Forensic analysis of novel SARS2r-CoV identified in game animal datasets in China shows evolutionary relationship to Pangolin GX CoV clade and apparent genetic experimentation

Adrian Jones
Independent bioinformatics researcher, Melbourne, Australia
Steven E Massey (stevenemassey@gmail.com)
University of Puerto Rico - Rio Piedras
Daoyu Zhang
Independent bioinformatics researcher, Sydney, Australia
Yuri Deigin
Youthereum Genetics Inc., Toronto, Ontario, Canada https://orcid.org/0000-0002-3397-5811
Steven C. Quay
Atossa Therapeutics, Inc., Seattle, WA USA https://orcid.org/0000-0002-0363-7651

Research Article

Keywords: pangolin, coronavirus, SARS-CoV-2, game animal, forensic, metagenomics, Guangxi

Posted Date: August 3rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1836803/v2

Forensic analysis of novel SARS2r-CoV identified in game animal datasets in China shows evolutionary relationship to Pangolin GX CoV clade and apparent genetic experimentation

Adrian Jones[1], Steven E. Massey[2*], Daoyu Zhang[3] , Yuri Deigin[4*] and Steven C. Quay[5]
[1] Independent Bioinformatics Researcher, Melbourne, Australia
[2] Biology Dept, University of Puerto Rico - Rio Piedras, PR USA
[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: stevenemassey@gmail.com

Abstract

Pangolins are the only animals other than bats proposed to have been infected with SARS-CoV-2 related coronaviruses (SARS2r-CoVs) prior to the COVID-19 pandemic. Here we examine the novel SARS2r-CoV we previously identified in game animal metatranscriptomic datasets sequenced by He et al. (2022) and find that sections of the partial genome phylogenetically group with Guangxi (GX) pangolin CoVs (GX PCoVs), while the full RdRp sequence groups with bat-SL-CoVZC45. While the novel SARS2r-CoV is found in 6 pangolin datasets, the same CoV is also found in 10 additional NGS datasets from 5 separate mammalian species and is likely related to contamination by a laboratory researched virus. Absence of bat mitochondrial sequences from the datasets, the fragmentary nature of the virus sequence and the presence of a partial sequence of a cloning vector attached to a SARS2r-CoV read suggests that it has been cloned. We find that NGS datasets containing the novel SARS2r-CoV are contaminated with significant *Homo sapiens* genetic material, and numerous viruses not associated with the host animals sampled. We further identify the dominant human haplogroup of the contaminating *H.sapiens* genetic material to be F1c1a1, which is of East Asian provenance. The association of this novel CoV with both bat CoV and the Guangxi pangolin CoV (GX PCoV) clades is an important step towards identifying the origin of the GX PCoVs.

Keywords: pangolin, coronavirus, SARS-CoV-2, game animal, forensic, metagenomics, Guangxi

Introduction

A zoonotic jump from animals has been proposed as a potential origin for SARS-CoV-2. The virus emerged in Wuhan in late September/early October [1] to as late as between mid-October to mid-November [2] and spread worldwide, leading to over 6 million deaths to date. The Huanan Seafood Market was implicated early in the pandemic as a potential source of the virus [3] but several of the earliest reported cases had no link to the market [4], and importantly no animals at the Huanan Seafood Market were found to test positive for the virus [5]. Furthermore, several early COVID-19 market cases may have occurred via human to human transmission rather than via a zoonotic jump [6]. In addition, the presence of only lineage B associated with human infections at the market [5] makes a market origin less likely, as lineage B is likely a derived lineage, while lineage A is likely ancestral [7]. An alternative hypothesis that a SARS-CoV-2 progenitor virus was present in one of the laboratories in Wuhan conducting SARSr-CoV research and accidentally escaped has not been sufficiently investigated [8].

Before the COVID-19 outbreak, bat-SL-CoVZC45 and bat-SL-CoVZXC21 were the only two published SARS2r-CoVs [9][10]. These sequences were obtained from *Rhinolophus pusillus* bats in Zhoushan city between June 2015 and February 2017, and were isolated through the direct inoculation of homogenized bat intestinal material into the brains of 3-day old suckling Balb/c mice [9]. However, the bat-SL-CoVZC45 and bat-SL-CoVZXC21 genomes are only 89.12% and 88.65% identical to SARS-CoV-2 respectively. Shortly after the publication of the SARS-CoV-2 genome, the RaTG13 genome was published [11], with a significantly higher nucleotide identity to SARS-CoV-2 (96.14%). The virus was sampled from a mineshaft in Mojiang, Yunnan where in 2012 six miners were infected with a SARSr-CoV, killing three [12].

In addition to RaTG13, the 3 other closest relatives to SARS-CoV-2 were sampled from Southern Yunnan [3] [13] and Northern Laos [14], which are located 1500km and 1700km from Wuhan respectively. Definitive routes for SARS-CoV-2 travel to Wuhan have not been identified via either natural zoonosis or research-related sampling, although a research conduit existed for the transportation of SARS2r CoVs from Southern China to the Wuhan Institute of Virology (WIV). Extensive sampling of bats and bat fecal material in Yunnan has been conducted by researchers affiliated with the WIV, Wuhan University and Guangdong Institute of Applied Biological Resources (GIABR) [15] [16] [17] [18] [19] (Fig. 1).

In addition to bat-hosted SARS2r-CoVs, two pangolin hosted SARS2r-CoV clades have been proposed: Guangdong (GD) pangolin CoVs (PCoVs) and Guangxi (GX) PCoVs. GD PCoV was first identified on the 31st January 2020 by Wong [20] as SARS-CoV-2-related, by finding that the receptor binding domain (RBD) amino acid sequence for a novel coronavirus found in pangolin organ tissues sequenced by [21] at the GIABR had high homology to the RBD for SARS-CoV-2. Consequently, it has been proposed that SARS-CoV-2 acquired its RBD via

recombination with, or from an ancestor in common with a GD PCoV [22]. However, the SARS2r-CoV in the Liu et al. (2019) datasets is likely contamination related rather than pangolin hosted given the low number of SARS2r-CoV reads, human genomic origin content, presence of non-pangolin hosted virus read content in similar abundance as SARS2r-CoV reads and correlation of the presence of SARS2r-CoV reads with high bacterial content. This is consistent with inadvertent contamination, with the SARS2r-CoV sequences misattributed to a pangolin host [23] [24].

The datasets sequenced by [21] were used in a series of papers published in March through June 2020 by [25] [26] [27] and [28] in support of GD PCoV genome assemblies. However, in addition to widespread contamination present in supplied RNA-Seq datasets, Xiao et al. (2020) provided PCoV GD_1 sequences in synthetic plasmids as evidence for natural infection of pangolins. Furthermore, the PCoV GD_1 genome cannot be assembled without these sequences [23]. Although Hassanin et al. [29] interpret circular read coverage pattern in sample M1 sequenced by Xiao et al. (2020) to indicate circular RNA molecules generated from GD PCoVs, Jones et al. [23] infer the circular pattern as resulting from circularized cDNA generated during the ligation step of molecular cloning. Coronaviruses have never been found to generate circular RNA and furthermore molecular cloning is consistent with the discovery of PCoV sequences inserted into plasmids in the same BioProproject. As such, molecular cloning of PCoV GD_1 is a more parsimonious explanation for this read coverage pattern [23] .

Datasets supporting assembly of the GD PCoV MP789 genome are also problematic. Two datasets, an amplicon dataset and dataset GZ1-2 were provided by [30] in support of MP789 genome generation in addition to two samples from [21]. Dataset GZ1-2 had a very low Pangolin CoV MP789 read count, with only 16 reads with unique coverage, including 6 SNVs and a 6nt insert relative to MP789. Importantly, even with the inclusion of the amplicon dataset and sample GZ1-2, PCoV M789 cannot be assembled, as the four combined datasets have 3nt gap in coverage of, and a SNV relative to the PCoV MP789 reference genome [23].

Guangxi (GX) PCoVs were first reported by Lam et al. [25] on the 18th February 2020, just over two weeks after [20] identified reads in [21] datasets with high amino acid identity to the SARS-CoV-2 RBD. GX PCoVs identified by Lam et al. form a separate clade to GD PCoVs and are more distantly related to SARS-CoV-2 (85.9%) than GD PCoVs (90.6%) [25] [26] [31]. The spike proteins of GX PCoVs have a fairly high amino acid similarity in the S1 N-Terminal Domain (NTD) to SARS-CoV-2, significantly higher than GD PCoVs, but conversely a lower similarity in the RBD to SARS-CoV-2 and GD PCoVs (Supp. Figs 1,2). GX PCoV sequencing data is far more restricted than for GD PCoVs and as such the potential for microbial profiling for viral host identification, viral contamination and potential synthetic plasmid contamination is limited.

To identify viruses representing a high risk of crossover to humans, including a search for potential reservoirs of SARS-CoV and SARS-CoV-2, He et al. undertook metatranscriptomic analysis of 1725 game animals from 16 species across China [32]. The study included 423 *Paguma larvata* (Masked palm civet) specimens, a species implicated in bidirectional exchange of SARS-CoV with humans [33]. Both *Manis javanica* (Malayan) and *Manis pentadactyla* (Chinese) pangolins were also sampled. Notably, of these 16 species, 7 species or related species were sold at the Huanan seafood market in November 2019 [34]. He et al. note "no viruses closely related to either SARS-CoV or SARS-CoV-2 (or other sarbecoviruses) were detected in any of animals examined", and no detection of SARSr-CoVs is mentioned anywhere in virus species classification of the samples. However Jones et al. unexpectedly identified a novel SARS2r-CoV in 6 pangolin (Malayan and Chinese) and 3 Malayan porcupine (*Hystrix brachyura*) metatranscriptomic datasets [23]. The SARS2r-CoV partial genome has high homology to the GX pangolin CoVs in the NSP 4 N and C terminal regions but is more closely related to bat-SL-CoVZC45 in the RdRp coding region.

Here we expand on our previous analysis to find GX_ZC45r-CoV, a novel bat-SL-CoVZC45-related coronavirus, in two coypu (*Myocastor coypus*), two Malayan porcupine, one each of Asian badger (*Meles leucurus*), Masked palm civet and Hoary bamboo rat (*Rhizomys pruinosus*) datasets sequenced by He et al., in addition to the 6 pangolin and three Malayan porcupine datasets described in [23]. We undertake more extensive mitochondrial alignments, human mitochondrial haplogroup analysis and more detailed phylogenetic analysis of genomic regions covered by the recovered SARS2r-CoV sequence and propose that the fragmentary nature of the genome indicates genetic manipulation.

Methods

All SRAs were trimmed using TrimGalore v0.6.7 using default adapter detection. All local blastn searches were made against a local copy of the NCBI nt database (Sayers et al. 2022) downloaded on the 21/01/2021. Multiple sequence alignment was conducted using the MUSCLE algorithm [35] in UGENE v42.0 [36].

Viral alignments

158 SRAs in PRJNA and PRJNA795267 were aligned to a set of viruses identified using fastv [37] and to GX CoV using bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) using the '--very-sensitive' alignment option. Where GX cov was identified, mimimap2 [38] was to align the dataset to a 'GX_ZC45r-CoV gap_filled' genome reference using minimap2 version 2.24 with the following parameters " -MD -c -eqx -x sr --sam-hit-only --secondary=no -t 32".

All SRAs with GX_ZC45r-CoV sequences were pooled and aligned to bat-SL-CoVZC45 (MG772933.1) with poly(A) tail removed and PCoV_GX-P4L (MT040333.1) using bwa-mem version 0.7.17 using default parameters.

Viral reference sequences were downloaded from NCBI (https://ftp.ncbi.nlm.nih.gov/refseq/release/viral/) on 11/05/2021. 64 SRAs from PRJNA793740 and PRJNA795267 (Supp. Fig. 18) were *de novo* assembled using MEGAHIT v.1.2.9 [39]. Final contigs were aligned to the viral reference set using minimap2 v2.22-r1101 with the following settings -MD -c -eqx -2 -t 32 --sam-hit-only --secondary=no. Coverage was calculated using bamdst v1.0.9 (https://github.com/shiquan/bamdst). Results were plotted using matplotlib v3.3.4.

Phylogenetic analyses
Partial RdRp section (297nt) maximum likelihood tree was generated using raxmlGUI 2.0 v2.0.8 [40] using the following parameters: --all --model GTR+I+G --seed 470171 --bs-metric tbe --tree rand{1} --bs-trees 1000.

Model testing in MEGA11 [41] for the 407nt partial RdRp section found a T92+G model was found to have the lowest BIC score.

SimPlot analyses
SimPlot++ groups for GX_ZC45r query plot, genome as named except:
ZXC21: bat-SL-CoVZXC21, ZC45: bat-SL-CoVZC45, PCoV_GX: PCoV_GX-P4L, PCoV_GD: PCoV_MP789, HKU3: HKU3-1, FJ2021: FJ2021D, AH2021: AH2021A.

SimPlot++ groups for PCoV GX (PCoV_GX: GX_P2V, PCoV_GX-P1E, PCoV_GX-P4L, PCoV_GX-P5E, PCoV_GX-P5L) query plot, single genomes except for these groups:
PCoV_GD: PCoV_A22-2, PCoV_MP789, PCoV_SM44-9, PCoV_SM79-9,
BANAL: BANAL-20-103/Laos/2020, BANAL-20-116/Laos/2020, BANAL-20-236/Laos/2020, BANAL-20-236/Laos/2020, BANAL-20-247/Laos/2020, BANAL-20-52/Laos/2020.

SimPlot++_groups for PCoV GD (PCoV_GD: PCoV_A22-2, PCoV_MP789, PCoV_SM44-9, PCoV_SM79-9) query plot, single genomes except for these groups:
PCoV_GX: GX_P2V, PCoV_GX-P1E, PCoV_GX-P4L, PCoV_GX-P5E, PCoV_GX-P5L,
BANAL: BANAL-20-103/Laos/2020, BANAL-20-116/Laos/2020, BANAL-20-236/Laos/2020, BANAL-20-236/Laos/2020, BANAL-20-247/Laos/2020, BANAL-20-52/Laos/2020.

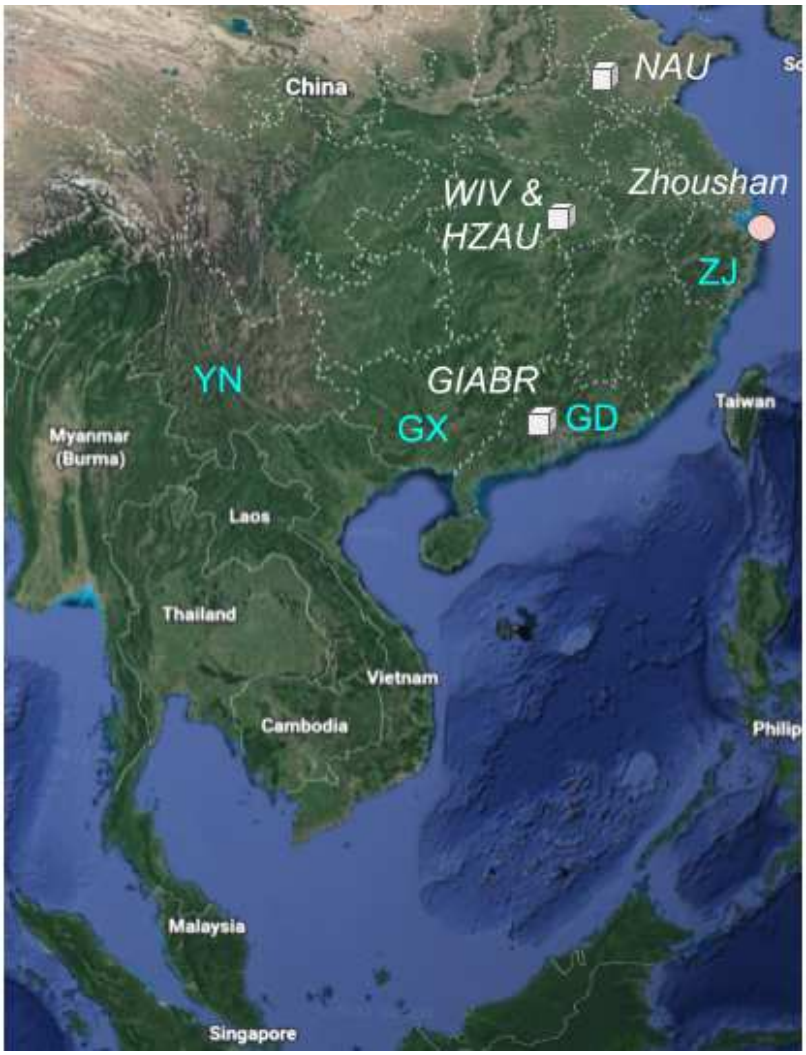


Fig. 1. Location of sequencing centers, provinces and sampling locations discussed here. Sequencing centers as cubes with text in italics: Nanjing Agriculture University (NAU); Wuhan Institute of Virology (WIV), Huazhong Agricultural University (HZAU); Guangdong Institute of Applied Biological Resources (GIABR). Provinces in blue text: Guangdong (GD); Guangxi (GX); Yunnan (YN); Zhejiang (ZJ); sampling locations as pink circles: Zhoushan city, Zhoushan Island, Zhejiang province, China.

Results

We undertook further analysis of all SRA data in BioProjects PRJNA793740 and PRJNA795267 and expand upon [23] to identify GX_ZC45r-CoV sequences in two coypu, two Malayan porcupine and one in each of Hoary bamboo rat, Asian badger and Masked palm civet SRA datasets not previously identified in [23].

We aligned each SRA dataset to a reference genome ‘GX_ZC45r-CoV gap_filled’ consisting of a GX_ZC45r-CoV derived from with empty regions replaced with bat-SL-CoVZC45 (MG772933.1). Coverage for each of the 16 game animal datasets is solely in the non structural protein 4 (NSP4), non structural protein 10 (NSP10) and RNA dependent RNA polymerase (RdRp) coding regions (Fig. 2). The number of reads mapping to ‘GX_ZC45r-CoV gap_filled’ in the seven additional game animal samples was very low at between 1 and 8 reads (Supp. Fig. 3).

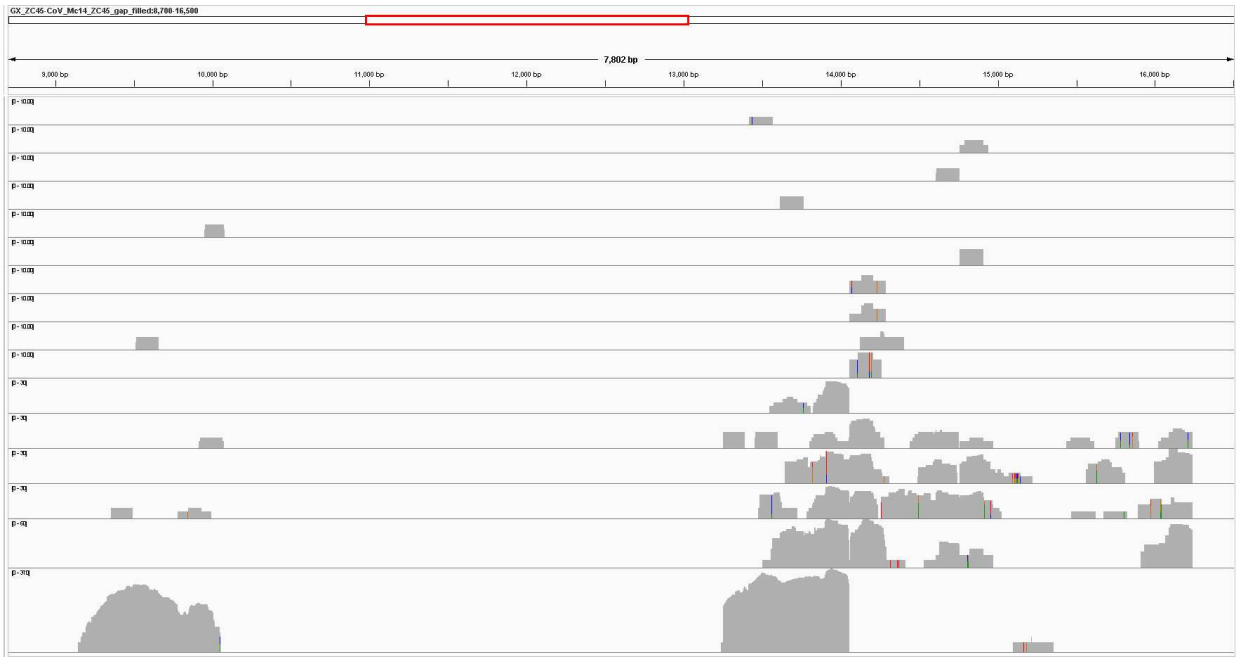


Fig. 2. GX_ZC45r-CoV gap_filled genome coverage by samples (top to bottom tracks): PL-AH-MO-5, MC-HeB-T-1, HB-HuB-A-2, MJ-ZJ-MO-4, MJ-ZJ-MO-6, ML-HeB-F-1, HB-HuB-A-1, MC-HuN-T-1, HB-FJ-NA-3, RP-JX-A-2 (0-10 range), HB-HuB-N-3, MJ-ZJ-MO-3, MJ-ZJ-MO-1, MP-ZJ-MO-4 (0-30 range), MJ-ZJ-MO-2 (0-60 range), HB-FJ-NA-7 (0-370 range). All tracks shown in log scale. Plotted using IGV.

Numerous SNVs relative to bat-SL-CoVZC45 are consistent across samples, with no consensus SNV not also found in other samples, indicating the same strain is found in each of the samples.

Pooled reads from the 16 SRAs containing GX_ZC45r-CoV were aligned to PCoV GX-P4L and bat-SL-CoVZC45. 47 SNVs are found in the NSP4 region when aligned to PCoV GX-P4L (Fig. 2). 104 SNVs and one 18nt missing section (15421-15438nt) are found in the NSP10/RdRp region when aligned to bat-SL-COVZC45 (Fig. 3). Coverage of the NSP4 coding region is incomplete, with complete coverage of the 3’ end, including complete coverage of the NSP4 C-terminus (NSP4_C) but only 50% coverage of the NSP4 transmembrane domain (NSP4_TM).

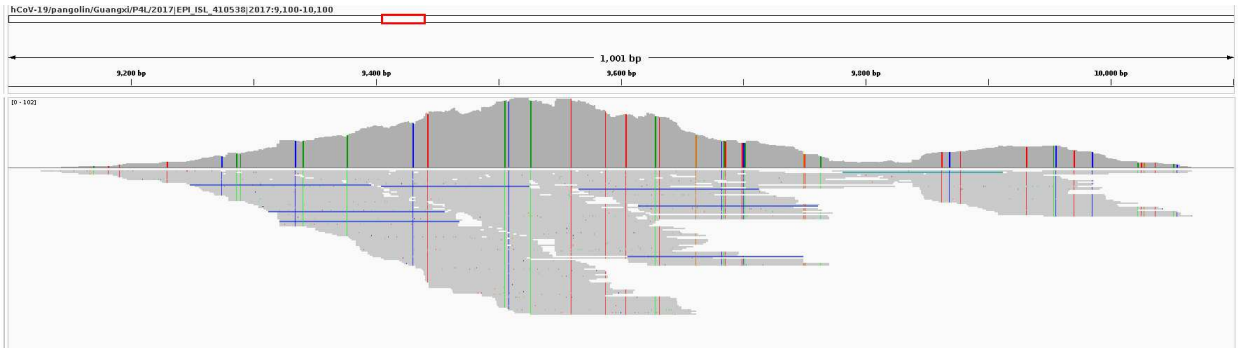


Fig. 3. Alignment of pooled reads to GX P4L, zoomed into the NSP4 region of ORF1a showing 952nt section of mapped reads. Reads coloured by strandness. SNV positions relative to GX P4L shown as vertical lines.

Coverage of the RdRp coding region is complete except for an 18nt missing section, however coverage of the NSP10 coding region is incomplete with 47% of the 5' end of NSP10 not covered by GX_ZC45r-CoV matching reads. We note that a 590nt region at the 5' end of the RdRp coding region (13467-14056 using bat-SL-CoVZC45 as a reference) has markedly higher read coverage than the rest of the RdRp, with an abrupt change at 14057nt. A second anomalous read coverage distribution occurs around position 14758nt relative to the bat-SL-CoVZC45 genome. These potentially indicate the genome may have been sequenced in fragments, potentially as parts of a reverse genetics system ([23], Supp. Figs. 4,5).

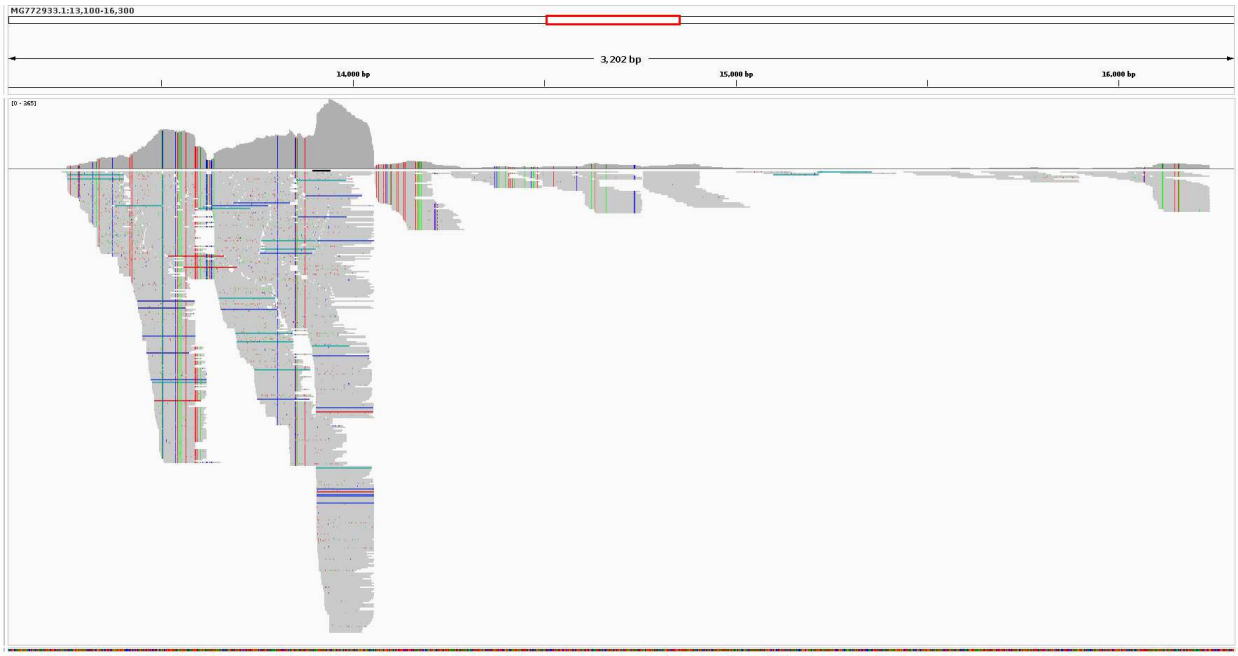


Fig. 4. Alignment of reads to bat-SL-CoVZC45, zoomed in to the NSP10 and RdRp regions. Reads coloured by strandness. SNV positions relative to bat-SL-CoVZC45 shown as vertical lines.

Mitochondrial mapping analysis

The 16 datasets analyzed in Fig. 2 were subjected to a systematic mitochondrial mapping analysis, which mapped the reads to all mitochondrial genomes present on NCBI (Supp. Info. 1 and Source Code). The datasets are heavily contaminated with a range of unexpected mammalian and other eukaryotic species (Supp. Fig. 6). Presumably, if GX_ZC45r-CoV was derived from a cell line or animal tissue, then the relevant mitochondrial genome should be present in all the datasets where the viral sequences are detected. Common mammalian species present in all 16 datasets are *Homo sapiens* (ranging from 16 to 98 % mitochondrial genome coverage), while *Mus musculus* (ranging from 13 to 71 % genome coverage) is present in all but MJ-ZJ-MO-1, PL-AH-MO-5, and ML-HeB-F-1 at 9, 6 and 2% respectively.

Interestingly, both *Homo sapiens* and Masked palm civet mitochondrial genome coverage are higher than the Asian badger mitochondrial genome in ML-HeB-F-1. The mitochondrial genome for *Laodelphax striatellus* an insect vector for Rice stripe virus [42] and *Malassezia restricta* are both found with high coverage in all samples except ML-HeB-F-1 and PL-AH-MO-5. *Malassezia restricta* is a human-hosted basidiomycetous yeast. However environmental *Malassezia* spp. with DNA similar to M. restricta may be widespread in the environment [43]. *Debaryomyces fabri*, a salt tolerant yeast found in varied environments [44], including human skin and is also common to 8 of the 16 datasets. Several species of Aspergillus and Candida yeasts are also found. The widespread yeast occurrences may be indicative of cell culture with yeast growth contaminating the samples, contamination during incubation [45], or yeast contamination of materials during the library preparation stage.

Identification of the human mitochondrial haplogroups

Those reads that mapped to the human mitochondrial genome were used to infer the human mitochondrial haplogroup in the 14 datasets that had significant human mitochondrial genome coverage. The datasets were mapped to the rCRS human reference mitochondrial genome using bowtie2 [46]. Then, mixemt [47] was used to infer the mitochondrial haplogroup. The results are displayed in Supp. Table 1.

A dominant haplogroup, F1c1(a1) was identified in 12 of the 14 datasets. Of the remaining two datasets, MJ-ZJ-MO-3 did not possess any reads that mapped to rCRS, while in MJ-ZJ-MO-2 100 % of human mitochondrial reads were attributed to the H27/H27e haplogroup. A number of minor haplogroups were observed in the remaining datasets : H1aw and H1t2 (MC-HuN-T-1), C (MC-HeB-T-1) and H27/H27e (MJ-ZJ-MO-4).

Haplogroup F1c1a1 is of East Asian origin, and its presence appears consistent with either worker contamination, or a human cell line of East Asian provenance. Haplogroup H27/H27e is of European / Central Asian origin, haplogroup C is found in Northeast Asia and the Americas,

while haplogroup H1 is found in Europe and North Africa. Given their geographical origin, aplogroups H27/H27e and H1 are inconsistent with worker contamination, and could represent cell lines. Given that haplogroup F1c1(a1) dominates the datasets, then if GX_ZC45r-CoV were associated with a human cell line, this haplogroup would be the likeliest candidate. Alternatively, if haplogroup F1c1(a1) derived from worker contamination then this indicates upstream contamination of sequencing reagents / samples rather than index hopping during sequencing. This observation may be useful when considering the source of the GX_ZC45r-CoV contamination.

Simplot analysis

Similarity plot analysis was conducted using SimPlot++ [48]to review the two sections of the genome which were recovered (Fig. 6). The entire recovered section of the NSP4 coding region has the highest similarity to GX PCoVs. The short recovered section of the NSP10 region (188nt when a19nt section of the non-coding region between NSP10 and the RdRp coding region is included) shows several genomes to have high similarity including Longquan-140 and bat-SL-CoVZC45.

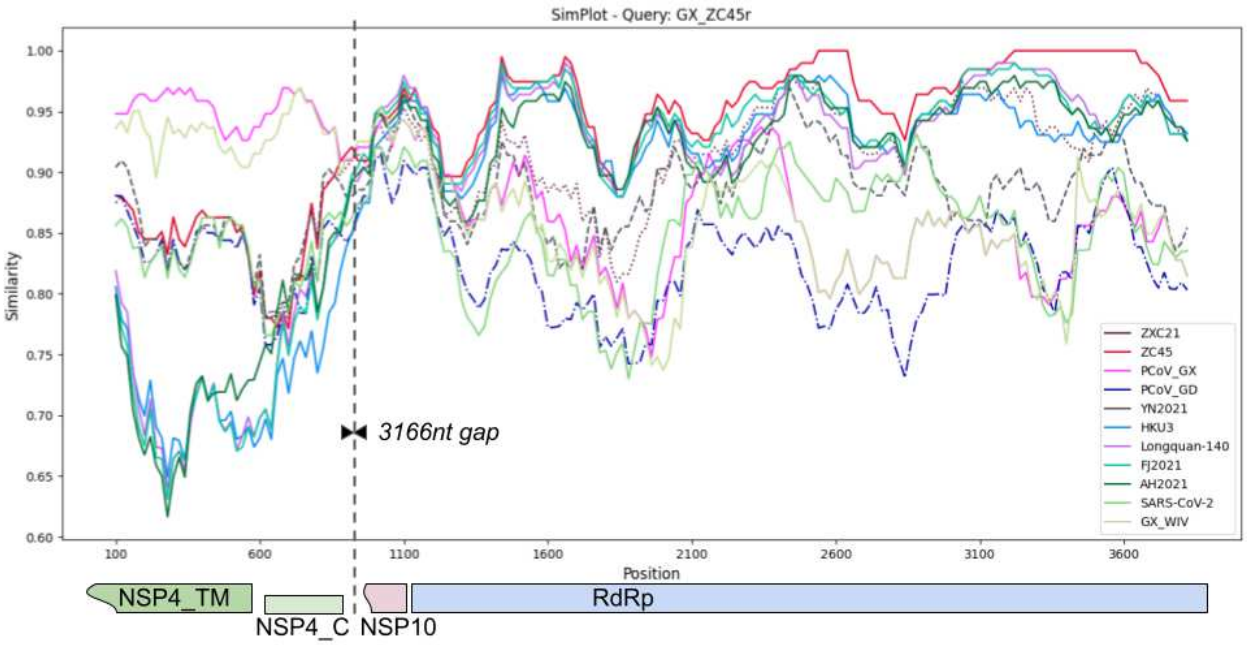


Fig. 6. Simplot analysis of NSP4 region spliced with NSP10 and RdRp gene regions. Plotted using Simplot++ using a 200bp window, 20bp step and default Kimura (2-parameter) distance model. 3166nt section (relative to multi-genome alignment) between covered regions was removed prior to analysis. Solid lines except: bat-SL-CoVZC21 - dotted, YN2021 - dashed, PCoV GD dash-dot. See methods for virus groupings.

Blastn analysis for the same 188nt NSP10 region (including a 19nt section of the 5' end of the non-coding region between NSP10 and the RdRp coding region) shows Sarbecovirus sp. isolate FJ2021D to have highest identity at 96% which includes 6 SNVs, one deletion and one insertion relative to GX_ZC45r-CoV. However PCoVs GX-P5E, GX-P5L, GX-P1E, GX-P4L, BtCoV Longquan-140, and BtCoV SC2018B all have 9 SNVs for 95% identity, for a more parsimonious match not requiring a deletion and insertion. Interestingly, although the absolute number of nucleotide differences is the same or very similar between the bat CoVs and PCoVs the positions of the differences are almost all completely different between the two groups (Supp. Fig. 7).

Over the RdRp coding region, the highest identity over almost the entire region is to bat-SL-CoVZC45 (ZC45), with two regions of 100% identity. However distinct regions of <95% similarity to bat-SL-CoVZC45 are evident, indicating significant evolutionary distance from bat-SL-CoVZC45.

GD PCoV MP789 and the PCoV GX group were also analyzed using SimPlot++ and queried against SARSr-CoVs with high identity over at least part of the GX_ZC45r-CoV partial genome. PCoV MP789 exhibits high identity to bat-SL-CoVZC45, bat-SL-CoVZXC21, RacCS271 and RaTG13 in the NSP4 region, but low identity to bat-SL-CoVZC45 and bat-SL-CoVZXC21 over the RdRp coding region (Supp. Fig. 8). The PCoV GX group exhibits a distinctly high identity to GX_ZC45r-CoV in the NSP4 coding region (Supp. Fig. 9). However, over the RdRp coding region of the genomes analyzed, the PCoV GX group only exhibits highest identity to GX_ZC45r-CoV over a 297nt region (2144-2440nt in Supp. Fig. 9). To quantify the match blastn was used with the 297nt section of GX_ZC45r-CoV as the query and PCoV GX-P4L as the subject and a 93.94% identity was found (279/297nt match). Blastn was again used to analyze this region of PCoV GX-P4L, which was located at 14432-14728nt, against the nt database, which resulted in the highest identity to any genome on NCBI of 91.84% (SARS-CoV-2 (OU470778.1)). This confirms that GX_ZC45r-Cov is the closest known match in this part of the RdRp gene.

Phylogenetic analysis

Phylogenetic trees were constructed for the NSP4, NSP10, RdRp and partial RdRp coding regions which were covered by GX_ZC45r-CoV reads. For the NSP4 region, which plays a role in assembling the viral double-membrane vesicles, a GTR+G+I model was estimated as having the lowest Bayesian information criterion (BIC) and 5 discrete gamma categories were used. GX_ZC45r-CoV exhibits a basal sister relationship to GX CoVs both with unanimous support. GX_WIV [31] exhibits a more diverged genome in this region than related GX_CoVs (Fig. 7).

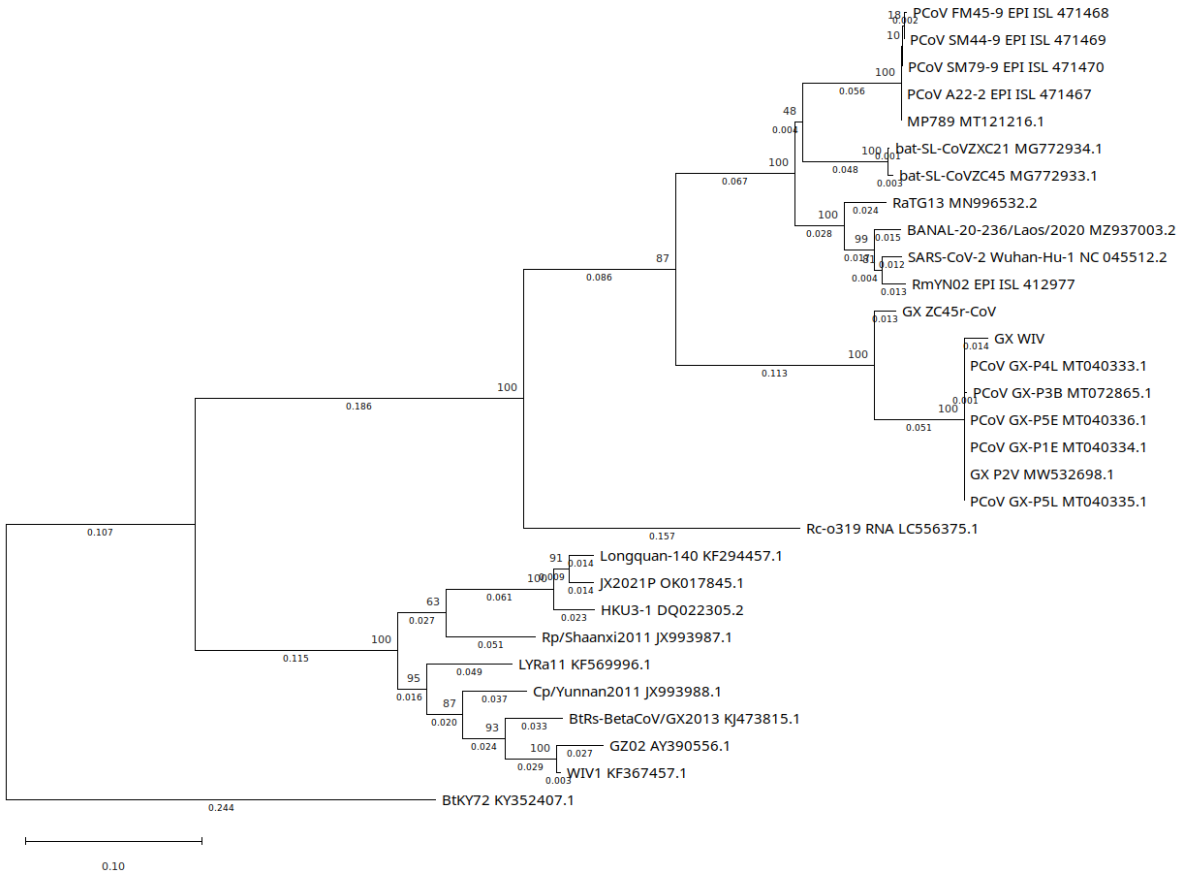


Fig. 7. Partial NSP4 region maximum likelihood tree constructed using a GTR+G+I model with 1000 bootstrap replicates. Branch support percentage is shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Several genomes only had partial coverage of this region: GX_WIV 77.8%; PCoV_GX-P3B 77%; PCoV_FM45-9 98.65%. MP20 had only 11.1% coverage and was excluded from analysis.

Phylogenetic analysis of the NSP10 region, which plays a role in mRNA cap methylation, shows GX_ZC45r-CoV to have a basal sister relationship to the GX CoV clade (Fig. 8). GD PCoVs however are more closely related to the SARS-CoV-2/BANAL clade and form a basal sister clade. Bat-SL-CoV-ZC45 is significantly more divergent from the GX, PD and SARS-CoV-2r CoV clades, and is located in the SARSr-CoV clade.

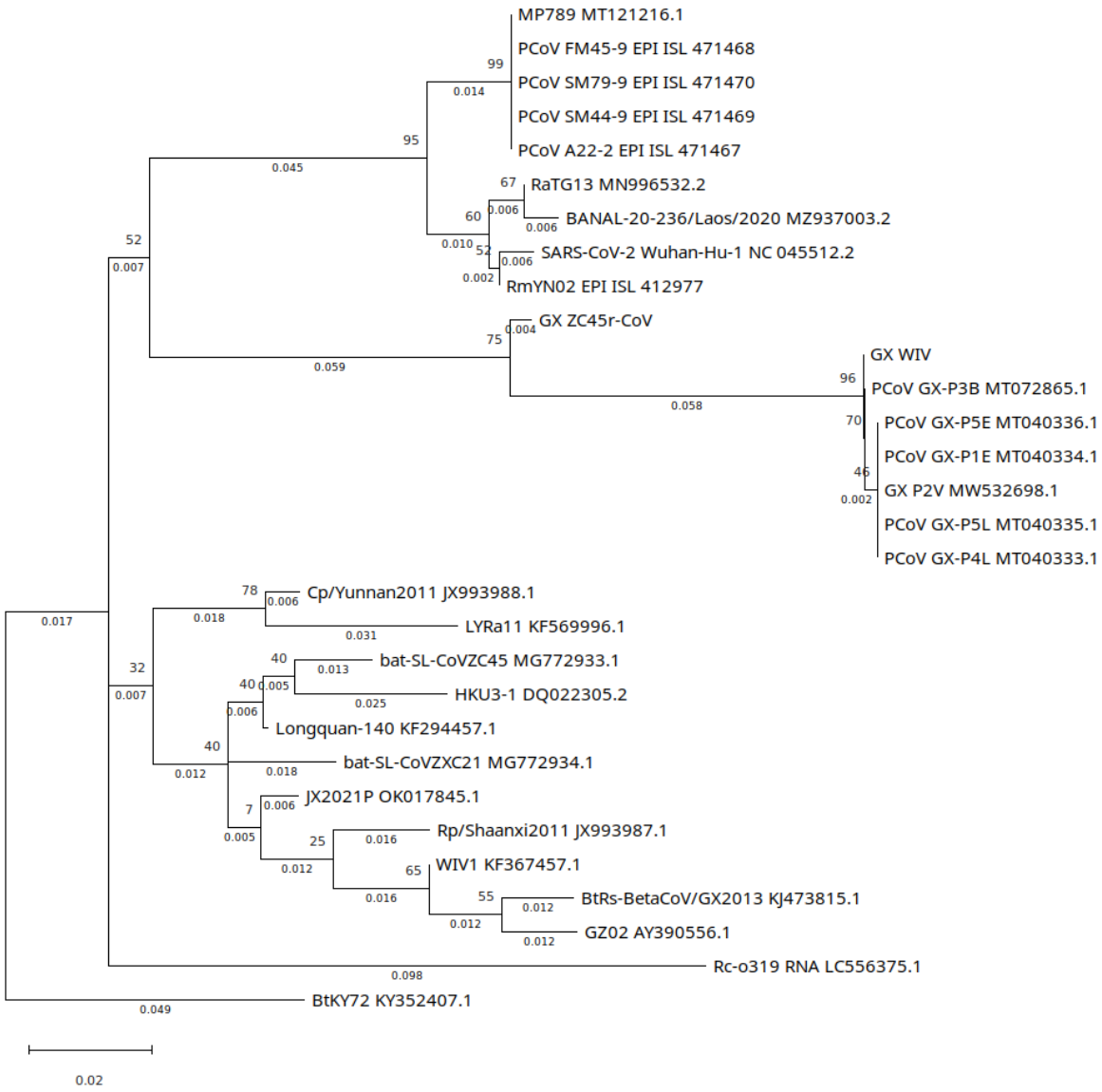


Fig. 8. NSP10 region maximum likelihood tree using Kimura 2 parameter model (K2+G) with 500 bootstrap replications. 5 discrete gamma categories. Branch support percentage is shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site.

In a maximum likelihood tree covering the RdRp coding region, bat-SL-CoVZC45 and GZ_ZC45r-CoV form a clade with an ancestral node in common with HKU3-1 and Longquan-140 and sit on the SARSr-CoV branch (Fig. 9). Similar to the NSP10 phylogenetic tree, all GD PCoVs form a sister clade to the SARS-CoV-2/BANAL clade. A blastn analysis of the RdRp of GX_ZC45r-CoV (including 18nt missing nucleotides) confirms bat-SL-CoVZC45 to have closest identity at 95.85%.

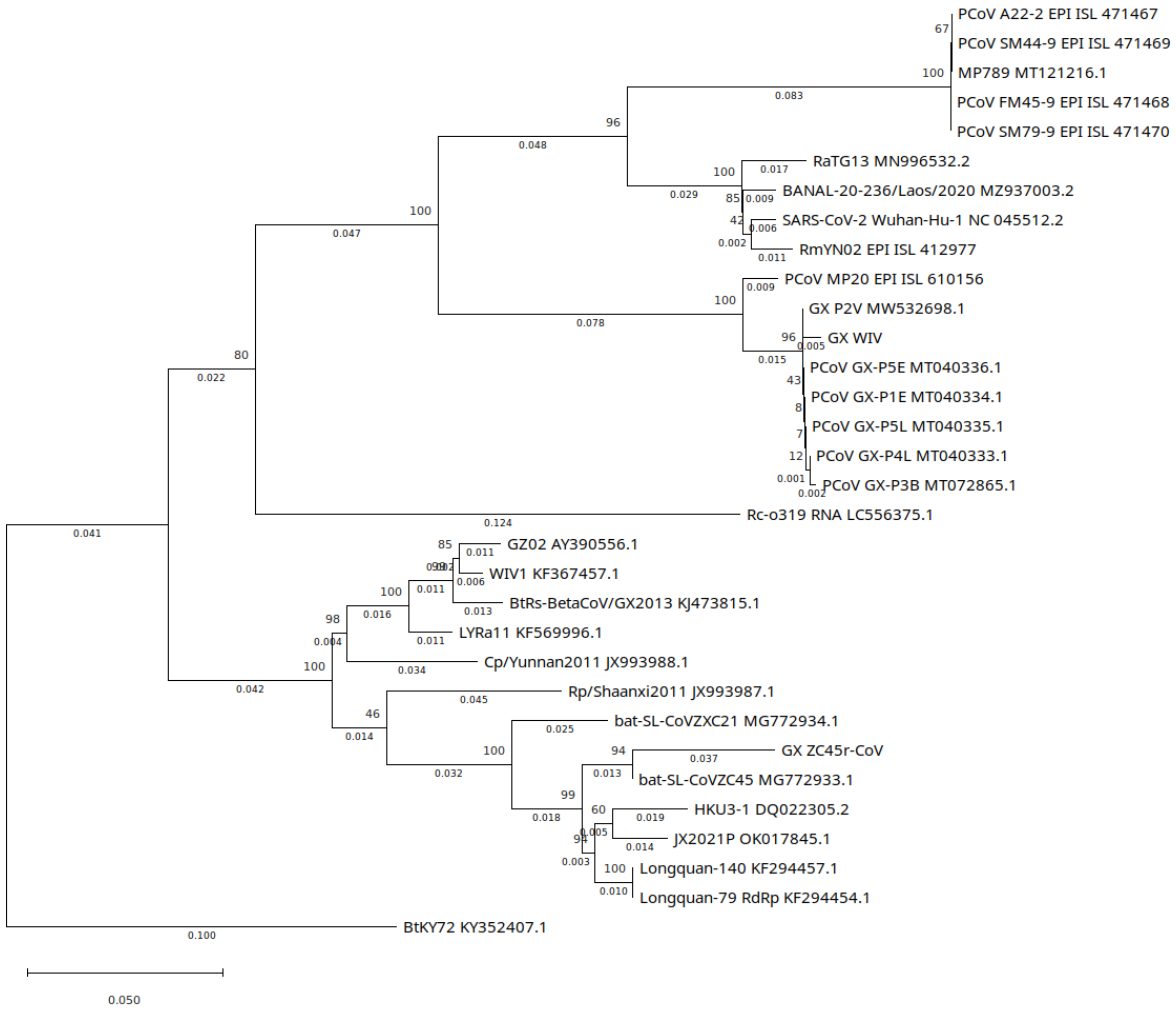


Fig. 9. RdRp region (2795nt) maximum likelihood phylogenetic tree using GTR+G+I model with 1000 bootstrap replicates. Branch support percentage is shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The following genomes have partial coverages: GX_ZC45r-CoV, GX_WIV, PCoV_GX-P3B , MP20, PCoV_FM45-9, comprising 99.4%, 96.5%, 93.6%, 42.6%, and 98.7% respectively.

The phylogenetic relationship of PCoV MP20 to the GX PCoVs in the RdRp gene is similar to the relative relationship of GX_ZC45r-CoV to the GX PCoVs in the NSP4 and NSP10 regions. To further compare the similarity of PCoV MP20 to GX_ZC45r-CoV, a similarity plot was constructed using SimPlot++ using MP20 as the query against selected SARS2-CoV genomes. As coverage of both GX_ZC45r-CoV and MP20 is limited, the four regions over which both genomes had coverage were spliced to form a single contiguous pseudo genome for analysis (Supp. Fig. 10). Where data is available, the two genomes exhibit moderate identity, ranging

between 70-90% except for an approximately 40nt section of the RdRp coding region with a 95-97% identity.

A maximum likelihood tree was generated to test the phylogenetic relationship of a 297nt section of the RdRp gene of PCoV GX-P4L which was found to have highest blastn identity to GX_ZC45r-CoV (Fig. 10). Model selection in MEGA11 was used to identify a T92+G model to have lowest BIC and a GTR+G+I model to have lowest Akaike information criterion (corrected for small same size) (AICc) score, and was found to generate a phylogenetic reconstruction with higher branch support. Similar to findings for the NSP4 and NSP10 coding regions, GX_ZC45r-CoV is situated in a basal sister relationship to GX PCoVs, but with poor branch support. A test on the effect of algorithm implementation, a maximum likelihood tree using a GTR+I+G model was also generated using raxmlGUI with the 297nt section of the RdRp gene of GX_ZC45r-CoV again found to form a basal sister relationship to GX PCoVs (Supp. Fig. 11).

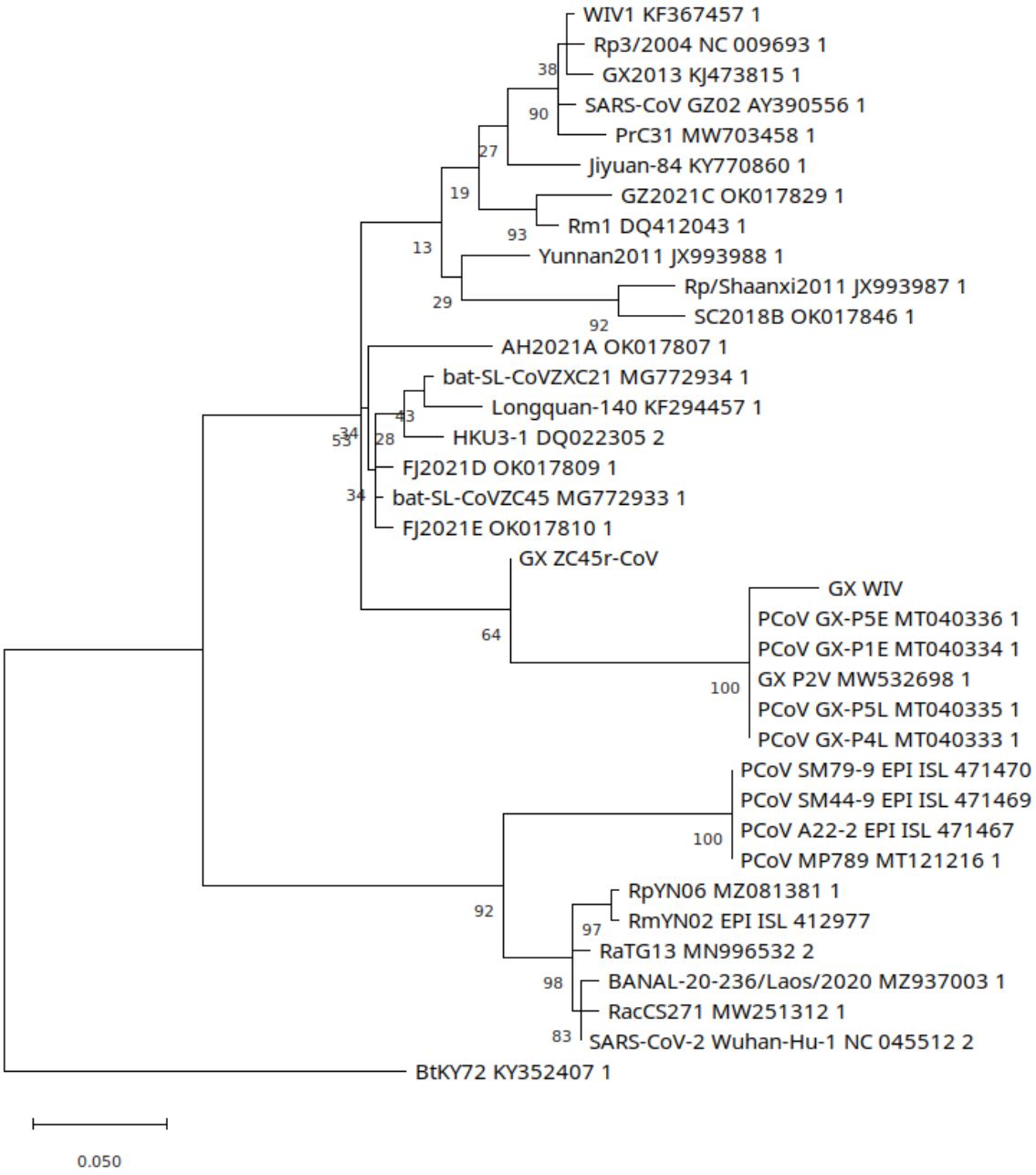


Fig. 10. Partial RdRp section (297nt) maximum likelihood phylogenetic tree using a GTR+G+I model with 100 bootstrap replicates. 7 discrete Gamma categories were used. Generated using MEGA11.

Hu et al. sampled 334 Zhoushan island bats and identified 89 to be carrying SARSr-CoVs, however only two full genomes were recovered, bat-SL-CoVZXC21 and bat-SL-CoVZC45 [9]. 89 partial RdRp amplicon sequences were submitted to GenBank (accessions MG772844 through MG772932). We aligned this set of partial sequences, together with GX_ZC45r-CoV and selected genomes and trimmed the multiple sequence alignment to 407nt. Unfortunately, the

440nt PCR targeted section of the RdRp covered the 18nt gap and a low read depth section of GX-ZC45r-CoV. CoVs bat-SL-CoVZC45, Z2 45 and ZC3 had lowest hanning dissimilarity, with 3nt SNVs relative to GX-ZC45. 11 other Zhoushan bat CoVs also exhibited low dissimilarity with 4 or 5 SNVs in this region. A maximum likelihood phylogenetic tree was constructed using the 23 partial RdRp sequences from Hu et al. [9] with lowest hamming dissimilarity to GX_ZC45r-CoV, and selected other genomes (Supp. Fig. 12). In general, the selected partial Zhoushan RdRp sequences exhibit relatively low differentiation, with GX_ZC45r-CoV clustering within this clade.

Recombination analysis

As the phylogenetic relationships for different parts of GX_ZC45r-CoV differ significantly, we undertook recombination analysis to determine if recombination breakpoints could be identified. We used RDP5 [49] which implements the RDP [50], GENECONV [51], Chimaera [52], MaxChi [53], BootScan [54]and SiScan [55] methods. Two potential recombination regions were detected by more than three methods (Supp. Table 2). The 3' breakpoint of recombination region 1 was detected at approximately the NSP4/NSP10 splice location where a 3166nt gap in genome coverage was removed from alignments, while the 5' end of this potential recombination event was not detected (Supp. Fig. 13). However, although possible recombination region 1 is wholly located within fragment 5 of Temmam et al. [14], the detected breakpoint could be artefactual given that it coincides with a missing section of the genome. Greater genome coverage is required to place higher confidence on, and identify breakpoint location for this potential recombination region.

Potential recombination region 2 is located within the RdRp coding region (Supp. Fig. 14). The major parent identified was bat-SL-CoVZC45 while the minor parent was not identified from the genomes used in alignment (Supp. Table 2). The recombination breakpoints for recombination region 2 were not identified in any of the other sequences analyzed. Neither (possible) recombination regions 1 or 2 were previously identified in other SARS2r-CoV genomes by [10],[56] or [57].

Complicating potential recombination analysis, as GX_ZC45r-CoV appears to have been sequenced as cDNA in plasmids, it is also possible that recombination could have been introduced artificially. However, two possible artificial splices indicated by read depth pattern were not detected as breakpoint sites using RDP5 and multi-sequence alignment review does not indicate an obvious genome change at these positions.

Synthetic vectors

Using a custom python script (Source Code) we searched for common primer and promoter sequences in 23 *de novo* assembled SRA datasets in PRJNA793740 and 15 *de novo* assembled datasets in PRJNA795267 including all SRAs with GX_ZC45r-CoV sequences. Contigs detected with vector sequences were reviewed using Addgene sequence analyzer (https://www.addgene.org/analyze-sequence/) and NCBI BLAST. Fragments of synthetic vectors were detected in all datasets. A partial vector containing an SV40 promoter and Neomycin selection marker [58] was found in MJ-ZJ-NA-2 and MJ-ZJ-F-1 and MH-ZJ-F-1, with the SV40 promoter sequence also found in the HB-FJ-L-2 dataset (Supp. Fig. 15). The partial vector sequence has a similar layout to pSV2 neo, a plasmid used for mammalian cell line expression [59]. A partial vector containing a 575nt sequence of the Woodchuck Hepatitis Virus (WHP) Posttranscriptional Regulatory Element (WPRE) and Puromycin resistance gene (puro) is found in sample MC-GX-A-1 (Supp. Fig. 16), with shorter partial vectors containing WPRE found in MJ-ZJ-MO-1, MP-ZJ-MO-4 and MC-GX-F-1. WPRE is widely used in viral vectors to increase viral expression and titres [60].

Highest coverage of a Tn5 transposase vector (4275-4570nt) was identified in 6 datasets (MJ-ZJ-NA-2, MJ-ZJ-F-1, MJ-ZJ-F-3, MH-ZJ-F-1, MJ-ZJ-F-6 and HB-FJ-F-4) with shorter sequences (478-3738nt) in 7 datasets (MJ-ZJ-MO-6, MJ-ZJ-MO-3, MC-HeB-T-1, MC-GX-A-1, MJ-ZJ-MO-1, MC-HuN-T-1 and HB-HuB-N-3) (Supp. Fig. 17). Contamination by library construction kit materials [61][62] may have been the source of this vector.

Human and murine hosted viruses

We used NCBI STAT to analyze the taxonomy of all 239 SRA datasets in BioProjects PRJNA793740 and PRJNA795267. We additionally ran fastv analysis against the Opengene viral genome kmer collection on 186 of the SRA datasets. We selected viruses that were commonly found across the SRA datasets and aligned 158 datasets to a reference set of viruses using bowtie2. We selected 46 datasets including all those with GX_ZC45r-CoV, Human orthorubaulavirus 2, and Woodchuck hepatitis virus sequences for analysis (Fig. 11, Supp. Figs. 18-20).

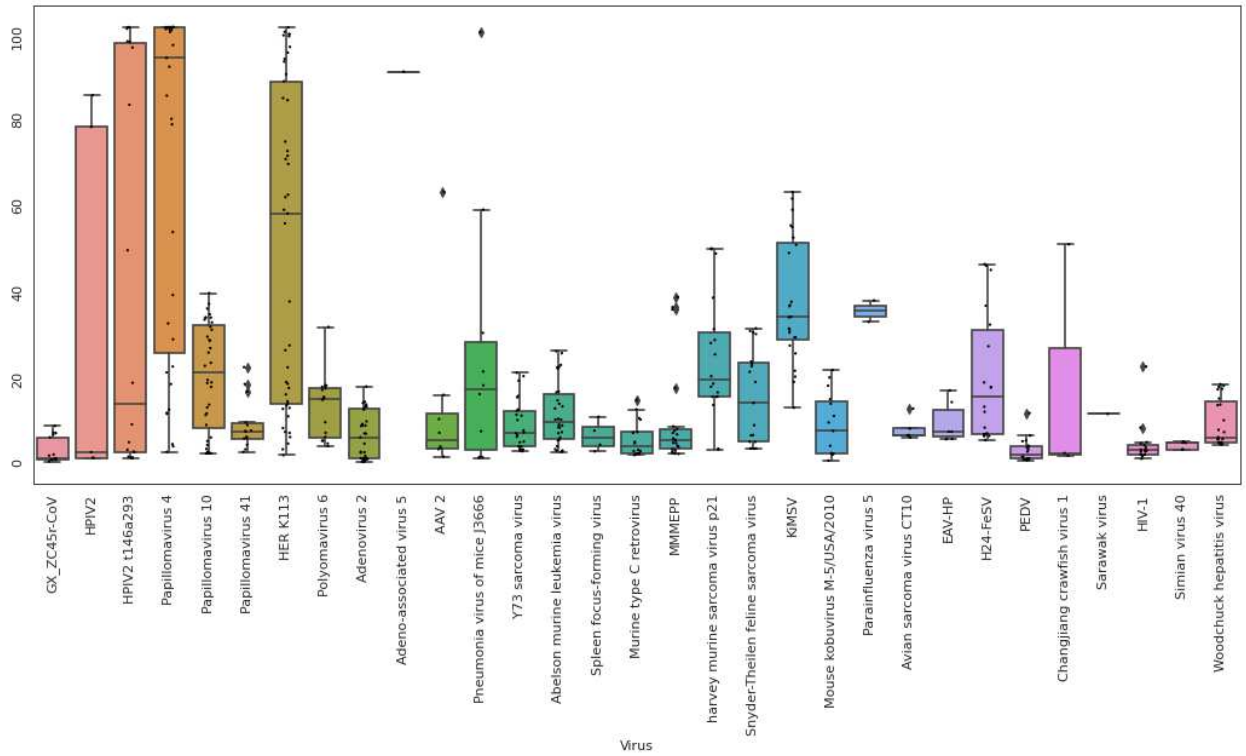


Fig. 11. Box plot of percent coverage of selected viruses for selected SRA's in BioProjects PRJNA793740 and PRJNA795267. Viruses were included only when each virus was present in at least one of the 46 datasets with 10% or greater genome coverage. However a cutoff was not applied to GX_ZC45r-CoV and Simian virus 40. See Supp. Info. 1 for full virus descriptions.

Human hosted viruses are found in high count numbers and coverage including Human papillomaviruses (HPVs) and Human endogenous retrovirus (HERV) K113. Human orthorubulavirus 2 (Human parainfluenza virus 2 (HPIV2)) strain t146a293_HPIV2 (Accession MH892406.1) is found at extremely high levels in pangolin samples MJ-ZJ-MO-4 and MJ-ZJ-MO-3 at 5.91% and 27.66% of the metagenome respectively and the same strain was identified in *Marmota himalayana* (Himalayan marmot) sample MH-HeB-NA-1 [23]. Using a Spearman correlation matrix, the presence of HPIV2 is moderately correlated with HPV and HERV K113 abundances (Supp. Fig. 21). Although He et al. proposed the HPIV2 to be a novel strain naturally infecting 2 Malayan pangolins, the presence of HPIV2 in samples from 5 animal species (Supp. Fig. 18) and extremely high levels in two samples shows the presence of HPIV2 is very likely related to upstream contamination of the datasets and not related to natural infection of the sampled animals. It is interesting to note that of the 16 datasets identified in the BioProjects sequenced by He et al. containing GX_ZC45r-CoV, 8 also contain HPIV2 indicating a possible causal link between the presence of GX_ZC45r-CoV and the presence of HPIV2.

To further determine potential contaminating viruses, *de novo* assembled contigs from 64 SRAs were aligned against the NCBI complete viral RefSeq set (Supp. Fig. 22). Similar to read alignment results, a correlation is evident between the presence of GX_ZC45r-CoV sequences

and high coverage of HPV 4, HPV 10, HERV K113 and fair correlation with human rubulavirus 2.

Numerous murine hosted viruses are also present in both read and contig alignments. In read analysis, pneumonia virus of mice J3666 in highest abundance and coverage in pangolin samples MJ-ZJ-MO-3, MJ-ZJ-MO-6, MJ-ZJ-MO-4, MJ-ZJ-MO-2 and MP-ZJ-MO-4. Kirsten murine sarcoma virus (KiMSV) is found with significant genome coverage in multiple Malayan porcupine, Hoary bamboo rat, coypu and Malayan pangolin samples (Supp. Fig. 20).

Discussion

Our findings are consistent with our previous analysis that the GX_ZC45r-CoV sequences identified in multiple game animal datasets, are contamination related [23]. Here we identify a small number of reads mapping to GX_ZC45r-CoV in seven additional metatranscriptomic datasets: two coypu, two Malayan porcupine, one each of Asian badger, Masked palm civet and Hoary bamboo rat datasets sequenced by [32]. The additional CoV reads map to the same three regions identified by [23]: the NSP4, NSP10 and RdRp coding regions.

The non-random mapping of NGS reads to the GX_ZC45r-CoV gap filled reference genome,solely to the NSP4, NSP10 and RdRp coding regions, across multiple game animal datasets is suggestive of both a single contamination source, and genetic manipulation of the coronavirus. If the coronavirus was present in the originating samples as a wild-type virus or 'live' isolate then the reads should map to the majority of the reference genome. Bat-SL-CoVZC45 was isolated from *R.pusillus* bats [9]. However, no reads mapped to any *Rhinolophus* sp. mitochondrial genome in the 16 datasets examined in Fig. 2, with the exception of MP-ZJ-MO4, which displays 0.88 % coverage of both the *Rhinolophus rex* and *Rhinolophus thomasi* mitochondrial genomes, an amount which is too low to be meaningful (Supp. Info. 2). While the possibility exists that an enrichment procedure was used to isolate the virus, this does not explain the fragmentary nature of the sequences. This observation supports the interpretation that the GX_ZC45r-CoV sequences are not wild-type, as they are not associated with *Rhinolophus* spp. reads. This implies it is either a 'live' isolate cultured in a non-Rhinolophid cell line, or represents some form of cloning experiment.

Interestingly, [9] infected live mice with bat-SL-CoVZC45 (reported as rat due to a mistranslation[1]). This indicates that bat-SL-CoVZC45 could be culturable in a mouse cell line. We note that all of the SRA datasets listed in Fig. 2 have reads mapping to the mouse mitochondrial genome. Alternatively, the presence of human mitochondrial reads in all of the datasets presents the possibility of the CoV being cultured in a human cell line. The majority of

[1] Noted by Andre Goffinet on Twitter https://archive.ph/wip/kBowy

the reads that map to the human mitochondrial reference genome belong to haplogroup F1c1a1, which is of East Asian origin, and may represent worker contamination, or alternatively a cell line of East Asian provenance. However, the non-random mapping of reads to the GX_ZC45r-CoV gap_filled genome appears inconsistent with a 'live' viral culture.

In coronaviruses, infectious clones are typically synthesized as fragments, which are ligated and then inserted into a variety of expression systems (Almazan et al. 2014). Multiple coronavirus reverse genetics methods are used including: *in vitro* ligation using type II restriction endonucleases [63] [64] [65] [66]; [67]. Vaccinia virus vectors [68] a bacterial artificial chromosome (BAC) system [69][16,70]; and transformation associated recombination (Thi Nhu Thao et al. 2020).

Virus rescue is conducted by transfection into a cell line that supports overexpression of the infectious clone RNA. The fragmentary nature of the mapping to the GX_ZC45r-CoV gap_filled genome might suggest that parts of an infectious clone are present in the datasets. This is further supported by the identification of a 21nt pUC57 (a common *Escherichia coli* high copy number plasmid) MCS sequence at the 5' end of reads comprising the NSP10/RdRp coding fragment, indicating the identification of its 5' end. It is interesting to note that the pUC57 plasmid is a widely used vector for coronavirus reverse genetic systems [71]. For comparison, the SARSr-CoV WIV1 BAC infectious clone was split into 8 fragments of lengths from 548nt to 4884nt [70]. However the 5' end of fragment D, which contained the NSP10 and RdRp coding regions, was located near the 3' end of NSP8. In addition, fragment C1 which contains the WIV1 NSP4 coding region is 2619nt, significantly longer than the 952nt long NSP4 coding section recovered in GX_ZC45.

If indeed the reads that map to the GX_ZC45r-CoV genome (Fig. 2) belong to components of an infectious clone then this implies that unpublished SARS2r-CoV genomes and reverse genetics systems are at present being studied in China. Unpublished Beta-CoV reverse genetics systems are not unprecedented, as an unpublished HKU4r-CoV infectious clone with highest homology to BtTp-BetaCoV/GX2012 was identified in a Huazhong Agricultural University sequenced agricultural dataset, with the WIV the likely source [72].

The variant positions revealed when the reads are mapped to the NSP10-RdRp region of the bat-SL-CoVZC45 genome (Fig. 4) are interesting as they indicate some evolutionary divergence. This could represent an independent viral lineage, or divergence of a ZC45r CoV under culture conditions. In this regard, the phylogenetic discordance between NSP4 and NSP10, and RdRp is unlikely to have arisen due to sequence divergence under culture conditions, and appears more consistent with a novel sarbecovirus. In addition, the divergence of the RdRp of GX_ZC45r-CoV from ZC45 (0.037 nucleotide substitutions per site, Fig. 9) seems too large to have occurred via culturing of a ZC45 isolate.

As several reads in the He et al. game animals datasets were found to map seamlessly from the 3' end of NSP10 to the 5' end of the RdRp of a novel SARS2r-CoV (Supp. Fig. 23), the simplest explanation is that the NSP10 and RdRp form a contiguous section of a single genome, which as discussed above appears to have been in the form of cDNA plasmids when sequenced. The recovered NSP4 and NSP10 regions of GX_ZC45r-CoV, as well as a 297nt section within the RdRp coding region all form a basal sister relationship to GX PCoVs. However the RdRp coding region taken as a whole groups with the bat-SL-CoVZC45/HKU3-1/Longquan-140 clade on the SARS branch of a maximum likelihood phylogenetic tree (Fig. 8). This is perplexing and could be explained if either the GX_ZC45r-CoV was a result of natural recombination, or if the genome was artificially manipulated to combine different parts of other genomes.

It is worth mentioning that another unpublished GX-PCoV related coronavirus has recently been recovered from a human cell sequencing dataset published by the WIV [31], however any relation of this genome with the GX_ZC45r-CoV identified here remains to be established.

Although a 21nt section matching the MCS of pUC57 plasmid was found in reads mapping to the novel GX_ZC45r-CoV discussed here, it is unknown if any of the other synthetic vectors identified here are associated with GX_ZC45r-CoV laboratory research. As HPIV2 was found at exceedingly high levels in two samples, it is possible that the vector containing SV40 promoter and Neomycin marker and the vector containing WPRE and puro sequences found here could be related to HPIV2 research. Sequences with a 99.7% identity to HPIV2 strain t146a293_HPIV2 (MH892406.1) were identified in SRA datasets from multiple different animal species sequenced by [32] [23]. That the same HPIV2 strain so closely matched a human strain, was found in multiple animal species, was associated with human genomic contamination and was found to comprise 28% of sample MJ-ZJ-MO-3 almost certainly indicates the virus was not related to a natural infection of pangolins as proposed by He et al. (2022), but stems from laboratory contamination [23]. Alternatively, it is also possible the synthetic vectors identified could have been sourced via contamination from unrelated research.

That GX_ZC45r-CoV exhibits strong phylogenetic grouping with the GX PCoVs in the NSP4, NSP10 and partial RdRp regions, raises questions as to the origin of the GX CoV clade. If the GX_ZC45r-CoV is a natural bat-hosted CoV, given the RdRp similarity to bat-SL-CoVZC45 and high identity of a 407nt region of the RdRp to several Zhoushan bat CoVs sampled by Hu et al., the likely host species is *R.pusillus* located in Zhoushan city on Zhoushan Island, Zhejiang Province (Fig. 1) [9]. If Malayan pangolins captured in Guangxi province [25] were infected with a bat-hosted virus sourced from Zhoushan Island, the question then arises as to how did this occur? Also, if the pangolins were smuggled from outside China the most likely country of origin is Vietnam [73], further distancing the pangolins from Zhejiang province. Additionally,

why has the GX_ZC45r-CoV genome and its relationship to the GX PCoVs not been made public?

It is of further concern that out of nine published GX PCoVs, only one unfiltered/non highly enriched pangolin tissue SRA dataset has been provided to support assembly of a GX PCoV, GX-P3B, a partial genome with 86% coverage of GX_P2V [23][24]. The dataset is of low quality with read lengths highly skewed to very short lengths, and is contaminated with SARS-CoV-2 reads. One other unfiltered/non highly enriched SRA dataset supporting GX PCoV assembly, GX/P2V a Vero-E6 cell culture sample, is also contaminated with SARS-CoV-2 reads.

How pangolins smuggled into China and captured in Guangxi province came to be infected with a CoV highly related to a bat CoV from Zhoushan Island is perplexing. High quality, unfiltered datasets, free of contamination, of infected animals should be provided to the international community to ascertain the veracity of true pangolin infection.

We further request that He et al. publish the full GX_ZC45r-CoV genome, and document the sampling and sequencing history of this CoV. We further ask He et al. to document the source of and processes that led to the contamination of BioProjects PRJNA793740 and PRJNA795267 with human genetic material, HPIV2 and GX_ZC45r-CoV and synthetic vectors.

Conclusion

We analyzed a novel SARS2r-CoV first identified by [23], GX_ZC45r-CoV, and found that sections of the partial genome shares the same ancestor as the GX PCoV clade. However, the RdRp phylogenetically groups with Zhoushan bat CoV bat-SL-CoVZC45, and a partial section of the RdRp groups with multiple Zhoushan bat SARSr-CoVs. As such, it is possible the novel CoV is a *R.pusillus* hosted virus from Zhoushan Island, Zhejiang province, China. We identified the novel SARS2r-CoV in 7 additional game animal RNASeq datasets not previously identified, for a total of 16 datasets from 5 different species, all of which contain significant *H.sapiens* genetic material, and numerous viruses not associated with the host animals sampled. We further identify the dominant human haplogroup of the contaminating *H.sapiens* genetic material to be F1c1a1, which is of East Asian provenance. Reads mapping to the CoV genome, in all animal datasets were solely located in the NSP4, NSP10 and RdRp coding regions. The common read mapping locations, marked truncation of read coverage at two locations in the genome, and the presence of a multiple cloning site sequence from a pUC57 plasmid at the 5' end of a read mapping to the NSP10 region are consistent with the presence of a laboratory cultured virus sequenced from plasmids, rather than from a natural infection of the animals sampled. The novel virus has important evolutionary implications for the GX PCoV clade and we request He et al. to

publish its complete genome and sampling details to help elucidate the origin of the GX PCoV clade of viruses.

Supplementary Information

Supp. Info. 1, containing GX_ZC45r-CoV mapping and virus alignment statistics, and virus names codes used in plots: Supp_Info_1.xlsx
doi: 10.5281/zenodo.6806806
Link: https://zenodo.org/record/6806806

Supp. Info. 2: https://github.com/semassey/Scanning-NGS-datasets-for-mitochondrial-and-coronavirus-contaminants/blob/main/Mito-mappings-16-datasets-GX_ZC45r.zip

Supplementary Data

Supplementary data containing the files listed below can be accessed at:
doi: 10.5281/zenodo.6806806
Link: https://zenodo.org/record/6806806

Novel GX_ZC45r-CoV genome: GX_ZC45r-CoV.fa

Gap filled GX_ZC45r-CoV genome: GX_ZC45-CoV_ZC45_gap_filled_no_polyA.fa

Read alignments to gap filled GX_ZC45r-CoV genome using minimap2: GX_ZC45-CoV_ZC45_gap_filled_no_polyA_minimap2_16_SRA.sam

Source Code

Systematic mitochondrial mapping procedure

The pipeline used for mapping NGS reads to all mitochondrial genomes in the NCBI database, and calculating their relative coverage (thus detecting contaminating eukaryotic species in an NGS dataset), can be found at:

https://github.com/semassey/Scanning-NGS-datasets-for-mitochondrial-and-coronavirus-contaminants

Primer search code

https://github.com/biosciencerесearch/Forensic_analysis_of_novel_SARS2r-CoV

Acknowledgements

We thank Jonathan Latham for feedback which helped improve the manuscript.

Author Contributions

Conceptualization, A.J., D.Z.,Y.D., S.Q.; Methodology, A.J., D.Z., S.M.; Software, S.M., A.J.; Formal Analysis, A.J., D.Z., S.M..; Investigation, A.J., S.M., D.Z., Y.D., S.Q.; Data Curation, A.J., D.Z., S.M.; Writing – Original Draft Preparation, A.J.; Writing – Review & Editing, A.J., S.M., D.Z., Y.D., S.Q.; Visualization, A.J.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Caraballo-Ortiz MA, Miura S, Sanderford M, Dolker T, Tao Q, Weaver S, et al. TopHap: rapid inference of key phylogenetic structures from common haplotypes in large genome collections with limited diversity. Bioinformatics. 2022;38: 2719–2726.

2. Pekar J, Worobey M, Moshiri N, Scheffler K, Wertheim JO. Timing the SARS-CoV-2 index case in Hubei province. Science. 2021;372: 412–417.

3. Zhou H, Chen X, Hu T, Li J, Song H, Liu Y, et al. A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. Curr Biol. 2020;30: 3896.

4. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med. 2020;382: 1199–1207.

5. Gao G, Liu W, Wong G, Wang J, Wang F, Li M. Surveillance of SARS-CoV-2 in the environment and animal samples of the Huanan Seafood Market. Research Square. 2022.

6. Courtier-Orgogozo V, de Ribera FA. SARS-CoV-2 infection at the Huanan seafood market. Environ Res. 2022;214: 113702.

7. Kumar S, Tao Q, Weaver S, Sanderford M, Caraballo-Ortiz MA, Sharma S, et al. An

Evolutionary Portrait of the Progenitor SARS-CoV-2 and Its Dominant Offshoots in COVID-19 Pandemic. Mol Biol Evol. 2021;38: 3046–3059.

8. Harrison NL, Sachs JD. A call for an independent inquiry into the origin of the SARS-CoV-2 virus. Proceedings of the National Academy of Sciences. 2022. doi:10.1073/pnas.2202769119

9. Hu D, Zhu C, Ai L, He T, Wang Y, Ye F, et al. Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. Emerg Microbes Infect. 2018;7: 154.

10. Zhu Z, Meng K, Meng G. Genomic recombination events may reveal the evolution of coronavirus and the origin of SARS-CoV-2. Sci Rep. 2020;10: 21617.

11. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579: 270–273.

12. Rahalkar MC, Bahulikar RA. Lethal Pneumonia Cases in Mojiang Miners (2012) and the Mineshaft Could Provide Important Clues to the Origin of SARS-CoV-2. Front Public Health. 2020;8: 581569.

13. Zhou H, Ji J, Chen X, Bi Y, Li J, Wang Q, et al. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. Cell. 2021;184: 4380–4391.e14.

14. Temmam S, Vongphayloth K, Baquero E, Munier S, Bonomi M, Regnault B, et al. Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. Nature. 2022;604: 330–336.

15. Ge X-Y, Wang N, Zhang W, Hu B, Li B, Zhang Y-Z, et al. Coexistence of multiple coronaviruses in several bat colonies in an abandoned mineshaft. Virologica Sinica. 2016. pp. 31–40. doi:10.1007/s12250-016-3713-9

16. Hu B, Zeng L-P, Yang X-L, Ge X-Y, Zhang W, Li B, et al. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLoS Pathog. 2017;13: e1006698.

17. Luo Y, Li B, Jiang R-D, Hu B-J, Luo D-S, Zhu G-J, et al. Longitudinal Surveillance of Betacoronaviruses in Fruit Bats in Yunnan Province, China During 2009-2016. Virol Sin. 2018;33: 87–95.

18. Yang X-L, Tan CW, Anderson DE, Jiang R-D, Li B, Zhang W, et al. Characterization of a filovirus (Měnglà virus) from Rousettus bats in China. Nat Microbiol. 2019;4: 390–395.

19. Graduate students in the Department of Ecology participate in wildlife science expeditions, field behavioral experiments, and genetic sample collection. 2019.

20. Wong M. nCoV-2019 Spike Protein Receptor Binding Domain Shares High Amino Acid Identity With a Coronavirus Recovered from a Pangolin Viral Metagenomic Dataset. 2020. Available: https://virological.org/t/ncov-2019-spike-protein-receptor-binding-domain-shares-high-amino-acid-identity-with-a-coronavirus-recovered-from-a-pangolin-viral-metagenomic-dataset/362

21. Liu P, Chen W, Chen J-P. Viral Metagenomics Revealed Sendai Virus and Coronavirus Infection of Malayan Pangolins (). Viruses. 2019;11. doi:10.3390/v11110979

22. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of

SARS-CoV-2. Nat Med. 2020;26: 450–452.

23. Jones A, Zhang D, Deigin Y, Quay S. Analysis of pangolin metagenomic datasets reveals significant contamination, raising concerns for pangolin CoV host attribution. arXiv:2108.08163. 2022.

24. Jones A, Massey SE, Zhang D, Deigin Y, Quay SC. Further analysis of metagenomic datasets containing GD and GX pangolin CoVs indicates widespread contamination, undermining pangolin host attribution. arXiv:2207.03288. 2022.

25. Lam TT-Y, Jia N, Zhang Y-W, Shum MH-H, Jiang J-F, Zhu H-C, et al. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature. 2020;583: 282–285.

26. Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou J-J, et al. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature. 2020;583: 286–289.

27. Zhang T, Wu Q, Zhang Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. Curr Biol. 2020;30: 1578.

28. Liu P, Jiang J-Z, Wan X-F, Hua Y, Li L, Zhou J, et al. Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? PLoS Pathog. 2020;16: e1008421.

29. Hassanin A, Jones H, Ropiquet A. SARS-CoV-2-like viruses from captive Guangdong pangolins generate circular RNAs. 2020. Available: https://hal.archives-ouvertes.fr/hal-02616966

30. Liu P, Jiang J-Z, Wan X-F, Hua Y, Li L, Zhou J, et al. Correction: Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? PLoS Pathog. 2021;17: e1009664.

31. Jones A, Massey S, Zhang D, Design, Y., Quay SC. Guangxi pangolin CoV-related virus identified in Wuhan sequenced dataset. In prep.

32. He W-T, Hou X, Zhao J, Sun J, He H, Si W, et al. Virome characterization of game animals in China reveals a spectrum of emerging pathogens. Cell. 2022;185: 1117–1129.e8.

33. Janies D, Habib F, Alexandrov B, Hill A, Pol D. Evolution of genomes, host shifts and the geographic spread of SARS-CoV and related coronaviruses. Cladistics. 2008;24: 111–130.

34. Worobey M, Levy JI, Serrano LM, Crits-Christoph A, Pekar JE, Goldstein SA, et al. The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic. Science. 2022; abp8715.

35. Edgar RC. MUSCLE: multiple sequence alignment with improved accuracy and speed. Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004. 2004. doi:10.1109/csb.2004.1332560

36. Okonechnikov K, Golosova O, Fursov M, UGENE team. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012;28: 1166–1167.

37. Chen S, He C, Li Y, Li Z, Melançon CE. A computational toolset for rapid identification of SARS-CoV-2, other viruses and microorganisms from sequencing data. Brief Bioinform. 2021;22: 924–935.

38. Li H. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 2021.

doi:10.1093/bioinformatics/btab705

39. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015. pp. 1674–1676. doi:10.1093/bioinformatics/btv033

40. Edler D, Klein J, Antonelli A, Silvestro D. raxmlGUI 2.0: A graphical interface and toolkit for phylogenetic analyses using RAxML. Methods in Ecology and Evolution. 2021. pp. 373–377. doi:10.1111/2041-210x.13512

41. Tamura K, Stecher G, Kumar S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol. 2021;38: 3022–3027.

42. Li S, Xiong R, Wang X, Zhou Y. Five proteins of Laodelphax striatellus are potentially involved in the interactions between rice stripe virus and vector. PLoS One. 2011;6: e26585.

43. Amend A. From dandruff to deep-sea vents: Malassezia-like fungi are ecologically hyper-diverse. PLoS Pathog. 2014;10: e1004277.

44. Michán C, Martínez JL, Alvarez MC, Turk M, Sychrova H, Ramos J. Salt and oxidative stress tolerance in Debaryomyces hansenii and Debaryomyces fabryi. FEMS Yeast Res. 2013;13: 180–188.

45. Cornelison CT, Stubblefield B, Gilbert E, Crow SA Jr. Recurrent Aspergillus contamination in a biomedical research facility: a case study. J Ind Microbiol Biotechnol. 2012;39: 329–335.

46. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359.

47. Vohr SH, Gordon R, Eizenga JM, Erlich HA, Calloway CD, Green RE. A phylogenetic approach for haplotype analysis of sequence data from complex mitochondrial mixtures. Forensic Sci Int Genet. 2017;30: 93–105.

48. Samson S, Lord É, Makarenkov V. SimPlot ++: a Python application for representing sequence similarity and detecting recombination. Bioinformatics. 2022. doi:10.1093/bioinformatics/btac287

49. Martin DP, Varsani A, Roumagnac P, Botha G, Maslamoney S, Schwab T, et al. RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. Virus Evol. 2021;7: veaa087.

50. Martin D, Rybicki E. RDP: detection of recombination amongst aligned sequences. Bioinformatics. 2000;16: 562–563.

51. Padidam M, Sawyer S, Fauquet CM. Possible emergence of new geminiviruses by frequent recombination. Virology. 1999;265: 218–225.

52. Posada D, Crandall KA. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci U S A. 2001;98: 13757–13762.

53. Smith JM. Analyzing the mosaic structure of genes. J Mol Evol. 1992;34: 126–129.

54. Martin DP, Posada D, Crandall KA, Williamson C. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res Hum Retroviruses. 2005;21: 98–102.

55. Gibbs MJ, Armstrong JS, Gibbs AJ. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics. 2000;16: 573–582.

56. Boni MF, Lemey P, Jiang X, Lam TT-Y, Perry BW, Castoe TA, et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat Microbiol. 2020;5: 1408–1417.

57. Lytras S, Hughes J, Martin D, Swanepoel P, de Klerk A, Lourens R, et al. Exploring the Natural Origins of SARS-CoV-2 in the Light of Recombination. Genome Biol Evol. 2022;14. doi:10.1093/gbe/evac018

58. Lanza AM, Kim DS, Alper HS. Evaluating the influence of selection markers on obtaining selected pools and stable cell lines in human cells. Biotechnol J. 2013;8: 811–821.

59. Southern PJ, Berg P. Transformation of mammalian cells to antibiotic resistance with a bacterial gene under control of the SV40 early region promoter. J Mol Appl Genet. 1982;1: 327–341.

60. Higashimoto T, Urbinati F, Perumbeti A, Jiang G, Zarzuela A, Chang L-J, et al. The woodchuck hepatitis virus post-transcriptional regulatory element reduces readthrough transcription from retroviral vectors. Gene Ther. 2007;14: 1298–1304.

61. Picelli S, Björklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. 2014;24: 2033–2040.

62. Li N, Jin K, Bai Y, Fu H, Liu L, Liu B. Tn5 Transposase Applied in Genomics Research. Int J Mol Sci. 2020;21. doi:10.3390/ijms21218329

63. Yount B, Curtis KM, Fritz EA, Hensley LE, Jahrling PB, Prentice E, et al. Reverse genetics with a full-length infectious cDNA of severe acute respiratory syndrome coronavirus. Proc Natl Acad Sci U S A. 2003;100: 12995–13000.

64. Scobey T, Yount BL, Sims AC, Donaldson EF, Agnihothram SS, Menachery VD, et al. Reverse genetics with a full-length infectious cDNA of the Middle East respiratory syndrome coronavirus. Proc Natl Acad Sci U S A. 2013;110: 16157–16162.

65. Menachery VD, Yount BL Jr, Debbink K, Agnihothram S, Gralinski LE, Plante JA, et al. A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. Nat Med. 2015;21: 1508–1513.

66. Xie X, Muruato A, Lokugamage KG, Narayanan K, Zhang X, Zou J, et al. An Infectious cDNA Clone of SARS-CoV-2. Cell Host & Microbe. 2020. pp. 841–848.e3. doi:10.1016/j.chom.2020.04.004

67. Xie X, Lokugamage KG, Zhang X, Vu MN, Muruato AE, Menachery VD, et al. Engineering SARS-CoV-2 using a reverse genetic system. Nat Protoc. 2021;16: 1761–1784.

68. van den Worm SHE, Eriksson KK, Zevenhoven JC, Weber F, Züst R, Kuri T, et al. Reverse genetics of SARS-related coronavirus using vaccinia virus-based recombination. PLoS One.

2012;7: e32857.

69. Almazán F, Sola I, Zuñiga S, Marquez-Jurado S, Morales L, Becares M, et al. Reprint of: Coronavirus reverse genetic systems: infectious clones and replicons. Virus Res. 2014;194: 67–75.

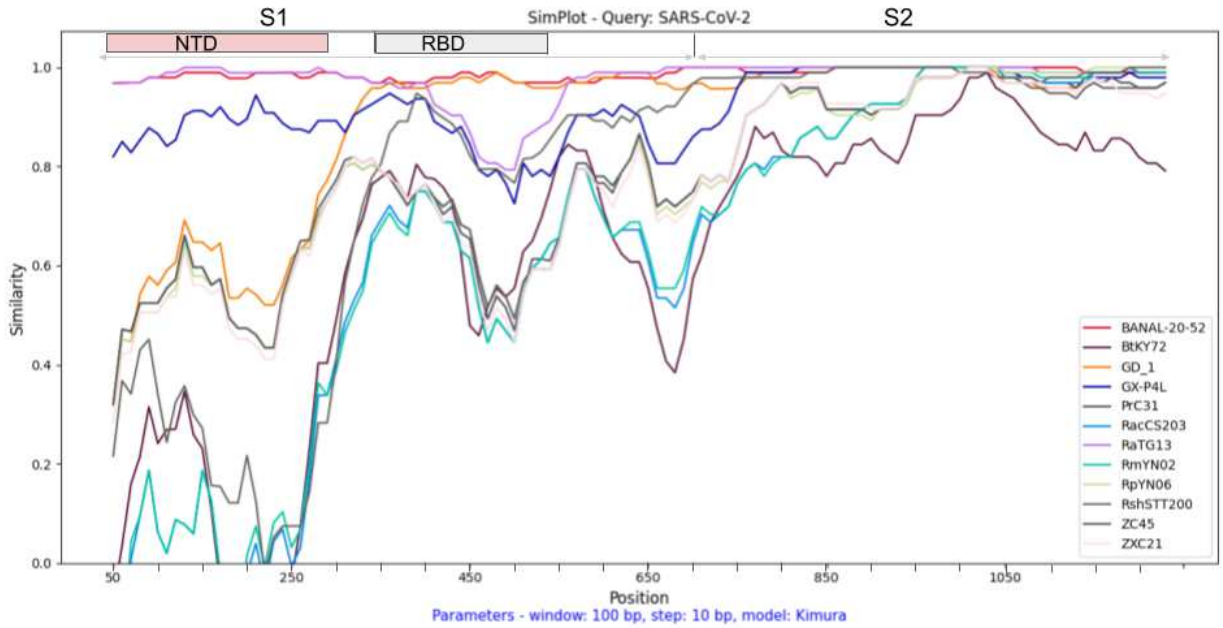
70. Zeng L-P, Gao Y-T, Ge X-Y, Zhang Q, Peng C, Yang X-L, et al. Bat Severe Acute Respiratory Syndrome-Like Coronavirus WIV1 Encodes an Extra Accessory Protein, ORFX, Involved in Modulation of the Host Immune Response. J Virol. 2016;90: 6573–6582.

71. Cockrell AS, Beall A, Yount B, Baric R. Efficient Reverse Genetic Systems for Rapid Genetic Manipulation of Emergent and Preemergent Infectious Coronaviruses. Methods Mol Biol. 2017;1602: 59–81.

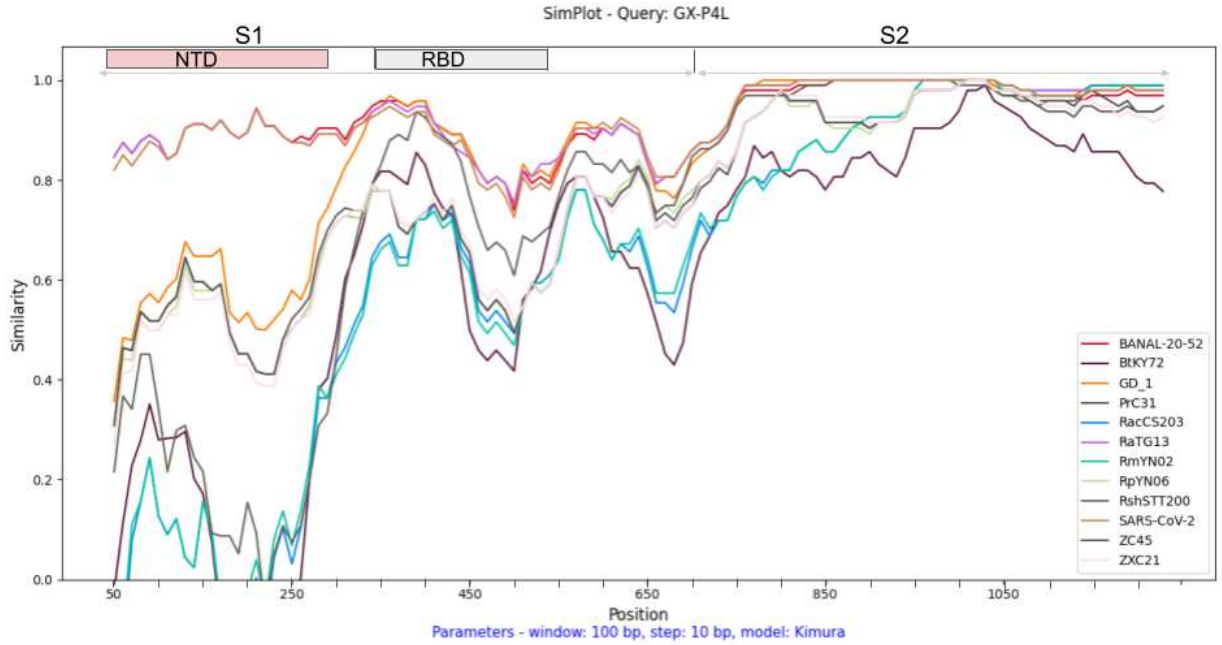
72. Zhang D, Jones A, Deigin Y, Sirotkin K, Sousa A. Unexpected novel Merbecovirus discoveries in agricultural sequencing datasets from Wuhan, China. arXiv:2104.01533v1. 2021.

73. Xu L, Guan J, Lau W, Xiao Y. An Overview of Pangolin Trade in China. 2016. Available: https://www.traffic.org/publications/reports/pangolin-trade-in-china/#:~:text=key%20findings,of%20pangolin%20trade%20in%20China.

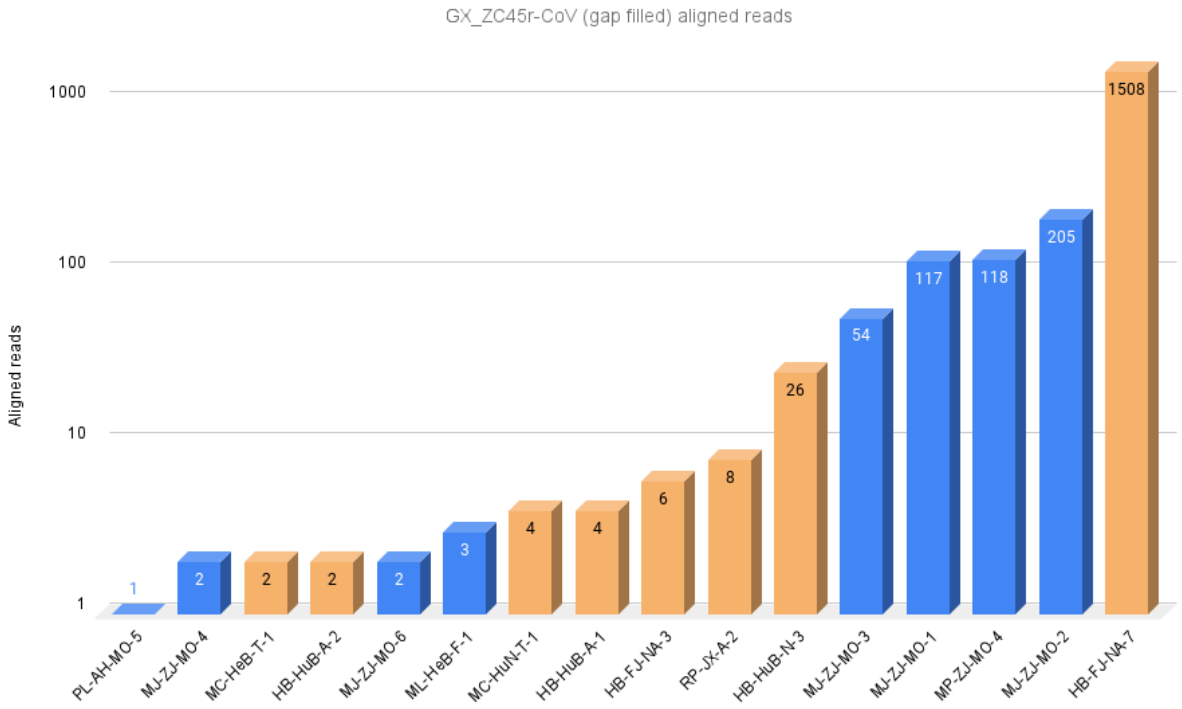
Supplementary Figures and Tables



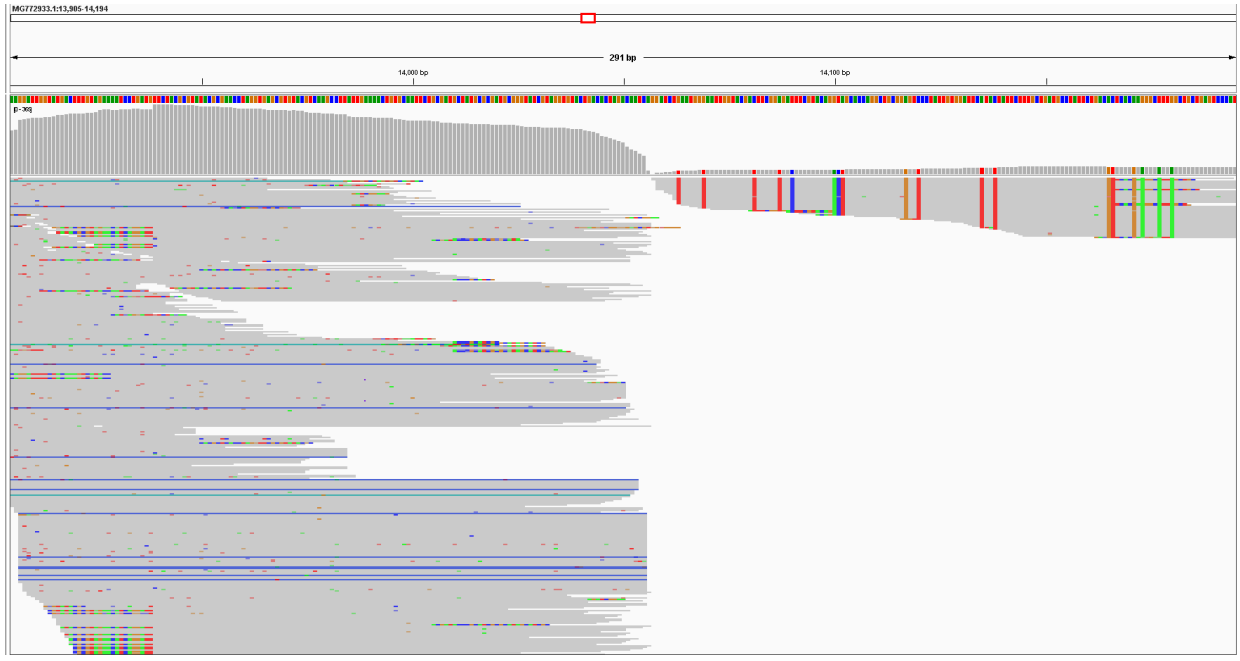
Supp. Fig. 1. SARS-CoV-2 spike protein amino acid similarity to SARS2r-CoVs. Aligned using the MUSCLE algorithm in UGENE, plotted using SimPlot++ using a Kimura 2 parameter model with a 100bp window and 10bp step size. Both PCoV GD_1 and BANAL-52 have high amino acid similarity to the SARS-CoV-2 RBD. Reference numbering relative to multiple-sequence alignment.



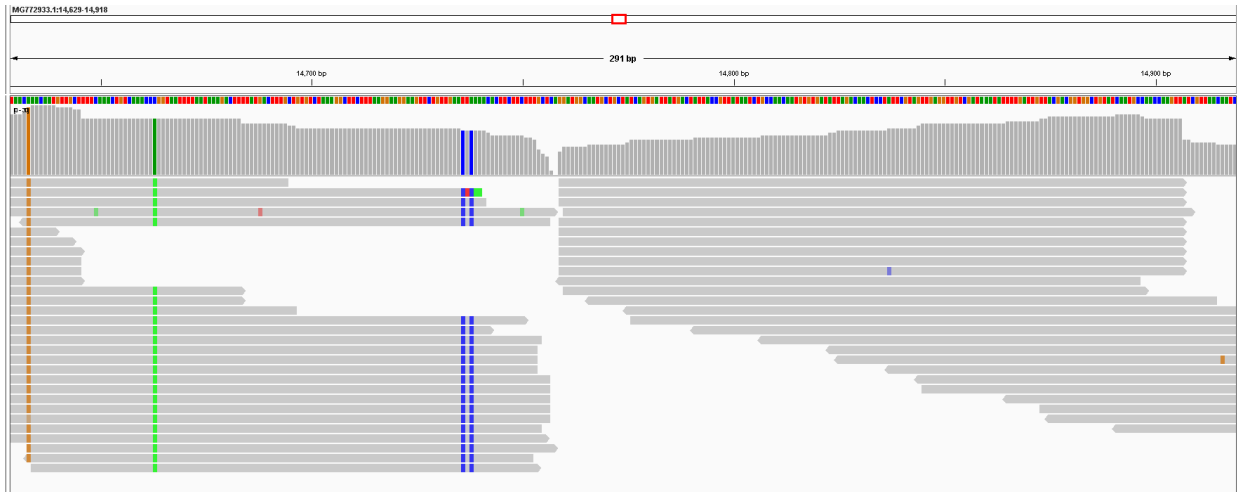
Supp. Fig. 2. PCoV GX-P4L spike amino acid similarity plot to SARS2r-CoVs. Aligned using MUSCLE algorithm in UGENE, plotted using SimPlot++ using a Kimura 2 parameter model with a 100bp window and 10bp step size. PCoV GX-P4L has highest overall amino acid similarity to RaTG13 and SARS-CoV-2. Reference numbering relative to multiple-sequence alignment.



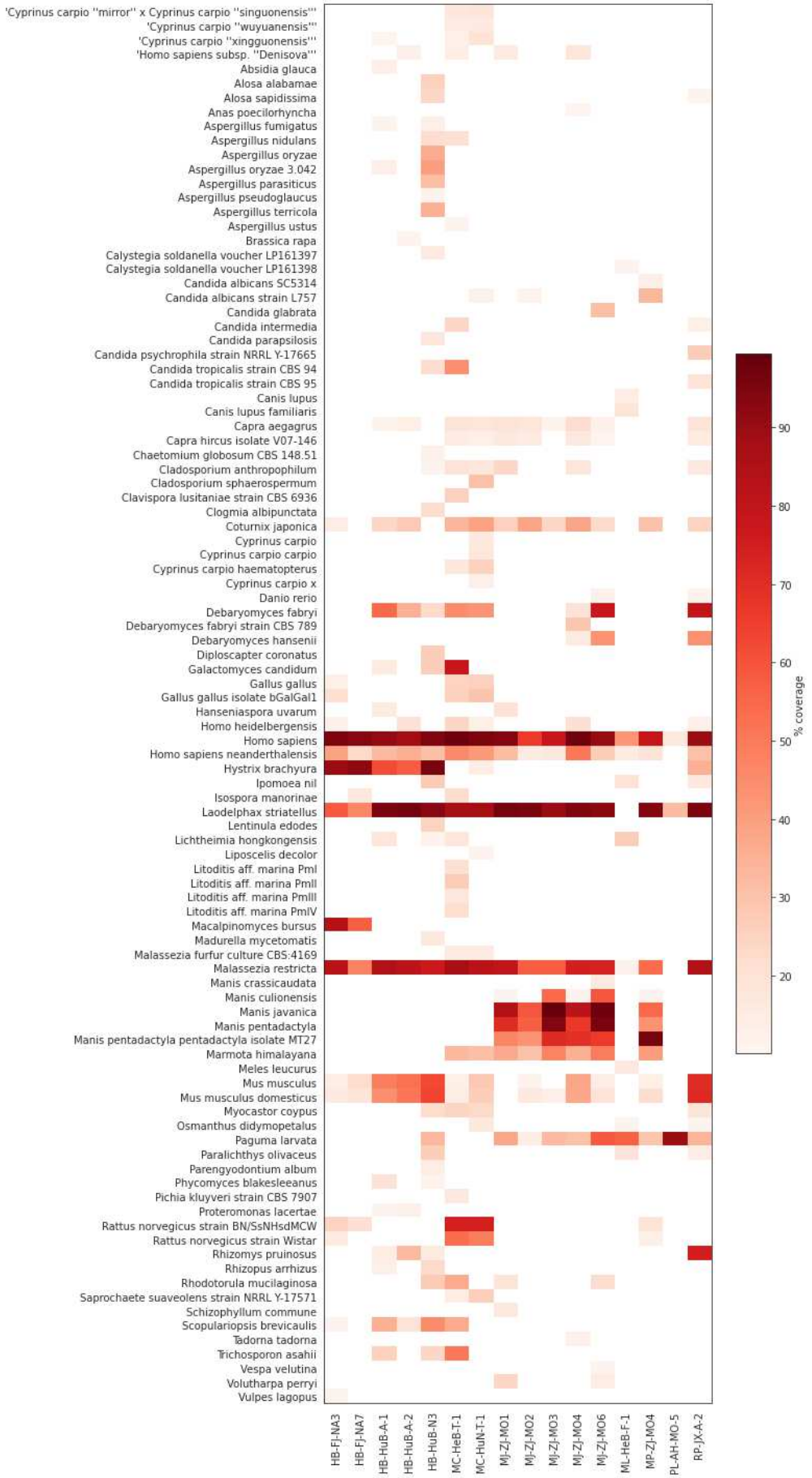
Supp. Fig. 3. Reads per sample aligning to GX_ZC45r-CoV gap filled with bat-SLCoV-ZC45. BioProject PRJNA793740 samples coloured blue, BioProject PRJNA795267 samples coloured orange. Read count in log scale. Datasets containing GX_ZC45r-CoV newly identified here: PL-AH-MO-5, MC-HeB-T-1, HB-HuB-A-2, ML-HeB-F-1, MC-HuN-T-1, HB-HuB-A-1 and RP-JX-A-2.



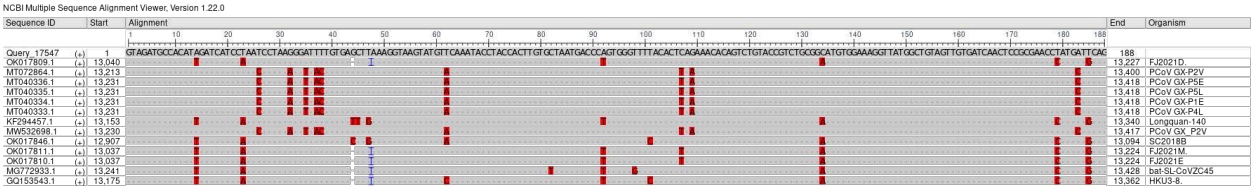
Supp. Fig. 4. Alignment of GX_ZC45r-CoV to bat-SL-CoVZC45 (MG772933.1) showing anomalous read coverage across position 14056nt potentially indicating artificial splicing.



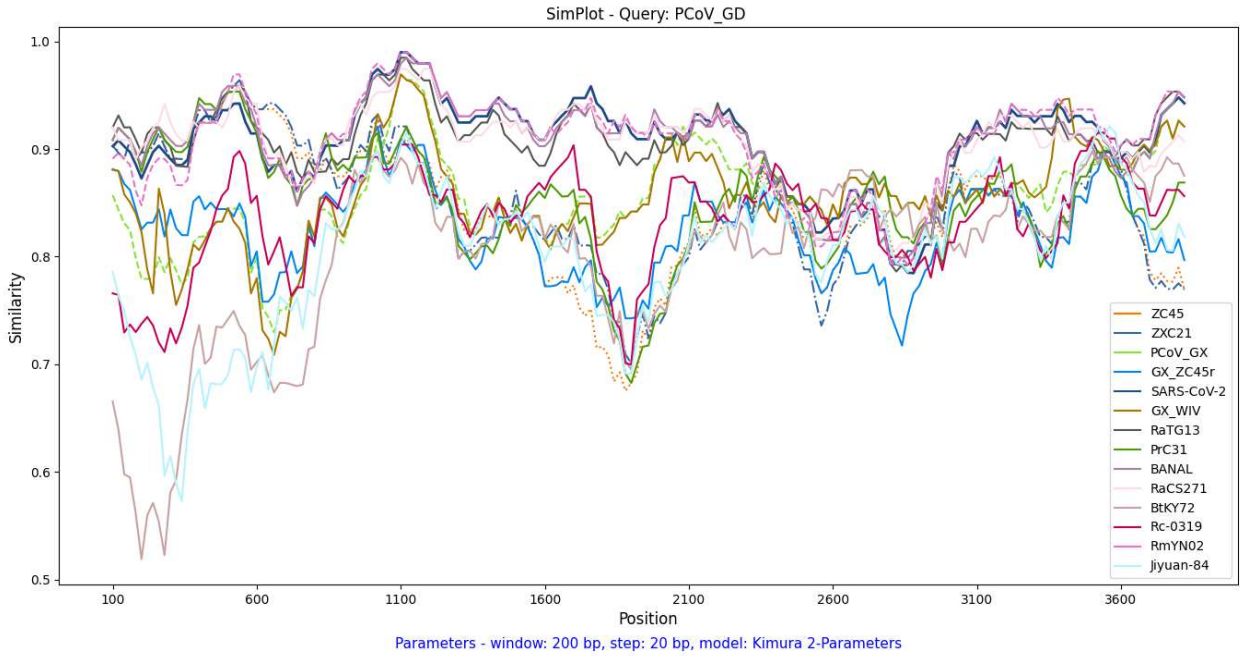
Supp. Fig. 5. Alignment of GX_ZC45r-CoV to bat-SL-CoVZC45 (MG772933.1) showing anomalous read coverage across position 14758nt potentially indicating artificial splicing.


'Cyprinus carpio ''mirror'' x Cyprinus carpio ''singuonensis'''
'Cyprinus carpio ''wuyuanensis'''
'Cyprinus carpio ''xingguonensis'''
'Homo sapiens subsp. ''Denisova'''
Absidia glauca
Alosa alabamae
Alosa sapidissima
Anas poecilorhyncha
Aspergillus fumigatus
Aspergillus nidulans
Aspergillus oryzae
Aspergillus oryzae 3.042
Aspergillus parasiticus
Aspergillus pseudoglaucus
Aspergillus terricola
Aspergillus ustus
Brassica rapa
Calystegia soldanella voucher LP161397
Calystegia soldanella voucher LP161398
Candida albicans SC5314
Candida albicans strain L757
Candida glabrata
Candida intermedia
Candida parapsilosis
Candida psychrophila strain NRRL Y-17665
Candida tropicalis strain CBS 94
Candida tropicalis strain CBS 95
Canis lupus
Canis lupus familiaris
Capra aegagrus
Capra hircus isolate V07-146
Chaetomium globosum CBS 148.51
Cladosporium anthropophilum
Cladosporium sphaerospermum
Clavispora lusitaniae strain CBS 6936
Clogmia albipunctata
Coturnix japonica
Cyprinus carpio
Cyprinus carpio carpio
Cyprinus carpio haematopterus
Cyprinus carpio x
Danio rerio
Debaryomyces fabryi
Debaryomyces fabryi strain CBS 789
Debaryomyces hansenii
Diploscapter coronatus
Galactomyces candidum
Gallus gallus
Gallus gallus isolate bGalGal1
Hanseniaspora uvarum
Homo heidelbergensis
Homo sapiens
Homo sapiens neanderthalensis
Hystrix brachyura
Ipomoea nil
Isospora manorinae
Laodelphax striatellus
Lentinula edodes
Lichtheimia hongkongensis
Liposcelis decolor
Litoditis aff. marina Pml
Litoditis aff. marina PmII
Litoditis aff. marina PmIII
Litoditis aff. marina PmIV
Macalpinomyces bursus
Madurella mycetomatis
Malassezia furfur culture CBS:4169
Malassezia restricta
Manis crassicaudata
Manis culionensis
Manis javanica
Manis pentadactyla
Manis pentadactyla pentadactyla isolate MT27
Marmota himalayana
Meles leucurus
Mus musculus
Mus musculus domesticus
Myocastor coypus
Osmanthus didymopetalus
Paguma larvata
Paralichthys olivaceus
Parengyodontium album
Phycomyces blakesleeanus
Pichia kluyveri strain CBS 7907
Proteromonas lacertae
Rattus norvegicus strain BN/SsNHsdMCW
Rattus norvegicus strain Wistar
Rhizomys pruinosus
Rhizopus arrhizus
Rhodotorula mucilaginosa
Saprochaete suaveolens strain NRRL Y-17571
Schizophyllum commune
Scopulariopsis brevicaulis
Tadorna tadorna
Trichosporon asahii
Vespa velutina
Volutharpa perryi
Vulpes lagopus
HB-FJ-NA3
HB-FJ-NA7
HB-HuB-A-1
HB-HuB-A-2
HB-HuB-N3
MC-HeB-T-1
MC-HuN-T-1
MJ-ZJ-MO1
MJ-ZJ-MO2
MJ-ZJ-MO3
MJ-ZJ-MO4
MJ-ZJ-MO6
ML-HeB-F-1
MP-ZJ-MO4
PL-AH-MO-5
RP-JX-A-2
% coverage
90
80
70
60
50
40
30
20

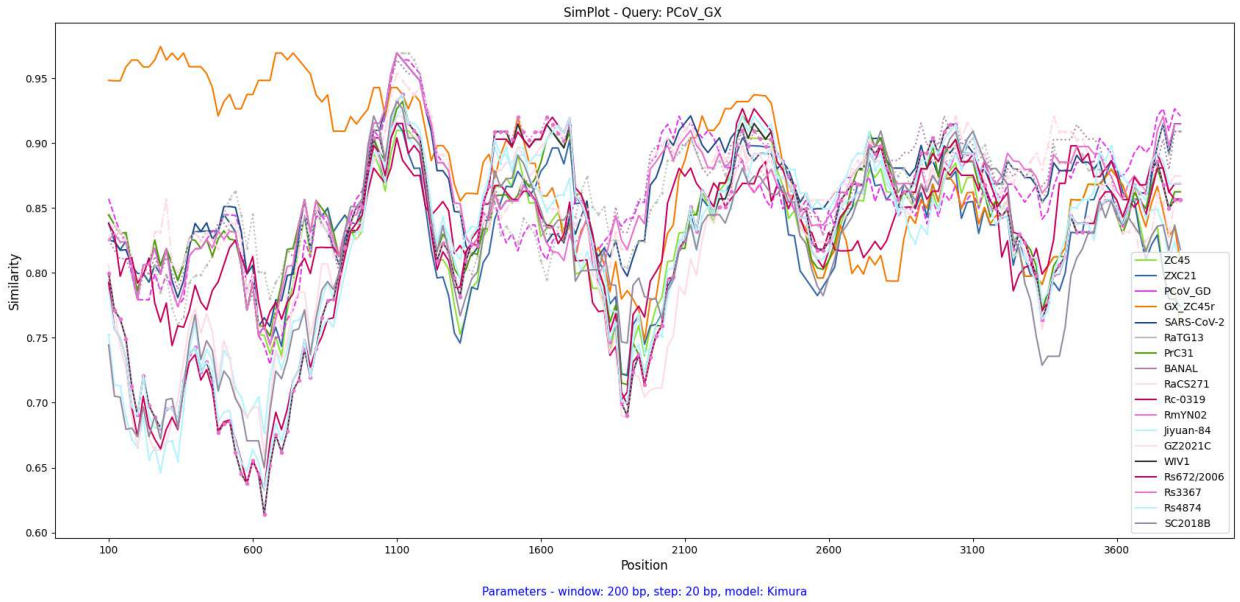
Supp. Fig. 6. Mitochondrial genome coverage >=10% for the 16 game animal datasets containing SARS2r-CoV reads. Alignment generated using bowtie2, plotted using matplotlib.



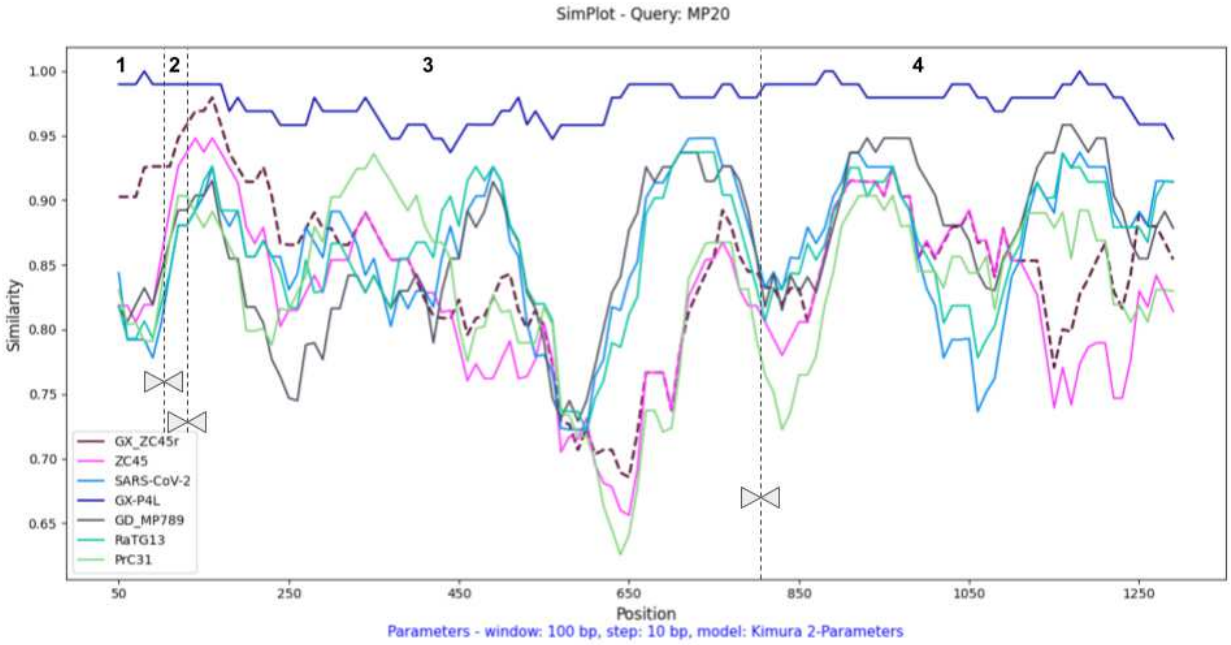
Supp. Fig. 7. Blastn analysis of the recovered NSP10 region of GX_ZC45r-CoV showing the 13 highest identity genome matches.



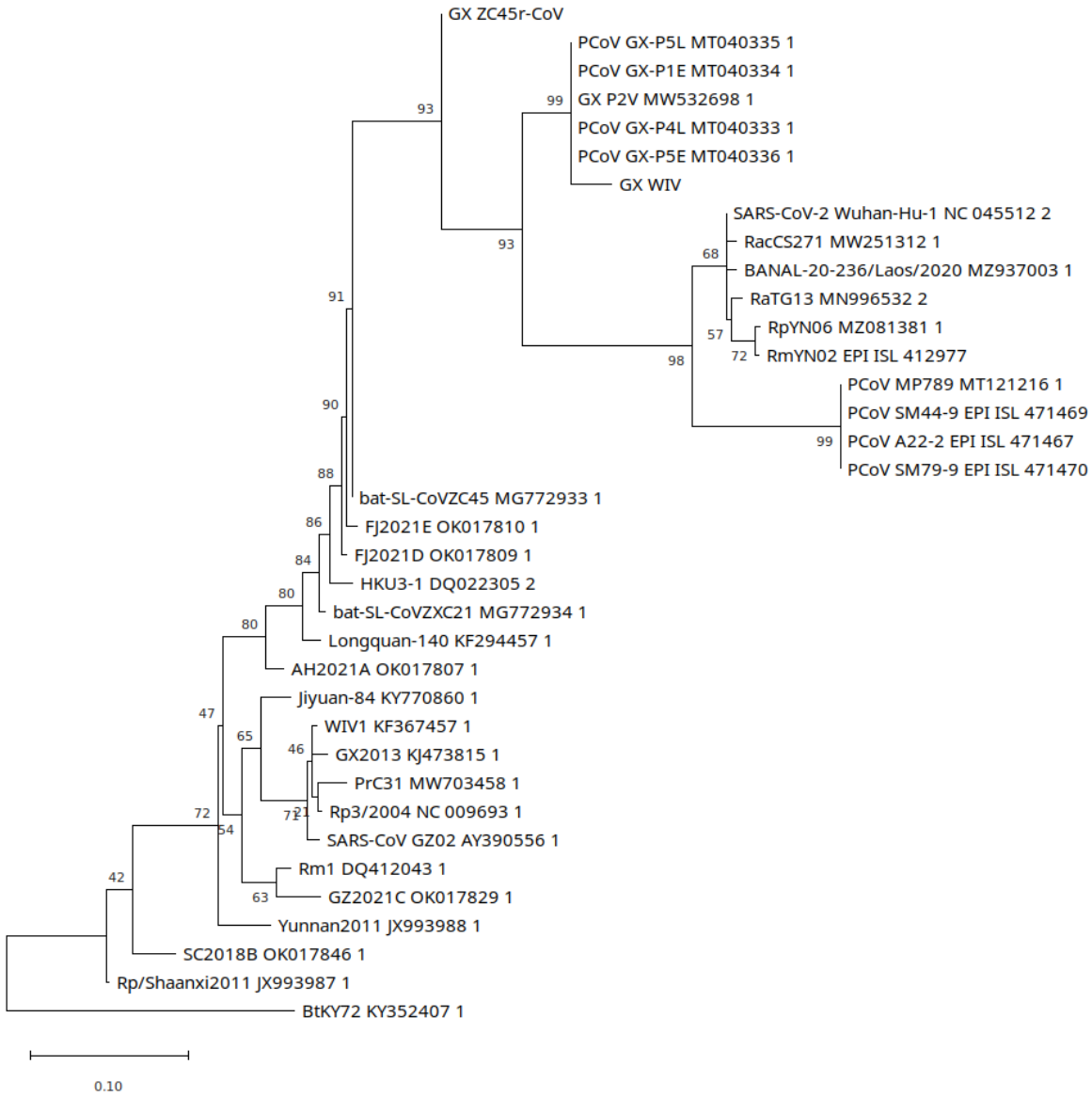
Supp. Fig. 8. Simplot analysis of selected SARSr-CoV genomes using PCoV MP789 as a query. Note the same 3166nt gap as per Fig. 15 is situated at 942-943nt. Plotted using Simplot++. Solid lines except: RmYN02 - dashed, PCoV GX - dashed, bat-SL-CoVZC45 - dotted, bat-SL-CoVZXC21 - dash - dotted, SARS-CoV-2 line weight 2 (all others line weight 1.5).



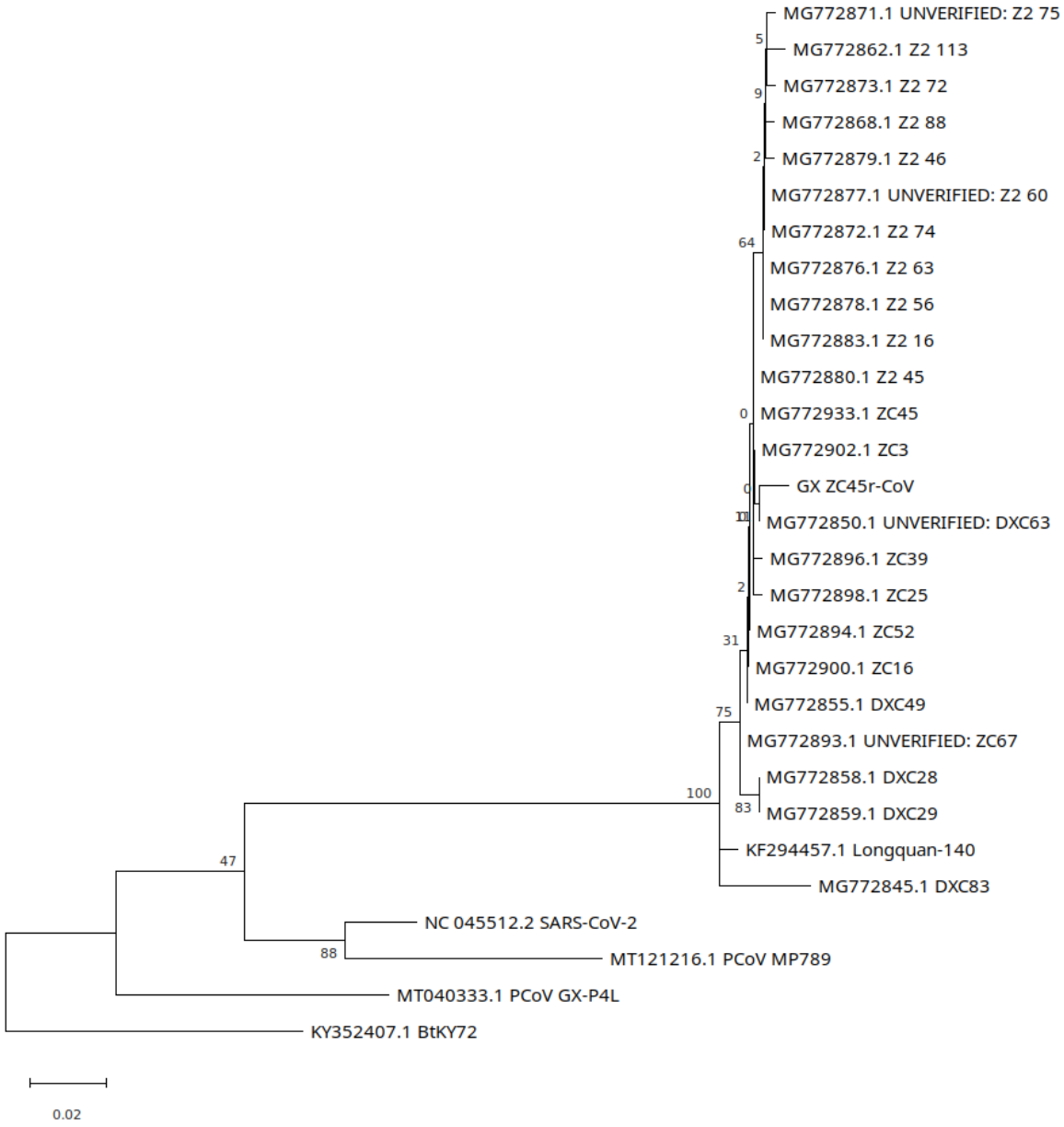
Supp. Fig. 9. Simplot analysis of selected SARSr-CoV genomes using PCoV GX group as a query. Note the same 3166nt gap as per Fig. 15 is situated at 942-943nt. Solid lines except: BANAL group - dotted, PCoV GD group - dashed, Rs3367 - dotted with circle markers, RaCS271 - dash dot with hexagon markers, Jiyuan-84 - solid with square markers, RaTG13 dotted with plus markers. Plotted using Simplot++. See methods for CoV grouping.



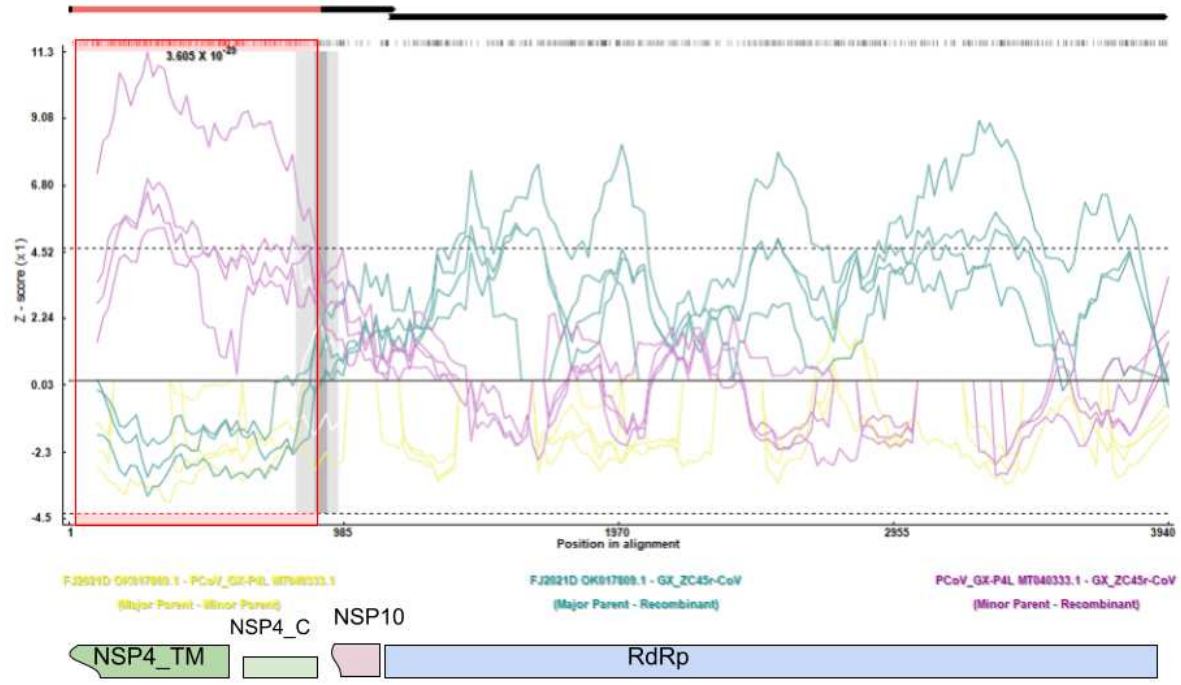
Supp. Fig. 10. Simplot analysis of four sections of PCoV MP20 which overlap with sections of GX_ZC45r-CoV after multi sequence alignment. PCoV MP20 queried against selected SARSr-CoV genomes. The four regions are: 1) 105nt at 3' end of NSP4, 2) 29nt at 5' end of the NSP10 gene covered by GX_ZC45r-CoV, 3) 675nt in RdRp and 4) 532nt at 3' end of RdRp gene coding region. GX_ZC45r-CoV plotted as dashed lines. Plotted using SimPlot++.



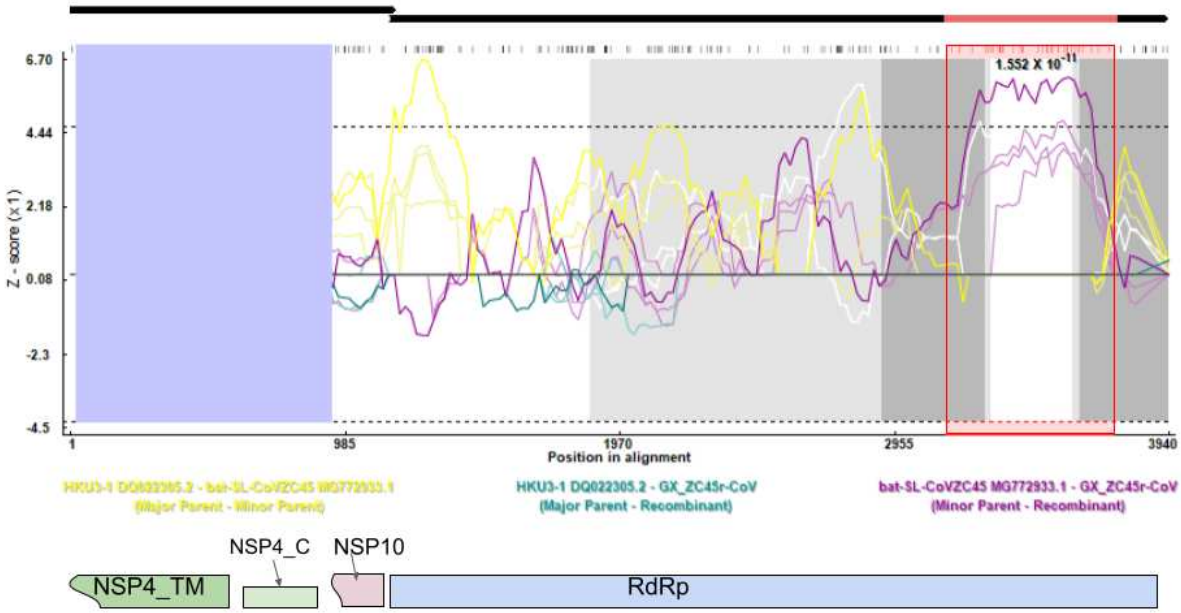
Supp. Fig. 11. Partial RdRp section (297nt) maximum likelihood phylogenetic tree using a GTR+I+G (GTR+FO+I+G4m) model with 1000 bootstrap replicates using raxmlGUI 2.0.



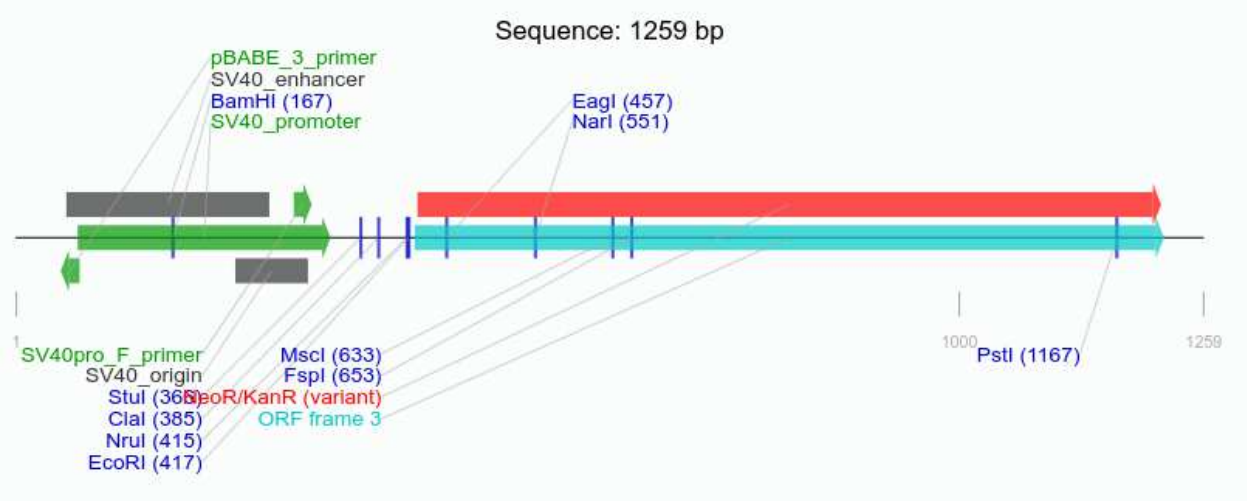
Supp. Fig. 12. Partial RdRp section (407nt) maximum likelihood phylogenetic tree using a T92+I model with 100 bootstrap replicates using MEGA11.



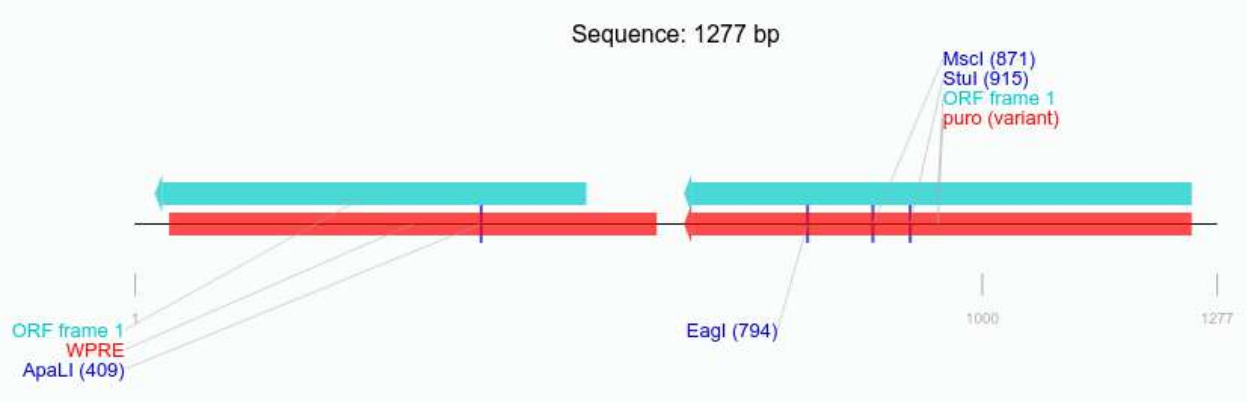
Supp. Fig.13. Potential recombination region 1 detected using SiScan model in RDP5. Note 5' end of recombination region not detected.



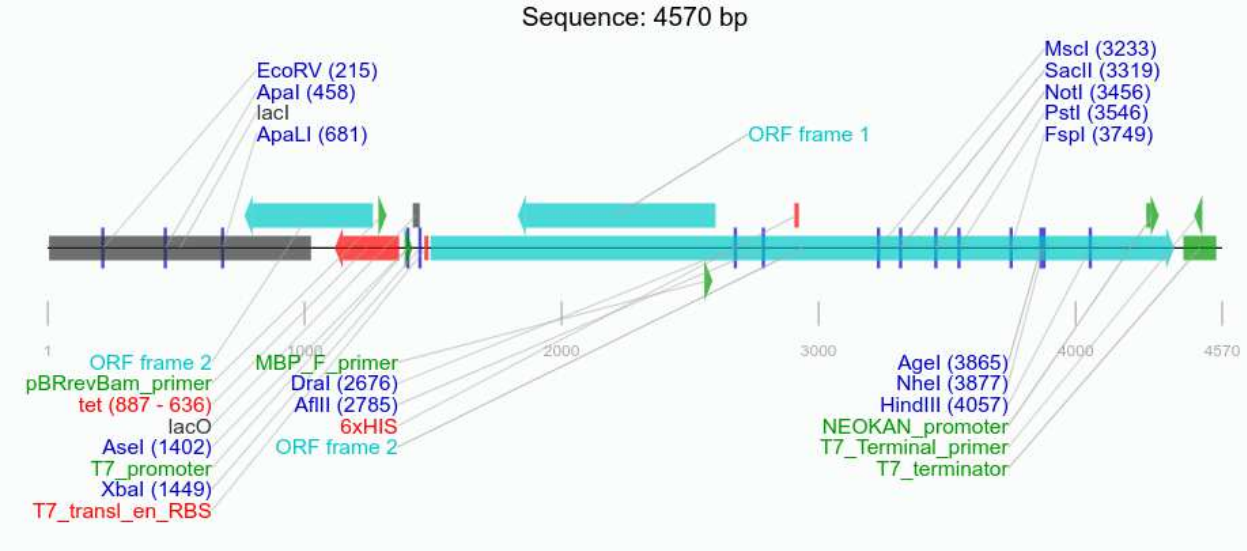
Supp. Fig. 14. Potential recombination region 2 detected using SiScan model in RDP5.



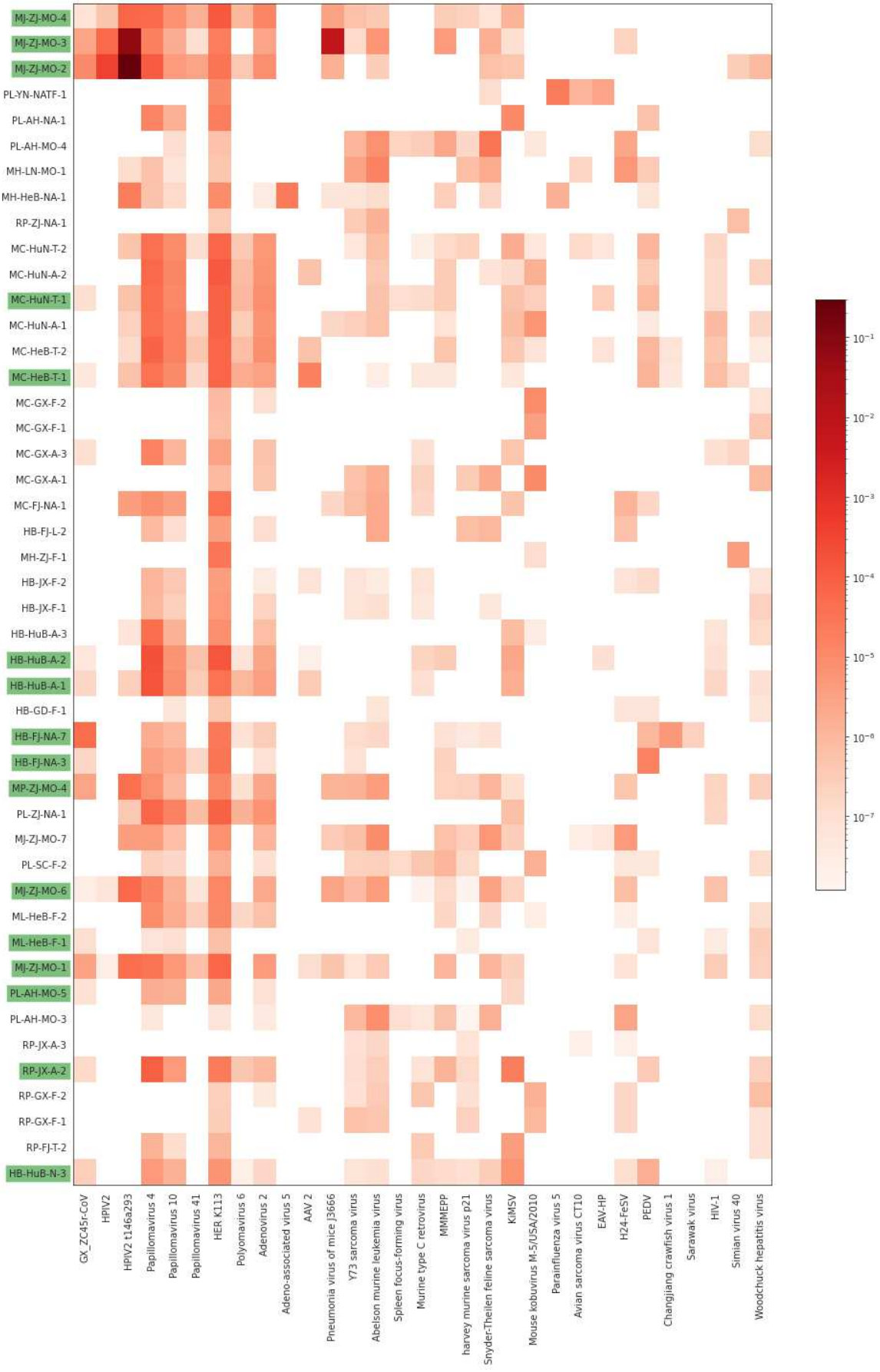
Supp. Fig. 15. Partial vector sequence identified in MJ-ZJ-F-1 with similar layout to mammalian expression vector pSV2neo.



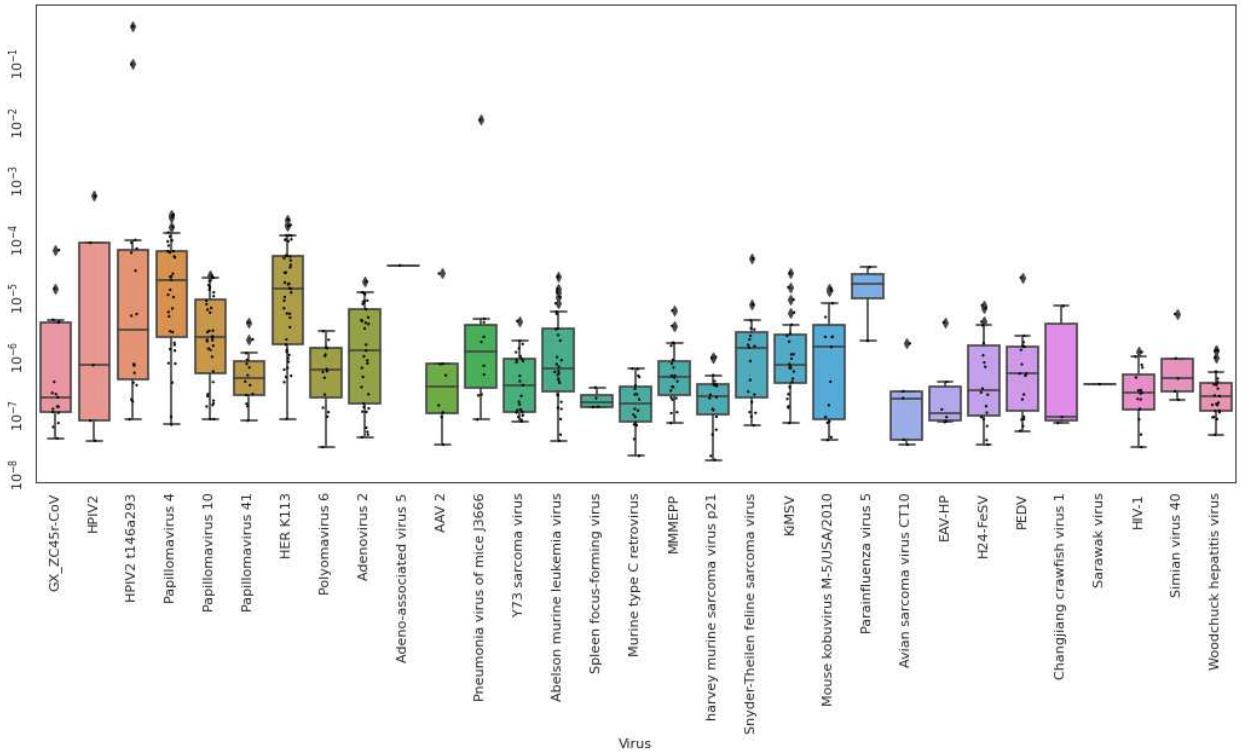
Supp. Fig. 16. Partial synthetic vector de novo assembled from the MC-GX-A-1 dataset. Highest blastn max score match is found to Lentiviral expression vector MCPyV (MZ648044.1) (98.08% identity).



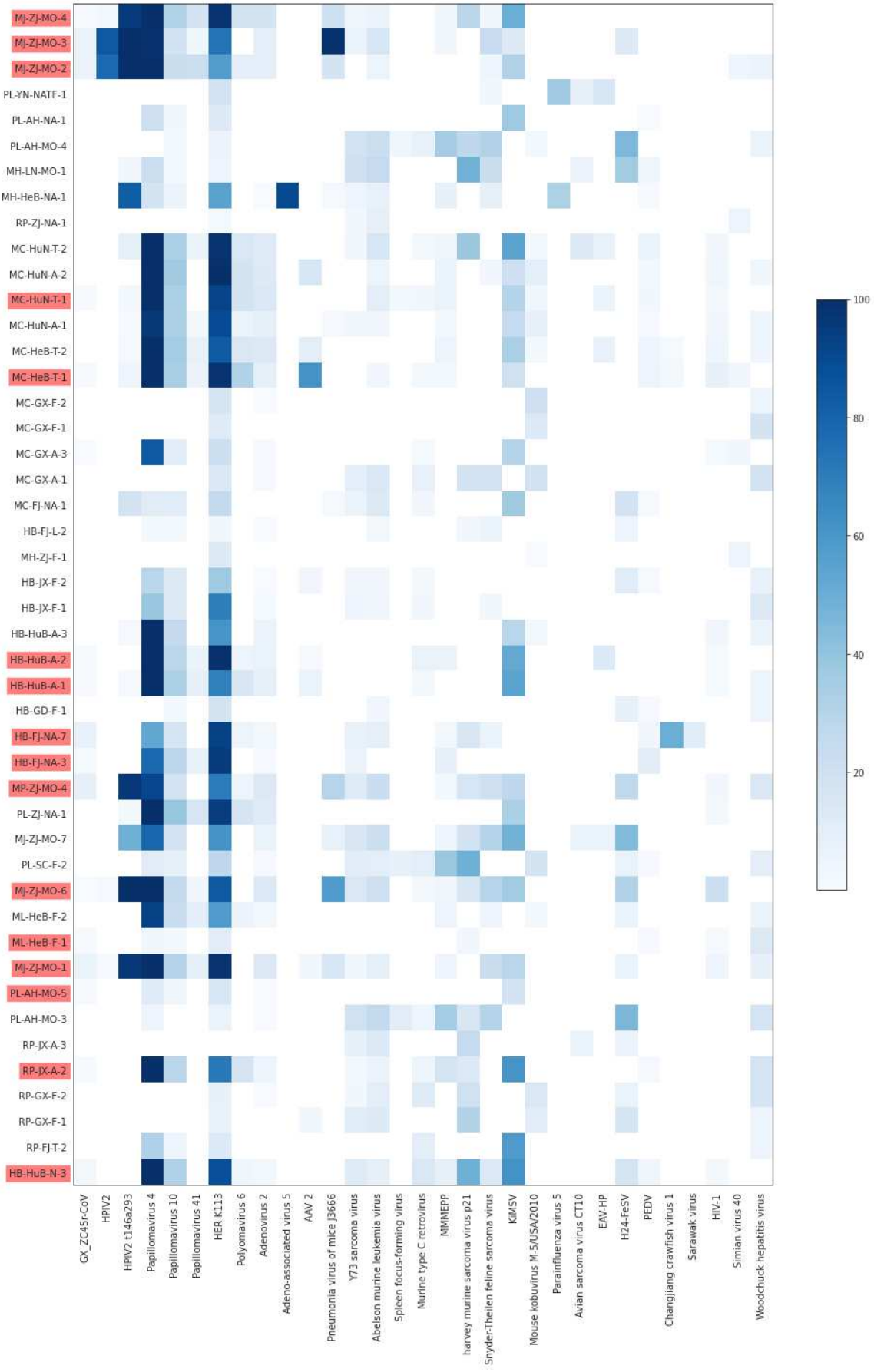
Supp. Fig. 17. Synthetic plasmid sequence de novo assembled from the MJ-ZJ-F-3 dataset. Highest blastn identity is found to Tn5 transposase cloning vectors pKRCPN1 (ALS31075.1), TnpS [Cloning vector pTHSSe_56] (AVR55160.1) and Tn5 hyperactive transposase [Cloning vector pBAM1] (ADY68344.1).


MJ-ZJ-MO-4
MJ-ZJ-MO-3
MJ-ZJ-MO-2
PL-YN-NATF-1
PL-AH-NA-1
PL-AH-MO-4
MH-LN-MO-1
MH-HeB-NA-1
RP-ZJ-NA-1
MC-HuN-T-2
MC-HuN-A-2
MC-HuN-T-1
MC-HuN-A-1
MC-HeB-T-2
MC-HeB-T-1
MC-GX-F-2
MC-GX-F-1
MC-GX-A-3
MC-GX-A-1
MC-FJ-NA-1
HB-FJ-L-2
MH-ZJ-F-1
HB-JX-F-2
HB-JX-F-1
HB-HuB-A-3
HB-HuB-A-2
HB-HuB-A-1
HB-GD-F-1
HB-FJ-NA-7
HB-FJ-NA-3
MP-ZJ-MO-4
PL-ZJ-NA-1
MJ-ZJ-MO-7
PL-SC-F-2
MJ-ZJ-MO-6
ML-HeB-F-2
ML-HeB-F-1
MJ-ZJ-MO-1
PL-AH-MO-5
PL-AH-MO-3
RP-JX-A-3
RP-JX-A-2
RP-GX-F-2
RP-GX-F-1
RP-FJ-T-2
HB-HuB-N-3
GX_ZC45r-CoV
HPIV2
HPIV2 t146a293
Papillomavirus 4
Papillomavirus 10
Papillomavirus 41
HER K113
Polyomavirus 6
Adenovirus 2
Adeno-associated virus 5
AAV 2
Pneumonia virus of mice J3666
Y73 sarcoma virus
Abelson murine leukemia virus
Spleen focus-forming virus
Murine type C retrovirus
MMMEPP
harvey murine sarcoma virus p21
Snyder-Theilen feline sarcoma virus
KiMSV
Mouse kobuvirus M-5/USA/2010
Parainfluenza virus 5
Avian sarcoma virus CT10
EAV-HP
H24-FeSV
PEDV
Changjiang crawfish virus 1
Sarawak virus
HiV-1
Simian virus 40
Woodchuck hepatitis virus
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}

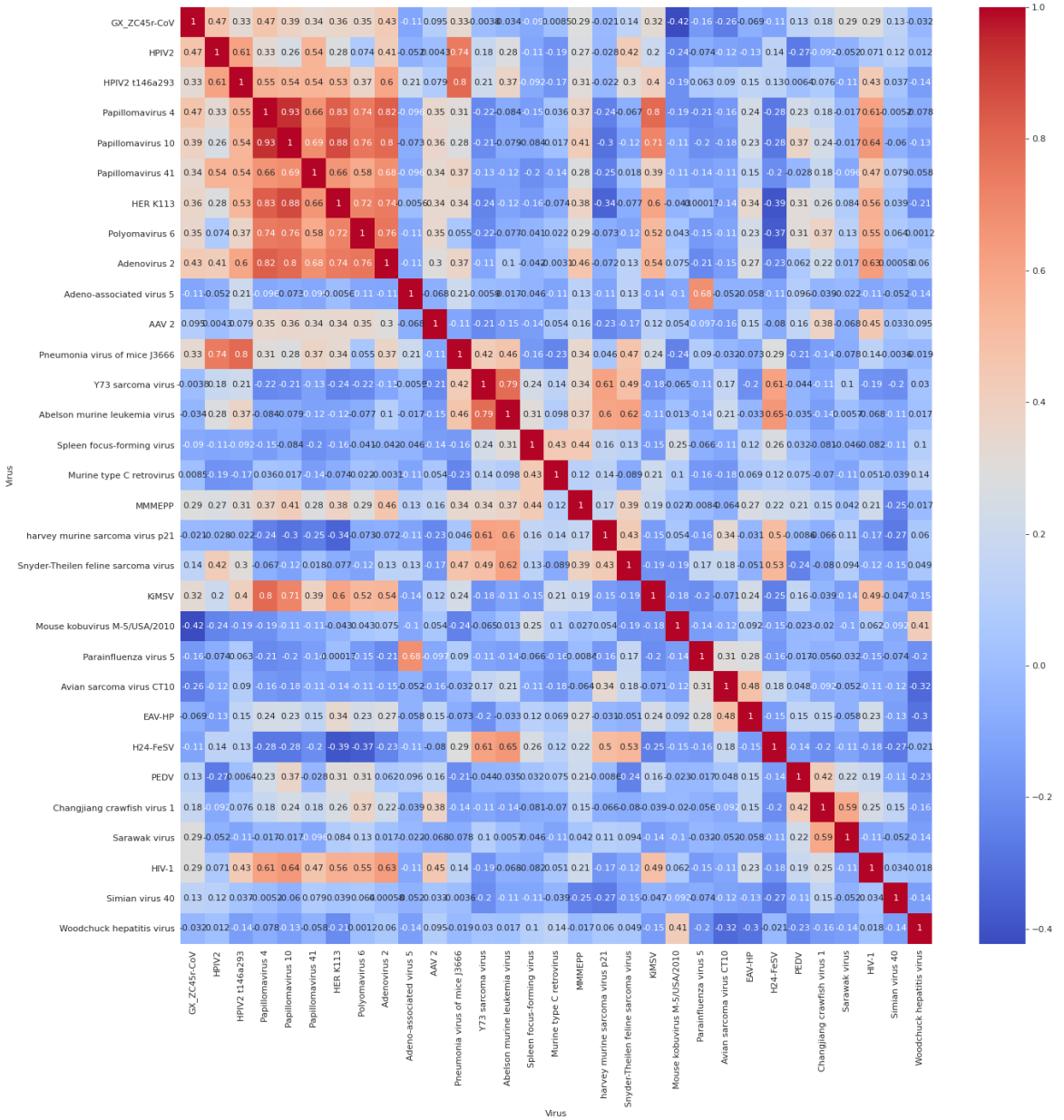
Supp. Fig. 18. Counts per read for reads mapping to selected viruses per SRA dataset for 46 SRA datasets from PRJNA793740 and PRJNA795267. For all but GZ_ZC45r-CoV and Simian Virus 40 (SV40) a 10% cutoff was applied whereby each virus is present in at least one of the 46 datasets with 10% or greater genome coverage. Normalized counts coloured in log scale. Names of samples containing GX_ZC45r-CoV sequences are highlighted in green. Counts per read calculated by dividing counts by total read count for each SRA. See Spp. Info. 1 for virus name abbreviations.



Supp. Fig. 19. Box plot distribution of counts per read mapping to selected viruses per SRA dataset for 46 SRA datasets from PRJNA793740 and PRJNA795267. For all but GZ_ZC45r-CoV and Simian Virus 40 (SV40) a 10% cutoff was applied whereby each virus is present in at least one of the 46 datasets with 10% or greater genome coverage. Read counts shown in log scale. See Spp. Info. 1 for virus name abbreviations.

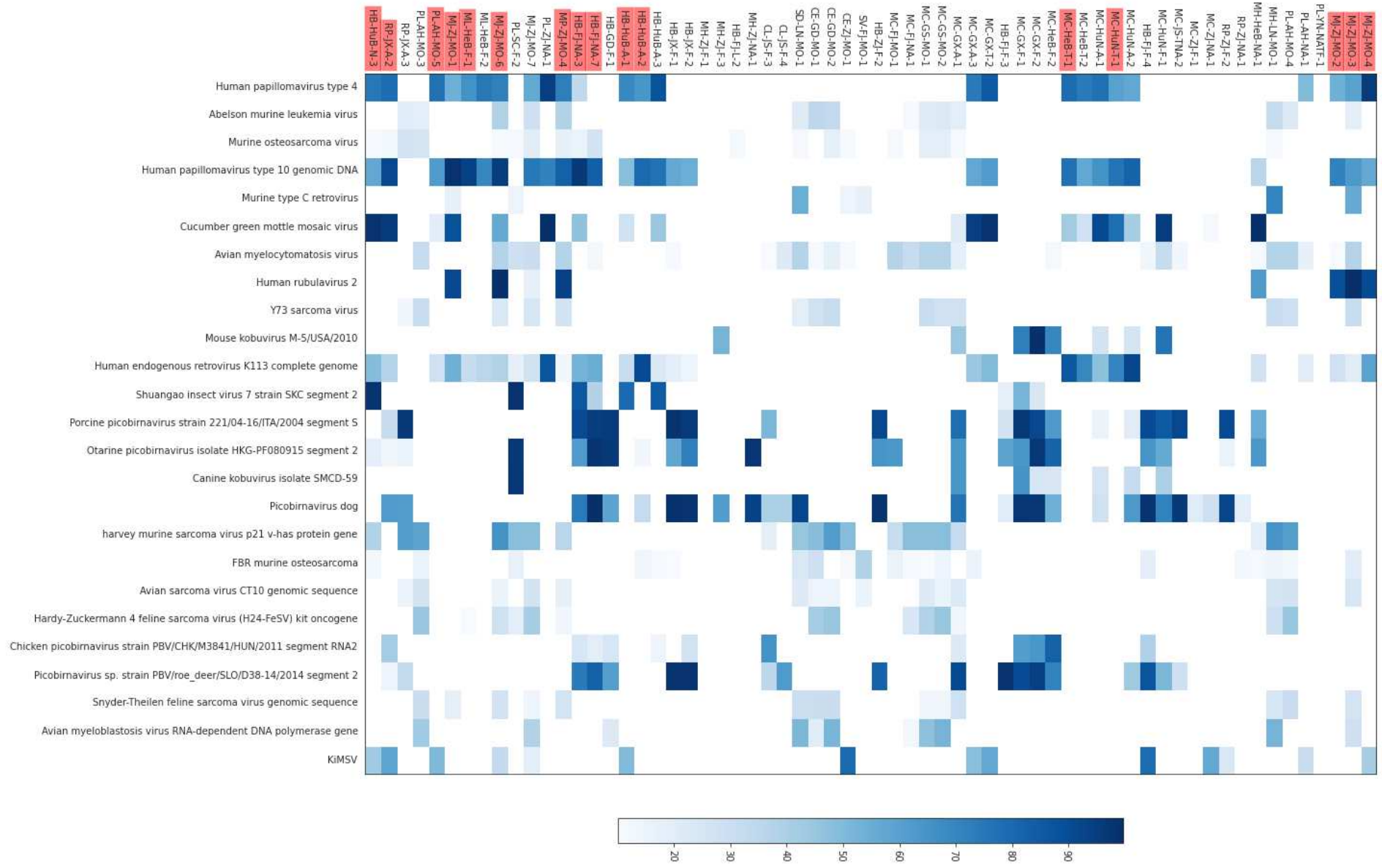

MJ-ZJ-MO-4
MJ-ZJ-MO-3
MJ-ZJ-MO-2
PL-YN-NATF-1
PL-AH-NA-1
PL-AH-MO-4
MH-LN-MO-1
MH-HeB-NA-1
RP-ZJ-NA-1
MC-HuN-T-2
MC-HuN-A-2
MC-HuN-T-1
MC-HuN-A-1
MC-HeB-T-2
MC-HeB-T-1
MC-GX-F-2
MC-GX-F-1
MC-GX-A-3
MC-GX-A-1
MC-FJ-NA-1
HB-FJ-L-2
MH-ZJ-F-1
HB-JX-F-2
HB-JX-F-1
HB-HuB-A-3
HB-HuB-A-2
HB-HuB-A-1
HB-GD-F-1
HB-FJ-NA-7
HB-FJ-NA-3
MP-ZJ-MO-4
PL-ZJ-NA-1
MJ-ZJ-MO-7
PL-SC-F-2
MJ-ZJ-MO-6
ML-HeB-F-2
ML-HeB-F-1
MJ-ZJ-MO-1
PL-AH-MO-5
PL-AH-MO-3
RP-JX-A-3
RP-JX-A-2
RP-GX-F-2
RP-GX-F-1
RP-FJ-T-2
HB-HuB-N-3
GX_ZC45r-CoV
HPIV2
HPIV2 t146a293
Papillomavirus 4
Papillomavirus 10
Papillomavirus 41
HER K113
Polyomavirus 6
Adenovirus 2
Adeno-associated virus 5
AAV 2
Pneumonia virus of mice J3666
Y73 sarcoma virus
Abelson murine leukemia virus
Spleen focus-forming virus
Murine type C retrovirus
MMMEPP
harvey murine sarcoma virus p21
Snyder-Theilen feline sarcoma virus
KiMSV
Mouse kobuvirus M-5/USA/2010
Parainfluenza virus 5
Avian sarcoma virus CT10
EAV-HP
H24-FeSV
PEDV
Changjiang crawfish virus 1
Sarawak virus
HIV-1
Simian virus 40
Woodchuck hepatitis virus
100
80
60
40
20

Supp. Fig. 20. Percent coverage of selected viruses by reads for 46 SRA datasets from PRJNA793740 and PRJNA795267. For all but GZ_ZC45r-CoV and Simian Virus 40 (SV40) a 10% cutoff was applied whereby each virus is present in at least one of the 46 datasets with 10% or greater genome coverage. Names of samples containing GX_ZC45r-CoV sequences are highlighted in red. See Spp. Info. 1 for virus name abbreviations.

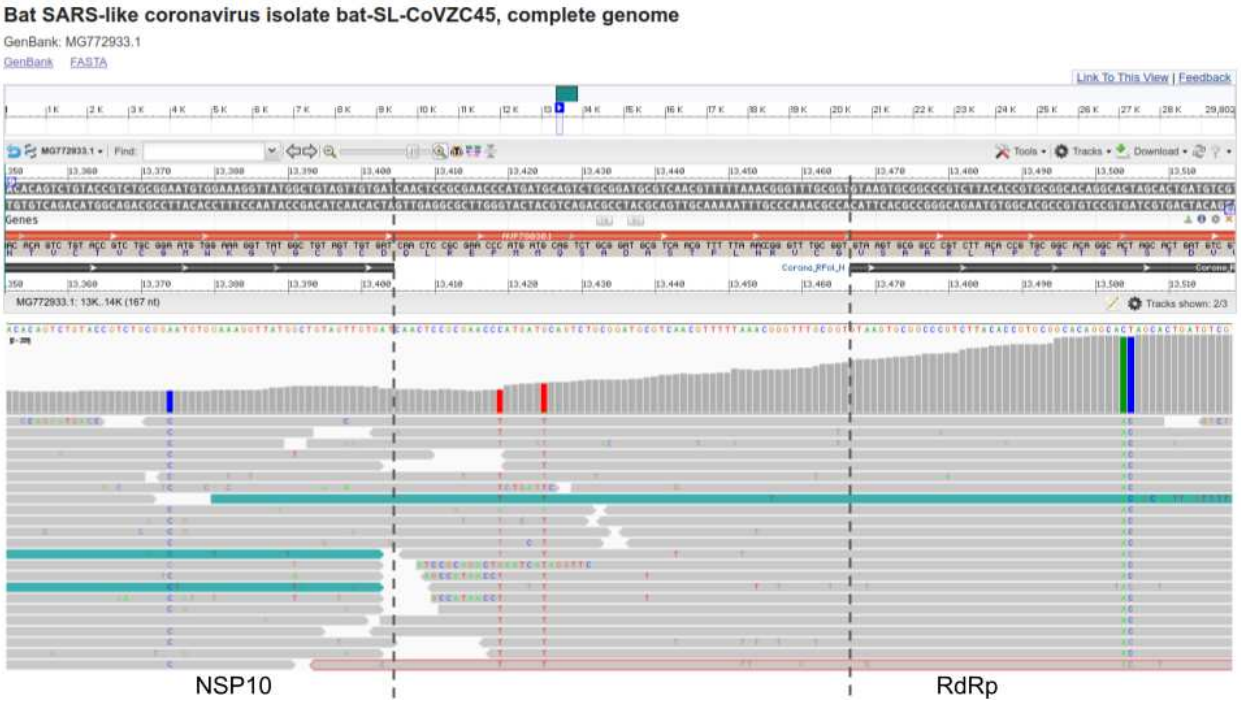


Supp. Fig. 21. Spearman correlation coefficient matrix for virus read counts for 46 SRA datasets from PRJNA793740 and PRJNA795267 (see Supp. Fig. 18 for sample names). SRAs were aligned to viruses using bowtie2 using the '--very-sensitive' parameter. For all but GZ_ZC45r-CoV and Simian Virus 40 (SV40) a 10% cutoff was applied whereby each virus is present in at least one of the 46 datasets with 10%

or greater genome coverage. Virus read counts were divided by total read counts per SRA prior to analysis.


HB-HuB-N-3
RP-JX-A-2
RP-JX-A-3
PL-AH-MO-3
PL-AH-MO-5
MJ-ZJ-MO-1
ML-HeB-F-1
ML-HeB-F-2
MJ-ZJ-MO-6
PL-SC-F-2
MJ-ZJ-MO-7
PL-ZJ-NA-1
MP-ZJ-MO-4
HB-FJ-NA-3
HB-FJ-NA-7
HB-GD-F-1
HB-HuB-A-1
HB-HuB-A-2
HB-HuB-A-3
HB-JX-F-1
HB-JX-F-2
MH-ZJ-F-1
MH-ZJ-F-3
HB-FJ-L-2
MH-ZJ-NA-1
CL-JS-F-3
CL-JS-F-4
SD-LN-MO-1
CE-GD-MO-1
CE-GD-MO-2
CE-ZJ-MO-1
SV-FJ-MO-1
HB-ZJ-F-2
MC-FJ-MO-1
MC-FJ-NA-1
MC-GS-MO-1
MC-GS-MO-2
MC-GX-A-1
MC-GX-A-3
MC-GX-T-2
HB-FJ-F-3
MC-GX-F-1
MC-GX-F-2
MC-HeB-F-2
MC-HeB-T-1
MC-HeB-T-2
MC-HuN-A-1
MC-HuN-T-1
MC-HuN-A-2
HB-FJ-F-4
MC-HuN-F-1
MC-JS-TNA-2
MC-ZJ-F-1
MC-ZJ-NA-1
RP-ZJ-F-2
RP-ZJ-NA-1
MH-HeB-NA-1
MH-LN-MO-1
PL-AH-MO-4
PL-AH-NA-1
PL-YN-NATF-1
MJ-ZJ-MO-2
MJ-ZJ-MO-3
MJ-ZJ-MO-4
Human papillomavirus type 4
Abelson murine leukemia virus
Murine osteosarcoma virus
Human papillomavirus type 10 genomic DNA
Murine type C retrovirus
Cucumber green mottle mosaic virus
Avian myelocytomatosis virus
Human rubulavirus 2
Y73 sarcoma virus
Mouse kobuvirus M-5/USA/2010
Human endogenous retrovirus K113 complete genome
Shuangao insect virus 7 strain SKC segment 2
Porcine picobirnavirus strain 221/04-16/ITA/2004 segment S
Otarine picobirnavirus isolate HKG-PF080915 segment 2
Canine kobuvirus isolate SMCD-59
Picobirnavirus dog
harvey murine sarcoma virus p21 v-has protein gene
FBR murine osteosarcoma
Avian sarcoma virus CT10 genomic sequence
Hardy-Zuckermann 4 feline sarcoma virus (H24-FeSV) kit oncogene
Chicken picobirnavirus strain PBV/CHK/M3841/HUN/2011 segment RNA2
Picobirnavirus sp. strain PBV/roe_deer/SLO/D38-14/2014 segment 2
Snyder-Theilen feline sarcoma virus genomic sequence
Avian myeloblastosis virus RNA-dependent DNA polymerase gene
KiMSV
20
30
40
50
60
70
80
90

Supp. Fig. 22. Percent coverage of selected viruses by contigs for 64 SRA datasets from PRJNA793740 and PRJNA795267. SRA datasets were *de novo* assembled then aligned to a NCBI viral reference set. Only viruses with 10% or greater coverage and which were found in a minimum of 7 of the 64 SRA datasets were plotted. Names of samples containing GX_ZC45r-CoV sequences are highlighted in red. See Spp. Info. 1 for virus name abbreviations.



Supp. Fig. 23. After alignment of the 16 datasets containing GX_ZC45 sequences to the bat-SL-CoVZC45 reference, 21 reads completely span the non-coding region between NSP10 and RdRp coding regions (indicated by vertical dashed lines). Here only a subset of reads mapping this region are displayed including 8 reads completely spanning the non-coding region. Displayed using NCBI MSA viewer and IGV.

Sample name	Haplogroup	Reads mapped
MC-HuN-T-1	78.7 % F1c1a1 14.7 % H1aw 0.1 % H1t2	464 50 26
MC-HeB-T-1	90.7 % F1c1a1 9.3 % C	1423 18
HB-HuB-A-1	100 % F1c1a1	1919
HB-HuB-A-2	100 % F1c1a1	4885
HB-FJ-NA-3	100 % F1c1a1	2058
HB-FJ-NA-7	100 % F1c1a1	1201
HB-HuB-N-3	100 % F1c1a1	1209
MP-ZJ-MO-4	100 % F1c1a1	1292
MJ-ZJ-MO-1	100% F1c1	22444
MJ-ZJ-MO-2	100 % H27/H27e	5142
MJ-ZJ-MO-3	-	0 reads mapped
MJ-ZJ-MO-4	85.7 % F1c1a1 14.3 % H27/H27e	764 149
MJ-ZJ-MO-6	100 % F1c1a1	7676
RP-JX-A-2	100 % F1c1a1	3150

Supp. Table 1. Haplogroup analysis of human mitochondrial reads identified in datasets with high CoV presence

Event	Begin	End	Recombinant Sequence(s)	Minor Parental Sequence(s)	Major Parental Sequence(s)	RDP	GENECONV	Bootscan	Maxchi	Chimaera	SiSscan	PhylPro	LARD	3Seq
1	20*	896	GX_ZC45r-CoV	PCoV_GX-P4L	FJ2021D	1.88E-33	8.18E-33	6.04E-31	5.16E-22	6.78E-23	1.18E-27	NS	NS	2.03E-40
1				GX_WIV										
2	3140	3748	^GX_ZC45r-CoV	bat-SL-CoVZC45	Unknown (HKU3-1)	1.51E-02	1.70E-04	NS	4.14E-04	4.02E-03	1.14E-10	NS	NS	7.17E-03

Supp. Table. 2. Potential recombination breakpoints detected using RDP5 detected by more than three methods. * The actual breakpoint position is undetermined. ^ The recombinant sequence may have been misidentified. Referenced to spliced NSP4 +NSP10, RdRp partial genome.