




Brief Report

Unwarranted Exclusion of Intermediate Lineage A-B SARS-CoV-2 Genomes Is Inconsistent with the Two-Spillover Hypothesis of the Origin of COVID-19

Steven E. Massey ^{1,*}, Adrian Jones ² , Daoyu Zhang ³, Yuri Deigin ⁴  and Steven C. Quay ⁵ 

¹ Biology Department, University of Puerto Rico-Rio Piedras, San Juan, PR 00925, USA

² Independent Researcher, Melbourne, VIC 3000, Australia

³ Independent Researcher, Sydney, NSW 2120, Australia

⁴ Youthereum Genetics Inc., Toronto, ON L4J 8G9, Canada

⁵ Atossa Therapeutics, Inc., Seattle, WA 98104, USA

* Correspondence: steven.massey@upr.edu

Abstract: Pekar et al. (2022) propose that SARS-CoV-2 was a zoonotic spillover that first infected humans in the Huanan Seafood Market in Wuhan, China. They propose that there were two separate spillovers of the closely related lineages A and lineage B in a short period of time. The two lineages are differentiated by two SNVs; hence, a single-SNV A-B intermediate must have occurred in an unsampled animal host if the two-spillover hypothesis is correct. Consequently, confirmation of the existence of an intermediate A-B genome from humans would falsify their hypothesis of two spillovers. Pekar et al. identified and excluded 20 A-B intermediate genomes from their analysis. A variety of exclusion criteria were applied, including low read depth and the assertion of repeated erroneous base calls at lineage-defining positions 8782 and 28144. However, data from GISAID show that most of the genomes were sequenced to high average sequencing depth, appearing inconsistent with these criteria. The decision to exclude the majority of genomes was based on personal communications, with raw data unavailable for inspection. Multiple errors, biases, and inconsistencies were observed in the exclusion process. For example, 12 intermediate genomes from one study were excluded; however, 54 other genomes from the same study were included, indicating selection bias. Puzzlingly, two intermediate genomes from Beijing were discarded despite an average sequencing depth of 2175X; however, four genomes from the same sequencing study were included in the analysis. Lastly, we discuss 14 additional possible intermediate genomes not discussed by Pekar et al. and note that genome sequence filtration is inappropriate when considering the presence or absence of a specific SNV pair in an outbreak. Consequently, we find that the exclusion of many of the intermediate genomes is unfounded, leaving the conclusion of two natural zoonoses unsupported.

Keywords: SARS-CoV-2; intermediate; zoonosis; Huanan seafood market; Wuhan; spillover



Citation: Massey, S.E.; Jones, A.; Zhang, D.; Deigin, Y.; Quay, S.C. Unwarranted Exclusion of Intermediate Lineage A-B SARS-CoV-2 Genomes Is Inconsistent with the Two-Spillover Hypothesis of the Origin of COVID-19. *Microbiol. Res.* **2023**, *14*, 448–453. <https://doi.org/10.3390/microbiolres14010033>

Academic Editor: Caijun Sun

Received: 5 February 2023

Revised: 15 March 2023

Accepted: 16 March 2023

Published: 22 March 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Recently, a widely reported analysis by Pekar et al. proposed that the COVID-19 pandemic originated via two independent zoonoses of lineage A and lineage B SARS-CoV-2 in the Huanan Seafood Market, Wuhan, China in late 2019 [1]. The study involved simulations of different evolutionary scenarios and used empirically observed SARS-CoV-2 genomes from the early stages of the pandemic to inform the analysis.

Lineage A and lineage B were the first two SARS-CoV-2 lineages to establish themselves in the first months of the pandemic [2]. Lineage A of SARS-CoV-2 possesses T8782 and C28144 (T/C), while lineage B possesses C8782 and T28144 (C/T) [3]. These two SNVs separated the two lineages early in the pandemic, and underwent subsequent divergence, with B rapidly becoming dominant. Lineage A appears ancestral, as T/C is found in a variety of closely related sarbecoviruses including RaTG13 [4] and BANAL-20-52 [5].

The transition of lineage A > lineage B would have involved two mutations at positions 8782 and 28144, and so genomes intermediate between lineage A and lineage B possessing a single mutation should have existed, either in the human population in Wuhan during the early outbreak or in a host animal, as proposed by [1]. Such intermediate genomes would either be C8782/C28144 (C/C) or T8782/T28144 (T/T), reflecting the two potential series of twin mutations that led to the conversion of lineage A into lineage B. The existence of intermediate lineage A-B genomes from humans would be inconsistent with two independent zoonoses of lineage A and lineage B, which requires that the A-B intermediate occurred in an unidentified host animal before the transmission of both lineages to the human population.

Pekar et al. identified 20 A-B intermediate genomes in their analysis but elected to exclude all of them, for a variety of reasons. This left 787 genomes remaining, which they proceeded to use for their analyses. We will go through their exclusion criteria and show that several genomes are potentially true intermediates, as follows.

2. Exclusion for Reasons of Contamination

Of the 20 potential A-B intermediate genomes identified by Pekar et al., 16 were C/C and 4 were T/T (Table 1). Pekar et al. claim that many of them share rare mutations with lineage A or lineage B viruses; this was used as a basis to exclude them as ‘artifacts of contamination or bioinformatics’. Curiously, the authors fail to define what an artifact of contamination is, and how they can be sure it is an artifact. The contamination analysis was not described, no results were reported, and it is not clear if the analysis was applied to the entire dataset of 787 genomes as well. In particular, the authors fail to identify which intermediate genome sequences were contaminated. Contaminating virus sequences are difficult to differentiate from within-host variants or co-infection with two strains. The best way they can be identified is through analysis of the background reads in order to detect anomalies with the stated sample source (for example, human haplogroup analysis may show if mitochondrial sequences present in the raw dataset are from more than one individual, indicating contamination). However, such analyses were not reported, not least because raw datasets were not available for the majority of the intermediate genomes.

Table 1. A-B intermediate genomes excluded from the analysis of Pekar et al. Shown are the genome GISAID accessions, with sequence source, average genome sequencing depth (from GISAID), and reasons given for exclusion by Pekar et al. [1].

GISAID Identifier	Intermediate Genotype	Source	Average Genome Sequencing Depth	Exclusion Criterion
EPI_ISL_452363	C/C	Beijing	2500X	‘whose additional mutations were not observed in early lineage A or B genomes and whose underlying data was not available’
EPI_ISL_452361	C/C	Beijing	1850X	‘whose additional mutations were not observed in early lineage A or B genomes and whose underlying data was not available’
EPI_ISL_1069206	C/C	Anhui	NA	Belongs to later A lineage
EPI_ISL_413017	C/C	South Korea	NA	(1) Belongs to both later A and B lineages
EPI_ISL_451325	C/C	Sichuan	759X	(2) $\leq 10X$ coverage at 28144 (1) Belongs to later A lineage (2) *

Table 1. Cont.

GISAID Identifier	Intermediate Genotype	Source	Average Genome Sequencing Depth	Exclusion Criterion
EPI_ISL_451394	C/C	Sichuan	2302X	(1) Belongs to both later A and B lineages (2) *
EPI_ISL_451390	C/C	Sichuan	1793X	(1) Belongs to later B lineage (2) *
EPI_ISL_451322	C/C	Sichuan	57X	*
EPI_ISL_451389	C/C	Sichuan	2388X	*
EPI_ISL_451377	C/C	Sichuan	2916X	*
EPI_ISL_451330	C/C	Sichuan	476X	*
EPI_ISL_451319	C/C	Sichuan	636X	*
EPI_ISL_451320	C/C	Sichuan	1335X	*
EPI_ISL_451353	C/C	Sichuan	496X	*
EPI_ISL_451076	C/C	Sichuan	NA	*
EPI_ISL_454919	C/C	Wuhan	NA	*
EPI_ISL_462306	T/T	Singapore	NA	≤10X read depth at positions 8782 and 28144 Low sequencing depth and mixed C/T bases at position 8782 (personal communication, Di Liu and Yi Yan, Table S1 of Pekar et al.)
EPI_ISL_493179	T/T	Wuhan	17378X	Low sequencing depth and mixed C/T bases at position 8782 (personal communication, Di Liu and Yi Yan, Table S1 of Pekar et al.)
EPI_ISL_493180	T/T	Wuhan	27852X	Low sequencing depth and mixed C/T bases at position 8782 (personal communication, Di Liu and Yi Yan, Table S1 of Pekar et al.)
EPI_ISL_493182	T/T	Wuhan	15274X	Low sequencing depth and mixed C/T bases at position 8782 (personal communication, Di Liu and Yi Yan, Table S1 of Pekar et al.)

* incorrect base calls, often due to 'low sequencing depth' and 'low sequencing depth at position 8782 led to the erroneous assignment of intermediate haplotypes' (personal communication, L.Chen).

3. Exclusion for Reasons of Low Sequencing Depth

Pekar et al. use 'low sequencing depth' as a reason for the exclusion of most of the intermediate genomes (Table 1). However, this exclusion criterion was reliant on personal communications from 'L.Chen' (for the exclusion of 12 C/C genomes from Wuhan and Sichuan) and 'Di Liu and Yi Yan' (for the exclusion of 3 T/T genomes). However, the high average sequencing depths reported by GISAID for the majority of the datasets (Table 1) appear inconsistent with the assertion that low read depth was responsible for erroneous base calls at position 8782 or 28144 leading to the incorrect assignment of an intermediate genotype. Although the read depth may vary throughout the genome, Pekar et al. fail to explain why these two positions were preferentially subject to error. While it is quite plausible for a specific study to have a low read depth at specific locations and, at the same time, a high average sequencing depth overall, the intermediate genomes came from seven different labs. In addition, if low read depth were a significant problem, then there should be an excess of unique SNVs throughout the genome, indicative of sequencing errors. In particular, unique SNVs should be observed immediately flanking positions 8782 and 28144 if these locations are particularly prone to errors; however, this was not observed.

Only two raw sequencing datasets were used to justify exclusion. A T/T genome from Singapore (EPI_ISL_462306) and a C/C genome from South Korea (EPI_ISL_413017) were

excluded for having a read depth $\leq 10X$ at positions 8782/28144 and 28144, respectively. However, this exclusion criterion was apparently not applied to the 787 genomes for which raw datasets were also available, representing selection bias. Presumably, if low read depth could lead to miscalls at positions 8782 and 28144, the same possibility exists for the apparent A and B lineage genomes comprising the final 787 genome dataset used in Pekar et al.'s analysis.

In addition, the authors fail to explain why a 10X read depth was chosen as a cutoff, rather than a cutoff based on dataset quality control and statistical error analysis to determine a more robust lower bound [6]. If a clear majority of nucleotides at the two key positions are either C or T, then this is unlikely to be artefactual given an overall error rate on Illumina Miseq machines of 0.47% [7] (the sequencing platforms used to sequence the intermediate genomes are shown in Table S1).

4. Exclusion for Reasons of Convergence

Seven intermediate genomes were excluded for possessing what were described as A, B, or a combination of A- and B-specific SNVs (Table 1). The rationale given was that these were A or B lineage genomes that acquired convergent mutations at positions 8782 and 28144 to produce C/C or T/T genotypes (note that if a B lineage genome underwent the mutation T28144C, converting it into a C/C intermediate, this would be classified as a reversion mutation rather than a convergent mutation). However, no data were provided to demonstrate that a particular SNV was indeed A- or B-specific.

Four of the seven genomes had only one A- or B-specific (however defined) SNV (EPI_ISL_1069206 had one 'A specific' SNV, while EPI_ISL_451390 and two unidentified genomes had one 'B specific' SNV each). If it were established that they are indeed true A- or B-specific SNVs, they could represent homoplasies that themselves arose via convergence. No caveat to this effect was included in Pekar et al.'s study. Whether the remaining three genomes which had more than one SNV (claimed to be A- or B-specific) are located at an intermediate position between lineage A and lineage B genomes on a phylogenetic tree of early SARS-CoV-2 genomes, or are placed amongst lineage A or lineage B genomes, was not reported by the authors. Placement at an intermediate position would imply that they are intermediate genomes that acquired additional SNVs, something that would not be surprising.

5. Exclusion for Lack of Underlying Data

Pekar et al. report excluding two C/C intermediates from Beijing (EPI_ISL_452361 and EPI_ISL_452363) for the reason that their additional mutations (both genomes have two SNVs compared to Wuhan-Hu-1) 'were not observed in early Lineage A or B genomes and underlying data was not available'. Data from GISAID indicate that the genomes were sequenced to a high average sequencing depth, 1850X and 2500X, respectively. Consequently, the possibility of sequencing errors is low. We note that the criterion of a lack of underlying data was not applied to the 787 genomes used for Pekar et al.'s analysis, and so was selectively applied. Indeed, Pekar et al. included four genomes in their study from the same sequencing batch as the two excluded Beijing genomes, but these also lacked underlying data (EPI_ISL_452357, EPI_ISL_452358, EPI_ISL_452359, and EPI_ISL_454417).

Regarding the first criterion that additional SNVs in the two intermediate genomes were not observed in early lineage A or B genomes, this is puzzling as it is not explained why this should be problematic. EPI_ISL_452363 has A2966C, which is unique in GISAID SARS-CoV-2 genomes, and C28253T, which is observed in a genome sampled on 10 April 2020 (MT907516), and so could be regarded as early. It is hard to understand why the presence of these two SNVs necessitate that EPI_ISL_452363 should be excluded.

6. Exclusion via Personal Communication

A key problem is that many of the intermediate genomes were excluded based on personal communications, which cannot be independently validated. It is unconventional

to rely on personal communications to exclude key data which have a significant impact on the conclusions of a paper. A total of 12 C/C intermediates from Sichuan and Wuhan were excluded on the basis of a personal communication from L.Chen, who appears to be the lead author of [8], in which the 12 C/C intermediate genomes were published. The exclusion criteria were summarized as ‘incorrect base calls, often due to low sequencing depth’ and ‘low sequencing depth at position 8782 led to the erroneous assignment of intermediate haplotypes’ (Table 1).

Bafflingly, however, 54 genomes from the same study in [8] were included in the 787 genomes analyzed by Pekar et al. (Supp Data S1). The basis for excluding 12 C/C intermediate genomes, but including 54 other genomes from the same study, was conspicuously not explained, representing another example of selection bias.

Di Liu and Yi Yan provided a table, via personal communication, of three possible T/T intermediate genomes (EPI_ISL_493179, EPI_ISL_493180, and EPI_ISL_493182) which show read depths at position 8782 of 64X, 40X, and 29X, respectively (from Table S1 of Pekar et al.’s study). These genomes are published in [9]. Two patient samples have 8782T at a significant minor allele fraction of 0.375. The third, EPI_ISL_493182, has an 8782T fraction of 0.66 at the 29X read depth, and a 28144T fraction of 0.936 at the 69289X read depth. The read depth at position 8782 of 29X exceeds the 10X cutoff applied to other genomes by Pekar et al. Consequently, this is clearly a T/T consensus intermediate genome.

We do not believe that the exclusion of genomes because they are not 100% pure at 8782 and 28144 is a valid criterion for ruling out the possibility that these genomes may be intermediates. Consistent with this, EPI_ISL_493182 is described as a T/T intermediate genome in the publication that reports its sequencing (denoted as ‘C100’) [9]. Unfortunately, no raw data were provided by Di Liu/Yi Yan or L.Chen, which would allow a further inspection of positions 8782 and 28144.

7. Non-Consideration of Additional Intermediates

Finally, we note that 14 additional potential intermediate genomes were not considered by Pekar et al. at all [10], with several meeting their genome filtration inclusion criteria. In general terms, the exclusion of genome sequences by filtration is inappropriate when addressing whether a particular genotype, represented by only a small number of SNVs, is present or absent in a genomic dataset. Thus, this procedure is not able to rule out the 14 additional genomes as true intermediates.

In conclusion, we find that the exclusion of the majority of the A-B intermediate genomes from the analysis of Pekar et al. was unwarranted, and at a minimum, they cannot be ruled out as true intermediates. We, therefore, urge Pekar et al. to revise their analysis and conclusions accordingly.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/microbiolres14010033/s1>, Supp Data S1: Genomes from the study by Lin et al. (2020) that were included in the analysis of Pekar et al.; Table S1: Sequencing platforms and facilities used to sequence the 20 intermediate genomes.

Author Contributions: Conceptualization, S.E.M., A.J., D.Z., Y.D. and S.C.Q.; methodology, S.E.M., A.J., D.Z., Y.D. and S.C.Q.; investigation, S.E.M., A.J., D.Z., Y.D. and S.C.Q.; writing—original draft preparation, S.E.M., A.J., D.Z., Y.D. and S.C.Q.; writing—review and editing, S.E.M., A.J., D.Z., Y.D. and S.C.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pekar, J.E.; Magee, A.; Parker, E.; Moshiri, N.; Izhikevich, K.; Havens, J.L.; Gangavarapu, K.; Malpica Serrano, L.M.; Crits-Christoph, A.; Matteson, N.L.; et al. The Molecular Epidemiology of Multiple Zoonotic Origins of SARS-CoV-2. *Science* **2022**, *377*, 960–966. [[CrossRef](#)] [[PubMed](#)]
2. Tang, X.; Ying, R.; Yao, X.; Li, G.; Wu, C.; Tang, Y.; Li, Z.; Kuang, B.; Wu, F.; Chi, C.; et al. Evolutionary analysis and lineage designation of SARS-CoV-2 genomes. *Sci. Bull.* **2021**, *66*, 2297–2311. [[CrossRef](#)] [[PubMed](#)]
3. 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*, 1012–1023. [[CrossRef](#)] [[PubMed](#)]
4. Zhou, P.; Yang, X.L.; Wang, X.G.; Hu, B.; Zhang, L.; Zhang, W.; Si, H.R.; Zhu, Y.; Li, B.; Huang, C.L.; et al. Discovery of a Novel Coronavirus Associated with the Recent Pneumonia Outbreak in Humans and Its Potential Bat Origin. *Nature* **2020**, *579*, 270–273. [[CrossRef](#)] [[PubMed](#)]
5. Temmam, S.; Vongphayloth, K.; Baquero, E.; Munier, S.; Bonomi, M.; Regnault, B.; Douangboubpha, B.; Karami, Y.; Chrétien, D.; Sanamxay, D.; et al. Bat Coronaviruses Related to SARS-CoV-2 and Infectious for Human Cells. *Nature* **2022**, *604*, 330–336. [[CrossRef](#)] [[PubMed](#)]
6. De Maio, N.; Walker, C.; Borges, R.; Weilguny, L.; Slodkowitz, G.; Goldman, N. Issues with SARS-CoV-2 Sequencing Data. *Virological.org*. 2022. Available online: <https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473> (accessed on 1 February 2023).
7. Stoler, N.; Nekrutenko, A. Sequencing Error Profiles of Illumina Sequencing Instruments. *NAR Genom. Bioinform.* **2021**, *3*, lqab019. [[CrossRef](#)] [[PubMed](#)]
8. Lin, J.W.; Tang, C.; Wei, H.C.; Du, B.; Chen, C.; Wang, M.; Zhou, Y.; Yu, M.X.; Cheng, L.; Kuivanen, S.; et al. Genomic Monitoring of SARS-CoV-2 Uncovers an Nsp1 Deletion Variant That Modulates Type I Interferon Response. *Cell Host Microbe* **2021**, *29*, 489–502.e8. [[CrossRef](#)] [[PubMed](#)]
9. Yan, Y.; Wu, K.; Chen, J.; Liu, H.; Huang, Y.; Zhang, Y.; Xiong, J.; Quan, W.; Wu, X.; et al. Rapid Acquisition of High-Quality SARS-CoV-2 Genome via Amplicon-Oxford Nanopore Sequencing. *Virol. Sin.* **2021**, *36*, 901–912. [[CrossRef](#)] [[PubMed](#)]
10. Washburne, A.; Jones, A.; Zhang, D.; Deigin, Y.; Quay, S.; Massey, S.E. Statistical Challenges for Inferring Multiple SARS-CoV-2 Spillovers with Early Outbreak Phylodynamics. *bioRxiv* **2022**. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.