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

Daoyu Zhang.

ABSTRACT

Recently, There were much hype about an alleged SARS-like coronavirus being found in samples of Malayan pangolins (Manis Javanica) possessing nearly identical RBD to the SARS-CoV-2 coronavirus. Prominent journals cite the alleged discovery to claim that pangolins may be one of 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 in serious risk of contamination. Here we also report that the presence of 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 an 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.

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 under NCBI GenBank. By Accession number, size and citation by thesis (if claimed to have SARS-CoV-2 related reads by paper).

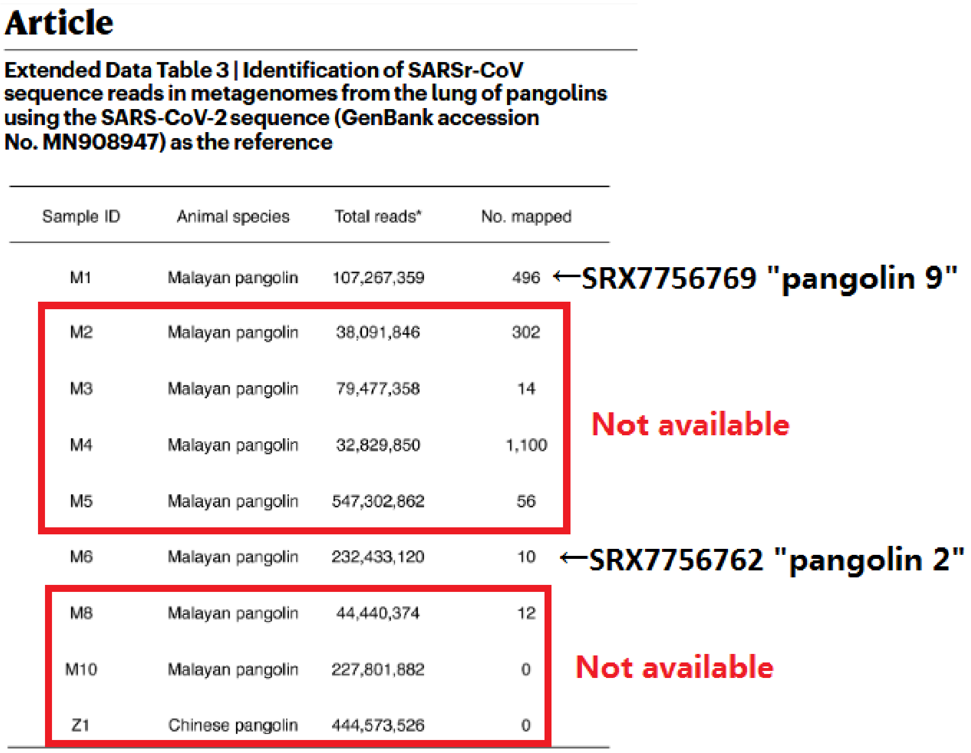
Accession number	Size	SARS-CoV-2-like Coronavirus Identified and Cited?
SRX6893158	16,491,648	
SRX6893157	9,275,501	Lung12 [3] SRR10168374
SRX6893156	22,220,187	Lung11 [1]
SRX6893155	18,067,615	Lung09 [1] [3] SRR10168376
SRX6893154	16,414,925	Lung08 [1] [3] [4] SRR10168377
SRX6893153	19,045,923	Lung07 [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
M2	Malayan pangolin	38,091,846	302
M3	Malayan pangolin	79,477,358	14
M4	Malayan pangolin	32,829,850	1,100
M5	Malayan pangolin	547,302,862	56
M6	Malayan pangolin	232,433,120	10
M8	Malayan pangolin	44,440,374	12
M10	Malayan pangolin	227,801,882	0
Z1	Chinese pangolin	444,573,526	0

←SRX7756769 "pangolin 9"

Not available

←SRX7756762 "pangolin 2"

Not available

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	Manis javanica: **1.28%**	Simiiformes 57084	N/D

20-Sep-2019			
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

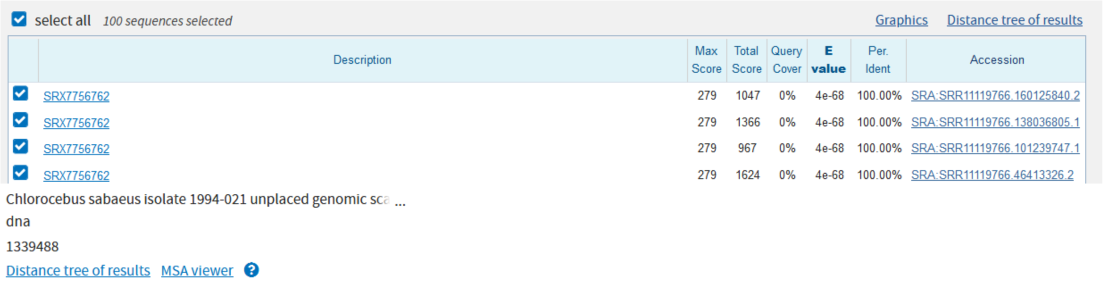
**:Not claimed as being SARS-CoV-2 related in the original publication. Likely unrelated.

***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.

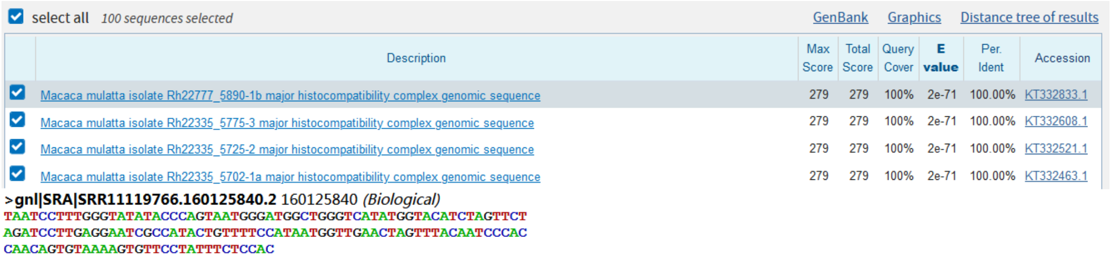


☑ select all *100 sequences selected* Graphics Distance tree of results

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX7756762	279	1047	0%	4e-68	100.00%	SRA:SRR11119766.160125840.2
☑	SRX7756762	279	1366	0%	4e-68	100.00%	SRA:SRR11119766.138036805.1
☑	SRX7756762	279	967	0%	4e-68	100.00%	SRA:SRR11119766.101239747.1
☑	SRX7756762	279	1624	0%	4e-68	100.00%	SRA:SRR11119766.46413326.2

Chlorocebus sabaeus isolate 1994-021 unplaced genomic sca ...
dna
1339488
Distance tree of results MSA viewer

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.

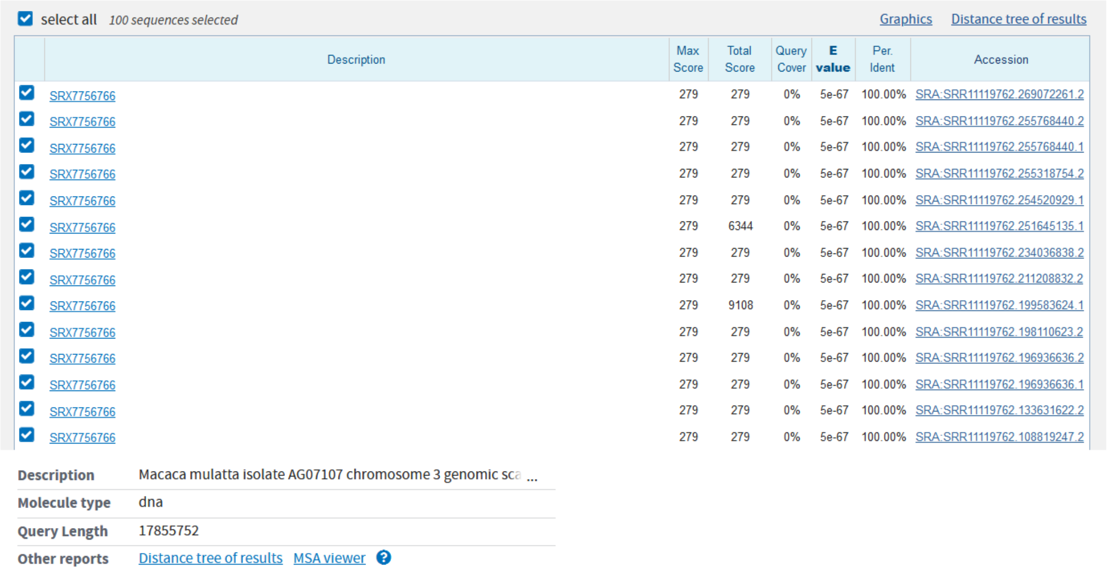


☑ select all *100 sequences selected* GenBank Graphics Distance tree of results

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Macaca mulatta isolate Rh22777_5890-1b major histocompatibility complex genomic sequence	279	279	100%	2e-71	100.00%	KT332833.1
☑	Macaca mulatta isolate Rh22335_5775-3 major histocompatibility complex genomic sequence	279	279	100%	2e-71	100.00%	KT332608.1
☑	Macaca mulatta isolate Rh22335_5725-2 major histocompatibility complex genomic sequence	279	279	100%	2e-71	100.00%	KT332521.1
☑	Macaca mulatta isolate Rh22335_5702-1a major histocompatibility complex genomic sequence	279	279	100%	2e-71	100.00%	KT332463.1

```
>gnl|SRA|SRR11119766.160125840.2 160125840 (Biological)
TAATCCTTTGGGTATATACCCAGTAATGGGATGGCTGGGTCATATGGTACATCTAGTTCT
AGATCCTTGAGGAATCGCCATACTGTTTTCCATAATGGTTGAACTAGTTTACAATCCCAC
CAACAGTGTAAAAGTGTTCCTATTTCTCCAC
```

Fig.2b BLAST result on the returned sequence revealed it as a Primate-derived MHC complex gene, confirming Primate origin.

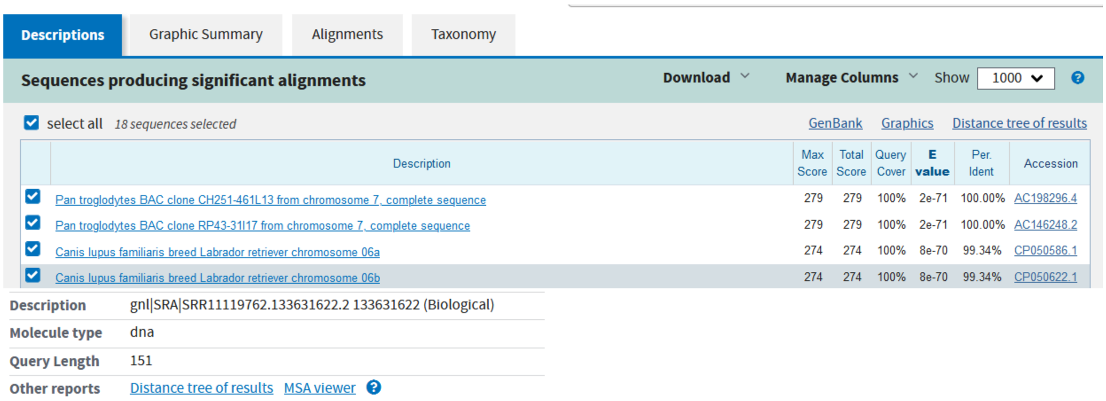


☑ select all *100 sequences selected* Graphics Distance tree of results

	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 sca ...
Molecule type	dna
Query Length	17855752
Other reports	Distance tree of results MSA viewer

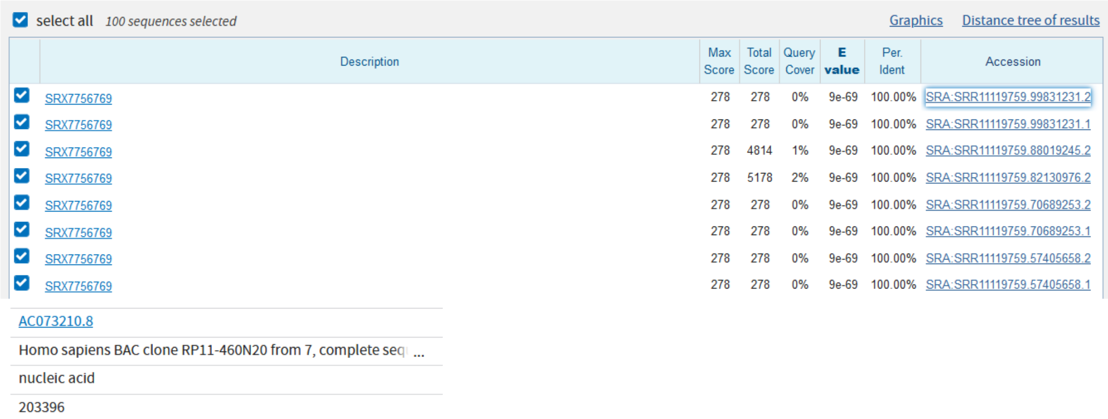
Fig.3a Specific BLAST analysis of **SRX7756766** revealed large amount of 100% full-length matches with Macaca Mulatta.



	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Pan troglodytes BAC clone CH251-461L13 from chromosome 7, complete sequence	279	279	100%	2e-71	100.00%	AC198296.4
☑	Pan troglodytes BAC clone RP43-31I17 from chromosome 7, complete sequence	279	279	100%	2e-71	100.00%	AC146248.2
☑	Canis lupus familiaris breed Labrador retriever chromosome 06a	274	274	100%	8e-70	99.34%	CP050586.1
☑	Canis lupus familiaris breed Labrador retriever chromosome 06b	274	274	100%	8e-70	99.34%	CP050622.1

Description	gnl|SRA|SRR11119762.133631622.2 133631622 (Biological)
Molecule type	dna
Query Length	151
Other reports	Distance tree of results MSA viewer

Fig.3b More intriguing—many of the reads showed only 100% matches to hominids—Chimpanzees and also clearly Macaca Mulatta itself. This indicate that SRX7756766 also contained significant amount of material derived from primates.



	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX7756769	278	278	0%	9e-69	100.00%	SRA:SRR11119759.99831231.2
☑	SRX7756769	278	278	0%	9e-69	100.00%	SRA:SRR11119759.99831231.1
☑	SRX7756769	278	4814	1%	9e-69	100.00%	SRA:SRR11119759.88019245.2
☑	SRX7756769	278	5178	2%	9e-69	100.00%	SRA:SRR11119759.82130976.2
☑	SRX7756769	278	278	0%	9e-69	100.00%	SRA:SRR11119759.70689253.2
☑	SRX7756769	278	278	0%	9e-69	100.00%	SRA:SRR11119759.70689253.1
☑	SRX7756769	278	278	0%	9e-69	100.00%	SRA:SRR11119759.57405658.2
☑	SRX7756769	278	278	0%	9e-69	100.00%	SRA:SRR11119759.57405658.1

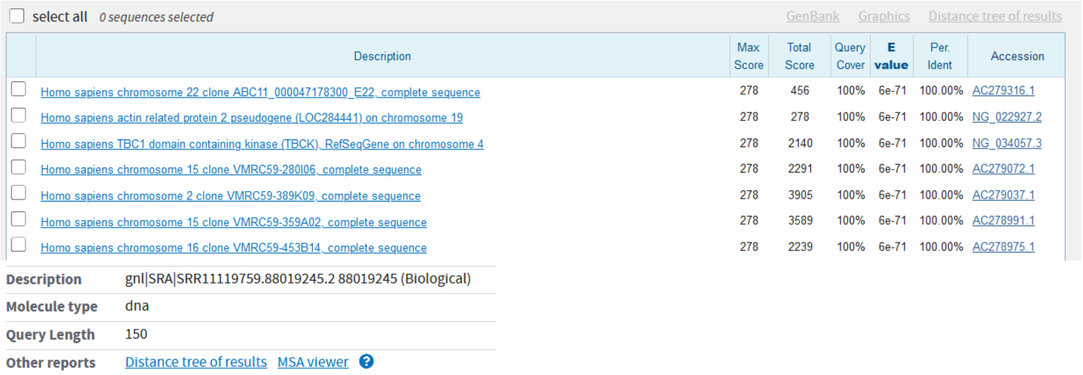
AC073210.8

Homo sapiens BAC clone RP11-460N20 from 7, complete seq ...

nucleic acid

203396

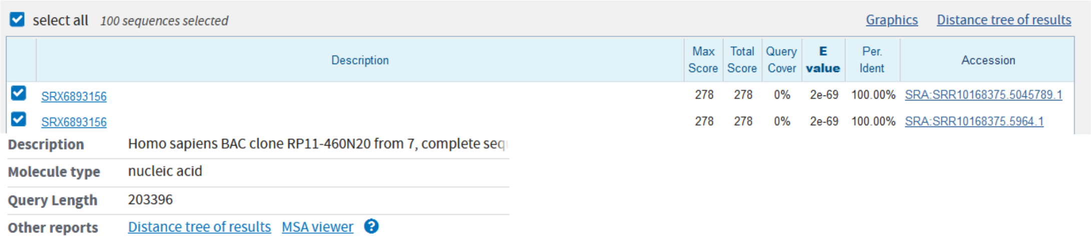
Fig.4a Similarly, SRX7756769 contained large amount of reads that are 100% full-length matches to Human genomic DNA.



	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
□	Homo sapiens chromosome 22 clone ABC11_000047178300_E22, complete sequence	278	456	100%	6e-71	100.00%	AC279316.1
□	Homo sapiens actin related protein 2 pseudogene (LOC284441) on chromosome 19	278	278	100%	6e-71	100.00%	NG_022927.2
□	Homo sapiens TBC1 domain containing kinase (TBCK), RefSeqGene on chromosome 4	278	2140	100%	6e-71	100.00%	NG_034057.3
□	Homo sapiens chromosome 15 clone VMRC59-280I06, complete sequence	278	2291	100%	6e-71	100.00%	AC279072.1
□	Homo sapiens chromosome 2 clone VMRC59-389K09, complete sequence	278	3905	100%	6e-71	100.00%	AC279037.1
□	Homo sapiens chromosome 15 clone VMRC59-359A02, complete sequence	278	3589	100%	6e-71	100.00%	AC278991.1
□	Homo sapiens chromosome 16 clone VMRC59-453B14, complete sequence	278	2239	100%	6e-71	100.00%	AC278975.1

Description	gnl|SRA|SRR11119759.88019245.2 88019245 (Biological)
Molecule type	dna
Query Length	150
Other reports	Distance tree of results MSA viewer

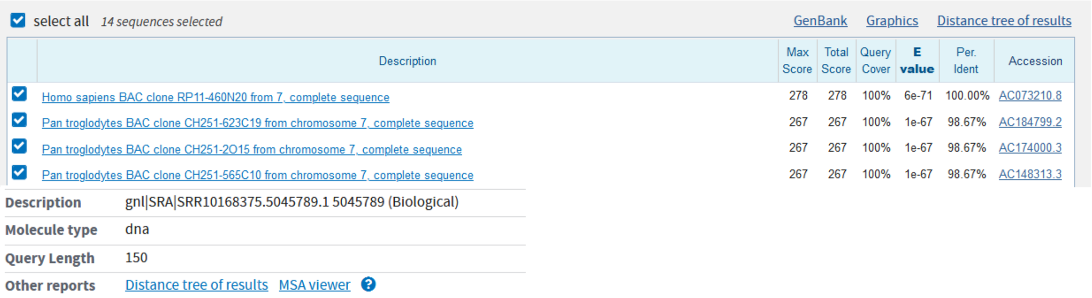
Fig.4b A BLAST analysis on reads sampled from the 100% hit results confirmed that it was found only in humans. Once again confirming human origin.



	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX6893156	278	278	0%	2e-69	100.00%	SRA:SRR10168375.5045789.1
☑	SRX6893156	278	278	0%	2e-69	100.00%	SRA:SRR10168375.5964.1

Description	Homo sapiens BAC clone RP11-460N20 from 7, complete seq
Molecule type	nucleic acid
Query Length	203396
Other reports	Distance tree of results MSA viewer

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

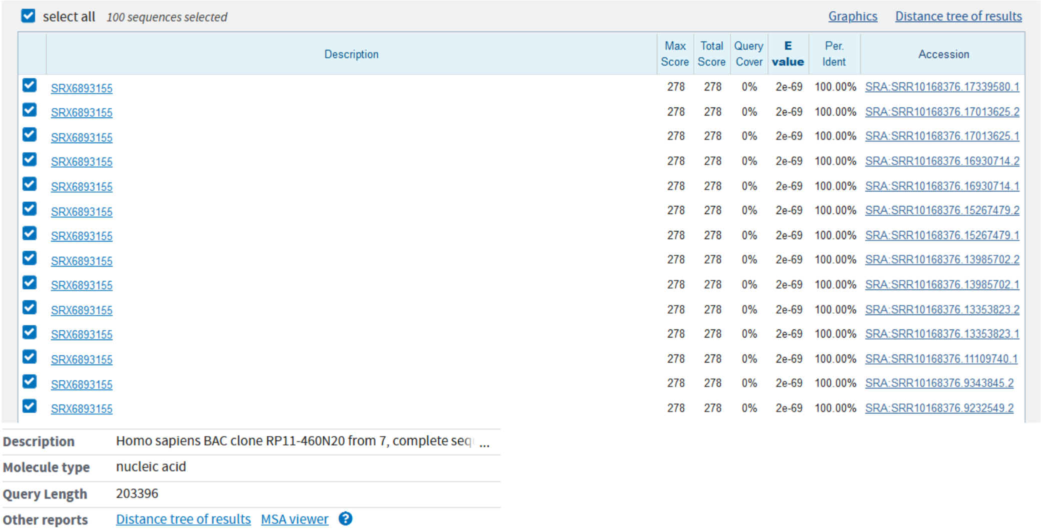


☑ select all *14 sequences selected* GenBank Graphics Distance tree of results

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Homo sapiens BAC clone RP11-460N20 from 7, complete sequence	278	278	100%	6e-71	100.00%	AC073210.8
☑	Pan troglodytes BAC clone CH251-623C19 from chromosome 7, complete sequence	267	267	100%	1e-67	98.67%	AC184799.2
☑	Pan troglodytes BAC clone CH251-2O15 from chromosome 7, complete sequence	267	267	100%	1e-67	98.67%	AC174000.3
☑	Pan troglodytes BAC clone CH251-565C10 from chromosome 7, complete sequence	267	267	100%	1e-67	98.67%	AC148313.3

Description	gnl|SRA|SRR10168375.5045789.1 5045789 (Biological)
Molecule type	dna
Query Length	150
Other reports	Distance tree of results MSA viewer ?

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

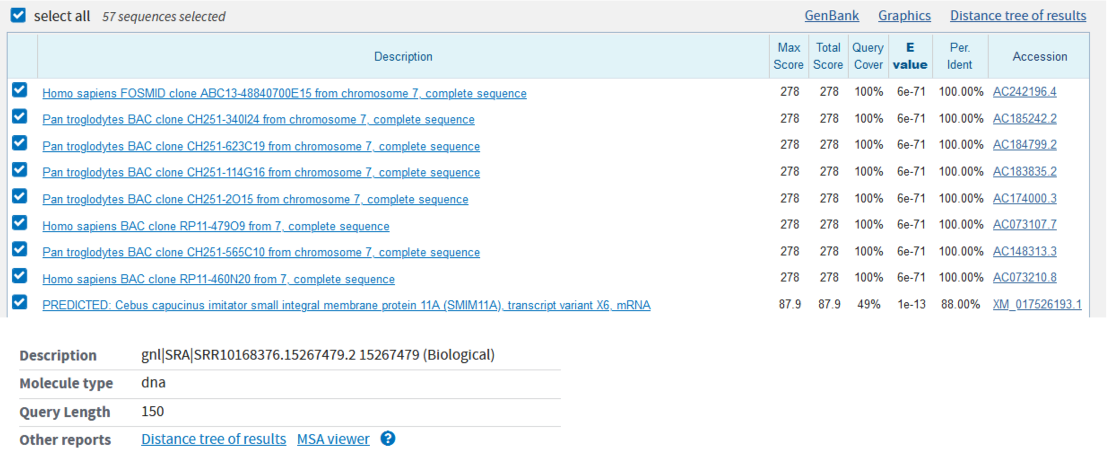


☑ select all *100 sequences selected* Graphics Distance tree of results

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX6893155	278	278	0%	2e-69	100.00%	SRA:SRR10168376.17339580.1
☑	SRX6893155	278	278	0%	2e-69	100.00%	SRA:SRR10168376.17013625.2
☑	SRX6893155	278	278	0%	2e-69	100.00%	SRA:SRR10168376.17013625.1
☑	SRX6893155	278	278	0%	2e-69	100.00%	SRA:SRR10168376.16930714.2
☑	SRX6893155	278	278	0%	2e-69	100.00%	SRA:SRR10168376.16930714.1
☑	SRX6893155	278	278	0%	2e-69	100.00%	SRA:SRR10168376.15267479.2
☑	SRX6893155	278	278	0%	2e-69	100.00%	SRA:SRR10168376.15267479.1
☑	SRX6893155	278	278	0%	2e-69	100.00%	SRA:SRR10168376.13985702.2
☑	SRX6893155	278	278	0%	2e-69	100.00%	SRA:SRR10168376.13985702.1
☑	SRX6893155	278	278	0%	2e-69	100.00%	SRA:SRR10168376.13353823.2
☑	SRX6893155	278	278	0%	2e-69	100.00%	SRA:SRR10168376.13353823.1
☑	SRX6893155	278	278	0%	2e-69	100.00%	SRA:SRR10168376.11109740.1
☑	SRX6893155	278	278	0%	2e-69	100.00%	SRA:SRR10168376.9343845.2
☑	SRX6893155	278	278	0%	2e-69	100.00%	SRA:SRR10168376.9232549.2

Description	Homo sapiens BAC clone RP11-460N20 from 7, complete seq ...
Molecule type	nucleic acid
Query Length	203396
Other reports	Distance tree of results MSA viewer ?

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

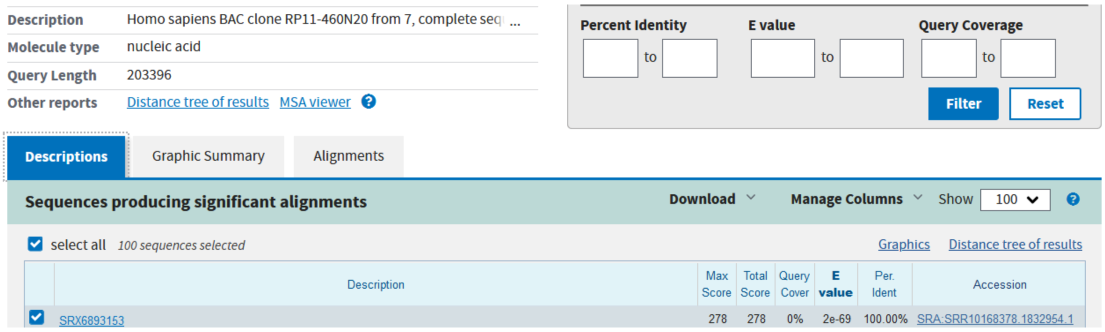


☑ select all *57 sequences selected* GenBank Graphics Distance tree of results

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Homo sapiens FOSMID clone ABC13-48840700E15 from chromosome 7, complete sequence	278	278	100%	6e-71	100.00%	AC242196.4
☑	Pan troglodytes BAC clone CH251-340I24 from chromosome 7, complete sequence	278	278	100%	6e-71	100.00%	AC185242.2
☑	Pan troglodytes BAC clone CH251-623C19 from chromosome 7, complete sequence	278	278	100%	6e-71	100.00%	AC184799.2
☑	Pan troglodytes BAC clone CH251-114G16 from chromosome 7, complete sequence	278	278	100%	6e-71	100.00%	AC183835.2
☑	Pan troglodytes BAC clone CH251-2O15 from chromosome 7, complete sequence	278	278	100%	6e-71	100.00%	AC174000.3
☑	Homo sapiens BAC clone RP11-479O9 from 7, complete sequence	278	278	100%	6e-71	100.00%	AC073107.7
☑	Pan troglodytes BAC clone CH251-565C10 from chromosome 7, complete sequence	278	278	100%	6e-71	100.00%	AC148313.3
☑	Homo sapiens BAC clone RP11-460N20 from 7, complete sequence	278	278	100%	6e-71	100.00%	AC073210.8
☑	PREDICTED: Cebus capucinus imitator small integral membrane protein 11A (SMIM11A), transcript variant X6, mRNA	87.9	87.9	49%	1e-13	88.00%	XM_017526193.1

Description	gnl|SRA|SRR10168376.15267479.2 15267479 (Biological)
Molecule type	dna
Query Length	150
Other reports	Distance tree of results MSA viewer ?

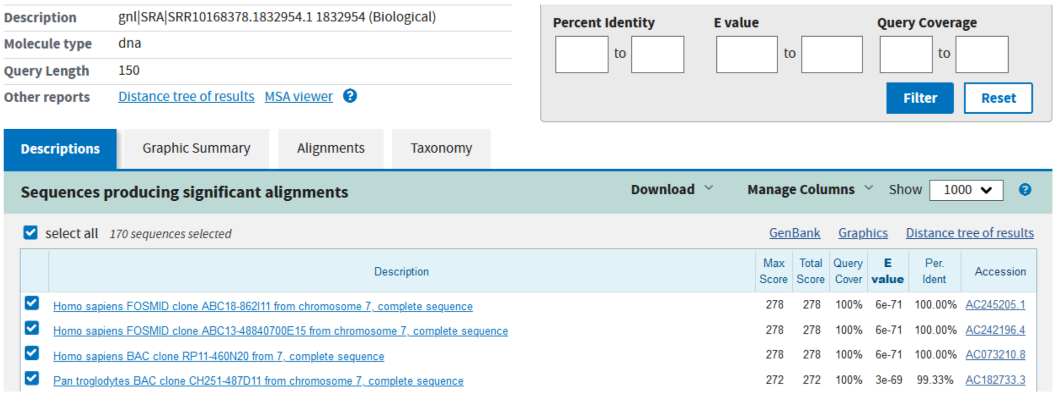
Fig.6b The results, when put through BLAST, confirms that the 100% matches are in fact derived from a Hominid origin.



Description	Homo sapiens BAC clone RP11-460N20 from 7, complete seq ...
Molecule type	nucleic acid
Query Length	203396
Other reports	Distance tree of results MSA viewer

☑	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX6893153	278	278	0%	2e-69	100.00%	SRA:SRR10168378.1832954.1

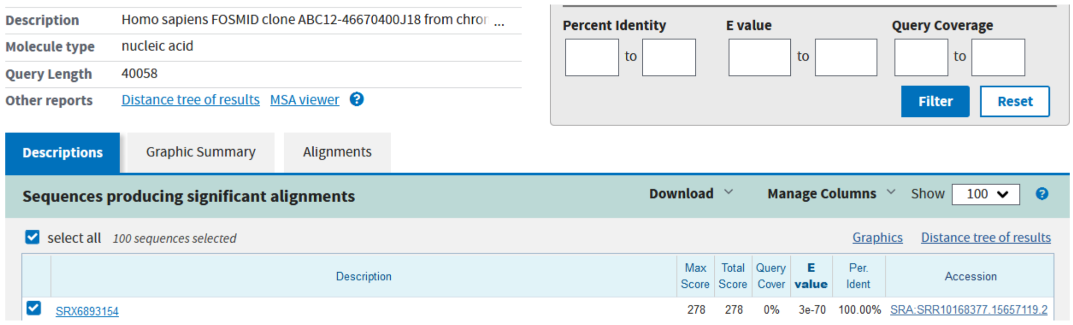
Fig.7a SRX6893153 have also returned 100% match full-length read on this tiny part of the human genome.



Description	gnl|SRA|SRR10168378.1832954.1 1832954 (Biological)
Molecule type	dna
Query Length	150
Other reports	Distance tree of results MSA viewer

☑	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Homo sapiens FOSMID clone ABC18-862I11 from chromosome 7, complete sequence	278	278	100%	6e-71	100.00%	AC245205.1
☑	Homo sapiens FOSMID clone ABC13-48840700E15 from chromosome 7, complete sequence	278	278	100%	6e-71	100.00%	AC242196.4
☑	Homo sapiens BAC clone RP11-460N20 from 7, complete sequence	278	278	100%	6e-71	100.00%	AC073210.8
☑	Pan troglodytes BAC clone CH251-487D11 from chromosome 7, complete sequence	272	272	100%	3e-69	99.33%	AC182733.3

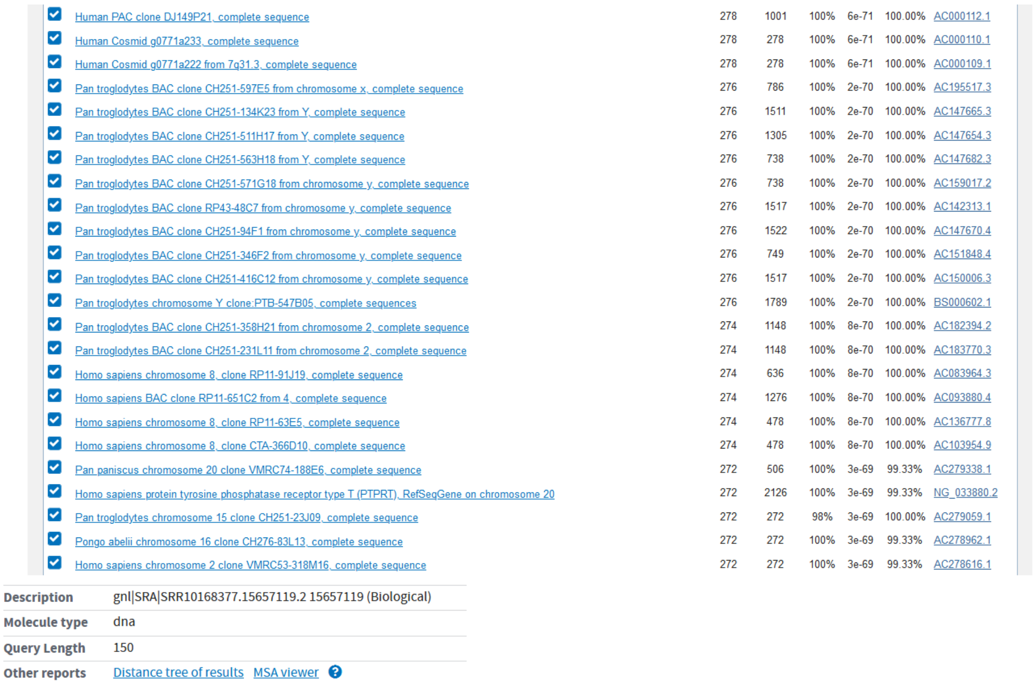
Fig.7b Similarly, the read is only found in humans—indicating the Homo Sapiens Trace result is accurate.



Description	Homo sapiens FOSMID clone ABC12-46670400J18 from chror ...
Molecule type	nucleic acid
Query Length	40058
Other reports	Distance tree of results MSA viewer

☑	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX6893154	278	278	0%	3e-70	100.00%	SRA:SRR10168377.15657119.2

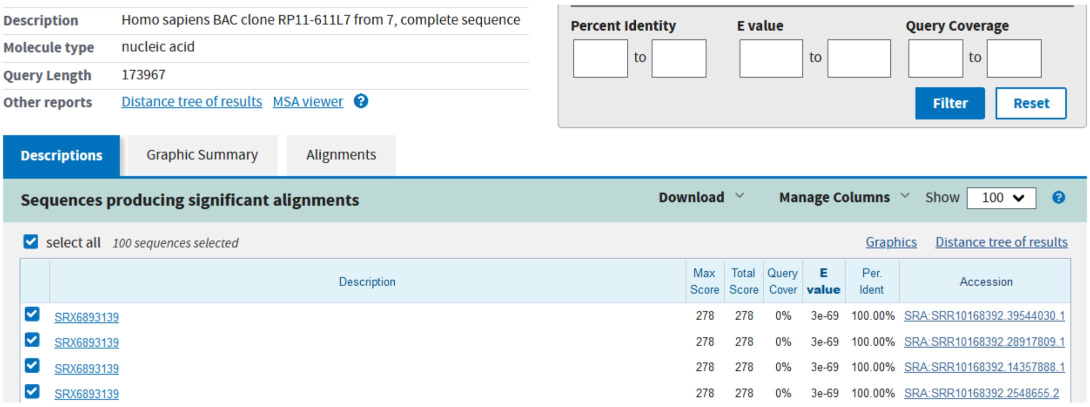
Fig.8a One read from the Human MHC gene is recovered from SRX6893154 with a query sequence only 40058bp in length.



☑	Human PAC clone DJ149P21, complete sequence	278	1001	100%	6e-71	100.00%	AC000112.1
☑	Human Cosmid g0771a233, complete sequence	278	278	100%	6e-71	100.00%	AC000110.1
☑	Human Cosmid g0771a222 from 7q31.3, complete sequence	278	278	100%	6e-71	100.00%	AC000109.1
☑	Pan troglodytes BAC clone CH251-597E5 from chromosome x, complete sequence	276	786	100%	2e-70	100.00%	AC195517.3
☑	Pan troglodytes BAC clone CH251-134K23 from Y, complete sequence	276	1511	100%	2e-70	100.00%	AC147665.3
☑	Pan troglodytes BAC clone CH251-511H17 from Y, complete sequence	276	1305	100%	2e-70	100.00%	AC147654.3
☑	Pan troglodytes BAC clone CH251-563H18 from Y, complete sequence	276	738	100%	2e-70	100.00%	AC147682.3
☑	Pan troglodytes BAC clone CH251-571G18 from chromosome y, complete sequence	276	738	100%	2e-70	100.00%	AC159017.2
☑	Pan troglodytes BAC clone RP43-48C7 from chromosome y, complete sequence	276	1517	100%	2e-70	100.00%	AC142313.1
☑	Pan troglodytes BAC clone CH251-94F1 from chromosome y, complete sequence	276	1522	100%	2e-70	100.00%	AC147670.4
☑	Pan troglodytes BAC clone CH251-346F2 from chromosome y, complete sequence	276	749	100%	2e-70	100.00%	AC151848.4
☑	Pan troglodytes BAC clone CH251-416C12 from chromosome y, complete sequence	276	1517	100%	2e-70	100.00%	AC150006.3
☑	Pan troglodytes chromosome Y clone:PTB-547B05, complete sequences	276	1789	100%	2e-70	100.00%	BS000602.1
☑	Pan troglodytes BAC clone CH251-358H21 from chromosome 2, complete sequence	274	1148	100%	8e-70	100.00%	AC182394.2
☑	Pan troglodytes BAC clone CH251-231L11 from chromosome 2, complete sequence	274	1148	100%	8e-70	100.00%	AC183770.3
☑	Homo sapiens chromosome 8, clone RP11-91J19, complete sequence	274	636	100%	8e-70	100.00%	AC083964.3
☑	Homo sapiens BAC clone RP11-651C2 from 4, complete sequence	274	1276	100%	8e-70	100.00%	AC093880.4
☑	Homo sapiens chromosome 8, clone RP11-63E5, complete sequence	274	478	100%	8e-70	100.00%	AC136777.8
☑	Homo sapiens chromosome 8, clone CTA-366D10, complete sequence	274	478	100%	8e-70	100.00%	AC103954.9
☑	Pan paniscus chromosome 20 clone VMRC74-188E6, complete sequence	272	506	100%	3e-69	99.33%	AC279338.1
☑	Homo sapiens protein tyrosine phosphatase receptor type T (PTPRT), RefSeqGene on chromosome 20	272	2126	100%	3e-69	99.33%	NG_033880.2
☑	Pan troglodytes chromosome 15 clone CH251-23J09, complete sequence	272	272	98%	3e-69	100.00%	AC279059.1
☑	Pongo abelii chromosome 16 clone CH276-83L13, complete sequence	272	272	100%	3e-69	99.33%	AC278962.1
☑	Homo sapiens chromosome 2 clone VMRC53-318M16, complete sequence	272	272	100%	3e-69	99.33%	AC278616.1

Description	gnl|SRA|SRR10168377.15657119.2 15657119 (Biological)
Molecule type	dna
Query Length	150
Other reports	Distance tree of results MSA viewer

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



Description	Homo sapiens BAC clone RP11-611L7 from 7, complete sequence
Molecule type	nucleic acid
Query Length	173967
Other reports	Distance tree of results MSA viewer

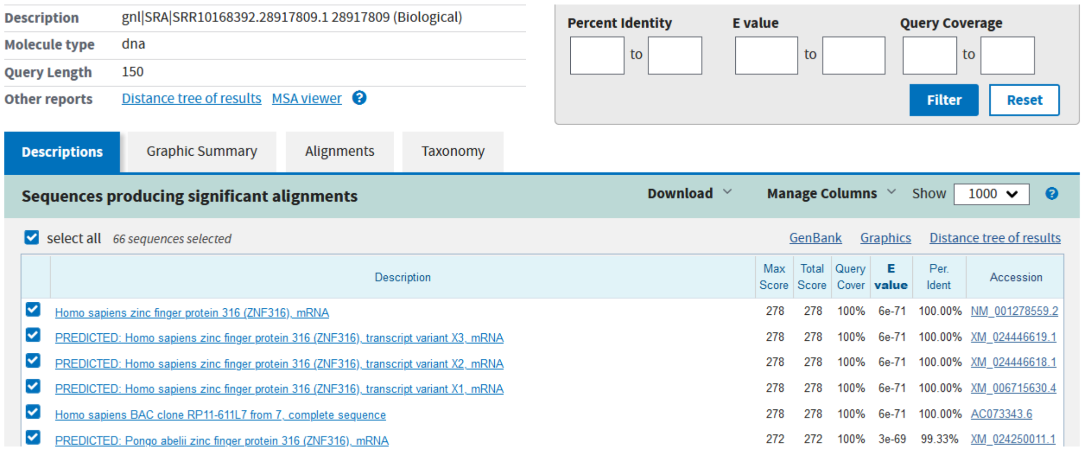
Percent Identity to E value to Query Coverage to

Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX6893139	278	278	0%	3e-69	100.00%	SRA:SRR10168392.39544030.1
☑	SRX6893139	278	278	0%	3e-69	100.00%	SRA:SRR10168392.28917809.1
☑	SRX6893139	278	278	0%	3e-69	100.00%	SRA:SRR10168392.14357888.1
☑	SRX6893139	278	278	0%	3e-69	100.00%	SRA:SRR10168392.2548655.2

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.



Description	gnl|SRA|SRR10168392.28917809.1 28917809 (Biological)
Molecule type	dna
Query Length	150
Other reports	Distance tree of results MSA viewer

Percent Identity to E value to Query Coverage to

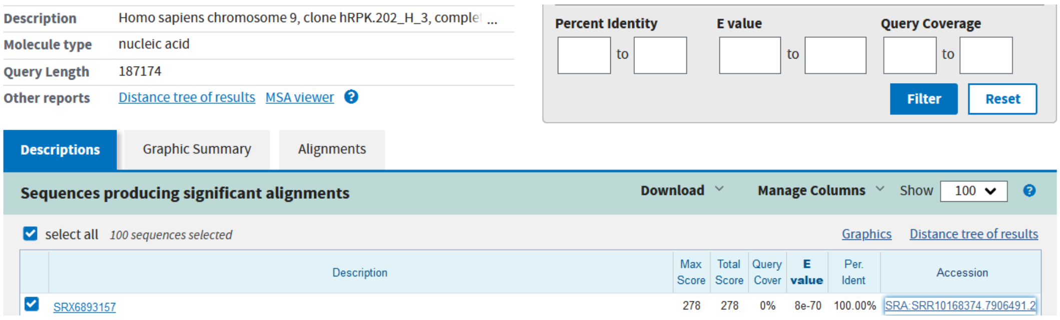
Sequences producing significant alignments

☑ select all *66 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Homo sapiens zinc finger protein 316 (ZNF316), mRNA	278	278	100%	6e-71	100.00%	NM_001278559.2
☑	PREDICTED: Homo sapiens zinc finger protein 316 (ZNF316), transcript variant X3, mRNA	278	278	100%	6e-71	100.00%	XM_024446619.1
☑	PREDICTED: Homo sapiens zinc finger protein 316 (ZNF316), transcript variant X2, mRNA	278	278	100%	6e-71	100.00%	XM_024446618.1
☑	PREDICTED: Homo sapiens zinc finger protein 316 (ZNF316), transcript variant X1, mRNA	278	278	100%	6e-71	100.00%	XM_006715630.4
☑	Homo sapiens BAC clone RP11-611L7 from 7, complete sequence	278	278	100%	6e-71	100.00%	AC073343.6
☑	PREDICTED: Pongo abelii zinc finger protein 316 (ZNF316), mRNA	272	272	100%	3e-69	99.33%	XM_024250011.1

Fig.9b Examining these reads revealed that they are only found in humans and apes. This is

therefore also clear evidence that there are Human/Hominid-derived contamination in SRX6893139.



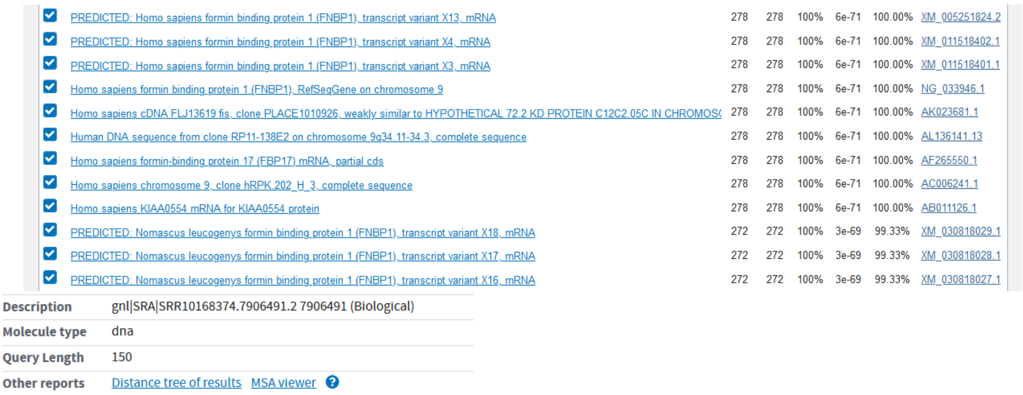
Description: Homo sapiens chromosome 9, clone hRPK.202_H_3, comple ...
Molecule type: nucleic acid
Query Length: 187174
Other reports: Distance tree of results MSA viewer

Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX6893157	278	278	0%	8e-70	100.00%	SRA:SRR10168374.7906491.2

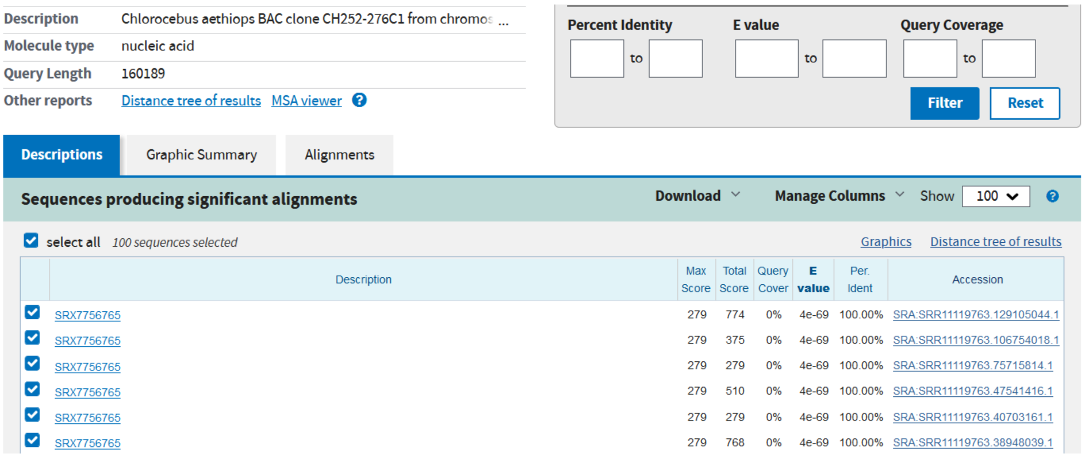
Fig.10a One read is also recovered from SRX6893157. From a query sequence only 187174nt in length.



☑	PREDICTED: Homo sapiens formin binding protein 1 (FNBP1), transcript variant X13, mRNA	278	278	100%	6e-71	100.00%	XM_005251824.2
☑	PREDICTED: Homo sapiens formin binding protein 1 (FNBP1), transcript variant X4, mRNA	278	278	100%	6e-71	100.00%	XM_011518402.1
☑	PREDICTED: Homo sapiens formin binding protein 1 (FNBP1), transcript variant X3, mRNA	278	278	100%	6e-71	100.00%	XM_011518401.1
☑	Homo sapiens formin binding protein 1 (FNBP1), RefSeqGene on chromosome 9	278	278	100%	6e-71	100.00%	NG_033946.1
☑	Homo sapiens cDNA FLJ13619 fis, clone PLACE1010926, weakly similar to HYPOTHETICAL 72.2 KD PROTEIN C12C2.05C IN CHROMOS(	278	278	100%	6e-71	100.00%	AK023681.1
☑	Human DNA sequence from clone RP11-138E2 on chromosome 9q34.11-34.3, complete sequence	278	278	100%	6e-71	100.00%	AL136141.13
☑	Homo sapiens formin-binding protein 17 (FBP17) mRNA, partial cds	278	278	100%	6e-71	100.00%	AF265550.1
☑	Homo sapiens chromosome 9, clone hRPK.202_H_3, complete sequence	278	278	100%	6e-71	100.00%	AC006241.1
☑	Homo sapiens KIAA0554 mRNA for KIAA0554 protein	278	278	100%	6e-71	100.00%	AB011126.1
☑	PREDICTED: Nomascus leucogenys formin binding protein 1 (FNBP1), transcript variant X18, mRNA	272	272	100%	3e-69	99.33%	XM_030818029.1
☑	PREDICTED: Nomascus leucogenys formin binding protein 1 (FNBP1), transcript variant X17, mRNA	272	272	100%	3e-69	99.33%	XM_030818028.1
☑	PREDICTED: Nomascus leucogenys formin binding protein 1 (FNBP1), transcript variant X16, mRNA	272	272	100%	3e-69	99.33%	XM_030818027.1

Description: gnl|SRA|SRR10168374.7906491.2 7906491 (Biological)
Molecule type: dna
Query Length: 150
Other reports: Distance tree of results MSA viewer

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



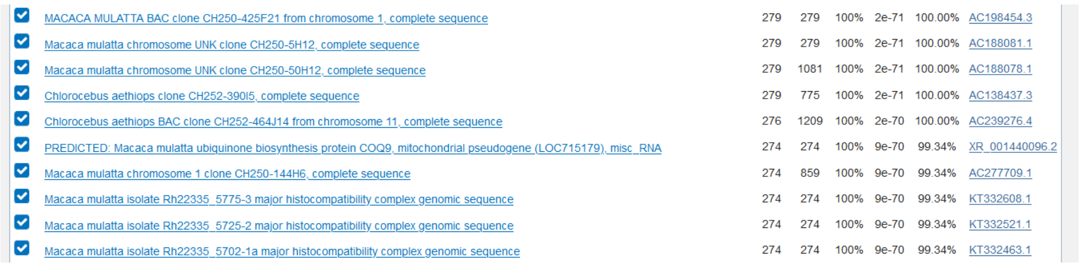
Description: Chlorocebus aethiops BAC clone CH252-276C1 from chromos ...
Molecule type: nucleic acid
Query Length: 160189
Other reports: Distance tree of results MSA viewer

Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX7756765	279	774	0%	4e-69	100.00%	SRA:SRR11119763.129105044.1
☑	SRX7756765	279	375	0%	4e-69	100.00%	SRA:SRR11119763.106754018.1
☑	SRX7756765	279	279	0%	4e-69	100.00%	SRA:SRR11119763.75715814.1
☑	SRX7756765	279	510	0%	4e-69	100.00%	SRA:SRR11119763.47541416.1
☑	SRX7756765	279	279	0%	4e-69	100.00%	SRA:SRR11119763.40703161.1
☑	SRX7756765	279	768	0%	4e-69	100.00%	SRA:SRR11119763.38948039.1

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



☑	MACACA MULATTA BAC clone CH250-425F21 from chromosome 1, complete sequence	279	279	100%	2e-71	100.00%	AC198454.3
☑	Macaca mulatta chromosome UNK clone CH250-5H12, complete sequence	279	279	100%	2e-71	100.00%	AC188081.1
☑	Macaca mulatta chromosome UNK clone CH250-50H12, complete sequence	279	1081	100%	2e-71	100.00%	AC188078.1
☑	Chlorocebus aethiops clone CH252-390I5, complete sequence	279	775	100%	2e-71	100.00%	AC138437.3
☑	Chlorocebus aethiops BAC clone CH252-464J14 from chromosome 11, complete sequence	276	1209	100%	2e-70	100.00%	AC239276.4
☑	PREDICTED: Macaca mulatta ubiquinone biosynthesis protein COQ9, mitochondrial pseudogene (LOC715179), misc_RNA	274	274	100%	9e-70	99.34%	XR_001440096.2
☑	Macaca mulatta chromosome 1 clone CH250-144H6, complete sequence	274	859	100%	9e-70	99.34%	AC277709.1
☑	Macaca mulatta isolate Rh22335_5775-3 major histocompatibility complex genomic sequence	274	274	100%	9e-70	99.34%	KT332608.1
☑	Macaca mulatta isolate Rh22335_5725-2 major histocompatibility complex genomic sequence	274	274	100%	9e-70	99.34%	KT332521.1
☑	Macaca mulatta isolate Rh22335_5702-1a major histocompatibility complex genomic sequence	274	274	100%	9e-70	99.34%	KT332463.1

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.

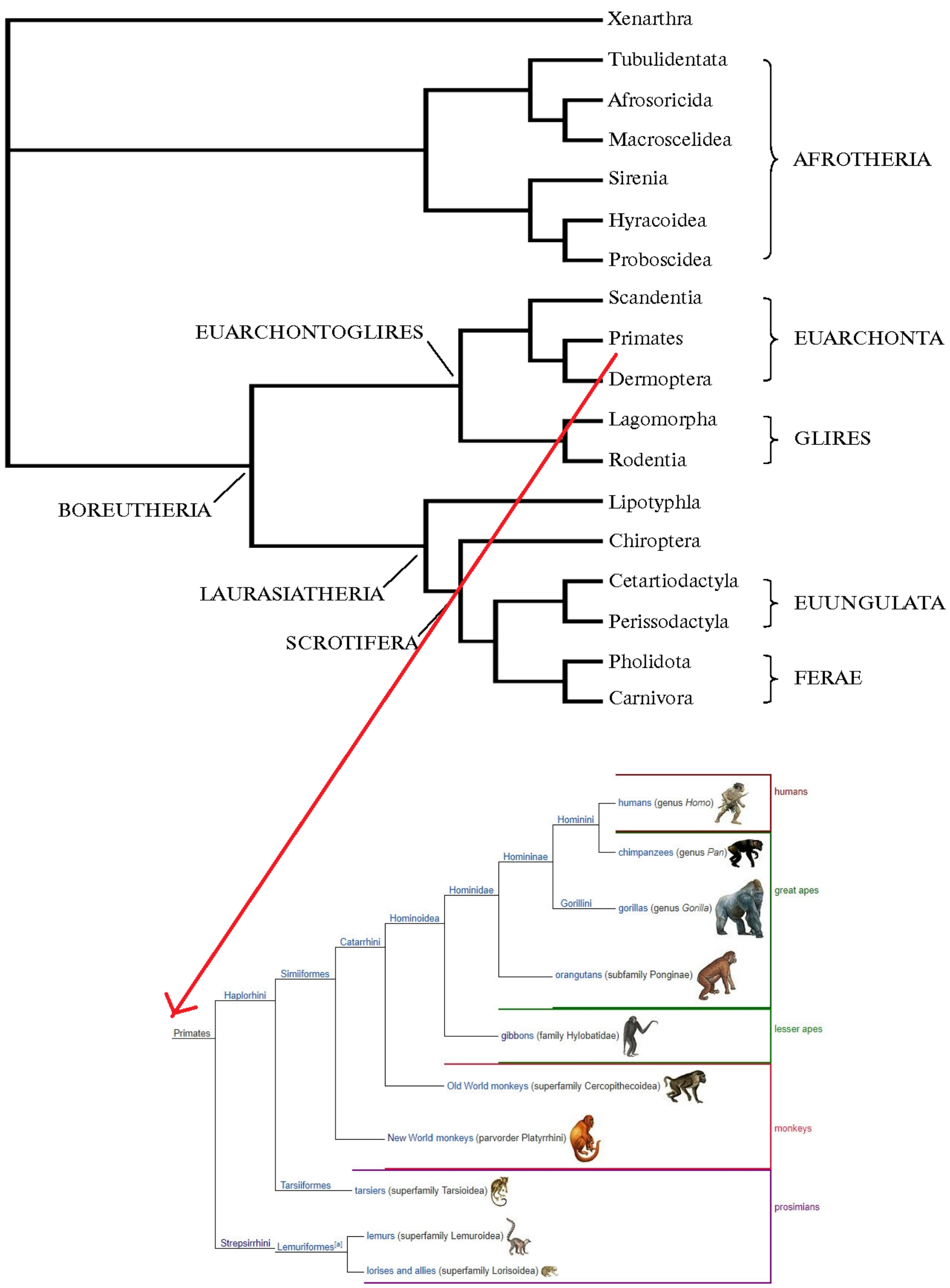


Fig. 12 Family tree of mammals, Including the position and classification of Primates in the lineage of Mammalia.

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
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]

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 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 **PRJNA696875**, 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

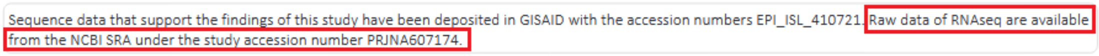
Accession number and date	Size	Non-Coronavirus reads?	Source Tissue Provider and Collected by	Virus Designation: GD or GX?	Design
SRX7732094 15-Feb-2020	2,633	No	N/A	GD	This dataset contains coronavirus-like sequence reads, based on BLAST search.
SRX7732093 15-Feb-2020	470,344	Yes	Intestine Yanling Hu 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]



Sequence data that support the findings of this study have been deposited in GISAID with the accession numbers EPI_ISL_410721. Raw data of RNAseq are available from the NCBI SRA under the study accession number PRJNA607174.

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.

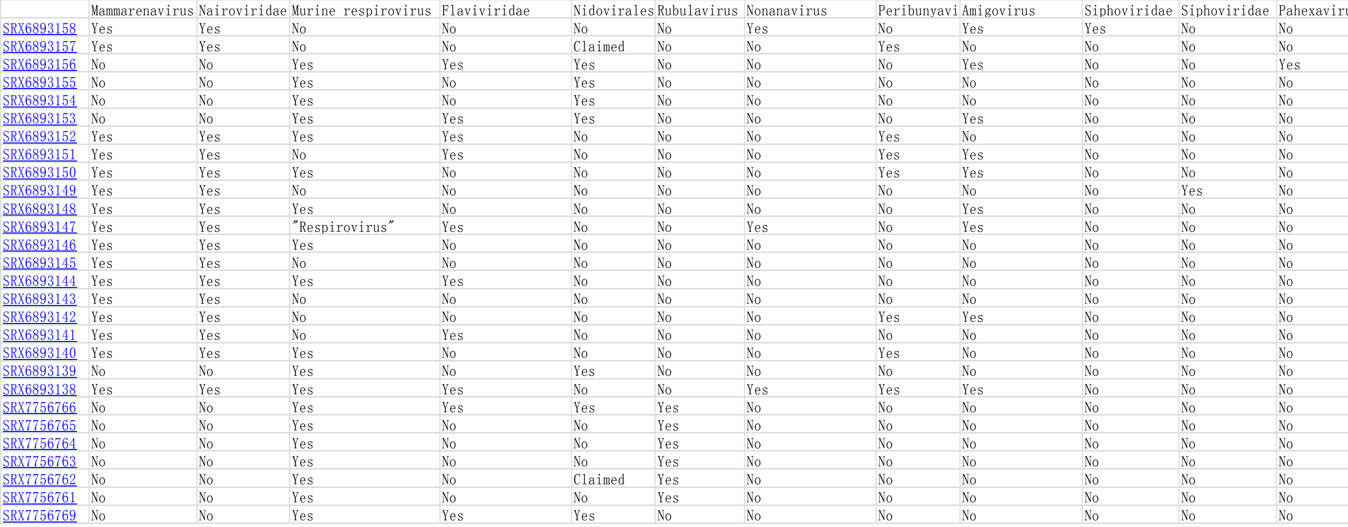
Most importantly, of all dead pangolins being sampled in the studies, only 9 out of a total of 29

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	Pahexaviru
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 Syncytiums 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]

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.

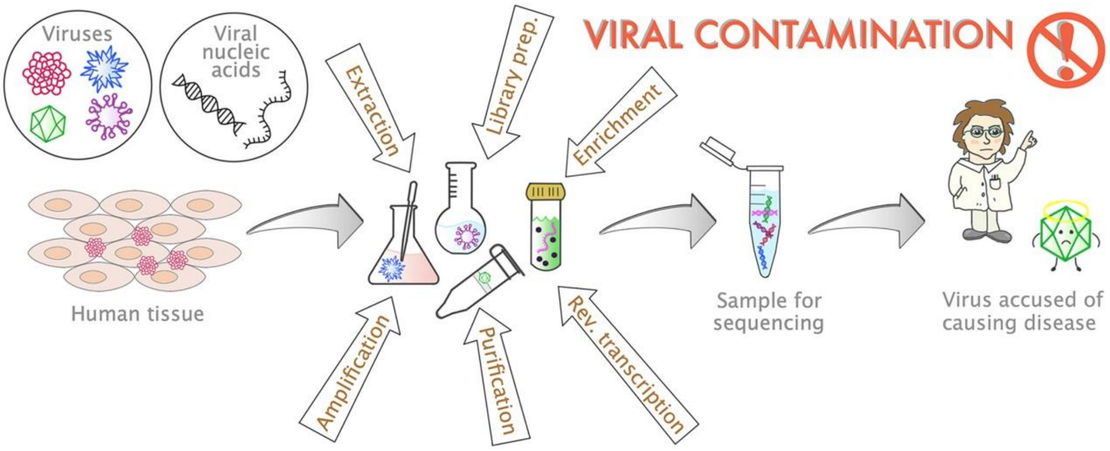


Figure 14. A cartoon diagram of contamination in sequencing experiment leading to false results and false “discoveries”.

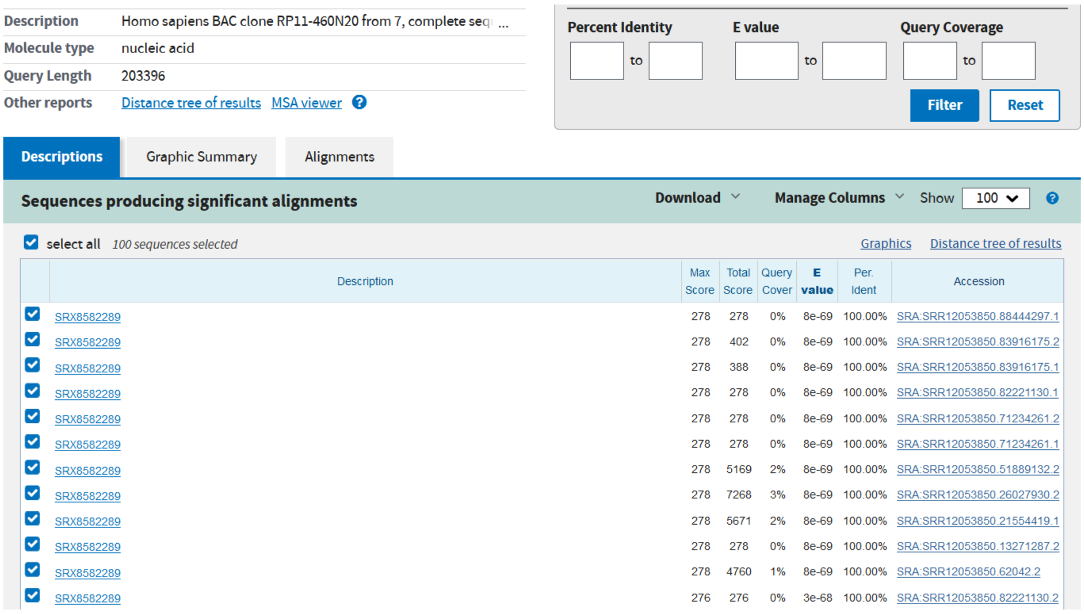
Note as in 2020/7/23

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].

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
SRX8582289 22-Jun-2020	Manis javanica: **43.52%**	Catarrhini 98913	Pangolin coronavirus 792

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

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



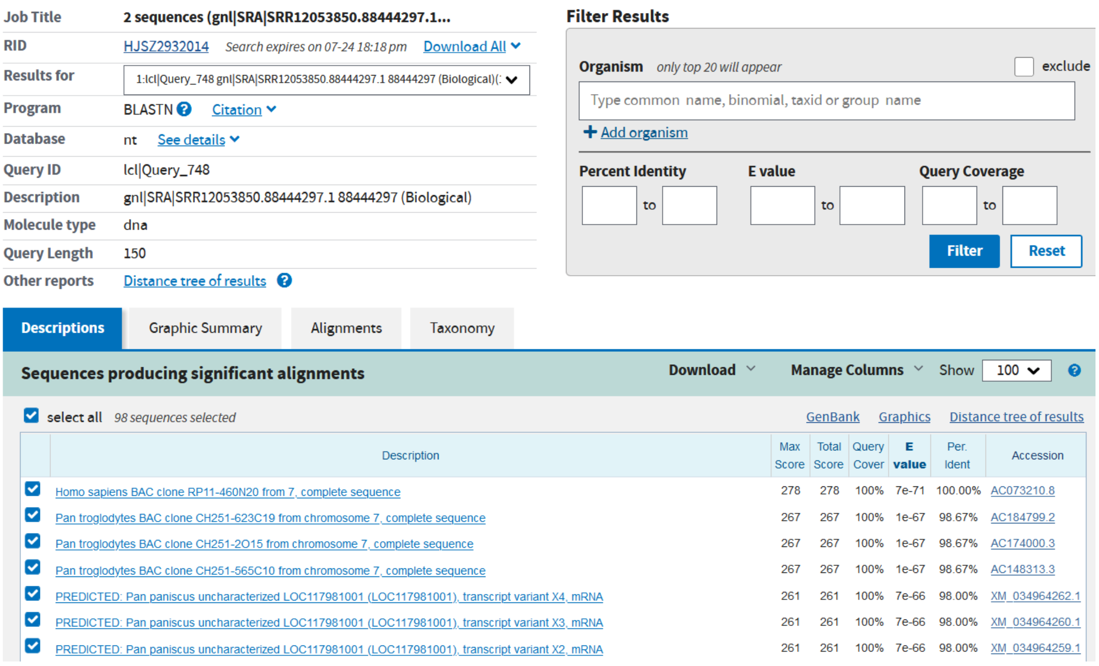
Description	Homo sapiens BAC clone RP11-460N20 from 7, complete seq ...
Molecule type	nucleic acid
Query Length	203396
Other reports	Distance tree of results MSA viewer

Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX8582289	278	278	0%	8e-69	100.00%	SRA:SRR12053850.88444297.1
☑	SRX8582289	278	402	0%	8e-69	100.00%	SRA:SRR12053850.83916175.2
☑	SRX8582289	278	388	0%	8e-69	100.00%	SRA:SRR12053850.83916175.1
☑	SRX8582289	278	278	0%	8e-69	100.00%	SRA:SRR12053850.82221130.1
☑	SRX8582289	278	278	0%	8e-69	100.00%	SRA:SRR12053850.71234261.2
☑	SRX8582289	278	278	0%	8e-69	100.00%	SRA:SRR12053850.71234261.1
☑	SRX8582289	278	5169	2%	8e-69	100.00%	SRA:SRR12053850.51889132.2
☑	SRX8582289	278	7268	3%	8e-69	100.00%	SRA:SRR12053850.26027930.2
☑	SRX8582289	278	5671	2%	8e-69	100.00%	SRA:SRR12053850.21554419.1
☑	SRX8582289	278	278	0%	8e-69	100.00%	SRA:SRR12053850.13271287.2
☑	SRX8582289	278	4760	1%	8e-69	100.00%	SRA:SRR12053850.62042.2
☑	SRX8582289	276	276	0%	3e-68	100.00%	SRA:SRR12053850.82221130.2

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



Job Title	2 sequences (gnl|SRA|SRR12053850.88444297.1...
RID	HJSZ2932014 *Search expires on 07-24 18:18 pm* Download All
Results for	1:lcl|Query_748 gnl|SRA|SRR12053850.88444297.1 88444297 (Biological)(:
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_748
Description	gnl|SRA|SRR12053850.88444297.1 88444297 (Biological)
Molecule type	dna
Query Length	150
Other reports	Distance tree of results

Sequences producing significant alignments

☑ select all *98 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Homo sapiens BAC clone RP11-460N20 from 7, complete sequence	278	278	100%	7e-71	100.00%	AC073210.8
☑	Pan troglodytes BAC clone CH251-623C19 from chromosome 7, complete sequence	267	267	100%	1e-67	98.67%	AC184799.2
☑	Pan troglodytes BAC clone CH251-2O15 from chromosome 7, complete sequence	267	267	100%	1e-67	98.67%	AC174000.3
☑	Pan troglodytes BAC clone CH251-565C10 from chromosome 7, complete sequence	267	267	100%	1e-67	98.67%	AC148313.3
☑	PREDICTED: Pan paniscus uncharacterized LOC117981001 (LOC117981001), transcript variant X4, mRNA	261	261	100%	7e-66	98.00%	XM_034964262.1
☑	PREDICTED: Pan paniscus uncharacterized LOC117981001 (LOC117981001), transcript variant X3, mRNA	261	261	100%	7e-66	98.00%	XM_034964260.1
☑	PREDICTED: Pan paniscus uncharacterized LOC117981001 (LOC117981001), transcript variant X2, mRNA	261	261	100%	7e-66	98.00%	XM_034964259.1

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.

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