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. The basis for their analysis is the hypothesis that there were two spillovers into humans that are recognized by a two-SNV difference, called Lineage A and B, and that the one-SNV intermediate A/B genomes found in numerous human infections are all sequencing errors, implying that the intermediate A/B genomes with a single SNV occurred in unsampled animal hosts. Consequently, confirmation of the existence of an intermediate A/B genome from humans would falsify their hypothesis. Pekar et al. identified and excluded 20 A/B intermediate genomes from their analysis. A variety of exclusion criteria were applied, including low sequencing depth, and the assertion of repeated sequencing errors at lineage defining positions 8782 and 28144. However, data from GISAID shows that most of the genomes were sequenced to high coverage, contradicting these criteria. The decision to exclude the majority of genomes was based on personal communications, with raw data not being available for inspection. Multiple errors and inconsistencies were observed in the exclusion process. Mapping analysis of a genome from Singapore, dismissed due to an arbitrary read depth cutoff, confirms it as a true intermediate, while an intermediate genome from Wuhan was discarded even though it conformed to the cutoff. Puzzlingly, two genomes from Beijing were discarded despite an average sequencing depth of 2175X. Lastly, we identify a new potential intermediate genome from Guangzhou. 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 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, using empirically observed SARS-CoV-2 genomes from the early stages of the pandemic to inform the analysis. The existence of intermediate lineage A/B from genomes from humans would be inconsistent with two independent zoonoses of lineage A and lineage B.

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, which 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 A -> B would have involved two mutations and so intermediate genomes should have existed, either in the human population in Wuhan during the early outbreak, or in an intermediate host, 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 A into B.

Pekar et al. identify 20 intermediate genomes in their analysis, but elect to exclude all of them, for a variety of reasons. We show that their exclusion criteria were flawed, and that several genomes are true intermediates, as follows.

Critique of the exclusion criteria

1. For reasons of contamination

Pekar et al. identified 20 potential intermediate genomes, 16 C/C and 4 T/T (Table 1), however decided to exclude them all 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. In addition, the authors fail to identify which genome sequences were contaminated. The only way this can be done with certainty is by analysis of background reads in order to detect anomalies with the stated sample source (for example, haplogroup analysis may show if mitochondrial sequences are from more than one individual). However, these analyses were not reported.

GISAID identifier	Intermediate genotype	Source	Sequencing depth	Exclusion criterion
EPI_ISL_452363	C/C	Beijing	2500X	Underlying data was not available
EPI_ISL_452361	C/C	Beijing	1850X	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 2) ≤ 10X coverage at 28144

EPI_ISL_451325	C/C	Sichuan	759X	1) Belongs to later A lineage 2) Low sequencing depth at position 8782 led to the erroneous calling (personal communication, L.Chen)
EPI_ISL_451394	C/C	Sichuan	2302X	1) Belongs to both later A and B lineages 2) Incorrect base calls, often due to low sequencing depth (personal communication, L.Chen)
EPI_ISL_451390	C/C	Sichuan	1793X	1) Belongs to later B lineage 2) 'Incorrect base calls, often due to low sequencing depth' (personal communication, L.Chen)
EPI_ISL_451322	C/C	Sichuan	57X	Incorrect base calls, often due to low sequencing depth (personal communication, L.Chen)
EPI_ISL_451389	C/C	Sichuan	2388X	Incorrect base calls, often due to low sequencing depth (personal communication, L.Chen)
EPI_ISL_451377	C/C	Sichuan	476X	Incorrect base calls, often due to low sequencing depth (personal communication, L.Chen)
EPI_ISL_451330	C/C	Sichuan	476X	Incorrect base calls, often due to low sequencing depth (personal communication, L.Chen),
EPI_ISL_451319	C/C	Sichuan	636X	Incorrect base calls, often due to low sequencing depth (personal communication, L.Chen)
EPI_ISL_451320	C/C	Sichuan	1335X	Incorrect base calls, often due to low sequencing depth (personal communication, L.Chen)
EPI_ISL_451353	C/C	Sichuan	496X	Incorrect base calls, often due to low sequencing depth (personal communication, L.Chen)
EPI_ISL_451076	C/C	Sichuan	NA	Incorrect base calls, often due to low sequencing depth (personal communication, L.Chen)
EPI_ISL_454919	C/C	Wuhan	NA	Incorrect base calls, often due to low sequencing depth (personal communication, L.Chen)
EPI_ISL_462306	T/T	Singapore	NA	≤ 10X coverage at 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.)

Table 1 Intermediate genomes excluded from the analysis of Pekar et al.
Shown are the genome GISAID accessions, with sequence source, sequencing depth (from GISAID) and reasons given for exclusion by (Pekar et al. 2022)

2. For reasons of low sequencing depth

Pekar et al. claim low sequencing depth as the reason for exclusion of most of the genomes (Table 1). However, this was reliant on personal communications from an ‘L.Chen’ (for the exclusion of 11 C/C genomes), and ‘Di Liu and Yi Yan’ (for the exclusion of 3 T/T genomes). However, no further identifying information is given regarding these individuals (discussed below). The high sequencing depths reported by GISAID for the majority of the datasets (Table 1) contradicts the assertion that low sequencing depth was responsible for an erroneous base call at position 8782 or 28144. While sequencing depth may vary throughout the genome, Pekar et al.

fail to explain why these two positions were preferentially subject to error. In addition, if low sequencing depth were a significant problem then there should be an excess of unique SNVs indicative of sequencing errors, which was not observed.

Mostly, the Oxford Nanopore sequencing platform was used for sequencing the 20 intermediate genome datasets (Table 2), which uses the MinION sequencer in different configurations (PromethION, GridION). The MinION has an error rate of 1.2–2.2 % for mismatches (Delahaye and Nicolas 2021). This relatively high error rate can be compensated for by increasing sequencing depth: the recommended sequencing depth for the SARS-CoV-2 genome is 60X (Bull et al. 2020), a criterion 17 of the intermediate genomes fulfill (Table 1).

Indeed, we note that the Sichuan C/C intermediate EPI_ISL_451320, sequenced to a depth of 1335X, is used by NextStrain as an intermediate for rooting purposes (to place the root of the phylogenetic tree between lineages A and B) (https://nextstrain.org/groups/blab/ncov/early-outbreak/root-AB).

Only two raw sequencing datasets were used to justify exclusion. The authors excluded a T/T genome from Singapore (EPI_ISL_462306) and a C/C genome from South Korea (EPI_ISL_413017) 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 available. Presumably, if low sequencing depth could lead to miscalls at positions 8782 and 28144, the same possibility exists for the apparent A and B lineage genomes comprisiing the 787 genome dataset (implying that some may be intermediate genomes misattributed as lineage A or B).

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. If a clear majority of nucleotides 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).

We mapped the Singapore (EPI_ISL_462306) SRA dataset to SARS-CoV-2 Wuhan-Hu-1. Position 8782 had 12/12 reads possessing a T, while position 28144 had 6/6 reads possessing a T. Using a depth of coverage at a site ≥5 and base quality of ≥30 to consider a variant (De Maio et al. 2020) we conclude that the EPI_ISL_462306 genome is supported by read data and is a true T/T intermediate genome (Fig 1 and Supporting Data). If low sequencing depth were indeed a problem in this genome, then a significant number of apparently novel SNVs should be observed in the rest of the genome (resulting from miscalling), which is not the case.

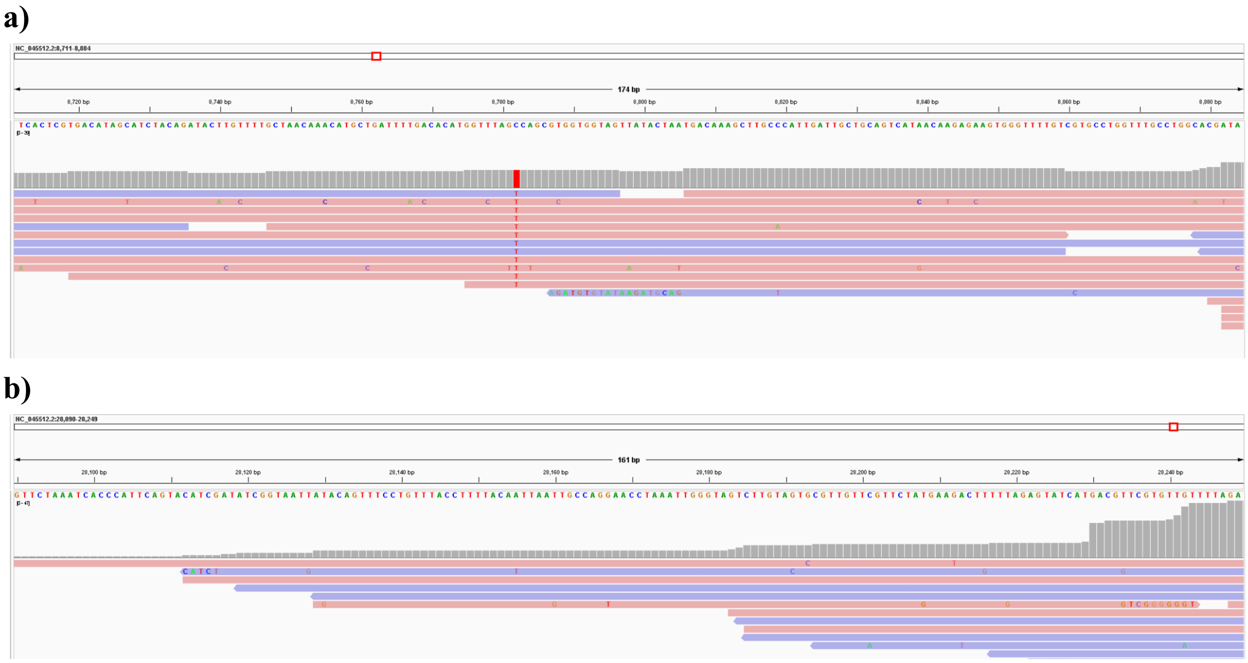


Fig. 1 Alignment of Singapore genome EPI_ISL_462306 reads to Wuhan-Hu-1 reference a) 12 reads with a unanimous 8782T coding, and b) 6 reads with a unanimous 28144T coding. Together these show an unambiguous T/T genotype.

We additionally identified a potential intermediate T/T genome with read depths >640X at defining sites in a patient sampled between January 25 and February 10 2020 by the Guangzhou Medical University (Wang et al. 2021) (Supporting Data). The consensus genome after alignment of sample GZMU0025 to SARS-CoV-2 Wuhan-Hu-1 has 5 mutations (Table 2). All SNVs other than C8782T have minor alleles at 10%-50% abundance. We additionally identified several lineage T/T and C/C subdominant mutations in other samples in the same BioProject (Supporting Data).

G6819T	A: 2, C: 5, G: 158 (10%), T: 1453 (90%)
C8782T	A: 3, C: 1, G: 1, T: 642 (99%)
C17373T	A: 2, C: 3284 (50%), G: 18, T: 3288 (50%)
T24371C	A: 11, C: 571 (66%), G; 16, T: 267 (31%)
A27006G	A: 880 (47%), C: 7, G: 968 (52%)

Table 2. Consensus SNVs in sample GZMU0025 after alignment to SARS-CoV-2 Wuhan-Hu-1 The table shows nucleotide counts. 28144T had a distribution of C: 688 (10%), T: 6519 (90%).

We further identify several potential intermediate T/T and C/C genomes in BioProject PRJNA612766 that cannot be conclusively excluded as intermediate genomes at position 8782 was not sequenced (Supplementary Data).

3. For reasons of convergence

7 intermediate genomes were excluded for possessing A, B or a combination of A and B specific SNVs (Table 1), the rationale being 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, 4 of the genomes had only one A or B specific SNV (EPI_ISL_1069206 had one A specific SNV, while EPI_ISL_451390 and two unidentified genomes had one B specific SNV each). These could just as easily represent intermediate genomes that have evolved a single lineage specific SNV via convergence. No caveat to this effect was included in Pekar et al.

4. For lack of 'underlying data'

Remarkably, Pekar et al. report excluding two C/C genomes from Beijing (EPI_ISL_452361 and EPI_ISL_452363), for the reason that 'underlying data was not available'. However, data from GISAID indicates that the genomes were sequenced to high coverage, 1850X and 2500X, respectively. Consequently, the possibility of sequencing errors is low. We note that the ill-defined criterion was not applied to the 787 genomes used for the analysis, and so was selectively applied.

5. Via 'personal communication'

A key problem with exclusion of many of the intermediate genomes is that they were excluded based on personal communications from an 'L.Chen', and 'Di Liu and Yi Yan', who are not clearly identified. 11 C/C genomes from Sichuan and Wuhan were excluded on the basis of a personal communication from L.Chen. The exclusion criterion was vaguely summarized as 'incorrect base calls, often due to low sequencing depth ' (Table 1). However, Table 1 shows that there are 12 C/C genomes from Sichuan and Wuhan, not 11 as stated in Pekar et al. In addition, the Sichuan and Wuhan genomes were sequenced at different sequencing facilities (West China Hospital of Sichuan University and the National Virus Resource Center, Chinese Academy of Sciences, Wuhan, respectively, Table S1), so it is unlikely that L.Chen sequenced them all. Indeed, it would appear that while L.Chen is associated with the Sichuan genomes, the person who provided a personal communication regarding the Wuhan sequence EPI_ISL_454919 is unidentified (who we term 'person X').

Di Liu and Yi Yan provide a table of three T/T 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.). This exceeds the 10X cutoff applied to other genomes by Pekar et al. and is sufficient to call a consensus SNV at this position.

EPI_ISL_493182 has a 8782T proportion of 66% and so is clearly a consensus T/T intermediate genome.

In contrast, the sequencing depth at position 28144 is high for all three genomes (61361X, 95374X and 69269X, respectively). GISAID reports overall genome sequencing depths of 17378X, 27852X and 15274X, respectively. While read depth typically varies around the genome, the observed very low read depth around position 8782 is unusual given an average read depth of 20168X for the three genomes. No explanation is given for the marked difference in read depths between position 8782 and 28144. Unfortunately, no raw data is provided by either Di Liu/Yi Yan, L.Chen or person X, which would allow further inspection of positions 8782 and 28144.

In conclusion, we find that the exclusion of the majority of the intermediate genomes from the analysis of Pekar et al. is unwarranted. We therefore urge Pekar et al. to revise their analysis and conclusions accordingly.

References

Bull, Rowena A., Thiruni N. Adikari, James M. Ferguson, Jillian M. Hammond, Igor Stevanovski, Alicia G. Beukers, Zin Naing, et al. 2020. “Analytical Validity of Nanopore Sequencing for Rapid SARS-CoV-2 Genome Analysis.” *Nature Communications* 11 (1): 6272.

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.

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.

Wang, Yanqun, Daxi Wang, Lu Zhang, Wanying Sun, Zhaoyong Zhang, Weijun Chen, Airu Zhu, et al. 2021. “Intra-Host Variation and Evolutionary Dynamics of SARS-CoV-2 Populations in COVID-19 Patients.” *Genome Medicine* 13 (1): 30.

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.

Supporting Data

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

SRA	reads mapped	average length	average quality	bases mapped (cigar)	mismatches	error rate
SRR17868030	1078118	238	35.7	256124676	895111	3.49483E-03
SRR18012762	3038	144	35.9	439320	1817	4.13594E-03
SRR13616010	649984	99	30	64427874	650981	1.01002E-02

Table S2 Summary statistics for key SRA datasets aligned to SARS-CoV-2 (NC_045512.2). Calculated using Samtools.

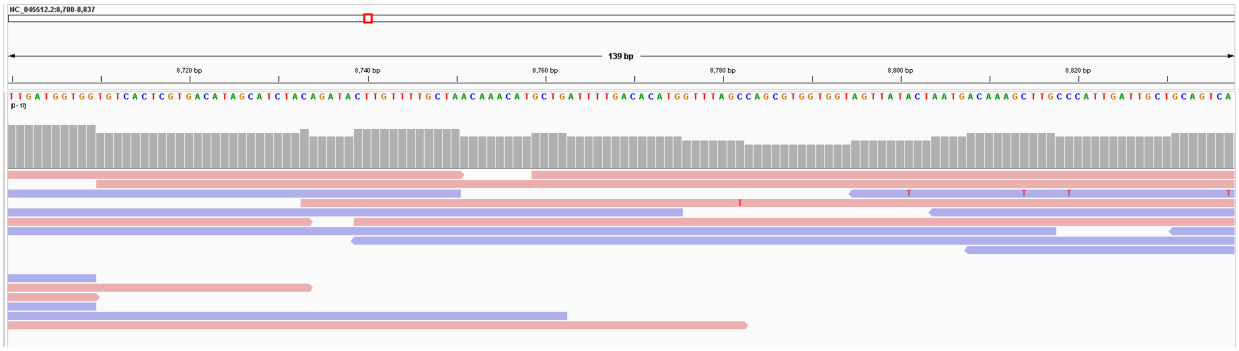
EPI_ISL_462306

We reviewed the dataset supporting the hCoV-19/Singapore/188/2020 genome (EPI_ISL_462306), SRR17868030 in BioProject PRJNA802993. The intermediate T/T genome was excluded as a valid intermediate genome by Pekar et al. (2022) solely because read depth at positions 8782 and 27144 were less than 10X. We aligned the dataset to SARS-CoV-2 Wuhan-Hu-1 (NC_045512.2) and found that dataset contains a dominant 382nt deletion in ORF7b and ORF8 27848-28229 previously documented in 11 early Singapore strains and one genome from Taiwan (Su et al. 2020). We clipped regions misaligned to SARS-CoV-2, and generated a consensus sequence using ivar (Grubaugh et al. 2018) using a minimum read depth of 5 and minimum quality of 30, and identified a subdominant intermediate lineage T/T genome lacking the 382nt deletion. As the deletion includes location 28144, as such the T/T genome pre-dates the 382nt deletion. The consensus genome has a unanimous 8782T with 12X read depth and unanimous 8144T with a 6X read depth. The genome has one additional mutation G27933T relative to SARS-CoV-2 Wuhan-Hu-1 (NC_045512.2).

EPI_ISL_413017

The dataset supporting the inferred consensus lineage C/C genome assembled by Kang et al. (EPI_ISL_413017), SRR18012762 in BioProject PRJNA806767 was sequenced by the Korea University College of medicine using a Illumina MiSeq platform. Mapping depth was low, at an average of 14.7 reads. However mapping statistics show an error rate of 1 base per 250 bases, which is an upper bound as it assumes any intra-host deviation from the reference as an error. The dataset was excluded from consideration as supporting an intermediate genome by Pekar et al. (2022) with reasons being: “belongs to both later A and B lineages” and a < 10X coverage at 28144. We identified that position 28144 is covered by 19 reads, and 8782 by 7 reads for a consensus lineage B genome. However a subdominant lineage C/C mutant genome exists at >30% frequency in the sample.

a)


NC_045512.2:8,700-8,837
139 bp
8,720 bp
8,740 bp
8,760 bp
8,780 bp
8,800 bp
8,820 bp
TTGATGGTGGTGTCACTCGTGACATAGCATCTACAGATACTTGTTTTGCTAACAAACATGCTGATTTTGACACATGGTTTAGCCAGCGTGGTGGTAGTTATACTAATGACAAAGCTTGCCCATTGATTGCTGCAGTCA
[0 - 17]
T
T
T
T
T

b)

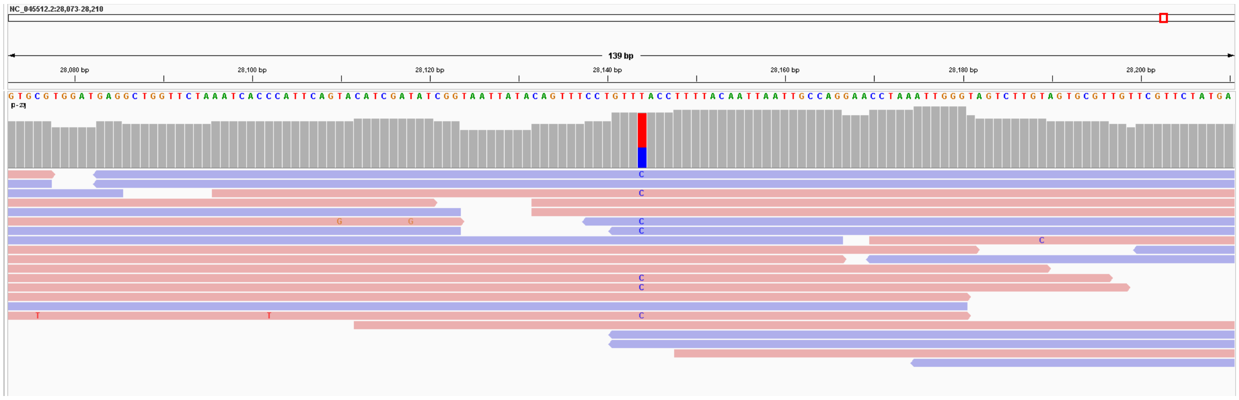

NC_045512.2:28,073-28,210
139 bp
28,080 bp
28,100 bp
28,120 bp
28,140 bp
28,160 bp
28,180 bp
28,200 bp
GTGCGTGGATGAGGCTGGTTCTAAATCACCCATTCAGTACATCGATATCGGTAATTATACAGTTTCCTGTTTACCTTTTACAATTAATTGCCAGGAACCTAAATTGGGTAGTCTTGTAGTGCGTTGTTCGTTCTATGA
[0 - 23]

Fig. S1. Alignment of South Koren genome EPI_ISL_413017 reads to Wuhan-Hu-1 reference **a)** Position 8782 is covered by 7 reads, 6 with 8782C and one with 8782T while **b)** Position 28144 is covered by 19 reads, 7 with 28144C and 12 with 28144T.

GZMU0025.capture

We aligned all SRA datasets in SARS-CoV-2 intra-host variation RNA-Seq BioProject PRJNA698267 by Wang et al. (2021) to the SARS-CoV-2 Wuhan-Hu-1 genome (NC_045512.2). We identified an 8782T/28144T possible intermediate T/T genome in dataset GZMU0025.capture (SRR13616010), a clinical sample of SARS-CoV-2 collected by the Guangzhou Institute of Respiratory Disease in February 2020.

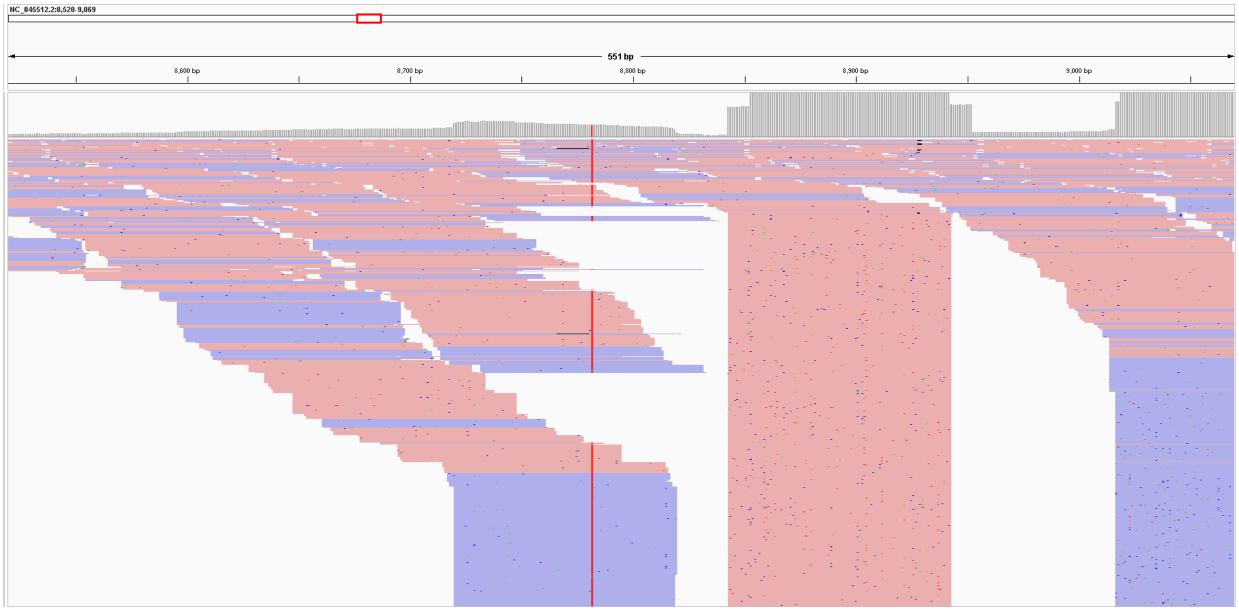


Fig. S2 Alignment of SRR13616010 to the SARS-CoV-2 Wuhan-Hu-1 genome (NC_045512.2) showing near unanimous (99%) 8782T SNV relative to the reference. Plotted in IGV.

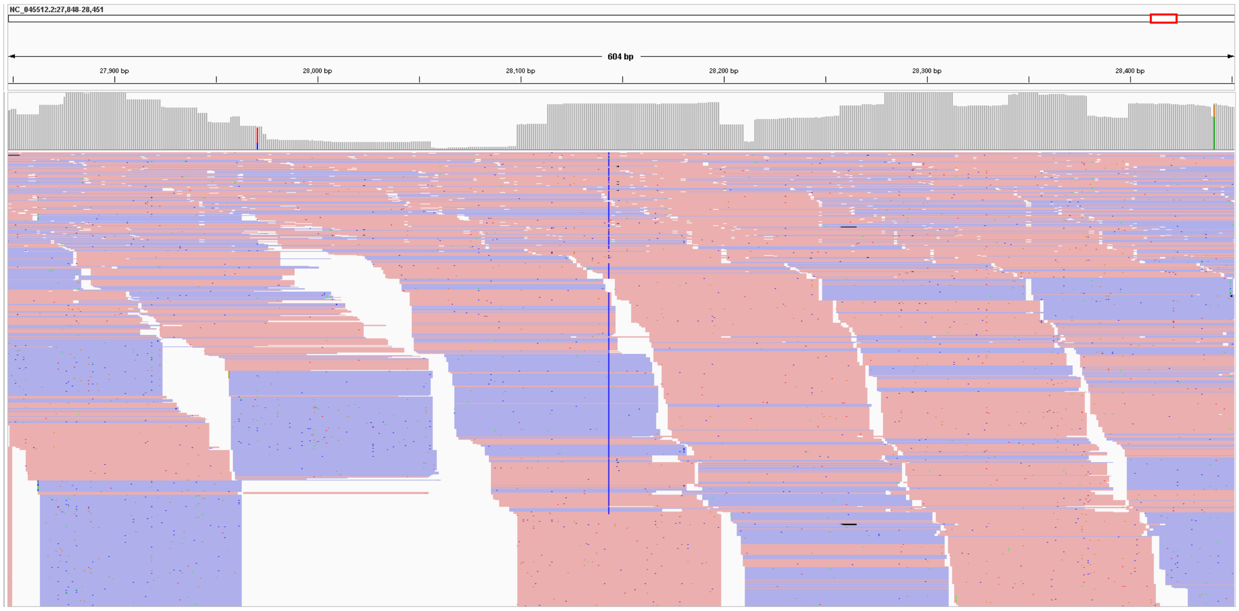


Fig. S3 Alignment of SRR13616010 to the SARS-CoV-2 Wuhan-Hu-1 genome (NC_045512.2) with 90% 28144T and 10% 28144C distribution relative to the reference. Plotted in IGV.

Four additional mutations relative to SARS-CoV-2 Wuhan-Hu-1 are shown in Table 2. We note the consensus genome shares G6819T, C17373T (and 8782T) with Guangdong/SZ-N59-P0049/2020 (EPI_ISL_413872), while the G29527A mutation in the parent branch for this genome is only found at 47% frequency (A: 971 (45%), G: 1166 (55%)). Interestingly, C8782T was the only SNV to reach a near unanimous frequency. In addition to the T/T genome identified PRJNA698267, 4 samples with subdominant C/C genomes (1316020, 131615986, 13615979 and 1315972) at 10% to 30% of reads and T/T (SRR13615966, SRR13615968) at c. 10-20% of reads (Supp. Data). Site-specific RNA editing by the host immune system plays a role in genome divergence (Azgari et al. 2021), and may have been especially important in the earliest months of the outbreak when SARS-CoV-2 was first exposed to a human immune system. The shared SNVs with EPI_ISL_413872 thus cannot be assumed to be random mutations indicating homoplasy.

Additionally we also aligned all early COVID-19 patient datasets from Renmin hospital in BioProject PRJNA612766 recovered by Bloom (2021) to the SARS-CoV-2 Wuhan-Hu-1 genome (NC_045512.2). Sequencing in this BioProject targeted local regions of the genome, dominantly upstream of ORF1ab (Wang et al. 2020) and although position 28144 was covered, 8782 was not and unambiguous determination of lineage A/B and intermediate C/C and T/T genomes was not possible. Multiple B or T/T and either A or C/C lineages are found. Given the identification of a T/T genome in Guangzhou patient samples, in the dataset supporting

EPI_ISL_462306 and in other genomes invalidly excluded as discussed above, the potential presence of intermediate T/T or C/C genomes in these patient samples cannot be ruled out.

Methods

Detailed methods including exact commands can be found at the github repository linked in code and data section below.

All SRA datasets in PRJNA698267 were trimmed using TrimGalore v0.6.7 using default adapter detection.

SRR17868030 and SRR18012762 were trimmed and filtered using fastp with default settings.

All SRA datasets were aligned to the SARS-CoV-2 Wuhan-Hu-1 (NC_045512.2) genome with poly(A) tail removed using minimap2 version 2.24.

A consensus genome for SRR13616010 generated using samtools and ivar v1.3.1 (Grubaugh et al. 2018) using a minimum read depth of 10 and default minimum quality of 20 (an ivar minimum quality of 30 was found to produce exactly the same sequence).

A consensus sequence for the SRR18012762 dataset (supporting EPI_ISL_413017) was generated using ngsutils to remove clipped ends of reads, samtools mpileup and then ivar was used to call a consensus using a minimum read depth of 5 and minimum read quality of 30.

A consensus sequence for the SRR17868030 dataset (supporting EPI_ISL_462306) was generated using samclip (https://github.com/tseemann/samclip) to remove reads with a softclip length of more than 10, then using ngsutils to remove clipped ends of reads. Samtools mpileup was used to generate a pileup file, and ivar was used to call a consensus using a minimum read depth of 5 and minimum read quality of 30.

Read mapping statistics for SRR18012762 were calculated after alignment to NC_045512.2 with poly(A) tail removed. NGSutils (Breese and Liu, 2013) was used to remove soft clipped ends of mapped reads. Samtools v 1.5.1 (Li et al. 2009) was then used to calculate statistics.

Nucleotide mappings at positions 8782, 28144 and 29095 was calculated on all trimmed bams for all all SRA datasets in PRJNA698267 and PRJNA612766 using the Mutation_stats.ipynb notebook at the github repository linked in code and data section below.

Code and Data

Coverage and mapping statistics for SNV positions 8782, 28144 and 29095 for all SRA datasets in BioProjects PRJNA698267 and PRJNA612766 aligned to SARS-CoV-2 Wuhan-Hu-1 as well as mutation scanning code and FastQC results can be found at the following location:

https://github.com/bioscienceresearch/Intermediate_Genomes

Supplementary References

Bloom, J. Recovery of Deleted Deep Sequencing Data Sheds More Light on the Early Wuhan SARS-CoV-2 Epidemic. Mol Biol Evol (2021) 38, 5211-5224.

Breese, MR and Liu Y. NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. Bioinformatics (2013) 29(4): 494-496. doi: 10.1093/bioinformatics/bts731

Grubaugh ND, Gangavarapu K, Quick J, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. bioRxiv. Published online 2018. doi:10.1101/383513

Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352

Su YCF, Anderson DE, Young BE, et al. Discovery and genomic characterization of a 382-nucleotide deletion in ORF7B and orf8 during the early evolution of SARS-CoV-2. MBio. 2020;11(4):1-9. doi:10.1128/mBio.01610-20