The Pan-SL-CoV/GD sequences may be from contamination.

Daoyu Zhang.

ABSTRACT

Recently, there were much hype about reports of SARS-like coronaviruses being found in samples of Malayan pangolins (Manis Javanica) collected in Guangdong 2019, which appeared to possess nearly identical RBD's to the SARS-CoV-2 coronavirus. Prominent journals cited these discoveries to claim that pangolins may be a possible intermediate host for the zoonotic transmission of SARS-CoV-2 to humans.

Here, we report that all databases used to support such a claim, upon which metagenomic analysis was possible, contained unexpected reads and was potentially contaminated. Here we also report that the presences of these unexpected reads are directly related to the presence of coronavirus reads. Finally, we deduced the actual causative agent of the death of the pangolins sampled in Guangdong 2019 where the claim of coronavirus detections was made.

METHODS

The NCBI Trace tool

The NCBI SRA archive come with it's own tool called Trace, which identifies the origin or reads within the SRA dataset through the recognition of unique K-mers within the nucleotide sequence. Multiple reads of 32 nucleotides is taken from each read to identify the reads toward an origin by comparison with a large database of reference sequences, which produces a classification signal. Then read of 64 nucleotides are taken from each of the read for definitive mapping toward species in the reference database. If any one of the 32nt or 64nt K-mers are found in more than one reference sequence, the reads are instead classified at the lowest phylogenetic classification node where reference sequences containing such a K-mer is found.

The 32nt TRACE generate a "strong signal" classification of sequence origin useful for the deduction of the content of the sample by organism of origin, accessed via the NCBI Krona charting tool,

While the 64nt TRACE generate a definitive classification signal used for the exact tracing of reads to the origin from a specific Species/Taxon, used for the exact classification of reads.

Both the 32nt and 64nt TRACE analysis classify their reads according to the lowest common taxonomical node where K-mers from said read are present in the reference sequence database, a strategy known as "lowest non-ambiguous mapping". Such a strategy avoids the problem with RNA degradation or sequencing errors by excluding potential errors in reads, without introducing potential ambiguous classification by clustering ambiguous reads under the lowest common

classification node such ambiguity is found.
Therefore, if TRACE gives identification to a specific taxonomical node for a sequence read, it could be from any of the taxonomical nodes and species classified under the node, but it could not be from a taxonomical node or species that is not under said node. E.g. if TRACE says hominoidea which was classified under Catarrhini; Simiiformes; Haplorrhini; Primates; Euarchontoglires, Then it can't be from a pangolin since pangolins (Manis Spp.) are classified under Pholidota; Laurasiatheria. The lowest common classification node between Primates and Pangolins is Boreoeutheria—reads from parts of the genomes shared between Primates and Pangolins will only be classified to Boreoeutheria, but not further classified down toward either Laurasiatheria or Euarchontoglires. And definitely will not be classified individually toward Pholidota or Primates, or any child nodes or phylogenetic nodes under them.

Specific BLAST analysis

Whenever a genus or species is provided by analysis, a specific BLAST analysis is performed to confirm the presence of reads toward the exact species by a search of the database in question with representative reference sequences of the specific species in question in look for matches that is either: 100% match, or: contained no 100% matches on BLAST when queried against the Pangolin reference sequences available on GanBank.

Genome alignment and analysis of non-NCBI datasets

If online BLAST analysis was unavailable to a dataset used, the dataset (as fastq) is downloaded and the sequences are aligned with the human genome using BowTie2[20] or MagicBLAST[21]. The Alignment results (if present) are then verified for their uniqueness to the human genome as per the Specific BLAST analysis method listed above.

Sequence assembly

Sequence assembly of PRJNA610466 short read data was performed using MEGAHIT[16] on paired-end sequencing data with default parameters.

RESULTS

The Accession numbers and contents of all Pan-SL-CoV/GD related sequencing experiments are listed under the following table.

Table 1: List of available GD Pangolin sample datasets as provided in the NCBI SRA. By Accession number, size and citation by thesis (if claimed to have SARS-CoV-2 related reads by paper).

Accession number	Size	SARS-CoV-2-like Coronavirus Identified and Cited?
SRX6893158	16,491,648	

SRX6893157	9,275,501	Lung12 [3] SRR10168374
SRX6893156	22,220,187	Lung11 [1]
SRX6893155	18,067,615	Lung09 [1] [3] SRR10168376
SRX6893154	16,414,925	Lung08 [1] [3] [4] SRR10168377
SRX6893153	19,045,923	Lung07 [1] [3] [4] SRR10168378
SRX6893152	13,527,964	
SRX6893151	16,068,654	
SRX6893150	12,967,281	
SRX6893149	12,590,769	
SRX6893148	15,273,939	
SRX6893147	15,975,904	
SRX6893146	19,038,817	
SRX6893145	19,055,973	
SRX6893144	15,350,468	
SRX6893143	11,527,782	
SRX6893142	20,045,443	
SRX6893141	18,903,834	
SRX6893140	19,986,780	
SRX6893139	39,738,679	Lung02 [3] SRR10168392
SRX6893138	22,900,426	
SRX7756769	107,267,359 PRJNA607174**	M1[2]***
SRX7756766	273,651,431 PRJNA607174**	
SRX7756765	196,761,202 PRJNA607174**	
SRX7756764	222,286,763 PRJNA607174**	
SRX7756763	212,161,250 PRJNA607174**	
SRX7756762	232,433,120 PRJNA607174**	M6[2]***
SRX7756761	113,900,941 PRJNA607174**	
SRX7732094	2,633*	“P2S”[3]

*: “Design: This dataset contains coronavirus-like sequence reads, based on BLAST search.”

**: All available SRA datasets from PRJNA607174

***:Actual SRA datasets identified from the “Extended Data Table 3” of [2]

Article

Extended Data Table 3 | Identification of SARSr-CoV sequence reads in metagenomes from the lung of pangolins using the SARS-CoV-2 sequence (GenBank accession No. MN908947) as the reference

Sample ID	Animal species	Total reads*	No. mapped	
M1	Malayan pangolin	107,267,359	496	←SRX7756769 "pangolin 9"
M2	Malayan pangolin	38,091,846	302	
M3	Malayan pangolin	79,477,358	14	
M4	Malayan pangolin	32,829,850	1,100	Not available
M5	Malayan pangolin	547,302,862	56	
M6	Malayan pangolin	232,433,120	10	←SRX7756762 "pangolin 2"
M8	Malayan pangolin	44,440,374	12	
M10	Malayan pangolin	227,801,882	0	Not available
Z1	Chinese pangolin	444,573,526	0	

Fig.1 the "Extended Data Table 3" of [2]. SRA datasets identified in the available database is pointed out by an arrow, while SRA "runs" that failed to be identified in known datasets are outlined in a red square.

Analysis of reads from The Available datasets using NCBI Trace.

Table 2. The Trace result of Known GD Pangolin datasets when examined using NCBI Trace SRA.

Accession number and registration date	Primary Mammalian Trace results and percentage	Primate-related results in Krona and read size by Kbp	Identification of "Coronaviridae" as by Trace and total read size
SRX6893158 20-Sep-2019	Manis javanica: **14.66%**	N/D	N/D
SRX6893157 20-Sep-2019	Boreoeutheria: **1.24%**	Catarrhini 644546	N/D***
SRX6893156 20-Sep-2019	Manis javanica: **7.51%** Homo sapiens: **0.03%**	Homo sapiens 81948	Pangolin coronavirus 2Kbp
SRX6893155 20-Sep-2019	Homo sapiens: **0.37%**	Homininae 3534150	Pangolin coronavirus 5Kbp
SRX6893154 20-Sep-2019	Homo sapiens: **0.02%**	Hominoidea 356003	Pangolin coronavirus 154Kbp
SRX6893153 20-Sep-2019	Homo sapiens: **0.01%**	Homo sapiens 162180	Pangolin coronavirus 41Kbp

SRX6893152 20-Sep-2019	Manis javanica: **2.87%** Euarchontoglires: **1.37%**	N/D	N/D
SRX6893151 20-Sep-2019	Manis javanica: **7.47%**	N/D	N/D
SRX6893150 20-Sep-2019	Boreoeutheria: **1.91%**	N/D	N/D
SRX6893149 20-Sep-2019	Manis javanica: **1%**	Simiiformes 313069	N/D
SRX6893148 20-Sep-2019	Manis javanica: **0.4%**	Catarrhini 194320	N/D
SRX6893147 20-Sep-2019	Manis javanica: **2.71%**	Catarrhini 69937	N/D
SRX6893146 20-Sep-2019	Boreoeutheria: **1.72%**	Hominoidea 231755	N/D
SRX6893145 20-Sep-2019	Homininae: **0.27%** Manis javanica: **1.01%**	Homininae 2536765	N/D
SRX6893144 20-Sep-2019	Manis javanica: **0.62%**	Hominoidea 166628	N/D
SRX6893143 20-Sep-2019	Manis javanica: **1.63%**	N/D	N/D
SRX6893142 20-Sep-2019	Manis javanica: **1.28%**	Simiiformes 57084	N/D
SRX6893141 20-Sep-2019	Boreoeutheria: **1.41%**	N/D	N/D
SRX6893140 20-Sep-2019	Boreoeutheria: **1.56%**	N/D	N/D
SRX6893139 20-Sep-2019	Homo sapiens: **0.01%**	Homo sapiens 491120	Pangolin coronavirus 2Kbp
SRX6893138 20-Sep-2019	Boreoeutheria: **1.67%**	Homininae 2761176	N/D
SRX7756769 18-Feb-2020	Homo sapiens: **0.03%**	Homo sapiens 5457929	Bat SARS-like coronavirus 2Kbp Wuhan seafood market pneumonia virus 2Kbp
SRX7756766 18-Feb-2020	Manis javanica: **78.6%**	Cercopithecidae 3116	Betacoronavirus 2Kbp**
SRX7756765 18-Feb-2020	Manis javanica: **87.17%**	Cercopithecinae 11339	N/D****
SRX7756764 18-Feb-2020	Manis javanica: **48.39%**	Cercopithecidae 22600	N/D
SRX7756763 18-Feb-2020	Manis javanica: **94.95%**	Cercopithecidae 5076	N/D

SRX7756762 18-Feb-2020	Manis javanica: **95.37%**	Catarrhini* 2831	Nidovirales 0Kbp
SRX7756761 18-Feb-2020	Manis javanica: **13.63%**	Chlorocebus sabaeus 498506	N/D
SRX7732094 15-Feb-2020	N/A***	N/A	Pangolin coronavirus***

*: Chlorocebus Sabaeus

**:Not claimed as being SARS-CoV-2 related in the original publication. Likely unrelated.

***Not analyzable. All Non-Coronavirus data filtered out. Leaving only 2,633 reads, all of which can be mapped to the SARS-CoV-2 reference genome.

****8 reads as claimed by [10]

Specific BLAST analysis

In order to determine the authenticity of the Primate-related reads in the datasets, Specific BLAST analysis is carried out for all datasets that possessed claimed or analyzed reads of coronaviridae-related viruses. An 100% full-length match that does not map to non-primates confirms Authenticity of read.

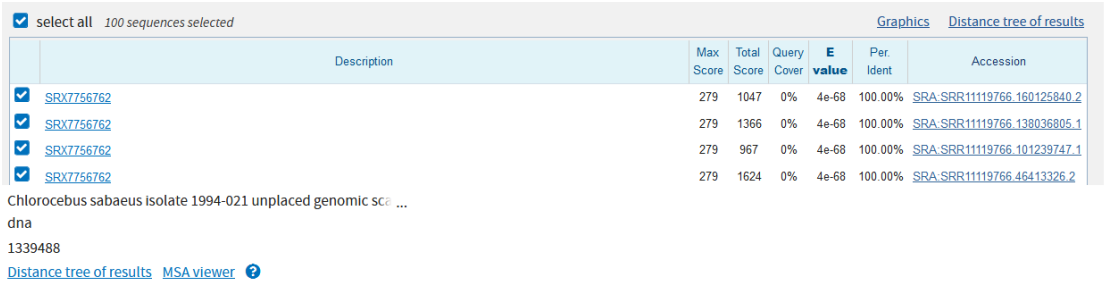


Fig.2a Specific BLAST analysis on the PRJNA607174 dataset, SRX7756762 ,that contained claimed SARS-CoV-2 related coronavirus reads. The 100% full-length matches clearly indicate presence of Primate-derived material.

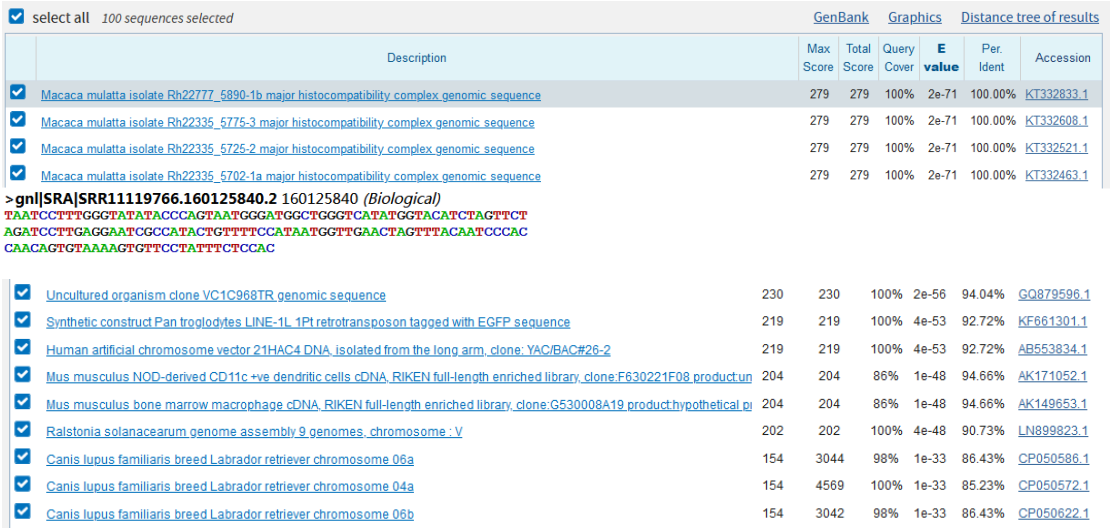
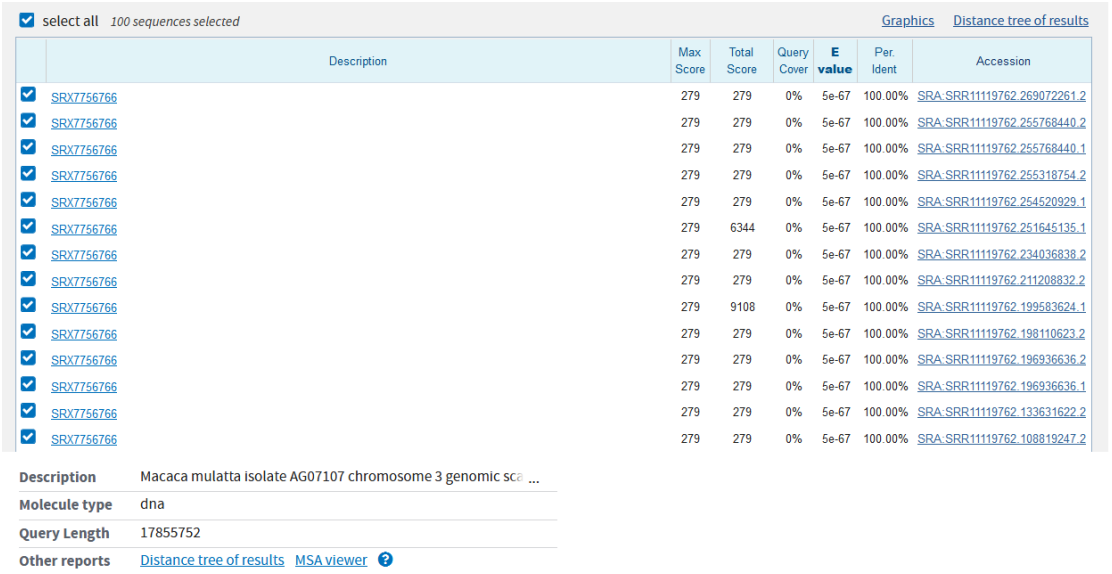


Fig.2b BLAST result on the returned sequence revealed it as a Primate-derived MHC complex gene that is not found in non-primates, confirming Primate origin.

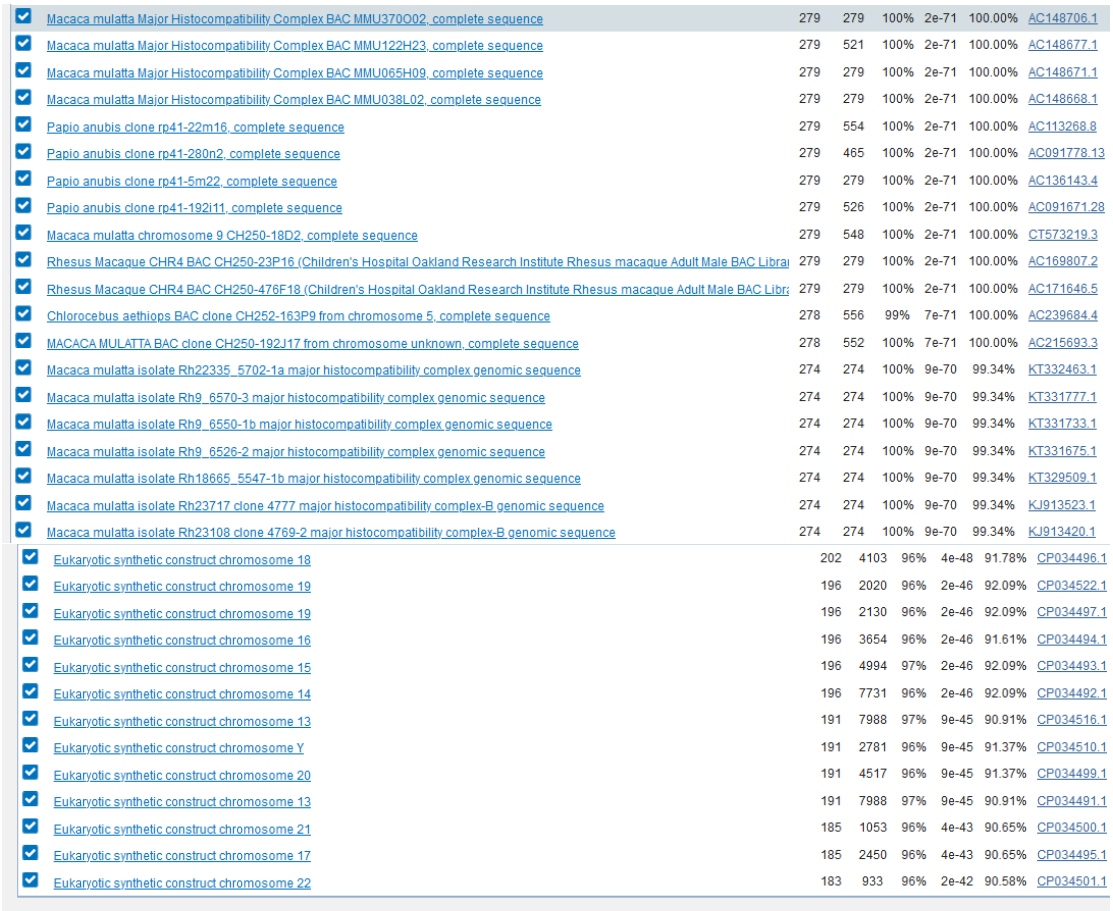


select all *100 sequences selected* Graphics Distance tree of results

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
SRX7756766	279	279	0%	5e-67	100.00%	SRA:SRR11119762.269072261.2
SRX7756766	279	279	0%	5e-67	100.00%	SRA:SRR11119762.255768440.2
SRX7756766	279	279	0%	5e-67	100.00%	SRA:SRR11119762.255768440.1
SRX7756766	279	279	0%	5e-67	100.00%	SRA:SRR11119762.255318754.2
SRX7756766	279	279	0%	5e-67	100.00%	SRA:SRR11119762.254520929.1
SRX7756766	279	6344	0%	5e-67	100.00%	SRA:SRR11119762.251645135.1
SRX7756766	279	279	0%	5e-67	100.00%	SRA:SRR11119762.234036838.2
SRX7756766	279	279	0%	5e-67	100.00%	SRA:SRR11119762.211208832.2
SRX7756766	279	9108	0%	5e-67	100.00%	SRA:SRR11119762.199583624.1
SRX7756766	279	279	0%	5e-67	100.00%	SRA:SRR11119762.198110623.2
SRX7756766	279	279	0%	5e-67	100.00%	SRA:SRR11119762.196936636.2
SRX7756766	279	279	0%	5e-67	100.00%	SRA:SRR11119762.196936636.1
SRX7756766	279	279	0%	5e-67	100.00%	SRA:SRR11119762.133631622.2
SRX7756766	279	279	0%	5e-67	100.00%	SRA:SRR11119762.108819247.2

Description	Macaca mulatta isolate AG07107 chromosome 3 genomic sc ...
Molecule type	dna
Query Length	17855752
Other reports	Distance tree of results MSA viewer

Fig.3a Specific BLAST analysis of **SRX7756766** revealed large amount of 100% full-length matches with Macaca Mulatta.



Macaca mulatta Major Histocompatibility Complex BAC MMU370O02, complete sequence	279	279	100%	2e-71	100.00%	AC148706.1
Macaca mulatta Major Histocompatibility Complex BAC MMU122H23, complete sequence	279	521	100%	2e-71	100.00%	AC148677.1
Macaca mulatta Major Histocompatibility Complex BAC MMU065H09, complete sequence	279	279	100%	2e-71	100.00%	AC148671.1
Macaca mulatta Major Histocompatibility Complex BAC MMU038L02, complete sequence	279	279	100%	2e-71	100.00%	AC148668.1
Papio anubis clone rp41-22m16, complete sequence	279	554	100%	2e-71	100.00%	AC113268.8
Papio anubis clone rp41-280n2, complete sequence	279	465	100%	2e-71	100.00%	AC091778.13
Papio anubis clone rp41-5m22, complete sequence	279	279	100%	2e-71	100.00%	AC136143.4
Papio anubis clone rp41-192i11, complete sequence	279	526	100%	2e-71	100.00%	AC091671.28
Macaca mulatta chromosome 9 CH250-18D2, complete sequence	279	548	100%	2e-71	100.00%	CT573219.3
Rhesus Macaque CHR4 BAC CH250-23P16 (Children's Hospital Oakland Research Institute Rhesus macaque Adult Male BAC Libra	279	279	100%	2e-71	100.00%	AC169807.2
Rhesus Macaque CHR4 BAC CH250-476F18 (Children's Hospital Oakland Research Institute Rhesus macaque Adult Male BAC Libr	279	279	100%	2e-71	100.00%	AC171646.5
Chlorocebus aethiops BAC clone CH252-163P9 from chromosome 5, complete sequence	278	556	99%	7e-71	100.00%	AC239684.4
MACACA MULATTA BAC clone CH250-192J17 from chromosome unknown, complete sequence	278	552	100%	7e-71	100.00%	AC215693.3
Macaca mulatta isolate Rh22335_5702-1a major histocompatibility complex genomic sequence	274	274	100%	9e-70	99.34%	KT332463.1
Macaca mulatta isolate Rh9_6570-3 major histocompatibility complex genomic sequence	274	274	100%	9e-70	99.34%	KT331777.1
Macaca mulatta isolate Rh9_6550-1b major histocompatibility complex genomic sequence	274	274	100%	9e-70	99.34%	KT331733.1
Macaca mulatta isolate Rh9_6526-2 major histocompatibility complex genomic sequence	274	274	100%	9e-70	99.34%	KT331675.1
Macaca mulatta isolate Rh18665_5547-1b major histocompatibility complex genomic sequence	274	274	100%	9e-70	99.34%	KT329509.1
Macaca mulatta isolate Rh23717 clone 4777 major histocompatibility complex-B genomic sequence	274	274	100%	9e-70	99.34%	KJ913523.1
Macaca mulatta isolate Rh23108 clone 4769-2 major histocompatibility complex-B genomic sequence	274	274	100%	9e-70	99.34%	KJ913420.1
Eukaryotic synthetic construct chromosome 18	202	4103	96%	4e-48	91.78%	CP034496.1
Eukaryotic synthetic construct chromosome 19	196	2020	96%	2e-46	92.09%	CP034522.1
Eukaryotic synthetic construct chromosome 19	196	2130	96%	2e-46	92.09%	CP034497.1
Eukaryotic synthetic construct chromosome 16	196	3654	96%	2e-46	91.61%	CP034494.1
Eukaryotic synthetic construct chromosome 15	196	4994	97%	2e-46	92.09%	CP034493.1
Eukaryotic synthetic construct chromosome 14	196	7731	96%	2e-46	92.09%	CP034492.1
Eukaryotic synthetic construct chromosome 13	191	7988	97%	9e-45	90.91%	CP034516.1
Eukaryotic synthetic construct chromosome Y	191	2781	96%	9e-45	91.37%	CP034510.1
Eukaryotic synthetic construct chromosome 20	191	4517	96%	9e-45	91.37%	CP034499.1
Eukaryotic synthetic construct chromosome 13	191	7988	97%	9e-45	90.91%	CP034491.1
Eukaryotic synthetic construct chromosome 21	185	1053	96%	4e-43	90.65%	CP034500.1
Eukaryotic synthetic construct chromosome 17	185	2450	96%	4e-43	90.65%	CP034495.1
Eukaryotic synthetic construct chromosome 22	183	933	96%	2e-42	90.58%	CP034501.1

Fig.3b BLASTing such matches gives 1005 matches to only Primates, and with no matches outside of Primates. This indicates that **SRX7756766** also contained significant amount of material derived from primates.

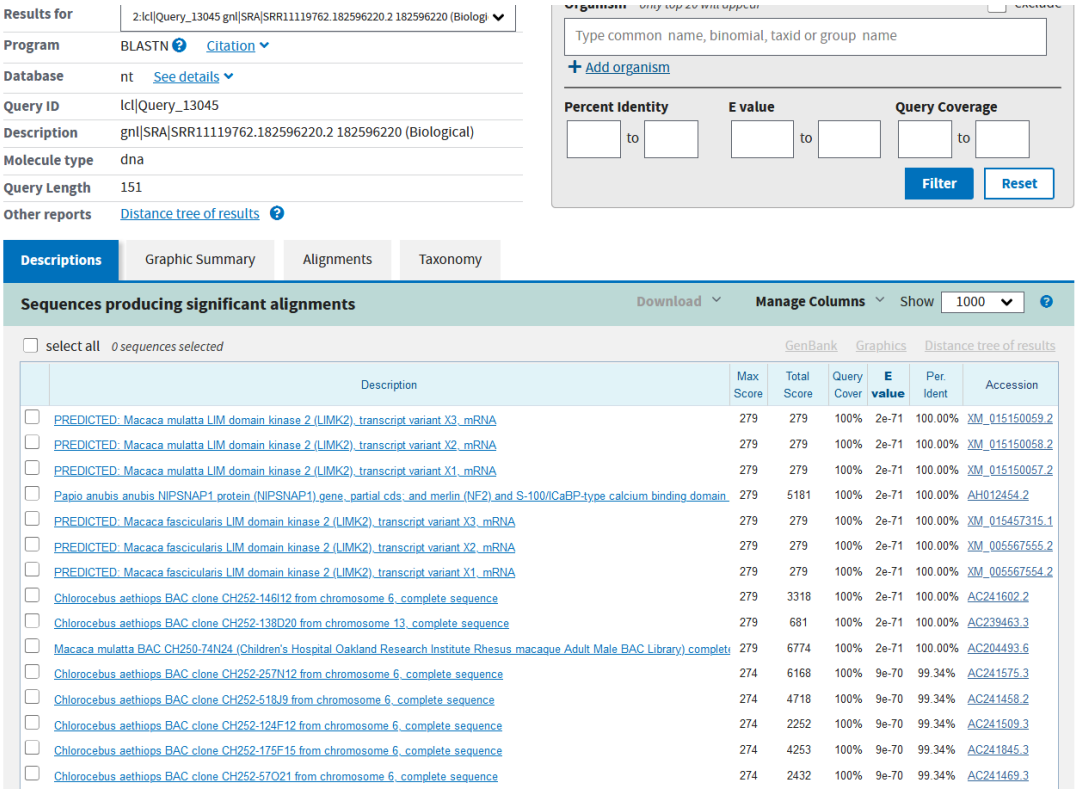


Fig.3C Presence of Primate-derived mRNA reads in **SRX7756766** confirms the Primate origin of these reads.

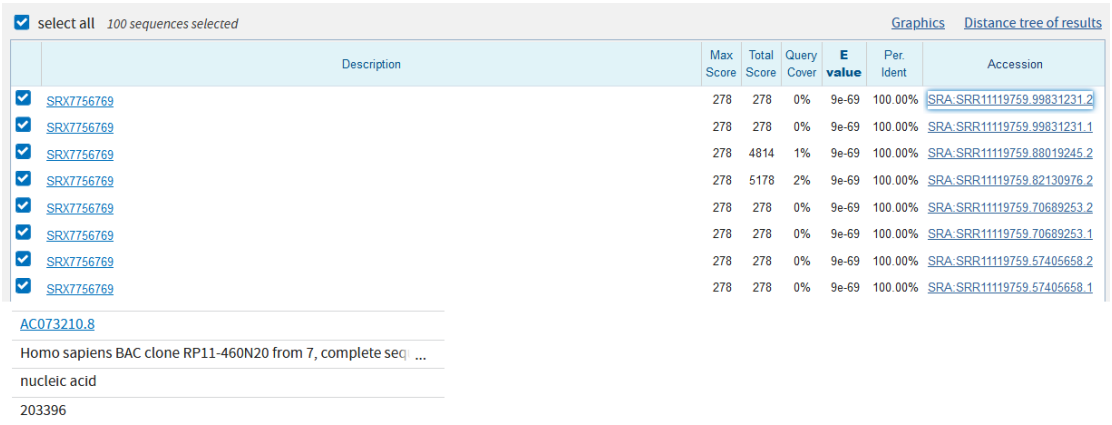
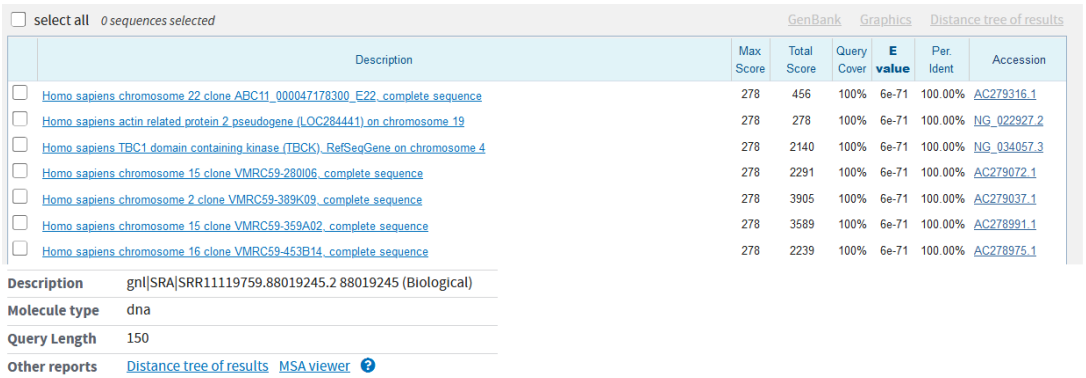


Fig.4a Similarly, **SRX7756769** contained large amount of reads that are 100% full-length matches to Human genomic DNA.



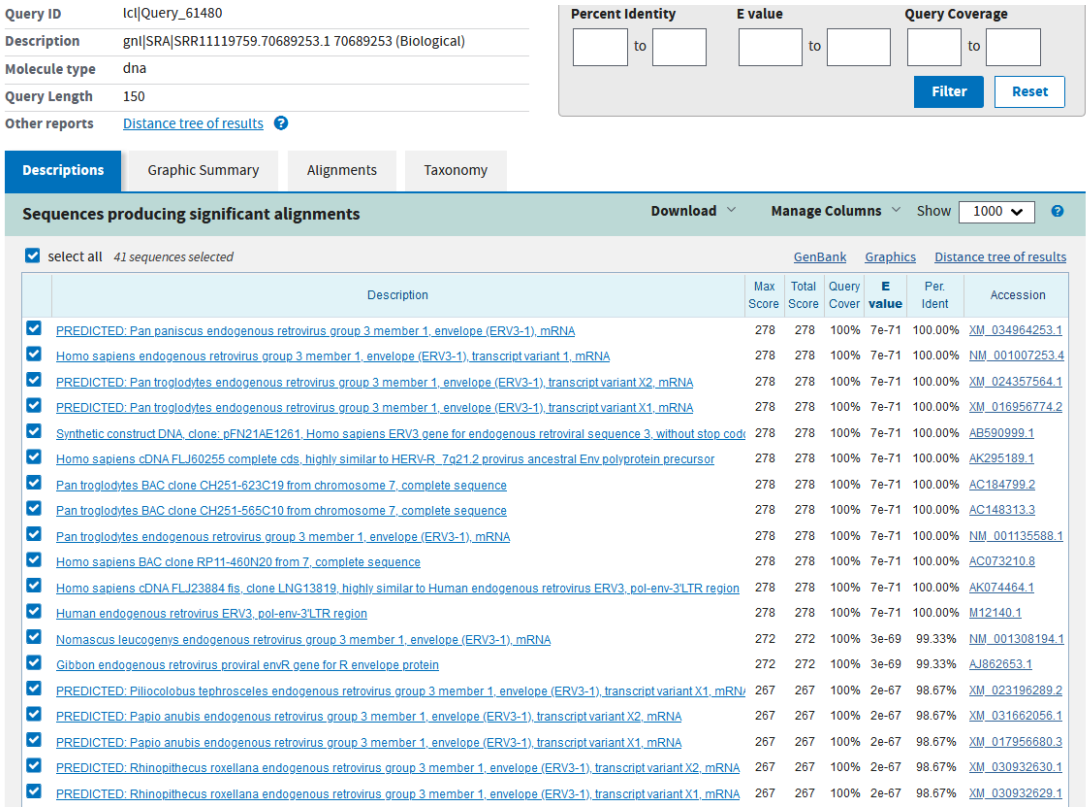


Fig.4b A BLAST analysis on reads sampled from the 100% hit results confirmed that it was found only in humans. Once again confirming human origin.

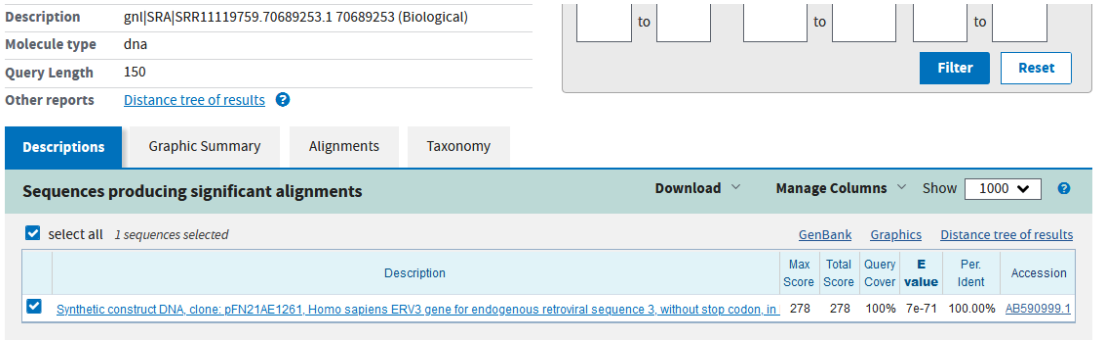


Fig.4c The sequence have no matches outside of Primates.

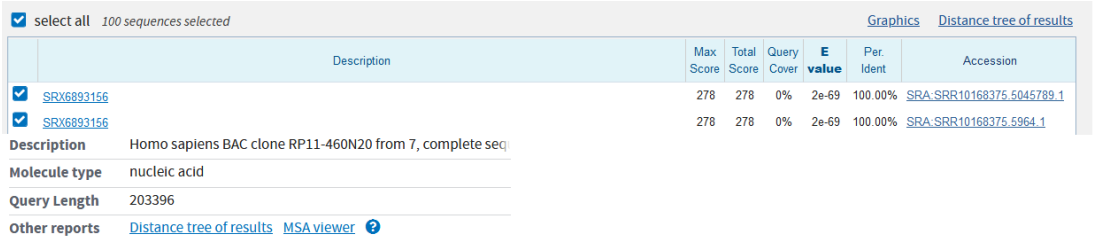


Fig.5a SRX6893156 also returned 100% matched results from the human Genome.

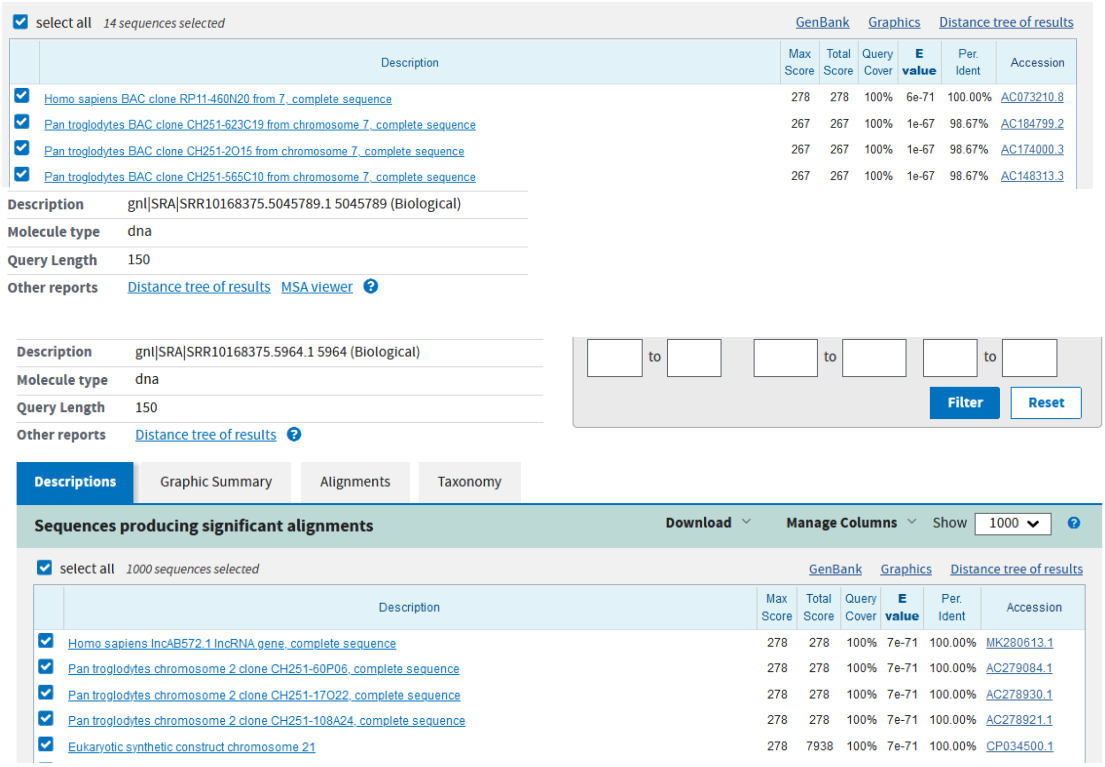


Fig.5b BLAST search on the result returned 100% match only found in humans, confirming origin in human-derived material.

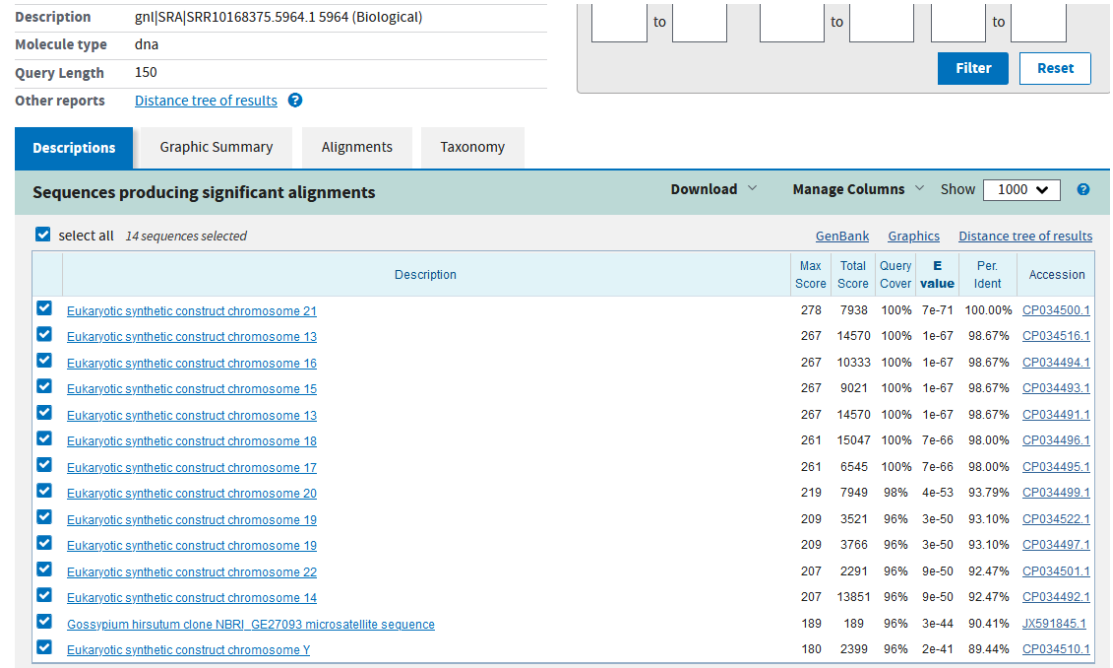


Fig.5c BLAST result of the sequences in question revealed that it is not found outside of Primates.

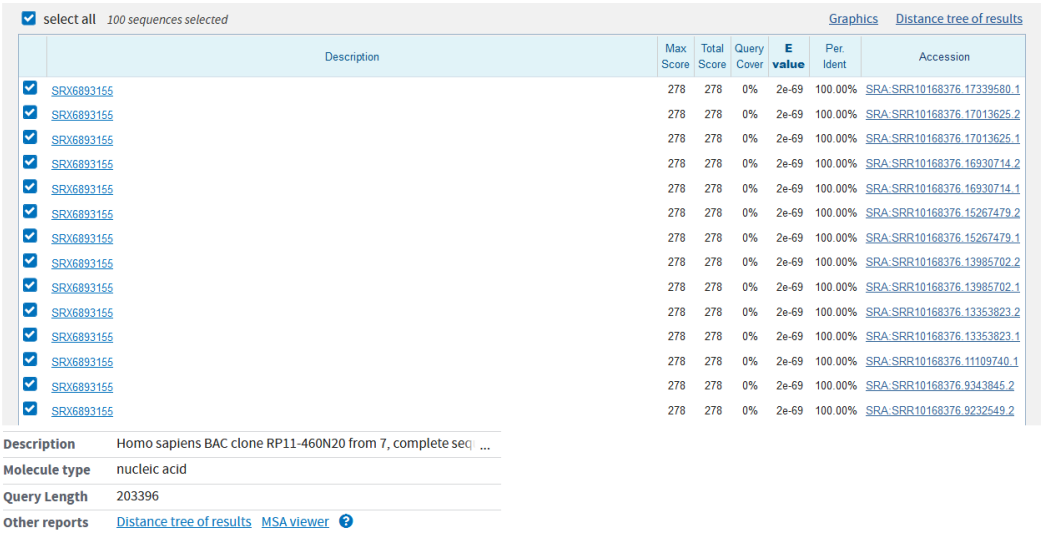
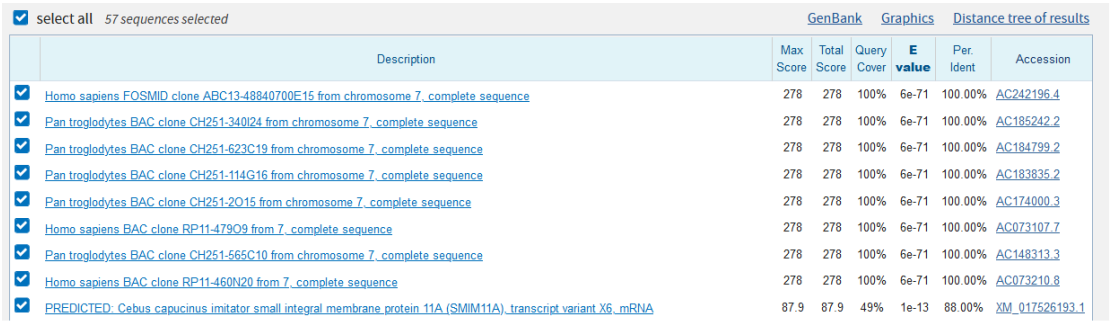
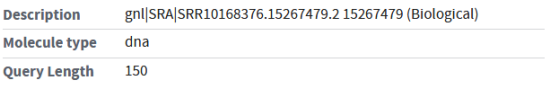
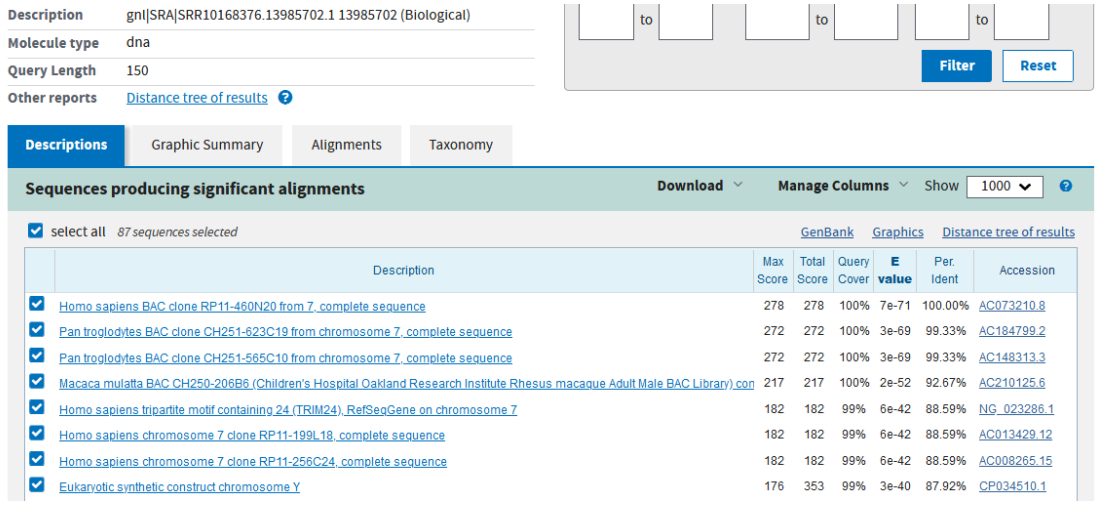


Fig.6a Similarly, BLAST research on SRX6893155 gives large number of full length 100% matches to the human genome.







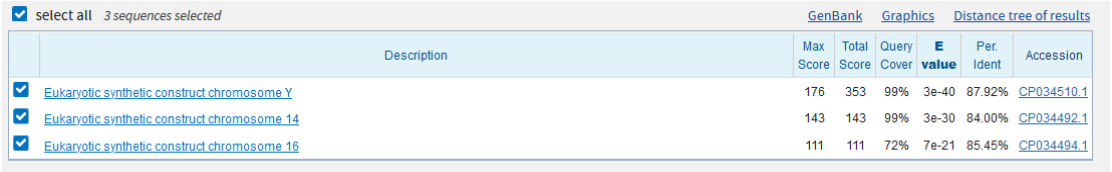


Fig.6b The results, when put through BLAST, confirms that the 100% matches are in fact derived from a Hominid origin.

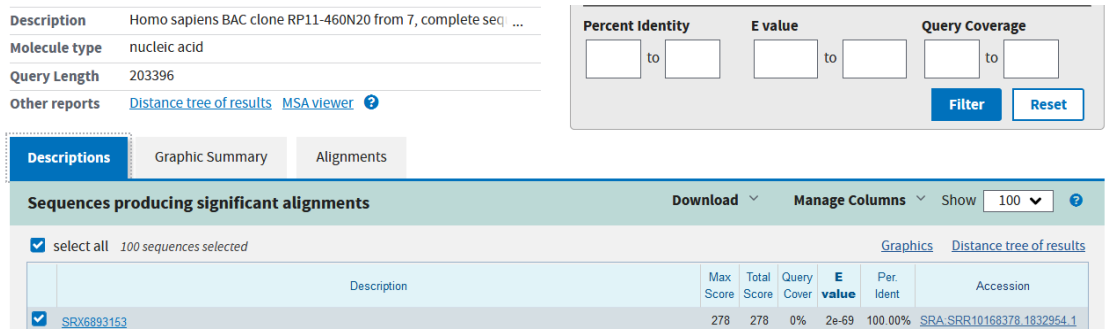


Fig.7a SRX6893153 have also returned 100% match full-length read on this tiny part of the human genome.

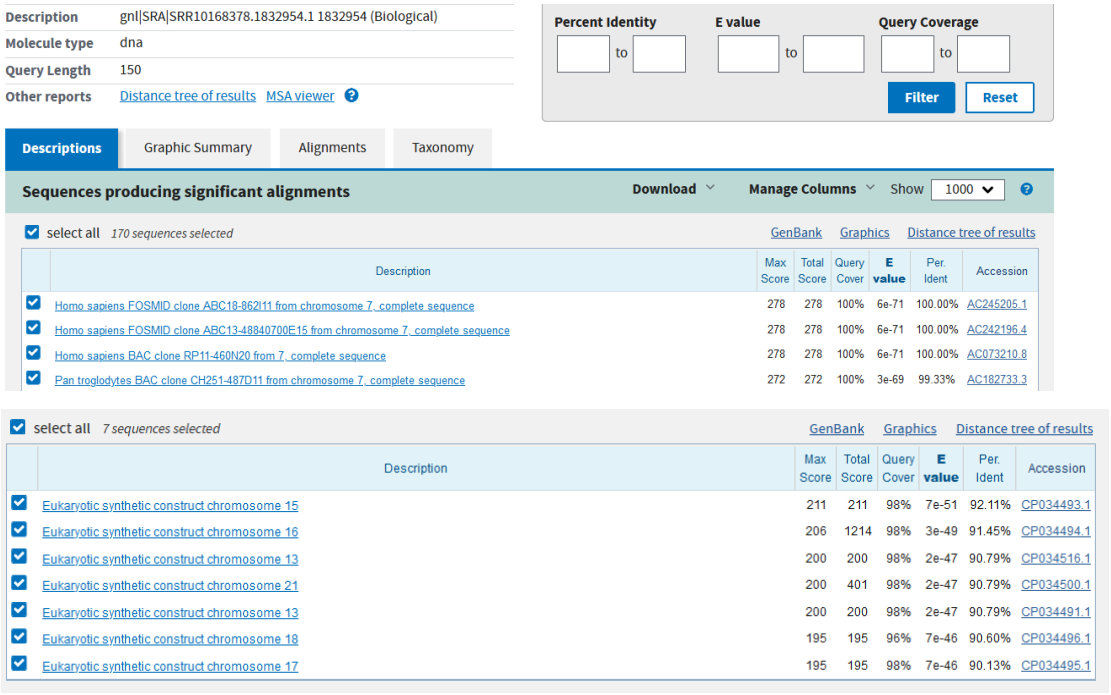


Fig.7b similarly, the read is only found in humans—indicating the Homo Sapiens Trace result is accurate.

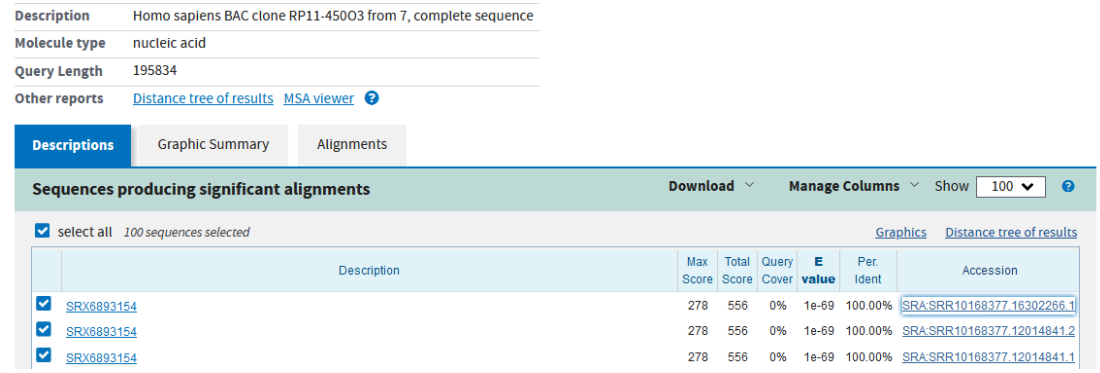


Fig.8a Reads from the Human PMS1 gene is recovered from SRX6893154 with a query sequence only 195834bp in length.

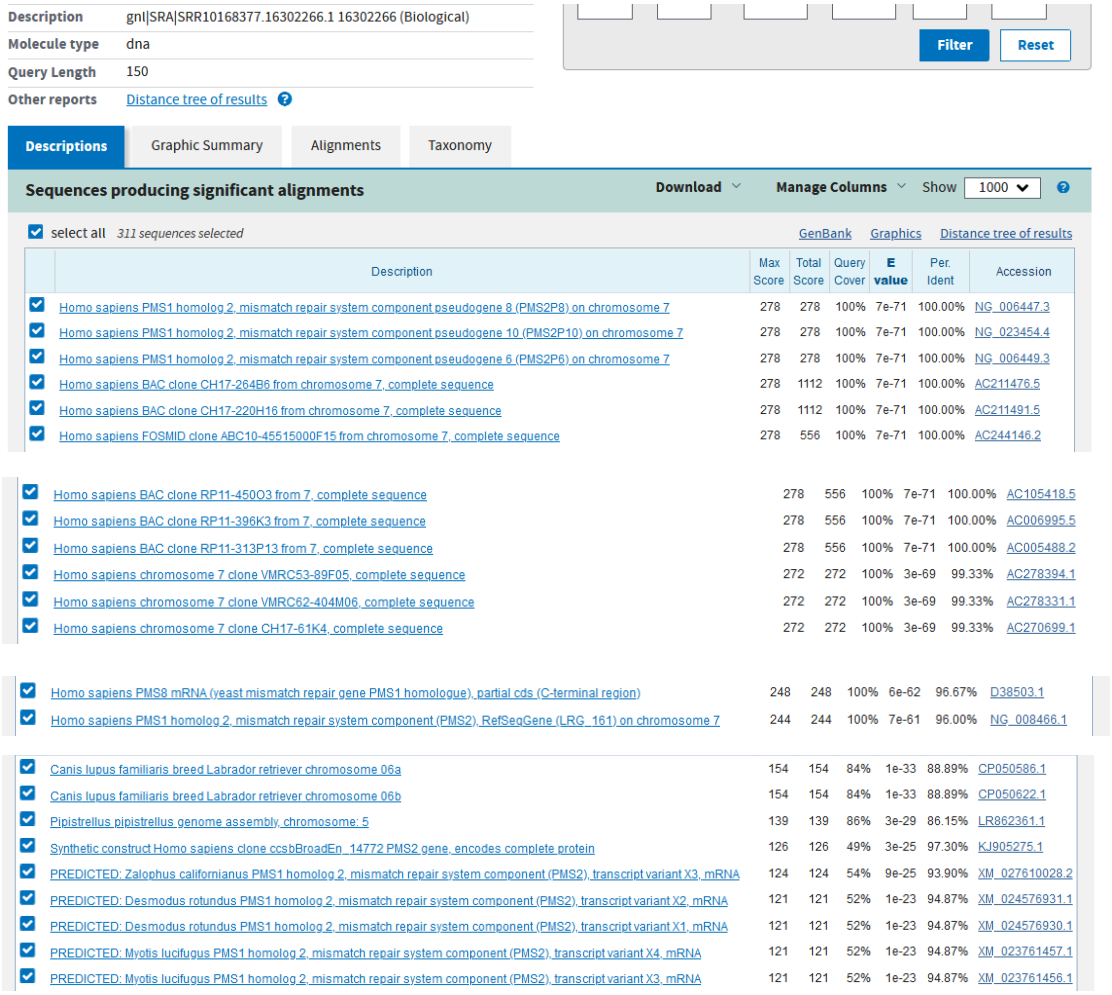


Fig.8b This PMS1 read is only found in Humans. This is clearly a contaminant from a hominid origin.

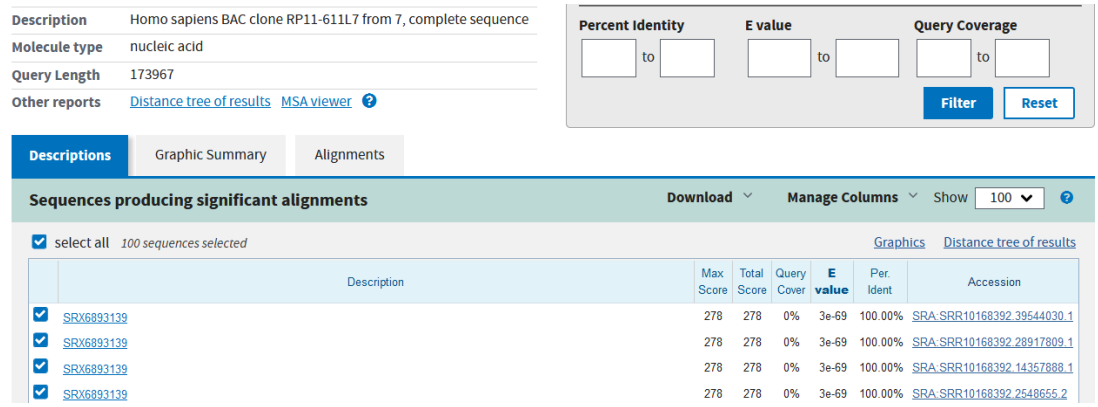


Fig.9a similarly, multiple 100% match Full length reads were obtained from **SRX6893139.** As this query sequence is only 173967 nucleotides in length, the real extent of Human-derived contamination is also extremely severe.

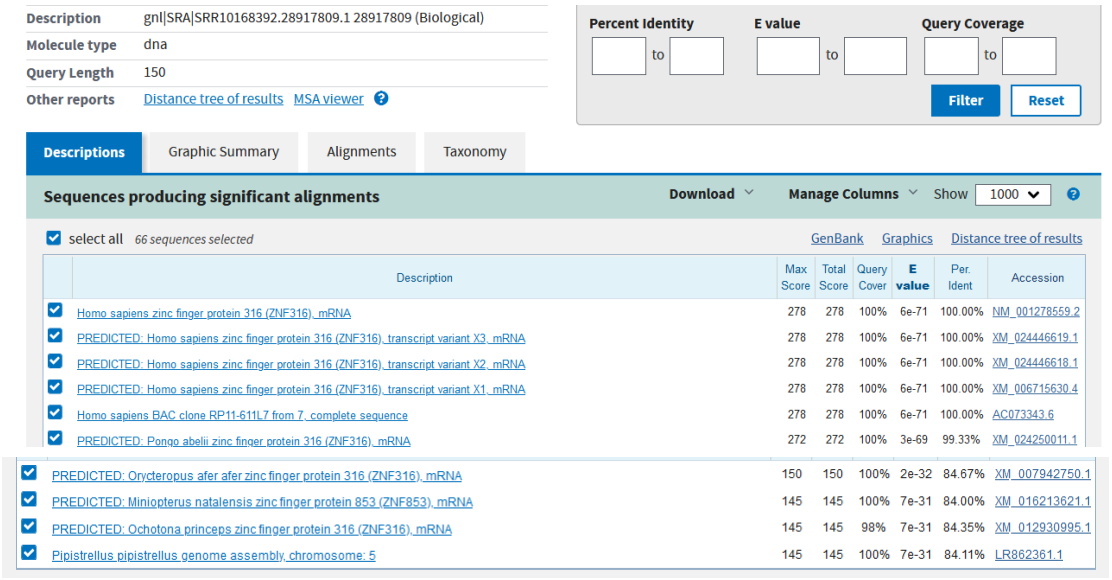


Fig.9b examining these reads revealed that they are only found in humans and apes. This is therefore also clear evidence that there is Human/Hominid-derived contamination in **SRX6893139**.

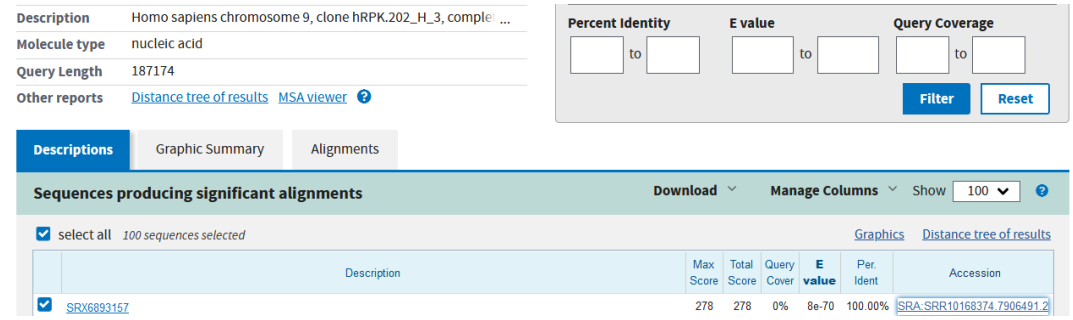


Fig.10a one read is also recovered from **SRX6893157**, from a query sequence only 187174nt in length.

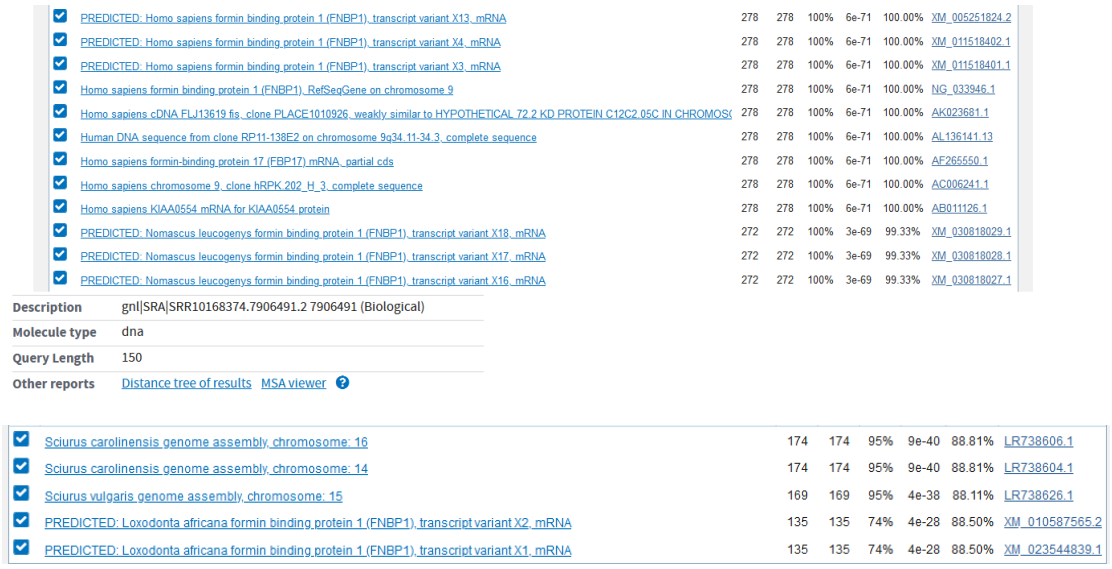


Fig.10b this particular sequence is only found in humans—indicating that even the **SRX6893157** dataset was contaminated by material of human origin.

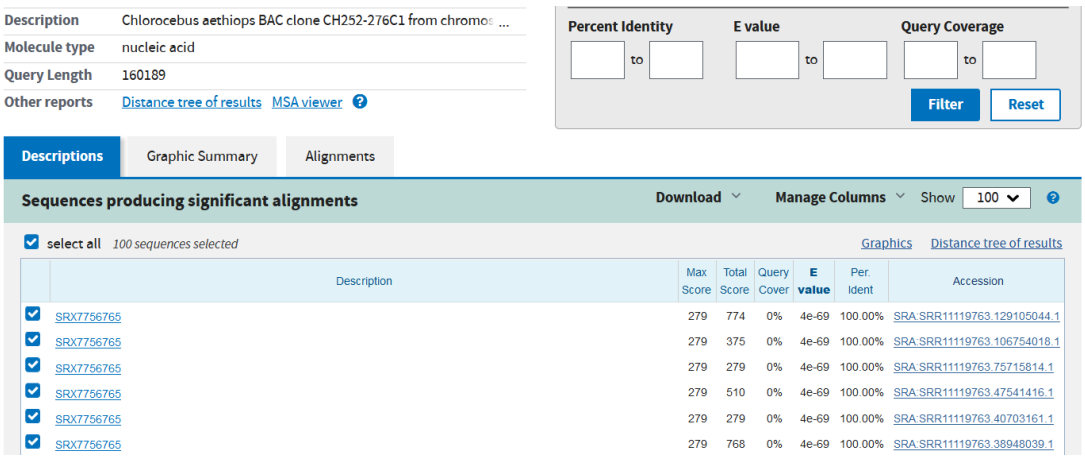


Fig.11a The presence of Reads from Somatic Chlorocebus aethiops in **SRX7756765** confirms the identity of the Cercopithecinae reads there.

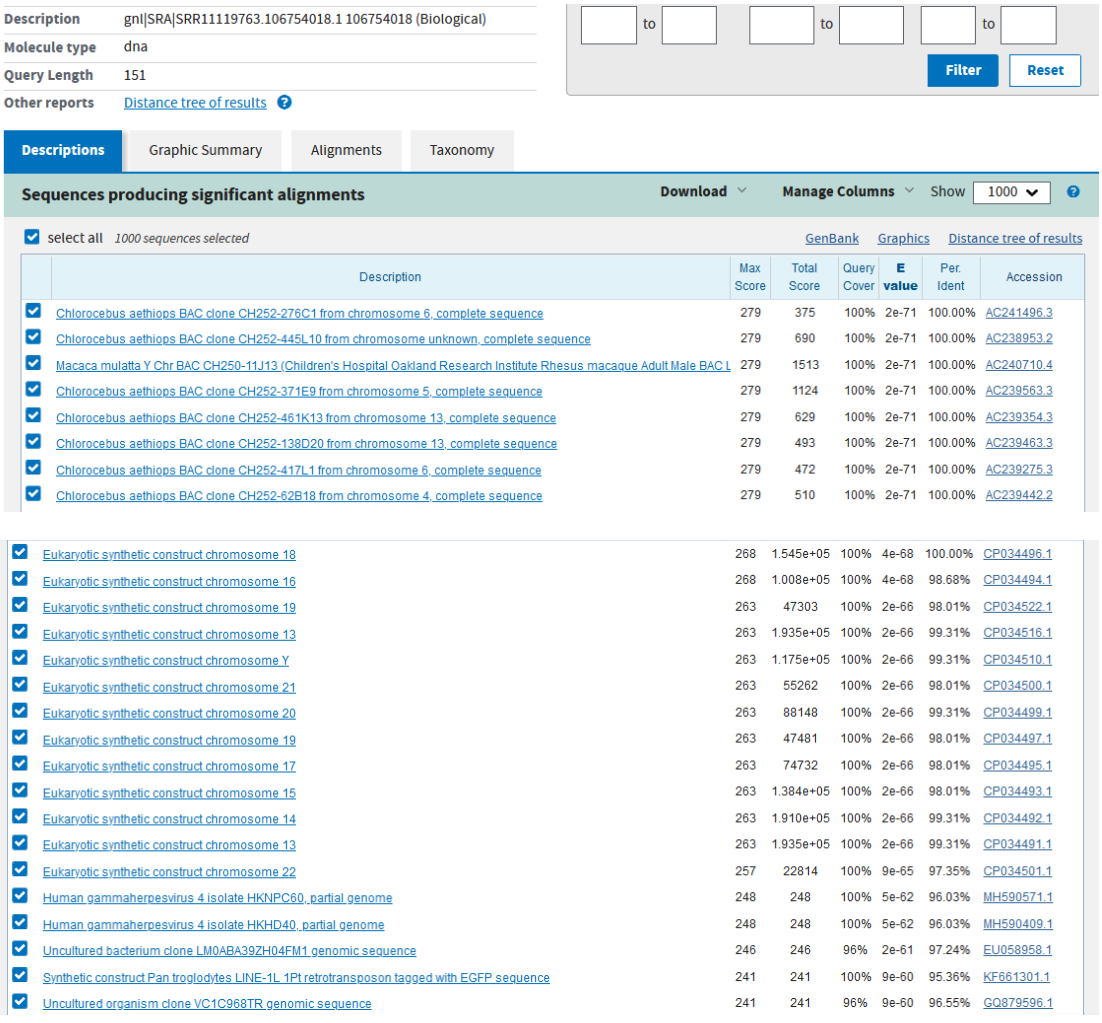


Fig.11b the sequences from the BLAST hits indicate that they were unique to the family Cercopithecinae. Confirming Primate origin.

Analyzing the extent of contamination.

As the Specific BLAST analysis confirmed significant level of Human-derived contamination in all samples positive for SARS-CoV-2 related Coronaviruses, The TRACE result can therefore be

trusted for the analysis on the extent of contamination.

The 32nt Krona Trace system is used for elucidating the ratio of different taxa within a sample. As Specific BLAST analysis confirmed the significant presence of Human and Primate derived Genetic material--The most basal group of primates detected in all Coronavirus-positive samples belong to Catarrhini—or Humans, Apes and Old-World Monkeys. Therefore, Trace classification results that can be classified into sister nodes of Catarrhini should be considered as Contamination by Primate-derived material.

Since Catarrhini is under Simiiformes; Haplorrhini; Primates; Euarchonta; Euarchontoglires and Manis is under Pholidota; Laurasiatheria, If a read is TRACEd down to Catarrhini, it can not be from a Pangolin, and it will have to be from a Primate-derived source—Contamination by material from the lab.

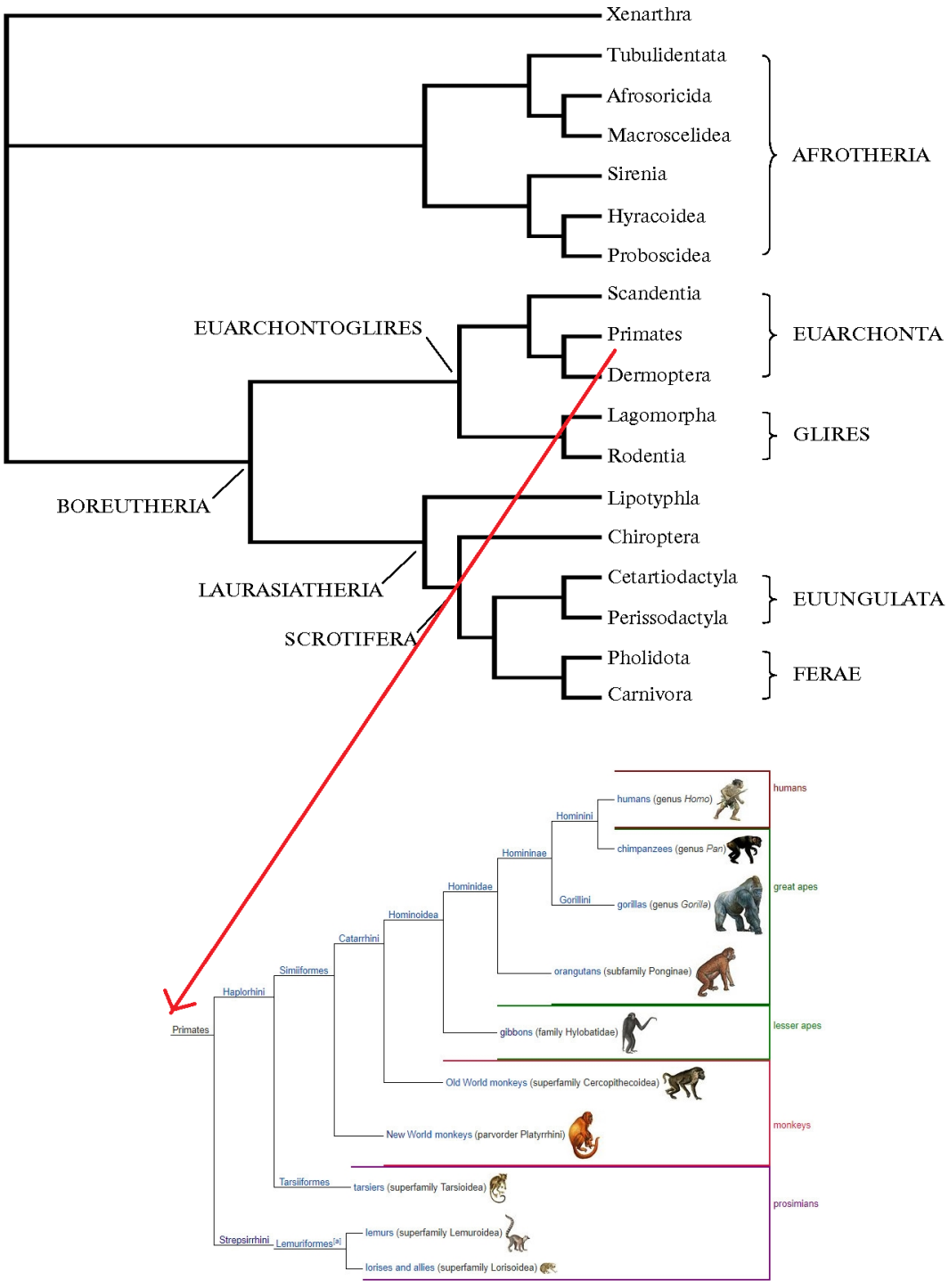


Fig. 12 Family tree of mammals, Including the position and classification of Primates in the lineage of Mammalia.

Table 3a Ratios of Hominid-traced reads to Pangolin-traced reads in the SRA datasets that contained reads of the GD- Pangolin-CoV sequence, and had Hominid reads.

Accession and date	Primate classification and total traced Kbps	Total traced Kbps to Manis Javanica (Pangolin)	Ratio of Primate to Pangolin	Virus classification and amount of reads by Kbps
SRX7756769 18-Feb-2020	Homo sapiens 5457929	15401134	0.35	Bat SARS-like coronavirus 2Kbp Wuhan seafood market pneumonia virus 2Kbp
SRX6893139 20-Sep-2019	Homo sapiens 491120	5301351	0.0926	Pangolin coronavirus 2Kbp
SRX6893157 20-Sep-2019	Catarrhini 644546	1889448	0.34	N/D***
SRX6893156 20-Sep-2019	Homo sapiens 81948	4765461	0.01719	Pangolin coronavirus 2Kbp
SRX6893155 20-Sep-2019	Homininae 3534150	525801	6.7214	Pangolin coronavirus 5Kbp
SRX6893154 20-Sep-2019	Hominoidea 356003	2232008	0.159	Pangolin coronavirus 154Kbp
SRX6893153 20-Sep-2019	Homo sapiens 162180	3110158	0.05214	Pangolin coronavirus 41Kbp

***: No trace result on Coronaviruses, despite claimed reads from [3]

Table 3b Ratios of Primate-traced reads to Coronavirus-traced reads in the SRA datasets that contained reads claimed to be traced to of the GD- Pangolin-CoV sequence, and lacked Hominid reads.

Accession and date	Primate classification and reads (in Kbp)	Virus classification and reads	Ratio of virus reads to Primate reads
SRX7756766 18-Feb-2020	Cercopithecidae 3116; BLAST to Macaca Mulatta	Betacoronavirus 2Kbp **	0.000642
SRX7756762 18-Feb-2020	Catarrhini 2831; BLAST to Chlorocebus sabaeus	Nidovirales 0Kbp Claimed 10x150bp reads	0.000530
SRX7756765 22-Apr-2020	Cercopithecinae 11339 BLAST to Chlorocebus Aethiops	N/D***	N/A
SRX7732094 15-Feb-2020	N/A*	Pangolin coronavirus	N/A*

*: No non-coronavirus reads available in the dataset with a total of 2,633 reads, making analysis impossible.

**: No claimed reads from [2]

***: Claimed 8 reads from [10]

DISCUSSIONS

The extent of contamination in the pangolin sequencing datasets

As the samples were supposed to be pangolin lung tissue, which will neither contact with nor be contaminated by non-pangolin derived mammalian tissues when still inside the animal, any non-pangolin mammalian reads within such a dataset can only be introduced to the sequencing process after the sample itself have been taken and brought into a lab.

As the classification Catarrhini itself is phylogenetically very deep down the Primate line which is itself distinguished from the Pangolin line at a very basal node (Boreoeutheria), and since we have already confirmed that the Primate line in PRJNA573298 traces mostly to humans by using Specific BLAST analysis, (**SRX6893157**, the only one of the claimed coronavirus read dataset that gives a classification just down to Catarrhini, contained 213 full length 100% matches to the Human Mitochondrial reference genome alone, which is only 16569 bp in length. All other datasets gives definitive TRACE mapping to Homo Sapiens and contained distinct 100% matched reads to even very small parts of the Human genome.), We can deduce the extent of contamination of the PRJNA573298 dataset by Primate-related materials as from a minimum of 1.6% to as high as 87% by sample mass—using the ratio of Primate reads to Pangolin reads on TRACE. Such high level of contamination with Primate-derived material is unacceptable for a sample that was supposed to be Lung tissue. And therefore, the virome data of such samples in PRJNA573298 no longer reflects the original virome of the animal, and an potential “novel” reads from these contaminated samples may have been from in-lab contamination instead.

Deducing the dynamic of contamination in PRJNA607174

Of all 7 PRJNA607174 datasets, only **SRX7756769** and **SRX7756762** is claimed by Xiao et. Al to contain SARS-CoV-2-like reads. However, TRACE results revealed low level of contamination by Cercopithecidae (Old World Monkey) reads across all the samples. In particular, the **SRX7756762** dataset contained definitive mappings to Chlorocebus sabaeus, or African Green Monkey, while **SRX7756766** which contained 2Kbp unclaimed reads of Betacoronaviruses on TRACE, contained 100% full-length definitive mappings to Macaca Mulatta that may also be mapped to Chlorocebus Aethiops and Homo Sapiens.

SRX7756769 genetically resembles other samples in PRJNA573298, in both the kind of contamination and the extent of contamination. It contained an large excess of homo sapiens reads in levels similar to the contaminated samples in PRJNA573298.

From the method section of Lam et.al, we knew that they have performed Virus isolation using VERO E6 cells—Species Chlorocebus Sabaeus on one of the samples that have a positive PCR test for coronaviruses. The low level of contamination by Cercopithecidae-related reads in all the samples in PRJNA607174 except for **SRX7756769** itself support the possibility that **SRX7756769** is the first sample to be sequenced, and it happens before the lab begun using VERO E6 cells in the experiment. They then isolated the virus from the contaminated **SRX7756769** in VERO E6 cells, characterized it but did not sequence it, and this cell culture material then contaminated **SRX7756762** and possibly **SRX7756766**, resulting the 10 reads in **SRX7756762** and the 2Kb Batacoronavirus reads in **SRX7756766.**

The exact nature of SRX7732094 needs to be further scrutinized.

The P2S dataset, SRX7732094, displays very unusual property when compared to other Datasets under the same BioProject. It is the only dataset with all Non-coronavirus reads being filtered out, and contained too little spots for it to be an ILLUMINA NextSeq 550 run. Furthermore, it was the only dataset that did not contain metadata with either an isolation source or a Library prep procedure, other than "This dataset contains coronavirus-like sequence reads, based on BLAST search."

Such a strange designation and the fact of the dataset being heavily filtered, Raises problems on whether such a dataset is an actual BioSample at all. If this sample is really as claimed by Lam et. Al, Why the dataset have to be put through such heavy filtering when the other sequencing runs was clearly not filtered as severely as this dataset? Why there was no BioSample metadata on either Biomaterial provider, Source Tissue or Collector when all other Sequencing runs clearly provided such metadata information?

Unless the complete, unfiltered sequencing reads are made available on **SRX7732094**, and the rest of **PRJNA606875**, this Dataset can not be considered to be a real, reliable sample, and it must be excluded as "evidence" of a SARS-CoV-2-like virus infecting

pangolins in Guangdong, 2019.

Table 4 Sequencing runs in PRJNA696875, Accession number, BioSample, Content and designation

Accession number and date	Size	Non-Coronavirus reads?	Source Tissue Provider and Collected by	Virus Designation: GD or GX?	Design
SRX7732094 15-Feb-2020	2,633	No	N/A	GD	This dataset contains coronavirus-like sequence reads, based on BLAST search.
SRX7732093 15-Feb-2020	470,344	Yes	Intestine Yanling Hu Wuchun Cao	GX	NEBNext Ultra II DNA Library Prep Kit, paired sequencing data has been integrated.
SRX7732092 15-Feb-2020	340,661	Yes	Lung Yanling Hu Wuchun Cao	GX	NEBNext Ultra II DNA Library Prep Kit, paired sequencing data has been integrated.
SRX7732091 15-Feb-2020	416,659	Yes	Intestine Yanling Hu Wuchun Cao	GX	NEBNext Ultra II DNA Library Prep Kit, paired sequencing data has been integrated.
SRX7732090 15-Feb-2020	520,254	Yes	Lung Yanling Hu Wuchun Cao	GX	NEBNext Ultra II DNA Library Prep Kit, paired sequencing data has been

					integrated.
SRX7732089 15-Feb-2020	19,607,536	Yes	Blood Yanling Hu Wuchun Cao	GX	Ion Total RNA-Seq Kit v2
SRX7732088 15-Feb-2020	4,550,437	Yes	lung and intestine Yanling Hu Wuchun Cao	GX	Ion Total RNA-Seq Kit v2

By closely examining the P2V dataset, SRX7732088, which claimed to be a culture sample in VERO E6 cells, Chlorocebus Sabaeus, the exact viral load in-culture when compared to Cellular mRNA can be deduced by dividing the total identifiable coronavirus signal to the total identifiable Primate signal within the dataset, 6943Kbp/451932Kbp, which correspond to 0.01536:1 Viral RNA to Cellular RNA.

This places the viral loads on the other datasets with Coronavirus-like reads from GD well within the threshold expected from cell culture contamination of the sequencing samples—including the samples in PRJNA607174.

Potential breach of data availability statement by Xiao et al.[2]

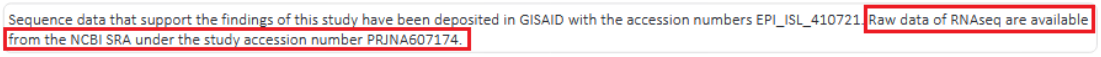


Fig 13. The Data Availability Statement of Xiao et al.

In the Data availability statement, the “Raw data of RNAseq” are clearly stated to be deposited under PRJNA607174. However, only 2 of the “Extended Data Table S3” datasets actually matches the datasets deposited on PRJNA607174. The other 7 datasets were completely unavailable. And the actual deposited datasets on PRJNA607174 does not match what have been claimed by Extended Data Table S3. As the RNA-seq Raw data was stated to be available within PRJNA607174, the failure to publish all the claimed data constitute a breach of the Data Availability statement on the article. Unless such datasets are published and independently examined, All such claimed reads from the strangely unpublished datasets can not be trusted as evidence of a SARS-CoV-2-like virus infecting pangolins in GuangDong, 2019.

Identifying the Etiological agent of the GuangDong 2019 incident.

By using an approach of both SRA TRACE analysis and specific BLAST Analysis, We have uncovered the fact that all samples that does not Contain confirmed Human-derived material, also lacked Claimed reads of a SARS-CoV-2 like virus that can be confirmed using NCBI Trace. All samples with claimed or traced reads of Coronaviruses in general, contained confirmed primate reads with the lowest common phylogenetic node Catarrhini. Samples that does not give a TRACE result on primate-derived material all lacked identifiable or claimed coronavirus reads.

This strongly imply that the Coronavirus-like reads are associated with human/Primate-sourced contamination material.

Most importantly, of all dead pangolins being sampled in the studies, only 9 out of a total of 29

Analyzable samples/datasets contained TRACEd or Claimed Coronavirus reads—despite all dead pangolins displayed similar symptoms in captivity. This imply that the alleged pangolin coronavirus is not the Etiological agent of the death of the pangolins being sampled in the studies. This is further supported by the fact that 4 out of 10 lung samples in PRJNA573298 and 4 out of 7 lung samples in PRJNA607174 lacked any claimed or TRACEd coronavirus reads—despite the same symptoms displayed and similar date of death.

In order to establish the Etiological agent of the dead pangolins in the single GuangDone Accident that leads to the sampling and studies. A full virome TRACE analysis is conducted on the available samples for the determining of the exact etiological agent.

Extended Data Table S1

Full virome TRACE results of all Analyzable datasets of the GD pangolin incident

	Mammarenavirus	Nairoviridae	Murine respirovirus	Flaviviridae	Nidovirales	Rubulavirus	Nonanavirus	Peribunyavi	Amigovirus	Siphoviridae	Siphoviridae	Pahexavir
SRX6893158	Yes	Yes	No	No	No	No	Yes	No	Yes	Yes	No	No
SRX6893157	Yes	Yes	No	No	Claimed	No	No	Yes	No	No	No	No
SRX6893156	No	No	Yes	Yes	Yes	No	No	No	Yes	No	No	Yes
SRX6893155	No	No	Yes	No	Yes	No	No	No	No	No	No	No
SRX6893154	No	No	Yes	No	Yes	No	No	No	No	No	No	No
SRX6893153	No	No	Yes	Yes	Yes	No	No	No	Yes	No	No	No
SRX6893152	Yes	Yes	Yes	Yes	No	No	No	Yes	No	No	No	No
SRX6893151	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	No	No
SRX6893150	Yes	Yes	Yes	No	No	No	No	Yes	Yes	No	No	No
SRX6893149	Yes	Yes	No	No	No	No	No	No	No	No	Yes	No
SRX6893148	Yes	Yes	Yes	No	No	No	No	No	Yes	No	No	No
SRX6893147	Yes	Yes	"Respirovirus"	Yes	No	No	Yes	No	Yes	No	No	No
SRX6893146	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No
SRX6893145	Yes	Yes	No	No	No	No	No	No	No	No	No	No
SRX6893144	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No
SRX6893143	Yes	Yes	No	No	No	No	No	No	No	No	No	No
SRX6893142	Yes	Yes	No	No	No	No	No	Yes	Yes	No	No	No
SRX6893141	Yes	Yes	No	Yes	No	No	No	No	No	No	No	No
SRX6893140	Yes	Yes	Yes	No	No	No	No	Yes	No	No	No	No
SRX6893139	No	No	Yes	No	Yes	No	No	No	No	No	No	No
SRX6893138	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	No	No
SRX7756766	No	No	Yes	Yes	Yes	Yes	No	No	No	No	No	No
SRX7756765	No	No	Yes	No	No	Yes	No	No	No	No	No	No
SRX7756764	No	No	Yes	No	No	Yes	No	No	No	No	No	No
SRX7756763	No	No	Yes	No	No	Yes	No	No	No	No	No	No
SRX7756762	No	No	Yes	No	Claimed	Yes	No	No	No	No	No	No
SRX7756761	No	No	Yes	No	No	Yes	No	No	No	No	No	No
SRX7756769	No	No	Yes	Yes	Yes	No	No	No	No	No	No	No

A full Virome TRACE result suggest all the dead pangolins were infected by either Mammarenaviruses or Murine Respirovirus, or both. Including both samples that contained Claimed or TRACEd Coronavirus reads and the samples that didn't.

Murine Respirovirus and Mammarenaviruses co-infect 7 out of 29 Available Analyzable datasets, while none of the 29 datasets lacked both—indicating that both viruses were prevalent in the location where the pangolins were captive at The Guangdong Wildlife Rescue Center.

Symptoms of Murine Respirovirus in animals resembles that of SARS-CoV-2 in humans—It forms massive Syncytia in Eukaryotic cells, suppresses the immune system and causes secondary bacterial infections. The virus causes necrosis of Lung tissue in 5 days, with similar inflammation and immunopathological effects in the lung tissues of infected animals [5]—creating the histopathological effect as reported by Xiao et al.

It should be worth pointing out that the only examined lung tissues were examined by Xiao et al. And all Lung tissue samples examined by Xiao et.al contained Reads from the Murine Respirovirus.

Similarly, Mammarenaviruses are also known to cause multi organ, lethal[7] infections, characterized by endothelial pathology and swelling of internal organs. [6] All of which were Symptoms reported in the incident. As these samples were not examined Histopathologically by either the authors of [4] nor by any of the authors of any other article who have used the

datasets/samples, leaving the only mean of elucidating the cause of death being the observed symptoms and the coarse examination of the organs during sampling. Mammarenavirus infection therefore remains the most likely cause of death of the Murine Respirovirus Negative samples in the available datasets.

Is the "GD pangolin CoV" really a virus of the pangolin?

The only examination of the binding affinity of the GD pangolin CoV RBD to different animal receptors was done by Xiao et al [2], which performed molecular dynamic simulation of the RBD docking to the Human ACE2 receptor, The Civet ACE2 receptor and the pangolin ACE2 receptor. If the RBD of GD pangolin CoV in deed evolved in pangolins, we should expect the binding affinity of the RBD toward the pangolin ACE2 receptor to be the highest binding affinity returned from the examination.

However, neither the GD pangolin CoV RBD, nor the RBD of SARS-CoV-2 which is highly similar, produced a higher binding affinity to the pangolin ACE2 receptor than to the human ACE2 receptor, and both binds the Human ACE2 receptor with the highest affinity across all 3 animal species (Human, Civet, Pangolin) examined.

This fact argues strongly against the RBD residues of the GD pangolin CoV being evolved in pangolins, and instead favoring the RBD and the virus being the result of a passage experiment of a possible virus of pangolin origin (The GX/P2V virus was isolated and passaged in VERO E6 cells during it's collection in 2017) in Primate-derived cell lines.

There are only 2 locations of Biological sample storage in Guangdong, the Guangdong Institute of Applied Biological Resources and the China National GeneBank.

As all Credible (Non-filtered and contained analyzable Non-Coronavirus reads) samples were collected in a single incident from the Guangdong Wildlife Rescue Center[1][4][2], which the initial sample collection and storage was carried out by the Guangdong Institute of Applied Biological Resources[4], this experimental culture likely contaminated the GD pangolin samples during their initial collection or Storage, Either by the lab worker doing the initial sampling, or during their storage in the facility.

Epidemiology analysis of SARS-CoV-2 and related viruses argues strongly against the existence of a Coronavirus with the claimed RBD residues and sequence similarity in or near the GuangDong Wildlife Rescue Center at the time and date of the incident and the collection of the samples.

The earliest collection date of the GD pangolin CoV available, MP789, GenBank MT084071.1, is displayed at 29 March 2019.

Since the original location of the animals and samples in question was inside the Guangdong Wildlife Rescue Center which is neither a certified Biosafety Laboratory nor possessed adequate PPE when handling the animals, from the Simulation results by Xiao et al[2] and the observed

high human transmissibility of SARS-CoV-2 which had a very similar RBD, Should the GD pangolin CoV genuinely exists at that date and within the unprotected Guangdong Wildlife Rescue Center, It would almost certainly infect one to multiple On-site workers (Rescue workers which lacked either the Biosafety training or the adequate PPEs required to handle tissues or animals infected with a virus as characterized by the GD pangolin CoV papers) in the Guangdong Wildlife Rescue Center, and caused a SARS-level epidemic in Guangdong 2013 beginning in or around April 2019.
However, no such epidemic was recorded, nor there have been any virus that genetically resembled the GD pangolin CoV sequence (which is only 90% similar to SARS-CoV-2) being isolated in humans anywhere in the world even till today.
Nor there is a possibility that the current SARS-CoV-2 pandemic may have stemmed from the 29 March incident with the GD pangolin CoV, since the estimated time of divergence between the current SARS-CoV-2 genome to the GD pangolin CoV Genome was estimated to be at least 100 years ago , ranging from 1851 [1730,1958] to 1877 [1746,1986] [8], for a genome that is only 90% similar to SARS-CoV-2 and possessed significant difference in the sequence and composition of the viral proteins they encodes.
As the Earliest time of discovery and the incident on the GD pangolin CoV is no earlier than the beginning of Year 2019, The time between the incident and the first isolate of SARS-CoV-2 is far too short for GD pangolin CoV incident to be involved in the formation of the current SARS-CoV-2 pandemic, since even the neutral sites on the RBD itself would have taken more than 19.8 years to drift/evolve into what we seen today on the actual SARS-CoV-2 genome. [9]

Homo Sapiens reads are also found in SRX7756766, SRX7756765 and SRX7756762

In addition to Chlorocebus Spp. Indicative of VERO E6 cells, we have also found trace amount of reads uniquely matched to Hominidae within SRX7756766, SRX7756765 and SRX7756762. The presence of such reads may indicate that the Coronavirus-like reads from within such dataset were the result of index-hopping from a more highly contaminated original sample dataset, such as SRX7756769, of which the Homo Sapiens reads within such datasets may have index hopped into SRX7756766, SRX7756765 and SRX7756762, alongside with extremely low level of Coronavirus-related reads. Alternatively, these Homo Sapiens reads may represent the Homo Sapiens host sequence from the original sample which were left in the cell culture medium after inoculation of the VERO E6 cells as indicated by the virus isolation procedure performed by Xiao et al[2]. The presence of low levels of Homo Sapiens sequences within these datasets also confirms the origin of the inoculum into the VERO E6 cells as being samples that had significant amount of Homo Sapiens genetic material within them, which agrees with the hypothesis that the Coronavirus-related reads within these 3 datasets were the result of contamination by the In-lab VERO cell culture used by Xiao et al for virus isolation.

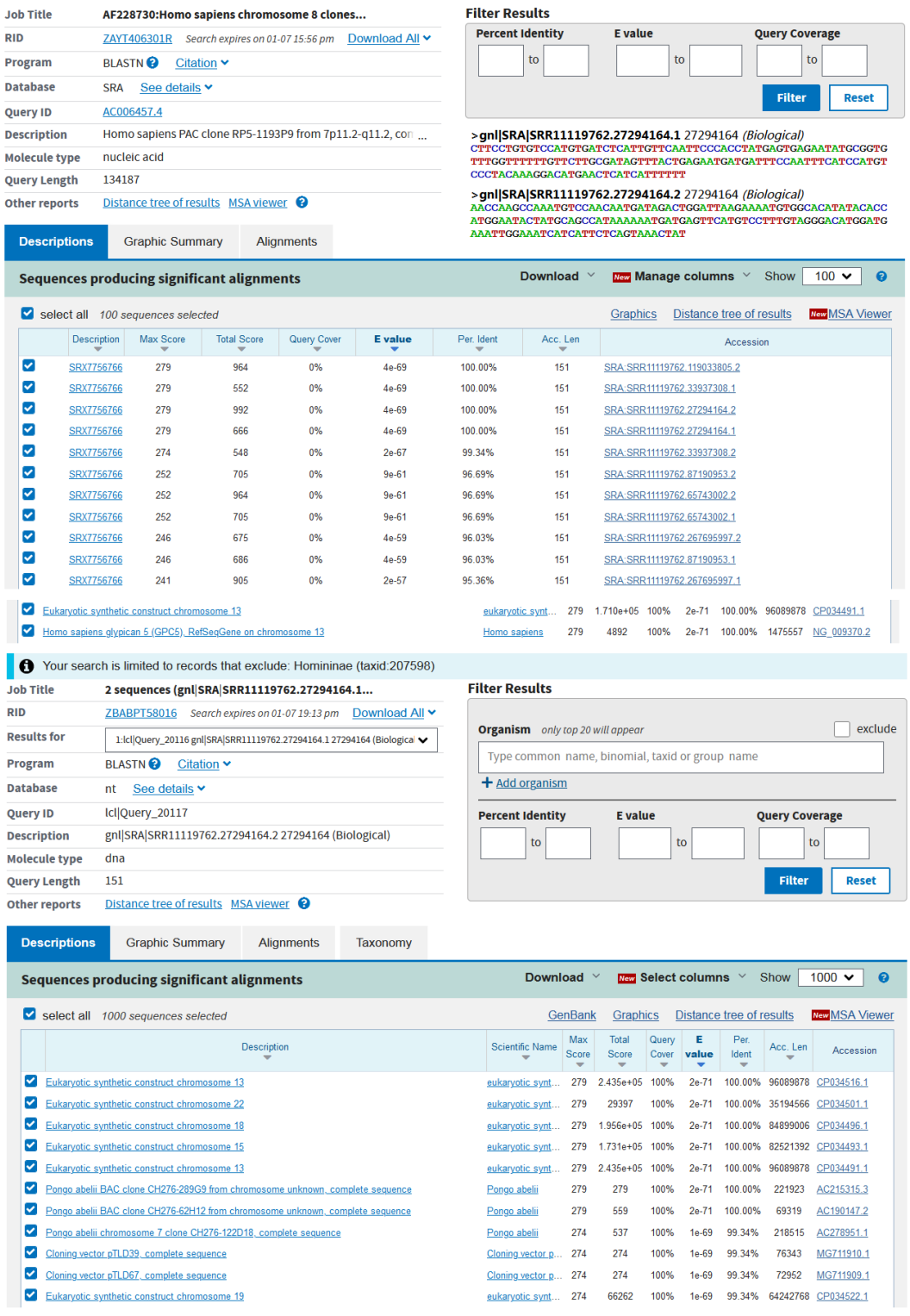


Fig 14: In addition to Chlorocebus Aethiops, reads uniquely matched to Hominidae have been found within SRX7756766.

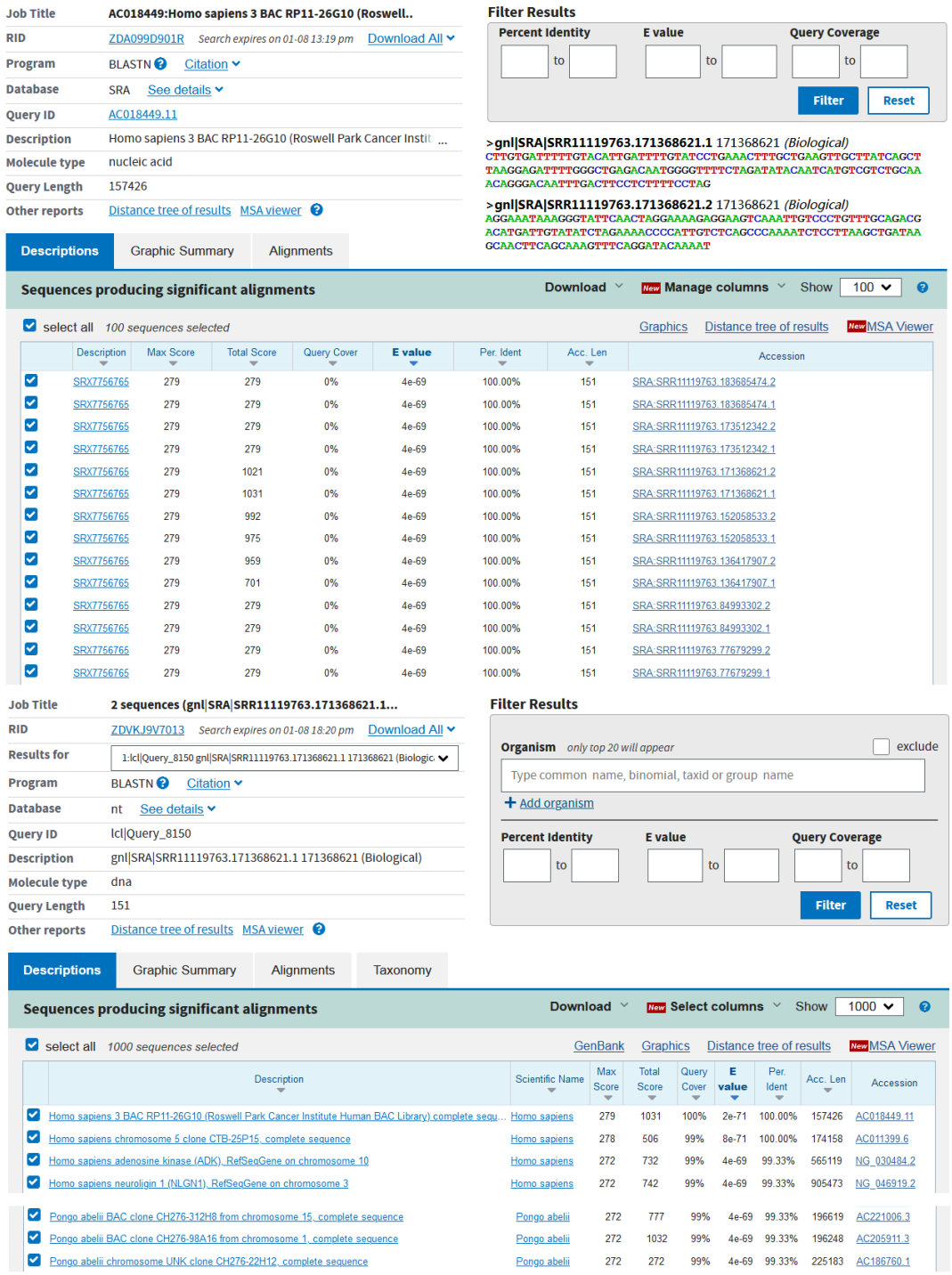


Fig 15: Reads uniquely matched to Homo Sapiens have been found within SRX7756765.

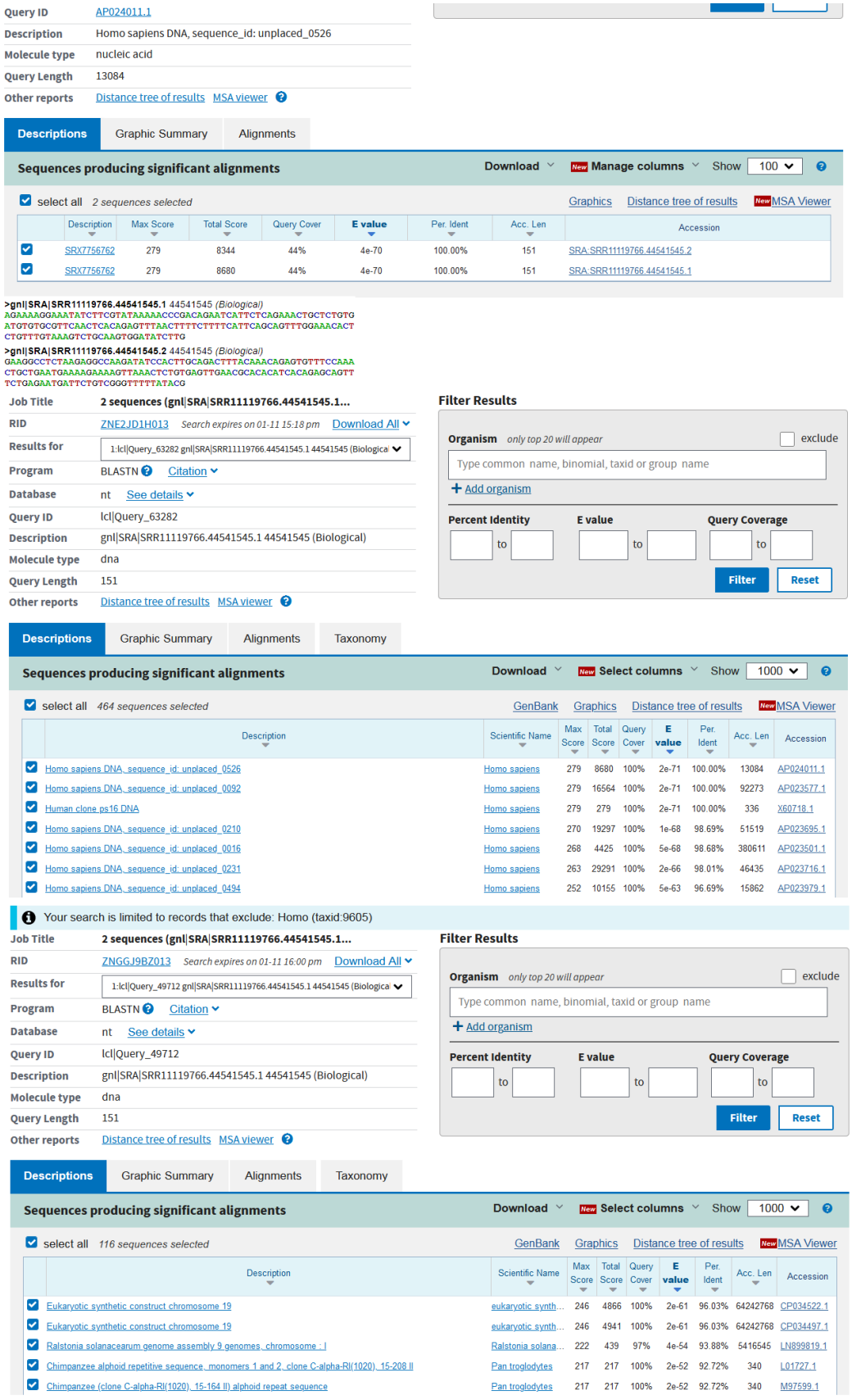


Fig.16: Reads uniquely matched to Homo Sapiens Alpha Satellite DNA in SRX7756762.

The Pan-SL-CoV-GD sequences can not be found in other pangolin sequencing datasets from China.

Recently, we are able to access and perform BLAST analysis on a large dataset of 93 pangolin samples deposited by Hu J *et al.,*[12] located under the BioProject PRJNA529540. These samples, alongside with an older sample, SRX1319167, represent a longitudinal survey spanning from between 1990 to 2017 of Both Manis Pentadactyla and Manis Javanica from China. We could not obtain any traces of reads resembling the Pan-SL-CoV-GD sequences from these datasets. Such discovery is in agreement with the conclusion of Hu J *et al.,*[12] which failed to find any evidence of SARS-CoV-2-like Coronaviruses within their sequencing study.

Considering that another longitudinal survey of 334 pangolins in Malaysia[13] have also failed to reveal any evidence of Coronaviruses or other potentially zoonotic viruses, the failure to isolate sequences of Coronaviruses from pangolin sequencing datasets are in good agreement that no natural infection of a Coronavirus can happen to a pangolin in the wild. This may be due to their solitary behavior[14] which keep them completely physically isolated from each other for up to 9 months for each year when the population is not in it's mating season which happens from May to July. As this is longer than the time of which a pangolin could stay infected before either clearing the infection or dying in any known incidence of viral infection in captive pangolins[15], any virus species that enters a pangolin population in the wild will either be cleared or kills all of its current hosts when the population is not in or have left its mating season, resulting in the viral population to go extinct. Therefore, the absence of viral infections in pangolin populations is the normal state of such population, and any datasets that claimed viral infection of pangolins must be subjected to the highest level of scrutiny to exclude any potential presence of contamination.

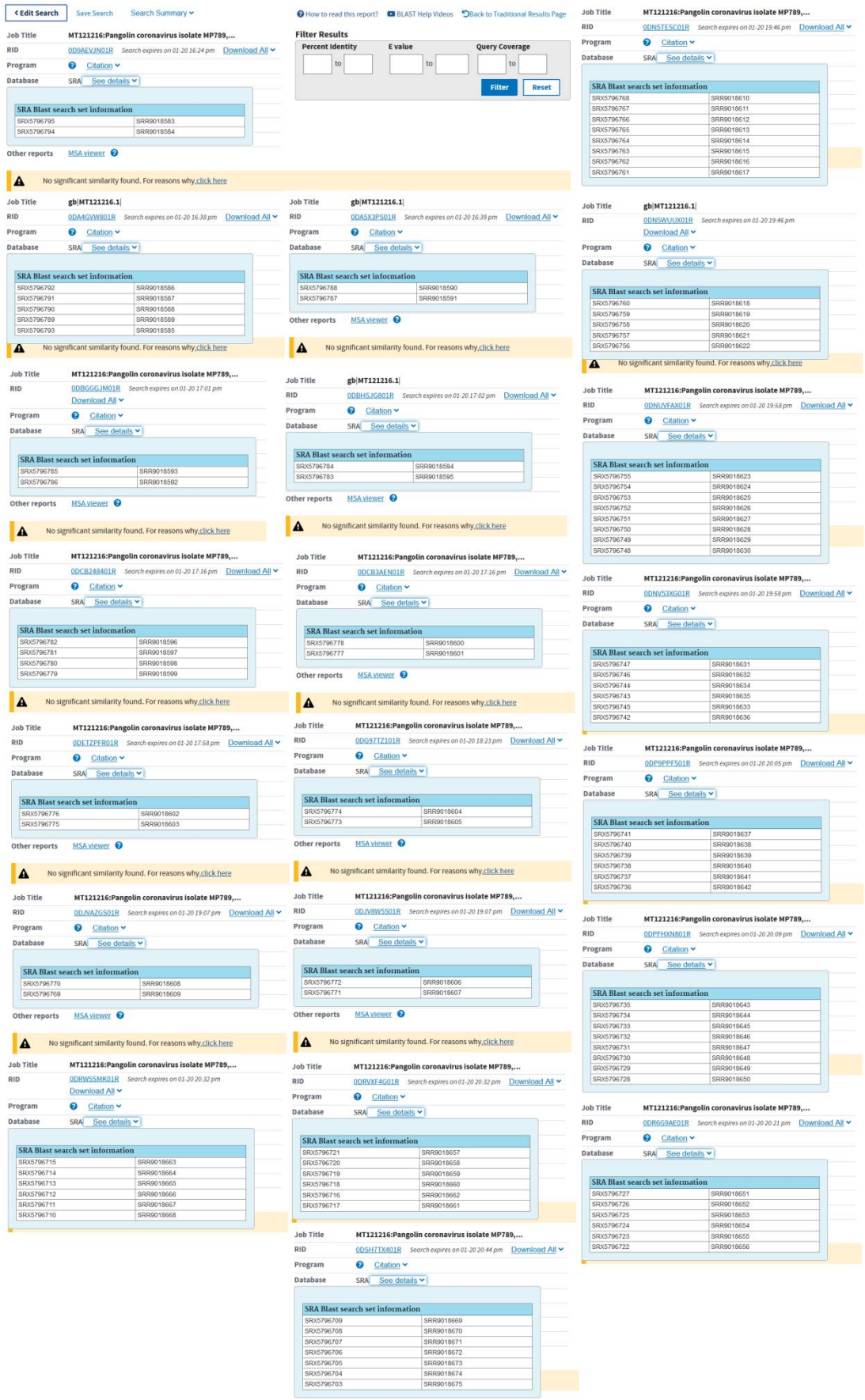


Fig.17a: No evidence of reads resembling the Pan-SL-CoV/GD sequences could be obtained from

PRJNA529540.

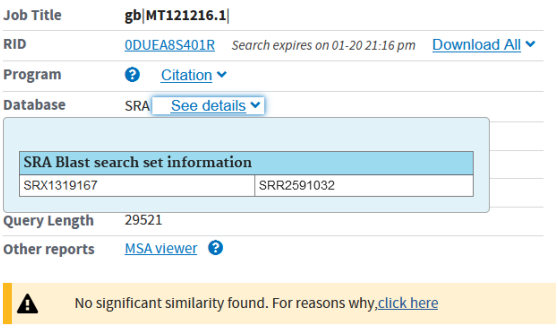


Fig17b: No evidence of reads resembling the Pan-SL-CoV/GD sequences could be found in SRX1319167.

Potential malpractice associated with Chinese pangolin sequencing data

Recently, we are able to obtain 20 RNA-seq datasets for the transcriptomic sequencing of both Manis Javanica and Manis Pentadactyla skin appendage (skin, scales), deposited by the Guangdong Institute of Applied Biological Resources from a project that is Separate from the Coronavirus-related sequencing project ongoing in the People's Republic of China. Located under the BioProject accession number PRJNA610466.

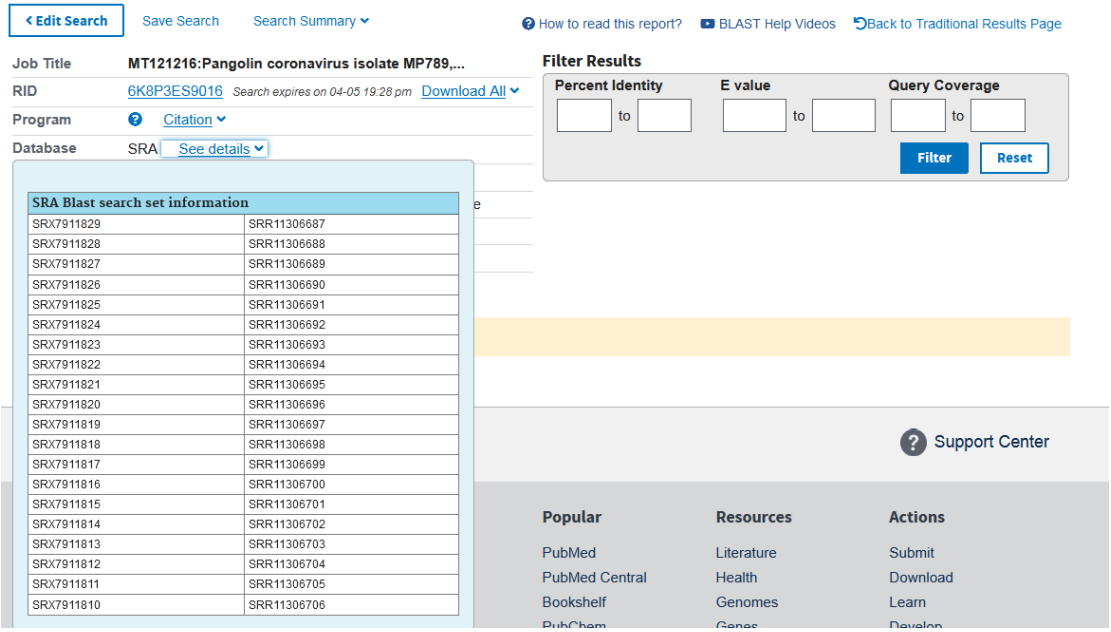


Fig 18. No evidence of the Pan-SL-CoV/GD sequences could be found within PRJNA610466.

Although we did not find any evidence of Coronaviruses from these datasets, We noticed that some of the datasets contained several sequences that were associated with vectors associated with the immortalization and engineering of mammalian cells, namely sequences resembling HIV-1, Macaca Mulatta polyomavirus 1 and Human betaherpesvirus 5.

HS14 (SRR11306689)

Metadata | Analysis | Reads | Data access

Taxonomy Analysis

Unidentified reads: 7.23%

Identified reads: 92.77%

- cellular organisms: **92.77%**
  - Eukaryota: **76.86%**
    - Opisthokonta: **76.69%**
      - Metazoa: **75.87%**
        - Boreoeutheria: **74.22%**
          - Laurasiatheria: **68.86%**
            - Manis javanica: **64.4%**
          - Euarchontoglires: **1.72%**
            - Simiiformes: **0.92%**
              - Catarrhini: **0.87%**
                - Hominoidea: **0.79%**
                  - Hominidae: **0.7%**
                    - Homininae: **0.63%**
                      - Homo sapiens: **0.3%**
      - Fungi: **0.79%**
    - Viridiplantae: **0.11%**
    - Sar: **< 0.01%** (188 Kbp)
  - Bacteria: **13.22%**
  - Archaea: **< 0.01%** (132 Kbp)
- Viruses: **< 0.01%** (222 Kbp)
  - Caudovirales: **< 0.01%** (100 Kbp)
    - Siphoviridae: **< 0.01%** (62 Kbp)
      - Tunavirinae: **< 0.01%** (33 Kbp)
        - Rtpvirus: **< 0.01%** (29 Kbp)
          - unclassified Rtpvirus: **< 0.01%** (29 Kbp)
            - Enterobacteria phage vB_EcoS_IME542: **< 0.01%** (16 Kbp)
            - Escherichia phage vB_Ecos_CEB_EC3a: **< 0.01%** (12 Kbp)
        - unclassified Tunavirinae: **< 0.01%** (5 Kbp)
          - Escherichia phage IMM-001: **< 0.01%** (5 Kbp)
      - Pahexavirus: **< 0.01%** (15 Kbp)
        - Propionibacterium phage Solid: **< 0.01%** (9 Kbp)
    - Myoviridae: **< 0.01%** (23 Kbp)
      - Eneladusvirus: **< 0.01%** (13 Kbp)
        - Cronobacter phage vB_CsaM_GAP32: **< 0.01%** (13 Kbp)
      - Tevenvirinae: **< 0.01%** (10 Kbp)
        - unclassified Tevenvirinae: **< 0.01%** (10 Kbp)
  - Riboviria: **< 0.01%** (74 Kbp)
    - Tymovirales: **< 0.01%** (35 Kbp)
      - Actinidia seed-borne latent virus: **< 0.01%** (35 Kbp)
    - Potyviridae: **< 0.01%** (29 Kbp)
      - Sugarcane mosaic virus: **< 0.01%** (29 Kbp)
    - Negarnaviricota: **< 0.01%** (9 Kbp)
      - Mononegavirales: **< 0.01%** (9 Kbp)
        - Filoviridae: **< 0.01%** (7 Kbp)
        - Paramyxoviridae: **< 0.01%** (2 Kbp)
          - Mammalian rubulavirus 5: **< 0.01%** (2 Kbp)
  - Polyomaviridae: **< 0.01%** (18 Kbp)
    - Macaca mulatta polyomavirus 1: **< 0.01%** (18 Kbp)
  - Ortervirales: **< 0.01%** (16 Kbp)
    - Orthoretrovirinae: **< 0.01%** (11 Kbp)
      - Gammaretrovirus: **< 0.01%** (9 Kbp)
        - RD114 retrovirus: **< 0.01%** (7 Kbp)
      - Lentivirus: **< 0.01%** (2 Kbp)
        - Human immunodeficiency virus 1: **< 0.01%** (2 Kbp)
  - Herpesvirales: **< 0.01%** (6 Kbp)
    - Herpesviridae: **< 0.01%** (5 Kbp)
      - Gammaherpesvirinae: **< 0.01%** (3 Kbp)
        - Human gammaherpesvirus 4: **< 0.01%** (3 Kbp)
      - Betaherpesvirinae: **< 0.01%** (2 Kbp)
        - Human betaherpesvirus 5: **< 0.01%** (2 Kbp)
  - Poxviridae: **< 0.01%** (4 Kbp)
    - Vaccinia virus GLV-1h68: **< 0.01%** (4 Kbp)

Fig.19: NCBI TRACE analysis result of SRR11306689

We downloaded the dataset with most concentrated occurrence of such sequences, SRR11306689, and performed sequence assembly using MEGAHIT[16]. Contiguous sequences with homology to these cellular engineering-associated sequences were identified using BowTie2[17] and their identities were elucidated through a combination of specific BLAST analysis and through sequence analysis using the Addgene sequence analysis tool.

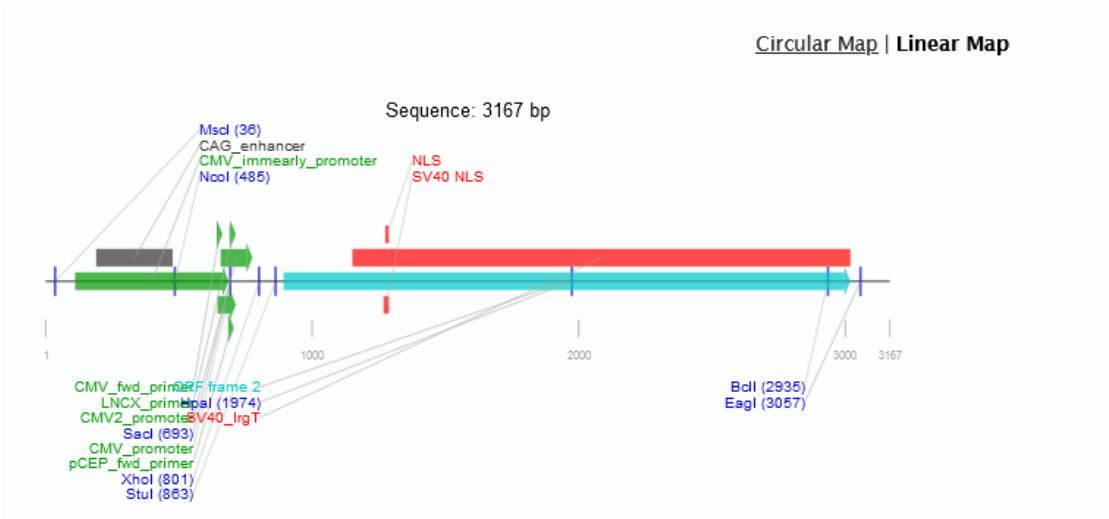


Fig.20: Addgene sequence analysis result of the single contig with detected homology to Macaca Mulatta polyomavirus 1 and Human betaherpesvirus 5.

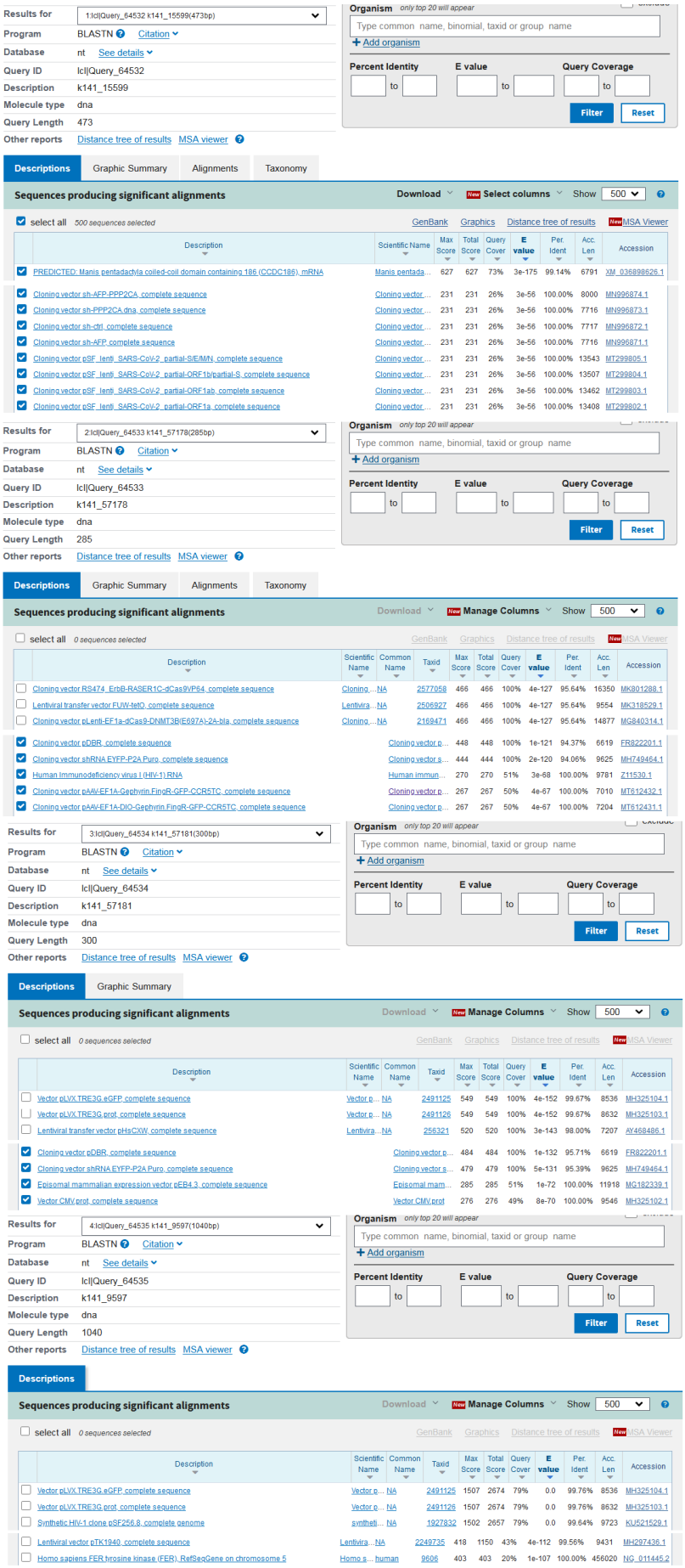
Sequence analysis of the largest contig with homology to cellular engineering related sequences revealed a Simian Virus 40(SV40) Large T antigen (LTA) placed behind a CMV promoter. Such a sequence is normally used to immortalize cells through the oncogenic properties of the Large T antigen, normally delivered into the cells using an integrative transfection technique, such as lentiviral vectors. This sequence appear to be a partial mRNA transcript.

☐	Murine retrovirus shuttle vector pZIPneoSV(TAg), complete sequence	syntheti...	NA	32630	4096	4096	70%	0.0	99.78%	7020	Z93724.1
☐	Synthetic construct clone pARVA, T-Ag, complete sequence	syntheti...	NA	32630	3849	5119	89%	0.0	99.30%	5690	MF174873.1
☐	Mammalian expression vector pSV529HIFNG, complete sequence	Mamma...	NA	1945111	3502	6652	70%	0.0	99.84%	9455	LT727634.1
☐	Mammalian expression vector pSV51E6Hf2, complete sequence	Mamma...	NA	1945103	3502	6652	70%	0.0	99.84%	9574	LT727623.1
☐	Mammalian expression vector pSV51E6Hf1, complete sequence	Mamma...	NA	1945102	3502	6652	70%	0.0	99.84%	9472	LT727622.1

☑	Mammalian expression vector LNXCO3, complete sequence	Mammalian exp...	1463	1463	25%	0.0	99.75%	7484	LT727330.1
☑	EIAV-based lentiviral vector, complete sequence	EIAV-based lent...	1461	1461	24%	0.0	100.00%	7941	GQ872121.1
☑	Mutant Human betaherpesvirus 5 clone AD169-BAC20, complete genome	Human betaher...	1448	1448	24%	0.0	100.00%	232314	MN920393.1
☑	Mutant Human betaherpesvirus 5 clone AD169-BAC2, complete genome	Human betaher...	1448	1448	24%	0.0	100.00%	233833	MN900952.1

Fig.21: BLAST result of this contig revealed that this sequence is associated with shuttle vectors (murine retrovirus, lentivirus) carrying the SV40 Large T antigen.

Four contigs within SRR11306689 are found with homology to Human Immunodefiency Virus 1 (HIV-1), which upon specific BLAST analysis reveal themselves to be sequences derived from lentiviral transfer vectors.



Results for 1:lcl|Query_64532 k141_15599(473bp)
Program BLASTN Citation
Database nt See details
Query ID lcl|Query_64532
Description k141_15599
Molecule type dna
Query Length 473
Other reports Distance tree of results MSA viewer

Organism only top 20 will appear
Type common name, binomial, taxid or group name
Add organism
Percent Identity to
E value to
Query Coverage to
Filter Reset

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments
Download New Select columns Show 500
select all 500 sequences selected
GenBank Graphics Distance tree of results New MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
PREDICTED: Manis pentadactyla coiled-coil domain containing 186 (CCDC186), mRNA	Manis pentada...	627	627	73%	3e-175	99.14%	6791	XM_036898626.1
Cloning vector sh-AFP-PPP2CA, complete sequence	Cloning vector...	231	231	26%	3e-56	100.00%	8000	MN996874.1
Cloning vector sh-PPP2CA,dna, complete sequence	Cloning vector...	231	231	26%	3e-56	100.00%	7716	MN996873.1
Cloning vector sh-ctrl, complete sequence	Cloning vector...	231	231	26%	3e-56	100.00%	7717	MN996872.1
Cloning vector sh-AFP, complete sequence	Cloning vector...	231	231	26%	3e-56	100.00%	7716	MN996871.1
Cloning vector pSF_lenti_SARS-CoV-2_partial-S/E/M/N, complete sequence	Cloning vector...	231	231	26%	3e-56	100.00%	13543	MT299805.1
Cloning vector pSF_lenti_SARS-CoV-2_partial-ORF1b/partial-S, complete sequence	Cloning vector...	231	231	26%	3e-56	100.00%	13507	MT299804.1
Cloning vector pSF_lenti_SARS-CoV-2_partial-ORF1ab, complete sequence	Cloning vector...	231	231	26%	3e-56	100.00%	13462	MT299803.1
Cloning vector pSF_lenti_SARS-CoV-2_partial-ORF1a, complete sequence	Cloning vector...	231	231	26%	3e-56	100.00%	13408	MT299802.1

Results for 2:lcl|Query_64533 k141_57178(285bp)
Program BLASTN Citation
Database nt See details
Query ID lcl|Query_64533
Description k141_57178
Molecule type dna
Query Length 285
Other reports Distance tree of results MSA viewer

Organism only top 20 will appear
Type common name, binomial, taxid or group name
Add organism
Percent Identity to
E value to
Query Coverage to
Filter Reset

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments
Download New Manage Columns Show 500
select all 0 sequences selected
GenBank Graphics Distance tree of results New MSA Viewer

Description	Scientific Name	Common Name	Taxid	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Cloning vector RS474_ErbB-RASER1C-dCas9VP64, complete sequence	Cloning...	NA	2577058	466	466	100%	4e-127	95.64%	16350	MK801288.1
Lentiviral transfer vector FUW-tetO, complete sequence	Lentivira...	NA	2506927	466	466	100%	4e-127	95.64%	9554	MK318529.1
Cloning vector pLenti-EF1a-dCas9-DNMT3B(E697A)-2A-bla, complete sequence	Cloning...	NA	2169471	466	466	100%	4e-127	95.64%	14877	MG840314.1
Cloning vector pDBR, complete sequence			Cloning vector p...	448	448	100%	1e-121	94.37%	6619	FR822201.1
Cloning vector shRNA EYFP-P2A Puro, complete sequence			Cloning vector s...	444	444	100%	2e-120	94.06%	9625	MH749464.1
Human immunodeficiency virus I (HIV-1) RNA			Human immun...	270	270	51%	3e-68	100.00%	9781	Z11530.1
Cloning vector pAAV-EF1A-Gephyrin.FingR-GFP-CCR5TC, complete sequence			Cloning vector p...	267	267	50%	4e-67	100.00%	7010	MT612432.1
Cloning vector pAAV-EF1A-DIO-Gephyrin.FingR-GFP-CCR5TC, complete sequence			Cloning vector p...	267	267	50%	4e-67	100.00%	7204	MT612431.1

Results for 3:lcl|Query_64534 k141_57181(300bp)
Program BLASTN Citation
Database nt See details
Query ID lcl|Query_64534
Description k141_57181
Molecule type dna
Query Length 300
Other reports Distance tree of results MSA viewer

Organism only top 20 will appear
Type common name, binomial, taxid or group name
Add organism
Percent Identity to
E value to
Query Coverage to
Filter Reset

Descriptions Graphic Summary

Sequences producing significant alignments
Download New Manage Columns Show 500
select all 0 sequences selected
GenBank Graphics Distance tree of results New MSA Viewer

Description	Scientific Name	Common Name	Taxid	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Vector pLVX.TRE3G.eGFP, complete sequence	Vector p...	NA	2491125	549	549	100%	4e-152	99.67%	8536	MH325104.1
Vector pLVX.TRE3G.prot, complete sequence	Vector p...	NA	2491126	549	549	100%	4e-152	99.67%	8632	MH325103.1
Lentiviral transfer vector pHsCXW, complete sequence	Lentivira...	NA	256321	520	520	100%	3e-143	98.00%	7207	AY468486.1
Cloning vector pDBR, complete sequence			Cloning vector p...	484	484	100%	1e-132	95.71%	6619	FR822201.1
Cloning vector shRNA EYFP-P2A Puro, complete sequence			Cloning vector s...	479	479	100%	5e-131	95.39%	9625	MH749464.1
Episomal mammalian expression vector pEB4.3, complete sequence			Episomal mam...	285	285	51%	1e-72	100.00%	11918	MG182339.1
Vector CMV.prot, complete sequence			Vector CMV.prot	276	276	49%	8e-70	100.00%	9546	MH325102.1

Results for 4:lcl|Query_64535 k141_9597(1040bp)
Program BLASTN Citation
Database nt See details
Query ID lcl|Query_64535
Description k141_9597
Molecule type dna
Query Length 1040
Other reports Distance tree of results MSA viewer

Organism only top 20 will appear
Type common name, binomial, taxid or group name
Add organism
Percent Identity to
E value to
Query Coverage to
Filter Reset

Descriptions

Sequences producing significant alignments
Download New Manage Columns Show 500
select all 0 sequences selected
GenBank Graphics Distance tree of results New MSA Viewer

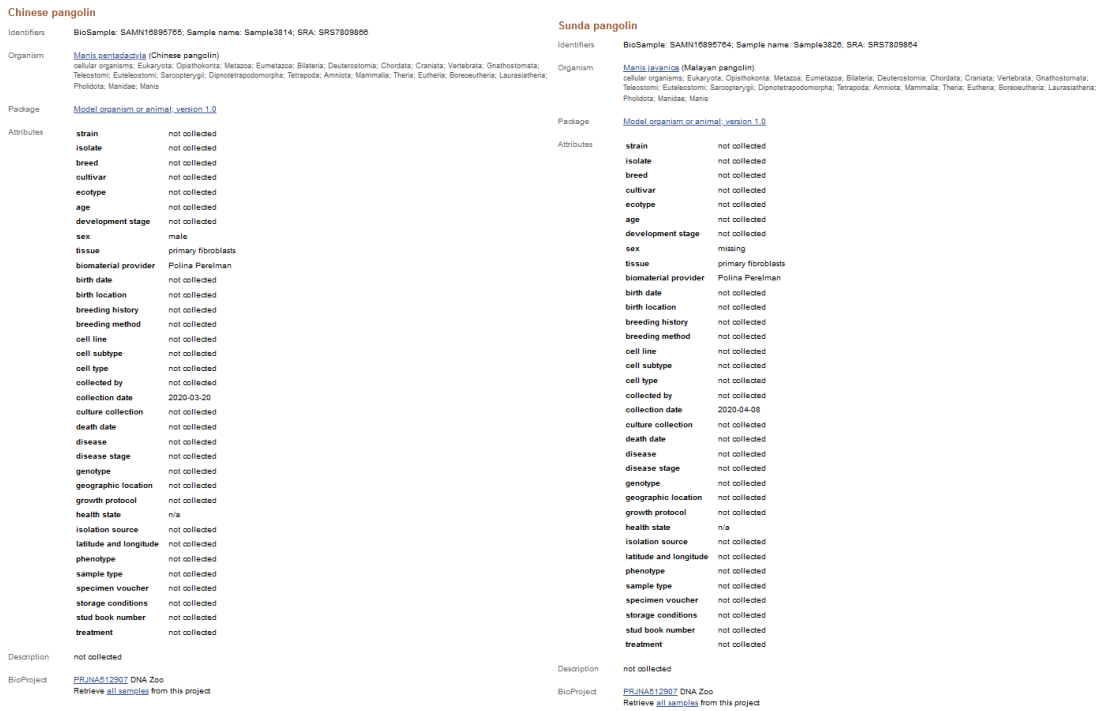
Description	Scientific Name	Common Name	Taxid	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Vector pLVX.TRE3G.eGFP, complete sequence	Vector p...	NA	2491125	1507	2674	79%	0.0	99.76%	8536	MH325104.1
Vector pLVX.TRE3G.prot, complete sequence	Vector p...	NA	2491126	1507	2674	79%	0.0	99.76%	8632	MH325103.1
Synthetic HIV-1 clone pSF256.8, complete genome	syntheti...	NA	1927832	1502	2657	79%	0.0	99.64%	9723	KU521529.1
Lentiviral vector pTK1940, complete sequence	Lentivira...	NA	2249735	418	1150	43%	4e-112	99.56%	9431	MH297436.1
Homo sapiens FER tyrosine kinase (FER), RefSeqGene on chromosome 5	Homo s...	human	9606	403	403	20%	1e-107	100.00%	456020	NG_011445.2

Fig.23: Specific BLAST analysis result of the 4 contigs associated with Human Immunodeficiency Virus 1 (HIV-1).

A deep analysis of these 4 contigs revealed 2 contigs spanning the vector-virus junction for common lentiviral transfer vectors, one contig displaying an Integration junction of the LTR into Manis Pentadactyla DNA and one contig displaying an integration junction of the LTR into Homo Sapiens DNA.

The presence of these fragmented sequences carrying both the payload (SV40 Large T antigen) and the vehicle (Lentiviral transfer vectors) with evidence of delivery into pangolin cells (integration junction of Vector LTR DNA into Manis Pentadactyla genomic DNA), suggesting an ongoing effort of immortalizing pangolin cells and keeping them in culture being conducted in the Guangdong Institute of Applied Biological Resources.

Indeed, the presence of Primary Fibroblast (Skin, Muscle) cells cultures from both Manis Pentadactyla and Manis Javanica and their availability to labs have been recently confirmed by two BioSamples of primary fibroblast cells “collected by Dr. Shujin Luo (Peking University, China)” placed under accession SAMN16895765 and SAMN16895764, with collection date of 20/03/2020 and 08/04/2020 respectively.



Chinese pangolin

Identifiers	BioSample: SAMN16895765; Sample name: Sample3814; SRA: SRS7809866
Organism	Manis pentadactyla (Chinese pangolin) cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Laurasiatheria; Pholidota; Manidae; Manis
Package	Model organism or animal; version 1.0
Attributes	

strain	not collected
isolate	not collected
breed	not collected
cultivar	not collected
ecotype	not collected
age	not collected
development stage	not collected
sex	male
tissue	primary fibroblasts
biomaterial provider	Polina Perelman
birth date	not collected
birth location	not collected
breeding history	not collected
breeding method	not collected
cell line	not collected
cell subtype	not collected
cell type	not collected
collected by	not collected
collection date	2020-03-20
culture collection	not collected
death date	not collected
disease	not collected
disease stage	not collected
genotype	not collected
geographic location	not collected
growth protocol	not collected
health state	n/a
isolation source	not collected
latitude and longitude	not collected
phenotype	not collected
sample type	not collected
specimen voucher	not collected
storage conditions	not collected
stud book number	not collected
treatment	not collected

Description	not collected
BioProject	PRJNA512907 DNA Zoo Retrieve all samples from this project

Sunda pangolin

Identifiers	BioSample: SAMN16895764; Sample name: Sample3826; SRA: SRS7809864
Organism	Manis javanica (Malayan pangolin) cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Laurasiatheria; Pholidota; Manidae; Manis
Package	Model organism or animal; version 1.0
Attributes	

strain	not collected
isolate	not collected
breed	not collected
cultivar	not collected
ecotype	not collected
age	not collected
development stage	not collected
sex	missing
tissue	primary fibroblasts
biomaterial provider	Polina Perelman
birth date	not collected
birth location	not collected
breeding history	not collected
breeding method	not collected
cell line	not collected
cell subtype	not collected
cell type	not collected
collected by	not collected
collection date	2020-04-08
culture collection	not collected
death date	not collected
disease	not collected
disease stage	not collected
genotype	not collected
geographic location	not collected
growth protocol	not collected
health state	n/a
isolation source	not collected
latitude and longitude	not collected
phenotype	not collected
sample type	not collected
specimen voucher	not collected
storage conditions	not collected
stud book number	not collected
treatment	not collected

Description	not collected
BioProject	PRJNA512907 DNA Zoo Retrieve all samples from this project

Figure 24: BioSample description of SAMN16895765 and SAMN16895764 as “primary fibroblast cells”.

As such a cell line could potentially be used to culture the Pan-SL-CoV/GD virus, should an isolate in VERO E6 [2] exist, these cell lines may potentially be used to contrive “novel” BioSamples and SRAs for this sequence through inoculation and serial passage of the cultured virus in order to eliminate the primate host sequences from the original samples, due to the central role of the RBD of this supposedly “wild” sequence in current publications regarding SARS-CoV-2 origin.

We therefore urge caution when adopting any short read sequencing (SRA) data or viral nucleotide sequences from pangolins with a date of deposition after the collection date for SAMN16895765 and SAMN16895764, especially after the publishing date of PRJNA610466,

24/11/2020, due to the identity of potential “tissue” samples being no longer restricted exclusively to living or dead wild animals once a primary or immortalized cell line of such a species have been established, potentially allowing malpractice when “sequencing” “new” samples from these two species.

The MEGAHIT result and the obtained vector sequences from SRR11306689 have been deposited as

Galaxy266-[Assembly_with_MEGAHIT_on_data_260] SRR11306689.fasta

HIV-1 from SRR11306689.fa

and

SV40 LTA+CMV from SRR11306689.fa

Conclusions

The Extreme lack of transparency and the sheer level of contamination from the original samples, the lack of epidemiological evidence of it’s existence at the location of it’s collection, and the receptor binding affinity of the Viral RBD itself indicating it as not being evolved nor adapted in pangolins, all strongly argue against the existence of a SARS-CoV-2 like virus infecting pangolins captive in Guangdong at 2019.

Moreover, it suggests that the GD pangolin CoV exists only as a culture in Primate-derived cells within the lab/facility used for the initial collection and/or storage of the samples of the pangolins in question, raising important issues on the serial passage Gain-Of-Function research of viral pathogens.

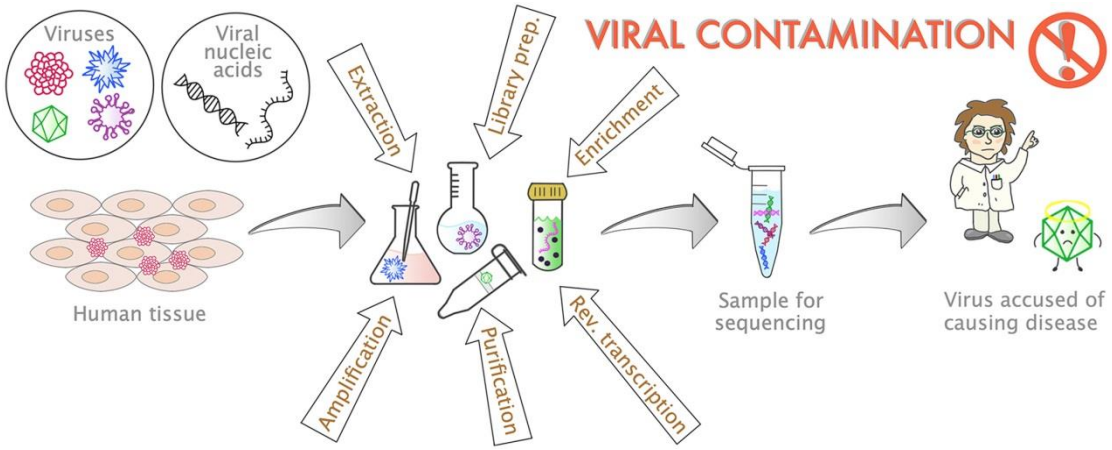


Figure 25. A cartoon diagram of contamination in sequencing experiment leading to false results and false “discoveries”.

Note as in 2020/7/23

A recent Dataset, seemingly unrelated to the Xiao et.al Nature dataset, **SRX8582289**, appeared under PRJNA607174. This dataset seems to be newly sequenced, and it was not referred in [2].

Table S2: TRACE analysis result of the **SRX8582289** dataset.

Accession number and	Primary Mammalian	Primate-related results	Identification of

registration date	Trace results and percentage	in Krona and read size by Kbp	“Coronaviridae” as by Trace and total read size
SRX8582289 22-Jun-2020	Manis javanica: **43.52%**	Catarrhini 98913	Pangolin coronavirus 792

Nevertheless, in-depth analysis revealed significant amount of contamination from the Human genome, with ratio of Virus to cell=0.8%.

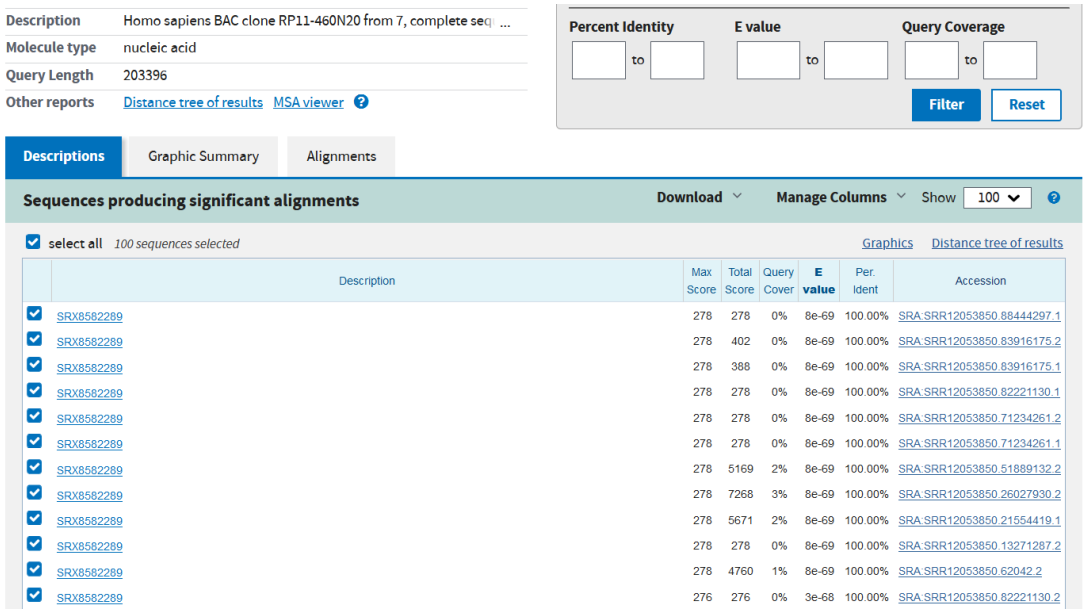


Figure S1A: Some BLAST hits out of a human Somatic BAC clone.

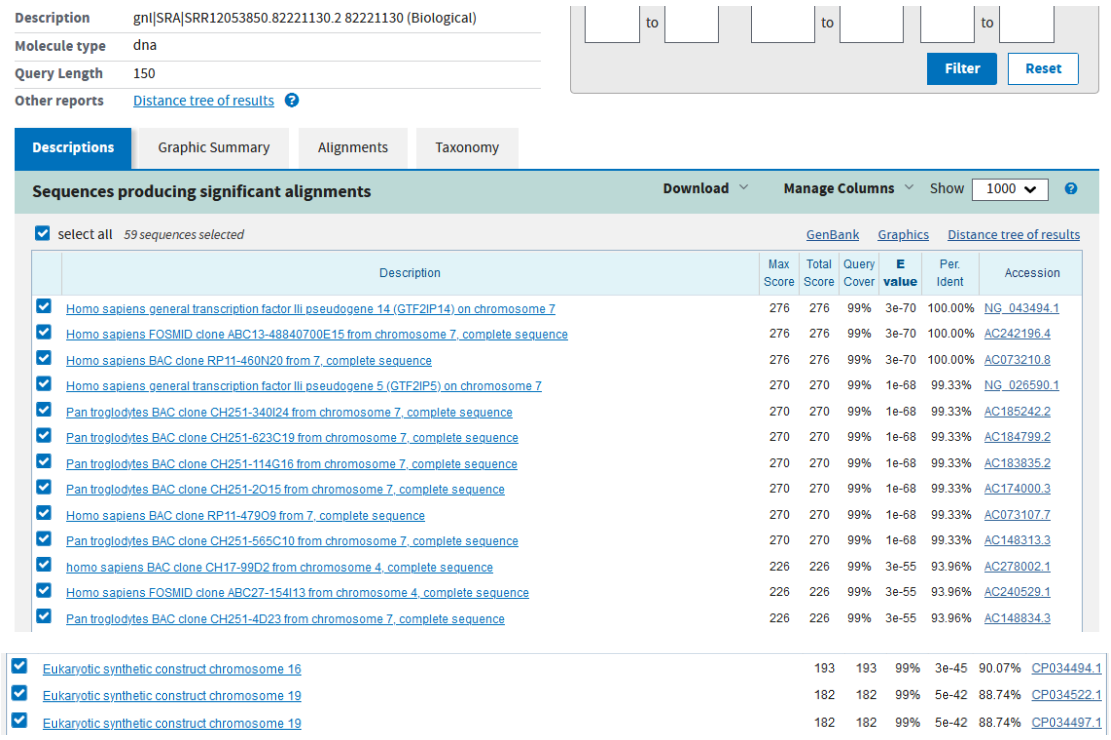


Fig. S1B: BLAST results returned only Homo Sapiens as 100% match. This indicate that the listed Catarrhini reads come from Homo Sapiens.

The significance of this particular dataset is yet unknown.

Note as in 2020/12/26

Two Recent SRAs, **SRX9714436** and **SRX9714921**, were recently deposited by the Guangdong Institute of Applied Biological Resources with a listed DOI connection to 10.1371/journal.ppat.1008421 [1]. Both samples have a depositor of LinMao Li, 2020-12-21 the same time as the specified BioProject registration date. Only one of the SRAs contained significant amount of Coronavirus-related reads.

Table S3: TRACE analysis result of **SRX9714436** and **SRX9714921**.

Accession number and registration date	Primary Mammalian Trace results and percentage	Primate-related results in Krona and read size by Kbp	Identification of "Coronaviridae" as by Trace and total read size
SRX9714436 21-Dec-2020	Manis javanica: **3.14%** Homo sapiens: **0.04%**	**Homo sapiens** 12332	Pangolin coronavirus 3 Kbp
SRX9714921 21-Dec-2020	Homo sapiens: **0.15%**	**Homo sapiens** 9923	N/D

As expected by TRACE results, Reads that are 100% full-length uniquely matched to Homo Sapiens were obtained from **SRX9714436** and **SRX9714921.**

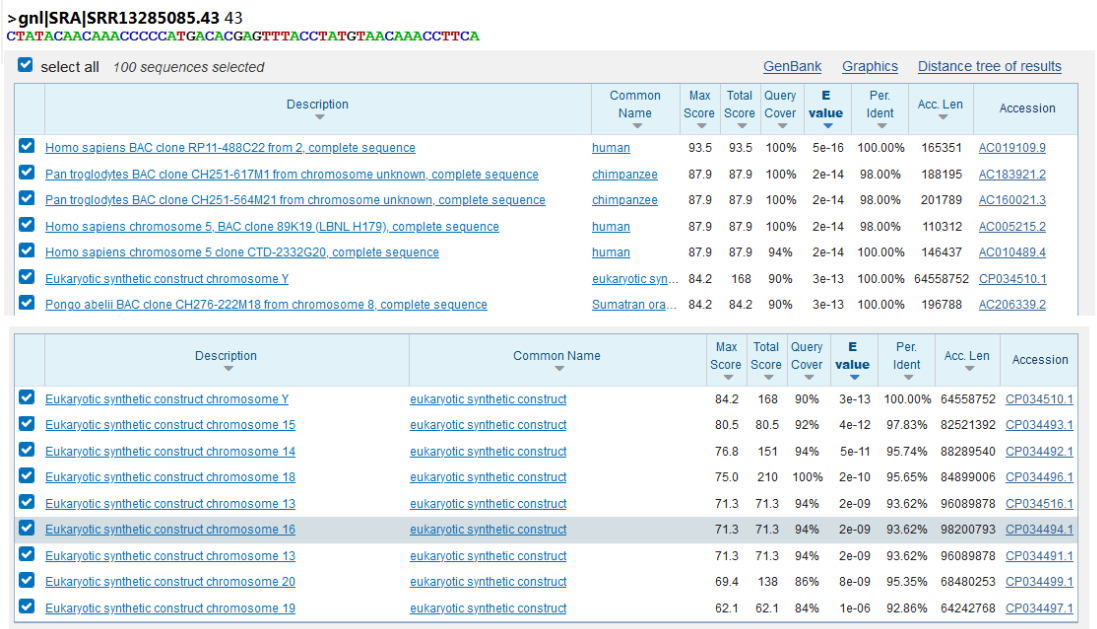


Fig.S2A: Homo Sapiens 100% full-length Unique matched read obtained from **SRX9714436**

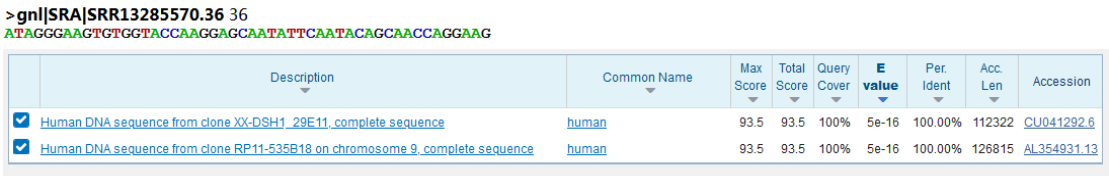


Fig.S2B: Homo Sapiens 100% full-length Unique matched read obtained from **SRX9714921**

A provided .fastq file was also found in **SRX9714436**. Analysis using stand-alone MagicBlast[11]

suggests significant presence of Homo Sapiens reads within this fastq file, similar to that of the Run itself.

NDX550397_RUO:309:H3FKWBGXH:1:11101:21446:1055 16 AC019109.9 87488 255
50M * 0 0 TGAAGGTTTGTTACATAGGTAAACTCGTGTCATGGGGGTTTGTTGTATAG *
NH:i:1 AS:i:50 NM:i:0

	Description	Common Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Homo sapiens BAC clone RP11-488C22 from 2, complete sequence	human	93.5	93.5	100%	5e-16	100.00%	165351	AC019109.9
☑	Pan troglodytes BAC clone CH251-617M1 from chromosome unknown, complete sequence	chimpanzee	87.9	87.9	100%	2e-14	98.00%	188195	AC183921.2
☑	Pan troglodytes BAC clone CH251-564M21 from chromosome unknown, complete sequence	chimpanzee	87.9	87.9	100%	2e-14	98.00%	201789	AC160021.3
☑	Homo sapiens chromosome 5, BAC clone 89K19 (LBNL H179), complete sequence	human	87.9	87.9	100%	2e-14	98.00%	110312	AC005215.2
☑	Homo sapiens chromosome 5 clone CTD-2332G20, complete sequence	human	87.9	87.9	94%	2e-14	100.00%	146437	AC010489.4
☑	Eukaryotic synthetic construct chromosome Y	eukaryotic syn...	84.2	168	90%	3e-13	100.00%	64558752	CP034510.1
☑	Pongo abelii BAC clone CH276-222M18 from chromosome 8, complete sequence	Sumatran ora...	84.2	84.2	90%	3e-13	100.00%	196788	AC206339.2
☑	Human DNA sequence from clone RP11-987D21 on chromosome X, complete sequence	human	84.2	84.2	96%	3e-13	97.92%	55442	BX119919.5
☑	PREDICTED: Callithrix jacchus uncharacterized LOC118154814 (LOC118154814), ncRNA	white-tufted-e...	82.4	82.4	94%	1e-12	97.87%	2719	XR_004745075.1
☑	Homo sapiens solute carrier family 26 member 3 (SLC26A3), RefSeqGene (LRG_683) on chromos...	human	82.4	82.4	100%	1e-12	96.00%	44767	NG_008046.1
☑	Eukaryotic synthetic construct chromosome Y	eukaryotic synthetic construct	84.2	168	90%	3e-13	100.00%	64558752	CP034510.1
☑	Eukaryotic synthetic construct chromosome 15	eukaryotic synthetic construct	80.5	80.5	92%	4e-12	97.83%	82521392	CP034493.1
☑	Eukaryotic synthetic construct chromosome 14	eukaryotic synthetic construct	76.8	151	94%	5e-11	95.74%	88289540	CP034492.1
☑	Eukaryotic synthetic construct chromosome 18	eukaryotic synthetic construct	75.0	210	100%	2e-10	95.65%	84899006	CP034496.1
☑	Eukaryotic synthetic construct chromosome 13	eukaryotic synthetic construct	71.3	71.3	94%	2e-09	93.62%	96089878	CP034516.1
☑	Eukaryotic synthetic construct chromosome 16	eukaryotic synthetic construct	71.3	71.3	94%	2e-09	93.62%	98200793	CP034494.1
☑	Eukaryotic synthetic construct chromosome 13	eukaryotic synthetic construct	71.3	71.3	94%	2e-09	93.62%	96089878	CP034491.1
☑	Eukaryotic synthetic construct chromosome 20	eukaryotic synthetic construct	69.4	138	86%	8e-09	95.35%	68480253	CP034499.1
☑	Eukaryotic synthetic construct chromosome 19	eukaryotic synthetic construct	62.1	62.1	84%	1e-06	92.86%	64242768	CP034497.1

Fig.S3: BLAST result of the read **NDX550397_RUO:309:H3FKWBGXH:1:11101:21446:1055 TGAAGGTTTGTTACATAGGTAAACTCGTGTCATGGGGGTTTGTTGTATAG** within the provided fastq file. The read is 100% full-length uniquely matched to Homo Sapiens.

Note as in 2021/05/19

We have recently obtained a SRA dataset on the Chinese National GeneBank(CNGB) of a claimed “skin sample” from a Pangolin individual with claimed infection by the Pan-SL-CoV/GD genome, DOI: http://dx.doi.org/10.26036/CNP0001573. Analysis of the read files provided have identified significant amount of human read contamination within the dataset. The alignment result to the human genome has been deposited as CNP0001573_human_alignments.sam.

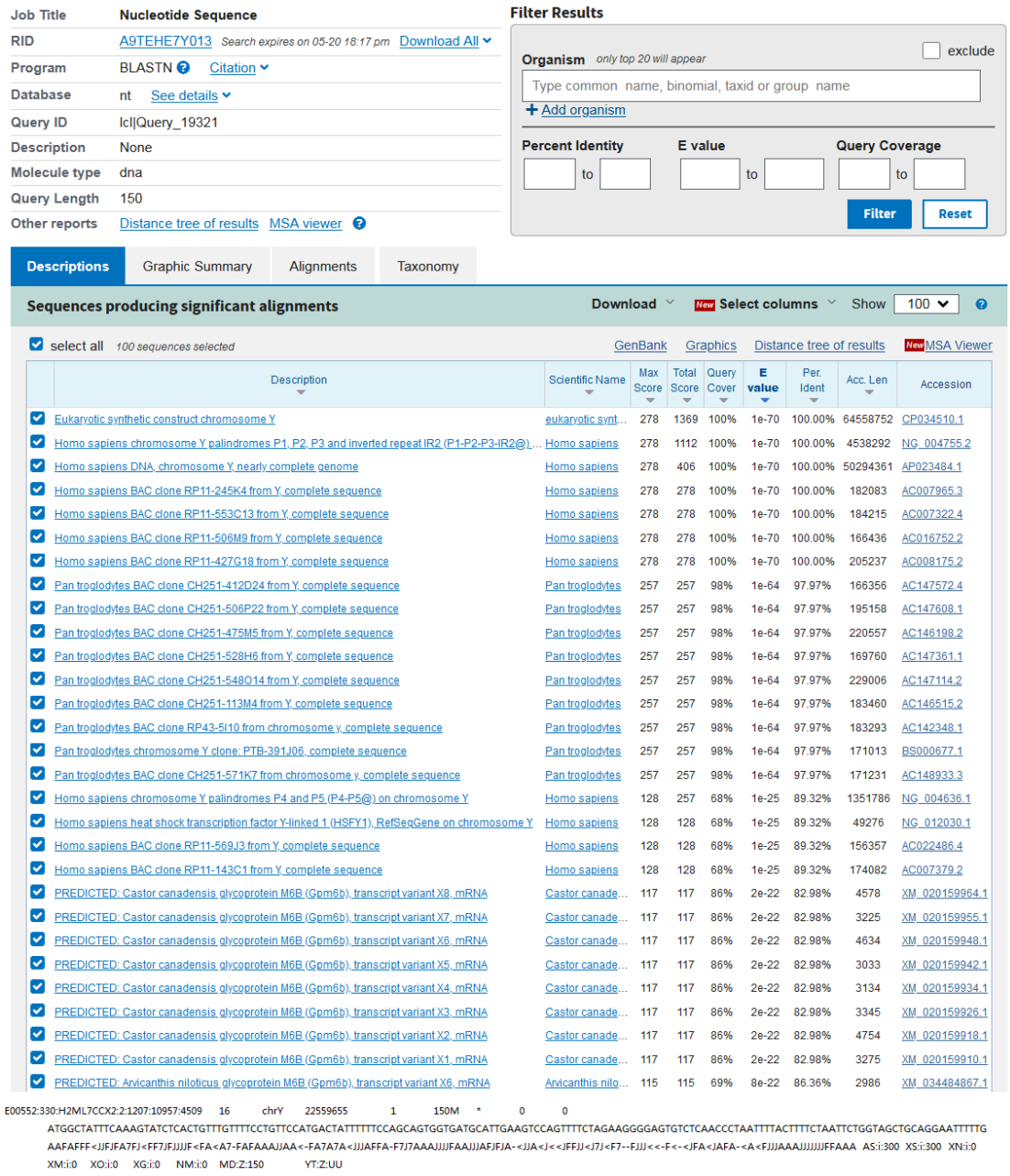


Fig.S4: An example of a read aligned to the human Y chromosome from CNP0001573. This could be linked to the human lung adenocarcinoma cell line A549, which originated from a male individual and is used for experimentation with SARS-like viruses.[18][19]

Note as in 2021/06/20

Two sequencing datasets under PRJNA607174, SRX7756767 and SRX7756768 have been published. These datasets are barely readable with very little sequence that resembles attributable, non-ribosomal RNA. Only one of these datasets are deemed analyzable by the NCBI TRACE tool.

Identification of reads were performed using BowTie2 against MN908947.3 and MT121216.1,

and only SRX7756767 contained one read from the Pan-SL-CoV/GD sequence.

Table S4: TRACE analysis result of SRX7756768 and SRX7756767.

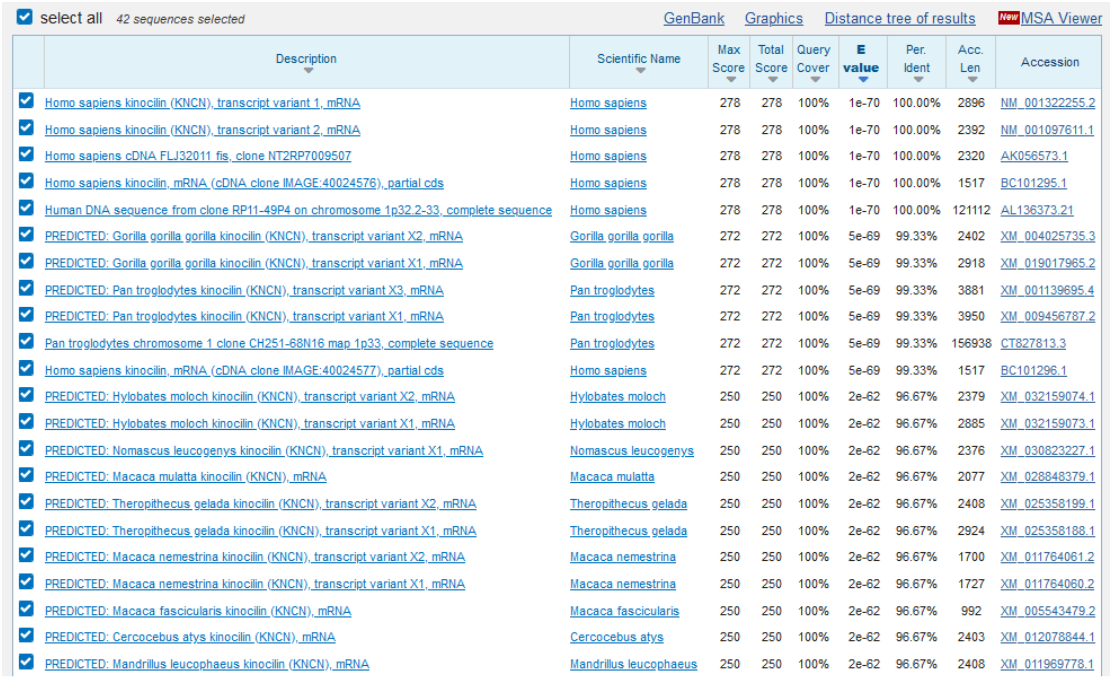
Accession number and registration date	Primary Mammalian Trace results and percentage	Primate-related results in Krona and read size by Kbp	Identification of "Coronaviridae" as by Trace and total read size
SRX7756768 18-Feb-2020	Homo sapiens: **0.03%**	**Homo sapiens** 6557937	Canine coronavirus 2 2Kbp
SRX7756767 18-Feb-2020	Homo sapiens: 8.69Mbp (BowTie2) Manis pentadactyla: 312.5Mbp (BowTie2) Mus Musculus: 12.39Kbp (BowTie2)	N/A (not analyzable)	One read matching MT121216.1 (BowTie2)

A00129:562:HWNFWDSXX:1:2267:15420:7247 16 MN908947.3 812 8 150M * 0 0 ACTCGCTATGTCGATAACACCTACTGTGGCCCTGATGGCTACCCTCTTGAGTGCATTAAAGACTTGCTG GCGCGTGCTGGTAAAGCTTCTTGCACTTTGTCCGAACAACTGGACTTTCTTGACACTAAGAGAGGTGTGTAC TGCTGCCGT FFFFFFFFFFFFFFFFFF,F:FFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFF:F:FFFFFFFFFFFFFFFFFFFFFFF FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,,FFFFFFFFFFFFFFFFFFFFFF:FF AS:i:-48 XN:i:0 XM:i:10 XO:i:0 XG:i:0 NM:i:10 MD:Z:19A2T40C1T2A2A17A27A13G5A12 YT:Z:UU

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Eukaryotic synthetic construct chromosome 19	eukaryotic synt...	278	278	100%	1e-70	100.00%	64242768	CP034522.1
☑	Eukaryotic synthetic construct chromosome 19	eukaryotic synt...	278	278	100%	1e-70	100.00%	64242768	CP034497.1
☑	Homo sapiens chromosome 19 open reading frame 12 (C19orf12), RefSeqGene on chromosome 19	Homo sapiens	278	278	100%	1e-70	100.00%	23904	NG_031970.2
☑	Homo sapiens isolate CHM13 chromosome 19	Homo sapiens	278	278	100%	1e-70	100.00%	61707364	CP068259.2
☑	Homo sapiens DNA, chromosome 19, nearly complete genome	Homo sapiens	278	278	100%	1e-70	100.00%	59105444	AP023479.1
☑	Homo sapiens chromosome 19 clone LLNLR-232H11, complete sequence	Homo sapiens	278	278	100%	1e-70	100.00%	37770	AC010513.6
☑	PREDICTED: Pongo abelii chromosome 19 C19orf12 homolog (C19H19orf12), transcript variant X4, ...	Pongo abelii	231	231	93%	8e-57	96.43%	2247	XM_009253101.2
☑	PREDICTED: Pan paniscus chromosome 19 C19orf12 homolog (C19H19orf12), transcript variant X7,...	Pan paniscus	152	152	54%	6e-33	100.00%	4099	XM_003816225.5
☑	PREDICTED: Pan paniscus chromosome 19 C19orf12 homolog (C19H19orf12), transcript variant X1,...	Pan paniscus	152	152	54%	6e-33	100.00%	4292	XM_003816222.4
☑	PREDICTED: Pan troglodytes chromosome 19 C19orf12 homolog (C19H19orf12), transcript variant X...	Pan troglodytes	152	152	54%	6e-33	100.00%	3594	XM_003953437.4
☑	PREDICTED: Pan troglodytes chromosome 19 C19orf12 homolog (C19H19orf12), transcript variant X...	Pan troglodytes	152	152	54%	6e-33	100.00%	3771	XM_003316249.4
☑	Homo sapiens chromosome 19 open reading frame 12 (C19orf12), transcript variant 1, mRNA	Homo sapiens	152	152	54%	6e-33	100.00%	4265	NM_001031726.4
☑	Homo sapiens chromosome 19 open reading frame 12 (C19orf12), transcript variant 5, mRNA	Homo sapiens	152	152	54%	6e-33	100.00%	4242	NM_001282929.1
☑	PREDICTED: Pan troglodytes chromosome 19 C19orf12 homolog (C19H19orf12), transcript variant X...	Pan troglodytes	150	150	54%	2e-32	100.00%	4176	XM_009435202.3
☑	PREDICTED: Hylobates moloch chromosome unknown C19orf12 homolog (CUNH19orf12), transcript...	Hylobates moloch	147	147	54%	3e-31	98.78%	3759	XM_032170324.1
☑	PREDICTED: Gorilla gorilla gorilla chromosome 19 C19orf12 homolog (C19H19orf12), transcript varia...	Gorilla gorilla g...	147	147	54%	3e-31	98.78%	3598	XM_031004299.1
☑	PREDICTED: Gorilla gorilla gorilla chromosome 19 C19orf12 homolog (C19H19orf12), transcript varia...	Gorilla gorilla g...	147	147	54%	3e-31	98.78%	3768	XM_031004297.1
☑	PREDICTED: Nomascus leucogenys chromosome 10 C19orf12 homolog (C10H19orf12), transcript v...	Nomascus leuc...	147	147	54%	3e-31	98.78%	3757	XM_030820282.1
☑	PREDICTED: Pongo abelii chromosome 19 C19orf12 homolog (C19H19orf12), transcript variant X9, ...	Pongo abelii	141	141	54%	1e-29	97.56%	2017	XM_009253106.2
☑	PREDICTED: Pongo abelii chromosome 19 C19orf12 homolog (C19H19orf12), transcript variant X2, ...	Pongo abelii	141	141	54%	1e-29	97.56%	2187	XM_009253103.2
☑	PREDICTED: Trachypithecus francoisi chromosome unknown C19orf12 homolog (CUNH19orf12), tra...	Trachypithecu...	130	130	54%	3e-26	95.12%	1637	XM_033225574.1
☑	PREDICTED: Theropithecus gelada chromosome 19 C19orf12 homolog (C19H19orf12), transcript va...	Theropithecus ...	130	130	54%	3e-26	95.12%	1532	XM_025368959.1
☑	PREDICTED: Cercocebus atys chromosome unknown open reading frame, human C19orf12 (LOC10...	Cercocebus atys	130	130	54%	3e-26	95.12%	1610	XM_012043383.1
☑	PREDICTED: Papio anubis chromosome 20 C19orf12 homolog (C20H19orf12), transcript variant X5, ...	Papio anubis	124	124	54%	1e-24	93.90%	1604	XM_009194058.4
☑	PREDICTED: Macaca mulatta chromosome 19 C19orf12 homolog (C19H19orf12), transcript variant X...	Macaca mulatta	124	124	54%	1e-24	93.90%	2485	XM_015123472.2
☑	PREDICTED: Macaca nemestrina chromosome unknown C19orf12 homolog (CUNH19orf12), transcri...	Macaca nemes...	124	124	54%	1e-24	93.90%	1628	XM_011751162.2

@A00129:562:HWNFWDSXX:1:1103:3875:24486
CATGCCTACGCCACTCCTGGGTGACAGCGAGCCAGGGCCCGGGACTCTAGTCCCAGCTCTGCTGCTTACTTGTGTGACTTCAGCCTCTCCATGCCTCAGTTTCTCCATAGGGACAAT
GACTACACTAACAGTGTCCACCCCGGCACCGGG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFF

Figure S5a: Read from Homo Sapiens from SRX7756767.



select all 42 sequences selected GenBank Graphics Distance tree of results New MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Homo sapiens kinocilin (KNCN), transcript variant 1, mRNA	Homo sapiens	278	278	100%	1e-70	100.00%	2896	NM_001322255.2
☑	Homo sapiens kinocilin (KNCN), transcript variant 2, mRNA	Homo sapiens	278	278	100%	1e-70	100.00%	2392	NM_001097611.1
☑	Homo sapiens cDNA FLJ32011 fis, clone NT2RP7009507	Homo sapiens	278	278	100%	1e-70	100.00%	2320	AK056573.1
☑	Homo sapiens kinocilin, mRNA (cDNA clone IMAGE:40024576), partial cds	Homo sapiens	278	278	100%	1e-70	100.00%	1517	BC101295.1
☑	Human DNA sequence from clone RP11-49P4 on chromosome 1p32.2-33, complete sequence	Homo sapiens	278	278	100%	1e-70	100.00%	121112	AL136373.21
☑	PREDICTED: Gorilla gorilla gorilla kinocilin (KNCN), transcript variant X2, mRNA	Gorilla gorilla gorilla	272	272	100%	5e-69	99.33%	2402	XM_004025735.3
☑	PREDICTED: Gorilla gorilla gorilla kinocilin (KNCN), transcript variant X1, mRNA	Gorilla gorilla gorilla	272	272	100%	5e-69	99.33%	2918	XM_019017965.2
☑	PREDICTED: Pan troglodytes kinocilin (KNCN), transcript variant X3, mRNA	Pan troglodytes	272	272	100%	5e-69	99.33%	3881	XM_001139695.4
☑	PREDICTED: Pan troglodytes kinocilin (KNCN), transcript variant X1, mRNA	Pan troglodytes	272	272	100%	5e-69	99.33%	3950	XM_009456787.2
☑	Pan troglodytes chromosome 1 clone CH251-68N16 map 1p33, complete sequence	Pan troglodytes	272	272	100%	5e-69	99.33%	156938	CT827813.3
☑	Homo sapiens kinocilin, mRNA (cDNA clone IMAGE:40024577), partial cds	Homo sapiens	272	272	100%	5e-69	99.33%	1517	BC101296.1
☑	PREDICTED: Hylobates moloch kinocilin (KNCN), transcript variant X2, mRNA	Hylobates moloch	250	250	100%	2e-62	96.67%	2379	XM_032159074.1
☑	PREDICTED: Hylobates moloch kinocilin (KNCN), transcript variant X1, mRNA	Hylobates moloch	250	250	100%	2e-62	96.67%	2885	XM_032159073.1
☑	PREDICTED: Nomascus leucogenys kinocilin (KNCN), transcript variant X1, mRNA	Nomascus leucogenys	250	250	100%	2e-62	96.67%	2376	XM_030823227.1
☑	PREDICTED: Macaca mulatta kinocilin (KNCN), mRNA	Macaca mulatta	250	250	100%	2e-62	96.67%	2077	XM_028848379.1
☑	PREDICTED: Theropithecus gelada kinocilin (KNCN), transcript variant X2, mRNA	Theropithecus gelada	250	250	100%	2e-62	96.67%	2408	XM_025358199.1
☑	PREDICTED: Theropithecus gelada kinocilin (KNCN), transcript variant X1, mRNA	Theropithecus gelada	250	250	100%	2e-62	96.67%	2924	XM_025358188.1
☑	PREDICTED: Macaca nemestrina kinocilin (KNCN), transcript variant X2, mRNA	Macaca nemestrina	250	250	100%	2e-62	96.67%	1700	XM_011764061.2
☑	PREDICTED: Macaca nemestrina kinocilin (KNCN), transcript variant X1, mRNA	Macaca nemestrina	250	250	100%	2e-62	96.67%	1727	XM_011764060.2
☑	PREDICTED: Macaca fascicularis kinocilin (KNCN), mRNA	Macaca fascicularis	250	250	100%	2e-62	96.67%	992	XM_005543479.2
☑	PREDICTED: Cercocebus atys kinocilin (KNCN), mRNA	Cercocebus atys	250	250	100%	2e-62	96.67%	2403	XM_012078844.1
☑	PREDICTED: Mandrillus leucophaeus kinocilin (KNCN), mRNA	Mandrillus leucophaeus	250	250	100%	2e-62	96.67%	2408	XM_011969778.1

@A00129:562:HWNFWDSXX:1:2507:3468:14888
TGGGGGTGGAGGGTCCCTGTCTGTGGTTCTTGTGGGCCTGTGGTTCACTTCTGTAGCCGGGTTAGAGCAAGGGACTGGGGCATCCTGGT
CATAGTCAGGGCCTTGGACCCCAAATTCCAGAGAAACAAATCAGTAGATTGAACAGGCAAC
+
,FFFFFFFFFFFFFFFFFF::FFFFFFFFFFFFF:FFF:FFFFFFFFFFFFFFFFFFFFFF:FFFFFF:FFFF:FFFFFFFFFFFFFF,,FFFFFFFFF,FFFF:FFF:FFF
FFF:FFFFFFFFFFF::FFFFFF:FFFFFFF:FF

Figure S5b: Read from Homo Sapiens from SRX7756768.

Human, pangolin and mouse genomic matches supporting table S4 have been deposited as Galaxy7-[Bowtie2_on_data_6_and_data_3__unaligned_reads_(L)]SRX7756767_Homo_Sapiens.fastqsanger.gz
Galaxy13-[Bowtie2_on_data_11__unaligned_reads_(L)_SRX7756767_Manis_Pentadactyla].fastqsanger.gz
And
Galaxy17-[Bowtie2_on_data_6_and_data_16__unaligned_reads_(L)]_Mus_Musculus.fastqsanger

Human genomic matches supporting figure S5b have been deposited as Galaxy499-[Bowtie2_on_data_237_and_data_494__unaligned_reads_(L)_SRX7756768]_Homo_Sapiens.fastqsanger

REFERENCES

[1] Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)?

Ping Liu ,
Jing-Zhe Jiang ,
Xiu-Feng Wan,
Yan Hua,
Linmiao Li,
Jiabin Zhou,

Xiaohu Wang,
Fanghui Hou,
Jing Chen,
Jiejian Zou,
Jinping Chen

Published: May 14, 2020
https://doi.org/10.1371/journal.ppat.1008421

[2] Xiao, K., Zhai, J., Feng, Y. *et al.* Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature* (2020). https://doi.org/10.1038/s41586-020-2313-x

[3] Lam, T.T., Shum, M.H., Zhu, H. *et al.* Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature* (2020). https://doi.org/10.1038/s41586-020-2169-0

[4] Liu, P.; Chen, W.; Chen, J.-P. Viral Metagenomics Revealed Sendai Virus and Coronavirus Infection of Malayan Pangolins (*Manis javanica*). *Viruses* **2019**, *11*, 979.

[5] Inducible epithelial resistance improves survival of Sendai virus pneumonia in mice by both inactivating virus and preventing CD8+ T cell-mediated immunopathology
S. Wali, J. R. Flores, A.M. Jaramillo, D. L. Goldblatt, J. Pantaleón García, M. J. Tuvim, B. F. Dickey, S. E. Evans
doi: https://doi.org/10.1101/2020.01.30.917195

[6] Jorlan Fernandes, Renata Carvalho de Oliveira, Alexandro Guterres, Débora Ferreira Barreto-Vieira, Ana Claudia Pereira Terças, Bernardo Rodrigues Teixeira, Marcos Alexandre Nunes da Silva, Gabriela Cardoso Caldas, Janice Mery Chicarino de Oliveira Coelho, Ortrud Monika Barth, Paulo Sergio D'Andrea, Cibele Rodrigues Bonvicino, Elba Regina Sampaio de Lemos,
Detection of Latino virus (Arenaviridae: Mammarenavirus) naturally infecting Calomys callidus,
Acta Tropica,
Volume 179,
2018,
Pages 17-24,
ISSN 0001-706X,
https://doi.org/10.1016/j.actatropica.2017.12.003.
(http://www.sciencedirect.com/science/article/pii/S0001706X17311749)

[7] Hemorrhagic Fever-Causing Arenaviruses: Lethal Pathogens and Potent Immune Suppressors
Morgan E. Brisse1,2 and Hinh Ly2,*

[8] Evolutionary origins of the SARS‐CoV‐2sarbecovirus lineage responsible for the COVID-19 pandemicMaciej F Boni1* , Philippe Lemey2* , Xiaowei Jiang3, Tommy Tsan-Yuk Lam4, Blair Perry5, Todd Castoe5, Andrew Rambaut6 and David L Robertson7

[9] Xiaolu Tang, Changcheng Wu, Xiang Li, Yuhe Song, Xinmin Yao, Xinkai Wu, Yuange Duan, Hong Zhang, Yirong Wang, Zhaohui Qian, Jie Cui, Jian Lu, On the origin and continuing evolution of SARS-CoV-2, *National Science Review*, , nwaa036, https://doi.org/10.1093/nsr/nwaa036

[10] SARS-CoV-2-like viruses from captive Guangdong pangolins generate circular RNAs
Alexandre Hassanin 1 Huw Jones 2 Anne Ropiquet 2

[11] Boratyn, G.M., Thierry-Mieg, J., Thierry-Mieg, D. et al. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. BMC Bioinformatics 20, 405 (2019). https://doi.org/10.1186/s12859-019-2996-x

[12] Jing-Yang Hu, Zi-Qian Hao, Laurent Frantz, Shi-Fang Wu, Wu Chen, Yun-Fang Jiang, Hong Wu, Wei-Min Kuang, Haipeng Li, Ya-Ping Zhang, Li Yu, Genomic consequences of population decline in critically endangered pangolins and their demographic histories, *National Science Review*, Volume 7, Issue 4, April 2020, Pages 798–814, https://doi.org/10.1093/nsr/nwaa031

[13] Lee, J., Hughes, T., Lee, MH. *et al.* No Evidence of Coronaviruses or Other Potentially Zoonotic Viruses in Sunda pangolins (*Manis javanica*) Entering the Wildlife Trade via Malaysia. *EcoHealth* **17,** 406–418 (2020). https://doi.org/10.1007/s10393-020-01503-x

[14] Hua L, Gong S, Wang F, Li W, Ge Y, Li X, Hou F. Captive breeding of pangolins: current status, problems and future prospects. Zookeys. 2015 Jun 8;(507):99-114. doi: 10.3897/zookeys.507.6970. PMID: 26155072; PMCID: PMC4490220.

[15] Wen-Hua Gao, Xian-Dan Lin, Yan-Mei Chen, Chun-Gang Xie, Zhi-Zhou Tan, Jia-Jun Zhou, Shuai Chen, Edward C Holmes, Yong-Zhen Zhang, Newly identified viral genomes in pangolins with fatal disease, *Virus Evolution*, Volume 6, Issue 1, January 2020, veaa020, https://doi.org/10.1093/ve/veaa020

[16] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, Tak-Wah Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph, *Bioinformatics*, Volume 31, Issue 10, 15 May 2015, Pages 1674–1676, https://doi.org/10.1093/bioinformatics/btv033

[17] Langmead B, Wilks C, Antonescu V, Charles R. **Scaling read aligners to hundreds of threads on general-purpose processors**. *Bioinformatics* Vol 35, Iss 3, 2019, pp 421–432.

[18] "A549 Cell Line: Human alveolar adenocarcinoma cell line -General Information". Retrieved 3 January 2012. http://a549.com/

[19] Ge, Xingyi & Li, Jia-Lu & Yang, Xinglou & Chmura, Aleksei & Zhu, Guangjian & H. Epstein, Jonathan & Mazet, Jonna & Hu, Ben & Zhang, Wei & Peng, Cheng & Zhang, Yunzhi & Luo, Chu-Ming & Tan, Bing & Wang, Ning & Zhu, Yan & Crameri, Gary & Zhang, Shu-Yi & Wang, Lin-Fa & Daszak, Peter & Shi, Zhengli. (2013). Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. Nature. 503. 10.1038/nature12711.

[20] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357-9. doi: 10.1038/nmeth.1923. PMID: 22388286; PMCID: PMC3322381.

[21] Boratyn, G.M., Thierry-Mieg, J., Thierry-Mieg, D. et al. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. BMC Bioinformatics 20, 405 (2019). https://doi.org/10.1186/s12859-019-2996-x