Unique SARS-CoV-2 variant found in public sequence data of Antarctic soil samples collected in 2018-2019

István Csabai (csabai@elte.hu)
Eötvös Loránd University
Krisztián Papp
Eötvös Loránd University
Dávid Visontai
Eötvös Loránd University
József Stéger
Eötvös Loránd University
Norbert Solymosi
University of Veterinary Medicine

Research Article

Keywords:

Posted Date: December 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1177047/v1

Unique SARS-CoV-2 variant found in public sequence data of Antarctic soil samples collected in 2018-2019

István Csabai[1*], Krisztián Papp[1], Dávid Visontai[1], József Steger[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

The COVID-19 pandemic has been going on for two years now and although many hypotheses have been put forward, its origin remain obscure. We investigated whether the huge public sequencing data archives' samples collected earlier than the earliest known cases of the pandemic might contain traces of SARS-CoV-2. Here we report the bioinformatic analysis of a metagenome sample set collected from soil on King George Island, Antarctica between 2018-12-24 and 2019-01-13. It contains sequence fragments matching the SARS-CoV-2 reference genome with altogether more than half million nucleotides, covering the complete genome on average $17\times$. Preliminary phylogeny analysis places the sample close to the known earliest cases. The high sequence coverage rules out chance alignments from other species but possible laboratory contamination cannot be excluded. The sequence harbours a unique combination of mutations, unseen in other samples, so whatever its origin, it can add important piece of information to the puzzle of the ongoing pandemic.

Introduction

The COVID-19 pandemic has been going on for two years now and although many hypotheses have been put forward, its origin remains obscure. Sequencing techniques have evolved at a tremendous pace over the past decade and have been used by researchers around the world to examine large number of samples and deposit them in international public sequence archives such as ENA or SRA. As reported by the International Nucleotide Sequence Database Collaboration[1], the total size of sequence data archive has exceeded 9 petabytes in 2020 and grew roughly 10 times in the last 4-year period. The uploaded samples often contain sequences not only from the species that the researchers originally intended to study, but also genetic material from the environment or other species[2]. This can be the either the result of contamination during the wet-lab processing or true biological signal; the true origin is often hard to identify.

Our hypothesis was that these huge "gold mines" of sequence archives may contain genome fragments from early human SARS-CoV-2 cases or from the hypothesised originator zoonotic host. We searched through metagenomic profiles of samples collected earlier than the earliest known cases of the pandemic for traces of SARS-CoV-2 genetic sequences.

We did find a number of samples and here we report the analysis of one of these data sets.

Materials and Methods

The analysed sequencing project can be found under project ID PRJNA692319[3]. It contains 12 samples, all were collected from soil at King George Island, Antarctica in a 3 week period after 2018-12-24, which is summer in the southern hemisphere. According to the European Nucleotide Archive's metadata the sequencing read data was submitted by University of Science and Technology of China on 2021-01-15. There is no information on the date of wet-lab processes and sequencing that should have happened some time between the sample collection (2019-01) and the upload dates. The metadata contains some information on the sequencing process. The WGS with average spot length of 300 was performed on an Illumina HiSeq 4000 in a paired library layout, 150nt long reads resulting on average 9 Gbases per sample. See Table 1 on page 2 for some details of the samples. The public data sets that can be accessed based on the project ID PRJNA692319 or the listed run accession numbers from SRA or ENA archives.

After downloading the FASTQ files we performed metagenomic analysis and aligned the reads to the SARS-CoV-2 reference genome NC_045512.2[4]. For the alignment and variant calling, (with minor modifications, detailed below) we followed the "VEO workflow"[5] developed by the VEO consortium for unified processing of the raw SARS-CoV-2 sequencing data uploaded

Table 1. Sample properties. The first 6 columns show basic metadata of the samples. See[3] for more information. The 2 extra last columns show the count of reads that align to SARS-CoV-2 genome from the raw R1 and R2 reads. Low complexity reads aligned to the poly-A tail were not counted.

Run	Library Name	Collection date	Isolation source	lat lon	R1	R2
SRR13441700	AKGI_BS1_2018_12_24	2018-12-24	Antarctic soil	62.13 S 58.95 W	112	65
SRR13441701	AKGI_PL3_2019_01_13	2019-01-13	Antarctic soil	62.21 S 58.93 W	0	0
SRR13441702	AKGI_PL2_2019_01_13	2019-01-13	Antarctic soil	62.21 S 58.93 W	0	0
SRR13441703	AKGI_PL1_2019_01_13	2019-01-13	Antarctic soil	62.21 S 58.93 W	0	0
SRR13441704	AKGI_PS3_2019_01_13	2019-01-13	Antarctic soil	62.21 S 58.92 W	387	4485
SRR13441705	AKGI_PS2_2019_01_13	2019-01-13	Antarctic soil	62.21 S 58.92 W	242	3800
SRR13441706	AKGI_PS1_2019_01_13	2019-01-13	Antarctic soil	62.21 S 58.92 W	0	0
SRR13441707	AKGI_SS3_2019_01_05	2019-01-05	Antarctic soil	62.21 S 59.01 W	0	0
SRR13441708	AKGI_BS3_2018_12_24	2018-12-24	Antarctic soil	62.13 S 58.95 W	349	3537
SRR13441709	AKGI_BS2_2018_12_24	2018-12-24	Antarctic soil	62.13 S 58.95 W	112	161
SRR13441710	AKGI_SS2_2019_01_05	2019-01-05	Antarctic soil	62.21 S 59.01 W	113	106
SRR13441711	AKGI_SS1_2019_01_05	2019-01-05	Antarctic soil	62.21 S 59.01 W	0	0

to ENA-EBI and presented at the COVID-19 data portal share site (https://www.covid19dataportal.org/[6]). In short, the raw sequencing reads were aligned to the SARS-CoV-2 reference genome with bwa mem[7] then the genome coverage was calculated by samtools mpileup[8]. The lofreq[9] software was used to call variants and the VEO workflow's custom python script created the final consensus sequences.

The initial analysis revealed that although all 12 samples contain some reads that match the SARS-CoV-2 genome, only 3 of them, SRR13441704, SRR13441705 and SRR13441708 had enough matching reads to cover the whole genome (see "# aligned R1/R2" columns of Table 1 on page 2); in the following sections we present the analysis only for these 3 abundant samples.

Since the samples contain large amounts of foreign genetic material and also the original project was not optimized for SARS-CoV-2 detection, the standard VEO analysis workflow was adjusted. Our initial alignment revealed that there is a strong asymmetry between the forward (R1) and the reverse (R2) sequencing reads. We dropped the initial trimming and strict filtering and to avoid possible problems that this asymmetry may cause, in the final analysis we applied single read based alignment and used only the more abundant R2 reads.

The few samples and the relatively short genome made possible, and the non-standard nature of the samples made necessary the visual inspection of the alignments which was done by the IGV[10] tool. The gene-wise quasispecies analysis was performed by aBayesQR by 0.15 SNV threshold[11].

The Pangolin: Phylogenetic Assignment of Named Global Outbreak Lineages[12] web application[13] and the UShER: Ultrafast Sample placement on Existing tRee[14] online tool was used for quick preliminary phylogeny analysis.

Results

The aligned reads cover the genome (see Fig. 1) on average 5.2×. Certain parts have much higher sequencing depth than others but this is not unusual for metagenomic sequencing. There is a strong correlation between the coverage of the three samples which may indicate common biological or contamination origin.

Though some parts of the genome are covered by only a handful of nucleotides it was possible to call mutations with reasonable confidence. An extract of the variant call VCF files is shown in Table 2. Some of the mutations (POS=13694, 16156, 18060, 21761, 23525, 28144) are found in all the samples, while others were called in only one (POS=8782, 17039, 17634, 18082, 25498, 26458, 26895) or two (POS=29449) samples. By the visual inspection of the read alignments (BAM files) with the IGV tool we could also find some mutated bases in the samples where the variant call returned null result. Table 3 shows an extract from the samtools mpileup results for the positions where mutations were not detected by the workflow in all samples. For all cases mutated bases indeed occur with about the same ratio in all samples, so all samples may carry the same virus variant(s), only the low sequencing depth did not make possible to pass the mutation call algorithm's quality requirements.

Following this assumption, we have rerun the analysis on the pooled set of the SRR13441704, SRR13441705 and SRR13441708 sample's R2 reads. The consensus genome sequence can be found as a Supplementary file, SRR134417_04_05_08_R2.fasta. Due to its low allele frequency the C17634G SNP did not get into the consensus sequence.

Pangolin[13] assigned lineage "A" as the most likely lineage of SARS-CoV-2 for the consensus sequence. UShER web service[15] was also used to explore the phylogeny. Figure 2 displays the closest 50 samples from the GISAID collection as of 2021-12-12. In agreement with the Pangolin tool, the closest neighbours are in lineage A and A.1 and by classification

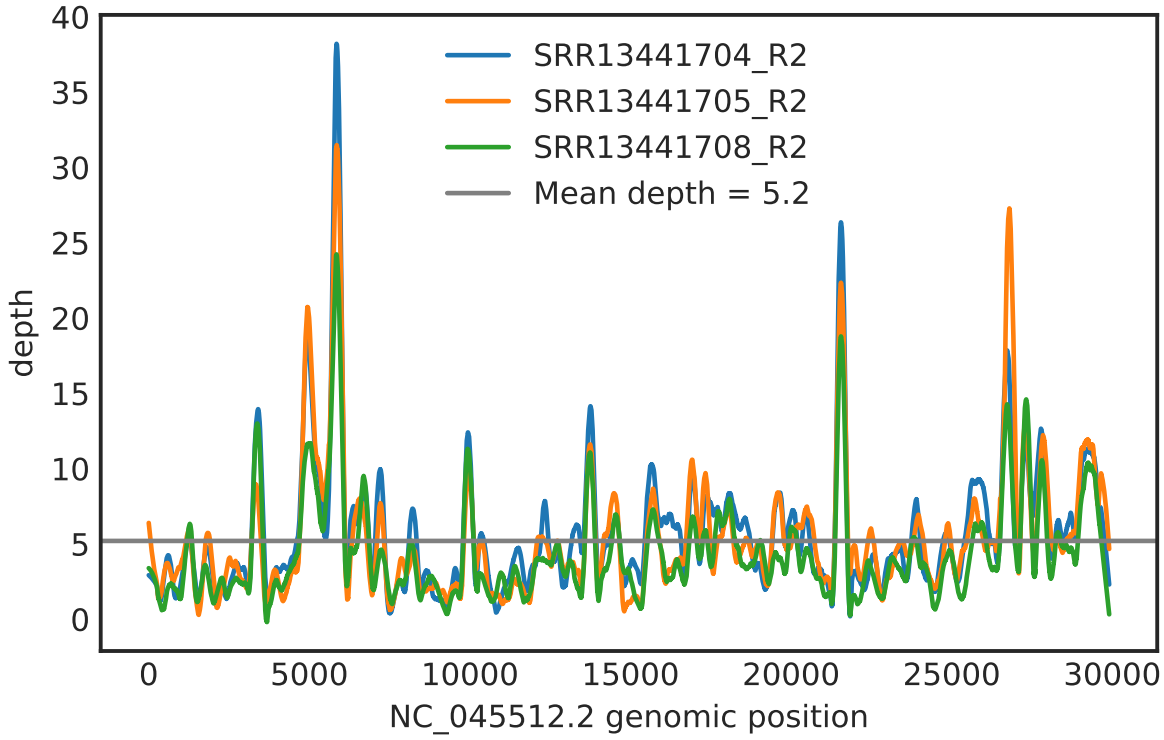

Figure 1. The smoothed coverage of the SARS-CoV-2 reference genome by the R2 reads from samples SRR13441704, SRR13441705 and SRR13441708. The average depth is 5.2 but there are large correlated fluctuations in the coverage.

Table 2. Mutations and deletions found in the samples, extracted from the VCF files. The last 3 columns show multiple allele frequencies (AF) and sequencing depth (DP) values, respectively, where the mutation occurred in multiple samples. The leading "SRR134417" was stripped from the RUN ID in the last column.

POS	REF	ALT	Annotation	Gene	AA change	AF	DP	RUN
8782	C	T	synonymous_variant	ORF1ab	Ser2839Ser	[0.66]	[9]	[08]
13694	C	T	missense_variant	ORF1ab	Thr4482Ile	[0.38, 0.29, 0.22]	[47, 34, 35]	[04, 05, 08]
16156	A	G	missense_variant	ORF1ab	Met5303Val	[0.36, 0.54, 0.5]	[25, 11, 10]	[04, 05, 08]
17039	A	G	missense_variant	ORF1ab	Asn5597Ser	[0.53]	[13]	[05]
17634	C	G	missense_variant	ORF1ab	Asp5795Glu	[0.25]	[20]	[08]
18060	C	T	synonymous_variant	ORF1ab	Leu5937Leu	[0.37, 0.38, 0.46]	[27, 26, 26]	[04, 05, 08]
18082	A	G	missense_variant	ORF1ab	Ile5945Val	[0.46]	[28]	[04]
21761	G	del27	disrupt.inframe.del	S	Ile68_Thr76del	[0.37, 0.33, 1.0]	[8, 9, 5]	[04, 05, 08]
23525	C	T	missense_variant	S	His655Tyr	[0.69, 0.61, 0.66]	[13, 13, 6]	[04, 05, 08]
25498	C	T	missense_variant	ORF3a	Pro36Ser	[0.45]	[11]	[05]
26458	G	T	missense_variant	E	Asp72Tyr	[0.5]	[14]	[05]
26895	C	T	missense_variant	M	His125Tyr	[0.51]	[60]	[05]
28144	T	C	missense_variant	ORF8	Leu84Ser	[0.64, 0.71, 0.66]	[17, 14, 15]	[04, 05, 08]
29449	G	T	synonymous_variant	N	Val392Val	[0.43, 0.24]	[32, 29]	[04, 08]

Nextstrain scheme, in B19 clade. We note that the UShER tool did not list the 27nt long deletion at 21761 as a difference from the reference genome. It also gave a warning for the unusually high parsimony score, i.e. the sample has a very unique mutation composition. The sample differs by 8 mutations plus the not counted deletion from the closest sample among the more than 6 million samples currently deposited to GISAID, GeneBank, COG-UK and CNCB.

The 27nt long deletion at 21761 (Spike_I68del, Spike_H69del, Spike_V70del, Spike_S71del, Spike_G72del, Spike_T73del, Spike_N74del, Spike_G75del, Spike_T76del in amino acid notation) is especially intriguing. Only 61 samples of the more than

Table 3. Pileup for the mutation positions which were not identified in all samples. The lower and upper case letters denote mutated nucleotides, while dot and comma the original reference nucleotide. (The meanings of other symbols are not relevant, they can be found in samtools' documentation.) We can observe that although mutations were not called by the workflow, mutated bases indeed occur with about the same ratio in all samples.

pos	ref	SRR13441704	SRR13441705	SRR13441708
8782	C	t.TTa..T	TT,..,	a.tTtTtT.
17039	A	.$....G..,,g..g,	Gg,Gg,,,GTgg^],	..,C.GG,,,,,,g,
17634	C	.......G..Gg,,,g,.,g,.,,,.G.,...	,G.,,,.,.,..G..	.g,.g,G.GG,.,,,.....
18082	A	g,,g,,G,GGg.g,,,g,,,,.g,gGgG	G,,,,n,,,,,GG,G,.g,,G*..,	.g,gg,,,g,g.,g,G,,,,,,,..c,.
25498	C	,.gT.TT,t,,,..T..^].	T.Tg,..T.tT	..t.,t.

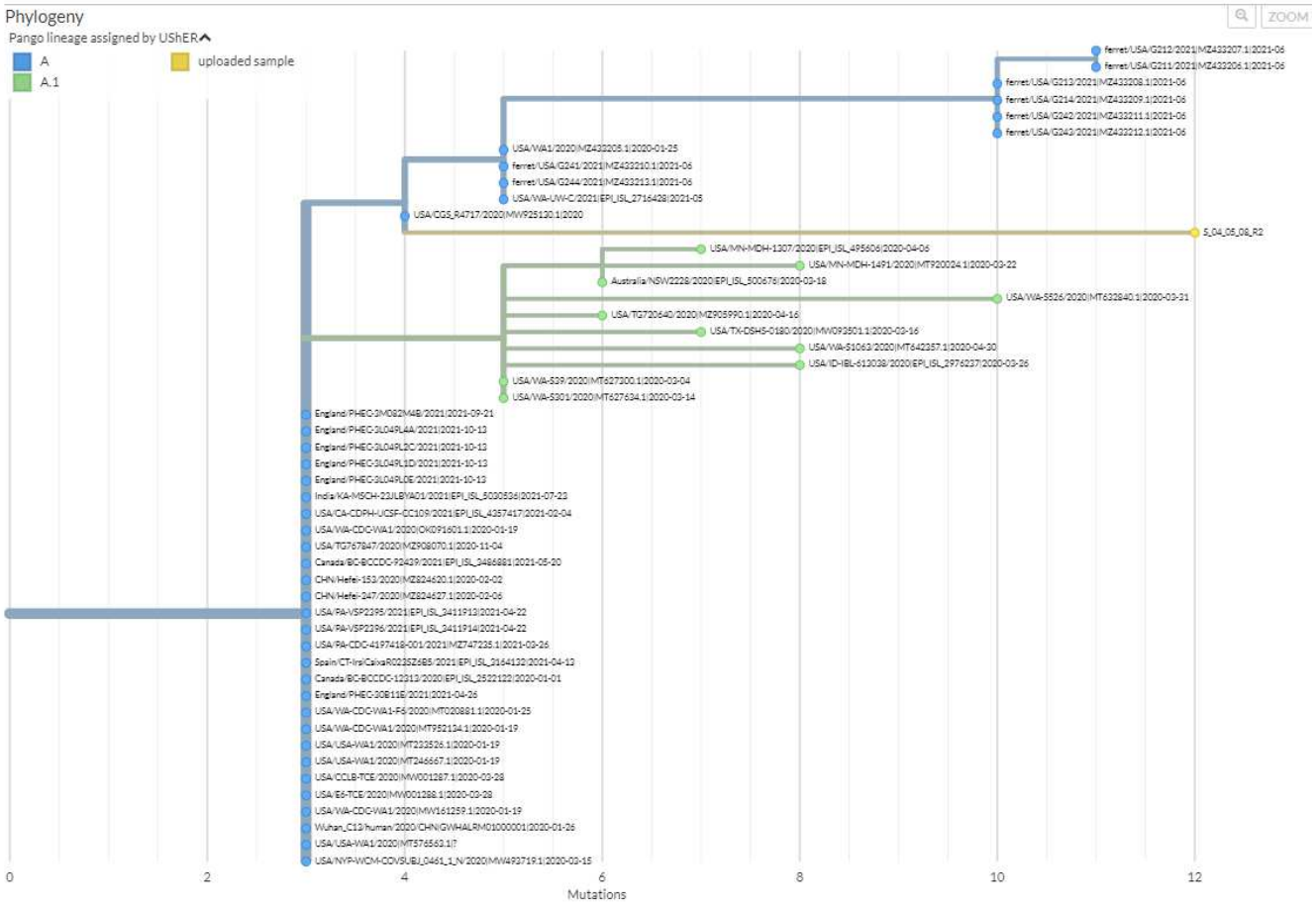

Figure 2. Phylogeny of the pooled sample combined from SRR13441704, SRR13441705 and SRR13441708. Our sample is placed close to the root of the SARS-CoV-2 phylogeny tree, branched from Pangolin lineage "A" with a large parsimony score of 8 mutations.

6 million consensus sequences uploaded as of 2021-12-12 to GISAID harbours this specific deletion, see Table 4. Of these samples, 19 Italian samples have very recent collection date, 2021-11-16 and 2021-11-17. Most probably there is no direct relation, but we mention that the new Omicron variant has a long but not identical overlapping deletion at this site and also Omicron shares the C23525T (H655Y) SNP with the samples studied here.

As the allele frequency (AF) column of Table 2 shows that none of the mutations occur with 100%. There are two mutations at positions 18060 and 18082 which are close enough to have the chance to frequently reside in the same 150nt long sequencing reads. Visual inspection with IGV suggested that this is not the case, these two mutations never occur in the same read, indicating that the two samples may contain multiple strains. The result of gene-wise quasispecies analysis that was performed with aBayesQR, shown in Figure 3, confirmed this finding.

Figure 3. aBayesQR gene-wise quasispecies analysis results. The selected region around mutations at genomic positions 18060 and 18082 shows the mutated bases with white background for sample SRR13441704. Other samples, (not shown) have similar composition. Where mutations occur, they are in distinct reads, none carries both of the mutations. This made possible for the software to separate the distinct variants.

Table 4. Statistics for samples in GISAID that contain 27nt deletion at position 21761. Half of the Italian samples (n=19) have very recent collection date of 2021-11-16 and 2021-11-17.

Host	Location	Pango lineage	Count	Date min	Date max
Canis lupus familiaris	Croatia	B.1.1	1	2021-04	2021-04
Human	Malaysia	B	1	2020-01-24	2020-01-24
Human	Taiwan	B	2	2020-10	2020-11
Human	Taiwan	B	1	2020-03-18	2020-03-18
Human	Belgium	A	1	2021-04-10	2021-04-10
Human	Belgium	A	1	2021-04-15	2021-04-15
Human	Croatia	B.1.1	1	2021-04	2021-04
Human	Czech Republic	B.1.258	1	2020-10-06	2020-10-06
Human	France	B.1.1	2	2020-03-26	2020-03-26
Human	France	P.1.15	1	2020-03-23	2020-03-23
Human	Germany	B.1.1.7	1	2021-02	2021-02
Human	Italy	B.1	38	2021-05-03	2021-11-17
Human	Italy	None	2	2021-06-08	2021-11-16
Human	Russia	None	1	2020-06-04	2020-06-04
Human	Russia	B.1	1	2020-04-19	2020-04-19
Human	Slovenia	B.1	2	2020-03-07	2020-03-07
Human	Slovenia	B.1.1	2	2020-03-09	2020-03-30
Human	Turkey	B.1.1	2	2020-03-17	2020-03-17

Discussion

The PRJNA692319 project's original objective was to apply shotgun metagenomics to tundra soils in maritime Antarctica to determine the effects of sea animal activities on the nitrogen cycle microbial community and function gene. At the sequence archive they did not report related scientific publication, but we could find two articles Dai et al. 2021[16] and Wang et al. 2019[17] with overlapping authors and affiliation at University of Science and Technology of China, Hefei, China. These publications list samples with the same identifiers and based on that SRR13441704 and SRR13441705 were collected from "Penguin colony soil" while SRR13441708 from "background tundra soils on the upland areas". Details of the wet-lab procedure, sequencing library preparation and the date and location of the sequencing are not recorded at the sequence archives. These would be crucial pieces of information to decide whether the detected SARS-CoV-2 content has real biological origin or it is the result of lab contamination or sequencing artifact.

In either case we find the samples very interesting. The variant seems to be quite different from all other known samples, but at the same time harbours only a few mutations with respect to the reference genome that suggests early origin. According to the epidemics reports China is almost free of COVID-19 except the time interval between late 2019 to April 2020 and also no widespread infection was reported outside of Wuhan and Hubei province. This makes the chance contamination from an infected person very unlikely or indicates wide unknown latent spread of infections. On the other hand, true presence of SARS-CoV-2 in the collected samples seems even more unlikely and intriguing.

Details on the sample processing methods and dates will be requested from the original authors and the preprint can be extended with this information. If part of the samples are still available, further investigation may give answers to many of the

open questions and help to either make sequencing procedures more reliable or to decipher the origin of SARS-CoV-2.

Acknowledgments

This work was financed by EU Horizon 2020 programs VEO No. 874735 and BY-COVID No. 101046203. The authors thank the GISAID database and all labs who contributed SARS-CoV-2 sequence data. A full acknowledgement table for all GISAID authors can be found at https://github.com/lgozasht/SARS-COV-2_COLLECTIVE_ANALYSIS/blob/master/gisaid_acknowledgements.tsv.gz. Special thanks for the original researchers for collecting and analysing the PRJNA692319 samples and for all others who upload full raw sequencing data and metadata to ENA and SRA public archives.

Additional information

The authors declare no conflict of interest.

References

1. Arita, M., Karsch-Mizrachi, I. & Cochrane, G. The international nucleotide sequence database collaboration. *Nucleic Acids Res.* **49**, D121–D124 (2021).
2. Spisak, S. *et al.* Complete genes may pass from food to human blood. *PLoS One* **8**, e69805 (2013).
3. Prjna692319 data set. https://www.ebi.ac.uk/ena/browser/view/PRJNA692319. Accessed: 2021-12-10.
4. Sars-cov-2 reference genome nc_045512.2. https://www.ncbi.nlm.nih.gov/nuccore/1798174254. Accessed: 2021-11-29.
5. Ena workflow for sars-cov-2 genome alignment and variant calling. https://github.com/enasequence/covid-sequence-analysis-workflow/blob/master/workflow.nf. Accessed: 2021-11-15.
6. Covid-19 data portal, share site. https://www.covid19dataportal.org/. Accessed: 2021-11-29.
7. Li, H. & Durbin, R. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics* **25**, 1754–1760 (2009).
8. Danecek, P. *et al.* Twelve years of samtools and bcftools. *Gigascience* **10**, giab008 (2021).
9. Wilm, A. *et al.* Lofreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic acids research* **40**, 11189–11201 (2012).
10. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. biotechnology* **29**, 24–26 (2011).
11. Ahn, S. & Vikalo, H. abayesqr: A bayesian method for reconstruction of viral populations characterized by low diversity. *J. Comput. Biol.* **25**, 637–648 (2018).
12. O'Toole, Á. *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.* **7**, veab064 (2021).
13. Pangolin: Phylogenetic assignment of named global outbreak lineages - web application. https://cov-lineages.org/resources/pangolin.html. Accessed: 2021-12-11.
14. Turakhia, Y. *et al.* Ultrafast sample placement on existing trees (usher) enables real-time phylogenetics for the sars-cov-2 pandemic. *Nat. Genet.* **53**, 809–816 (2021).
15. Usher: Ultrafast sample placement on existing tree service. https://hgwdev.gi.ucsc.edu/cgi-bin/hgPhyloPlace. Accessed: 2021-12-11.
16. Dai, H.-T., Zhu, R.-B., Sun, B.-W., Che, C.-S. & Hou, L.-J. Effects of sea animal activities on tundra soil denitrification and nirs-and nirk-encoding denitrifier community in maritime antarctica. *Front. microbiology* **11**, 2537 (2020).
17. Wang, Q., Zhu, R., Zheng, Y., Bao, T. & Hou, L. Effects of sea animal colonization on the coupling between dynamics and activity of soil ammonia-oxidizing bacteria and archaea in maritime antarctica. *Biogeosciences* **16**, 4113–4128 (2019).