# 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], Steven C. Quay[5]

[1] Biology Dept, University of Puerto Rico - Rio Piedras, San Juan, PR USA

[2] Independent Bioinformatics Researcher, Melbourne, Australia

[3] Independent Genetics Researcher, Sydney, Australia

[4] Youthereum Genetics Inc., Toronto, Ontario, Canada; ORCID 0000-0002-3397-5811

[5] Atossa Therapeutics, Inc., Seattle, WA USA; ORCID 0000-0002-0363-7651

[*]Correspondence to: 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 shows 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 4 genomes from the same sequencing study were included in their 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 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

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 (Pekar et al. 2022). 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 of SARS-CoV-2 possesses T8782 and C28144 (T/C), while lineage B possesses C8782 and T28144 (C/T) (Tang et al. 2020). These two SNVs separated the two lineages early in the pandemic, and underwent subsequent divergence, with B becoming dominant over time. Lineage A appears ancestral, as T/C is found in a variety of closely related sarbecoviruses including RaTG13 (Zhou et al. 2020) and BANAL-20-52 (Temmam et al. 2022).

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 should have existed, either in the human population in Wuhan during the early outbreak, or in a host animal, as proposed by (Pekar et al. 2022). Such intermediate genomes would either be C8782 / C28144 (C/C) or T8782 / T28144 (T/T), reflecting the two potential series of 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.

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

*1. 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 by 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 datasetcare 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.

| 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<br>2) ≤ 10X coverage at 28144 |
| EPI_ISL_451325 | C/C | Sichuan | 759X | 1) Belongs to later A lineage<br>2) * |
| EPI_ISL_451394 | C/C | Sichuan | 2302X | 1) Belongs to both later A and B lineages<br>2) * |
| EPI_ISL_451390 | C/C | Sichuan | 1793X | 1) Belongs to later B lineage<br>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 |
| EPI_ISL_493179 | T/T | Wuhan | 17378X | Low sequencing depth and mixed C/T bases at position 8782, Table S1 (personal communication, Di Liu and Yi Yan, Table S1 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 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 Ya, Table S1 Pedar 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)

**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. 2022)

*2. 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) appears 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 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 low read depth at specific locations and at same time a high average sequencing depth overall - the intermediate genomes come from 7 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 ≤ 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 (implying that some of these may be true intermediate genomes, misattributed as lineage A or B due to errors at positions 8782 and 28144).

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 (De Maio et al. 2020). 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% (Stoler and Nekrutenko 2021) (the sequencing platforms used to sequence the intermediate genomes is shown in Table S2).

*3. 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. However, 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. In addition, no data was provided to demonstrate that a particular SNV was indeed A or B specific.

Four of the 7 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. Whether the remaining 3 genomes, which had more than one SNV (claimed as 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.

*4. 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 2 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 indicates that the genomes were sequenced to high average sequencing depth, 1850X and 2500X, respectively. Consequently, the possibility of sequencing errors is low. We note that the criterion of 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 4 genomes in their study from the same sequencing batch as the 2 Beijing genomes, but these also lacked underlying data (EPI_ISL_452357, EPI_ISL_452358, EPI_ISL_452359, 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 10th 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.

*5. 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. 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 (Lin et al. 2021), in which the 12 C/C intermediate genomes are 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 by (Lin et al. 2021) 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, is 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 shows read depths at position 8782 of 64X, 40X and 29X, respectively (from Table S1 of Pekar et al.). These genomes are published in (Yan et al. 2021). Two patient samples have 8782T at a significant minor allele fraction of 0.375. The third, EPI_ISL_493182 has a 8782T fraction of 0.66 at 29X read depth, and a 28144T fraction of 0.936 at 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 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') (Yan et al. 2021). Unfortunately, no raw data was provided by Di Liu/Yi Yan or L.Chen, which would allow further inspection of positions 8782 and 28144.

*6) Non-consideration of additional intermediates*

Finally, we note that 14 additional potential intermediate genomes were not considered by Pekar et al. at all (Washburne et al. 2022), with several not meeting their genome filtration 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. is unwarranted, and at minimum they cannot be ruled out as true intermediates. We therefore urge Pekar et al. to revise their analysis and conclusions accordingly.

**References**

Delahaye, Clara, and Jacques Nicolas. 2021. "Sequencing DNA with Nanopores: Troubles and Biases." *PLOS ONE*. https://doi.org/10.1371/journal.pone.0257521.

De Maio, Nicola, Conor Walker, Rui Borges, Lukas Weilguny, Greg Slodkowicz, and Nick Goldman. 2020. "Issues with SARS-CoV-2 Sequencing Data." *Virological.org*. https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.

Lin, Jing-Wen, Chao Tang, Han-Cheng Wei, Baowen Du, Chuan Chen, Minjin Wang, Yongzhao Zhou, et al. 2021. "Genomic Monitoring of SARS-CoV-2 Uncovers an Nsp1 Deletion Variant That Modulates

Type I Interferon Response." *Cell Host & Microbe* 29 (3): 489–502.e8.
Pekar, Jonathan E., Andrew Magee, Edyth Parker, Niema Moshiri, Katherine Izhikevich, Jennifer L. Havens, Karthik Gangavarapu, et al. 2022. "The Molecular Epidemiology of Multiple Zoonotic Origins of SARS-CoV-2." *Science*, July, eabp8337.
Stoler, Nicholas, and Anton Nekrutenko. 2021. "Sequencing Error Profiles of Illumina Sequencing Instruments." *NAR Genomics and Bioinformatics* 3 (1): lqab019.
Tang, Xiaolu, Changcheng Wu, Xiang Li, Yuhe Song, Xinmin Yao, Xinkai Wu, Yuange Duan, et al. 2020. "On the Origin and Continuing Evolution of SARS-CoV-2." *National Science Review* 7 (6): 1012–23.
Temmam, Sarah, Khamsing Vongphayloth, Eduard Baquero, Sandie Munier, Massimiliano Bonomi, Béatrice Regnault, Bounsavane Douangboubpha, et al. 2022. "Bat Coronaviruses Related to SARS-CoV-2 and Infectious for Human Cells." *Nature* 604 (7905): 330–36.
Washburne, Alex, Adrian Jones, Daoyu Zhang, Yuri Deigin, Steven Quay, and Steven E. Massey. 2022. "Statistical Challenges for Inferring Multiple SARS-CoV-2 Spillovers with Early Outbreak Phylodynamics." *bioRxiv*. https://doi.org/10.1101/2022.10.10.511625.
Yan, Yi, Ke Wu, Jun Chen, Haizhou Liu, Yi Huang, Yong Zhang, Jin Xiong, et al. 2021. "Rapid Acquisition of High-Quality SARS-CoV-2 Genome via Amplicon-Oxford Nanopore Sequencing." *Virologica Sinica* 36 (5): 901–12.
Zhou, Peng, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, et al. 2020. "Discovery of a Novel Coronavirus Associated with the Recent Pneumonia Outbreak in Humans and Its Potential Bat Origin." https://doi.org/10.1101/2020.01.22.914952.

# Supplementary Data

| GISAID identifier | Sequencing platform | Collection date | Sequencing facility |
|---|---|---|---|
| EPI_ISL_452363 | Oxford Nanopore GridION | 31 Jan 2020 | Laboratory of Infectious Diseases Center of Beijing Ditan Hospital |
| EPI_ISL_452361 | Oxford Nanopore GridION | 11 Feb 2020 | Laboratory of Infectious Diseases Center of Beijing Ditan Hospital |
| EPI_ISL_1069206 | Oxford Nanopore MinION | 26 Jan 2020 | Microbiology Laboratory,Lu`an Center for Disease Control and Prevention, Anhui |
| EPI_ISL_413017 | Illumina MiSeq | 8 Feb 2020 | Department of Microbiology, Institute for Viral Diseases, College of Medicine, Korea University |
| EPI_ISL_451325 | Oxford Nanopore MinION | 31 Jan 2020 | West China Hospital of Sichuan University |
| EPI_ISL_451394 | Oxford Nanopore MinION | 3 Feb 2020 | West China Hospital of Sichuan University |
| EPI_ISL_451390 | Oxford Nanopore MinION | 23 Jan 2020 | West China Hospital of Sichuan University |
| EPI_ISL_451322 | Oxford Nanopore MinION | 1 Feb 2020 | West China Hospital of Sichuan University |
| EPI_ISL_451389 | Oxford Nanopore MinION | 30 Jan 2020 | West China Hospital of Sichuan University |
| EPI_ISL_451377 | Oxford Nanopore MinION | 8 Feb 2020 | West China Hospital of Sichuan University |
| EPI_ISL_451330 | Oxford Nanopore MinION | 29 Jan 2020 | West China Hospital of Sichuan University |
| EPI_ISL_451319 | Oxford Nanopore MinION | 27 Jan 2020 | West China Hospital of Sichuan University |
| EPI_ISL_451320 | Oxford Nanopore MinION | 27 Jan 2020 | West China Hospital of Sichuan University |
| EPI_ISL_451353 | Oxford Nanopore MinION | 27 Jan 2020 | West China Hospital of Sichuan University |
| EPI_ISL_451076 | Oxford Nanopore MinION | 8 Feb 2020 | West China Hospital of Sichuan University |
| EPI_ISL_454919 | Oxford Nanopore MinION | 3 Feb 2020 | Wuhan Chain Medical Labs (CMLabs) |
| EPI_ISL_462306 | Illumina MiSeq | 14 Feb 2020 | National Public Health Laboratory, National Centre for Infectious Diseases, Singapore |
| EPI_ISL_493179 | Oxford Nanopore PromethION | 26 Jan 2020 | National Virus Resource Center, Chinese Academy of Sciences, Wuhan |
| EPI_ISL_493180 | Oxford Nanopore PromethION | 26 Jan 2020 | National Virus Resource Center, Chinese Academy of Sciences, Wuhan |
| EPI_ISL_493182 | Oxford Nanopore PromethION | 26 Jan 2020 | National Virus Resource Center, Chinese Academy of Sciences, Wuhan |

**Table S1** Sequencing platforms and facilities used to sequence the 20 intermediate genomes Sequencing information was derived from the GISAID entry for each genome.

EPI_ISL_451313, EPI_ISL_451314, EPI_ISL_451315, EPI_ISL_451316, EPI_ISL_451321, EPI_ISL_451326, EPI_ISL_451327, EPI_ISL_451328, EPI_ISL_451329, EPI_ISL_451331, EPI_ISL_451334, EPI_ISL_451338, EPI_ISL_451344, EPI_ISL_451345, EPI_ISL_451346, EPI_ISL_451354, EPI_ISL_451356, EPI_ISL_451357, EPI_ISL_451359, EPI_ISL_451360, EPI_ISL_451365, EPI_ISL_451369, EPI_ISL_451370, EPI_ISL_451371, EPI_ISL_451374, EPI_ISL_451376, EPI_ISL_451378, EPI_ISL_451379, EPI_ISL_451380, EPI_ISL_451381, EPI_ISL_451382, EPI_ISL_451384, EPI_ISL_451385, EPI_ISL_451386, EPI_ISL_451387, EPI_ISL_451388, EPI_ISL_451391, EPI_ISL_451392, EPI_ISL_451395, EPI_ISL_451398, EPI_ISL_454930, EPI_ISL_454931, EPI_ISL_454932, EPI_ISL_454933, EPI_ISL_454934, EPI_ISL_454935, EPI_ISL_454936, EPI_ISL_454937, EPI_ISL_454980, EPI_ISL_454981, EPI_ISL_455374, EPI_ISL_455375, EPI_ISL_455386,EPI_ISL_455391

**Supp Data S1** Genomes from the study by Lin et al (2020) that were included in the analysis of Pekar et al