences were described as inconsistent with Wang et al.'s (2020b) text ("on or after January 30, 2020, *rather than* 'early in the epidemic' as originally described in Wang et al. (2020b)"; Bloom 2021, emphasis added). Bloom further wrote that it was "impossible to […] determine exactly when *[the recovered sequences]* were collected." Uncertainty in collection dates and the possibility that they could be among the "earliest" samples was also communicated in the popular and scientific news media (Chang and Langreth 2021; Cohen 2021; Pease 2021).

While compiling sequencing data and metadata for another project, we found that January 30, 2020 sample collection dates for Wang et al. (2020b)'s sequences were available on the SRA. According to the SRA team (email to FD, December 2023), these collection dates are identical to those provided by Wang et al. in March 2020. Here, we provide multiple lines of evidence showing that the January 30, 2020 collection date was correct. We show that the datasets from which Bloom obtained Wang et al. (2020b)'s data included collection date metadata (Farkas et al. 2020; Lifebit 2020). We additionally show how Guangdong patient sequences with A+C29095T are linked to the same Wuhan hospital, and that this genotype is not unique in its association with Wuhan. Lastly, we discuss how robust pandemic origins scenarios must rigorously account for available data, including sample collection dates, available metadata, and refined estimates of mutation rates.

Results

The January 30, 2020 Collection Date was Present in the Dataset from Wang *et al.* (2020)

In his article, Bloom claimed that the recovered sequences "lack[ed] full metadata," and that as a result it was "impossible to […] determine exactly when they were collected." However, while Bloom cited a press conference and two blog posts indirectly reporting January 30, 2020 collection dates (Wang 2021a, 2021b), collection dates were published earlier in July 2021 in a dataset cited by Bloom (PRJCA005725; Fig. 1a). The same date is present in the Farkas et al. (2020) table (Fig. 1b), from which Bloom identified Wang et al. (2020b) sequencing data. We further found the same collection dates in another dataset used in Bloom's study, a mirror of early pandemic sequencing data from Lifebit Life Sciences (Lifebit 2020). Bloom replaced collection dates of January 30, 2020 by "early in epidemic" while processing metadata during data analysis in a script listed in his Materials and Methods section, and did not report the collection dates. Ultimately, this resulted in labeling the recovered sequences as "deleted early Wuhan" (in his Fig. 5) rather than showing their collection date.

Following the logic of Bloom's paper, a January 30, 2020 collection date is too late to shift the likelihoods of proposed SARS-CoV-2 progenitor genotypes. For instance, the third proposed root (A+T3171C) was considered less plausible by him in part because "the first sequence identical to its progenitor was not collected until January 24." In addition, full genome sequences from samples collected before February 2020 were not rare: there were 507 such sequences in data considered by Bloom (2021). Wang et al. (2020b)'s sequences are therefore not exceptional.

Collection dates on or after January 30, 2020 have consistently been in Wang et al. 's sample metadata since March 2020. There is no evidence that this collection date could be inaccurate. This type of information is commonly extracted from metadata. Finally, collection dates being unreported or imprecisely described in corresponding scientific papers (or the lack of a corresponding paper) was not an exclusion criteria for other sequences in Bloom's study (e.g. his inclusion of late January Wuhan sequences from Yan et al. 2021). Nonetheless, occasional errors in chronological metadata have complicated accurately inferring phylogenetic trees during the pandemic (e.g. Sanderson 2024). We therefore investigated whether the composition of recovered sequences is consistent with January 30 collection dates, with analyses limited to Bloom (2021)'s dataset (i.e. sequences available in mid-2021).

Recovered Sequences are Consistent with Contemporaneous Data, Supporting Late January Sample Collection

We compared the data from Wang et al. to contemporaneous data in two different ways. First, we turn to Wuhan sequencing data generated via a similar nanopore-based technology as Wang et al. (2020b), reported in the context of an article by Yan et al. (2021). The samples were collected from "various Wuhan health care facilities" on January 25, 2020 and 26; consensus sequences were deposited on GISAID; they are included in Bloom's study. Two sequences from the Yan et al. (2021) dataset are present in proposed progenitor nodes in Bloom (2021): `C13` in the A+C18060T root, and `C31` in the A+C29095T root.

The distribution of substitutions in sequences from Yan et al. (2021) is similar to that of Wang et al. (2020b) (Fig. 2). In particular, the proportions of the lineage-A defining mutation T28144C are indistinguishable in the two datasets [17/42 in the Yan et al. (2021) data and 5/13 in the recovered sequences; Fisher's Exact Test, $P = 1$]; so are also the proportions of the C29095T mutation highlighted by Bloom [1/42 in the Yan et al. (2021) data and 1/13 in the recovered sequences; Fisher's Exact Test, $P = 0.42$]. The two distributions remain similar when the outgroup comparator is changed (supplementary fig. S2, Supplementary Material online).

Substitutions towards the chosen outgroup are not necessarily signs of their ancestral nature. The −1 positions of three sequences in Fig. 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 supplementary fig. S3, Supplementary Material online). 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 genotype in lineage A.

The comparison can be extended to the whole dataset used by Bloom (2021) (supplementary fig. S4, Supplementary Material

(a) Screenshot of CNCB metadata

Accession	SAMC430284
Sample name	C2-4h
Title	C2-4h
Sample type	Clinical or host-associated pathogen
Organism	Severe acute respiratory syndrome coronavirus 2
Description	Nanopore Targeted Sequencing for the Detection of SARS-CoV-2
Attributes	Collected by: Renmin Hospital of Wuhan University Collection date: 2020-01-30 Geographic location: China: Wuhan

(b) Screenshot of the Farkas *et al.* (2020) table

Sample Name	SRA Study	Collection_Date	geo_loc_name	collected_by
D12-4h	SRP252977	30-Jan-2020	China: Wuhan	Renmin Hospital of Wuhan University
D12-10min	SRP252977	30-Jan-2020	China: Wuhan	Renmin Hospital of Wuhan University
A1-4h	SRP252977	30-Jan-2020	China: Wuhan	Renmin Hospital of Wuhan University
D10-4h	SRP252977	30-Jan-2020	China: Wuhan	Renmin Hospital of Wuhan University
D10-10min	SRP252977	30-Jan-2020	China: Wuhan	Renmin Hospital of Wuhan University
C2-4h	SRP252977	30-Jan-2020	China: Wuhan	Renmin Hospital of Wuhan University
C2-10min	SRP252977	30-Jan-2020	China: Wuhan	Renmin Hospital of Wuhan University
C1-4h	SRP252977	30-Jan-2020	China: Wuhan	Renmin Hospital of Wuhan University

Fig. 1. The 30-January collection dates were available to Bloom and unchanged since first published in March 2020. a) Screenshot of metadata on CNCB (https://ngdc.cncb.ac.cn/biosample/browse/SAMC430284). b) Screenshot of metadata in Farkas et al. (2020) table (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 at https://github.com/jbloom/SARS-CoV-2_PRJNA612766/blob/main/manual_analyses/PRJNA612766/Supplementary_Table_1.xlsx). The file was available since 2020 and was used by Bloom to discover the Wang et al. sequencing data. Bloom replaced the date in his https://github.com/jbloom/SARS-CoV-2_PRJNA612766/blob/main/manual_analyses/PRJNA612766/extract_accessions.ipynb script.

online). Focusing on sequences collected ± 7 d around January 30, 2020, the proportions of key mutations such as T28144C and C29095T are, again, indistinguishable between Bloom's dataset and the recovered sequences (T28144C: 226/650 in Bloom (2021)'s dataset; Fisher's Exact Test, $P = 0.78$; C29095T: 30/650; Fisher's Exact Test, $P = 0.47$). Given the proportion of C29095T in the available sequences (4.6% in Bloom's dataset), there was a high chance of finding this specific mutation in at least one of the 13 additional sequences from Wang et al.. The recovered sequences dataset is, therefore, unremarkable; it is consistent with expectations for samples collected in Wuhan around January 30, 2020. Thus, identifying one example of A+C29095T in the recovered sequence dataset does not support the conclusion that A+C29095T was underrepresented in the earliest sequences.

The Guangdong Sequences with a Wuhan Exposure Were not Independent

In addition to identifying A+C29095T in one of Wang et al. (2020b)'s sequences, Bloom (2021) argued that A+C29095T was also supported as a progenitor by "many of the sequences […] from early patients who were infected in Wuhan but then sequenced in and attributed to Guangdong." This Wuhan link in a well-described cluster (Chan et al. 2020; Kang et al. 2020) had previously been identified in an early analysis of possible progenitor genotypes (Yu et al. 2020). Initially described by Bloom (2021) as "two different clusters of patients who traveled to Wuhan in late December of 2019," a correction now notes that there was one cluster rather than two (Bloom 2023), after we and others pointed it out. The change is significant, because it decreases the corresponding estimated prevalence of A+C29095T in Wuhan. We identified additional sequences from this cluster, sometimes multiple samples from the same patients, that were included in Bloom's dataset without this annotation (see supplementary tables S3 and S4, Supplementary Material online for details). Further, we note that patients in the 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 people infected by SARS-CoV-2 in Wuhan at that early date.

Analyzing early exports from Wuhan might characterize the Wuhan outbreak with less risk of ascertainment bias. We find that epidemiological links to Wuhan are very common in case reports from January 2020, and are not limited to A+C29095T sequences. Many other sequences in Bloom's trees were from direct exports from Wuhan, but were not labeled as such (see our annotations in supplementary fig. S1C, Supplementary Material online). 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+C18060T (Holshue et al. 2020). This is also true of early international exports of lineage A (Eden et al. 2020) and lineage B (Okada et al. 2020). Importantly, the earliest known export is a lineage B case with symptom onset predating any

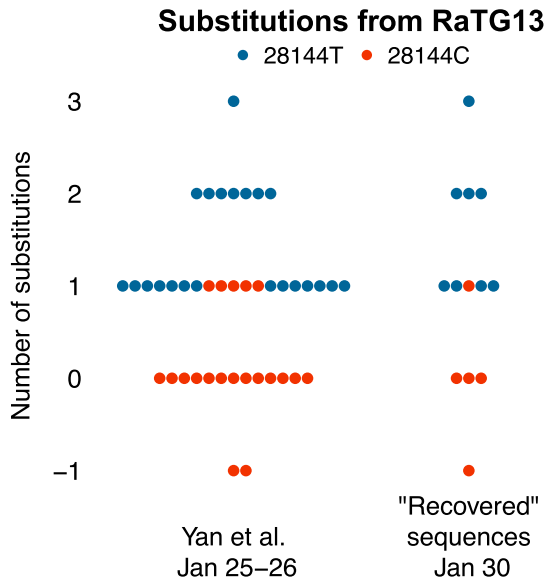

Fig. 2. Number of substitutions from bat SARS-like coronavirus RaTG13 in the region between nucleotides 21,570 to 29,550 that is considered in Fig. 4 in Bloom (2021) [relative to lineage A, which is equivalent to lineage A+C18060T {"proCoV2"} in this region, as position 18,060 is not covered]. Sequences 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.)

case in the Guangdong cluster by almost 2 weeks, and it is linked to the Huanan market (Liu 2020).

New Data Published Since Mid-2021 Fail to Support an A+C29095T Progenitor

First, we checked whether our conclusions still held using a second dataset compiled using stringent quality control (Pekar et al. 2022), to which we added recently published sequences (Lv et al. 2024), totaling 448 sequences collected before February 2020. We find that the number of substitutions in the Wang et al. (2020b) dataset remains consistent with those observed in other sequences with similar collection dates (supplementary figs. S5 and S6, Supplementary Material online).

We also investigated whether sequences reported by Lv et al. (2024) would support Bloom's conclusions, which were based on identifying early Wuhan sequences with the proposed progenitor genotypes. None of the eight high-coverage sequences collected by Lv et al. (2024) before February 2020 would fall into a progenitor node in Bloom's phylogenetic trees: the one sequence with C29095T, collected on January 29, 2020, also has G25947T and would therefore be placed in another, derived, node with sequences from samples collected in Wuhan and Shanghai. We found that another Lv et al. (2024) sequence with C29095T, sampled on February 3, 2020, had an identical genotype and comparable metadata to an already existing sequence from Shanghai. Further inspection showed that early sequences in Lv et al. (2024) were independent samples obtained from the same cohort of patients (laboratory-confirmed COVID-19 patients hospitalized at Shanghai Public Health Clinical Center between January 20, 2020 and February 25, 2020) as another study (Zhang et al. 2020). Lv et al. (2024)'s sequences therefore do not bring support to Bloom (2021)'s conclusions.

Lv et al. (2024) also report direct Wuhan links for 8 of 13 Shanghai patients with epidemiological history and collection dates before February 2020. This further emphasizes the need to thoroughly annotate epidemiological links rather than draw conclusions from partial annotation.

The history of the Guangdong cluster indicated that the C29095T substitution was present in Wuhan in late December 2019; it is therefore unsurprising that C29095T was detected in late January 2020 in Wuhan by Yan et al. (2021) and by Wang et al. (2020b). Methods in Bloom (2021) did not account for the fact 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). Subsequent work on this topic (Bloom et al. 2023; Ruis et al. 2023) has even specifically identified C29095T as occurring much more frequently than a typical C→T mutation (Bloom and Neher 2023, supplementary data `nt_fitness.csv`). Of all mutations considered by Bloom and Neher (2023)(three possible SNPs at each of 29295 positions), C29095T is the ninth most frequent one, making it more likely to be derived than ancestral compared to the other potential progenitors considered by Bloom (2021). In fact, C29095T recurs in Bloom's phylogenetic trees, where this position mutates three times.

Finally, Bloom (2021) noted that all sequences linked to the Huanan market were of lineage B, three mutations away from his proposed progenitors. At the time, this observation had led to the suggestion that the market had been a place of secondary amplification, but not the source of the outbreak. Later analyses and data challenged this conclusion. The two lineage-A cases with onset in December 2019 were shown to be geographically associated with the market (Worobey 2021; Worobey et al. 2022), leading to the prediction that lineage A was in the market. Then, lineage A was detected in an environmental sample from the market (Liu et al. 2024): in a stall with a suspected case with mid-December 2019 onset, and trace evidence in one sample from a different stall (Crits-Christoph et al. 2024). Lastly, the predominance of lineage B among Wuhan samples was observed beyond potential ascertainment bias linked to the Huanan market. A study genotyping random samples from Wuhan patients admitted to five hospitals found a consistent predominance of lineage B (Hu et al. 2021), mirroring what was observed for early pandemic sequences collected in Wuhan and also globally.

Discussion

The facts that we present do not support Bloom (2021)'s conclusion that Wang et al. (2020b)'s sequences demonstrate that A+C29095T was underrepresented in the earliest sequences, nor his conclusion that they increase the plausibility of A+C29095T as the progenitor genotype of SARS-CoV-2. Wang et al. (2020b)'s samples were collected in late January 2020, and mutations identified in these samples, including C29095T, are unsurprising to find again in Wuhan. Further, links to Wuhan were common, and annotating them just for sequences in one cluster does not distinguish the A+C29095T sequences as more likely to be ancestral than others.

There was no contradiction in Wang et al. (2020b)'s 2020 and 2021 statements on collection dates: the date they published in 2021 (January 30, 2020) was the same as the date

they had submitted to the SRA in March 2020, and the term "early in epidemic" is not inconsistent with the date. The meaning of "early in the epidemic" is context-dependent: "early" at Renmin Hospital is later than at hospitals closer to the epicenter of the outbreak, but likely earlier than at hospitals outside of Wuhan. Wang et al. (2020b) used samples collected while fever clinics at Renmin Hospital were suddenly overwhelmed with demand for molecular testing of suspected Covid-19 1 week after the beginning of Wuhan's lockdown (Liu et al. 2020) and only a few days after Renmin hospital was still sending staff to supplement Jinyintan hospital where the earliest patients were sent for treatment (Hubei Daily 2020; Yang 2024). "Early," however, does not necessarily mean "the earliest." Late January is not so early that the Wang et al. (2020b) sequences can support the conclusion that A+C29095T was underpreresented in the earliest samples collected from Covid-19 patients. It is also critical to understand the context in which scientists in Wuhan were working in early 2020. While Bloom (2021) criticized Wang et al. (2020b) for not fully sequencing their samples, the fact that full genome sequencing was not prioritized in late January 2020 in Wuhan realistically reflected prioritizing capabilities for patient diagnosis, isolation, and treatment.

Although partial SARS-CoV-2 sequences are of limited value for phylogenetic studies, the data shared by Wang et al. (2020b) contained information. The mutations reported in Wang et al. (2020b)'s Table 1 include additional mutations from early SARS-CoV-2 lineages, notably one mutation (A24325G, sample B9) found in a sample from a Huanan market vendor (World Health Organization 2021). In addition, other Wang et al. samples are in sublineages derived from lineage A and lineage B, indicating that these are not samples from patients with the earliest infections of the pandemic. To our knowledge, no subsequent phylogenetic analysis since Bloom (2021)'s paper has used Wang et al. (2020b)'s data. Had assembled versions of Wang et al. (2020b)'s sequences been shared on GISAID, they would have been excluded in Bloom (2021)'s analysis, because of their partial coverage.

The question of the precise identity of SARS-CoV-2's root remains unresolved. Using a method that accounts for both sample collection dates and the sequences of related bat coronaviruses, lineage A without additional mutations was identified as the most likely genotype of the common ancestor of SARS-CoV-2 sequences from humans (Crits-Christoph et al. 2024; Pekar et al. 2022). This analysis strongly rejects A+C18060T and A+C29095T as ancestral haplotypes (Crits-Christoph et al. 2024, supplementary Table S1, Supplementary Material online). To address the rooting conundrum, Pekar et al. (2022) considered another SARS-CoV-2 origin scenario, involving multiple SARS-CoV-2 spillovers from animals to humans, with lineage A spillover likely occurring after lineage B. Such a scenario of multiple transmissions close in time and space, from a group of animals to humans, also occurred later in the pandemic (AHAW et al. 2023), notably with pet hamsters in Hong Kong, for which a genomic investigation identified multiple zoonotic spillovers (Yen et al. 2022). Low diversity in coronavirus genomes identified in samples from bats at the same time and place is also common; for example, RshSTT182 and RshSTT200 genomes differ by only three nucleotides (Delaune et al. 2021).

Bloom (2021) addressed the conundrum that "more ancestral" genotypes are not among the SARS-CoV-2 sequences with the earliest collection dates in Wuhan, by considering that an A+C29095T or A+C18060T progenitor may be underrepresented because of uneven sampling and/or reporting. We have shown that additional evidence considered by Bloom—the identification of one A+C29095T sequence collected in Wuhan on January 30, 2020, and the annotation of a Wuhan link for four sequences in one cluster—does not actually increase support for an A+C29095T progenitor. Our analysis was robust when considering recently published sequences (Lv et al. 2024). It is critical that conclusions be continuously tested by considering all available data, including sample collection dates, and justifying (meta)data exclusion.

Materials and Methods

We followed the same methods as Bloom (2021) to compare sequences to outgroups.

Source Data

We used data shared by Bloom on Github at https://github.com/jbloom/SARS-CoV-2_PRJNA612766.

For the expanded dataset, we used outputs of a dataset curated by Pekar et al. (2022), complemented by data recently shared by Lv et al. (2024) (selecting the earliest informative sequence for each patient as described in Crits-Christoph et al. (2024); Patients IDs are provided in Lv et al. (2024)'s appendices. Selected accessions are listed in the associated Zenodo repository.) 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. GISAID 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 and are included in the data analyzed by Bloom (2021).

Collection dates are available in SRA metadata in the `Collection_Date` field.

Rank of C29095T

The rank of rate of C29095T substitution was extracted from the results of a previously reported analysis (Bloom and Neher 2023) using an updated analysis from 06-Nov-2024 (commit `067fce1`) https://github.com/jbloomlab/SARS2-mut-fitness/blob/main/results/nt_fitness/ntmut_fitness_all.csv. In this analysis, the frequencies of occurrences of individual SNPs are extracted from the UShER tree of public SARS-CoV-2 sequences (Turakhia et al. 2021), excluding a small fraction of masked sites and likely sequencing artifacts. We ranked SNPs by the number of occurrences. The corresponding code is available in the Zenodo repository.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

F.D. thanks the SRA team for their answers to her questions. We thank Alex Crits-Christoph, Jonathan Pekar, Joel Wertheim, and Mike Worobey for comments and discussions. Zhihua Chen first spotted the two Guangdong clusters error in Bloom (2021). F.D. thanks Dake Kang for providing details on the Chinese law behind the destruction of samples (State Council of the People's Republic of China 2004), and for

explaining that it was an older regulation, not specific to Covid-19. F.D. thanks Wiley for sharing details on the Wang et al. (2020b) publication process and on their investigation. 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).

Funding

Z.H. was supported by FCT - Fundação para a Ciência e a Tecnologia, I.P., through MOSTMICRO-ITQB R&D Unit (https://doi.org/10.54499/UIDB/04612/2020; https://doi.org/10.54499/UIDP/04612/2020) and LS4FUTURE Associated Laboratory (https://doi.org/10.54499/LA/P/0087/2020).

Data Availability

Data and code are available on Zenodo (https://zenodo.org/doi/10.5281/zenodo.10665464). We used the same sequence datasets as Bloom (2021) (accessions listed in https://github.com/jbloom/SARS-CoV-2_PRJNA612766/blob/main/results/early_sequences/deltadist.csv). Additional analysis with more recent data used the same accessions as Crits-Christoph et al. (2024) (listed in `data/accessions_P22-Lv24.csv` in the Zenodo repository.) The Wang et al. (2020b) sequences are available in project PRJCA005725 on CNCB.

References

Azgari C, Kilinc Z, Turhan B, Circi D, Adebali O. The mutation profile of SARS-CoV-2 is primarily shaped by the host antiviral defense. *Viruses*. 2021:13(3):394. https://doi.org/10.3390/v13030394.

Berman A, Boykin L, Ceasar M, Sowa A, Twigger S. NIH/NLM: root cause analysis: removal of SRA sequence data records. Technical Report. BioTeam, Inc; 2022. https://ftp.ncbi.nlm.nih.gov/sra/doc/BioTeam-RCA-RedactedReport.pdf.

Bloom JD. Recovery of deleted deep sequencing data sheds more light on the early Wuhan SARS-CoV-2 epidemic. *Mol Biol Evol*. 2021:38(12):5211–5224. https://doi.org/10.1093/molbev/msab246.

Bloom JD. Correction to: recovery of deleted deep sequencing data sheds more light on the early Wuhan SARS-CoV-2 epidemic. *Mol Biol Evol*. 2023:40(4):msad201. https://doi.org/10.1093/molbev/msad085.

Bloom JD, Beichman AC, Neher RA, Harris K. Evolution of the SARS-CoV-2 mutational spectrum. *Mol Biol Evol*. 2023:40(4):msad085. https://doi.org/10.1093/molbev/msad085.

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

Brunak S, Danchin A, Hattori M, Nakamura H, Shinozaki K, Matise T, Preuss D. Nucleotide sequence database policies. *Science*. 2002:298(5597):1333–1333. https://doi.org/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.* A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet*. 2020:395(10223):514–523. https://doi.org/10.1016/S0140-6736(20)30154-9.

Chang R, Langreth R. U.S. confirms removal of Wuhan virus sequences from database. Bloomberg; 2021. https://www.bloomberg.com/news/articles/2021-06-24/u-s-confirms-removal-of-wuhan-virus-sequences-from-database.

Cohen J. Claim that Chinese team hid early SARS-CoV-2 sequences to stymie origin hunt sparks furor. Science. 2021. https://www.science.org/content/article/claim-chinese-team-hid-early-sars-cov-2-sequences-stymie-origin-hunt-sparks-furor.

Crits-Christoph A, Levy JI, Pekar JE, Goldstein SA, Singh R, Hensel Z, Gangavarapu K, Rogers MB, Moshiri N, Garry RF, *et al.* Genetic tracing of market wildlife and viruses at the epicenter of the COVID-19 pandemic. *Cell*. 2024:187(19):5468–5482.e11. https://doi.org/10.1016/j.cell.2024.08.010.

Débarre F, Hensel Z. A critical reexamination of recovered SARS-CoV-2 sequencing data. bioRxiv 580500. https://doi.org/10.1101/2024.02.15.580500, 2024, preprint: not peer reviewed.

Delaune D, Hul V, Karlsson EA, Hassanin A, Ou TP, Baidaliuk A, Gámbaro F, Prot M, Tu VT, Chea S, *et al.* A novel SARS-CoV-2 related coronavirus in bats from Cambodia. *Nat Commun*. 2021:12(1):6563. https://doi.org/10.1038/s41467-021-26809-4.

De Maio N, Walker CR, Turakhia Y, Lanfear R, Corbett-Detig R, Goldman N. Mutation rates and selection on synonymous mutations in SARS-CoV-2. *Genome Biol Evol*. 2021:13(5):evab087. https://doi.org/10.1093/gbe/evab087.

Eden JS, Rockett R, Carter I, Rahman H, De Ligt J, Hadfield J, Storey M, Ren X, Tulloch R, Basile K, *et al.* An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus Evol*. 2020:6(1):veaa027. https://doi.org/10.1093/ve/veaa027.

Farkas C, Fuentes-Villalobos F, Garrido JL, Haigh J, Barría MI. Insights on early mutational events in SARS-CoV-2 virus reveal founder effects across geographical regions. *PeerJ*. 2020:8(1):e9255. https://doi.org/10.7717/peerj.9255.

Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, Spitters C, Ericson K, Wilkerson S, Tural A, *et al.* First case of 2019 novel coronavirus in the United States. *N Engl J Med*. 2020:382(10):929–936. https://doi.org/10.1056/NEJMoa2001191.

Hu B, Liu R, Tang X, Pan Y, Wang M, Tong Y, Ye G, Shen G, Ying R, Fu A, *et al.* The concordance between the evolutionary trend and the clinical manifestation of the two SARS-CoV-2 variants. *Natl Sci Rev*. 2021:8(8):nwab073. https://doi.org/10.1093/nsr/nwab073.

Hubei Daily. To All the Angels in White, Thank You for Your Hard Work!. 2020. https://news.cri.cn/20200127/615de051-1480-73bf-1d61-7385deaa3ca6.html — https://archive.is/7yYKb.

Jiang XL, Zhang XL, Zhao XN, Li CB, Lei J, Kou ZQ, Sun WK, Hang Y, Gao F, Ji SX, *et al.* Transmission potential of asymptomatic and paucisymptomatic severe acute respiratory syndrome coronavirus 2 infections: a 3-family cluster study in China. *J Infect Dis*. 2020:221(12):1948–1952. https://doi.org/10.1093/infdis/jiaa206.

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

Lifebit. Lifebit SARS-CoV-2 Dataset. https://blog-assets-lifebit.s3-eu-west-1.amazonaws.com/SARS-CoV-2+Public+Dataset.pdf (archived at https://web.archive.org/web/20240615185317/https://blog-assets-lifebit.s3-eu-west-1.amazonaws.com/SARS-CoV-2+Public+Dataset.pdf); sample metadata directly downloaded from SRA available at https://lifebit-sars-cov-2.s3-eu-west-1.amazonaws.com/reads/SraRunTable.txt (archived at https://web.archive.org/web/20240825203435/https://lifebit-sars-cov-2.s3-eu-west-1.amazonaws.com/reads/SraRunTable.txt).2020.

Liu J. Epidemiological, clinical and viral gene evolution characteristics of important emerging infectious diseases (SFTS and COVID-19) [PhD thesis]. 2020.

Liu R, Han H, Liu F, Lv Z, Wu K, Liu Y, Feng Y, Zhu C. Positive rate of RT-PCR detection of SARS-CoV-2 infection in 4880 cases from one hospital in Wuhan, China, from Jan to Feb 2020. *Clin Chim Acta*. 2020:505:172–175. https://doi.org/10.1016/j.cca.2020.03.009.

Liu WJ, Liu P, Lei W, Jia Z, He X, Shi W, Tan Y, Zou S, Wong G, Wang J, *et al.* Surveillance of SARS-CoV-2 at the Huanan seafood market. *Nature*. 2024:631(8020):402–408. https://doi.org/10.1038/s41586-023-06043-2.

Lv JX, Liu X, Pei YY, Song ZG, Chen X, Hu SJ, She JL, Liu Y, Chen YM, Zhang YZ. Evolutionary trajectory of diverse SARS-CoV-2 variants at the beginning of COVID-19 outbreak. *Virus Evol*. 2024:10(1):veae020. https://doi.org/10.1093/ve/veae020.

AHAW, Nielsen SS, Alvarez J, Bicout DJ, Calistri P, Canali E, Drewe JA, Garin-Bastuji B, Gonzales Rojas JL, Gortázar C, *et al.* SARS-CoV-2

in animals: susceptibility of animal species, risk for animal and public health, monitoring, prevention and control. *EFSA J*. 2023:21(2): e07822. https://doi.org/10.2903/j.efsa.2023.7822.

Okada P, Buathong R, Phuygun S, Thanadachakul T, Parnmen S, Wongboot W, Waicharoen S, Wacharapluesadee S, Uttayamakul S, Vachiraphan A, *et al*. Early transmission patterns of coronavirus disease 2019 (COVID-19) in travellers from Wuhan to Thailand, January 2020. *Euro Surveill*. 2020:25(8):2000097. https://doi.org/10.2807/1560-7917.ES.2020.25.8.2000097.

Pease R. Tales of unexpected DNA data. 2021-06-24, 2021. https://www.bbc.co.uk/sounds/play/w3ct1l3t.

Pekar JE, Magee A, Parker E, Moshiri N, Izhikevich K, Havens JL, Gangavarapu K, Malpica Serrano LM, Crits-Christoph A, Matteson NL, *et al*. The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2. *Science*. 2022:377(6609):960–966. https://doi.org/10.1126/science.abp8337.

Rambaut A, Holmes EC, O'Toole Á, McCrone JT, Ruis C, Du Plessis L, Pybus OG. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*. 2020:5(11): 1403–1407. https://doi.org/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*. A lung-specific mutational signature enables inference of viral and bacterial respiratory niche. *Microb Genom*. 2023:9(5):mgen001018. https://doi.org/10.1099/mgen.0.001018.

Sanderson T. Chronumental: time tree estimation from very large phylogenies. 2024. https://doi.org/10.1101/2021.10.27.465994.

State Council of the People's Republic of China. Regulations on the biosafety management of pathogenic microbiology laboratories. 2004. https://www.gov.cn/gongbao/content/2019/content_5468882.htm.

Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, *et al*. On the origin and continuing evolution of SARS-CoV-2. *Natl Sci Rev*. 2020:7(6):1012–1023. https://doi.org/10.1093/nsr/nwaa036.

Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, Haussler D, Corbett-Detig R. Ultrafast sample placement on existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat Genet*. 2021:53(6):809–816. https://doi.org/10.1038/s41588-021-00862-7.

Wang M, Fu A, Hu B, Tong Y, Liu R, Gu J, Liu J, Jiang W, Shen G, Zhao W, *et al*. Nanopore target sequencing for accurate and comprehensive detection of SARS-CoV-2 and other respiratory viruses. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2020.03.04.20029538, 2020a, preprint: not peer reviewed.

Wang M, Fu A, Hu B, Tong Y, Liu R, Liu Z, Gu J, Xiang B, Liu J, Jiang W, *et al*. Nanopore targeted sequencing for the accurate and comprehensive detection of SARS-CoV-2 and other respiratory viruses. *Small*. 2020b:16(32):2002169. https://doi.org/10.1002/smll.202002169.

Wang M, Fu A, Hu B, Tong Y, Liu R, Liu Z, Gu J, Xiang B, Liu J, Jiang W, *et al*. Correction: Nanopore targeted sequencing for the accurate and comprehensive detection of SARS-CoV-2 and other respiratory viruses. *Small*. 2021:17(32):2104078. https://doi.org/10.1002/smll.202104078.

Wang Z. The Chinese side of the COVID data withdrawal controversy. 2021a. https://www.pekingnology.com/p/the-chinese-side-of-the-covid-data (Interview conducted by Yang Liu).

Wang Z. Why did Wuhan University researchers delete COVID-19 data at NIH? 2021b. https://www.pekingnology.com/p/why-did-wuhan-university-researchers.

World Health Organization. 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.2021.

Worobey M. Dissecting the early COVID-19 cases in Wuhan. *Science*. 2021:374(6572):1202–1204. https://doi.org/10.1126/science.abm4454.

Worobey M, Levy JI, Serrano LM, Crits-Christoph A, Pekar JE, Goldstein SA, Rasmussen AL, Kraemer MUG, Newman C, Koopmans MPG, *et al*. The Huanan seafood wholesale market in Wuhan was the early epicenter of the COVID-19 pandemic. *Science*. 2022:377(6609):951–959. https://doi.org/10.1126/science.abp8715.

Yan Y, Wu K, Chen J, Liu H, Huang Y, Zhang Y, Xiong J, Quan W, WuX, Liang Y, *et al*. Rapid acquisition of high-quality SARS-CoV-2 genome via amplicon-Oxford Nanopore sequencing. *Virol Sin*. 2021:36(5):901–912. https://doi.org/10.1007/s12250-021-00378-8.

Yang D. *Wuhan: how the COVID-19 outbreak in China spiraled out of control*. Oxford: Oxford University Press; 2024.

Yen HL, Sit THC, Brackman CJ, Chuk SSY, Gu H, Tam KWS, Law PYT, Leung GM, Peiris M, Poon LLM, *et al*. Transmission of SARS-CoV-2 delta variant (AY.127) from pet hamsters to humans, leading to onward human-to-human transmission: a case study. *Lancet*. 2022:399(10329):1070–1078. https://doi.org/10.1016/S0140-6736(22)00326-9.

Yu WB, Tang GD, Zhang L, Corlett RT. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2/HCoV-19) using whole genomic data. *Zool Res*. 2020:41(3): 247–257. https://doi.org/10.24272/j.issn.2095-8137.2020.022.

Zhang X, Tan Y, Ling Y, Lu G, Liu F, Yi Z, Jia X, Wu M, Shi B, Xu S, *et al*. Viral and host factors related to the clinical outcome of COVID-19. *Nature*. 2020:583(7816):437–440. https://doi.org/10.1038/s41586-020-2355-0.

Zimmer C. Those Virus Sequences That Were Suddenly Deleted? They're Back. The New York Times. 2021. https://www.nytimes.com/2021/07/30/science/coronavirus-sequences-lab-leak.html.