Anomalies in BatCoV/RaTG13 sequencing and provenance

Daoyu Zhang

ABSTRACT

To this date, the most critical piece of evidence on the purposed “natural origin” theory of SARS-CoV-2, was the sequence known as RaTG13, allegedly collected from a single fecal sample from Rhinolophus Affinis. Understanding the provenance of RaTG13 is critical on the ongoing debate of the Origins of SARS-CoV-2. However, this sample is allegedly “used up” and therefore can no longer be accessed nor sequenced independently [1], and the only available data was the 3 related Genbank accessions: MN996532.1, SRX7724752 and SRX8357956.

We report these datasets possessed multiple significant anomalies, and the provenence of the promised claims of RaTG13 or it’s role in proving a “probable bat origin”[2] of SARS-CoV-2 can not be satisfied nor possibly be confirmed.

RESULTS

Anomalous enrichment of telomere-like repeat sequences in the dataset SRX7724752

>gnl|SRA|SRR11085797.3.1 3 *(Biological)*
CTAACCCTAACCCTAGCACTATCCTGTTTCCAACCCCAACCCTAACCCTCACCCTAACCC
TAACCCCAGCCTGTTTCATACCTTAACTCGCACCTCATCGCTAACCCCAGCCCTCACCCG
ATCCTGTTTCCTCTCCCGAACATAACCCT

>gnl|SRA|SRR11085797.3.2 3 *(Biological)*
GGTTAGGGTTAGGGTTAGGGTTGGAAACAGGATAGGGTTAGGGTTAGGGTGAGGGTTAGG
GTTAGAGTTAGGGTGGGAAACAGGATAGGGGTAGGGTTAGGGCGAGGGATAGGGATAGGG
AGGGAAACAGGATAGTGGGAGGGCTAGGGGT

>**gnl|SRA|SRR11085797.8.1** 8 *(Biological)*
GTTAGGGTTAGGGTTAGGGTTAGGGTTGGGGTTGGATACAGGATATGGTTAGGGTTAGGG
GTAGGGTCAGGGTTAGGATTGGAAACGAGATAGGTTACGTGATAGGGTTAGCGTTAGGGT
TAGGTTTAGTAATCCGCAACGGCTTAGGGTT

>**gnl|SRA|SRR11085797.8.2** 8 *(Biological)*
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTATCCTGTTCCCAACCCTAACC
CTAACCCTAACCCTAACCCTAACACAAAACATAACCCTAACCCCAACCCAAACCCTAACC
CCATCTTTACTCACACCCTAACCCAAAACTC

>**gnl|SRA|SRR11085797.10.1** 10 *(Biological)*
GTTAGGGTTAGGGTTAGCGTTAGGGTTAGAAACAGGATAGGGTTAGGGTTAGGGTTAGGG
TTAGGGTTTTGGTTGGTCACAGTGTTGCGCTAGGCATAGGGATAGGGTTCGCGTTAGGGT
TAGGGTTAGGATTCGGAAGAGCTAGCTAAA

>**gnl|SRA|SRR11085797.10.2** 10 *(Biological)*
GTTCCCAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTTTCCTTTTTCCAACCCTA
ACACTAACCCTAACACTAAACCTAACCCCAACCCTACCACTATACTATATCCGACTCTCA
CGCTAACACTAAACATAAGTAATCACAAATT

>**gnl|SRA|SRR11085797.13.1** 13 *(Biological)*
TAACCCTAACCCTAAGCGTAAACCTAATCCAATCCTGTTCCCAACCCTAACCCTAACCCT
GACCCTAAGCTTTTCCCGACCCGAACGCCGACCCGATCCGCCACCCTAACCGTAACCCGT
TCCCAACCCCTCCTACTGCTCGATCCGCCT

>**gnl|SRA|SRR11085797.13.2** 13 *(Biological)*
GGAACAGGATAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTGGGAACAGGATAGG
GTTAGGGTTAGGGTTGGGATCAGGATAGGGATAGGGATAGGGATAGGGATAGGGTTAGGG
TGGGGAACAGGAGAGCGTTAGGCAAGG

>**gnl|SRA|SRR11085797.14.1** 14 *(Biological)*
GGTTAGGGTTAGGGTGGAGAAGAGGATAGGTTTAGGTTTAGGGTTAGGGTTAGGGTGAGG
GTTAGGGAGAGGGTTAGCTACACGATAGGAGTAGGGTAACGATTAGGGTTAGGGTTAGGT
TTGGAAAAAGCATAGGCTATGAGGTACGGT

>**gnl|SRA|SRR11085797.14.2** 14 *(Biological)*
CTGCTTCCAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
CTATCCTGTTCCCAACCCTAACCCTAACCCTAACCCCAACCCTAACCCCACCCAAACCCT
AACCCTAACCCCAACCCATACCCCAACCAT

>**gnl|SRA|SRR11085797.15.1** 15 *(Biological)*
TGTTCCCAACCCTAAACCTAAACCTAAGCCGATCCTGTTCCCAACCCTAACCCTAACCCT
ATCCTGTAAACAACCCCCACCCTAAAAACATCCTCGTACAAACCCTAACCCAACCCCCAT
CCCAAACCACATACCCGTCACGAACCCACCC

>**gnl|SRA|SRR11085797.15.2** 15 *(Biological)*
GTTGGGGTTAGGGTTGGGAACAGGATAGGGTTAGGGTTAGGGTTAGGGTTGGGGTCAGGG
TTGGGACAAGGACGGGTAGGGGGTTAGGGTAGTGCACAGGGTAGGGAGTGGGGTAGGGTT
GGGAACAGAGAAGGGACAGTGGGGGGAGTGG

>**gnl|SRA|SRR11085797.16.1** 16 *(Biological)*
CTAACCCTAACCCTACCCCTATCCTGTTCCTAACCCGAACCCTAACCCTAACCCTAACCC
TAACCCTCACCTGTTCCAGACCGTAATGCTAACCTTAACACTATCCTGTGCGCTACCCCG
ACCCTAACCCTCAGCCGACGCGTCACGCCCG

>**gnl|SRA|SRR11085797.16.2** 16 *(Biological)*
GGTAAGGGTTAGGGTTAGGATAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
GTTGGGAACAGGATAGGGTTAGGGTTAGGGTTAGGGTTTGGAACAGGATAGGGTTAGGGA
TAGGGGTAGAGATAGGGTGAGGTGGTGGAA

>**gnl|SRA|SRR11085797.17.1** 17 *(Biological)*
AACCCTAACCCTAACCCTAACGCTATCATGATCCCATCCCTAACCCTAACCCTAACCCTA
ACCCTAATACTAACCCTACCCTTTTCATCTCCCTTACACTACCCCCAACACGCCACCCAT
CCCCAACCACTATGCATGCACTGTCCTAAAC

>**gnl|SRA|SRR11085797.17.2** 17 *(Biological)*
GGTTAGGGTTGGGAACAGGATAGGGTTAGGGTTAGGGTTAGGGTTAGGAACAGGATAGGG
GTAGGGTTAGGGCTAGGGTGAGAAACAGGGTAGGGGTAGGGTGAGGATAAGGGATAGGGT
TGGGGTTGGGAACAGAGAAGGGGAAGGGCA

>**gnl|SRA|SRR11085797.18.1** 18 *(Biological)*
CTAACCTGTTCCCAAACTTAAATCCAATCCTAACCCTATCCTGTTCCCAACCCTAACCCT
AAACCTATACCTATCCTGCCCCACACACCGACCCTATACACCACCCTAAACGCAACCCTA
ACCCCATCCTGTTATCGAAGCATACCCCCAC

>**gnl|SRA|SRR11085797.18.2** 18 *(Biological)*
GTTAGGGTTAGGGTTAGGGTTGGTAACAGGATAGGGTTAGGGTTAGGGTTAGTGTTGGGA
ACAGGATAGGGGGAGGGAGAGGGTTAGGGTTGGGAAGAGGATAGGGATAGGGGTAGGGAG
AGGGTTGGGGATAGGGAAGAGAGAGGAG

>**gnl|SRA|SRR11085797.19.1** 19 *(Biological)*
CCTAACCCTAAACCTAACCCTCTCCTGTTTCCAACCATAACCCTAACCCTATCCCTAACC
CTAACCCACTCCTGTTCTTAACACTAACCTTAACTCTGAGCTCATCCCCAAACCTAACCA
TAACCCCACCAGTTCCGATACCATCACCCCC

>**gnl|SRA|SRR11085797.19.2** 19 *(Biological)*
AGGGTTAGGGTTGGAAACAGGATAGGGTTAGGGTTAGGGTTAGGGTTAGGTTTAGGGTTG
GGGACAGGATAGGGTTAGGGTTAGGGTTAGGGTTGGGGTTAGGGTTAGGTGTGGGGTTGG
GATTGGGTATGGGTAGTGGTCAGGGATAGTG

Figure 1: The reads that contained Telomere-like repeat sequences within the first 20 reads of SRX7724752.

Despite the theoretical presence of traces of Telomere-like repeats in total RNA of most cells, such repeats comprise only a tiny fraction of the total cellular RNA within real biological samples, and normally does not show up in the first 100 reads. RaTG13 contained an anomalous amount of such repeats, which comprises 63% of the dataset and exist in nearly any set of 10 reads within this dataset. In comparison, the next highest content of such repeats within any other sample of similar context on NCBI, contained merely 4% of these repeats, which does not show up in the first 20 reads of the dataset. Telomere-like repeats are not detected in the first 100 reads of any other datasets examined.

In comparison, the related SRX7724693 lacked such reads within the first 100 reads of the dataset.

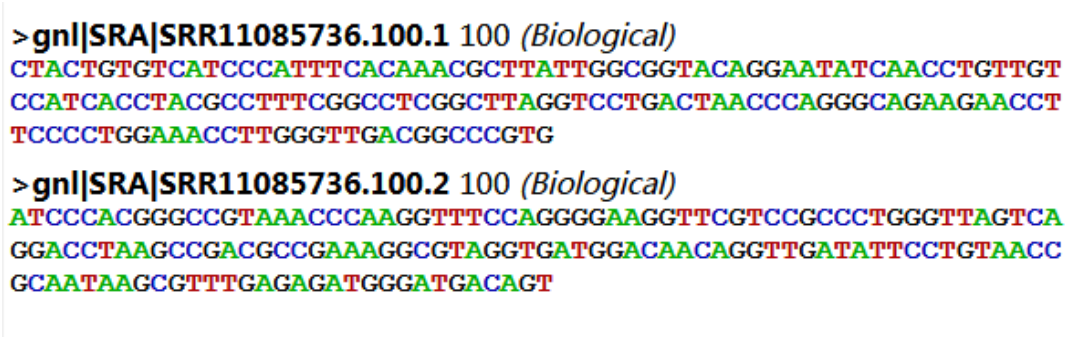


Figure 2: the first 100 reads in SRX7724693 did not show any Telomere-like repeats.

In addition, SRX7724752 contained 6% all-N sequences that were exactly 35nt long, which is not found at levels any close in other datasets that had the same design section.

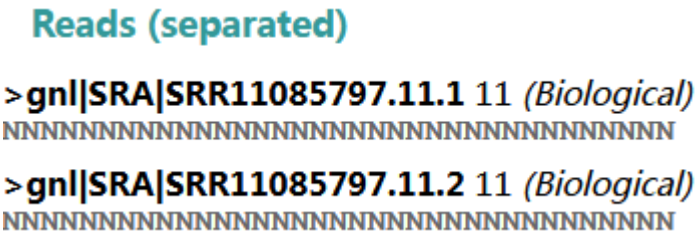


Figure 3: an example of All-N read in SRX7724752.

Anomalous enrichment of non-attributable and low-match data within SRX7724752

no match	42
mouse	1
ape	1
Tea	3
dog	10
Rhi bat	9
Fungus	1
worm	2
cotton	3
Human	4
cat	2
Ribosome	6
fish	8
otter	1
sheep	2
Apicomple	1
squirrel	1
Myo bat	1
rat	1
horse	1
mouse	1
orca	2

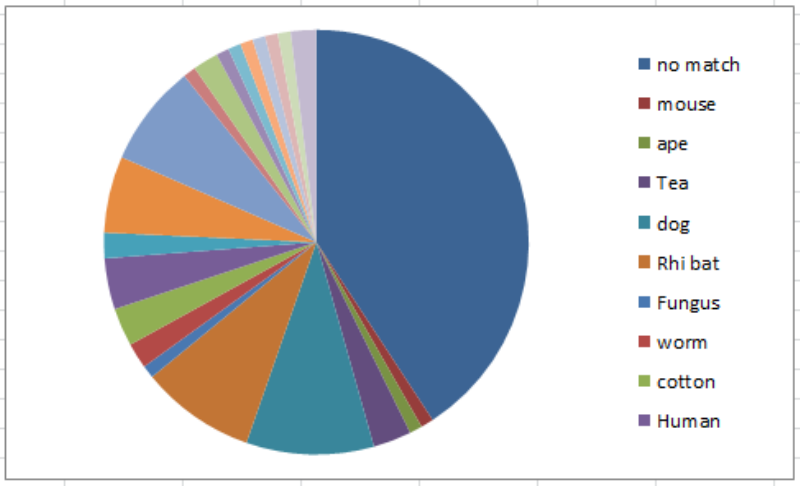
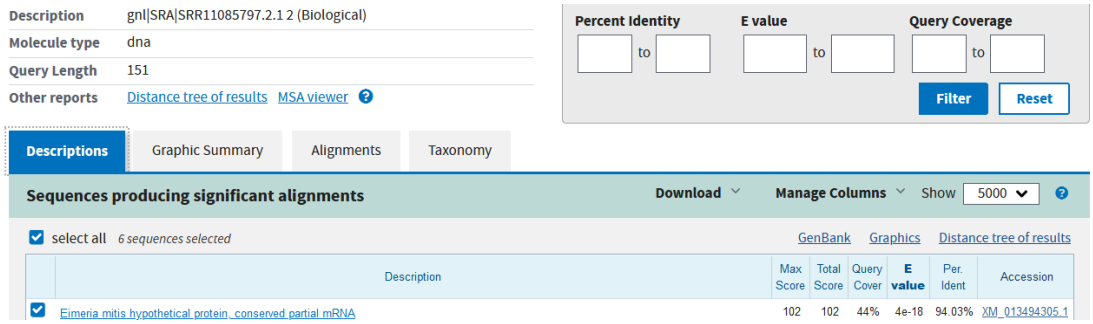


Figure 4: BLAST result of 100 random reads obtained from RaTG13 using BLASTn.

In addition to the anomalous enrichment of repeats, The vast majority of the non-repeat sequences in SRX7724752 does not show any clear matches when examined using BLASTn. With matching results ranging from nearly all domains of life—all of which were partial and low-quality matches, including that of bats.

Only 2 out of 7 Non-repeat and non-PolyN sequences from the first 20 reads from SRX7724752 had any matches, and the match was only partial matches to certain hypothetical proteins



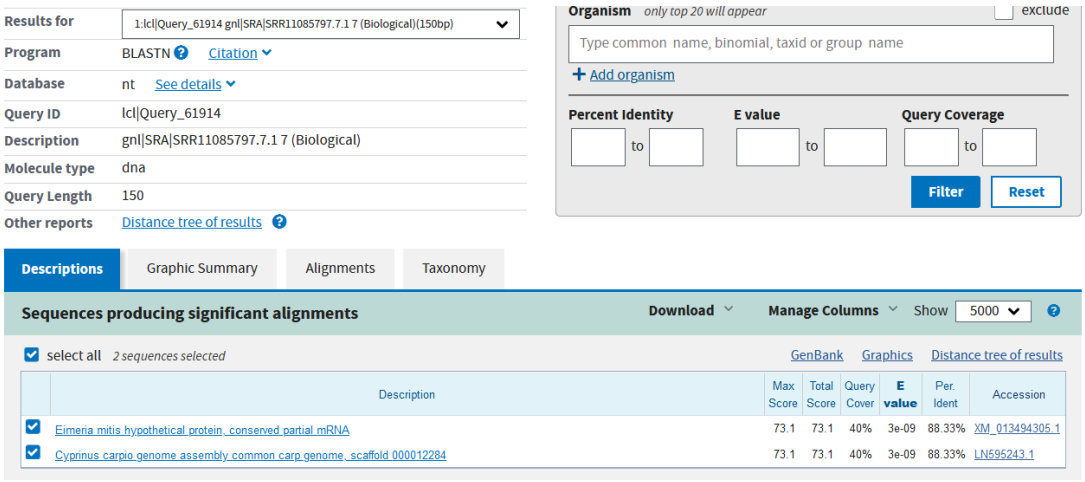


Figure 5: the BLASTn result of the 2 non-repeat and non-PolyN sequences in the first 20 reads of SRX7724752. The rest can not be matched to any known organisms.

Depletion of bacterial-like reads in SRX7724272 which is inconsistent with fecal samples prepared using the methods as indicated by the “Design” section of the SRX7724752 metadata.

Fecal matter [3], is primarily bacteria by composition. All other fecal swabs prepared using the methods indicated by the metadata correctly showed the presence of bacteria as the majority of the reads. In contrast, SRX7724272 contained only 0.65% bacteria-like reads, all of which were 16S rRNA.

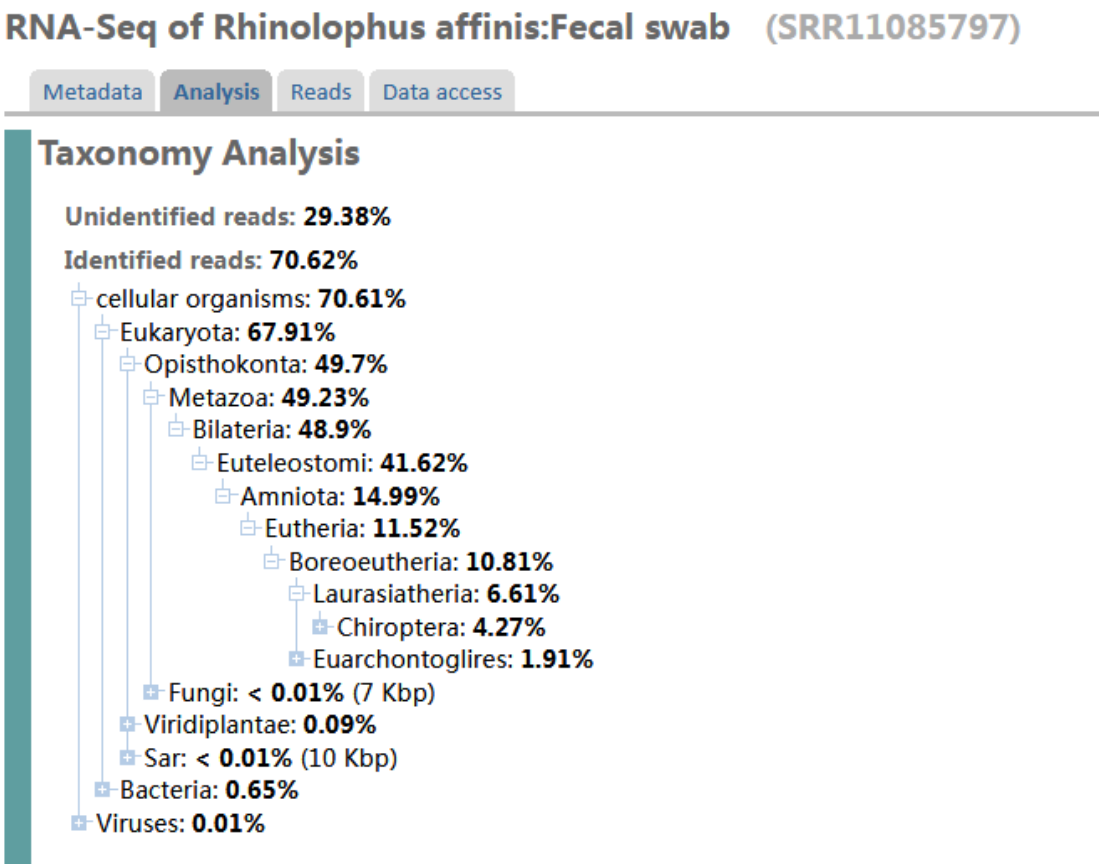


Figure 6: Phylogenetic analysis of SRX7724272.

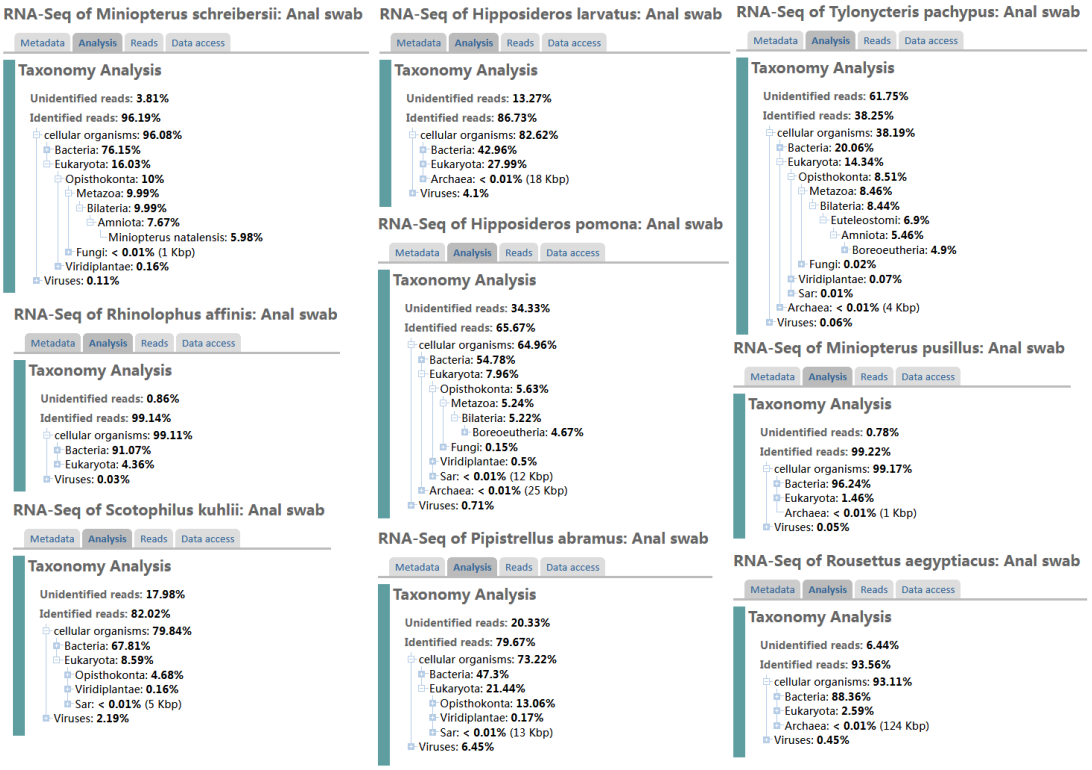


Figure 7: A set of 9 Swabs. The only ones that matches RaTG13 by metadata on Genbank. None of them had more Eukarya-like reads than Bacteria-like reads.

In addition to the anomalous depletion of bacterial-like reads, SRX7724272 also lacked

discernible reads from bacterial mRNA.

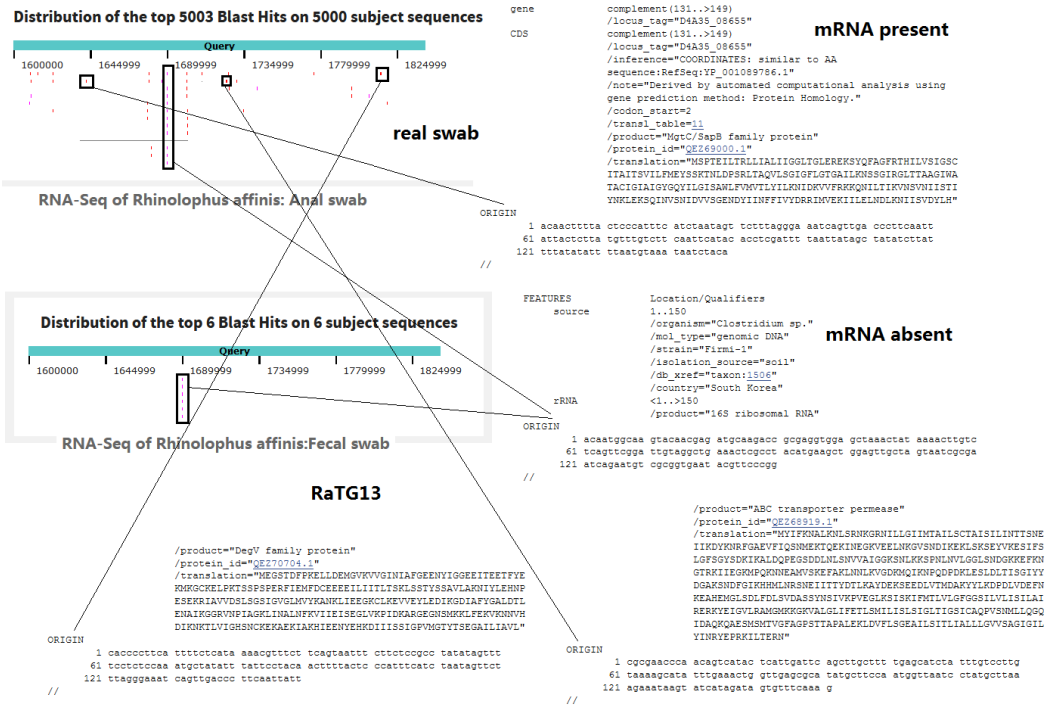


Figure 8: BLAST hits of bacterial non-ribosomal RNA genome on SRX7724272 and another swab from Rhinolophus Affinis under the same library preparation section.

Observation of anomalous and unexpected data within SRX7724752

>gnl|SRA|SRR11085797.11550005.1 11550005 *(Biological)*
GCCCGTATTTAGCCTTAGATGGAGTTTACCACCCGCTTTGGGCTGCATTCCCAAGCAACC
CGACTCCGGGAAGACCCGGGCCCGGCGCGCCGGGGGCCGCTACCGGCCTCACACCGTCCA
CGGGCTGGGCCTCGATCAGAAGGACTTGGGC

>gnl|SRA|SRR11085797.11550005.2 11550005 *(Biological)*
CGGTGGGGCGCGGGACATTTGGCGTACGGAAGACCCACTCCCCGGCGCCGCTCGTGGGGG
CCCAAGTCCTTCTGATCGAGGCCCAGCCCGTGGACGGTGTGAGGCCGGTAGCGGCCCCCG
GCGCGCCGGGCCCGGGTCTTCCCGGAGTCGG

Description

gnl|SRA|SRR11085797.11550005.1 11550005 (B ...

Molecule type

dna

Query Length

151

Other reports

Distance tree of results MSA viewer

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Manage Columns Show 100

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

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
PREDICTED: Phyllostomus discolor 28S ribosomal RNA (LOC114512504), rRNA	279	279	100%	2e-71	100.00%	XR_003685809.1
PREDICTED: Phyllostomus discolor basic proline-rich protein-like (LOC114512442), mRNA	279	279	100%	2e-71	100.00%	XM_028531404.1
Homo sapiens lncAB370.3 lncRNA gene, complete sequence	274	274	100%	9e-70	99.34%	MK280359.1
Homo sapiens lncAB366.1 lncRNA gene, complete sequence	274	274	100%	9e-70	99.34%	MK280356.1
Felis catus Senzu DNA, chromosome: E1, American Shorthair breed	274	1372	100%	9e-70	99.34%	AP023165.1
PREDICTED: Marmota flaviventris 28S ribosomal RNA (LOC117794687), rRNA	274	274	100%	9e-70	99.34%	XR_004618536.1
PREDICTED: Arvicanthis niloticus 28S ribosomal RNA (LOC117704856), rRNA	274	274	100%	9e-70	99.34%	XR_004606369.1
PREDICTED: Arvicanthis niloticus 28S ribosomal RNA (LOC117704855), rRNA	274	274	100%	9e-70	99.34%	XR_004606368.1

Description gnl|SRA|SRR11085797.11550005.1 11550005 (Biological)

Molecule type dna

Query Length 151

Other reports Distance tree of results MSA viewer

Percent Identity to **E value** to **Query Coverage** to Filter Reset

Descriptions | Graphic Summary | Alignments

Sequences producing significant alignments Download Manage Columns Show 100

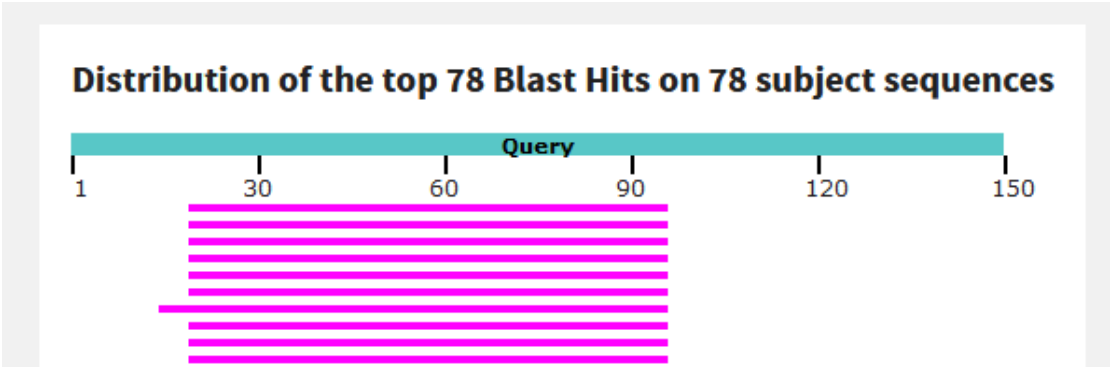
select all *100 sequences selected* Graphics Distance tree of results

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
SRX7724752	279	279	100%	2e-73	100.00%	SRA:SRR11085797.11601488.1
SRX7724752	279	279	100%	2e-73	100.00%	SRA:SRR11085797.11599592.1
SRX7724752	279	279	100%	2e-73	100.00%	SRA:SRR11085797.11598985.1
SRX7724752	279	279	100%	2e-73	100.00%	SRA:SRR11085797.11598757.1
SRX7724752	279	279	100%	2e-73	100.00%	SRA:SRR11085797.11588533.1
SRX7724752	279	279	100%	2e-73	100.00%	SRA:SRR11085797.11585765.1
SRX7724752	279	279	100%	2e-73	100.00%	SRA:SRR11085797.11585419.1
SRX7724752	279	279	100%	2e-73	100.00%	SRA:SRR11085797.11583582.1
SRX7724752	279	279	100%	2e-73	100.00%	SRA:SRR11085797.11582051.1
SRX7724752	279	279	100%	2e-73	100.00%	SRA:SRR11085797.11580259.1
SRX7724752	279	279	100%	2e-73	100.00%	SRA:SRR11085797.11579956.1
SRX7724752	279	279	100%	2e-73	100.00%	SRA:SRR11085797.11579609.1
SRX7724752	279	279	100%	2e-73	100.00%	SRA:SRR11085797.11573984.1
SRX7724752	279	279	100%	2e-73	100.00%	SRA:SRR11085797.11570618.1
SRX7724752	279	279	100%	2e-73	100.00%	SRA:SRR11085797.11568464.1
SRX7724752	279	279	100%	2e-73	100.00%	SRA:SRR11085797.11567362.1
SRX7724752	279	279	100%	2e-73	100.00%	SRA:SRR11085797.11566985.1
SRX7724752	279	279	100%	2e-73	100.00%	SRA:SRR11085797.11558788.1

Figure 9: Phyllostomus Discolor, or Pale spear-nosed bat, a species of bat native to the Americas, is found in high abundance within SRX7724752.

>gnl|SRA|SRR11085797.11550023.2 11550023 *(Biological)*

TTTGTTTTGTTTTTTATAATTTATTTTTAAAATTTATTGGGGTGACAATTGTTAGTAAAA
TTACATAGATTTCAGGTGTACAATTCTGTATTACATGTGGACGGTCCAGCCGCCACGAGT
TCAACGTTTTACATGAAAGGGGGTGTGGGA



Download GenBank Graphics

Rhinolophus ferrumequinum clone VMRC7-331J24, complete sequence

Sequence ID: AC150242.3 **Length:** 120894 **Number of Matches:** 1

Range 1: 32707 to 32786 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
122 bits(66)	3e-24	76/80(95%)	3/80(3%)	Plus/Plus

```
Query  20     tttatttttaaa---atttattGGGGTGACAATTGTTAGTAAAATTACATAGATTTCAGG  76
              ||| ||||||||    ||||||||||||||||||||||||||||||||||||||||||
Sbjct  32707  TTTTTTTTTAAATTTATTTATTGGGGTGACAATTGTTAGTAAAATTACATAGATTTCAGG  32766

Query  77     TGTACAATTCTGTATTACAT  96
              ||||||||||||||||||||
Sbjct  32767  TGTACAATTCTGTATTACAT  32786
```

Figure 10: A sequence which was matched to a bat gDNA BAC clone in the first 96 nucleotides, but then matching nothing on the later nucleotides. This match end with a T.

Figure 11: a viral sequence fused to a mRNA-like sequence. Again overlapping on an A.

Of the only 3 sequences within the viral reads within SRX7724752 that displays fusion of different sequences, only one sequence matches that of a canonical coronavirus subgenomic mRNA leader, another one was the read illustrated in Figure.11, while the third one was a non-canonical fusion of two non-TRS regions in the RaTG13 genome.

>gnl|SRA|SRR11085797.10676687.1 10676687 *(Biological)*

GGTCCTTGATGTCACAGCGTCCTAGATGGTGTCCAGCAATACGAAGATGTCCACGAAGGA
TGACAGCTCCGATTACAAGTTCACTCTCTAGAAGCGGTCTGGTCAAAATAGTGCCATGGA
GTGGCACGTTGAGCAAAATGTTAGTTTCTGG

>gnl|SRA|SRR11085797.10676687.2 10676687 *(Biological)*

ATGAAGGCAATTCACCATTCCATCCTCTAGCTGATAATAAATTTGCACTGACTTGCTTTA
GCACTGATGTGGCTGAGCTACTTCATTGCTTCTTTCAGGCTATTTGCACGTACGCGTTCC
ATGTGGTCATTCAATCCAGAAACTAACATTT

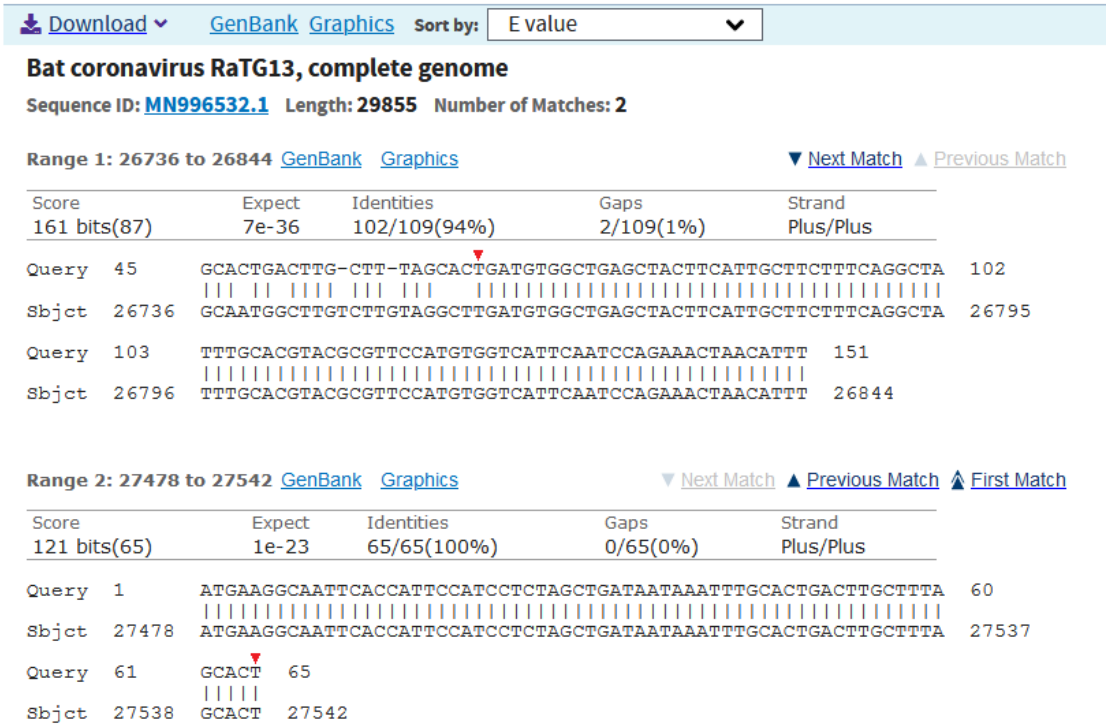


Figure 12: an anomalous fusion of two non-canonical regions of the RaTG13 genome. The fusion again happens on a T.

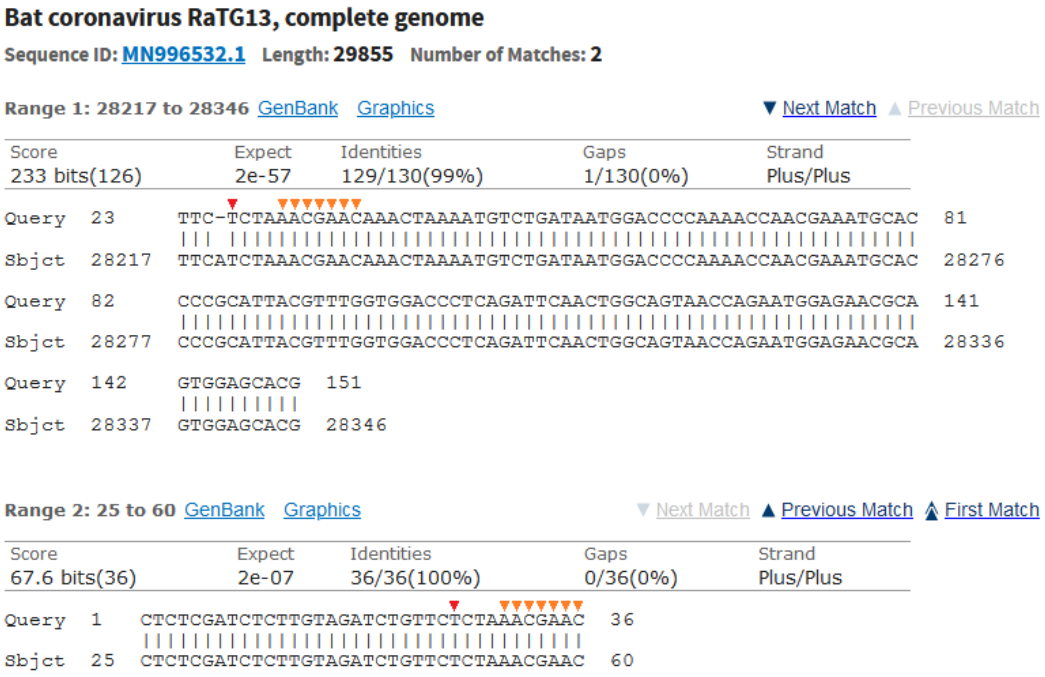


Figure 13: the only canonical sgRNA-like read* in SRX7724752.

Furthermore, SRX7724752 contained significant amount of reads that had higher query coverage on the DNA sequence than on the corresponding mRNA. This most likely indicate a clonal, rather than cDNA, library, was responsible for most of the bat-like reads observed in SRX7724752.

>**gnl|SRA|SRR11085797.76.1** 76 *(Biological)*

CATCAAACTGAGGTTTCAGCAAGGCAAAGATAGCCAGCAACAAAACAAAAAGGCATCCTA
CTGAATGGAAGCAGATAATTGCCAATAGTACATCAGTAAGGAGTTAATATTAAGAATTAG
TTTTTAAAAAGCTCTATATGATGTCAGAAAT

>**gnl|SRA|SRR11085797.76.2** 76 *(Biological)*

GTTTTCACTTGCATTTCTCTAATAATTAGTGATGTTGAGCATCTTTTCATATGTCTATTG
GCCATCTGTATGTCGTCTTTGGAGAAATGTCTATTCAGATTTCTGCCCAATTTTTAATTG
GCTTGTTTGTTTTTTGTTTTTGAATTGAGTT

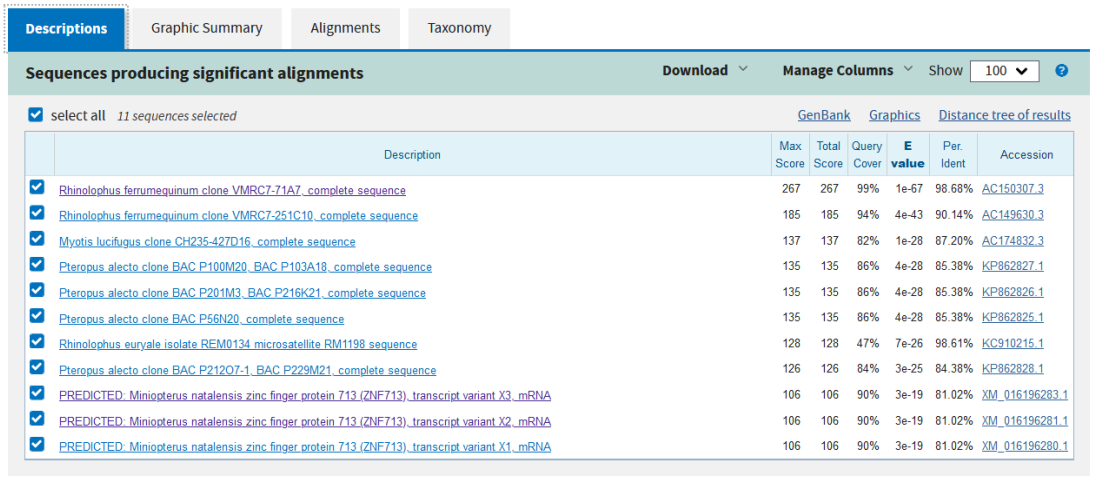


Figure 14: a read from SRX7724752 which have higher coverage on the clone than on the corresponding mRNA. E.g. the read contained nucleotide sequences that were not supposed to be transcribed in actual cells/bats.

Inability of SRX8357956 to prove the promises claimed in [5]

Date	RaTG13		SARS-CoV-2		Sequence	Blast Archive	Name
	Identity	Query Cover	Identity	Query Cover			
14-Oct-18	95.53%	56%	93.15%	56%	24	archive.is/H107n	gnl|SRA|SRR11806578.24 RaTG13-9-5-5_9-5-f1_2018-10-14_B02
14-Oct-18	97.31%	76%	93.18%	75%	23	archive.is/8phs4	gnl|SRA|SRR11806578.23 RaTG13-9-5-4_9-5-r1_2018-10-14_C02
14-Oct-18	97.55%	53%	93.00%	52%	25	archive.is/ycQ89	gnl|SRA|SRR11806578.25 RaTG13-9-5-5_9-5-r1_2018-10-14_D02
14-Oct-18	99.43%	97%	96.39%	97%	22	archive.is/abSp6	gnl|SRA|SRR11806578.22 RaTG13-9-5-4_9-5-f1_2018-10-14_A02
11-Oct-18	97.37%	98%	92.23%	98%	20	archive.is/B20Et	gnl|SRA|SRR11806578.20 RaTG13-9-5-1_21230-F_2018-10-11_A12
11-Oct-18	98.54%	99%	88.26%	99%	21	archive.is/L2pTq	gnl|SRA|SRR11806578.21 RaTG13-9-5-1_23258-R_2018-10-11_B12
08-Oct-18	98.64%	98%	96.19%	99%	5	archive.is/W7Fxp	gnl|SRA|SRR11806578.5 RaTG13-11-2_18297-F_TSS20181008-027-0303_G10
08-Oct-18	99.19%	98%	91.83%	98%	7	archive.is/h8810	gnl|SRA|SRR11806578.7 RaTG13-12-2_24144-R_TSS20181008-027-0303_C11
08-Oct-18	99.89%	99%	87.86%	99%	6	archive.is/jVJY2	gnl|SRA|SRR11806578.6 RaTG13-12-2_22717-F_TSS20181008-027-0303_H10
30-Sep-18	99.50%	99%	92.71%	99%	9	archive.is/ON8UX	gnl|SRA|SRR11806578.9 RaTG13-2-3_RaTG13-2-R1_2018-09-30_B11
30-Sep-18	99.79%	99%	92.89%	99%	8	archive.is/udSil	gnl|SRA|SRR11806578.8 RaTG13-2-3_RaTG13-2-F_2018-09-30_A02
29-Sep-18	99.00%	99%	94.20%	99%	10	archive.is/jdzvN	gnl|SRA|SRR11806578.10 RaTG13-2-3_RaTG13-2-R2_2018-09-29_D05
29-Sep-18	99.09%	98%	94.97%	98%	3	archive.is/Il96Y	gnl|SRA|SRR11806578.3 RaTG13-10-3_RaTG13-10-F_2018-09-29_G04
29-Sep-18	99.72%	98%	98.06%	98%	11	archive.is/7Kioa	gnl|SRA|SRR11806578.11 RaTG13-20-1_RaTG13-F_2018-09-29_H04
29-Sep-18	99.72%	98%	95.19%	98%	4	archive.is/tKTXg	gnl|SRA|SRR11806578.4 RaTG13-10-3_RaTG13-10-R_2018-09-29_E05
27-Sep-18	95.03%	98%	90.88%	98%	14	archive.is/NNfnm	gnl|SRA|SRR11806578.14 RaTG13-4-2_RaTG13-4-R_2018-09-27_G06
27-Sep-18	95.82%	98%	93.46%	93%	13	archive.is/Rmrhq	gnl|SRA|SRR11806578.13 RaTG13-4-2_RaTG13-4-F_2018-09-27_G05
27-Sep-18	98.08%	98%	94.50%	98%	1	archive.is/kcHAi	gnl|SRA|SRR11806578.1 RaTG13-1-2_RaTG13-1-F_2018-09-27_E05
27-Sep-18	98.81%	99%	96.90%	99%	17	archive.is/nhvd2	gnl|SRA|SRR11806578.17 RaTG13-6-2_RaTG13-6-R_2018-09-27_H06
27-Sep-18	98.91%	99%	94.54%	99%	2	archive.is/veLPW	gnl|SRA|SRR11806578.2 RaTG13-1-2_RaTG13-1-R_2018-09-27_F06
27-Sep-18	99.09%	99%	96.92%	99%	16	archive.is/0tqMp	gnl|SRA|SRR11806578.16 RaTG13-6-2_RaTG13-6-F_2018-09-27_A06
27-Sep-18	99.28%	98%	96.81%	98%	12	archive.is/ZHJmY	gnl|SRA|SRR11806578.12 RaTG13-3-2_RaTG13-3-F_2018-09-27_F05
27-Sep-18	99.46%	98%	96.74%	98%	15	archive.is/Epig7	gnl|SRA|SRR11806578.15 RaTG13-5-2_RaTG13-5-F_2018-09-27_H05
27-Sep-18	99.50%	98%	98.49%	98%	18	archive.is/NdyHK	gnl|SRA|SRR11806578.18 RaTG13-7-2_RaTG13-7-F_2018-09-27_B06
27-Sep-18	99.53%	99%	95.67%	97%	19	archive.is/2qg0a	gnl|SRA|SRR11806578.19 RaTG13-8-2_RaTG13-8-F_2018-09-27_C06
20-Jun-17	99.10%	99%	96.61%	99%	28	archive.is/ve7nN	gnl|SRA|SRR11806578.28 RaTG13-R-1-1_7896-1-F1_2017-06-20_E03
20-Jun-17	99.61%	99%	97.43%	99%	32	archive.is/ehzBr	gnl|SRA|SRR11806578.32 RaTG13-R-4-1_7896-4-F_2017-06-20_F03
20-Jun-17	99.87%	98%	97.42%	98%	33	archive.is/do9Rt	gnl|SRA|SRR11806578.33 RaTG13-R-4-1_7896-4-R_2017-06-20_H03
20-Jun-17	99.90%	98%	97.44%	98%	29	archive.is/HjQD8	gnl|SRA|SRR11806578.29 RaTG13-R-1-1_7896-1-R1_2017-06-20_G03
17-Jun-17	98.56%	99%	95.85%	99%	26	archive.is/fqWWF	gnl|SRA|SRR11806578.26 RaTG13-ORF8-1-1_ORF8-F_2017-06-17_A05
17-Jun-17	98.99%	98%	96.52%	98%	27	archive.is/N01Ah	gnl|SRA|SRR11806578.27 RaTG13-ORF8-1-1_ORF8-R1_2017-06-17_A06
03-Jun-17	99.07%	97%	97.49%	97%	30	archive.is/WwyWy	gnl|SRA|SRR11806578.30 RaTG13-R-2-1_7896-2-F1_2017-06-03_A07
03-Jun-17	99.46%	99%	98.01%	99%	31	archive.is/tCLHu	gnl|SRA|SRR11806578.31 RaTG13-R-2-1_7896-2-R1_2017-06-03_A08

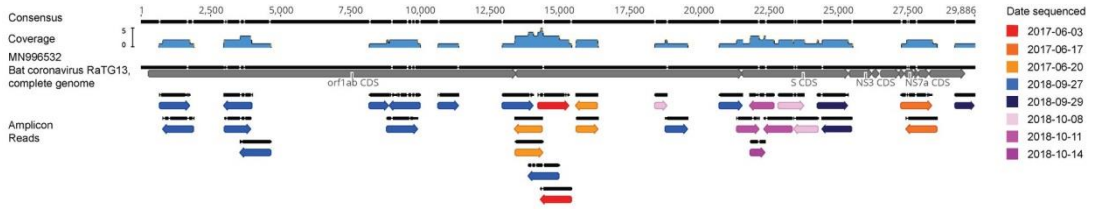


Figure 15: A complete analysis [4] of all Amplicon sequences in SRX8357956. Including the location of these amplicons and the similarity of such amplicon to the RaTG13 and SARS-CoV-2 genome.

Chuan Xiao et.al claimed that RaTG13 contained all the 3 S1 variable loops that were previously considered unique in SARS-CoV-2. [5] However, such claims can not be verified using the amplicons listed in SRX8357956.

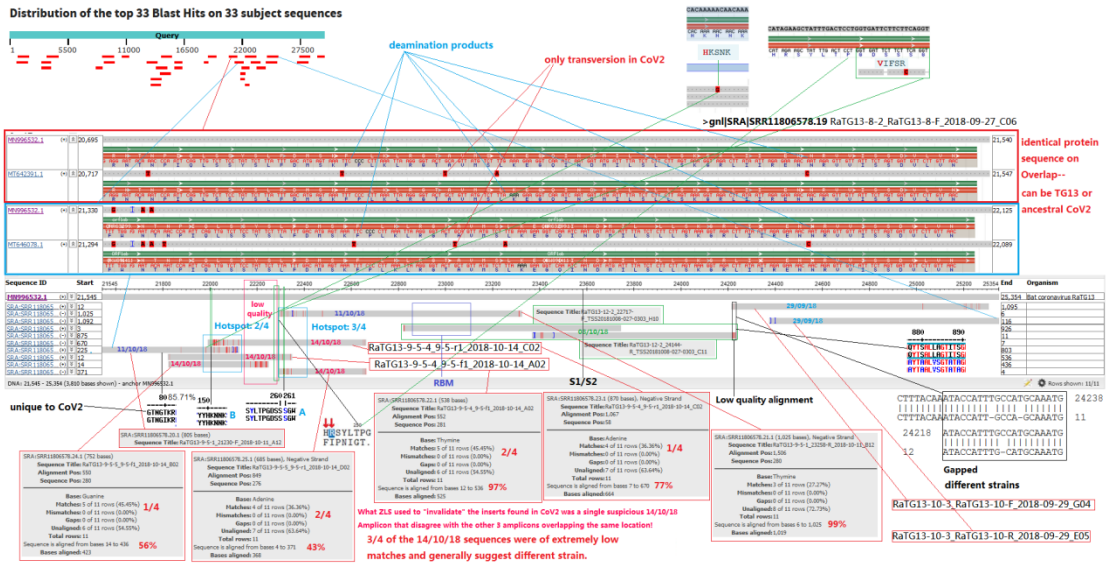


Figure 16: a thorough analysis of the amplicons located on the S locus of RaTG13 in SRX8357956. Notice that the last 4 amplicons sequenced in 14/10/2018 was of very low quality matches, and matched other organisms—including Mouse(mus musculus).

select all *99 sequences selected* GenBank Graphics Distance tree of results

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Gadus morhua genome assembly, chromosome: 16	60.8	105	27%	4e-05	78.05%	LR633958.1
Mus musculus BAC clone RP24-498P8 from chromosome 9, complete sequence	59.9	59.9	47%	1e-04	72.92%	AC168217.2
Mus musculus chromosome 9, clone RP24-484G16, complete sequence	59.9	59.9	47%	1e-04	72.92%	AC137678.11

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

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Homo sapiens BAC clone RP11-792A8 from 7, complete sequence	54.5	54.5	44%	0.006	71.15%	AC027644.9
Coregonus sp. 'balchen' genome assembly, chromosome: 7	52.7	52.7	29%	0.022	73.68%	LR778259.1
Coregonus sp. 'balchen' genome assembly, chromosome: 15	50.9	50.9	16%	0.078	83.33%	LR778267.1
Salmo trutta genome assembly, chromosome: 21	50.9	50.9	37%	0.078	71.90%	LR584437.1
Xanthophyllomyces dendrorhous genome assembly Xden1, scaffold Scaffold_79	50.9	50.9	9%	0.078	96.88%	LN483167.1
Coregonus sp. 'balchen' genome assembly, chromosome: 20	50.0	50.0	72%	0.078	67.49%	LR778272.1
Aquila chrysaetos chrysaetos genome assembly, chromosome: 14	50.0	141	42%	0.078	69.50%	LR606194.1
Bos mutus isolate yakQH1 chromosome 16	50.0	50.0	24%	0.078	74.68%	CP027084.1
Mus musculus BAC clone RP23-128D11 from 7, complete sequence	50.0	50.0	36%	0.078	71.90%	AC122222.6
Mus musculus BAC clone RP23-66E21 from 7, complete sequence	50.0	50.0	36%	0.078	71.90%	AC131741.4

select all *24 sequences selected* GenBank Graphics Distance tree of results

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Mus musculus targeted KO-first, conditional ready, lacZ-tagged mutant allele Fabp4:tm1a(KOMP)Wts	50.9	50.9	23%	0.045	85.11%	JN963014.1
Mus musculus targeted non-conditional, lacZ-tagged mutant allele Fabp4:tm1e(KOMP)Wtsi; transger	50.9	50.9	23%	0.045	85.11%	JN947213.1
Mus musculus chromosome 3, clone RP23-436F15, complete sequence	50.9	50.9	23%	0.045	85.11%	AC123726.11
Mus musculus chromosome 3, clone RP24-137C19, complete sequence	50.9	50.9	23%	0.045	85.11%	AC113990.10

Figure 16: BLAST result of the non-RaTG13 matched parts of Amplicons 25, 24 and 23 in SRX8357956

Using the remaining amplicons, the 3 variable loops, GTNGIKR, HKSNK and VIFSQ was obtained.

This is vastly different from the variable loops possessed by SARS-CoV-2, which were GTNGTKR, HKNNK and GDSSSG. Therefore, the promise of Chuan Xiao et. Al does not hold upon raw data analysis.

Probable discontinuities in RaTG13 sequencing in SRX8357956

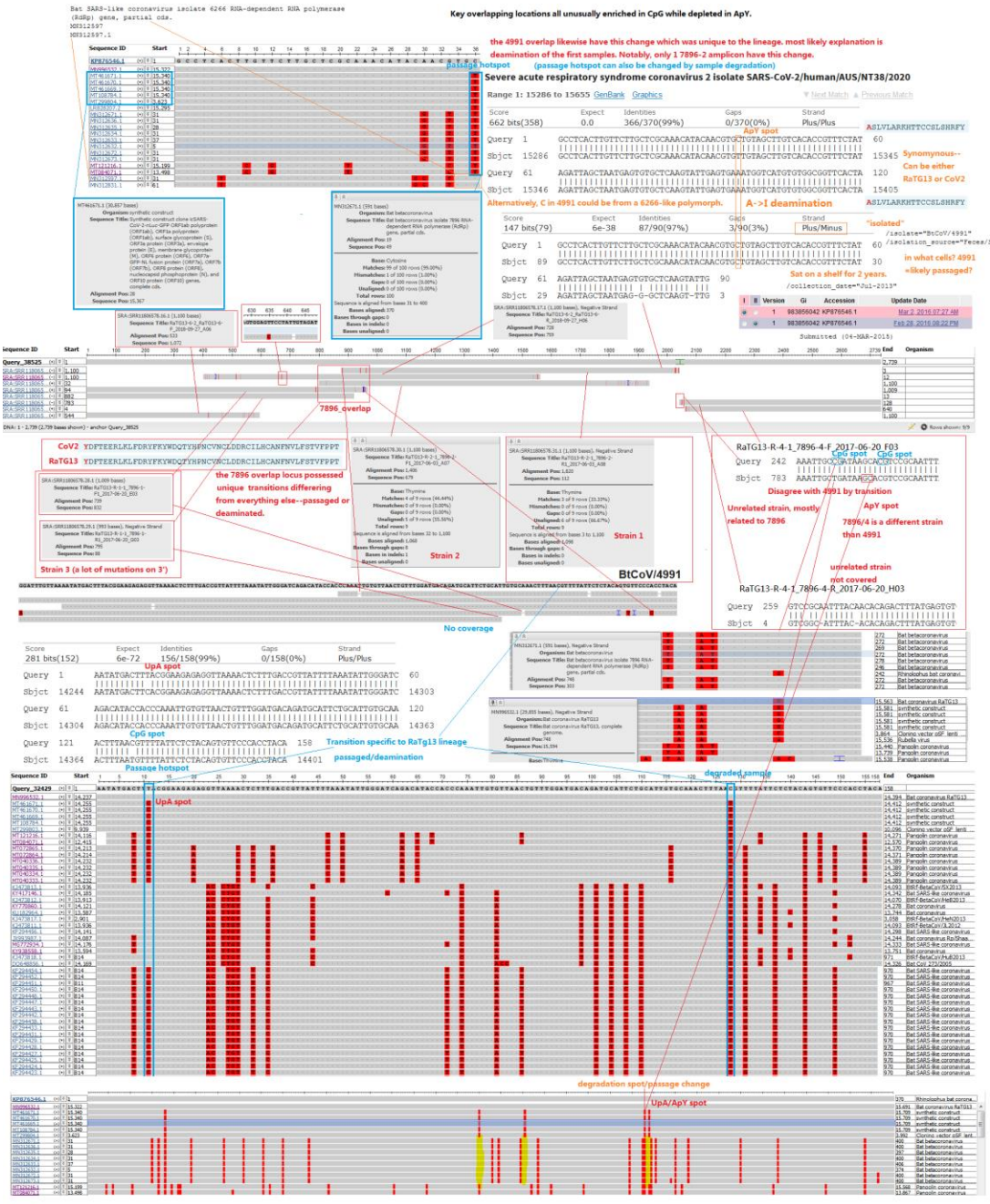


Figure 17: Detailed analysis of the early amplicons located in the nsp12 RdRp region of RaTG13 in SRX8357956.

Within the amplicons labeled “7896”, there were 2 sites of overlap—the first overlap, a region 158bp in length, contained only 2nt difference—all C-T transitions—to SARS-CoV-2. Such transitions easily arise in passage, and are probable sequencing errors from a degraded/passaged sample of DNA.

The second overlap, one with BtCoV/4991, contained only 1 C-T transition, which have a probable origin in the primers used to generate the amplicons in the first place.

DISCUSSIONS

Origins of the anomalies in SRX7724752

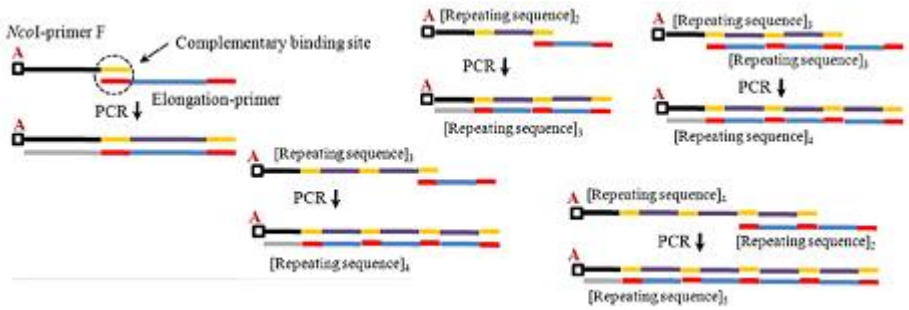


Figure 18A: Origin of repeating sequences in SRX7724752.

The only satiable explanation for the anomalous enrichment of the Telomere-like repeats in SRX7724752, involves the self-amplification of such sequences in a PCR reaction with little to no template.

Normally, with significant amount of template, the random primers normally used in RT-PCR amplifies most sequences evenly and outcompetes the repeat sequences in the reaction, and the result was an accurate reflection of such repeats within cellular samples—extremely poor. However, in samples that have little to no template, such that the random primers/random hexamers used in the reaction were not able to prime the amplification of most sequences—e.g. the amount of normal templates within the reaction falls below the timescale needed for the amplification of the repeating sequences, Repeating sequences, of which telomere-like repeats forms the vast majority of it in the environment and in most samples, can self-amplify in a primer-independent fashion, eventually reaching very high dominance, through repeated denaturing, sliding, reannealing and extension.

As this is a linear process, the self-amplification process is very slow, and is normally outcompeted by the normal amplicons as long as any usable amount of templates were present. Therefore, the presence of anomalously enriched telomere-like repeats within SRX7724752 indicate that the original sample couldn't have contained enough templates for the generation of the complete genome, through any means possible.

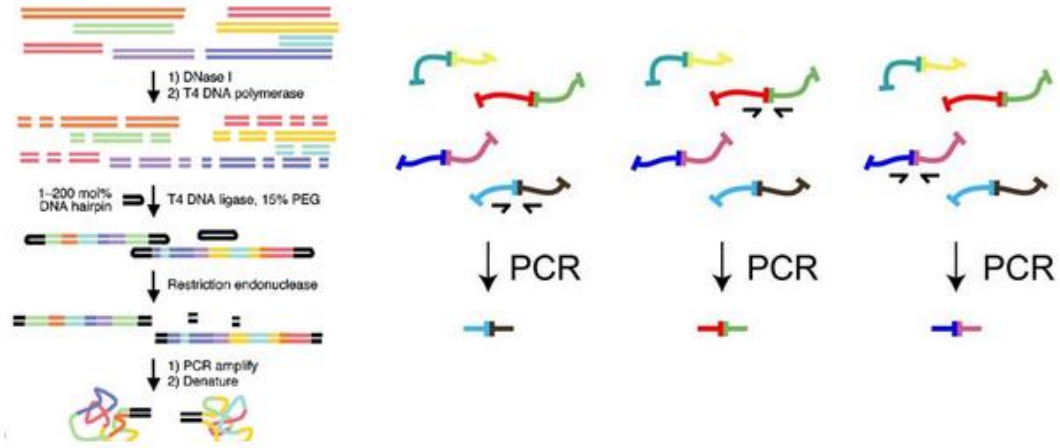


Figure 18B: Origin of the random matched sequences and partial sequences in SRX7724752

The majority of the sequences that were not repeats, when BLASTed, does not match any known organisms. There were also many sequences that matches—only partially, to many diverse organisms. What was striking, is, however, is that these matches often ends with an “A” or a “T”. The most possible explanation of this anomaly is that the Library preparation process of ILLUMINA RNA-seq, which include strand synthesis and A-T ligation to adaptors, were fed dsDNA rather than ssRNA, as input. Such dsDNA input may be PCR products, or it may be a pre-made ILLUMINA sequencing library—Certain mRNA-like reads were inadvertently inverted, indicating double-stranded cDNA was likely used instead of single stranded mRNA.

RID	J85DMESK016 *Search expires on 08-01 20:59 pm* Download All
Results for	2:lcl|Query_5937 gnl|SRA|SRR11085797.66.2 66 (Biological)(150bp)
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_5937
Description	gnl|SRA|SRR11085797.66.2 66 (Biological)
Molecule type	dna
Query Length	150
Other reports	Distance tree of results

Download GenBank Graphics

PREDICTED: Hipposideros armiger putative P2Y purinoceptor 10 (LOC109385656), mRNA

Sequence ID: XM_019648164.1 Length: 2682 Number of Matches: 1

Range 1: 2162 to 2293 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
154 bits(83)	1e-33	116/132(88%)	1/132(0%)	Plus/Minus

```
Query  12    TTT-TCATTATTAAGTATTATGTACTGTACATAATTGTATGTACTATACTTTTATACAAC  70
             ||| ||||||| |||| ||||||||||||||   |||||||||  ||||||||||||   ||
Sbjct  2293  TTTATCATTATCAAGTGTTATGTACTGTACAGTATTGTATGTGTTATACTTTTATATGAC  2234

Query  71    TGGCAGCACAGCAGGTTTGTTTATACCAGCATCACCACAAAAATGTGAGTAATGCATTAC  130
             || ||||| || ||| ||||||| |||||||||||||||||||||||||||||||||||
Sbjct  2233  TGACAGCATAGTAGGCTTGTTTACACCAGCATCACCACAAAAATGTGAGTAATGCATTAC  2174

Query  131   ACTACAATGTTA  142
             ||||  ||||||
Sbjct  2173  ACTATGATGTTA  2162
```

Figure 19: An inverted mRNA-like read.

*: Analysis of the sole sgRNA-like read reveal the usage of a leader/F primer and the mispriming of Amplicon DNA

```
                    ttagatttcatctaaacgaacaaactaaaatgtctgataatggaccccaaaaccaacgaaatgcaccccgcattacgtttggtggaccct
          CTCTCGATCTCTTGTAGATCTGTTC TCTAAACGAAC
ACAAACCAACGAACTCTCGATCTCTTGTAGATCTGT
                                   TAACCTCCTTTTGTCCCCTAGTTCACCGATCTCATCGCCCCTGCCAGGGGTCCATGGACTGTGTTGATCCTGTCCTCTGCTCCTCC
```

Figure 20a: the match analysis between different genomic fragments of RaTG13, of the sole sgRNA-like read* in SRX7724752. *:Figure 13

Despite being sgRNA-like in the first glance, analysis of the exact overlapping region of this particular read reveal that this region is identical to BtCoV/ZC45 and BtCoV/ZXC21—indicating it's identity as likely being a consensus primer.

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

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Pangolin coronavirus isolate PCoV_GX-P3B genomic sequence	67.6	67.6	100%	8e-09	100.00%	MT072865.1
Pangolin coronavirus isolate PCoV_GX-P2V, complete genome	67.6	67.6	100%	8e-09	100.00%	MT072864.1
Pangolin coronavirus isolate PCoV_GX-P5E, complete genome	67.6	67.6	100%	8e-09	100.00%	MT040336.1
Pangolin coronavirus isolate PCoV_GX-P5L, complete genome	67.6	67.6	100%	8e-09	100.00%	MT040335.1
Pangolin coronavirus isolate PCoV_GX-P1E, complete genome	67.6	67.6	100%	8e-09	100.00%	MT040334.1
Pangolin coronavirus isolate PCoV_GX-P4L, complete genome	67.6	67.6	100%	8e-09	100.00%	MT040333.1
Bat coronavirus RaTG13, complete genome	67.6	67.6	100%	8e-09	100.00%	MN996532.1
Bat SARS-like coronavirus isolate bat-SL-CoVZXC21, complete genome	67.6	67.6	100%	8e-09	100.00%	MG772934.1
Bat SARS-like coronavirus isolate bat-SL-CoVZC45, complete genome	67.6	67.6	100%	8e-09	100.00%	MG772933.1
Bat coronavirus (BtCoV/279/2005), complete genome	65.8	65.8	97%	3e-08	100.00%	DQ648857.1
Mutant SARS coronavirus Urbani clone SARS-Urbani-MA_SHC014-spike, complete genome	63.9	63.9	94%	1e-07	100.00%	MT308984.1
Coronavirus BtRs-BetaCoV/YN2018D, complete genome	63.9	63.9	94%	1e-07	100.00%	MK211378.1

Figure 20b: BLAST result of CTCTCGATCTCTTGTAGATCTGTTCTCTAAACGAAC.

This particular sequence have extended overlap to the beginning of the N gene, which was coincidentally at the end of the last 2017/06/17 amplicon. This indicate it was most likely the product of mispriming, rather than a true sgRNA-like read.

SRX7724752 is a mixed library consists of a matrix of dried American bat guano, a bat WGS/RNA-seq ILLUMINA library, a synthetic 16S library and megaprimer PCR products from the SRX8357956 Amplicons and a degraded sample of SARS-CoV-2 cDNA

Phyllostomus Discolor, a species of bat native to Mexico and southern United states, leaves numerous Full-length 100% matched reads that don't match anything else. Coincidentally, Mexico is one of the major supplier of bat guano used for fertilizer and other commercial purposes[6]. The confirmed presence of this particular bat species, suggest the use of a commercial dried bat guano matrix as the bulk of the sample being sequenced. As in PRJNA494391[7] which synthetic metagenome samples were constructed using cDNA amplicons and a specific material matrix to simulate realistic metagenomic reads of a desired virus in a sample.

Traces of the original template used in the megaprimer PCR process can be seen as traces of low-matched virus-like reads within this dataset, which are found across the entire RaTG13 genome.

Figure 21: Read coverage of SRX7724752 on the RaTG13 genome. The red pixels represent significant mismatches on the reads in the dataset.

The Bacterial-like reads in SRX7724752 is also likely a synthetic 16S library—as the only other dataset with Telomere-like repeats(4%), still contained significant amount of bacterial mRNA.

>gnl|SRA|SRR11085733.2232944.1 2232944 *(Biological)*
GCCTTCGTTTGTATATAGTTTTAATGCAAATCCCCTAACATCTCTTTCAGCATCTGCTGC
ACCTCTTTCACCAGCAACTGTAGAAAATCTTAAAAGGGCTTTTGTTTTTTTTACCAACTTT
GTTAAAAATATCTGCTTTAGAATATTTTGT

>gnl|SRA|SRR11085733.2232944.2 2232944 *(Biological)*
AGAGGTCCTACTCTTTTACAAGATACTTGGCTTTTAGAAAAACTTGCACATTTCGATAGG
GAAAGGATACCAGAAAGAGTTGTGCACGCTAAAGGAAGTGCTGCATACGGCGAATTAACA
ATTACTAATGATATTACAAAATATTCTAAA

Helicobacter hepaticus ATCC 51449, complete genome

Sequence ID: AE017125.1 **Length:** 1799146 **Number of Matches:** 1

Range 1: 47720 to 47856 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
176 bits(95)	3e-40	123/137(90%)	0/137(0%)	Plus/Minus

```
Query  7      CCTACTCTTTTACAAGATACTTGGCTTTTAGAAAAACTTGCACATTTCGATAGGGAAAGG  66
              ||||| ||||||||||||||||||||||||||||||||||||||| ||||| || ||
Sbjct  47856  CCTACACTTTTACAAGATACTTGGCTTTTAGAAAAACTTGCACATTTTGATAGAGAGAGA  47797

Query  67     ATACCAGAAAGAGTTGTGCACGCTAAAGGAAGTGCTGCATACGGCGAATTAACAATTACT  126
              || ||||| ||||| ||||||||||||||||||| ||||| || ||||||||||||||
Sbjct  47796  ATCCCAGAGAGAGTGGTGCACGCTAAAGGAAGTGCAGCATATGGTGAATTAACAATTACA  47737

Query  127    AATGATATTACAAAATA  143
              |||||||||||  ||||
Sbjct  47736  AATGATATTACTCAATA  47720
```

```
     gene            complement(<1..>137)
                     /gene="katA"
                     /locus_tag="HH_0043"
                     /old_locus_tag="HH0043"
     CDS             complement(<1..>137)
                     /gene="katA"
                     /locus_tag="HH_0043"
                     /old_locus_tag="HH0043"
                     /EC_number="1.11.1.6"
                     /codon_start=1
                     /transl_table=11
                     /product="catalase"
                     /protein_id="AAP76640.1"
                     /translation="MSKKFTTATGTPLGDNQNSITAGKKGPTLLQDTWLLEKLAHFDR
                     ERIPERVVHAKGSAAYGELTITNDITQYTKAELFNKVGKKTKAFLRFSVVAGERGAAD
                     AERDVRGFALKLYTNEGNWDIVGNNTPVFFIKDAIKFPDFIHTQKRDPKTNMKSPTAM
                     WDFWSLHPESLHQVTILMSDRGIPRSYREMHGFGSHTYSFINAKNERFWVKFHFVCLQ
                     GIHNLTNKESEAVIAKDRESHQKDLFENIEKGNFPKWRFCIQVMSEKEAENYRFNPFD
                     LTKVWSHKDYPLIEVGILELNKNPENYFAEVEQAAFNPANIVPGVGYSPDKVLQGRLF
                     SYGDTQRYRLGINHTQLPVNAPIVPVNNTHRDGFMQQGQFGDRRNYEPSYFNDYVEDK
                     NALEPPLFVQEGDVMYKYDHREYEDDYFVQAGDLYRLMTAEQKEALCQNIKESMEGVP
                     DEIKKRQLEHFKKADKAYGKRVAELLGL"
ORIGIN
        1 tattgagtaa tatcatttgt aattgttaat tcaccatatg ctgcacttcc tttagcgtgc
       61 accactctct ctgggattct ctctctatca aaatgtgcaa gtttttctaa aagccaagta
      121 tcttgtaaaa gtgtagg
//
```

Figure 22a: a bacterial mRNA read in SRX7724696, the only other dataset on NCBI that contained Telomere-like repeats in the first 100 reads of the dataset. Total amt. of repeats=4%

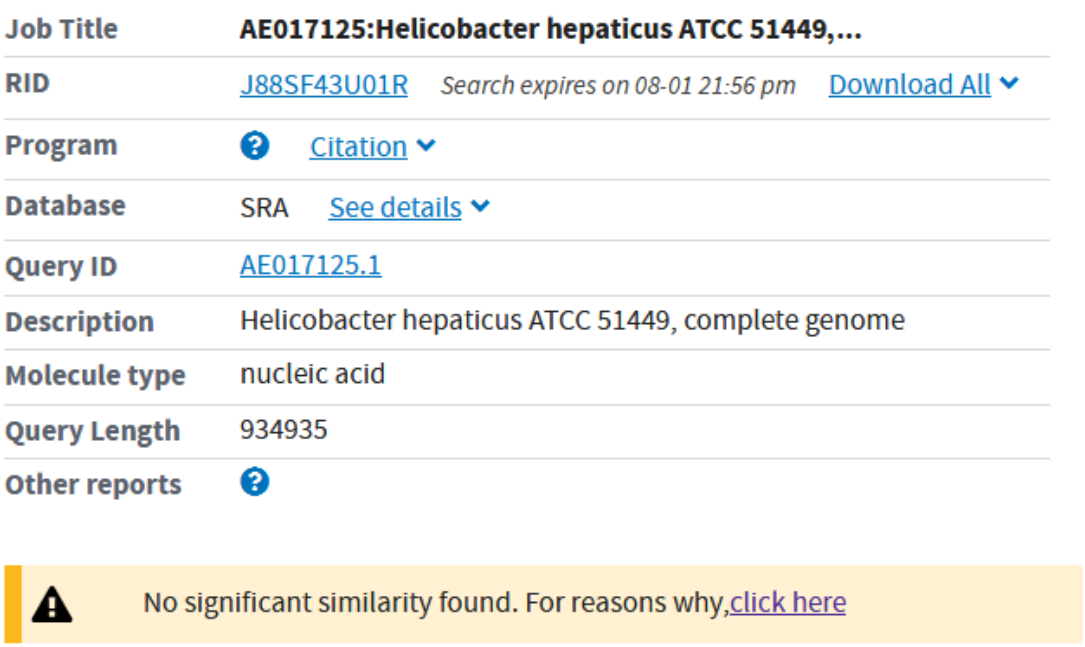
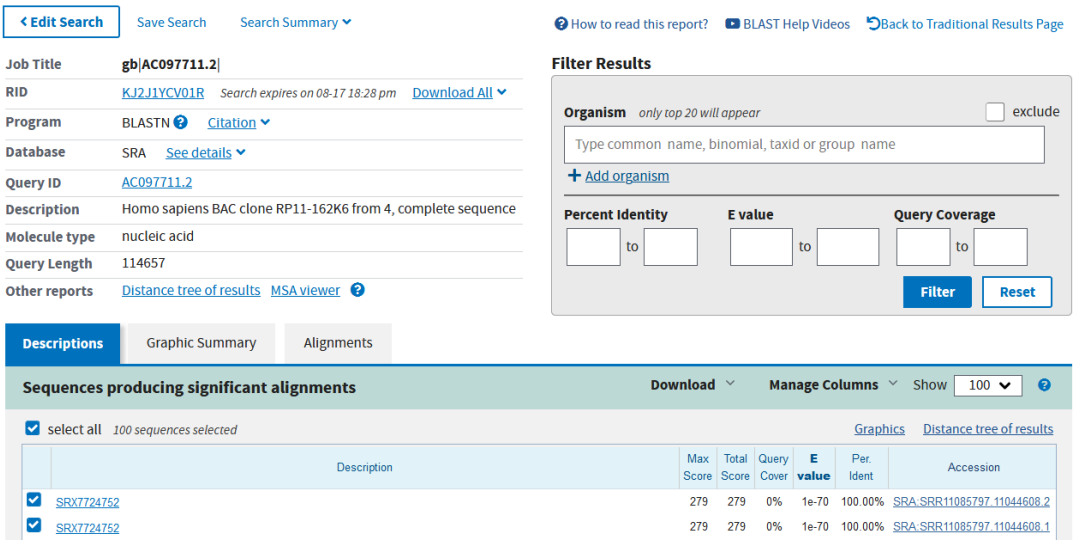


Figure 22b: the same species of bacteria in SRX7724752. No significant matches were found.

This dataset is likely subjected to probe-capture sequencing similar to these other datasets—the use of a positive-sense CoV probe resulted in the selective presentation of the negative ssDNA strand of the ligation products to show up. This is supported by the observation that while most of the virus-like reads were on the negative strand, the Repeats does not show a bias in strand polarity, and the mRNA-like reads have a much higher chance of being on the wrong polarity for RNA-seq. This is likely due to the ligation process being used.

Probable signs of laboratory manipulation of SRX7724752



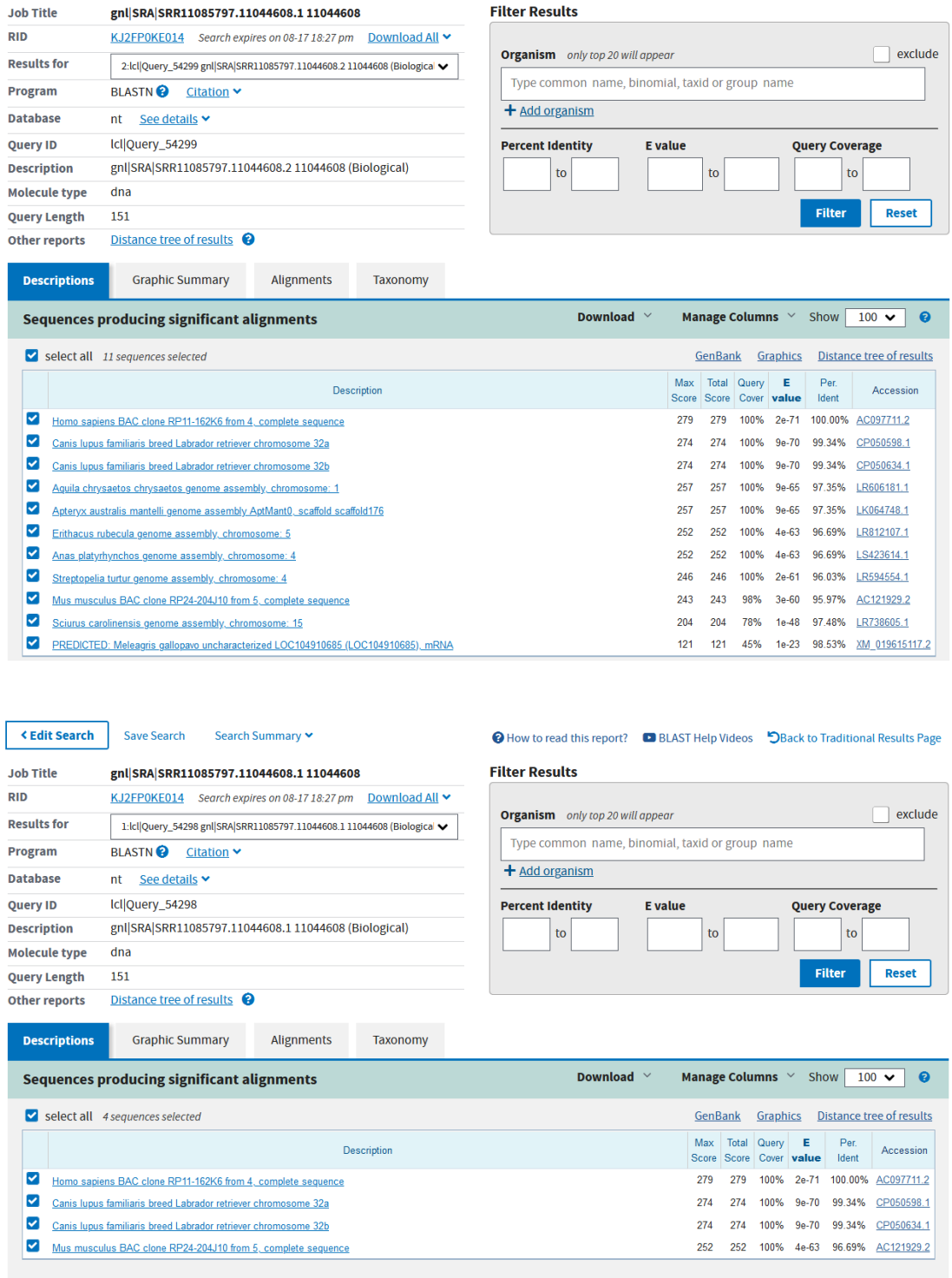
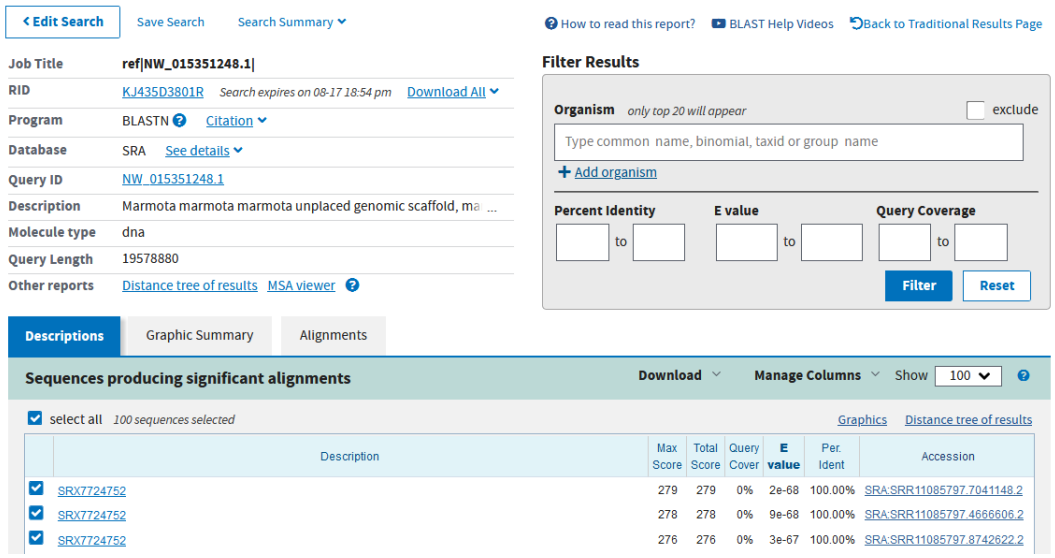
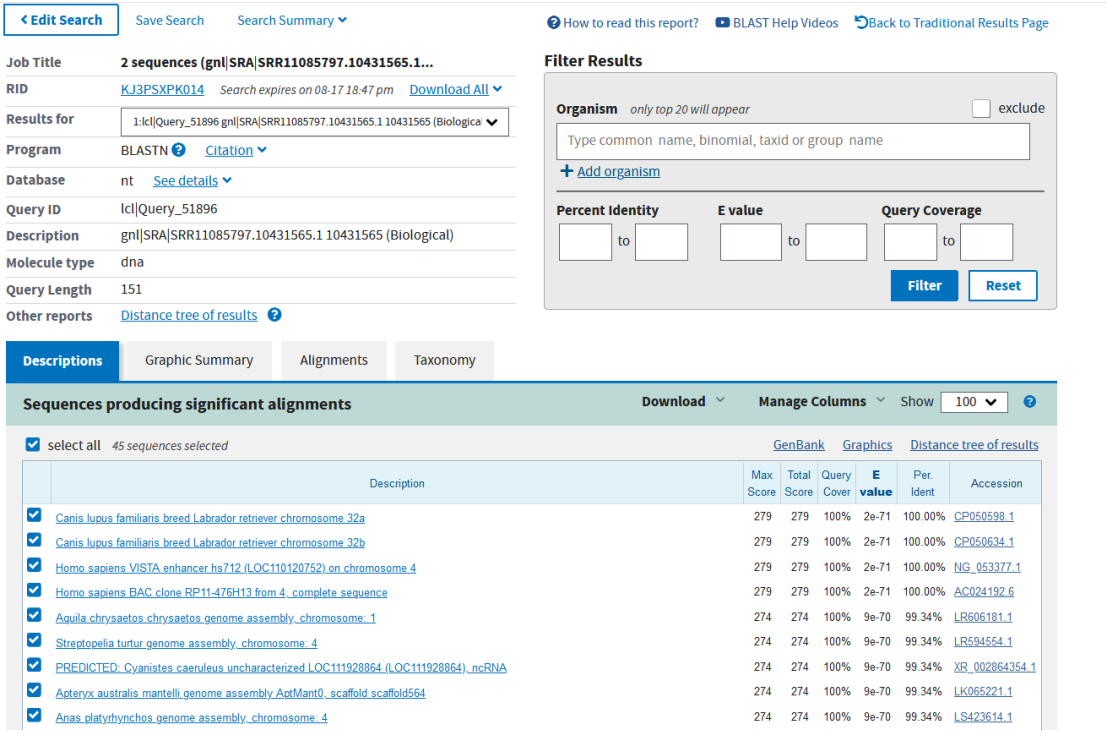


Figure 23: Unique, fully-matched 100% read from Homo Sapiens is recovered from the dataset SRX7724752.


Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page
Job Title ref|NW_015351248.1|
RID KJ435D3801R Search expires on 08-17 18:54 pm Download All
Program BLASTN Citation
Database SRA See details
Query ID NW_015351248.1
Description Marmota marmota marmota unplaced genomic scaffold, ma ...
Molecule type dna
Query Length 19578880
Other reports Distance tree of results MSA viewer
Filter Results
Organism only top 20 will appear
exclude
Type common name, binomial, taxid or group name
Add organism
Percent Identity
E value
Query Coverage
to
Filter
Reset
Descriptions
Graphic Summary
Alignments
Sequences producing significant alignments
Download
Manage Columns
Show 100
select all 100 sequences selected
Graphics
Distance tree of results
Description
Max Score
Total Score
Query Cover
E value
Per. Ident
Accession
SRX7724752 279 279 0% 2e-68 100.00% SRA:SRR11085797.7041148.2
SRX7724752 278 278 0% 9e-68 100.00% SRA:SRR11085797.4666606.2
SRX7724752 276 276 0% 3e-67 100.00% SRA:SRR11085797.8742622.2


Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page
Job Title 2 sequences (gnl|SRA|SRR11085797.10431565.1...
RID KJ3PSXPK014 Search expires on 08-17 18:47 pm Download All
Results for 1:lcl|Query_51896 gnl|SRA|SRR11085797.10431565.1 10431565 (Biological
Program BLASTN Citation
Database nt See details
Query ID lcl|Query_51896
Description gnl|SRA|SRR11085797.10431565.1 10431565 (Biological)
Molecule type dna
Query Length 151
Other reports Distance tree of results
Filter Results
Organism only top 20 will appear
exclude
Type common name, binomial, taxid or group name
Add organism
Percent Identity
E value
Query Coverage
to
Filter
Reset
Descriptions
Graphic Summary
Alignments
Taxonomy
Sequences producing significant alignments
Download
Manage Columns
Show 100
select all 45 sequences selected
GenBank
Graphics
Distance tree of results
Description
Max Score
Total Score
Query Cover
E value
Per. Ident
Accession
Canis lupus familiaris breed Labrador retriever chromosome 32a 279 279 100% 2e-71 100.00% CP050598.1
Canis lupus familiaris breed Labrador retriever chromosome 32b 279 279 100% 2e-71 100.00% CP050634.1
Homo sapiens VISTA enhancer hs712 (LOC110120752) on chromosome 4 279 279 100% 2e-71 100.00% NG_053377.1
Homo sapiens BAC clone RP11-476H13 from 4, complete sequence 279 279 100% 2e-71 100.00% AC024192.6
Aquila chrysaetos chrysaetos genome assembly, chromosome: 1 274 274 100% 9e-70 99.34% LR606181.1
Streptopelia turtur genome assembly, chromosome: 4 274 274 100% 9e-70 99.34% LR594554.1
PREDICTED: Cyanistes caeruleus uncharacterized LOC111928864 (LOC111928864), ncRNA 274 274 100% 9e-70 99.34% XR_002864354.1
Apteryx australis mantelli genome assembly AptMant0, scaffold scaffold564 274 274 100% 9e-70 99.34% LK065221.1
Anas platyrhynchos genome assembly, chromosome: 4 274 274 100% 9e-70 99.34% LS423614.1

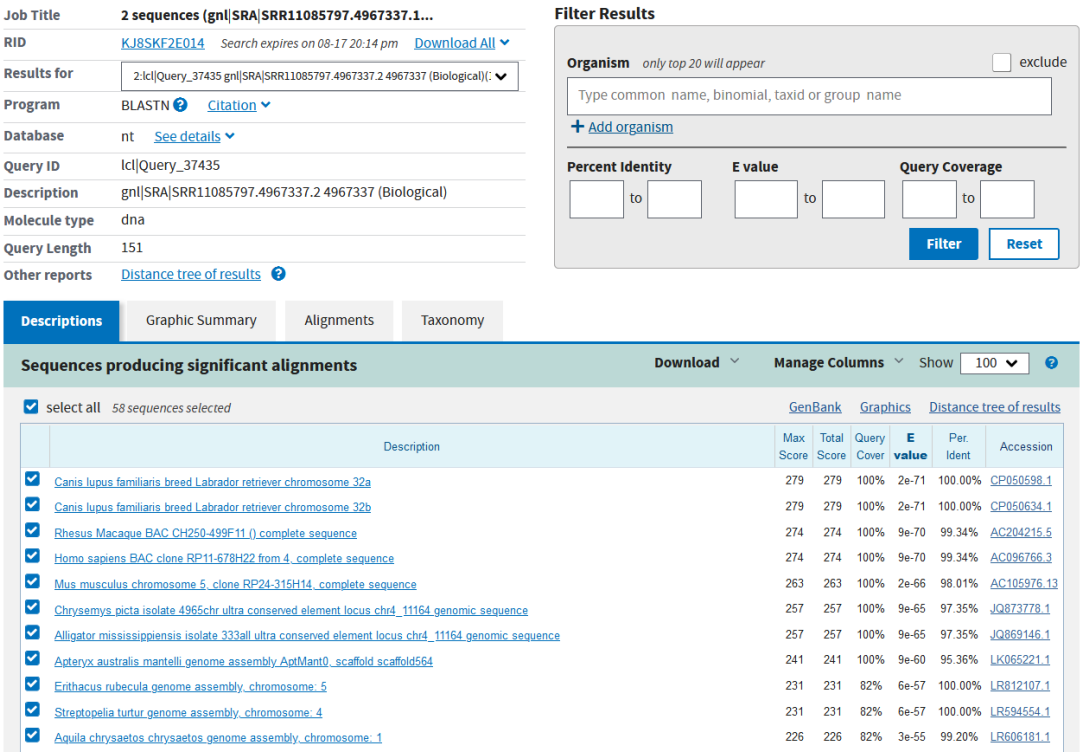


Figure. 24: Marmota Marmota genetic scaffold assemblies returned significant amount of 100% full-length matched reads that were sometimes also found in Homo Sapiens and Canis Lupus Famillaris.

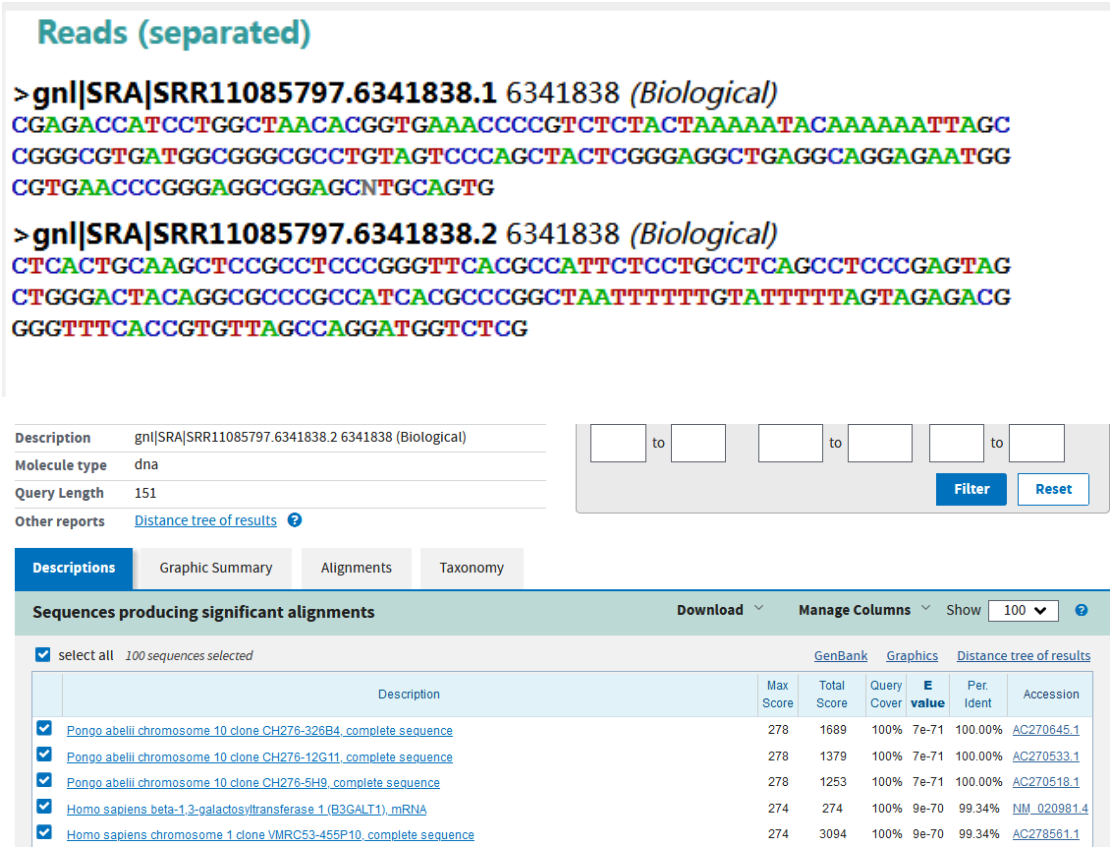
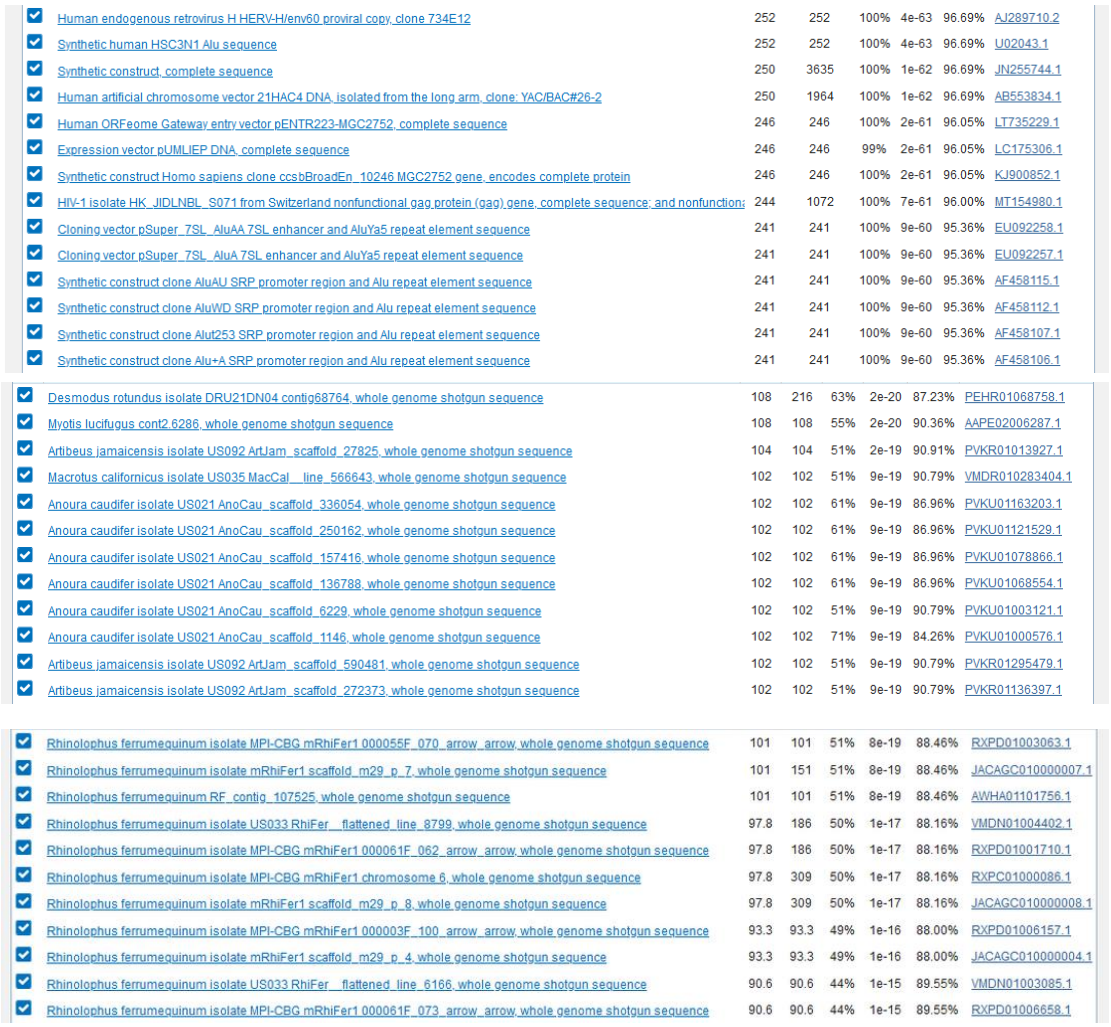


Figure.25a: 100% full-length matched reads to Hominid(Pongo Albelii) genomic DNA.



Human endogenous retrovirus H HERV-H/env60 proviral copy, clone 734E12	252	252	100%	4e-63	96.69%	AJ289710.2
Synthetic human HSC3N1 Alu sequence	252	252	100%	4e-63	96.69%	U02043.1
Synthetic construct, complete sequence	250	3635	100%	1e-62	96.69%	JN255744.1
Human artificial chromosome vector 21HAC4 DNA, isolated from the long arm, clone: YAC/BAC#26-2	250	1964	100%	1e-62	96.69%	AB553834.1
Human ORFeome Gateway entry vector pENTR223-MGC2752, complete sequence	246	246	100%	2e-61	96.05%	LT735229.1
Expression vector pUMLIEP DNA, complete sequence	246	246	99%	2e-61	96.05%	LC175306.1
Synthetic construct Homo sapiens clone ccsbBroadEn_10246 MGC2752 gene, encodes complete protein	246	246	100%	2e-61	96.05%	KJ900852.1
HIV-1 isolate HK_JIDLNBL_S071 from Switzerland nonfunctional gag protein (gag) gene, complete sequence; and nonfunction	244	1072	100%	7e-61	96.00%	MT154980.1
Cloning vector pSuper_7SL_AluAA 7SL enhancer and AluYa5 repeat element sequence	241	241	100%	9e-60	95.36%	EU092258.1
Cloning vector pSuper_7SL_AluA 7SL enhancer and AluYa5 repeat element sequence	241	241	100%	9e-60	95.36%	EU092257.1
Synthetic construct clone AluAU SRP promoter region and Alu repeat element sequence	241	241	100%	9e-60	95.36%	AF458115.1
Synthetic construct clone AluWD SRP promoter region and Alu repeat element sequence	241	241	100%	9e-60	95.36%	AF458112.1
Synthetic construct clone Alut253 SRP promoter region and Alu repeat element sequence	241	241	100%	9e-60	95.36%	AF458107.1
Synthetic construct clone Alu+A SRP promoter region and Alu repeat element sequence	241	241	100%	9e-60	95.36%	AF458106.1
Desmodus rotundus isolate DRU21DN04 contig68764, whole genome shotgun sequence	108	216	63%	2e-20	87.23%	PEHR01068758.1
Myotis lucifugus cont2.6286, whole genome shotgun sequence	108	108	55%	2e-20	90.36%	AAPE02006287.1
Artibeus jamaicensis isolate US092 ArtJam_scaffold_27825, whole genome shotgun sequence	104	104	51%	2e-19	90.91%	PVKR01013927.1
Macrotus californicus isolate US035 MacCal__line_566643, whole genome shotgun sequence	102	102	51%	9e-19	90.79%	VMDR010283404.1
Anoura caudifer isolate US021 AnoCau_scaffold_336054, whole genome shotgun sequence	102	102	61%	9e-19	86.96%	PVKU01163203.1
Anoura caudifer isolate US021 AnoCau_scaffold_250162, whole genome shotgun sequence	102	102	61%	9e-19	86.96%	PVKU01121529.1
Anoura caudifer isolate US021 AnoCau_scaffold_157416, whole genome shotgun sequence	102	102	61%	9e-19	86.96%	PVKU01078866.1
Anoura caudifer isolate US021 AnoCau_scaffold_136788, whole genome shotgun sequence	102	102	61%	9e-19	86.96%	PVKU01068554.1
Anoura caudifer isolate US021 AnoCau_scaffold_6229, whole genome shotgun sequence	102	102	51%	9e-19	90.79%	PVKU01003121.1
Anoura caudifer isolate US021 AnoCau_scaffold_1146, whole genome shotgun sequence	102	102	71%	9e-19	84.26%	PVKU01000576.1
Artibeus jamaicensis isolate US092 ArtJam_scaffold_590481, whole genome shotgun sequence	102	102	51%	9e-19	90.79%	PVKR01295479.1
Artibeus jamaicensis isolate US092 ArtJam_scaffold_272373, whole genome shotgun sequence	102	102	51%	9e-19	90.79%	PVKR01136397.1
Rhinolophus ferrumequinum isolate MPI-CBG mRhiFer1 000055F_070_arrow_arrow, whole genome shotgun sequence	101	101	51%	8e-19	88.46%	RXPD01003063.1
Rhinolophus ferrumequinum isolate mRhiFer1 scaffold_m29_p_7, whole genome shotgun sequence	101	151	51%	8e-19	88.46%	JACAGC010000007.1
Rhinolophus ferrumequinum RF_contig_107525, whole genome shotgun sequence	101	101	51%	8e-19	88.46%	AWHA01101756.1
Rhinolophus ferrumequinum isolate US033 RhiFer__flattened_line_8799, whole genome shotgun sequence	97.8	186	50%	1e-17	88.16%	VMDN01004402.1
Rhinolophus ferrumequinum isolate MPI-CBG mRhiFer1 000061F_062_arrow_arrow, whole genome shotgun sequence	97.8	186	50%	1e-17	88.16%	RXPD01001710.1
Rhinolophus ferrumequinum isolate MPI-CBG mRhiFer1 chromosome 6, whole genome shotgun sequence	97.8	309	50%	1e-17	88.16%	RXPC01000086.1
Rhinolophus ferrumequinum isolate mRhiFer1 scaffold_m29_p_8, whole genome shotgun sequence	97.8	309	50%	1e-17	88.16%	JACAGC010000008.1
Rhinolophus ferrumequinum isolate MPI-CBG mRhiFer1 000003F_100_arrow_arrow, whole genome shotgun sequence	93.3	93.3	49%	1e-16	88.00%	RXPD01006157.1
Rhinolophus ferrumequinum isolate mRhiFer1 scaffold_m29_p_4, whole genome shotgun sequence	93.3	93.3	49%	1e-16	88.00%	JACAGC010000004.1
Rhinolophus ferrumequinum isolate US033 RhiFer__flattened_line_6166, whole genome shotgun sequence	90.6	90.6	44%	1e-15	89.55%	VMDN01003085.1
Rhinolophus ferrumequinum isolate MPI-CBG mRhiFer1 000061F_073_arrow_arrow, whole genome shotgun sequence	90.6	90.6	44%	1e-15	89.55%	RXPD01006658.1

Fig.25b: BLAST search of this sequence revealed it to be a Homo Sapiens endogenous Retrovirus most similar to HIV-1, and is not found in any known bat genomic assemblies. This sequence is also found in several cloning vectors for mammalian DNA. Significance of these sequences are currently unknown.

SRX7724752 contained Traces of confirmed contamination from other organisms, in particularly that of order Carnivora, Rodentia and Homo Sapiens. As such DNA contamination mostly happen during extensive manipulation of samples in the labs, This indicate that SRX7724752 Contained traces of laboratory manipulation, including Canis Lupus Famillaris DNA contamination which could not have been present in a fecal sample of a bat, even assuming normal lab manipulation for sequencing purposes.

This indicate the sample may have been subjected to in-vitro manipulation.

No evidence of methodological reasons for the generation of anomalies in SRX7724752

In Order to test whether a specific sequencing technique was used for the sequencing of SRX7724752 which may have generated the anomalies observed above, we decided to use the sequencing depth of the Coronaviruses within SRX7724752 and compare it against another set of

mNGS sequencing data of identical sample, origin, institute and submitted at the same date, located in PRJNA606159.

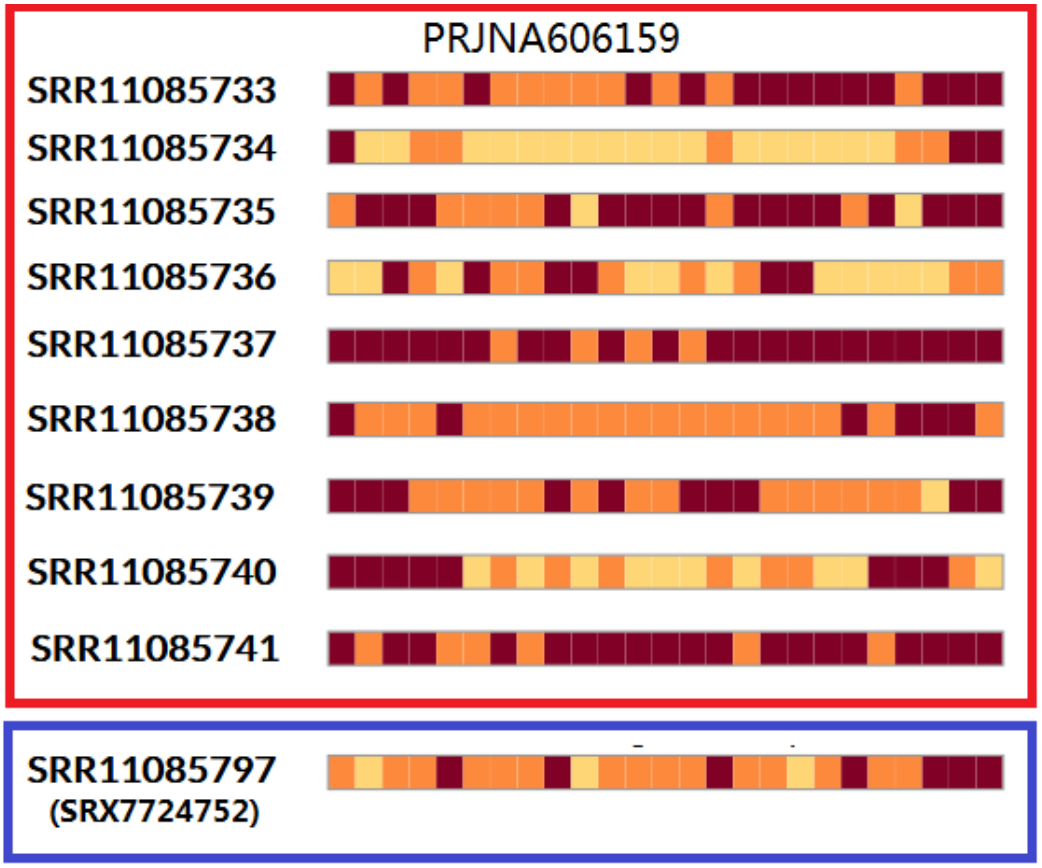


Fig. 26a: the Coverage map of Coronaviridae within the datasets located in PRJNA606159, compared against SRR11085797.

We generated the sequencing depth Heatmap [8] of all datasets located within PRJNA606159, and the sequencing depth pattern of the Coronavirus reads within such dataset does not show any statistical differences from that of SRR11085797.

SRX7724696: **RNA-Seq of Hipposideros larvatus: Anal swab**
1 ILLUMINA (Illumina HiSeq 3000) run: 13.5M spots, 3.9G bases, 1.8Gb downloads

Design: Total RNA was extracted from bronchoalveolar lavage fluid using the QIAamp Viral RNA Mini Kit following the manufacturers instructions. An RNA library was then constructed using the TruSeq Stranded mRNA Library Preparation Kit (Illumina, USA). Paired-end (150 bp) sequencing of the RNA library was performed on the HiSeq 3000 platform (Illumina).

Submitted by: Wuhan Institute of Virology, Chinese Academy of Sciences

SRX7724752: **RNA-Seq of Rhinolophus affinis:Fecal swab**
1 ILLUMINA (Illumina HiSeq 3000) run: 11.6M spots, 3.3G bases, 1.7Gb downloads

Design: Total RNA was extracted from bronchoalveolar lavage fluid using the QIAamp Viral RNA Mini Kit following the manufacturers instructions. An RNA library was then constructed using the TruSeq Stranded mRNA Library Preparation Kit (Illumina, USA). Paired-end (150 bp) sequencing of the RNA library was performed on the HiSeq 3000 platform (Illumina).

Submitted by: Wuhan Institute of Virology, Chinese Academy of Sciences

Fig.26b: the experimental design section of the datasets within PRJNA606159 is identical to that of SRX7724752.

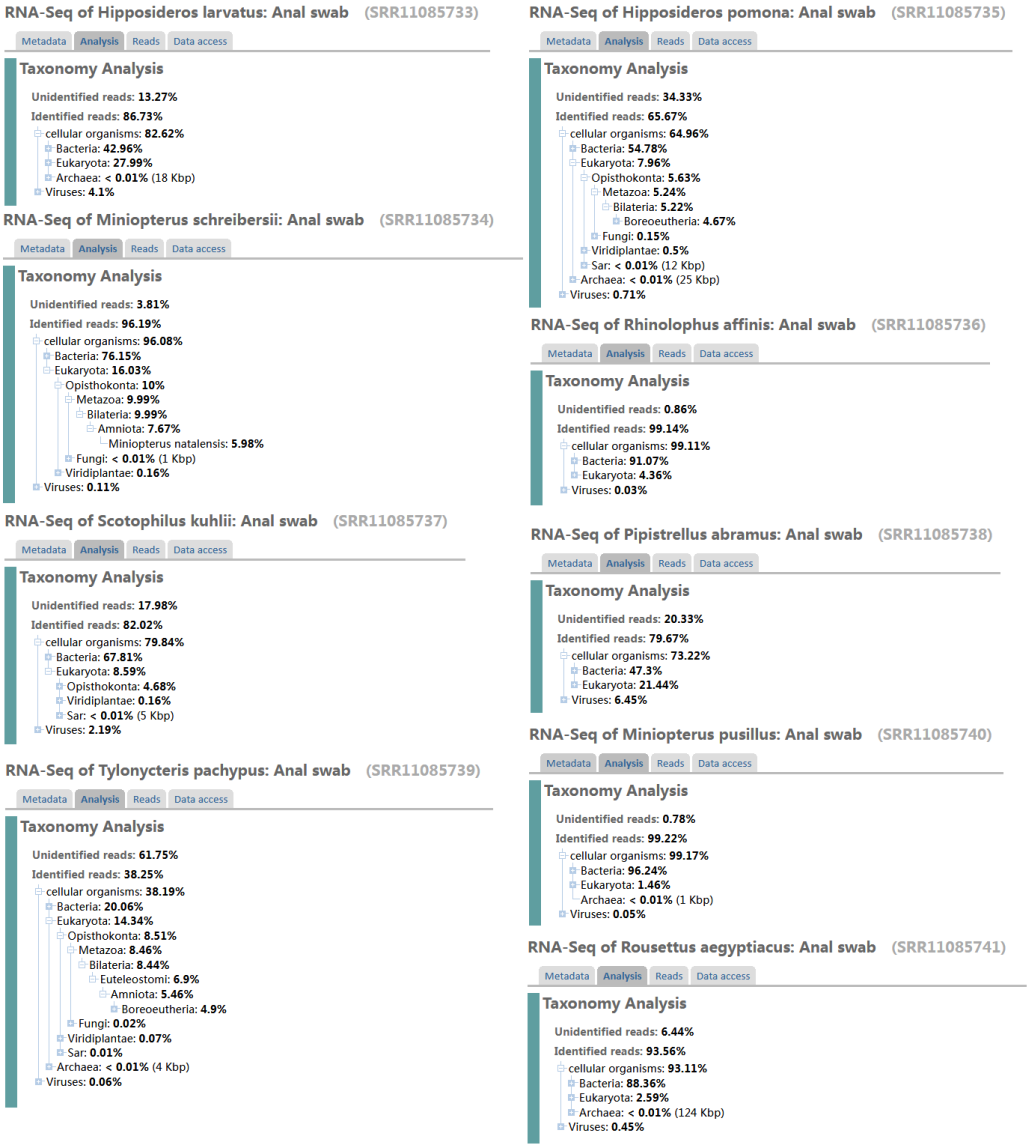


Fig.26c: No evidence of any anomalies were found within the datasets presented in PRJNA606159

We recently obtained a set of viral mNGS coverage data from a sequencing experiment that Uses PolyA enrichment for the selection of sequences [9].

Despite being isolated from the total RNA of freshly dissected and cleaned Bee Tissue samples, these PolyA enriched datasets displayed a heavy bias toward the 3'-end for all viral genomes that contained a polyA tail, and did not obtain any coverage past 8000nt to the 3'-end of such viral genomes. This is consistent with the fact that viral genomic RNA obtained from samples, even when freshly prepared, will always suffer from numerous RNA strand breaks, and therefore will be heavily biased toward the 3'-end as the enrichment process would have kept mostly the RNA that contained an intact polyA tail. As Coronaviruses have a PolyA tail, this is in sharp contrast to that found in SRX7724752, which does not show signs of such bias.

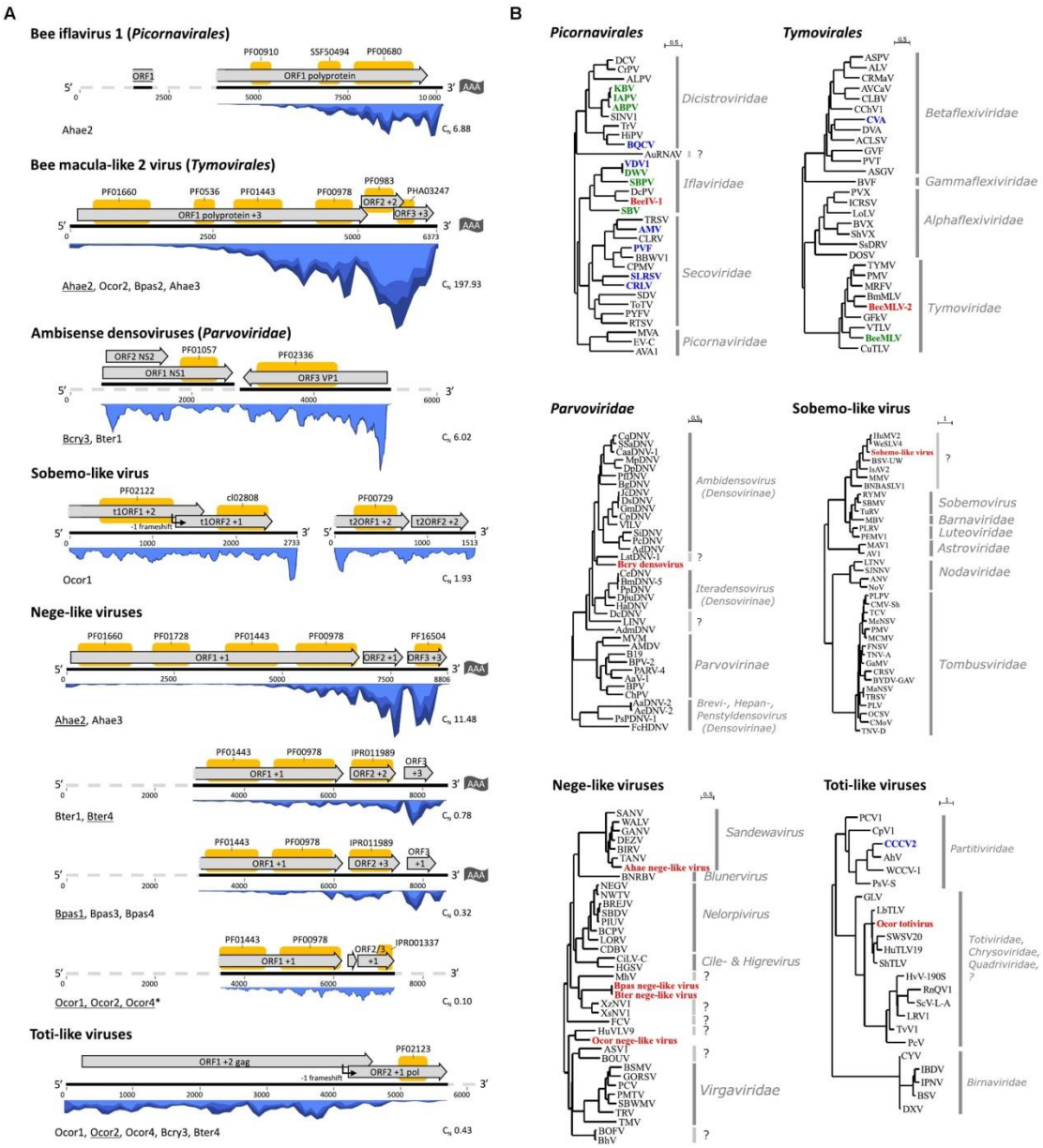


Figure 2 from [9]. A clear bias toward the 3'-end of RNA viral families that contained a polyA tail, was noticed.

The anomalies in SRX7724752 is associated with the absence of RNA viruses.

In order to further analyze the implications of the observed anomalies in SRX7724752, we performed a Keyword search on NCBI SRA using the Keyword "Bat feces" and "Bat fecal".
We did not find any evidence of an RNA virus (Riboviria) within any of the returned datasets that contained less than 2.5% bacteria in total cellular organisms that can be confirmed by BLAST.

Description	Phaseolus vulgaris endornavirus 1 isolate PvEV-1_Brazil poly ...
Molecule type	nucleic acid
Query Length	14072
Other reports	

⚠ No significant similarity found. For reasons why, click here

Figure 27: an example of a TRACE result that does not actually exist when BLASTed against the reference sequences of said virus.

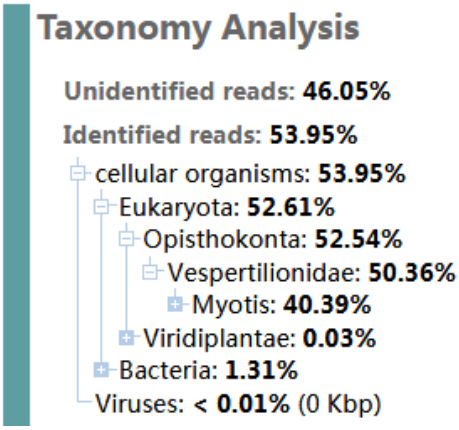


Figure 28a: an example of a bacteria-depleted dataset. An absence of Riboviria reads was noted.

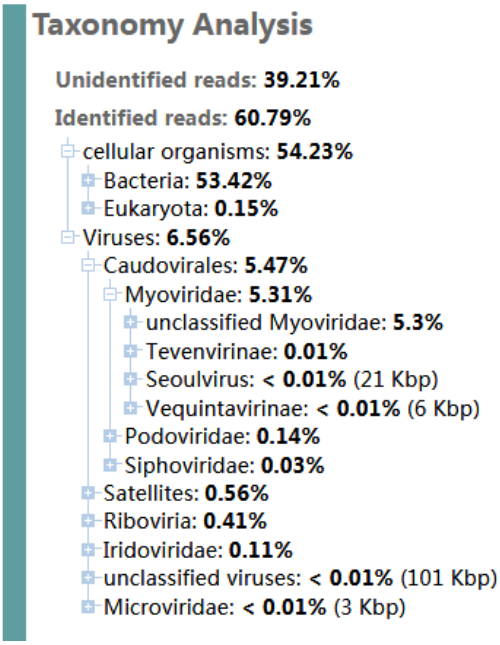


Figure 28b: in contrast, Riboviria reads are found only in datasets that contained a significant amount of bacteria.

In addition, We found only 1 dataset that contained any significant levels of a Telomere-like repeat sequence. However, this dataset does not contain any evidence of an RNA virus(Riboviria).

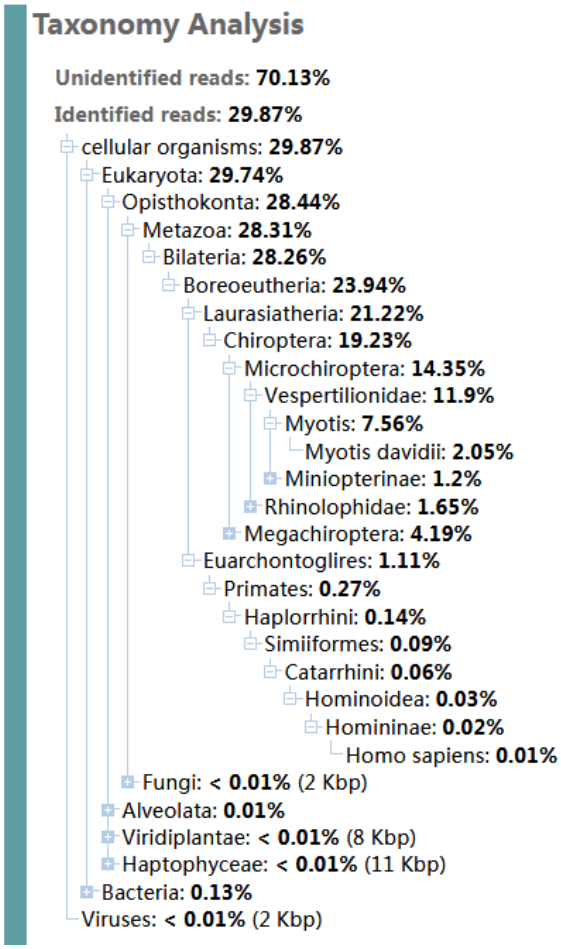


Figure 29: the only dataset with significant level of Telomere-like repeats (2%). There are no evidence of Riboviria(RNA viruses) within this dataset.

Signature of likely attenuation of the RaTG13 RBD.

The RaTG13 RBD has reduced binding affinity to ACE2 compared to other viruses of the same clade.

Recently, a publication which tested the binding affinity and infection efficiency of RaTG13 S to human ACE2[10] have been published, which suggest that unlike the other RBDs within this clade (namely SARS-CoV-2 and pCoV_GX), RaTG13 can not bind to human ACE2 efficiently, and is incapable of entering cells through ACE2 that is expressed at physiological levels.

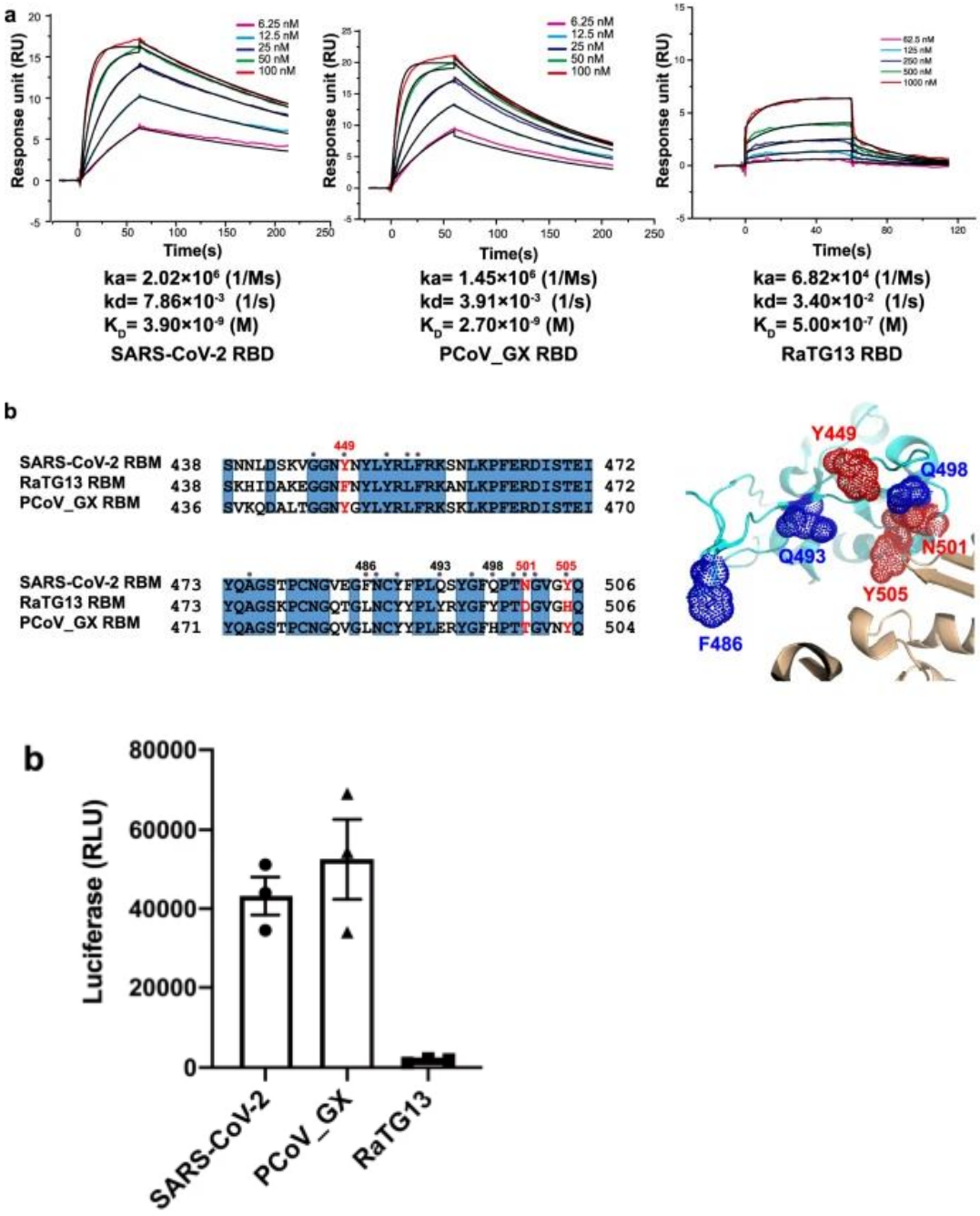


Fig.5a and 6b from [10]: RaTG13 RBD bound to human ACE2 very inefficiently, and did not show entry into hACE2-HEK293T cells at physiological level of ACE2 expression.

Initially considered as evidence of “bat specificity” for RaTG13, An recent test [11] did not find any higher binding affinity of the RaTG13 RBD to R.affinis ACE2 than to human ACE2—in fact, both the flow cytometry data and the pseudovirus entry data into HEK293T with Overexpressed ACE2 suggest a binding affinity of RaTG13 RBD to R.affinis ACE2 that is slightly lower than that of RaTG13 RBD to human ACE2.

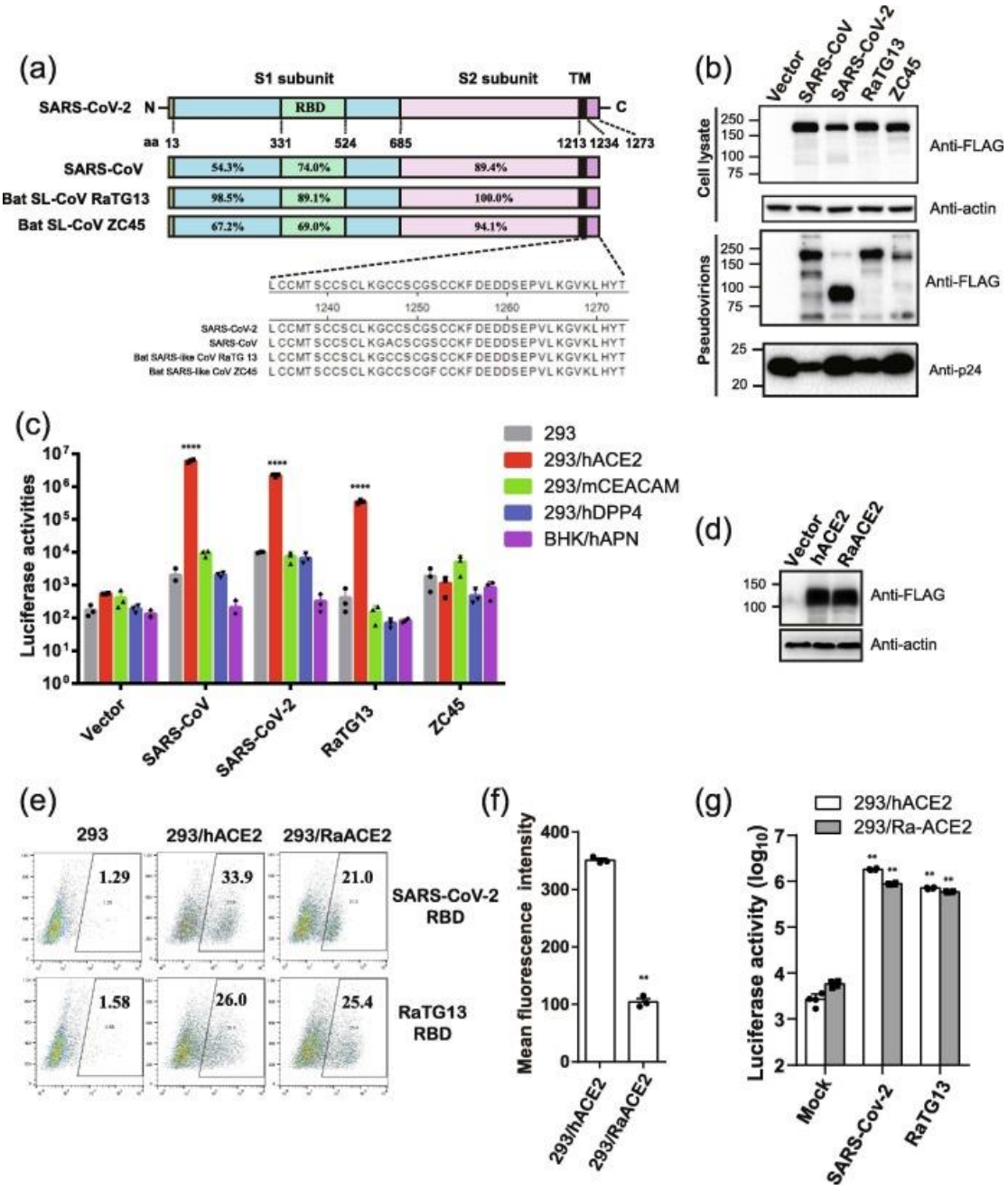
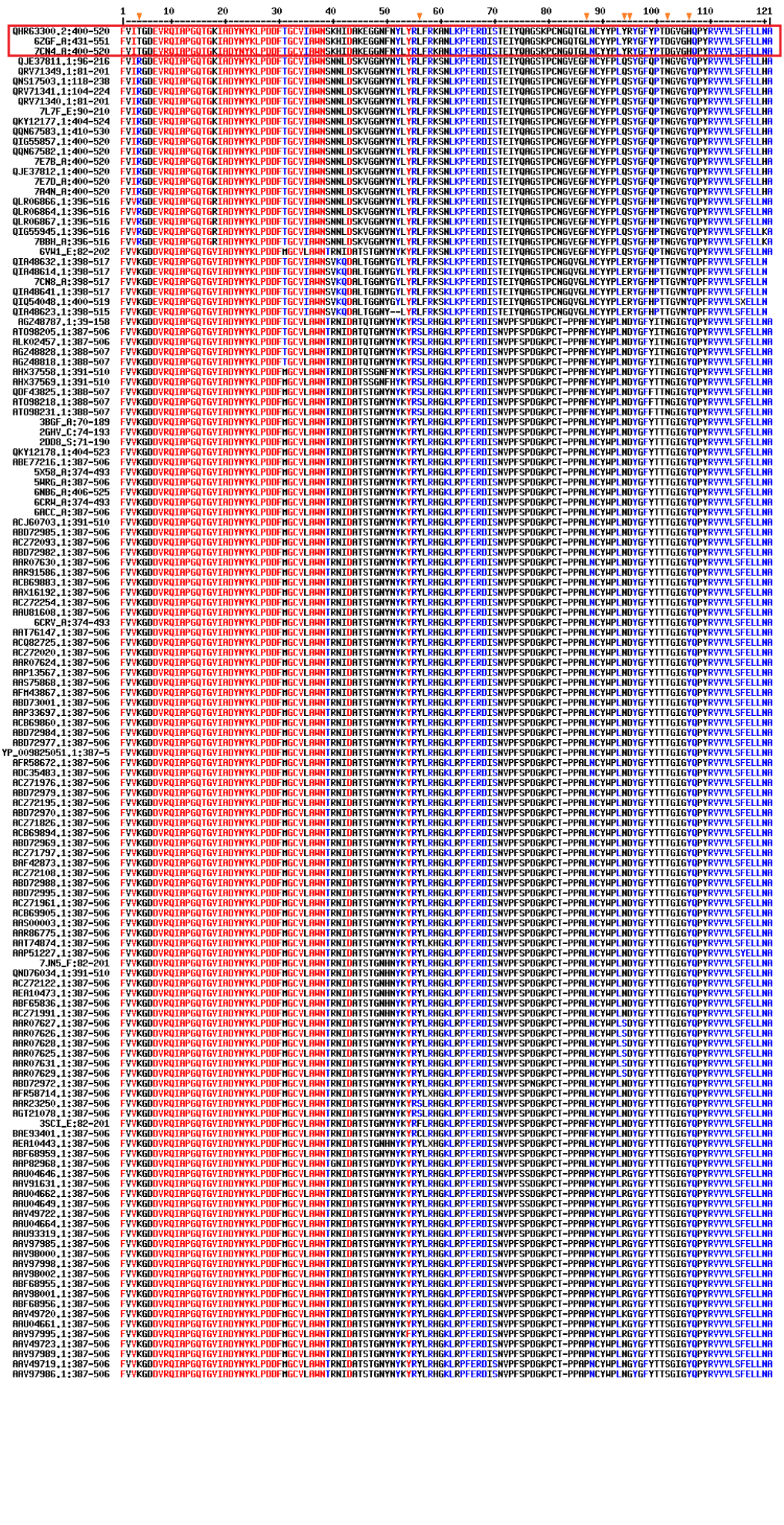


Fig.5 from [11]: d: The overexpression level for human and R.affinis ACE2 on the HEK293T cells are the same. e: The binding affinity of RaTG13 RBD-Fc on R.affinis ACE2(Lower Right) is slightly lower than the binding affinity of RaTG13 RBD-Fc on human ACE2(Lower Middle), and is significantly lower than the binding affinity of SARS-CoV-2 RBD to human ACE2(Upper middle). g: Pseudovirus entry assay using Lentivirus on HEK293T with Overexpressed ACE2 show that SARS-CoV-2-hACE2 > SARS-CoV-2-R.affinis ACE2 > RaTG13-hACE2 > RaTG13-R.affinis ACE2 by pseudovirus entry efficiency.

A multiple alignment of the RaTG13 Spike protein RBD to other Sarbecovirus Spike protein RBDs quickly revealed two specific residues—T403 and D501—that were never found in other Sarbecoviruses. In addition, H505 is found to be absent in all previous R.affinis infecting Sarbecoviruses.



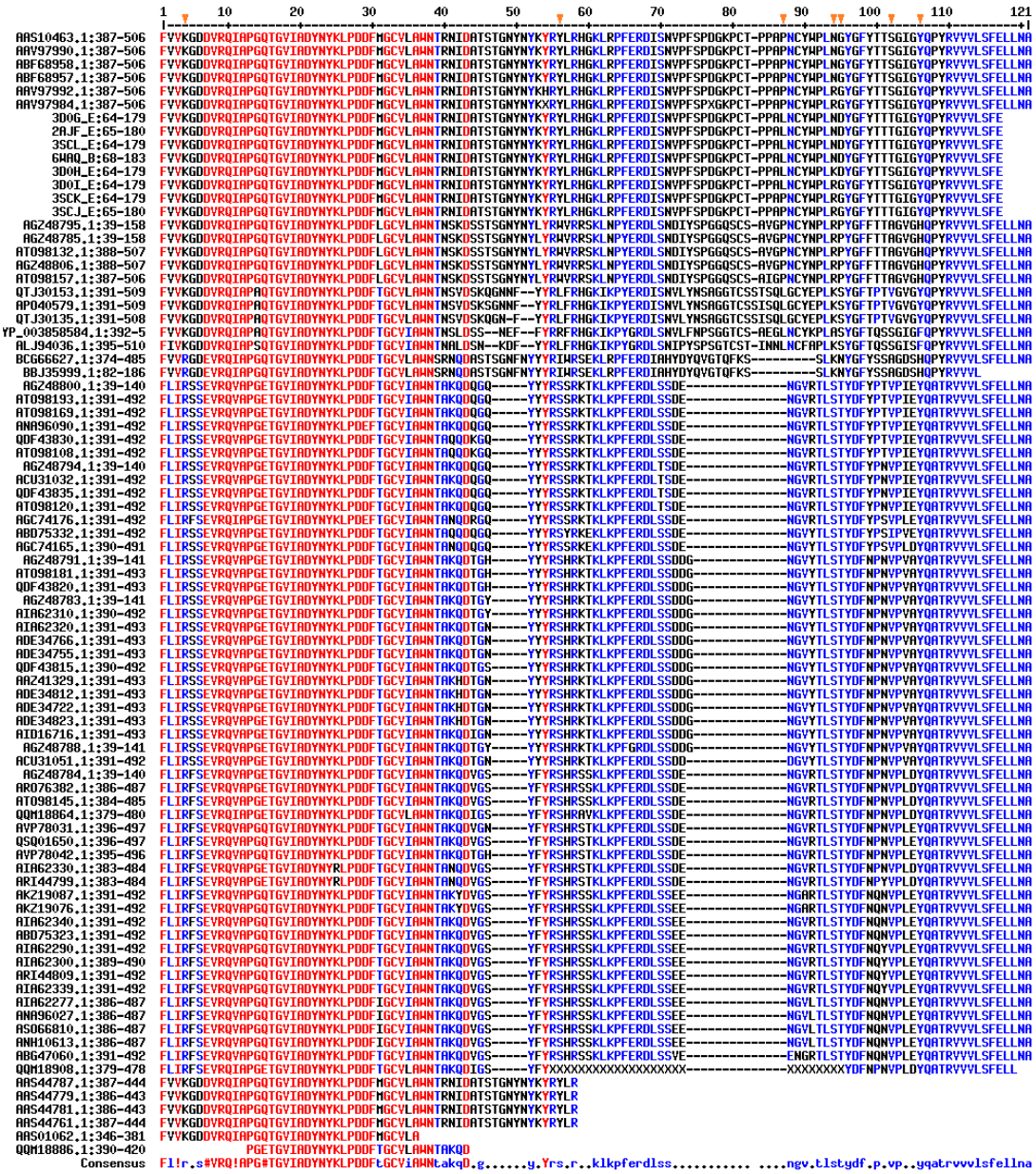


Fig.30a: multiple sequence alignment of all current known Sarbecovirus RBM sequences on NCBI. Orange arrows indicating critical residues for infection using ACE2[12][13]. Red square indicates nr and PDB sequences for the RaTG13 S.

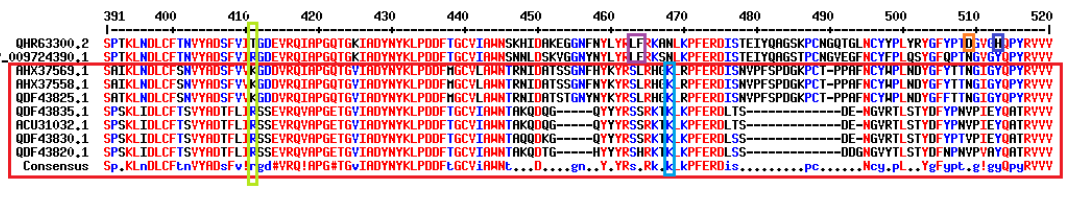


Fig.30b: multiple sequence alignment of R.affinis infecting Sarbecoviruses indicating that H505 is not found in other known Sarbecovirus strains infecting this species. Red square indicates previous Sarbecovirus with a host listed as “Rhinolophus Affinis”.

As these sites were found to be unique in RaTG13, and since these 2 sites resulted in a significant change in the residue’s general properties compared to the analogous position on all other

Sarbecovirus RBD known (Basic in all other RBDs->Neural polar for R403T, Neutral in all other RBDs -> Acidic for N501D), we set to deduce their effect in the binding of the RBD to ACE2. A publication using deep mutational scanning analysis[14] suggest that Y449F, N501D and Y505H resulting in the highest reduction of binding affinity to hACE2 when applied to SARS-CoV-2.

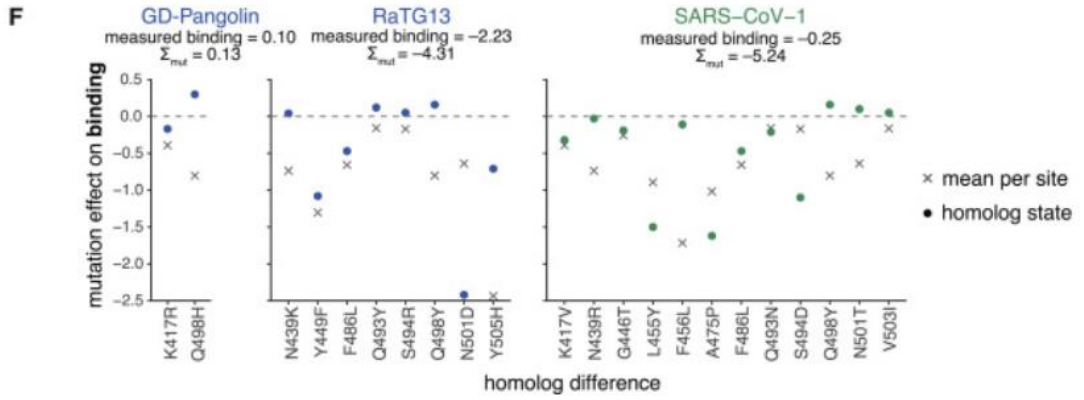
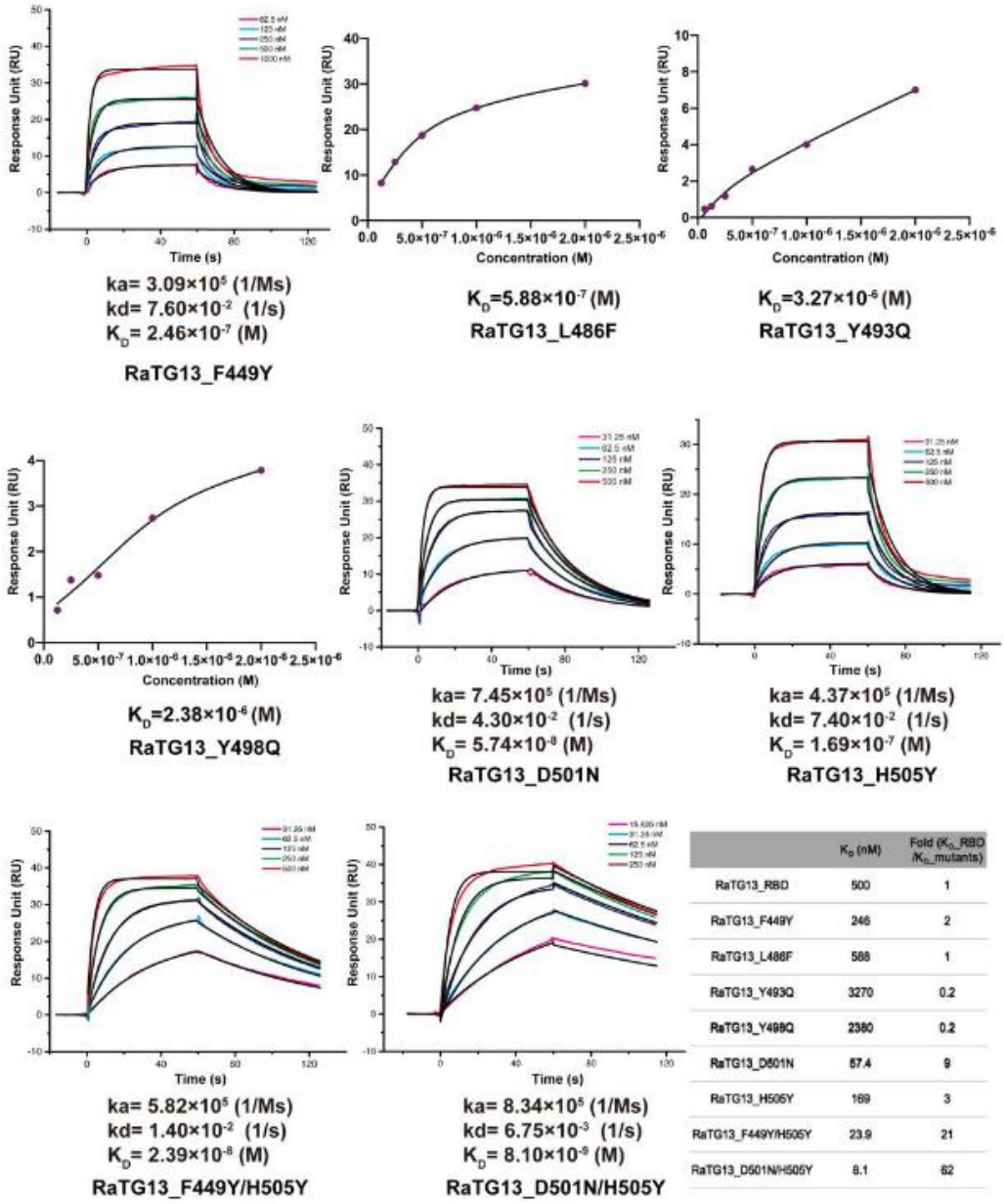


Figure 5F from [14] with different mutations from RaTG13 RBD applied to SARS-CoV-2 RBD. N501D resulted in the most severe reduction in binding affinity to ACE2, followed by Y449F and Y505H.

Using structural analysis, we discovered that the residues on ACE2 surrounding Y449, N501 and Y505 in SARS-CoV-2 are identical between Human and R.affinis ACE2, indicating that the reduction of binding affinity conferred by N501D, Y449F and Y505H would also cause the same reduction in binding affinity to R.affinis ACE2. Indeed, no sequence from R.affinis contained D501 or H505, implying that these 2 residues are also avoided in viruses that naturally circulating in this species, indicating that they cause a substantial reduction of viral fitness if introduced.



	K_D (nM)	Fold (K_D_RBD /K_D_mutants)
RaTG13_RBD	500	1
RaTG13_F449Y	246	2
RaTG13_L486F	588	1
RaTG13_Y493Q	3270	0.2
RaTG13_Y498Q	2380	0.2
RaTG13_D501N	57.4	9
RaTG13_H505Y	169	3
RaTG13_F449Y/H505Y	23.9	21
RaTG13_D501N/H505Y	8.1	62

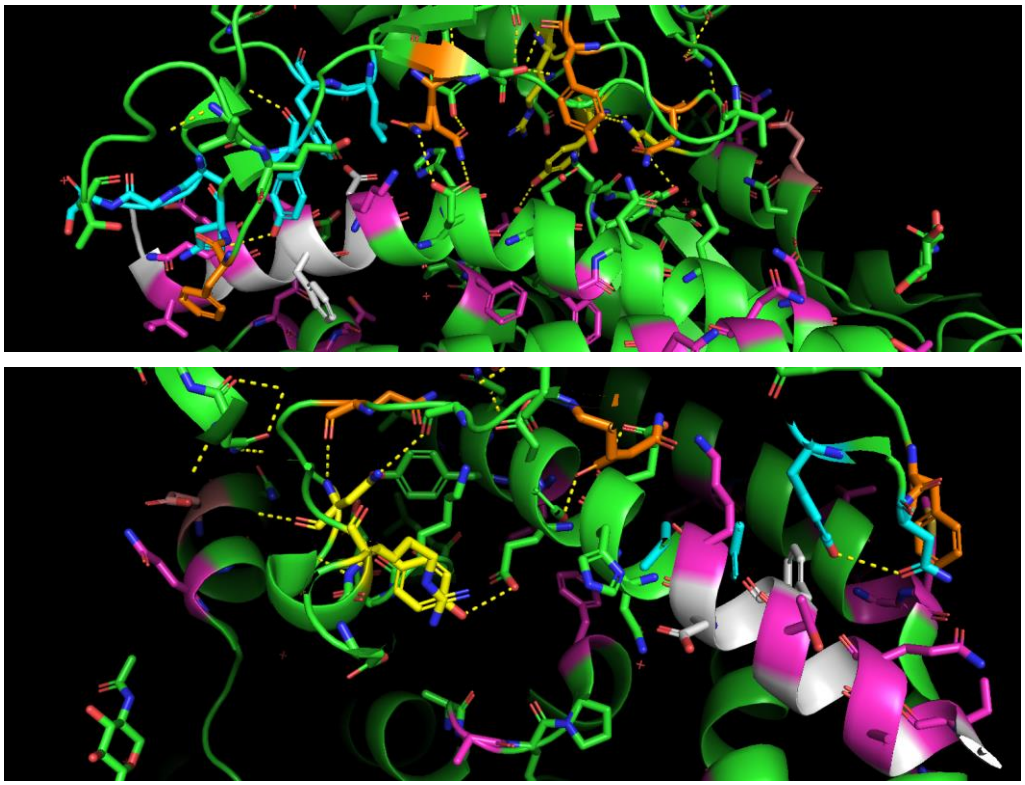
Fig. S10 from [10] denoting the effect of different amino acid substitutions on the RaTG13 RBD and their effects on binding to hACE2.

Indeed, upon mutating the position 501 and position 505 into N and Y, the consensus sequence within the R.affinis species, the binding affinity of RaTG13 RBD to ACE2 is fully restored, with an 82-fold increase compared to QHR63300.2 in term of ACE2 binding affinity. As these 2 residues interfaces with residues that were conserved between human and R.affinis ACE2, it is predicted that the same change will also improve the binding of the RaTg13 RBD to R.affinis ACE2, potentially to similar levels as SARS-CoV-2 to human ACE2. Such a prediction of the critical requirement of a neutral 501, and in a lesser extent, Y505, on R.affinis ACE2 have been confirmed through Surface Plasmon Resonance testing of the RaTG13 RBD with mutaion D501N and H505Y introduced into it, where the binding affinity was found to be massively improved on R.affinis ACE2 by 57.83x and 6.29x respectively.[30]

Interestingly, the authors of [11] did not test the effect on ACE2 binding affinity of D501N or H505Y in the RaTG13 RBD, nor did they test any of the RBD mutants on R.affinis ACE2.

```
                 1       10        20        30        40        50        60        70        80        90       100       110       120       130
                 |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
NP_001358344.1   MSSSSWLLLSLVAVTAAQSTIEEQAKTFLDKFNHEAEDLFYQSSLASWNYNTNITEENVQNMNNAGDKWSAFLKEQSTLAQMYPLQEIQNLTVKLQLQALQQNGSSVLSEDKSKRLNTILNTMSTIYSTG
QMQ39244.1       MSGSSWLLLSLVAVTAAQSTTEDRAKIFLDNFNHEAEDLSYQSSLASWEYNTNISDENVQKMDEAGAKWSAFYEEQSKLAKNYPLEEIQTVPVKLQLQILQQSGSPVLSEDKSKRLNSILNAMSTIYSTG
Consensus        MSgSSWLLLSLVAVTAAQSTiE#rAKiFLDnFNHEAEDLfYQSSLASW#YNTNIs#ENVQnM##AGaKWSAFleEQSkLAqnYPL#EIQnlpVKLQLQaLQQnGSpVLSEDKSKRLNsILNaMSTIYSTG

                 131     140       150       160       170       180       190       200       210       220       230       240       250       260
                 |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
NP_001358344.1   KVCNPDNPQECLLLEPGLNEIMANSLDYNERLWAWESWRSEVGKQLRPLYEEYVVLKNEMARANHYEDYGDYWRGDYEVNGVDGYDYSRGQLIEDVEHTFEEIKPLYEHLHAYVRAKLMNAYPSYISPIG
QMQ39244.1       KVCKPNNPQECFLLEPGLDNIMGTSKDYNERLWAWEGWRAEVGKQLRPLYEEYVALKNEMARGYHYEDYGDYWRRDYETEESSGSGYSRDQLMKDYDRIFTEIKPLYEHLHAYVRTKLMDTYPFHISPTG
Consensus        KVCnP#NPQEClLLEPGL##IManSlDYNERLWAWEgWRaEVGKQLRPLYEEYVaLKNEMARanHYEDYGDYWRrDYEt#esdGsdYSRdQLieDV#riFeEIKPLYEHLHAYVRaKLM#aYPfhISPiG

                 261     270       280       290       300       310       320       330       340       350       360       370       380       390
                 |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
NP_001358344.1   CLPAHLLGDMWGRFWTNLYSLTVPFGQKPNIDVTDAMVDQAWDAQRIFKEAEKFFVSVGLPNMTQGFWENSMLTDPGNVQKAVCHPTAWDLGKGDFRILMCTKVTMDDFLTAHHEMGHIQYDMAYAAQPF
QMQ39244.1       CLPAHLLGDMWGRFWTNLYPLTVPFGQKPNIDVTDAMVNQGWDANRIFKEAEKFFVSVGLPNMTEGFWNNSMLTEPGDGRKVVCHPTAWDLGKGDFRIKMCTKVTMEDFLTAHHEMGHIQYDMAYATQPY
Consensus        CLPAHLLGDMWGRFWTNLYpLTVPFGQKPNIDVTDAMV#QaWDA#RIFKEAEKFFVSVGLPNMT#GFW#NSMLT#PG#grKaVCHPTAWDLGKGDFRIlMCTKVTM#DFLTAHHEMGHIQYDMAYAaQP%

                 391     400       410       420       430       440       450       460       470       480       490       500       510       520
                 |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
NP_001358344.1   LLRNGANEGFHEAVGEIMSLSAATPKHLKSIGLLSPDFQEDNETEINFLLKQALTIVGTLPFTYMLEKWRWMVFKGEIPKDQWMKKWWEMKREIVGVVEPVPHDETYCDPASLFHVSNDYSFIRYYTRTL
QMQ39244.1       LLRNGANEGFHEAVGEVMSLSVATPKHLKTMGLLSPDFLEDNETEINFLLKQALNIVGTLPFTYMLEKWRWMVFRGEIPKEEWMKKWWEMKRDLVGVVEPVPHDETYCDPASLFHVRNDYSFIRYYTRTI
Consensus        LLRNGANEGFHEAVGE!MSLSaATPKHLKsiGLLSPDFqEDNETEINFLLKQALnIVGTLPFTYMLEKWRWMVFrGEIPK##mMKKWWEMKR#iVGVVEPVPHDETYCDPASLFHVaNDYSFIRYYTRTi

                 521     530       540       550       560       570       580       590       600       610       620       630       640       650
                 |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
NP_001358344.1   YQFQFQEALCQAAKHEGPLHKCDISNSTEAGQKLFNMLRLGKSEPWTLALENVVGAKNMNVRPLLNYFEPLFTWLKDQNKNSFVGWSTDWSPYADQSIKVRISLKSALGDKAYEWNDNEMYLFRSSVAYA
QMQ39244.1       FEFQFHEALCRIAQHDGPLHKCDISNSTDAGKKLHQMLSVGKSQPWTVTLKDIVDSRNMDVGPLLRYFEPLYTWLQEQNRKSYVGWNTDWSPYSDQSIKVRISLKSALGEKAYEWNDNEMYLFRSSVAYA
Consensus        %#FQFqEALCraAqH#GPLHKCDISNST#AGqKLh#MLrlGKS#PWTlaLe#!YdarNM#VrPLLrYFEPL%TWLq#QNrnS%VGWnTDWSPYaDQSIKVRISLKSALG#KAYEWNDNEMYLFRSSVAYA

                 651     660       670       680       690       700       710       720       730       740       750       760       770       780
                 |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
NP_001358344.1   MRQYFLKVKNQMILFGEEDVRVANLKPRISFNFFVTAPKNVSDIIPRTEVEKAIRMSRSRINDAFRLNDNSLEFLGIQPTLGPPNQPPVSIWLIVFGVVMGVIVVGIVILIFTGIRDRKKKNKARSGENP
QMQ39244.1       MREYFSK-KNQPILFGVEMVMVSNLKPRISFNFHVTSPGNVSDIIPRSEVEGAIRMSRSRINDAFRLDDNSLEFLGIQPTLGPPYQPPVTIWLIVFGVVMAVVVVGIVVLIITGIRDRRKTDQARSEENP
Consensus        MR#YFlK.KNQmILFGeE#VrVaNLKPRISFNFhVTaPgNVSDIIPRsEVEgAIRMSRSRINDAFRL#DNSLEFLGIQPTLGPPnQPPVsIWLIVFGVVMaV!VVGIV!LIiTGIRDRrKk#qARSeENP

                 781     790       800  805
                 |--------+---------+----|
NP_001358344.1   YASIDISKGENNPGFQNTDDVQTSF
QMQ39244.1       YSSVDLNKGENNPGFQNGDDVQTSF
Consensus        YaS!DinKGENNPGFQNgDDVQTSF
```

Fig.31: Alignment of Human and R.affinis ACE2.



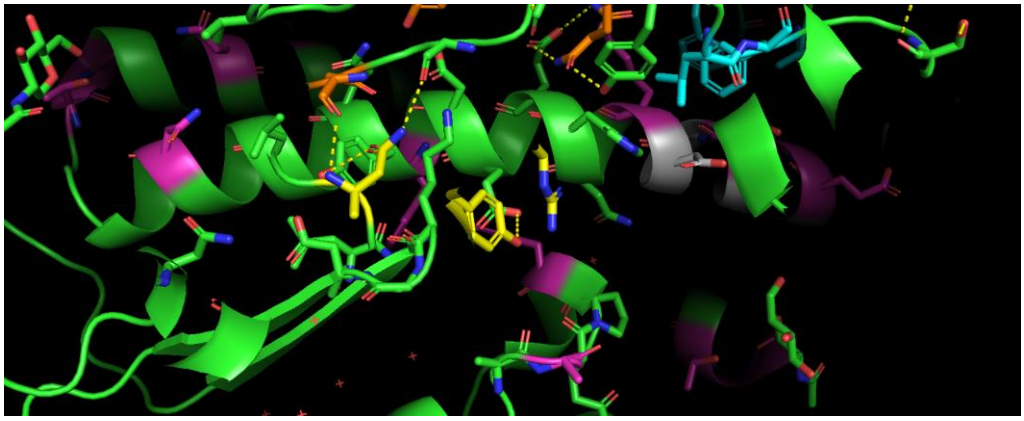


Fig.32: Structure of the pocket surrounding R403, Y449, N501 and Y505 on the SARS-CoV-2 RBD protein. Yellow is the R403, N501 and Y505 on SARS-CoV-2 RBD, green and white sticks denote residues that are identical between Human and R.affinis ACE2, Orange sticks denote residues that were different between SARS-CoV-2 and RaTG13 RBD and Magenta sticks denote residues that were different between human and R.affinis ACE2.

Consistent with the discovery of two residues (T403 and D501) with unique chemical properties that have never been recorded in any Sarbecoviruses, This reduction of binding affinity was found to be general for all animal ACE2 tested, with the highest recorded pseudovirus entry (RLU) being ~10^1 times lower than the entry efficiency of SARS-CoV-2 S on hACE2.

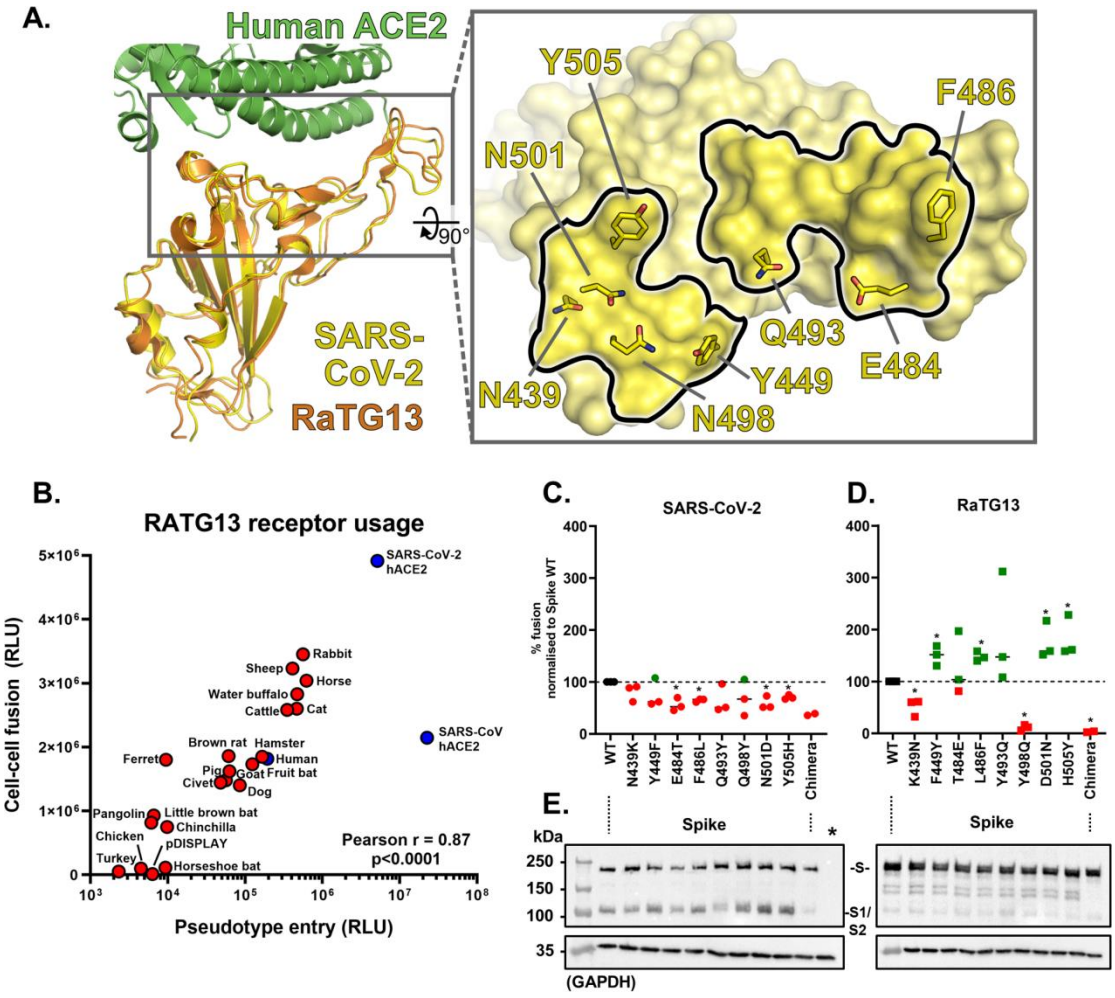


Fig.4 from [16]: Cell-to-cell fusion and pseudovirus entry assay of SARS-CoV-2, SARS-CoV and

RaTG13 on different ACE2 orthologues overexpressed on HEK293T cells.

The RaTG13 S exhibits a restricted tropism and is specific to Immortalized Kidney cells.

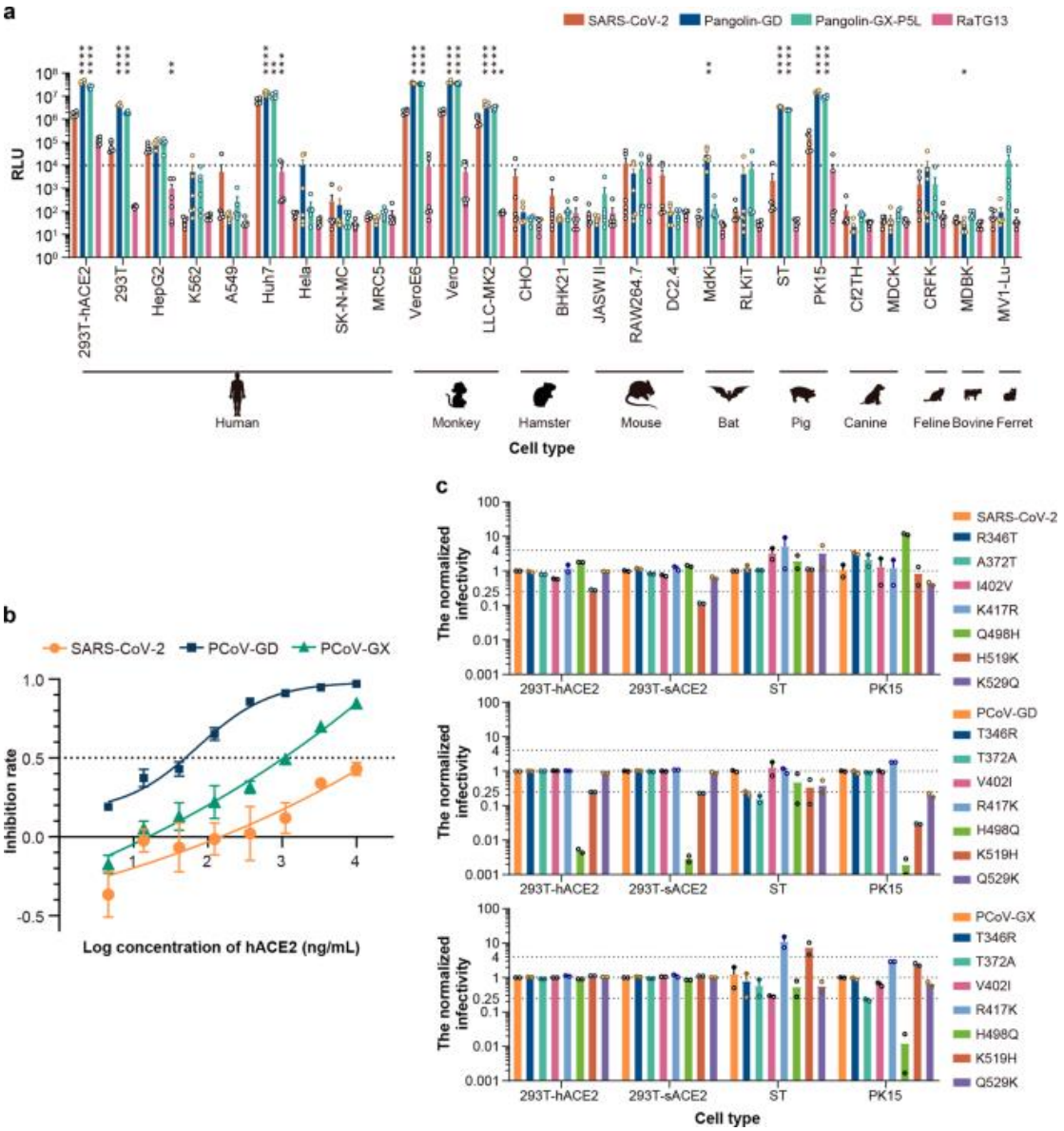


Fig.1 from [15] comparing the tropism of SARS-CoV-2, pCoV-GD, pCoV-GX and RaTG13. While only HEK293T-ACE2 displayed an infectivity for RaTG13 pseudovirus with an RLU of over 10^4, both HEK293T-ACE2 and PK15 displayed a roughly 10^1 difference between the infectivity for SARS-CoV-2 and RaTG13. All other cell line where there RaTG13 show above-background infectivity, except for mouse Macrophage cell line RAW264.7, had a difference of 10^2 to 10^3 in term of pseudovirus entry for SARS-CoV-2 and RaTG13, all with infectivity of SARS-CoV-2 significantly above that of RaTG13.

As ACE2 bind integrins through the KGD motif on position 353-355 which is conserved in human and R.affinis ACE2[17][18][19], It is modeled that ACE2 in physiological concentrations is bound to Integrin α5β1 and is inaccessible to binding by the RBD of Sarbecovirus Spike proteins[18], unless it is displaced by an KGD/RGD motif that is found in all RBD sequences that lacked the two deletions in SL-CoVs RBDs that does not use ACE2 for entry.[20]

The primary feature of the HEK293T cells (ACE2) used in [11] is their substantially higher expression of ACE2 over their constitutional ITGB1 expression, which will result in large amount of free ACE2 that is physiologically unrealistic in real tissues like bat intestines or human lungs.

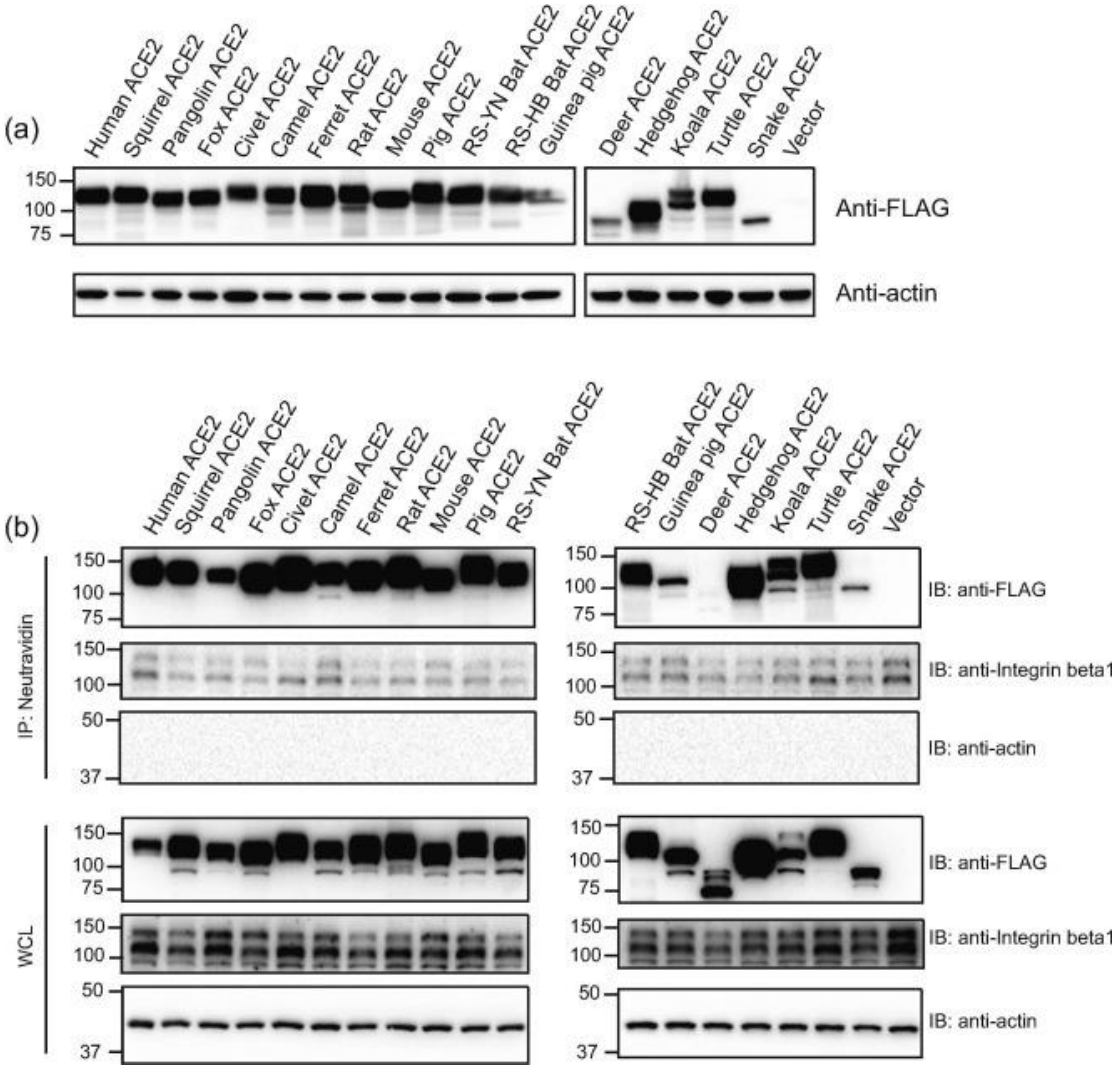


Fig.2 from [11]: Overexpression of ACE2 orthologues on HEK293T result in an overwhelmingly high amount of both whole-cell and surface ACE2 molecules comparing to the amount of ITGB1 molecules available for binding to ACE2.

Indeed, Inhibiting the interaction of the Spike to α5β1integrins using an integrin-inhibiting peptide have been found to reduce the binding and entry efficiency of SARS-CoV-2 to hACE2 and VERO E6 cells[18], and Integrins have been speculated as a co-receptor for SARS-CoV-2[19][21].

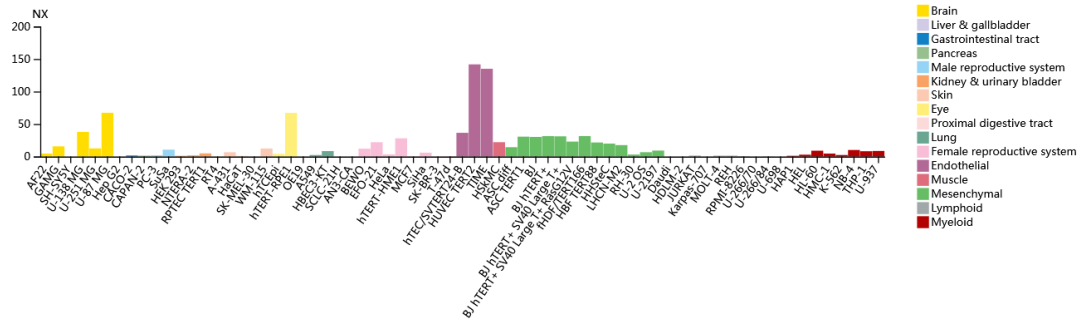


Fig.33a: expression level of ITGA5 on different cell lines.

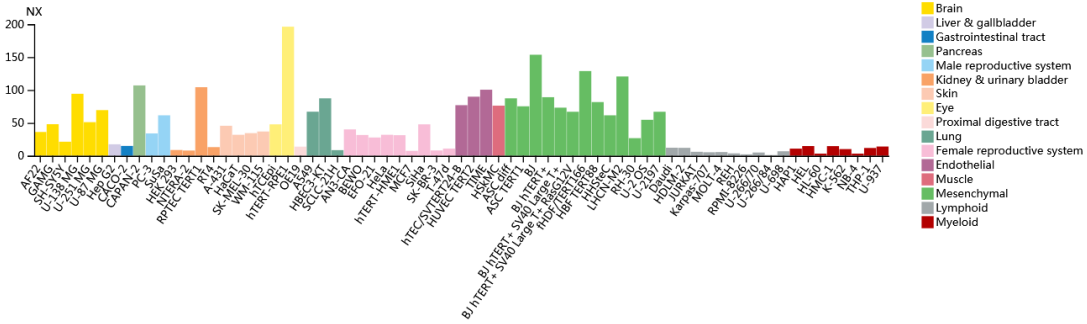


Fig.33b: expression level of ITGB1 on different cell lines.

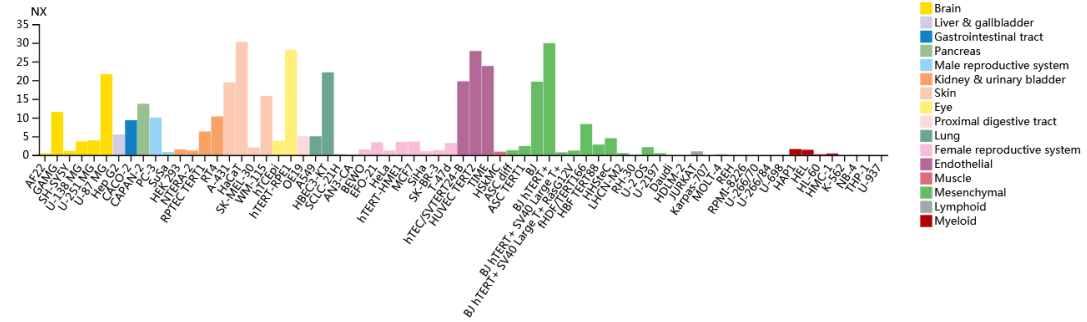


Fig.33c: expression level of ITGA2 on different cells lines.

The effect of integrin expression on the entry efficiency of RaTg13 and SARS-CoV-2 S pseudotyped lentivirus is demonstrated by the difference between the entry efficiency of the 3 Spike proteins to HeLa-ACE2 and HEK293T cells[22].

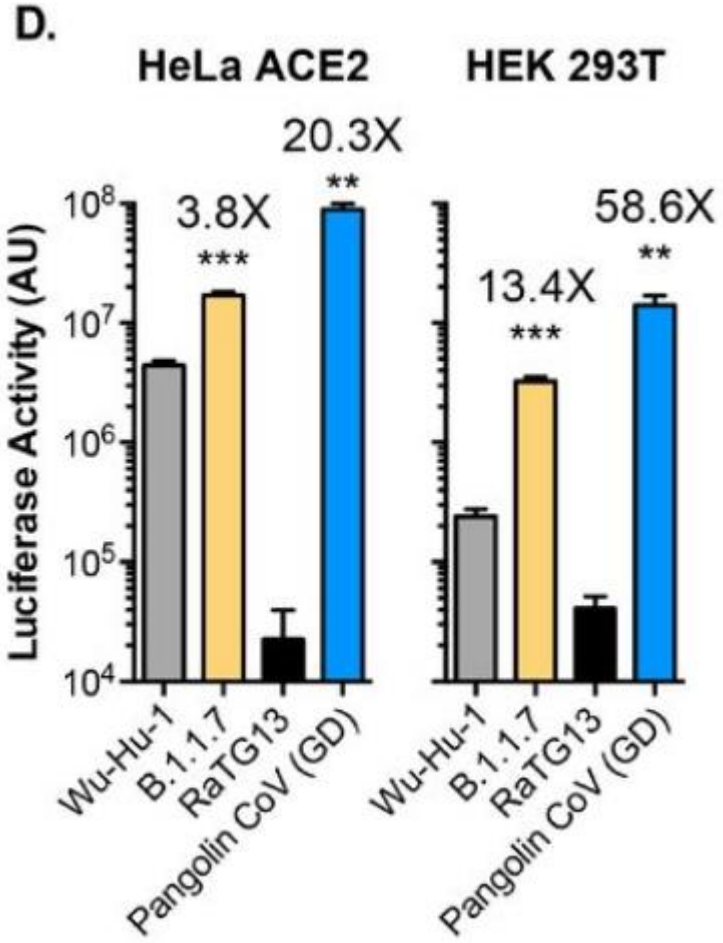


Fig.3d from[22]: HeLa ACE2 show substantially higher ratios between WuHu-1 or B1.1.7 entry and RaTG13 entry, compared to HEK293T.

As immortalized Kidney cells (HEK293T) are found to have much lower level of ITGB1, ITGA2 or ITGA5 expression, compounded with the fact that HEK293T-ACE2 cells is the only human cell line that support substantial entry by RaTG13 pseudotyped lentivirus with RLU above 10^4[15] despite all cell lines from humans expressing only human ACE2, This highly restricted tropism of RaTg13 toward Immortalized Kidney cells likely indicate serial passage within such cells.

Incidentally, the only cell line from R.affinis in possession of the Wuhan Institute Of Virology (WIV) was RaK4324 cells from the Kidneys, which would have been the laboratory passage host for RaTG13 if it have been cultured prior to sequencing.[23]

As both the neutral 501 and the basic (R/K)403 are found to be 100% conserved in all Sarbecovirus RBD proteins except for RaTG13, these 2 positions are likely indispensable for the in-vivo fitness of Sarbecoviruses in both reservoir hosts, in other animals and in humans.

Since experimental evidence have validated the broadly detrimental effect of both D501 and T403 on the RaTG13 S on viral fitness (RBD binding to ACE2, Spike entry into cells)[14][31](although [31] did not find an correlation between ATN-161 and SARS-CoV-2 as opposed to [18], it has been found that the CaCo-2 cells they used expresses mainly ITGA2 instead of ITGA5, and both CaLu-3 and Caco-2 expresses ITGA2 significantly[32], which does not interact with ATN-161[33] and therefore will not be inhibited by the peptide.), these two

positions can be considered as signature of attenuation in the RaTG13 RBD protein.

RaTG13 is an attenuated vaccine strain cultured in immortalized bat kidney cells?

The only known cell line from R.affinis in possession of the Wuhan Institute of Virology was RaK4324 Primary Kidney cells[23], which were used in the isolation and culture of bat Coronaviruses.

Should the SRA dataset of RaTG13, SRX7724752, have been a cell culture of an attenuated virus within an Immortalized version of the RaK4324 cells, It would simultaneously explain nearly all the known anomalies associated with the raw read data and the nucleotide sequence of RaTG13, MN996532.

Bat telomeres are known to not shorten with age[24], which indicate that the mechanism of telomere erosion is likely absent in bat cells. Should a traditional telomerase based immortalization strategy being used on a culture of R.affinis Kidney cells, one of the likely outcome is that telomeric sequences will grow uncontrollably and accumulate to very high fractions after extensive passage of the cell line due to the TERT activity not being balanced by telomere erosion mechanisms that were found in other mammalian cells but not bat cells.

In addition, Cell cultures are normally kept under sterile conditions using a cocktail of antibiotics in combination with aseptic techniques to minimize microbial growth, which would have resulted in a sample that is mostly sterile with minimal to no bacterial sequences.

One of the defining feature of the SARS-CoV-2 S is the optimization of ACE2 binding and folding at 37℃, a feature that it shares with the RaTG13 S.[26][27] However, the body temperature of a Horseshoe bat can reach up to 41℃[28], where substantial unfolding of the RaTG13 Spike happens according to Differential Scanning Fluorimetry on the Spike trimer[27]. This is incompatible with the high body temperature of a horseshoe bat (as the virus will be inactivated by the heat generated by bat flight), but is compatible with a cell culture as most cells in laboratories are cultured at 37℃.

Traces of lentivirus- and HERV-like fragments found in SRX7724752 likely indicate the usage of retroviral- and lentiviral- vectors on the sample, which are frequently used for the delivery of a TERT gene for the immortalization of cell lines in-vitro, a pre-requisite for the subsequent culture and attenuation for a vaccine strain.

Indeed, the combined features of the RaTG13 genome resemble that of a Live Attenuated Vaccine (LAV) almost suspiciously, to the point that there are actual proposals for the usage of this sequence as a candidate vaccine against SARS-CoV-2[25].

In addition, evidence of mutagen usage during the divergent evolution of RaTG13 and SARS-CoV-2 from a common ancestor can be found in the Spike protein CDS of RaTG13, manifesting as a very large excess of C:T transitions compared to the substitutional pattern of SARS-CoV-2 WIV04:ZC45, RaTG13:ZC45 or SARS-CoV Tor2:WIV1 on the aligned sections of their Spike protein CDS sequences.[29] As Mutagens like 5-fluorouracil is not present in wild bats, The RaTG13 Spike protein CDS would have to be grown in a cell culture to be influenced by 5-fluorouracil and show such a peculiar substitutional pattern.

CONCLUSION

The raw data of BtCoV/RaTG13 Contained multiple anomalies that signifies that the original sample could not have contained enough RNA template for the extraction of a complete viral genome as in MN996532.1

Furthermore, many of these anomalies points toward the fraudulent use of a mixed DNA library, rather than genuine mRNA, for the sequencing of SRX7724752, evident by the presence of widespread A-T ligation of unrelated dsDNA fragments that can only happen if the same library preparation process have been ran on dsDNA instead of ssRNA. which would constitute Academic fraud.

The Spike glycoprotein of RaTG13 does not resemble that of a wild virus but instead possessed multiple signatures of artificial attenuation in a cell culture when compared to the SARS-CoV-2 Spike and the Spike sequences of other related viruses, indicating that the sequence did not derive from what the Wuhan Institute Of Virology claimed to be.

Therefore, the sequencing of BtCoV/RaTG13 cannot be considered to be valid or honest as is, and any publications, including [2], and other publications that cites or use RaTG13 as critical pieces of evidence or proof, must be immediately invalidated and retracted.

Supplementary data

The Pymol session file (pse) used to generate fig.32 have been deposited as “broken ACE2.pse”.

REFERENCES

[1] https://www.thetimes.co.uk/article/seven-year-covid-trail-revealed-l5vxt7jqp

[2] Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579(7798):270-273. doi:10.1038/s41586-020-2012-7

[3] Rose C, Parker A, Jefferson B, Cartmell E. The Characterization of Feces and Urine: A Review of the Literature to Inform Advanced Treatment Technology. *Crit Rev Environ Sci Technol*. 2015;45(17):1827-1879. doi:10.1080/10643389.2014.1000761

[4] https://telegra.ph/RaTG13-07-06

[5] Xiao C, Li X, Liu S, Sang Y, Gao SJ, Gao F. HIV-1 did not contribute to the 2019-nCoV genome. *Emerg Microbes Infect*. 2020;9(1):378-381. Published 2020 Feb 14. doi:10.1080/22221751.2020.1727299

[6] https://www.planetnatural.com/product/mexican-bat-guano/

[7] Paskey, A.C., Frey, K.G., Schroth, G. *et al.* Enrichment post-library preparation enhances the sensitivity of high-throughput sequencing-based detection and characterization of viruses from complex samples. *BMC Genomics* **20,** 155 (2019). https://doi.org/10.1186/s12864-019-5543-2

[8] Edgar, R. C. *et al*. Petabase-scale sequence alignment catalyses viral discovery. *bioRxiv* 2020.08.07.241729 (2020) doi:10.1101/2020.08.07.241729

[9] Study of the Metatranscriptome of Eight Social and Solitary Wild Bee Species Reveals Novel Viruses and Bee Parasites

Karel Schoonvaere1,2*, Guy Smagghe3, Frédéric Francis2 and Dirk C. de Graaf1

1Laboratory of Molecular Entomology and Bee Pathology, Department of Biochemistry and Microbiology, Faculty of Sciences, Ghent University, Ghent, Belgium

2Functional and Evolutionary Entomology, Gembloux Agro-Bio Tech, University of Liege, Gembloux, Belgium

3Laboratory of Agrozoology, Department of Crop Protection, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium

[10] Zhang, S., Qiao, S., Yu, J. *et al.* Bat and pangolin coronavirus spike glycoprotein structures provide insights into SARS-CoV-2 evolution. *Nat Commun* **12,** 1607 (2021). https://doi.org/10.1038/s41467-021-21767-3

[11] Pei Li, Ruixuan Guo, Yan Liu, Yingtao Zhang, Jiaxin Hu, Xiuyuan Ou, Dan Mi, Ting Chen, Zhixia Mu, Yelin Han, Zihan Chen, Zhewei Cui, Leiliang Zhang, Xinquan Wang, Zhiqiang Wu, Jianwei Wang, Qi Jin, Zhaohui Qian,

The Rhinolophus affinis bat ACE2 and multiple animal orthologs are functional receptors for bat coronavirus RaTG13 and SARS-CoV-2,

Science Bulletin,

2021,

,

ISSN 2095-9273,

https://doi.org/10.1016/j.scib.2021.01.011.

[12] Andersen, K.G., Rambaut, A., Lipkin, W.I. *et al.* The proximal origin of SARS-CoV-2. *Nat Med* **26,** 450–452 (2020). https://doi.org/10.1038/s41591-020-0820-9

[13] Makowski, L.; Olson-Sidford, W.; W. Weisel, J. Biological and Clinical Consequences of Integrin Binding via a Rogue RGD Motif in the SARS CoV-2 Spike Protein. *Viruses* **2021**, *13*, 146. https://doi.org/10.3390/v13020146

[14] Tyler N. Starr, Allison J. Greaney, Sarah K. Hilton, Daniel Ellis, Katharine H.D. Crawford, Adam S. Dingens, Mary Jane Navarro, John E. Bowen, M. Alejandra Tortorici, Alexandra C. Walls, Neil P. King, David Veesler, Jesse D. Bloom,

Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding,

Cell,

Volume 182, Issue 5,

2020,

Pages 1295-1310.e20,

ISSN 0092-8674,

https://doi.org/10.1016/j.cell.2020.08.012.

[15] Nie, J., Li, Q., Zhang, L. *et al.* Functional comparison of SARS-CoV-2 with closely related pangolin and bat coronaviruses. *Cell Discov* **7,** 21 (2021). https://doi.org/10.1038/s41421-021-00256-3

[16] Conceicao C, Thakur N, Human S, Kelly JT, Logan L, Bialy D, Bhat S, Stevenson-Leggett P, Zagrajek AK, Hollinghurst P, Varga M, Tsirigoti C, Tully M, Chiu C, Moffat K, Silesian AP, Hammond JA, Maier HJ, Bickerton E, Shelton H, Dietrich I, Graham SC, Bailey D. The SARS-CoV-2 Spike protein has a broad tropism for mammalian ACE2 proteins. PLoS Biol. 2020 Dec 21;18(12):e3001016. doi: 10.1371/journal.pbio.3001016. PMID: 33347434; PMCID: PMC7751883.

[17] Clarke NE, Fisher MJ, Porter KE, Lambert DW, Turner AJ. Angiotensin converting enzyme (ACE) and ACE2 bind integrins and ACE2 regulates integrin signalling. PLoS One. 2012;7(4):e34747. doi: 10.1371/journal.pone.0034747. Epub 2012 Apr 16. PMID: 22523556; PMCID: PMC3327712.

[18] Brandon J. Beddingfield, Naoki Iwanaga, Prem P. Chapagain, Wenshu Zheng, Chad J. Roy, Tony Y. Hu, Jay K. Kolls, Gregory J. Bix,

The Integrin Binding Peptide, ATN-161, as a Novel Therapy for SARS-CoV-2 Infection,

JACC: Basic to Translational Science,

Volume 6, Issue 1,

2021,

Pages 1-8,

ISSN 2452-302X,

https://doi.org/10.1016/j.jacbts.2020.10.003.

[19] Factors Associated with Emerging and Re-emerging of SARS-CoV-2 Variants

Austin N. Spratt, Saathvik R. Kannan, Lucas T. Woods, Gary A. Weisman, Thomas P. Quinn, Christian L. Lorson, Anders Sönnerborg, Siddappa N. Byrareddy, Kamal Singh

bioRxiv 2021.03.24.436850; doi: https://doi.org/10.1101/2021.03.24.436850

[20]

Difference in Receptor Usage between Severe Acute Respiratory Syndrome (SARS) Coronavirus and SARS-Like Coronavirus of Bat Origin

Wuze Ren, Xiuxia Qu, Wendong Li, Zhenggang Han, Meng Yu, Peng Zhou, Shu-Yi Zhang, Lin-Fa Wang, Hongkui Deng, Zhengli Shi

Journal of Virology Jan 2008, 82 (4) 1899-1907; DOI: 10.1128/JVI.01085-07

[21] Tikam Chand Dakal,

SARS-CoV-2 attachment to host cells is possibly mediated via RGD-integrin interaction in a calcium-dependent manner and suggests pulmonary EDTA chelation therapy as a novel treatment for COVID 19,

Immunobiology,

Volume 226, Issue 1,

2021,

152021,

ISSN 0171-2985,

https://doi.org/10.1016/j.imbio.2020.152021.

[22] Characterisation of B.1.1.7 and Pangolin coronavirus spike provides insights on the evolutionary trajectory of SARS-CoV-2

Samuel J. Dicken, Matthew J. Murray, Lucy G. Thorne, Ann-Kathrin Reuschl, Calum Forrest, Maaroothen Ganeshalingham, Luke Muir, Mphatso D. Kalemera, Machaela Palor, Laura E. McCoy, Clare Jolly, Greg J. Towers, Matthew B. Reeves, Joe Grove

bioRxiv 2021.03.22.436468; doi: https://doi.org/10.1101/2021.03.22.436468

[23] Wang N, Luo C, Liu H, Yang X, Hu B, Zhang W, Li B, Zhu Y, Zhu G, Shen X, Peng C, Shi Z. Characterization of a New Member of Alphacoronavirus with Unique Genomic Features in Rhinolophus Bats. Viruses. 2019 Apr 24;11(4):379. doi: 10.3390/v11040379. PMID: 31022925; PMCID: PMC6521148.

[24] Growing old, yet staying young: The role of telomeres in bats' exceptional longevity

By Nicole M. Foley, Graham M. Hughes, Zixia Huang, Michael Clarke, David Jebb, Conor V. Whelan, Eric J. Petit, Frédéric Touzalin, Olivier Farcy, Gareth Jones, Roger D. Ransome, Joanna Kacprzyk, Mary J. O'Connell, Gerald Kerth, Hugo Rebelo, Luísa Rodrigues, Sébastien J. Puechmaille, Emma C. Teeling

Science Advances07 Feb 2018 : eaao0926

[25] Chougle, Afreen and Chougle, Humera and Sayyed, Rajat, Prospect of Using RaTG13 Sarbecovirus As a Candidate Vaccine for COVID-19 (September 2, 2020). Available at SSRN: https://ssrn.com/abstract=3685067 or http://dx.doi.org/10.2139/ssrn.3685067

[26] Zhonghua Zhou, Ziyi Yang, Junxian Ou, Hong Zhang, Qiwei Zhang, Ming Dong, Gong Zhang, Temperature dependence of the SARS-CoV-2 affinity to human ACE2 determines COVID-19 progression and clinical outcome,
Computational and Structural Biotechnology Journal,
Volume 19,
2021,
Pages 161-167,
ISSN 2001-0370,
https://doi.org/10.1016/j.csbj.2020.12.005.

[27] Wrobel, A.G., Benton, D.J., Xu, P. *et al.* SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. *Nat Struct Mol Biol* **27,** 763–767 (2020). https://doi.org/10.1038/s41594-020-0468-7

[28] O'Shea TJ, Cryan PM, Cunningham AA, Fooks AR, Hayman DT, Luis AD, Peel AJ, Plowright RK, Wood JL. Bat flight and zoonotic viruses. Emerg Infect Dis. 2014 May;20(5):741-5. doi: 10.3201/eid2005.130539. PMID: 24750692; PMCID: PMC4012789.

[29] Lv Longxian, Li Gaolei, Chen Jinhui, Liang Xinle, Li Yudong Comparative Genomic Analyses Reveal a Specific Mutation Pattern Between Human Coronavirus SARS-CoV-2 and Bat-CoV RaTG13 Frontiers in Microbiology vol.11 2020
https://www.frontiersin.org/article/10.3389/fmicb.2020.584717
https://doi.org/10.3389/fmicb.2020.584717
ISSN: 1664-302X

[30] Liu K, Pan X, Li L, Yu F, Zheng A, Du P, Han P, Meng Y, Zhang Y, Wu L, Chen Q, Song C, Jia Y, Niu S, Lu D, Qiao C, Chen Z, Ma D, Ma X, Tan S, Zhao X, Qi J, Gao GF, Wang Q. Binding and molecular basis of the bat coronavirus RaTG13 virus to ACE2 in humans and other species. Cell. 2021 May 24:S0092-8674(21)00661-9. doi: 10.1016/j.cell.2021.05.031. Epub ahead of print. PMID: 34139177; PMCID: PMC8142884.

[31] Spike mutation T403R allows bat coronavirus RaTG13 to use human ACE2
Fabian Zech, Daniel Schniertshauer, Christoph Jung, Alexandra Herrmann, Qinya Xie, Rayhane Nchioua, Caterina Prelli Bozzo, Meta Volcic, Lennart Koepke, Jana Krüger, Sandra Heller, Alexander Kleger, Timo Jacob, Karl-Klaus Conzelmann, Armin Ensser, Konstantin M.J. Sparrer, Frank Kirchhoff
bioRxiv 2021.05.31.446386; doi: https://doi.org/10.1101/2021.05.31.446386

[32] Fossum SL, Mutolo MJ, Yang R, Dang H, O'Neal WK, Knowles MR, Leir SH, Harris A. Ets homologous factor regulates pathways controlling response to injury in airway epithelial cells.

Nucleic Acids Res. 2014 Dec 16;42(22):13588-98. doi: 10.1093/nar/gku1146. Epub 2014 Nov 20. PMID: 25414352; PMCID: PMC4267623.
[33] Khalili P, Arakelian A, Chen G, Plunkett ML, Beck I, Parry GC, Doñate F, Shaw DE, Mazar AP, Rabbani SA. A non-RGD-based integrin binding peptide (ATN-161) blocks breast cancer growth and metastasis in vivo. Mol Cancer Ther. 2006 Sep;5(9):2271-80. doi: 10.1158/1535-7163.MCT-06-0100. PMID: 16985061.