

Supplementary Figures and Tables

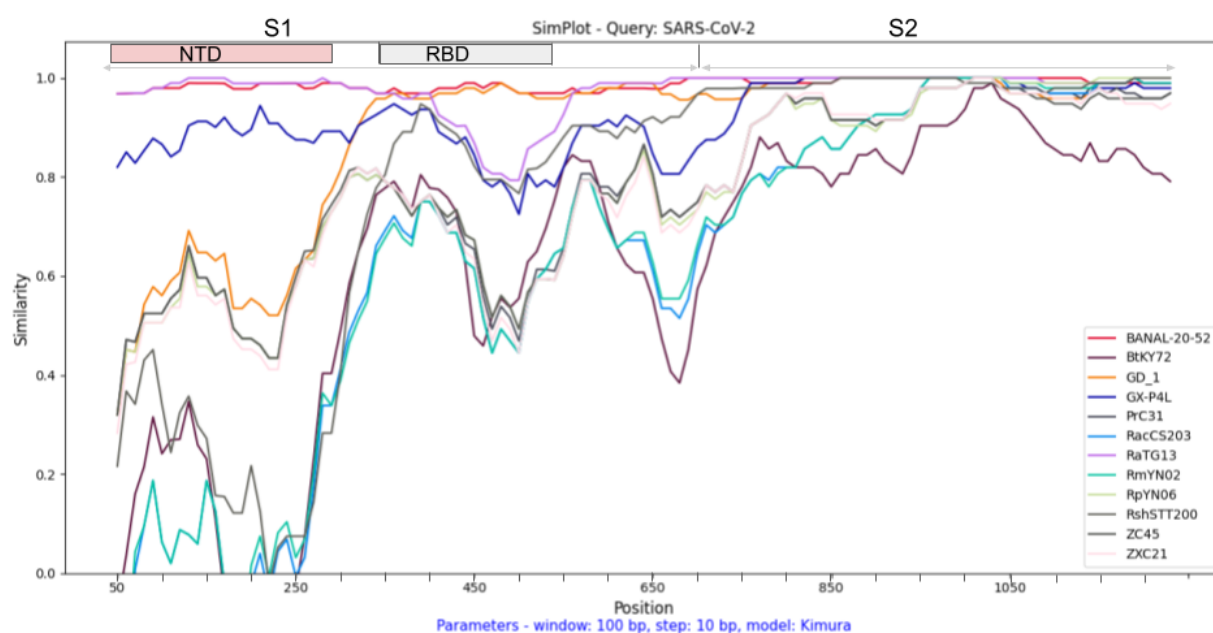


Figure S1. SARS-CoV-2 spike protein amino acid similarity to SARS2r-CoVs. Aligned using the MUSCLE algorithm in UGENE, plotted using SimPlot++ using a Kimura 2 parameter model with a 100bp window and 10bp step size. Both PCoV GD_1 and BANAL-52 have high amino acid similarity to the SARS-CoV-2 RBD. Reference numbering relative to multiple-sequence alignment.

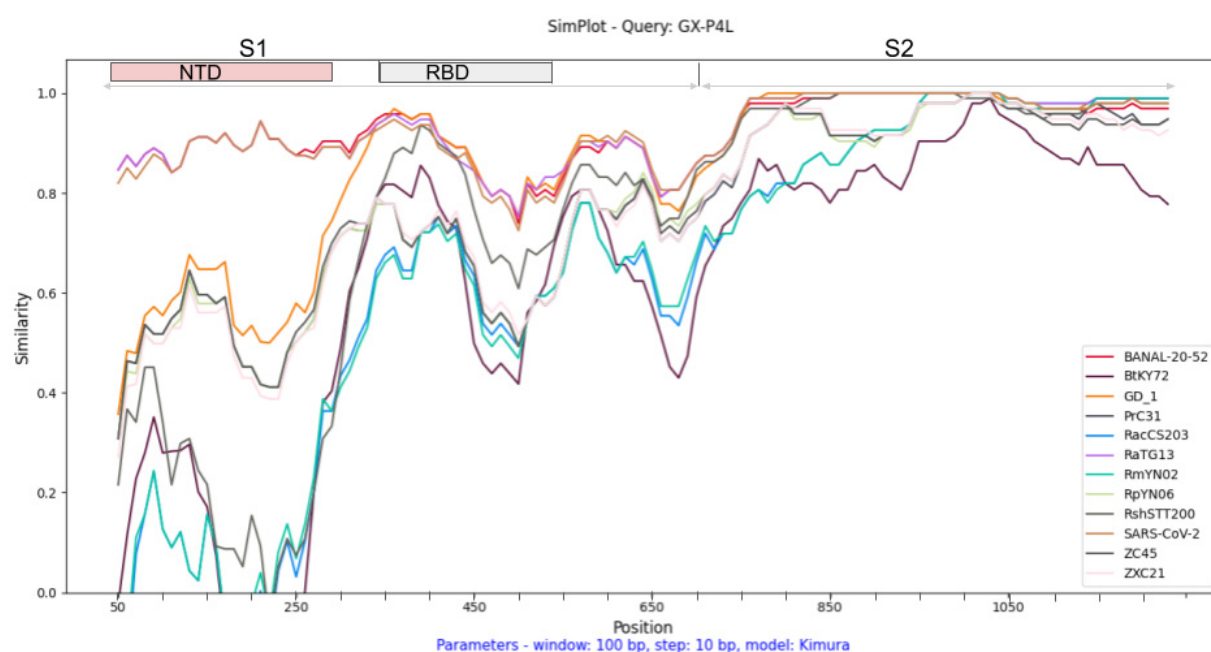


Figure S2. PCoV GX_P4L spike amino acid similarity plot to SARS2r-CoVs. Aligned using MUSCLE algorithm in UGENE, plotted using SimPlot++ using a Kimura 2 parameter model with a 100bp window and 10bp step size. PCoV GX_P4L has highest overall amino acid

similarity to RaTG13 and SARS-CoV-2. Reference numbering relative to multiple-sequence alignment.

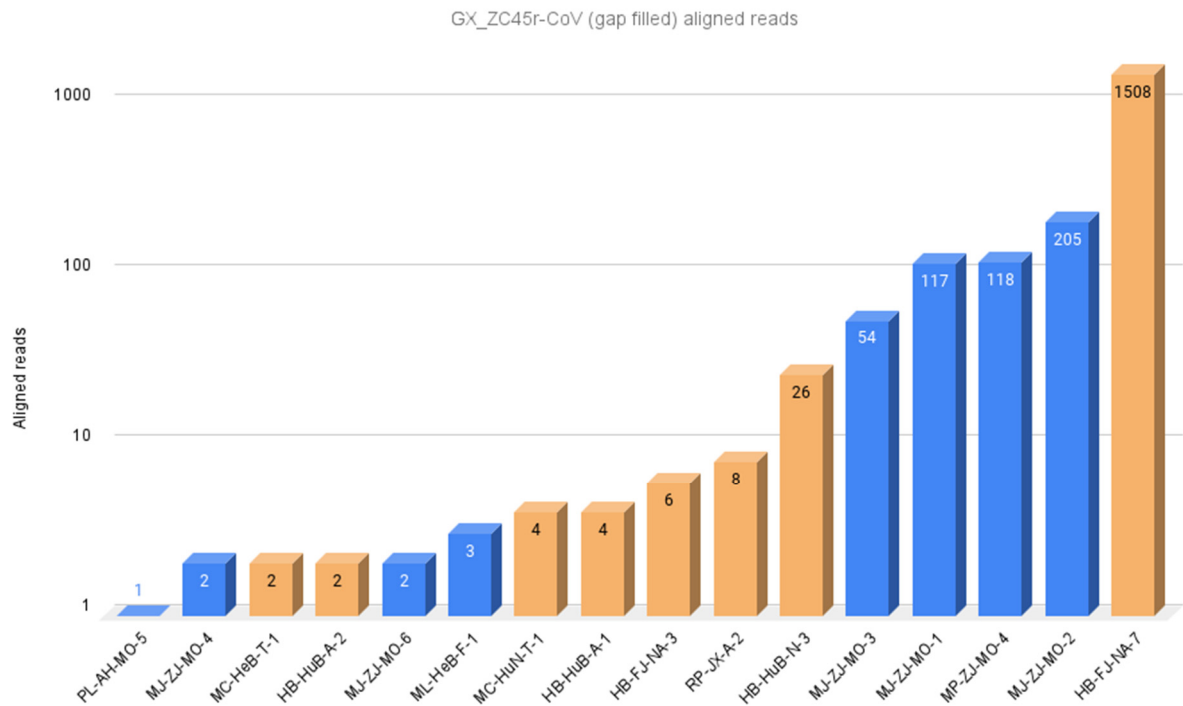


Figure S3. Reads per sample aligning to GX_ZC45r-CoV gap-filled genome. BioProject PRJNA793740 samples coloured blue, BioProject PRJNA795267 samples coloured orange. Read count in log scale. Datasets containing GX_ZC45r-CoV newly identified here: PL-AH-MO-5, MC-HeB-T-1, HB-HuB-A-2, ML-HeB-F-1, MC-HuN-T-1, HB-HuB-A-1 and RP-JX-A-2.

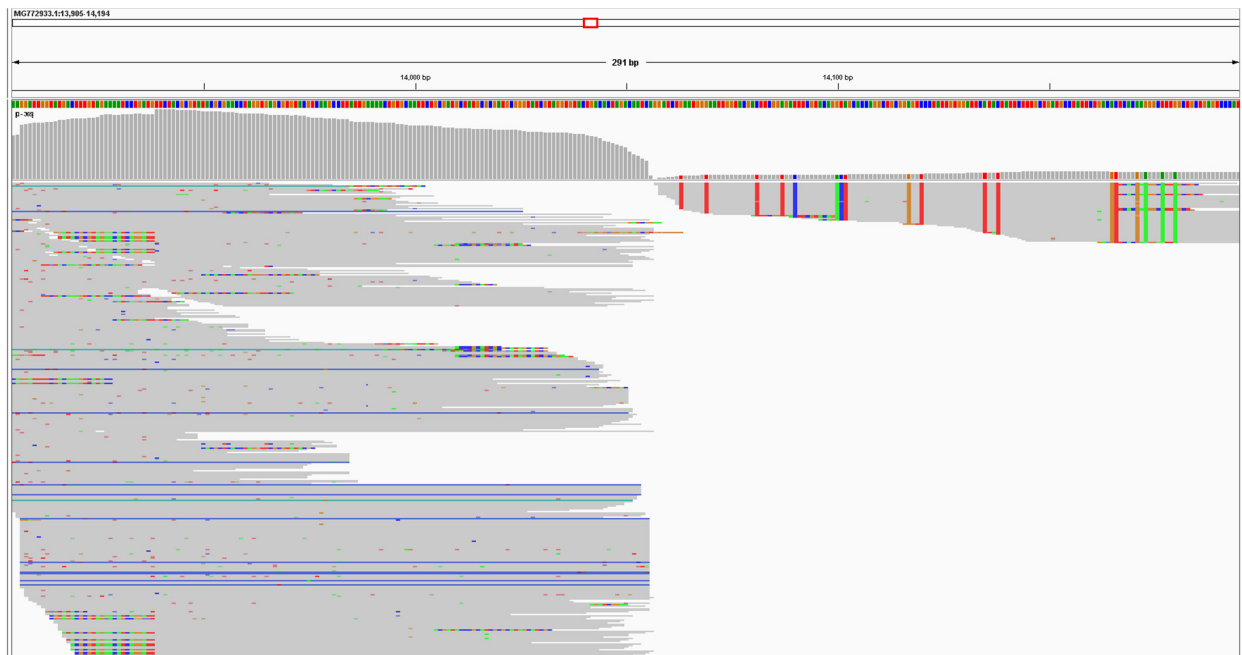


Figure S4. Alignment of GX_ZC45r-CoV to bat-SL-CoVZC45 (MG772933.1) showing anomalous read coverage across position 14056nt potentially indicating artificial splicing.

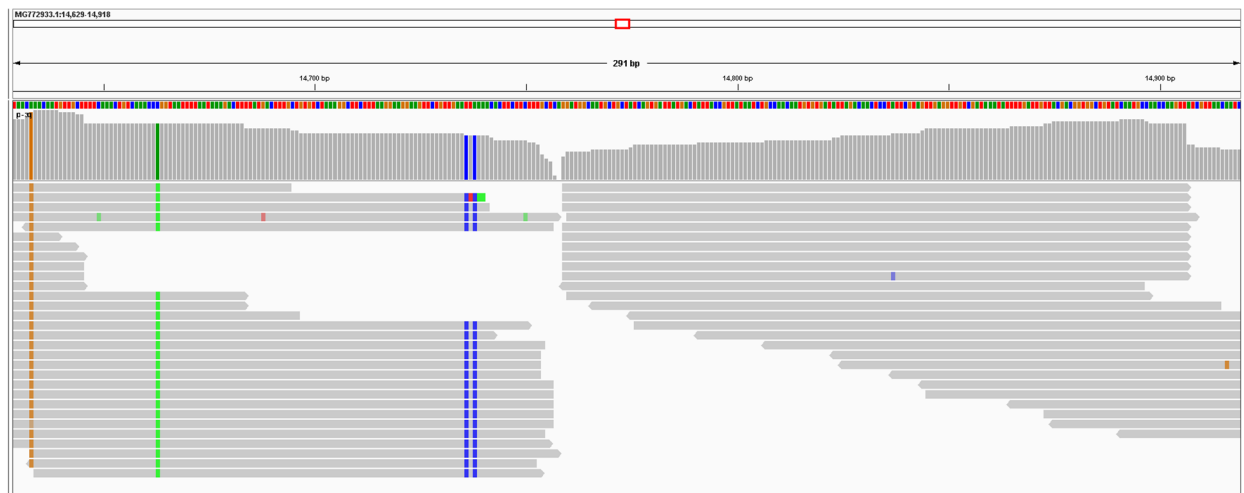


Figure S5. Alignment of GX_ZC45r-CoV to bat-SL-CoVZC45 (MG772933.1) showing anomalous read coverage across position 14758nt potentially indicating artificial splicing.

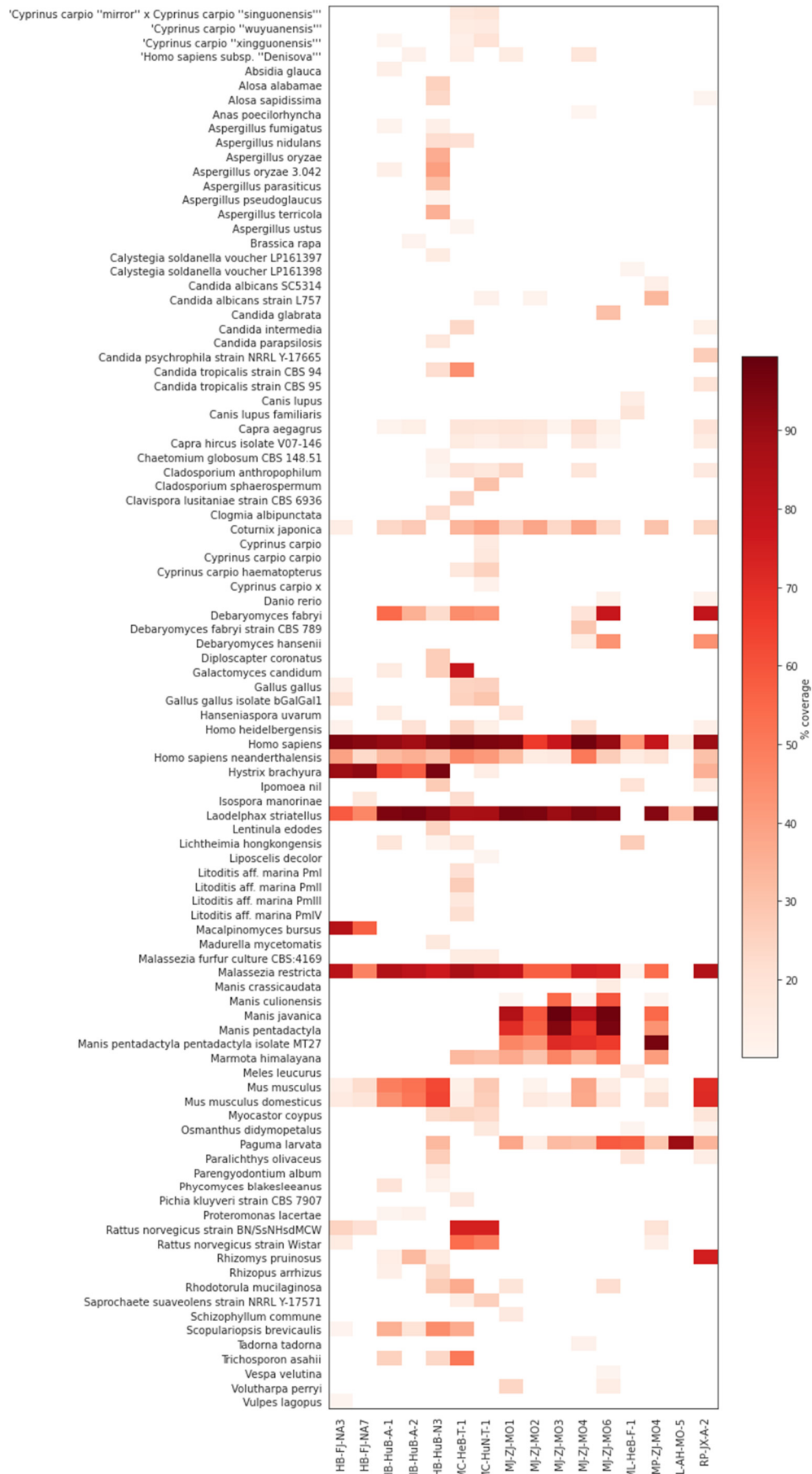


Figure S6. Mitochondrial genome coverage $\geq 10\%$ for the 16 game animal datasets containing SARS2r-CoV reads. Alignment generated using bowtie2, plotted using matplotlib.



Figure S7. Blastn analysis of the recovered NSP10 region of GX_ZC45r-CoV showing the 13 highest identity genome matches.

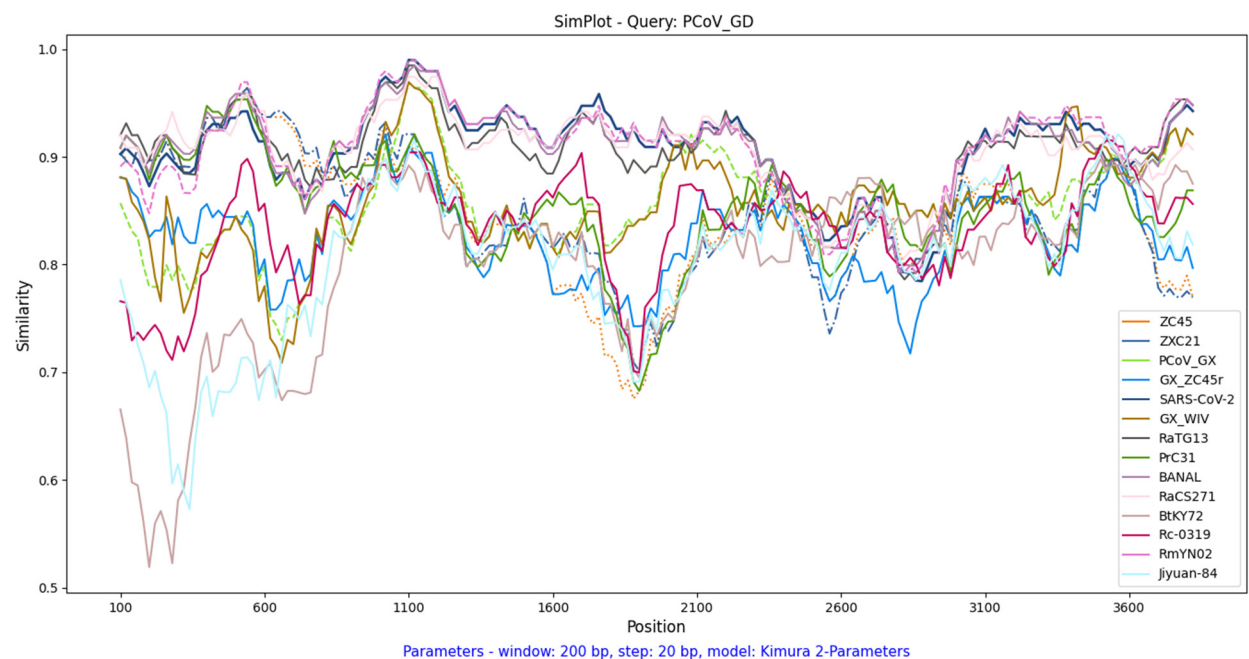


Figure S8. Simplot analysis of selected SARSr-CoV genomes using PCoV MP789 as a query. Note the same 3166nt gap as per Figure 6 is situated at 942-943nt. Plotted using Simplot++. Solid lines except: RmYN02 - dashed, PCoV GX - dashed, bat-SL-CoVZC45 - dotted, bat-SL-CoVZXC21 - dash - dotted, SARS-CoV-2 line weight 2 (all others line weight 1.5).

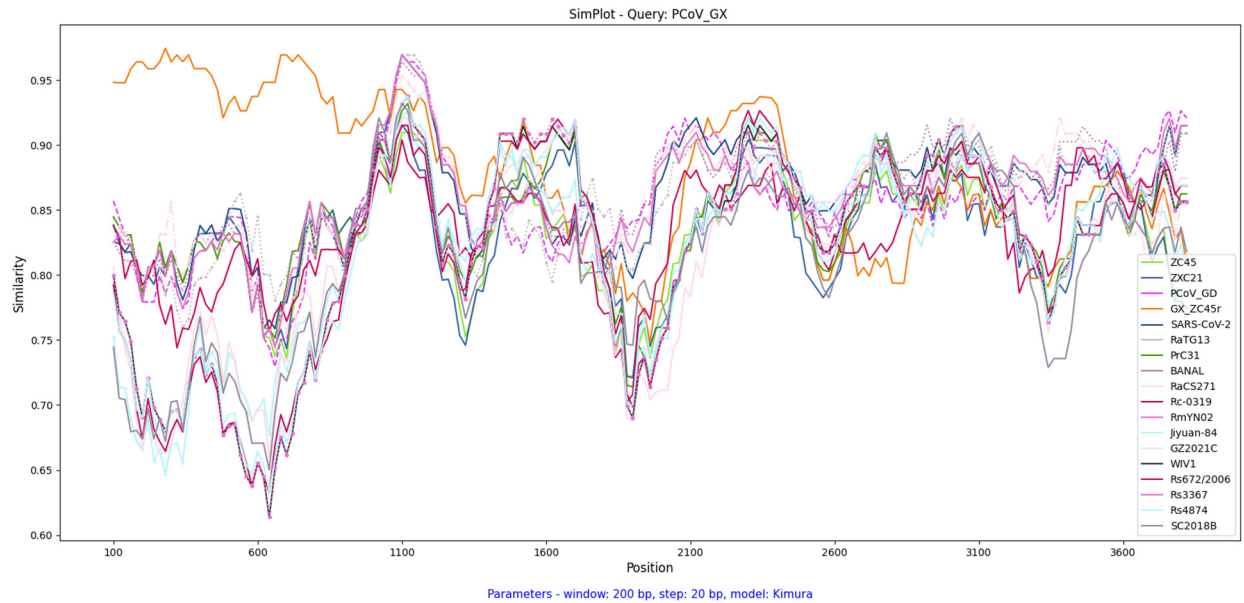


Figure S9. Simplot analysis of selected SARSr-CoV genomes using GX PCoVgroup as a query. Note the same 3166nt gap as per Figure 6 is situated at 942-943nt. Solid lines except: BANAL group - dotted, GD PCoV group - dashed, Rs3367 - dotted with circle markers, RaCS271 - dash dot with hexagon markers, Jiyuan-84 - solid with square markers, RaTG13 dotted with plus markers. Plotted using Simplot++. See methods for CoV grouping.

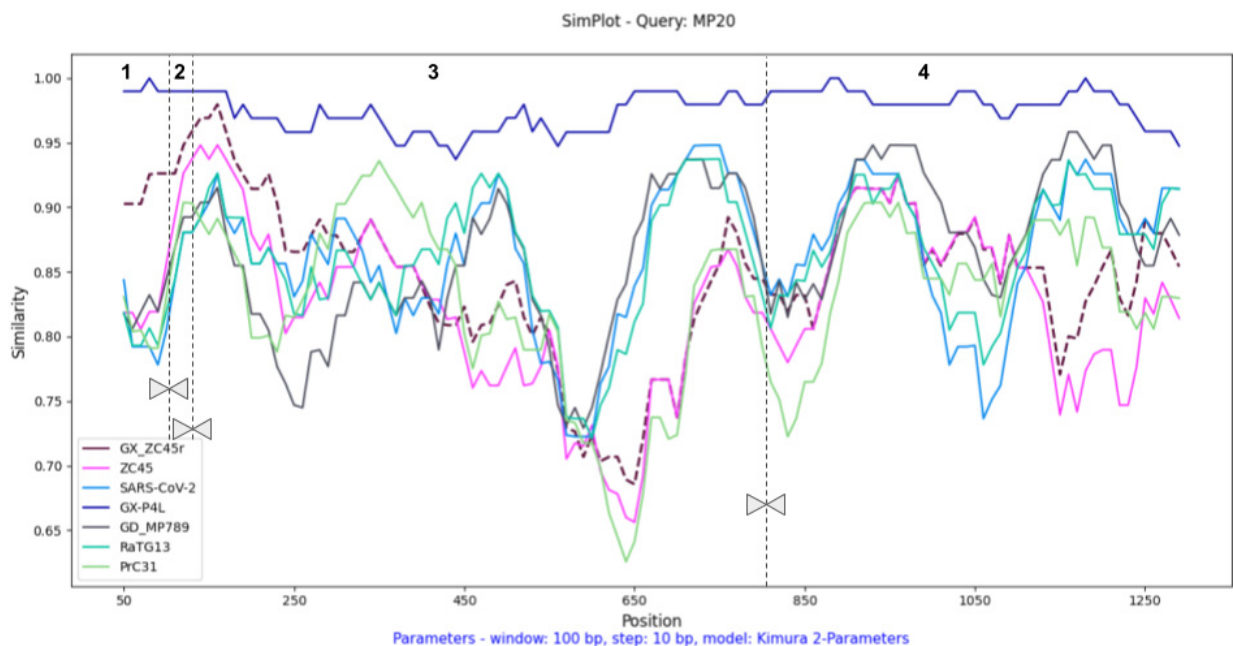


Figure S10. Simplot analysis of four sections of PCoV MP20 which overlap with sections of GX_ZC45r-CoV after multi sequence alignment. PCoVMP20 was queried against selected SARSr-CoV genomes. The four regions are: 1) 105nt at 3' end of NSP4, 2) 29nt at the 5' end of the NSP10 gene covered by GX_ZC45r-CoV, 3) 675nt in RdRp and 4) 532nt at the 3' end of the RdRp gene coding region. GX_ZC45r-CoV plotted as dashed lines. Plotted using Simplot++.



Figure S11. 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.

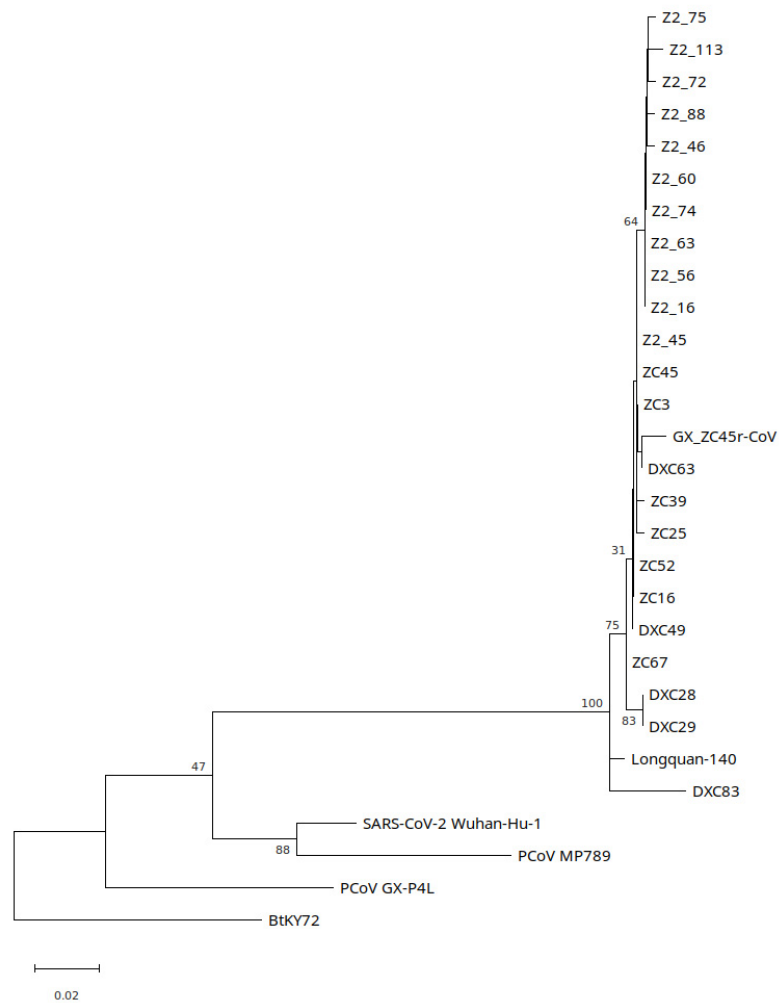


Figure S12. Partial RdRp section (407nt) maximum likelihood phylogenetic tree using a T92+I model with 100 bootstrap replicates using MEGA11.

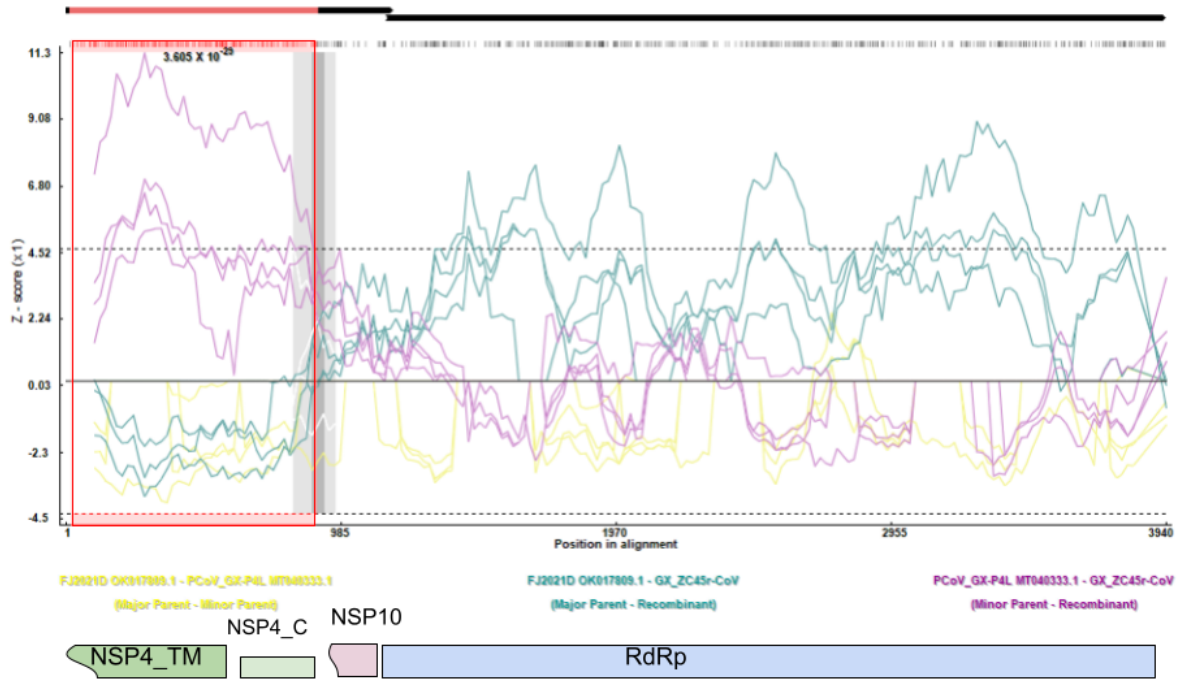


Figure S13. Potential recombination region 1 detected using SiScan model in RDP5. Note 5' end of recombination region not detected.

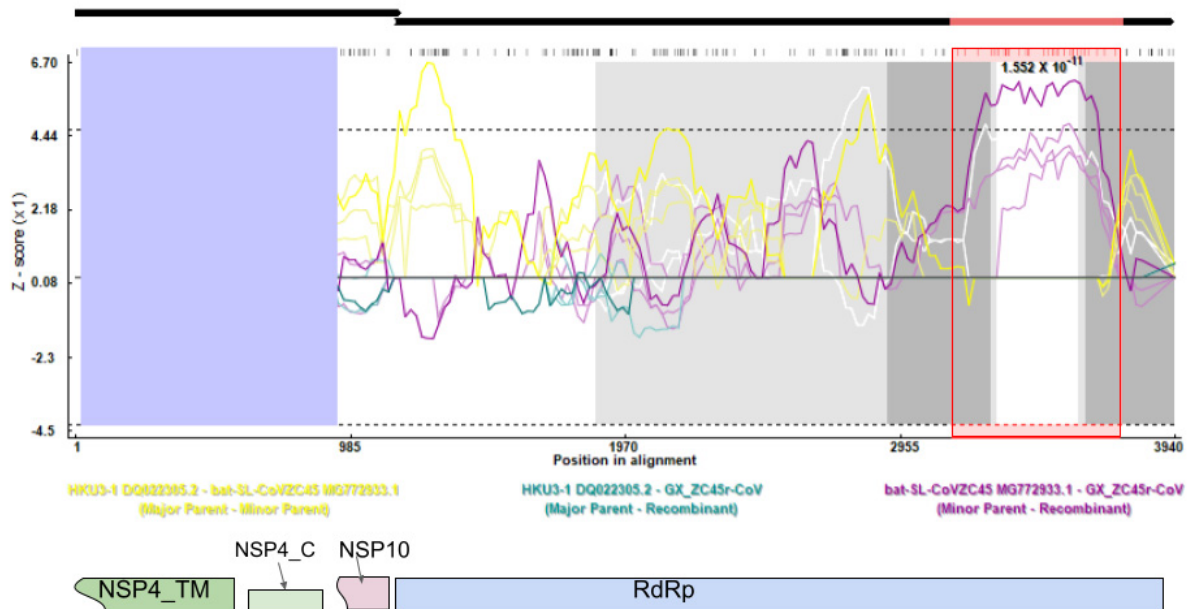


Figure S14. Potential recombination region 2 detected using SiScan model in RDP5.

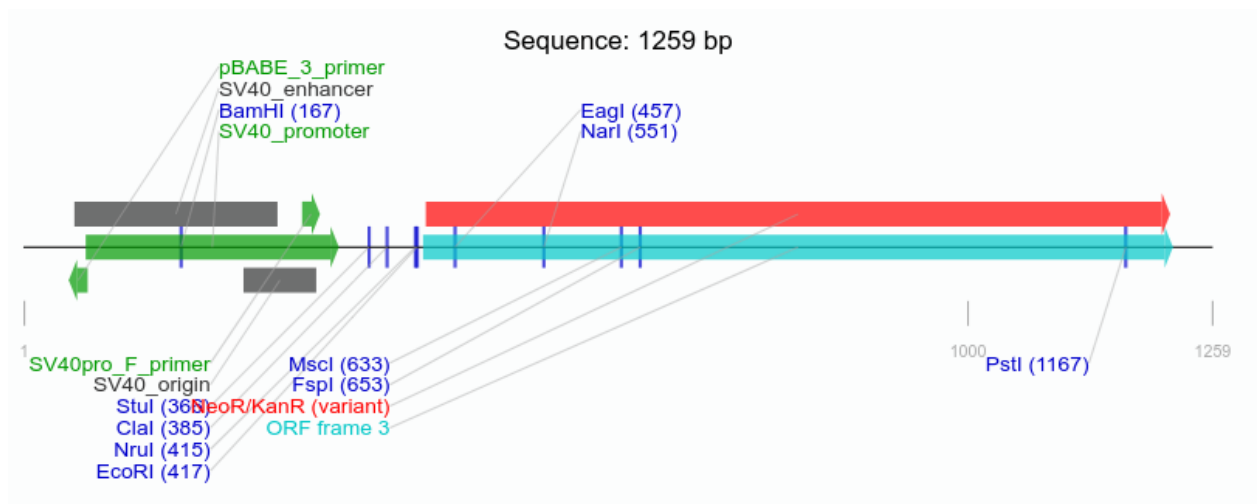


Figure S15. Partial vector sequence identified in MJ-ZJ-F-1 with similar layout to mammalian expression vector pSV2neo.

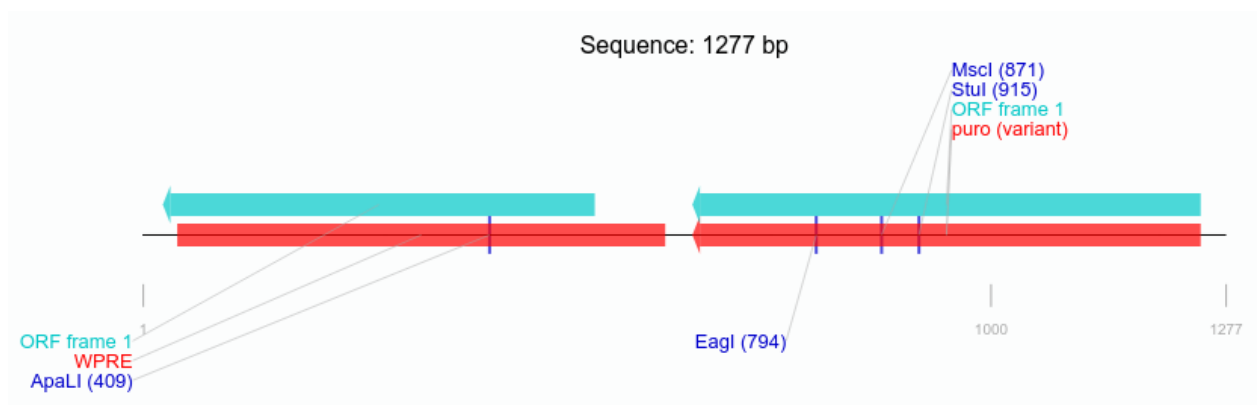


Figure S16. Partial synthetic vector de novo assembled from the MC-GX-A-1 dataset. Highest blastn max score match is found to Lentiviral expression vector MCPyV (MZ648044.1) (98.08% identity).

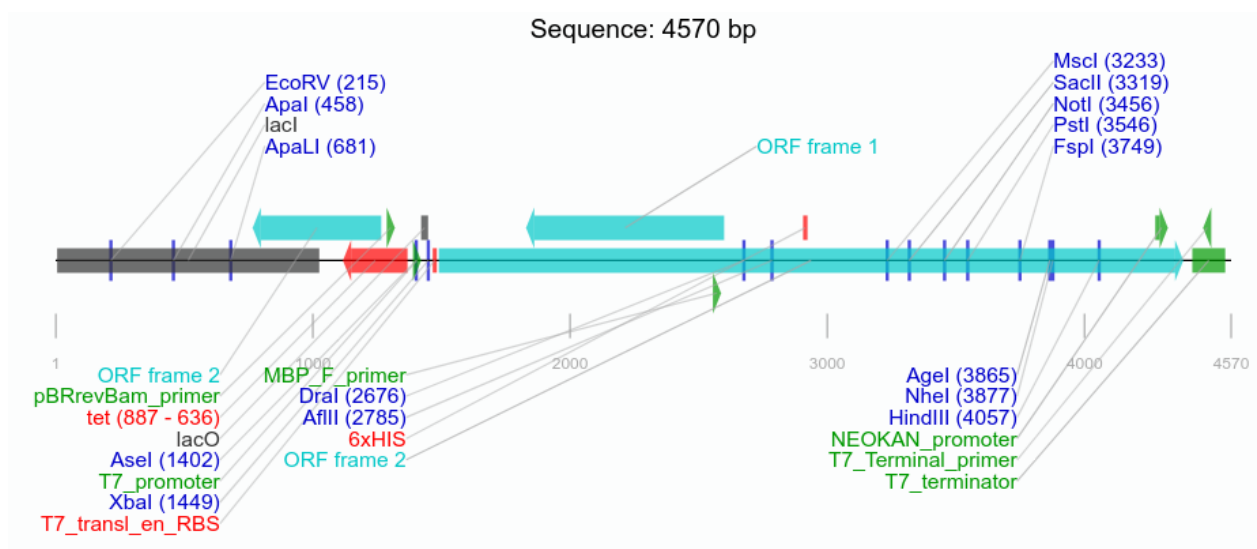


Figure S17. Synthetic plasmid sequence *de novo* assembled from the MJ-ZJ-F-3 dataset. Highest blastn identity is found to Tn5 transposase cloning vectors pKRCPN1 (ALS31075.1), TnpS [Cloning vector pTHSSe_56] (AVR55160.1) and Tn5 hyperactive transposase [Cloning vector pBAM1] (ADY68344.1).

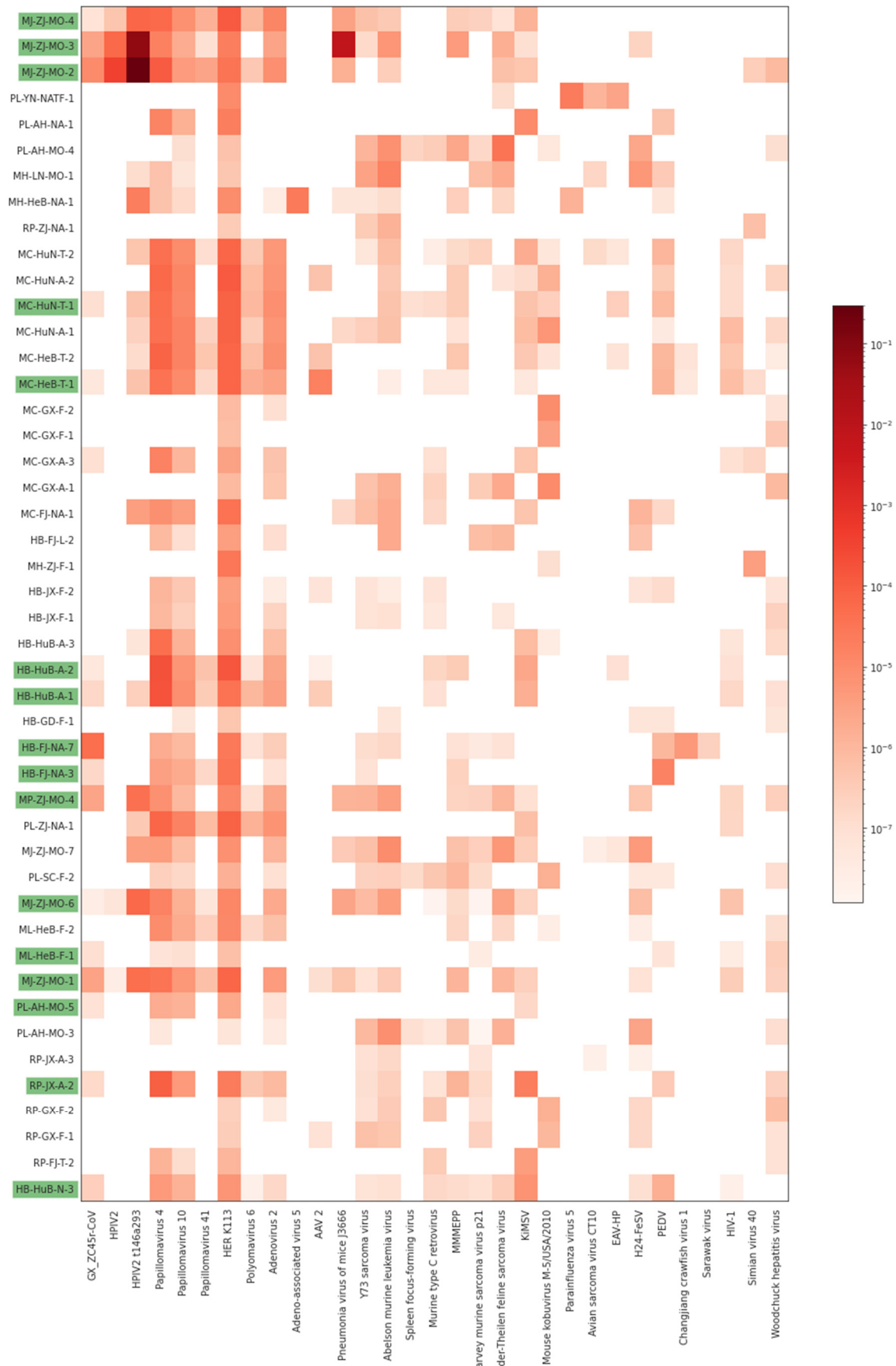


Figure S18. Counts per read for reads mapping to selected viruses per SRA dataset for 46 SRA datasets from PRJNA793740 and PRJNA795267. For all but GX_ZC45r-CoV and SV40 a 10% cutoff was applied whereby each virus is present in at least one of the 46 datasets with 10% or greater genome coverage. Normalized 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 Supp. Info. 2 for virus name abbreviations.

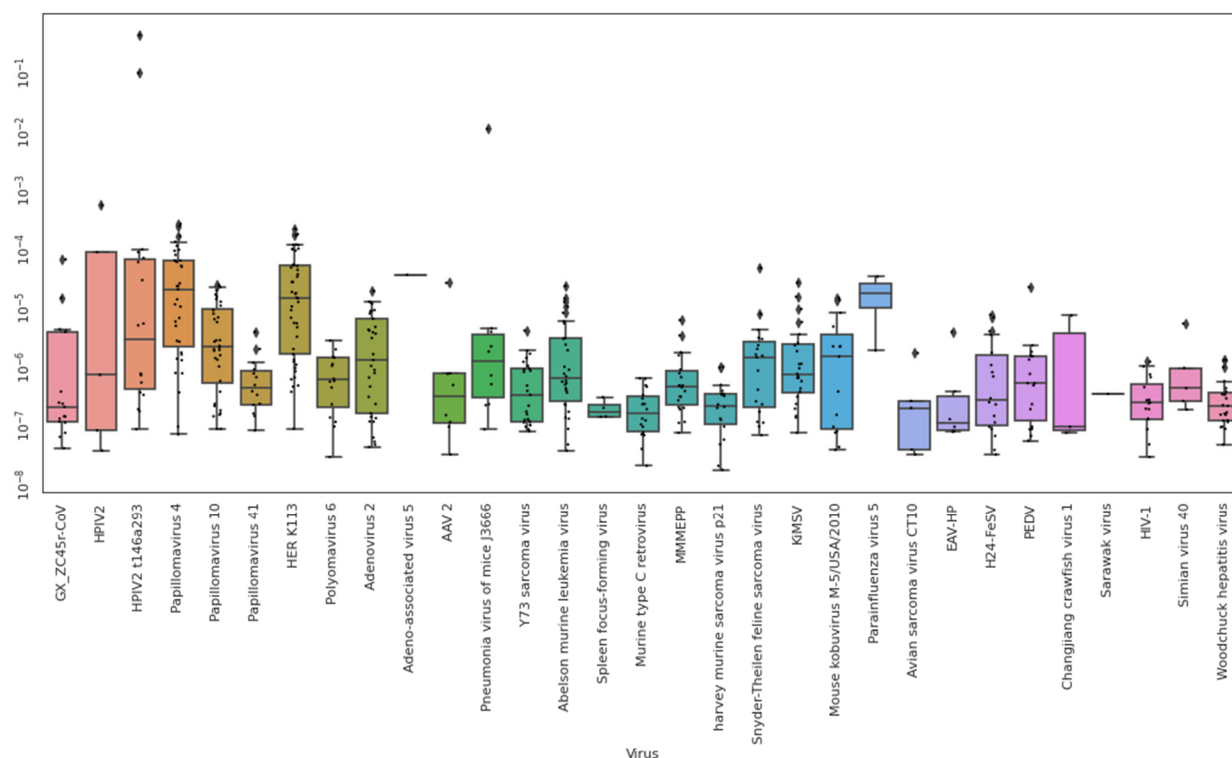


Figure S19. Box plot distribution of counts per read mapping to selected viruses per SRA dataset for 46 SRA datasets from PRJNA793740 and PRJNA795267. For all but GZ_ZC45r-CoV and SV40 a 10% cutoff was applied whereby each virus is present in at least one of the 46 datasets with 10% or greater genome coverage. Read counts shown in log scale. See Supp. Info. 2 for virus name abbreviations.

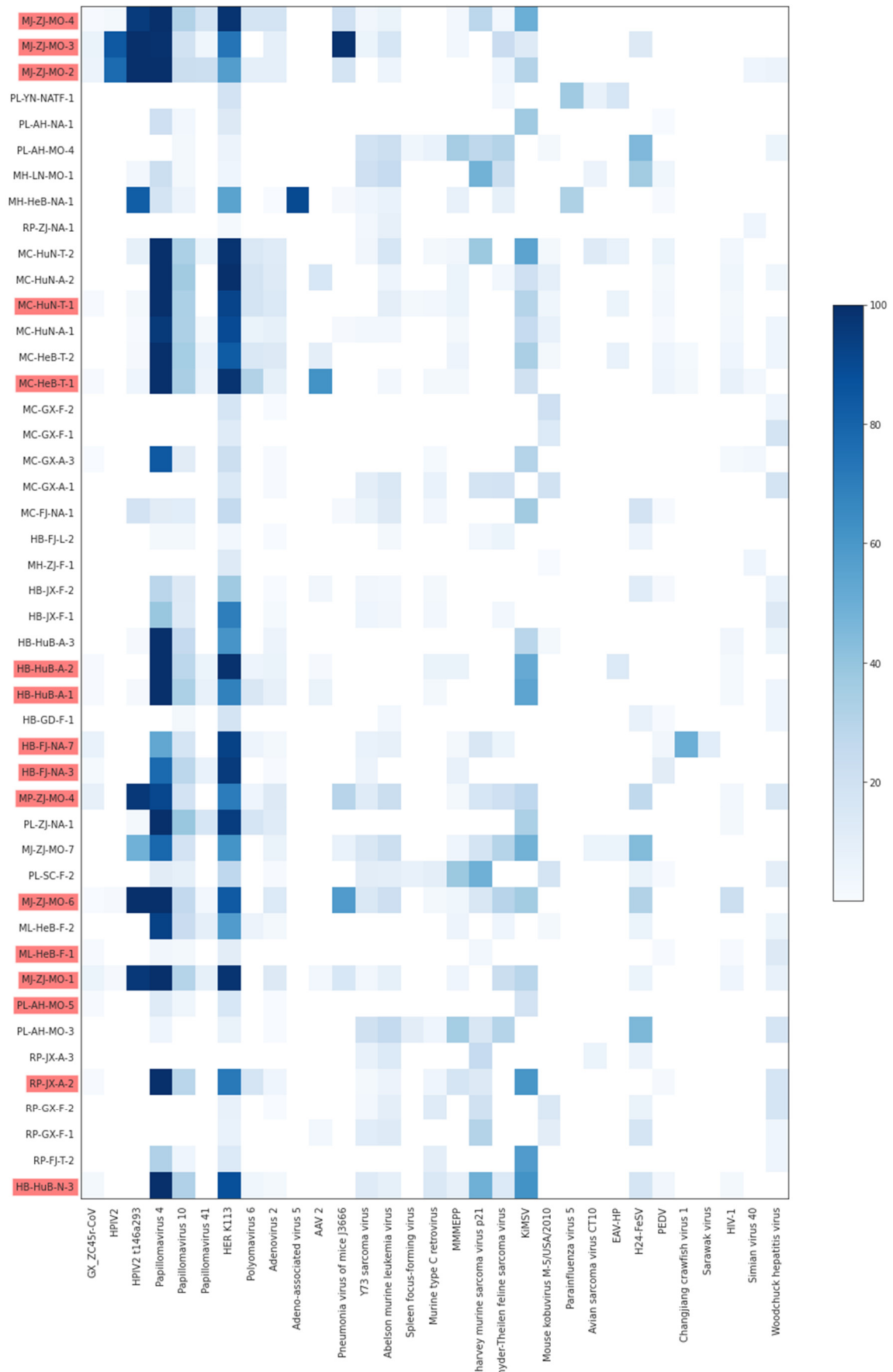


Figure S20. Percent coverage of selected viruses by reads for 46 SRA datasets from PRJNA793740 and PRJNA795267. For all but GX_ZC45r-CoV and SV40 a 10% cutoff was applied whereby each virus is present in at least one of the 46 datasets with 10% or greater genome coverage. Names of samples containing GX_ZC45r-CoV sequences are highlighted in red. See Supp. Info. 2 for virus name abbreviations.

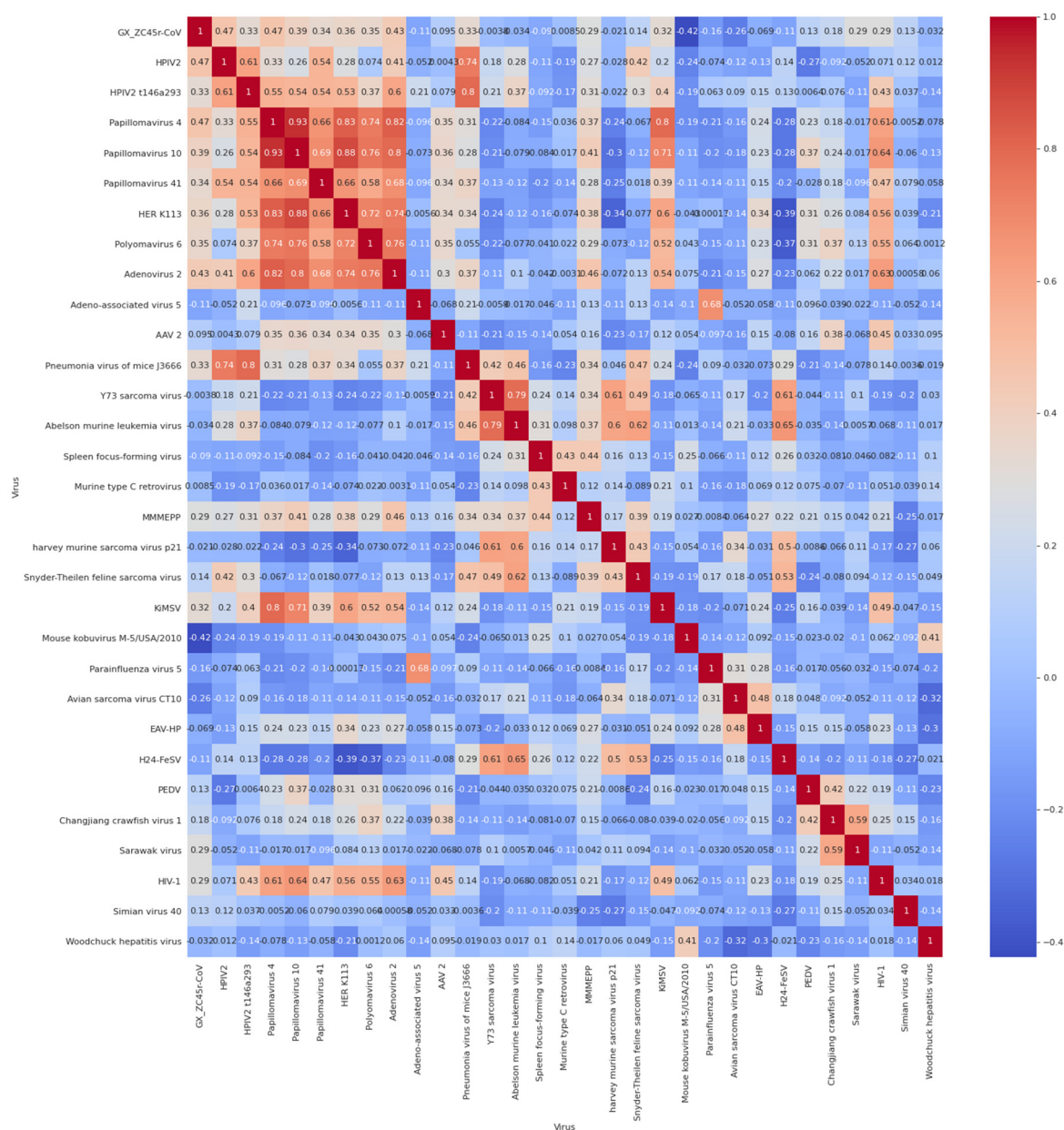


Figure S21. Spearman correlation coefficient matrix for virus read counts for 46 SRA datasets from PRJNA793740 and PRJNA795267 (see Figure S18 for sample names). SRAs were aligned to viruses using bowtie2 using the '--very-sensitive' parameter. For all but GX_ZC45r-CoV and SV40 a 10% cutoff was applied 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.

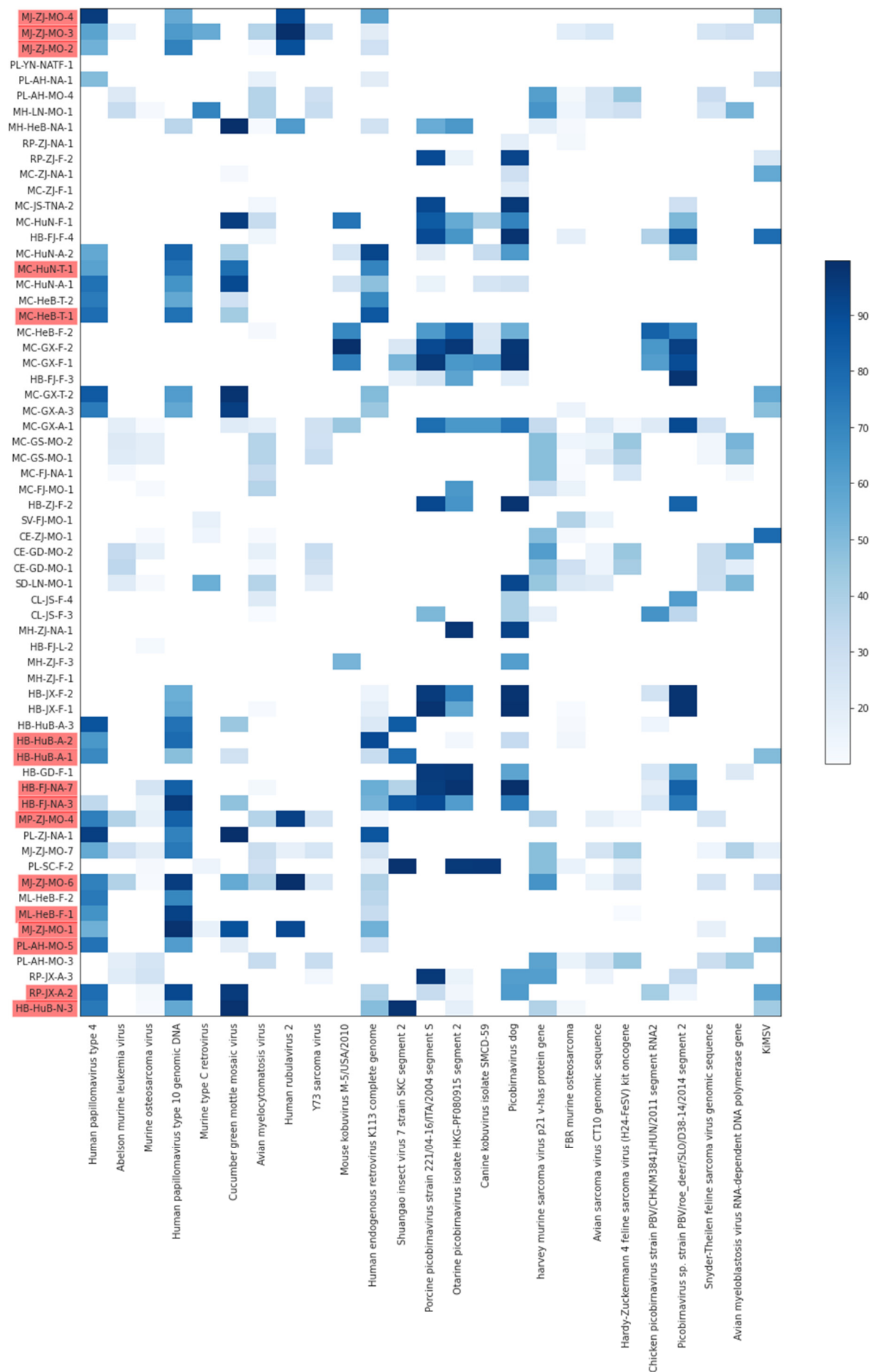


Figure S22. Percent coverage of selected viruses by contigs for 64 SRA datasets from PRJNA793740 and PRJNA795267. SRA datasets were *de novo* assembled then aligned to a NCBI viral reference set. Only viruses with 10% or greater coverage and which were found in a minimum of 7 of the 64 SRA datasets were plotted. Names of samples containing GX_ZC45r-CoV sequences are highlighted in red. See Supp. Info. 2 for virus name abbreviations.

Bat SARS-like coronavirus isolate bat-SL-CoVZC45, complete genome

GenBank: MG772933.1

GenBank FASTA



Figure S23. After alignment of the 16 datasets containing GX_ZC45r-CoV 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.

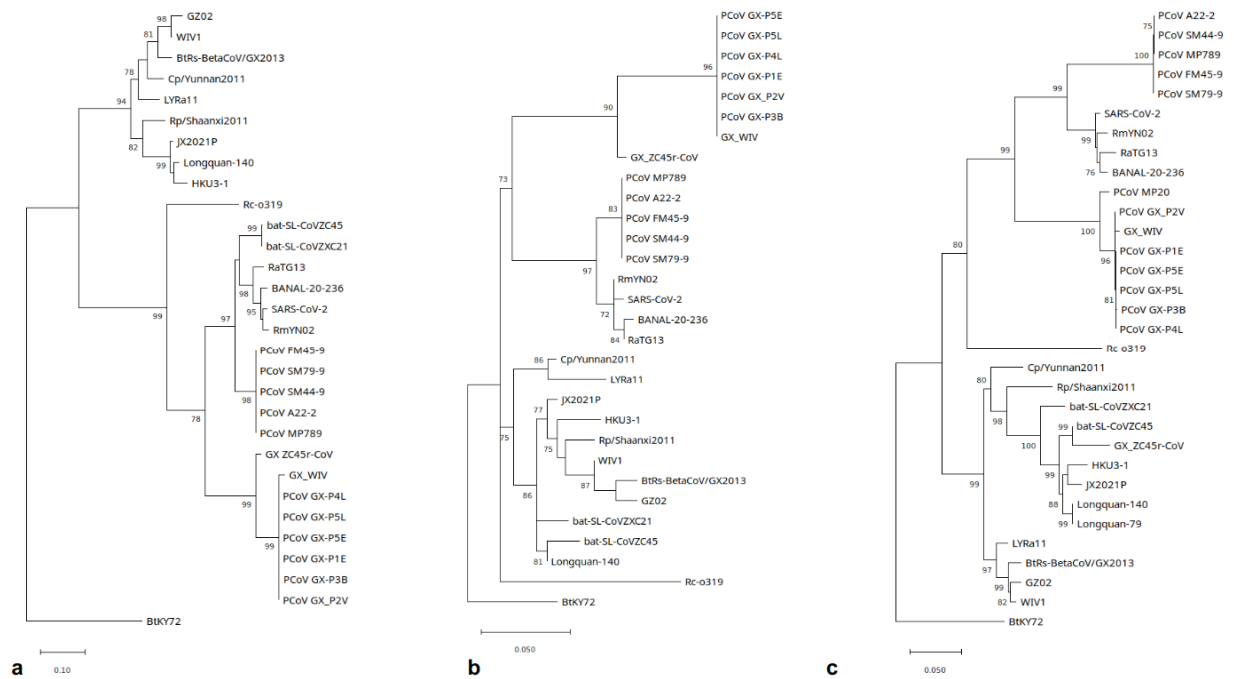


Figure S24. Maximum likelihood phylogenetic trees generated using PhyML with model selected using smart model section for: a) partial NSP4 region using GTR+G+I model; b) NSP10 region using TN93+G model; and c) RdRp region using GTR+G model. Support values of 70% or greater are shown.

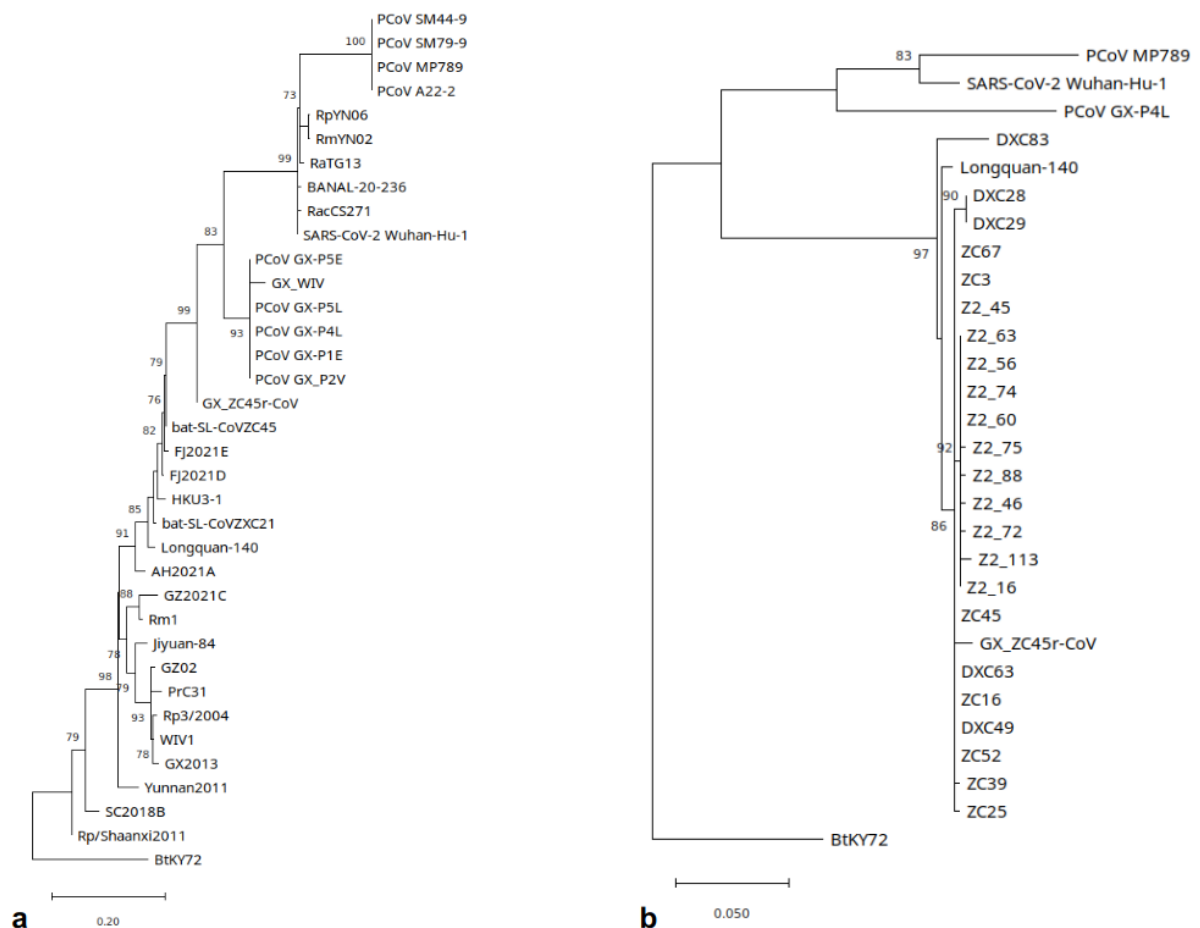


Figure S25. Maximum likelihood phylogenetic trees generated using PhyML with model selected using smart model section for: a) 297nt partial RdRp region using HKY85+G model; b) 407nt partial RdRp region using T93+G model. Support values of 70% or greater are shown.

Sample name	Haplogroup	Reads mapped
MC-HuN-T-1	78.7 % F1c1a1	464
	14.7 % H1aw	50
	0.1 % H1t2	26
MC-HeB-T-1	90.7 % F1c1a1	1423
	9.3 % C	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	764
	14.3 % H27/H27e	149
MJ-ZJ-MO-6	100 % F1c1a1	7676
RP-JX-A-2	100 % F1c1a1	3150

Table S1. Haplogroup analysis of human mitochondrial reads identified in datasets with high GX_ZC45r-CoV presence.

Event	Begin	End	Recombinant Sequence(s)	Minor Parental Sequence(s)	Major Parental Sequence(s)	RDP	GENECONV	Bootscan	Maxchi	Chimaera	SiScan	PhyPro	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

Table S2. 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.