Additional foreign sequences in the "Antarctica SARS-CoV-2" PRJNA692319 samples

István Csabai[1*] **and Norbert Solymosi**[1,2]

[1]Department of Physics of Complex Systems, ELTE Eötvös Loránd University, Budapest, Hungary
[2]Centre for Bioinformatics, University of Veterinary Medicine, Budapest, Hungary
[*]csabai@elte.hu

ABSTRACT

Recently Csabai et al.[1] have found a metagenomic sample set originally collected at Antarctica that most likely as a result of contamination contains traces of unique SARS-CoV-2 variants. Later on they identified putative host genomes[2]. The preprints resulted a wide discussion in the news and social media and some comments on the preprint server or via private emails. Here we try to briefly reflect on some of them.
Caveats: This is not intended to be a full article, some references are missing, and some arguments are not final.

Flow cell capacity and number of the parallel samples

In his comments "0-2"[2] Alexander Crits-Christoph calculates that if the Antarctica samples have been sequenced on Illumina HiSeq 4000, they more probably occupied only around 10% of the flow cell. The revealed SARS-CoV-2 sequencing was among the other simultaneously sequenced samples but it most probably not filled the remaining 90% part of flow cell. There may have been multiple SARS-CoV-2 sequencings which observation is consistent with our analysis of the SARS-CoV-2 mutations that indicate the presence of multiple strains. The coverage is low for both SARS-CoV-2 and the mammal mitochondria to reliably decompose them, estimate the number of parallel genomes or decide if they are from a single sample with minority variants or from multiple samples. These are very good and important observations and we agree that they should be considered especially when the phylogeny of the sequences is analysed.

Sequencing instrument and index hopping

After the preprints got viral in social media Steve Massey (@stevenemassey) posted a tweet: https://twitter.com/stevenemassey/status/1491539892811313155. He noticed that the read identifiers in the uploaded FASTQ files for the Antarctica project PRJNA692319 indicate that despite the SRA metadata (that was filled in by the submitters) shows Illumina HiSeq 4000 as the sequencing instrument, the SARS-CoV-2 containing samples were sequenced on BGI MGISEQ. In FASTQ files that contain the sequencing reads each read has a unique identifier. The exact format of the identifiers depends on the manufacturer of the sequencing instrument and may contain type and serial number of the chip, information on lane and position, etc. The same observation was made through email by Kevin McKernan with the additional comment that MGISEQ instruments are less prone to index hopping, the type of error that is suspected as the contamination mechanism. The documentation by MGI Tech[3] indeed lists as an advantage of this platform that the unique library prep and RCR amplification results much lower index hopping rates compared with other platforms, at a rate of 0.0001% 0.0004%. Combined with the extreme 10^{12}nt capacity of these instruments this still may result tens of thousands of erroneously assigned reads which may or may not be enough to explain the contamination for our case. Since SARS-CoV-2 is an RNA virus direct contamination (e.g. from an infected lab assistant) is not feasible. If index hopping is ruled out the virus RNA should have been present in the same laboratory as cDNA to be able to contaminate the metagenomic DNA library. In this case the R1/R2 mate asymmetry requires further explanation.

Flow cell id and possible method for dating of the samples

In a reply to the above mentioned tweet, Daoyu Zhang (@Daoyu15) and other Twitter users have further analysed the flow cell IDs (FCID): https://twitter.com/Daoyu15/status/1492024118157344768 and https://twitter.com/Daoyu15/status/1492495520153047042. They noticed that the FCID for the SARS-CoV-2 containing Antarctica data sets starts with the V300043327 chip ID while for example for WIV07-2 (accession at SRA: SRR11092059, one of the first known SARS-CoV-2 sequences published in March 2020[4]) the FCID starts with V300043428 (see the "Reads" tab

at https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR11092059). If the time of use of the flow cells approximately follows the same order as the FCIDs assigned, the sequencing of the Antarctica samples could have happened before the sequencing of the first published SARS-CoV-2 genomes.

Suid alphaherpesvirus 1 (SuHV1) contamination and assignment of host genomes

In his comments "3-4"[2] Alexander Crits-Christoph also notices that some of the Antarctica data sets contain not only SARS-CoV-2 virus sequences, but also *Suid alphaherpesvirus 1* that cause Aujeszky's disease, usually called pseudorabies. This most likely again came not from the Antarctica soil, but from similar contamination as SARS-CoV-2. As SuHV1 does not infect humans, and traces of pig (the most common host) genome cannot be detected in the samples, this may also imply, that the identified mtDNA (green monkey and Chinese hamster cell lines) may partly or in all belong to hosts of SuHV1 and not of SARS-CoV-2.

We have aligned the Antarctica reads to reference genome of SuHV1 in a similar way as we did for SARS-CoV-2. Figure 1 shows the genome coverage. We have chosen to use logarithmic scale as there are two large peaks around 120knt where the genome has low complexity repetitive regions. This demonstrates that instead of the mean depth (average number of reads covering a genomic position) it is better to use the coverage (ratio of the covered vs not covered positions). We have listed the SuHV1 coverage values for all the Antarctica samples, separately for R1 and R2 reads in Table 1. For both SARS-CoV-2 (see Table 1 in the first preprint[1]) and for the mammalian mtDNA (see Table 1 in the second preprint[2]) there is a strong asymmetry between the R1 and R2 reads. Though for SuHV1 we also see the highest coverage values for the same samples SRR13441704, SRR13441705 and (somewhat less for) SRR13441708 that were the most abundant in SARS-CoV-2 and mtDNA, there is no asymmetry between the R1 and R2 mates.

Also we note, that in contrast to the SARS-CoV-2 RNA virus, SuHV1 is a double-stranded DNA virus. With Fiugure 1 and 2 in the second preprint[2] we argued that the mtDNA sequences are from RNA sequencing and not DNA sequencing, since the internal parts of the genes have higher coverage depth than the gene ends. For SuHV1 (see Figure 1) we could not detect such clear trend.

As the exact mechanism of the contamination is not known we cannot decide with absolute certainty that the presumed hosts are related to SuHV1 or SARS-CoV-2 (or neither of them) but both the R1/R2 asymmetry and the RNA/DNA sequencing signatures makes our original hypothesis more likely.

Human host and the 27nt deletion at 21761

In the first preprint[1] we have mentioned that one of the most characteristic variation in the recovered SARS-CoV-2 genome is the 27nt long deletion at genomic position 21761 (Δ 68-IHVSTGTNGT-76). In a private email Gergely Szöllősi pointed out that the closest known bat viruses RaTG13 and BANAL-52 does not contain this deletion. By searching for the flanking regions around the deletion he has found a sample that contains the same deletion. The sample is described in the article by Ramirez et al.[5]. In this study the deletion carrying sample is not from a human subject but from a serial passage experiment on Huh7 cell line and the deletion is result of the adaptation. Huh7 is among the relatively few cell lines that have Asian origin, it comes from hepatoma tissue of a Japanese male. Further analysis is required but according to our preliminary analysis the mutation profile of Huh7 mtDNA is consistent with the mutation profile of human mtDNA found in the Antarctica samples. If this were the case and mtDNA originated from the SARS-CoV-2 samples, all hosts would be of cell line origin. (While writing this brief report we have found that Steven Massey has also noticed this: https://twitter.com/stevenemassey/status/1492987865998966788.)

CRISPR-Cas9 guide RNA fragments

To search for traces of genetic engineering in the studied samples we have aligned them against the following two target independent guide RNA sequence fragments:

```
>gRNA1
GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGC
>gRNA2
GTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGC
```

Similarly to the SARS-CoV-2 alignment, the same samples SRR13441704, SRR13441705 and SRR13441708 gave significant matches. Altogether 88 and 96 reads matched gRNA1 and gRNA2, respectively in these three samples. An illustration of the alignments is shown in Figure 2. The short 20nt long flanking regions that are soft clipped by the aligner are the variable regions for the CRISPR-Cas9 system's guide RNA that matches the targeted region where genome modification is

aimed. The short target sequences do not match the SARS-CoV-2 genome and further investigation would be needed to decide if the revealed genetic editing is related to the SARS-CoV-2 experiments or similar contamination from other studies analysed in the same sequencing run.

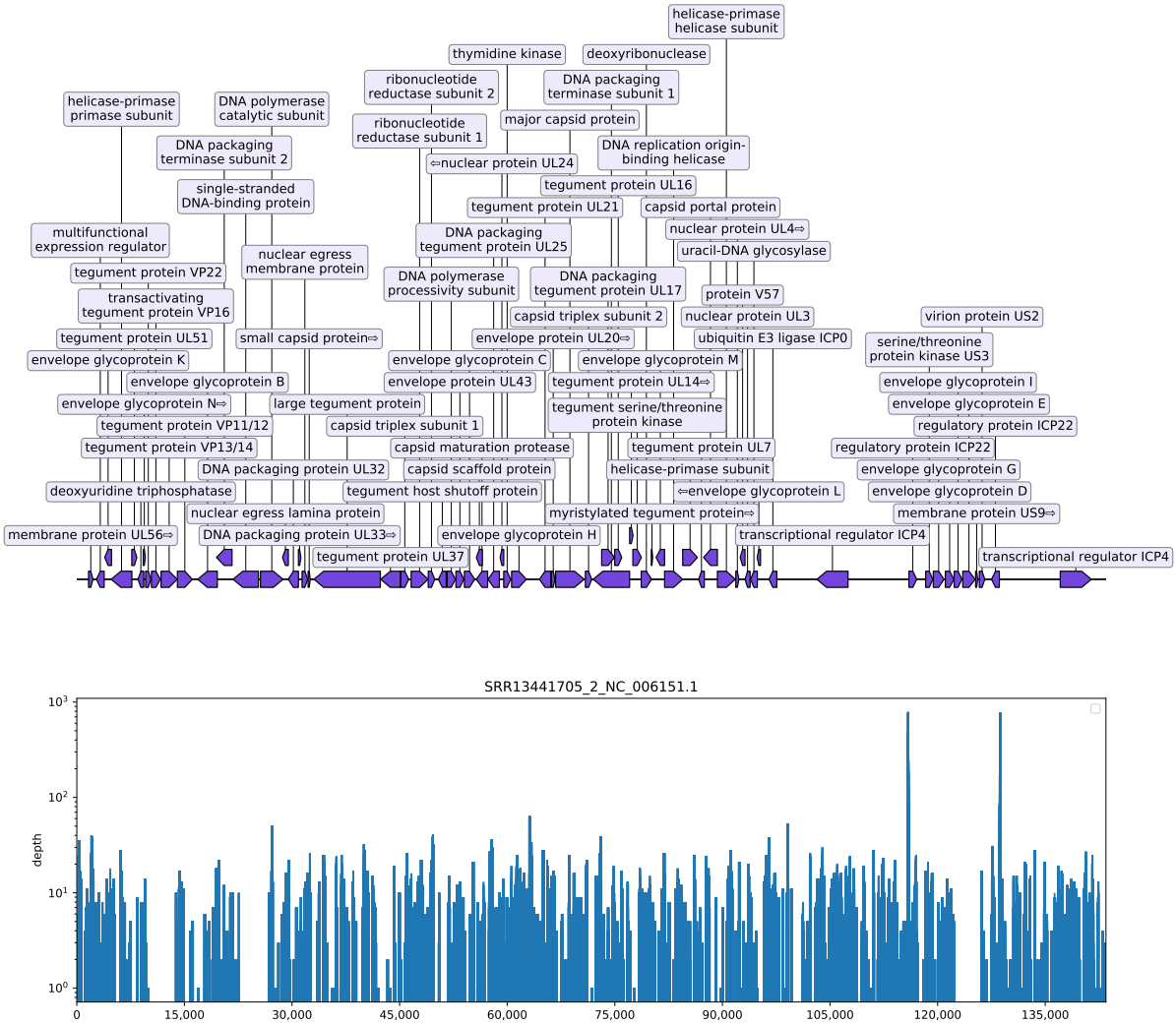

Figure 1. The coverage of the *Suid alphaherpesvirus 1* reference genome for the SRR13441705 sample's R2 reads plotted on logarithmic scale. The two large peaks are due to low complexity repetitive regions.

Acknowledgments

The authors thanks all the researchers, including the ones mentioned in this short write-up and many others who contribute with sound analysis. We also thank Jesse Bloom for continued insightful discussions. Special thanks for the original researchers for collecting and analysing the PRJNA692319 samples and for all others who upload full raw sequencing data and reliable metadata to ENA and SRA and other public archives. This work was financed by EU Horizon 2020 programs VEO No. 874735 and BY-COVID No. 101046203.

References

1. Csabai, I., Papp, K., Visontai, D., Stéger, J. & Solymosi, N. Unique sars-cov-2 variant found in public sequence data of antarctic soil samples collected in 2018-2019. *Submitted, DOI:10.21203/rs.3.rs-1177047/v1* **0**, 0 (2022).

2. Csabai, I. & Solymosi, N. Host genomes for the unique sars-cov-2 variant leaked into antarctic soil metagenomic sequencing data. *DOI:10.21203/rs.3.rs-1330800/v1* **0**, 0 (2022).

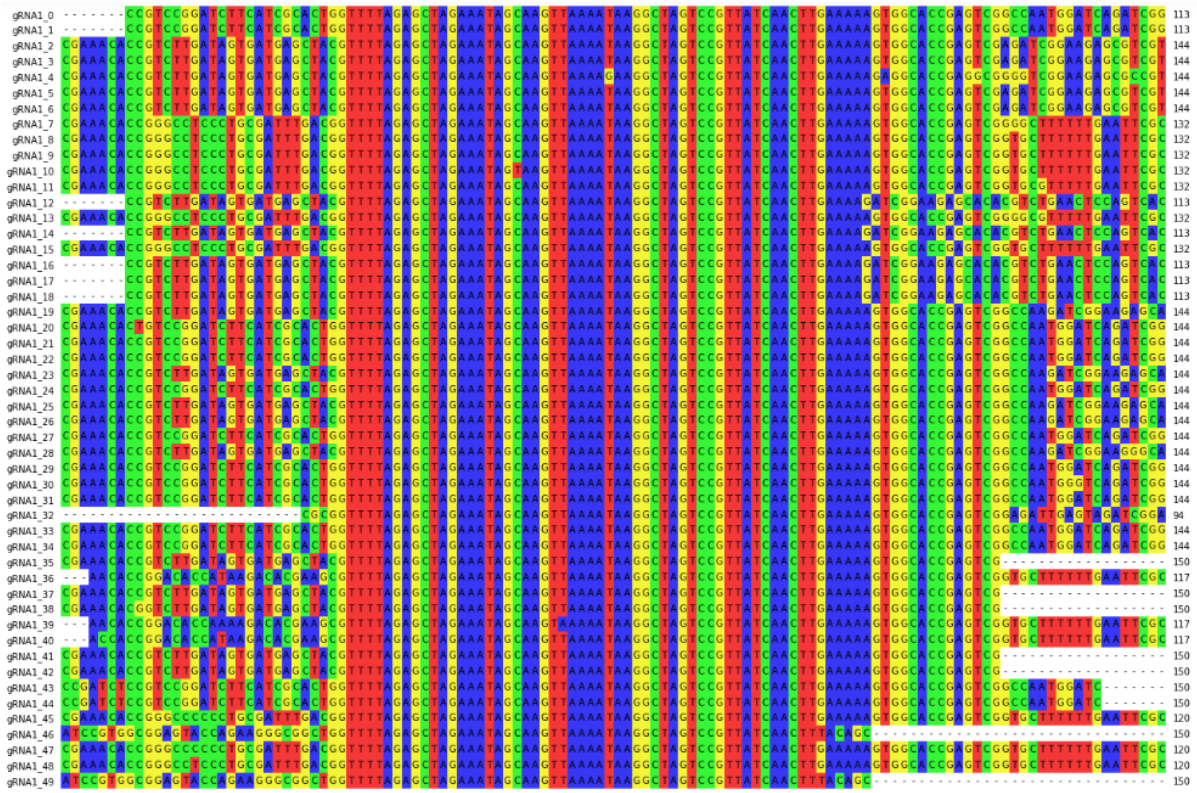

Figure 2. Some CRISPR-Cas9 guide RNA containing fragments in the merged SRR13441704, SRR13441705 and SRR13441708 sample set. The aligned colored stripes starting with `GTTTTAGAG...` at column 31 are for the CRISPR target independent gRNA1 scaffold sequences. The 20 flanking bases on the left constitute the target sequence.

3. Genetic Sequencer MGISEQ-2000. https://en.mgitech.cn/Uploads/Temp/file/20200115/5e1e68f7779a5.pdf. Accessed: 2022-02-13.

4. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).

5. Ramirez, S. *et al.* Overcoming culture restriction for sars-cov-2 in human cells facilitates the screening of compounds inhibiting viral replication. *Antimicrobial Agents Chemotherapy* **65**, e00097–21 (2021).

Table 1. *Suid alphaherpesvirus 1* genome coverage and depths for the samples. R1 and R2 mates were aligned separately. Green background highlights the samples that contained significant amount of SARS-CoV-2 reads. Blue background highlights coverage values above 5. For SARS-CoV-2 the R2 mates had significantly higher coverage, which is not the case for SuHV1. Note that the high average depth values are due to large peaks in low complexity repetitive regions.

Run	Mate	Coverage	Depth
SRR13441700	R1	0.95	0.19
SRR13441700	R2	0.94	0.39
SRR13441701	R1	0.99	0.80
SRR13441701	R2	1.29	109.36
SRR13441702	R1	0.72	0.62
SRR13441702	R2	0.85	8.53
SRR13441703	R1	0.92	2.33
SRR13441703	R2	1.28	25.72
SRR13441704	R1	**27.52**	1.34
SRR13441704	R2	**26.96**	2.33
SRR13441705	R1	**44.19**	3.93
SRR13441705	R2	**43.77**	4.44
SRR13441706	R1	0.82	0.33
SRR13441706	R2	0.85	12.45
SRR13441707	R1	0.64	0.57
SRR13441707	R2	0.84	20.41
SRR13441708	R1	**5.83**	0.47
SRR13441708	R2	**6.01**	2.41
SRR13441709	R1	3.71	0.23
SRR13441709	R2	3.81	0.46
SRR13441710	R1	0.57	0.08
SRR13441710	R2	0.63	0.24
SRR13441711	R1	0.99	0.50
SRR13441711	R2	1.31	27.91